Fox Module 4 Bivariate displays practice problems on ozone co-plots

(The attached PDF file has better formatting.)

CO-PLOTS FOR OZONE

The data relate to ozone levels in the atmosphere. The variables in the data sample are

- *rad:* solar radiation
- *temp:* daily temperature
- *wind:* wind speed
- *ozone:* ozone level in atmosphere

The ranges of the four variables are

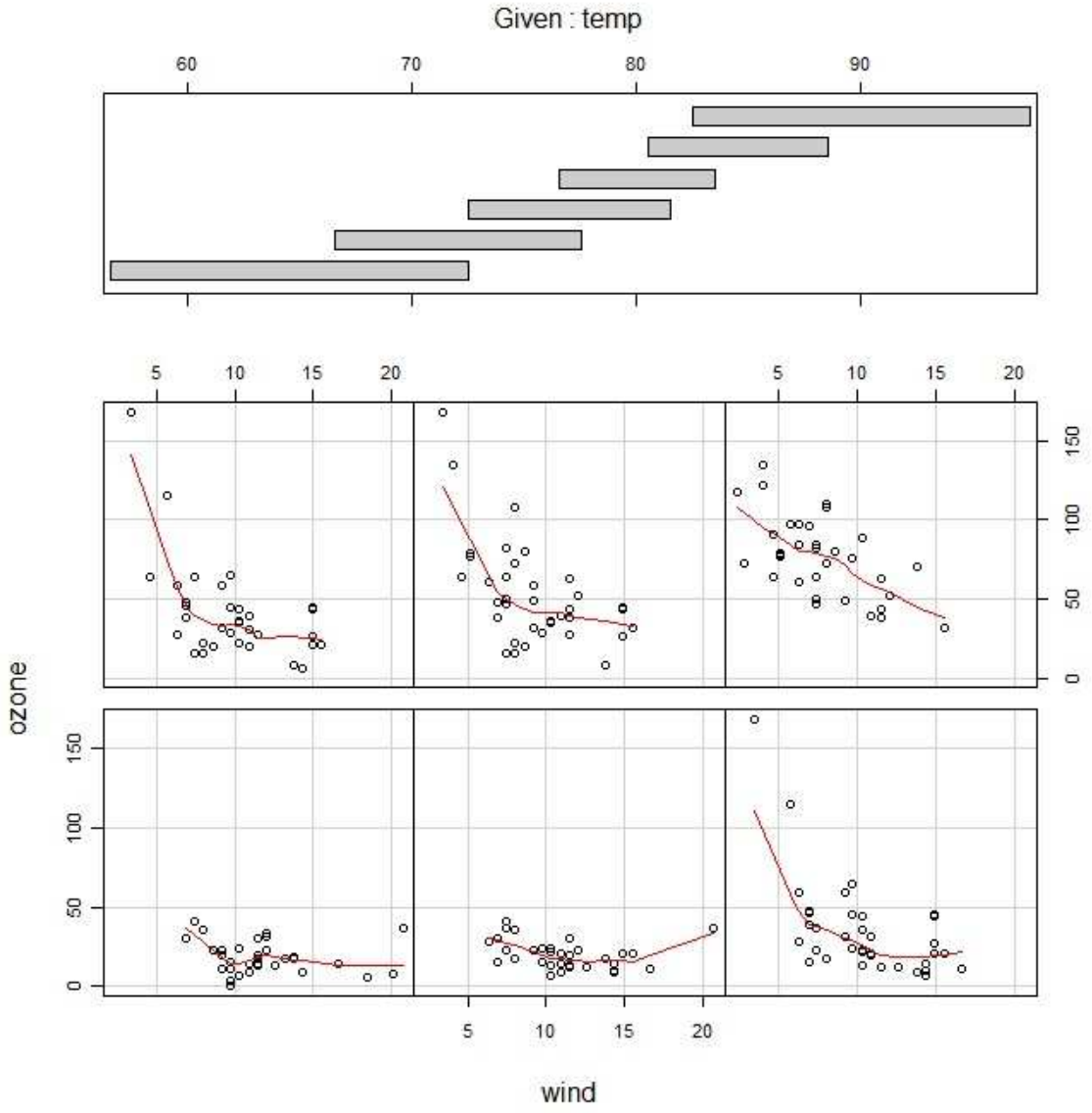
- Daily temperature ranges from 57 to 97 degrees.
- Wind speed ranges from 2.3 to 20.7 miles per hour.
- Solar radiation ranges from 7 to about 334 units.
- Ozone levels range from 1 to 168 units (particles in a given volume of air).

The panels below are ordered from lower-left to upper-right by the values of the conditioning variable in the upper panel (temp) from left to right. The lower-left plot is for the lowest temperatures (56-72 degrees) and the upper right plot is for the highest temperatures (82-96 degrees). Solar radiation is not used in these plots.

The temperatures overlap. Each panel shares 50% of its points with the next higher panel and 50/% of its points with the next lower panel. For example, the first panel is 57°-73°and the second panel is 66°-77°. The overlap avoids distortions from random fluctuations near the end of a panel. For example, suppose the panels were for temperatures of 57°-65° vs 66°-77° and ozone levels were unusually high on a day with 65° temperature. The graph might show a spike in the first panel but not in the second panel.

Jacob: How can we tell the overlap percentage from the plot?

Rachel: One can not see the percentage from the plot, since one needs to know the distribution of points in each temperature group. The 50% is the R default, which Fox uses for his plots. One can change the default to another overlap percentage (if desired). The final exam problems do not ask anything not clear in the plot.


Given : temp
60
70
80
90
5
10
15
20
0
50
100
150
ozone
wind

** Exercise 4.1: Variables

A. What is the explanatory variable in these plots?
B. What is the response variable in these plots?
C. What is the conditioning variable in these plots?

Part A: The explanatory variable is wind speed.

Part B: The response variable is ozone level.

Part C: The conditioning variable is temperature.

** Exercise 4.2: Correlations

A. At what temperatures is higher wind speed positively correlated with ozone levels?
B. At what temperatures is higher wind speed negatively correlated with ozone levels?

Part A: At no temperatures are ozone levels positively correlated with wind speed. The second plot in the lower row shows a slight uptick at the highest wind speed caused by a single outlier. It is a random fluctuation, not the expected correlation of ozone levels and wind speed.

Part B: At temperatures of about 85° or higher, ozone levels are positively correlated with wind speed. At temperatures of about 70° to 90°, the negative correlation is strong at low wind speeds and weaker at high wind speeds.

Fox Module 4 Bivariate displays practice problems on ozone scatter-plot matrices

(The attached PDF file has better formatting.)

** Exercise 4.1: Scatterplot matrix

The data are from a study of ozone levels in the atmosphere. The four variables in the scatterplot matrix are

- *rad:* solar radiation
- *temp:* daily temperature
- *wind:* wind speed
- *ozone:* ozone level in atmosphere

A. What are the ranges of the four variables?
B. What is the thin red line in each plot?
C. Two plots show the relation of temperature and solar radian. One plot shows a humped curve and the other plot shows a positively sloped curve. Explain what each curve implies.
D. Which variables are negatively correlated over their entire ranges?

Part A: The exact ranges are hard to read from the axes of the scatterplot matrix. They are

- Daily temperature ranges from 57 to 97 degrees.
- Wind speed ranges from 2.3 to 20.7 miles per hour.
- Solar radiation ranges from 7 to about 334 units.
- Ozone levels range from 1 to 168 units (particles in a given volume of air).

Jacob: Do we read the ranges along the horizontal axis or the vertical axis?

Rachel: For the panel with the name of the variable, the scales are the same along the two axes.

Take heed: Final exam problems about the ranges of the variables test if you look at the proper scales. They do not test the exact values.

Part B: The thin red line is a *loess* curve.

Jacob: The *loess* curve uses linear regression, but it is curved, not straight. Why is this?

Rachel: The linear regression differs at each point. For wind speed of 5, it may usespoints of wind speed from 0 to 10; at wind speed 10, it may use points of wind speed from 5 to 15.

Jacob: Random fluctuations might distort the loess curve. Suppose a wind speed = 11, ozone levels were high (by random fluctuation). For a regression from 1 to 11, β is high; for a regression from 0 to 10, β is lower.

Rachel: The loess curve uses weighted regressions. For a regression from 1 to 11, the point 11 receives little weight.

Part C: The humped curve in the first column and second row has solar radian as the explanatory variable and temperature as the response variable. For low solar radian, as solar radian increases, temperature increases; for high solar radian, as solar radian increases, temperature decreases.

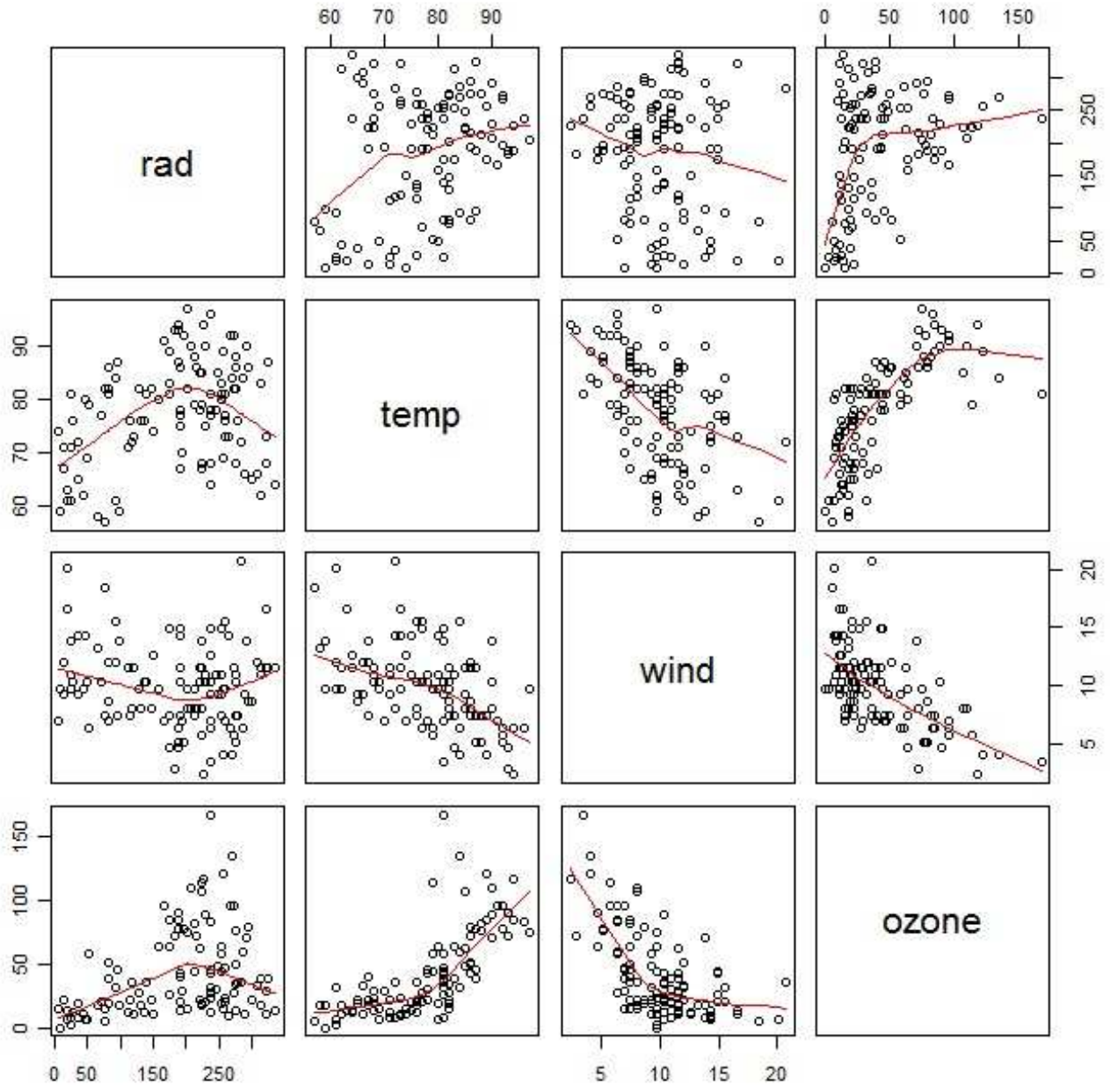
The positively sloped curve in the first row and second column has temperature as the explanatory variable and solar radian as the response variable. As temperature increases, solar radian increases.

Jacob: Does temperature cause the change in solar radiation, or does solar radiation cause the change in temperature?

Rachel: The graph does not imply causation. Neither directly affects the other.

Part D: Temperature and wind speed are negatively correlated. In California, when wind speeds are higher, the days are cooler.

Wind speed and ozone levels are also negatively correlated, at least at lower wind speeds. Higher wind disperses the ozone.


rad
temp
wind
ozone
60 70 80 90
0 50 100 150
0 50 150 250
5 10 15 20
60 70 80 90
0 50 100 150
0 50 150 250
5 10 15 20

Fox Module 5: Multivariate displays

Practice problems: Conditioning plots

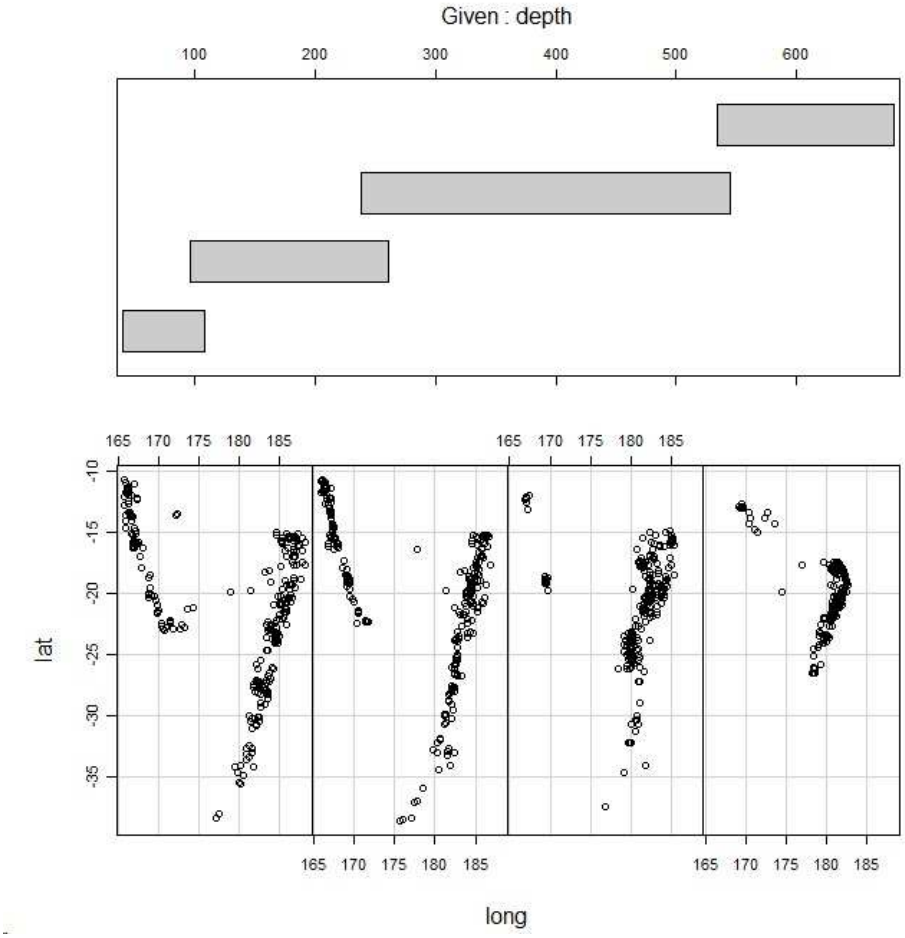
For Section 3.3.4, “Conditioning plots,” on pages 46-47.

Conditioning plots are scatterplots conditioned on values of one or more other variables. The conditioning plot below is from a book by William Cleveland. It shows the number of Tonga Trench earthquakes by latitude and longitude conditioned on the depth of the quake.

The depth values overlap 10%. The depths are chosen so that each group has about the same number of earthquakes.

- Earthquakes are common at depths of 0 to 100, so this depth has the smallest range.
- Earthquakes are rare at depths of 250 to 550, so this depth has the smallest range.

Look at the latitudes and longitudes carefully. You can see how the depths change along the Tonga Trench. Depths of 250 to 550 are not common, so a greater range is needed. The size of the range reflects both the area at that depth and earthquake frequency.



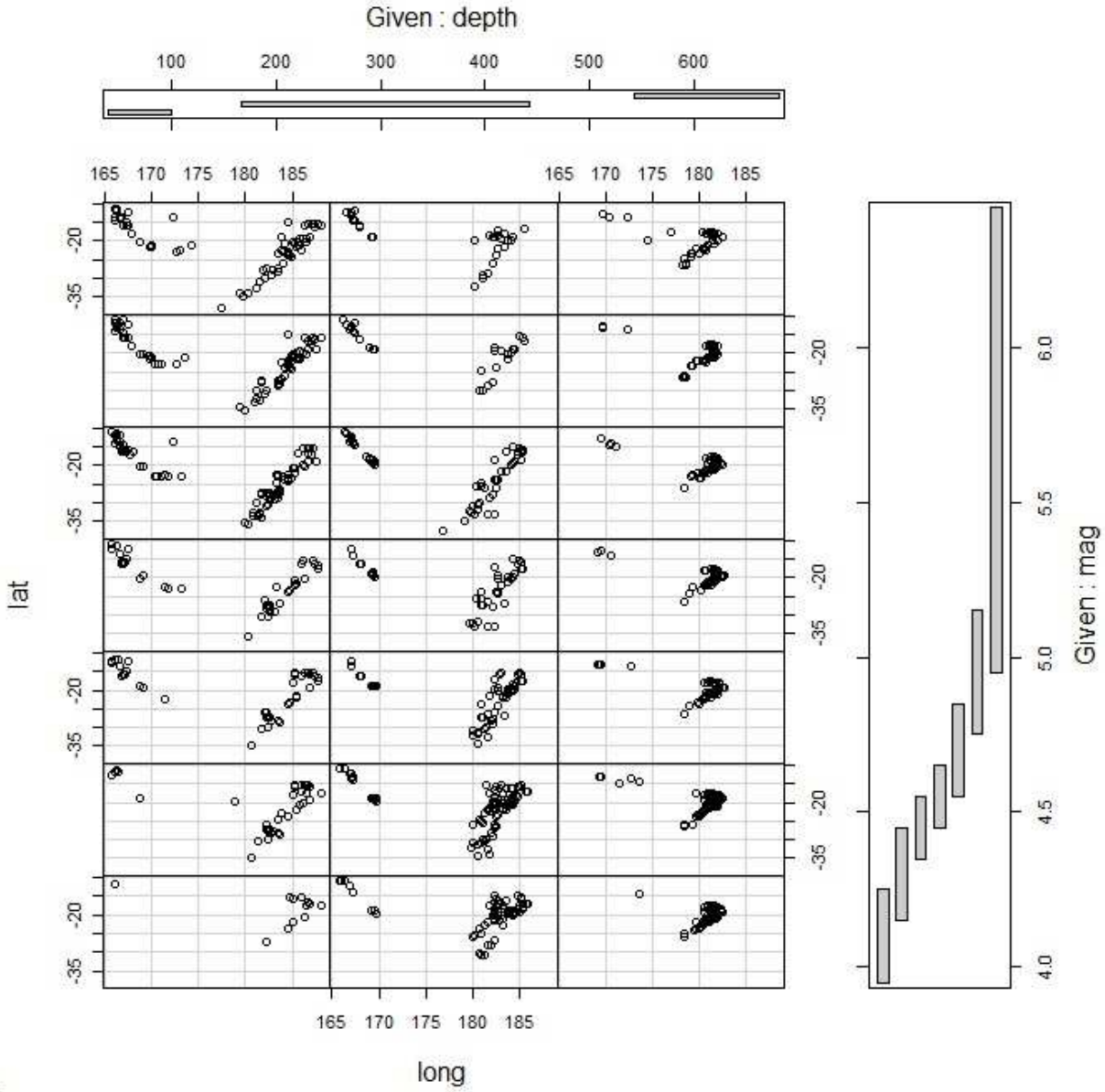
Question 1.1: Conditioning plots

The graphic below shows Tonga Trench earthquakes by longitude (horizontal axis) and latitude (vertical axis), conditioned on depth (below sea level) and magnitude.

Which of the following is true?

A. Low magnitude (4.0) earthquakes are more likely at depths of 0-100; high magnitude (5.0 to 6.5) earthquakes are more likely at depths of 600.
B. High magnitude (5.0 to 6.5) earthquakes are more likely at depths of 0-100; high magnitude (4.0) earthquakes are more likely at depths of 600.
C. All earthquakes are more likely at depths of 0-100.
D. All earthquakes are more likely at depths of 600.
E. One can not relate depth to magnitude from this conditioning plot.

Answer 1.1: B


Given : depth
100
200
300
400
500
600
165
170
175
180
185
-20
-35
lat
long
Given : mag
4.0
4.5
5.0
5.5
6.0

Module 8: Simple linear regression practice problems

(The attached PDF file has better formatting.)

LINEAR REGRESSION: PRACTICE EXAM PROBLEMS

This posting illustrates linear regression exam problems covering the basic formulas. On the final exam, expect a scenario with five pairs of points similar to the exercise below. The problem derives the ordinary least squares estimators, their standard errors, t-values, levels of significance, and F-statistic. Some statistical items are taught in later modules; these practice problems covers many items in basic regression analysis.

An actuary fits a two-variable regression model ($Y_i = \alpha + \beta \times X_i + \epsilon_i$) to the relation between the incurred loss ratio (x) and the retrospective ratio (y), using the data below:

Policy Year	*(x)*	*(y)*	$(x - \bar{x})$	$(x - \bar{x})^2$	$(y - \bar{y})$	$(y - \bar{y})^2$	$(x-\bar{x})(y-\bar{y})$
20X1	61.00%	15.00%	0.00%	0.00%	0.32%	0.0010%	0.0000%
20X2	62.00%	13.20%	1.00%	0.01%	-1.48%	0.0219%	-0.0148%
20X3	63.00%	14.00%	2.00%	0.04%	-0.68%	0.0046%	-0.0136%
20X4	60.00%	15.20%	-1.00%	0.01%	0.52%	0.0027%	-0.0052%
20X5	59.00%	16.00%	-2.00%	0.04%	1.32%	0.0174%	-0.0264%
Average	61.00%	14.68%	0.00%	0.02%	0.00%	0.009536%	-0.01200%

The column captions use lower case *x* and *y* for the variables; the deviations are shown explicitly as $(x - \bar{x})$ and $(y - \bar{y})$. Some statistician use upper case letter for the variables and lower case letters for the deviations.

Take heed: The notation in the John Fox regression analysis text differs sightly from the notation in some of the discussion forum postings.

- John Fox uses the symbols *A* and *B* as the least squares estimators for α and β.
- He uses *RSS* for the residual sum of squares; other authors use *ESS*, the error sum of squares.
- He uses *RegSS* for the regression sum of squares; other authors use *RSS*.

The final exam problems use Fox's notation.

Question 8.1: Ordinary Least Squares Estimator of β

What is the value of *B*, the ordinary least squares estimator of β?

A. –0.600
B. –0.120
C. –0.020
D. –0.019
E. –0.012

Answer 8.1: A

The table gives the sum of the cross-product terms and of the squared deviations of X.

$B = \sum(x_i - \bar{x})(y_i - \bar{y}) / \sum(x_i - \bar{x})^2 = -0.012 / 0.020 = -0.600$

Note: The last row of the table shows averages. The ratio of the averages is the ratio of the sums.

Question 8.2: Ordinary Least Squares Estimator of α

What is the value of A, the ordinary least squares estimator of α?

A. –0.6100
B. –0.1468
C. +0.1468
D. +0.5128
E. +0.6100

Answer 8.2: D

Use the relation: $A = \overline{y} - B \times \overline{x} = 14.68\% - (-0.60) \times 61.00\% = 0.5128$

Question 8.3: Total Sum of Squares (TSS)

What is the total sum of squares (TSS)?

A. 0.0117%
B. 0.0360%
C. 0.0477%
D. 0.0833%
E. 0.1310%

Answer 8.3: C

The total sum of squares can be found two ways.

(1) We subtract the mean of Y from each observed value and square the deviations:

$$\sum(y_i - \overline{y})^2 = 0.32\%^2 + (-1.48\%)^2 + (-0.68\%)^2 + 0.52\%^2 + 1.32\%^2 = 0.04768\%$$

(2) We square the observed values of Y and subtract N times the square of the mean:

$$\sum Y^2 - N\overline{Y}^2 \quad \sum Y^2 - \left(\sum Y\right)^2 / N$$

= 15% + 13.2% + 14% + 15.2% + 16% – 14.68%^{2} / 5 = 0.04768%

The table in the exam problem gives the TSS as 5 × 0.009536% = 0.04768%

Question 8.4: Regression Sum of Squares (RegSS)

What is the regression sum of squares (RegSS)?

A. 0.0117%
B. 0.0360%
C. 0.0477%
D. 0.0833%
E. 0.1310%

Answer 8.4: B

Find the fitted Y value at each observation as $A + B \times X$. Subtract the mean of Y and square the result. The sum of these is the regression sum of squares.

Policy Year	*(x)*	*(y)*	$\hat{y}$	$(\hat{y} - \bar{y})$	$(\hat{y} - \bar{y})^2$
20X1	61.00%	15.00%	14.68%	0.00%	0.0000%
20X2	62.00%	13.20%	14.08%	-0.60%	0.0036%
20X3	63.00%	14.00%	13.48%	-1.20%	0.0144%
20X4	60.00%	15.20%	15.28%	0.60%	0.0036%
20X5	59.00%	16.00%	15.88%	1.20%	0.0144%
Average	61.00%	14.68%	14.68%	0.00%	0.007200%

$5 \times 0.0072\% = 0.0360\%$

A quick formula: regression sum of squares (RegSS) = $B^2 \times \sum(x_i - \bar{x})^2$

$\sum(x_i - \bar{x})^2 = 1\%^2 + 2\%^2 + (-1\%)^2 + (-2\%)^2 = 0.10\%$

RegSS = $B^2 \times 0.10\% = 0.6^2 \times 0.10\% = 0.0360\%$

Question 8.5: Error Sum of Squares (ESS) or Residual Sum of Squares (RSS)

What is the error sum of squares (ESS) or residual sum of squares (RSS)

A. 0.0117%
B. 0.0360%
C. 0.0477%
D. 0.0833%
E. 0.1310%

Answer 8.5: A

We compute the residual sum of squares two ways.

(1) The ESS (RSS) is the TSS minus the RegSS.

0.04768% – 0.0360% = 0.01168%

(2) We determine residuals as the observed Y minus the fitted Y. The sum of the squared residuals is the residual sum of squares (RSS).

Policy Year	*(x)*	*(y)*	$\hat{y}$	$(\hat{y} - \bar{y})$	$(\hat{y} - \bar{y})^2$
20X1	61.00%	15.00%	14.68%	-0.32%	0.0010%
20X2	62.00%	13.20%	14.08%	0.88%	0.0077%
20X3	63.00%	14.00%	13.48%	-0.52%	0.0027%
20X4	60.00%	15.20%	15.28%	0.08%	0.0001%
20X5	59.00%	16.00%	15.88%	-0.12%	0.0001%
Average	61.00%	14.68%	14.68%	0.00%	0.002336%

5 × 0.002336% = 0.011680%

Question 8.6: Standard Error

What is s^2, the estimated variance of the regression?

A. 0.0036%
B. 0.0039%
C. 0.0360%
D. 0.0389%
E. 0.0117%

Answer 8.6: B

The estimated variance of the regression is the residual sum of squares divided by the degrees of freedom (the number of observations minus the number of explanatory variables). The explanatory variables are the independent variables plus the constant term. For a simple linear regression, this is N–2: 0.01168% / 3 = 0.003893%. This is an unbiased estimate of σ^2.

Question 8.7: Variance of Ordinary Least Squares Estimator of β

What is the *variance* of the ordinary least squares estimator of β? (This is the variance, not the standard error.)

A. 0.36%
B. 0.39%
C. 3.60%
D. 3.89%
E. 1.17%

Answer 8.7: D

The variance of *B* is the σ^2 (or its unbiased estimate) divided by $\sum(x_i - \overline{x})^2$:

0.00389% / 0.10% = 3.890%

Question 8.8: *t* Statistic

What is the *t* statistic for testing the null hypothesis that $\beta = 0$?

A. –3
B. –2
C. –1
D. +1
E. +2

Answer 8.8: A

The *t* statistic is (the difference between *B* and the null hypothesis) divided by the standard deviation of *B*, which is the square root of the variance of *B*:

$-0.6 / 3.890\%^{1/2} = -3.042$

Question 8.9: *p*-value

The *p*-value for β for this regression equation is 0.0558. Which of the following is true, assuming the classical regression assumptions hold?

A. The true β is within ±5.58% (multiplicative) of the ordinary least squares estimator.
B. The true β is within ±0.0558 (additive) of the ordinary least squares estimator.
C. The probability is 95% that the true β is within ±0.0558 of the ordinary least squares estimator.
D. If the true value of β is zero, the probability that the absolute value of the ordinary least squares estimator of β is at least as great as in this regression equation is 5.58%.
E. If the true value of β is zero, the probability is 95% that the absolute value of the ordinary least squares estimator of β is no more than 0.0558.

Answer 8.9: D

To test hypotheses, we consider the probability that we would observe an ordinary least squares estimator as far from the null hypothesis (or farther) because of sampling error. The *p*-value gives this probability, as stated in Statement D.

Question 8.10: *F* Statistic

What is the *F* statistic for testing the null hypothesis that $\beta = 0$?

A. –4
B. –1
C. +1
D. +4
E. +9

Answer 8.10: E

We compute the F statistic two ways.

(1) For a two-variable regression model, the *F* statistic is the square of the *t* statistic:

$$-0.6^2 / 3.893\% = 9.247$$

(2) The *F* statistic is the ratio of the regression sum of squares divided by its degrees of freedom to the error sum of squares divided by its degrees of freedom:

$$[0.0360\% / 1] / [0.01168\% / 3] = 9.247$$

Question 8.11: R^2

What is the value of R^2, the coefficient of determination?

A. 55%
B. 65%
C. 75%
D. 85%
E. 95%

Answer 8.11: C

R^2 = 0.036% / 0.04768% = 75.50%

Fox Module 10 R^2 practice problems

(The attached PDF file has better formatting.)

** Exercise 10.1: R^2

A simple linear regression with an intercept and one explanatory variable fit to 18 observations has a total sum of squares (TSS) = 256 and s^2 (the ordinary least squares estimator for σ^2) = 4.

A. How many degrees of freedom does the regression equation have?
B. What is RSS, the residual sum of squares?
C. What is RegSS, the regression sum of squares?
D. What is the R^2 of the regression equation?
E. What is the adjusted (corrected) R^2 of the regression equation?
F. What is the correlation of the explanatory variable and the response variable?
G. What is the F-value for the omnibus F-test?
H. What is the t-value for the explanatory variable?

Part A: The regression equation has N – k – 1 = 18 – 1 – 1 = 16 degrees of freedom.

Take heed: In this equation, *k* is the number of explanatory variables not including the intercept α.

Part B: The estimate of the variance of the error term (s^2) is the *residual (error)* sum of squares divided by the number of degrees of freedom, or N – *k*: s^2 = RSS / df, so the residual sum of squares (RSS) = s^2 × degrees of freedom = 4 × 16 = 64.

Part C: The regression sum of squares (RegSS) = TSS – RSS.

- The *total* sum of squares TSS is the sum of the squared residuals, given in the problem as 256.
- RSS = s^2 × (N – 2) = 4 × 16 = 64.
- RegSS = 256 – 64 = 192.

Part D: The R^2 = RegSS / RSS = 1 – RSS/TSS = 192 / 256 = 75%.

Part E: Adjusted R^2 = 1 – (1 – R^2) × (N – 1) / (N – k) = 1 – (1 – 75%) × 17 / 16 = 73.44%

Part F: The correlation $\rho(x,y)$ = r = $\sqrt{R^2}$ = $\sqrt{75\%}$ = 0.866.

*Part G:*Fox, Chapter 6, page 109: for the omnibus F-test in a simple linear regression, R_0^2 = 0 and *k* = 1, so

F = (N – 2) × R^2 / (1 – R^2) = (18 – 2) × 0.75 / (1 – 0.75) = 48.000

Part H: The *t*-value for simple linear regression is the square root of the F value: $\sqrt{48}$ = 6.928

[This practice problem is an essay question, reviewing the meaning of the significance tests, goodness-of-fit tests, and measures of predictive power. It relates the statistical tests to the form of the regression line, emphasizing the intuition. Final exam problems test specific items in a multiple choice format.]

** Exercise 10.2: Measures of significance

The R^2, the adjusted (corrected) R^2, the s^2 (the ordinary least squares estimator for σ^2), the *t*-value, and the *F*-value measure the significance, goodness-of-fit, or predictive power of the regression.

A. What does the R^2 measure?
B. What does the adjusted (corrected) R^2 measure?
C. When is it important to use the adjusted (corrected) R^2 instead of the simple R^2?
D. If the $R^2 \approx 0$, what can one say about the regression?
E. If the $R^2 \approx 1$, what can one say about the regression?
F. What does the s^2 measure?
G. Given R^2, what is the F-value for the omnibus F-test?
H. What does the F-value measure?
I. If the F-value ≈ 0, what can one say about the regression?

Part A: R^2 measures the percentage of the total sum of squares explained by the regression, or RegSS / TSS.

Jacob: Why does the textbook show the R^2 as 1 – RSS / TSS? This is equivalent, since RSS + RegSS = TSS.

Rachel: To adjust for degrees of freedom (for the corrected R^2), we adjust RSS and TSS. The format R^2 = 1 – RSS / TSS makes it easier to understand the adjustment for degrees of freedom.

Jacob: Does the R^2 measure if the regression analysis is significant? The textbook gives significance levels for t-values and F-values (and associated confidence intervals for the regression coefficients), but it does not give significant levels for R^2.

Rachel: R^2 combines two items: whether the explanatory variables have predictive power and whether the regression coefficients are significantly different from zero (or from another null hypothesis). This exercise reviews the concepts and explains what R^2 implies vs what s^2 and the F-value imply.

Part B: R^2 does not adjust for degrees of freedom. If the regression has N data points and uses N explanatory variables (or N-1 independent variables + 1 intercept), all points are fit exactly, and the R^2 = 100%. This is true even if the explanatory variables have no predictive power: that is, each explanatory variable is independent of the response variable.

The same problem exists even if the number of explanatory variables is less than the number of data points. Even if the explanatory variables are independent of the response variable and have no predictive power, the R^2 is always more than zero.

The adjusted (corrected) R^2 adjusts for degrees of freedom. The degree of freedom apply to RSS and TSS, not to RegSS. With N data points and *k* independent variables (= *k+1* explanatory variables including the intercept), the TSS has N-1 degrees of freedom and the RSS has N-k-1 degrees of freedom.

Fox explains: R^2 is 1 – RSS / RSS = the complement of (the residual sum of squares / total sum of squares). The adjusted R^2 is the complement of (the residual variance / the total variance).

The adjusted (corrected) R^2 = 1 – (RSS / N-k-1) / (TSS / N-1).

The R^2 is a ratio of sums of squares and the adjusted (corrected) R^2 is a ratio of variances.

Part C: For most regression analyses, the R^2 is fine. It says what percentage of the variation in the sample values is explained by the regression. This percentage is not used for tests of significance, so a slight over-statement is not a problem.

Jacob: Is the R^2 over-stated? The textbook does not say that is over-stated.

Rachel: The R^2 says what percentage of the variation in the sample values is explained by the regression. It is the correct percentage, not over- or under-stated. Some of the explanation is spurious, caused by random fluctuations in small data samples. The adjusted R^2 says: What would the R^2 be if we had an infinite number of data points?

Jacob: This adjustment seems proper; why do we still use the simple R^2?

Rachel: We have a simple data set; we don't know what the R^2 would be if we had an infinite number of data points. We estimate the expected correction. This estimate is unbiased, but it is sometimes too high and sometimes too low.

To compare regression equations with different degrees of freedom, one must use the adjusted R^2. For example, suppose one regresses a response variable Y on several explanatory variables. One might say that the best regression equation is the one which explains the largest percentage of the variation in the response variable. R^2 is not a valid measure, since adding an explanatory variable always increases the R^2, even if the explanatory variable is unrelated to the response variable. Instead, we choose the regression equation with the highest adjusted R^2.

Part D: If R^2 is close to zero, the explanatory variables explain almost none of the variance in the response variable. For a simple linear regression with one explanatory variable, the correlation of X and Y is close to zero.

Jacob: Suppose we draw a scatterplot of Y against X. If R^2 is close to zero, is the scatterplot a cloud of points with no pattern?

Rachel: The R^2 reflects two things: the variance of the error term and the slope of the regression line. The variance of the error term compared to the dispersion of the response variable determines whether the scatterplot is a cloud of points with no clear pattern or a set of points lying next to the regression line. The slope of the regression line (the β coefficient) determines whether the explanatory variable much affects the response variable.

The units of measurement are important. Suppose we regress personal auto claim frequency on the distance the car is driven.

- If the slope coefficient is β when the distance is in miles (or kilometers), the slope coefficient is $\beta \times 1{,}000$ when the distance is thousands of miles (kilometers).
- If the slope coefficient is β when the claim frequency is in claims per car, the slope coefficient is $\beta / 100$ when the claim frequency is claims per hundred cars.

Illustration: Suppose the regression line is $Y = 1 + 0 \times X + \epsilon$. N (number of points) = 1,000, the explanatory variables are the integers from 1 to 1,000, and $\sigma^2_\epsilon = 1$. The scatterplot is a horizontal line Y = 1 with slight random fluctuations above and below the line. The scatterplot shows a clear pattern; it is not a cloud of points. But R^2 is close to zero, since the values of X have no effect on the values of Y.

Now suppose the true regression line is $Y = 1 + 1 \times X + \epsilon$, with N (number of points) = 1,000, the explanatory variables are the integers from 1 to 1,000, and $\sigma^2_\epsilon = 1$ million. The scatterplot is a 45° diagonal line Y = X with much random fluctuations above and below the line. The scatterplot does not show a clear pattern; it appears as a cloud of points, and only by looking carefully does one see the pattern. But R^2 is not close to zero, since the values of X have a strong effect on the values of Y. The exact value of R^2 depends on the error terms.

Some statisticians do not much use R^2, since it is a mix of two values: the slope of the regression line and the ratio of σ_ε to the dispersion of the Y values. We do not use R^2 for goodness-of-fit tests or tests of significance, since it mixes two items. We use the *t*-value (or the *F*-value) for the significance of the explanatory variables.

Part E: If R^2 is close to 1, the correlation of the explanatory variable and the response variable (X and Y) is close to 1 or –1. Almost all the variation in the response variable is explained by the explanatory variables.

An R^2 is close to 1 implies that the ratio of σ_ε to the dispersion of the Y values (the variance of Y) is low. Three things affect the R^2.

- RSS and σ^2_ε are low.
- β is not low.
- TSS (the variance of Y) is high.

Part F: s^2 is the ordinary least squares estimator of σ^2_ε. Most importantly, s^2 is an unbiased estimator of σ^2_ε.

Jacob: Does this imply that *s* is an unbiased estimator of σ_ε?

Rachel: If s^2 is an unbiased estimator of σ^2_ε, *s* is not an unbiased estimator of σ_ε. To grasp the rationale for this, suppose σ^2_ε is 4 and s^2 is 2, 3, 4, 5, or 6, with 20% probability of each.

- σ_ε is $\sqrt{4} = 2$.
- s is $\sqrt{2}$, $\sqrt{3}$, $\sqrt{4}$, $\sqrt{5}$, or $\sqrt{6}$, with a 20% probability of each.
- The mean of s is $(\sqrt{2} + \sqrt{3} + \sqrt{4} + \sqrt{5} + \sqrt{6}) / 5 = 1.966$.

s is a reasonable estimator of σ_ε, but it is not unbiased.

Part G: Use the relation $F = [(N – k – 1) / q] \times R^2 / (1 – R^2)$, where *k* is the number of explanatory variables (not including the intercept) and q is the number of variables in the group being tested.

Jacob: How is this relation derived?

Rachel: Use the expression for the F-value in terms of RSS and divide numerator and denominator by TSS.

Jacob: Fox has a *q* in his formula (page 109) and an R_0. What is the difference between *k* and *q*, and what is R_0?

Rachel: Fox shows the general form of the F-value. For the *omnibus* F-test, the null hypothesis is that all ß's are zero, so $k = q$ and R_0^2 (the R^2 for the null hypothesis) = 0.

Jacob: Can you explain the intuition for that last statement?

Rachel: If all ß's are zero, RSS = TSS, and RegSS = 0.

Part H: The *F*-value measures if a group of explanatory variables in combination is significant. The omnibus *F*-test measures if all the explanatory variables in combination are significant.

Jacob: Is that the same as *at least one explanatory variable is significant*? After all, if the explanatory variables in combination are significant, at least one of them must be significant.

Rachel: No, that is not correct. A clear example is a regression analysis on a group of correlated explanatory variables. Suppose an actuary regresses the loss cost trend for workers' compensation on three inflation indices: monetary inflation (the change in the CPI), wage inflation, and medical inflation. All three inflation indices are highly correlated. If any one were used in the regression equation alone, it would significantly affect the loss cost trend. If all three are used, we may not be able to discern which affects the loss cost trend, and none might be significant.

Jacob: If the regression equation has only one explanatory variable, are the *t*-value and the *F*-value the same?

Rachel: They have the same p-values, and they are equivalent significance tests, but they have different units. The F-value is the square of the t-value.

Part I: If the F-value is close to zero, the slope coefficient is not significantly different from zero. This means one of three things:

1. The slope coefficient is close to zero. The slope coefficient β depends on the units of measurement, so the term *close to zero* depends on the units of measurement. To avoid problems with the units of measurement, assume the X and Y values are normalized: deviations from the mean in units of the standard deviation.

2. The variance of the error term σ^2_ε is large relative to the variance of the response variable. The random fluctuation in the residual variance overwhelms the effect of the explanatory variable.

3. The data sample has so few points that the regression pattern is spurious. For example, one can draw a straight line connecting any two points, so the regression analysis means nothing. The F-value has zero degrees of freedom and is not significant no matter how large it is.

[The following exercise explains some intuition for R^2, adjusted R^2, F values, and significance.]

** Exercise 10.3: Measures of significance

Two regression equations Y and Z regress inflation rates on interest rates using data from different periods. The true population distributions of the explanatory variable and the response variable are the same in the two equations.

- Equation Y has the higher R^2 and an estimated slope coefficient of $\beta_{Y.}$
- Equation Z has the higher adjusted (corrected) R^2 and an estimated slope coefficient of $\beta_{Z.}$

A. Which regression equation uses a larger data set?
B. Which regression equation has a greater F-value?
C. Which is the better estimate of the slope coefficient: β_Y or β_Z?

Part A: Equation Y has the higher R^2 and the lower adjusted (corrected) R^2. This implies that Equation Y has fewer data points, and more of its R^2 is spurious.

Part B: The F-test uses the same adjustment for degree of freedom as the adjusted R^2, so Equation Z has the higher F-value.

Part C: β_Z has the higher t-value (the square root of the F-value), so it is the better estimate. In practice, we would use a weighted average of the two ß's, with more weight given to Equation Z.

** Exercise 10.4: R^2

A simple (two-variable) linear regression model $Y_i = \alpha + \beta \times X_i + \epsilon_i$ is fit to the 5 points:

$$(0, 0), (1, 1), (2, 4), (3, 4), (4, 6)$$

A. What is the mean X value?
B. What is the mean Y value?
C. What are the five points in deviation form?
D. What is $\sum(x_i - \overline{x})^2$?
E. What is $\sum(y_i - \overline{y})^2$?
F. What is $\sum(x_i - \overline{x})(y_i - \overline{y})$?
G. What is R^2?
H. What is the adjusted (corrected) R^2?

Part A: The mean X value ($\overline{x}$) = (0 + 1 + 2 + 3 + 4) / 5 = 2

Part B: The mean Y value ($\overline{y}$) = (0 + 1 + 4 + 4 + 6) / 5 = 3

Part C: For the deviations from the mean, subtract 2 from each X value and 3 from each Y value to get

$$(–2, –3), (–1, –2), (0, 1), (1, 1), (2, 3)$$

Part D: $\sum(x_i - \overline{x})^2 = 4 + 1 + 0 + 1 + 4 = 10$

Part E: $\sum(y_i - \overline{y})^2 = 9 + 4 + 1 + 1 + 9 = 24$

Part F: $\sum(x_i - \overline{x})(y_i - \overline{y}) = 6 + 2 + 0 + 1 + 6 = 15$

Part G: The total sum of squares (TSS) = $\sum(y_i - \overline{y})^2 = 9 + 4 + 1 + 1 + 9 = 24$

The regression sum of squares (RegSS) = $[\sum(x_i - \overline{x})(y_i - \overline{y})]^2 / \sum(x_i - \overline{x})^2 = 15^2 / 10 = 22.5$

The R^2 = RegSS / TSS = 22.5 / 24 = 93.75%

Part H: Adjusted $R^2 = 1 - (1 - R^2) \times (N - 1) / (N - k) = 1 - (1 - 0.9375) \times (5 - 1) / (5 - 2) = 0.917$

** Question 10.5: Adjusted R^2

We fit the model $Y_i = \alpha + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \beta_4 X_{4i} + \epsilon_i$ to N observations.

- Y = the expected value of R^2
- Z = the expected value of the adjusted R^2.

As N increases, which of the following is true?

A. Y increases and Z increases
B. Y increases and Z decreases
C. Y decreases and Z increases
D. Y decreases and Z decreases
E. Y decreases and Z stays the same

Answer 10.5: E

If N = 2, R^2 = 100%, since we can fit a straight line connecting two points. As N increases, R^2 declines to the square of the correlation between the population variables X and Y.

The adjusted R^2 is corrected for degrees of freedom, so its expected value is the square of the correlation between the variables X and Y, regardless of N.

Intuition: R^2 is correct for large samples and overstated for small samples.

The adjusted (corrected) R^2 is an unbiased estimate for all samples.

** Question 10.6: Adjusted R^2

We estimate two regression equations, S and T, with a different number of observations and a different number of independent variables in each regression equation.

- R^2_s and R^2_t are the R^2 for equations S and T.
- N_s and N_t are the number of observations for equations S and T.
- K_s and K_t are the number of independent variables for equations S and T.

$R^2_s = R^2_t$. Under what conditions is the adjusted R^2 for equation S definitely greater than the adjusted R^2 for equation T?

A. $N_s > N_t$ and $K_s > K_t$
B. $N_s < N_t$ and $K_s < K_t$
C. $N_s > N_t$ and $K_s < K_t$
D. $N_s < N_t$ and $K_s > K_t$
E. In all scenarios, the adjusted R^2 for equation S may be more or less than the adjusted R^2 for equation T.

Answer 10.6: C

Use the formula for the adjusted R^2 in terms of R^2, N, and *k*. Intuitively, the difference between the R^2 and the adjusted R^2 decreases as the degrees of freedom increase.

Adjusted $R^2 = 1 - (1 - R^2) \times (N - 1) / (N - k)$.

N is more than *k*. The value of (N-1)/(N-k)

- decreases as N increases
- increases as *k* increases

As (N-1)/(N-k) *decreases*, the adjusted R^2 *increases*. Choice C has these relations.

Fox Module 10 Advanced multiple regression

REGRESSION ANALYSIS SUM OF SQUARES AND R^2 *PRACTICE PROBLEMS*

(The attached PDF file has better formatting.)

Know the three types of sums of squares: total, residual, and regression.

- Ordinary least squares estimators minimize the sums of squared residuals.
- The estimator for σ is a sum of squares adjusted for degrees of freedom.
- The R^2 is a ratio of two sums of squares.
- The adjusted (corrected) R^2 adjusts this ratio for degrees of freedom.
- The *F*-statistic is a similar ratio, also adjusted for degrees of freedom.

Most regression concepts are based on sums of squares. Standardized coefficients and generalized linear models adjust the sum of squares for the conditional distributions of the explanatory and response variables. GLMs use maximum likelihood estimation, which is similar to (not identical to) minimizing a normalized sum of squares.

Final exam problems are of two types.

- Quantitative problems compute the various sums of squares, R^2, adjusted R^2, analysis of variance, *F*-statistic, and similar items.
- Qualitative problems ask how these items change with units of measurement, number of observations, and displacement.

** Exercise 10.1: R^2

Ten pairs of observations (X_i, Y_i) are fit to the model $Y_i = \alpha + \beta \times X_i + \epsilon_i$, where ε_i are independent, normally distributed random variables with mean 0 and variance σ^2.

$\sum x_i = 50$
$\sum x^2_i = 1{,}050$
$\sum y_i = 60$
$\sum y^2_i = 3{,}560$
$\sum x_i\, y_i = 1{,}260$

A. What is TSS, the total sum of squares?
B. What is RegSS, the regression sum of squares?
C. What is R^2, the coefficient of determination?
D. What is the correlation of X and Y?
E. What is RSS, the residual sum of squares?
F. What is the (omnibus) F-value for this regression?

Part A: $TSS = \sum(y_i - \overline{y})^2 = 3{,}560 - 60^2 / 10 = 3{,}200$

Part B: $RegSS = [\, \sum(x_i - \overline{x})(y_i - \overline{y})\,]^2 / \sum(x_i - \overline{x})^2 = 960^2 / 800 = 1{,}152$

$\sum(x_i - \overline{x})^2 = 1{,}050 - 50^2 / 10 = 800$
$\sum(y_i - \overline{y})^2 = 3{,}560 - 60^2 / 10 = 3{,}200$
$\sum(x_i - \overline{x})(y_i - \overline{y}) = 1{,}260 - 50 \times 60 / 10 = 960$

Jacob: What is the rationale for this formula?

Rachel: The regression sum of squares RegSS =

$\sum(\, \hat{y}_i - \overline{y})^2 = \sum [\, (\alpha + \beta x_i) - (\alpha + \beta \overline{x})\,]^2 = \beta^2 \sum\, (x_i - \overline{x})^2$

$\beta = \sum(x_i - \overline{x})(y_i - \overline{y}) / \sum(x_i - \overline{x})^2$, so $RegSS = [\, \sum(x_i - \overline{x})(y_i - \overline{y})\,]^2 / \sum(x_i - \overline{x})^2$

Part C: $R^2 = 1{,}152 / 3{,}200 = 0.360 = 36\%$

Part D: The correlation of X and Y is

$(\sum x_i\, y_i - N\, \sum x_i\, \sum y_i) / [\, (\sum x^2_i - N\, \sum x_i) \times (\sum y^2_i - N\, \sum y_i)\,]^{0.5}$

$= (1{,}260 - 50 \times 60 / 10) / [\, (1{,}050 - 50^2 / 10) \times (3{,}560 - 60^2 / 10)\,]^{½} = 0.600$

Using deviations from the means, the correlation is

$\sum(x_i - \overline{x})(y_i - \overline{y}) / (\sum(x_i - \overline{x})^2 \times \sum(y_i - \overline{y})^2\,)^{0.5} = = 960 / (800 \times 3{,}200)^{½} = 0.600$

Note: R^2 is the square of the correlation $= 0.600^2 = 0.360$

Part E: The residual sum of squares RSS = TSS – RegSS = 3,200 – 1,152 = 2,048

Part F: The omnibus *F*-value is (RegSS / *k*) / (RSS / N-*k*-1) = 1,152 / (2,048 / (10 - 1 - 1) = 4.500

Module 11: Statistical inference for simple linear regression

(The attached PDF file has better formatting.)

Intuition: *P-VALUES VS CRITICAL VALUES*

Jacob: We often speak of rejecting a null hypothesis at a 95% or a 90% confidence interval. We can also phrase hypothesis testing with *p*-values. Which is better?

Rachel: Statisticians prefer *p*-values. If we reject a null hypothesis at a 5% significance level, we don't know if the *p*-value is 5.1%, and the null hypothesis probably ought to be rejected (or viewed with suspicion) or the *p*-value is 50%, and the null hypothesis should not be rejected. A 5% significance level is an arbitrary choice; it has no greater justification than a 6% level or a 4% level or any other level. Yet social scientists sometimes speak of regression results as absolutes; they say a certain result is significant or is not significant. This is misleading.

Jacob: If the *p*-value is better, why do we use arbitrary confidence intervals?

Rachel: A lay person may have trouble interpreting a *p*-value. Suppose we want to know if women or more likely than men to vote for one of two candidates in an election. If we say "the *p*-value is 8.2%," the listener says: "What does that mean?" Explaining the statistical meaning to a lay person may not help. So we choose a significance level and say: "yes" or "no." The listener may not realize that we could change "yes" to "no" by changing the significance level.

Jacob: For actuaries, is the *p*-value a good measure?

Rachel: It is a better measure than a significant test, but it suffers from the same problems. Listeners thinks we are testing the observed relation between the X and Y variables, but we are only testing the null hypothesis. In many regression analyses, we are confident that β is not zero, but we don't know its true value.

Illustration: Suppose we are determining the inflation rate, the interest rate, or a loss cost trend. We know that the trend is not 0%, but we don't know its true value, such as 8%, 9%, or 10%. A *p*-value is no help. If the observed trend is 8.7%, the *p*-value may be 0.01%. This doesn't tell us that the trend is 8.7%; it says that the trend is not 0%, which we know.

Jacob: Is a confidence interval better?

Rachel: It is better to say that we are P% confident that the true trend is between 8.7% – z and 8.7% + z.

Jacob: This seems like a good statement; it answers our concerns about the true trend.

Rachel: Not necessarily. We want to know the current trend. But the statistical statement says the following: “If the trend has been stable over the experience period, and any observed differences over the years stem solely from sampling error, then the true trend is between 8.7% – z and 8.7% + z.” Our listeners respond: “We do not assume the trend is the same every year. It may change from year to year. We want to know the best estimate of the current trend.”

Jacob: Isn’t the ordinary least squares estimator the best estimate of the current trend?

Rachel: Suppose we have 11 years with trends of 8.0%, 8.2%, 8.4%, …, 9.8%, and 10.0%. The standard trend analysis gives an ordinary least squares estimator of 9.0%. Our listeners are likely to reject this in favor of a 10.0% current trend.

Jacob: For trend analyses, should we should use the most recent value?

Rachel: Suppose we examine 11 years, and we find trends of

9.0%, 8.2%, 8.8%, 9.8%, 8.4%, 9.6%, 8.6%, 9.4%, 8.0%, and 10.0%.

We ascribe the differences to sampling error, and we choose a trend of 9%, not 10%.

Jacob: How do we choose between these two scenarios?

Rachel: The time series course deals with this choice. The first scenario is a random walk, and the second scenario is white noise. The time series question is “How much of the observed annual differences is the drift of a random walk and how much is sampling error of white noise?”

Fox Module 11: Statistical inference for simple linear regression

(The attached PDF file has better formatting.)

REGRESSION ANALYSIS UNITS OF MEASUREMENT PRACTICE PROBLEMS

Know how ordinary least squares estimators, their standard errors, t-values, and p-values depend on the units of measurement and displacement from the origin. The principles are

- Multiplying the explanatory variable by *k* multiplies its β by 1/*k*.
- Multiplying the response variable by *k* multiplies all the β's by *k*.
- Displacements of explanatory variables and the response variable from the origin changes α, not the ß's.

Intuition: β is in units of response variable / explanatory variable.

Illustration: Suppose claim frequency = α + β × kilometers driven.

- α is in units of claim frequency.
- β is in units of claim frequency / kilometers driven

If we write the regression equation as claim frequency = α + 1000β × meters driven.

- α is in units of claim frequency.
- β is in units of claim frequency / kilometers driven

Intuition: The ß's depend on the deviations of the values from their means. A constant displacement of all the values doesn't affect the deviations. But a constant displacement of *k* raises the response variable Y by k × β. α has the same displacement as the response variable, so it also rises by k × β.

Elasticities, standardized coefficients, and *t*-values are unit-less.

- Elasticities are percentage changes: $\partial Y/Y \,/\, \partial X/X$.
- The change in a value has the same units as the value itself.
  - If X is kilometers driven, then ∂X is also measured in kilometers driven.
  - If Y is claim frequency, then ∂Y is also measured in claim frequency.

Standardized coefficients are $\beta \times \sigma_x \,/\, \sigma_y$.

- β is in units of Y / X.
- σ_x is in units of X.
- σ_y. Is in units of Y.

⇛ The standardized coefficient is unit-less.

Measures of significance are not affected by units of measurement.

- The *t*-value is the ordinary least squares estimator divided by its standard deviation.
- The estimator and its standard deviation have the same units, so the *t*-value is unit-less.

The correlation between two random variables is unrelated to units of measurement, so the R^2 statistic is also unit-less.

*Question 11.1: Goodness-of-fit and Units of Measurement

We use least squares regression with N pairs of observations (X_i, Y_i) to estimate average *annual* claims cost in dollars per average *miles driven* each week, giving $Y= 50 + 40X + \epsilon$.

If we change the parameters to annual claims costs in *Euros* and *kilometers* driven per week, which of the following is true?

A. The R^2 increases and the *t* value for kilometers driven increases
B. The R^2 increases and the *t* value for kilometers driven decreases
C. The R^2 decreases and the *t* value for kilometers driven increases
D. The R^2 decreases and the *t* value for kilometers driven decreases
E. The R^2 stays the same and the *t* value for kilometers driven stays the same

Answer 11.1: E

The R^2 and the *t* statistic are both unit-less.

- The R^2 is a proportion. If we double the units of Y, the TSS, RegSS, and RSS all increase by a factor of $2^2 = 4$. The R^2 doesn't change.
- The *t* statistic is the ordinary least squares estimator divided by its standard deviation. If we double the units of X, both the estimator and its standard deviation decrease by 50%.

*Question 11.2: Miles Driven and Annual Claim Costs

We use least squares regression with N pairs of observations (X_i, Y_i) to estimate average *annual* claims cost in dollars per average *miles driven* per day, giving $Y = 50 + 40X + \epsilon$. For instance, a policyholder who drives an average of 25 miles a day has average claim costs of 50 + 40 × 25 = 1,050 dollars a year.

If we change the parameters to annual claims costs in Euros and kilometers driven a day, what is the revised regression equation? For this problem, assume €1.00 = $1.25 and 1 kilometer = ⅝ mile (five eighths of a mile).

A. $Y = 40 + 40X + \epsilon$
B. $Y = 40 + 20X + \epsilon$
C. $Y = 40 + 64X + \epsilon$
D. $Y = 62.5 + 25X + \epsilon$
E. $Y = 62.5 + 64X + \epsilon$

Answer 11.2: B

The estimate of β is the covariance $\rho(x,y)$ divided by the variance of X.

- Using euros multiplies each Y value by 1.00 / 1.25 = 0.80.
- Using kilometer multiplies each X value by 8/5 = 1.60.

Illustration: \$10.00 = 10 × 0.80 = €8.00, and 10 miles = 10 × 1.60 = 16 kilometers.

Multiplying the Y values by 0.80 and the X values by 1.60

- Multiplies the covariance by 0.80 × 1.60 = 1.280
- Multiplies the variance of X by $1.60^2 = 2.560$

This multiplies β by 1.280 / 2.560 = 0.500.

α is not affected by the units of X, since the product $\beta \times X$ is not affected by the units of X. But α varies directly with the units of Y: if Y is multiplied by 0.80, α is multiplied by 0.80.

Jacob: Is the product $\beta \times X$ unit-less?

Rachel: No; the product is in the units of Y.

We can check our result numerically:

- Before the change, if X = 0 miles, Y = \$50. Now X = 0 gives Y = €40, so α is 40.
- Before the change, if X = 5 miles, Y = \$250. Now X = 8 kilometers gives Y = \$250 × 0.8 = €200. Since α = 40, β is (200 – 40) / 8 = 20.

*Question 11.3: Displacement

We regress Y on X with a two-variable regression model $Y_i = \alpha + \beta \times X_i + \varepsilon_i$.

- X is the number of hours studied as a deviation from its mean.
- Y is the exam score as a deviation from its mean.

We change the values of X and Y to

- X is the actual number of hours studied (mean = 80 hours)
- Y is the actual exam score (mean score = 80)

Which of the following is true?

A. The R^2 increases and the adjusted R^2 increases
B. The R^2 increases and the adjusted R^2 stays the same
C. The R^2 decreases and the adjusted R^2 increases
D. The R^2 decreases and the adjusted R^2 stays the same
E. The R^2 stays the same and the adjusted R^2 stays the same

Answer 11.3: E

The displacement of X and Y does not affect the correlation between the random variables, so it does not affect the R^2 or the adjusted R^2.

*Question 11.4: Displacement

We regress Y on X with a two-variable regression model $Y_i = \alpha + \beta \times X_i + \varepsilon_i$. Which of the following is true?

A. If we double each X value and decrease each Y value by 1, α increases.
B. If we double each X value but don't change the Y values, α decreases.
C. If we double each X value and increase each Y value by 1, α decreases.
D. If we double each X value and increase each Y value by 1, α increases.
E. If we double each X value and decrease each Y value by 1, α stays the same.

Answer 11.4: D

- Doubling each X value reduces β by 50% but does not change α.
- Increasing each Y value by 1 increases α by 1 but does not change β.

*Question 11.5: Standardized Coefficients and Elasticities

We regress the average auto insurance loss costs in *dollars* (the Y dependent variable) on the number of hours the auto is driven each week (the X independent variable). We estimate the ordinary least squares estimator $\hat{\beta}$, the standardized coefficient $\hat{\beta}^*$, and the elasticity η.

If we use Euros for the loss costs instead of dollars, which of the following is true? Assume that one Euro is 1.25 dollars.

A. $\hat{\beta}$ increases; $\hat{\beta}^*$ and η stay the same.

B. $\hat{\beta}$ decreases; $\hat{\beta}^*$ and η stay the same.

C. $\hat{\beta}$ and $\hat{\beta}^*$ stay the same; and η increases.

D. $\hat{\beta}$ and $\hat{\beta}^*$ stay the same; and η decreases.

E. $\hat{\beta}$ and η stay the same, and $\hat{\beta}^*$ increases.

Answer 11.5: B

If an hour of driving each week increases loss costs by $10, it increases loss costs by €8, so β decreases.

The standardized coefficient and elasticity are unit-less, so they are not affected by a change in the units of measurement.

Module 11: Statistical inference for simple linear regression

(The attached PDF file has better formatting.)

Basic regression principles practice problems

\
*Question 11.1: Key assumptions

Classical regression analysis is based on the statistical model

$$y_j = \alpha + \beta_1 x_{j1} + \beta_2 x_{j2} + \ldots + \beta_k x_{jk} + \varepsilon_j$$

All but which of the following are among the assumptions of classical regression analysis?

A. Linearity: $E(x_j) = \bar{x}$ for each x_i
B. Constant variance: $Var(\varepsilon_j) = \sigma_\varepsilon^2$
C. Normality: $\varepsilon_j \sim N(0, \sigma_\varepsilon^2)$
D. Independence: ϵ_j, ϵ_k are independent for $j \neq k$
E. The X values are fixed or, if random, are measured without error and are independent of the errors.

Answer 11.1: A

Choice A should be one of the following:

- Linearity: $E(\epsilon_i) = 0$
- Linearity: $\bar{Y} = \alpha + \beta_1 \bar{X}_1 + \beta_2 \bar{X}_2 + \ldots + \beta_k \bar{X}_k$

*Question 11.2: Conditional Means

A regression equation is given by $y_j = \alpha + \beta_1 x_{j1} + \beta_2 x_{j2} + \ldots + \beta_k x_{jk} + \varepsilon_j$

The textbook refers to the conditional means of the Y_j values as $\mu_1, \mu_2, \ldots, \mu_n$ for $1 \leq j \leq n$.

These means are conditional on what values?

A. α
B. β
C. α and β
D. x_j
E. ϵ_j

Answer 11.2: D

*Question 11.3: Least-squares coefficients

A regression model with N observations and k explanatory variables is

$$y_j = \alpha + \beta_1 x_{j1} + \beta_2 x_{j2} + \ldots + \beta_k x_{jk} + \varepsilon_j$$

Under the assumptions of classical regression analysis, the least-squares coefficients are

A. Linear functions of the data
B. Unbiased estimators of the population regression coefficients
C. The most efficient unbiased estimators of the population regression coefficients
D. The same as the χ-squared estimators
E. Normal distributed

Answer 11.3: D

Choice D should be: The same as the maximum likelihood estimators

Module 12: Statistical inference for multiple linear regression

(The attached PDF file has better formatting.)

Intuition: R^2 *VS STANDARD ERROR*

Jacob: We have two measures of goodness-of-fit: R^2 and the standard error of the regression. Do they measure the same thing? Is one a function of the other, like the *t* statistic and the *F* statistic for the two-variable regression model?

Rachel: The standard error measures the *unexplained variance*. This is the residual sum of squares, or the error sum of squares: what is left over after the regression. The R^2 measures the *percentage* of the total sum of squares that is explained by the regression.

If the total sum of squares is held constant, the R^2 and the standard error measure the same thing. If the total sum of squares differs for two regression equations, the equation with the larger total sum of squares may have the same R^2 but a higher standard error.

Jacob: Why would the total sum of squares be the same for two regression equations? The total sum of squares is the deviation of the Y values from their mean. The Y values are stochastic, so why should the total sum of squares be the same for different regressions?

Rachel: Suppose we seek to explain a phenomenon, such as the scores on an actuarial exam. We consider several explanatory variables, such as the hours the candidate studies, the college grades, the actuarial courses taken, or the years of work experience.

- Given a regression equation, we assume the Y values are stochastic. In theory, the regression is an experiment: we choose values for the independent variable, and we observe the resulting values of the dependent variable.
- In practice, we begin with the observed scores. We hypothesize explanations, and we test each explanatory variable.

If the total sum of squares is fixed before we begin forming regression equations, the R^2 and the standard error measure the same thing. If we start in the opposite direction – first choosing values of the independent variable and then observing the Y values – the total sum of squares is not fixed.

Jacob: If the total sum of squares is not fixed, which is the better measure of the goodness-of-fit, the R^2 or the standard error?

Rachel: If the residuals have a normal distribution with a constant variance, the standard error should not depend on the dispersion of the X values. The standard error is an unbiased estimate of σ^2, which is assumed to be constant. If the variance is not constant, a wider dispersion of X values may give a higher σ^2.

Module 12: Statistical inference for multiple regression

(The attached PDF file has better formatting.)

Multiple regression practice problems

*Question 12.1: Variance of beta

A multiple regression model is $y_j = \alpha + \beta_1 x_{j1} + \beta_2 x_{j2} + \ldots + \beta_k x_{jk} + \varepsilon_j$

- σ_ε is the standard error of the regression.
- S_j is the variance of explanatory variable X_j.
- R^2_j is the R^2 for explanatory variable *j* regressed on the other explanatory variables.

Which is the correct expression for the variance of the estimator for β_j?

A. $V(B_j) = \frac{1}{1-R_j^2} \times \frac{\sigma_\varepsilon^2}{S_j^2}$

B. $V(B_j) = \frac{1}{1-R_j^2} \times \frac{\sigma_\varepsilon^2}{(n-2)S_j^2}$

C. $V(B_j) = \frac{1}{1-R_j^2} \times \frac{\sigma_\varepsilon^2}{(n-1)S_j^2}$

D. $V(B_j) = \frac{1}{R_j^2-1} \times \frac{\sigma_\varepsilon^2}{(n-1)S_j^2}$

E. $V(B_j) = \frac{R_j^2}{1-R_j^2} \times \frac{\sigma_\varepsilon^2}{(n-1)S_j^2}$

Answer 12.1: C

Know the formulas for the variance and standard deviation of the least squares estimators of the regression coefficients. Focus on the meaning of each variable and the effects, such as “What does R_j mean? If R_j increases, does the variance of B_j increase or decrease?”

*Question 12.2: *F*-Test

- RegSS is the regression sum of squares in Fox's text (other authors use RSS)
- RSS is the residual (error) sum of squares in Fox's text (other authors use ESS)
- TSS is the total sum of squares
- n is the number of data points in the sample
- k is the number of explanatory variables (not including the intercept)

An F-statistic testing the hypothesis that all the slopes (ß's) are zero has the expression

A. $\frac{\mathrm{Re}gSS / k}{RSS / (n-k-1)}$

B. $\frac{\mathrm{Re}gSS / (n-k-1)}{RSS / k}$

C. $\frac{RSS / (n-k-1)}{\mathrm{Re}gSS / k}$

D. $\frac{\mathrm{Re}gSS / k}{RSS / (n-k-2)}$

E. $\frac{RSS / k}{\mathrm{Re}gSS / (n-k-1)}$

Answer 12.2: A

Take heed: The formula for the F statistic can be written using RSS, RegSS, or R^2. The three formulas are equivalent. Know all three for the final exam.

*Question 12.3: Degrees of freedom of F-statistic

A regression model has 14 data points, 3 explanatory variables (ß's), and an intercept.

An F-test for the null hypothesis that **2** slopes are 0 has how many degrees of freedom?

A. 3 and 10
B. 2 and 10
C. 4 and 11
D. 3 and 11
E. 2 and 11

Answer 12.3: B

Degrees of freedom = q and (n – k – 1) (p119)

*Question 12.4: Bias

A statistician regresses Y on two explanatory variables X_1 and X_2 but does not use a third explanatory variable X_3. Under which of the following conditions will β_2 be biased?

A. $\rho(Y, X_3) = 0$ and $\rho(X_2, X_3) \neq 0$
B. $\rho(Y, X_3) \neq 0$ and $\rho(X_2, X_3) = 0$
C. $\rho(Y, X_3) \neq 0$ and $\rho(X_2, X_3) \neq 0$
D. $\rho(Y, X_2) \neq 0$ and $\rho(X_2, X_3) \neq 0$
E. $\rho(Y, X_2) = 0$ and $\rho(X_2, X_3) \neq 0$

Answer 12.4: C

Module 12: Statistical inference for multiple regression

(The attached PDF file has better formatting.)

Homework assignment: F test and analysis of variance

This homework assignment continues the scenario in Module 9.

We regress the Y values on the X_1 and X_2 values in the table below.

X_1	X_2	Y	X_1	X_2	Y	X_1	X_2	Y	X_1	X_2	Y
1	1	-0.395	1	2	-1.705	1	3	-2.942	1	4	-3.634
2	1	1.942	2	2	0.964	2	3	-2.463	2	4	-1.349
3	1	1.717	3	2	0.206	3	3	0.397	3	4	-0.982
4	1	2.258	4	2	2.908	4	3	-0.092	4	4	-0.235

A. What is the null hypothesis for the omnibus *F* test?
B. What are the total sum of squares (TSS), regression sum of squares (ResSS), and residual sum of squares (RSS)?
C. What are the degrees of freedom for the residual sum of squares and regression sum of squares?
D. What is the value of the *F* statistic?
E. What is the *p* value for this F statistic?

- Show the formulas and the computations for Parts A through D.
- Use Excel or other statistical software to find the *p* value in Part E.

Module 12: Statistical inference for multiple linear regression

(The attached PDF file has better formatting.)

REGRESSION ANALYSIS PRACTICE PROBLEMS F TEST

F tests are used in the student project as well as on the final exam. The practice problems show the material needed for the final exam. (Fox uses RSS instead of ESS.)

*Question 1.1: F Tests

An insurer uses 3 independent variables and one constant term to forecast auto insurance rates. The 3 independent variables are driver age, territory, and driving record.

The insurer presumes that driving record is a significant rating variable, but driver age and territory might not be significant. The insurer uses an *F* test to determine if driver age and territory are not significant (in combination). There are 35 observations. The appropriate test statistic is

$$F_{q,N-k} = \frac{(ESS_R - ESS_{UR})/q}{ESS_{UR}/(N-k)}$$

or

$$F_{q,N-k} = \frac{(RSS_R - RSS_{UR})/q}{RSS_{UR}/(N-k)}$$

TSS is the total sum of squares and ESS (RSS) is the error (residual) sum of squares.

Which of the following is true?

A. ESS is greater for the restricted equation, but TSS is the same for both equations.
B. ESS is greater for the unrestricted equation, but TSS is the same for both equations.
C. ESS and TSS are both greater for the restricted equation.
D. ESS and TSS are both greater for the unrestricted equation.
E. ESS may be greater for either equation, but TSS is the same for both equations.

Answer 1.1: A

- The total sum of squares (TSS) depends only on the Y values, so it is the same for both regression equations.
- The error (residual) sum of squares is the unexplained variation. Adding an independent variable explains more of the variation. Even if the additional variable is not related to the dependent variable, it "explains" some of the variance just by sampling error.

Jacob: If the additional variable is unrelated to the dependent variable, how much more of the variance do we expect it to explain?

Rachel: This is degrees of freedom. Suppose a sample of N data points has a variance of σ^2. The total sum of squares of these N points is $(N-1)\sigma^2$. If we use an unrelated independent variable to help explain the variance, the unexplained variance remains σ^2, so the unexplained variation is $(N-2)\sigma^2$. The additional variation explained is σ^2, so the additional variance explained is $\sigma^2/(N-2)$.

Jacob: Is this true in all cases? Are these figures exact?

Rachel: These are expected figures. In any scenario, the additional variance explained may be greater or smaller. In most regressions, even seemingly unrelated variables may have some correlation. For example, if we regress the price of cars in Japan on the price of bananas in Spain, what is the expected value of β?

Jacob: These two variables are unrelated; β should be zero.

Rachel: The two items are unrelated, but both prices are affected by inflation, and inflation in Japan is related to inflation in Spain. β will be positive, even if it is close to zero.

Jacob: What if we use deflated prices for both cars and bananas?

Rachel: If the prices of cars and bananas follow random walks, we still expect a non-zero beta. Suppose we examine years 20X0 through 20X9, and 20X0 was a high price year for cars.

- If 20X0 was a high price year for bananas, 20X1 is expected to be a high price year for cars, though both prices drift back towards their means. We expect a positive β.
- If 20X0 was a low price year for bananas, 20X1 is expected to be a high price year for cars and a low price year for bananas, though both prices drift back towards their means. We expect a negative β.

Jacob: This is disconcerting. Random walks are common. If the independent variable follows a random walk, we are over-stating the significance of the regression.

Rachel: The textbook mentions this in the time series section. In the regression section, it says that the statistical tests are exact only if the classical regression assumptions are met. One of these assumptions is that the X values are *not* stochastic.

*Question 1.2: *F* Ratio and *t* Statistic

We fit 21 observations to $Y_i = \alpha + \beta \times X_i + \varepsilon_i$ with $\hat{\alpha}$ = 5.

- The means of X and Y are 1 and 3.
- The *F* statistic for testing the relation between X and Y is 1.96.

What is the *t* statistic for the ordinary least squares estimator of β?

A. –3.84
B. –1.96
C. –1.40
D. +1.40
E. +3.84

Answer 1.2: C

In the two-variable regression model, the *t* statistic is the square root of the *F* ratio: $1.96^{½}$ = ±1.400.

The means of X and Y are 1 and 3, so $3 = 5 + \beta \times 1 \Rightarrow \beta < 0$, so the *t* statistic is *negative*.

*Question 1.3: *F* Distribution

We estimate a multiple regression model with an *intercept*, *four independent* variables, and *55 observations*. We use the *F* statistic to test the hypothesis that *two* of the β coefficients are not both zero. What are the proper parameters for the *F* statistic?

A. F(2, 55)
B. F(3, 53)
C. F(4, 50)
D. F(2, 50)
E. F(4, 53)

Answer 1.3: D

The degrees of freedom are (q, N-k), where k is the number of explanatory variables including the constant term (the intercept).

q = 2
N = 55
k = 4 + 1 = 5
N – k = 50

*Question 1.4: *F* Distribution

Which of the following is true regarding the *F* Distribution?

A. The *F* Distribution is symmetrical and ranges in value from 0 to infinity.
B. The *F* Distribution is skewed and ranges in value from 0 to infinity.
C. The *F* Distribution is symmetrical and ranges in value from $-\infty$ to $+\infty$.
D. The *F* Distribution is skewed and ranges in value from $-\infty$ to $+\infty$.
E. The *F* Distribution is used to test hypotheses about a ordinary least squares estimator only when the variance of the estimator is known.

Answer 1.4: B

A sum of squares is non-negative, so the ratio of sums of squares is non-negative.

If all the sample points lie on a straight line (for a two dimensional model), the standard error of the regression coefficient is zero, the *t*-value is infinity, and the *F*-statistic is infinity.

Fox Module 13 Dummy variables

(The attached PDF file has better formatting.)

REGRESSION ANALYSIS DUMMY VARIABLES PRACTICE PROBLEMS

Much actuarial work uses dummy variables, such as male = 1 and female = 0 or urban = 1 and rural = 0 or normal blood pressure = 0 and high blood pressure = 1. John Fox applies regression to social issues and medical research, which has similar dummy variables, such as vote = 1 and did not vote = 0 or ill = 1 and healthy = 0.

*Question 13.1: Dummy Variables

To forecast auto insurance rates, an actuary uses a multiple regression model with dummy variables for territory and driver age. The state has 15 territories and 6 driver age groups. Each car is in one and only one territory and each driver is in one and only one age group.

The premium rate is the base rate plus the territorial relativity plus the age group relativity. For example, if the base rate is $1,200, territory 01 has a relativity of +$500, and age group 21-25 has a relativity of +$600, the premium rate for drivers age 21-25 in territory 01 is $2,300.

There are no interaction terms. For example, if the age group relativity for 21-25 year old drivers is +$750 in territory 01, it is +$750 in all territories.

How many dummy variables are needed for this regression?

A. 19
B. 20
C. 21
D. 88
E. 89

Answer 13.1: A

For mutually exclusive qualitative values with no N/A or abstain choice, we need one less dummy variable than the number of choices: (15 – 1) + (6 – 1) = 19

Illustration: For male vs female, we need one dummy variable, since if the driver is male, he is not female, and if she is female, she is not male.

If the state has 12 territories, and every car is in one and only one territory, we need 11 dummy variables. If the car is in one of the 11 territories represented by these 11 dummy variables, it has a 1 for that variable and a 0 for the other variables. If it has a 0 for all 11 variables, it must be in the twelfth territory.

Jacob: What does an interaction term mean?

Rachel: Suppose we have two qualitative dimensions, sex and age, with two values in each dimension: male vs female and youthful vs adult.

- For males, annual premium rates are $2,000 for adult and $5,000 for youthful.
- For females, annual premium rates are $1,000 for adult and $2,000 for youthful.

The youthful vs adult difference is higher for males than for females. We need three dummy variables, or 2 × 2 – 1, not two dummy variables, or (2 – 1) × (2 – 1).

Jacob: Interaction terms can be important. Do we maximize the number of interaction terms?

Rachel: Interactions terms make the regression analysis less efficient. We want orthogonal class dimensions, so each variable is measuring something different and no variables overlap. Orthogonal class dimensions means the variables are proxies for the same item.

Illustration: More aggressive drivers have higher accident frequencies. Young males have more testosterone than older males, and young males have more testosterone than young females. A two-class system is more efficient than a one class system. We want two classes that are not proxies for the same thing.

Actuaries related sex and age to character traits like risk-taking and aggression over sixty years ago, but they could not use hormone tests to quantify these relations. Insurers can't test their policyholders for testosterone. Over the past fifteen years, scientists have re-examined these relations with hormone levels, studying high- vs low-testosterone males and females. Many sex differences affecting behavior, such as competitiveness and risk-taking, seem to be a function hormone levels.

*Question 13.2: Dummy Variables

To forecast auto insurance rates, John uses a multiple regression model with dummy variables for five class dimensions, each of which has two values:

- males vs female
- young vs adult
- married vs unmarried
- urban resident vs suburban
- good credit score vs poor credit score

Nancy uses dummy variables for a single class dimension with $2^5 = 32$ values, such as adult, married woman living in the suburbs with a good credit score.

Let Y be the number of dummy variables used by Nancy, and Y be the number of dummy variables used by John. What is Y – Z?

A. –26
B. –22
C. 0
D. 22
E. 26

Answer 13.2: E

- Z: John needs 5 dummy variables: one for each rating dimension: $5 \times (2 - 1) = 5$.
- Y: Nancy needs 31 dummy variables, or $32 - 1$.

$Y - Z = 31 - 5 = 26$

*Question 13.3: Dummy Variables

An actuary regresses personal auto claim frequency on miles driven, deriving an intercept α and a slope parameter β. The actuary believes the intercept or the slope may differ for male vs female drivers.

The actuary wants to test three possible scenarios:

1. The intercept differs for men vs women, but the slope is assumed to be the same.
2. The slope differs for men vs women, but the intercept is assumed to be the same.
3. Both the intercept and the slope differ for men vs women.

For which of these three scenarios might the actuary use dummy variables?

A. 1 and 2 only
B. 1 and 3 only
C. 2 and 3 only
D. 1, 2, and 3
E. None of A, B, C, or D is correct

Fox shows graphics for each of these.

If the intercept differs but the slope is the same, the two regression lines are parallel. If D is the dummy variable, the two intercepts are α_1 and $\alpha_1 + D \times \alpha_2$.

If the intercept is the same but the slope differs, the two regression lines intersect at the intercept. If D is the dummy variable, the two slopes are β_1 and $\beta_1 + D \times \beta_2$

In much social research, the intercept is arbitrary. In much actuarial work, the intercept has meaning.

Answer 13.3: D

Regression analysis Module 15: Advanced interactions

(The attached PDF file has better formatting.)

Selecting the optimal model using sums of squares and degrees of freedom (*F* test)

- Tables 7.1 and 7.2 on page 139 are tested on the final exam.
- This posting explains the computations for the *F* test in these tables.

The variables are: I = income, E = education, and T = type

The regression sums of squares are

Model	*Terms*	*Sum of Squares*	*df*
1	I, E, T, I × T, E × T	24,794	8
2	I, E, T, I × T	24,556	6
3	I, E, T, E × T	23,842	6
4	I, E, T	23,666	4
5	I, E	23,074	2
6	I, T, I × T	23,488	5
7	E, T, E × T	22,710	5

For each model,

- The residual sum of squares is $\sum\left(Y-\hat{\hat{Y}}\right)^2$.
- The regression sum of squares is $\sum\left(\overline{Y}-\hat{Y}\right)^2$.
- The total sum of squares is $\sum\left(\overline{Y}-Y\right)^2$.

The total sum of squares does not depend on the model; it is 28,347 in this illustration.

Jacob: All three formulas for the sums of squares use only Y values, not X value or ß's.

Rachel: The regression sum of squares and the residual sum of squares use the fitted Y values, which depend on the X values. They vary by model.

The degrees of freedom in Table 7.1 on page 139 are the number of explanatory variables in the model (k). The degrees of freedom are actually N-k-1. This illustration shows the degrees of freedom for the numerator of the F test, which is the difference in the number of variables in the full vs reduced models. N-1 is the same for all models, so it drops out of the difference.

For the number of explanatory variables:

- I and E are one explanatory variable each.
- T, I × T, and E × T are two explanatory variables each.

Table 7.2 shows the degrees of freedom and sum of squares in the numerator of the F test.

Source	*Models*	*Sum of Squares*	*df*	*F*
Income	3 – 7	1,132	1	28.35
Education	2 – 6	1,068	1	26.75
Type	4 – 5	592	2	7.41
Income × Type	1 – 3	952	2	11.92
Education × Type	1 – 2	238	2	2.98
Residuals		3,553	89	
Total		28,347	97	

The total sum of squares is 28,347. The sample has 98 data points, so the total sum of squares has 98 – 1 = 97 degrees of freedom.

The full model (Model 1) has a regression sum of squares of 24,794, so it has a residual sum of squares of 28,347 – 24,794 = 3,553. This residual sum of squares has 98 – 8 – 1 = 89 degrees of freedom.

The denominator of the F ratio (for all tests) is 3,553 / 89 = 39.921.

Illustration: To test the significance of income, we contrast models 3 and 7.

The sum of squares is 23,842 for Model 3 and 22,710 for Model 7. The difference in the sum of squares is 23,842 – 22,710 = 1,132.

Model 3 has 6 explanatory variables and Model 7 has 5 explanatory variables. The degrees of freedom in the numerator of the F test is 6 – 5 = 1.

- The numerator of the F ratio is 1,132 / 1 = 1,132.
- The F ratio is 1,132 / 39.921 = 28.356.

Illustration: To test the significance of education × type, we contrast models 1 and 2.

The sum of squares is 24,794 for Model 1 and 24,556 for Model 2. The difference in the sum of squares is 24,794 – 24,556 = 238.

Model 1 has 8 explanatory variables and Model 2 has 6 explanatory variables. The degrees of freedom in the numerator of the F test is 8 – 6 = 2.

- The numerator of the F ratio is 238 / 2 = 119.
- The F ratio is 119 / 39.921 = 2.981.

To find the *p*-values in Table 7.2, use a table of the F-distributions or statistical software, such as Excel. If an exam problem asks for a *p*-value, it will give a table.

Module 15: Advanced interactions

(The attached PDF file has better formatting.)

Practice problems for Interactions, dummy variables, F tests

(This posting covers Modules 14 and 15.)

*Question 15.1: Hypothesis testing

We use regression analysis to compare personal auto claim frequency in urban, suburban, and rural areas.

Claim frequency = $\alpha + \beta_1 D_1 + \beta_2 D_2 + \epsilon$, where

- D_1 = 1 for urban and 0 otherwise
- D_2 = 1 for sub-urban and 0 otherwise

How do we show that territory (urban vs sub-urban vs rural) is significant?

A. The *t*-values for both β_1 and β_2 are greater than their critical values.
B. The *t*-value for either β_1 or β_2 is greater than its critical value.
C. The *t*-values for both β_1 and β_2 are less than their critical values.
D. The *F*-value for β_1 plus β_2 is greater than its critical value.
E. The *F*-value for β_1 plus β_2 is less than its critical value.

Answer 15.1: D

By "β_1 plus β_2" we mean the F test examining the combination of these two coefficients. Fox uses qualitative explanatory variables in many examples, just as actuaries use age, sex, and other attributes of the insured.

*Question 15.2: Parameters

We regress personal auto claim frequency on (i) annual driving and (ii) urban, suburban, and rural areas.

Claim frequency = $\alpha + \beta_1 D_1 + \beta_2 D_2 + \beta_3 DD + \epsilon$, where

- D_1 = 1 for urban and 0 otherwise
- D_2 = 1 for sub-urban and 0 otherwise
- DD = annual driving distance in kilometers

What is the predicted difference in claim frequency for urban vs sub-urban insureds driving 30,000 kilometers a years?

A. $\beta_1 + \beta_2 + \beta_3 \times 30{,}000$
B. $\beta_1 - \beta_2 + \beta_3 \times 30{,}000$
C. $\beta_1 + \beta_2$
D. $\beta_1 - \beta_2$
E. $\beta_1 D_1 + \beta_2 D_2$

Answer 15.2: D

*Question 15.3: Principle of marginality

An actuary regresses personal auto claim frequency on (i) amount of driving, (ii) income of driver, and (iii) territory. Driving and income are quantitative explanatory variables, and territory is a qualitative factor with three levels: urban, sub-urban, and rural.

Which of the following correctly reflects the principle of marginality?

A. The income by territory interaction is marginal to the income effect.
B. The amount of driving effect is marginal to the amount of driving by territory interaction.
C. We do not test the income by territory interaction until we test the income effect.
D. If we can rule out a main effect on theoretical grounds, we test the interaction effect.
E. None of A, B, C, or D is true.

Answer 15.3: B

(P135)

Fox emphasizes the principle of marginality. Some other textbooks do not discuss this topic. It is useful for actuarial analyses, which looks at the interactions of policyholder attributes.

Fox Module 17: Hat values practice problem

**Exercise 17.1: Hat values

A statistician regresses the nine Y values on the nine X values.

Y	6.87	6.58	7.69	6.96	7.39	14.14	7.90	15.62	13.23
X	1	2	3	4	5	6	7	8	9

A. What are the hat values at each point?
B. What is the minimum hat value in any regression equation?
C. What is the maximum hat value in any regression equation?

Part A: The hat values are shown in the table below.

# Pts	*9*	*Deviance*	*Deviance Squared*	*Hat Value*
values	1	-4	16	0.378
	2	-3	9	0.261
	3	-2	4	0.178
	4	-1	1	0.128
	5	0	0	0.111
	6	1	1	0.128
	7	2	4	0.178
	8	3	9	0.261
	9	4	16	0.378
ave/tot	5	0	60	0.222

Each hat value is: (Fox page 245) $h_i = \frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum_{j=1}^{n}(x_j - \bar{x})^2}$.

Part B: If $x_j = \bar{x}$, the hat value is 1/n; this is the minimum.

Part C: Suppose the explanatory variable has N independent points, of which (N-1) are zero and 1 is N.

The points $x_j = 0$ have hat values of $1/n + 1/n^2 / ((n-1)/n^2 + (n-1)^2/n^2) = 1/n + 1 / ((n-1) + (n-1)^2)$.

As $n \rightarrow \infty$, the hat values $\rightarrow 1/n$.

The point $x_j = N$ has a hat value of $1/n + (n-1)^2/n^2 / ((n-1)/n^2 + (n-1)^2/n^2) = 1/n + (n-1)^2 / ((n-1) + (n-1)^2)$

$= 1/n + (n-1) / n = 1$.

As $n \rightarrow \infty$, the hat value $\rightarrow 1$.

RA module 21: Structure of GLMs practice problems

(The attached PDF file has better formatting.)

Fox Regression analysis Chapter 15 Structure of Generalized linear models

** Exercise 21.1: Components of generalized linear models

A generalized linear model has three components: a linear predictor, a link function, and a random component (a conditional distribution of the response variable).

A. What is a linear predictor?
B. What is a link function?
C. What is the random component?
D. For classical regression analysis, what are these three elements?

Part A: The linear predictor is a linear function of regressors, $\eta_j = \alpha + \beta_1 X_{1j} + \beta_2 X_{2j} + \ldots + \beta_k X_{kj}$. η_j is a function of the fitted value, not necessarily the fitted value itself.

Part B: The link function is smooth and invertible linearizing function $g(\cdot)$, which transforms the expectation of the response variable, $\mu_j = E(Y_j)$, to the linear predictor $g(\mu_j) = \eta_j$.

Illustration: For a log-link function, if the linear predictor $\eta_j = \alpha + \beta_1 X_{1j} + \beta_2 X_{2j} + \ldots + \beta_k X_{kj} = 2$, the fitted value is $e^2 = 7.389$, since *ln*(7.389) = 2.

Part C: The random component specifies the conditional distribution of the response variable Y_j (for the j^{th} of n independently sampled observations), given the values of the explanatory variables in the model.

Illustration: For Poissson GLM with a log-link function, if the linear predictor $\eta_j = 2$, Y_j has a Poisson distribution with a mean of $e^2 = 7.389$.

Part D: For classical regression, the linear predictor is the same.
The link function is the identity function: $\eta_j = \mu_j = E(Y_j)$.
The random component is a normal distribution with the same variance at every point.

Jacob: What types of link functions should we know?

Rachel: Know the log link and logit link functions.

Jacob: What types of conditional distributions should we know?

Rachel: Know the Poisson, Gamma, and binomial distributions.

Jacob: The textbook does not give formulas for solving GLMs. Do we have to solve GLMs for the final exam?

Rachel: One can't solve GLMs by pencil and paper. The final exam tests the GLM concepts in the practice problems on the discussion forum; it does not give data and ask for the GLM coefficients.

** Exercise 21.2: Fitting generalized linear models

A. How are generalized linear models fit to observed data?
B. How does this differ from classical regression analysis?

Part A: Fitting a distribution to observed values has two parts:

- Choose the distribution, such as normal, Poisson, Gamma, binomial. This is the conditional distribution of the response variable.
- Choose the parameters of the distribution to maximize the likelihood (the probability) of observing the empirical data.

Jacob: For classical regression analysis, do we choose a conditional distributions of the response variable?

Rachel: Yes, we choose a normal distribution with a constant variance.

Jacob: Are there statistical methods to choose the conditional distribution?

Rachel: We use intuition and the relation of the variance to the mean.

Intuition: For probabilities, we use a binomial distribution. For counts, we might use a Poisson distribution or a negative binomial distribution. For stock prices or claim severities, we might use a lognormal distribution or a Gamma distribution.

Part B: Classical regression analysis assumes the distribution is a normal distribution with the same variance at every point. With this assumption, maximizing the likelihood is the same as minimizing the squared error of the residuals. Ordinary least squares estimation for a normal distribution is maximum likelihood estimation.

Jacob: Do we maximize the likelihood or the loglikelihood?

Rachel: The loglikelihood is a monotonic function of the loglikelihood. If we have a points $f(x_j)$, where f is a function of x, the value x_j which maximizes $f(x_j)$ is also the value which maximizes $ln(f(x_j))$.

Jacob: Is this the same as minimizing the residual deviance?

Rachel: The residual deviance is 2 × (K – loglikelihood $(x_{1,j}, x_{2,j}, \dots , x_{n,j})$. The set of $(x_{1,j}, x_{2,j}, \dots , x_{n,j})$ which maximize the likelihood also maximize the loglikelihood and minimize the residual deviance.

** Exercise 21.3: Link function

An actuary uses a generalized linear model to relate the response variable Y_j to two explanatory variables (covariates), X_1 and X_2.

- Let μ_j be the expected value for the response variable at observation *j*.
- Let η_j be the linear predictor at observation *j*

The intercept of the GLM is α, and the coefficients of X_1 and X_2 are β_1 and β_2.

A. What is the linear predictor η_j?
B. For a log-link function *g*(), what is the relation between μ_j and the independent variables?
C. For a logit link function *g*(), what is the relation between μ_j and the independent variables?

Part A: The linear predictor at observation *j* = $\alpha + \beta_1 X_{1j} + \beta_2 X_{2j}$.

Jacob: GLMs differ from classical regression they are used for multiplicative models, probability models with dichotomous random variables, and models of skewed distributions. Yet this linear predictor is the same as the formula in classical regression analysis.

Rachel: GLMs have three parts: linear predictor, link function, and conditional distribution of the response variable. The linear predictor is the same as for classical regression analysis.

Part B: $g(\mu_j) = ln(\mu_j) = \eta_j = \alpha + \beta_1 X_{1j} + \beta_2 X_{2j}$

Jacob: For classical regression analysis, do we say $Y_j = \alpha + \beta_1 X_{1j} + \beta_2 X_{2j}$?

Rachel: The observed value Y_j is a random variable; it is not equal to an expression of scalars. For classical regression analysis, we write $Y_j = \alpha + \beta_1 X_{1j} + \beta_2 X_{2j} + \epsilon_j$, where ϵ_j has a normal distribution with a mean of zero and the same variance for all values of the explanatory variables. GLMs use an *identity link function* for classical regression analysis, where g(x) = x:

$g(\mu_j) = \mu_j = \alpha + \beta_1 X_{1j} + \beta_2 X_{2j}$

Jacob: Why is the log-link function so often used in GLMs?

Rachel: Many relations are multiplicative models. For example, personal auto insurance premiums depend on driver characteristics (like male vs female) and territory (like urban vs rural). The insurance rates are a multiplicative model: the male rate may be twice the female rate and the urban rate may be three times the rural rate. The log-link function gives a multiplicative model:

$ln(\mu_j) = \eta_j = \alpha + \beta_1 X_{1j} + \beta_2 X_{2j} \Rightarrow$
$\mu_j = \exp(\alpha + \beta_1 X_{1j} + \beta_2 X_{2j}) \Rightarrow$
$\mu_j = \exp(\alpha) \times \exp(\beta_1 X_{1j}) \times \exp(\beta_2 X_{2j})$

Define new parameters:

- $\alpha' = \exp(\alpha)$ = the base rate
- $\beta_1' = \exp(\beta_1)$ = the male/female relativity
- $\beta_2' = \exp(\beta_2)$ = the urban/rural relativity

Part C: $g(\mu_j) = ln[\, \mu_j / (1 - \mu_j) \,] = \eta_j = \alpha + \beta_1 X_{1j} + \beta_2 X_{2j}$

Jacob: What is the rationale for the logit link function? Log odds may be relevant to horse racing or Las Vegas casinos, but they have no intuitive relation to actuarial distributions.

Rachel: That is true, and the logit link function is not appropriate for all actuarial distributions. But the logit link function has the proper form; it converts a range from 0 to 1 to a range from $-\infty$ to $+\infty$.

** Exercuse 21.4: Link function

An actuary uses a generalized linear model with a log-link function to relate the response variable Y_j to two explanatory variables, X_1 and X_2. Let μ_j be the expected value for the response variable at observation *j*. The intercept of the GLM is α, and the coefficients of X_1 and X_2 are β_1 and β_2.

A. What is the relation of the explanatory variables to the response variable using the link function?
B. What is the relation of the explanatory variables to the response variable using the inverse of the link function?

Part A: $ln(\mu_j) = \alpha + \beta_1 X_1 + \beta_2 X_2$

Part B: $\mu_j = \exp(\alpha + \beta_1 X_1 + \beta_2 X_2)$

Jacob: Why don't we use $ln(Y_j) = \alpha + \beta_1 X_1 + \beta_2 X_2$ and $Y_j = \exp(\alpha + \beta_1 X_1 + \beta_2 X_2)$?

Rachel: Y is a random variable: the linear predictor adjusted by the inverse of the link function plus a random component.

** Exercise 21.5: Likelihoods

A. What is the range of a likelihood?
B. What is the range of a log-likelihood?
C. What is meant by a saturated model?
D. What is the relation of the likelihood for the saturated model vs any other model?
E. What is the relation of the log-likelihood for the saturated model vs any other model?

Part A: Suppose Y is a function of X. As an example, let Y be the Poisson probability for X: $y = \mu^x e^{-\mu}/x!$

- $\text{pdf}(x \mid \mu) = (\mu^x e^{-\mu}) / x!$
- $\text{likelihood}(\mu \mid x) = (\mu^x e^{-\mu}) / x!$

The pdf (probability density function) and the likelihood have a range of [0, 1].

Part B: The logarithm of 0 is $-\infty$ and the logarithm of 1 is 0, so the range of the loglikelihood is $(-\infty, 0]$.

Part C: A saturated model has the fitted equal to the observed value at every point. If a regression equation or a GLM has N points, the saturated model has N parameters and 0 degrees of freedom.

Part D: The likelihood is greatest when μ = the observed value. For a given x, the value of $(\mu^x e^{-\mu}) / x!$ is maximized for $\mu = x$. If L_s.is the likelihood for the saturated model and L_m is the likelihood for any other model (with the same type of conditional distribution for the response variable but different parameters), then

$0 \le L_m \le L_s \le 1$

Part E: If LL_s.is the log-likelihood for the saturated model and LL_m is the log-likelihood for any other model (with the same type of conditional distribution for the response variable but different parameters), then

$-\infty \le LL_m \le LL_s \le 0$

Note: This exercise uses $\le$ (less than or equal to). Some exam problems use < (less than). Both relations are correct, as long as the model in question is not the saturated model.

Fox Module 22: *INTUITION: NORMAL, POISSON, AND EXPONENTIAL DISTRIBUTIONS*

(The attached PDF file has better formatting.)

Fox's textbook teaches you how to use GLMs. This on-line course covers the concepts of GLMs, not the details of forming and running GLMs. You don't have the software needed to run GLMs, unless you use SAS or R (or similar packages).

This posting explains the intuition of normal, Poisson, and exponential distributions. These are conditional distributions of the response variable, conditioned on the fitted value. We show how to determine the GLM values using Excel's *SOLVER* add-in.

Classical linear regression assumes a normal distribution with a constant variance. The distribution of the errors does not depend on the response variable.

GLMs speak of the conditional distribution of the response variable.

- The distribution is conditional on the explanatory variables.
- The variance of the distribution is not constant. It depends on the mean, which is a function of the explanatory variables.

The distribution of errors is replaced by a distribution of observations about their means.

- The fitted line depends on the relation of the variances of observations to their means.
- We use examples of three points fit to a straight line to clarify the intuition.

We examine normal, Poisson, and exponential distributions in this posting.

- The exponential distribution has a variance proportional to the square of the mean.
- The Poisson distribution has a variance proportional to the mean.

THE ILLUSTRATION

Illustration: We fit a linear model to three points: { (1,1), (5,5), and (9,9) }.

- The points lie on a straight line, and the linear model is $Y = 0 + X + \epsilon$.
- The fit does not depend on the distribution of the residuals. The residuals are zero, and their variance is zero regardless of the conditional distribution of the response variable.

We revise the three points and re-fit the linear model.

- We add 1 point to Y at X = 1, subtract two points at X = 5, and add one point at X = 9.
- The revised three points are { (1,2), (5,3), and (9,10) }.

Classical regression analysis assumes

- The distribution of the error terms at each point has the same variance.
- The expected (fitted) value of Y does not affect the variance of Y about its mean.

The three points still have a mean of (5,5), but the observed line is bent in the middle.

- The differences from the mean for the original points are { (–4, –4), (0, 0), (4, 4) }.
- The differences from the mean for the revised points are { (–4, –3), (0, –2), (4, 5) }.

The original line with a constant slope of 1 is now two line segments.

- From (1,2) to (5,3), the slope is 0.25.
- From (5,3) to (9,10), the slope is 1.75.

A linear model means the true slope is constant. The β coefficient is the estimated slope of the full line, so it is a weighted average of 0.25 and 1.75. The conditional distribution of the response variable determines how much weight is given to each line segment.

- For a normal distribution with the same variance at each point, adding 1 point to a Y value of 9 has the same weight as adding 1 point to a Y value of 1.
- If the standard deviation σ of the distribution is constant, the probability of being 1 point away from the mean is the same whether the mean is 1 or 9.

If the true slope is 1, the likelihood of random errors decreasing the first slope by 0.75 is the same as that of random errors increasing the second slope by 0.75.

For other distributions of the response variable, the variance is not the same at Y=1 as at Y=9, and the weights for the two slopes are not equal. For a Poisson distribution:

- The variance is proportional to its mean and the standard deviation is proportional to the square root of its mean.
- A random error of one point from a Y value of 1 is less likely than a random error of one point from a Y value of 9.

We examine first the likelihood of residuals in a Poisson distribution, and then we show how a Poisson distribution rotates the regression line in this illustration.

Exercise 1.1: Poisson distributions

We examine Poisson distributions with means of 1 vs 9, comparing

- The probability that Y = the mean μ
- The probability that Y = μ + 1

These are the residuals at fitted line Y = Z in the illustration above at the two end points:

- Fitted Y = 1 and observed Y = 1 vs 2
- Fitted Y = 9 and observed Y = 9 vs 10

What is the relation of these two probabilities for each Poisson distribution?

Solution 1.1: We compare

- The likelihood of a Y value of 2 vs 1 when the mean is 1 with
- The likelihood of a Y value of 10 vs 9 when the mean is 9.

Mean	μ = 1		μ = 9	
Observation	y = 1	y = 2	y = 9	y = 10
Probability	0.368	0.184	0.132	0.119

- The likelihood of y=2 when μ=1 is 50% of the likelihood of y=1 when μ=1.
- The likelihood of y=10 when μ=9 is 90% of the likelihood of y=9 when μ=9.

Take heed: The final exam ask about the relative likelihoods of residuals given conditional distributions of the response variable, similar to the exercise above.

Exercise 1.2: Line of Best fit

Three observed points are { (1,2), (5,3), and (9,10) }. We fit straight lines with three GLMs:

- Normal, Poisson, and exponential distributions of the observed Y values.
- Identity link functions ⇛ the fitted Y value is a linear function of the X value.

Note: The link function is important when we have two or more dimensions. This exercise has one dimension, and the solution is the same for a log-link function. We use the identity link function: that is, the fitted Y value is a linear function of the X value. This exercise does not deal with link functions. It shows how the assumed distribution affects the fitted line. *Do not worry about link functions for this posting or the accompanying homework assignment.*

- The distribution gives the variance of the response variable about its mean.
- Assume the normal distribution has the same variance at each point.

Exercise 1.3: Weighting the Residuals

The weights for the residuals depend on the variance of the response variable at the point. These weights determine the estimated slopes and intercepts.

- The β coefficient is the slope of the regression line.
- The α coefficient is the Y-intercept of the regression line.

A. Which distribution gives the most weight to the residual at X = 1? At X = 9?
B. Which distribution gives the most weight to the slope from X = 1 to X = 5? From X = 5 to X = 9?
C. Which distribution gives the highest β? The lowest β?
D. Which distribution gives the highest α? The lowest α?

Intuition: No straight line passes through all three points.

- A straight line through (1,2) and (5,3) passes through (9,4).
- A straight line through (9,10) and (5,3) passes through (1,–4).

Ordinary least squares minimizes the sum of squared errors. The straight lines in the bullet points above have residuals of zero at two points and a residual of 6 at the third point.

- The sum of squared errors for both lines above is $0^2 + 0^2 + 6^2 = 36$.
- This sum of squared errors gives equal weight to all points. Equal weights assume the variances are the same at all points.

With GLMs, the weights are not equal at all the points since the variances are not equal.

Part A: The weights depend on the *fitted Y values* at each point.

- We *don't know* the fitted Y value until we fit the line.
- The fitted line depends on the distribution of the response variable about its mean.

This exercise focuses on the GLM concepts, not the mathematics. We do not know the exact fitted value at each sample point, so we do not know the exact weights, so we can not minimize the weighted residuals by an algebraic formula. GLMs uses iterative weighted least squares: repeated numerical estimates which converge to the solution.

This exercise, as a proxy, uses the observed Y value at each point. The observed values are spread out and the residuals are small, so the proxy works well.

Take heed: The final exam may ask which point receives the most weight in a GLM.

The final exam does not require you to solve complex GLMs. The exam problems focus on the concepts of GLM analysis, not the numerical solutions to complex equations.

For a Poisson distribution of the response variable:

- The observed values are Y = 2 at X = 1 and Y = 10 at X = 9.
- The variances are approximately $\sigma^2 = 2$ at X = 1 and $\sigma^2 = 10$ at X = 9.
- The standard deviations are approximately $\sigma = 1.414$ at X = 1 and $\sigma = 3.162$ at X = 9.

The variances are approximations. The fitted Y value is near 2 but not exactly 2, so the variance is near 2, not exactly 2.

With a Poisson distribution of the response variable, the estimate is more precise at X=1 than at X=9 because the variance is smaller at X=1 than at X=9.

HETEROSCEDASTIC DATA, STANDARDIZED RESIDUALS, AND WEIGHTED LEAST SQUARES

A GLM with a Poisson distribution is like linear regression with heteroscedastic data.

- We use standardized residuals, or the residuals divided by the standard deviation.
- We give more weight to the residual at X = 1 when fitting the regression line.
- We give less weight to the residual at X = 9.

For an exponential distribution of the residuals:

- The observed values are Y = 2 at X = 1 and Y = 10 at X = 9.
- The variances are approximately $\sigma^2 = 4$ at X = 1 and $\sigma^2 = 100$ at X = 9.
- The standard deviations are approximately $\sigma = 2$ at X = 1 and $\sigma = 10$ at X = 9.

A GLM with an exponential distribution has an even greater difference of the variances.

- We give much more weight to the residual at X=1 when fitting the regression because the variance is so small at X=1 (compared to the variance at other points).
- We give even less weight to the residual at X = 9 because its variance is so large.

Of the three distributions of the error term:

- The exponential distribution gives the most weight to the observed value at X = 1.
- The normal distribution gives the most weight to the observed value at X = 9.

Take heed: The ratio of observed values (10 to 2 at X=9 compared to X=1) is not the ratio of fitted values. For the exponential distribution:

- The low variance at X=1 causes the fitted Y value to be close to Y=2.
- The high variance at X=9 causes the fitted Y value to be farther from Y=10.

Illustration: If the GLM causes the fitted values to be (1, 1.8) and (9, 8.1), the *ratio of fitted values* is 8.1 / 1.8 = 4.500 to 1.

Parts B and C: The distribution of the residuals changes the slope of the fitted line.

The fitted line for the Poisson and exponential distributions passes

- Closest to the point (1, 2).
- Less closely to the point (5,3).
- Least closely to the point (9, 10).

For the Poisson distribution compared to the normal distribution with constant variance:

- The slope of the fitted line is closer to the slope from (1,2) to (5,3) than to the slope from (5,3) to (9,10).
- We give more weight to the slope 0.25 than to the slope 1.75.
- The fitted β is closer to 0.25 than to 1.75
- $\Rightarrow$ The fitted line is flatter than a 45° diagonal line.

For an exponential distribution compared to the Poisson distribution:

- We give even more weight to the slope 0.25 than to the slope 1.75.
- β is even lower.
- The fitted line is even flatter.

Take heed: Fitting the curve changes the means of the Y values at X = 1 vs X = 9. See the fitted values at the end of this posting.

Part D: The relative values of the intercept α reflect the relative slopes.

- The normal distribution with constant variance has a slope of 1 and an α of zero.
- The Poisson distribution has a flatter line (lower slope) and a higher α.
- The exponential distribution has the flattest line (lowest slope) and the highest α.

The Poisson distribution rotates the fitted line clockwise, making it flatter. The points lie in the first quadrant, so the clockwise rotation raise the Y-intercept.

Choosing other points has different effects on α.

- If we make all the X vales negative (–1, –5, and –9) and keep the same Y values (2, 3, and 10), the fitted line slopes downward and the Y-intercept is on the right side. The Poisson distribution rotates the line counter-clockwise and raises the Y intercept.

- If we change the points to (1, 10), (5, 3), and (9, 2), the fitted line slopes downward and the Y-intercept is on the right side. The Poisson distribution rotates the line counter-clockwise and lowers the Y intercept.

- If we change the points to (–1, 10), (–5, 3), and (–9, 2), the fitted line slopes upward and the Y-intercept is on the right side. The Poisson distribution rotates the line clockwise and lowers the Y intercept.

{The homework assignment for this Module is similar to the exercise below, but it does not require computations. Read this exercise and then complete the homework assignment.}

Exercise 1.4: Line of Best fit

Three observed points are { (1,2), (5,3), and (9,10) }.

We fit straight lines with normal, Poisson, and exponential distributions of the residuals. We fit a straight line Y = a + b × X with a Poisson distribution of the residuals.

A. What is the loglikelihood at the three observed points as a function of *a* and *b*?
B. What is the total loglikelihood?
C. What two equations do we use to maximize this loglikelihood?
D. Solve the two equations using Excel's *SOLVER* add-in. Some statistical packages solve GLMs directly, including SAS and R.
E. What are the residuals and the squared residuals at each of the three points?
F. Is the sum of squared residuals more or less than with classical regression analysis?
G. What is the ratio of the fitted values at the points X = 1 and 9?
H. What is the ratio of the variances of the error terms at the points X = 1 and 9?
I. What is the ratio of the residuals at the points X = 1 and 9?
J. Compute the sum of squared residuals divided by their variances. Show that this sum is less than the sum with an ordinary least squares regression line.

Part A: For a Poisson distribution of error terms, the loglikelihood = y $ln(\mu) - \mu - ln(y!)$, where y is the observation and μ is the fitted mean.

Each Y value is a linear function of the X value: Y = a + b × X = intercept + slope × X

The loglikelihood of observing each of the three points is

- (1,2) → 2 $ln(a + 1b) - (a + 1b) - ln(2!)$
- (5,3) → 3 $ln(a + 5b) - (a + 5b) - ln(3!)$
- (9,10) → 10 $ln(a + 9b) - (a + 9b) - ln(10!)$

Part B: The loglikelihood of observing all three points is the sum of the loglikelihoods.

$$\begin{aligned} & 2\, ln(a + 1b) - (a + 1b) - ln(2!) \\ + \; & 3\, ln(a + 5b) - (a + 5b) - ln(3!) \\ + \; & 10\, ln(a + 9b) - (a + 9b) - ln(10!) \end{aligned}$$

Part C: We set the partial derivatives with respect to *a* and *b* equal to zero.

Setting partial derivative with respect to *a* equal to 0 gives

$$2 / (a + b) + 3 / (a + 5b) + 10 / (a + 9b) - 3 = 0$$

Setting partial derivative with respect to *b* equal to 0 gives

$$2 \times 1 / (a + b) + 3 \times 5 / (a + 5b) + 10 \times 9 / (a + 9b) - (1 + 5 + 9) = 0$$

Part D: We have two equations in two unknowns. They are not linear equations, so we have no closed form solution.

We use Excel's *SOLVER* add-in to find the values of *a* and *b*.

- Type the name *intercept* in Cell A11 and an estimate (such as –1) in Cell B11.
- Type the name *slope* in Cell A12 and an estimate (such as 1) in Cell B12.
- In Cell A13, enter the formula

= 2 / (intercept + slope) + 3 / (intercept + 5 * slope) + 10 / (intercept + 9 * slope) - 3

- In Cell A14, enter the formula

= 2 × 1 / (intercept + slope) + 3 * 5 / (intercept + 5 * slope) + 10 * 9 / (intercept + 9 * slope) - (1 + 5 + 9)

- In Cell A15, enter the formula = A13^2 + A14^2

Click on the *TOOLS* menu and choose *SOLVER*.

Set Cell A15 to 0 by choosing values for *intercept, slope*.

SOLVER returns the solution: intercept = 0.834517 and slope = 0.833119.

The slope of the fitted line is 0.833, so it is flatter than the least squares regression line.

Part E: The fitted values are

X	*Observed Value*	*Fitted Value*	*Residual*	*Residual Squared*
1	2	1.668	0.332	0.110
5	3	5.000	-2.000	4.000
9	10	8.333	1.667	2.780
Total	15	15.000	0.000	6.891

We show the solution using Excel's *SOLVER* add-in because most candidates use *SOLVER*. The most versatile software package for GLMs is R, which is freely available. With R:

- specify the x values as xv <- c(1,5,9)
- specify the y values as yv <- c(2,3,10)
- run the GLM as glm(yv ~ xv)

The variables names xv and yv are arbitrary; any names are fine. Specify the probability distribution as family = Poisson and the link function as link(log).

Part F: With ordinary least squares estimation, the residuals are 1, –2, and 1, and the sum of squared residuals is $1^2 + 2^2 + 1^2 = 6$.

With the Poisson distribution of error terms, the sum of squared residuals is 6.891, which is greater.

Parts G, H, I:

We compare the ratios of fitted values, variances, and residuals at the points X = 1 and 9.

- Fitted Values: 8.333 / 1.668 = 4.996 ≈ 5.
- The variances are proportional to the fitted values, so the ratio is also 5.
- Residual: 1.667 / 0.332 = 5.021 ≈ 5.

Part J: We compute the residual divided by the variance of the error term, which equals the fitted value.

X	*Observed Value*	*Fitted Value*	*Residual*	*Residual Squared*	*Rsd / Var*
1	2	1.668	0.332	0.110	0.199
5	3	5.000	-2.000	4.000	-0.400
9	10	8.333	1.667	2.780	0.200
Total	15	15.000	0.000	6.891	-0.001

The total (residual / variance) is zero, ignoring the rounding error in the table.

The Poisson distribution of error terms causes the fitted line to rotate about its mean four fifths of the way back to the observed values.

Part J: We compute the residual divided by the variance of the error term. We use two approximations for the variance.

- The variance of the error term is the fitted value (the mean of the distribution).
- The variance is the average of the observed and fitted values.

X	*Observed Value*	*Fitted Value*	*Residual*	*Residual Squared*	*Rsd / Var1*	*Rsd / Var2*
1	2	1.668	0.332	0.110	0.199	0.181
5	3	5.000	-2.000	4.000	-0.400	-0.500
9	10	8.333	1.667	2.780	0.200	0.182
Total	15	15.000	0.000	6.891	-0.001	-0.137

If we know the fitted value with certainty, we use the ratio of the residual to the fitted value.

- In this exercise, we estimate the fitted value from the observed value.
  - If we know nothing about the relation of the X and Y values, the best estimate of Y is the observed value.

- The residual + the fitted value = the observed value.
- The residuals are in the same proportion as the fitted values at X=1 and X=9.
  - The residuals are in the same proportion as the observed values at X=1 and X=9.
  - The residuals are in the same proportion as the average of the observed and fitted values at X=1 and X=9.

The Poisson distribution of error terms causes the fitted line to rotate about its mean.
It rotates four fifths of the way back to the observed values.

Exercise 1.5: Line of Best fit

Three observed points are { (1,2), (5,3), and (9,10) }.

We fit a straight line Y = a + b × X with an exponential distribution of the error terms.

A. What is the loglikelihood at the three observed points as a function of *a* and *b*?
B. What is the total loglikelihood?
C. What two equations do we use to maximize this loglikelihood?
D. Solve the two equations using Excel's *SOLVER* add-in.
E. What are the residuals and the squared residuals at each of the three points?
F. Is the sum of squared residuals more or less than with classical regression analysis?
G. What is the ratio of the fitted values at the points X = 1 and 9?
H. What is the ratio of the variances of the error terms at the points X = 1 and 9?
I. What is the ratio of the residuals at the points X = 1 and 9?
J. Compute the sum of squared residuals divided by their variances. Show that this sum is less than the sum with an ordinary least squares regression line.

Part A: For an exponential distribution of the error terms, the loglikelihood is $-y/\mu - ln(\mu)$, where y is the observation and μ is the fitted mean.

Each Y value is a linear function of the X value:

Y = a + b × X = intercept + slope × X

The loglikelihood of observing each of the three points is

- (1,2) → $- ln(a + 1b) - 2 / (a + 1b)$
- (5,3) → $- ln(a + 5b) - 3 / (a + 5b)$
- (9,10) → $- ln(a + 9b) - 10 / (a + 9b)$

Part B: The loglikelihood of observing all three points is the sum of the loglikelihoods.

$$
\begin{aligned}
& - ln(a + 1b) - 2 / (a + 1b) \\
+ \; & - ln(a + 5b) - 3 / (a + 5b) \\
+ \; & - ln(a + 9b) - 10 / (a + 9b)
\end{aligned}
$$

Part C: We set the partial derivatives with respect to *a* and *b* equal to zero.

Setting partial derivative with respect to *a* equal to 0 gives

$$-1 / (a + b) + -1 / (a + 5b) + -1 / (a + 9b) + 2 / (a + b)^2 + 3 / (a + 5b)^2 + 10 / (a + 9b)^2 = 0$$

Setting partial derivative with respect to *b* equal to 0 gives

$-1 / (a + b) + -5 / (a + 5b) + -9 / (a + 9b) + 2 * 1 / (a + b)^2 + 3 * 5 / (a + 5b)^2 + 10 * 9 / (a + 9b)^2 = 0$

Part D: We have two equations in two unknowns. They are not linear equations, so we have no closed form solution.

We use Excel's *SOLVER* add-in to find the values of *a* and *b*.

- Type the name *intercept* in Cell A11 and an estimate (such as –1) in Cell B11.
- Type the name *slope* in Cell A12 and an estimate (such as 1) in Cell B12.
- In Cell A13, enter the formula for the partial derivative with respect to *a* (*intercept*).
- In Cell A14, enter the formula for the partial derivative with respect to *b* (*slope*).
- In Cell A15, enter the formula = A13^2 + A14^2

Click on the *TOOLS* menu and choose *SOLVER*.

Set Cell A15 to 0 by choosing values for *intercept, slope*.

SOLVER returns the solution: intercept = 1.137295 and slope = 0.728246.

The slope of the fitted line is 0.728, so it is flatter than the fitted line with a Poisson distribution of error terms.

Part E: The fitted values are

X	*Observed Value*	*Fitted Value*	*Residual*	*Residual Squared*
1	2	1.866	0.134	0.018
5	3	4.779	-1.779	3.163
9	10	7.692	2.308	5.329
Total	15	14.336	0.664	8.510

We compare the ratios of fitted values, variances, and residuals at the points X = 1 and 9.

- Fitted Values: 7.692 / 1.866 = 4.122.
- Variances are proportional to the squares of the fitted values: $4.122^2 = 16.991 \approx 17$
- Residuals: 2.308 / 0.134 = 17.224 ≈ 17.

We used an exponential distribution in this exercise to keep the mathematics simple. GLM software uses a Gamma distribution. (The exponential distribution is the one-parameter version of the Gamma distribution.) The Gamma distribution is the exponential family proxy for distributions whose variance is proportional to the square of the mean.

Part F: For the three distributions of the error term:

- Normal with constant variance: the sum of squared residuals is $1^2 + 2^2 + 1^2 = 6$.

- Poisson: the sum of squared residuals is 6.891.
- Exponential: the sum of squared residuals is 8.510

Parts G, H, I: We compare the ratios of fitted values, variances, and residuals at the points X = 1 and 9.

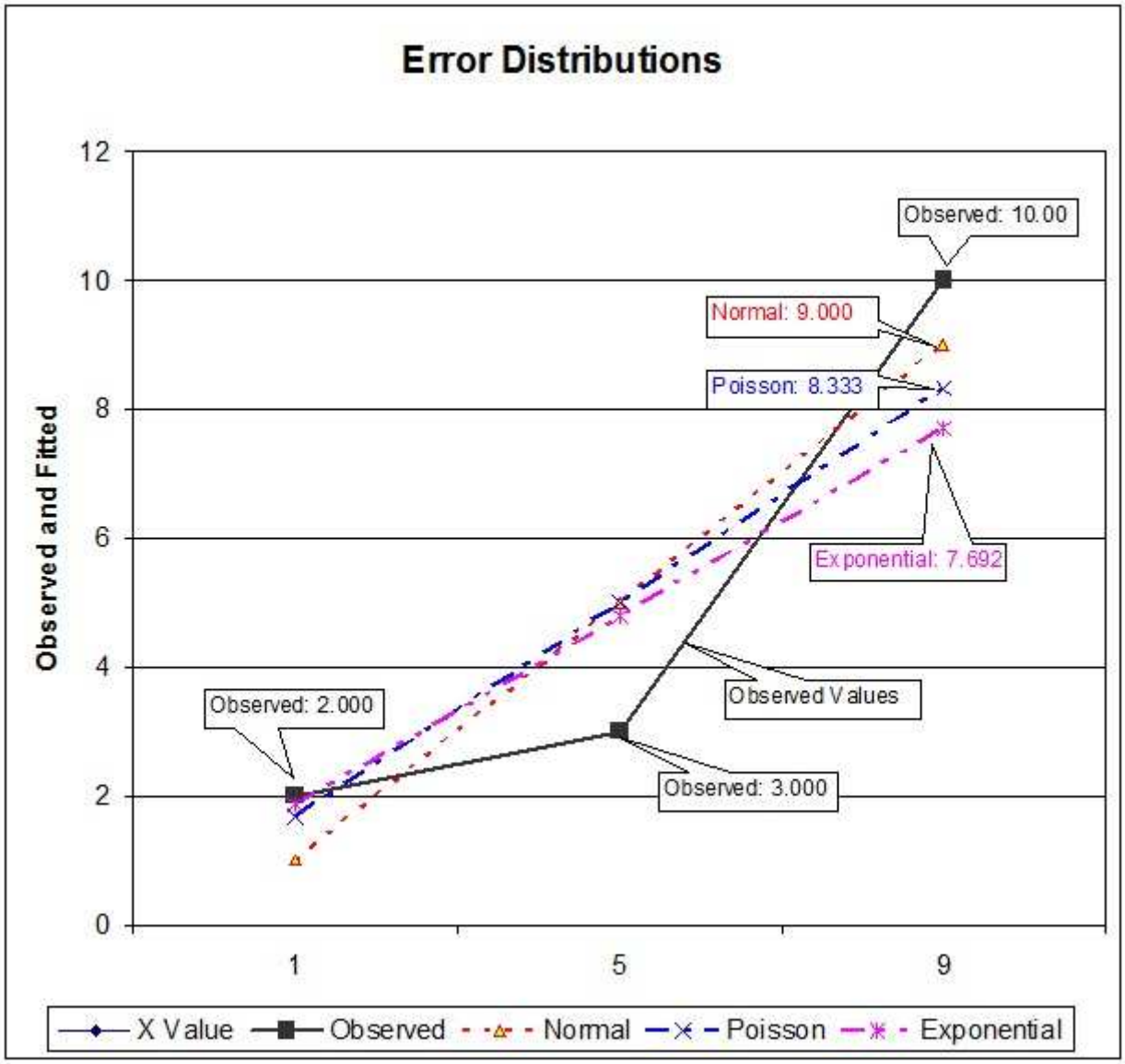
- Fitted Values: 7.692 / 1.866 = 4.122.
- Variances are proportional to the squares of the fitted values: $4.122^2 = 16.991 \approx 17$
- Residuals: 2.308 / 0.134 = 17.224 ≈ 17.

Part J: We compute the residual divided by the variance of the error term, which is the square of the fitted value.

X	*Observed Value*	*Fitted Value*	*Residual*	*Residual Squared*	*Rsd / Var*
1	2	1.866	0.134	0.018	0.038
5	3	4.779	-1.779	3.163	-0.078
9	10	7.692	2.308	5.329	0.039
Total	15	14.336	0.664	8.510	0.000

The graphic below shows the three observed values and the fitted lines. The picture helps you understand the effect of the conditional distribution of the response variable.

- The three observed values do not fall on a straight line.
  - We assume the observed values are distorted by random fluctuation.
  - Each value is an expected value plus a random error.
  - The line of best fit depends on the variance of the error terms.



The relation of the variance to the fitted value rotates the fitted line.

- If the variances are equal at all points, the residuals are equal at X=1 and X=9 (Normal distribution with constant variance).

- If the variances are proportional to the fitted values, the residuals at X=1 and X=9 are proportional to the fitted values at those points (Poisson distribution).
- If the variances are proportional to the squares of the fitted values, the residuals at X=1 and X=9 are proportional to the squares of the fitted values (exponential distribution).

		Normal		*Poisson*		*Exponential*	
X Value	*Observed Value*	*Fitted*	*Residual*	*Fitted*	*Residual*	*Fitted*	*Residual*
1	2	1.000	1.000	1.668	0.332	1.866	0.134
5	3	5.000	-2.000	5.000	-2.000	4.779	-1.779
9	10	9.000	1.000	8.333	1.667	7.692	2.308
Total			0		0		0.663

You may want to run GLMs using R. The section below provides the code.

Excel's *SOLVER* add-in shows how GLMs derive the fitted values. Statistical software, such as SAS or R, use more efficient iterative techniques. R is perhaps the most powerful and flexible GLM package. The R code for this exercise is below.

The *glm* function runs a GLM of the response variable on the covariate. The link function does not matter in this exercise. We use a Gamma distribution instead of an exponential distribution, which gives a slightly better fit, though the difference is less than 0.1%.

```
covariate <- c(1,5,9)
response <- c(2,3,10)

glm.pois.iden <- glm(response ~ covariate, family= poisson (link = "identity"))
glm.gamm.iden <- glm(response ~ covariate, family= Gamma (link = "identity"))

## list element [[3]] gives the fitted values of the GLM.

glm.pois.iden[[3]]
glm.gamm.iden[[3]]
```

R output		*Normal*		*Poisson*		*Exponential*	
X Value	*Observed Value*	*Fitted*	*Residual*	*Fitted*	*Residual*	*Fitted*	*Residual*
1	2	1.000	1.000	1.667	0.333	1.865	0.135
5	3	5.000	-2.000	5.000	-2.000	4.779	-1.779
9	10	9.000	1.000	8.333	1.667	7.694	2.306
Total			0		0		0.6618

Exercise 1.6: Maximum likelihood estimation

We use maximum likelihood to fit a linear model to three points (1, 2), (5, 3), (9, 10).

We have a choice of three distributions for the error term:

- Normal distribution with a constant variance.
- Poisson distribution
- Exponential distribution

A. Which distribution gives the lowest TSS (total sum of squares)?
B. Which distribution gives the lowest ESS (residual sum of squares)?

Part A: The mean Y value is ⅓ × (2 + 3 + 10) = 5.

The total sum of squares (TSS) is $(2 – 5)^2 + (3 – 5)^2 + (10 – 5)^2 = 38$.

- The TSS measures the dispersion of the observed values from the overall mean, not from their fitted values.
- It does not depend on the distribution of the error terms about their means.

All three distributions have the same TSS.

Part B: We explain intuitively the relative sizes of the error sum of squares.

- For a normal distribution with a constant variance, the ordinary least squares estimators are the maximum likelihood estimators.
- The ordinary least squares estimators minimize the error sum of squares.

⇛ The ESS is lowest for the normal distribution with a constant variance.

Intuition: Suppose we have two points at opposite sides of the fitted line and we rotate the fitted center about its center.

- Moving one fitted value (FV_1) closer to its observed value (OV_2) moves the other fitted value (FV_2) farther from its observed value (OV_2).
- Linear regression (a normal distribution with a constant variance) does not favor any point above others.
- It gives the same weight to X=1 as to X=9.

If the distance from the mean to the observed value is d, the ESS is $N \times d^2$.

- A GLM moves one fitted value *k* units closer to its observed value at a cost of *k* units for another fitted value.
- The GLM changes its ESS from $2d^2$ to $(d – k)^2 + (d + k)^2 = 2d^2 + 2k^2$.

RA module 22: Poisson and Gamma GLMs practice problems

(The attached PDF file has better formatting.)

Fox Regression analysis Chapter 15 Structure of Generalized linear models

** Exercise 22.1: Conditional distribution

The range of the dependent variable depends on the conditional distribution in the generalized linear model.

What are the ranges of the following conditional distributions?'

A. Gaussian (normal) distribution
B. Binomial distribution
C. Poisson distribution
D. Gamma distribution
E. Lognormal distribution

Part A: The Gaussian (normal) distribution has a range of $(-\infty, +\infty)$. This distribution is not appropriate for variables which don't have negative values or distributions which are skewed.

Part B: The binomial distribution has a range of $((0, 1, \ldots, n_j) / n_j)$, where n_j is the number of exposures.

Illustration: A study of retention rates asks how many policyholders renew. If there are n_j policyholders, 0 to n_j may renew. (Property-casualty insurance uses *renewal rate* instead of *retention rate*.) Studies of new drugs (medications) ask whether the patient recovers from illness or does not recover (or dies vs does not die).

Part C: The Poisson distribution has a range of $(0, 1, 2, \ldots)$. Claim counts may be 0, 1, 2, ... (any integer).

Illustration: Studies of insurance claim counts (or claim frequencies) use Poisson distributions.

Parts D and E: The Gamma distribution and the lognormal distribution have ranges of $(0, +\infty)$, and they are positively skewed.

Illustration: The Gamma and lognormal distributions are used for stock prices and insurance claim severities.

Jacob: Fox doesn't discuss the lognormal distribution. This seems strange, as lognormal distributions are used for many actuarial and financial distributions. For example, the Black-Scholes formula assumes stock prices have a lognormal distribution.

Rachel: The lognormal distribution is not a member of the exponential family of distributions. The Gamma distribution is similar to the lognormal distribution and is a member of the exponential family of distributions. For generalized linear models, we use Gamma distributions instead of lognormal distributions.

Jacob: The Gamma distribution differs from the lognormal distribution. With modern computers, we can do maximum likelihood estimates using either distribution. Why not use the distribution that fits the data better?

Rachel: The important item is the relative variance as a function of the expected value. For both the Gamma and lognormal distributions, the variance is proportional to the square of the mean. Maximum likelihood estimation gives (nearly) the same result for these two distributions.

** Exercise 22.2: Residual deviance

The likelihood for the model being tested is 8% and the likelihood for the saturated model is 10%.

A. What is the loglikelihood for the model being tested?
B. What is the loglikelihood for the saturated model?
C. What the residual deviance for the model being tested?

Part A: $ln(0.08) = -2.52573$

Part B: $ln(0.10) = -2.30259$

Part C: $2 \times [\, ln(0.10) - ln(0.08) \,] = 0.44629$

RA module 23: Logit and probit models practice problems

(The attached PDF file has better formatting.)

Fox Regression analysis Chapter 14 Logit and Probit Models

** Exercise 23.1: Categorical response variables

Classical regression analysis is not appropriate for models with dichotomous response variables for three reasons: normal distribution, constant variance, range of response variable.

A. What is a dichotomous response variable?
B. Why is the distribution of the error terms not normal?
C. Why is the variance of the error terms not constant?
D. What is the range of the response variable?

Part A: A dichotomous response variable takes one of two values, such a True vs False. Many medical studies have dichotomous response variable. A researcher tests the optimal dosage of a new medication. The response variable may be

- patient dies or does not die
- patient recovers or does not recover

Jacob: Does the model predict whether the patient will die or not die (recover or not recover)?

Rachel: The model predicts a death rate or a recovery rate, such as a 30% death rate. The observed values are 0 (lives) or 1 (dies). The error term has only two possible values: 0 – 30% = –30% and 1 – 30% = 70%. This distribution is not normal.

Jacob: Is this a Bernoulli distribution?

Rachel: For a single patient (a single trial), this is a Bernoulli distribution. The study may have 1,000 patients, and the distribution is a binomial distribution.

Part C: Suppose the death rate for a dosage of 10 is 30% and for a dosage of 20 is 50%. The distribution of the error term is a binomial distribution with a mean of 30% or 50%. The variance of the error term is

$(\pi \times (1 - \pi)) / N$, where π is the death rate and N is the number of patients. This variance differs for each value of the explanatory variable.

Part D: The range of the response variable is [0, 1]: the death rate ranges from 0% to 100%. If the response variable had a normal distribution, its range would be ($-\infty$ to $+\infty$).

** Exercise 23.2: Link functions and conditional distributions

An actuary examines the relation of retention rates (renewal rates) to several explanatory variables (the time since the policy was first issued to this insured, the attributes of the insured, such as sex and age, and so forth).

A. What conditional distribution should the actuary use?
B. What four link functions might the actuary use?
C. How would you choose among these four link functions?

Part A: The response variable has two values: renew or not renew; this is a Bernoulil distribution. If the data point has N exposures, the response variable has N+1 possible values: 0 renewals, 1 renewal, … , N renewals: this is a binomial distribution.

Part B: The textbook recommends four link functions: logit, probit, log-log, and complementary log log.

Part C: The logit and probit link functions are symmetric and give about the same conditional distribution of the response variable. Many statisticians prefer the logit link function because it has a simple interpretation: it is the log odds of the probability. For a likelihood of P%, the logit is ln(P% / (1 – P%)). But this rationale doesn't mean the logit link function is better or worse than the probit link function.

The log log and complementary log log link functions are skewed: one to the right and one to the left. If the observed values are skewed, one of these link functions may be better.

** Exercise 23.3: Variance

An actuary uses a generalized linear model to relate the retention rate (the probability that the policyholder renews the policy) to the time since the policy was first issued and characteristics of the policyholder.

- The dependent variable has a Bernoulli distribution: the policyholder either renews or does not renew.
- The expected value of the dependent variable is a probability of renewal.

The actuary uses a binomial distribution with a logit link function. Policyholders with longer durations since the policy was first issued have higher retention rates. The data have 1,000 policyholders at each duration since the policy was first issued with 20,000 total policyholders.

- The average observed retention for all durations is 82%.
- The predicted retention for all durations is 80%.
- The average observed retention for duration = 10 years is 90%.
- The predicted retention for duration = 10 years is 92%.

What is the variance of the retention rate for duration = 10 years?

Solution 23.3: The variance of a binomial distribution is (p × (1 – p)) × N

92% × (1 – 92%) × 1,000 = 73.6

The variance of the retention rate (= the binomial distribution / N) is (p × (1 – p)) / N

92% × (1 – 92%) / 1,000 = 0.00007360

See Fox, *Regression analysis*, Chapter 15, Structure of Generalized linear models, page 381