Regression analysis, Module 1, “Statistical models”

(The attached PDF file has better formatting.)

Homework assignment: probabilities

Fox uses the data in Table 1.1 on page 5 to infer that judges grant leave at different rates.

A. If all judges grant leave in 25% of cases, and the differences among judges are random fluctuations, what is the probability a judge (Desjardins) grants leave in 49% or more of cases?
B. If all judges grant leave in 25% of cases, and the differences among judges are random fluctuations, what is the probability that a judge (Pratte) grants leave in 9% or fewer of cases?

Write an algebraic expression for the solution. You need not compute a numerical solution.

Note: Judge Desjardins heard 47 cases and granted leave in 49% × 47 = 23 cases.

- Write the expression for 23 successes in 47 cases with a probability of 25%.
  - This is a binomial probability with π = 25%.
- Write the summation for 23 through 47 successes. You need not evaluate the sum.
  - The sum goes from 23 successes to 47 successes.

Judge Pratte heard 57 cases and granted leave in 9% × 57 = 5 cases.

- Write the expression for 5 successes in 57 cases with a probability of 25%.
- Write the summation for 0 through 5 successes. You need not evaluate the sum.
  - The sum goes from 0 successes to 5 successes.

Note: The PMF of the binomial distribution is $\binom{n}{k} p^k (1-p)^{n-k}$

where n is the number of trials and p is the probability of success on each trial.

You do not have to compute any figures for this homework assignment.

Module 2: Basics of regression analysis

(The attached PDF file has better formatting.)

Homework Assignment: attributes of classical regression analysis

CLAIM SEVERITY AND SPEED

Suppose a regression of Y = the logarithm of claim severity on X = the speed of the car satisfies the five attributes of classical regression analysis on pages 15-17. Explain whether of regression of Y' = claim severity on X = the speed of the car satisfies each attribute.

Jacob: What is this homework assignment asking?

Rachel: $Y' = e^{Y}$. If the conditional distribution of Y, given X, is symmetric, is the conditional distribution of Y', given X, symmetric or skewed? Answer this question for each of the five attributes on page 15-17:

- symmetric vs skewed
- single mode vs multiple modes
- normal vs heavy tailed
- equal vs unequal spread
- linear vs non-linear

For four of these five attributes, the relation assumed in classical regression analysis does not hold for Y' if it holds for Y.

Jacob: Are the five attributes explicitly listed?

Rachel: The five attributes are implicit in Fox's discussion: symmetric, unimodal, normal distribution, constant variance, and linear relation.

Module 3: Univariate displays

(The attached PDF file has better formatting.)

Homework assignment: stem and leaf display

A stem and leaf display for assault rates in the fifty U.S. states appears below. The assault rates range from 4.5% to 33.7%.

```
 4 | 568
 5 | 367
 6 |
 7 | 2
 8 | 136
 9 |
10 | 2699
11 | 035
12 | 000
13 |
14 | 59
15 | 1699
16 | 1
17 | 48
18 | 8
19 | 0
20 | 14
21 | 1
22 |
23 | 68
24 | 99
25 | 2459
26 | 3
27 | 69
28 | 5
29 | 4
30 | 0
31 |
32 |
33 | 57
```

A. What is the median assault rate? There are 50 states, so average two points.
B. What is the lower hinge (the 25^{th} percentile)?
C. What is the upper hinge (the 75^{th} percentile)?
D. What is the value of $(H_U - \text{Median}) / (\text{Median} - H_L)$, where H_U is the upper hinge and H_L is the lower hinge?

E. This ratio indicates the skewness of the distribution. Is this distribution positively skewed or negatively skewed?

Module 4: Bivariate Displays

(The attached PDF file has better formatting.)

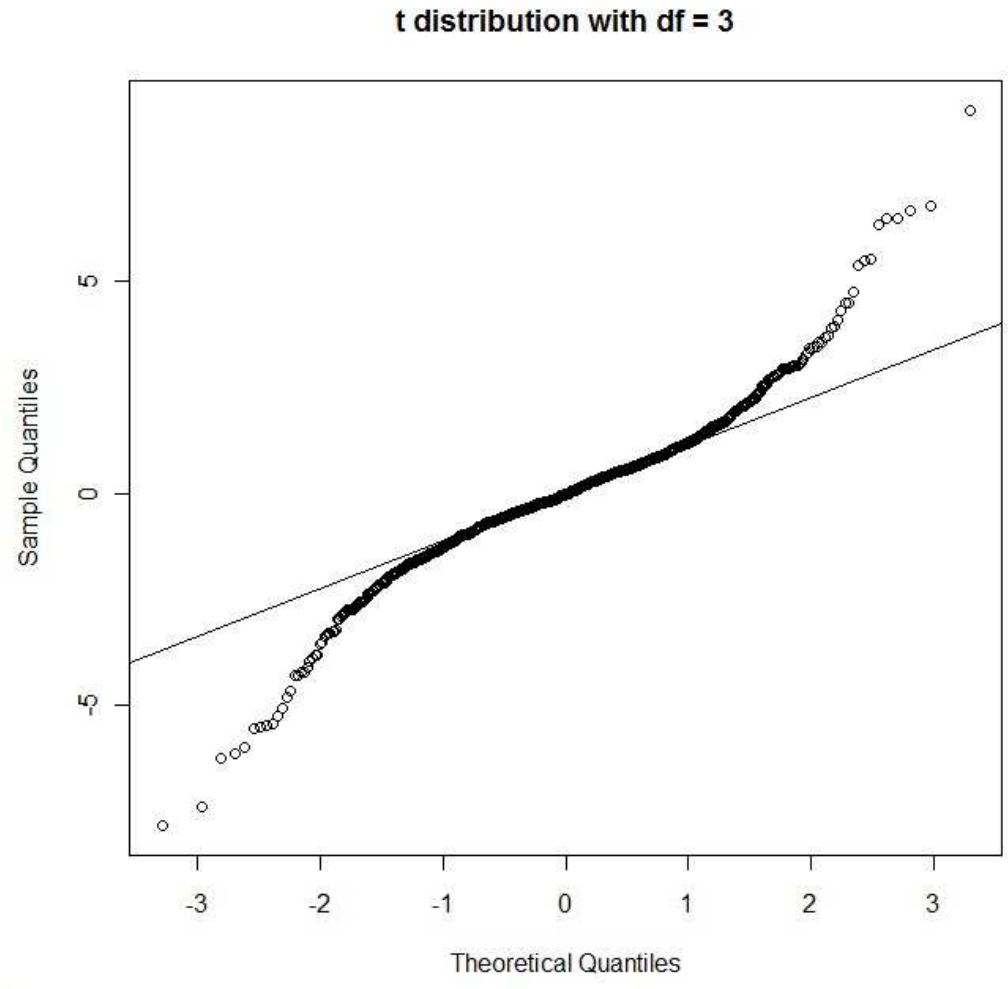
Homework Assignment: quantile comparison plots

Quantile comparison plots are discussed in Module 3 and are used later in the text. This homework assignment discusses quantile comparison plots, not bivariate displays

We compare quantile comparison plots for two distributions:

- Figure 3.9 on page 37: A t-distribution with 3 degrees of freedom.
- Figure 3.8 on page 37: A χ-squared distribution with 2 degrees of freedom.

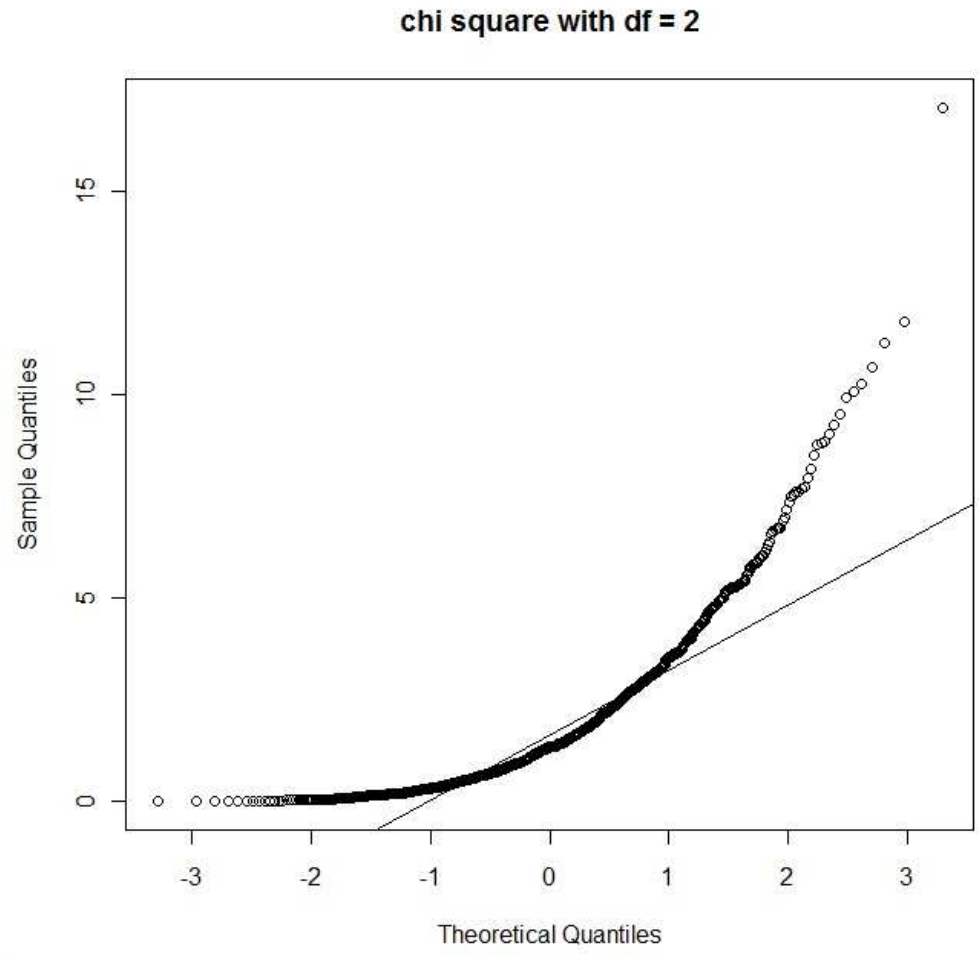
Below is a quantile comparison plot for 1,000 random draws from a *t*-distribution with 3 degrees of freedom.



The quantile comparison plot for a *t*-distribution with 2 degrees of freedom is shaped like an S-curve.

A. At the upper tail, are values more or less extreme than in a normal distribution?
B. At the lower tail, are values more or less extreme than in a normal distribution?
C. Is the *t*-distribution with 2 degrees of freedom (i) symmetric thin-tailed, (ii) symmetric thick-tailed, (iii) positively skewed, or (iv) negatively skewed?

Below is a quantile comparison plot for 1,000 random draws from a χ-squared distribution with 2 degrees of freedom.



The quantile comparison plot for a χ-squared distribution with 2 degrees of freedom is shaped like a convex banana.

A. At the upper tail, are values more or less extreme than in a normal distribution?
B. At the lower tail, are values more or less extreme than in a normal distribution?
C. Is a χ-squared distribution with df = 2 (i) symmetric thin-tailed, (ii) symmetric thick-tailed, (iii) positively skewed, or (iv) negatively skewed?

Module 4: Bivariate Displays

(The attached PDF file has better formatting.)

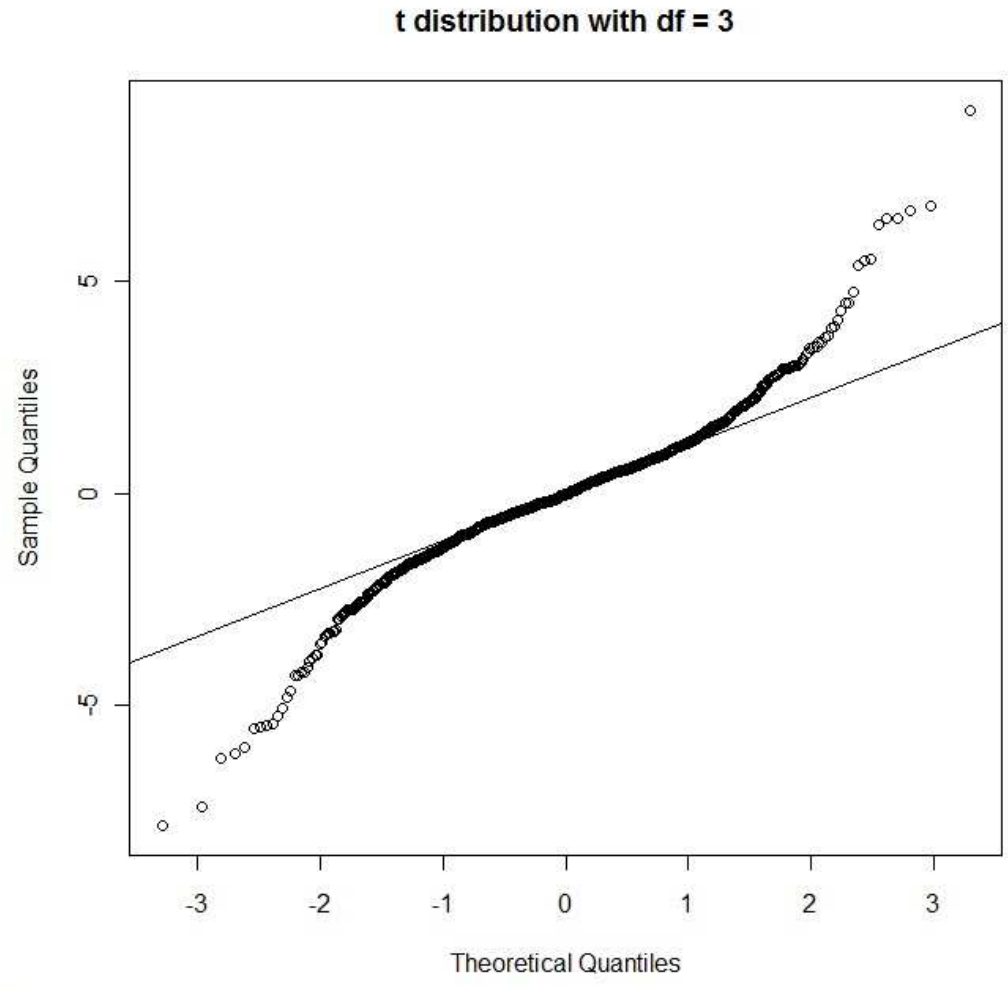
Homework Assignment: quantile comparison plots

Quantile comparison plots are discussed in Module 3 and are used later in the text. This homework assignment discusses quantile comparison plots, not bivariate displays

We compare quantile comparison plots for two distributions:

- Figure 3.9 on page 37: A t-distribution with 3 degrees of freedom.
- Figure 3.8 on page 37: A χ-squared distribution with 2 degrees of freedom.

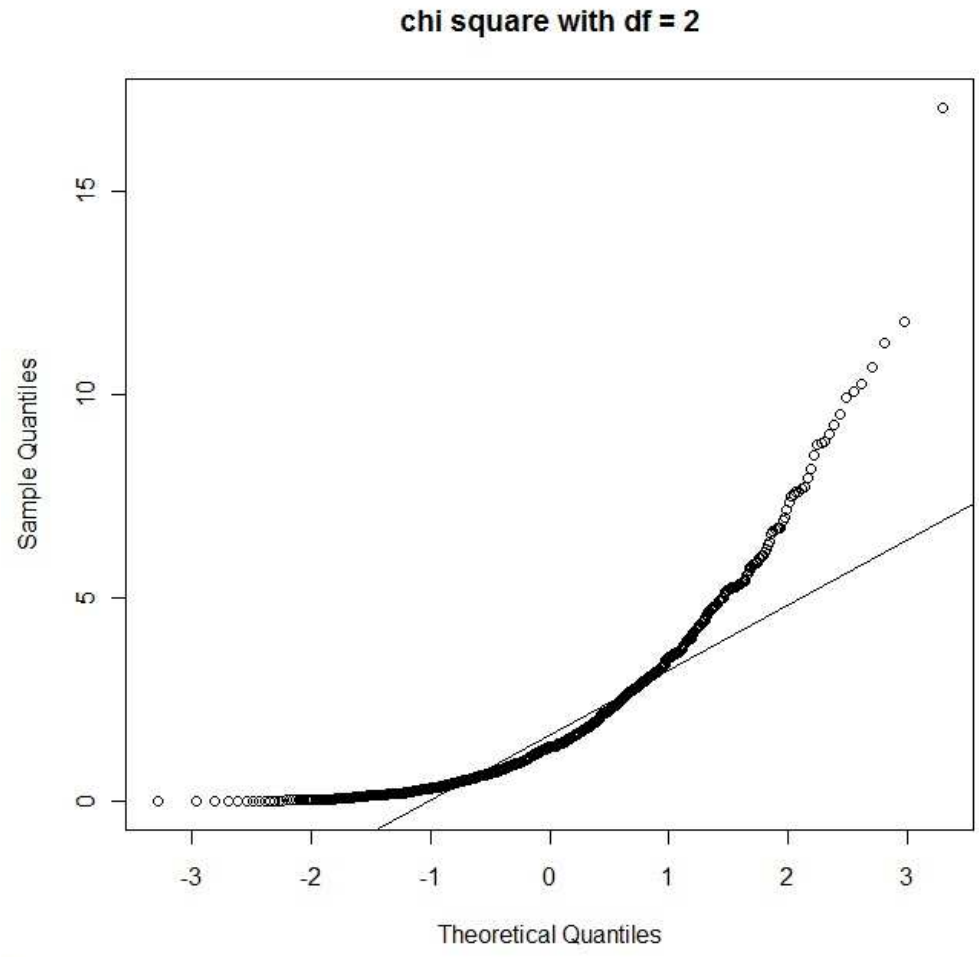
Below is a quantile comparison plot for 1,000 random draws from a *t*-distribution with 3 degrees of freedom.



The quantile comparison plot for a *t*-distribution with 2 degrees of freedom is shaped like an S-curve.

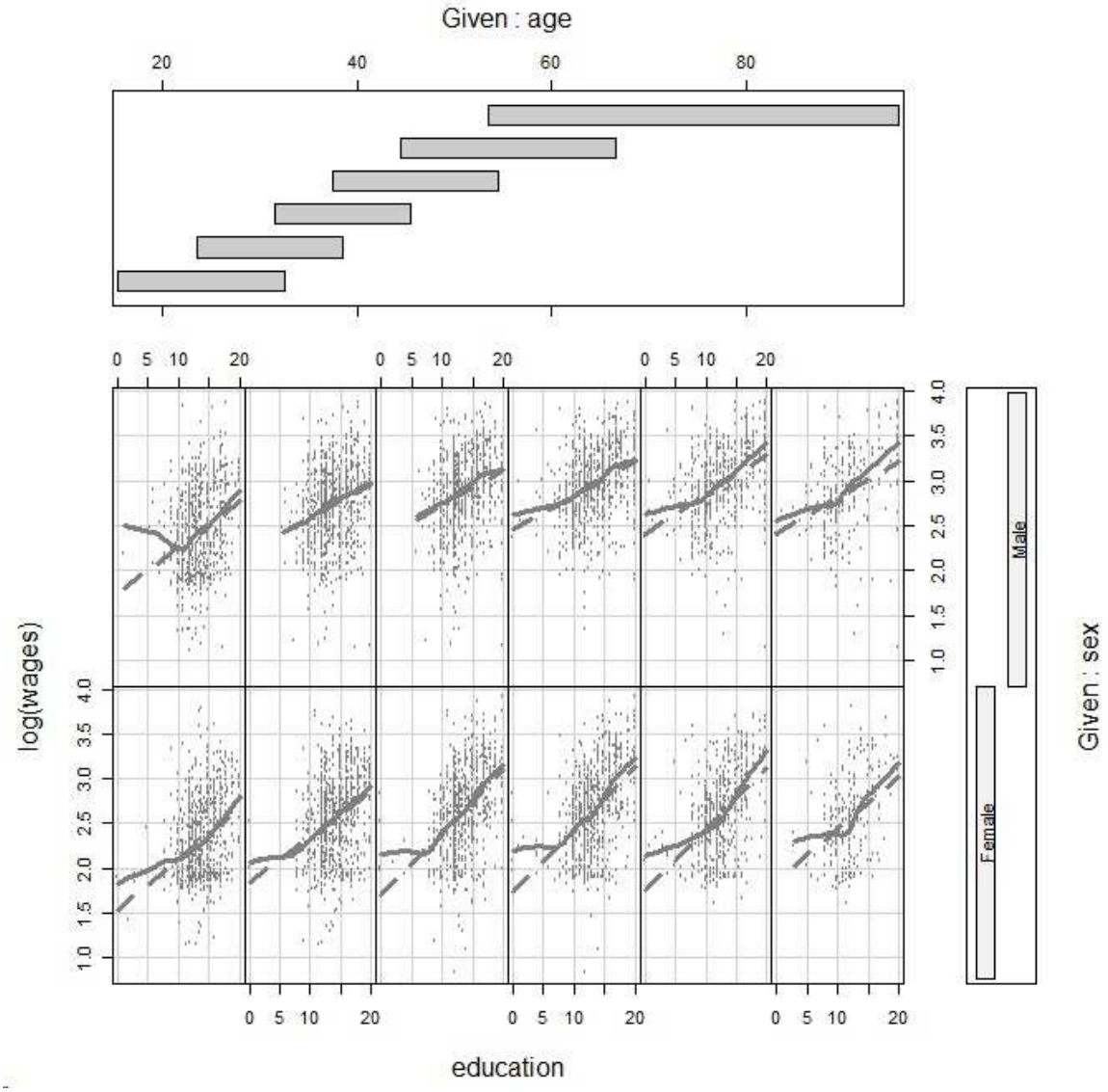
A. At the upper tail, are values more or less extreme than in a normal distribution?
B. At the lower tail, are values more or less extreme than in a normal distribution?
C. Is the *t*-distribution with 2 degrees of freedom (i) symmetric thin-tailed, (ii) symmetric thick-tailed, (iii) positively skewed, or (iv) negatively skewed?

Below is a quantile comparison plot for 1,000 random draws from a χ-squared distribution with 2 degrees of freedom.



The quantile comparison plot for a χ-squared distribution with 2 degrees of freedom is shaped like a convex banana.

A. At the upper tail, are values more or less extreme than in a normal distribution?
B. At the lower tail, are values more or less extreme than in a normal distribution?
C. Is a χ-squared distribution with df = 2 (i) symmetric thin-tailed, (ii) symmetric thick-tailed, (iii) positively skewed, or (iv) negatively skewed?


Given : age
20
40
60
80
0 5 10 20
log(wages)
1.0 1.5 2.0 2.5 3.0 3.5 4.0
education
Given : sex
Female
Male

Module 5: Multivariate displays

(The attached PDF file has better formatting.)

Homework assignment: conditioning plots

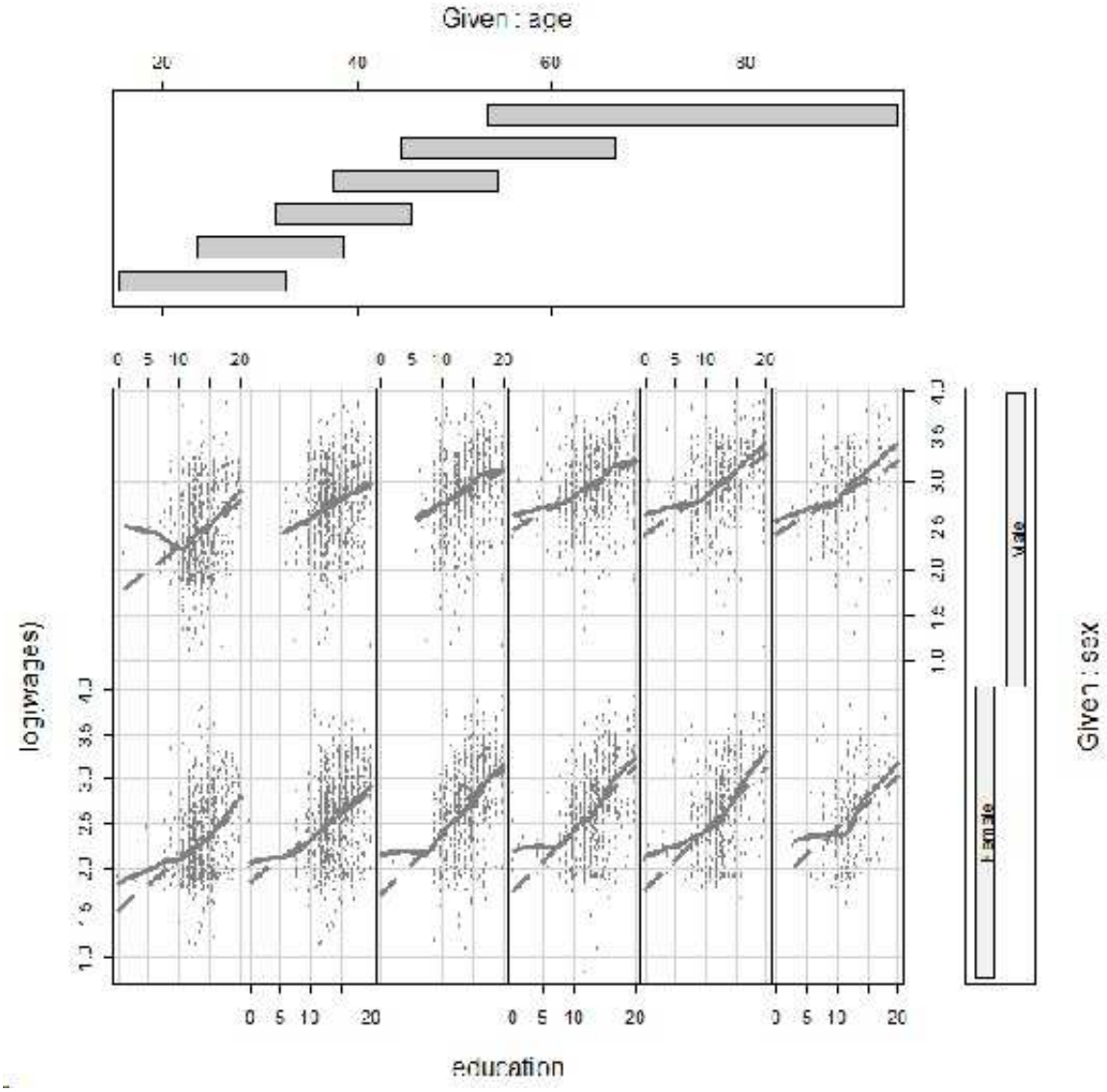
This data set has 7,425 observations and 4 variables: age, sex, education, and wages.

- The dummy variable for sex has values 1 = female and 2 = male.
- The ages range from 16 to 95.
- The education is years of schooling (from 0 to 20)
- The vertical axis is the logarithm of wages (2,300 to 49,920 Canadian dollars).

Each graph has two lines.

- The straight (diagonal) line is a simple regression line.
- The curved line is a lowess curve.

A. What is the dependent variable? What is the independent variable? What are the conditioning variables?
B. How much do the ages overlap? Note that upper bound of one age is the midpoint of the next age and the lower bound of the next age.
C. The graph has 12 panes. Explain what the graph in the top row and second column from the left shows. (A one sentence explanation is sufficient.)
D. The lowess curves match the regression lines at higher education levels but lie above the regression lines at lower education levels. What does this imply?
E. What group has a stronger relation of wages to education level: women age 38 to 54 or women age 23 to 38?
F. What group has a stronger relation of wages to education level: women age 38 to 54 or men age 38 to 54?


Given : age
20
40
60
80
0 5 10 20
logiwages)
education
Given : sex
Female
Male

Module 6: Transforming data

(The attached PDF file has better formatting.)

Homework assignment: choosing a transformation

A random distribution of 10,000 values has the following characteristics:

- Minimum value: 0.0186
- First quartile: 0.4977
- Median: 0.9822
- Mean: 1.6280
- Third quartile: 1.9480
- Maximum value: 45.180

A. Is this distribution symmetric, left-skewed, or right-skewed? Calculate $(H_U\text{-}M)/(M\text{-}H_L)$, (the upper hinge minus the median) divided by (the median minus the lower hinge) to justify your answer.
B. Should you transform the value up or down the ladder of powers and roots to make the distribution symmetric?
C. You are choosing among five transformations to remove the skewness: X^2, $\sqrt{X}$, *ln*(X), $1/\sqrt{X}$, and $1/X$. Which transformation would you choose? Use the table below to justify your answer. You may eliminate some choices are moving the wrong way up or down the ladder of powers and roots.

	H_L	Median	H_U	$(H_U\text{-}M)/(M\text{-}H_L)$
X				
X^2				
$\sqrt{X}$				
ln(X)				
$1/\sqrt{X}$				
$1/X$				

Module 7: Advanced transformations

(The attached PDF file has better formatting.)

Homework assignment: Logit and probit transformations

The logit transformation is tested on the final exam; the probit transformation is not tested. This homework assignment shows their practical equivalence for transforming data.

The textbook says that "once their scales are equated, the logit and probit transformations are, for practical purposes, indistinguishable: logit $\approx (\pi/\sqrt{3}) \times$ probit."

A. Explain the logit and probit transformations. A one sentence explanation is sufficient.
B. Fill in the table below to compare the two transformations.
C. In what range are the two transformations practically equivalent? In what ranges might the two transformations give different results? (The formula for the logit transformation is in the textbook. Excel gives the probit transformation as the inverse of the CDF of the standard normal distribution.)

P	*Logit*	*Probit*	*P*	*Logit*	*Probit*
0.001			0.5		
0.002			0.6		
0.01			0.8		
0.02			0.9		
0.1			0.98		
0.2			0.99		
0.4			0.998		
0.5			0.999		

Module 8: Simple linear regression

(The attached PDF file has better formatting.)

Homework assignment: Estimating regression parameters

Some final exam problems give a set of points and ask to compute ordinary least squares estimators, sums of squares, *t* values, confidence intervals, and other regression statistics. The final exam may give one set of points and ask about several statistics and estimates, or separate points for each statistic. This homework assignment reviews the material you must know to solve the final exam problems. This module has another posting with worked out solutions to a similar problem.

An actuary fits a two-variable regression model $Y_i = \alpha + \beta \times X_i + \varepsilon_i$
to the relation between the explanatory variable X and the response variable Y:

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	66.00%	22.50%	0.00%	0.00%	0.60%	0.0036%	0.0000%
20X2	67.00%	19.50%	1.00%	0.01%	-2.40%	0.0576%	-0.024%
20X3	68.00%	21.00%	2.00%	0.04%	-0.90%	0.0081%	-0.018%
20X4	65.00%	22.50%	-1.00%	0.01%	0.60%	0.0036%	-0.006%
20X5	64.00%	24.00%	-2.00%	0.04%	2.10%	0.0441%	-0.042%
Average	66.00%	21.90%	0.00%	0.02%	0.00%	0.0234%	-0.0180%

The column captions in the table use lower case *x* and *y* for the variables; the deviations are shown explicitly as $(x - \bar{x})$ and $(y - \bar{y})$. The last line has averages, not totals.

A. What is the value of $\hat{\beta}$ (B), the ordinary least squares estimator of β? (P 81)
B. What is the value of $\hat{\alpha}$ (A) , the ordinary least squares estimator of α? (P 81)
C. What is the total sum of squares (TSS)? (P 83-86)
D. What is the regression sum of squares (RegSS)? (P 83-86)
E. What is the residual sum of squares (RSS), or error sum of squares (ESS)? (P 83-86)
F. What is s^2, the estimated variance of the regression? (P 82)
G. What is the value of R^2, the coefficient of determination? (P 83-86)

Show the computations for the homework assignment, not just the solution. You can check your solutions with Excel or other statistical software.

Module 9: Multiple regression

(The attached PDF file has better formatting.)

Homework assignment: Two independent variables

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 least squares estimator of α?
B. What is the least squares estimator of β_1, the coefficient of X_1?
C. What is the least squares estimator of β_2, the coefficient of X_2?

Show the formulas and the computations. You can check your work with Excel or other statistical software.

Module 10: Advanced multiple regression

(The attached PDF file has better formatting.)

Homework assignment: Two correlated independent variables

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
1	1	1.016	6	8	-1.076
2	6	-3.429	7	4	3.461
3	2	0.049	8	9	-2.525
4	7	-3.099	9	5	4.195
5	3	0.359	10	10	-0.746

A. What is the correlation of X_1 and X_2?
B. What is the least squares estimator of α?
C. What is the least squares estimator of β_1, the coefficient of X_1?
D. What is the least squares estimator of β_2, the coefficient of X_2?
E. What is the standard error of the least squares estimator of β_1, the coefficient of X_1?
F. What is the standard error of the least squares estimator of β_2, the coefficient of X_2?

Show the formulas and the computations. You can check your work with Excel or other statistical software.

Module 11: Statistical inference for simple linear regression

(The attached PDF file has better formatting.)

Homework assignment: Estimating regression parameters

Some final exam problems give a set of points and ask to compute ordinary least squares estimators, sums of squares, *t* values, confidence intervals, and other regression statistics. The final exam may give one set of points and ask about several statistics and estimates, or separate points for each statistic.

This homework assignment continues the exercise from Module 8.

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

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	66.00%	22.50%	0.00%	0.00%	0.60%	0.0036%	0.0000%
20X2	67.00%	19.50%	1.00%	0.01%	-2.40%	0.0576%	-0.024%
20X3	68.00%	21.00%	2.00%	0.04%	-0.90%	0.0081%	-0.018%
20X4	65.00%	22.50%	-1.00%	0.01%	0.60%	0.0036%	-0.006%
20X5	64.00%	24.00%	-2.00%	0.04%	2.10%	0.0441%	-0.042%
Average	66.00%	21.90%	0.00%	0.02%	0.00%	0.0234%	-0.0180%

~ The column captions in the table use lower case *x* and *y* for the variables; the deviations are shown explicitly as $(x - \bar{x})$ and $(y - \bar{y})$.
~ The last line has averages, not totals. Some formulas in the textbook use totals.

We regress retrospective premium ratios on reported loss ratios to estimate premium assets for retrospectively rated business. Actuarial issues of retrospectively rated business are not important; the homework deals with the regression analysis only.

The Module 8 homework assignment solves for the least squares estimators of α and β.

A. What is the *variance* of the ordinary least squares estimator of β?
B. What is the *t* statistic for testing the null hypothesis that $\beta = 0$?
C. What is the 95% confidence interval for the true value of β?
D. What is the *p* value for testing the null hypothesis that $\beta = 0$?
E. What is the *variance* of the ordinary least squares estimator of α?
F. What is the *t* statistic for testing the null hypothesis that $\alpha = 21.90\%$?
G. What is the 95% confidence interval for the true value of α?

H. What is the *p* value for testing the null hypothesis that $\alpha = 21.90\%$?

Show the computations for the homework assignment, not just the solution. You can check your solutions with Excel or other statistical software.

The null hypotheses for α and β depend on the scenario.

- A null hypothesis of $\beta = 0$ means that retrospective rating has no effect on the premium.
- A null hypothesis of $\beta = 1$ means that a dollar of loss causes a dollar of premium.

The null hypothesis for the value of α depends on the situation.

- A null hypothesis of $\alpha = 0$ means that retrospective rating is the standard premium.
  - This hypothesis is not realistic, since large insureds get premium discounts.
- A null hypothesis of $\alpha = 21.90\%$ means retrospective rating has no effect on premium.
  - The average observed discount is the expected discount regardless of losses.

These retrospective rating issues are not part of the regression analysis course.

The confidence intervals require a *t* distribution for the appropriate degrees of freedom. Use Excel to find the *t* values for a 95% confidence interval.

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.

Fox Module 13: Dummy variable regression HW

(The attached PDF file has better formatting.)

Homework assignment: auto insurance rating territories

An insurer examines claim frequencies for 15 territories: 5 urban, 5 suburban, and 5 rural.

Urban		*Sub-urban*		*Rural*	
Territory	*Claim Frequency*	*Territory*	*Claim Frequency*	*Territory*	*Claim Frequency*
1	14.30%	6	10.88%	11	9.59%
2	11.00%	7	16.58%	12	8.54%
3	20.90%	8	12.72%	13	10.32%
4	16.85%	9	9.40%	14	10.01%
5	12.85%	10	12.92%	15	9.24%

A. How many dummy variables does this regression use?
B. What are the values of the dummy variables for urban, sub-urban, and rural? Assume rural is the base territory, with dummy variables equal to zero.
C. Use Excel or other statistical software to run the regression. What are the values of α, β_1, and β_2? (Fox uses γ_1 and γ_2 instead of β_1 and β_2 for dummy variables.) Explain what each of the coefficients means.

Jacob: Fox uses both γ_1 and γ_2 as well as β_1 and β_2.

Rachel: We do this is the territory number has a quantitative value. But the territory numbers here are just indicators; they have no quantitative meaning. The regression equation is

$$\text{Frequency} = \alpha + \beta_1 \times D_1 + \beta_2 \times D_2$$

In Fox's notation, this is $\text{Frequency} = \alpha + \gamma_1 \times D_1 + \gamma_2 \times D_2$

Module 14: Modeling interactions

(The attached PDF file has better formatting.)

Homework assignment: rating territories and mileage

This homework assignment continues the exercise in Module 13.

An insurer examines claim frequencies for 15 territories: 5 urban, 5 suburban, and 5 rural. The insurer also has the average miles driven per car in each territory (in thousands). If you live outside the U.S., replace mileage with thousands of kilometers.

	Urban			*Sub-urban*			*Rural*	
Territory	*Mileage*	*Claim Freq'y*	*Territory*	*Mileage*	*Claim Freq'y*	*Territory*	*Mileage*	*Claim Freq'y*
1	5	8.45%	6	20	6.99%	11	10	3.83%
2	10	10.90%	7	40	12.94%	12	20	5.06%
3	15	13.45%	8	60	19.01%	13	30	6.00%
4	20	16.04%	9	80	25.06%	14	40	6.94%
5	25	18.49%	10	100	31.11%	15	50	7.92%

A. How many dummy variables does this regression use?
B. What are the values of the dummy variables for urban, sub-urban, and rural? Assume rural is the base territory, with dummy variables equal to zero.
C. Write the regression equation with all interactions. You should have six terms.
D. Use Excel or other statistical software to run the regression. What are the values of the six regression parameters?

The claim frequencies are chosen so that the standard error of the regression is small. The observed values are very close to the fitted values, so you can tell if your solution is right.

Fox Module 15: Advanced interactions

(The attached PDF file has better formatting.)

Homework assignment: F test with interactions

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

The variables mean: 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

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	

For each model,

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

A. Why does the total sum of squares (TSS) not depend on the model? What is the TSS in this illustration?
B. Which model has the smallest residual sum of squares (RSS)? How do we know this even without computing any figures?
C. How do we test the significance of income? What is the null hypothesis? How the F-ratio is computed? (Show the calculations.)
D. How do we test the significance of education × type? What is the null hypothesis? How the F-ratio is computed? (Show the calculations.)

The following comments may help you understand the exhibits:

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. But this illustration focuses on 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.

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.

Show the calculation of the F-ratio for Parts C and D.

Fox Module 16: One way ANOVA

(The attached PDF file has better formatting.)

Homework assignment: insurance renewals

(This homework assignment is a simpler version of the previous assignment, which was too difficult. This homework assignment is not needed for the course ending January 2011; only 17 homework assignments are needed.)

An insurer examines policy renewals by territory.

Urban		*Sub-urban*		*Rural*	
Territory	*Renewal Rate*	*Territory*	*Renewal Rate*	*Territory*	*Renewal Rate*
1	65.00%	6	80.00%	13	89.00%
2	80.00%	7	75.00%	14	90.00%
3	75.00%	8	90.00%	15	91.00%
4	75.00%	9	90.00%		
5	80.00%	10	90.00%		
		11	82.00%		
		12	88.00%		

Use a one way analysis of variance to determine if renewal rates differ by territory. *Renewal rates are percentages, so use log odds (logits), not the observed renewal rate.*

A. Why would a linear regression of renewal rates on territory not be appropriate?
B. What are the log odds by territory?
C. Why might a linear regression of renewal rates on territory be appropriate?

Part A: Renewal rates are percentages; the observed values have a binomial distribution. If a territory has 1,000 drivers, and the mean renewal rate is P, the variance of the renewal rate is $P \times (1 - P) / 1{,}000$. Classical regression analysis assumes the response variable has a normal distribution with a constant variance in all territories. What is the range of a normal distribution? What is the range of a binomial distribution? If two binomial distributions have different expected values but the same N, can they have the same variance?

Part B: The log odds are $ln(\pi / (1 - \pi))$. Calculate the log odds by territory.

Part C: What is the range of the log odds? As $\pi \rightarrow 1$, what are the log odds? As $\pi \rightarrow 0$, what are the log odds?

Fox shows how to regress renewal rates on explanatory variables in chapter 14. This homework assignment converts renewal rates to log odds, which is the first step in the procedure.

Classical regression analysis is still not ideal, since the variance in rural territories is greater than the variance in urban territories. In later modules, we show better ways of doing the statistical analysis (GLMs).

Fox Module 17: Unusual and influential data

(The attached PDF file has better formatting.)

Homework assignment: hat values

- A statistician regresses the points Y = (5, 4, 3, 2, 1) on X = (1, 2, 3, 4, 15).
- The five points are (1,5), (2,4), (3, 3,), (4, 2), and (15,1).

A. What is the mean hat value for these five points?
B. What is the range of hat values for these five points?
C. What are the hat values of the five points?

Fox Module 18: Outliers and influence, advanced

(The attached PDF file has better formatting.)

Homework assignment: studentized residuals

A data sample has five points.

explanatory variable	1	2	3	4	5
response variable	1	3	2	3	0

A. Regress the response variable on the explanatory variable. What is the residual at the last point (5,0)?
B. Regress the first four response variables on the first four explanatory variables. What is the studentized residual at the last point (5,0)?
C. Explain intuitively why the residual is correlated with the response variable.
D. Is the studentized residual correlated with the response variable? Why or why not?

Fox Module 19: Heteroscedasticity

(The attached PDF file has better formatting.)

Homework assignment: residual plots and heteroscedasticity

This homework assignment simulates heteroscedastic data and forms residual plots. Use Excel of other statistical software.

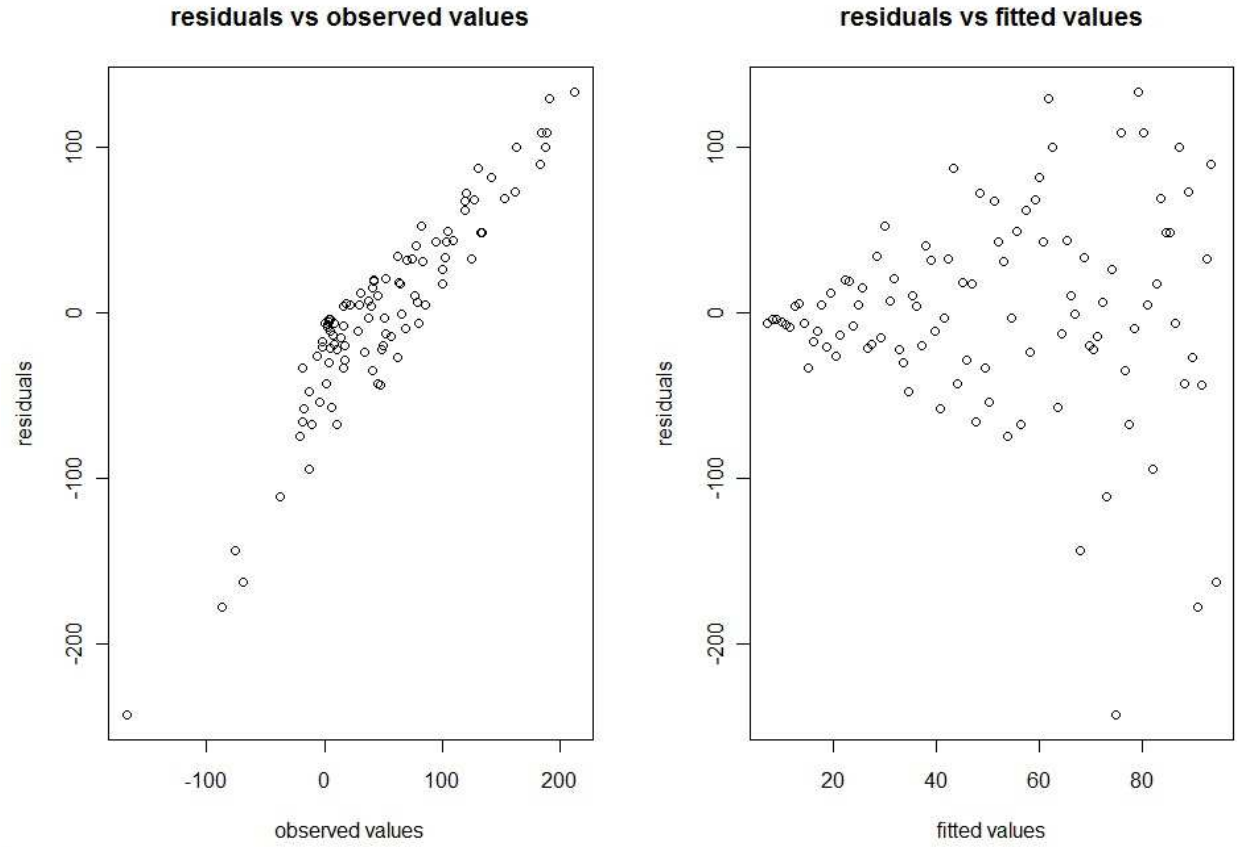
You must use Excel or other statistical software for the student project. Learn the statistical tools in your software. Simulations are the quickest way to learn the statistical concepts.

A. Choose a sample for the explanatory variables. The exhibits below use 1 to 100.
B. Simulate response variables whose observed values are heteroscedastic. The exhibits below use $Y_j = X_j + X_j \times \Phi(0,1)$: the variance of the error term is proportional to the value of the X_j.
C. Form the regression equation and derive the fitted values at each point.
D. Plot the residuals against the observed values (left pane below).
E. Pot the residuals against the fitted values (right pane below).
F. Explain why the two plots are so different.

We show sample exhibits below. Your exhibits will look different, since you simulate points.

If your variance is too small, you won't see the effect that the textbook discusses.
We should use $Y_j = X_j + \hat{y}_j \times \Phi(0,1)$: the variance of the error term is proportional to the fitted value of Y_j. This is harder to simulate, so we use the simpler formula above.

Form the exhibits in Excel, SAS, R, or any other software. Sample exhibits are below.




```
xvv <- 1:100
yvv <- xvv + xvv * rnorm(100,0,1)
lm.hsce <- lm(yvv ~ xvv)
rsds <- residuals(lm.hsce)
fits <- fitted(lm.hsce)
par(mfcol=c(1,2))
plot(yvv, rsds, xlab="observed values", ylab = "residuals", main="residuals vs observed values")
plot(fits, rsds, xlab="fitted values", ylab = "residuals", main="residuals vs fitted values")
```

Fox Module 20: Collinearity

(The attached PDF file has better formatting.)

Homework assignment: Interest rates and inflation

Financial economists often model nominal interest rates as expected inflation plus the real interest rate plus an error term.

- Real interest rates are relatively steady over time.
- Expected inflation is often modeled as actual inflation in the previous period.
- The error term is both random fluctuation and changes in monetary policy.

A statistician has monthly data for interest rates, wage inflation, and general inflation.

A. What is the estimated β for a regression of interest rates on general inflation?
B. What is the standard error of the β coefficient?
C. What is the estimated β for a regression of interest rates on wage inflation?
D. What is the standard error of the β coefficient?
E. What are the estimated β′s for a regression of interest rates on both inflation rates?
F. What is the correlation of wage inflation with general inflation?
G. What are the standard errors of the β coefficients?

Month	*Wage Inflation*	*General Inflation*	*Interest Rates*	*Month*	*Wage Inflation*	*General Inflation*	*Interest Rates*
January	6.06%	5.12%	7.85%	July	6.48%	5.70%	8.39%
February	6.18%	5.20%	8.39%	August	6.76%	5.60%	8.84%
March	6.17%	5.14%	8.17%	September	6.74%	5.73%	9.11%
April	6.26%	5.35%	8.20%	October	7.00%	5.97%	9.04%
May	6.40%	5.36%	8.25%	November	6.95%	6.00%	8.95%
June	6.42%	5.51%	8.25%	December	7.28%	6.14%	9.60%

Fox Module 21: Generalized linear models, concepts

(The attached PDF file has better formatting.)

Homework assignment: maximum likelihood estimation: exponential decay

[Note: The homework assignment is at the bottom of this posting, after the explanation of the method.]

Health claims occurring in month *0* and settled in month *j* are a percentage P of the claims open at the end of month *j-1*.

- No claims are settled in the month they occur.
- P% of claims open at the end of a month are expected to settle in the next month.
- Actual claims settlements are distorted by random fluctuation.

A reserving actuary estimates the percentage P.

To simplify the mathematics, we assume that if 100 claims occur in December 20X1,

- 100 × P claims close in January,
- 100 × (1 – P) × P claims close in February, and so forth.

This is not exact, since if more claims close in January, fewer claims are left open, and fewer claims close in later months. But it is roughly correct, and it gives a simple solution.

For claims occurring in December 20X1, the number of claims closed by month in 20X2 are

Jan	*Feb*	*Mar*	*Apr*	*May*	*Jun*	*Jul*	*Aug*	*Sep*	*Oct*	*Nov*	*Dec*	*>>*
29	16	17	4	3	9	4	5	1	1	1	3	7

- 100 claims occur in December 20X1 (the sum of the claims closed and still open).
- Jan, Feb, Mar, … = January, February, March, …
- >> = still open at December 31, 20X2.

We compute the maximum likelihood estimate for the percentage P.

A. Suppose there were no random fluctuation. Using the January figures and the 100 claims in December, what is the percentage P?

Part A: Were there no random fluctuation, the percentage P would be constant.

- The claims occurring in December 20X1 is 100 = 29 + 16 + 17 + … + 7.
- The percentage of claims closed in January is P × 100 = 29 ⇛ P = 29%.

This estimator is not optimal. It ignores the claims closed in later months, so it doesn't use all available information.

This estimator is unbiased, but it is not optimal.

- 71 claims are open at Jan 30, of which 16 close in February ⇛ P = 16 / 71 = 22.54%.
- 55 claims are open at Feb 28, of which 17 close in March ⇛ P = 17 / 55 = 30.91%.

We use a maximum likelihood estimator. For claims occurring December 20X1:

- The probability of closing in January 20X8 is P.
- The probability of closing in February 20X8 is (1–P)(P).
- The probability of closing in March 20X8 is $(1–P)^2(P)$.
- ….
- The probability of being open at December 31, 20X8 is $(1–P)^{12}$.

The probability of observing the actual figures is

$$C \times P^{29} \times [(1–P)(P)]^{16} \times [(1–P)^2(P)]^{17} \times [(1–P)^{12}]^7 \times \ldots = C \times P^{93} \times (1–P)^{322}$$

C is the combinatorial constant: the number of ways to arrange 100 claims so that 29 are in January, 16 are in February, 17 are in March, and so forth.

Illustration: Of four claims, 2 close in January, 1 closes in February, and 1 closes in March.

We label the claims A, B, C, and D. The possible combinations are

- A and B close in January; C closes in February; and D closes in March.
- A and B close in January; D closes in February; and C closes in March.
- …

The illustration has 12 possibilities: The claim closing in March has 4 possibilities, for each of which there are three possibilities for the claim closing in February. Computing this constant is more difficult with 100 claims and 12 months. But this constant doesn't affect the maximum likelihood estimation, so we can ignore it.

Homework assignment

To maximize the likelihood, set the derivative of the likelihood (or the loglikelihood) with respect to P equal to zero. Solve for P.

[If you have difficulty, ask a question on the discussion forum.]

Fox Module 22: Generalized linear models, discrete and continuous data

(The attached PDF file has better formatting.)

Homework assignment: Education and Auto Accidents

The homework assignment follows the discussion forum reading for this module.

We fit a linear model to three groups of drivers:

Exposures	*Years of Schooling*	*Auto Accidents per 100 Drivers*
1,000	8	15
1,000	12	8
1,000	16	3

- The X value is the years of schooling.
- The Y value is the number of auto accidents per 100 drivers.

The table shows that drivers with

- 8 years of schooling (elementary school) have claim frequencies of 15%.
- 12 years of schooling (high school) have claim frequencies of 8%.
- 16 years of schooling (college) have claim frequencies of 3%.

We compare GLMs with different distributions of the error term.

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

Assume each year of schooling has the *same linear effect* on claim frequency.

- We fit a straight line to the three points.
- The variance of the error term depends on the GLM.

A. Which model gives the higher claim frequency for drivers with eight years of schooling?
B. Which model gives the higher claim frequency for college educated drivers?
C. Why might a linear model not be proper for these data? How does decreasing marginal utility affects the slopes? If a driver with 9 years of schooling has an expected claim frequency 1 percentage point less than a driver with 8 years of schooling, should the difference from 12 to 13 years of schooling be more or less than 1 percentage point?
D. How do actuaries treat class dimensions like years of schooling? Do actuaries treat this as a quantitative or qualitative class dimension?

Fox Module 23 Generalized linear models probabilities HW

(The attached PDF file has better formatting.)

Homework assignment: Renewal Rates and Years Insured

Policy renewal rates increase the longer the policyholder has been insured.

- New policyholders often switch to another insurer at the end of the policy term.
- Policyholders who have stayed for ten years are very likely to stay another year.

The type of non-renewal affects the statistical modeling. At older ages, some policyholders do not renew because they die or they no longer have any exposure. Auto policyholders may give up driving and Homeowners policyholders may move to a retirement home.

- For pricing, it may not matter why the policyholder fails to renew. The insurer loses the value of future business whether the policyholder is alive or dead, driving or not driving, or still living in the home.
- For statistical modeling, the type of non-renewal affects the relation. The renewal rate for auto insurance rises with each renewal. As the number of years insured increases beyond the expected driving life of the policyholder, the renewal ratio may decrease. If the renewal rate increases and then decreases, a logit GLM may not work well.

The logit GLM in this exercise uses the percentage of renewals with the existing insurer vs all insurers. The data are from a telephone questionnaire with policyholders who did not renew. One question was whether the policyholder has coverage with another insurer or does not have insurance. The renewal rate is the number of policyholders who renew divided by the number who still have a policy with any insurer.

We model renewal rates as a function of years insured using a logit link function.

We examine *six month* policies that come up for renewal in 20X1. Each record shows

- The years already insured at the renewal data, ranging from 0.5 to 30.
  - 0.5 years means the policy is at its first renewal.
  - 30 years means the policy is at its 60^{th} renewal.
- Whether the insured renews the policy: True = renews and False = does not renew

From the individual records, we form an aggregate data base with 60 records. Each record has three fields.

- The policy age, ranging from 0.5 to 30 (1 half year to 60 half years).
- The exposures at that policy age, from 10,000 at 0.5 years to *** at 30 years.
- The percentage of policies that renew.

We relate the renewal rate to years insured. We examine regressions of the renewal rate to years insured and of the log odds of the renewal rate to years insured.

The exposures indicate the quality of the empirical data at each renewal date.

- At 0.5 years insured, the data are highly credible (many exposures).
- At 30 years, the data are less credible (few exposures).

GLMs use *weighted* regressions. The weights depend on the exposures and the conditional distribution function.

This homework assignment does not require you to fit the GLM. Fitting the GLM is hard by pencil and paper, but it requires only a single function in R: glm(renewals ~ years, …).

The homework assignment asks which points are more influential. The answer depends on the exposures at each point and the variance of the distribution at each point.

The data are in an Excel spread-sheet and data files. Use the format that is convenient. You can complete the homework using Excel and its *REGRESSION* add-in. If you want to fit the GLM, use the data files and R.

A. Graph the renewal rate as a function of years insured. Is the curve convex or concave? [You can answer this intuitively, since the renewal rate is bounded by 100%.]
B. Form a regression line linking renewal rates to years insured. What are the least squares estimates for α and β? [Ignore the exposures for this part.]
C. Does the regression line over-estimate or under-estimate the renewal rate for (i) 1 to 2 years-insured, (ii) 29 to 30 years-insured, (iii) 14.5 to 15.5 years-insured? [You can answer by comparing the observed vs fitted values or by comparing the curve with the regression line.]
D. The first point looks like an outlier. It is an influential point, so it skews the regression line. If we exclude this point from the regression, are α and β higher or lower? Which regression line has the higher R^2? Which regression line has the lower estimated σ^2? You can answer all these questions intuitively. If you want, check your work with Excel.
E. A simple regression gives the same weight to each point. Based on the exposures, which renewal rates have more random fluctuation: high or low years insured? Should we give more weight to high or low years insured in fitting the regression line? You don't have to do a weighted regression for the homework assignment.
F. The GLM uses link functions, distributions, and exposures. Form logits of the renewal rates. (The logits are the log odds.) Graph these logits as a function of years insured. What is the shape of this curve: convex, concave, or straight? Are there any outliers?
G. Regress the logit of the renewal rate on the years insured. Solve for α and β.

Form the graphs with Excel, R, or other software. *You don't have to submit the graphs with your homework assignment.* The graphs help you visualize the data and spot outliers. The Excel spreadsheet attached to this homework assignment has the graphs. Your graphs should look similar (or the same).

You should always graph data for a statistical study, such as the student project. Excel, R, and most statistical packages have excellent graphing tools.

Jacob: In this homework assignment, the logit transformation creates a linear curve. Is this generally true in real applications? What if the logit does not form a linear curve?

Rachel: The logit transformation of probabilities often creates linear curves, but not always. For some skewed distributions, we might us *complementary log log* transformations. If the distribution is symmetric but the thickness of the tails doesn't fit the logit distribution, we might try a probit distribution. We use whatever transformation creates a linear relation. The logit is a simple transformation, and it works well in many scenarios.

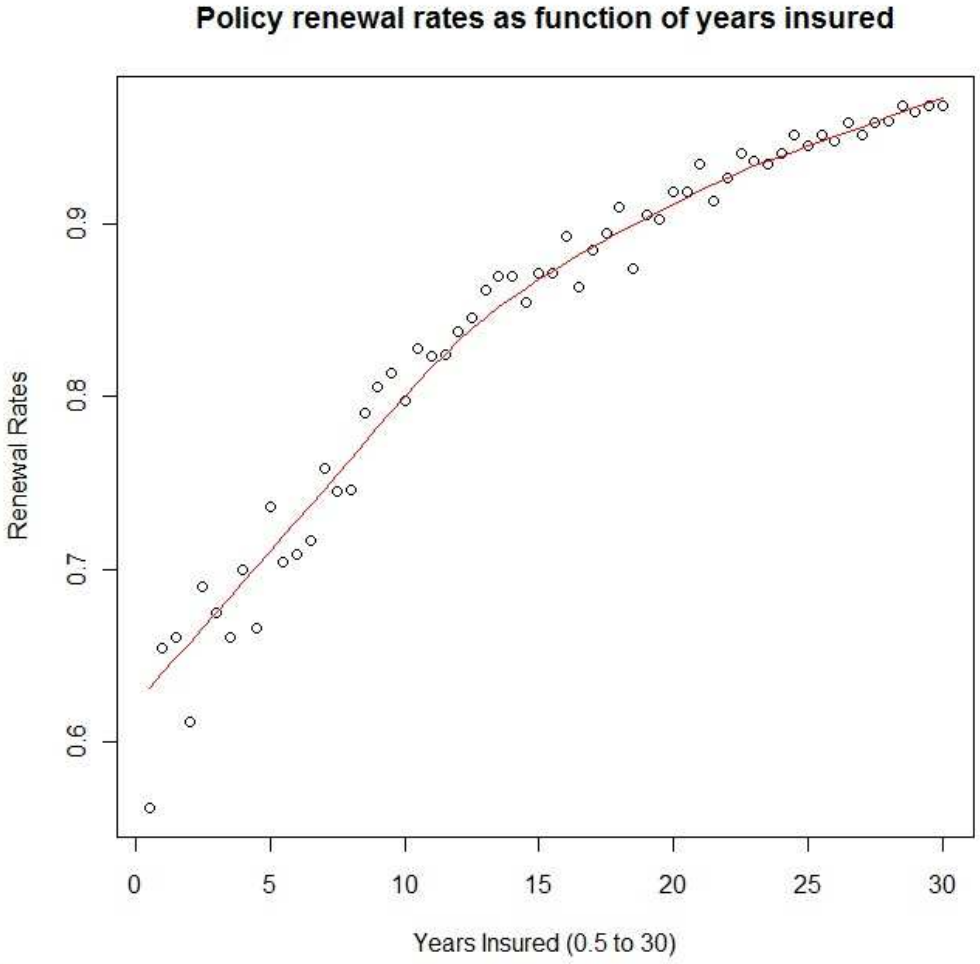
Jacob: How does the binomial distribution affect the GLM? We didn't use the distribution in this homework assignment.

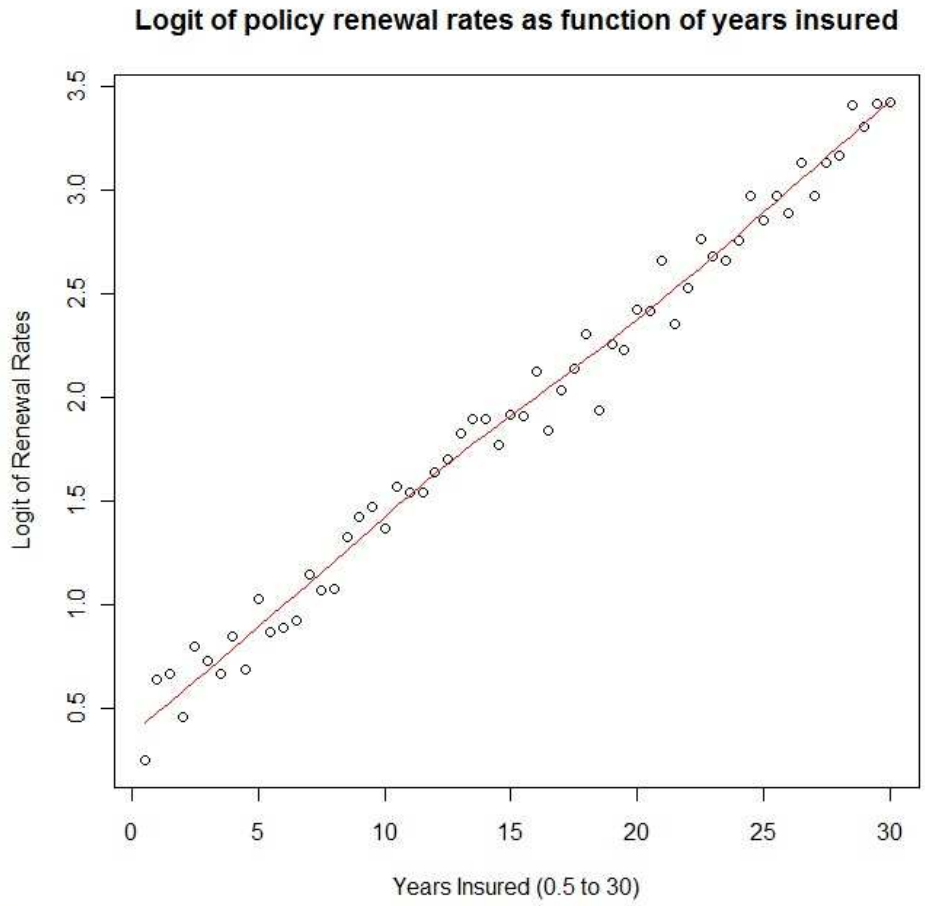
Rachel: The distribution is important, but it is harder to grasp. The homework assignment for the previous module deals with Poisson and Gamma distributions, and the same logic applies to binomial distributions. The variance of a binomial distribution is highest at the center and lowest in the tails.

Jacob: The GLM predicts the logits of the renewal rate. How do we get the renewal rate?

Rachel: Given the logit, the renewal rates is $1/(1 + e^{-\text{logit}}) = e^{\text{logit}} /(1 + e^{\text{logit}})$.

{The graphs below are what you should find.}


Policy renewal rates as function of years insured
Renewal Rates
0.6
0.7
0.8
0.9
0
5
10
15
20
25
30
Years Insured (0.5 to 30)


Logit of policy renewal rates as function of years insured
Logit of Renewal Rates
0.5
1.0
1.5
2.0
2.5
3.0
3.5
0
5
10
15
20
25
30
Years Insured (0.5 to 30)

Fox Module 23 Generalized linear models probabilities HW

(The attached PDF file has better formatting.)

Homework assignment: Renewal Rates and Years Insured

Policy renewal rates increase the longer the policyholder has been insured.

- New policyholders often switch to another insurer at the end of the policy term.
- Policyholders who have stayed for ten years are very likely to stay another year.

The type of non-renewal affects the statistical modeling. At older ages, some policyholders do not renew because they die or they no longer have any exposure. Auto policyholders may give up driving and Homeowners policyholders may move to a retirement home.

- For pricing, it may not matter why the policyholder fails to renew. The insurer loses the value of future business whether the policyholder is alive or dead, driving or not driving, or still living in the home.
- For statistical modeling, the type of non-renewal affects the relation. The renewal rate for auto insurance rises with each renewal. As the number of years insured increases beyond the expected driving life of the policyholder, the renewal ratio may decrease. If the renewal rate increases and then decreases, a logit GLM may not work well.

The logit GLM in this exercise uses the percentage of renewals with the existing insurer vs all insurers. The data are from a telephone questionnaire with policyholders who did not renew. One question was whether the policyholder has coverage with another insurer or does not have insurance. The renewal rate is the number of policyholders who renew divided by the number who still have a policy with any insurer.

We model renewal rates as a function of years insured using a logit link function.

We examine *six month* policies that come up for renewal in 20X1. Each record shows

- The years already insured at the renewal data, ranging from 0.5 to 30.
  - 0.5 years means the policy is at its first renewal.
  - 30 years means the policy is at its 60^{th} renewal.
- Whether the insured renews the policy: True = renews and False = does not renew

From the individual records, we form an aggregate data base with 60 records. Each record has three fields.

- The policy age, ranging from 0.5 to 30 (1 half year to 60 half years).
- The exposures at that policy age, from 10,000 at 0.5 years to *** at 30 years.
- The percentage of policies that renew.

We relate the renewal rate to years insured. We examine regressions of the renewal rate to years insured and of the log odds of the renewal rate to years insured.

The exposures indicate the quality of the empirical data at each renewal date.

- At 0.5 years insured, the data are highly credible (many exposures).
- At 30 years, the data are less credible (few exposures).

GLMs use *weighted* regressions. The weights depend on the exposures and the conditional distribution function.

This homework assignment does not require you to fit the GLM. Fitting the GLM is hard by pencil and paper, but it requires only a single function in R: glm(renewals ~ years, …).

The homework assignment asks which points are more influential. The answer depends on the exposures at each point and the variance of the distribution at each point.

The data are in an Excel spread-sheet and data files. Use the format that is convenient. You can complete the homework using Excel and its *REGRESSION* add-in. If you want to fit the GLM, use the data files and R.

A. Graph the renewal rate as a function of years insured. Is the curve convex or concave? [You can answer this intuitively, since the renewal rate is bounded by 100%.]
B. Form a regression line linking renewal rates to years insured. What are the least squares estimates for α and β? [Ignore the exposures for this part.]
C. Does the regression line over-estimate or under-estimate the renewal rate for (i) 1 to 2 years-insured, (ii) 29 to 30 years-insured, (iii) 14.5 to 15.5 years-insured? [You can answer by comparing the observed vs fitted values or by comparing the curve with the regression line.]
D. The first point looks like an outlier. It is an influential point, so it skews the regression line. If we exclude this point from the regression, are α and β higher or lower? Which regression line has the higher R^2? Which regression line has the lower estimated σ^2? You can answer all these questions intuitively. If you want, check your work with Excel.
E. A simple regression gives the same weight to each point. Based on the exposures, which renewal rates have more random fluctuation: high or low years insured? Should we give more weight to high or low years insured in fitting the regression line? You don't have to do a weighted regression for the homework assignment.
F. The GLM uses link functions, distributions, and exposures. Form logits of the renewal rates. (The logits are the log odds.) Graph these logits as a function of years insured. What is the shape of this curve: convex, concave, or straight? Are there any outliers?
G. Regress the logit of the renewal rate on the years insured. Solve for α and β.

Form the graphs with Excel, R, or other software. *You don't have to submit the graphs with your homework assignment.* The graphs help you visualize the data and spot outliers. The Excel spreadsheet attached to this homework assignment has the graphs. Your graphs should look similar (or the same).

You should always graph data for a statistical study, such as the student project. Excel, R, and most statistical packages have excellent graphing tools.

Jacob: In this homework assignment, the logit transformation creates a linear curve. Is this generally true in real applications? What if the logit does not form a linear curve?

Rachel: The logit transformation of probabilities often creates linear curves, but not always. For some skewed distributions, we might us *complementary log log* transformations. If the distribution is symmetric but the thickness of the tails doesn't fit the logit distribution, we might try a probit distribution. We use whatever transformation creates a linear relation. The logit is a simple transformation, and it works well in many scenarios.

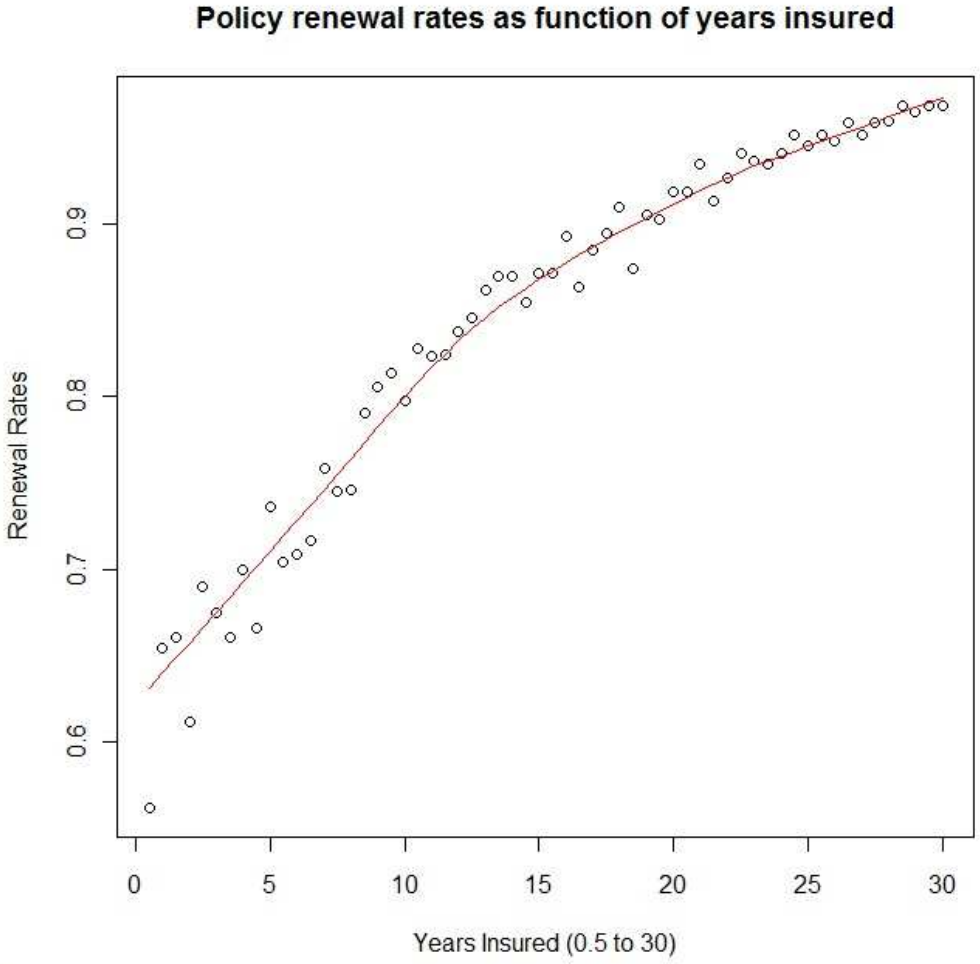
Jacob: How does the binomial distribution affect the GLM? We didn't use the distribution in this homework assignment.

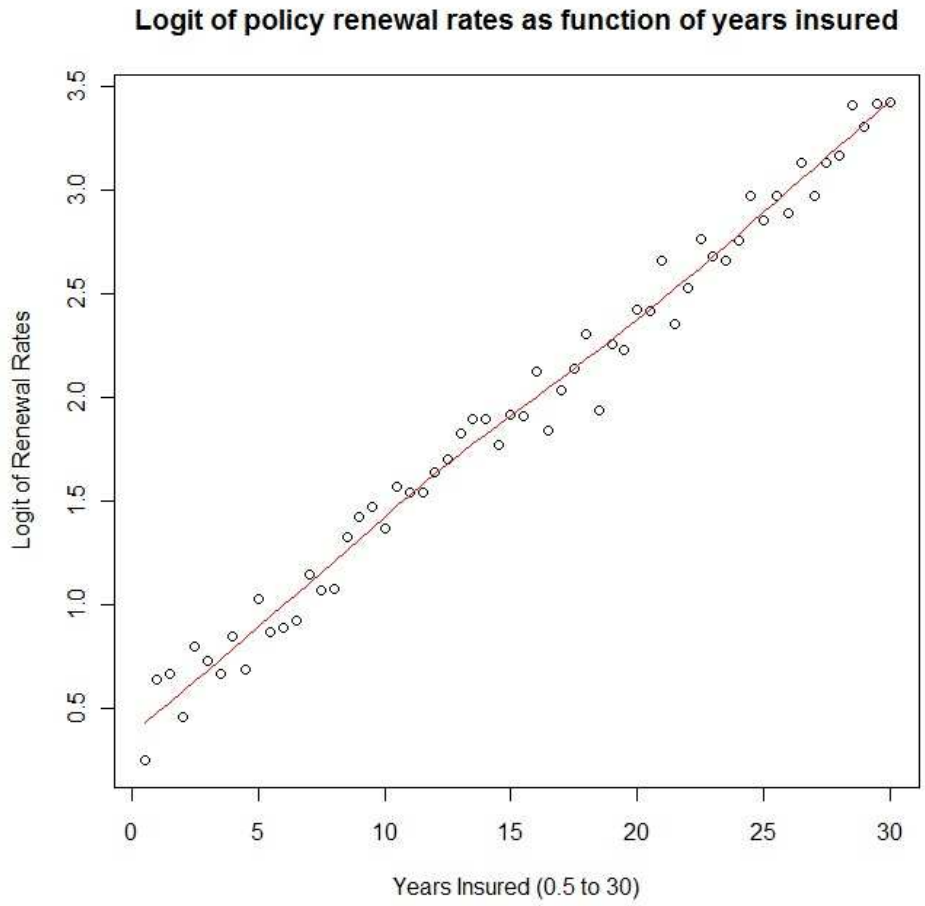
Rachel: The distribution is important, but it is harder to grasp. The homework assignment for the previous module deals with Poisson and Gamma distributions, and the same logic applies to binomial distributions. The variance of a binomial distribution is highest at the center and lowest in the tails.

Jacob: The GLM predicts the logits of the renewal rate. How do we get the renewal rate?

Rachel: Given the logit, the renewal rates is $1/(1 + e^{-\text{logit}}) = e^{\text{logit}} /(1 + e^{\text{logit}})$.

{The graphs below are what you should find.}


Policy renewal rates as function of years insured
Renewal Rates
0.6
0.7
0.8
0.9
0
5
10
15
20
25
30
Years Insured (0.5 to 30)


Logit of policy renewal rates as function of years insured
Logit of Renewal Rates
0.5
1.0
1.5
2.0
2.5
3.0
3.5
0
5
10
15
20
25
30
Years Insured (0.5 to 30)