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 σ^2_ϵ = 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.