MS Modules 18 and 19 – units of measurement practice problems

(The attached PDF file has better formatting.)

Variables and parameters have two parts: an amount and a unit of measurement. Regression parameters and the least squares estimates depend on the units of the explanatory and response variables.

Illustration: Suppose $Y_j = \beta_0 + \beta_1 X_{1j} + \beta_2 X_{2j} + \epsilon_j$

The items in the sum – Y_j, β_0, $\beta_1 \times X_{1j}$, $\beta_2 \times X_{2j}$, and ϵ_j – are in the same units, so

- β_0 (and its estimate $\hat{\beta}_0$) has the same units as Y
- ϵ (the error term) and the regression residuals have the same units as Y
- $\beta_1 \times X_1$ has the same units as Y, so β_1 (and its estimate $\hat{\beta}_0$) has the same units as Y/X_1

A change in the units of a parameter causes an offsetting change in the magnitude of the parameter.

Illustration: Suppose an explanatory variable is X = 2,500 meters and $\beta_1 = 0.03$.

If the units of the explanatory variable change to kilometers (1 kilometer = 1,000 meters) and the units of the response variable do not change, then

- the amount of X changes to 2.5 (divide by 1,000)
- β_1 changes to 30 (multiply by 1,000) so that $\beta \times X$ does not change

$\hat{\beta}_1$, the ordinary least squares estimator of β_1, changes the same way as β_1. If $\hat{\beta}_1 = 0.03$ when X is in meters, it is 30 when X is in kilometers.

The square of a random variable is in units-squared of the random variable.

Illustration: If Y is measured in meters, SST, SSE, and SSR, are in meters-squared. If the units of Y change from meters to kilometers, the values of Y are divided by 1,000, and the value of SST, SSE, and SSR are divided by $1{,}000^2 = 1{,}000{,}000$.

Ratios are in units of the numerator divided by units of the denominator.

Illustration: The R^2 is the regression sum of squares (SSR) divided by the total sum of squares (SST). SSR and SST are measured in the same units, so the R^2 is unit-less.

The number of observations (N) and the number of parameters (k) in the regression analysis are unit-less. If the explanatory variable changes from meters to kilometers, the number of observations and the number of parameters do not change. The adjusted $R^2 = 1 - (1 - R^2) \times (N - 1) / (N - k - 1)$, so the adjusted R^2 is also unit-less.

Standard deviations and standard errors are in the same units as the random variable they apply to.

Illustration: The standard error of $\hat{\beta}_1$ is in the same units as $\hat{\beta}_1$; the standard error of $\hat{\beta}_0$ is in the same units as $\hat{\beta}_0$; the standard error of the error term ϵ is in the same units as ϵ (which is the same units as the response variable Y and the residual *e*).

Many statistical measures are unit-less: they do not depend on the units of measurement of the explanatory or response variables.

Standardized values, such as standardized residuals, are the values divided by their standard errors, so they are unit-less.

Illustration: If the response variable changes from dollars to euros or from degrees Fahrenheit to degrees Centigrade, the standardized residuals do not change.

The *t* value for a least squares estimate is a standardized value, so it is unit-less. The *t* value measures the significance of the regression equation: changing the units of measurement does not change the significance.

Illustration: The *t* value is the estimated parameter divided by its standard error. A random variable and its standard error have the same units, so the *t* value is unit-less.

p values are measures of significance, so they are unit-less. They measure the percentage of scenarios with outcomes at least as extreme as the observed data (assuming random fluctuations and a null hypothesis). The *p* values are independent of the units of measurement of explanatory and response variables.

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_j 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 only β_0, not the slope coefficients.

Intuition: Knowing the units of the regression coefficients helps determine how a change in the units affects the coefficients.

- β_0 is in units of the response variable, not in units of the explanatory variables.
- β_1 is in units of response variable / explanatory variable.

Illustration: Suppose claim frequency = $\beta_0 + \beta_1 \times$ kilometers driven.

- β_0 is in units of claim frequency.
- β_1 is in units of claim frequency / kilometers driven

If we write the regression equation as claim frequency = $\beta_0 + (\beta_1/1{,}000) \times$ meters driven.

- β_0 is in units of claim frequency.
- β_1 is in units of claim frequency / kilometers driven $\Rightarrow$ β_1 / 1,000 is in units of claim frequency / meters.

Intuition: β_1 depends on the deviations of the values from their mean. A constant displacement of all the values doesn't affect the deviations. But a constant displacement of *k* raises the response variable Y by $k \times \beta_1$. β_0 has the same displacement as the response variable, so it also rises by $k \times \beta_1$.

Standardized coefficients and *t*-values are unit-less.

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.

$\Rightarrow$ 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.

Exercise 18.1: Sample size and expected values

A statistician regresses the response variable (Y_j) on the explanatory variable (X_j), $Y_j = \beta_0 + \beta_1 X_j + \epsilon_j$, with N = 5 observations, gives $\hat{\beta}_1$ (the least squares estimate for β_1) = 1. The null hypothesis is $\beta_1 = 0$. The X values here are randomly sampled from a population (e.g., X and Y are a bivariate normal distribution).

As N increases, what happens to the *expected values* of the following?

A. R^2
B. The adjusted R^2
C. σ_X, the standard deviation of X, or $\sqrt{(S_{xx}/(n-1))}$
D. $\sigma(\hat{\beta}_1)$, the standard error of $\hat{\beta}_1$
E. The *t* value for the null hypothesis H_0: $\beta_1 = 0$
F. The *p* value for the significance of the *t* test above
G. The width of the 95% confidence interval for β_1

Part A: As N increases, the expected value of R^2 decreases. The change in R^2 is noticeable for small values of N. For larger values of N, the change in the expected value of R^2 is small.

Part B: The adjusted R^2 is adjusted for degrees of freedom. A change in the number of observations changes the degrees of freedom, so the formula for the adjusted R^2 has an offsetting change in (N – 1) / (N – k) and the expected value of the adjusted R^2 does not change.

Intuition: With N = 2, the correlation of the explanatory variable and the response variable is +1 or –1, and the R^2 of the regression line is 100%, even if the two variables are independent. With N = 3 and a independent variables, the correlation is ±1 just by random fluctuations in one third of scenarios, and the expected value of R^2 is 50%.

Question: How do we derive the expected value of R^2?

Answer: If the explanatory variable and the response variable are independent, the expected value of the adjusted R^2 is zero. Using the formula

$$\text{Adjusted } R^2 = 1 - (1 - R^2) \times (N - 1) / (N - k - 1)$$

we derive

$$R^2 = 1 - (1 - \text{adjusted } R^2) \times (N - k - 1) / (N - 1)$$

For N = 3 and k = 1, we have

$$R^2 = 1 - (1 - 0) \times (3 - 1 - 1) / (3 - 1) = ½ = 50\%.$$

Part C: The population variance and standard deviation of X are not affected by the number of observations in the regression analysis. The sample variance and standard deviation of X may be higher or lower than the population variance and standard deviation, but the difference is random and has an expected value of zero.

Question: Why does the practice problem say that the X values are randomly sampled from a population? Isn't this always true?

Answer: The dependent variable Y is a random variable with mean μ_j (= the fitted value at $X = x_j$) and variance σ^2 (equal at all points). The independent variable X is not necessarily a random variable. It may be chosen for the experiment, such as the first N positive integers. (The variance of the first N integers depends on N.)

Question: Does S_{xx}, the sum of squared deviations of the X values, depend on the number of observations if the X values are a random sample from a population?

Answer: The variance is the sum of squared deviations divided by N-1. The variance is independent of N, so the expected value of the sum of squared deviations is proportional to N-1.

Part D: The variance of $\hat{\beta}_1$ is the estimated value of σ^2 divided by the sum of squared deviations of X. Since the estimated value of σ^2 is independent of N and the sum of squared deviations of X is proportional to N-1, the variance of $\hat{\beta}_1$ is proportional to 1/(N-1), and the standard error of $\hat{\beta}_1$ is proportional to $1/\sqrt{(N-1)}$. As N increases, the expected value of the standard error of $\hat{\beta}_1$ decreases.

Part E: The *t* value for $\hat{\beta}_1$ is the value of $\hat{\beta}_1$ divided by its standard error.

- The expected value of $\hat{\beta}_1$ does not change with the number of observations.
- The expected value of the standard error of $\hat{\beta}_1$ is proportional to $1/\sqrt{(N-1)}$.
- ⇛ The expected value of the *t* value for $\hat{\beta}_1$ is proportional to $\sqrt{(N-1)}$.

Part F: The *p* value for the significance of $\hat{\beta}_1$ decreases as the *t* value for $\hat{\beta}_1$ increases. This *p* value depends on the cumulative distribution function of the *t* distribution, which does not have a simple algebraic form.

Question: Do the degrees of freedom also affect the change in the *p* value?

Answer: More observations give a more compact (less heavy tailed) *t* distribution and a lower *p* value.

- As the *t* value increases, the *p* value decreases, even if N does not change.
- As the *N* increases, the *p* value decreases, even if the *t* value does not change.

Part G: The width of the 95% confidence interval for $\hat{\beta}_1$ is $2 \times t_{5\%/2,\, n-2} \times \sigma(\hat{\beta}_1)$. More observations increase the degrees of freedom and reduce the standard deviation of $\hat{\beta}_1$ and reduce the width of the confidence interval.

Exercise 18.2: Miles Driven and Annual Claim Costs

A U.S. actuary uses 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. The null hypothesis is H_0: $\beta_1 = 0$.

A European actuary uses the relation to estimate annual claims costs in *Euros* based on *kilometers* driven a day. For this problem, assume €1.00 = $1.25 and 1 kilometer = ⅝ mile (five eighths of a mile).

A. What is the percentage change in the intercept parameter β_0?
B. What is the percentage change in the slope coefficient β_1?
C. What is the percentage change in the variance σ^2?
D. What is the percentage change in the standard error of $\hat{\beta}_1$?
E. What is the percentage change in the *t* value to test the null hypothesis?
F. What is the percentage change in the *F* value to test the null hypothesis?
G. What is the percentage change in the width of the 95% confidence interval for β?
H. What is the percentage change in the *p* value for the test of the null hypothesis?
I. What is the percentage change in the total sum of squares?
J. What is the percentage change in the regression sum of squares?
K. What is the percentage change in the error sum of squares?
L. What is the percentage change in the R^2?

Part A: The intercept parameter β_0 is the value of the response variable (dependent variable) when the explanatory variable is zero. It does not depend on the value of the explanatory variable, so the units of measurement for the explanatory variable are not relevant.

The intercept parameter β_0 has the same units as the response variable. If these units increase by a factor of UM, β_0 increases by a factor of UM.

Illustration: An intercept of β_0 dollars = β_0 × 1.00 / 1.25 Euros (assuming €1.00 = $1.25) = β_0 × 0.80 Euros.

If β_0 = $10 for the regression model in dollars, then β_0 = €8 for the regression model in euros.

Part B: If the explanatory variable changes by 1, the response variable changes by β_1. The slope coefficient β_1 has units of dollars / miles. If the explanatory variable changes by 1 mile, the response variable changes by β_1 dollars.

If the explanatory variable changes by 1 kilometer, it changes by ⅝ × 1 mile, so the response variable changes by ⅝ × β_1 dollars = ⅝ × β_1 × 0.80 euros. The slope coefficient β_1 for the regression equation in kilometers and euros changes by a factor of ⅝ × 0.80 = 0.50.

Illustration: Suppose β_1 = $0.20 per mile for the regression model in dollars. An insured who drives 400 miles *per annum* has an annual cost that is $0.20 per mile × 100 miles per annum = $20 more *per annum* than an insured who drives 0 miles *per annum*.

- With the $1.25 = €1.00 exchange rate, $20 = €20/1.25 = €16.
- 100 miles = (1 / ⅝) × 100 = 160 kilometers.

160 kilometers *per annum* increases annual costs by €16, so β_1 = €16 / 160 kilometers *per annum* = 0.10.

Part C: The units of ϵ (the error term) are the same as the units of the response variable and of β_0, so the units of σ are these units as well and the units of σ^2 are the square of these units. If the change in the units causes

β_0 to be multiplied by 0.80, each error term is multiplied by 0.80, σ is multiplied by 0.80, and σ^2 is multiplied by 0.80 × 0.80 = 0.64.

Illustration: If σ^2 = 25 dollars^2 for the regression model in dollars, then σ^2 = 25 dollars^2 × (0.8 dollars/euros)^2 = 16 euros^2 for the regression model in euros.

Parts E, F, L: Tests of significance do not change. If the regression equation in miles and dollars is significant, the regression equation in kilometers and euros is significant. The *t* value, *p* value, and *F* value do not change.

The *t* value is $\hat{\beta}_0$ divided by its standard error. The standard error (standard deviation) of a random variable is in the same units of measurement as the random variable itself, so the *t* value does not change when the units of measurement change. The *F* value is the square of the *t* value, so it also does not change. The *p* value is the significance level of the *t* value, so it also does not change.

Part G: The width of the 95% confidence interval for β_1 is 2 × $\sigma(\beta_1)$ × $t_{5\%, df}$, where $t_{5\%, df}$ is the *t* value for a two-sided 95% confidence interval for a *t* distribution with degrees of freedom = *df*. The *t* value does not depend on the units of measurement, so the width of the 95% confidence interval is proportional to $\sigma(\beta_1)$. In this example, β_1 is multiplied by 0.50, so $\sigma(\beta_1)$ is multiplied by 0.50, and the width of the 95% confidence interval is multiplied by 0.50.

Part I: The total sum of squares SST depends on the values of the response variable. All these values are squared, so if the response variable changes by a factor of UM, the SST changes by a factor of UM^2.

Illustration: If the total sum of squares for the regression model in dollars is 1000 dollars^2, it is 1000 dollars^2 × (0.8 euros/dollar)^2 = 640 euros^2 for the regression model in euros.

Part J: The regression sum of squares SSR (= SST – SSE) is the portion of the total sum of squares explained by the explanatory variable. The proportion does not depend on the units of measurement: the units of SST, SSE, and SSR are the same. If the response variable changes by a factor of UM, SSR changes by a factor of UM^2.

Part K: SSE = SST – SSR, so it changes by the same ratio as SST and SSR.

SSE = σ^2 × the degrees of freedom. The degrees of freedom does not change when units of measurement change. In this example, σ^2 is multiplied by 0.64, so SSE is multiplied by 0.64.

Part L: SST, SSR, and SSE all change by the same percentage factor. The R^2 is SSR / SST = 1 – SSE/SST, so it does not change.

Exercise 18.3: Fahrenheit vs Centigrade

An American meteorologist regresses annual global warming (the response variable) in degrees Fahrenheit on coal burned (the independent variable) in billions of pounds:

$$\text{global warming} = \beta_0 + \beta_1 \times \text{energy use} + \epsilon$$

The response variable (global warming) is the *change* in the daily temperature over the year.

The least squares estimate for β_0 is 0.1 and the least squares estimate for β_1 is 0.01.

The meteorologist modifies the analysis to use degrees Centigrade and billions of kilograms for a presentation in Europe. The formula to convert degrees Fahrenheit to degrees Centigrade is

$$\text{degrees Centigrade} = (\text{degrees Fahrenheit} - 32^\circ) \times 5/9$$

For the weights, use 1 kilogram = 2.2 pounds (an approximate formula).

A. If no coal is burned, what is the modeled global warming in degrees Fahrenheit?
B. What is β_0 for the regression model in degrees Centigrade?
C. For each billion kilograms of coal burned, what is the extra global warming in degrees Centigrade?

Part A: If no coal is burned, the expected global warming = $\beta_0 + \beta_1 \times 0 = 0.1$ degrees Fahrenheit.

Part B: Warming is the change in temperature:

$$\text{degrees Centigrade} = (\text{degrees Fahrenheit} - 32^\circ) \times 5/9 \Rightarrow$$
$$\text{change in degrees Centigrade} = (\text{change in degrees Fahrenheit}) \times 5/9$$

β_0 for the regression model in degrees Centigrade = $0.1 \times 5/9 = 0.55556$.

Part C: For each billion pounds of coal burned, the extra global warming in degrees Fahrenheit is 0.01, so

β_1 = 0.01 degrees Fahrenheit / pounds =

$\beta_1 = 0.01 \times 5/9$ degrees Centigrade / (1 / 2.2 kilograms) = $0.01 \times (5/9) \times (1/2.2) = 0.00253$

Exercise 18.4: 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:

A. What is the revised value of β_1?
B. What is the revised value of β_0?

Assume €1.00 = $1.25 and 1 kilometer = ⅝ mile (five eighths of a mile).

Part A: The estimate of β_1 is the covariance of X and 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_1 by 1.280 / 2.560 = 0.500.

Part B: β_0 is not affected by the units of X, since the product $\beta_1 \times X$ is not affected by the units of X. But β_0 varies directly with the units of Y: if Y is multiplied by 0.80, β_0 is multiplied by 0.80.

Question: Is the product $\beta_1 \times X$ unit-less?

Answer: 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 β_0 is 40.
- Before the change, if X = 5 miles, Y = $250. Now X = 8 kilometers gives Y = $250 × 0.8 = €200. Since $\beta_0 = 40$, β_1 is (200 – 40) / 8 = 20.