Regression analysis Module 8: variances and means (practice problems)

(The attached PDF file has better formatting.)

** Exercise 8.1: Variances and means

A regression equation of Y on X, $Y_i = \alpha + \beta \times X_i + \epsilon_i$, with N=5 observations, has

- RSS, the residual sum of squares, = 5.10
- σ^2_B, the variance of B, the ordinary least squares estimator of β, = 0.17
- σ^2_A, the variance of A, the ordinary least squares estimator of α, = 1.87

A. What is the σ^2_ε, the variance of the error term?
B. What is $\sum x_i^2$, the sum of the squared *x* values?
C. What is $\sum(x_i - \overline{x})^2$, the sum of the squared *x* residuals?
D. What is $\overline{x}$, the mean of the X values?

Part A: σ^2_ε = RSS / degrees of freedom of the regression equation, which is N – k – 1, where k is the number of explanatory variables: 5.1 / (5 – 1 – 1) = 1.700.

Part B: The variance of B, σ^2_B, is $\sigma^2_\varepsilon / \sum(x_i - \overline{x})^2$, and the variance of A, σ^2_A, is $[\sigma^2_\varepsilon / \sum(x_i - \overline{x})^2] \times [\sum x_i^2 / N] \Rrightarrow$

$$\sum x_i^2 = N \times \sigma^2_A / \sigma^2_B = 5 \times 1.87 / 0.17 = 55.$$

Part C: $\sum(x_i - \overline{x})^2 = \sigma^2_\varepsilon / \sigma^2_B = 1.70 / 0.17 = 10.$

Part D: $\sum(x_i - \overline{x})^2 = \sum x_i^2 - 2 \sum x_i \overline{x} + N \overline{x}^2 = \sum x_i^2 - N \overline{x}^2$, since $\sum x_i \overline{x} = N \overline{x}^2$, so

$$\overline{x} = \{ [\sum x_i^2 - \sum(x_i - \overline{x})^2] / N \}^{1/2} = \{ [55 - 10] / 5 \}^{1/2} = 3$$

** Exercise 8.2: Sampling variance

A regression has N observations, with a standard error of σ^2_ε and a variance of the explanatory variable of S^2_x.

Explain how the following affect the sampling variances of the slope estimate B and the intercept estimate A.

A. σ^2_ε
B. sample size N
C. variance of the explanatory variable S^2_x
D. The closeness of the X values to zero

The formulas for the sampling variances are

$\sigma^2_B = \sigma^2_\varepsilon / \sum(x_i - \overline{x})^2 = \sigma^2_\varepsilon / ((N - 1) \times S^2_x)$

$\sigma^2_A = \sigma^2_B \times (\sum x^2_i / N)$

Part A: As σ^2_ε increases, σ^2_B and σ^2_A increase.

Intuition: As the standard error of the regression increases, the estimates of the regression coefficients are less certain.

Part B: As N increases, σ^2_B and σ^2_A decrease.

Intuition: With only a few observed values, the estimated regression line is uncertain. Both the slope and the intercept may be distorted by one or two outlying values. With more observed values, the estimated regression line is more certain.

Part C: As S^2_x increases, σ^2_B and σ^2_A decrease.

Intuition: The regression line passes through ($\overline{x}$, $\overline{y}$), the means of the X and Y values. Think of the regression line as a bar hinged at the point ($\overline{x}$, $\overline{y}$) but with unknown slope. Random fluctuations in the observed values of the response variable Y may distort the slope. If the X values are widely dispersed, some of them are far from the mean $\overline{x}$. An incorrect slope coefficient causes a large squared error at that point, so incorrect slope coefficients are less likely. An incorrect slope coefficient causes a large error in the intercept, so if the slope coefficient is more accurate, so is the intercept.

Part D: The closeness of the X values to zero has no effect on σ^2_B. B, the estimate of the slope coefficient β, depends on $(x_i - \overline{x})$, not on x_i, so adding a constant to all the x values doesn't change B. But if $\overline{x}$ is far from zero, an error in the slope coefficient greatly affects the intercept. As $(\sum x^2_i / N)$ increases, σ^2_A increases.

Intuition: If $\overline{x} = 0$, the intercept is $\overline{y}$, with no uncertainty. No matter what value B has, A is $\overline{y}$. If $\overline{x}$ is 100, an error of *k* in the estimate of B causes an error of 100 × *k* in the estimate of A.

** Exercise 8.3: Standard errors of ordinary least squares estimators for α (A) and β (B)

A statistician uses a regression on the X values {-1, -0.9, -0.8, -0.7, ..., -0.1, 0, 0.1, …, 0.7, 0.8, 0.9, 1) to test null hypotheses that $\alpha = 0$ and that $\beta = 0$. The ordinary least squares estimators of α and β are both 1.000.

A. Which estimator has the higher standard error?
B. Which estimator has the higher *t*-value?
C. Which estimator has the higher *p*-value for the test of the null hypothesis?

Part A: We don't know the standard errors of A or B, since we don't know the standard error of the regression S^2_ε. But we know the ratios of these standard errors. N (the number of data points) = 21, and $\sum x^2_i = 7.7$, so $\sum x^2_i$ / N = 7.7 / 21 = 0.367. B has the higher standard error.

Part B: The *t*-value is the regression coefficient divided by its standard error. B has the higher standard error, so A has the higher *t*-value.

Part C: A higher *t*-value means a more significant coefficient so a lower *p*-value. B has the higher *p*-value.