MS Module 17: Regression analysis confidence intervals and hypothesis testing – practice problems

(The attached PDF file has better formatting.)

Exercise 17.1: Prediction interval and confidence interval (intuition)

Let W be the ratio of the width of the prediction interval to the width of the confidence interval for a given (i) confidence level, (ii) x-value, (iii) and linear regression model.

What happens to this ratio (the width of the prediction interval to the width of the confidence interval) as

A. N, the number of observations, increases, but the mean and standard deviation of X do not change.
B. $S_{xx}$, the sum of squared residuals of the X values, increases, but N does not change.
C. $|x^* - \overline{x}|$, the absolute value of the X deviation at which we are measuring the widths, increases.
D. The variance of the error term ($\sigma^2$) increases
E. The least squares estimate of $\beta_1$ increases
F. The confidence level ($1 - \alpha$) increases

*Part A:* The widths of the confidence interval and of the prediction interval are proportional to three items:

! the *z* value (or *t* value)
! the standard deviation of the error term $\sigma_\varepsilon$
! the square root of an expression that is one unit greater for the prediction interval than for the confidence interval: $(1 + 1/n + (x^* - \overline{x})^2 / S_{xx})$ vs $(1/n + (x^* - \overline{x})^2 / S_{xx})$

! Confidence interval: $\hat{\beta}_0 + \hat{\beta}_1 x^* \pm t_{\alpha/2,n-2} \times s \times (1/n + (x^* - \overline{x})^2 / S_{xx})^{½}$
! Prediction interval: $\hat{\beta}_0 + \hat{\beta}_1 x^* \pm t_{\alpha/2,n-2} \times s \times (1 + 1/n + (x^* - \overline{x})^2 / S_{xx})^{½}$

The number of observations N appears as 1/N under the square root symbol. The standard deviation of X does not change, so $S_{xx}$ is proportional to N-1. As N increases, 1/n and $(x^* - \overline{x})^2 / S_{xx}$ decrease, and the widths of the confidence interval and the prediction interval decrease.

The prediction interval has a constant 1 under the square root symbol as well, which doesn't change as N increases, so the width of the prediction interval decreases proportionately less than the width of the confidence interval. Increasing N raises the width-ratio W.

The ratio W is = $[ (1 + 1/n + (x^* - \overline{x})^2 / S_{xx}) / (1/n + (x^* - \overline{x})^2 / S_{xx}) ]^{½}$

As N increases, the numerator of W decreases proportionately less than the denominator

*Question:* Is the solution valid only if the mean and standard deviation of the X values does not change?

*Answer:* If the X values are a sample from a population, raising the number of observations doesn't change the mean and standard deviation of the population or the expected mean and expected standard deviation of the sample. But regression analysis does not assume the X values are a random sample from a population. The X values may be the first N integers, whose mean and standard deviation change as N changes.

*Question:* What is the intuition for the reduction of the ratio W as N increases?

*Answer:* The ratio W reflects two sources of uncertainty:

! The width of the confidence interval reflects the uncertainty in the regression line.
! The width of the prediction interval reflects the uncertainty in the regression line and the stochasticity of the data points (the random error term).

A larger number of observations reduces the uncertainty in the regression line but not the stochasticity of the data points (the random error term). As $N \to \infty$, the uncertainty in the regression line approaches zero and the width of the confidence interval approaches zero, but the width of the prediction interval approaches the *z* value times the standard deviation of the error term.

*Part B:* As $S_{xx}$, the sum of squared residuals of the X values, increases (with all else remaining the same), the value of $(x^* - \overline{x})^2 / S_{xx}$ decreases. The denominator of the W ratio decreases proportionately more than the numerator, so the W ratio increases.

*Question:* What is the intuition for this?

*Answer:* The uncertainty in the regression line decreases but the stochasticity of the data points (the random error term) remains the same, so W increases.

*Part C:* As $(x^* - \overline{x})^2$ increases, $(x^* - \overline{x})^2 / S_{xx}$ increases, but the other terms in the expressions for the widths of the confidence interval and the prediction interval do not change, so the ratio W decreases.

*Question:* What is the intuition for this?

*Answer:* The regression line passes through $(\overline{x}, \overline{y})$ and has the slope $\beta_1$. Think of the regression line as rotating through the point $(\overline{x}, \overline{y})$. The uncertainty in the slope of $\beta_1$ is the standard deviation of $\beta_1$.

If $x^*$ is far from the mean $\overline{x}$, the uncertainty in the slope of $\beta_1$ causes great uncertainty in the value of $\hat{y}$ at $x = x^*$. This uncertainty affects the confidence interval and the prediction interval equally. The extra uncertainty in the prediction interval caused by the stochasticity of the error term is independent of the distance of $x^*$ from $\overline{x}$, so the ratio W decreases.

*Part D:* The width of the confidence interval and the width of the prediction interval are both proportional to s, the estimate for $\sigma$, so the ratio W does not change.

*Question:* What is the intuition for this?

*Answer:* The standard deviation of the error term is *s*, the estimate of $\sigma$. The standard deviation of the slope of the regression line is *s* divided by the estimate of $\beta_1$. Both widths in the ratio W are proportional to *s*.

*Part E:* The least squares estimate of $\beta_1$ affects the location of the confidence and prediction intervals, not the widths of the intervals. Even if the standard deviation of $\beta_1$ is proportional to the magnitude of $\beta_1$, the ratio of the widths W does not change, since the two widths change by the same factor.

*Part F:* A higher confidence level, such as 99% instead of 95%, lengthens the widths of the two intervals by the same factor; the ratio of the widths does not change.

Exercise 17.2: Confidence interval and prediction interval

- The independent (X) values for a linear regression are {1, 2, …, 11}.
- The width of the $(1 - \alpha)$ confidence interval at the point x = 2 is 1.

A. What is the formula for the width of the $(1 - \alpha)$ confidence interval at the point x = 2?
B. What is the formula for the width of the $(1 - \alpha)$ prediction interval at the point x = 9?
C. What is the width of the $(1 - \alpha)$ prediction interval at the point x = 9?

*Part A:* The $(1 - \alpha)$ confidence interval is

$$\hat{\beta}_0 + \hat{\beta}_1 x^* \pm t_{\alpha/2,n-2} \times s \times (1/n + (x^* - \overline{x})^2 / S_{xx})^{½}$$

The width of the confidence interval is

$$2 \times t_{\alpha/2,n-2} \times s \times (1/n + (x^* - \overline{x})^2 / S_{xx})^{½}$$

The values of $\overline{x}$ and $S_{xx}$ are 6 and 110 (worked out in other exercises).

At $x^* = 2$, the width is

$$t_{\alpha/2,n-2} \times s \times 2 \times (1/n + (x^* - \overline{x})^2 / S_{xx})^{½} =$$

$$t_{\alpha/2,n-2} \times s \times 2 \times (1/11 + (2 - 6)^2 /110)^{½} = 0.972345 \times t_{\alpha/2,n-2} \times s$$

*Part B:* The $(1 - \alpha)$ prediction interval is

$$\hat{\beta}_0 + \hat{\beta}_1 x^* \pm t_{\alpha/2,n-2} \times s \times (1 + 1/n + (x^* - \overline{x})^2 / S_{xx})^{½}$$

The width of the prediction interval is

$$2 \times t_{\alpha/2,n-2} \times s \times (1 + 1/n + (x^* - \overline{x})^2 / S_{xx})^{½}$$

The values of $\overline{x}$ and $S_{xx}$ are 6 and 110 (worked out in other exercises).

At $x^* = 9$, the width is

$$t_{\alpha/2,n-2} \times s \times 2 \times (1 + 1/n + (x^* - \overline{x})^2 / S_{xx})^{½} =$$

$$t_{\alpha/2,n-2} \times s \times 2 \times (1 + 1/11 + (9 - 6)^2 /110)^{½} = 2.165851 \times t_{\alpha/2,n-2} \times s$$

*Part C:* The width of the prediction interval at the point x = 9 is

$$1 \times 2.165851 / 0.972345 = 2.227451$$

Exercise 17.3: Width of confidence interval and width of prediction interval

A linear regression analysis relates Y to X.

- The X values are {1, 2, …, 11}
- The error sum of squares (SSE) is 36.

A. What is $\overline{x}$, the average X value?
B. What is $S_{xx}$, the sum of squared residuals for the X values?
C. What is $s^2$, the estimate of $\sigma^2$?
D. What is the *t* value for a 95% two-sided confidence interval?
E. What is the width of the 95% confidence interval at x = 8?
F. What is the width of the 95% prediction interval at x = 8?

*Part A:* The mean X value is (1 + 2 + 3 + 4 + 5 + 6 + 7 + 8 + 9 + 10 + 11) / 11 = 6

*Part B:* $S_{xx}$, the sum of squared residuals for the X values, is

$(1-6)^2 + (2-6)^2 + (3-6)^2 + (4-6)^2 + (5-6)^2 + (6-6)^2 + (7-6)^2 + (8-6)^2 + (9-6)^2 + (10-6)^2 + (11-6)^2 = 110$

*Part C:* The value of $s^2$, the estimate of $\sigma^2$, is SSE/(N-2) = 36 / (11-2) = 4.

*Part D:* The degrees of freedom for the *t* value is N-2 = 11-2 = 9. The *t* value for a 95% two-sided confidence interval is $z_{\alpha/2,n-2} = z_{0.025,9} = 2.262157$ (table look-up or spreadsheet function).

*Part E:* The confidence interval is $\hat{\beta}_0 + \hat{\beta}_1 x^* \pm t_{\alpha/2,n-2} \times s \times (1/n + (x^* - \overline{x})^2 / S_{xx})^{½}$. The width of the confidence interval is

$$2 \times t_{\alpha/2,n-2} \times s \times (1/n + (x^* - \overline{x})^2 / S_{xx})^{½}$$

At $x^* = 8$, this width is $2 \times 2.262157 \times 2 \times (1/11 + (8-6)^2 / 110)^{0.5} = 3.22813$

*Part F:* The prediction interval is $\hat{\beta}_0 + \hat{\beta}_1 x^* \pm t_{\alpha/2,n-2} \times s \times (1 + 1/n + (x^* - \overline{x})^2 / S_{xx})^{½}$. The width of the prediction interval is

$$2 \times t_{\alpha/2,n-2} \times s \times (1 + 1/n + (x^* - \overline{x})^2 / S_{xx})^{½}$$

At $x^* = 8$, this width is $2 \times 2.262157 \times 2 \times (1 + 1/11 + (8-6)^2 / 110)^{0.5} = 9.60721$

Exercise 17.4: Confidence interval for predicted value

- Y is a linear function of X: $Y_j = \beta_0 + \beta_1 X_j + \epsilon_j$
- Using data for the 7 points {1, 2, 3, 4, 5, 6, 7}, we estimate $\beta_0 = 1$, $\beta_1 = 3$, and $s^2 = 2$

We observe another point {X, Y}, where X = 5.50

A. What is $\hat{y}$ at X = 5.50?
B. What is $\overline{x}$?
C. What is $S_{xx}$?
D. What is the estimated standard deviation of the statistic $\hat{y}$ at X = 5.50?
E. What are the degrees of freedom for the *t* value?
F. What is the *t* value for the 95% confidence interval?
G. What is the 95% confidence interval for $\hat{y}$ at X = 5.50?
H. What is the estimated standard deviation of the prediction for Y at X = 5.50?
I. What is the 95% confidence interval for the predicted Y value at X = 5.50?

*Part A:* $\hat{y} = 1 + 3 \times 5.5 = 17.50$

*Part B:* $\overline{x} = (1 + 2 + 3 + 4 + 5 + 6 + 7) / 7 = 4$

*Part C:* $S_{xx} = (1 - 4)^2 + (2 - 4)^2 + (3 - 4)^2 + (4 - 4)^2 + (5 - 4)^2 + (6 - 4)^2 + (7 - 4)^2 = 28$

*Part D:* The estimated standard deviation of the statistic $\hat{y}$ at X = 5.5 is

$$s \times (1/n + (x^* - \overline{x})^2 / S_{xx})^{½} =$$

$$2^{0.5} \times (1/7 + (5.5 - 4)^2 / 28)^{0.5} = 0.668153$$

*Part E:* The degrees of freedom for the *t* value is 7 – 2 = 5.

*Part F:* The *t* value for the two-sided 95% confidence interval with 5 degrees of freedom is 2.570582 (table look-up or spread-sheet function).

*Part G:* The confidence interval is the fitted value ± the *t* value × the standard deviation of the fitted value.

- Lower bound: 17.50 – 2.570582 × 0.66815 = 15.782466
- Upper bound: 17.50 + 2.570582 × 0.66815 = 19.217534

*Part H:* The estimated standard deviation for the predicted Y value at X = 5.5 is

$$s \times (1 + 1/n + (x^* - \overline{x})^2 / S_{xx})^{½} =$$

$$2^{0.5} \times (1 + 1/7 + (5.5 - 4)^2 / 28)^{0.5} = 1.564106$$

*Part I:* The prediction interval is the fitted value ± the *t* value × the standard deviation of the predicted value.

- Lower bound: 17.50 – 2.570582 × 1.564106 = 13.479337
- Upper bound: 17.50 + 2.570582 × 1.564106 = 21.520663

Know the formulas for the confidence interval and the prediction interval:

- Confidence interval: $\hat{\beta}_0 + \hat{\beta}_1 x^* \pm t_{\alpha/2,n-2} \times s \times (1/n + (x^* - \overline{x})^2 / S_{xx})^{½}$
- Prediction interval: $\hat{\beta}_0 + \hat{\beta}_1 x^* \pm t_{\alpha/2,n-2} \times s \times (1 + 1/n + (x^* - \overline{x})^2 / S_{xx})^{½}$