Inference for Simple Linear Regression and Interpretation

J. Dittenber, M.S.

October 31, 2021

Inference Procedures for Simple Linear Regression Parameters

A: Confidence Interval for β_1

We can calculate the 95% confidence interval manually or by using R. To calculate it manually, we use the general template

$$b_1 \pm t_{\frac{\alpha}{2}} \times s\{b_1\}$$

where we have

$$b_1 = \frac{\sum(x_i - \overline{x})(y_i - \overline{x})}{\sum(x_i - \overline{x})^2}$$

$t_{\frac{\alpha}{2}}$ calculated using $df = n - 2$ and

$$s\{b_1\} = \sqrt{\frac{MSE}{\sum(x_i - \overline{x})^2}}$$

and

$$MSE = \frac{\sum(y_i - \overline{y})^2}{n-2}$$

where for each sum we have i run from 1 to n.

$$b_1 = 0.03883 \qquad s\{b_1\} = 0.01277 \qquad t_{\frac{\alpha}{2}} = 1.9803$$

Therefore a 95% confidence interval can be constructed as such

$$0.03883 \pm 1.9803 * 0.01277$$

95% **Confidence Interval**

A 95% confidence interval for the parameter β_1 is given by

$$(0.0135, 0.064)$$

Interpretation of the Confidence Interval

We can say with 95% confidence that the true population parameter, β_1 lies in the interval

$$(0.0135, 0.064)$$

What can we learn from this interval

Since 0 is not in the 95% confidence interval it is not probable that the true parameter is zero. That is because the hypothesis test has a null hypothesis that $\beta_1 = 0$, which, if true, would indicate that there is no linear association between ACT and GPA. Since 0 is not in the interval, this means that the ACT is a significant predictor of GPA at the 5% significance level. However, this alone, cannot tell us anything more about the capability of ACT to predict GPA. We need to also determine many other things such as the R-value, R-squared value, examine the residuals plot and also check that the assumptions are met before generalizing a regression model to a population. It is likely the case that ACT in addition to some other predictors can create a more capable model.

B. Conduct t-test to determine correlation between ACT and GPA at the 5% significance level

Recall that β_1 is a parameter in the regression model

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$

Given that any linear regression model assumes normality of the probability distributions of the response variable along with constant variance, it follows, that if $\beta_1 = 0$ then the probability distributions of Y are all identical.

$$E(Y) = E(\beta_0 + 0 * X) + E(\varepsilon) = \beta_0$$

This would indicate that there is no linear (or any other) relationship between the variables ACT and GPA. To determine if there is a linear relationship between ACT and GPA, we can conduct a t-test. We first set our two-tailed hypothesis test as follows:

$$H_0 : \beta_1 = 0 \qquad\qquad H_A : \beta_1 \neq 0$$

To calculate our test statistic, we follow the general formula

$$t^* = \frac{statistic - parameter}{standard \quad error} = \frac{b_1 - \beta_1}{s\{b_1\}} = \frac{0.03883 - 0}{0.01277} = 3.040$$

Here we conduct our test according to the following criteria:

If $|t^*| \leq t(1 - \frac{\alpha}{2}; n-2) \rightarrow$ do not reject H_0 and if $|t^*| > t(1 - \frac{\alpha}{2}; n-2) \rightarrow$ reject H_0

Decision: t-critical method

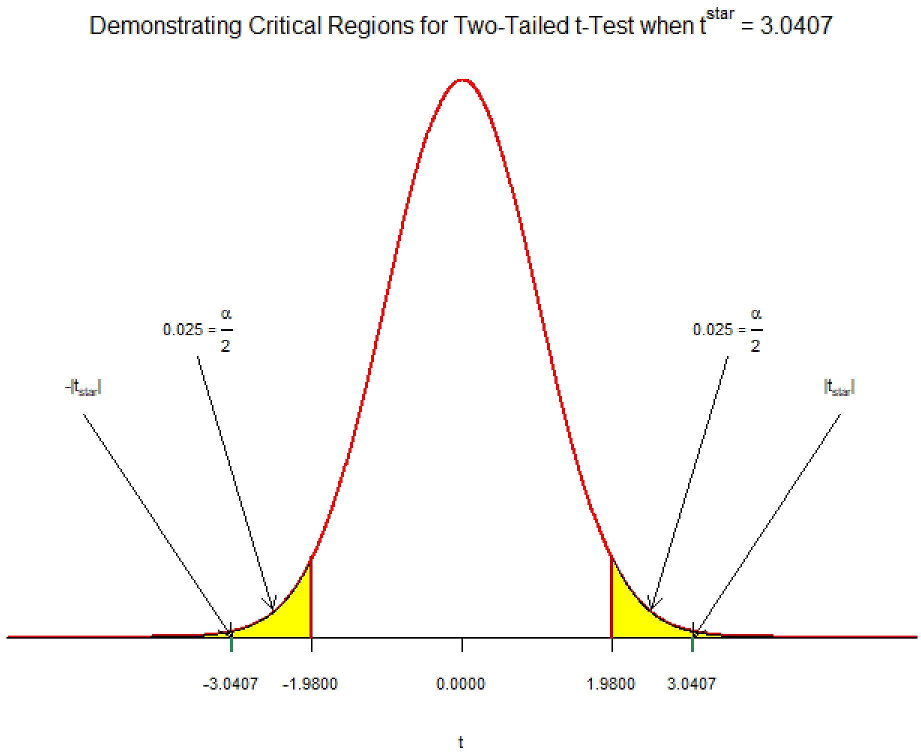
In this case, we have

$$t^* = 3.040$$

and

$$t(.975, 118) = 1.9803$$

since $|t^*| > t(1 - \frac{\alpha}{2}; n-2)$, we **reject the null hypothesis.**



P-value method

We can calculate our p-value as well if we would like to directly compare it to the alpha-level. Here we calculate

$$P(t(118) > t^* = 3.040) = 2 * 0.00145 = 0.0029 = p - value$$

and so in this way, we can confirm the decision that was made based on comparison of the t-scores since $p - value < \alpha$.

Conclusion

Given that we found $t^* > t(.975, 118)$ we may reject the null hypothesis. That is, at the 5% significance level, there is a statistically significant linear relationship between ACT and GPA.

C. 95% Confidence interval for ACT=28

We need to find a point estimate for this level of $x_h = 28$ and then construct a confidence interval following the typical template

$$estimate \pm t_{\frac{\alpha}{2}} \times SE$$

To calculate an estimate we use

$$\hat{Y}_h = b_0 + b_1 X_h$$

We have that the

$$\hat{Y} \sim N\left(E(Y_h), MSE\left[\frac{1}{n} + \frac{(x_h - \overline{x})^2}{\sum(x_i - \overline{x})^2}\right]\right)$$

and so

$$E(\hat{Y}_h) = E(Y_h) = E(b_0 + b_1 X_h)$$

taking the expected value of the RHS

$$E(Y_h) = b_0 + b_1 X_h$$

Previously we calculated b_1 and $b_0 = \overline{Y} - b_1\overline{X}$, so we have that

$$b_1 = 0.03883 \qquad b_0 = 2.11405 \qquad X_h = 28 \qquad \hat{Y}_{28} = 2.114050 + 0.03883(28) = 3.20$$

We want to estimate $E(Y_h)$ when the ACT score is 28, therefore we will calculate

$$\hat{Y}_h \pm t_{\frac{\alpha}{2}} \times s\{\hat{Y}_h\}$$

Here, we calculate $t_{\frac{\alpha}{2}} = t(1 - \frac{\alpha}{2}; n-2) = t(.975, 118) = 1.9803$ as we did in the previous part of the problem. And we calculate

$$MSE\left[\frac{1}{n} + \frac{(x_h - \overline{x})^2}{\sum(x_i - \overline{x})^2}\right]$$

First, calculate

$$MSE = \frac{\sum \varepsilon^2}{n-2} = \frac{45.81761}{118} = 0.3883$$

and then

$$\left[\frac{1}{n} + \frac{(x_h - \overline{x})^2}{\sum(x_i - \overline{x})^2}\right] = \left[\frac{1}{120} + \frac{10.72562}{2379.925}\right] = 0.01284$$

$$s^2(\hat{Y}_{28}) = 0.3883 \times 0.01284 = 0.00499 \to s(\hat{Y}_{28}) = 0.0706$$

$$\hat{Y}_h \pm t_{\frac{\alpha}{2}} \times s\{\hat{Y}_h\} = 3.20 \pm 1.9803 \times 0.0706$$

Therefore, the 95% confidence interval is $(3.06, 3.34)$. **Interpretation** We can say with 95% confidence that the expected value of the distribution of Y_h is in the interval $(3.06, 3.34)$ at the level $X_h = 28$. We can say with 95% confidence that a student with an ACT score of 28 will have a GPA between 3.06 and 3.34 using this model.