Lab 11: Regression Inference

Objective: Last lab, we looked at a linear regression model for Stats 250 scores, and started to examine if it was an appropriate to use a linear model. In this lab, you will learn how to perform hypothesis tests and compute confidence intervals in regression and assess if the linear relationship is significant, as well as learn how to check the assumptions needed for these inference procedures to be valid.

Application: Recall our example from last lab -- Pam believes that the number of Facebook friends a person has could interfere with the number of hours they spends socializing offline. She wants to be able to predict the number of hours a person spends socializing offline for a given number of Facebook friends. She will collect data from her friends and examine the relationship between these two variables. Pam will be able to fit a simple linear regression model if the relationship between these variables is linear.

Overview: To do this, Pam has fit a **linear regression model,** which is what will also be fit during the lab. Pam's regression is based on a linear model relating the number of Facebook friends to the mean amount of time spend socializing offline as follows: E(*time spent socializing offline*) = $\beta_0 + \beta_1$(*number of Facebook friends)*. Here β_0 and β_1 are parameters – fixed but unknown constants. Specifically, β_0 is the population *y*-intercept (the amount of time spent socializing offline when the number of Facebook friends is zero) and β_1 is the population slope (the change in the mean time spent socializing offline for every additional Facebook friend). These two values are unknown, but can be estimated using the least squares criterion. The resulting estimated regression line is generally written as $\hat{y} = b_0 + b_1(x)$. The estimates, b_0 and b_1 are referred to as the **least squares estimates** of β_0 and β_1.

There are also several assumptions that Pam must check in order for inferences to be valid. First, the time spent socializing offline must be normal with a mean that varies linearly with the number of Facebook friends and a standard deviation that does not depend on the predicted socializing time. To check her assumption Pam would create a residuals vs fitted plot: a plot with residuals on the y-axis and the predicted social time on the x-axis. For this assumption to hold, the residuals should be randomly scattered.

Pam must also assume that the error terms are normally distributed and are identically distributed. To check this assumption about the normal distribution she will create a QQ plot of the residuals. To check if the errors are identically distributed, she will create a time plot of the residuals and look for stability.

Warm-Up: Check Your Understanding

A linear regression model was built to assess the relationship between the weight (pounds) and the height (inches) of men aged 18 to 29. The estimate of the least squares regression line is given to be:

Predicted weight = -250 + 6*(Height)

a. What is the predicted weight for a male with a height of 70 inches?

250 ***170*** ***420***

b. We know that the average height of a male aged 18 to 29 is 72 inches. If we were to construct a 95% confidence interval for the weight of a male that is 72 inches tall, how would the width of that interval compare to a 95% confidence interval for the weight of a male that is 70 inches tall?

NARROWER ***WIDER*** ***STAYS THE SAME***

ILP: Is There a Significant Linear Relationship Between Final Exam Score and Our Selected Predictor?

Recall our previous lab on the Stats 250 data. Is the explanatory variable that we chose a useful linear predictor for the Final Exam score? That is, do we observe a significant, non-zero linear relationship between the Final Exam Score and our selected predictor variable?

Remember that another way to make inferences about the significance of the linear relationship is through a confidence interval for the population slope. Further, recall the basic form of a confidence interval: point estimate ± (a few) standard errors.

Most standard computer regression output provides the slope estimate and its standard error, and the "few" will correspond to a t^* value for the corresponding confidence level with degrees of freedom for regression of $n - 2$. Since a confidence interval provides a range of reasonable values for the parameter, it can be used to perform two-sided hypothesis tests by seeing whether the hypothesized value falls in the interval or not.

Task: Assess if the chosen explanatory variable is significant in the linear model.

> **NOTE:** You may have to re-run some of the output that we ran in our last class, specifically creating the regression model. Refer to the last lab for instructions on how to do this.

Hypothesis Test:

1. **State the Hypotheses:** H_0: _____ and H_a: _____ ,

   where _____ represents:

   Determine Alpha: We were told the significance level was 5%.

   > **Remember:** Your hypotheses and parameter definition should always be a statement about the **population(s)** under study.

2. **Checking the Assumptions** - **Covered in the next activity.**

3. **Compute the Test-Statistic and calculate the *p*-value:**
   Test-Statistic

a. Using the computer output generated by the regression model, which two test statistics could you use to test these hypotheses? Give the value for each test statistic.

t = _____ F = _____

Check it out: What happens if you square your t statistic value? Think about the shape of the t distribution and if you squared all the t values, it would look like an F distribution.

Calculate the *p*-Value:

b. The *p*-value for both test statistics is the same. What is that *p*-value?_____

Note: this *p*-value is for testing the two-sided alternative H_a: $\beta_1 \neq 0$. If we want to conduct a one-sided version of the alternative hypothesis we could only use the t test statistic and would need to work with the two-sided *p*-value in the output to find the appropriate one-sided *p*-value. Drawing a quick sketch of what that one-sided *p*-value looks like can help.

c. We could also get the F-test statistic value from running an ANOVA table to summarize the variance in our regression model. To do this, go to **Models > Hypothesis Tests > ANOVA**. To ensure that we see the table in a familiar format, select **Sequential "Type I"** as our option and click OK. Use this ANOVA table to calculate the r^2 value, and verify that this is the same as last lab. Remember that SSTotal = SSRegressionModel + SSError(or residuals).

$$r^2 = \frac{SSRegression\ model}{SSTotal} =$$

4. Evaluate the *p*-value and Conclusion

Evaluate the *p*-value:

What is your decision at a 5% significance level? **Reject H_0** **Fail to Reject H_0**

Remember:	Reject H_0	⇔ Results statistically significant
	Fail to Reject H_0	⇔ Results not statistically significant

Conclusion:

What is your conclusion *in the context of the problem*?

Note: Conclusions should always include a reference to the population parameter(s) of interest. They should not be too strong; you can say that you have sufficient evidence, but do NOT say that we have *proven* anything true or false.

5. Confidence Intervals (CI):

a. Generate the confidence interval for estimating the population slope using **Models > Confidence Intervals**. Give the 95% confidence interval for the population slope.

b. Provide an interpretation of the resulting interval in context.

c. Based on the confidence interval, would you reject the null hypothesis at a 5% significance level? Circle one: **Yes** **No**
Explain.
Did your conclusion here match the one you made in part 4?

ILP: Is the Linear Model Appropriate? Are Assumptions Met for Inference?

You will produce and examine the residuals from the regression line as well as create some plots to assess the fit of the linear model. This will also serve to evaluate the validity of the testing and confidence intervals performed in the earlier activities.

Regression assumptions may be stated in terms of the response variable or in terms of the error terms. The statistical model for simple linear regression assumes that for each value of x, the observed values of the response are normally distributed with some mean (that may depend on x in a linear way) and a standard deviation σ that does not depend on x. **For each x, Y is N(E(Y), σ), where E(Y) = $\beta_0 + \beta_1 x$.**

Thinking about the error terms, we can say the **true error terms** (those that we do not observe) are the difference of the response and the true mean (for a given x). These errors are to have a normal distribution with a mean of 0 and a standard deviation of σ (that does not depend on x).

***Task*:** Create diagnostic plots for regression and check that assumptions are met.
To create diagnostic plots for regression, go to **Models > Graphs > Basic diagnostic plots**.

1. The graph generated in the upper left shows a scatterplot of residuals on the y-axis against fitted y values on the x-axis. This is called a residuals vs fitted plot. Sketch the general pattern of the plot.

   What assumption of the error terms does this plot help assess?

   What conclusion can you draw from this plot?

2. The graph generated in the upper right shows a QQ plot of residuals.
   Sketch the general pattern of the plot.

   What assumption of the error terms does this plot help assess?

   Based on the plot what is your conclusion about this assumption?

Cool Down: Check Your Understanding

1. Which is the only test statistic you can use if you want to test if the Final Exam score is a significant positive linear predictor of your chosen explanatory variable?

2. Some other Stats 250 students were interested in predicting their final exam score, so you calculated some prediction intervals for your curious friends. Some of your friends Exam 2 scores are given below. Which score will have the narrowest prediction interval? (Hint: You may need to run some summary statistics to do this)

   52 ***55*** ***58*** ***61*** ***64*** ***67***