Stat 230 - Formulas and Details on Inferences for SLR

P.B.Matheson adapted from Prof. Wagaman

1. t-test for slope (2.1) and Intervals for Predictions (2.4)
2. Partitioning Variability in Regression ANOVA (2.2)
3. t-test for Correlation (2.3) and more about Rsquared

3 questions

1 - Is X a significant predictor of Y? (t-test for slope)

2 - Is my SLR model significant? (ANOVA for Regression - F test)

3 - Is there a significant relationship between X and Y? (t-test for correlation coefficient)

IN SLR these are all the same question; this changes in Multiple Linear Regresssion (MLR) where we have more than one predictor.

Question 1: t-test for Slope (Is X a significant predictor of Y?)

To recap, the theoretical model we are using in SLR is

$$Y = \beta_0 + \beta_1 X + \epsilon,$$

where the conditions include that ϵ is distributed as $N(0, \sigma_\epsilon)$. (Remember this is notation to summarize the normality, zero mean, and constant variance conditions.) The β's and the σ_ϵ are model parameters here.

When we get the fitted model, we write it as:

$$\hat{Y} = \hat{\beta_0} + \hat{\beta_1} X,$$

where the $\hat{\beta}$'s are estimates of the β's and σ_ϵ is estimated by $\hat{\sigma_\epsilon}$, our residual standard error.

The inference procedure we have focused on in intro stats is a t-test for the slope coefficient, β_1. We usually test for whether this coefficient is equal to 0 or not (a two-sided test).

The hypotheses would be: $H_0 : \beta_1 = 0$ vs. $H_A : \beta_1 \neq 0$.

The conditions to perform the test are the same as those outlined for regression generally (technically the last 2 in the book are needed in addition to the first 4, which are for generally fitting the line). With those conditions met, the sampling distribution of the sample slope has a t-distribution, which is the theoretical basis for the test.

The test statistic is $t = \hat{\beta_1}/SE(\hat{\beta_1})$, and the p-value is found based on a t distribution with $n - 2$ degrees of freedom. This is the default output in most computer packages, including R.

While R gives you the value of t and it's associated p value. You can find p values (the exact probability of getting a t value as large as the one you calculated) and t critical values (the size of t required for your test to be significant) by using the following r commands:

t critical values For a ONE SIDED test (is Beta1 <0?), use qt(alpha, df, lower.tail=FALSE) where alpha usually equals 0.05, df equals n-2).

For a TWO SIDED test (is Beta1 <> 0?), use qt(alpha/2, df, lower.tail=FALSE).

exact p values for a given t value (t observed or your calculated t test statistic) pt (q, df, lower.tail=TRUE) where q = your t value.

Remember that coefficients from our model are estimates; if we took another sample we would get different estimates. The variability in model coefficients can be described by confidence intervals.

To construct a confidence interval, the multiplier is drawn from the t distribution with $n-2$ df, and the estimate and standard error of the estimate are $\hat{\beta_1}$ and $SE(\hat{\beta_1})$ respectively.

It is possible to do this test one-sided (e.g., only interested if beta is >0) or for different null values (not zero), in which case, use a CI instead of re-doing the test procedure.

This test will continue to be used in multiple linear regression but will have a slightly different use (TBD later).

More about intervals (confidence and prediction): You can generate confidence intervals for mean responses (for example, what are typical values for the average response at this x value?) and prediction intervals for individual responses (for example, what are typical values for the response at this x value?). The formulas are provided in the text. We can use the predict function in R to have the computer do the computational work.

Important concepts:

- CIs for mean responses are narrower than prediction intervals for individual responses when considering the same predictor value and confidence level (see Figure 2.3 on pg 72 in book).

- Both types of intervals get wider as the X value under consideration moves further from the mean of the X variable.

- We should use these intervals because we know there is additional variability in the predicted value from the model that we can pass on information about. In other words, these can be better for someone to use than just the single $\hat{y}$ from the model.

R will give you the 95% CI using confint(fm) where fm is the name of your model (could be different) or

Question 2: ANOVA for Regression (Is my SLR model significant?)

ANOVA stands for Analysis of Variance. There is an ANOVA procedure that allows you to test for a significant linear relationship in SLR (simple linear regression). The idea is that the total variance in the response can be broken down into a piece explained by the model and a piece that is leftover in the residuals. The more variance we can explain using the model, the better.

$$\text{Total Variation in Y} = \text{Variation Explained By Model} + \text{Unexplained Variation in Residuals}$$

$$\sum(y-\bar{y})^2 = \sum(\hat{y}-\bar{y})^2 + \sum(y-\hat{y})^2$$

$$SSTotal = SSModel + SSE$$

We standardize each sum of squares into mean squares, and take their ratio to obtain an F test statistic. F refers to the distribution of the test statistic under the null hypothesis that the model is not useful for predicting the response (all predictor slopes are 0; in SLR, just that the predictor slope is 0).

$$F = \frac{MSModel}{MSE} = \frac{SSModel/1}{SSE/n-2}$$

The appropriate p-value can be obtained from an F distribution with 1 and $n-2$ degrees of freedom, provided the SLR conditions are met. These df for the F distribution will be different when we move to MLR (multiple linear regression).

While R gives us the F value and it's corresponding p value in the anova(fm) command, we can get the actual probability of a given F statistic or the critical value of F that can be used for the cutoff of what is significant.

For the probability to the left of F use qf(alpha, dfnum, dfdenom, lower.tail=TRUE)

if you want the actual probability of getting the F ratio you obtained just subtract it from 1. 1 - qf(alpha, dfnum, dfdenom, lower.tail=TRUE)

if the p value is less than your a priori stated alpha, your F test is significant, the model is significant.

NOTE: This test procedure is used differently than the t-test for an individual slope once we get to MLR. We will use it to test for whether any individual predictor slope is different from 0, and then use the individual t-tests to narrow down which predictors are useful for modeling the response. For SLR though, these procedures are equivalent. You may hear the F test referred to as the overall model utility test, overall F-test, test for overall model effectiveness or ANOVA for regression, or something similar. We will learn about different F-tests for MLR.

Question 3: t-test for rho (Is there a significant relationship between X and Y?

More about testing correlation coefficients and computing R-squared In SLR, the R-squared is the proportion of variability in the response (Y) explained by its linear relationship with the predictor variable (X). It is computed based on the breakdown of variance from the ANOVA and also happens to be the correlation squared (just for SLR).

$$Rsquared = \frac{SSModel}{SSTotal} = 1 - \frac{SSE}{SSTotal}$$

Larger values of R-squared indicate a better model fit (ranges from 0 to 1).

You can test for whether or not the population correlation coefficient ρ is non-zero between the predictor and response using the sample correlation coefficient, r. The conditions are the same as for running inference for regression. The test yields a t test statistic where the p-value is found based on a t distribution with $n-2$ df if the SLR conditions are met.

$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$$

NOTE: In the case of SLR, this test is equivalent to the ANOVA F-test for a significant relationship and the t-test for slope. When we move to MLR, these 3 procedures will be used differently. The test for correlation will always be just a pairwise test examining the relationship between any two quantitative variables.