Linear Regression: Inference

Reading: Ch. 8.4

We are going to focus on inference for linear regression for this activity. Outliers (Ch. 8.3) are not a focus of the exam, so we will not be doing additional practice. You should be aware that they exist and that they can only be thrown out in special cases.

Inference Overview

When we talk about inference in the context of linear regression, we are thinking about the uncertainty associated with our parameter estimates. In the case of simple linear regression (a single predictor) in particular, we are interested in the uncertainty associated with the estimate of the "slope" ($\hat{\beta}_1$), which is the coefficient associated with our predictor (explanatory variable).

Review: Anatomy of a Simple Linear Regression Model

In a simple linear regression model, we assume that the relationship between two variables, x (explanatory/predictor) and y (response) is linear:

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i, \ i = 1, ..., n$$

Once the model has been fit, we get estimates of the parameters β_0 and β_1, which are $\hat{\beta}_0$ and $\hat{\beta}_1$, respectively. The fitted model has the form

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i, \ i = 1, ..., n$$

Inference for Regression Parameters

Here is some example R output from Lab 3. This is for the model where we predict first year college GPA (FYGPA) from high school GPA (HSGPA).

```
Call:
lm(formula = FYGPA ~ HSGPA, data = satGPA)

Residuals:
     Min       1Q   Median       3Q      Max
-2.30544 -0.37417  0.03936  0.41912  1.75240

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.09132    0.11789   0.775    0.439
HSGPA        0.74314    0.03635  20.447   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.6222 on 998 degrees of freedom
Multiple R-squared:  0.2952,Adjusted R-squared:  0.2945
F-statistic: 418.1 on 1 and 998 DF,  p-value: < 2.2e-16
```

We have already discussed some aspects of this output, primarily the `Estimate` column in the `Coefficients` portion of the output. We will employ the rest of the output (`Std. Error`, `t value`, and `Pr(>|t|)` for performing inference.

The summary output from the linear model automatically conducts some hypotheses for us. Specifically, the test statistics (`t value`) and p-values (`Pr(>|t|)`) correspond to the following tests:

- **Inference for Slope**
  - Hypotheses: $H_0 : \beta_1 = 0$ versus $H_A : \beta_1 \neq 0$
  - t-test with $n - 2$ degrees of freedom
  - Recall using the output above, the estimate, $\hat{\beta}_1$ is denoted by the name of the explanatory variable. Here it is HSGPA.
  - This is a test for **an association between the predictor and the response variables**.
    - If the p-value is greater than α, then there is insufficient evidence of a relationship between the two.
    - If the p-value is less than α, then there is sufficient evidence of a relationship between the two. Furthermore the sign of the test statistic (`t value`) tells us whether the relationship is positive or negative.
  - Note: we could also conduct these tests using confidence intervals, but we are not going to focus on that in this class.
- **Inference for Intercept**
  - Hypotheses: $H_0 : \beta_0 = 0$ versus $H_A : \beta_0 \neq 0$
  - t-test with $n - 2$ degrees of freedom
  - Recall using the output above, the estimate, $\hat{\beta}_0$ is denoted (Intercept).
  - Note: we are going to focus on inference for the slope, but you should be aware that we could conduct inference for the intercept, too.

Recall the following problem:

Predicting Success in College. During the college admissions process, many factors are considered to assess whether an applicant will be successful in college. Specifically, these factors are assumed to be good determinants of how an applicant will perform in the first year of college (assessed by GPA). Common metrics used to predict first year GPA include high school GPA and SAT score. In small groups, you will explore the relationships between first year GPA (units: first year GPA points) and high school GPA (units: high school GPA points), and first year GPA and SAT score (units: SAT points). The data set you will be using is available through the openintro library.

- **Problem 1:** Model relationship between first year college GPA and high school GPA.
- **Problem 2:** Model relationship between first year college GPA and SAT score.

General Instructions

1. Continue this exercise in your R Markdown file for Lab 3. Use the model that you fit previously to these data (either Model 1 or Model 2).

2. Identify the response variable and the predictor.

3. Write out the linear model.

4. If you are working on Problem 1, suppose you have a high school GPA of 3.5. If you are working on Problem 2, suppose you have an SAT score of 120. Calculate the predicted first year GPA.

5. Suppose the true first year GPA is 3.2. What is the residual? Does your model overestimate or underestimate this first year GPA?

6. Using the regression output, conduct a hypothesis test to determine whether there is convincing evidence that your predictor (either high school GPA or SAT score) is a useful predictor of first year GPA. Do this using:
   (a) A t-test and p-value ($\alpha = 0.05$)
   (b) A confidence interval (confidence level of 0.95)

   In both cases, you should be sure to draw conclusions and interpret them in the context of the problem. Note, the standard error for your confidence interval can be obtained from the `Std. Error` column of the coefficients summary output from your linear model.

   Note: in (a) and (b) we use the t-distribution because the standard deviation is unknown.

7. Why are the associated degrees of freedom for this test 998?

8. What is the value of R^2 (Multiple R squared in the regression output)? Interpret this value. Do you find this to be compelling evidence that your predictor sufficiently explains the variability in first year college GPA?