Lab 1: Simple linear regression

1 Review

We have discussed simple linear regression. The goal of regression is to describe the behavior of some response variable by the behavior of predictor (or explanatory) variables. For simple linear regression, we have n independent subjects with a single predictor. We will extend these methods to incorporate more predictor variables later. Here are some review questions about simple linear regression.

1. What is the model that we fit?

Estimation

2. What are two methods for fitting this model?

3. What assumptions do these methods make?

4. How do they differ and how are they the same?

Inference

5. What test do we use to test if the coefficient β_1 (the regression coefficient of the predictor) is different from 0? Why would we want to test if β_1 is equal to 0?

Confidence and Prediction Intervals

6. Confidence interval is for the *true mean value of the response variable* given a specific value of the predictor. The prediction interval is for the *individual response variable value* given a specific value of the predictor, and thus involves additional variability compared to the confidence interval and is much wider than the confidence interval.

2 Example

We will use the low birth weight data set as the example. Recall that the data set contains information from a random sample of 100 low birth weight infants born in Boston, MA in 1990s. The response (outcome) variable of interest is `headcirc`, head circumference measurements in centimeters. Other variables in the dataset are described in the following table. The data set named `lbw.sas7bdat` is posted on course web page in Moodle.

Name	**Variable**
birthwt	birth weight, in grams
length	infant length in centimeter
momage	Age of the Mother in Years
gestage	gestational age in weeks
toxemia	mother's diagnosis of toxemia during pregancy 1=Yes, 0=No.

1. First, Numerical summary of the data.

2. Then, We will begin by analyzing the effect of `gestage` on `headcirc`.

(a) Draw a scatter plot of `gestage` versus `headcirc`.

(b) What is the model we will fit?

(c) What is the estimate for the effect of `gestage` on `headcirc`? How do you interpret this?

3. We will now proceed to make inferences based on the fitted model.

(a) Perform the appropriate t-test to determine if there is a significant relationship between `gestage` and `headcirc`?

(b) How does this compare to the F-test result given in the output?

(c) What if we were only interested in testing if increased `gestage` lead to an increase in `headcirc`? Perform this test.

(d) Find a two-sided 95% confidence interval for β_1, the regression coefficient of `gestage`.

(e) Find a two-sided 95% confidence interval of the mean value of `headcirc` for those with a `gestage` of 33 weeks.

(f) How do you interpret this interval?

(g) Calculate the prediction interval of `headcirc` for a future observation with `gestage` of 33 weeks.

(h) How do you interpret this prediction interval?