This week we will consider regression analysis. We will examine relationships between paired numerical variables. For example, suppose we are interested in comparing age and income. Each person in the study has one number for age and one number for income. Since the numbers came from the same person, those numbers are paired together.

Last week we used scatterplots to display this type of data and correlation to measure how strong the relationship is. This week we will use regression to describe the relationship. Then we will see how such a relationship can be used to predict one variable based on the other.

**Linear Regression**

1. Load again the grades data from last week. We have determined that there is a weak linear relationship between HW Ave and Final. Let's describe that relationship with a mathematical function that can be used to predict the Final grade.

*gradesdata2 <- gradesdata |> filter(is.na(FinalExam)==FALSE)*

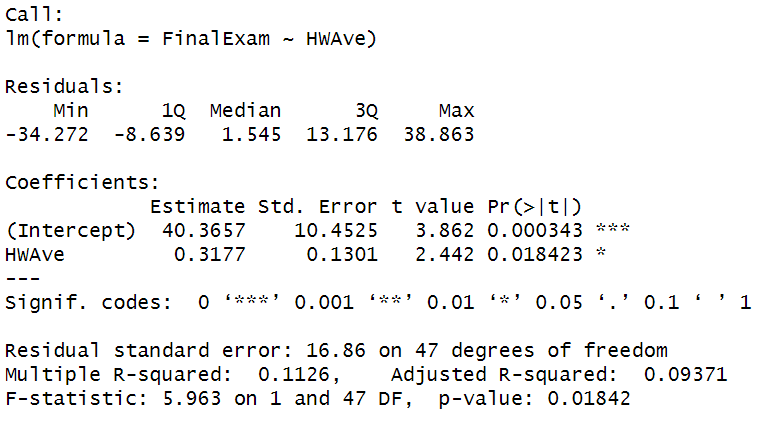
*attach(gradesdata2)*

*lm.grades<-lm(FinalExam~HWAve) #creates a linear regression model*

*#lm stands for linear model*

*summary(lm.grades)*

Paste the output below.



There is a lot to talk about with this command. First, notice the expression *FinalExam ~ HWAve* in the argument of the *lm()* command. This is telling R to think of a model with FinalExam as an output, and HWAve as an input. Because we are using the function lm(), we are looking at a linear model. (lm stands for linear model) So we are fitting a straight line to the data. By “fitting,” we mean finding the “best” and for the following equation of a line:

Based on the output you should have gotten above, the regression equation is FinalExam = 40.3657 + 0.3177 HWAve . These numbers are under “Estimate” in the “Coefficients:” section of the output. There are two numbers in this equation. The first, 40.3657, is the intercept. You can see it in the “Intercept” row of the table labeled “Coefficients.” It is interpreted as the value of the response if the predictor is zero. In this example, if you got a zero homework grade, we predict that your final exam grade would be 40.3. In some examples, the intercept makes no sense. For example, a diamond with zero carats doesn't exist. Similarly, a baby with zero length doesn’t exist. So in a model that relates carats of a diamond to price or a model that relates the length of a baby to the weight, the intercept wouldn’t make any sense. But in this model, it is easy to see what a student with a score of zero on homework would be, so the interpretation of the final exam grade for that student makes sense. So the interpretation of the intercept depends on what your data look like.

The second number, 0.3177, is the slope. You can see it in the “HWAve” row of the same table. It is interpreted as the amount the response changes if the predictor changes by 1. In this example, if we increase the HWAve by one point, we expect the FinalExam grade to increase by 0.3177 points.

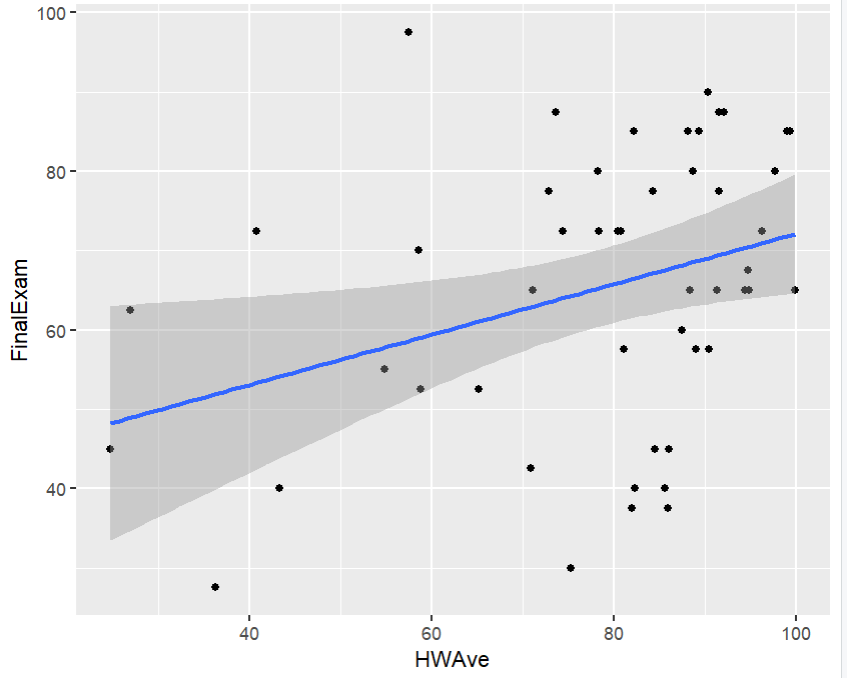
2. Let's look at a graph of the regression equation on the scatter plot of the data.

*ggplot(gradesdata2, aes(x=HWAve, y=FinalExam))+*

*geom\_point()+*

*geom\_smooth(method=`lm`)*

Paste the graph below.



The geom\_smooth command also adds confidence bands around the fitted line. We will talk about those bands later.

We can use the regression equation to make predictions. If we want to predict the final exam grade for someone who made a 70 on the homework, we replace the HWAve variable with 70, getting Final Exam = 40.3657 + 0.3177 \* 70 = 62.6047.

3. What final exam grade do we expect someone who got an 85 homework average to get?

67.3702

Of course, you can have the computer do the predicting for you.

*newdata <- data.frame(HWAve=70)*

*# the predict function requires that the new input be in a data frame, so the*

*# function data.frame forces the HWAve=70 to be interpreted as a data frame.*

*predict(lm.grades,newdata)*

*predict(lm.grades,newdata,level=0.95,interval="prediction") #adds a prediction interval*

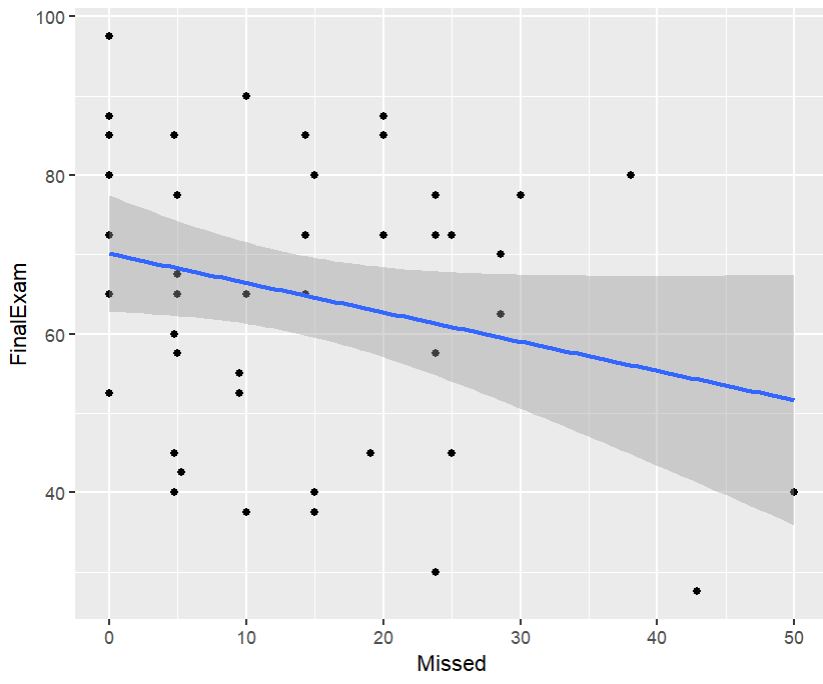
Notice we get almost the same value that we got above. The predict() function is not rounding the estimates of the slope and intercept as much as I did above, so we get a slight difference due to having more decimal places. By adding the two extra arguments of *level* and *interval*, we get an interval that describes the uncertainty in that prediction. The interval is called a "prediction interval", and we are 95% sure that the actual value will be within this interval. Notice our prediction interval is pretty wide. This is because HWAve is actually not that great of a predictor for the FinalExam grade, which we observed before when we looked at the scatterplot.

4. What is the 95% prediction interval for the final exam grade for someone who got an 85 on homework?

67.36854

**Percent Missed**

5. Create a linear model of the Final Exam vs. Percent Missed. (You will have to create a variable for Percent Missed as we did in the last assignment.) What is the slope of the line? Paste your code as well as your answer here.

   
gradesdata2 <- mutate(gradesdata2, Missed = 100-Attendance)

attach(gradesdata2)

lm.grades<-lm(FinalExam~Missed)

ggplot(gradesdata2, aes(x=Missed, y=FinalExam))+

geom\_point()+

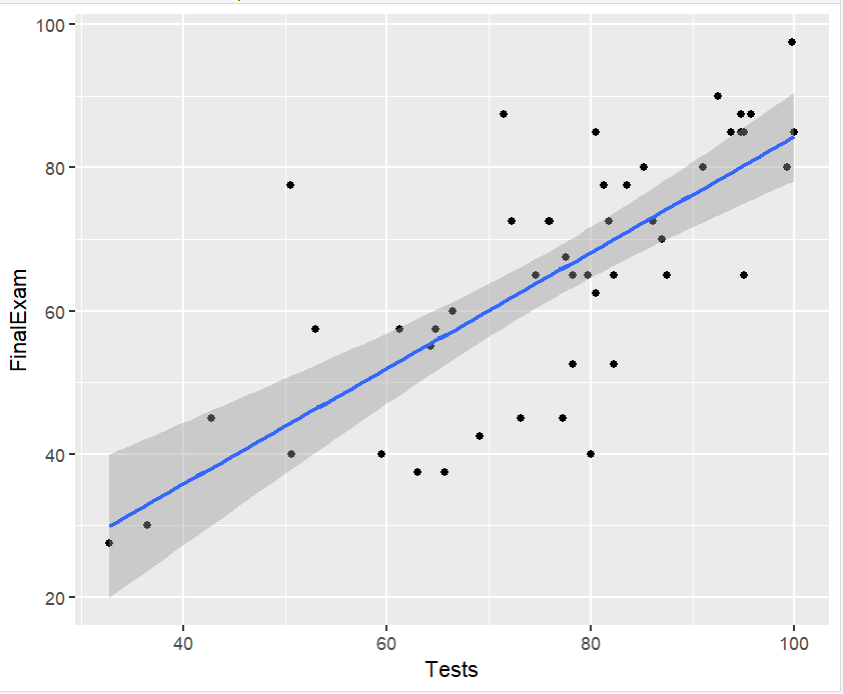
geom\_smooth(method=`lm`)

6. What would the model predict the final exam to be for someone who missed 15% of the class? What would a 95% prediction interval for the final exam grade be? Paste your code as well as your answer here.

64.5818

**Tests**

7. Create a linear model of the Final Exam vs. Tests. What is the slope of the line? Paste your code as well as your answer here.



lm.grades<-lm(FinalExam~Tests)

ggplot(gradesdata2, aes(x=Tests, y=FinalExam))+

geom\_point()+

geom\_smooth(method=`lm`)

8. What would the model predict the final exam to be for someone who had a test average of 75? What would a 95% prediction interval for the final exam grade be? Paste your code as well as your answer here.

newdata3 <- data.frame(Tests=75)

predict(lm.grades,newdata3,level=0.95,interval="prediction")

64.12753

**Comparing the Models**

9. Compare the lengths of the prediction intervals (biggest point minus smallest point) for the three models for predicting the final exam that you have created. Which model gives the shortest prediction interval?

The missed model has the shortest prediction interval since there wasn’t any entry greater than 50 while the other two are twice that

It makes sense that a model that uses both the HW average and the percent of class missed would be more accurate than one that uses only one predictor variable. We will discuss multivariate regression and how to measure how good a model is next week.