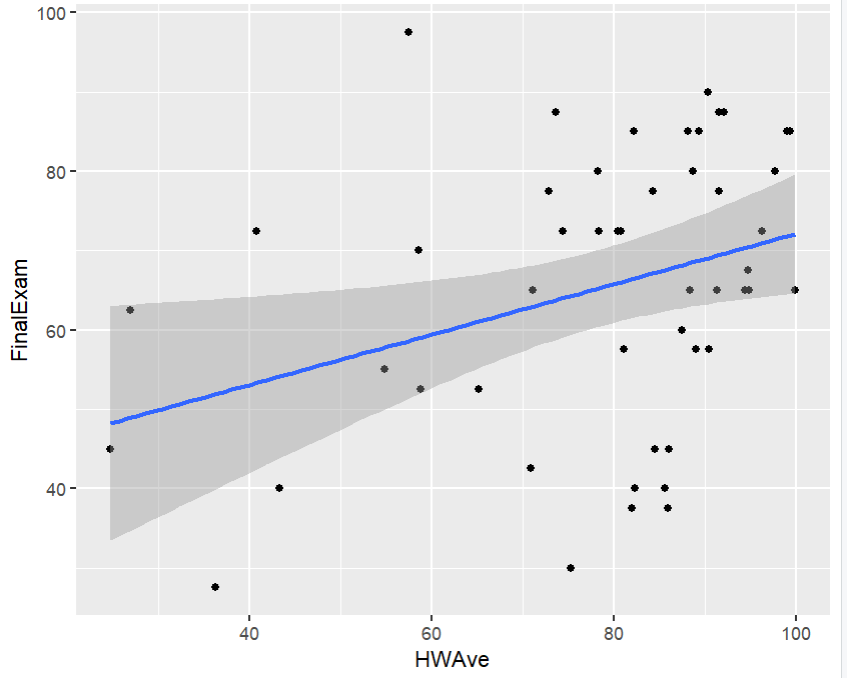
Last week we studied regression analysis with one variable. Now we will look at a regression example with two variables. The model we are using will be

Y = a1X1 + a2X2 + b + error

**Scatter plots**

1. Load and attach the file gradesdata.csv. We used this file last week to look at final exam grades as predicted by HW averages, tests, and attendance. Create scatterplots, one with HW average on the x-axis, and final exam grades on the y axis; one with tests on the x-axis and final exam on the y-axis; and one with attendance on the x-axis and final exam grades on the y axis. You did this last week.

Paste your graphs below.



2. It makes sense that a model that both HWAve and Tests would be more accurate than one that uses only one predictor variable. Let’s first look at the models that only have one predictor:

*lm.gradesHW<-lm(FinalExam~ HWAve) #saw this last week*

*summary(lm.gradesHW)*

*lm.gradesTests<-lm(FinalExam~ Tests) #using Tests*

*summary(lm.gradesTests)*

Paste the summaries here:

> lm.gradesHW<-lm(FinalExam~ HWAve)

> summary(lm.gradesHW)

Call:

lm(formula = FinalExam ~ HWAve)

Residuals:

Min 1Q Median 3Q Max

-34.272 -8.639 1.545 13.176 38.863

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 40.3657 10.4525 3.862 0.000343 \*\*\*

HWAve 0.3177 0.1301 2.442 0.018423 \*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 16.86 on 47 degrees of freedom

Multiple R-squared: 0.1126, Adjusted R-squared: 0.09371

F-statistic: 5.963 on 1 and 47 DF, p-value: 0.01842

> lm.gradesTests<-lm(FinalExam~ Tests)

> summary(lm.gradesTests)

Call:

lm(formula = FinalExam ~ Tests)

Residuals:

Min 1Q Median 3Q Max

-28.171 -4.991 1.290 7.400 33.187

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 3.4712 8.3623 0.415 0.68

Tests 0.8088 0.1072 7.543 1.23e-09 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 12.03 on 47 degrees of freedom

Multiple R-squared: 0.5477, Adjusted R-squared: 0.538

F-statistic: 56.9 on 1 and 47 DF, p-value: 1.233e-09

3. Now we want a model that uses both variables.

*lm.gradesboth<-lm(FinalExam~ HWAve+Tests)*

*summary(lm.gradesboth)*

Paste the summary here:

> lm.gradesboth<-lm(FinalExam~ HWAve+Tests)

> summary(lm.gradesboth)

Call:

lm(formula = FinalExam ~ HWAve + Tests)

Residuals:

Min 1Q Median 3Q Max

-28.387 -4.040 1.042 7.199 32.924

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 1.85635 9.48924 0.196 0.846

HWAve 0.03823 0.10267 0.372 0.711

Tests 0.79075 0.11852 6.672 2.83e-08 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 12.15 on 46 degrees of freedom

Multiple R-squared: 0.549, Adjusted R-squared: 0.5294

F-statistic: 28 on 2 and 46 DF, p-value: 1.111e-08

**Coefficients**

4. In the summaries, you have values of the coefficients of the variables. So, in the HWAve-only model, the coefficient of HWAve is 0.3177. This is the slope, and it means that if the HWAve value increases by 1, the Final Exam grade will increase by .3177. What is the coefficient in the Tests-only model, and what does it mean?

0.8088, it means that if test scores increase by 1 final exam scores will increase by 0.8088

5. In the model with both predictors, there are two coefficients, one for HWAve, and one for Tests. The one for HWAve tells us that if HWAve increases by one, and Tests stays the same, FinalExam will increase by this amount. What are the values for each variable? Interpret them.

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 1.85635 9.48924 0.196 0.846

HWAve 0.03823 0.10267 0.372 0.711

Tests 0.79075 0.11852 6.672 2.83e-08 \*\*\*

Going from left to right by column the first value is the estimated value for the coefficients, then the standard error is next which measures the stddev of those estimates, the t-value is next and the larger it is the more it suggests the coefficient is decently higher than 0, lastly the Pr(>|t|) column and this one is just our p-value.

Since the tests t-value is high for Tests then we can assume it has a coefficient higher significantly greater than 0, the miniscule Tests p-value compared to HWAve’s shows that it is a much greater predictor of Final Exam scores.

**Predictions**

6. We can make predictions of the final exam grades with each model. Last week, we predicted the final exam grade if the HWAve was 70:

*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.gradesHW,newdata)*

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

Use the model with only Tests to predict the final exam grade if the test average was 85%. Give a 95% prediction interval. Paste the output here.

> newdata4 <- data.frame(Tests=85)

> predict(lm.gradesHW,newdata4)

1 2 3 4 5 6 7 8

65.93665 69.62774 63.52726 71.93190 69.10582 68.52317 59.03069 70.95525

9 10 11 12 13 14 15 16

66.41555 48.90933 71.40883 67.54621 66.13928 70.48989 57.79724 65.20949

17 18 19 20 21 22 23 24

64.00141 63.76610 58.98134 69.04648 51.88886 64.27243 69.37486 53.32227

25 26 27 28 29 30 31 32

62.89110 67.66297 70.35972 67.14051 62.96112 69.45352 65.23925 68.72965

33 34 35 36 37 38 39 40

48.22929 67.70243 58.63732 68.43802 70.44159 68.62675 68.35109 71.82387

41 42 43 44 45 46 47 48

68.15255 69.43230 54.11683 61.07161 66.03260 66.51873 72.10685 66.47186

49

67.20041

7. When we have a model with two variables, we need to give the values of both variables in order to make a prediction.

*newdata <- data.frame(HWAve=70,Tests=85)*

*predict(lm.gradesboth,newdata)*

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

Paste the output here.

> newdata5 <- data.frame(HWAve=70,Tests=85)

> predict(lm.gradesboth,newdata5)

1

71.74605

> predict(lm.gradesboth,newdata5,level=0.95,interval="prediction") #adds a prediction interval

fit lwr upr

1 71.74605 46.8472 96.6449

**Which model is better?**

8. We’d like to compare the models and decide which one is better. R2 measures the percent of variation in Y that is explained by the relationship with X. In the multiple regression case, it is the percent of variation in Y that is explained by the model that includes both X values. We will look in particular at the adjusted R2 value. It has been adjusted for the number of predictors to allow us to compare models with different numbers of predictors. What is the adjusted R2 for each of the three models?

Model R2 value

HW only 0.09371271

Tests only 0.5380299

Both 0.5294051

9. The model with the highest adjusted R2 is the one that explains the variation in final exam grades the best. Which model is that?

Tests only

10. Based on the R2 of the “best” model, is all the variation in final exam grades explained now? If we could collect any data we wanted on the students in the class, brainstorm at least two variables that we might want to include in a model to get a better fit.

I hypothesize that the students overall GPA, gender, race, or economic background could all be variables that evaluate academic performance.

11. Load and attach the dataset in NFL2007Standings. Create a multiple regression model that predicts the winning percent (WinPct) based on the points the team scored (PointsFor) and the points that were scored against them(PointsAgainst). What are the coefficients for the two predictors? Interpret these numbers.

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 0.4172230 0.1394480 2.992 0.00561

PointsFor 0.0017662 0.0001870 9.445 2.37e-10

PointsAgainst -0.0015268 0.0002751 -5.551 5.50e-06

------------------------------------------------------------------------

The variables both have low p-values meaning they’re statistically significant and they also both have extremely low estimates and stderror scores. Also, the t-value for PointsFor is high indicating that the coefficient is much higher than 0.

12. Predict the win percentage for a team that scored 450 points and had 250 points scored against them.

0.8303038

13. What is the adjusted R2 value for the model you found in number 11? How do we interpret that number?

0.8763931

There a many parts of a regression model that we haven’t talked about. For example, we didn’t check any assumptions to make sure that linear regression was the best choice. We also didn’t talk about how to determine if there is any interaction between the predictors. Regression is a topic that has a lot of substance. STAT 4360 is a course on regression here at MTSU.