Test 1

Acts 170

Solution

Use the data file ‘GradSurvey’. A list of questions that were asked to graduate students is below.

1.)

> grad <- read.csv(choose.files(), header=TRUE)

> attach(grad)

> dim(grad)

[1] 44 15

a. Compute descriptive statistics for UGPA and Texts. Looking at their mean and median, what type of skewness do you expect and why?

> summary(cbind(UGPA, Texts))

UGPA Texts

Min. :2.800 Min. : 0.0

1st Qu.:3.075 1st Qu.: 27.5

Median :3.400 Median : 100.0

Mean :3.368 Mean : 232.5

3rd Qu.:3.625 3rd Qu.: 362.5

Max. :3.900 Max. :1250.0

The mean and median for UGPA are almost equal, implying that they are symmetric.

The mean > median for Texts, so it should be right skewed.

b. Create boxplots, one on top of the other for UGPA and Texts. Does the shape of the distributions correspond with what you expected above? Comment on anything else noticeable in the plots.

> par(mfrow=c(2,1))

> boxplot(UGPA, horizontal=TRUE)

> boxplot(Texts, horizontal=TRUE)

UGPA is almost completely symmetric as expected above.

Texts is right skewed since the right side of the box is larger and the right whisker are longer. There are also a couple of large outliers on the right end.



c. Create a side by side histogram and a (normal) qq plot for Salary. Comment on each plot.

> par(mfrow=c(1,2))

> hist(Salary)

> qqnorm(Salary)

> qqline(Salary)



Salary is right skewed in the histogram and not normally distributed in the qq plot since several points veer far away from the line in the top of the qq plot.

d. Take a log transformation of Salary and make a side by side histogram of the original variable and the log one. Comment on the plots.

> lnSalary = log(Salary)

> hist(Salary)

> hist(lnSalary)



Both the original Salary and the lnSalary are right skewed, but the log version a little less so.

e. Make boxplots of Undergraduate GPA by their Undergraduate major. Comment on the plots.

> boxplot(UGPA ~ UMajor)



UGPA for Biological Sciences and Engineering have about the same spread and median, as do Business and Other. Biological Sciences and Engineering are right skewed and Other and Business are pretty symmetric. There are no outliers in any of the plots.

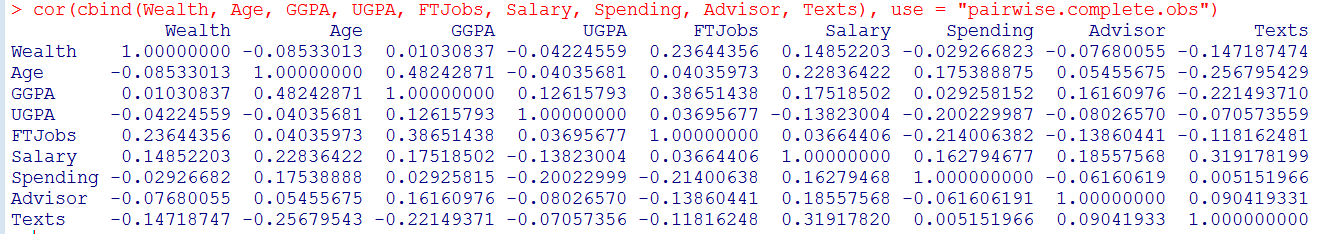
f. Make a scatterplot of undergraduate GPA and Salary, with Salary as the y variable. One student, Sebastian, has a very unusual expectation of his future Salary. Put his name close to his point as a label on the plot.

> plot(UGPA, Salary)

> text(3.4, 290, "Sebastian")



g. Produce a table of correlations for all of your quantitative variables. What two variables are most highly correlated with Wealth?

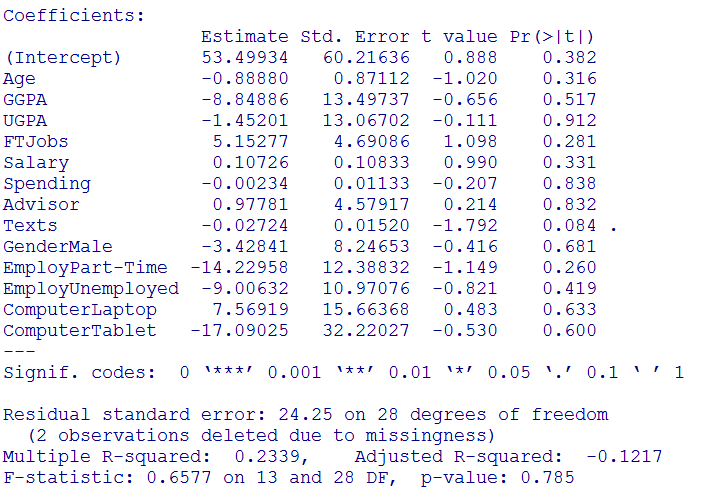


FTJobs and Salary are the 2 most highly correlated variables with Wealth at 0.236 and 0.148.

h. Fit a regression model using Wealth as the outcome variable and all of the quantitative variables, as well as Gender, Employ, and Computer as the explanatory variables. Comment on R² and Ra². Interpret Ra².

> model1 = lm(Wealth ~ Age + GGPA + UGPA + FTJobs + Salary + Spending + Advisor + Texts + Gender + Employ + Computer)

> summary(model1)



R² is 0.2339 and Ra² is -0.1217. They are both low, which is bad. They are both far apart, which is also bad. Ra² is also negative, which is unusually terrible which means that the penalty for taking the explanatory variables is far outweighing the benefit in explaining Wealth they are providing.

Ra² means that -12.17% of the variation in Wealth is being explained by the explanatory variables once we have taken a penalty for the number of X variables included.

i. Do a hypothesis test to determine whether Advisor rating is a significant predictor of Wealth. Provide hypothesis statements, t test statistic, t critical value, and a decision. Use alpha = 0.05.

> # df = n-(k+1) = 44-(13+1) = 30 – 2 observations due to missingness = 28

> qt(0.975, 30)

[1] 2.042272

> qt(0.975, 28)

[1] 2.048407

H0: β7 = 0

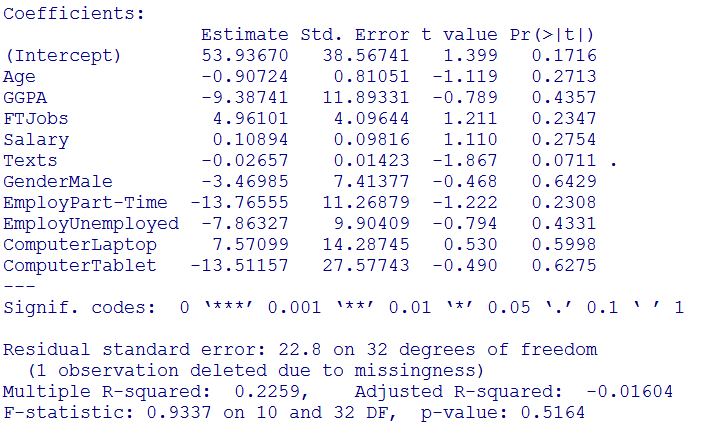
H1: β7 ≠ 0

Since our t test statistic = 0.214 < t critical value = 2.048407, we do Not Reject Ho. This means we believe that Advisor is not a significant predictor of Wealth.

j. Build a new model removing the three variables UGPA, Spending, and Advisor since they were especially terrible. Compare this new model’s s to the s from the first model. Did you expect this change given the p-value of the variables that you dropped? Explain.

> model2 = lm(Wealth ~ Age + GGPA + FTJobs + Salary + Texts + Gender + Employ + Computer)

> summary(model2)



The old s = 24.25, the new s = 22.8. The new s is better because it is a lower error, I did expect a lower overall error since I dropped 3 insignificant values which were not helping to improve overall model fit and explain y successfully.

k. Using the model from part (j) and an alpha = 0.10, which variables are significant? How do you know?

At alpha = 0.10, only Texts is significant with a p-value = 0.0711. All others are insignificant.

l. Using the model from part (j), interpret the slope for the full time jobs variable.

For every 1 full time job added, a person’s level of Wealth to feel rich increases by 4.961 million dollars, holding all other X variables constant.

m. Using the model from part (j), interpret the slope for the gender variable.

If someone is a Male, their level of Wealth to feel rich decreases by 3.470 million dollars when compared to a female, holding all other X variables constant.

n. Suppose you have a 30 year old female, who has a graduate GPA of 3.5, has had 2 full time jobs, sends 100 texts a week, an expected salary of 90 thousand dollars, is currently employed part time, and prefers to use a laptop. Use the model from part (j) and find a 95% prediction interval for her wealth needed to say she is rich. Interpret this interval. (Hint: You may need to use “ “ for your categorical variables to work. Like “Female”.)

> newdata = data.frame(Age=30, Gender="Female", GGPA=3.5, FTJobs=2, Texts=100, Salary=90, Employ="Part-Time", Computer="Laptop")

> predict(model2, newdata, interval="prediction", level=.95)

fit lwr upr

1 4.738252 -47.6702 57.14671

95% prediction interval = (-47.67, 57.15)

If a person has all of the qualities listed above, we are 95% confident that they have an individual predicted Wealth to feel rich between -47.67 and 57.15 million dollars.

o. What is the difference between a confidence interval and a prediction interval? You do not need to find a confidence interval, just explain to me the difference between what the 2 find.

A confidence interval is the range for what the estimated group average should expect to be.

A prediction interval is the range for what an individual should expect to have happen to them.

p. Produce a scatterplot matrix of all quantitative variables you used from your model in part (j) and Wealth. Is there anything noticeable in your plots? If you could square only 1 term, which one would you choose? (You do not need to actually do this.)

> Xs = data.frame(Age, GGPA, FTJobs, Salary, Texts, Wealth)

> pairs(Xs, upper.panel=NULL)

There is not a strong linear pattern in any of the plots with Wealth. There also seem to be 3-4 high outlier points in all of the plots.

I would square Age or Texts.



q. Take GGPA from the model in part (j) and compute its partial correlation coefficient. What is the interpretation of this coefficient?

> model3 = lm(Wealth ~ Age + FTJobs + Salary + Texts + Gender + Employ + Computer)

> model4 = lm(GGPA ~ Age + FTJobs + Salary + Texts + Gender + Employ + Computer)

> e3 = residuals(model3)

> e4 = residuals(model4)

> cor(e3, e4)

[1] -0.1381915

The correlation between Wealth and GGPA is -0.1382 once the other explanatory variables have been controlled for.

r. Produce an added variable plot for the partial correlation coefficient above. Comment on the difference of the original correlation and scatterplot and the partial correlation and added variable plot now.

> plot(e3, e4)



We can see that this added variable plot does correspond to the weak negative partial correlation which we found above, whereas we see below that the original correlation is positive, but virtually 0 and also corresponds to its plot which has no obvious linear pattern present.

> cor(GGPA, Wealth)

[1] 0.01030837

> plot(GGPA, Wealth)



s. Take a subset of your original dataset by using only students who have greater than a 3.0 graduate GPA. Redo the model from part (j). Comment on any notable changes.

> grad1 <- subset(grad, GGPA>3)

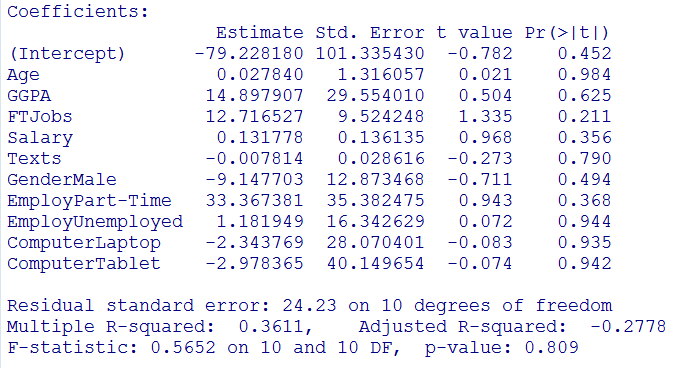
> dim(grad1)

[1] 21 15

> attach(grad1)

> model4 = lm(Wealth ~ Age + GGPA + FTJobs + Salary + Texts + Gender + Employ + Computer)

> summary(model4)



The R² is higher here at 0.3611, as compared to 0.2259 before, which is better.

The R²a is lower here at -.2778, as compared to -.016 before, which is worse.

S has gotten worse by increasing from 22.8 to 24.23.

The overall p-value for model fit has also gotten worse by increasing from 0.5164 to 0.809.

At al alpha = 0.10 we had 1 significant variable in Texts before and there are no significant variables now.

