**STAT 462 – Applied Regression Analysis**

**Fall 2017, Lab 7**

Prepare a short report with relevant output, your comments, and answers to the questions (this does not need to be exhaustive or polished, but should contain enough to show that you completed all tasks and analyses).

Submit the report at the end of the lab session.

Consider the dataset *cement.txt*, that contains data about 13 different cement mixtures. The following variables are measured:

*Aluminum:* amount of Ca3Al2O6

*Silicate*: amount of C2S

*Aluminum\_ferrite*: amount of 4CaO Al2O3 Fe2O3

Silicate\_bic: amount of C3S

*Hardness\_cement*: hardness of cement produced mixing the four components

Build a simple linear regression model in order to predict the hardness of cement given the amount of aluminum.

Compute the 0.50 and 0.90 quantiles of the variable aluminum in the dataset (use the function *quantile*).

* Construct the point estimate for the mean cement hardness corresponding to each of the two quantiles. Which of the two predictions you trust more? Why?

> quantile(Aluminum,probs=c(0.5,0.9))

50% 90%

7 11

> predict(model1,data.frame(Aluminum=c(7,11)))

1 2

94.56058 102.03557

0.5 quantile has mean cement hardness = 94.56058

0.9 quantile has mean cement hardness = 102.03557

> mean(Hardness\_cement)

[1] 95.42308

I trust 0.5 quantile because it has the mean = 95.42308 that is closer to the sample mean.

* Construct a 95% confidence interval for the mean cement hardness corresponding to each of the two quantiles. Compute their width. Why do they have different width?

> predict(model1,data.frame(Aluminum=c(7,11)),interval='confidence',level=0.95)

fit lwr upr

1 94.56058 87.99073 101.1304

2 102.03557 94.30998 109.7612

Confidence interval for mean cement hardness of 0.5 quantile is (87.99073, 101.1304), and it has width=101.1304-87.99073=13.13967

Confidence interval for mean cement hardness of 0.9 quantile is (94.30998, 109.7612), and it has width=109.7612-94.30998=15.45122

They have different width.

This is because different point estimates for each quantile have different variance.

* Construct a 95% prediction interval for the cement hardness of a new observation corresponding to each of the two quantiles. Compute their width and compare them with confidence intervals for the mean cement hardness. Why are they wider than the confidence intervals?

> predict(model1,data.frame(Aluminum=c(7,11)),interval='prediction',level=0.95)

fit lwr upr

1 94.56058 70.05417 119.0670

2 102.03557 77.19436 126.8768

Confidence interval for mean cement hardness of 0.5 quantile is (70.05417, 119.0670), and it has width=119.0670-70.05417=49.01283

Confidence interval for mean cement hardness of 0.9 quantile is (77.19436, 126.8768), and it has width=126.8768-77.19436=49.68244

Compare with confidence intervals for the mean cement hardness, the widths of prediction intervals are larger than the widths of prediction intervals.

This is because when we use the prediction model, we will bring in more error for each predicted value. Thus, the prediction intervals are wider than confidence interval.

R code:

setwd("//udrive.win.psu.edu/Users/j/q/jql5883/Desktop/math462")

getwd()

cement=read.csv("cement.txt",header=T,sep="")

attach(cement)

quantile(Aluminum,probs=c(0.5,0.9))

model1=lm(Hardness\_cement~Aluminum)

predict(model1,data.frame(Aluminum=c(7,11)))

mean(Hardness\_cement)

predict(model1,data.frame(Aluminum=c(7,11)),interval='confidence',level=0.95)

predict(model1,data.frame(Aluminum=c(7,11)),interval='prediction',level=0.95)

The following part will give you 0.5 extra points.

Build a multiple linear regression model in order to predict the hardness of cement given the amount of the four components.

Compute the 0.50 and 0.75 quantile of the four components in the dataset (use the function *quantile*).

* Construct the point estimate for the mean cement hardness corresponding to each of the two quantiles. Compare the estimates with the ones you computed using only the Aluminum. Do you trust more the ones that use all predictors?
* Construct a 95% confidence interval for the mean cement hardness corresponding to each of the two quantiles. Compute their width.
* Construct a 95% prediction interval for the cement hardness of a new observation corresponding to each of the two quantiles. Compute their width and compare them with confidence intervals for the mean cement hardness. Why are they wider than the confidence intervals?