Jonothan Meyer
Linear Regression - Dr. Dyhr - MW 2:4pm
2/18/21
Project 3

Exercise 1

The 'prostate' data set from the Faraway library was choosen for analysis in this project. Cancer volume, transformed logorithmiclly was choosen as the prediction variable, denoted 'lcavol', and prostate weight transformed logorithmically and denoted 'lcp' was choosen as the response variable.

(1) Compute point estimates of Beta1, Beta0, and sigma. Y = Beta0 + Beta1(x) + E

A linear model was made of the predictor and response data using the lm() function in R-Studio. The point estimates of Beta1 = .80, Beta0 = -1.26, and sigma = 1.04.

(2) Identify a predictor variable value in the range of your original data that is not actually sampled in your original data set; call this x*. Construct a point estimate of MewY|x=x*|

4 was choosen as the unknown lcavol value. Plugging x*=4 into our model equation y = -1.26 + .80 * 4 yields a value of 1.94. So the estimated response of having cancer volume at a logarithmic level of 4 will be 1.94 logarithmic units of prostate weight based off of our model.

(3) Identify a confidence level of your choice and construct a confidence interval estimate of Beta1 that uses that confidence level.

Using a 95% confidence interval and the confint(linear model) function on R, we found the confidence interval estimate of Beta1 to be (.62,.98). Meaning we can say that there is a 95% that the true slope for modeling this data lies between .62 and .98 if our assumptions surrounding the model are correct. These numbers were confirmed by doing manual calculation of the confidence interval on R as well.

(4) Using the same confidence level, construct a confidence interval estimate of MewY|x=x*| The estimated reponse of an x*=4 according to our model equation is 1.94. Using R, a direct calculation of a 95% CI at x*=4 was found to be (1.43,2.46) using the following lines of code:

```
xframe<-as.data.frame(4)
colnames(xframe)<-"lcavol"
predict(mv.lm,xframe,interval="conf",level=.95) # This function produced the CI: (1.43, 2.46)
```

This confidence interval was confirmed doing a manual calculation on R by using the necessary formulas that calculate the mean of x, Sxx, response value, sd of residuals, and Syhat. This was done using the following R code:

```
xstar<- 4
mean.x<-mean(prostate$lcavol)
s.res<- sd(residuals(mv.lm)) #1.031255
sxx<-sum((prostate$lcavol-mean.x)^2)
ypred<-b0+b1*xstar
s.xstar<-s.res*(1/n.samp+(xstar-mean.x)^2/sxx)^.5
c(ypred-qt(1-.05/2,n.samp-2)*s.xstar,
```

```
ypred+qt(1-.05/2,n.samp-2)*s.xstar) #<-produces the lower and upper bounds of CI
```

(5) Using the same confidence level, construct a prediction interval for a new response variable value, y|x = x*|.

Using the same methodology I desired to estimate a 95% prediction interval for a response of x*=5. At this response level we can say that 95% of the time the response will be between .57 and 4.91 units of prostate weight or, (.57,4.91) units of lcavol. This was found using the following code from R:

```
xframe<-as.data.frame(5)
colnames(xframe)<-"lcavol"
predict(mv.lm,xframe,interval="pred",level=.95)
```

Exercise 2:
(1) Randomly partition the data set into a model calibration part consisting of 80% of the data set and a model validation component consisting of the other 20% of the data set.

The prostate data was randomly partitioned into a model calibration with 80% of the data contained in a variable called 'development' and a variable called 'holdout' consisting of the other 20%.

(2) Compute point estimates of Beta1, Beta0, and sigma using the model calibration part of the data set.

Constructing a linear model of the 'development' data set produced point estimates Beta1 = .73, Beta0 = 1.48, and sigma = .74

(3) Identify a confidence level of your choice and construct a confidence interval estimate of Beta1 using the model calibration part of the data set.

For a confidence level of 95% the range for Beta1 is estimated between (.59,.88). Meaning that if our model is correct there is 95% chance that the true slope for modeling this data lies between .59 to .88 units of lcp. This was found using confint(development.lm, level=.95) in R.

(4) Construct a 95% prediction interval estimate of Mew|lcavol|lpsa for all of the lpsa values in your validation data set. What proportion of these prediction intervals capture the observed response?

We used the following R code to find a 95% prediction interval estimate for all of the lpsa values in the smaller data set (n=20):

```
xframe<-as.data.frame(holdout$lpsa)
sample.lpsa.ci<-data.frame(predict(holdout.lm,xframe, interval="pred",level=.95))
```

This was used to find the following output of prediction intervals:

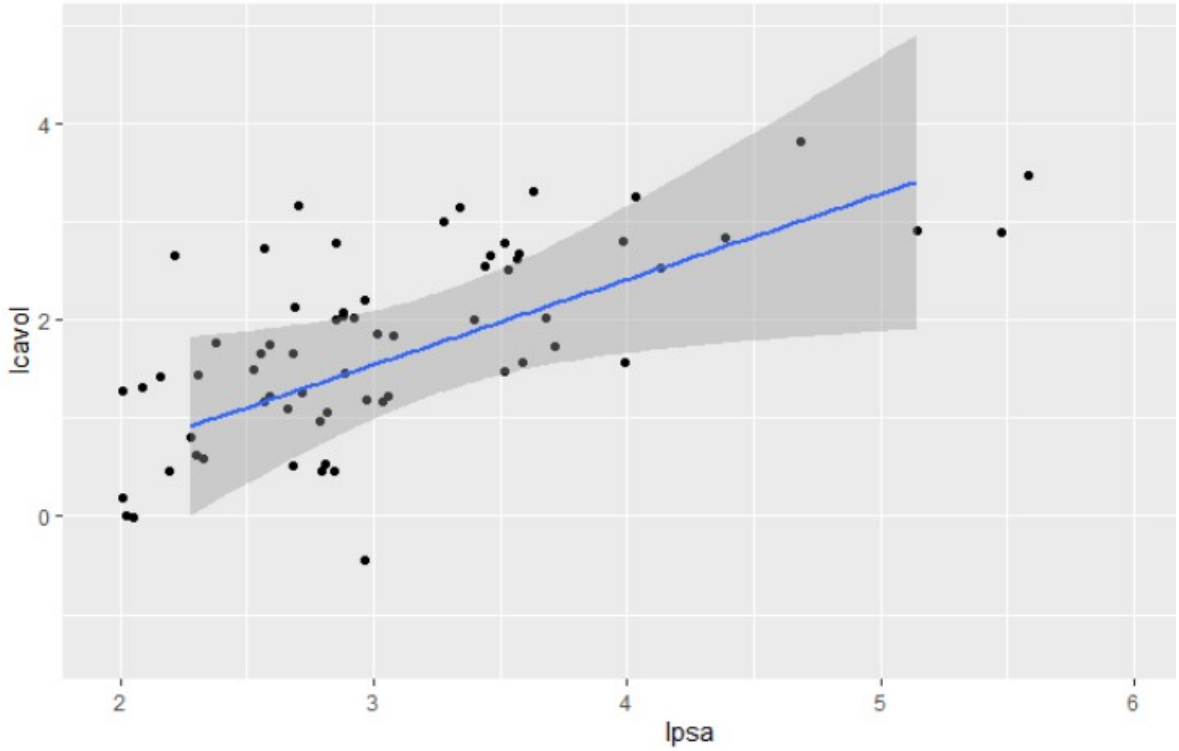
```
> xframe<-as.data.frame(holdout$lpsa)
> colnames(xframe)<-"lpsa"
> predict(holdout.lm,xframe, interval="pred",level=.95)
           fit       lwr      upr
1  -1.44483277 -4.598028 1.708362
2  -0.43614721 -3.263182 2.390887
3  -0.35785767 -3.169028 2.453313
4  -0.31540058 -3.118596 2.487795
5  -0.20367297 -2.988026 2.580680
6  -0.06322473 -2.828374 2.701925
7   0.16412209 -2.580767 2.909011
8   0.40196734 -2.336310 3.140245
9   0.41007732 -2.328240 3.148395
10  0.48206900 -2.257364 3.221502
11  0.49964953 -2.240266 3.239565
12  0.52341505 -2.217282 3.264112
13  0.57135603 -2.171373 3.314085
14  0.90010435 -1.872822 3.673031
15  0.90886788 -1.865247 3.682983
16  0.92938940 -1.847585 3.706364
17  0.93274034 -1.844712 3.710192
18  0.96841832 -1.814292 3.751128
19  1.01675737 -1.773591 3.807105
20  1.86683191 -1.148222 4.881886
```

We then used the following R code to see if the response fell within each data sample's predictive range:

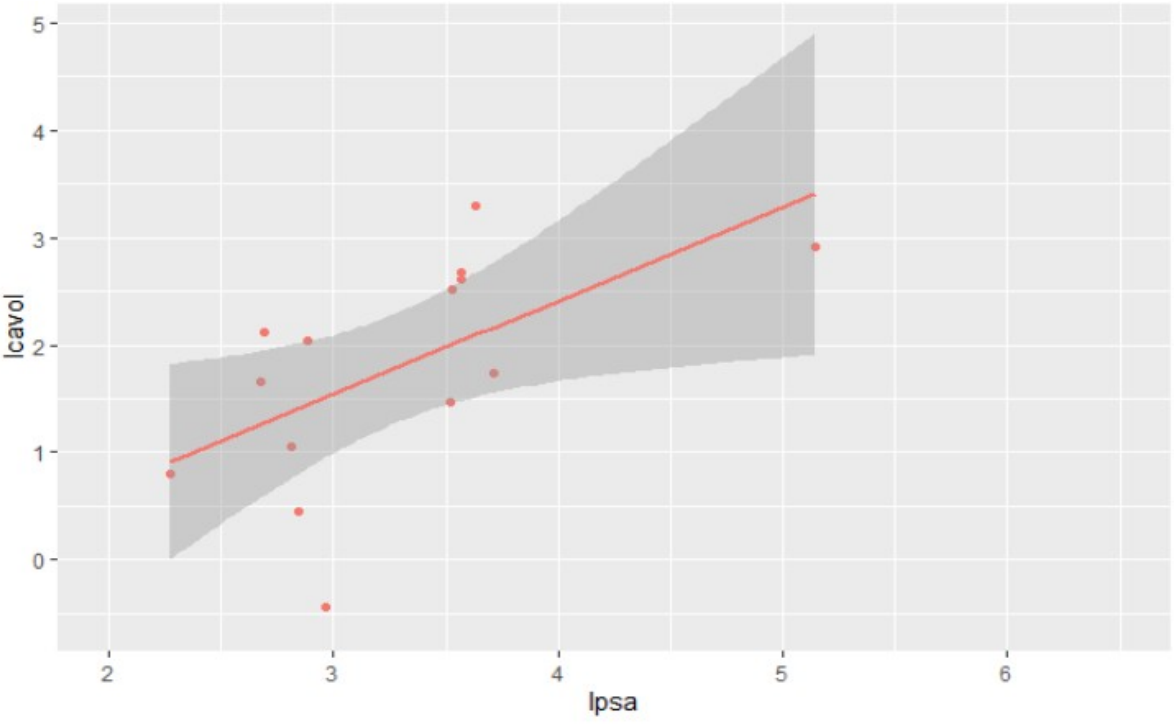
```
Observed.cavol<-data.frame(holdout$lcavol) #store actual response value in a data frame
range<-c(sample.lpsa.ci$lwr,sample.lpsa.ci$upr,Observed.cavol)  #store lwr/upper bounds of
                                                                #lpsa in data.frame
Observed.cavol > range$lwr & Observed < range$upr               #calculate if the response fell
                                                                #in that range.
```

This produced an output of 'TRUE' for each calculation. So 100% of the response's fell within the 95% predictive range.

(5) Plot a 95% confidence band computed from the model calibration part of the original data set on a scatterplot of that part of the calibration data set.



(6) Plot a 95% confidence band computed from the model calibration part of the data set on a scatterplot of the model validation part of the data set that includes a regression line computed using the model validation part of the data set.



Exercise 3:

(a) Fit a regression model with taste as the response and the three chemical con-tents as predictors. Identify the predictors that are statistically significant at the 5% level.

```
cheddar.lm<-lm(taste~Acetic+H2S+Lactic, cheddar)
summary(cheddar.lm)
```

	Estimate	Standard Error	t value	P(>|t|)
(Intercept)	-28.88	19.74	-1.46	0.16
Acetic	0.33	4.46	0.07	0.94
H2S	3.91	1.25	3.13	0.0#425
Lactic	19.67	8.63	2.28	0.03

When fitting a regression model with taste as the response and the other three chemicals as predictors we can see that at the 5% significance level both H2S and Lactic Acid are statistically significant. This is because for null hypothesis H0=Bi at 95% H2S has a p-value of .004, and Lactic has a p-value of .03. Both of these are less than .05 so we must reject the null hypothesis that these predictors have no significant determination in the response, and instead fail to reject the alternative hypothesis that Ha!=Bi for H2S and Lactic.

(b) Acetic and H2S are measured on a log scale. Fit a linear model where all three predictors are measured on their original scale. Identify the predictors that are statistically significant at the 5% level for this model.

```
cheddar2.lm<-lm(taste~exp(Acetic)+exp(H2S)+Lactic, cheddar)
summary(cheddar2.lm)
```

	Estimate	Standard Error	t value	P(>|t|)
(Intercept)	-1.90xe+001	1.13xe+01	-1.68	0.1
Acetic^e	1.891xe-02	1.562xe-02	1.21	0.24
H2S^e	7.668xe-04	4.188xe-04	1.83	0.08
Lactic	2.501xe+01	9.062xe+00	2.76	0.01

When fitting the data to a regression model after transforming the logarithmic data to it's original form the only predictor variable that is still statistically significant at the 5% level is Lactic acid. Lactic resulted in a p-value of .01, the only p-value of the predictor variables below .05. This implies that potentially we could produce a linear regression model with a response of taste and only use Lactic as our predictive variable that worked as well as a model that uses all three variables. However, this is not a certainty, and more testing is required to confirm this as a reliable model.

The transformation of the logarithmic predictor variables was confirmed by the following R code:

```
exAcetic<-exp(cheddar$Acetic)      #take the Acetic data and transform it by multiplying each
                                   #data point by data.point^e
```

```
exAcetic[1]     #check value of 1 data point
# = 93.97
log(93.97229)
# = 4.43, which is the same as the first data point in the original data set
```

(d) If H2S is increased 0.01 for the model used in (a), what change in the taste would be expected?

```
incH2s<- cheddar$H2S + .01
cheddar4.lm<-lm(taste~Acetic+incH2s+Lactic, cheddar)
summary(cheddar4.lm)
```

If we were to increase H2S by 0.01 for the model used in part (a) we would expect a slight change in statistical significance, to the degree that it would be found to be significantly significant at even the 1% significance level. If each H2S were incremented by .01 and the response of taste was the same, then H2S would be determined to have the highest impact on taste for this model.

(e) What is the percentage change in H2S on the original scale corresponding to an additive increase of 0.01 on the (natural) log scale?

```
incH2S<- cheddar$H2S + .01
exincH2S<-exp(incH2s)
exH2S<-exp(cheddar$H2S)
mean.logincH2S<-mean(logincH2S)
mean.logH2S<-mean(logH2s)
difference<-mean.logincH2S/mean.logH2S
difference #1.01005
```

After incriminating H2S by .01 and then transforming the H2S (cheddar$H2S) and incremented H2S (incH2S) transforming them to the original data set and finding the proportion between the two, we can see that the data set is 1.01005x larger.