Jonothan Meyer
Linear Regression
Project 8
04/27/21

Exercise (5 points): Use the your group project results for Project for this exercise. Consider two of the models you considered in this project that have a nested (sub-model) structure. Compare the two models using an F-test.

For this portion I wanted to look at two of the models I worked on independently early on in the process. The first is a linear model consists of all of the available predictor variables, other than two that have high colinearity with the response ('dream' and 'nondream'). The model this is compared to only contains the predictors that intuitively seem like they would have the largest significance on the response. For the sake of simplicity I wanted to use models that did not contain transformations. I felt that by doing this the interpretation of the F-test would be more clear and accurate.

First model:
Sleep ~ body + brain + lifespan + gestation + predation + exposure + danger

'Intuitive' model:
Sleep ~ gestation + predation + danger

Computation:
anova(mams.lm, reduced_mams.lm)

R Computation:
```
mams <- mammalsleep[complete.cases(mammalsleep),]
mams.lm <- lm(sleep~body+brain+lifespan+gestation+predation+exposure+danger, data = mams)
reduced_mams.lm <- lm(sleep~gestation+predation+danger, data=mams)
anova(mams.lm, reduced_mams.lm)
```

Conclusion:
Comparison of the first model and nested model resulted in a F value of 1.097, and a subsequent p-value of .3738. Using any reasonable alpha we would surprisingly fail to reject the null hypothesis that there is a significant difference between these models. The results lead me to believe that these two models are not parsimonious, and conclude that in this instance it would be better to use the reduced model. However, this conclusion is only limited to these specific models and should not be expanded to say that all variations of more complex models are not parsimonious.

Exercise (5 points): Use the sat data frame for this exercise. This exercise is based on Exercises 4, Chapter 3, in Linear Models with R.

(1) Fit a model with total SAT score as the response and expend, ratio and salary as predictors. Test the hypothesis that Beta_salary = 0. Test the hypothesis that salary = 0. Test the hypothesis that β_salary = β_ratio = β_expend = 0. Do any of these predictors have an effect on the response?

After building a linear model with total as the response and expend, ratio, and salary as predictors we reject the null hypothesis that β_salary = 0 with an alpha of .90. This is in favor of the alternative hypothesis that β_salary does have a significant effect on the response. The p-value of β_salary was found to be .07 for this linear model. Using the same alpha we also reject the null hypothesis that β_salary = β_ratio = β_expend = 0 with a p-value of .012. Collectively all of the predictors were estimated to have a significant effect on the response. The results from this linear model lead me to believe that teachers salary seems to have the most significant impact in determining students test scores in a rather surprising and unexpected way. Intuitively you would assume a positive relationship between these two variables, but it appears as though higher teacher pay generally results in lower student test scores. However, collectively all together, teachers salary, the amount of expense per pupil, and the ratio of students/teachers all have an estimated positive significant effect on students total sat test scores.

(2) Now add takers to the model. Test the hypothesis that β_takers = 0. Compare this model to the previous one using an F-test. Demonstrate that the F-test and t-test here are equivalent.

After adding 'takers' to the linear model we reject the null hypothesis that β_takers = 0 with a p-value of 2.61x10^-16 for the 'takers' variable. We estimate that the percent of students that are eligible to take the SATs has a significant effect on total SAT scores. To demonstrate that the F-test and t-test are equivalent here we compare the output of R's anova comparing the two models, and the t-test value of 'taken' in the summary output of the linear model that includes taken. Since we know that the F-test statistic equals t-value^2 we can confirm this ourselves. The t-value of 'taken' from the summary of the larger linear model is -12.559, and the F-value from the anova comparison that removes taken from the model is 157.74. We can therefore confirm that -12.559^2 = 157.73, so t^2 = F. More explicitly this was found using the following commands in R:

```
sat.lm <- lm(total~ expend + ratio + salary, data = sat)
sat_more.lm <- lm(total ~ expend + ratio + salary + takers, data = sat)
anova <- anova(sat.lm, sat_more.lm)
F_value <- round(as.numeric(anova$F), 4)
summary.lm <- summary(sat_more.lm)
T_value <- round(as.numeric(summary.lm$coefficients[5,3]), 4)
T_squared <- T_value * T_value
F_value - T_squared # = ~0. So we conclude they are the same value.
```

Exercise 3 (5 points): Using transformations to obtain a good fit, fit a linear model to one of the following response variables and predictor variable(s) combinations. Use a goodness-of-fit test to check your model, if appropriate; if a test for lack of fit is not possible for your data, explain why. Decide if it is reasonable to leave the response untransformed. This exercise is based on Exercises 2-7, Chapter 9, in Linear Models with R.

(1) Use yield as a response and nitrogen as a predictor from the cornit data set.

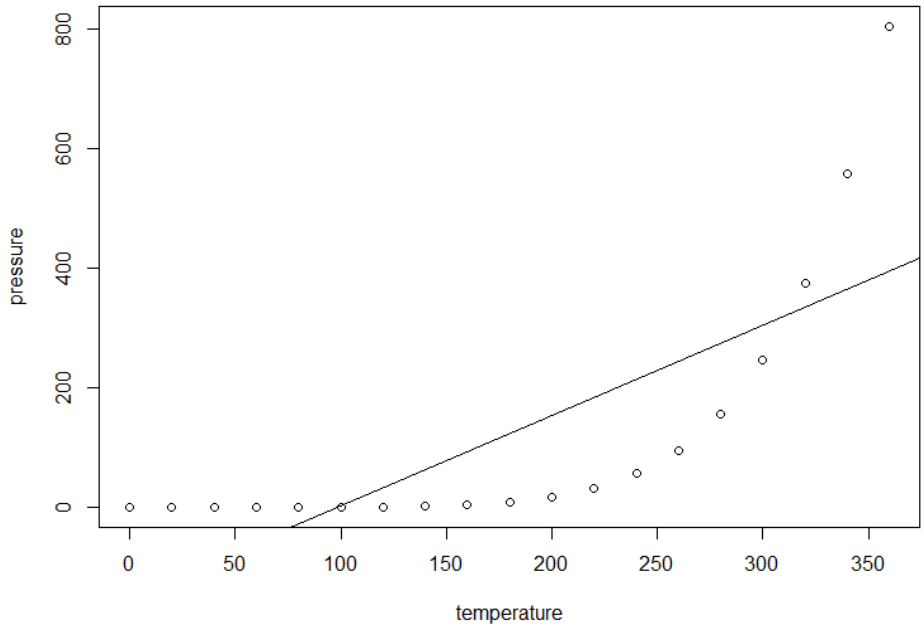
(2) Use 03 as a response and temp,humidity, and ibh as predictors from the ozone data set.

(3) Use pressure as a response and temperature as a predictor from the pressure data set.

(4) Use volume as a response and girth, humidity, and height as predictors from the trees data set.

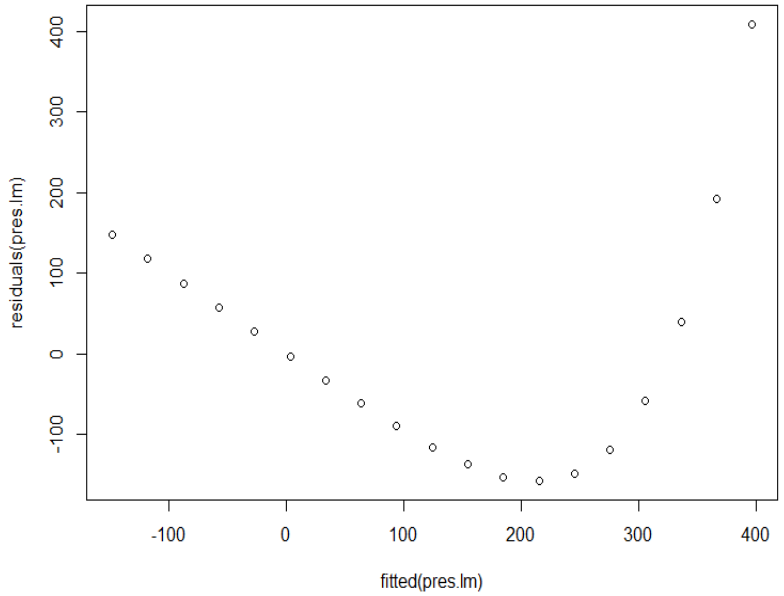
(5) Use cheddar as a response and three other variables as predictors from the cheddar data set.

(3)
The simple linear model for the pressure data set without any transformations fits the model relatively poorly. Intuitively this was to be expected after looking at the scatterplot temperature on pressure. The relationship appears to clearly be exponential, so the fit without transformations will always be rather poor. The coefficient of determination without transformations was found to be .5742, with a residual standard error of 150.8. The scatterplot and corresponding least squares regression line can be observed below.

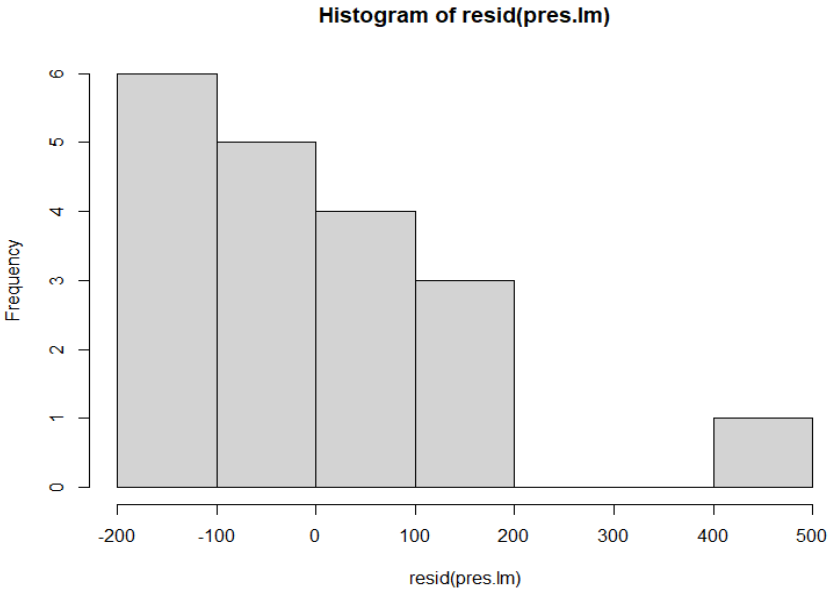


Clearly this duo of variables is modeled without transformations very poorly. This on its own would be good enough evidence to reevaluate the model, but for the sake of hammering it home, its worthwhile to look at plots relating to residual normality and constant variance below.

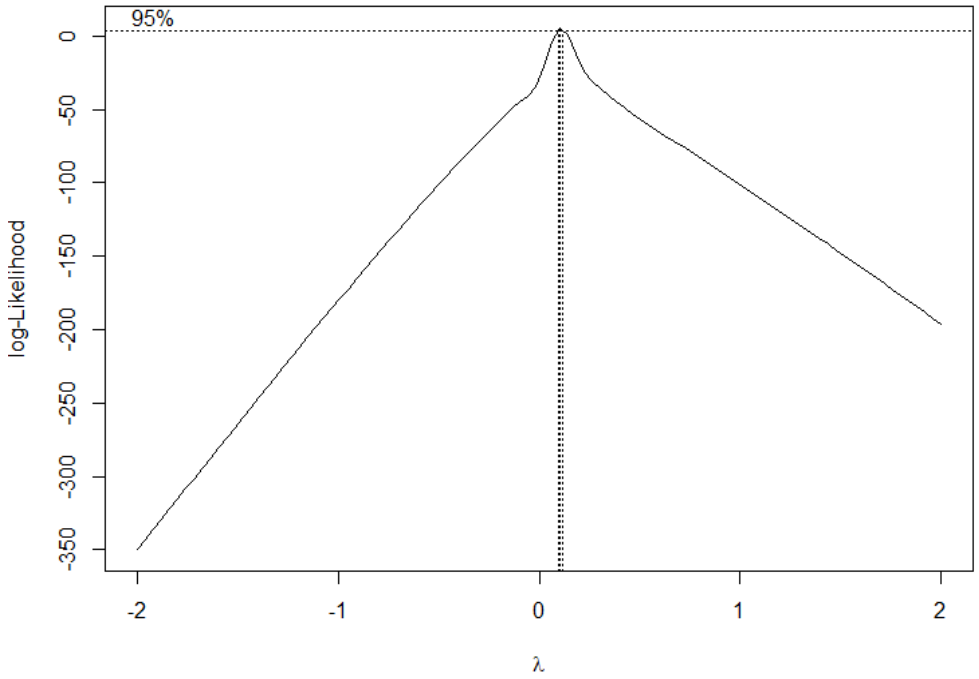
Residuals vs. Fitted Values



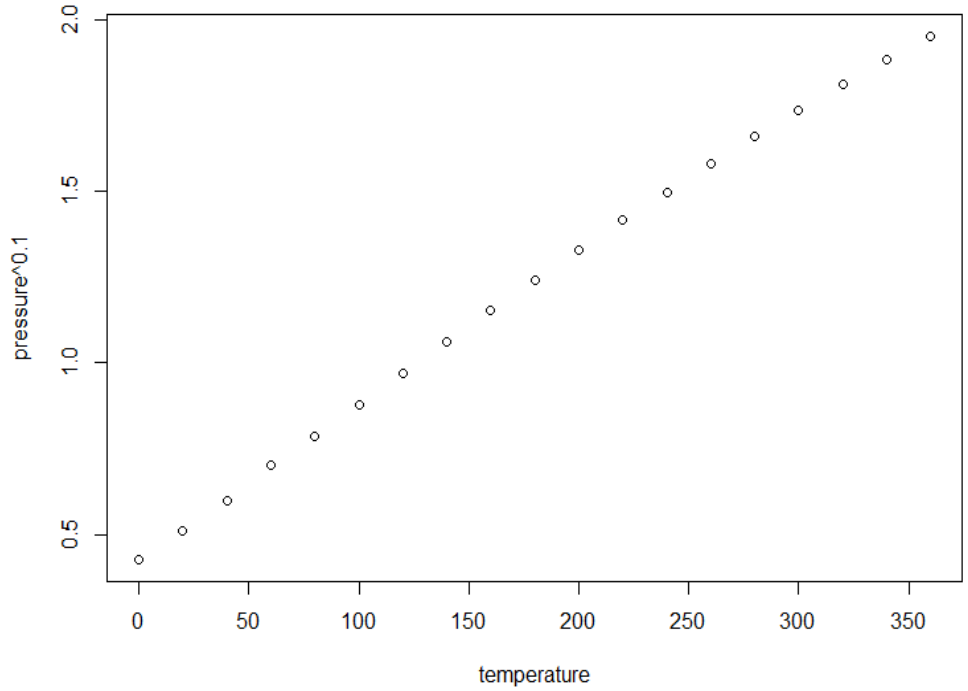
Frequency of Residuals



All of the model assumptions were violated when using this model. A logarithmic transformation could dramatically help the model to fit better. To find which log transformation would work best the BoxCox method was used. Based on the results of this that can be observed below it seems as though the best value to use would be $\gamma = .1$.



Scatterplot with Transformed Response:



After transforming the response the linear model has a coefficient of determination of .9984, and a residual standard error of .02. When looking at the model assumptions it appears as though both constant variance and normality are violated. However, I believe this to be the product of the model matching the data almost too perfectly. So the appearance of lack of normality of the residuals and the pattern in the fitted vs. residuals plot should be taken with a grain of salt. Overall, a transformation on the response of $\gamma = .1$ is appropriate and drastically improves the model.

Goodness-of-fit:
The first linear model and transformed model were build with the following R commands, and then

compared for fit using anova.

```
pres.lm <- lm(pressure~temperature, data = pressure)
pres.tras.lm <- lm(pressure^.1~temperature, data = pressure)
anova(pres.lm, pres.tras.lm)
```

The resulting F-test p-value was .000171, which we use as further evidence that the predictor and response have a lack of fit with a linear model.

Exercise (5 points): This exercise uses a method called Ridge Regression that penalizes estimates of coefficients for being large. Ridge regression, covered in Section 11.3, makes the assumption that the regression coefficients should not be very large. This exercise is based on Exercises 3-4, Chapter 11, in Linear Models with R. For one of the following response variables and predictor variables combinations, (a) build a linear regression model with all predictors, (b) build a linear regression with variables selected according to the AIC procedure, and (c) build a linear regression with variables selected using a ridge regression procedure.

(1) Fit models using the seatpos data with hipcenter as the response and all other variables as possible predictors.
(2) Fit models using the fat data with siri as the response and all other variables except brozek and density as possible predictors.

(a)
The second data set was used to build a linear model using the following R functions:

```
fattt <- fat %>% select(-brozek, -density)
head(fattt)
names(fattt)
fat.lm <- lm(siri ~., data = fattt)
```

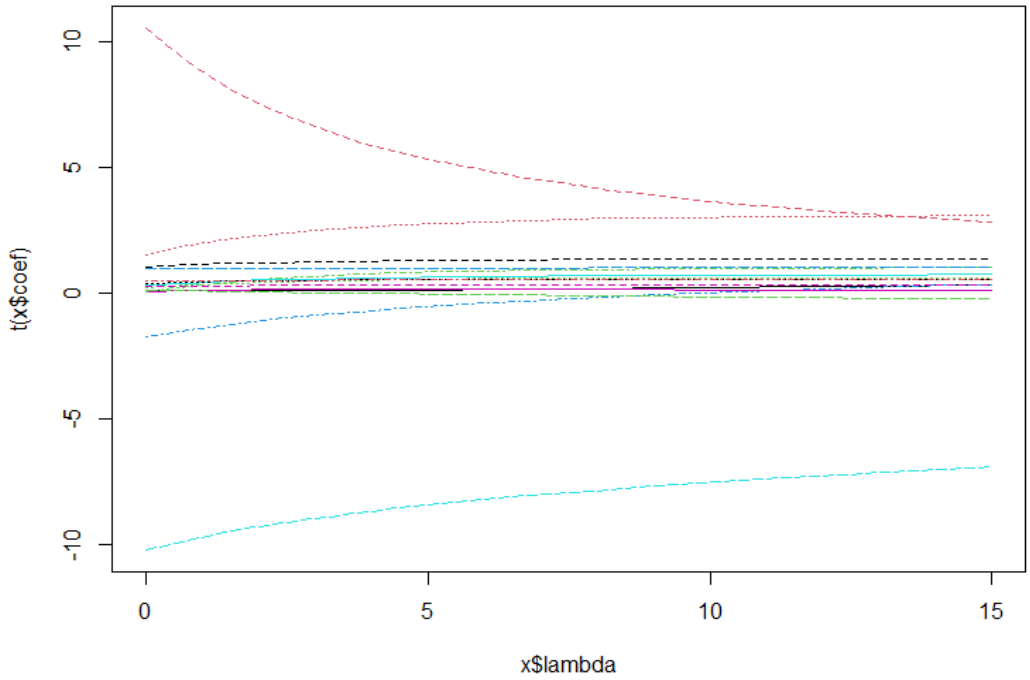
(b)
The AIC procedure was then conducted on the linear model using step(fat.lm, trace = 1). The initial AIC for the full model was 222.5 and iterated 5 times before removal of additional predictors no longer lowered the AIC. The first predictor removed was 'hip' with an AIC contribution of 220.50, the second was 'neck' with a AIC contribution of 218.57, the third was 'age' with a AIC contribution of 216.93, and the fourth and final predictor removed was 'wrist' with an AIC contribution of 216.07. The final AIC for the model was 216.07, an improvement of 6.43. The final model after using this proceedure is as follows: siri ~ weight + height + adipos + free + chest + abdom + thigh + knee + ankle + biceps + forearm

(c)
Using the ridge regression procedure a $\gamma = 5$ was chosen after observing the output and plot of the following commands:

```
lambda = seq(-5, 15, len=21)
fat.ridge.lm <- lm.ridge(siri~.,data = fattt, lambda = lambda )
plot(lm.ridge(siri~., data=fattt, lambda = lambda))
select(fat.ridge.lm <- lm.ridge(siri~.,data = fattt, lambda = lambda))
coef(fat.ridge.lm) #The best lambda was found to be 0
```

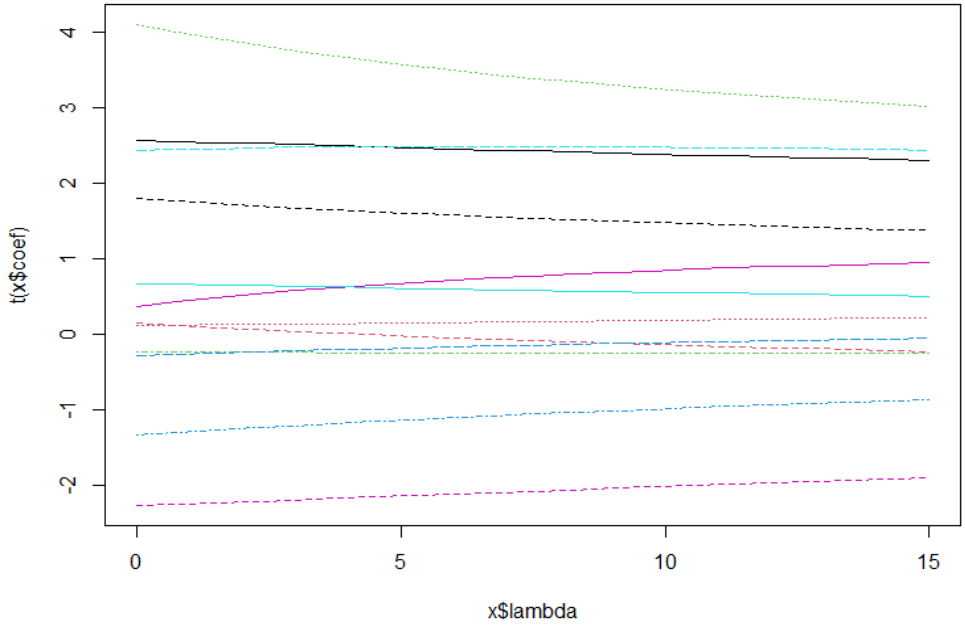
```
fat.ridge.lm <- lm.ridge(siri~.,data=fattt, lambda = 0)
coef(fat.ridge.lm)
```

Ridge regression from 0 to 15 using all predictors:



After spending a good deal of time searching for the predictors that matched with the lines at (0, +/- 10) I was able to figure out that they represent 'weight', 'free', and 'abdom'. The smallest GCV when including the full linear model was found to be $\gamma = 0$. To further improve the linear model I wanted to see what would happen if I ran a ridge regression again after removing the previous three predictors to see if it resulted in a lambda value other than zero to better understand the process of ridge regression. After performing this process the smallest value of GCV was found to be at $\gamma = 10.5$ which can be observed on the plot below.

Ridge Regression from 0 to 15 without 'weight', 'free', and 'abdom'



This appears to be the least sensitive model equation based off the results from the plot and ridge regression sequence. Therefore the final ridge regression model is ridge.lm(siri ~ age + height + adipos + neck + chest + hip + thigh + knee + ankle + biceps + forearm + wrist, data = fat, lambda = 10.5)