Wolfe\_HW\_4

Tyler Wolfe

9/21/2020

library(alr4)

## Loading required package: car

## Loading required package: carData

## Loading required package: effects

## Registered S3 methods overwritten by 'lme4':  
## method from  
## cooks.distance.influence.merMod car   
## influence.merMod car   
## dfbeta.influence.merMod car   
## dfbetas.influence.merMod car

## lattice theme set by effectsTheme()  
## See ?effectsTheme for details.

The first two problems in this homework are on the MinnLand data from the alr4 package. Imagine yourself working for an organization such as one that is interested in food production, that needs to understand land prices of farms. We need to be able to communicate the drivers of these prices and the effects of time and region. To understand trends but also where we may find land that is undervalued. In the next two questions you’ll begin to explore how the price per acre has changed over time, and also investigate if there was a region effect. A region effect would be important in detecting an area where prices may be more robust.

Problem 1: Using the MinnLand data in the alr4 package (read the help file for more info on data), fit a model using log(acrePrice) as the response and year as the predictor (use year as a factor). In general housing prices in the US were increasing from 2002-2006, and then began to fall in 2007. Is this true for Minnesota? Use a boxplot and your regression coefficents to explain. Report the results any other statistical test we have learned to give insight into the implications of your model. (Make sure you check your residual plots).

First step will be to explore the data and help file to ensure understanding of the data.

help(MinnLand)  
head(MinnLand)

## acrePrice region improvements year acres tillable financing crpPct  
## 1 766 Northwest 0 2002 82 94 title\_transfer 0  
## 2 733 Northwest 0 2003 30 63 title\_transfer 0  
## 3 850 Northwest 4 2002 150 47 title\_transfer 0  
## 4 975 Northwest 0 2003 160 86 title\_transfer 0  
## 5 886 Northwest 62 2002 90 NA title\_transfer 0  
## 6 992 Northwest 30 2003 120 83 title\_transfer 0  
## productivity  
## 1 NA  
## 2 NA  
## 3 NA  
## 4 NA  
## 5 NA  
## 6 NA

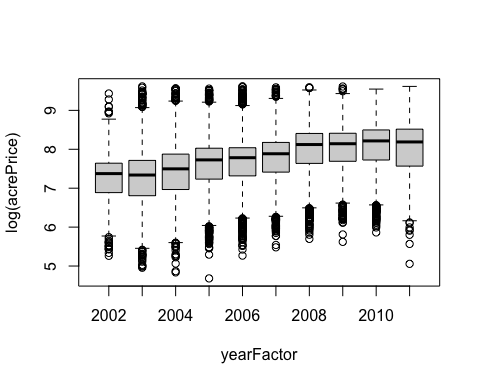
Due to the specific question and variables given in the prompt, we know that we need to convert the vector “year” to a factor.

MinnLand$yearFactor<-as.factor(MinnLand$year)  
table(MinnLand$yearFactor)

##   
## 2002 2003 2004 2005 2006 2007 2008 2009 2010 2011   
## 566 2114 2490 2321 2187 2248 2431 1459 1823 1061

Now that “year” has been converted to read as a factor (yearFactor), lets look at the comparison of the variables in a boxplot.

boxplot(log(acrePrice)~yearFactor,data=MinnLand)

 When comparing yearFactor to the log(acrePrice), it seems that the mean of the log(acrePrice) increases slightly every year, with the possible exception of 2003, which seems to stay the same or decrease by a small amount. Another interesting observation, is that 2011, box seems to taller in comparision to its preceding years, which a lower lower limit to occurances. In order to better understand the visualization, lets perform a linear regression.

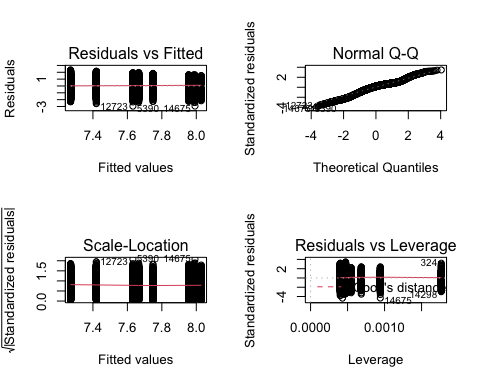
model<-lm(log(acrePrice)~yearFactor, data=MinnLand)  
summary(model)

##   
## Call:  
## lm(formula = log(acrePrice) ~ yearFactor, data = MinnLand)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -2.9499 -0.3785 0.1301 0.4354 2.3456   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 7.27175 0.02848 255.345 < 2e-16 \*\*\*  
## yearFactor2003 -0.00155 0.03207 -0.048 0.961   
## yearFactor2004 0.14794 0.03155 4.689 2.76e-06 \*\*\*  
## yearFactor2005 0.36026 0.03176 11.343 < 2e-16 \*\*\*  
## yearFactor2006 0.39392 0.03195 12.329 < 2e-16 \*\*\*  
## yearFactor2007 0.47682 0.03186 14.965 < 2e-16 \*\*\*  
## yearFactor2008 0.68364 0.03162 21.620 < 2e-16 \*\*\*  
## yearFactor2009 0.71407 0.03355 21.284 < 2e-16 \*\*\*  
## yearFactor2010 0.75733 0.03260 23.231 < 2e-16 \*\*\*  
## yearFactor2011 0.72071 0.03526 20.437 < 2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.6775 on 18690 degrees of freedom  
## Multiple R-squared: 0.1293, Adjusted R-squared: 0.1289   
## F-statistic: 308.5 on 9 and 18690 DF, p-value: < 2.2e-16

Based on the summary of the regression, we can observe that as the year increases by 1, the price per acre of land also increases by .14-.75% for all years with the exception of 2003, where there was a .00155 decrease (this year was also the only coefficient without a significant pvalue). I also noticed that the rate at price per acre increased was steady over that period until 2011, where the mean of increase is lower than the preceding 2010 price per acre of land.

These are interesting findings, but because we want to do our due diligence, we should check our residual plots.

par(mfrow=c(2,2))  
plot(model)



We see in the Residuals vs. Fitted plot that the variance of the residuals seems to be standardized and without a curve. This indicates constant variance and confirms that we can trust our model.

I also felt that it was interesting that we see banding in the data points, due to the factored year vector.

Problem 2: Using the MinnLand data again, fit two models (with year as a factor in both)

M1<-lm(log(acrePrice)~year+region, data=MinnLand)

M2<-lm(log(acrePrice)~year+region+year\*region, data=MinnLand)

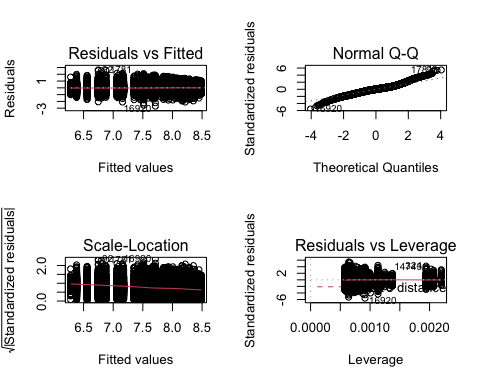
Explain the difference between the two models, then provide a comparison of the models. Using the model you define as “best” provide an explanation of the coefficients. (HINT: An EFFECTS PLOT makes this easier visually).

As we have already examined the data in some capacity in problem 1, we do not need to re-examine here. I will begin by adjusting the models to include the year variable as a factor, executing both models in seperate chunks.

M1<-lm(log(acrePrice)~yearFactor+region, data=MinnLand)  
summary(M1)

##   
## Call:  
## lm(formula = log(acrePrice) ~ yearFactor + region, data = MinnLand)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -2.70107 -0.27600 0.00925 0.25007 2.64400   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 6.28055 0.02180 288.054 < 2e-16 \*\*\*  
## yearFactor2003 0.10140 0.02293 4.422 9.85e-06 \*\*\*  
## yearFactor2004 0.27903 0.02257 12.362 < 2e-16 \*\*\*  
## yearFactor2005 0.46728 0.02271 20.573 < 2e-16 \*\*\*  
## yearFactor2006 0.51865 0.02285 22.695 < 2e-16 \*\*\*  
## yearFactor2007 0.59520 0.02279 26.120 < 2e-16 \*\*\*  
## yearFactor2008 0.78307 0.02261 34.629 < 2e-16 \*\*\*  
## yearFactor2009 0.80345 0.02399 33.494 < 2e-16 \*\*\*  
## yearFactor2010 0.87669 0.02332 37.596 < 2e-16 \*\*\*  
## yearFactor2011 0.87674 0.02524 34.730 < 2e-16 \*\*\*  
## regionWest Central 0.79307 0.01153 68.809 < 2e-16 \*\*\*  
## regionCentral 1.13194 0.01086 104.253 < 2e-16 \*\*\*  
## regionSouth West 1.07670 0.01238 87.004 < 2e-16 \*\*\*  
## regionSouth Central 1.29009 0.01204 107.194 < 2e-16 \*\*\*  
## regionSouth East 1.32975 0.01340 99.212 < 2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.4841 on 18685 degrees of freedom  
## Multiple R-squared: 0.5556, Adjusted R-squared: 0.5553   
## F-statistic: 1669 on 14 and 18685 DF, p-value: < 2.2e-16

par(mfrow=c(2,2))  
plot(M1)

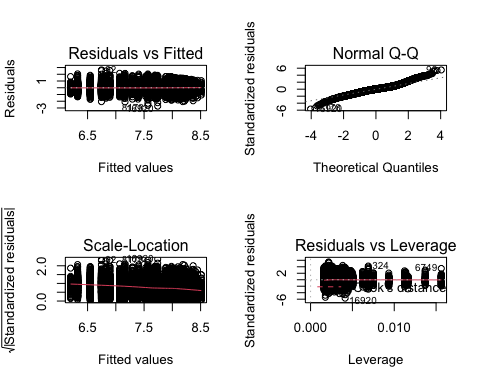


After examining model 1 we see that we’ve included the regions, but this information isn’t necessarily helpful as it’s only taking into account the means of the factors of different regions, not the means of regions by year. Perhaps the second model will allow us to make more insights. (although the second provided model was using predictors: "yearF+region+yearF*region" I will be executing the second model as "yearFactor*region, as shown in the collaborate.)

M2<-lm(log(acrePrice)~yearFactor\*region, data=MinnLand)  
summary(M2)

##   
## Call:  
## lm(formula = log(acrePrice) ~ yearFactor \* region, data = MinnLand)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -2.73006 -0.27521 0.01157 0.25561 2.64607   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 6.19911 0.06022 102.939 < 2e-16 \*\*\*  
## yearFactor2003 0.12445 0.06476 1.922 0.05467 .   
## yearFactor2004 0.34837 0.06367 5.471 4.52e-08 \*\*\*  
## yearFactor2005 0.54665 0.06425 8.508 < 2e-16 \*\*\*  
## yearFactor2006 0.62531 0.06432 9.722 < 2e-16 \*\*\*  
## yearFactor2007 0.69422 0.06419 10.815 < 2e-16 \*\*\*  
## yearFactor2008 0.86828 0.06417 13.532 < 2e-16 \*\*\*  
## yearFactor2009 0.94283 0.06679 14.115 < 2e-16 \*\*\*  
## yearFactor2010 0.95188 0.06505 14.634 < 2e-16 \*\*\*  
## yearFactor2011 0.96351 0.06723 14.333 < 2e-16 \*\*\*  
## regionWest Central 0.89062 0.07915 11.253 < 2e-16 \*\*\*  
## regionCentral 1.20484 0.07230 16.664 < 2e-16 \*\*\*  
## regionSouth West 1.09079 0.07613 14.328 < 2e-16 \*\*\*  
## regionSouth Central 1.45223 0.07896 18.392 < 2e-16 \*\*\*  
## regionSouth East 1.48043 0.08250 17.945 < 2e-16 \*\*\*  
## yearFactor2003:regionWest Central -0.06041 0.08631 -0.700 0.48400   
## yearFactor2004:regionWest Central -0.14535 0.08493 -1.711 0.08703 .   
## yearFactor2005:regionWest Central -0.10822 0.08573 -1.262 0.20685   
## yearFactor2006:regionWest Central -0.10811 0.08568 -1.262 0.20706   
## yearFactor2007:regionWest Central -0.06810 0.08572 -0.794 0.42693   
## yearFactor2008:regionWest Central -0.09024 0.08551 -1.055 0.29126   
## yearFactor2009:regionWest Central -0.15673 0.08981 -1.745 0.08099 .   
## yearFactor2010:regionWest Central -0.04117 0.08698 -0.473 0.63601   
## yearFactor2011:regionWest Central -0.13921 0.09123 -1.526 0.12706   
## yearFactor2003:regionCentral 0.03938 0.07877 0.500 0.61715   
## yearFactor2004:regionCentral -0.06105 0.07766 -0.786 0.43179   
## yearFactor2005:regionCentral -0.01894 0.07830 -0.242 0.80887   
## yearFactor2006:regionCentral -0.04535 0.07878 -0.576 0.56486   
## yearFactor2007:regionCentral -0.11180 0.07888 -1.417 0.15636   
## yearFactor2008:regionCentral -0.13345 0.07872 -1.695 0.09004 .   
## yearFactor2009:regionCentral -0.16203 0.08284 -1.956 0.05049 .   
## yearFactor2010:regionCentral -0.15092 0.08074 -1.869 0.06160 .   
## yearFactor2011:regionCentral -0.12382 0.08462 -1.463 0.14342   
## yearFactor2003:regionSouth West -0.02205 0.08522 -0.259 0.79580   
## yearFactor2004:regionSouth West -0.06516 0.08377 -0.778 0.43671   
## yearFactor2005:regionSouth West -0.11040 0.08394 -1.315 0.18842   
## yearFactor2006:regionSouth West -0.11492 0.08399 -1.368 0.17127   
## yearFactor2007:regionSouth West -0.02079 0.08352 -0.249 0.80339   
## yearFactor2008:regionSouth West 0.05438 0.08296 0.656 0.51215   
## yearFactor2009:regionSouth West -0.01820 0.08741 -0.208 0.83508   
## yearFactor2010:regionSouth West 0.13339 0.08534 1.563 0.11805   
## yearFactor2011:regionSouth West 0.15965 0.09689 1.648 0.09942 .   
## yearFactor2003:regionSouth Central -0.06014 0.08766 -0.686 0.49271   
## yearFactor2004:regionSouth Central -0.11564 0.08592 -1.346 0.17838   
## yearFactor2005:regionSouth Central -0.13846 0.08626 -1.605 0.10848   
## yearFactor2006:regionSouth Central -0.26222 0.08656 -3.029 0.00246 \*\*   
## yearFactor2007:regionSouth Central -0.24577 0.08595 -2.859 0.00425 \*\*   
## yearFactor2008:regionSouth Central -0.15184 0.08513 -1.784 0.07451 .   
## yearFactor2009:regionSouth Central -0.24561 0.08923 -2.753 0.00592 \*\*   
## yearFactor2010:regionSouth Central -0.15730 0.08721 -1.804 0.07129 .   
## yearFactor2011:regionSouth Central -0.10230 0.09262 -1.105 0.26938   
## yearFactor2003:regionSouth East -0.06330 0.09128 -0.693 0.48806   
## yearFactor2004:regionSouth East 0.01118 0.08993 0.124 0.90108   
## yearFactor2005:regionSouth East -0.15057 0.09083 -1.658 0.09741 .   
## yearFactor2006:regionSouth East -0.16132 0.09183 -1.757 0.07897 .   
## yearFactor2007:regionSouth East -0.15648 0.09113 -1.717 0.08598 .   
## yearFactor2008:regionSouth East -0.21373 0.09131 -2.341 0.01926 \*   
## yearFactor2009:regionSouth East -0.28636 0.09530 -3.005 0.00266 \*\*   
## yearFactor2010:regionSouth East -0.25598 0.09252 -2.767 0.00567 \*\*   
## yearFactor2011:regionSouth East -0.22985 0.09718 -2.365 0.01803 \*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.4818 on 18640 degrees of freedom  
## Multiple R-squared: 0.5609, Adjusted R-squared: 0.5596   
## F-statistic: 403.6 on 59 and 18640 DF, p-value: < 2.2e-16

par(mfrow=c(2,2))  
plot(M2)



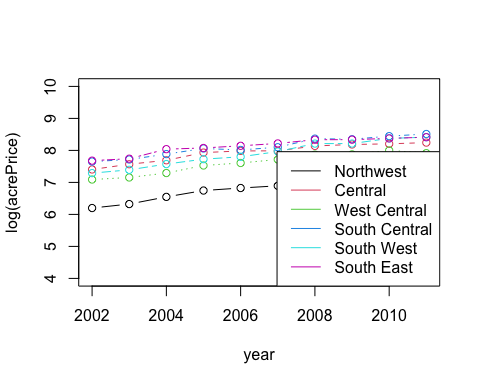
plot(log(acrePrice)~year,data=MinnLand, type="n", ylim=c(4,10))  
  
years=unique(MinnLand$year)  
years

## [1] 2002 2003 2004 2005 2006 2007 2008 2009 2010 2011

regions=unique(MinnLand$region)  
regions

## [1] Northwest Central West Central South Central South West   
## [6] South East   
## 6 Levels: Northwest West Central Central South West ... South East

for(i in 1:length(regions)){  
 newdata=MinnLand[MinnLand$region==regions[i],]  
 means=tapply(log(newdata$acrePrice),newdata$yearFactor, mean)  
 lines(means~years, col=i, lty=i,type="b")  
}  
  
legend("bottomright",col=(1:length(regions)), legend=regions,lty=1)



After running both provided models, checking residuals and using an all effects plot, we can make a few observations. First, the coefficient breakdown of regions by year provides us with much more information and allows us to compare not only by year, or region, but by region separated into year as well.

Although I feel that model 2 is the better choice, let’s check that assertion with an anova test.

anova(M1,M2)

## Analysis of Variance Table  
##   
## Model 1: log(acrePrice) ~ yearFactor + region  
## Model 2: log(acrePrice) ~ yearFactor \* region  
## Res.Df RSS Df Sum of Sq F Pr(>F)   
## 1 18685 4379.0   
## 2 18640 4326.3 45 52.613 5.0374 < 2.2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

Based on the anova test, and the significant p-value caused by the added variables in the second model, I would say that model 2 is the “best” model to use of the two provided models. Also, by examining the all effects plot for model two, we can see a positive trend in almost all regions over the course of 2002 to 2011. The only discrepancy would be in the West Central region from 2010 to 2011. We can also see that the South East region showed noticeably more growth between 2003 and 2004, and the South Central and South West regions showed stronger growth from 2007 to 2008.

Problem 3: The salary data in the alr4 package concerns salary at a small midwestern college for a legal case concerning pay discrimination against women. That data refers to tenure track faculty only. We have been asked to look at this claim based on the data we have avaliable. The variables are degree (MS/PhD), rank (Assistant Prof/ Assoc Prof/ Prof), sex (M/F), Year (number of years in current rank), ysdeg (years since highest degree), and salary (academic year salary in dollars). The code:

t.test(salarysex==“Male”], salarysex==“Female”], alternative=“greater”)

produces a two sample t-test, with hypothesis

NH: The mean salary of males in this population is less than or equal to the mean salary of females

AH: The mean salary of males in this population is greater than the mean salary of females

Run this code, and comment on the results of this test comparing your p-value to a 0.05 level, does anything change at a 0.01 level?

We know there could be other factors that effect salary, implement a model using all the variables in the data with salary as the response and the rest as predictors, comment on the effect of sex. Fit a second model removing rank as a factor, now comment on the effect of sex in the model? What insight does this provide on the data? (Hint: Assistant Prof salary < Assoc Prof Salary < Prof Salary in the same field)

First lets explore the data and open the help file.

help(salary)  
head(salary)

## degree rank sex year ysdeg salary  
## 1 Masters Prof Male 25 35 36350  
## 2 Masters Prof Male 13 22 35350  
## 3 Masters Prof Male 10 23 28200  
## 4 Masters Prof Female 7 27 26775  
## 5 PhD Prof Male 19 30 33696  
## 6 Masters Prof Male 16 21 28516

dim(salary)

## [1] 52 6

summary(salary)

## degree rank sex year ysdeg   
## Masters:34 Asst :18 Male :38 Min. : 0.000 Min. : 1.00   
## PhD :18 Assoc:14 Female:14 1st Qu.: 3.000 1st Qu.: 6.75   
## Prof :20 Median : 7.000 Median :15.50   
## Mean : 7.481 Mean :16.12   
## 3rd Qu.:11.000 3rd Qu.:23.25   
## Max. :25.000 Max. :35.00   
## salary   
## Min. :15000   
## 1st Qu.:18247   
## Median :23719   
## Mean :23798   
## 3rd Qu.:27258   
## Max. :38045

We see that their are 52 observations of college staff and 6 vectors. We will now run the provided code to execute a t-test with the Null Hypothesis that mean salary of males in this population is less than or equal to that of females. Our Alternative Hypothesis will be that mean salary collected by males is greater then that of females.

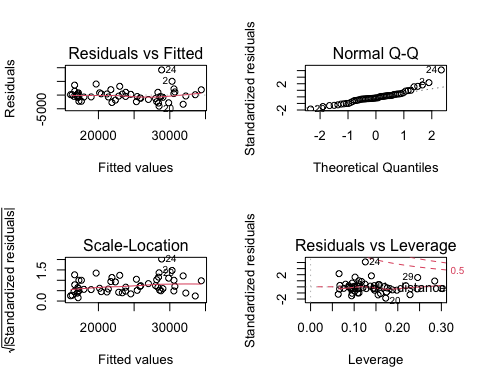
(I should also note that upon review of the summary for the salary data, we can see that there are 38 males and only 14 female staff. This could potentially create problems with the t-test, as we are comparing the means of two groups, where one group is over twice the size of the other. Also, this presents a question as to whether their are female staff represented across all positions. The lack of female representation may create a problem with disperate treatment or disperate impact and should be examined.)

t.test(x=salary$salary[salary$sex=="Male"],y=salary$salary[salary$sex=="Female"],alternative = "greater")

##   
## Welch Two Sample t-test  
##   
## data: salary$salary[salary$sex == "Male"] and salary$salary[salary$sex == "Female"]  
## t = 1.7744, df = 21.591, p-value = 0.04505  
## alternative hypothesis: true difference in means is greater than 0  
## 95 percent confidence interval:  
## 105.1409 Inf  
## sample estimates:  
## mean of x mean of y   
## 24696.79 21357.14

Examining the results of the t-test, we can see that their is a difference in the average salary based on sex when using .05 as the alpha (fail to reject the Null that males make less than or equal to females. However, when we adjust the alpha to .001, we reject the null, concluding that males mean salaries are greater than female mean salaries. Due to the differences between the two alpha’s we should continue our analysis by exploring other variables through a regression model.

salaryModel1<-lm(salary~year+ysdeg+degree+rank+sex, data=salary)  
par(mfrow=c(2,2))  
plot(salaryModel1)



summary(salaryModel1)

##   
## Call:  
## lm(formula = salary ~ year + ysdeg + degree + rank + sex, data = salary)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -4045.2 -1094.7 -361.5 813.2 9193.1   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 15746.05 800.18 19.678 < 2e-16 \*\*\*  
## year 476.31 94.91 5.018 8.65e-06 \*\*\*  
## ysdeg -124.57 77.49 -1.608 0.115   
## degreePhD 1388.61 1018.75 1.363 0.180   
## rankAssoc 5292.36 1145.40 4.621 3.22e-05 \*\*\*  
## rankProf 11118.76 1351.77 8.225 1.62e-10 \*\*\*  
## sexFemale 1166.37 925.57 1.260 0.214   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 2398 on 45 degrees of freedom  
## Multiple R-squared: 0.855, Adjusted R-squared: 0.8357   
## F-statistic: 44.24 on 6 and 45 DF, p-value: < 2.2e-16

After running a regression model with all variables, we can see that sex is not a significant variable to determine salary. The variables that display significance across all alpha’s are rank, and years in current rank.

Alternatively I wanted to also take a look at the representation across the various ranks, as we see that rank is a determining factor to salary. If there are less females represented in the higher ranks, the mean of female salary will be lower. Also, their may be potential for disparate impact or potential legal ramifications if it appears that females are not adequately represented across all ranks.

table(salary$rank,salary$sex)

##   
## Male Female  
## Asst 10 8  
## Assoc 12 2  
## Prof 16 4

As suspected, female staff members become less common in relation to male staff members as you progress through the ranks. This could become a legal liability for the university if not addressed. Because less females are making the pay accounted for in the higher ranks, the females are making less on average.

## Problem 4: In many manufacturing setting understanding levels in which machines can be tuned to produce a given product is very important. Even more important is understanding the ability of those products strength and ability to fail. In this setting we are going to look at the strength of yarn, based on the length of the yarn (len), the time the yarn was asked to hold per cycle (amp), and the amount the yarn was asked to hold (load). The response is the number of times/cycles the yarn made it through without breaking (cycles). The data can be found in the wool dataset in the alr4 package. Fit the full main effects model using cycles as the response and all other variables as predictors, and interpret the Tukey HSD intervals and test with regard to load, to describe the differences between the levels of load.

To begin, as always, I will explore the data to familiarize myself.

help(Wool)  
dim(Wool)

## [1] 27 4

head(Wool)

## len amp load cycles  
## 1 250 8 40 674  
## 2 250 8 45 370  
## 3 250 8 50 292  
## 4 250 9 40 338  
## 5 250 9 45 266  
## 6 250 9 50 210

We see that their are 27 observations across four variables: length, amplitude, load, and cycles.

summary(Wool)

## len amp load cycles   
## Min. :250 Min. : 8 Min. :40 Min. : 90.0   
## 1st Qu.:250 1st Qu.: 8 1st Qu.:40 1st Qu.: 312.0   
## Median :300 Median : 9 Median :45 Median : 566.0   
## Mean :300 Mean : 9 Mean :45 Mean : 861.4   
## 3rd Qu.:350 3rd Qu.:10 3rd Qu.:50 3rd Qu.:1105.0   
## Max. :350 Max. :10 Max. :50 Max. :3636.0

In our summary of the data set we see a magitude of more than 10 in our cycles vector, which suggests we should use a log transformation. We can also see in the summary and in the help file that there are only three distinct outcomes in each of the other vectors. Because of this, I will transform the other variables of length, amplitude, and load into factors.

Wool$lenF<-as.factor(Wool$len)  
Wool$ampF<-as.factor(Wool$amp)  
Wool$loadF<-as.factor(Wool$load)