Homework #3

Samuel Jones

11 March 2018

Part 1

Problem 1. Conduct the ordinary least squares regressions and compare your results to those reported in the paper. (1 point)

- I downloaded the supplemental data from the Journal of Applied Econometrics because the HousePrices dataset from the AER library was not producing identical results.

  The variables are defined as follows:
  sell = sale price of a house
  lot = the lot size of a property in square feet
  bdms = the number of bedrooms
  fb = the number of full bathrooms
  sty = the number of stories excluding basement
  drv = 1 if the house has a driveway
  rec = 1 if the house has a recreational room
  ffin = 1 if the house has a full finished basement
  ghw = 1 if the house uses gas for hot water heating
  ca = 1 if there is central air conditioning
  gar = the number of garage places
  reg = 1 if the house is located in the preferred neighbourhood of the city (Anglin, et al, 1996)

```
#### Part 1 #### Question 1 ####
ag.data <- read.csv("ag-data.fil.csv")
#### Table 2 Regression ####
table2.lm <- lm(log(sell) ~ drv + rec + ffin + ghw + ca + gar + reg + log(lot) +
    log(bdms) + log(fb) + log(sty), data = ag.data)
table2.tidy <- tidy(table2.lm)
table2 <- format(table2.tidy, big.mark = ",", digits = 3)
#### Table 3 Regression ####
table3.lm <- lm(log(sell) ~ drv + rec + ffin + ghw + ca + gar + reg + log(lot) +
    bdms + fb + sty, data = ag.data)
table3.tidy <- tidy(table3.lm)
table3 <- format(table3.tidy, big.mark = ",", digits = 3)
```

Table 1: OLS estimation of parametric benchmark model dependent variable: log(P)

term	estimate	std.error	statistic	p.value
(Intercept)	7.9205	0.2192	36.14	1.46e-145
drv	0.1095	0.0284	3.86	1.26e-04
rec	0.0597	0.0262	2.28	2.28e-02
ffin	0.0956	0.0217	4.42	1.21e-05
ghw	0.1734	0.0441	3.93	9.63e-05
ca	0.1708	0.0213	8.03	6.24e-15
gar	0.0489	0.0115	4.24	2.65e-05
reg	0.1297	0.0228	5.69	2.07e-08
log(lot)	0.3129	0.0269	11.62	5.36e-28
log(bdms)	0.0888	0.0437	2.03	4.28e-02
log(fb)	0.2638	0.0312	8.45	2.78e-16
log(sty)	0.1652	0.0249	6.63	8.41e-11

Table 2: OLS estimation of parametric model w/ discrete variables in levels: dependent variable: log(P)

term	estimate	std.error	statistic	p.value
(Intercept)	7.7451	0.2163	35.80	5.01e-144
drv	0.1102	0.0282	3.90	1.07e-04
rec	0.0580	0.0261	2.23	2.65e-02
ffin	0.1045	0.0217	4.82	1.90e-06
ghw	0.1790	0.0439	4.08	5.22e-05
ca	0.1664	0.0213	7.80	3.29e-14
gar	0.0480	0.0115	4.18	3.43e-05
reg	0.1319	0.0227	5.82	1.04e-08
log(lot)	0.3031	0.0267	11.36	6.41e-27
bdms	0.0344	0.0143	2.41	1.63e-02
fb	0.1658	0.0203	8.15	2.52e-15
sty	0.0917	0.0126	7.27	1.30e-12

- Other than rounding differences. Everything is the same in both tables from the paper with one exception. Table III shows a difference for bedrooms. The paper says "0.344", but my regression gave "0.0344". This is probably due to a typo by the authors.

Problem 2. Correctly interpret the coefficients for the DRV and LOT variables for both regressions. Would you recommend standardizing or mean-centering any variables? (1 point)

- The dependent variable for both regressions is the log of the sell price. For the driveway variable (drv), the intrepretation is log-level. This means that, ceteris paribus, the driveway is correllated with an approximate 11% increase in sell price in both tables. For the lot size variable (lot), the interpretation is log-log. This means that, ceteris paribus, for every percent increase in the lot size, the sell price would increase by approximately 0.3% in both tables.

- There is no need to standardize or mean-center any of the binomial dummy variables that just tell us if an option is included: drv, rec, ffin, ghw, ca and reg.
- The lot size (lot) cannot be zero. So, it probably makes the most sense to mean center that variable. Similarly, I should mean center number of bedrooms (bdms), number of full bathrooms (fb), number of stories (sty) and number of garage places (gar). However, I wouldn't want a partial number for any of these so it would be best to use the mode. The mode for garage is actually already 0.

```
#### Question 2 ####
summary(ag.data$lot)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    1650    3600    4600    5150    6360   16200
```

```
summary(ag.data$bdms)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   1.000   2.000   3.000   2.965   3.000   6.000
```

```
summary(ag.data$fb)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   1.000   1.000   1.000   1.286   2.000   4.000
```

```
summary(ag.data$sty)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   1.000   1.000   2.000   1.808   2.000   4.000
```

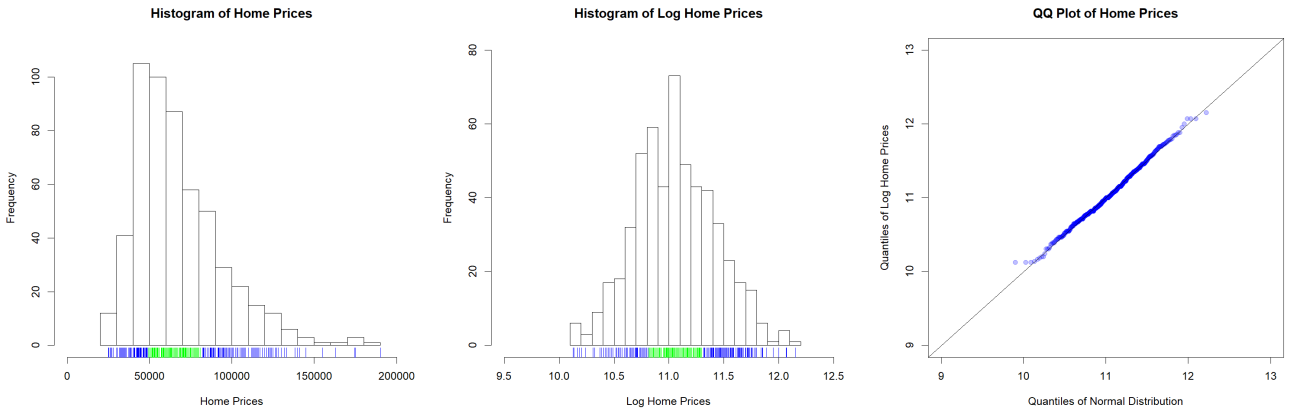
```
summary(ag.data$gar)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.0000  0.0000  0.0000  0.6923  1.0000  3.0000
```

Part 2

Problem 1. Create histograms of the house prices, log house prices, lot size, number of bedrooms, number of full bathrooms, and number of stories, with a summary of what you're seeing. (1 point)

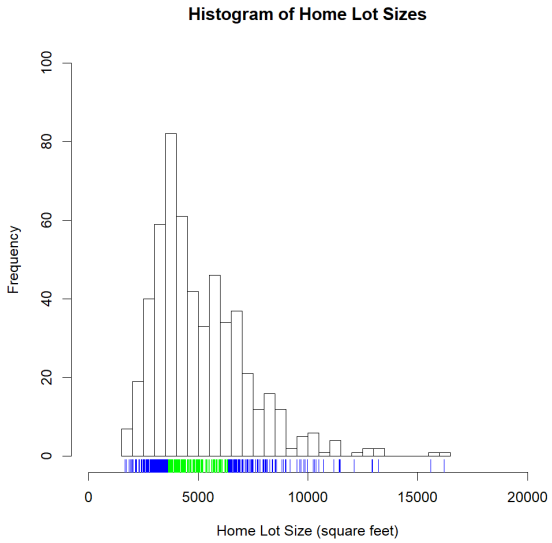
```
#### Part 2 #### Question 1 Histograms #### Took this code from week3.r Home Prices
#### Histogram
summary(ag.data$sell)
png(filename = "HomePrices.png", width = 1000, height = 1000, res = 100)
par(cex = 1.5)
hist(ag.data$sell, breaks = "FD", ylim = c(0, 110), xlab = "Home Prices", main = "Histog
    xlim = c(0, 2e+05))
rug(ag.data$sell, col = rgb(0, 1, 0, 1))
rug(ag.data$sell[ag.data$sell > 81999], col = rgb(0, 0, 1, 1))
rug(ag.data$sell[ag.data$sell < 49126], col = rgb(0, 0, 1, 1))
dev.off()
# Log of Home Prices Histogram
summary(log(ag.data$sell))
png(filename = "LogHomePrices.png", width = 1000, height = 1000, res = 100)
par(cex = 1.5)
hist(log(ag.data$sell), breaks = "FD", ylim = c(0, 80), xlab = "Log Home Prices",
    main = "Histogram of Log Home Prices", xlim = c(9.5, 12.5))
rug(log(ag.data$sell), col = rgb(0, 1, 0, 1))
rug(log(ag.data$sell)[log(ag.data$sell) > 11.3], col = rgb(0, 0, 1, 1))
rug(log(ag.data$sell)[log(ag.data$sell) < 10.81], col = rgb(0, 0, 1, 1))
dev.off()
# QQplot of Log Home Prices Adapted code from Section 4.1.2 of Alex Davis'
# Lecture Notes
k <- dim(ag.data)[1]
p <- (1:k - 0.5)/k
norm.q <- qnorm(p, mean = mean(log(ag.data$sell)), sd = sd(log(ag.data$sell)))
prices.q <- quantile(log(ag.data$sell), p, type = 5)
png(filename = "LogHomePricesQQ.png", width = 1000, height = 1000, res = 100)
par(cex = 1.5)
min.q <- min(floor(min(prices.q, norm.q)))
max.q <- max(ceiling(max(prices.q, norm.q)))
plot(norm.q, prices.q, xlim = c(min.q, max.q), ylim = c(min.q, max.q), ylab = "Quantiles
    xlab = "Quantiles of Normal Distribution", main = "QQ Plot of Home Prices", pch = 19
    col = rgb(0, 0, 1, 0.25))
abline(0, 1)
dev.off()
```

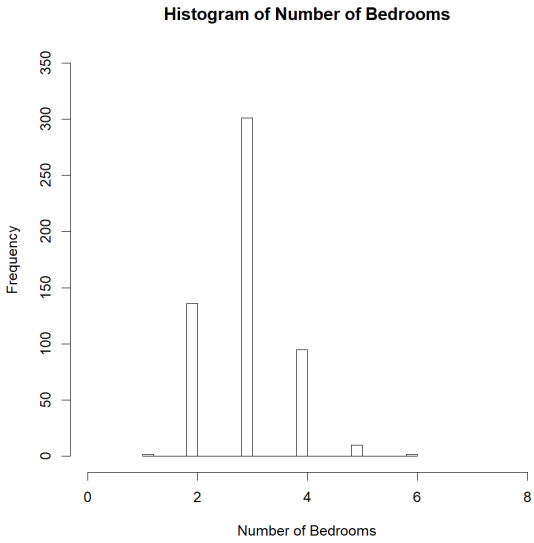


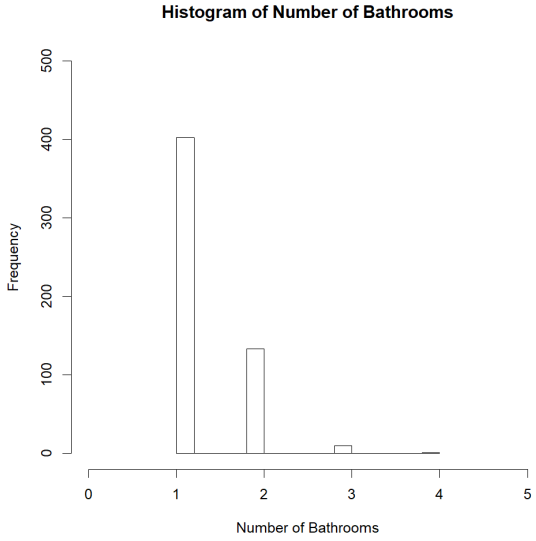
- In the first two figures, the green lines at the bottom represent the 2nd and 3rd quartiles of the data while the blue lines represent the first and fourth quartiles. The first figure shows that the distribution of home sell prices is positively skewed. The middle figure shows that once it is logged, it becomes a log normal distribution with some possibly heavy tails. The QQ plot on the left shows that the log-normal distribution is pretty well behaved. There do appear to be tails on both ends that could be compression.

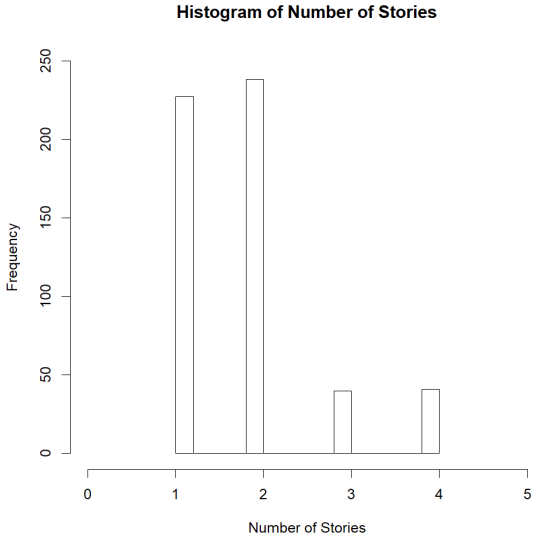
```
#### Home Lot Size Histogram ####
summary(ag.data$lot)
png(filename = "HomeLotSizes.png", width = 1000, height = 1000, res = 100)
par(cex = 1.5)
hist(ag.data$lot, breaks = "FD", ylim = c(0, 100), xlab = "Home Lot Size (square feet)",
    main = "Histogram of Home Lot Sizes", xlim = c(0, 20000))
rug(ag.data$lot, col = rgb(0, 1, 0, 1))
rug(ag.data$lot[ag.data$lot > 6359], col = rgb(0, 0, 1, 1))
rug(ag.data$lot[ag.data$lot < 3601], col = rgb(0, 0, 1, 1))
dev.off()

#### Number of Bedrooms Histogram ####
summary(ag.data$bdms)
png(filename = "HomeBedrooms.png", width = 1000, height = 1000, res = 100)
par(cex = 1.5)
hist(ag.data$bdms, breaks = "FD", ylim = c(0, 350), xlab = "Number of Bedrooms",
    main = "Histogram of Number of Bedrooms", xlim = c(0, 8))
dev.off()

#### Number of Bathrooms Histogram ####
summary(ag.data$fb)
png(filename = "HomeBathrooms.png", width = 1000, height = 1000, res = 100)
par(cex = 1.5)
hist(ag.data$fb, breaks = "FD", ylim = c(0, 500), xlab = "Number of Bathrooms", main = "
    xlim = c(0, 5))
dev.off()

#### Number of Stories Histogram ####
```

```
summary(ag.data$sty)
png(filename = "HomeStories.png", width = 1000, height = 1000, res = 100)
par(cex = 1.5)
hist(ag.data$sty, breaks = "FD", ylim = c(0, 250), xlab = "Number of Stories", main = "H
    xlim = c(0, 5))
dev.off()
```









- The lot sizes look positively skewed similar to the way the sell prices were skewed. The distribution looks similar as well. It might be best to perform a log transform on them as well which is exactly what I did in both of the regressions I ran. The majority of homes have 3 bedrooms. This seems like it's a reasonable average amount of rooms. Especially considering that it's a similar number to the number of 2- and 4-bedroom houses. It looks like $\frac{3}{4}$ of the homes have one full bath and the remining quarter have more than one. About 500 homes are nearly evenly split between 1- and 2-story homes. There are a little less than 100 homes with more than two stories.

Problem 2. Create tables (using the table() function) of the driveway, recreational room, full and finished basement, gas for hot water heating, central air conditioning, and preferred neighborhood dummy variables, with a summary of what you're seeing. (1 point)

```
#### Question 2 Tables ####
table(ag.data$drv)
```

```
##
##   0   1
##  77 469
```

```
table(ag.data$rec)
```

```
##
##   0   1
## 449  97
```

```
table(ag.data$ffin)
```

```
##
##   0   1
## 355 191
```

```
table(ag.data$ghw)
```

```
##
##   0   1
## 521  25
```

```
table(ag.data$ca)
```

```
##
##   0   1
## 373 173
```

```
table(ag.data$gar)
```

```
##
##   0   1   2   3
## 300 126 108  12
```

```
table(ag.data$reg)
```

```
##
##   0   1
## 418 128
```

- The majority of homes have driveways. Very few homes have recreational rooms. About $\frac{3}{5}$ of the homes have a fully-finished basement. Very few of the homes use gas for the hot water heater. Only about 31% of the homes have air conditioning. About $\frac{3}{5}$ of the homes do not have a garage. About 100 have a one-car garage, another 100 have a two-car garage and 12 have a three-car garage. Finally, about $\frac{1}{4}$ of the homes are in the preferred neighborhood of the city. So, the only variables that probably actually matter for the sake of regression are: fully-finished basement, garages, and air conditioning.

Problem 3. Comment on whether you think the Gauss-Markov assumptions are met for the benchmark model in Table III. Which assumptions are likely to be met or not met and why? (1 point)

- The Gauss-Markov assumptions are:

1. Linear in parameters - As the data has shown, if I log transform the sell prices and the lot sizes have fairly linear parameters. So, I should be able to make a pretty good model.

2. Random sampling - This depends on how we define our question. If we are asking about houses sold in July-September of 1987, then we have the entire population and not just a random sample. If we are trying to predict future sales or past sales or sales for any other location, then we do not have a randome sample. So, we are dealing with an unknown sampling error. But this assumption is definitley not met.

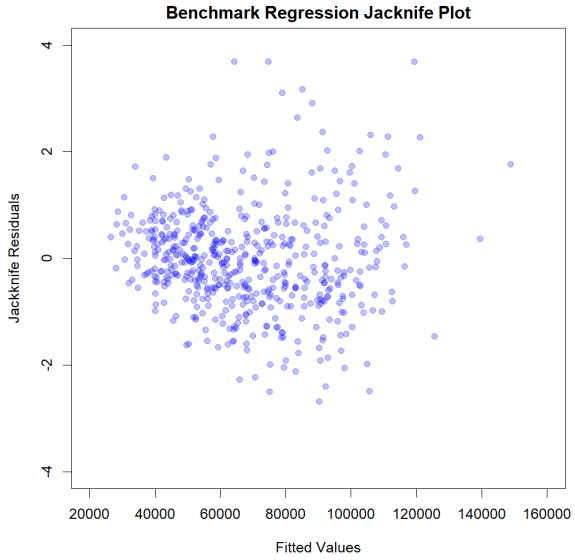
3. Zero conditional mean of the errors - I'll need to examine the residuals to see if they have a zero conditional mean. It's entirely possible that there are omitted variables. Without random assignment of the features, it's impossible to know and this is a very difficult condition to meet.

4. No perfect collinearity - I can be fairly certain of this condition becasue the regressions I ran earlier didn't drop any variables. So there is no perfect collinearity.

Problem 4. Plot the jackknife residuals versus fitted values for the benchmark regression using house prices rather than log house prices. What do you see? Using the Box-Cox scheme, what value of λ would you suggest for a transformation of the house prices? Is the log transform appropriate for the dependent variable? Why or why not? (1 point)

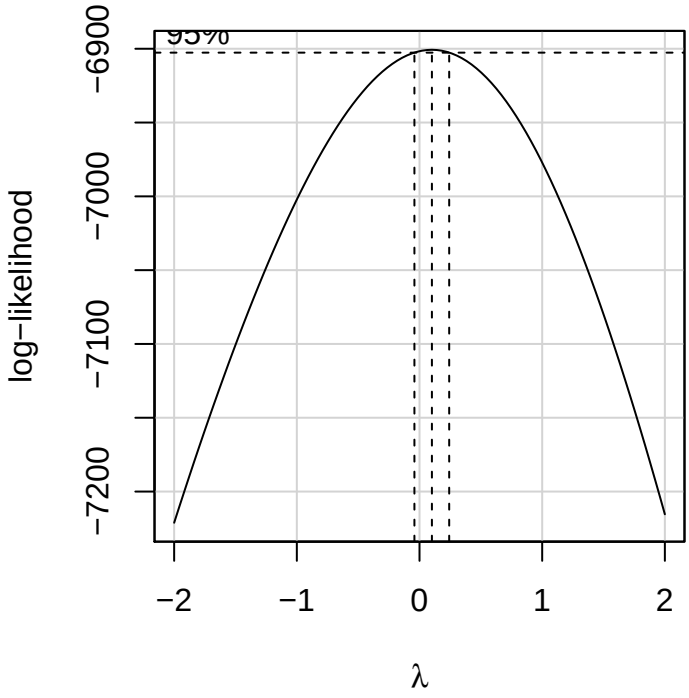
```
#### Question 4 Jackknives #### Modeled off code from week67.r from class
table3.1 <- lm(sell ~ drv + rec + ffin + ghw + ca + gar + reg + log(lot) + bdms +
    fb + sty, data = ag.data)
png(filename = "BenchmarkJK.png", width = 1000, height = 1000, res = 100)
par(cex = 1.5, mar = c(5, 4, 2, 1))
plot(fitted(table3.1), rstudent(table3.1), xlab = "Fitted Values", ylab = "Jackknife Res
```

```
    main = "Benchmark Regression Jacknife Plot", ylim = c(-4, 4), xlim = c(20000,
        160000), pch = 19, col = rgb(0, 0, 1, 0.25))
dev.off()
```



- The jackknife residuals appear to slope downward at first, but then slope upward. They appear to start positive dip into the negative and then go positive again. The residuals seem to disperse more as the residuals increase.

```
#### Box-Cox ####
benchmark.bc <- boxCox(ag.data$sell ~ ag.data$drv + ag.data$rec + ag.data$ffin +
    ag.data$ghw + ag.data$ca + ag.data$gar + ag.data$reg + log(ag.data$lot) + ag.data$bo
    ag.data$fb + ag.data$sty, family = "yjPower")
```




```
# Find the value of lambda to maximize log-likelihood
benchmark.bc$x[benchmark.bc$y == max(benchmark.bc$y)]
```

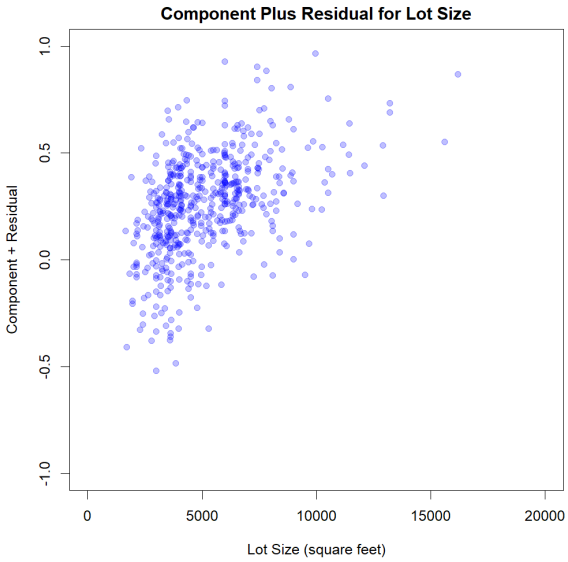
```
## [1] 0.1010101
```

- The value of λ that best maximizes teh log-likelihood is 0.1010101. That appears to match the previous analysis I did of the log transform of the sell price.

Problem 5. Create a component plus residual plot for the untransformed lot size variable in the benchmark regression model, using log transformed house prices. What transformation does the component plus residual plot suggest? Compare this to the value of λ that the Box-Tidwell function suggests. What transformation of lot size should we use (if any)? (1 point)

```
#### Question 5 #### Modeled after code in week67.r from class
cpr.lot <- lm(log(sell) ~ drv + rec + ffin + ghw + ca + gar + reg + lot + bdms +
    fb + sty, data = ag.data)
png(filename = "CPRlot.png", width = 1000, height = 1000, res = 100)
par(cex = 1.5, mar = c(5, 4, 2, 1))
plot(ag.data$lot, coef(cpr.lot)["lot"] * ag.data$lot + resid(cpr.lot), xlab = "Lot Size
```

```
    ylab = "Component + Residual", main = "Component Plus Residual for Lot Size",
    ylim = c(-1, 1), xlim = c(0, 20000), pch = 19, col = rgb(0, 0, 1, 0.25))
dev.off()
```



- This plot appears to be close to linear, but perhaps the Box-Tidwell will tell us more.

```
#### Box-Tidwell #####
benchmark.bt <- boxTidwell(log(sell) ~ lot, other.x = ~drv + rec + ffin + ghw + ca +
    gar + reg + bdms + fb + sty, data = ag.data)
benchmark.bt
```

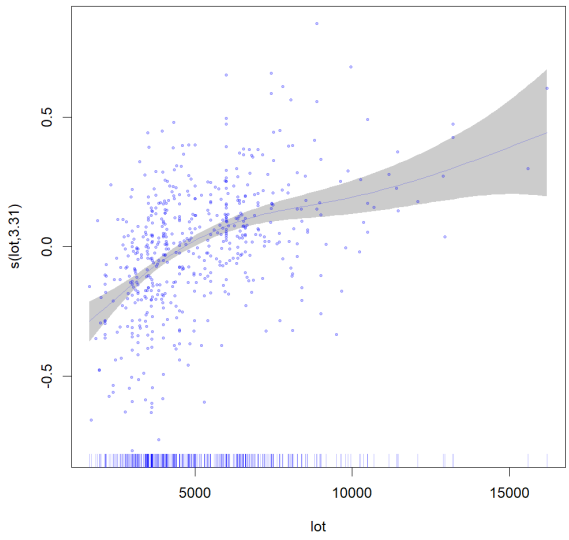
```
##  Score Statistic   p-value MLE of lambda
##        -3.886939 0.0001015    -0.3326529
##
## iterations =  3
```

- The Box-Tidwell analysis gives a value of -0.33 for the MLE of λ. This is not equivalent to a log transform, but a log transform may be easier to interpret.

Problem 6. Create a semi-parametric Generalized Additive Model (GAM) as an extension of the benchmark model with log transformed house prices. Use smoothing only for the lot size variable. Look at the partial residual plot of the lot size variable from the GAM. What transformation to the lot size variable looks appropriate? How does this compare to your conclusion from the component plus residual plot and Box-Tidwell transformation? (1 point)

```
#### Question 6 GAM ####
benchmark.gam <- gam(log(sell) ~ drv + rec + ffin + ghw + ca + gar + reg + s(lot) +
```

```
    bdms + fb + sty, data = ag.data)
png(filename = "GAMlot.png", width = 1000, height = 1000, res = 100)
par(cex = 1.5, mar = c(5, 4, 2, 1))
plot(benchmark.gam, residuals = TRUE, shade = TRUE, pch = 19, col = rgb(0, 0, 1,
    0.25), cex = 0.5)
dev.off()
```



- The GAM suggest a monotonic transformation is needed. This confirms the CPR analysis I just performed. There is no evidence of a need for a polynomial transformation. So, I can probably trust the Box-Tidwell analysis.

Problem 7. For the forecasting story, compare the semi-parametric GAM used in the previous section against the benchmark model with whatever you think is the most appropriate transformation for the lot size variable (based on what you've learned from the component plus residual plot, Box-Tidwell transform, and partial residual plot from the GAM). Does your model or the semi-parametric model perform better in terms of cross-validation? Summarize what you think this means. (1 point)

```
#### Question 7 Forecasting Story #### Based on week4.r and week5.r code from class
benchmark <- c()
cube.root <- c()
benchmark.gam2 <- c()

ag.data$lot.cubrt <- (ag.data$lot^(-1/3) - 1)/(-1/3)

forecast <- function(folds = 5) {
```

```
    # Construct the folds
    fold.num <- rep(1:folds, length.out = nrow(ag.data))
    fold.ran <- sample(fold.num)
    for (i in 1:folds) {
        # Construct training and test sets
        train <- ag.data[fold.ran != i, ]
        test <- ag.data[fold.ran == i, ]
        # Fit models to training data
        benchmark.lm <- lm(log(sell) ~ drv + rec + ffin + ghw + ca + gar + reg +
            log(lot) + bdms + fb + sty, data = train)
        cuberoot.lm <- lm(log(sell) ~ drv + rec + ffin + ghw + ca + gar + reg + lot.cubr
            bdms + fb + sty, data = train)
        gam2.lm <- gam(log(sell) ~ drv + rec + ffin + ghw + ca + gar + reg + s(lot) +
            bdms + fb + sty, data = train)
        # Test error
        benchmark.test <- (log(test$sell) - predict(benchmark.lm, newdata = test))^2
        cuberoot.test <- (log(test$sell) - predict(cuberoot.lm, newdata = test))^2
        gam2.test <- (log(test$sell) - predict(gam2.lm, newdata = test))^2
        # Store results
        benchmark <- append(benchmark, benchmark.test)
        cube.root <- append(cube.root, cuberoot.test)
        benchmark.gam2 <- append(benchmark.gam2, gam2.test)
    }
    # Test rmse
    rMSEbenchmark.lm <- sqrt(sum(benchmark)/(length(benchmark)))
    rMSEcuberoot.lm <- sqrt(sum(cube.root)/(length(cube.root)))
    rMSEgam2.lm <- sqrt(sum(benchmark.gam2)/(length(benchmark.gam2)))
    return(list(rMSEbenchmark.lm, rMSEcuberoot.lm, rMSEgam2.lm))
}

# Replicate the 5-fold cross-validation 100 times
cvs <- replicate(100, forecast())
# Pull out the average rMSE for each model
cv.benchmark <- mean(sapply(cvs[1, ], mean))
cv.cuberoot <- mean(sapply(cvs[2, ], mean))
cv.gam2 <- mean(sapply(cvs[3, ], mean))
cv.benchmark
```

```
## [1] 0.2135511
```

```
cv.cuberoot
```

```
## [1] 0.2133382
```

```
cv.gam2
```

```
## [1] 0.2142089
```

- According to the average rMSE of each model show that the benchmark (0.2137) and the cubedroot (0.2135) model are better than the GAM model which has a slightly higher rMSE of 0.2144. The cubed root model is the best, but it's so close to the benchmark that it's much easier to explain using the benchmark log transform.

Problem 8. Compare heteroskedasticity robust standard error estimates to "classical" standard error estimates for the benchmark model in Table III with log transformed house prices. Is there any difference in the standard errors? Comment on what this says about potential model misspecification. (1 point)

```
#### Question 8 Hetero/Homoskedasticity ####
classical <- coeftest(table3.lm)
robust <- coeftest(table3.lm, vcov = vcovHC(table3.lm, type = "HC0"), df = df.residual(
classical.se <- classical[, 2]
robust.se <- robust[, 2]
ratio <- robust.se/classical.se
ratio
```

```
## (Intercept)         drv         rec        ffin         ghw          ca
##   0.9706409   1.0142732   0.9814721   1.0790799   1.1688034   0.9694029
##         gar         reg    log(lot)        bdms          fb         sty
##   1.0133640   0.9302575   0.9827216   1.0215618   1.0462885   0.8655062
```

- The ratio of the robust to classical standard errors is nearly 1 for all of the variables. The variables with the greatest difference are the gas-powered water heater (~16.8% larger) and the stories (~13.5% smaller). As I showed in problem 2 there are only 25 homes with gas-powered water heaters. So, it's not going to tell us much.

Problem 9. Briefly comment on the statistical inference and causality stories. (1 point)

- Statistical inference - this goes back to the same question as in the Gauss-Markov assumption about a random sample. If the sample is the population, there is nothing to infer. If, however, we are trying to use this data to infer something about other times or places, we have no way of knowing how applicable it may be to those cases.
- Causality - There is no reason to think that any of the effects can be interpreted causally. There was no random assignment and therefore it would be irresponsible to make any causal inferences. There may

also be omitted variables that may have more causal explanatory power than those for which we have data and there's no way for us to find out or test it.