Applied Regression Analysis Final:

Predicting Homeprices

Wynetta Hererra
Demetrios Papakostas
STP 530

December 4th, 2018

3.31 Answer: We first fit a reduced linear regression of the usual form with the price as the response and house square footage as the explanatory X variable. We picked a random sample with size $n = 200$ and set the seed=123. The coefficient of correlation for the fit was 0.675. Despite this being a relatively "high" coefficient of correlation, figure 1 indicates a linear fit may not be appropriate, due to the fanning out of the residuals at higher square footage. In figure (1), we plot the fit and the residuals vs square footage. The regression model is

$$\hat{Y} = -80794 + 159X$$

which is clearly flawed just from the interpretation of β_0 alone. It is pretty clear that the variance does

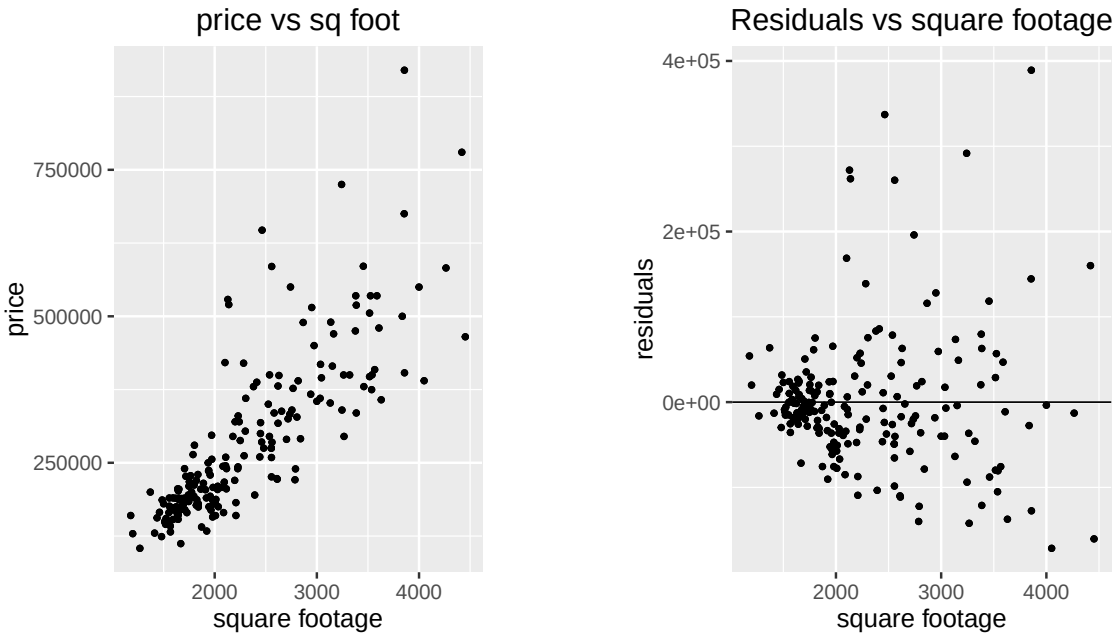

Figure 1: Initial diagnostic plot

not appear to be constant, as for larger homes, the price appears to increase non-linearly, which is not entirely surprising.

Expected values under normality assumption

A qq-normal plot would show assumptions are violated in our regression, but we can show this beyond visual aids, by conducting various tests. First, our null hypothesis is that distribution of our error terms do not depart substantially from a normal distribution. Then, we obtain the correlation of coefficient between the expected values of the residuals under normality versus the residuals. The expected values are obtained from

$$\sqrt{MSE}[z(\frac{k-0.375}{n+0.25})]$$

Controlling the α risk at 0.05, we find for n=100, the critical value (obtained from table B.6) is 0.987. For n=200, this value would be closer to 1. Our observed coefficient,0.727, is well below the critical value, hence we reject our hypothesis that error terms do not depart substantially from a normal distribution. Relevant R-code is inserted.

```
library(MASS)
library(ggplot2)
library(tidyverse)
theme_set(theme_gray(base_size = 16))
data=read.table('RealEstate.txt',    header=FALSE,
```

```
 col.names = c("ID", "price", "sqft", "nbed", "nbath","AC",
"garagesize", "pool", "year", "quality","style",
"lotsize", "highway"))

set.seed(123)
idx=sample(522, 200, replace=FALSE)

subdata = data[idx,]
 x=subdata$sqft
 fit1=lm(price~sqft,data=subdata)
#summary(fit1)
yhat=fitted(fit1)
n<-length(subdata$price)
SSE = t(resid)%*%resid  #datay is residuals
MSE=SSE/(n-2)  #mean square error

#p=2 because we only have simple linear regression


####Expected residual values####
ExpVals= sapply(1:n, function(k) sqrt(MSE) *qnorm((k-.375)/(n+.25)))

oderedExpVals<- ExpVals[rank(resid)]

summary(lm(resid~orderedExpVals))
####semi-studentized t
semistudenttval<-resid/sqrt(MSE)
##using MASS package as verification##
studresid=studres(fit1)
studresid[studresid>3.34]
#semistudenttval[semistudenttval>1|semistudenttval< -1]
c<-qplot(yhat, semistudenttval) +ggtitle('semi-student residual vs fitted price')+
geom_hline(yintercept=0)+theme(plot.title = element_text(hjust = 0.5))

length(semistudenttval[semistudenttval>1|semistudenttval< -1])/length(subdata$price)
```

Semi-studentized Residuals

A plot of the semistudentized residuals against the fitted values is presented in figure (2)

Search for outliers

The null hypothesis is that there are no outliers, and the alternative is there are outliers in our set. To test this, we compare the studentized deleted residuals vs the Bonferroni critical value, noting that we include n tests for all deleted residuals, because if the regression model is appropriate, so that no case is outlying because of a change in the model, then each studentized deleted residual will follow the t distribution with $n-p-1$ degrees of freedom. The studentized deleted residuals are calculated from (where h_{ii} is the diagonal elements of the hat matrix.

$$t_i = e_i \left[\frac{n-p-1}{SSE(1-h_{ii})-e_i^2} \right]^{1/2}$$

The Bonferroni simultaneous test procedure with family significance level $\alpha = 0.10$ requires

$$t(1-\alpha/2n; n-p-1) = 3.34$$

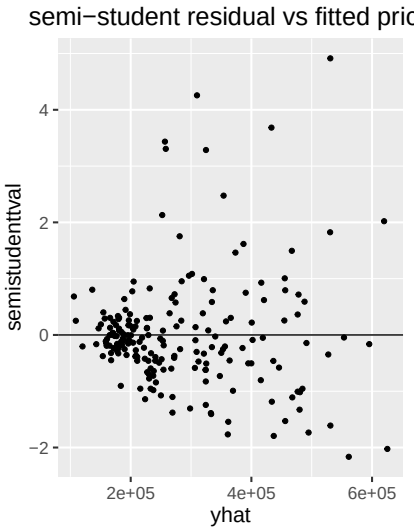

Figure 2: Plot of the semi-studentized residuals vs the expected y-values.

Therefore, we compare the studentized residuals vs 3.34, ie

$$|t_i| \leq 3.34$$

we conclude there are outliers at $\alpha = 0.10$ significance level, as in our random subsample, 6 of the residuals are greater than 3.34. The indices are 73, 69, 108, 203, 80, and 79.

We can also test for X outliers. We can obtain the elements of h_{ii} and if these leverage values are larger than $2p/n = 0.02$. The diagonal elements are obtained from

$$h_{ii} = X_i^T (X^T X)^{-1} X_i$$

where

$$X_i = \begin{pmatrix} 1 \\ X_{i,1} \\ \vdots \\ X_{i,p-1} \end{pmatrix}$$

In this subsample, there are 13 leverage values larger than $2p/n$, thus we conclude those are outlying X observations. The indices are 125,73,97,174,99,191,81, 202, 129, 70, 144, 210, 128, 94, and 153. However, these correspond to square footage of 3858,3857, 4264, 3630, 3608, 3566, 4419, 3536, 4453, 4000, 3855, 4050, 3836, 3588, 3540. These are all large homes, so we may not want to remove these points, as otherwise we will remove large houses from our model, which could be problematic.

Some relevant code is below:

```
####calculate the H-matrix####

hdiag=lm.influence(fit1)$hat
#indices where X is outlier (X is square footage)
hdiag[hdiag>0.02]
#to find the square footages
subdata$sqft[hdiag>0.02]
```

F-Test for lack of fit

Our full model is given by $Y_{ij} = \mu_j + \varepsilon_{ij}$. However, the general linear test approach requires consideration of the reduced model under H_0.

$$H_0 : E\{Y\} = \beta_0 + \beta_1 X$$
$$H_a : E\{Y\} \neq \beta_0 + \beta_1 X$$

We note that the F^* test statistic is calculated from

$$F^* = \frac{SSLF}{c-2} / \frac{SSPE}{n-c} = \frac{MSLF}{MSPE}$$

The decision test is

$$\text{If} \quad F^* \leq F(1-\alpha; c-2, n-c) \quad \text{Conclude } H_a$$
$$\text{If}^* \quad F^* > F(1-\alpha; c-2, n-c), \quad \text{Conclude } H_a$$

With confidence coefficient 0.95, we have an F^* statistic of 3.48, which is greater than 1.99, the decision rule value, therefore we conclude H_a.

Relevant R-code is inserted below:

```
library(alr3)
####F test for lack of fit####
alpha=.05
#full model
# with as.factor(x), we treate x as categorical vairable with 10 levels
# the model is "y~0+...", meaning that there is no intercept
fullfit = lm(price~0+as.factor(sqft), data=subdata)


#lack-of-fit test to compare the reduced and full models
anova1=anova(fit1,fullfit)



betternova=pureErrorAnova(lm(log(price)~log(sqft), data=subdata))
SSLF=betternova[3,2]

SSPE=betternova[4,2]
#SSLF=SSE-SSPE




c=length(unique(x))

#alpha=0.01
#critval=qf(1-alpha, c-2, n-c )
MSPE=SSPE/(n-c)


MSLF=SSLF/(c-2)

Fstat=MSLF/MSPE

Fcrit=qf(1-alpha, c-2, n-c)
```

Brown-Forsythe Test

Dividing data into two groups, $X \leq 2000$ and $X > 2000$ (where X is the square footage, and 2000 is motivated from the visual cue in figure (1) and using $\alpha = 0.05$, we perform the Brown-Forsythe test. We note that if we break data into groups, then we use e_{i1} to the denote the ith residual for group 1, and e_{i2} to denote the ith residual for group 2. (We do a usual linear regression but split the residuals according to which group they fall in, ie which X_i they "belong to".) We use n_1 and n_2 (which are 92 and 108 respectively) to denote sample sizes of the two groups. The two-sample t test statistic becomes

$$t^*_{BF} = \frac{\bar{d}_1 - \bar{d}_2}{s\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

where $\bar{d}_1$ and $\bar{d}_2$ are the sample samples of d_{i1} and d_{i2} respectively. The pooled variance is given by

$$s^2 = \frac{\sum(d_{i1} - \bar{d}_1)^2 + \sum(d_{i2} - \bar{d}_2)^2}{n-2}$$

If error terms have constant variance and n_1 and n_2 are not extremely small, t^*_{BF} follows approximately t distribution with $n-2$ degrees of freedom.

Our $|t^*_{BF}| = 6.08$. The t-value with 200-2 degrees of freedom at .95 confidence coefficient is 1.97. Since $|t^*_{BF}| > 1.97$, we conclude the error variance is not constant and does vary with the level of X.

Relevant code is listed below:

```
####Brown-Forsythe test####
alpha=0.05
brownnew<-subdata[subdata$sqft<=2000,]

n1<-length(brownnew$sqft)

brownnew2<-subdata[subdata$sqft>2000,]
n2<-length(brownnew2$sqft)


e1<-resid[subdata$sqft<=2000]#take the residuals which correspond to X<=2000
e2<-resid[subdata$sqft>2000]


d11=abs(e1-median(e1))
d12=abs(e2-median(e2))
d1=mean(d11)
d2=mean(d12)
numbrown<-d1-d2

denombrown<-sqrt((1/n1)+(1/n2))
d1sum<-sum((d11-d1)^2)
d2sum<-sum((d12-d2)^2)
s=sqrt((d1sum+d2sum)/(n-2))

tbf=(numbrown)/((s*denombrown))
critvalue<-qt(1-alpha/2,n-2)
```

Box-Cox Transformation

We first perform a Box-Cox transformation on the data to remedy the issue with the prices skewing outward at larger square footages. We use box-cox to find a good lambda for a transformation. The R Box-cox uses log-likelihood instead of SSE. The plot if represented in figure (3). Relevant code is inserted below:

Table 1: Log Likelihood

λ	-0.5	✓-0.3	-0.2	-0.1	0	0.1	0.2	0.3
Log Likelihood	-234.22	-232.57	-232.79	-233.71	-235.57	-237.983	-241.13	-245.02

```
box=boxcox(lm(price~sqft, data=subdata),lambda=c(-0.7,-0.6,-0.5,-.4,-.3,-.2,-.1,0,
.1,.2,0.3,0.4,0.5,0.6,0.7))

lambda = box$x;
loglikelihood = box$y;
lamlist<-c(loglikelihood[lambda==-.5],loglikelihood[lambda<=-.293&lambda>=-.305],
loglikelihood[lambda<=-.196&lambda>=-.208],
loglikelihood[lambda<=-.091&lambda>=-.11],
loglikelihood[lambda>=-0.002&lambda<=0.01],
loglikelihood[lambda>=0.094&lambda<=0.107],
loglikelihood[lambda>=0.194&lambda<=0.204],
loglikelihood[lambda>=0.299&lambda<=0.301])
lamlist
plot(lambda,loglikelihood, pch=16,main="original␣Y")
abline(0,0)
#plot(xdat, e2, pch=16, main="Y^(-0.5)")
#abline(0,0)
#plot(xdat, e3, pch=16, main="Log10(Y)")
#abline(0,0)

qqnorm(e1, main="original␣Y")
qqline(e1)

newfit<-lm(price^(-0.25)~sqft, data=subdata)

d<-qplot(lambda,loglikelihood)+geom_point()+geom_hline(yintercept=c(-235.57,-232.57))
+ggtitle('log␣likelihood␣vs␣lambda')+
theme(plot.title = element_text(hjust = 0.5))# pch=16,main="original Y")
```

Final Model

In the final model, we remove outliers and use the Box-Cox transformation price=ln(price). We use the ln transformation because λ is fairly stable from -0.5 to 0, and the value $\lambda = 0$ corresponding to a natural log transformation because of the ease in interpreting it. This is represented in figure (4), which indicates a better fit than previously seen in the first regression model we used.

We use this to predict the price when the square footage, X, takes on values 1100 and 4900. Additionally, we removed the outliers according to Bonferroni procedure compared with studentized deleted residuals, this time performed on the transformed Y's, leading to 3 outliers being removed, indices 108, 203, and 80.

On this new model, we again perform the Brown-Forsythe test, again with n_1 and n_2 corresponding to same cut-off points, just transformed by the ln transformation, but with the residuals from our ln-ln transformed linear regression. Again controlling the α risk at 0.05, this time our $\left|t^*_{BF}\right| = 1.36 < 1.97$, so this time we conclude H_0, that the error variance is constant. (Note that our critical value is the same to the hundredths place, despite losing three values).

Additionally, the R^2 of our new regression is 0.745, an improvement over the initial model. Of course this is not super telling on its own right, but in conjunction with the visual improvement, and the different conclusion from the Brown Forsythe test, this increases our confidence in this model. Therefore, we use this model to predict the final housing price.

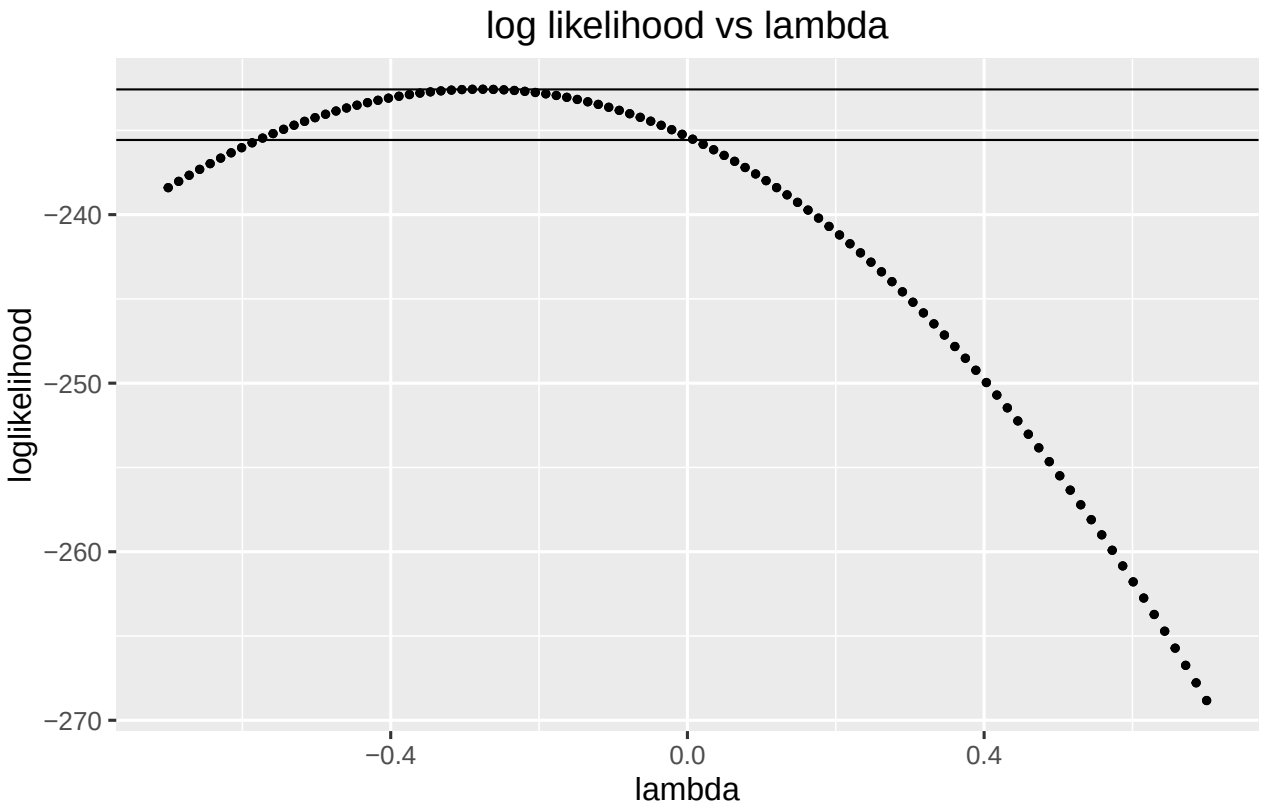

Figure 3: max log likelihood around $\lambda = -0.3$

We obtain a prediction interval with 95% confidence using the Bonferroni procedure for a new observation Y_h(new) when $X = 1100$ and $X = 4900$. We find

$$\hat{Y}_n \pm t(1-\alpha/2; n-p)s\{\text{pred}\}$$

where

$$s^2\{\text{pred}\} = \text{MSE} + s^2\{\hat{Y}_h\}$$

and

$$s^2\{\hat{Y}_h\} = \text{MSE}(X_h^T(X^TX)^{-1}X_h) = X_h^T s^2\{\hat{\beta}\}X_h$$

Therefore, the two cases are, after transforming back from ln to normal variables, that the prediction intervals are (with 0.95)

$$\hat{Y}_{1100} = 105873$$
$$\hat{Y}_{4900} = 707858$$
$$68186 \le Y_{1100} \le 165380$$
$$455887 \le Y_{4900} \le 1110144$$

Relevant R-code is below.

```
##outliers in the log case##
subdatalog2=subdata[c(-108,-203, -80),]

xnew=subdatanew$sqft
newfit=lm(log(price)~sqft, data=subdatalog2)
newfit.summary=summary(newfit)
```

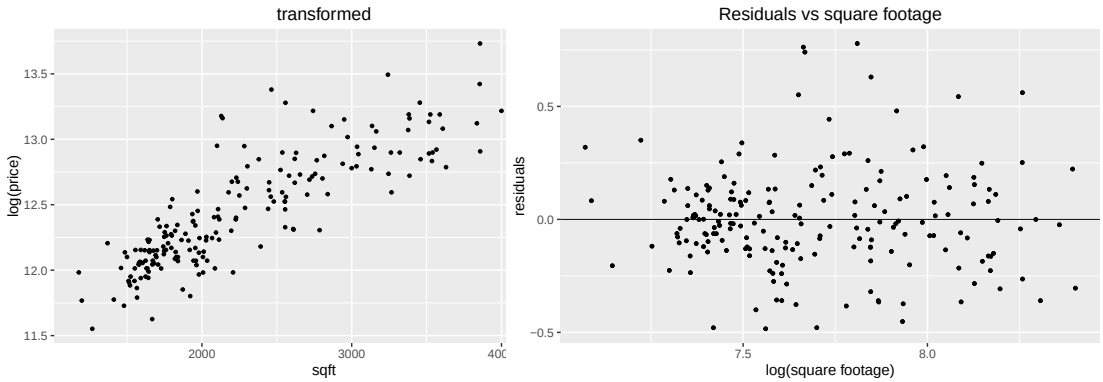

Figure 4: Transformed y-axis (taking the ln of the price) vs the square footage and the residuals versus the square footage. Indicative of a better fit.

```
beta0=newfit.summary$coefficients[1,1]
beta1=newfit.summary$coefficients[2,1]

Xnew=cbind(1,xnew)
XtransXinv=solve((t(Xnew)%*%Xnew))
residnew=residuals(newfit)
SSEnew=t(residnew)%*%residnew
MSEnew=SSEnew/(n-2)


#ssquared found for estimator of hat(beta),
ssquared=as.numeric(MSE)*XtransXinv

xh=data.frame(sqft=1100)#, sqft=4900)  #xh to get yhat for each h

yh=predict(newfit,xh, se.fit=TRUE, interval="predict",level = 0.95)


##manual method##
Xhat=as.matrix(xh)

Xhat=cbind(1,Xhat)

Xhat<-t(as.matrix(Xhat))
#ssquared found for estimator of hat(beta),

syhat=t(Xhat)%*%ssquared%*%Xhat  #s{hat{y}}

spred=syhat+MSE

p=2
alpha-0.05
crit=qt(1-alpha, n-p)


#ssquared found for estimator of hat(beta),
syhat=t(Xhat)%*%ssquared%*%Xhat  #s{hat{y}}
#syhat
spred=sort(syhat+MSE)
spred*crit
```

10.31 Answer: As evidenced by table (2), no obvious multicollinearity issues are apparent in the best fit model, as no VIF values are above 10. The best model was ln(price) as the response variable and square footage, number bath, pool, year built, quality, and style as predictors. Note, we made pool, quality, and style categorical variables.

Table 2: Test for multicollinearity issues.

Predictor variable	VIF	Df
sqft	3.879	1
nbath	2.961	1
pool	1.064	1
year	2.144	1
quality	3.535	
2 style	3.171	7
lotsize	1.259	1

Cook's distance is the influence on all fitted values, and is found from

$$D_i = \frac{\sum_{j=1}^{n}\left(\hat{Y}_j - \hat{Y}_{j(i)}\right)^2}{p \cdot \text{MSE}}$$

In figure (5), there is evidence to suggest we need to remove the indices 120, 135, 201, 38, 96, 160, 103, 129, 211, 281,177,24,37, 514,361, 127 as they are above the Cooks distance criteria. DFFITS gives

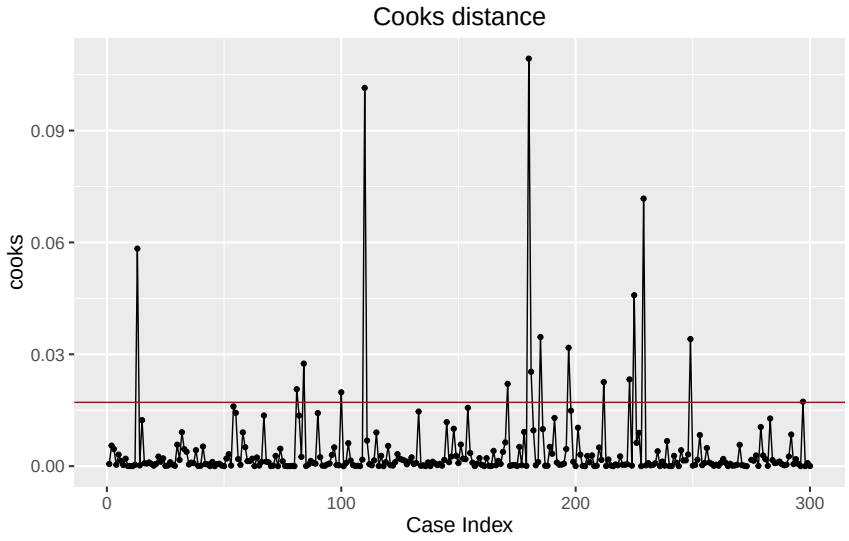

Figure 5: Cooks plot with outliers above red line. The outlying indices are 120, 135, 201, 38, 96, 160, 103, 129, 211, 281, 177, 24, 37, 514, 361, 127

influence on single fitted value. The $(\text{DFFITS})_i$ value is given by

$$(DFFITS)_i = \frac{\hat{Y}_i - \hat{Y}_{i(i)}}{\sqrt{MSE_i h_{ii}}}$$

We use the dffits criteria of $2 \cdot \sqrt{\frac{p+1}{n-p+1}}$. Note that in the best subset model, $p = 14$, because we have 13 predictors after categorizing certain variables, so of course we $p = 13 + 1 = 14$ parameters.

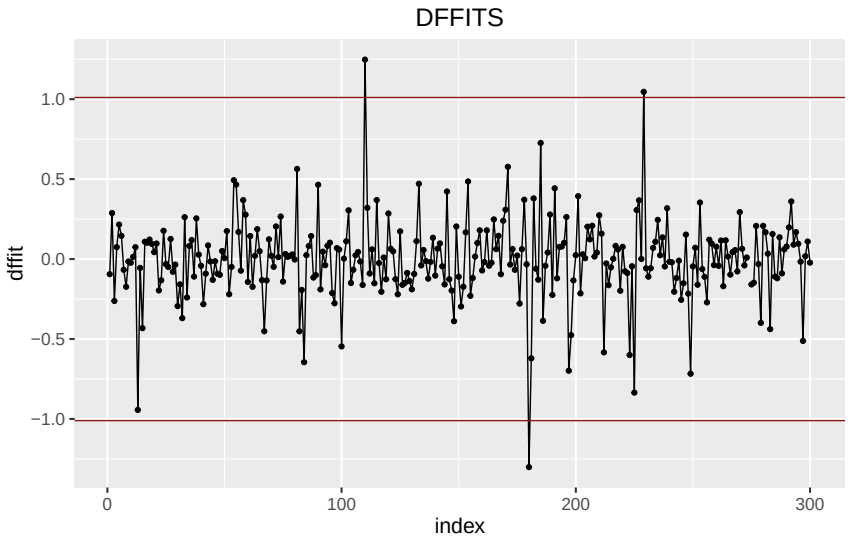

Figure 6: Dffits plot with outliers above and below red line. The outlying indices are 96, 103, 514.

Final Model Answer: After removing the appropriate influential outliers identified by the DFFITs test in 10.31, we fit our final model to the real estate data. We using the best subset from the exercise 9.33. The estimated parameters for each variable are given in table (3)

Table 3: Estimated parameters. Multiple R squared of 0.8452. Note, we are the predicting the natural log of the price.

Variable	Estimate	t-value	p-value
Intercept	1.68	1.402	0.162
sqft	3.22e-04	14.28	<2e-16
nbath	4.18e-02	3.40	0.00073
pool1	6.90e-02	2.22	0.027
year	5.13e-03	8.48	2.63e-16
quality2	-2.58e-01	-8.25	1.38e-15
quality3	-3.72e-01	-8.82	<2e-16
style2	-6.81e-02	2.495	0.013
style3	-1.55e-02	-0.596	0.55
style4	3.28e-02	0.580	0.56
style5	-4.29e-02	-0.969	0.33
style6	3.02e-02	0.677	0.499
style7	-1.06e-01	-4.15	3.97e-05
style9	-7.96e-02	-0.46	0.647
style10	-2.96e-01	-1.67	0.095
style11	-3.87e-01	-2.24	0.026
lotsize	4.50e-06	6.39	3.83e-10

Our assumptions appear to be met, as evidenced by figure (7).

We again use a variance inflation factor test once again shows no collinearity even on the full (not sampled) data set. The largest VIF factor is 4.27, well below the criteria of 10 established in the textbook and in class.

To interpret table (3), note that the estimate represents how much changing one unit of the variable affects the (natural log) of the price. We first begin by explaining β_0, the y-intercept coefficient.

For example, increasing square footage by 1 feet increases the natural log of the price by 0.000322 (the appropriate scaling since we log transformed the y-axis). This is not super intuitive. Rather, note that if we exponentiate both sides, our model becomes

$$\hat{Y} = \exp(\beta_0) \cdot \exp(\beta_1 x_1) \cdot \dots$$

So, with this, the interpretation of the coefficients is a little easier to swallow. For example, if we increase square footage by 1 foot, and hold all else constant, we find

$$\exp(\beta_1) = \frac{\exp(\beta_0) + \exp(\beta_1)}{\exp(\beta_0)} = 1.0003$$

Ie, for every 1 square foot we add, there is a .03% increase in price of home. Further, we can generalize this to any β_i coefficient, with $i = 1, \dots, p-1$. We have already discussed the β_0 case, and again noting that y represents the predicted price of a home,

$$\Delta\hat{Y} = \exp(\beta_1) \cdot \hat{Y}_{\text{old}} - \hat{Y}_{\text{old}} = \hat{Y}_{\text{old}}(\exp(\beta_1) - 1) \longrightarrow \frac{\Delta\hat{Y}}{\hat{Y}_{\text{old}}} = \exp(\beta_i) - 1$$

If we multiply $\frac{\Delta\hat{Y}}{\hat{Y}_{\text{old}}}$ by 100%, then we have the percent change in y. Therefore, a 1 square footage increase has a .003% increase on home price. Some variables have a negative effect, for example some

styles negatively affect the final price. For example, if your home is of style 2, our model suggests that, from the coefficient of style 2 being -6.81e-02, from our expression for percent change in y, we have a 6.58 % decrease in the price of home.

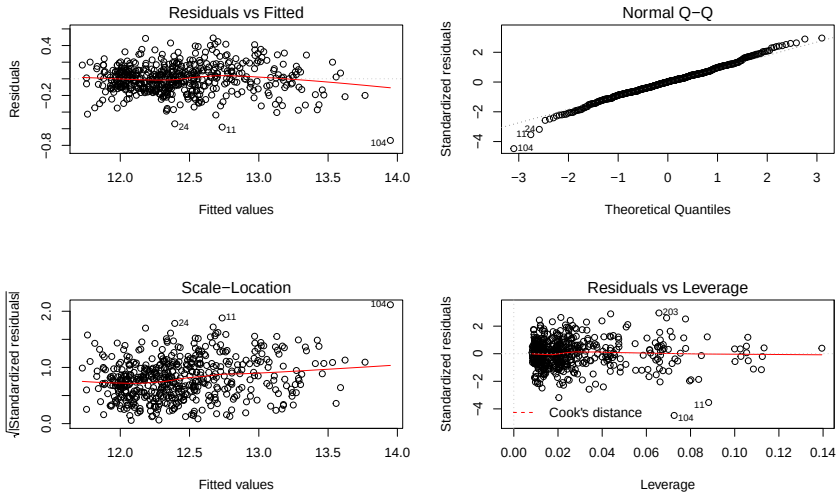

Figure 7: Assumptions appear to be met.

https://stats.idre.ucla.edu/other/mult-pkg/faq/general/faqhow-do-i-interpret-a-regression-model-when-some-variables-are-log-transformed/

good reference for interpreting transformed predictor and response variables.

We did not have garage size in our best model, so we did not include it when predicting our final prices. We transform the price back to US dollars by exponentiating the results from our prediction intervals. g is the number of confidence coefficients in the family. Code is listed below.

```
####very final part####

xh1=expand.grid(sqft=c(1500, 3000,4500),nbath=c(3,3,3), pool=levels(data$pool),
year=c(1988,1988,1988),  quality=levels(data$quality),style=levels(data$style),
lotsize=c(20000,20000,20000))#, sqft=4900)
 #xh to get yhat for each h
xh1new<-
xh1%>%
filter(style==5)%>%
filter(quality==1)%>%
filter(pool==0)%>%
head(3)# because of repetition


xmat1=matrix(c(1,1500,3,0,1988, 1,5,20000),nrow=8, ncol=1)
xmat2=matrix(c(1,3000,3,0,1988, 1,5,20000),nrow=8, ncol=1)
xmat3=matrix(c(1,4500,3,0,1988, 1,5,20000),nrow=8, ncol=1)

bigx=cbind(1,data$sqft,data$nbath, data$pool, data$year,
 data$quality, data$style, data$lotsize)

XtransXinvfin=solve(t(bigx)%*%bigx)
XtransXinvfin

fin=t(xmat3)%*%XtransXinvfin%*%xmat3



level=1-(alpha/g)
yh1=predict(model12new,xh1new, se.fit=TRUE, interval="predict",level = level)

hdiagfin=lm.influence(model12new)$hat

alpha = .05; g = 3, p=16
S = sqrt(g*qf(1-alpha, g, nfin-p))


## the quantile (B) for Bonferroni procedure when alpha = .1
B=qt(1-(alpha/(2*g)), nfin-p)
#is B>S
B>S
```

Since $S > B$, we use the Bonferroni procedure. With confidence coefficient 0.95, our final results are listed in table (4)

Table 4: Final Results

Condition	lower $	fit $	Upper $
1	180,737	280,304	434,708
2	293,607	454,148	697,557
3	400,312	735,804	1,113,972

Equation 10.29 is

$$h_{\text{new,new}} = X_{\text{new}}^T (X^T X)^{-1} X_{\text{new}}$$

We note that X_{new} is of the form in equation 10.18 (for the three conditions)

$$X_{\text{new}} = \begin{pmatrix} 1 \\ 1500 \\ 3 \\ 0 \\ 1988 \\ 1 \\ 5 \\ 20000 \end{pmatrix} \qquad X_{\text{new}}^T = \begin{pmatrix} 1 & 1500 & 3 & 0 & 1988 & 1 & 5 & 20000 \end{pmatrix} = 0.0728$$

$$X_{\text{new}} = \begin{pmatrix} 1 \\ 3000 \\ 3 \\ 0 \\ 1988 \\ 1 \\ 5 \\ 20000 \end{pmatrix} \qquad X_{\text{new}}^T = \begin{pmatrix} 1 & 3000 & 3 & 0 & 1988 & 1 & 5 & 20000 \end{pmatrix} = 0.0502$$

$$X_{\text{new}} = \begin{pmatrix} 1 \\ 4500 \\ 3 \\ 0 \\ 1988 \\ 1 \\ 5 \\ 20000 \end{pmatrix} \qquad X_{\text{new}}^T = \begin{pmatrix} 1 & 4500 & 3 & 1988 & 1 & 5 & 20000 \end{pmatrix} = 0.091$$

The computer tells us 10.29 yields For the three conditions respectively, we find

$$h_{\text{new1,new1}} = 0.0728$$
$$h_{\text{new2,new2}} = 0.0502$$
$$h_{\text{new3,new3}} = 0.091$$

We safely conclude there is no hidden extrapolation, as even 0.091, the largest value, is within the range of leverage values $h_{ii}f$ for the cases in the data set.