BTRY 6020 Lab IV Solutions

February 27, 2017

##Question 1: Multiple Linear Regression Example

In this example we look at how a fireplace is related to the selling price of a home. Specifically, can we quantify the monetary value of a fireplace as it results to the selling price of a home. The easy way to compare the value of a fireplace is to do a 2-sample t-test of the selling price of homes with and without fireplaces.

The data appears in the `Lab4q1Dat.xlsx` file. `Value` is the selling price of the house in thousands of dollars, and `Size` is the square footage of the apartment in thousands.

A) Plot the data, showing the relationship of `Size` and `Value` where `Firepl` data is also encoded. Be sure to include a legend with your plot by using the `legend` function.

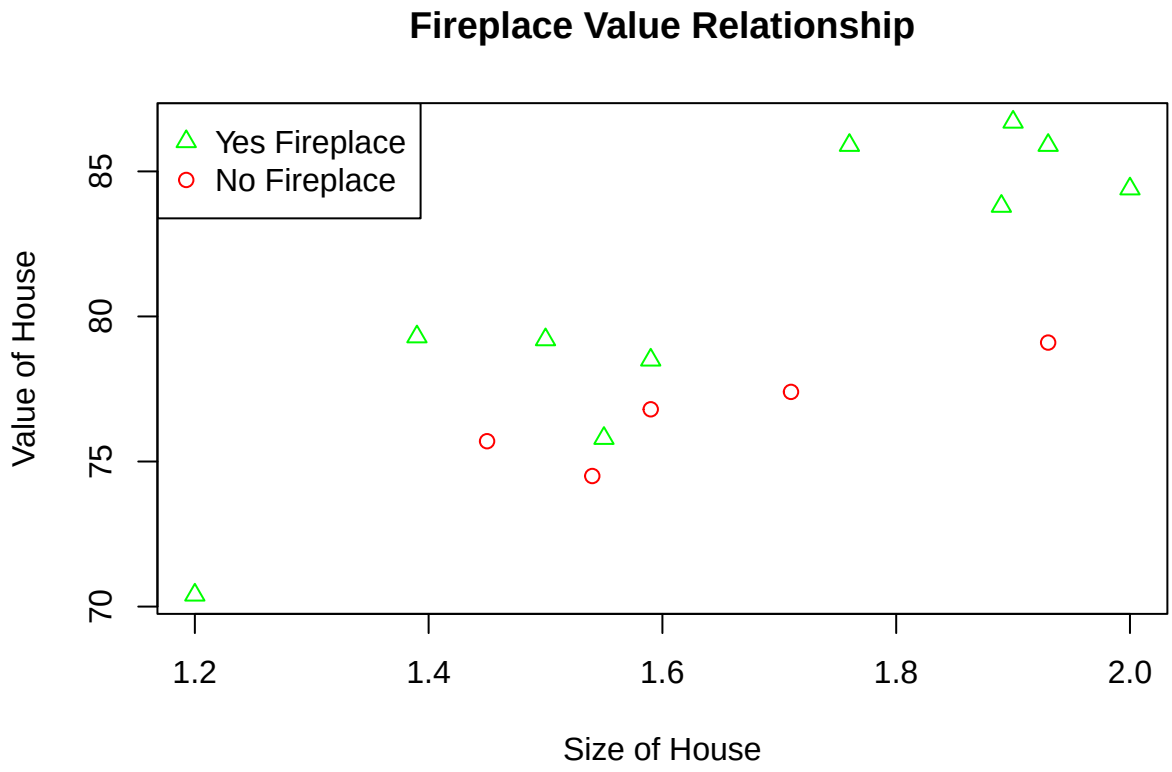
Below we plot the relationship between Size of House and Value of House. We set "Yes" fireplace observations to green triangles, and "No" fireplace variables to red circles.

```
#load data
library(readxl)
fireDat <- read_excel("Lab4q1Dat.xlsx")
head(fireDat)
```

```
## # A tibble: 6 x 3
##   Value  Size Firepl
##   <dbl> <dbl> <chr>
## 1  84.4  2    Yes
## 2  77.4  1.71 No
## 3  75.7  1.45 No
## 4  85.9  1.76 Yes
## 5  79.1  1.93 No
## 6  70.4  1.2  Yes
```

```
#ensure the Firepl variable will be treated as a factor in R
fireDat$Firepl = factor(fireDat$Firepl)

#the pch argument gives a symbol.  you can see
#http://www.endmemo.com/program/R/pic/pchsymbols.png
#for the different symbol options

#first make a symbols variable and color variable.
firesymbols = c()
firecolors = c()
#then for each observation, save the symbol option
for (i in 1:(dim(fireDat)[1])) {
#check if there is a fireplace, symbol 2 corresponds to triangles
  if (fireDat$Firepl[i] == "Yes") {
    firesymbols[i] = 2
    firecolors[i] = 'green'
  }
#check if there is no firpelace. symbol 1 corresponds to circle
```

```
  if (fireDat$Firepl[i] == "No") {
    firesymbols[i] = 1
    firecolors[i] = 'red'
  }
}

#plot everything
plot(fireDat$Size, fireDat$Value, pch = firesymbols, col = firecolors,
     xlab = "Size of House", ylab = "Value of House",
     main = "Fireplace Value Relationship")

#for the legend, you first argument is the placement of the legend.
#in general, topleft, bottomleft, bottomright, bottomleft work well
#but you can also give the exact xy coordinates.
legend("topleft", legend = c("Yes Fireplace", "No Fireplace"),
       pch = c(2, 1), col = c('green', 'red'))
```



B) Compute a 95% confidence interval for this difference *without controlling for the effect of size of home* using the `t.test` function while assuming equal variances. What is your 95% confidence interval for the difference in the selling price of homes with and without fireplaces.
Below we load the data and obtain the confidence interval.

```
#get house values for "No" fire place and "Yes" fireplace houses
noFire = subset(fireDat, Firepl == 'No')
yesFire = subset(fireDat, Firepl == "Yes")

#run t-test, with variances equal assumption
t.test(noFire$Value, yesFire$Value, var.equal = TRUE)
```

```
##
```

```
##  Two Sample t-test
##
## data:  noFire$Value and yesFire$Value
## t = -1.7388, df = 13, p-value = 0.1057
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -9.62022  1.04022
## sample estimates:
## mean of x mean of y
##     76.70     80.99
```

Therefore we are 95% confident that the average difference in selling price of a house with no Fireplace versus a house with a fireplace is between -9.62 and 1.04 thousand dollars.

C) Now do a simple linear regression using ONLY an indicator variable for FirePlace. Be sure to keep in mind the dummy coding of the categorical variable. What does the coefficient of Fireplace mean? (hint: the summary table of the linear regression can provide insight here) From this create a 95% confidence interval for the value of a fireplace WITHOUT controlling for hte effect of size of house.

First we fit the model

```
#fit the model, and summary
fireDat.lm1 = lm(Value~Firepl, data = fireDat)
summary(fireDat.lm1)
```

```
##
## Call:
## lm(formula = Value ~ Firepl, data = fireDat)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -10.590  -1.995   0.100   3.110   5.710
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   76.700      2.015  38.074 1.01e-14 ***
## FireplYes      4.290      2.467   1.739    0.106
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.505 on 13 degrees of freedom
## Multiple R-squared:  0.1887, Adjusted R-squared:  0.1263
## F-statistic: 3.023 on 1 and 13 DF,  p-value: 0.1057
```

The coefficient of `Firepl` corresponds to the estimated average change in the value of the house from going from no fireplace to including a fireplace.

Comment: By default, in R, the variable that comes first alphabetically in the categorical variable will be treated as the baseline variable. In our case, "No" comes first and "Yes" comes second. In the summary table we can see that the coefficient is named `FireplYes` – that is another way to know that this coefficient corresponds to the change in `Value` when `Firepl` goes from the baselines ("No" in this case) to "Yes".

A 95% confidence interval can be calculated as follows

```
tcrit = qt(0.975, 13)

lowerCI = 4.290 -2.467*tcrit
upperCI = 4.290 + 2.467*tcrit
```

```
c(lowerCI, upperCI)
```

```
## [1] -1.039629  9.619629
```

Therefore we are 95% confidence that a fireplace adds between -1.039 and 9.620 thousand dollars to the selling price of a house.

D) Include the Size variable in the data in the linear model, and compute a 95% confidence interval. Note the discreprancy in parts B and C to the confidence interval computere here. What causes this to happen? You should answer this according to two items.

i) look at the descriptive statistics for size of homes with and without fireaplces;
ii) look at the estimated variance for these two different procedures (really, two different models). Why does this difference exist? How does this difference impact your confidence intervals?

First we fit the linear model with `Size` included and calculate a 95% confidence interval.

```
fireDat.lm2 = lm(Value~Size + Firepl, data = fireDat)

#95% confidence interval, here we use the confint function instead of
#calculating it by hand. param = 3 is chosen because FireplYes is the
#3rd row of the summary table
confint(fireDat.lm2, parm = 3)
```

```
##              2.5 %   97.5 %
## FireplYes 1.148591 6.557374
```

Therefore, when keeping size constant, we are 95% confident that adding a fireplace to a house will increase the average price of the house by between 1.15 thousand and 6.56 thousand.

This confidence interval is about half the size of the one calculated in parts B and C.

For Part i: Using the `noFire` and `yesFire` variables defined in part A) we can look at sumary statistics within group.

```
summary(noFire)
```

```
##      Value           Size       Firepl
##  Min.   :74.5   Min.   :1.450   No :5
##  1st Qu.:75.7   1st Qu.:1.540   Yes:0
##  Median :76.8   Median :1.590
##  Mean   :76.7   Mean   :1.644
##  3rd Qu.:77.4   3rd Qu.:1.710
##  Max.   :79.1   Max.   :1.930
```

```
summary(yesFire)
```

```
##      Value            Size       Firepl
##  Min.   :70.40   Min.   :1.200   No : 0
##  1st Qu.:78.67   1st Qu.:1.512   Yes:10
##  Median :81.55   Median :1.675
##  Mean   :80.99   Mean   :1.671
##  3rd Qu.:85.53   3rd Qu.:1.897
##  Max.   :86.70   Max.   :2.000
```

We see that homes with Fireplaces are on average larger than homes without fireplaces by about 27 square feet.

For Part ii: The estimate variance ise the MSE from the fit. This can be calculated by squaring the residual standard error in the summary table for each liner model.

The MSE for the model without Size is $4.505^2 = 20.30$ versus the MSE for the model with Size included $2.263^2 = 5.12$. Therefore there is approximate 4 times reduction in variance. Therefore `Size` of home is explaining a lot of the variation in the sale price thereby reducing the amount of unexplained error.

E) In light of part D)i) above, explain again in a few sentences how multiple regression controls for the effect of one variable before evaluating the effects of another. Also, explain why adding significant controlling variables makes your estimates more precise in light of part D)ii) above.

Multiple regression will control for the differences in size by making the comparison for homes with and without fireplaces for homes of equal size, which the 2-sample t-procedure did not do, nor did the regression without size in the model. This makes the comparison more accurate. It also uses size as an explanatory variable, moving SSs from error into regression, reducing the estimated variance making our standard error smaller and hence our estimate of fireplace effects are more accurate (CIs narrower).

##Question 2: Interaction Example

A developer working in the Midwest and South is trying to predict selling price based on type of home (Single family (SF) or Townhoue (T)), the region built (South (S) or Midwest(M)), and the cost of the lot (which is pro-rated for the number of townhouses built on the lot). He randomly selects 167 homes from the 987 that he has built over the last 10 years, and adjusts the selling price for inflation. Data appears in `Lab4q2Dat.xlsx`. Unless otherwise specified, use $\alpha = 0.05$ or a 95% confidence interval.

A) Plot the data of Lot Cost against Selling Price. Include all relevant categories with a legend.

Here we plot the relationship between Lot Cost and Selling Price

```
houseDat = read_excel("Lab4q2Dat.xlsx")
head(houseDat)
```

```
## # A tibble: 6 x 4
##   Region Type  SellingPrice LotCost
##   <chr>  <chr>        <dbl>   <dbl>
## 1 M      SF          348744   53000
## 2 M      SF          274455   41000
## 3 M      SF          277720   44650
## 4 M      SF          307373   41292
## 5 M      SF          271105   45000
## 6 M      SF          262740   44900
```

```
#ensure R treats categorical variables as such
houseDat$Region = factor(houseDat$Region)
houseDat$Type = factor(houseDat$Type)
```

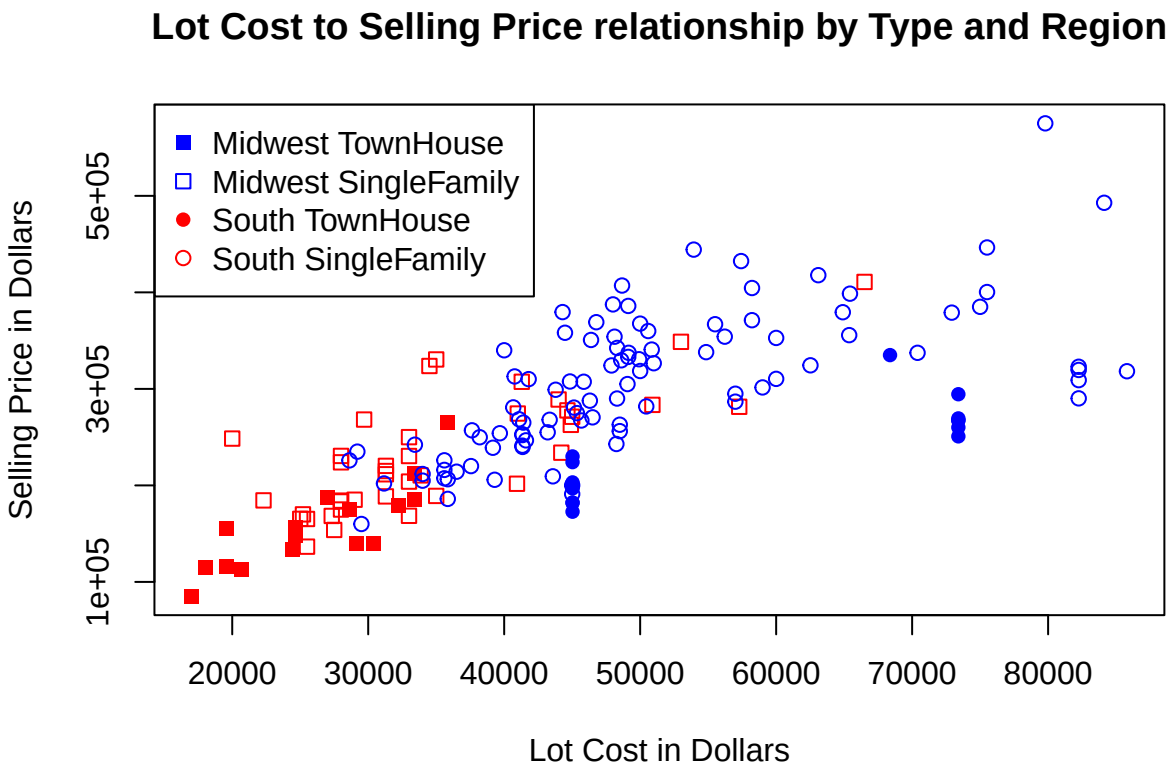
Note that the pch symbols for 0 and 15 are an outline and a filled in square, and the pch symbols for 1 and 16 are an outline and a filled in circle.

In our plot, MidWest will be squares, and South will be circles, where in both cases the the Single Family homes will be outlines and and Townhouses will be filed in.

Lastly, we'll color the Midwest red and the South blue.

```
#we want to make a vector that has the above properties.
#first save a symbols and colors variable
housesymbols = c()
housecolors = c()
#then run a loop.  For each observation, check which region and type it is
#and then select the correct symbol
for (i in 1:(dim(houseDat)[1])) {
  #we go through the 4 cases and choose colors and symbols for each one
  #Midwest Townhouse
```

```
  if ((houseDat$Region[i] == 'M')&(houseDat$Type[i] == 'T')) {
    housesymbols[i] = 15
    housecolors[i] = 'red'
  }

  #Southern Townhouse
  if ((houseDat$Region[i] == 'S')&(houseDat$Type[i] == "T")) {
    housesymbols[i] = 16
    housecolors[i] = 'blue'
  }

  #Midwest Single Family Home
  if ((houseDat$Region[i] == 'M')&(houseDat$Type[i] == "SF")) {
    housesymbols[i] = 0
    housecolors[i] = 'red'
  }

  #Southern Single Family Home
  if ((houseDat$Region[i] == 'S')&(houseDat$Type[i] == "SF")) {
    housesymbols[i] = 1
    housecolors[i] = 'blue'
  }
}



plot(houseDat$LotCost, houseDat$SellingPrice, pch = housesymbols, col = housecolors,
     xlab = "Lot Cost in Dollars", ylab = "Selling Price in Dollars",
     main = "Lot Cost to Selling Price relationship by Type and Region")

legend("topleft", legend = c("Midwest TownHouse", "Midwest SingleFamily",
                             "South TownHouse", "South SingleFamily"),
       pch = c(15, 0, 16, 1), col = c('blue', "blue", 'red', 'red'))
```



B) Run a multiple linear regrssion on selling price versus the three predictor variables Region, Type, and Lot Cost. For Region and Type, what are the baseline categories?

```
#fit model
houseDat.lm = lm(SellingPrice ~ Region + Type + LotCost, data = houseDat)
summary(houseDat.lm)
```

```
##
## Call:
## lm(formula = SellingPrice ~ Region + Type + LotCost, data = houseDat)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -124242  -33387   -5813   30390  169311
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.170e+05  1.206e+04   9.703  < 2e-16 ***
## RegionS      9.082e+03  9.891e+03   0.918     0.36
## TypeT       -7.257e+04  9.755e+03  -7.439  5.4e-12 ***
## LotCost      3.505e+00  2.982e-01  11.756  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 48850 on 164 degrees of freedom
## Multiple R-squared:  0.653,  Adjusted R-squared:  0.6467
## F-statistic: 102.9 on 3 and 164 DF,  p-value: < 2.2e-16
```

For Region, Midwest is the baseline category. For Type, Single Family is the baseline category.

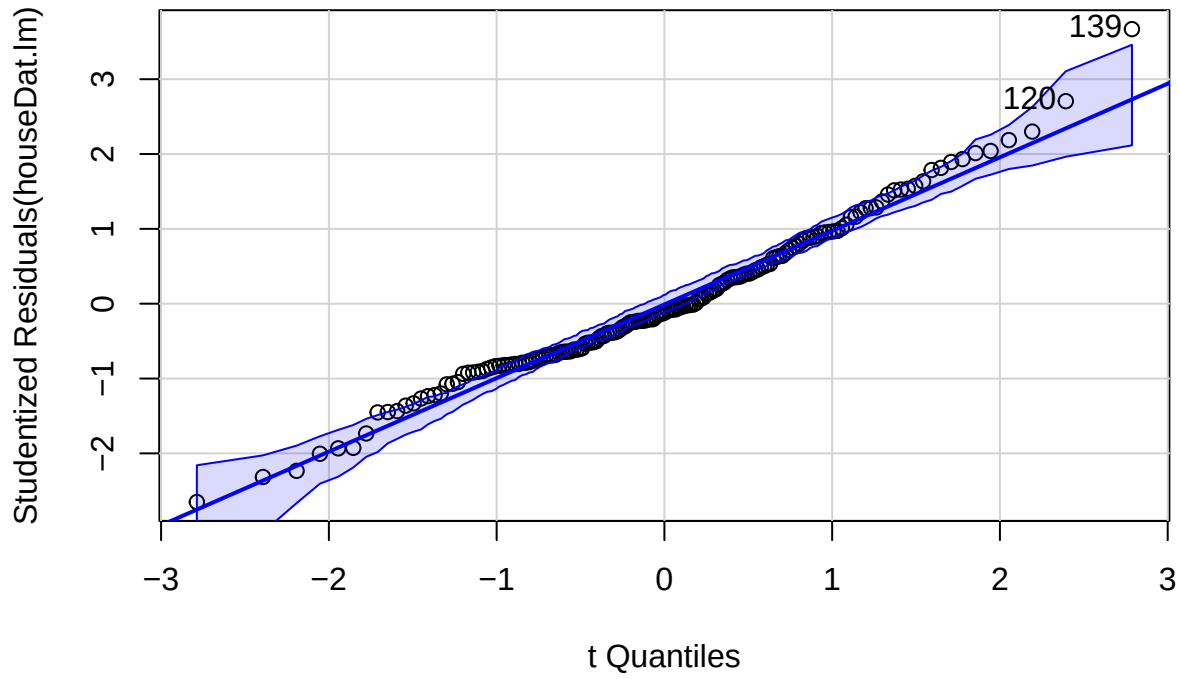
C) Look at diagnostic plots. What must be done before you proceed?

Below we look at qqplots, residual plolts

```
library(car)
```

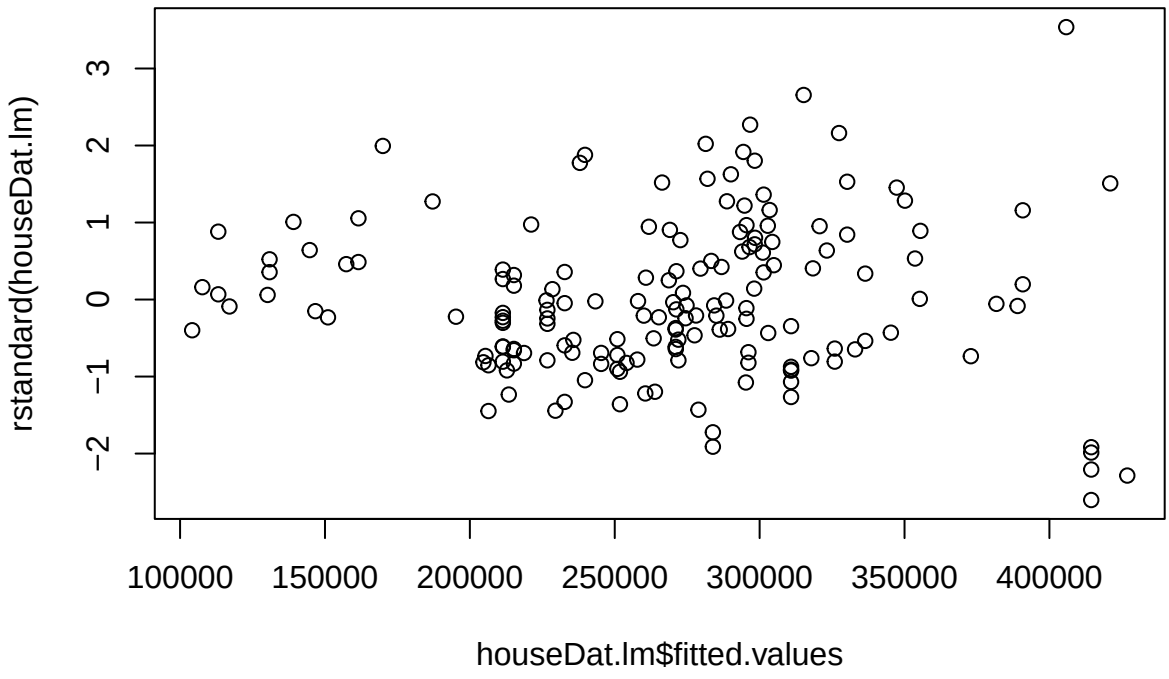
```
## Loading required package: carData
```

```
qqPlot(houseDat.lm)
```




```
## [1] 120 139
```

```
plot(houseDat.lm$fitted.values, rstandard(houseDat.lm))
```



We can see in the residual plot that we have increasing variance as predicted price increases, which is natural.

We us a natural log transformation of selling price to stabilize variance.

```
#Doing log transformatoin
houseDat$logSell = log(houseDat$SellingPrice)
```

Comment: The symbols and color variables can be still used if we want to color code the residual plot.

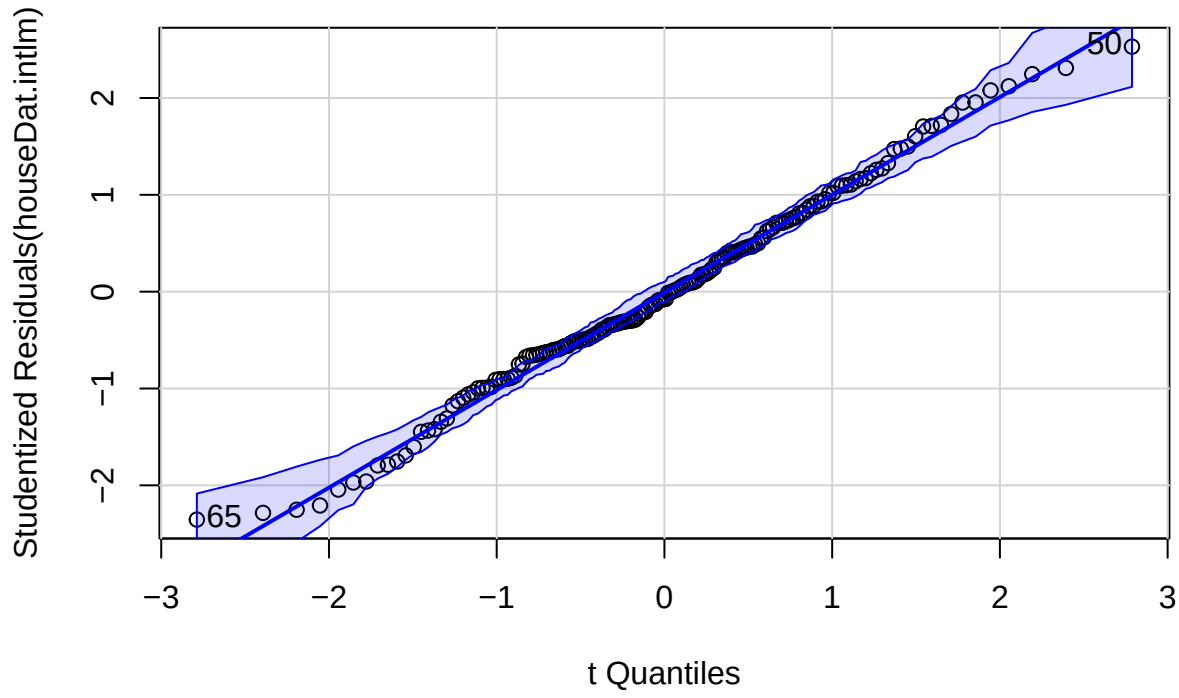
D) Fit a full interaction model with all first-order pairwise interactions. Check diagnostic plots including Cook's distance plot and determine if asumptions are met for the inference on this model. The builder would really like to simplify his model. Using this one multiple regression, test if all the interactions can be simultaneously dropped from the model. State hypotheses, test statistic, p-value, and conclusions.

We can fit the full model with

```
houseDat.intlm = lm(logSell ~ Region + Type + LotCost +
                      Region:Type + Region:LotCost + Type:LotCost, data = houseDat)
summary(houseDat.intlm)
```

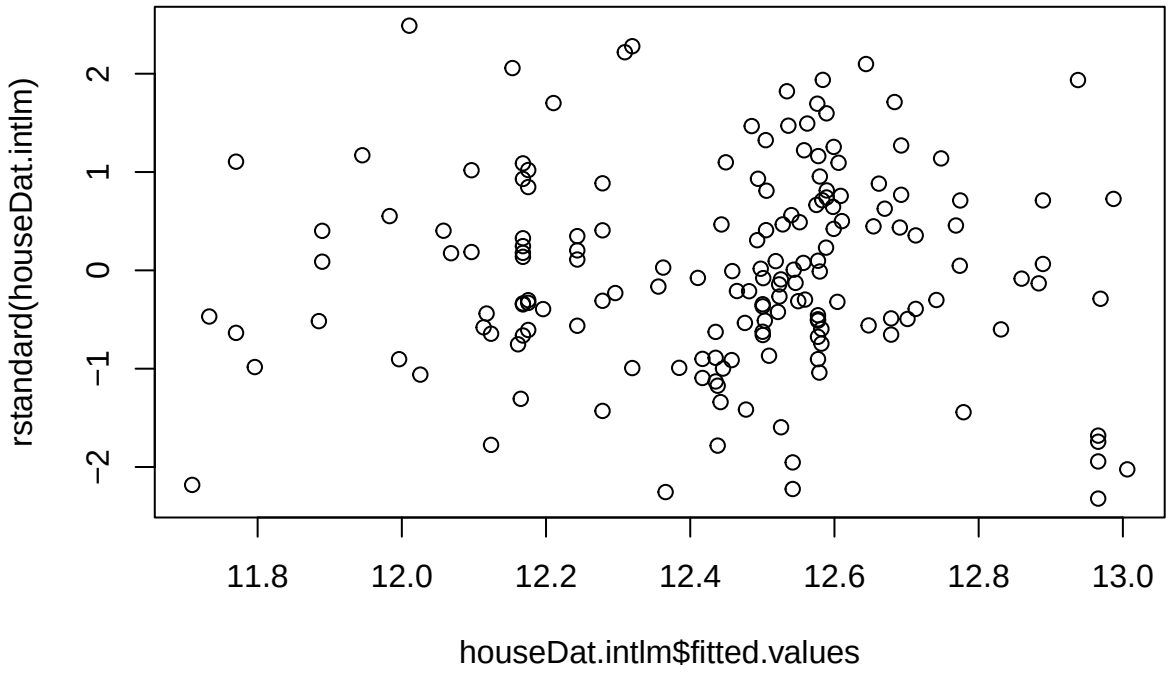
```
##
## Call:
## lm(formula = logSell ~ Region + Type + LotCost + Region:Type +
##     Region:LotCost + Type:LotCost, data = houseDat)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.3875 -0.1033 -0.0133  0.1203  0.4128
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)      1.160e+01  9.594e-02 120.894  < 2e-16 ***
## RegionS          4.323e-01  1.142e-01   3.787 0.000215 ***
## TypeT           -2.918e-01  1.019e-01  -2.864 0.004741 **
## LotCost          2.062e-05  2.622e-06   7.862 5.13e-13 ***
## RegionS:TypeT   -2.200e-01  1.162e-01  -1.892 0.060262 .
## RegionS:LotCost -9.243e-06  2.858e-06  -3.234 0.001479 **
## TypeT:LotCost    3.050e-06  3.119e-06   0.978 0.329605
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1728 on 161 degrees of freedom
## Multiple R-squared:  0.7203, Adjusted R-squared:  0.7099
## F-statistic:  69.1 on 6 and 161 DF,  p-value: < 2.2e-16
```

```
#diagnostic plots and cook's distnace
qqPlot(houseDat.intlm)
```

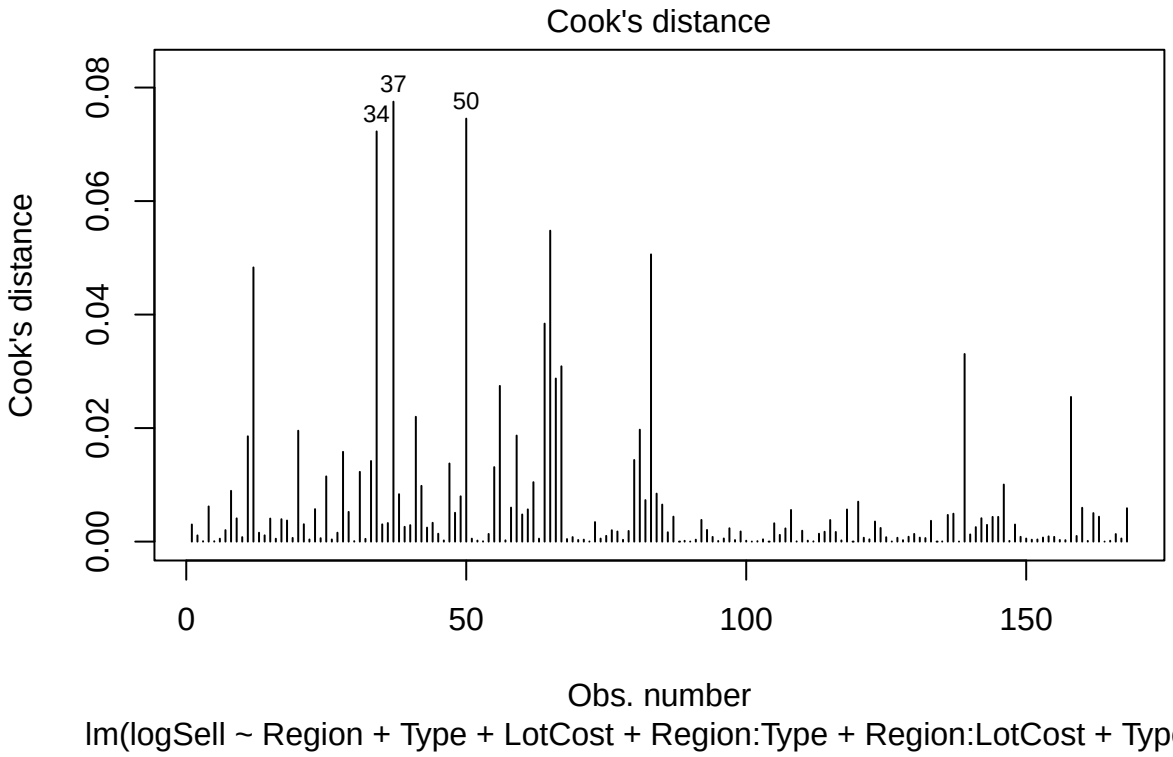



```
## [1] 50 65
```

```
plot(houseDat.intlm$fitted.values, rstandard(houseDat.intlm))
```




```
plot(houseDat.intlm, which = 4)
```



The assumptions for inference appear to be satisfied by this model.

To test if the interactions are significant, we test $H_0: $interaction coefficients are all equal to 0, vs the alternative $H_a: $ at least one interaction coefficient is not equal to 0.

We can get the change in the Residual Sum of Squares by adding together the Sequential Sum of Squares term from the anova output.

```
anova(houseDat.intlm)
```

```
## Analysis of Variance Table
##
## Response: logSell
##                 Df Sum Sq Mean Sq  F value Pr(>F)
## Region           1 4.7849  4.7849 160.3353 <2e-16 ***
## Type             1 2.5777  2.5777  86.3744 <2e-16 ***
## LotCost          1 4.6397  4.6397 155.4698 <2e-16 ***
## Region:Type      1 0.0337  0.0337   1.1297 0.2894
## Region:LotCost   1 0.3076  0.3076  10.3075 0.0016 **
## Type:LotCost     1 0.0285  0.0285   0.9562 0.3296
## Residuals      161 4.8048  0.0298
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

From this we get a test statistic as $F_{test} = \frac{0.0285+0.3076+0.0337)/3}{0.1728^2} = 4.138$. The pvalue can then be calculated as

```
1-pf(4.138, 3, 161)
```

```
## [1] 0.00739866
```

Therefore with a pvalue of 0.007 we canr reject the null hypothesis and conclude that not all interaction terms can be dropped.

E) Look at the significance of the interaction terms from your six-predictor model in Part C. Test the set of those which are not significant by themselves at $\alpha = 0.05$ with a simultaneous test by re-running the regression without these non-significant interaction terms and getting the test statistic by subtracting the SSRs from the two models.

The two non significant interactions were `Region:Type` and `Type:LotCost`. Below we fit a model without these and output the Sum of Squares Regression. Sum of Squares regression can be calculated by summing all values in the `Sum Sq` column in the anova table except for the sum of squares residual.

```
houseDat.intlmR = lm(logSell ~ Region + Type + LotCost +
                       Region:LotCost, data = houseDat)

#getting sum of squares residual from full model and reduced models
#note the -7 and -5 tell R not to include the residual sum of square
#when summing to obtain residual sum of squares.
SSRF = sum(anova(houseDat.intlm)[-7,2])
SSRR = sum(anova(houseDat.intlmR)[-5,2])
MSE = anova(houseDat.intlm)[7, 3]

#calculating F stat
Fstat = (SSRF-SSRR)/2/(anova(houseDat.intlm)[7,3])
#calculating p value
1-pf(Fstat, 2, 161)
```

```
## [1] 0.1272158
```

At the 0.05 significance level we do not have evidence that these interaction terms have an effect on the mean.

F) Based on your results in parts D and E, determine an appropriate model (dropping sets of non-significant predictors) and use it to predict the selling price of a single family home in the south on a lot which cost $42,500.

I choose the model above, with the three predictors and Region-Lot Cost interaction. The plots remain fine and we get the following output for a prediction interval.

```
newdata = data.frame(Region = "S", Type = "SF", LotCost = 42500)
prediction = predict(houseDat.intlm, newdata, interval = "prediction")
exp(prediction)
```

```
##        fit      lwr      upr
## 1 272032.5 192955.1 383517.6
```

We took the exponent in order to reverse the natural log transformation. This gives a 95% prediction interval for the sale of the given house to be between $192955 and $383516.

G) Based on the model you chose in E, does the region of the country have a fixed effect on selling price or does it depend on the other two variables? Explain in two sentences or less.

No, since the region of the country interacts with lot cost, the effect on selling price depends on the lot cost, and is not fixed.

H) Based on the model you chose in E, does th type of house have a fixed effect on selling price or does it depend on the other two variables? Explain in two sentences of less.

Yes, since the interactions of type of house are not significant, type of house has a fixed effect which does not depend on lot cost or region of the country.

I) Is the increase in selling price per dollar increase in lot cost greater in the Midwest than in the South? State hypotheses, test statistic, p-value, and conclusion.

Here we wish to test the interaction term for Region and lot cost. Since Midwest was baseline, we wish to test if the interaction term is negative (that is, does effect of lot cost derease as we go from Midwest to South). Hence, we test $H_0 : \beta_4 = 0$ against $H_a : \beta_4 < 0$.

Since we're testing against 0, the test statistic can be directly obtained from `summary(houseDat.intlmR)` as -2.847. This gives a pvalue as

```
pt(-2.847, 163)
```

```
## [1] 0.002490679
```

Therefore we reject H_0 and conclude that in fact selling price increases faster per dollar increase in lot cost in the Midwest than in the South.

J) What propotion of the variance in selling price (untransformed!) does the model you chose in E explain?

The untransformed fitted values are obtained as `exp(houseDat.intlmR$fitted.values)`. Therefore the total sum of squares and error sum of squares can be calculated as follows

```
fitval = exp(houseDat.intlmR$fitted.values)

SSE = sum((houseDat$SellingPrice-fitval)^2)
SST = sum((houseDat$SellingPrice - mean(houseDat$SellingPrice))^2)

1-SSE/SST
```

```
## [1] 0.6432374
```

That is, our model explains 64.3% of the variance in selling price measured in dollars (measured in ln(dollars) we got an r^2 of 71.3%)