Case Project

Jader Martins

01/05/2021

(a) Consider a linear model where the sale price of a house is the dependent variable and the explanatory variables are the other variables given above. Perform a test for linearity. What do you conclude based on the test result?

```
houses.lm <- lm("sell ~ lot + bdms + fb + sty + drv + rec + ffin + ghw + ca + gar + reg", data)
summary(houses.lm)
```

```
##
## Call:
## lm(formula = "sell ~ lot + bdms + fb + sty + drv + rec + ffin + ghw + ca + gar + reg",
##     data = data)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -41389  -9307   -591   7353  74875
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -4038.3504  3409.4713  -1.184 0.236762
## lot             3.5463     0.3503  10.124  < 2e-16 ***
## bdms         1832.0035  1047.0002   1.750 0.080733 .
## fb          14335.5585  1489.9209   9.622  < 2e-16 ***
## sty          6556.9457   925.2899   7.086 4.37e-12 ***
## drv          6687.7789  2045.2458   3.270 0.001145 **
## rec          4511.2838  1899.9577   2.374 0.017929 *
## ffin         5452.3855  1588.0239   3.433 0.000642 ***
## ghw         12831.4063  3217.5971   3.988 7.60e-05 ***
## ca          12632.8904  1555.0211   8.124 3.15e-15 ***
## gar          4244.8290   840.5442   5.050 6.07e-07 ***
## reg          9369.5132  1669.0907   5.614 3.19e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 15420 on 534 degrees of freedom
## Multiple R-squared:  0.6731, Adjusted R-squared:  0.6664
## F-statistic: 99.97 on 11 and 534 DF,  p-value: < 2.2e-16
```

We can check in the last column that most explanatory variables have high significance (***) and a p-value below 2.2e-16 so we reject the null hypothesis.

```
houses.res <- residuals.lm(houses.lm)
hist(houses.res, xlab = "Residuals", main = "")
```

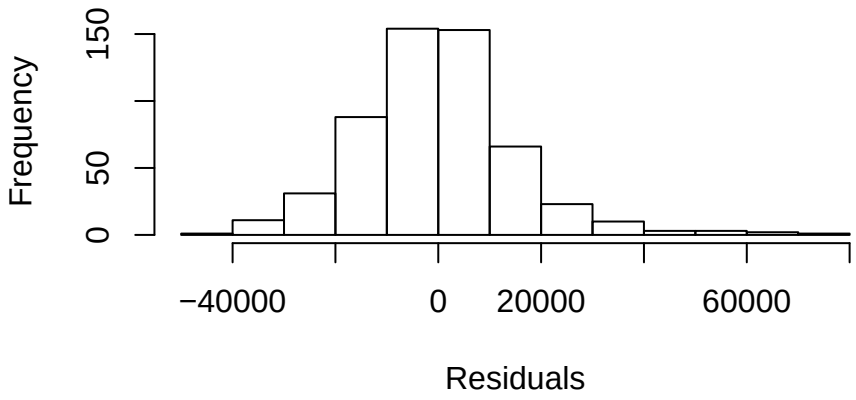

Figure 1: Histogram of regression residuals.

In the Figure 1 we see that the residuals are approximately normally distributed with a heavy tail on the right side. So our estimation is not highly biased in any point of the price scale.

```
par(mfrow = c(2, 2))
plot(houses.lm)
```

In the Figure 2 "Residual vs Fitted" plot we check an almost linear relation with no distinctive pattern, which is a good indication that our model capture the variation well. In the "Normal Q-Q" plot we check as above that out model is not biased in most of points and a little biased in extreme points of price. The "Scale-Location" plot is used to check if we have no heteroscedasticity so with an horizontal line and assuming that our points are ordered in time (no guarantee), we have a homoscedasticity indication. Finally, the "Residuals vs Leverage" indicates if extreme values in our regression are causing problems to the estimated line and with an horizontal line we have no problem with those points.

Concluding, as described above a linear regression is a good estimation for our dependent variable given out explanatory variables.

(b) Now consider a linear model where the log of the sale price of the house is the dependent variable and the explanatory variables are as before. Perform again the test for linearity. What do you conclude now?

```
houses.lm <- lm("log(sell) ~ lot + bdms + fb + sty + drv + rec + ffin + ghw + ca + gar + reg", data)
summary(houses.lm)
```

```
##
## Call:
## lm(formula = "log(sell) ~ lot + bdms + fb + sty + drv + rec + ffin + ghw + ca + gar + reg",
##     data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.67865 -0.12211  0.01666  0.12868  0.67737
```

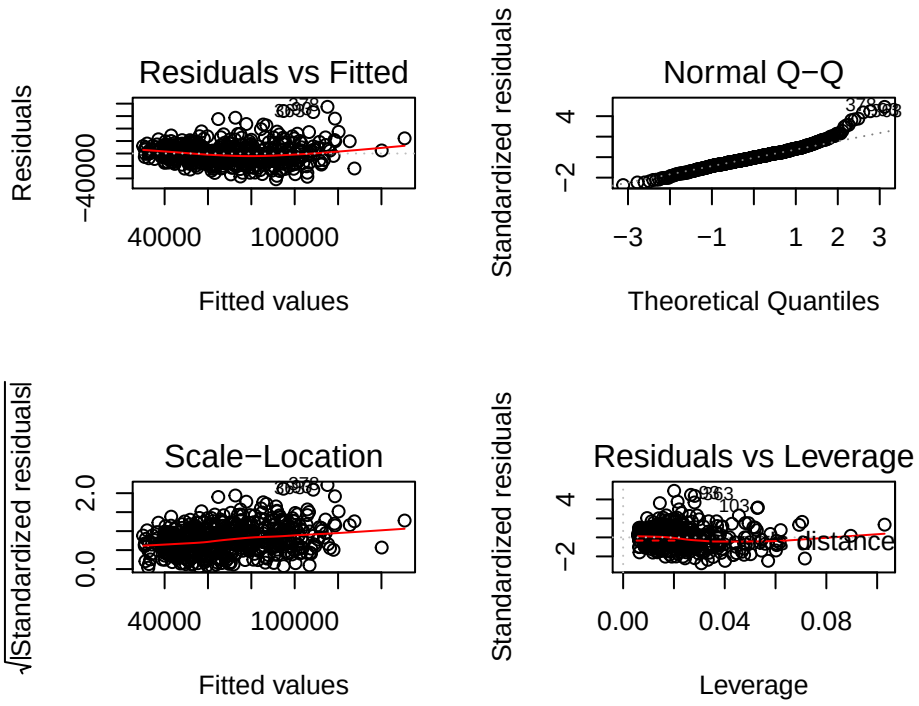

Figure 2: Regression diagnosis plot.

```
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 1.003e+01  4.724e-02 212.210  < 2e-16 ***
## lot         5.057e-05  4.854e-06  10.418  < 2e-16 ***
## bdms        3.402e-02  1.451e-02   2.345  0.01939 *
## fb          1.678e-01  2.065e-02   8.126 3.10e-15 ***
## sty         9.227e-02  1.282e-02   7.197 2.10e-12 ***
## drv         1.307e-01  2.834e-02   4.610 5.04e-06 ***
## rec         7.352e-02  2.633e-02   2.792  0.00542 **
## ffin        9.940e-02  2.200e-02   4.517 7.72e-06 ***
## ghw         1.784e-01  4.458e-02   4.000 7.22e-05 ***
## ca          1.780e-01  2.155e-02   8.262 1.14e-15 ***
## gar         5.076e-02  1.165e-02   4.358 1.58e-05 ***
## reg         1.271e-01  2.313e-02   5.496 6.02e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2137 on 534 degrees of freedom
## Multiple R-squared:  0.6766, Adjusted R-squared:  0.6699
## F-statistic: 101.6 on 11 and 534 DF,  p-value: < 2.2e-16
```

We've got a better adderence to the data with the log transformation.

```
houses.res <- residuals.lm(houses.lm)
hist(houses.res, xlab = "Residuals", main = "")
```

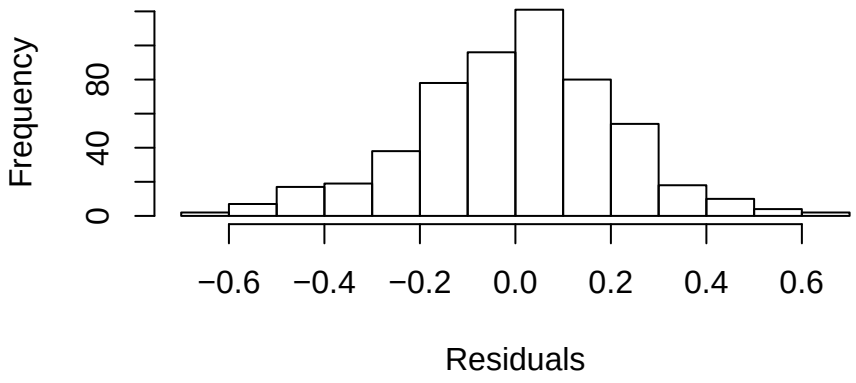

Figure 3: Histogram of regression residuals.

In the Figure 3 we can see that the heavy tail disappeared in the normal distribution.

```
par(mfrow = c(2, 2))
plot(houses.lm)
```

The problems from a biased estimation in extreme values also disappeared with the log transform, as shown in Figure 4, so we can conclude that this transformation helps to estimate better our data.

(c) Continue with the linear model from question (b). Estimate a model that includes both the lot size variable and its logarithm, as well as all other explanatory variables without transformation. What is your conclusion, should we include lot size itself or its logarithm?

```
houses.lm <- lm("log(sell) ~ log(lot) + bdms + fb + sty + drv + rec + ffin + ghw + ca + gar + reg", data
summary(houses.lm)
```

```
##
## Call:
## lm(formula = "log(sell) ~ log(lot) + bdms + fb + sty + drv + rec + ffin + ghw + ca + gar + reg",
##     data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.68355 -0.12247  0.00802  0.12780  0.67564
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  7.74509    0.21634  35.801  < 2e-16 ***
## log(lot)     0.30313    0.02669  11.356  < 2e-16 ***
## bdms         0.03440    0.01427   2.410 0.016294 *
## fb           0.16576    0.02033   8.154 2.52e-15 ***
```

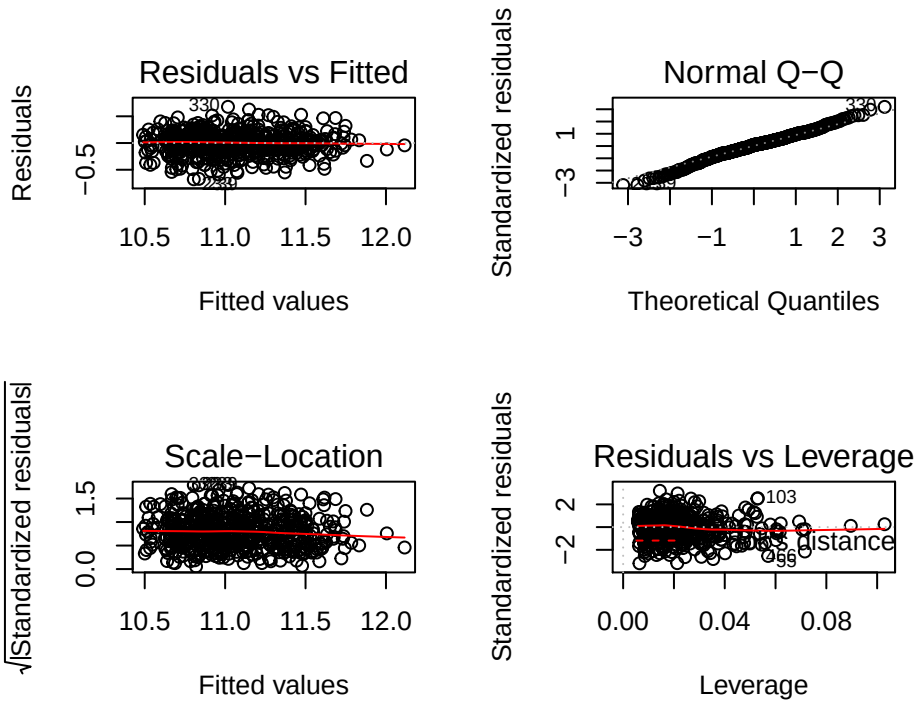

Figure 4: Regression diagnosis plot.

```
## sty          0.09169    0.01261   7.268 1.30e-12 ***
## drv          0.11020    0.02823   3.904 0.000107 ***
## rec          0.05797    0.02605   2.225 0.026482 *
## ffin         0.10449    0.02169   4.817 1.90e-06 ***
## ghw          0.17902    0.04389   4.079 5.22e-05 ***
## ca           0.16642    0.02134   7.799 3.29e-14 ***
## gar          0.04795    0.01148   4.178 3.43e-05 ***
## reg          0.13185    0.02267   5.816 1.04e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2104 on 534 degrees of freedom
## Multiple R-squared:  0.6865, Adjusted R-squared:  0.6801
## F-statistic: 106.3 on 11 and 534 DF,  p-value: < 2.2e-16
```

Include both variables, lot and log.lot, does not add significance to the model and as a result we got a worse F-test, so we conclude that is better to add only log.lot to the model.

(d) Consider now a model where the log of the sale price of the house is the dependent variable and the explanatory variables are the log transformation of lot size, with all other explanatory variables as before. We now consider interaction effects of the log lot size with the other variables. Construct these interaction variables. How many are individually significant?

```
houses.lm <- lm("log(lot) ~ log(sell) + bdms + fb + sty + drv + rec + ffin + ghw + ca + gar + reg", data
summary(houses.lm)
```

```
##
## Call:
## lm(formula = "log(lot) ~ log(sell) + bdms + fb + sty + drv + rec + ffin + ghw + ca + gar + reg",
##     data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.76460 -0.21127 -0.01328  0.18551  1.06217
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.414830   0.577180   2.451 0.014554 *
## log(sell)    0.641711   0.056509  11.356  < 2e-16 ***
## bdms         0.009402   0.020877   0.450 0.652651
## fb          -0.043323   0.031309  -1.384 0.167022
## sty         -0.083426   0.018898  -4.414 1.23e-05 ***
## drv          0.146762   0.041163   3.565 0.000396 ***
## rec          0.067969   0.037968   1.790 0.073993 .
## ffin        -0.128082   0.031759  -4.033 6.31e-05 ***
## ghw         -0.125899   0.064622  -1.948 0.051909 .
## ca           0.002993   0.032768   0.091 0.927265
## gar          0.062895   0.016749   3.755 0.000192 ***
## reg          0.007322   0.034011   0.215 0.829619
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3061 on 534 degrees of freedom
## Multiple R-squared:  0.4201, Adjusted R-squared:  0.4082
## F-statistic: 35.17 on 11 and 534 DF,  p-value: < 2.2e-16
```

The variables log.sell (transformed sale price), sty (number of stories excluding basement), drv (driveway in house), ffin (full finished basement), gar (number of covered garage places), are highly correlated with the log.lot which seems to be intuitively right, as they are related with the terrain size.

(e) Perform an F-test for the joint significance of the interaction effects from question (d).

The F-test gave us 35.17 which is an small value, so many variables are uncorrelated with log.lot.

(f) Now perform model specification on the interaction variables using the general-to-specific approach. (Only eliminate the interaction effects.)

```
houses.lm <- lm("log(lot) ~ log(sell) + sty + drv + ffin + gar", data)
summary(houses.lm)
```

```
##
## Call:
## lm(formula = "log(lot) ~ log(sell) + sty + drv + ffin + gar",
##     data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.89016 -0.21014 -0.00165  0.19588  1.06617
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.48596    0.48096   3.090 0.002108 **
## log(sell)    0.63237    0.04614  13.705  < 2e-16 ***
## sty         -0.08378    0.01765  -4.746 2.67e-06 ***
## drv          0.15797    0.04021   3.928 9.67e-05 ***
## ffin        -0.10824    0.02968  -3.647 0.000291 ***
## gar          0.05918    0.01670   3.543 0.000429 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3072 on 540 degrees of freedom
## Multiple R-squared:  0.4095, Adjusted R-squared:  0.404
## F-statistic: 74.89 on 5 and 540 DF,  p-value: < 2.2e-16
```

The value of F-test doubled so we have a much better model. But a lot of variability isn't explained yet with only those variables.

(g) One may argue that some of the explanatory variables are endogenous and that there may be omitted variables. For example, the 'condition' of the house in terms of how it is maintained is not a variable (and difficult to measure) but will affect the house price. It will also affect, or be reflected in, some of the other variables, such as whether the house has an air conditioning (which is mostly in newer houses). If the condition of the house is missing, will the effect of air conditioning on the (log of the) sale price be over- or underestimated? (For this question no computer calculations are required.)

Without this proxy the prices will be underestimated as newer houses will not be accounted and they are, generally, more expensive.

(f) Finally we analyze the predictive ability of the model. Consider again the model where the log of the sale price of the house is the dependent variable and the explanatory variables are the log transformation of lot size, with all other explanatory variables in their original form (and no interaction effects). Estimate the parameters of the model using the first 400 observations. Make predictions on the log of the price and calculate the MAE for the other 146 observations. How good is the predictive power of the model (relative to the variability in the log of the price)?

```
houses.lm <- lm("log(sell) ~ lot + bdms + fb + sty + drv + rec + ffin + ghw + ca + gar + reg", data[1:4
houses.yhat <- predict(houses.lm, data[400:nrow(data),])
error <- houses.yhat - log(data[400:nrow(data),]$sell)
mae <- mean(abs(error))
mae
```

```
## [1] 0.1373367
```

With a MAE of 0.1373367 we have a good estimation of house prices as it only misses exp(MAE) $\sim= 1$, with a standard deviation of prices given by 22532.76.