DATA 557: HW Assignment 6

Hriday Baghar

March 3, 2022

Data: "Sales_sample.csv" (same one as used in HW 5).

```
data <- read.csv("../HW5/Sales_sample.csv")
str(data)
```

```
## 'data.frame':	1000 obs. of  5 variables:
##  $ BEDS           : int  4 4 4 3 6 4 3 5 5 3 ...
##  $ BATHS          : num  2.5 2 2.25 2 2.5 1.75 2.75 3.25 2.5 2 ...
##  $ LOT_SIZE       : int  22578 4000 5000 6400 7431 7200 5500 12345 4000 7000 ...
##  $ LAST_SALE_PRICE: int  678000 888000 682000 1600000 750000 682000 896000 425000 911000 425000 ...
##  $ SQFT           : int  2410 2660 2800 3790 2940 2240 3230 4550 3800 1820 ...
```

```
summary(data)
```

```
##       BEDS           BATHS          LOT_SIZE     LAST_SALE_PRICE
##  Min.   :1.000   Min.   :0.75   Min.   :  653   Min.   :  87050
##  1st Qu.:3.000   1st Qu.:1.75   1st Qu.: 4000   1st Qu.: 475000
##  Median :3.000   Median :2.00   Median : 5502   Median : 632134
##  Mean   :3.388   Mean   :2.12   Mean   : 6635   Mean   : 735809
##  3rd Qu.:4.000   3rd Qu.:2.75   3rd Qu.: 7634   3rd Qu.: 859250
##  Max.   :6.000   Max.   :6.00   Max.   :80791   Max.   :4325000
##       SQFT
##  Min.   : 510
##  1st Qu.:1640
##  Median :2185
##  Mean   :2285
##  3rd Qu.:2760
##  Max.   :8820
```

1. Fit the linear regression model with sale price as response variable and SQFT, LOT_SIZE, BEDS, and BATHS as predictor variables (Model 1 from HW 5). Calculate robust standard errors for the coefficient estimates. Display a table with estimated coefficients, the usual standard errors that assume constant variance, and robust standard errors.

```
#Creating model same as in HW5
model.1 <- lm(LAST_SALE_PRICE ~ .,  data = data)
summary(model.1)
```

```
##
## Call:
## lm(formula = LAST_SALE_PRICE ~ ., data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1364578  -166436    -9884   122468  2964364
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)   5982.604  40023.271   0.149 0.881207
## BEDS        -60884.742  14461.536  -4.210 2.78e-05 ***
```

```
## BATHS       178177.446  17107.532  10.415  < 2e-16 ***
## LOT_SIZE         6.844      1.858   3.684 0.000242 ***
## SQFT           224.502     14.794  15.175  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 322100 on 995 degrees of freedom
## Multiple R-squared:  0.4691, Adjusted R-squared:  0.467
## F-statistic: 219.8 on 4 and 995 DF,  p-value: < 2.2e-16
```

```
#Calculating Robust SEs
library(sandwich)

robust.se <- sqrt(diag(sandwich::vcovHC(model.1)))
(ests.table <- data.frame(cbind(summary(model.1)$coefficients[,c("Estimate","Std. Error")], robust.se)))
```

```
##                 Estimate   Std..Error    robust.se
## (Intercept)   5982.604259 40023.271418 49655.792470
## BEDS        -60884.742104 14461.536156 17255.919552
## BATHS       178177.446061 17107.531726 22796.269233
## LOT_SIZE         6.844143     1.857731     7.734398
## SQFT           224.502066    14.793972    24.394722
```

2. Which set of standard errors should be used? Explain by referring to HW 5.

Since the constant variance assumption is violated in Model 1 we should use the robust standard errors.

3. Perform the Wald test for testing that the coefficient of the LOT_SIZE variable is equal to 0. Use the usual standard errors that assume constant variance. Report the test statistic and p-value.

```
reduced.model <- lm(LAST_SALE_PRICE ~ . -LOT_SIZE, data = data)
anova(model.1, reduced.model)
```

```
## Analysis of Variance Table
##
## Model 1: LAST_SALE_PRICE ~ BEDS + BATHS + LOT_SIZE + SQFT
## Model 2: LAST_SALE_PRICE ~ (BEDS + BATHS + LOT_SIZE + SQFT) - LOT_SIZE
##   Res.Df        RSS Df   Sum of Sq      F    Pr(>F)
## 1    995 1.0320e+14
## 2    996 1.0461e+14 -1 -1.4078e+12 13.573 0.0002418 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

From the above table we can see that the F-test statistic for LOT_SIZE = 13.573 and the p-value = 0.000242. Based on this we reject the null hypothesis that there is no linear relation between lot size and sale price.

Note: This test is equivalent to conducting a t-test (with n-p d.o.f) on estimate/SE for the parameter LOT_SIZE in model 1, as we can see from the p-value

4. Perform the robust Wald test statistic for testing that the coefficient of the LOT_SIZE variable is equal to 0. Report the test statistic and p-value.

```
library(lmtest)
```

```
## Loading required package: zoo
```

```
##
## Attaching package: 'zoo'
```

```
## The following objects are masked from 'package:base':
##
##     as.Date, as.Date.numeric
```

```
waldtest(model.1, reduced.model, test = "Chisq", vcov = vcovHC)
```

```
## Wald test
##
## Model 1: LAST_SALE_PRICE ~ BEDS + BATHS + LOT_SIZE + SQFT
## Model 2: LAST_SALE_PRICE ~ (BEDS + BATHS + LOT_SIZE + SQFT) - LOT_SIZE
##   Res.Df Df Chisq Pr(>Chisq)
## 1    995
## 2    996 -1 0.783     0.3762
```

```
#Checking if t-test gives same p-value

2 * (1 - pt(ests.table["LOT_SIZE", "Estimate"]/ ests.table["LOT_SIZE", "robust.se"], df = nrow(data) - nrow(ests.t
```

```
## [1] 0.3764262
```

```
# Q: Close enough but not exact?
# A: This is because we pass Chisq and not F to waldtest()
```

Test statistic = 0.783
P-value = 0.3762

We fail to reject the null hypothesis based on the Robust Wald Test.

5. Use the jackknife to estimate the SE for the coefficient of the LOT_SIZE variable. Report the jackknife estimate of the SE.

```
n <- nrow(data)
fit.jack.model <- function(i){
  lmi <- lm(LAST_SALE_PRICE ~ ., data = data, subset = -i)
  return(lmi$coef[4])
}
beta.jack <- sapply(1:n, fit.jack.model)

(se.jack <- (n - 1)*sd(beta.jack)/sqrt(n))
```

```
## [1] 7.730455
```

6. Use the jackknife estimate of the SE to test the null hypothesis that the coefficient of the LOT_SIZE variable is equal to 0. Report the test statistic and p-value.

```
#Which estimate should be used in numerator?
2 * (1 - pt(mean(beta.jack)/se.jack, df = nrow(data) - nrow(ests.table)))
```

```
## [1] 0.376258
```

7. Do the tests in Q3, Q4, and Q6 agree? Which of these tests are valid?

The tests in Q4 and Q6 agree and are the valid tests out of the three tests performed. This is because the constant variance assumption is violated in Model 1.

8. Remove the LOT_SIZE variable from Model 1 (call this Model 1A). Fit Model 1A and report the table of coefficients, the usual standard errors that assume constant variance, and robust standard errors.

```r
model.1a <- reduced.model
summary(model.1a)
```

```
##
## Call:
## lm(formula = LAST_SALE_PRICE ~ . - LOT_SIZE, data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1381156  -162981   -16906   119043  2960229
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  29034.46   39779.87   0.730    0.466
## BEDS        -59374.56   14546.68  -4.082 4.83e-05 ***
## BATHS       176027.85   17205.16  10.231  < 2e-16 ***
## SQFT           234.04      14.66  15.968  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 324100 on 996 degrees of freedom
## Multiple R-squared:  0.4619, Adjusted R-squared:  0.4602
## F-statistic: 284.9 on 3 and 996 DF,  p-value: < 2.2e-16
```

```r
robust.se.1a <- sqrt(diag(sandwich::vcovHC(model.1a)))
(ests.table.1a <- data.frame(cbind(summary(model.1a)$coefficients[,c("Estimate","Std. Error")], robust.se.1a)))
```

```
##                Estimate  Std..Error robust.se.1a
## (Intercept)  29034.4577 39779.87314   43389.5085
## BEDS        -59374.5563 14546.67942   16282.8349
## BATHS       176027.8543 17205.15513   22791.6266
## SQFT           234.0418    14.65724      27.3657
```

9. Add the square of the LOT_SIZE variable to Model 1 (call this Model 1B). Fit Model 1B and report the table of coefficients, the usual standard errors that assume constant variance, and robust standard errors.

```r
model.1b <- lm(LAST_SALE_PRICE ~ . + I(LOT_SIZE^2), data = data)
summary(model.1b)
```

```
##
## Call:
## lm(formula = LAST_SALE_PRICE ~ . + I(LOT_SIZE^2), data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1337125  -158642   -18611   117468  2980686
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)    9.870e+04  4.135e+04   2.387 0.017179 *
## BEDS          -4.850e+04  1.425e+04  -3.405 0.000689 ***
## BATHS          1.688e+05  1.677e+04  10.064  < 2e-16 ***
## LOT_SIZE      -1.704e+01  3.904e+00  -4.364 1.41e-05 ***
## SQFT           2.281e+02  1.447e+01  15.769  < 2e-16 ***
## I(LOT_SIZE^2)  4.666e-04  6.752e-05   6.910 8.66e-12 ***
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 314800 on 994 degrees of freedom
## Multiple R-squared:  0.4934, Adjusted R-squared:  0.4909
## F-statistic: 193.6 on 5 and 994 DF,  p-value: < 2.2e-16
```

```
robust.se.1b <- sqrt(diag(sandwich::vcovHC(model.1b)))
(ests.table.1b <- data.frame(cbind(summary(model.1b)$coefficients[,c("Estimate","Std. Error")], robust.se.1b)))
```

```
##                   Estimate   Std..Error robust.se.1b
## (Intercept)    9.870353e+04 4.135269e+04 6.963976e+04
## BEDS          -4.850262e+04 1.424650e+04 1.561273e+04
## BATHS          1.688097e+05 1.677417e+04 2.469718e+04
## LOT_SIZE      -1.704054e+01 3.904434e+00 1.114149e+01
## SQFT           2.281414e+02 1.446783e+01 2.466558e+01
## I(LOT_SIZE^2)  4.665612e-04 6.752152e-05 3.263620e-04
```

10. Perform the F test to compare Model 1A and Model 1B. Report the p-value.

```
anova(model.1a, model.1b)
```

```
## Analysis of Variance Table
##
## Model 1: LAST_SALE_PRICE ~ (BEDS + BATHS + LOT_SIZE + SQFT) - LOT_SIZE
## Model 2: LAST_SALE_PRICE ~ BEDS + BATHS + LOT_SIZE + SQFT + I(LOT_SIZE^2)
##   Res.Df        RSS Df  Sum of Sq      F    Pr(>F)
## 1    996 1.0461e+14
## 2    994 9.8474e+13  2 6.1379e+12 30.978 8.893e-14 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

11. State the null hypothesis being tested in Q10 either in words or by using model formulas.

$$H_0 : \beta_{LOT\hat{\_}SIZE} = \beta_{LOT\hat{\_}SIZE^2} = 0$$

12. Perform the robust Wald test to compare Model 1A and Model 1B. Report the p-value.

```
waldtest(model.1b, model.1a, test = "Chisq", vcov = vcovHC)
```

```
## Wald test
##
## Model 1: LAST_SALE_PRICE ~ BEDS + BATHS + LOT_SIZE + SQFT + I(LOT_SIZE^2)
## Model 2: LAST_SALE_PRICE ~ (BEDS + BATHS + LOT_SIZE + SQFT) - LOT_SIZE
##   Res.Df Df  Chisq Pr(>Chisq)
## 1    994
## 2    996 -2 2.3397     0.3104
```

13. Compare the results of the tests in Q10 and Q12. Which test is valid?

In Q10 we reject the null hypothesis and in Q12 we fail to reject the null hypothesis. The test in Q12 should be valid because as discussed above the constant variance assumption is violated in both the models (as established in HW5).

The following questions use the LOG_PRICE variable as in HW 5. Fit models corresponding to Model 1A and Model 1B with LOG_PRICE as the response variable. Call these models Model 1A_Log and Model 1B_Log.

```
model.1a.log <- lm(log10(LAST_SALE_PRICE) ~ . - LOT_SIZE, data = data)
model.1b.log <- lm(log10(LAST_SALE_PRICE) ~ . + I(LOT_SIZE^2), data = data)

summary(model.1a.log)
```

```
##
## Call:
## lm(formula = log10(LAST_SALE_PRICE) ~ . - LOT_SIZE, data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.94836 -0.08403  0.00746  0.09383  0.56812
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  5.455e+00  1.921e-02 283.903   <2e-16 ***
## BEDS        -1.369e-02  7.026e-03  -1.949   0.0516 .
## BATHS        8.548e-02  8.310e-03  10.287   <2e-16 ***
## SQFT         9.754e-05  7.080e-06  13.777   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1565 on 996 degrees of freedom
## Multiple R-squared:  0.4413, Adjusted R-squared:  0.4396
## F-statistic: 262.3 on 3 and 996 DF,  p-value: < 2.2e-16
```

```
summary(model.1b.log)
```

```
##
## Call:
## lm(formula = log10(LAST_SALE_PRICE) ~ . + I(LOT_SIZE^2), data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.94016 -0.07838  0.00200  0.08263  0.57133
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)    5.508e+00  2.004e-02 274.875  < 2e-16 ***
## BEDS          -7.129e-03  6.903e-03  -1.033    0.302
## BATHS          8.020e-02  8.128e-03   9.867  < 2e-16 ***
## LOT_SIZE      -1.392e-05  1.892e-06  -7.356 3.95e-13 ***
## SQFT           1.024e-04  7.010e-06  14.603  < 2e-16 ***
## I(LOT_SIZE^2)  2.292e-10  3.272e-11   7.005 4.56e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1525 on 994 degrees of freedom
## Multiple R-squared:  0.4707, Adjusted R-squared:  0.4681
## F-statistic: 176.8 on 5 and 994 DF,  p-value: < 2.2e-16
```

14. Perform the F test to compare Model 1A_Log and Model 1B_Log. Report the p-value.

```r
anova(model.1b.log, model.1a.log)
```

```
## Analysis of Variance Table
##
## Model 1: log10(LAST_SALE_PRICE) ~ BEDS + BATHS + LOT_SIZE + SQFT + I(LOT_SIZE^2)
## Model 2: log10(LAST_SALE_PRICE) ~ (BEDS + BATHS + LOT_SIZE + SQFT) - LOT_SIZE
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1    994 23.121
## 2    996 24.406 -2   -1.2848 27.618 2.124e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

15. State the null hypothesis being tested in Q14 either in words or by using model formulas.

$$H_0 : \beta_{LOT\hat{\_}SIZE} = \beta_{LOT\hat{\_}SIZE^2} = 0$$

16. Perform the robust Wald test to compare Model 1A_Log and Model 1B_Log. Report the p-value.

```r
waldtest(model.1a.log, model.1b.log, test = "Chisq", vcov = vcovHC)
```

```
## Wald test
##
## Model 1: log10(LAST_SALE_PRICE) ~ (BEDS + BATHS + LOT_SIZE + SQFT) - LOT_SIZE
## Model 2: log10(LAST_SALE_PRICE) ~ BEDS + BATHS + LOT_SIZE + SQFT + I(LOT_SIZE^2)
##   Res.Df Df  Chisq Pr(>Chisq)
## 1    996
## 2    994  2 44.081  2.678e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

17. Compare the results of the tests in Q14 and Q16. Do they give the same conclusion?

The results in Q14 and Q16 arrive at the same conclusion. We reject the null hypothesis.

18. Based on all of the analyses performed, answer the following question. Is there evidence for an association between the size of the lot and sales price? Explain.

Yes, there is evidence for association between lot size and sale price. More specifically, the association appears to exhibit some non-linear characteristics based on the results in Q14 and Q16 using LOT_SIZE^2.

We should draw statistical inference from the results of the log Model 1A and 1B over the regular Model 1A and 1B because even though we are able to account for the violation of the constant variance assumption through robust statistics in regular model 1A and 1B we still find that the linearity assumption is violated, which along with independence is the most important assumption in linear regression.

We can rely on the results of non-robust Wald tests (Q14) in case of log Model 1A and 1B because all the assumptions of linear regression are being met, and there is no need to use robust statistics.