Homework Assignment 6

Shrusti Ghela

March 03, 2022

Data: "Sales_sample.csv"

The data are a random sample of size 1000 from the "Sales" data (after removing observations with missing values).

Variables:

LAST_SALE_PRICE: the sale price of the home SQFT: area of the house (sq. ft.) LOT_SIZE: area of the lot (sq. ft.) BEDS: number of bedrooms BATHS: number of bathrooms

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.

```r
model_1 <- lm(LAST_SALE_PRICE ~ SQFT + LOT_SIZE + BEDS + BATHS, data=sales_data)
summary(model_1)
```

```
##
## Call:
## lm(formula = LAST_SALE_PRICE ~ SQFT + LOT_SIZE + BEDS + BATHS,
##     data = sales_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
## SQFT           224.502     14.794  15.175  < 2e-16 ***
## LOT_SIZE         6.844      1.858   3.684 0.000242 ***
## BEDS        -60884.742  14461.536  -4.210 2.78e-05 ***
## BATHS       178177.446  17107.532  10.415  < 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
```

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

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

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

The output from vcovHC is the estimated variance-covariance matrix of variances and covariances of the parameter estimates.

```
round(vcovHC(model_1),6)
```

```
              (Intercept)          SQFT      LOT_SIZE          BEDS
(Intercept) 2465697725.9 -284340.78988 -201398.36774   94829856.12
SQFT           -284340.8     595.10248       7.07667    -183849.89
LOT_SIZE       -201398.4       7.07667      59.82092     -78143.04
BEDS          94829856.1 -183849.89154  -78143.04417  297766759.60
BATHS       -502412112.2 -188824.59898   38129.93669 -103733107.04
                    BATHS
(Intercept) -502412112.23
SQFT           -188824.60
LOT_SIZE         38129.94
BEDS        -103733107.04
BATHS        519669890.96
```

The diagonal elements of the variance-covariance matrix are the variances of the coefficients, so their square-roots are the SEs.

Let's compare them to the standard SEs from the lm function.

```
v <- vcovHC(model_1)
robust.se <- sqrt(diag(v))
round(cbind(summary(model_1)$coef,robust.se),4)
```

```
##                Estimate Std. Error t value Pr(>|t|)  robust.se
## (Intercept)   5982.6043 40023.2714  0.1495   0.8812 49655.7925
## SQFT           224.5021    14.7940 15.1752   0.0000    24.3947
## LOT_SIZE         6.8441     1.8577  3.6841   0.0002     7.7344
## BEDS        -60884.7421 14461.5362 -4.2101   0.0000 17255.9196
## BATHS       178177.4461 17107.5317 10.4151   0.0000 22796.2692
```

We can see that the robust SEs are larger than the standard SEs.

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

For large sample sizes we usually use robust SEs. If we are confident about the homoscedasticity (constant variance) assumption, we can use the usual SEs. For small sample sizes they can be more accurate than the robust SEs (as long as the constant variance assumption holds) - the reason is that the robust SEs can be somewhat unstable with very small samples.

From HW5(1.4), we know that the constant variance assumption is not met for model_1. Furthermore, our sample size is large enough, so we should use the robust SEs instead of usual SEs.

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.

```r
reduced.model_1 <- lm(LAST_SALE_PRICE ~ SQFT + BEDS + BATHS, data=sales_data)
anova(reduced.model_1, model_1)
```

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

```r
waldtest(reduced.model_1, model_1)
```

```
## Wald test
##
## Model 1: LAST_SALE_PRICE ~ SQFT + BEDS + BATHS
## Model 2: LAST_SALE_PRICE ~ SQFT + LOT_SIZE + BEDS + BATHS
##   Res.Df Df      F    Pr(>F)
## 1    996
## 2    995  1 13.573 0.0002418 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

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.

```r
waldtest(reduced.model_1, model_1, test="Chisq",vcov=vcovHC)
```

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

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

```r
SE.jack <- (n-1)*sd(b.jack)/sqrt(n)
SE.jack
```

```
[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.

```r
test_statistic <- (model_1$coef[3] - 0)/SE.jack
data.frame(test_statistic,p=1-pf(test_statistic, 1,995))
```

```
##          test_statistic         p
## LOT_SIZE       0.885348 0.3469694
```

7. Do the tests in Q3, Q4, and Q6 agree? Which of these tests are valid? The p-value is greater than 0.05 for Q4 and Q6, SO, we would not reject the null hypothesis that the coefficient of the LOT_SIZE variable is equal to 0. However, for Q3, the p-value is less than 0.05, so we reject the null hypothesis. There are also robust Wald Tests for composite hypotheses in linear regression that can be used in place of the F-test, when model assumptions about the variance do not hold. So, the robust Wald test is a valid test. Jackknife estimate of the SE to test the null hypothesis is also a valid test because it is resistant to the violation of assumption of constant variance.

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 <- lm(LAST_SALE_PRICE ~ SQFT + BEDS + BATHS, data=sales_data)
#summary(model_1A)
```

```r
v <- vcovHC(model_1A)
robust.se <- sqrt(diag(v))
round(cbind(summary(model_1A)$coef,robust.se),4)
```

```
##                Estimate Std. Error t value Pr(>|t|)  robust.se
## (Intercept)  29034.4577 39779.8731  0.7299   0.4656 43389.5085
## SQFT           234.0418    14.6572 15.9677   0.0000    27.3657
## BEDS        -59374.5563 14546.6794 -4.0817   0.0000 16282.8349
## BATHS       176027.8543 17205.1551 10.2311   0.0000 22791.6266
```

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 ~ SQFT + LOT_SIZE + BEDS + BATHS + I(LOT_SIZE^2), data=sales_data)
#summary(model_1B)
```

```r
v <- vcovHC(model_1B)
robust.se <- sqrt(diag(v))
round(cbind(summary(model_1B)$coef,robust.se),4)
```

```
##                  Estimate Std. Error t value Pr(>|t|)  robust.se
## (Intercept)    98703.5276 41352.6927  2.3869   0.0172 69639.7586
## SQFT             228.1414    14.4678 15.7689   0.0000    24.6656
## LOT_SIZE         -17.0405     3.9044 -4.3644   0.0000    11.1415
## BEDS          -48502.6157 14246.4991 -3.4045   0.0007 15612.7258
## BATHS         168809.7119 16774.1743 10.0637   0.0000 24697.1788
## I(LOT_SIZE^2)      0.0005     0.0001  6.9098   0.0000     0.0003
```

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

```r
anova(model_1A, model_1B)
```

```
## Analysis of Variance Table
##
## Model 1: LAST_SALE_PRICE ~ SQFT + BEDS + BATHS
## Model 2: LAST_SALE_PRICE ~ SQFT + LOT_SIZE + BEDS + BATHS + 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.

model_1B(Full-model): $(LAST\_SALE\_PRICE) = \beta_0 + \beta_1 SQFT + \beta_2 LOT\_SIZE + \beta_3 BEDS + \beta_4 BATHS + \beta_5 LOT\_SIZE^2$

Null hypothesis: $H_0 : \beta_2 = \beta_5 = 0$.

model_1A(Reduced-model): $(LAST\_SALE\_PRICE) = \beta_0 + \beta_1 SQFT + \beta_3 BEDS + \beta_4 BATHS$

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

```r
waldtest(model_1A, model_1B, test="Chisq",vcov=vcovHC)
```

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

13. Compare the results of the tests in Q10 and Q12. Which test is valid? For Q12, we would not reject the null hypothesis. While for Q10, We reject the null hypothesis. There are also robust Wald Tests for composite hypotheses in linear regression that can be used in place of the F-test, when model assumptions about the variance do not hold. So, the robust Wald test is a valid test.

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.

```r
sales_data$LOG_PRICE <- log10(sales_data$LAST_SALE_PRICE)
head(sales_data)
```

```
##   BEDS BATHS LOT_SIZE LAST_SALE_PRICE SQFT LOG_PRICE
## 1    4  2.50    22578          678000 2410  5.831230
## 2    4  2.00     4000          888000 2660  5.948413
## 3    4  2.25     5000          682000 2800  5.833784
## 4    3  2.00     6400         1600000 3790  6.204120
## 5    6  2.50     7431          750000 2940  5.875061
## 6    4  1.75     7200          682000 2240  5.833784
```

```r
model_1A_LOG<- lm(LOG_PRICE ~ SQFT + BEDS + BATHS, data=sales_data)
model_1B_LOG <- lm(LOG_PRICE ~ SQFT + LOT_SIZE +  BEDS + BATHS + I(LOT_SIZE^2), data=sales_data)
```

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

```r
anova(model_1A_LOG, model_1B_LOG)
```

```
## Analysis of Variance Table
##
## Model 1: LOG_PRICE ~ SQFT + BEDS + BATHS
## Model 2: LOG_PRICE ~ SQFT + LOT_SIZE + BEDS + BATHS + I(LOT_SIZE^2)
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1    996 24.406
## 2    994 23.121  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. model_1B_LOG(Full-model): $(LOG\_PRICE) = \beta_0 + \beta_1 SQFT + \beta_2 LOT\_SIZE + \beta_3 BEDS + \beta_4 BATHS + \beta_5 LOT\_SIZE^2$

Null hypothesis: $H_0 : \beta_2 = \beta_5 = 0$.

model_1A_LOG(Reduced-model): $(LOG\_PRICE) = \beta_0 + \beta_1 SQFT + \beta_3 BEDS + \beta_4 BATHS$

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: LOG_PRICE ~ SQFT + BEDS + BATHS
## Model 2: LOG_PRICE ~ SQFT + LOT_SIZE + BEDS + BATHS + 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?

For both Q14 and Q16, we reject the null hypothesis that there is no linear relation of LOT_SIZE and LOT_SIZE^2 WITH LOG_PRICE as the p-value is significantly less than 0.05.

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. Throughout this assignment, we went through multiple combinations to find out the association between lot size and the sales price. In the last segment, we reject the null hypothesis that there is no linear relationship between LOG_PRICE and LOT_SIZE and LOT_SIZE^2. So, there is some evidence for an association between the size of the lot and sales price. However, that association is not strictly linear so to speak.