DATA 557: Applied Statistics and Experimental Design
Homework 6

Abhishek Saini

February 2022

Question 1

0.1 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

```
> library("sandwich")
>
> sales = read.csv("Sales_sample.csv")
>
> model1 = lm(formula = LAST_SALE_PRICE ~ SQFT + LOT_SIZE + BEDS + BATHS, data = sales)
> pred = model1$fitted.values
> res = model1$residuals
>
> v = vcovHC(model1)
> robust.se = sqrt(diag(v))
> round(cbind(summary(model1)$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
```

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

We should use the robust standard errors since not all assumptions of linear regression were valid as we saw in HW5. In particular, the constant variance assumption was not valid which would lead to incorrect inferences if the usual standard errors are used.

0.3 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.

The test statistic is 3.6841412 and p-value is 2.418418e-04

0.4 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.

Changing variable to LOT_SIZE because otherwise Q7 doesn't make sense. The test statistic for LOT_SIZE is 0.8848967 and the p-value is 0.3764259

```
> p_val = 2*(1-pt(abs(0.8848967), 1000-4));p_val
[1] 0.3764259
```

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

The jackknife estimate of the SE is 7.730455.

```
> n <- nrow(sales)
> b.jack <- rep(0,n)
> for(i in 1:n){
+   lmi <- lm(formula = LAST_SALE_PRICE ~ SQFT + LOT_SIZE + BEDS + BATHS, data=sales, subset=-
+   b.jack[i] <- lmi$coef[3]
+ }
>
> SE.jack <- (n-1)*sd(b.jack)/sqrt(n);SE.jack
[1] 7.730455
```

0.6 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.

The test statistic is 0.885348 and p-value is 0.3761827

```
> t_jack = 6.844143/SE.jack;t_jack
[1] 0.885348
> p_jack = 2*(1-pt(abs(t_jack), 1000-4));p_jack
[1] 0.3761827
```

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

The results in Q4 and Q6 agree and disagree with Q3. The tests 4 and 6 are valid because the constant variance assumption doesn't hold in model 1. So we cannot rely on the usual standard error for inferences.

0.8 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.

```
> model1_A = lm(formula = LAST_SALE_PRICE ~ SQFT + BEDS + BATHS, data = sales)
> v = vcovHC(model1_A)
> robust_A.se = sqrt(diag(v))
> round(cbind(summary(model1_A)$coef,robust_A.se),4)
               Estimate Std. Error t value Pr(>|t|) robust_A.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
```

0.9 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.

```
> model1_B = lm(formula = LAST_SALE_PRICE ~ SQFT + LOT_SIZE + I(LOT_SIZE^2) + BEDS + BATHS, da
> v = vcovHC(model1_B)
> robust_B.se = sqrt(diag(v))
> round(cbind(summary(model1_B)$coef,robust_B.se),4)
                 Estimate Std. Error t value Pr(>|t|) robust_B.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
I(LOT_SIZE^2)      0.0005     0.0001  6.9098   0.0000      0.0003
BEDS          -48502.6157 14246.4991 -3.4045   0.0007  15612.7258
BATHS         168809.7119 16774.1743 10.0637   0.0000  24697.1788
```

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

The p-value of the F-test to compare Model 1A and Model 1B is 8.892886e-14

```
> anova_result = anova(model1_A, model1_B)
> SSER = anova_result$RSS[1]
> SSEF = anova_result$RSS[2]
> dfR = anova_result$Res.Df[1]
> dfF = anova_result$Res.Df[2]
> F = (SSER-SSEF)/SSEF/(dfR-dfF)*dfF
> p_val = 1-pf(F, 2, dfF);p_val
[1] 8.892886e-14
```

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

The null hypothesis says that the regression coefficient of the terms LOT_SIZE and LOT_SIZE^2 in the regression model are zero. This implies that LAST_SALE_PRICE has no linear association with LOT_SIZE or LOT_SIZE^2.

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

The p-value is 0.3104

```
> waldtest(model1_A, model1_B, test="Chisq",vcov=vcovHC)
Wald test

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

0.13 Compare the results of the tests in Q10 and Q11. Which test is valid?

The results of the tests in Q10 and Q11 have different conclusions. Robust Wald test is valid because the constant variance assumption doesn't hold so we cannot rely on F-test to give us the correct inference.

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.

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

The p-value for the F-test to compare Model 1A_Log and Model 1B_Log is 2.124079e-12

```
> anova_result = anova(model1_A_LOG, model1_B_LOG)
> SSER = anova_result$RSS[1]
> SSEF = anova_result$RSS[2]
> dfR = anova_result$Res.Df[1]
> dfF = anova_result$Res.Df[2]
> F = (SSER-SSEF)/SSEF/(dfR-dfF)*dfF
> p_val = 1-pf(F, 2, dfF);p_val
[1] 2.124079e-12
```

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

The null hypothesis says that the regression coefficient of the terms LOT_SIZE and LOT_SIZE^2 in the regression model are zero. This implies that LOG_PRICE has no linear association with LOT_SIZE or LOT_SIZE^2.

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

The p-value is 2.678e-10.

```
> waldtest(model1_A_LOG, model1_B_LOG, test="Chisq",vcov=vcovHC)
Wald test

Model 1: LOG_PRICE ~ SQFT + BEDS + BATHS
Model 2: LOG_PRICE ~ SQFT + LOT_SIZE + I(LOT_SIZE^2) + BEDS + BATHS
  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
```

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

The results of the tests in Q14 and Q16 agree with each other. This is because all assumptions of linear regression model hold in LOG_PRICE models as we saw in HW5. Hence we can rely on the F-test to produce valid inference.

0.18 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.

In the first part of HW6, we did not find any evidence for association between the size of the lot and sale price when the statistical inferences were valid with robust tests. In the final part, we found evidence for association between the size of the lot and logarithm of the sale price. Hence, we conclude that there is a linear association between logarithm of the sale price and size of the lot.