Assignment6

Amrit Bhat

3/3/2022

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.

```
model_1 <- lm(LAST_SALE_PRICE ~ SQFT + LOT_SIZE + BEDS + BATHS,data = sales_df)
vCovMatrix <- vcovHC(model_1)
robust.se <- sqrt(diag(vCovMatrix))
combined_summary <- round(cbind(summary(model_1)$coef,robust.se),3)
combined_summary
```

```
               Estimate Std. Error t value Pr(>|t|) robust.se
(Intercept)    5982.604  40023.271   0.149    0.881 49655.792
SQFT            224.502     14.794  15.175    0.000    24.395
LOT_SIZE          6.844      1.858   3.684    0.000     7.734
BEDS         -60884.742  14461.536  -4.210    0.000 17255.920
BATHS        178177.446  17107.532  10.415    0.000 22796.269
```

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

Since the regression model 1 doesn't satisfy the constant variance assumption and because the sample size is sufficiently large (1000), we should resort to using 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.

$H0 : \beta(LOT\_SIZE) = 0$

$H1 : \beta(LOT\_SIZE)! = 0$

```
lot_size_beta <- summary(model_1)$coefficients[3]
lot_size_beta_se <- summary(model_1)$coefficients[8]
n=nrow(sales_df)
n_coefficients_estd = 5 ## No. of coefficients estimated is 5, because there are
## coefficients of 4 predictor variables and one intercept being estimated
wald_statistic = lot_size_beta / lot_size_beta_se
p = 2*(1-pt(abs(wald_statistic),df=n-n_coefficients_estd))
data.frame(wald_statistic,p)
```

```
  wald_statistic            p
1       3.684141 0.0002418418
```

As $p < 0.05$, we have evidence for rejecting the null hypothesis using Wald test.

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.

H0 : $\beta(LOT\_SIZE) = 0$

H1 : $\beta(LOT\_SIZE)! = 0$

```
lot_size_beta <- combined_summary[3]
lot_size_beta_se <- combined_summary[23]
robust_wald_statistic = lot_size_beta / lot_size_beta_se
p = 2*(1-pt(abs(robust_wald_statistic),df=n-n_coefficients_estd))
data.frame(robust_wald_statistic,p)
```

```
  robust_wald_statistic         p
1             0.8849237 0.3764116
```

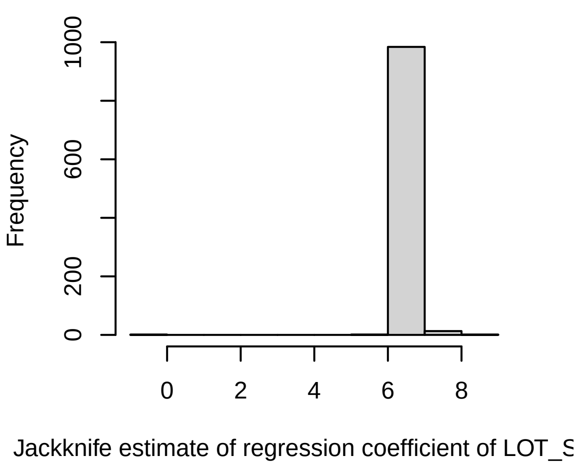
As $p > 0.05$, we do not have evidence for rejecting the null hypothesis using Robust Wald test.

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

```
par(mar=c(5,4,4,1))
n <- nrow(sales_df)
b.jack <- rep(0,n)
for(i in 1:n){
  lmi <- lm(LAST_SALE_PRICE ~ SQFT + LOT_SIZE + BEDS + BATHS,data = sales_df, subset=-i)
  b.jack[i] <- lmi$coef[3]
}
```

The distribution of the jackknife estimates

```
hist(b.jack,main="",xlab="Jackknife estimate of regression coefficient of LOT_SIZE")
```



Jackknife estimate of standard error

```
lot_size_SE.jack <- (n-1)*sd(b.jack)/sqrt(n)
lot_size_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.

```
jackknife_beta = mean(b.jack)
jackknife_statistic = jackknife_beta / lot_size_SE.jack
p = 2*(1-pt(abs(jackknife_statistic),df=n-n_coefficients_estd))
data.frame(jackknife_statistic,p)
```

```
    jackknife_statistic        p
  1           0.8852087 0.376258
```

As $p > 0.05$, we do not have evidence for rejecting the null hypothesis using jackknife test.

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

The test in Q3 disagrees with both Q4 and Q6 (which agree with each other to a high extent). As we know that the linear regression model doesn't hold the constant variance assumption, the results from the robust wald test and the jackknife test are valid.

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.

```
model_1A <- lm(LAST_SALE_PRICE ~ SQFT + BEDS + BATHS,data = sales_df)
vCovMatrix <- vcovHC(model_1A)
robust.se <- sqrt(diag(vCovMatrix))
combined_summary_1A <- round(cbind(summary(model_1A)$coef,robust.se),3)
combined_summary_1A
```

```
                Estimate Std. Error t value Pr(>|t|) robust.se
  (Intercept)  29034.458  39779.873   0.730    0.466 43389.508
  SQFT           234.042     14.657  15.968    0.000    27.366
  BEDS        -59374.556  14546.679  -4.082    0.000 16282.835
  BATHS       176027.854  17205.155  10.231    0.000 22791.627
```

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.

```
model_1B <- lm(LAST_SALE_PRICE ~ SQFT + LOT_SIZE + I(LOT_SIZE^2) + BEDS + BATHS,data = sales_df)
vCovMatrix <- vcovHC(model_1B)
robust.se <- sqrt(diag(vCovMatrix))
combined_summary_1B <- round(cbind(summary(model_1B)$coef,robust.se),3)
combined_summary_1B
```

```
                  Estimate Std. Error t value Pr(>|t|) robust.se
  (Intercept)    98703.528  41352.693   2.387    0.017 69639.759
  SQFT             228.141     14.468  15.769    0.000    24.666
  LOT_SIZE         -17.041      3.904  -4.364    0.000    11.141
  I(LOT_SIZE^2)      0.000      0.000   6.910    0.000     0.000
  BEDS          -48502.616  14246.499  -3.405    0.001 15612.726
  BATHS         168809.712  16774.174  10.064    0.000 24697.179
```

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

Full Model (Model 1B): $E(Y) = \beta_0 + \beta_1 SQFT + \beta_2 LOT\_SIZE + \beta_3 LOT\_SIZE^2 + \beta_4 BEDS + \beta_5 BATHS$

Null hypothesis: $H_0 : \beta_2 = \beta_3 = 0$.

Reduced Model (Model 1A): $E(Y) = \beta_0 + \beta_1 SQFT + \beta_4 BEDS + \beta_5 BATHS$

```
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 + I(LOT_SIZE^2) + BEDS + BATHS
  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
```

Since the p-value (8.893e-14) $<$ 0.05, therefore we reject the null hypothesis that LOT_SIZE and square of LOT_SIZE do not have an effect in calculating LAST_SALE_PRICE of house.

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

Null hypothesis: LOT_SIZE and square of LOT_SIZE do not have an effect or influence in calculating LAST_SALE_PRICE of house.

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

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

Since the p-value (0.3104) $>$ 0.05, we fail to reject the null hypothesis that LOT_SIZE and square of LOT_SIZE do not have an effect in calculating LAST_SALE_PRICE of house.

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

Q10 and Q12 generate contrary results to the hypothesis. Since the non-constant variance assumption doesn't hold for the full model, we can conclude that the robust test is valid.

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.

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

```
sales_df$LOG_PRICE <- log10(sales_df$LAST_SALE_PRICE)
model_1A_log <- lm(LOG_PRICE ~ SQFT + BEDS + BATHS,data = sales_df)
model_1B_log <- lm(LOG_PRICE ~ SQFT +LOT_SIZE + I(LOT_SIZE^2) + BEDS + BATHS,data = sales_df)
```

Full Model (Model 1B): $E(log(Y)) = \beta_0 + \beta_1 SQFT + \beta_2 LOT\_SIZE + \beta_3 LOT\_SIZE^2 + \beta_4 BEDS + \beta_5 BATHS$

Null hypothesis: $H_0 : \beta_2 = \beta_3 = 0$.

Reduced Model (Model 1A): $E(log(Y)) = \beta_0 + \beta_1 SQFT + \beta_4 BEDS + \beta_5 BATHS$

```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 + I(LOT_SIZE^2) + BEDS + BATHS
  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.

Null hypothesis: LOT_SIZE and square of LOT_SIZE do not have an effect or influence in calculating logarithmic estimates of LAST_SALE_PRICE of houses.

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 + 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
```

Since the p-value (2.678e-10) < 0.05, we have strong evidence to reject the null hypothesis that LOT_SIZE and square of LOT_SIZE do not have an effect in calculating logarithm of LAST_SALE_PRICE of houses.

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

Both tests conclude with having a strong evidence to reject the null hypothesis that LOT_SIZE and square of LOT_SIZE do not have an effect in estimating logarithm of LAST_SALE_PRICE of houses.

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.

Since the robust tests rejects the null hypothesis we can conclude that there is an association between the size of the lot and sales price