AMS 317 HW3 #2

Lucy Lin

2022-09-15

(a) Intercept estimate: -1104.2215

Intercept standard error: 687.0137

Slope estimate: 7.7942

Slope standard error: 0.3022

```
housedata <- read.csv("kc_house_data.csv")
fit <- lm(sqft_lot ~ sqft_living,data = housedata)
summary(fit)
```

```
##
## Call:
## lm(formula = sqft_lot ~ sqft_living, data = housedata)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
##  -65216  -10277   -5808    -883 1642331
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1104.2215   687.0137  -1.607    0.108
## sqft_living     7.7942     0.3022  25.795   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 40800 on 21611 degrees of freedom
## Multiple R-squared:  0.02987,    Adjusted R-squared:  0.02982
## F-statistic: 665.4 on 1 and 21611 DF,  p-value: < 2.2e-16
```

(b) For every 1 unit increase in sqft_living, "beta hat 1" is how much the sqft_lot increases. As the living space in square feet increases by 1, the lot size increases by about 7.7942.

(c) Reject the null hypothesis because p-value is lower than .05 here

```
#t.test(data$sqft_living,mu=7,alternative = "greater")
# R markdown ran into a render error so I will copy and paste the result from the line above
#Code from TA Yicong
#?t.test shows that the confidence level is 0.95 by default
```

```
One Sample t-test
```

data: data$sqft_living t = 331.81, df = 21612, p-value < 2.2e-16 alternative hypothesis: true mean is greater than 7 95 percent confidence interval: 2069.623 Inf sample estimates: mean of x 2079.9

(d) To find t value: t value = Estimate/Std.Error =

Example: sqft_living t value
= 7.7942/0.3022 = 25.795

To find Pr(>|t|):

##########Help#################

```
summary(fit)
```

```
##
## Call:
## lm(formula = sqft_lot ~ sqft_living, data = housedata)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
##  -65216  -10277   -5808    -883 1642331
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1104.2215   687.0137  -1.607    0.108
## sqft_living     7.7942     0.3022  25.795   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 40800 on 21611 degrees of freedom
## Multiple R-squared:  0.02987,    Adjusted R-squared:  0.02982
## F-statistic: 665.4 on 1 and 21611 DF,  p-value: < 2.2e-16
```

(e) ANOVA Table:

```
anova(fit)
```

```
## Analysis of Variance Table
##
## Response: sqft_lot
##                Df     Sum Sq    Mean Sq F value    Pr(>F)
## sqft_living     1 1.1075e+12 1.1075e+12  665.37 < 2.2e-16 ***
## Residuals   21611 3.5971e+13 1.6645e+09
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The ANOVA table is testing the null hypothesis for two-tailed test that beta_1 is equal to 0, with the F statistic with 1 degree of freedom for numerator and 199 for the denominator. The t value squared is the F statistic (Lecture 2 slide 38).

(f) R squared is SSR/TSS: SSR + SSE = TSS = 1.1075e+12 + 3.5971e+13. SSR = 1.1075e+12 SSR/TSS = 0.02987 Yes, the R squared matches the one from summary().

(g) The average lot area is 13704.79 when there is 1900 sqft living area. The 90% CI is [13239.63,14169.95].

```
new = data.frame(sqft_living = 1900)
predict(fit, new, interval = "confidence", level = 0.90)
```

```
##        fit      lwr      upr
## 1 13704.79 13239.63 14169.95
```

(h) (I am assuming that this is based on the R tutorial for the predict() function.) The average lot area is exactly the same but the lower and upper bounds are much more extreme with the predictions because it is a prediction of just a particular house. Negative values also do not make sense here but they are included in the interval. Average: 13704.79 Prediction Interval: [-53406.73,80816.31]

```
predict(fit, new, interval = "prediction", level = 0.90)
```

```
##        fit       lwr      upr
## 1 13704.79 -53406.73 80816.31
```