Lab 04 Regression Analysis (Part 1)

Lusine Zilfimian

March 11 (Wednesday), 2020

Contents

Needed packages

- These packages are required for this Session

```
library(knitr) # for kable()
library(dplyr) # for data manipulation
library(ggplot2) # for visualization
library(ggthemes) # for theme_pender()
library(stargazer) # for stargazer()
library(MASS) # for stepAIC()
library(car) # for vif()
```

- Install these packages if you do not have them

Data preparation

- Let's load and use the subset of the data

```
hd <- read.csv("housing.csv")[100:200,]; colnames(hd)
```

```
##  [1] "id"                "date"              "price"
##  [4] "bedrooms"          "bathrooms"         "sqft_living"
##  [7] "floors"            "waterfront"        "view"
## [10] "condition"         "grade"             "zipcode"
## [13] "Sqft_with_garden"
```

```
hd <- dplyr::select(hd, c("price", "sqft_living", "condition",
  "Sqft_with_garden"))
hd$price = hd$price/1000
table(hd$condition)
```

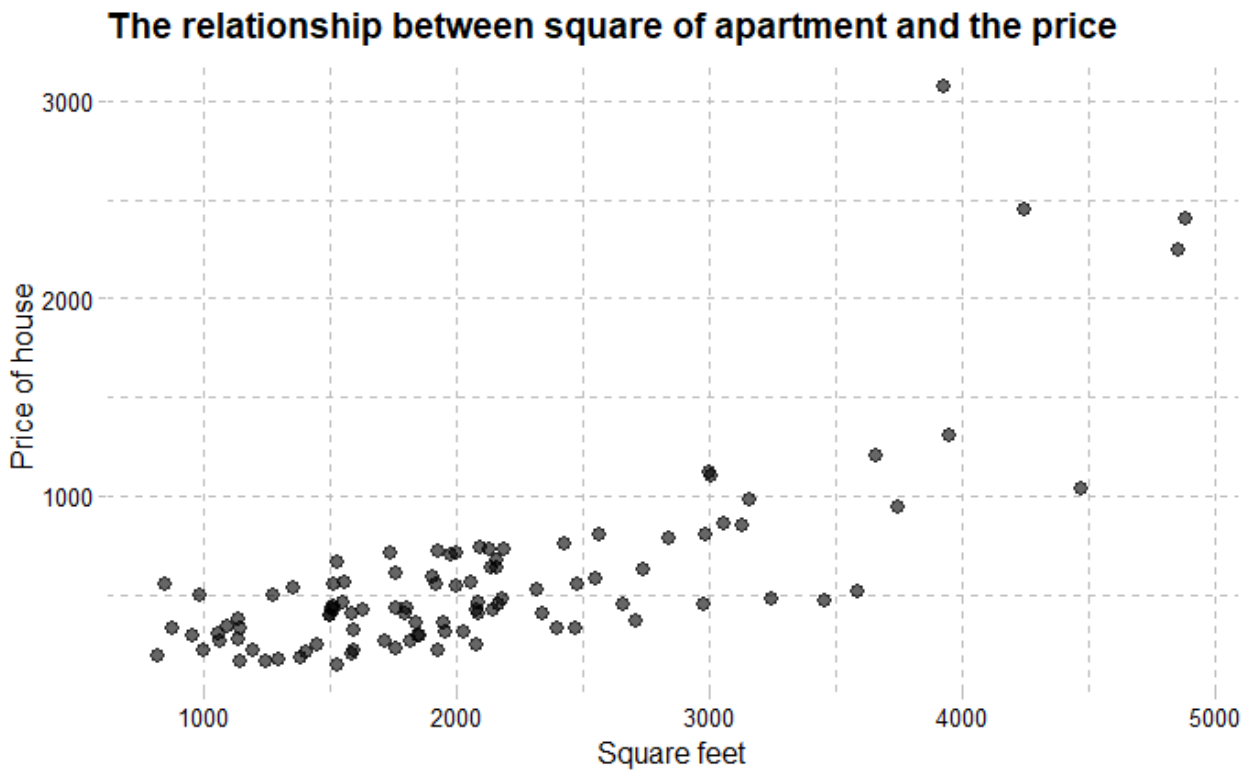
```
##
##  2  3  4  5
##  1 56 34 10
```

```
hd$condition <- factor(ifelse(hd$condition == 2 | hd$condition == 3,
  "fair", ifelse(hd$condition == 4, "good", "excellent")))
```

Undestanding the data

- Visualization of the main numeric variables:

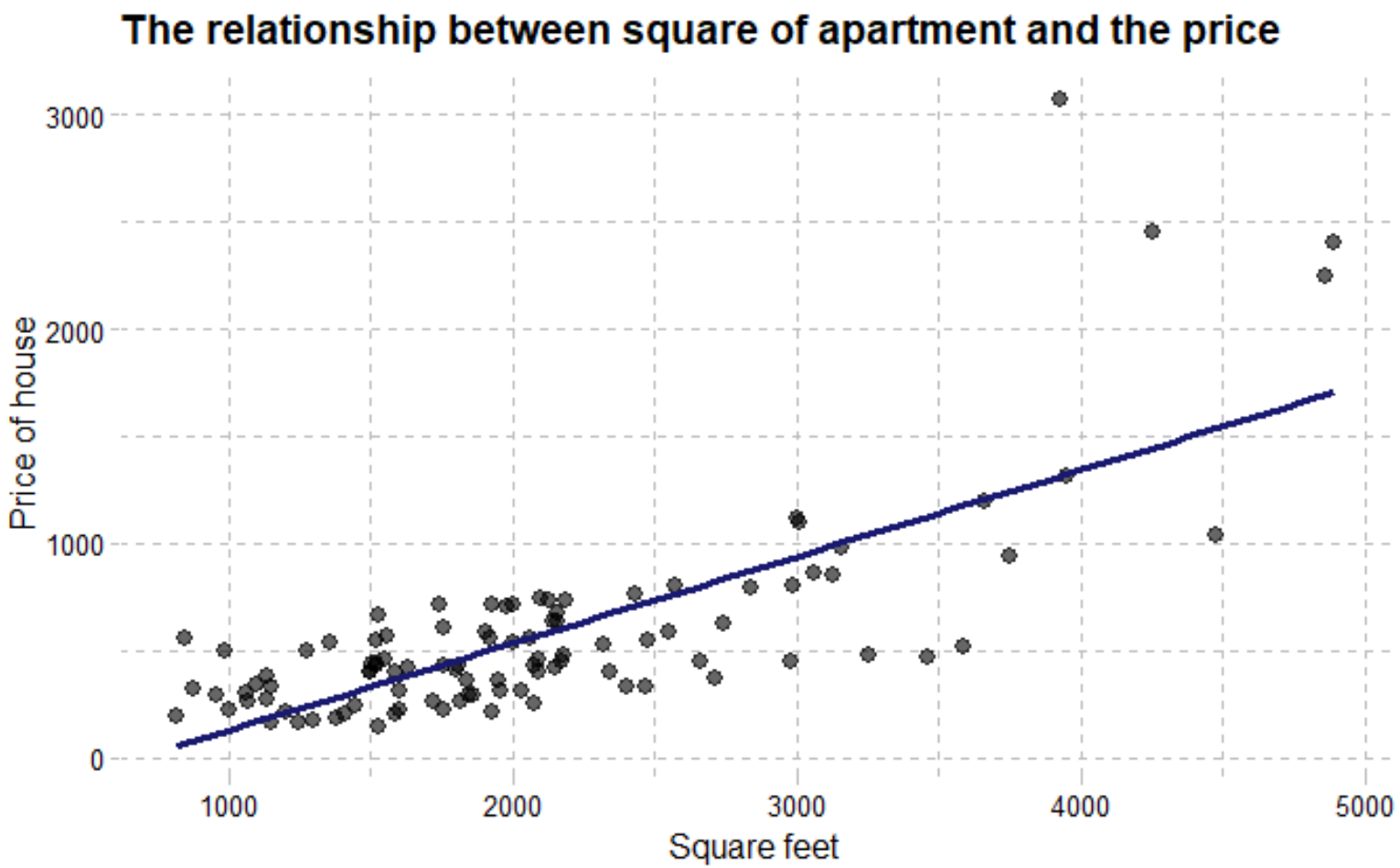
```
g1 <- ggplot(hd, aes(x = sqft_living, y = price))+
        geom_point(size = 2.5, alpha = 0.6)+
        ggtitle("The relationship between square of apartment and the price
        xlab("Square feet")+
        ylab("Price of house")+
        theme_pander()
```



Undestanding the data

- Adding the regression line

```
g2 <- g1+
  geom_smooth(method = "lm", se = F, col = "midnightblue", size = 1.2)
```



Intercept-only model

```
model0 <- lm(price ~ 1, data = hd)
names(model0)
```

```
##  [1] "coefficients"  "residuals"     "effects"       "rank"
##  [5] "fitted.values" "assign"        "qr"            "df.residual"
##  [9] "call"          "terms"         "model"
```

```
model0$coefficients
```

```
## (Intercept)
##    575.0682
```

Intercept-only model

```
summary(model0)
```

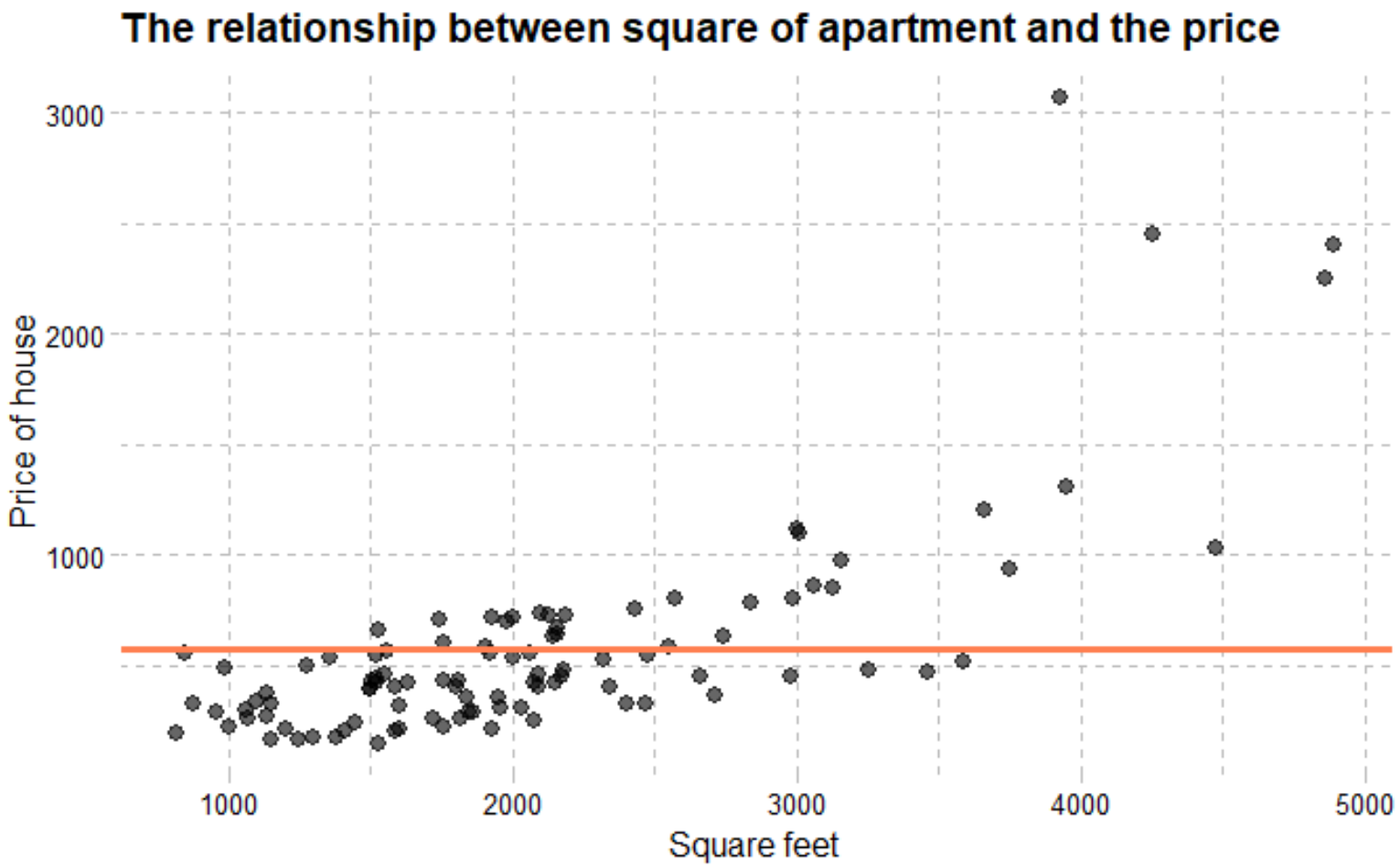
```
##
## Call:
## lm(formula = price ~ 1, data = hd)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -427.57 -258.57 -125.07   88.93 2494.93
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   575.07      47.06   12.22   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 473 on 100 degrees of freedom
```

```
mean(hd$price)
```

```
## [1] 575.0682
```

Intercept-only model

```
g3 <- g1 + geom_hline(yintercept = model0$coefficients ,
  col = "coral", size = 1.2)
```



Regression with one explanatory variable

```
model1 <- lm(price ~ sqft_living, data = hd)
coef(model1)
```

```
##  (Intercept)  sqft_living
## -275.6591736    0.4049002
```

- $\hat{\beta}_0 =$ -275.6591736
- $\hat{\beta}_1 =$ 0.4049002
- $\hat{Price} = \hat{\beta}_0 + \hat{\beta}_1 sqrt\_living$

Regression with one explanatory variable

```
summary(model1)
```

```
##
## Call:
## lm(formula = price ~ sqft_living, data = hd)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -662.93 -167.58   -7.02  140.02 1754.40
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -275.65917   80.41911  -3.428 0.000888 ***
## sqft_living    0.40490    0.03532  11.465  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 311.6 on 99 degrees of freedom
## Multiple R-squared:  0.5704, Adjusted R-squared:  0.5661
## F-statistic: 131.4 on 1 and 99 DF,  p-value: < 2.2e-16
```

Understanding the output of regression in R

- Let's derive the `Resudials` table:

```
min(resid(model1))
```

```
## [1] -662.9326
```

```
max(resid(model1))
```

```
## [1] 1754.401
```

```
median(resid(model1))
```

```
## [1] -7.015203
```

```
quantile(resid(model1), probs = c(0.25,0.75))
```

```
##       25%       75%
## -167.5823  140.0229
```

Understanding the output of regression in R

- Beta coefficients and SE

```
kable(head(X <- data.frame("X0" = 1,  "X1" = hd[,"sqft_living"])))
```

X0	X1
1	2340
1	2160
1	2320
1	1384
1	1820
1	2130

```
X <- data.matrix(X) # to apply the solve() function
Y <- data.matrix(hd[,"price"])
(b <- solve(t(X) %*% X) %*% t(X) %*% Y)
```

```
##             [,1]
## X0 -275.6591736
## X1    0.4049002
```

Understanding the output of regression in R

- Beta coefficients and SE

```
(RSE <- sqrt(sum((resid(model1)^2))/(dim(hd)[1] - 2)))
```

```
## [1] 311.5822
```

```
(bse <- RSE/sqrt(sum((hd$sqft_living - mean(hd$sqft_living))^2)))
```

```
## [1] 0.03531637
```

Understanding the output of regression in R

- t and p values (RSE is the same)

```
(t <- coef(model1)[2]/bse)
```

```
## sqft_living
##    11.46494
```

```
(p.value_t <- 2*pt(-t, df = 99))
```

```
##  sqft_living
## 7.221469e-20
```

```
(p.value_t <- 2*pt(-3.428, df = 99))
```

```
## [1] 0.0008873304
```

```
(RSE <- sqrt(sum((resid(model1)^2))/(dim(hd)[1] - 2)))
```

```
## [1] 311.5822
```

```
summary(model1)$sigma
```

```
## [1] 311.5822
```

Understanding the output of regression in R

- R^2 and F

```
(Rsquare <- sum((predict(model1) - mean(hd$price))^2)/
    sum(((hd$price - mean(hd$price))^2)))
```

```
## [1] 0.5703962
```

```
var(model1$fitted.values)/var(hd$price)
```

```
## [1] 0.5703962
```

```
cor(model1$fitted.values, hd$price)^2
```

```
## [1] 0.5703962
```

```
summary(model1)$r.sq
```

```
## [1] 0.5703962
```

```
(Fstat <- t^2)
```

```
## sqft_living
##    131.4449
```

Understanding the output of regression in R

- p values and confidence intervals

```
(p.valuef <- 2*pf(-Fstat,df1 = 2-1, df2 = 101-2))
```

```
## sqft_living
##           0
```

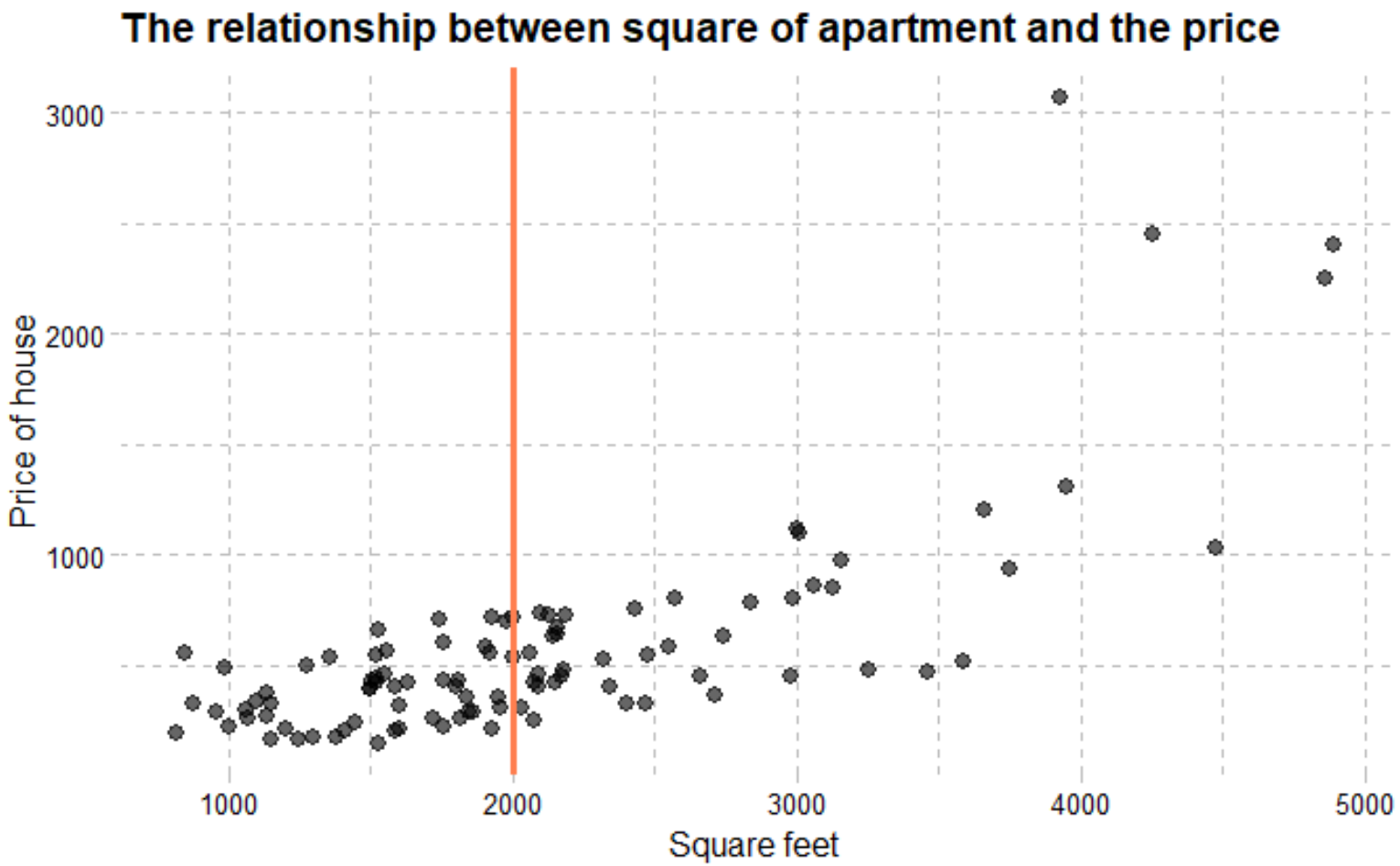
```
confint(model1)
```

```
##                     2.5 %       97.5 %
## (Intercept) -435.2281421 -116.0902050
## sqft_living    0.3348249    0.4749756
```

Interpretation

- The average or expected value given corresponding X

```
g4 <- g1 + geom_vline(xintercept = 2000, col = "coral", size = 1.2)
```



Prediction

- What will be the predicted price for the first three observations of the data?

```
pred.dat <- hd[1:3,]
pred.dat
```

```
##      price sqft_living condition Sqft_with_garden
## 100 404.95        2340      good              2351
## 101 671.50        2160 excellent              2269
## 102 530.00        2320      fair              2477
```

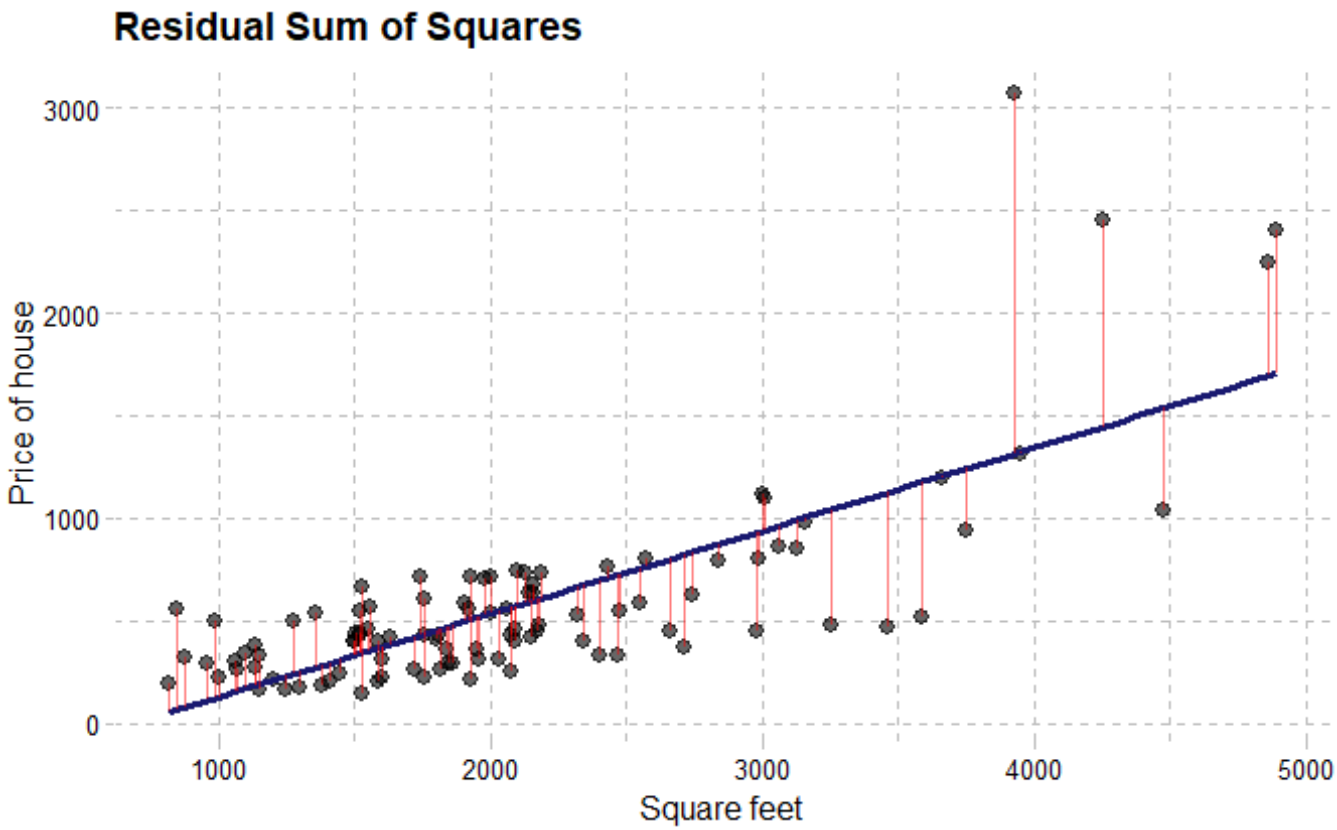
```
predict(model1, newdata = pred.dat)
```

```
##      100      101      102
## 671.8073 598.9253 663.7093
```

TSS, ESS, RSS

- Visualisation of RSS

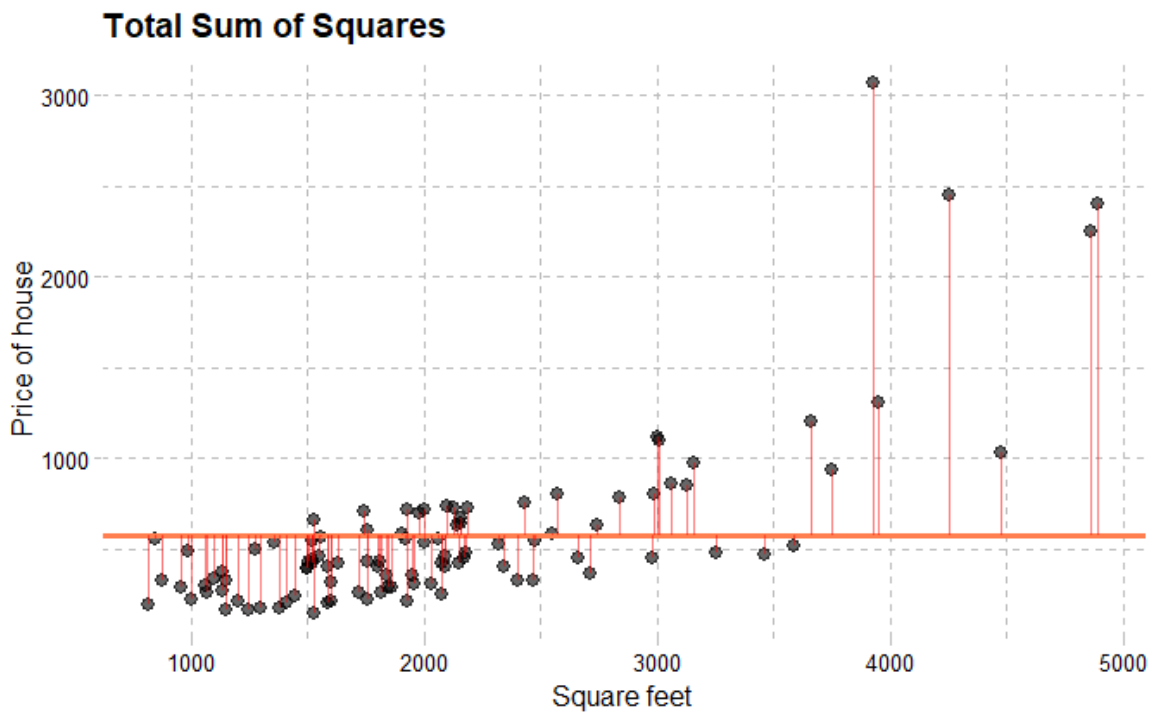
```
g5 <- g2 +
    geom_segment(aes(xend = sqft_living, yend = predict(model1)),
      alpha=0.5, col = "red")+
   ggtitle("Residual Sum of Squares")
```



TSS, ESS, RSS

- Visualisation of TSS

```
g6 <- g1 +
  geom_hline(yintercept = model0$coefficients , col = "coral", size = 1.2)+
  geom_segment(aes(xend = sqft_living, yend = mean(price)),
    alpha=0.5, col = "red")+
  ggtitle("Total Sum of Squares")
```



TSS, ESS, RSS

- Visualisation of ESS

```
g7 <- ggplot(hd, aes(x = sqft_living, fitted(model1)))+
  geom_point()+
  geom_hline(yintercept = model0$coefficients , col = "coral", size = 1.2)+
  geom_segment(aes(xend = sqft_living, yend = mean(price)),
    alpha=0.5, col = "red") +
    xlab("Square feet") + ylab("Price of house") +
        theme_pander() + ggtitle("Estimated Sum of Squares")
```