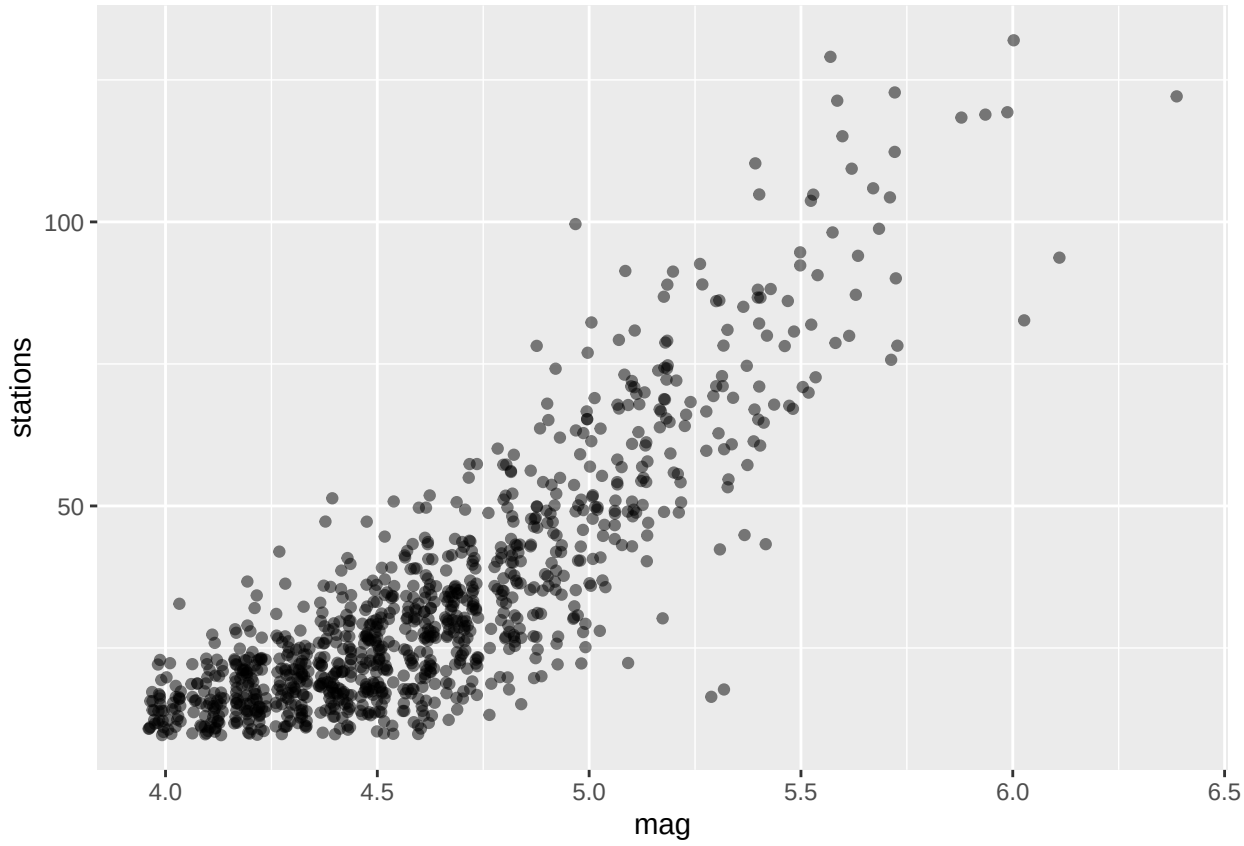
Lab 2: Linear Regression

An Island Never Cries

YILIN Li

Exercise 1

```
library(ggplot2)
p1 = ggplot(data = quakes, aes(x = mag, y = stations)) +
  geom_jitter(alpha = 0.5)
plot(p1)
```



From the graph above, I think number of stations which report the Earthquakes is positively correlated to the magnitude of the Earthquake.

Exercise 2

```
b0 = mean(quakes$stations)
b1 = 0
b1
```

```
## [1] 0
```

```
b0
```

```
## [1] 33.418
```

If in fact there's no relationship between magnitude of Earthquakes and the number of stations that report it, then the slope will be 0 and the intercept will be the average value of the number of the stations, which is around 33.

Exercise 3

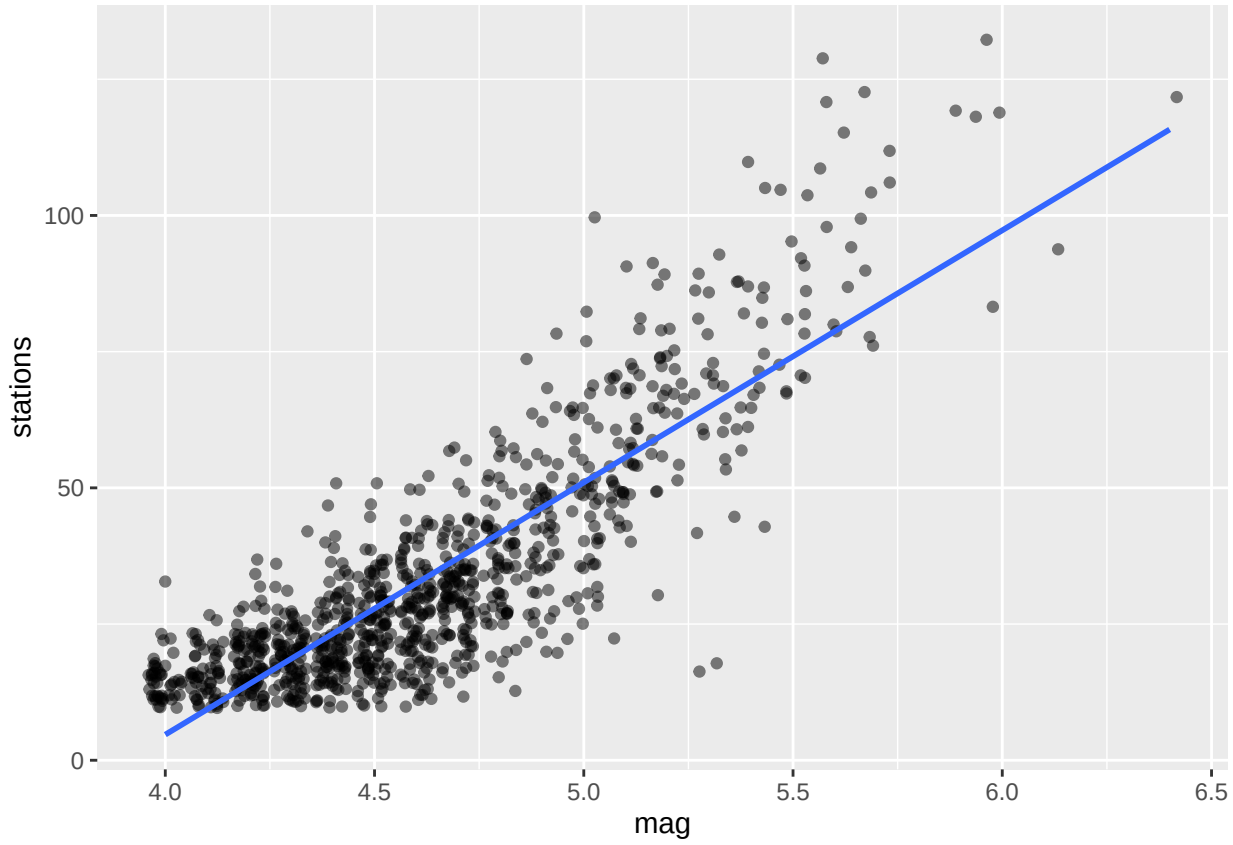
```
m1 = lm(stations ~ mag, data = quakes)
b0 = m1$coef[1]
b1 = m1$coef[2]
p2 = p1 + geom_smooth(method = 'lm', se = FALSE)
b0
```

```
## (Intercept)
##   -180.4243
```

```
b1
```

```
##      mag
## 46.28221
```

```
plot(p2)
```



1. The intercept is -180.4243, meaning if the magnitude of the Earthquake is 0, then about -180 stations will report. Even without practical meaning at a first glance, the negative value of the intercept indicates that no station will report if there's no Earthquake.

2. The slope is 46.28221, which means that as the magnitude of the Earthquake increases for 1 level, about 46 more stations will report the Earthquake.

Exercise 4

```
x = quakes$mag
y = quakes$stations
slope = (sd(y)/sd(x))*cor(x,y)
(b1 - slope) < 0.0001
```

```
##  mag
## TRUE
```

lm() function calculates the same value of slope as we calculate using the formula.

Exercise 5

```
n = nrow(quakes)
slope = m1$coefficients[2]
SE_slope = summary(m1)$coef[2,2]
t_stat = qt(0.025, df= n-2)

LB = slope + t_stat * SE_slope
UP = slope - t_stat * SE_slope
LB - confint(m1)[2,1] < 0.0001
```

```
##  mag
## TRUE
```

```
UP - confint(m1)[2,2] < 0.0001
```

```
##  mag
## TRUE
```

Two methods indeed construct the same confidence interval.

Exercise 6

```
pred_stations =predict(m1, data.frame(mag = 7.0))
pred_stations
```

```
##        1
## 143.5511
```

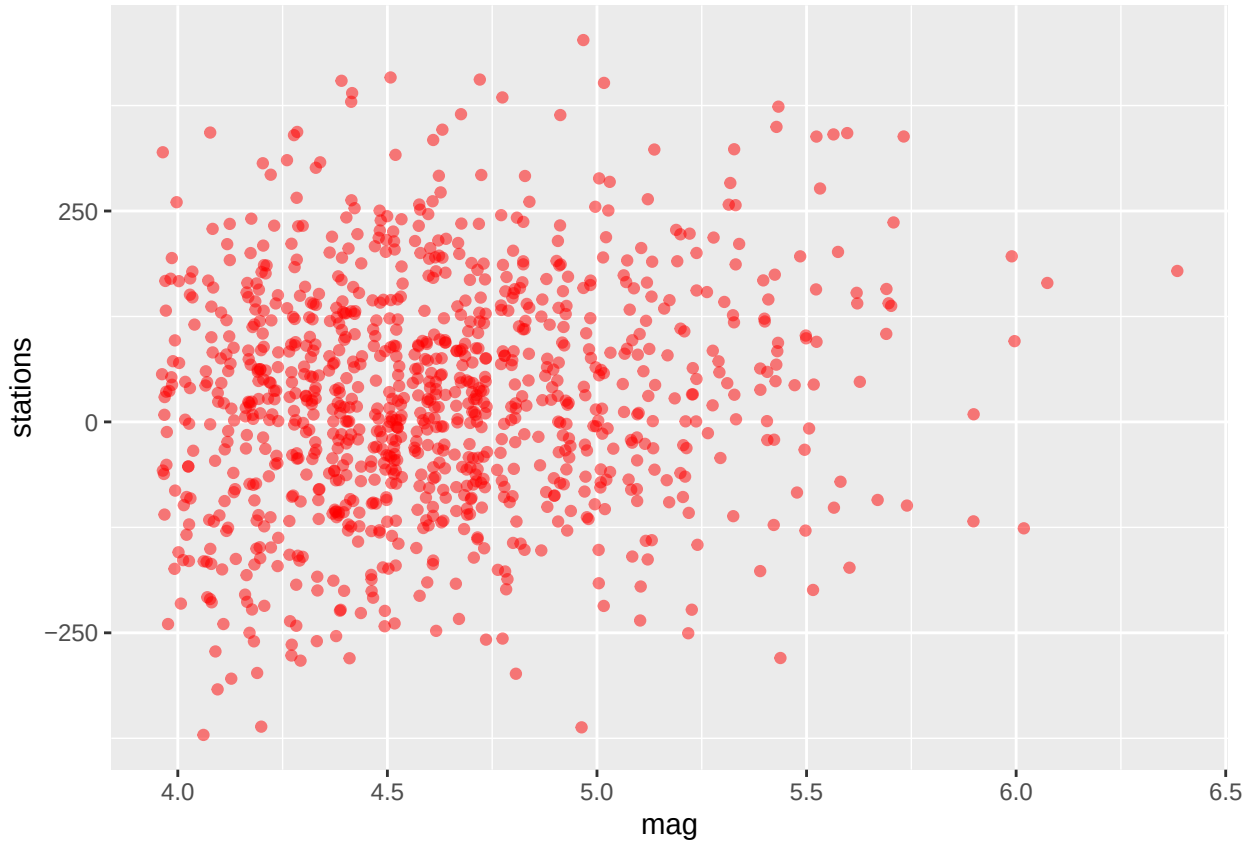
If the magnitude is 7.0, then about 144 stations will report the Earthquake.

Exercise 7

1. data description by plotting and inference of the relation bewteen magnitude and stations by checking the plot.
2. Inference
3. Inference by explaining the meaning of intercept and slope.
4. Inference.
5. Inference.
6. Prediction of future Earthquakes.

Simulation

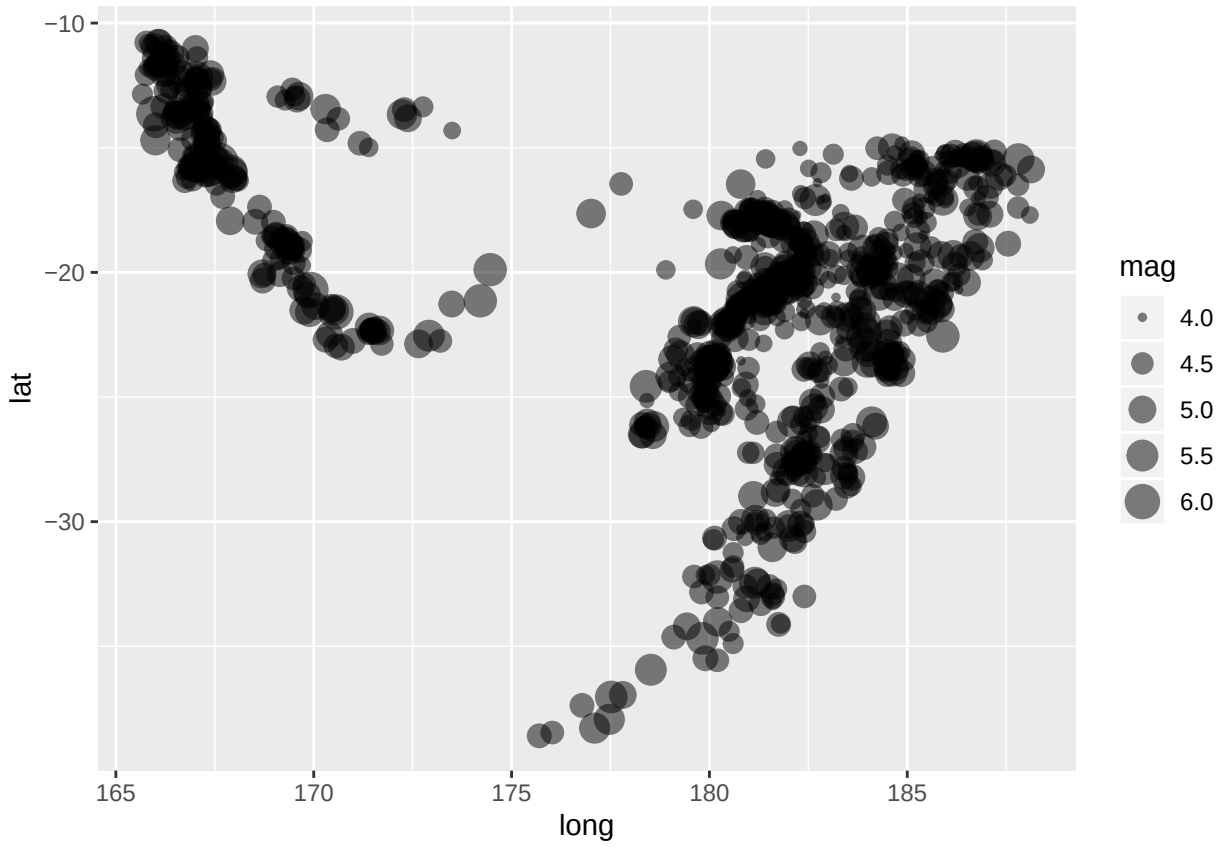
```
x = quakes$mag
f_hat = function(x){
  predict(m1, data.frame(mag = x))
}
y_hat = f_hat(x)
sigma_square = sum(m1$res^2)/(n - 2)
y = y_hat + rnorm(n, mean = 0, sd = sigma_square)
newdata = data.frame('mag' = x, 'stations' = y)
ggplot(data = newdata, aes(x = mag, y = stations)) +
  geom_jitter(alpha = 0.5, color = 'red')
```



Actually, it's not quite similar to the original data. The simulated data seems losing the linearity between magnitude and stations that report it. I think one of the reason is caused by the large residuals from the model, which create a large fluctuation of the irreducible error term. One solution could be adding mag^2 term to the model and fitting it.

Challenge Problem

```
ggplot(data = quakes, aes(x = long, y = lat, size = mag)) +
  geom_point(alpha = 0.5)
```


−10
−20
−30
lat
165
170
175
180
185
long
mag
4.0
4.5
5.0
5.5
6.0