

# Lab 2 – MATH 243

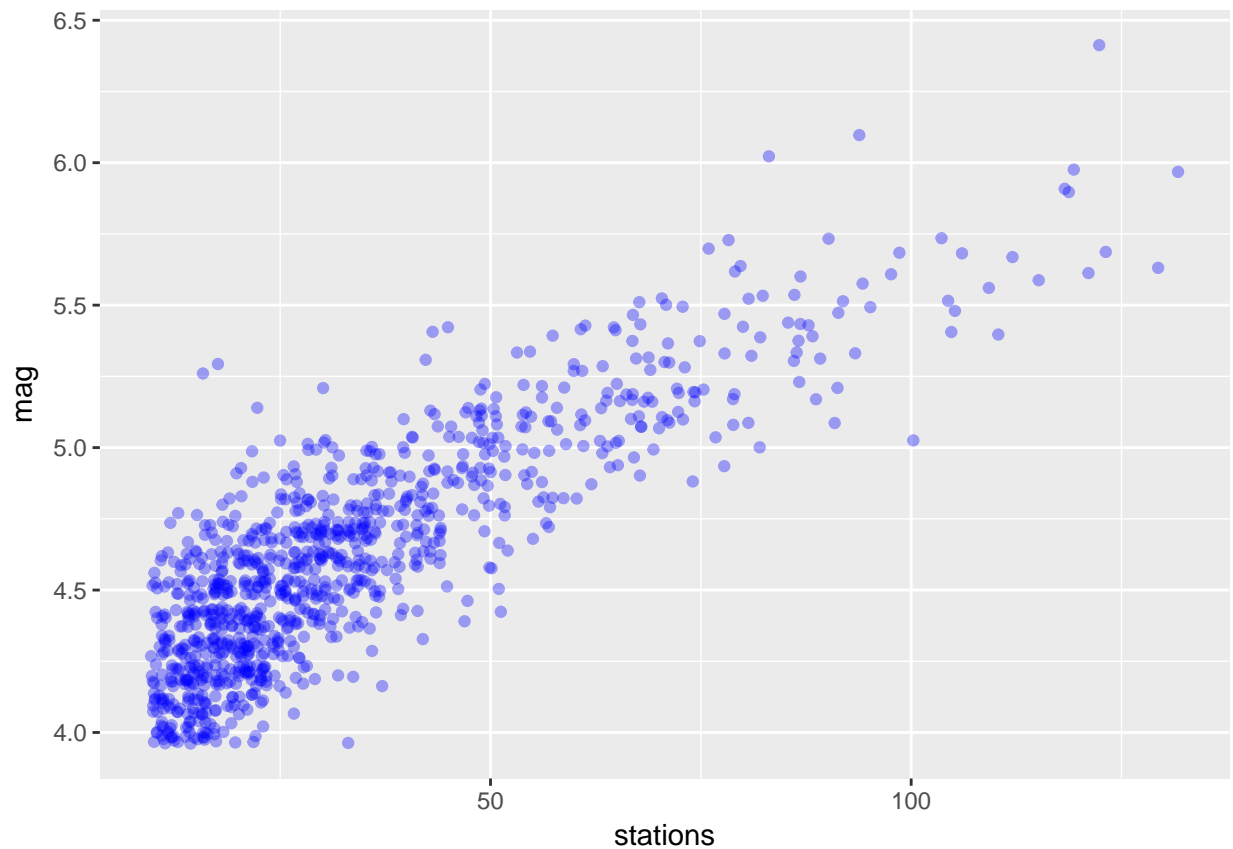
*Theodore Dounias*

*September 11, 2017*

## Earthquake Detection

1

```
#Using ggplot
ggplot(quakes, aes(x = stations, y = mag)) +
  geom_jitter(alpha = .35, colour = "blue")
```



I would characterize this relationship as fairly linear, with maybe a slight cubic element to it.

2

If there was no relationship, I would expect the slope to be zero, and the intercept to be some constant number, which if using the model described in (3.4) in our textbook would be equal to the sample response mean.

3

```
m1 <- lm(mag~stations, data = quakes)
summary(m1)
```

```
##
## Call:
```

```
## lm(formula = mag ~ stations, data = quakes)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.66269 -0.15124 -0.00428  0.13661  0.95227
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  4.0972676  0.0122067   335.66  <2e-16 ***
## stations     0.0156542  0.0003056    51.23  <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2115 on 998 degrees of freedom
## Multiple R-squared:  0.7245, Adjusted R-squared:  0.7242
## F-statistic: 2625 on 1 and 998 DF,  p-value: < 2.2e-16
```

We reject the null hypothesis of no relation based on the p-values we have here. Based on this summary, we conclude that an earthquake of magnitude equal to the intercept (~4.097) would not be reported by any station. For each new station reporting, we expect the earthquake to be of a magnitude .016 more severe.

4

```
#Using model in (3.4)
attach(quakes)
b1 <- sum((stations - mean(stations))*(mag - mean(mag)))/sum((stations - mean(stations))^2)
b1
```

```
## [1] 0.01565421
```

5

```
#Predicted conf_int
cf_b1 <- c(summary(m1)$coefficients[2, 1] - 2*summary(m1)$coefficients[2, 2],
          summary(m1)$coefficients[2, 1] + 2*summary(m1)$coefficients[2, 2])
cf_b1
```

```
## [1] 0.01504309 0.01626533
```

```
#Confidence interval
confint(m1)
```

```
##              2.5 %      97.5 %
## (Intercept) 4.0733138 4.12122135
## stations    0.0150546 0.01625382
```

6

```
#We are trying to find x for y = 7
st_mag7 <- (7 - summary(m1)$coefficients[1, 1])/summary(m1)$coefficients[2, 1]
st_mag7
```

```
## [1] 185.4282
```

7

Almost all of the questions, save for 6, involve some sort of inference, i.e. calculation or interpretation of parameters describing a population. Questions 1 and 3 ask us to use some tools to describe a dataset. Questions 2 and 6 ask us to predict some values (although 2 is much more about inference than prediction in the way it is interpreted in ISLR).

## Simulation

9

```
x <- as.numeric(quakes$stations)
```

10

```
f_hat <- function(x) {  
  summary(m1)$coefficients[1, 1] + x*summary(m1)$coefficients[2, 1]  
}
```

```
f_mean <- f_hat(x)
```

11

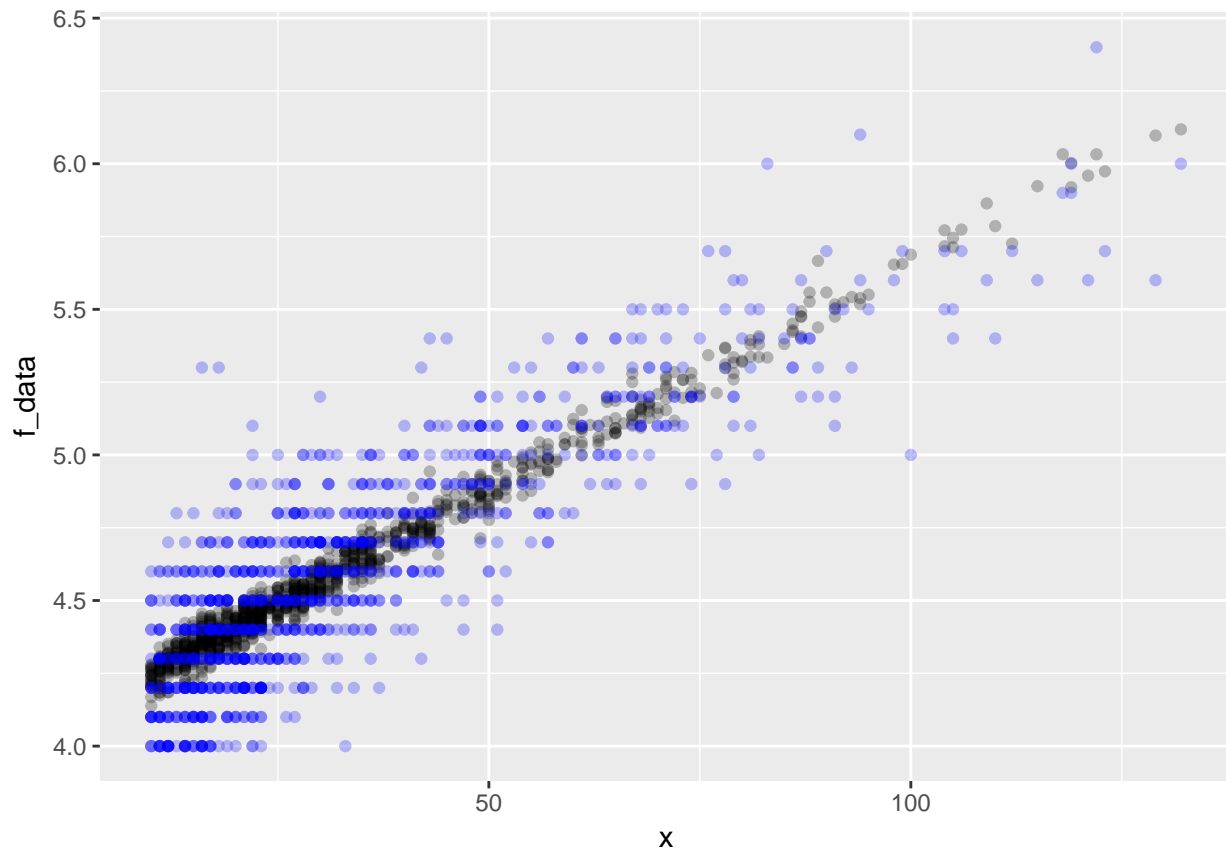
```
sigma <- sum((m1$res)^2)/(length(x) - 2)
```

```
f_data <- f_mean + rnorm(length(x), mean = 0, sd = sigma)
```

12

```
df_sim <- data.frame(x, f_data, quakes$mag)
```

```
ggplot(df_sim, aes(x, f_data)) +  
  geom_point(alpha = .25) +  
  geom_point(aes(x, quakes$mag), col = "blue", alpha = .25)
```



The simulated data is shown in black. The actual data we had is shown in blue. The simulated data seems to be far more concentrated around the linear path that we used to generate them, but seems to be following

the same trajectory overall, and has the same concentration along the x-axis.