SDS 323 Exercises 1

Kyle Carter, Jacob Rachiele, Crystal Tse, Jinfang Yan

2/14/2020

Problem 1: Flights at ABIA

Let's investigate the median departure delays by month. We look at the *median* since the data is highly skewed, as you see here.

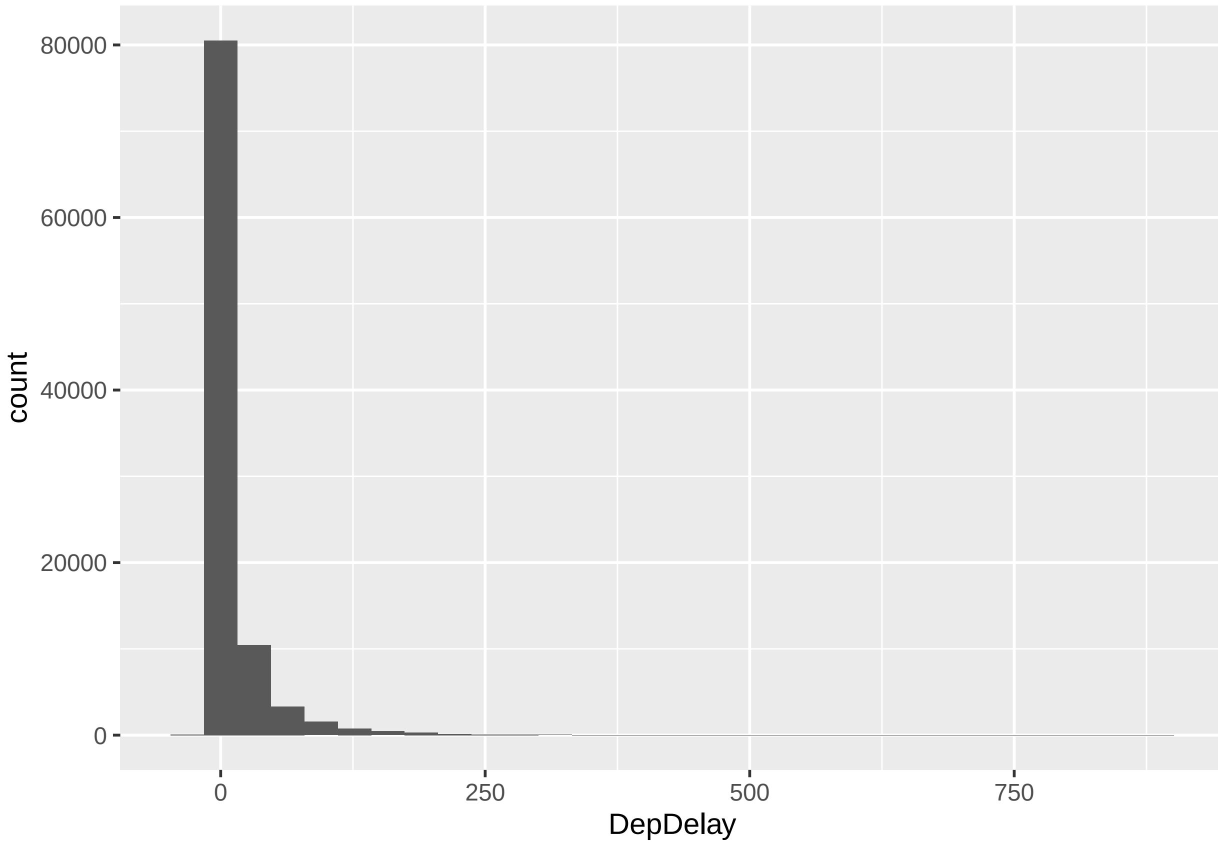
```
# Read in the Austin 2008 flights data

abia <- read.csv("./data/ABIA.csv", header = TRUE)

ggplot(data = abia) +
  geom_histogram(mapping = aes(x = DepDelay))
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

```
## Warning: Removed 1413 rows containing non-finite values (stat_bin).
```



departure_delays_by_month <- abia %>% group_by(Month) %>% filter(DepDelay > 0) %>% # Take the median since the distribution is highly skewed. summarize(delay = median(DepDelay, na.rm = TRUE)) %>% select(delay) %>% mutate(month = month.name)

ggplot(data = departure_delays_by_month) + geom_point(mapping = aes(x = delay, y = month), color = "red", size = 3, alpha = 0.6) + geom_vline(xintercept = 0, size = .25) + xlim(c(0, 20)) + scale_y_discrete(limits = rev(month.name)) + labs(title = "Median Departure Delay by Month", y = "", x ="Delay in Minutes")

arrival_delays_by_month <- abia %>% group_by(Month) %>% filter(ArrDelay > 0) %>% # Take the median since the distribution is highly skewed. summarize(delay = median(ArrDelay, na.rm = TRUE)) %>% select(delay) %>% mutate(month = month.name)

ggplot(data = arrival_delays_by_month) + geom_point(mapping = aes(x = delay, y = month), color = "red", size = 3, alpha = 0.6) + geom_vline(xintercept = 0, size = .25) + xlim(c(0, 20)) + scale_y_discrete(limits = rev(month.name)) + labs(title = "Median Arrival Delay by Month", y = "", x ="Delay in Minutes")

Problem 2: Regression Practice (Creatinine)

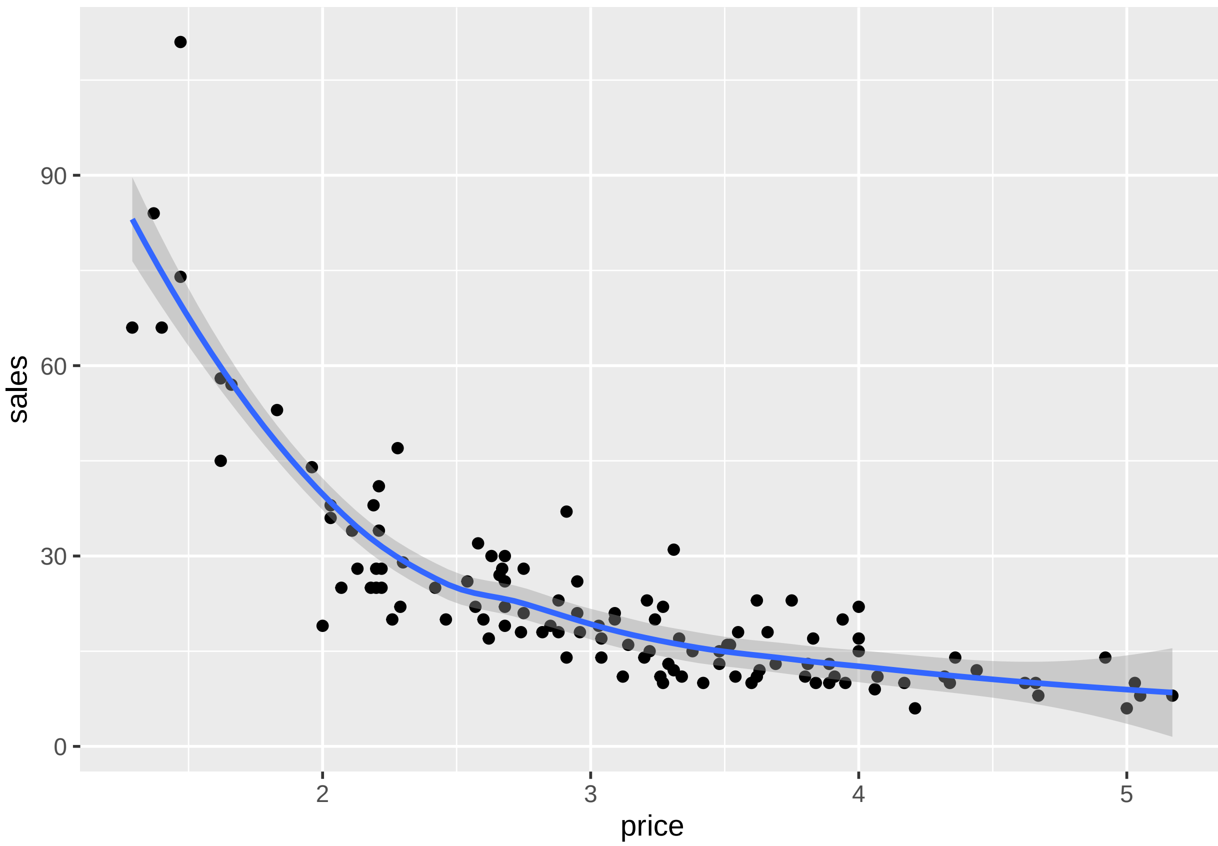
Problem 3: Green Buildings

Problem 4: Milk Prices

```
# Read the data file
milk <- read.csv("./data/milk.csv")
```

First, graph the data.

```r
ggplot(data = milk) +
  geom_point(aes(x = price, y = sales)) +
  geom_smooth(mapping = aes(x = price, y = sales))
```



Notice that this is not a linear relationship, which makes sense since quantity demanded is modeled in microeconomics using a Power Law: $Q = KP^E$, where Q is the quantity demanded, P is the price, E is the price elasticisty of demand and K is a constant.

Step 1: Write an equation that expresses net profit N in terms of both Q and P (and cost c)

$$N = (P - c)Q$$

Step 2: Use the microeconomic model of quanity demanded, which is a function of the price.

$Q = f(P) = KP^E$, so that $N = (P - c)f(P) = (P - c)(KP^E)$

The values of K and E are uknown, so we must estimate them from the data.

We can do this using linear regression using the product and power rules of logarithms, which tell us that $ln(Q) = ln(KP^E) = ln(K) + E(ln(P))$.

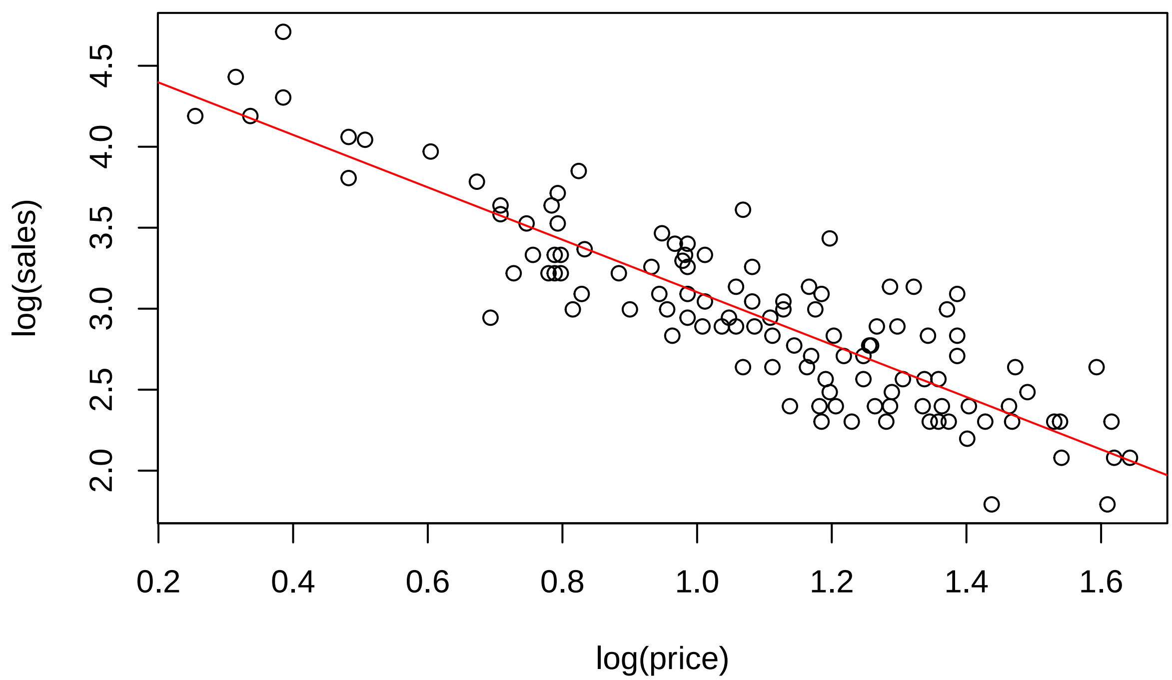
This has the form of a simple linear regression, where $\beta_0 = ln(K)$ and $\beta_1 = E$.

Step 3: Use simple linear regression to estimate the unknown coefficients.

```
model <- lm(log(sales) ~ log(price), data = milk)
```

Confirm the linearity of the logarithm of the data by plotting.

```
plot(log(sales) ~ log(price), data = milk)
abline(model, col = "red")
```

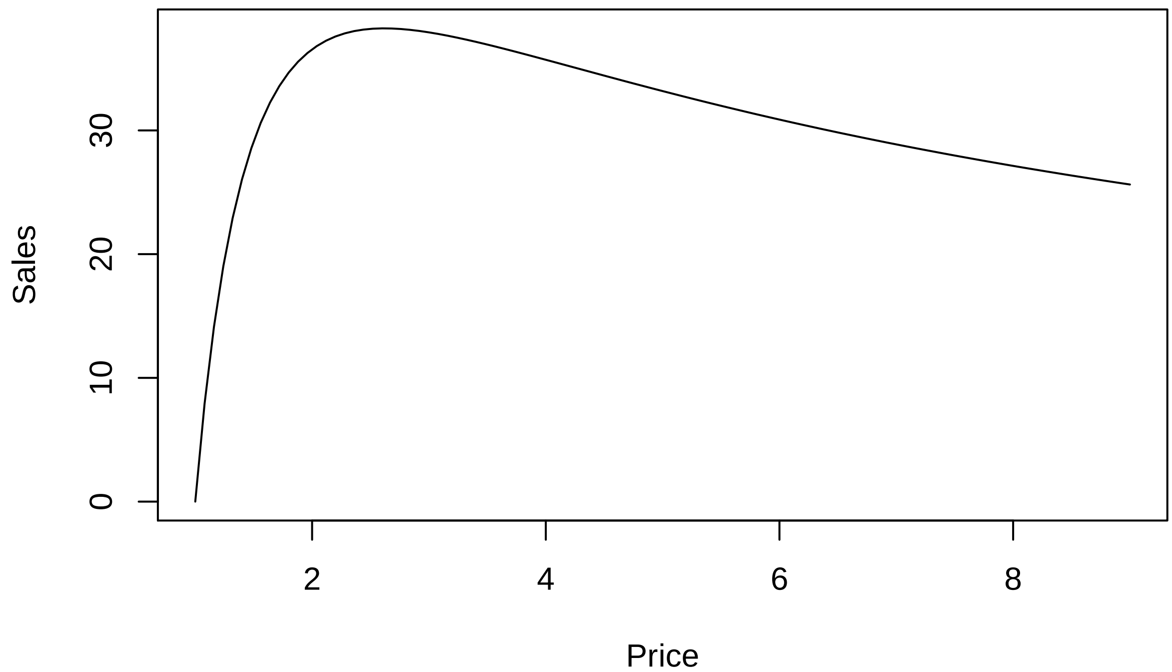


Now we have an estimate of ln(K) in the form of the intercept of the model, 4.72, and of E in the form of the slope of the model, -1.62

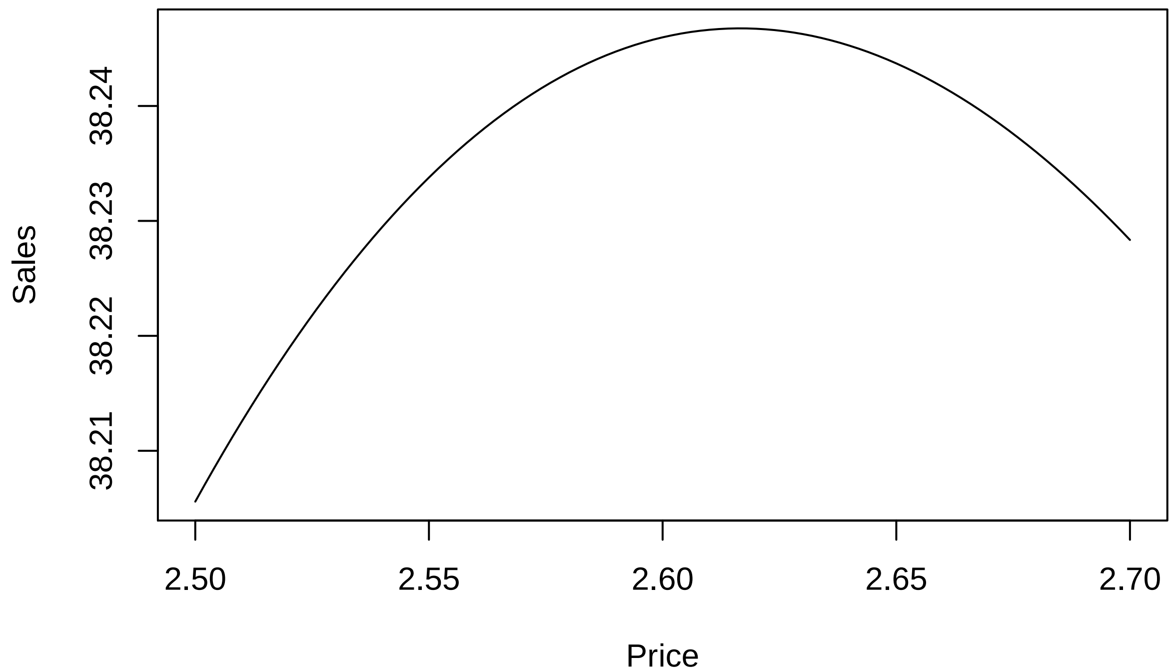
Taking the exponential of both sides gives us net profit in terms of P and c alone, $N \approx (P - c)(112P^{-1.62})$

Let's assume c = 1.

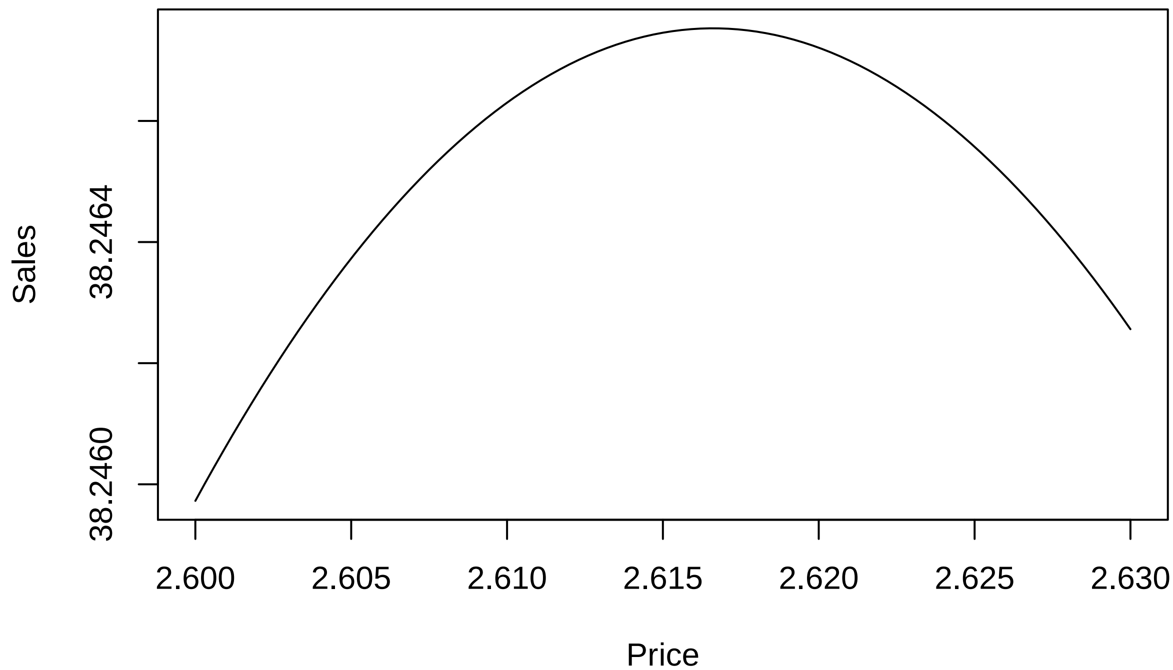
```
x <- milk$price
c <- 1
curve((x - c) * K * x^(E), from = 1, to = 9, xlab = "Price", ylab = "Sales")
```




```r
#Zoom in
curve((x - c) * K * x^(E), from = 2.5, to = 2.7, xlab = "Price", ylab = "Sales")
```




```r
#Zoom in more
curve((x - c) * K * x^(E), from = 2.60, to = 2.63, xlab = "Price", ylab = "Sales")
```



From the final plot, we see that the price that maximizes net profit is close to \$2.62.