170s Midterm

Jun Ryu - UID: 605574052

2023-02-19

```
# Problem 1

data <- data.frame("Class" = c("0-9","10-19","20-29","30-39","40-49","50-59"),
                   "Frequency" = c(13,15,9,14,16,8))
data$midpts <- c(4.5,14.5,24.5,34.5,44.5,54.5)
data
```

```
##   Class Frequency midpts
## 1   0-9        13    4.5
## 2 10-19        15   14.5
## 3 20-29         9   24.5
## 4 30-39        14   34.5
## 5 40-49        16   44.5
## 6 50-59         8   54.5
```

```
mean <- sum(data$Frequency*data$midpts)/sum(data$Frequency)
var <- sum(data$Frequency*(data$midpts - mean)^2)/sum(data$Frequency)
sd <- sqrt(var)

sum(data$Frequency)
```

```
## [1] 75
```

```
#the median is the 38th value since there are a total of 75 values in our dataset

median <- 29.5 + (38-37)*10/14
first_q <- 9.5 + (19-13)*10/15
third_q <- 39.5 + (57-51)*10/16
tenth_per <- (7.5-0)*10/13
IQR <- third_q - first_q

#final answers:
median
```

```
## [1] 30.21429
```

```
mean
```

```
## [1] 28.36667
```

```
sd
```

```
## [1] 16.56449
```

```
tenth_per
```

```
## [1] 5.769231
```

```
first_q
```

```
## [1] 13.5
```

```
IQR
```

```
## [1] 29.75
```

```
# Problem 2

df <- USArrests
head(df)
```

```
##            Murder Assault UrbanPop Rape
## Alabama      13.2     236       58 21.2
## Alaska       10.0     263       48 44.5
## Arizona       8.1     294       80 31.0
## Arkansas      8.8     190       50 19.5
## California    9.0     276       91 40.6
## Colorado      7.9     204       78 38.7
```

```
# part a)
summary(df$Murder)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.800   4.075   7.250   7.788  11.250  17.400
```

```
summary(df$Assault)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    45.0   109.0   159.0   170.8   249.0   337.0
```

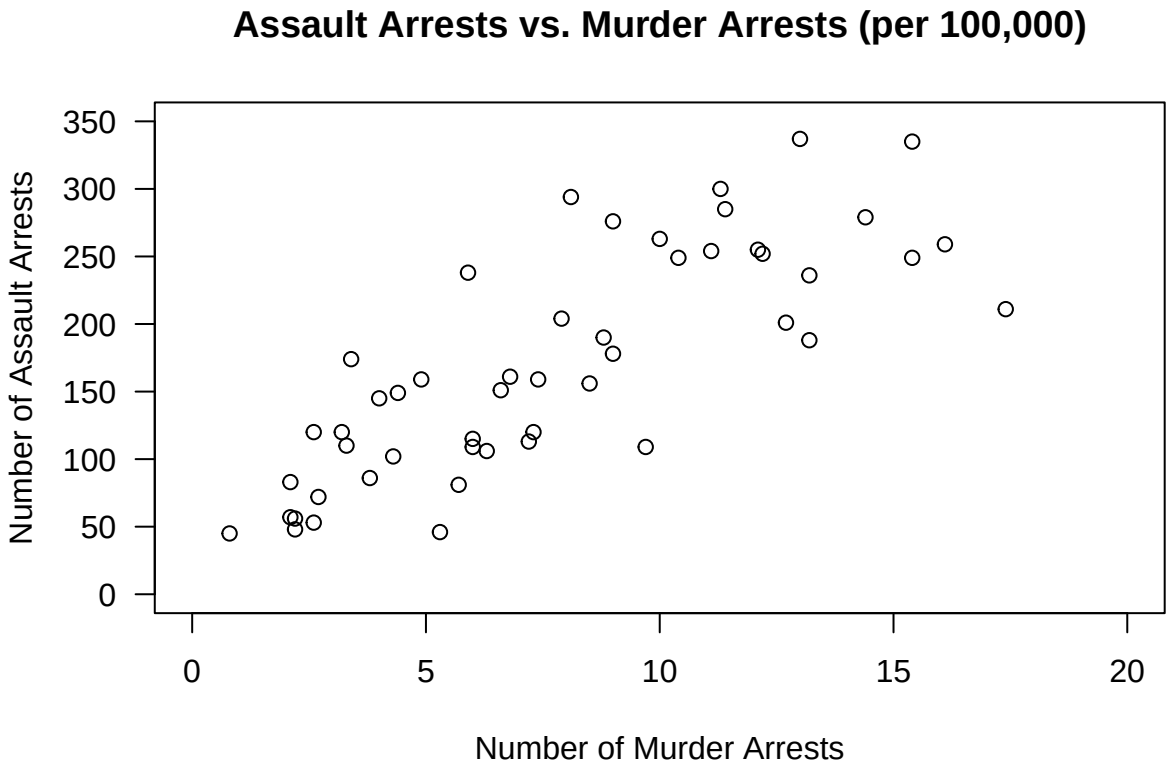
```
summary(df$UrbanPop)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   32.00   54.50   66.00   65.54   77.75   91.00
```

```
summary(df$Rape)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    7.30   15.07   20.10   21.23   26.18   46.00
```

```
# part b)
plot(df$Murder, df$Assault, xlab = "Number of Murder Arrests",
     ylab = "Number of Assault Arrests",
     main = "Assault Arrests vs. Murder Arrests (per 100,000)", xlim = c(0,20),
     ylim = c(0,350), las = 1)
```

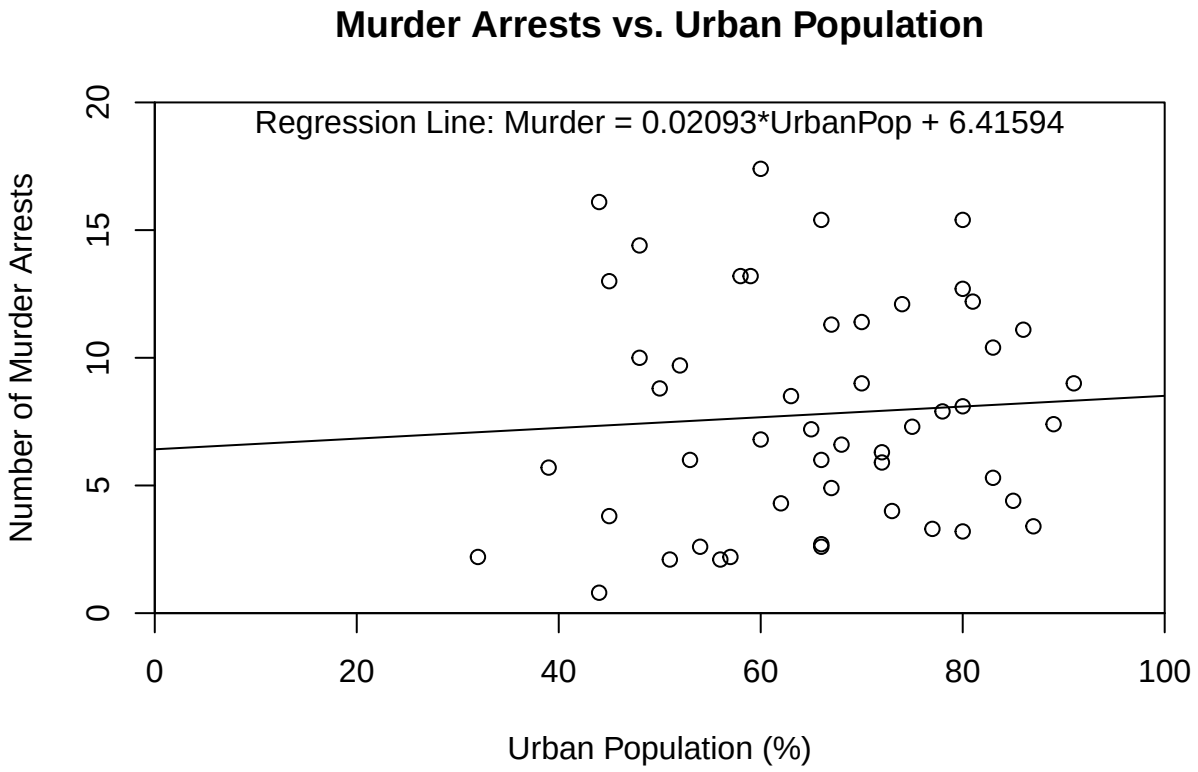


The above scatterplot shows a relatively positive correlation between the two types of arrests. This means that the more number of murder arrests there are, it is also likely to indicate a higher number of assault arrests (per 100,000).

```
#part c)
SLR <- lm(df$Murder ~ df$UrbanPop)
SLR
```

```
##
## Call:
## lm(formula = df$Murder ~ df$UrbanPop)
##
## Coefficients:
## (Intercept)  df$UrbanPop
##     6.41594      0.02093
```

```
plot(df$Murder ~ df$UrbanPop, xlim = c(0,100), ylim = c(0,20),
xlab = "Urban Population (%)", ylab = "Number of Murder Arrests",
xaxs = "i", yaxs = "i", main = "Murder Arrests vs. Urban Population")
abline(SLR)
mtext("Regression Line: Murder = 0.02093*UrbanPop + 6.41594", line = -1)
```



As we notice a lot of plot points are not near our fitted line, the variables show a weak correlation between each other. Nonetheless, the fitted line shows if the urban population percentage becomes greater, a greater number of murder arrests are predicted; however, the rate of increase is not too steep.

3) a) Poisson(λ)

$$P(X=x) = \begin{cases} \frac{\lambda^x e^{-\lambda}}{x!}, & x = 0, 1 \dots \\ 0, & \text{elsewhere} \end{cases}$$

Our likelihood function is:

$$L(\lambda; x_1 \dots x_n) = \prod_{i=1}^{n} \frac{\lambda^{x_i} e^{-\lambda}}{x_i!}$$

Log-likelihood is:

$$\ln(L(\lambda)) = \ln \prod_{i=1}^{n} \lambda^{x_i} + \ln \prod_{i=1}^{n} \frac{1}{x_i!} + \ln \prod_{i=1}^{n} e^{-\lambda}$$

$$= \sum_{i=1}^{n} x_i \ln\lambda + \ln \prod_{i=1}^{n} \frac{1}{x_i!} - \lambda n$$

Now, we take the derivative with respect to our parameter λ and set it equal to 0:

$$\frac{\partial}{\partial\lambda}(\ln(L(\lambda))) = \frac{\sum_{i=1}^{n} x_i}{\lambda} + 0 - n = 0$$

$$\Rightarrow \boxed{\hat{\lambda} = \frac{\sum_{i=1}^{n} x_i}{n} = \bar{x}}$$

b) Normal(μ, σ^2)

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

Our likelihood function:

$$L(\mu, \sigma^2; x_1 \dots x_n) = \prod_{i=1}^{n} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x_i-\mu}{\sigma}\right)^2}$$

$$= \left(\frac{1}{\sigma\sqrt{2\pi}}\right)^n e^{\sum_{i=1}^{n} -\frac{1}{2}\left(\frac{x_i-\mu}{\sigma}\right)^2}$$

$$\ln\left(L(\mu,\sigma^2)\right) = \ln\left(\frac{1}{\sigma\sqrt{2\pi}}\right)^n - \frac{1}{2}\sum_{i=1}^{n}\left(\frac{x_i-\mu}{\sigma}\right)^2$$

We first find the MLE for the parameter μ:

$$\frac{\partial}{\partial\mu}\left(\ln\left(L(\mu,\sigma^2)\right)\right) = \frac{1}{\sigma^2}\sum_{i=1}^{n}(x_i-\mu)$$

$$= \frac{1}{\sigma^2}\left(\sum_{i=1}^{n} x_i - n\mu\right) = 0$$

$$= \boxed{\hat{\mu} = \frac{\sum_{i=1}^{n} x_i}{n} = \bar{x}}$$

Now the parameter σ^2:

But, first, we simplify our log-likelihood formula a bit:

$$\ln\left(L(\mu,\sigma^2)\right) = \ln\left(2\pi\sigma^2\right)^{-n/2} - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2$$

$$= -\frac{n}{2}\ln(2\pi) - \frac{n}{2}\ln(\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2$$

Now, we take the derivative:

$$\frac{\partial}{\partial\sigma^2}\left(\ln\left(L(\mu,\sigma^2)\right)\right) = -\frac{n}{2\sigma^2} - \frac{1}{2}\sum_{i=1}^{n}(x_i-\mu)^2 \cdot \left(-\frac{1}{(\sigma^2)^2}\right)$$

$$= -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4}\sum_{i=1}^{n}(x_i-\mu)^2$$

$$= \frac{1}{2\sigma^2}\left(-n + \frac{1}{\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2\right) = 0$$

$$\Rightarrow \frac{1}{\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2 = n$$

$$= \boxed{\hat{\sigma}^2 = \frac{\sum_{i=1}^{n}(x_i-\hat{\mu})^2}{n}}$$

c) Gamma(α)

$$f(x) = \frac{x^{\alpha-1}e^{-x}}{\Gamma\alpha}, \quad 0<x<\infty$$

Our likelihood function:

$$L(\alpha; x_1 \dots x_n) = \prod_{i=1}^{n}\frac{x_i^{\alpha-1}e^{-x_i}}{\Gamma\alpha}$$

$$\ln(L(\alpha)) = \ln\prod_{i=1}^{n}x_i^{\alpha-1} + \ln\prod_{i=1}^{n}e^{-x_i} - \ln\prod_{i=1}^{n}\Gamma\alpha$$

$$(\alpha-1)\sum_{i=1}^{n}\ln x_i - \sum_{i=1}^{n}x_i - n\ln\Gamma\alpha$$

$$\frac{\partial}{\partial\alpha}\left(\ln(L(\alpha))\right) = \sum_{i=1}^{n}\ln x_i - n\frac{\Gamma'(\alpha)}{\Gamma(\alpha)}$$

$$= \sum_{i=1}^{n}\ln x_i - n\psi(\alpha)$$

ψ ← Digamma function

$$= n\left(\overline{\ln x} - \psi(\alpha)\right) = 0$$

$$\Rightarrow \boxed{\psi(\hat{\alpha}) = \overline{\ln x}}$$

We arrive here, but from here, we would have to use approximation methods to predict the MLE of α.

d) Binomial (p)

$$P(X=x) = \begin{cases} \binom{n}{x} p^x (1-p)^{n-x}, & x = 0, 1 \ldots \\ 0, & \text{elsewhere} \end{cases}$$

$$L(p; x_1 \ldots x_m) = \prod_{i=1}^{m} \binom{n}{x_i} p^{x_i} (1-p)^{n-x_i}$$

$$= \prod_{i=1}^{m} \binom{n}{x_i} \cdot p^{\sum_{i=1}^{m} x_i} (1-p)^{\sum_{i=1}^{m} n-x_i}$$

$$\ln(L(p)) = \ln\left(\prod_{i=1}^{m} \binom{n}{x_i}\right) + \sum_{i=1}^{m} x_i \ln p + \sum_{i=1}^{m} (n-x_i) \ln(1-p)$$

$$\frac{\partial}{\partial p}(\ln(L(p))) = 0 + \frac{\sum_{i=1}^{m} x_i}{p} - \frac{\sum_{i=1}^{m} (n-x_i)}{1-p} = 0$$

$$\Rightarrow \frac{1-p}{p} = \frac{nm - \sum_{i=1}^{m} x_i}{\sum_{i=1}^{m} x_i}$$

$$\Rightarrow \frac{1}{p} - \cancel{1} = \frac{nm}{\sum_{i=1}^{n} x_i} - \cancel{1}$$

$$\Rightarrow \boxed{\hat{p} = \frac{\overline{x_i}}{n}}$$

e) Geometric (p)

$$P(X=x) = \begin{cases} (1-p)^{x-1} p & , x = 0, 1 \dots \\ 0 & , \text{elsewhere} \end{cases}$$

$$L(p; x_1 \dots x_n) = \prod_{i=1}^{n} (1-p)^{x_i - 1} \cdot p$$

$$= (1-p)^{\sum_{i=1}^{n} (x_i - 1)} \cdot p^n$$

$$\ln(L(p)) = \sum_{i=1}^{n} (x_i - 1) \ln(1-p) + n \ln p$$

$$\frac{\partial}{\partial p} (\ln(L(p))) = \frac{n}{p} - \frac{\sum_{i=1}^{n} (x_i - 1)}{1-p} = 0$$

$$\Rightarrow \frac{1-p}{p} = \frac{\sum_{i=1}^{n} x_i - n}{n}$$

$$\Rightarrow \frac{1}{p} - \cancel{1} = \overline{x} - \cancel{1} \Rightarrow \boxed{\hat{p} = \frac{1}{\overline{x}}}$$

f) Exponential (λ)

$$f(x) = \lambda e^{-\lambda x} , \quad x \geq 0$$

$$L(\lambda; x_1 \dots x_n) = \prod_{i=1}^{n} \lambda e^{-\lambda x_i}$$

$$= \lambda^n e^{-\lambda \sum_{i=1}^{n} x_i}$$

$$\ln(L(\lambda)) = n \ln \lambda - \lambda \sum_{i=1}^{n} x_i$$

$$\frac{\partial}{\partial \lambda} \left(\ln(L(\lambda)) \right) = \frac{n}{\lambda} - \sum_{i=1}^{n} x_i = 0$$

$$\Rightarrow \boxed{\hat{\lambda} = \frac{n}{\sum_{i=1}^{n} x_i} = \frac{1}{\bar{x}}}$$

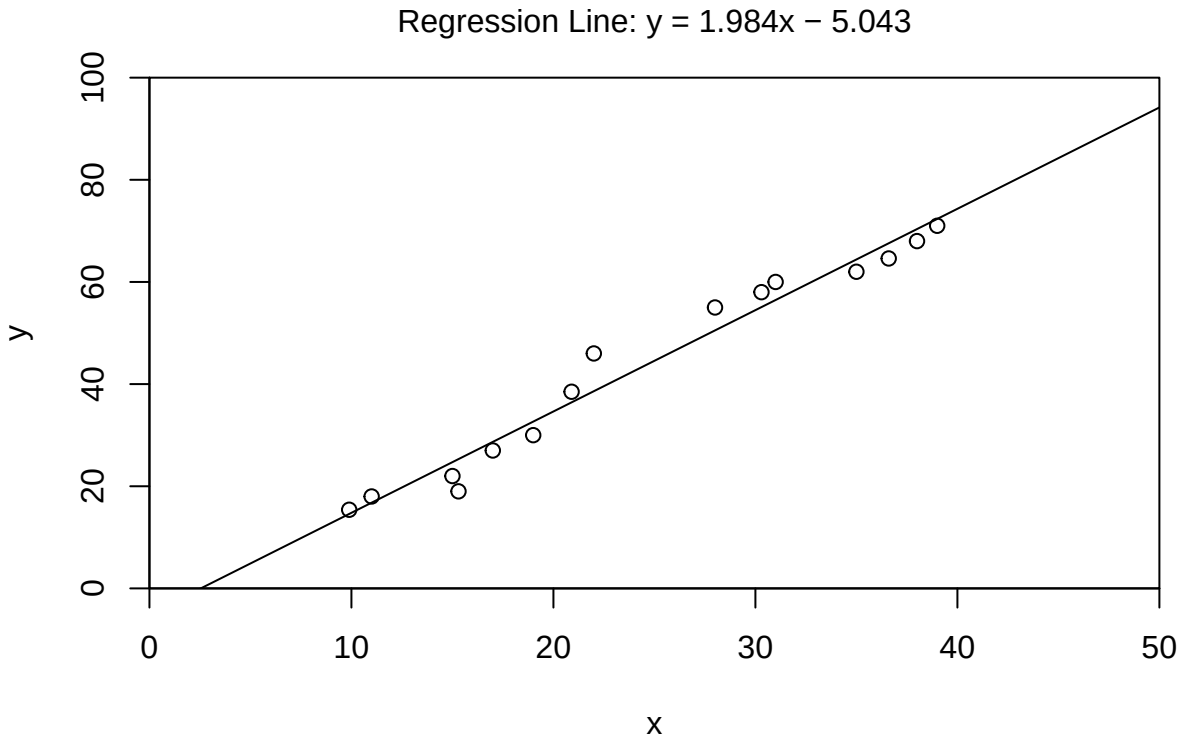
```r
# Problem 4

# part a)
x = c(9.89, 11, 15.3, 15, 17, 19, 20.9, 22, 28, 30.3, 31, 35, 36.6, 38, 39)
y = c(15.4, 18, 19, 22, 27, 30, 38.5, 46, 55, 58, 60, 62, 64.6, 68, 71)
df2 <- data.frame(x,y)

SLR2 <- lm(y ~ x, data = df2)
SLR2
```

```
##
## Call:
## lm(formula = y ~ x, data = df2)
##
## Coefficients:
## (Intercept)            x
##      -5.043        1.984
```

```r
plot(y ~ x, data = df2, xlim = c(0,50), ylim = c(0,100), xaxs = "i", yaxs = "i")
abline(SLR2)
mtext("Regression Line: y = 1.984x - 5.043", line = 1)
```




```r
# part b)
cor(x, y)^2
```

```
## [1] 0.9684332
```

The R-squared value (the coefficient of determination) shows a value extremely close to 1, which indicates that, given this simple regression model with the given data, the model fits the data really well and is able to make accurate predictions. Also, it indicates that the data shows a good correlation between its explanatory and response variables.

```r
# part c)
predict_data <- data.frame(x = c(16,25))
predict(SLR2, predict_data)
```

```
##        1        2
## 26.70313 44.56060
```

```r
# Problem 5

# part a)
library(MASS)
head(Boston)
```

```
##      crim zn indus chas   nox    rm  age    dis rad tax ptratio  black lstat
## 1 0.00632 18  2.31    0 0.538 6.575 65.2 4.0900   1 296    15.3 396.90  4.98
## 2 0.02731  0  7.07    0 0.469 6.421 78.9 4.9671   2 242    17.8 396.90  9.14
## 3 0.02729  0  7.07    0 0.469 7.185 61.1 4.9671   2 242    17.8 392.83  4.03
## 4 0.03237  0  2.18    0 0.458 6.998 45.8 6.0622   3 222    18.7 394.63  2.94
## 5 0.06905  0  2.18    0 0.458 7.147 54.2 6.0622   3 222    18.7 396.90  5.33
## 6 0.02985  0  2.18    0 0.458 6.430 58.7 6.0622   3 222    18.7 394.12  5.21
##   medv
## 1 24.0
## 2 21.6
## 3 34.7
## 4 33.4
## 5 36.2
## 6 28.7
```

```r
mean(Boston$crim)
```

```
## [1] 3.613524
```

```r
mean(Boston$rad)
```

```
## [1] 9.549407
```

```r
median(Boston$crim)
```

```
## [1] 0.25651
```

```r
median(Boston$rad)
```

```
## [1] 5
```

```r
sd(Boston$crim)
```

```
## [1] 8.601545
```

```r
sd(Boston$rad)
```

```
## [1] 8.707259
```

```r
IQR(Boston$crim)
```

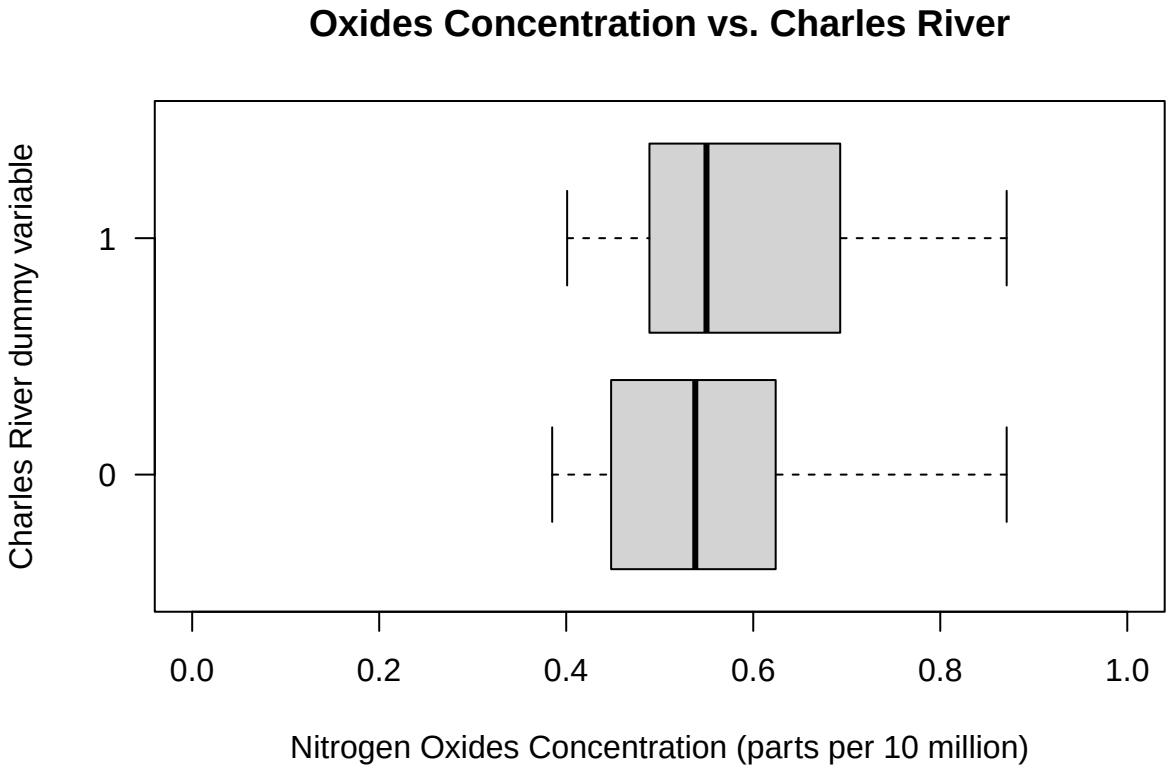
```
## [1] 3.595038
```

```r
IQR(Boston$rad)
```

```
## [1] 20
```

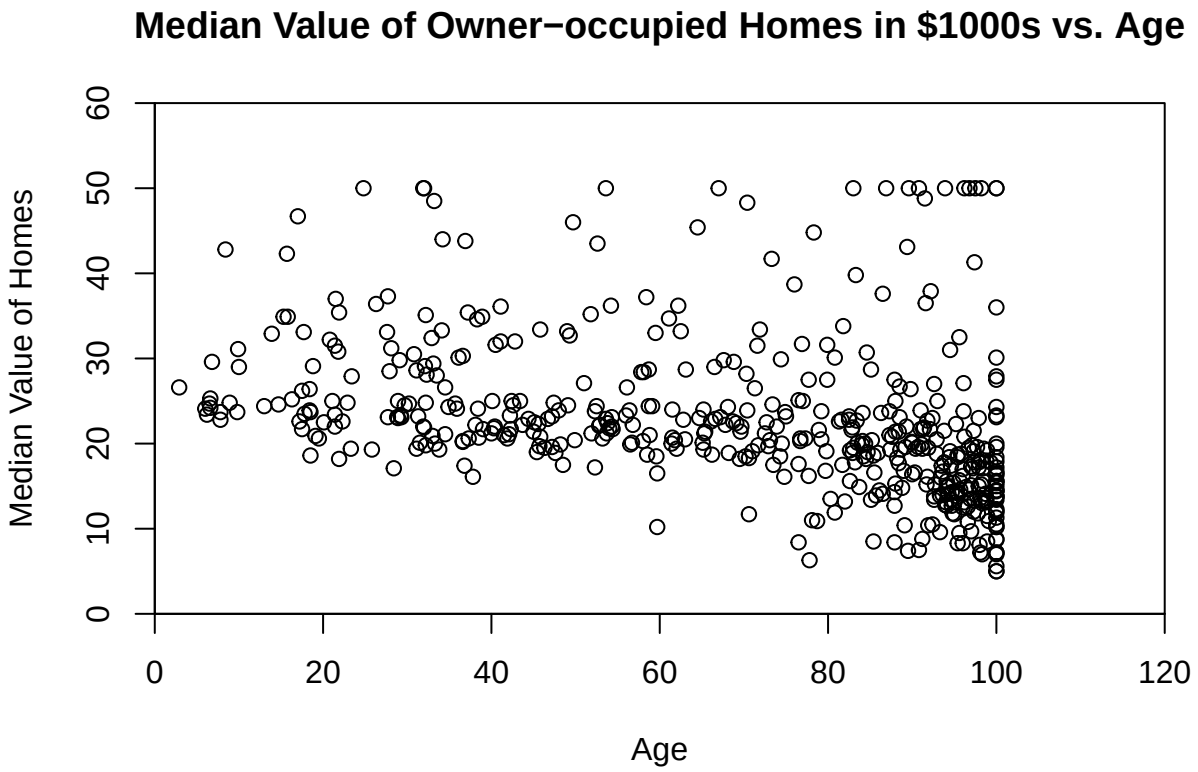
```r
# part b)
boxplot(Boston$nox ~ Boston$chas, horizontal = T,
```

```
        xlab = "Nitrogen Oxides Concentration (parts per 10 million)",
        ylab = "Charles River dummy variable", las = 1, ylim = c(0,1),
        main = "Oxides Concentration vs. Charles River")
```



Looking at the above plot, there does not seem to be much difference between the two conditions of our explanatory variables (1 = tract bounds river, 0 = otherwise). The minimum, the median, and the maximum look closely aligned between the two conditions. However, the 1st and the 3rd quartiles seem to be greater if tract bounds river.

```
# part c)
plot(medv ~ age, data = Boston, xlim = c(0,120), ylim = c(0,60),
xlab = "Age", ylab = "Median Value of Homes",
xaxs = "i", yaxs = "i", main = "Median Value of Owner-occupied Homes in $1000s vs. Age")
```



The above scatterplot doesn't really show an easily identifiable correlation between the two variables as the plot points are scattered all throughout the plot. However, we do notice a large cluster of points near the bottom right of our plot, indicating that as people near the last stages of life, their median value of homes is a bit lower than that of the other age groups.

```
# part d)
c(mean(Boston$crim) - qnorm(0.975)*sd(Boston$crim)/sqrt(length(Boston$crim)),
  mean(Boston$crim) + qnorm(0.975)*sd(Boston$crim)/sqrt(length(Boston$crim)))
```

```
## [1] 2.864062 4.362985
```

```
c(mean(Boston$tax) - qnorm(0.975)*sd(Boston$tax)/sqrt(length(Boston$tax)),
  mean(Boston$tax) + qnorm(0.975)*sd(Boston$tax)/sqrt(length(Boston$tax)))
```

```
## [1] 393.5523 422.9220
```

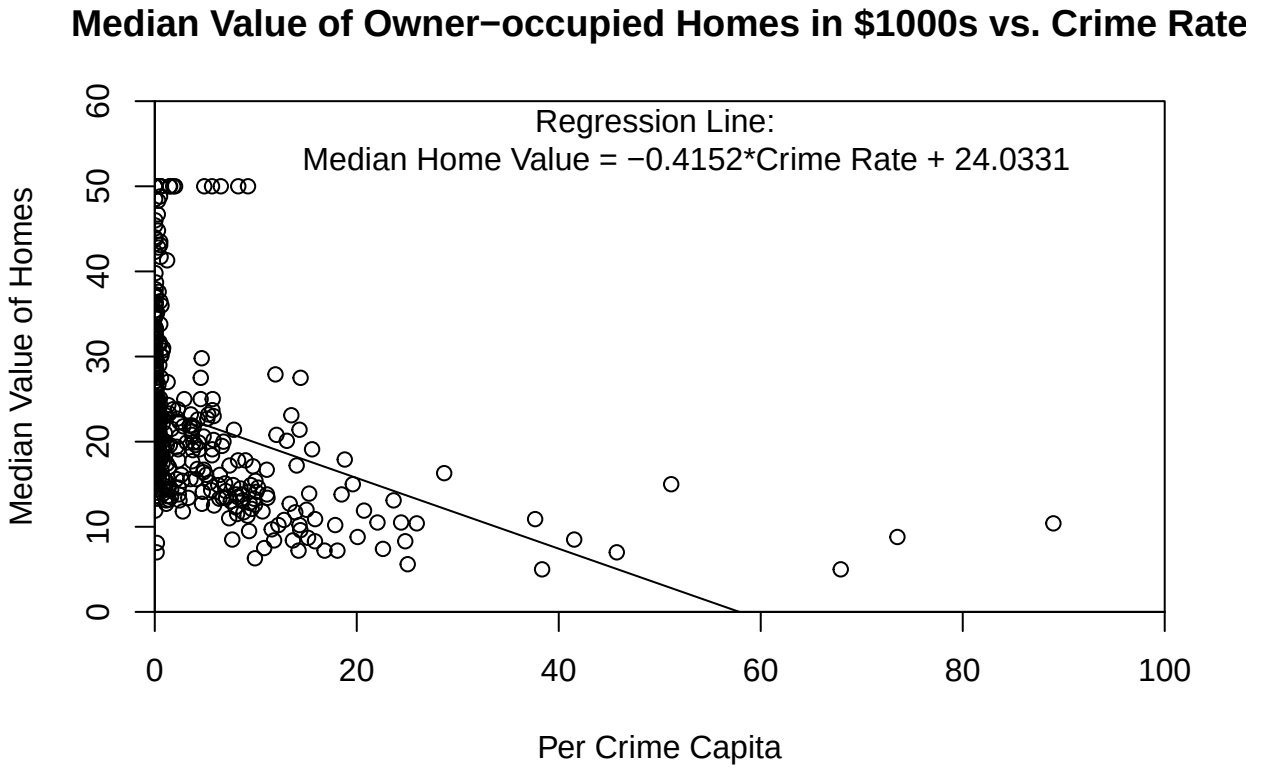
```
# part e)

SLR3 <- lm(Boston$medv ~ Boston$crim)
SLR3
```

```
##
## Call:
## lm(formula = Boston$medv ~ Boston$crim)
##
## Coefficients:
## (Intercept)  Boston$crim
##     24.0331      -0.4152
```

```
plot(Boston$medv ~ Boston$crim, xlim = c(0,100), ylim = c(0,60),
xlab = "Per Crime Capita", ylab = "Median Value of Homes", xaxs = "i",
yaxs = "i", main = "Median Value of Owner-occupied Homes in $1000s vs. Crime Rate")
```

```
abline(SLR3)
mtext("Regression Line:
      Median Home Value = -0.4152*Crime Rate + 24.0331", line = -2)
```




```
cor(Boston$crim, Boston$medv)
```

```
## [1] -0.3883046
```

```
cor(Boston$crim, Boston$medv)^2
```

```
## [1] 0.1507805
```

With a R-squared value of 0.15, this tells us that the simple linear regression (SLR) model does not fit our data very well and is not able to make accurate predictions. Observing the scatterplot, there are too many values that are vertically scattered on the left side of the plot and not close to the SLR line. Regardless, the correlation coefficient and the slope of the SLR line shows a negative correlation, which means that the higher the crime rate, the lower the median value of homes will become.

6) We have $X \sim \text{Normal}(\mu, \sigma_0^2)$ with $X = \{x_1 \dots x_n\}$ i.i.d. and prior distribution is $M \sim \text{Normal}(\mu_0, \rho_0^2)$.

We use the concept of proportionality ($\propto$):

we have $g(x_i | \mu, \sigma_0^2) \propto \kappa e^{-\frac{(x_i - \mu)^2}{2\sigma_0^2}}$,

where κ is just a normalizing constant.

So, $g(x | \mu, \sigma_0^2) = g(x_1 | \mu, \sigma_0^2) \cdot g(x_2 | \mu, \sigma_0^2) \dots$

$$\propto \kappa e^{-\sum_{i=1}^{n} (x_i - \mu)^2 / 2\sigma_0^2}$$

So, our posterior pdf is:

$$Pr_M(\mu | x_1 \dots x_n) = g(x | \mu, \sigma_0^2) \cdot h(\mu | \mu_0, \rho_0^2)$$

$$\propto \kappa e^{-\sum_{i=1}^{n} (x_i - \mu)^2 / 2\sigma_0^2 - (\mu - \mu_0)^2 / 2\rho_0^2}$$

Now, we just need to simplify the exponent to the form of $-(\mu - c)^2 / 2\tau^2$:

$$-\frac{1}{2}\left(\frac{\sum_{i=1}^{n} (x_i - \mu)^2}{\sigma_0^2} + \frac{(\mu - \mu_0)^2}{\rho_0^2} \right)$$

$$= -\frac{1}{2}\left(\frac{\sum_{i=1}^{n} x_i^2 - 2n\bar{x}\mu + n\mu^2}{\sigma_0^2} + \frac{\mu^2 - 2\mu\mu_0 + \mu_0^2}{\rho_0^2} \right)$$

From here, we can drop all terms without μ (because those are part of our normalizing constant):

$$-\frac{1}{2}\left(\frac{n\mu^2\rho_0^2 - 2n\bar{x}\mu\rho_0^2 + \mu^2\sigma_0^2 - 2\mu\mu_0\sigma_0^2}{\sigma_0^2\rho_0^2}\right)$$

$$= -\frac{1}{2}\left(\frac{(n\rho_0^2 + \sigma_0^2)\mu^2 - 2(n\bar{x}\rho_0^2 + \mu_0\sigma_0^2)\mu}{\sigma_0^2\rho_0^2}\right)$$

$$= -\frac{1}{2}\left(\frac{\mu^2 - 2\mu\,\frac{(n\bar{x}\rho_0^2 + \mu_0\sigma_0^2)}{(n\rho_0^2 + \sigma_0^2)}}{\frac{\sigma_0^2\rho_0^2}{n\rho_0^2 + \sigma_0^2}}\right)$$

Now, we use the complete the square method by bringing in a constant term (not involving μ) that will help us complete the square:

$$= -\frac{1}{2}\left(\frac{\left(\mu - \frac{(n\bar{x}\rho_0^2 + \mu_0\sigma_0^2)}{(n\rho_0^2 + \sigma_0^2)}\right)^2}{\frac{\sigma_0^2\rho_0^2}{n\rho_0^2 + \sigma_0^2}}\right)$$

Now, we have this in the form that we want: $-(\mu-c)^2/2\tau^2$

we have $\tau^2 = \dfrac{\sigma_0^2\rho_0^2}{n\rho_0^2+\sigma_0^2}$ and

$$c = \frac{\rho_0^2 \sum_{i=1}^{n} x_i + \mu_0\sigma_0^2}{n\rho_0^2+\sigma_0^2}$$

Since our posterior pdf of $ke^{-(\mu-c)^2/2\tau^2}$ is consistent with the normal distribution formula, we conclude that our posterior distribution is Normal (c, τ^2). ✓

```r
#problem 7
difference <- 13.75 - 14.50
c(difference - qnorm(0.995)*sqrt(5.2^2/40 + 3.98^2/49),
  difference + qnorm(0.995)*sqrt(5.2^2/40 + 3.98^2/49))
```

```
## [1] -3.324893  1.824893
```

From the values above, we can conclude that the mean price of coffee in big cities is likely to be more expensive as our 99% confidence interval has more values within the negative. Of course, we cannot state this with certainty as our interval dives into the positive as well (indicating a case where the median price in small towns is greater).