Directions.

The purpose of homework is three-fold: (1) to provide you with an additional opportunity for learning the material; (2) to provide you with exercises to practice applying the concepts; (3) and to check your understanding of the material. Problems that are marked with a $\star$ should be submitted for a grade. All others are strongly recommended but not required.

3. $\star$ Show that:

   (a) $\frac{\partial Q}{\partial \beta_0} = 0 \rightarrow nb_0 + \sum_i X_i b_1 = \sum_i Y_i$

$$\begin{aligned}
\frac{\partial Q}{\partial \beta_0} &= \frac{\partial}{\partial \beta_0} \sum_i (Y_i - \beta_0 - \beta_1 X_i)^2 \\
&= 2 \sum_i (Y_i - \beta_0 - \beta_1 X_i)(-1) \\
\rightarrow 0 &= -2 \sum_i (Y_i - b_0 - b_1 X_i) \\
\rightarrow 0 &= \sum_i Y_i - nb_0 - b_1 \sum_i X_i \\
\rightarrow \sum_i Y_i &= nb_0 + b_1 \sum_i X_i
\end{aligned}$$

   (b) $\frac{\partial Q}{\partial \beta_1} = 0 \rightarrow \sum_i X_i b_0 + \sum_i X_i^2 b_1 = \sum_i X_i Y_i$

$$\begin{aligned}
\frac{\partial Q}{\partial \beta_1} &= \frac{\partial}{\partial \beta_1} \sum_i (Y_i - \beta_0 - \beta_1 X_i)^2 \\
&= 2 \sum_i (Y_i - \beta_0 - \beta_1 X_i)(-X_i) \\
\rightarrow 0 &= -2 \sum_i X_i (Y_i - b_0 - b_1 X_i) \\
\rightarrow 0 &= \sum_i X_i Y_i - b_0 \sum_i X_i - b_1 \sum_i X_i^2 \\
\rightarrow \sum_i X_i Y_i &= b_0 \sum_i X_i + b_1 \sum_i X_i^2
\end{aligned}$$

4. $\star$ (Continued from 3.) Show that:

   (a) $b_0 = \bar{Y} - b_1 \bar{X}$

$$\begin{aligned}
\sum_i Y_i &= nb_0 + b_1 \sum_i X_i \\
\rightarrow b_0 &= \bar{Y} - b_1 \bar{X}
\end{aligned}$$

(b) $b_1 = \frac{SSXY}{SSX} = \frac{\sum_i (X_i - \bar{X})(Y_i - \bar{Y})}{\sum_i (X_i - \bar{X})^2}$

$$\begin{aligned}
\sum_i X_i Y_i &= b_0 \sum_i X_i + b_1 \sum_i X_i^2 \\
\rightarrow \sum_i X_i Y_i &= (\bar{Y} - b_1 \bar{X}) \sum_i X_i + b_1 \sum_i X_i^2 \\
\rightarrow b_1 \big(\sum_i X_i^2 - \sum_i X_i \bar{X} \big) &= \sum_i X_i Y_i - \sum_i X_i \bar{Y} \\
\rightarrow b_1 &= \frac{\sum_i X_i (Y_i - \bar{Y})}{\sum_i X_i (X_i - \bar{X})} \\
\rightarrow b_1 &= \frac{SSXY}{SSX}
\end{aligned}$$

The last step follows from the property $\sum_i (X_i - \bar{X}) = 0$:

$$\begin{aligned}
\sum_i (X_i - \bar{X}) &= \sum_i X_i - \sum_i \bar{X} \\
&= \frac{n}{n} \sum_i X_i - \frac{n^2}{n} \bar{X} \\
&= n\bar{X} - n\bar{X} = 0
\end{aligned}$$

so it follows that:

$$\begin{aligned}
SSX &= \sum_i (X_i - \bar{X})^2 \\
&= \sum_i (X_i - \bar{X}) X_i - \sum_i (X_i - \bar{X}) \bar{X} \\
&= \sum_i (X_i - \bar{X}) X_i - \bar{X} \sum_i (X_i - \bar{X}) \\
&= \sum_i (X_i - \bar{X}) X_i
\end{aligned}$$

Select written answers from Kutner.

1.12 See p. 34 for problem description.

a. The data were not the result of a statistical experiment since the investigator did not exert control over the explanatory variable through random assignment to study participants. The explanatory variable, weekly time spent exercising, was only observed for participants.

b. This observational study cannot conclude a causal relationship between exercise and frequency of colds in senior citizens because randomization was not used for the explanatory variable. The purpose of randomization is to balance out the effects of any other variables that might affect the frequency of colds in senior citizens (see (c.)).

c. Factors that could be alternative explanations for the relationship include: (1) climate or season (individuals in colder climates may have more frequent colds because of more time spent indoors and dryer air, and those same individuals may not exercise as much because of limited outdoor activity), or (2) some other measures of overall health status (individuals with other health conditions might be more prone to colds than others, and those same individuals may not exercise as much).

d. If participants had been randomly assigned to an amount of weekly time to spend on exercising, the study would have been experimental and conclude cause-and-effect.

1.13 See p. 34 for problem description.

a. The data were not the result of a statistical experiment since the investigator did not exert control over the explanatory variable through random assignment to study participants. The explanatory variable, time spent preparing, was only observed for participants.

b. This observational study cannot conclude a causal relationship between preparation time and productivity because randomization was not used for the explanatory variable. The purpose of randomization is to balance out the effects of any other variables that might affect productivity (see (c.)).

c. Factors that could be alternative explanations for the relationship include: (1) drive or motivation (individuals with more motivation to learn may also be more driven to produce) and (2) other job responsibilities (individuals in management, for example, may have more administrative job responsibilities that take away time from programming and class preparation).

d. If participants had been randomly assigned to an amount of preparation time, the study would have been experimental and conclude cause-and-effect.

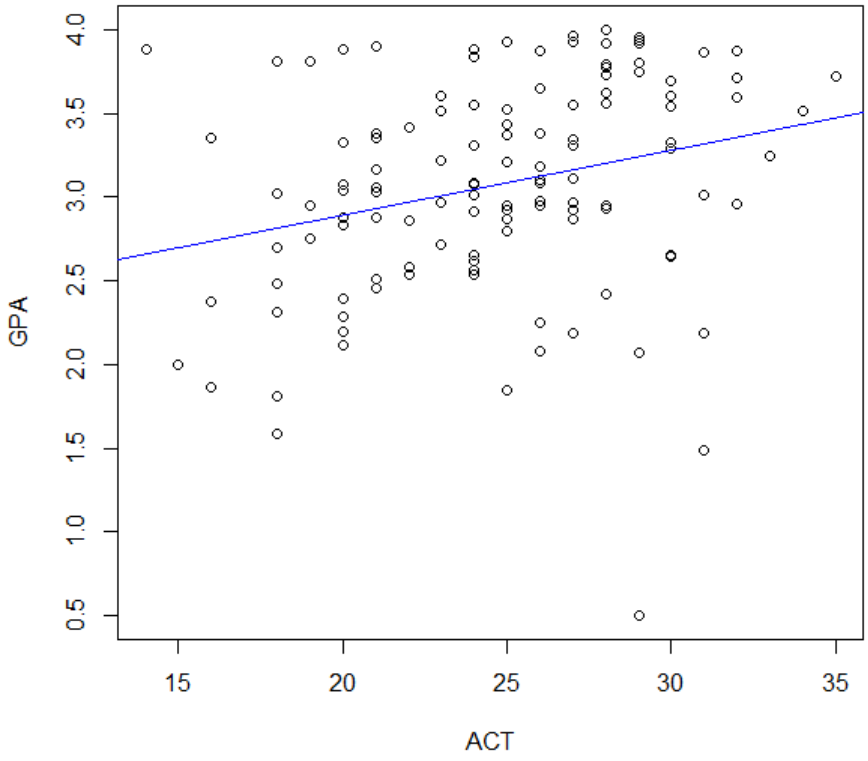
1.16 Least squares estimation does not require normality of the errors, so Y is not required to be normally distributed.

1.17 This is a matter of proper terminology–we only estimate unknown parameters like β_0 and β_1. b_0 and b_1 are random variables/estimators that, when applied to data, produce numeric summaries of the data.

1.19 See p. 35 for problem description. R code can be found in the Appendix.

a. The least squares estimates are $b_0 = 2.11$ and $b_1 = 0.039$, with estimated regression function $\hat{Y}_i = 2.11 + 0.039X_i$.

b. The estimated regression function does not appear to fit the data well. It appears that the slope of the fitted line is smaller than it should be. There are also observations that the model doesn't explain well (e.g., see observations with high ACT and low GPA).

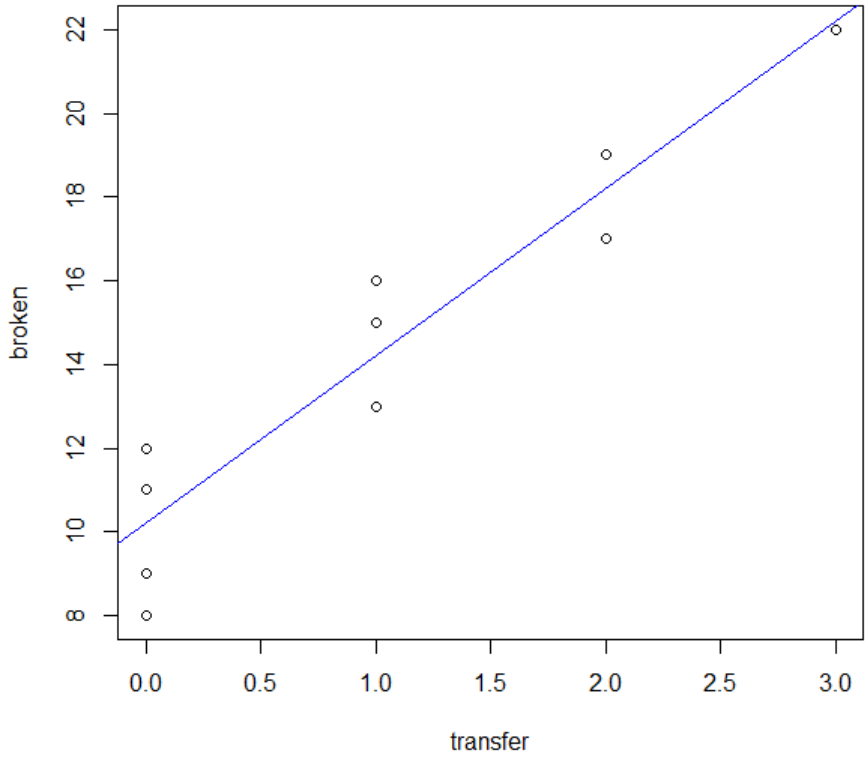


c. The estimate of the mean freshman GPA for students with ACT test score $X_h = 30$ is $\hat{Y}_h = 3.28$.

d. The estimate of the change in mean freshman GPA for a one point increase in ACT score is $b_1 = 0.039$ points.

1.21 See p. 35 for problem description. R code can be found in the Appendix.

a. The estimated regression function is $\hat{Y}_i = 10.2 + 4.0X_i$. The fitted line does appear to do a good job estimating the mean of Y across different values of X.



b. The estimate of the expected number of broken ampules in $X_h = 1$ transfers is $\hat{Y}_h = 14.2$.

c. The estimate of the increase in the expected number of broken ampules for a one-unit increase in X is $b_1 = 4$.

d. When $X_h = \bar{X} = 1$, $\hat{Y}_h = \bar{Y} = 14.2$.

1.45 See p. 39 for problem description. R code can be found in the Appendix.

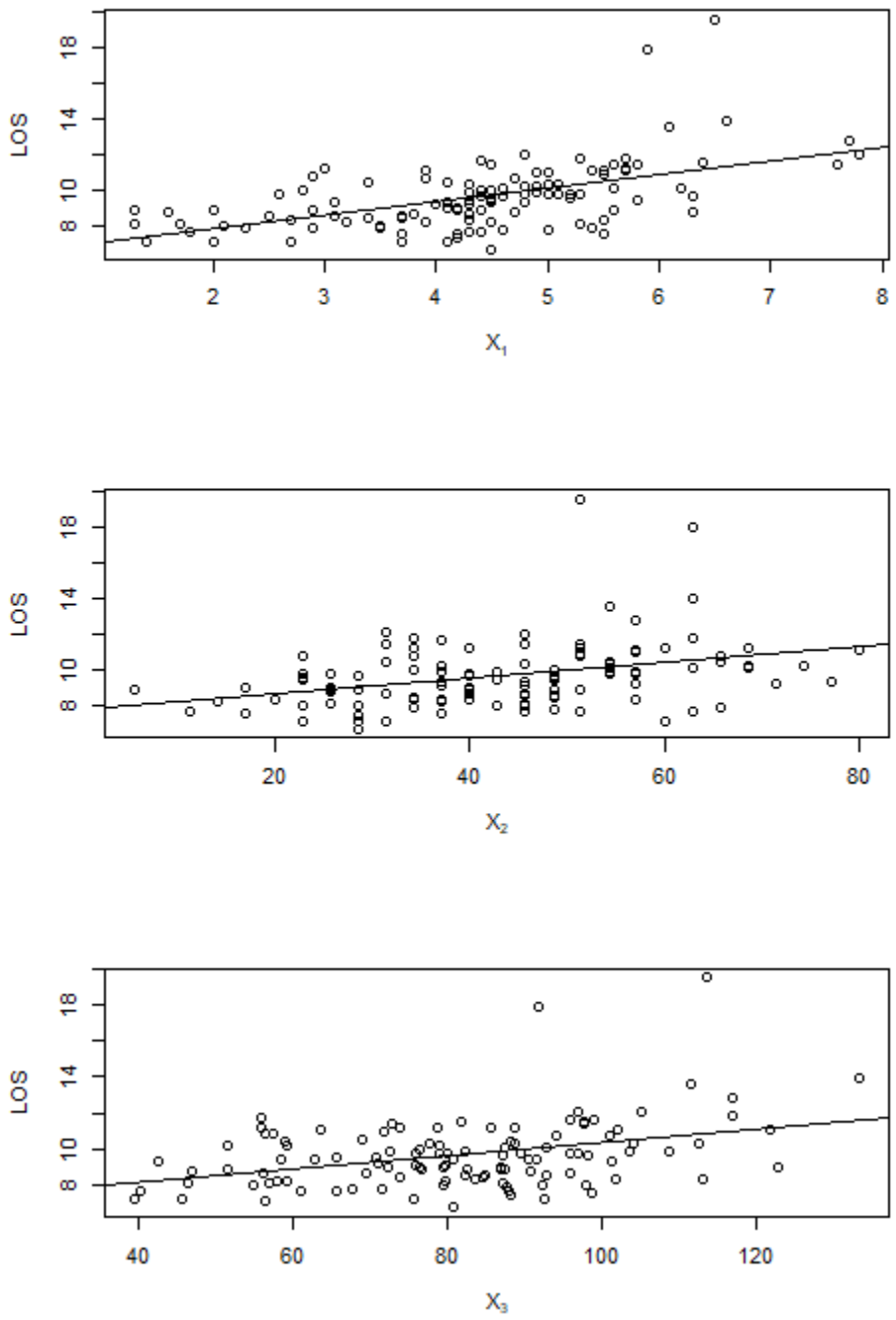
a. The estimated regression functions for the three simple linear regression models for $Y_i =$ Length of stay are:

$$\hat{Y}_i = 6.34 + 0.76X_{i1} \text{ where } X_{i1} = \text{ Infection Risk}$$

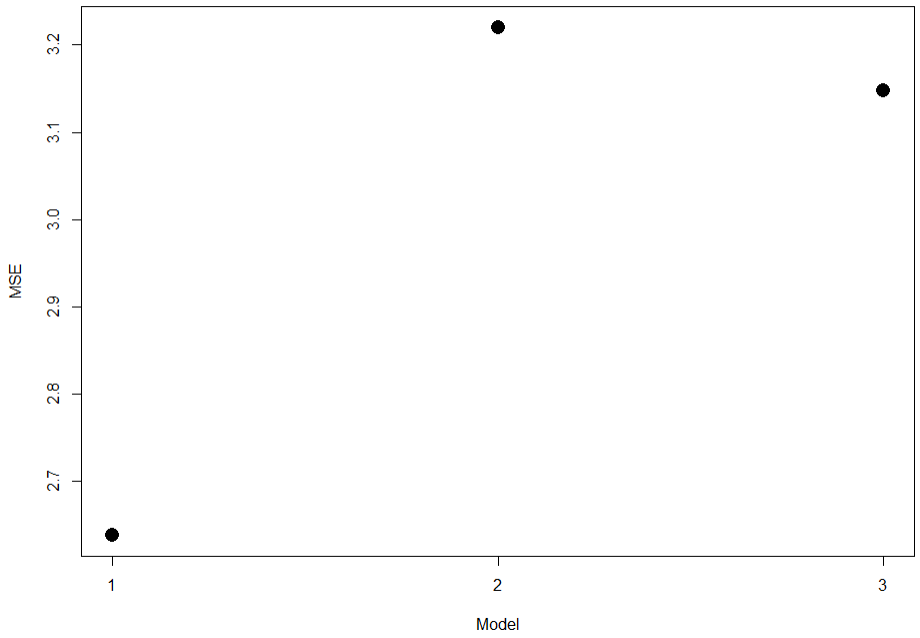
$$\hat{Y}_i = 7.72 + 0.05X_{i2} \text{ where } X_{i2} = \text{ Available facilities and services}$$

$$\hat{Y}_i = 6.57 + 0.04X_{i3} \text{ where } X_{i3} = \text{ Routine chest X-ray ratio}$$

b. A linear relation appears to provide a good fit from each of the three predictor variables, though there are a few observations that the line does not fit well (see unusually large values of length of stay).



c. Model 1 with X_{i1} = Infection Risk has the lowest amount of unexplained variation, with $MSE = 2.64$. Models 2 and 3 have higher amounts of unexplained variation, with $MSE = 3.15$ for X_{i3} = Routine chest X-ray ratio and $MSE = 3.22$ for X_{i2} = Available facilities and services.


MSE
2.7
2.8
2.9
3.0
3.1
3.2
1
2
3
Model

Homework #2 R Code

Jo Wick

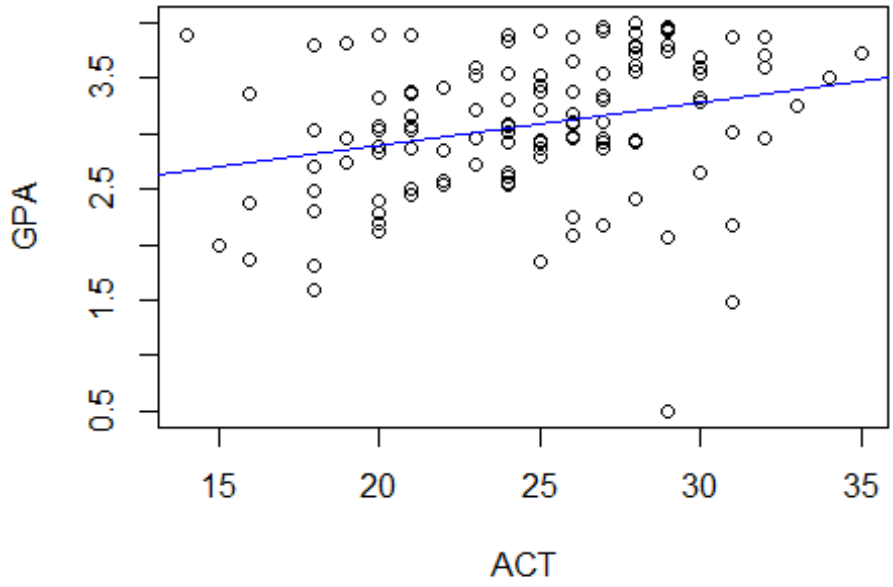
9/8/2021

Problem 1.19

$X = \text{ACT}\ Y = \text{GPA}$

```
# Load data in to R
library(readxl)
CH01PR19 <- read_excel("G:/Teaching/BIOS 840/Fall
2017/Datasets/CH01PR19.xlsx", col_names = FALSE)

# Clean the data
names(CH01PR19) <- c("GPA","ACT")

# GPA = Beta_0 + Beta_1 * ACT + epsilon
# GPA ~ ACT

# A. Obtain b0 and b1
# b0 = 2.11405 estimates beta_0
# b1 = 0.03883 estimates beta_1
m1 <- lm(GPA~ACT,data=CH01PR19)
summary(m1)

##
## Call:
## lm(formula = GPA ~ ACT, data = CH01PR19)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.74004 -0.33827  0.04062  0.44064  1.22737
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.11405    0.32089   6.588  1.3e-09 ***
## ACT          0.03883    0.01277   3.040  0.00292 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6231 on 118 degrees of freedom
## Multiple R-squared:  0.07262,    Adjusted R-squared:  0.06476
## F-statistic:  9.24 on 1 and 118 DF,  p-value: 0.002917

# B. Plot estimated regression function and data
# Plot to make sure a linear relationship makes sense
```

```r
plot(GPA~ACT,data=CH01PR19)
abline(m1,col="blue")
```




```r
# C. Estimate mean GPA for ACT = 30
(Y.hat <- predict(m1,data.frame(ACT=30)))
```

```
##        1
## 3.278863
```

```r
# D. Point estimate of change in mean response when entrance test score
increases by 1 point?
coef(m1)[2]
```

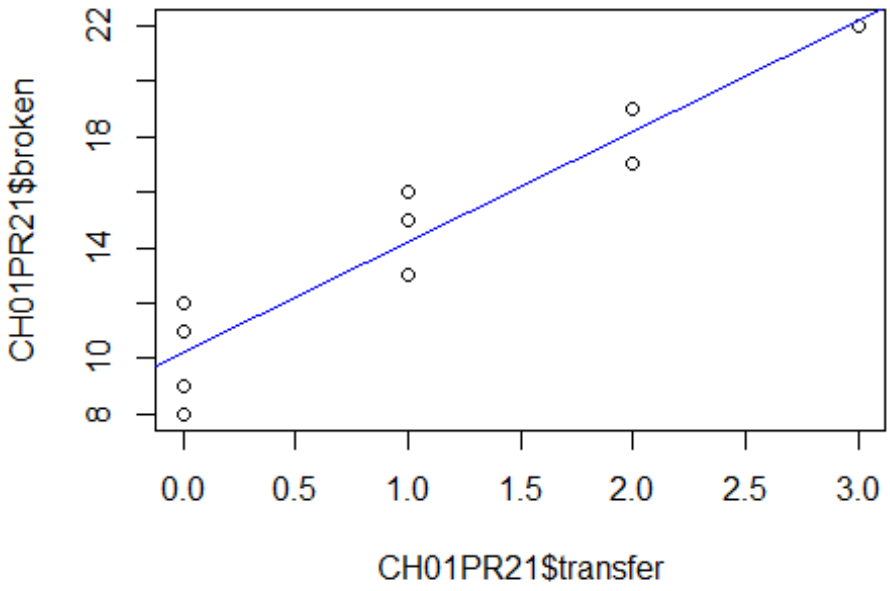
```
##        ACT
## 0.03882713
```

Problem 1.21

$X =$ number of transfers $Y =$ number of broken ampules

```r
# Load data in to R
CH01PR21 <- read_excel("G:/Teaching/BIOS 840/Fall
2017/Datasets/CH01PR21.xlsx",
                       col_names = FALSE)

# Clean the data
names(CH01PR21) <- c("broken","transfer")
```

```r
# broken = Beta_0 + Beta_1 * transfer + epsilon
# broken ~ transfer

# A. Obtain b0 and b1 and plot estimated regression function and data
# b0 =   estimates beta_0
# b1 =   estimates beta_1
m1 <- lm(broken~transfer,data=CH01PR21)
sm1 <- summary(m1)
print(sm1)

##
## Call:
## lm(formula = broken ~ transfer, data = CH01PR21)
##
## Residuals:
##    Min     1Q Median     3Q    Max
##   -2.2   -1.2    0.3    0.8    1.8
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  10.2000     0.6633  15.377 3.18e-07 ***
## transfer      4.0000     0.4690   8.528 2.75e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.483 on 8 degrees of freedom
## Multiple R-squared:  0.9009, Adjusted R-squared:  0.8885
## F-statistic: 72.73 on 1 and 8 DF,  p-value: 2.749e-05

# Plot to make sure a linear relationship makes sense
plot(CH01PR21$broken~CH01PR21$transfer)
abline(m1,col="blue")
```




```
# B. Estimate Y when X = 1 transfers
(Y.hat <- predict(m1,data.frame(transfer=1)))

##    1
## 14.2

# C. Estimate increase in Y when there are 2 transfers as compared to 1
transfer
coef(m1)[2]

## transfer
##        4

# D. Verify line goes through bar(X) and bar(Y)
(Y.bar <- mean(CH01PR21$broken))

## [1] 14.2

(Y.hat <- predict(m1, data.frame(transfer=mean(CH01PR21$transfer))))

##    1
## 14.2
```

Problem 1.45

X_1 = Infection Risk X_2 = Available Facilities and Services X_3 = Routine Chest X-ray Ratio
Y = Length of Stay

```r
# Load data in to R
APPENC01 <- read_excel("G:/Teaching/BIOS 840/Fall
2017/Datasets/APPENC01.xlsx",
                       col_names = FALSE)

# Clean the data

# Length of Stay (Y) is V2
# Infection Risk (X1) is V4
# Available Facilities and Services (X2) is V12
# Routine Chest X-ray Ratio (X3) is V6

APPENC01$LOS <- APPENC01$...2
APPENC01$InfRisk <- APPENC01$...4
APPENC01$FacServ <- APPENC01$...12
APPENC01$XRay <- APPENC01$...6

# Y = Beta_0 + Beta_1 * X + epsilon
# Y ~ X
# A. Fit three first order models and state the estimated regression
functions
m1 <- lm(LOS~InfRisk,data=APPENC01)
(sm1 <- summary(m1))

##
## Call:
## lm(formula = LOS ~ InfRisk, data = APPENC01)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.0587 -0.7776 -0.1487  0.7159  8.2805
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   6.3368     0.5213  12.156  < 2e-16 ***
## InfRisk       0.7604     0.1144   6.645 1.18e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.624 on 111 degrees of freedom
## Multiple R-squared:  0.2846, Adjusted R-squared:  0.2781
## F-statistic: 44.15 on 1 and 111 DF,  p-value: 1.177e-09

coef(m1)

## (Intercept)     InfRisk
##   6.3367865   0.7604209

# Fitted regression line for Infection Risk and Length of Stay is:
# LOS = 6.337 + 0.76*InfRisk
```

```r
m2 <- lm(LOS~FacServ,data=APPENC01)
(sm2 <- summary(m2))
```

```
##
## Call:
## lm(formula = LOS ~ FacServ, data = APPENC01)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.2712 -1.0716 -0.2816  0.7584  9.5433
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  7.71877    0.51020  15.129  < 2e-16 ***
## FacServ      0.04471    0.01116   4.008 0.000111 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.795 on 111 degrees of freedom
## Multiple R-squared:  0.1264, Adjusted R-squared:  0.1185
## F-statistic: 16.06 on 1 and 111 DF,  p-value: 0.0001113
```

```r
coef(m2)
```

```
## (Intercept)     FacServ
##  7.71876716  0.04470767
```

```r
# LOS = 7.719 + 0.045*FacServ
```

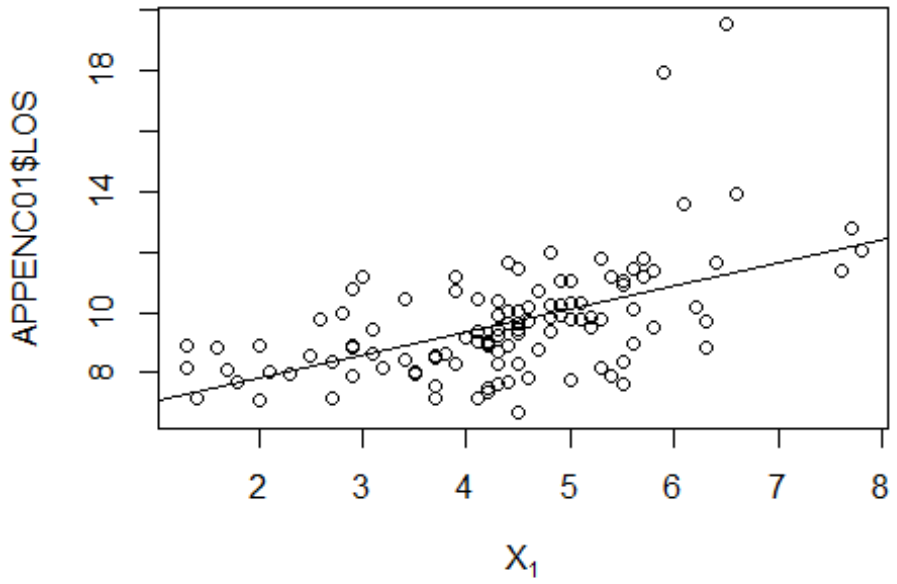
```r
m3 <- lm(LOS~XRay,data=APPENC01)
(sm3 <- summary(m3))
```

```
##
## Call:
## lm(formula = LOS ~ XRay, data = APPENC01)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.9226 -1.0810 -0.2708  0.8200  8.7008
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 6.566373   0.726094   9.043 5.67e-15 ***
## XRay        0.037756   0.008657   4.361 2.91e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.774 on 111 degrees of freedom
## Multiple R-squared:  0.1463, Adjusted R-squared:  0.1386
## F-statistic: 19.02 on 1 and 111 DF,  p-value: 2.906e-05
```

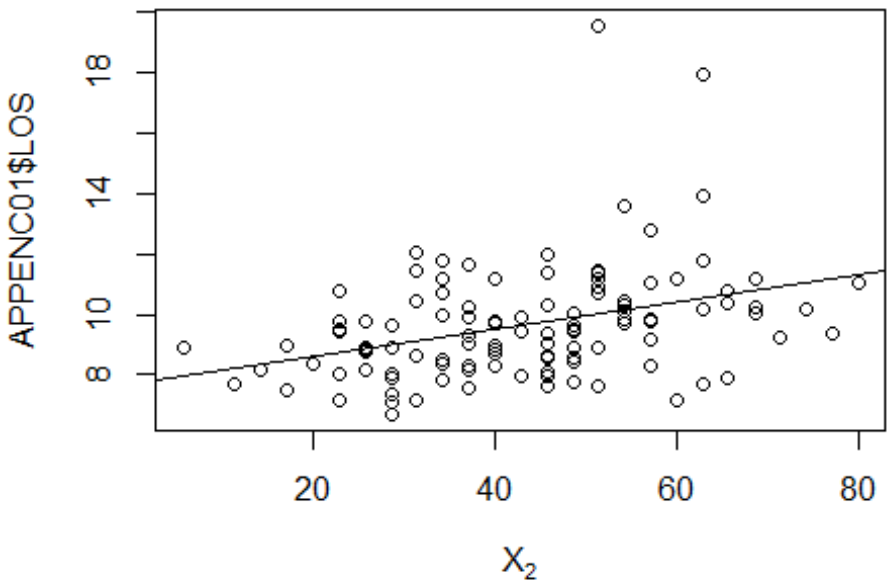
```r
coef(m3)
```

```
## (Intercept)        XRay
##  6.56637341  0.03775583
```

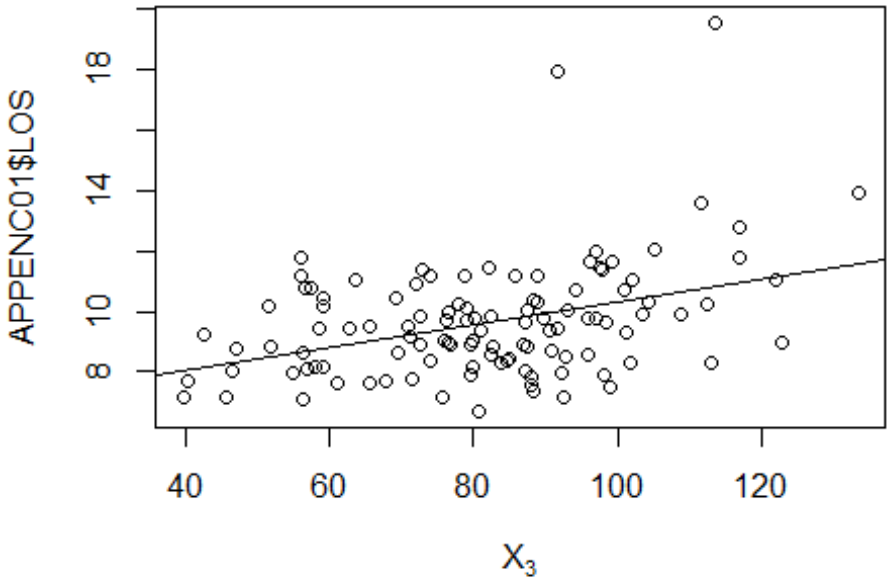
```r
# LOS = 6.566 + 0.038*XRay


# B. Plot the three estimated regression functions and data on separate
graphs
#    Does a linear relation appear to provide a good fit for each of the Xs?

plot(APPENC01$LOS~APPENC01$InfRisk,xlab=expression(X[1]))
abline(m1)
```




```r
plot(APPENC01$LOS~APPENC01$FacServ,xlab=expression(X[2]))
abline(m2)
```




```r
plot(APPENC01$LOS~APPENC01$XRay,xlab=expression(X[3]))
abline(m3)
```



```r
# C. Calculate MSE for each X, which leads to smallest variability about the
fitted line?
MSE1 <- sm1$sigma^2
MSE2 <- sm2$sigma^2
MSE3 <- sm3$sigma^2
MSE <- c(MSE1,MSE2,MSE3)
plot(MSE~c(1,2,3),xlab="Model",axes=F,pch=16,cex=2)
axis(1,c(1,2,3),c(1,2,3))
axis(2,c(seq(2.7,3.3,0.1)))
box(lty=1)
```