HOMEWORK 1

Yiqiao Yin [YY2502]

Contents

PROBLEM 1

Let us consider residual $\epsilon_i = y_i - \hat{y}_i$. We want to prove

- $\sum_{i=1}^n \hat{\epsilon}_i = 0$
- $\sum_{i=1}^n y_i = \sum_{i=1}^n \hat{Y}_i$
- $\sum_{i=1}^k X_i \epsilon_i = 0$
- $\sum_{i=1}^n \hat{Y}_i \epsilon_i = 0$

Let us prove the result in the following.

- Let us start by writing down $\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X_i$. This can be further written into $\hat{Y} = \bar{Y} - \hat{\beta}_1 \bar{X} + \hat{\beta}_1 X_i = \bar{Y} + \hat{\beta}(X_i - \bar{X})$. Now we can proceed with the left hand side of the equation below:

$$\begin{aligned}
\sum \epsilon &= \sum (Y_i - \hat{Y}) \\
&= \sum (Y_i - (\bar{Y} + \hat{\beta}_1 (X_i - \bar{X}))) \\
&= \sum Y_i - n\bar{Y} - \hat{\beta}_1 \sum (X_i - \bar{X}) \\
&= \sum Y_i - \sum Y_i - 0 \\
&= 0
\end{aligned}$$

and we are done.

- Recall from above we have $\sum \epsilon_i = 0$ and the definition of residual $\epsilon_i = Y_i - \hat{Y}_i$. Then we can derive the following

$$\begin{aligned} \sum \epsilon_i &= \sum Y_i - \hat{Y}_i = 0 \\ \Rightarrow \sum Y_i - \hat{Y}_i &= 0 \\ \Rightarrow \sum Y_i &= \sum \hat{Y}_i \end{aligned}$$

and the proof is complete.

- Recall that $\hat{Y}_i = \beta_0 + \beta_1 X_i$. We consider the following

$$\begin{aligned} \sum_i X_i \epsilon_i &= \sum (X_i(Y_i - (\beta_0 + \beta_1 X_i))) \\ &= \sum X_i Y_i - \beta_0 \sum X_i - \beta_1 \sum X_i^2 \\ &= \beta_0 \sum X_i + b\beta_1 \sum X_i^2 - \beta_0 \sum X_i - \beta_1 \sum X_i^2 \\ &= 0 \end{aligned}$$

  since $\sum X_i Y_i = \beta_0 \sum X_i + \beta_1 \sum X_i^2$ from page 17 of textbook.

- We show

$$\begin{aligned} \sum Y_i \epsilon_i &= \sum Y_i^2 - \sum Y_i(\beta_0 + \beta_1 X_i) \\ &= \sum Y_i^2 - \beta_0 \sum Y_i - \beta_1 \sum Y_1 X_i \\ &= \sum (\beta_0 + \beta_1 X_1)\epsilon_i \\ &= \beta_0 \sum \epsilon_i + \beta_1 \sum X_i \epsilon_i \end{aligned}$$

Done.

PROBLEM 2

Set working directory:

```
# Data
setwd("C:/Users/eagle/OneDrive/Course/CU Stats/STATS GR6101 - Applied Statistics I/Data")
data <- data.frame(read.delim("CH01PR19.txt", header = FALSE, sep = " "))[, c(2,6)]
# dim(data); head(data)

# Define Variables
Y <- data[, 1]
X <- data[, 2]

# Linear Model
linearModel <- lm(Y~X)
summary(linearModel)
```

```
##
## Call:
## lm(formula = Y ~ X)
##
## 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 ***
```

```
## X            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
```

From results above, we have the least square estimates of parameters to be $\hat{\beta}_0 = 2.11405$ and $\hat{\beta}_1 = 0.3883$. Thus, the linear regression model is

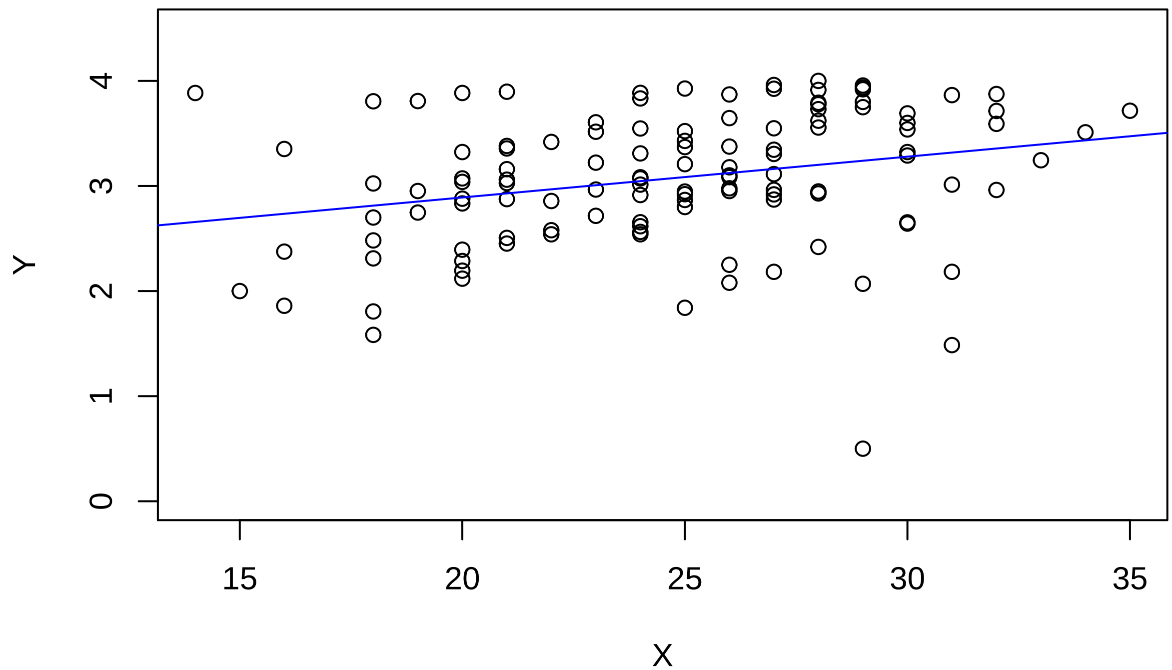
$$Y = 2.11405 + 0.03883 \cdot X$$

```r
# Data
setwd("C:/Users/eagle/OneDrive/Course/CU Stats/STATS GR6101 - Applied Statistics I/Data")
data <- data.frame(read.delim("CH01PR19.txt", header = FALSE, sep = " "))[, c(2,6)]
# dim(data); head(data)

# Define Variables
Y <- data[, 1]
X <- data[, 2]

# Linear Model
linearModel <- lm(Y~X)
# summary(linearModel)

# Plot
plot(X, Y, ylim = c(0,4.5));
abline(a = coef(linearModel)[1],
       b = coef(linearModel)[2], col = "blue")
```



From observing the plot above, it does not seem like the model is that good of a fit.

Given $X = 30$, we can compute estimated response variable to be the following:

```
estimate = 2.11405 + 0.03883*30
estimate
```

```
## [1] 3.27895
```

If the entrance test score increases by one point, we expect the point estimate to increase by 0.03883 on average.

PROBLEM 3

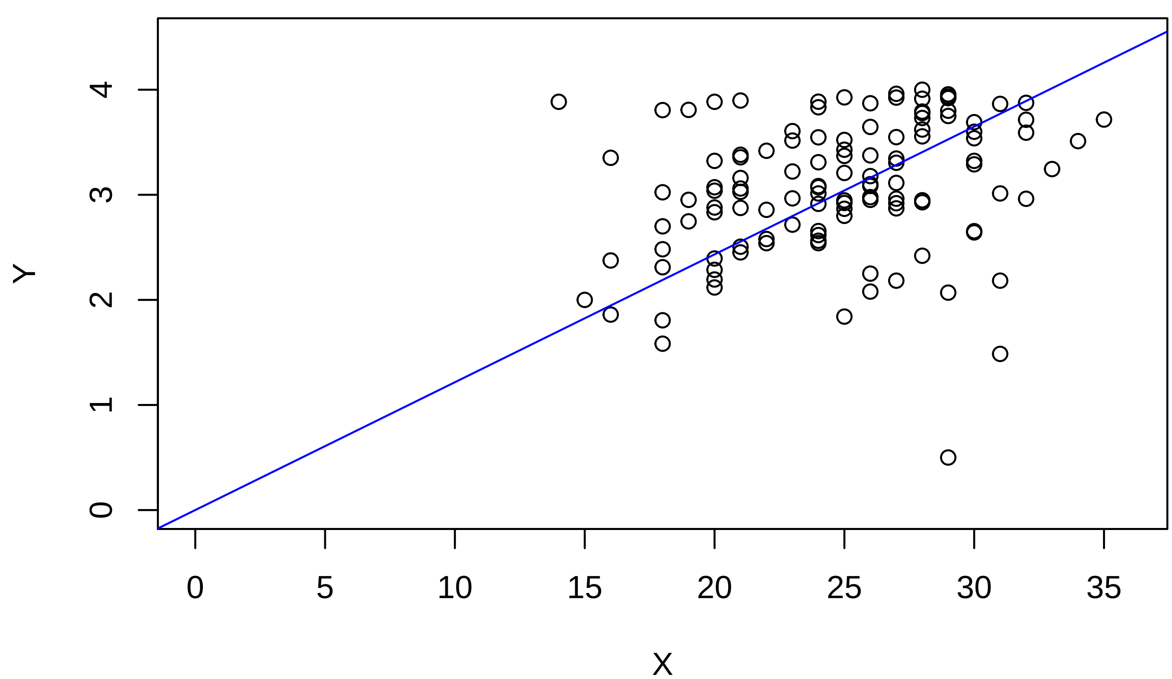
Suppose $\beta_0 = 0$, then model (1.1) from textbook becomes $Y_i = \beta_1 X_i + \epsilon_i$. In this case, the intercept of the model is zero, which means the fitted line goes through origin at $X = 0$. This action is allowed because $X = 0$ exists.

The regression line would go through origin of the Cartesian plane. Let me use the example in PROBLEM 2 below.

```
# Linear Model
linearModel <- lm(Y~X-1)
summary(linearModel)
```

```
##
## Call:
## lm(formula = Y ~ X - 1)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.0276 -0.2737  0.1077  0.4754  2.1820
##
## Coefficients:
##   Estimate Std. Error t value Pr(>|t|)
## X 0.121643   0.002637   46.13   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7257 on 119 degrees of freedom
## Multiple R-squared:  0.947,  Adjusted R-squared:  0.9466
## F-statistic:  2128 on 1 and 119 DF,  p-value: < 2.2e-16
```

```r
# Plot
plot(X, Y, ylim = c(0,4.5), xlim = c(0,36))
abline(a = 0, b = coef(linearModel)[1], col = "blue")
```



From theoretical point of view, consider regression model (1.1) from text, we have $Y_i = \beta_0 + \beta_1 X + \epsilon_i$. Assuming $\beta_0 = 0$ and $X = 0$ is within range. Then we expect to have a model $Y_i = \beta_1 X + \epsilon_i$. At the point $X = 0$, we havea $\mathbb{E}Y_i = \mathbb{E}(\beta_1 X + \epsilon)|_{X=0} = \mathbb{E}(0 + \epsilon_i) = \mathbb{E}\epsilon_i = 0$ but each experiment does not have to be zero, which means the line does not necessarily go through (0,0). Realistically speaking, the intercept of the model $Y_i\ \beta_1 X + \epsilon_i$ has $\mathbb{E}Y_i = 0$ and $\text{var}(Y_i) = \text{var}\epsilon_i = \sigma^2$.

PROBLEM 4

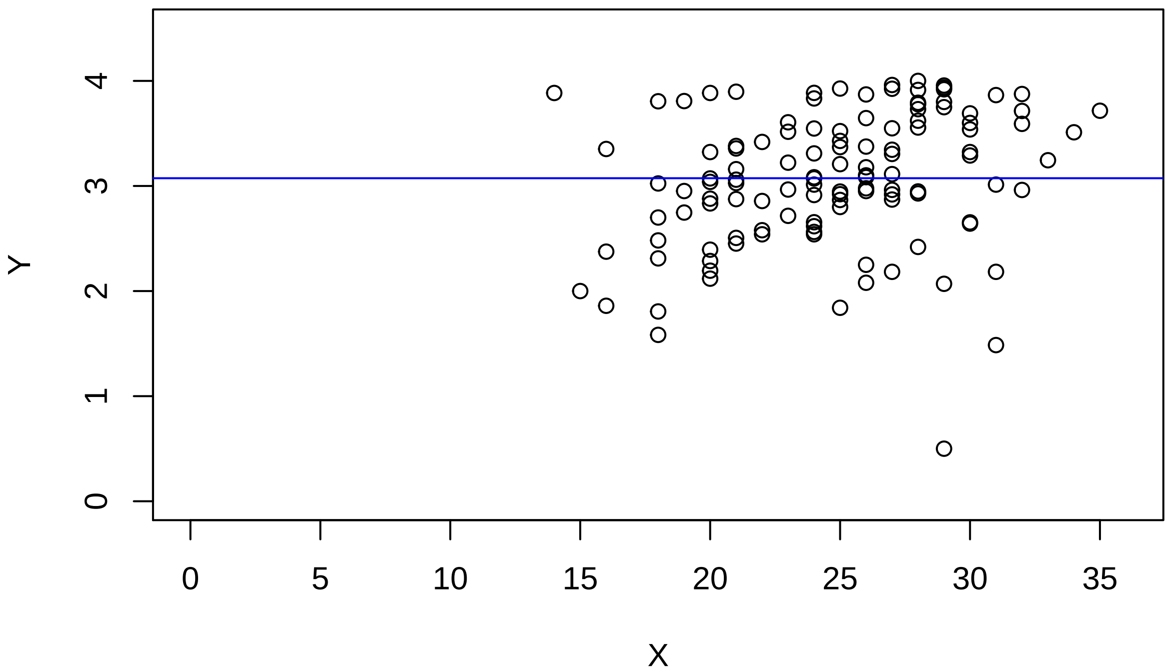
Suppose now that $\beta_1 = 0$ as opposed to $\beta_0 = 0$ in PROBLEM 3. We have a model without any covariate because the parameter β_1 is not there. In this case, we are really modeling Y by a constant vector, say a vector of 1's (e.g. $\mathbb{1}$s). Here the model $Y_i = \beta_0 + \epsilon_i$ will be a poor fit because it is as if we are modeling using $X_i = [1, 1, ..., 1]$ only.

Let me use the same example above and set covariate to 1 only. Below let us examine the model and the plot.

```
# Linear Model
linearModel <- lm(Y~1)
summary(linearModel)
```

```
##
## Call:
## lm(formula = Y ~ 1)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.57405 -0.38530  0.00345  0.51920  0.92595
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  3.07405    0.05882   52.26   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6443 on 119 degrees of freedom
```

```
# Plot
plot(X, Y, ylim = c(0,4.5), xlim = c(0,36))
# abline(a = 0, b = coef(linearModel)[1], col = "blue") # <= this is wrong!!!!
abline(a = coef(linearModel)[1], b = 0, col = "blue")
```



PROBLEM 5

Let us load the data and run linear model.

```
# Data
setwd("C:/Users/eagle/OneDrive/Course/CU Stats/STATS GR6101 - Applied Statistics I/Data")
data <- read.csv("APPENC02.csv", header = FALSE)
colnames(data) <- c(
  "ID",
  "Country",
  "State",
  "Land_Area",
  "Total_Population",
  "Perc_Popu_18_34",
  "Perc_Popu_Over_65",
  "Num_Active_Phy",
  "Num_Hospital_Beds",
  "Total_Serious_Crimes",
  "Percent_High_School",
  "Percent_Bachelor_Deg",
  "Percent_Below_Poverty",
  "Percent_Unemployment",
  "Per_Capita_Income",
  "Total_Personal_Income",
```

```
  "Geographic_Region"
)
dim(data); head(data)
```

```
## [1] 440  17
```

```
##   ID     Country State Land_Area Total_Population Perc_Popu_18_34
## 1  1 Los_Angeles    CA      4060          8863164            32.1
## 2  2        Cook    IL       946          5105067            29.2
## 3  3      Harris    TX      1729          2818199            31.3
## 4  4   San_Diego    CA      4205          2498016            33.5
## 5  5      Orange    CA       790          2410556            32.6
## 6  6       Kings    NY        71          2300664            28.3
##   Perc_Popu_Over_65 Num_Active_Phy Num_Hospital_Beds Total_Serious_Crimes
## 1               9.7          23677             27700               688936
## 2              12.4          15153             21550               436936
## 3               7.1           7553             12449               253526
## 4              10.9           5905              6179               173821
## 5               9.2           6062              6369               144524
## 6              12.4           4861              8942               680966
##   Percent_High_School Percent_Bachelor_Deg Percent_Below_Poverty
## 1                70.0                 22.3                  11.6
## 2                73.4                 22.8                  11.1
## 3                74.9                 25.4                  12.5
## 4                81.9                 25.3                   8.1
## 5                81.2                 27.8                   5.2
## 6                63.7                 16.6                  19.5
##   Percent_Unemployment Per_Capita_Income Total_Personal_Income
## 1                  8.0             20786                184230
## 2                  7.2             21729                110928
## 3                  5.7             19517                 55003
## 4                  6.1             19588                 48931
## 5                  4.8             24400                 58818
## 6                  9.5             16803                 38658
##   Geographic_Region
## 1                 4
## 2                 2
## 3                 3
## 4                 4
## 5                 4
## 6                 1
```

```
# Model 1:
linearModel1 <- lm(data$Num_Active_Phy~data$Total_Population)
summary(linearModel1)
```

```
##
## Call:
## lm(formula = data$Num_Active_Phy ~ data$Total_Population)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
```

```
## -1969.4  -209.2   -88.0    27.9  3928.7
##
## Coefficients:
##                         Estimate Std. Error t value Pr(>|t|)
## (Intercept)           -1.106e+02  3.475e+01  -3.184  0.00156 **
## data$Total_Population  2.795e-03  4.837e-05  57.793  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 610.1 on 438 degrees of freedom
## Multiple R-squared:  0.8841, Adjusted R-squared:  0.8838
## F-statistic:  3340 on 1 and 438 DF,  p-value: < 2.2e-16
```

```r
# Model 2:
linearModel2 <- lm(data$Num_Active_Phy~data$Num_Hospital_Beds)
summary(linearModel2)
```

```
##
## Call:
## lm(formula = data$Num_Active_Phy ~ data$Num_Hospital_Beds)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3133.2  -216.8   -32.0    96.2  3611.1
##
## Coefficients:
##                         Estimate Std. Error t value Pr(>|t|)
## (Intercept)            -95.93218   31.49396  -3.046  0.00246 **
## data$Num_Hospital_Beds   0.74312    0.01161  63.995  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 556.9 on 438 degrees of freedom
## Multiple R-squared:  0.9034, Adjusted R-squared:  0.9032
## F-statistic:  4095 on 1 and 438 DF,  p-value: < 2.2e-16
```

```r
# Model 3:
linearModel3 <- lm(data$Num_Active_Phy~data$Total_Personal_Income)
summary(linearModel3)
```

```
##
## Call:
## lm(formula = data$Num_Active_Phy ~ data$Total_Personal_Income)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1926.6  -194.5   -66.6    44.2  3819.0
##
## Coefficients:
##                             Estimate Std. Error t value Pr(>|t|)
## (Intercept)                -48.39485   31.83333   -1.52    0.129
## data$Total_Personal_Income   0.13170    0.00211   62.41   <2e-16 ***
## ---
```

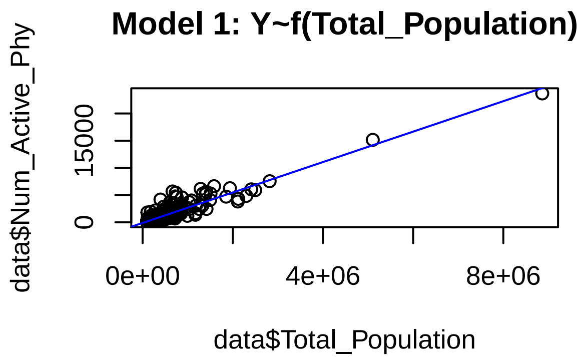
```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 569.7 on 438 degrees of freedom
## Multiple R-squared:  0.8989, Adjusted R-squared:  0.8987
## F-statistic:  3895 on 1 and 438 DF,  p-value: < 2.2e-16
```

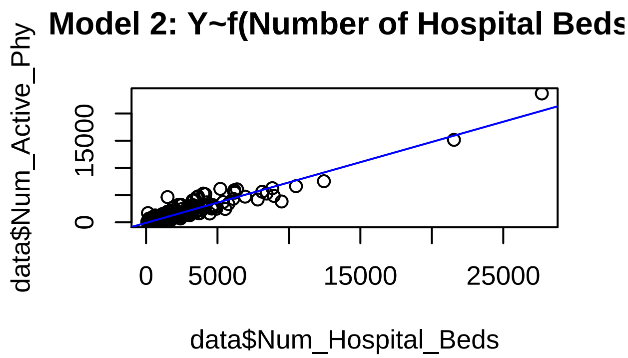
Let us state the model (using outputs from R above):

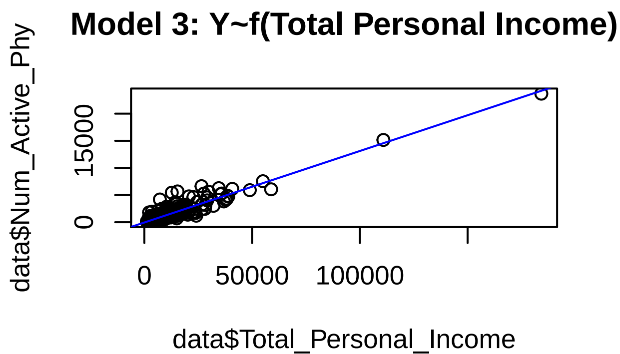
- Model 1: $Y = -110.6 + 0.00279 \cdot \text{Total Population}$
- Model 2: $Y = -95.9322 + 0.7431 \cdot \text{Number of Hospital Beds}$
- Model 3: $Y = -48.3949 + 0.1317 \cdot \text{Total Personal Income}$

Let us use the following R code to examine the plots:

```
# Plot
par(mfrow=c(2,2))
plot(data$Total_Population, data$Num_Active_Phy,
     main="Model 1: Y~f(Total_Population)");
abline(a = coef(linearModel1)[1],
       b = coef(linearModel1)[2], col = "blue")
plot(data$Num_Hospital_Beds, data$Num_Active_Phy,
     main="Model 2: Y~f(Number of Hospital Beds)");
abline(a = coef(linearModel2)[1],
       b = coef(linearModel2)[2], col = "blue")
plot(data$Total_Personal_Income, data$Num_Active_Phy,
     main="Model 3: Y~f(Total Personal Income)");
abline(a = coef(linearModel3)[1],
       b = coef(linearModel3)[2], col = "blue")
```







From above comparison plots, we can observe that all three models fit the data relatively the same. We observe most of the data points clustered in the left bottom corner of the Cartesian axis with a few outliers. From how wide the cluters are, we observe that the second model fits the data points slightly better than the first model and the second model. We can confirm this with Mean Square Error (MSE) which is calculated below.

```
# MSE (Use formula)
MSE1 <- mean((data$Num_Active_Phy -
                (coef(linearModel1)[1] +
                   coef(linearModel1)[2] * data$Total_Population))^2)
MSE2 <- mean((data$Num_Active_Phy -
                (coef(linearModel2)[1] +
                   coef(linearModel2)[2] * data$Num_Hospital_Beds))^2)
MSE3 <- mean((data$Num_Active_Phy -
                (coef(linearModel3)[1] +
                   coef(linearModel3)[2] * data$Total_Personal_Income))^2)

# Or (Use package)
mean(linearModel1$residuals^2)
```

```
## [1] 370511.7
```

```
mean(linearModel2$residuals^2)
```

```
## [1] 308781.9
```

```r
mean(linearModel3$residuals^2)
```

```
## [1] 323064.2
```

As computed above, the residuals are the following:

Risk	Results	Formula
MSE1	370511.7	Y~f(Total_Population)
MSE2	308781.9	Y~f(Number_of_Hospital_Beds)
MSE3	323064.2	Y~f(Total_Personal_Income)

PROBLEM 6

Referring to the same example, let us build mixed model of $Y = \beta_0 + \beta_1 \cdot \text{Percent of at least one Bachelor Degree}$ under different Geographical Region. There are 4 different levels for Geographical Region: NE, NC, S, W.

Let us examine the data using the following R code:

```r
# Mixed Model
head(data, 2)
```

```
##   ID     Country State Land_Area Total_Population Perc_Popu_18_34
## 1  1 Los_Angeles    CA      4060          8863164            32.1
## 2  2        Cook    IL       946          5105067            29.2
##   Perc_Popu_Over_65 Num_Active_Phy Num_Hospital_Beds Total_Serious_Crimes
## 1               9.7          23677             27700               688936
## 2              12.4          15153             21550               436936
##   Percent_High_School Percent_Bachelor_Deg Percent_Below_Poverty
## 1                70.0                 22.3                  11.6
## 2                73.4                 22.8                  11.1
##   Percent_Unemployment Per_Capita_Income Total_Personal_Income
## 1                  8.0             20786                184230
## 2                  7.2             21729                110928
##   Geographic_Region
## 1                 4
## 2                 2
```

```r
table(data$Geographic_Region)
```

```
## 
##   1   2   3   4 
## 103 108 152  77
```

```r
data_geo_1 <- data[data$Geographic_Region == 1, ]; dim(data_geo_1)
```

```
## [1] 103  17
```

```r
linearModel_Geo_1 <- lm(data_geo_1$Per_Capita_Income~data_geo_1$Percent_Bachelor_Deg)
summary(linearModel_Geo_1)
```

```
##
## Call:
## lm(formula = data_geo_1$Per_Capita_Income ~ data_geo_1$Percent_Bachelor_Deg)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -10613.5  -1276.2    -68.9   1256.6   6790.4
##
## Coefficients:
##                                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)                       9223.82     851.77   10.83   <2e-16 ***
## data_geo_1$Percent_Bachelor_Deg   522.16      37.13   14.06   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2708 on 101 degrees of freedom
## Multiple R-squared:  0.6619, Adjusted R-squared:  0.6586
## F-statistic: 197.8 on 1 and 101 DF,  p-value: < 2.2e-16
```

```r
data_geo_2 <- data[data$Geographic_Region == 2, ]; dim(data_geo_2)
```

```
## [1] 108  17
```

```r
linearModel_Geo_2 <- lm(data_geo_2$Per_Capita_Income~data_geo_2$Percent_Bachelor_Deg)
summary(linearModel_Geo_2)
```

```
##
## Call:
## lm(formula = data_geo_2$Per_Capita_Income ~ data_geo_2$Percent_Bachelor_Deg)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -7167.6  -915.4   105.6   886.6  6159.2
##
## Coefficients:
##                                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)                      13581.41     575.14  23.614  < 2e-16 ***
## data_geo_2$Percent_Bachelor_Deg   238.67      27.23   8.765 3.34e-14 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2100 on 106 degrees of freedom
## Multiple R-squared:  0.4202, Adjusted R-squared:  0.4147
## F-statistic: 76.83 on 1 and 106 DF,  p-value: 3.344e-14
```

```r
data_geo_3 <- data[data$Geographic_Region == 3, ]; dim(data_geo_3)
```

```
## [1] 152  17
```

```r
linearModel_Geo_3 <- lm(data_geo_3$Per_Capita_Income~data_geo_3$Percent_Bachelor_Deg)
summary(linearModel_Geo_3)
```

```
##
## Call:
## lm(formula = data_geo_3$Per_Capita_Income ~ data_geo_3$Percent_Bachelor_Deg)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9724.7 -1362.8   114.9  1255.6  9883.8
##
## Coefficients:
##                                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)                      10529.79     612.48   17.19   <2e-16 ***
## data_geo_3$Percent_Bachelor_Deg   330.61      27.13   12.19   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2734 on 150 degrees of freedom
## Multiple R-squared:  0.4975, Adjusted R-squared:  0.4941
## F-statistic: 148.5 on 1 and 150 DF,  p-value: < 2.2e-16
```

```r
data_geo_4 <- data[data$Geographic_Region == 4, ]
linearModel_Geo_4 <- lm(data_geo_4$Per_Capita_Income~data_geo_4$Percent_Bachelor_Deg)
summary(linearModel_Geo_4)
```

```
##
## Call:
## lm(formula = data_geo_4$Per_Capita_Income ~ data_geo_4$Percent_Bachelor_Deg)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -8684.3 -1477.3   191.7  1557.8  9552.1
##
## Coefficients:
##                                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)                       8615.05    1052.20   8.188 5.24e-12 ***
## data_geo_4$Percent_Bachelor_Deg   440.32      45.37   9.705 6.86e-15 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2866 on 75 degrees of freedom
## Multiple R-squared:  0.5567, Adjusted R-squared:  0.5508
## F-statistic: 94.19 on 1 and 75 DF,  p-value: 6.856e-15
```

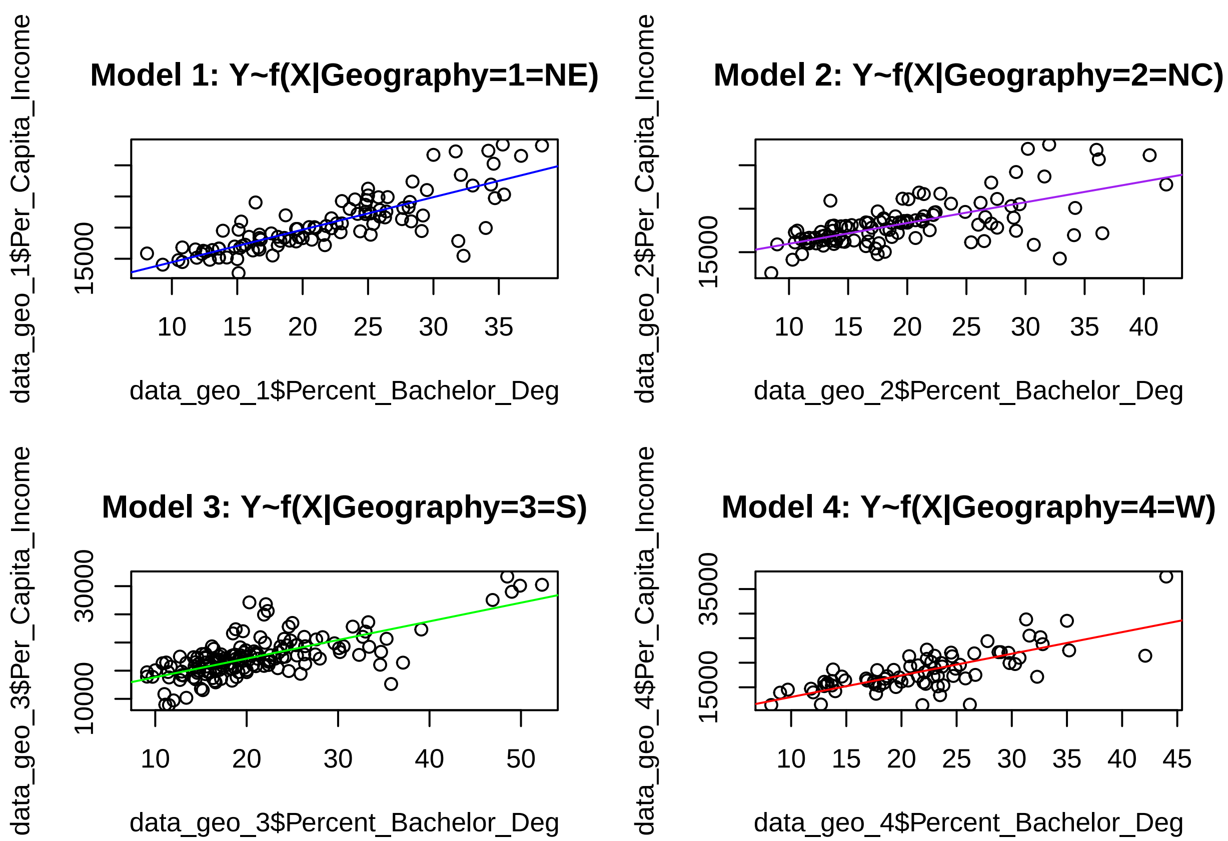
In total, we have 4 models:

- Model 1: $Y = 9223.82 + 522.16X$ given region is NE (or coded as 1);
- Model 2: $Y = 13581.41 + 238.67X$ given region is NC (or coded as 2);
- Model 3: $Y = 10529.79 + 330.61X$ given region is S (or coded as 3);

- Model 4: $Y = 8615.05 + 440.32X$ given region is W (or coded as 4).

Let X be *Percent_Bachelor_Deg* in the data.

```
# Plot
par(mfrow=c(2,2))
plot(data_geo_1$Percent_Bachelor_Deg, data_geo_1$Per_Capita_Income,
     main="Model 1: Y~f(X|Geography=1=NE)");
abline(a = coef(linearModel_Geo_1)[1],
       b = coef(linearModel_Geo_1)[2], col = "blue")
plot(data_geo_2$Percent_Bachelor_Deg, data_geo_2$Per_Capita_Income,
     main="Model 2: Y~f(X|Geography=2=NC)");
abline(a = coef(linearModel_Geo_2)[1],
       b = coef(linearModel_Geo_2)[2], col = "purple")
plot(data_geo_3$Percent_Bachelor_Deg, data_geo_3$Per_Capita_Income,
     main="Model 3: Y~f(X|Geography=3=S)");
abline(a = coef(linearModel_Geo_3)[1],
       b = coef(linearModel_Geo_3)[2], col = "green")
plot(data_geo_4$Percent_Bachelor_Deg, data_geo_4$Per_Capita_Income,
     main="Model 4: Y~f(X|Geography=4=W)");
abline(a = coef(linearModel_Geo_4)[1],
       b = coef(linearModel_Geo_4)[2], col = "red")
```



As we can observe above, the last geographical region is probaby a better fit.

Let us examine the MSE below.

```r
mean(linearModel_Geo_1$residuals^2)
```

```
## [1] 7192580
```

```r
mean(linearModel_Geo_2$residuals^2)
```

```
## [1] 4329650
```

```r
mean(linearModel_Geo_3$residuals^2)
```

```
## [1] 7376003
```

```r
mean(linearModel_Geo_4$residuals^2)
```

```
## [1] 8000959
```

Risk	Results	Formula
MSE1	7192580	Y~f(X|Geography=1(NE)), Y = 9223.82 + 522.16 X
MSE2	4329650	Y~f(X|Geography=2(NC)), Y = 13581.41 + 238.67 X
MSE3	7376003	Y~f(X|Geography=3(S)), Y = 10529.79 + 330.61 X
MSE3	8000959	Y~f(X|Geography=4(W)), Y = 8615.05 + 440.32 X