STAT 5385: Lab 1

Willliam Ofosu Agyapong

1/21/2022

Table 1: Variables in the SENIC dataset

Variable Name	Coded As
Identification Number	ID
Length of stay	LOS
Age	age
Infection risk	infec_risk
Routine culturing ratio	cul_ratio
Routine chest X-ray ratio	xray_ratio
Number of beds	beds
Medical school affiliation	med
Region	region
Average daily census	ADC
Number of nurses	nurses
Available facilities and services	AFS

Reading in the SENIC dataset

```
# Get the file path
dir_path <-  dirname(rstudioapi::getSourceEditorContext()$path)
# set the current working directory to this path
setwd(dir_path)

# Import the dataset from local drive
senic <- read.table(file = "../Data Sets/Appendix C Data Sets/APPENC01.txt",
                    header = FALSE)

# View first few observations; Everything looks good.
head(senic)
##   V1    V2   V3  V4   V5    V6  V7 V8 V9 V10 V11 V12
## 1  1  7.13 55.7 4.1  9.0  39.6 279  2  4 207 241  60
## 2  2  8.82 58.2 1.6  3.8  51.7  80  2  2  51  52  40
## 3  3  8.34 56.9 2.7  8.1  74.0 107  2  3  82  54  20
## 4  4  8.95 53.7 5.6 18.9 122.8 147  2  4  53 148  40
## 5  5 11.20 56.5 5.7 34.5  88.9 180  2  1 134 151  40
## 6  6  9.76 50.9 5.1 21.9  97.0 150  2  2 147 106  40

# Rename columns
names(senic) <- var_table$coded_names
```

```
# Obtain summary report
senic %>% dfSummary() %>% view()
```

With the help of the ***summarytools*** *package we observe that the data has 12 variables with 113 observations as expected, with no missing values. The data is pre-cleaned, so we dive straight into the modeling.*

Problem 1.45

For this part, the average length of stay in a hospital (Y, `LOS`) is the response variable, while the infection risk (`infec_risk`), available facilities and services (`AFS`), and routine chest X-ray (`xray_ratio`) constitute the individual predictor variables.

Part (a): Estimating the models

Infection risk

```
mod_infec_risk <- lm(LOS ~ infec_risk, data = senic)
summary(mod_infec_risk)
##
## Call:
## lm(formula = LOS ~ infec_risk, data = senic)
##
## 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 ***
## infec_risk    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
```

From the above output, for infection risk, the estimated regression function is $\hat{y} = 6.34 + 0.76x$, where x is the infection risk. There is a positive linear relationship between infection risk and LOS.

Available facilities and services (AFS)

```
mod_AFS <- lm(LOS ~ AFS, data = senic)
summary(mod_AFS)
##
## Call:
## lm(formula = LOS ~ AFS, data = senic)
##
## 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 ***
## AFS          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
```

From the above output, for AFS , the estimated regression function is $\hat{y} = 7.72 + 0.04x$, where x is AFS. There is a positive linear relationship between AFS and LOS.

Chest X-ray Ratio

```
mod_xray_ratio <- lm(LOS ~ xray_ratio, data = senic)
summary(mod_xray_ratio)
##
## Call:
## lm(formula = LOS ~ xray_ratio, data = senic)
##
## 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_ratio  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
```

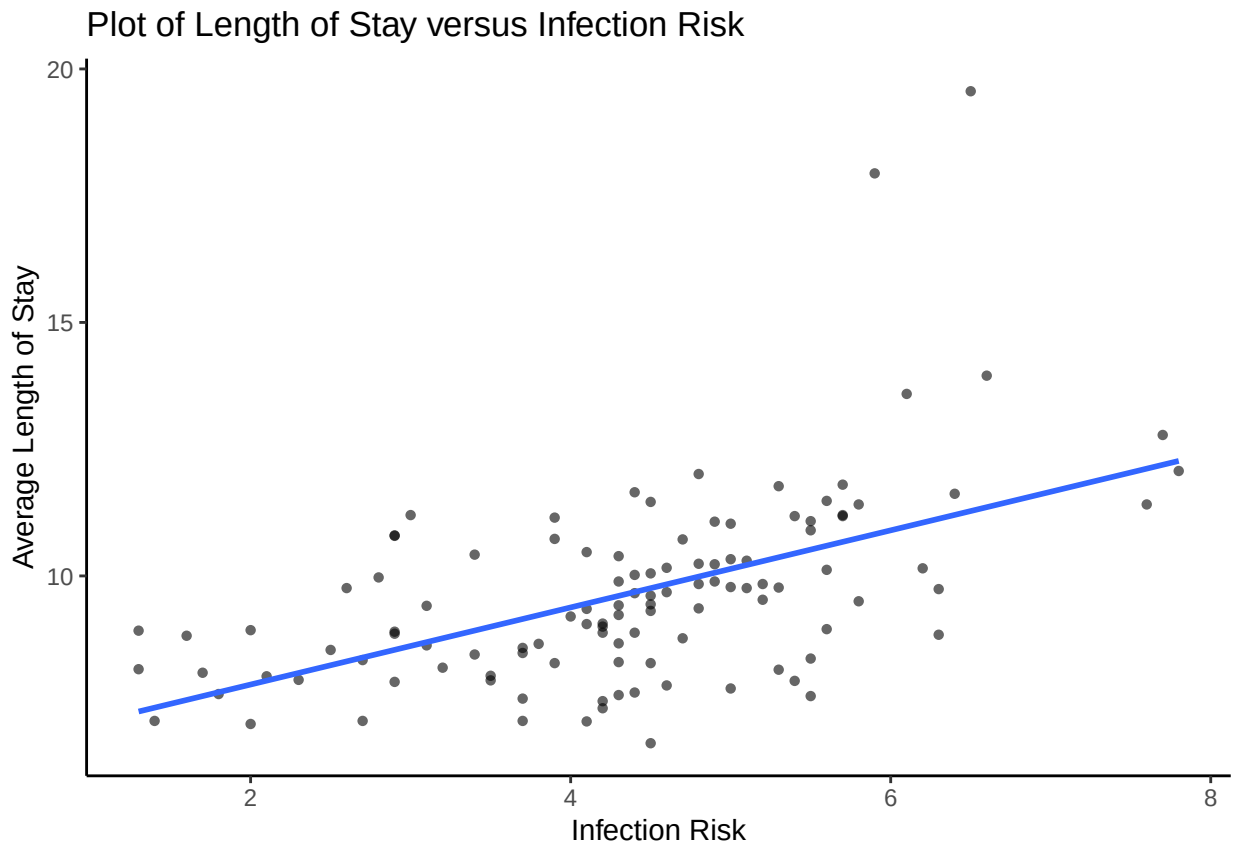
From the above output, for X-ray ratio, the estimated regression function is $\hat{y} = 6.57 + 0.04x$, where x is the X-ray ratio. There is a positive linear relationship between chest X-ray ratio and LOS.

Part (b): Plot the model

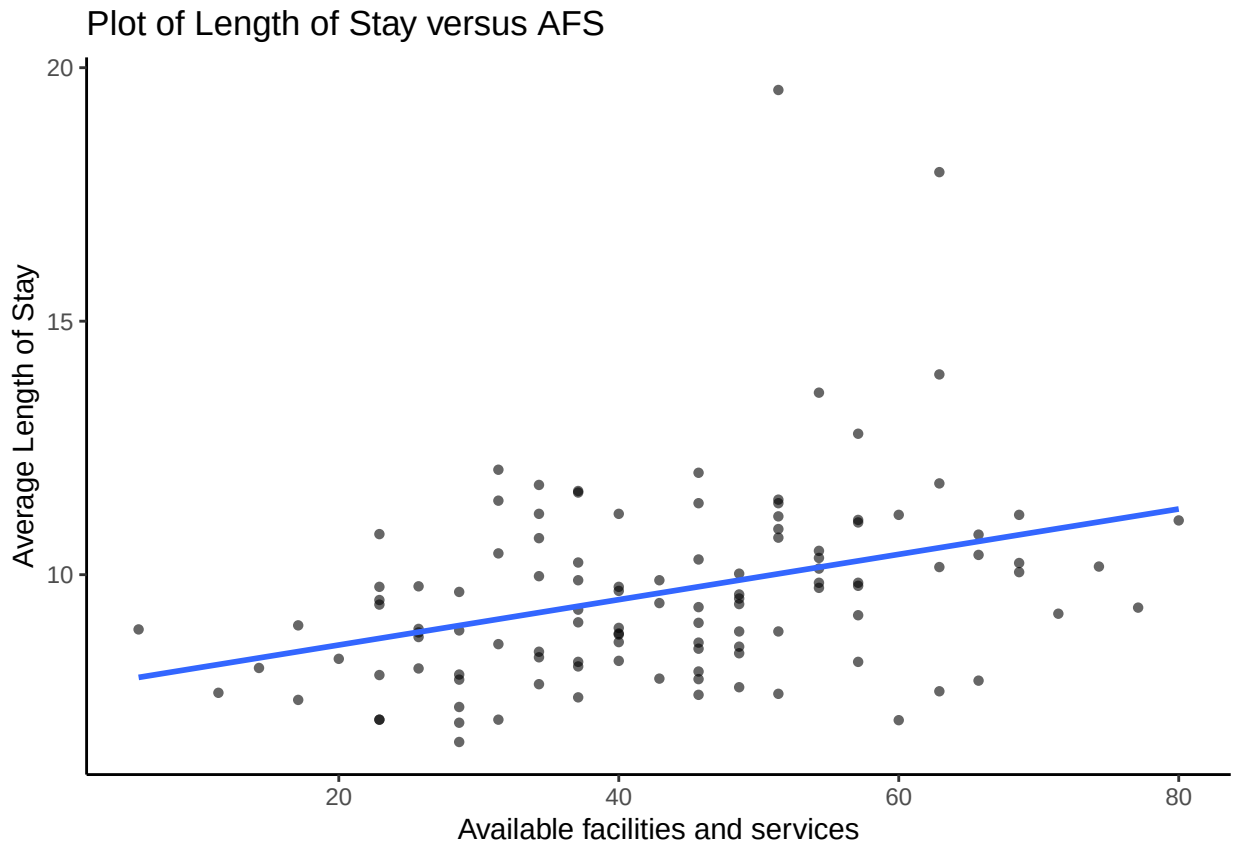
The following plots suggest that a linear relation appear to provide a good fit for each of the three predictor variables since, in all three plots, majority of the data points with the exception of few outliers are centered around the fitted regression line.

```
library(ggplot2)

# with(senic,plot(infec_risk,LOS))
# abline(lm(LOS~infec_risk,data=senic))

ggplot(senic, aes(infec_risk, LOS)) +
  geom_point(alpha = 0.6, shape = 16) +
  labs(x="Infection Risk", y="Average Length of Stay",
       title = "Plot of Length of Stay versus Infection Risk") +
```

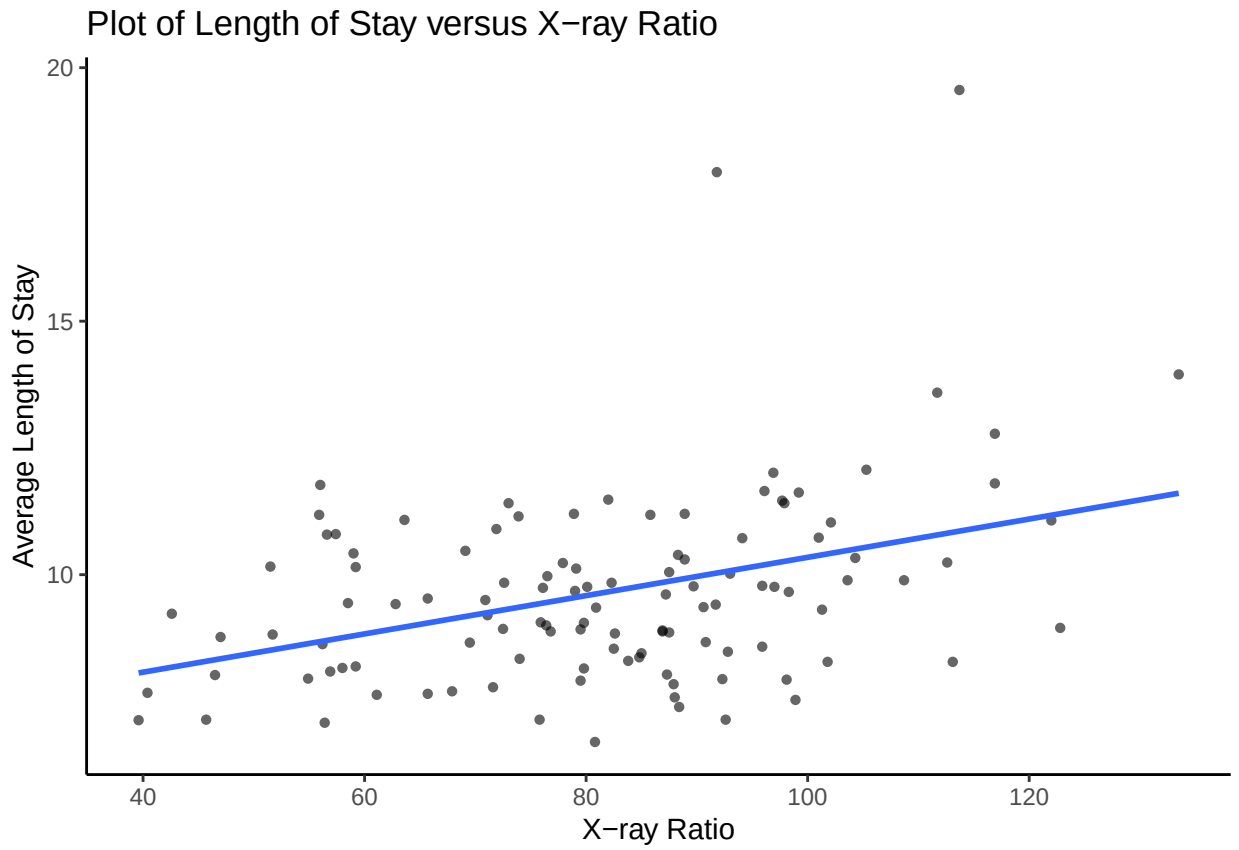
```
  theme_classic() +
  geom_smooth(method = "lm", se=FALSE)
```




```
# LOS vs. AFS
ggplot(senic, aes(AFS, LOS)) +
  geom_point(alpha = 0.6, shape = 16) +
  labs(x="Available facilities and services", y="Average Length of Stay",
       title = "Plot of Length of Stay versus AFS") +
  theme_classic() +
  geom_smooth(method = "lm", se=FALSE)
```




```
# LOS vs. xray_ratio
ggplot(senic, aes(xray_ratio, LOS)) +
  geom_point(alpha = 0.6, shape = 16) +
  labs(x="X-ray Ratio", y="Average Length of Stay",
       title = "Plot of Length of Stay versus X-ray Ratio") +
  theme_classic() +
  geom_smooth(method = "lm", se=FALSE)
```



Part (c): Calculate MSE

```
mse <- c(round(anova(mod_infec_risk)$`Mean Sq`[2], 4),
         round(anova(mod_AFS)$`Mean Sq`[2], 4),
         round(anova(mod_xray_ratio)$`Mean Sq`[2], 4)
         )
data.frame(
  Predictor = c("Infection Risk", "AFS", "X-ray Ratio"),
  MSE = mse
) %>%
  kable(caption = "MSE for each of the three predictor variables")
```

Table 2: MSE for each of the three predictor variables

Predictor	MSE
Infection Risk	2.6375
AFS	3.2206
X-ray Ratio	3.1473

From the table, it is seen that the predictor variable, `infection risk`, with the lowest MSE of `2.6375` leads to the smallest variability around the fitted regression line.

Problem 1.46

Part (a)

```
models <- list() # initialize empty list

# Get the various regions
regions <- as.integer(levels(as.factor(senic$region)))

# fit individual models for each region
for(i in regions) {
  models[[i]] = senic %>%
    filter(region == i) %>%
    lm(LOS ~ infec_risk, data = .)
  print(paste("Estimated model summary for region", i))
  print(summary(models[[i]]))
}
```

```
## [1] "Estimated model summary for region 1"
##
## Call:
## lm(formula = LOS ~ infec_risk, data = .)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.1887 -1.0713 -0.3449  0.6822  6.2617
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   4.5379     1.5852   2.863 0.008193 **
## infec_risk    1.3477     0.3159   4.267 0.000233 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.086 on 26 degrees of freedom
## Multiple R-squared:  0.4118, Adjusted R-squared:  0.3892
## F-statistic:  18.2 on 1 and 26 DF,  p-value: 0.0002326
##
## [1] "Estimated model summary for region 2"
##
## Call:
## lm(formula = LOS ~ infec_risk, data = .)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.1998 -0.6047  0.1244  0.7435  1.8283
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   7.5605     0.6267  12.063 4.89e-13 ***
## infec_risk    0.4832     0.1366   3.536  0.00134 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## Residual standard error: 1.019 on 30 degrees of freedom
## Multiple R-squared:  0.2942, Adjusted R-squared:  0.2707
## F-statistic: 12.51 on 1 and 30 DF,  p-value: 0.001341
##
## [1] "Estimated model summary for region 3"
##
## Call:
## lm(formula = LOS ~ infec_risk, data = .)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.9748 -0.4921 -0.2071  0.2900  2.4954
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   7.1293     0.4632  15.393  < 2e-16 ***
## infec_risk    0.5251     0.1107   4.742 3.49e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.9694 on 35 degrees of freedom
## Multiple R-squared:  0.3911, Adjusted R-squared:  0.3737
## F-statistic: 22.48 on 1 and 35 DF,  p-value: 3.494e-05
##
## [1] "Estimated model summary for region 4"
##
## Call:
## lm(formula = LOS ~ infec_risk, data = .)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.4158 -0.6716 -0.3077  0.6918  2.0425
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  8.03805    1.36481   5.889 3.94e-05 ***
## infec_risk   0.01728    0.30583   0.056    0.956
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.038 on 14 degrees of freedom
## Multiple R-squared:  0.0002279,  Adjusted R-squared:  -0.07118
## F-statistic: 0.003192 on 1 and 14 DF,  p-value: 0.9557
```

Estimated Regression Functions for each Region

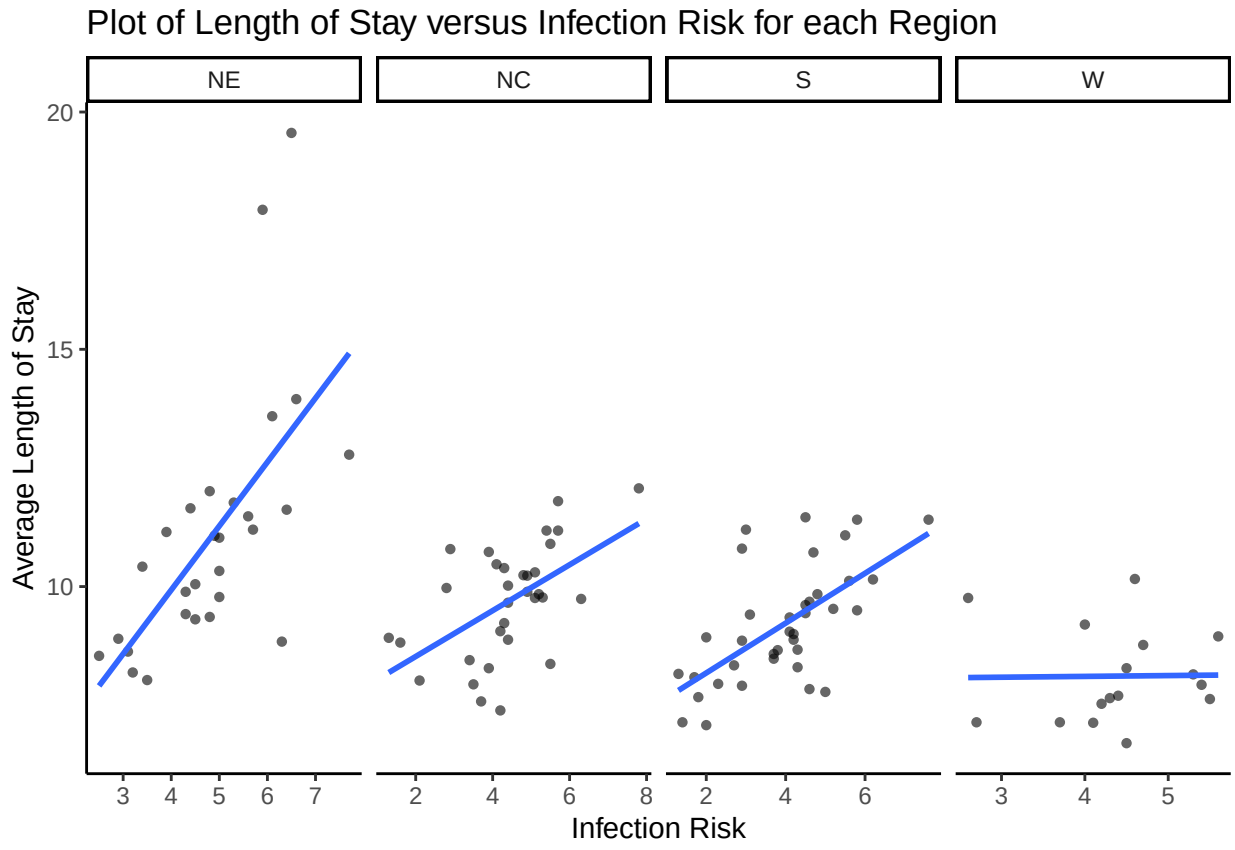
From the above outputs, for each region, the estimated regression function of average length of stay and infection risk (x) is:

- Region 1 (NE): $\hat{y} = 4.54 + 1.35x$.
- Region 2 (NC): $\hat{y} = 7.56 + 0.48x$.
- Region 3 (S): $\hat{y} = 7.13 + 0.53x$.
- Region 4 (W): $\hat{y} = 8.04 + 0.02x$.

Part (b)

With the help of the estimated models in part (a) and the graph below, we can say that the estimated regression functions are not similar for the four regions. This is because although there is positive linear relationship between infection risk and LOS across all regions, the effect of infection risk on LOS differ substantially from region to region. For instance, for every unit increase in infection risk we expect LOS to increase by `1.35` for the `NE` region, while the same unit increase in infection risk leads to only `0.02` increase in LOS on average for the `NC` region.

```
senic %>%
  mutate(region2 = factor(region,
                          levels = 1:4,
                          labels = c("NE", "NC", "S", "W"))
  ) %>%
  ggplot(aes(infec_risk, LOS)) +
  geom_point(alpha = 0.6, shape = 16) +
  facet_grid(cols = vars(region2), scales = "free_x")+
  labs(x="Infection Risk", y="Average Length of Stay",
       title = "Plot of Length of Stay versus Infection Risk for each Region") +
  theme_classic() +
  geom_smooth(method = "lm", se=FALSE)
```



Part (c): Calculate MSE

```
mse <- c(round(anova(models[[1]])$`Mean Sq`[2], 4),
         round(anova(models[[2]])$`Mean Sq`[2], 4),
         round(anova(models[[3]])$`Mean Sq`[2], 4),
```

```
        round(anova(models[[4]])$`Mean Sq`[2], 4)
        )
data.frame(
  Region = c("NE", "NC", "S", "W"),
  MSE = mse
) %>%
  kable(caption = "MSE for each of the four regions")
```

Table 3: MSE for each of the four regions

Region	MSE
NE	4.3531
NC	1.0379
S	0.9398
W	1.0779

Clearly, the variability around the fitted regression line is not the same for all the four regressions. However, we observe that three regions (`NC`, `S`, `W`) have approximately the same variability around `1` which differ greatly from that of the `NE` region. In all, the S region provides the smallest amount of variability.