STAT 561: HW1: Linear Regression Model

Prashanna Raj Pandit | Nazma Vali Shaik | Hema Sai Paruchuri

2025-04-02

Question Number 1:

Load and Visualize Data

```
pat_sat <- read.table("pat_stat.txt", header = TRUE)
head(pat_sat)
```

```
##   pat_sat pat_age severity anxiety
## 1      48      50       51     2.3
## 2      57      36       46     2.3
## 3      66      40       48     2.2
## 4      70      41       44     1.8
## 5      89      28       43     1.8
## 6      36      49       54     2.9
```

```
str(pat_sat)
```

```
## 'data.frame':    46 obs. of  4 variables:
##  $ pat_sat : int  48 57 66 70 89 36 46 54 26 77 ...
##  $ pat_age : int  50 36 40 41 28 49 42 45 52 29 ...
##  $ severity: int  51 46 48 44 43 54 50 48 62 50 ...
##  $ anxiety : num  2.3 2.3 2.2 1.8 1.8 2.9 2.2 2.4 2.9 2.1 ...
```

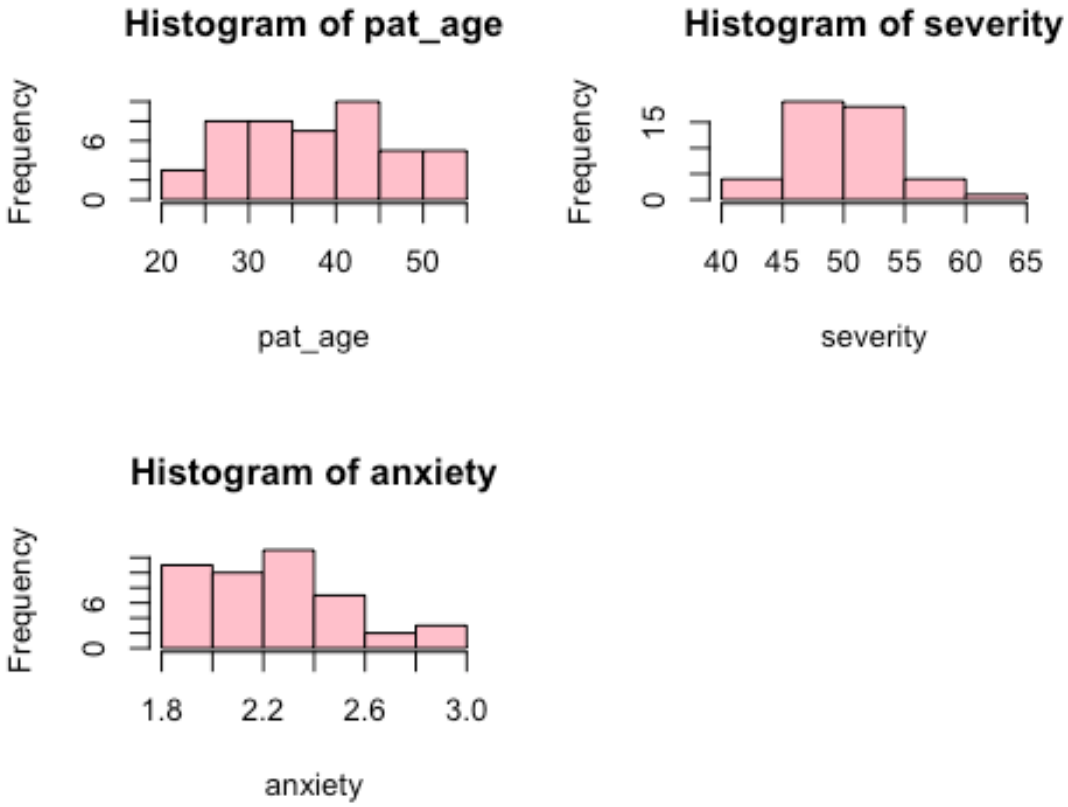
1. (a) Histogram and Box Plot

Prepare a histogram and box plot for each of the predictor variables using the hist() and boxplot() functions in R. Also use summary() to generate summaries for each of the predictor variables (do not produce the summary results). Are any noteworthy features revealed by these plots and your exploration?

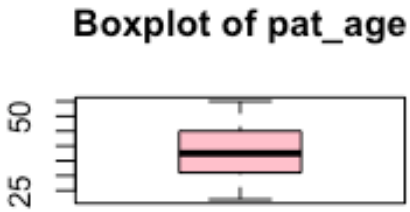
In histogram, Severity is approximately symmetric (normal distribution), centered around 50–55. The patient age somehow appears slightly right-skewed, with most patients aged between 20 and 45. The anxiety score looks right skew but they are symmetrically distributed with low variability. The mean (2.287) and median (2.300) are very close, reinforcing symmetry.

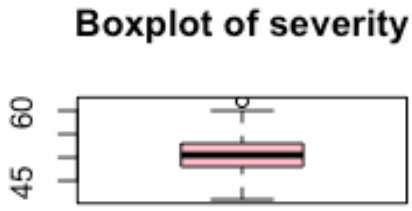
In box plot, The patient age ranges from 22 to 55 with median around 37.The value of severity ranges from approximately 40 to 60 with median at 50. There is also an outlier present in severity. The values of anxiety ranges from approximately 1.8 to 2.9 with median at 2.3.

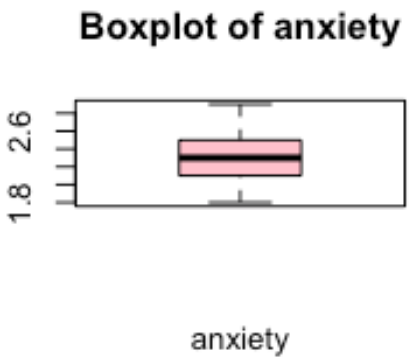
```r
predictor <- colnames(pat_sat)
par(mfrow = c(2, 2))  # Set layout to display multiple plots
for (col in predictor[2:length(predictor)]) {
  hist(pat_sat[[col]], main = paste("Histogram of", col),
       xlab = col, col = "pink", border = "black")
}
```




```r
# Box Plot
par(mfrow = c(2, 2))
for (col in predictor[2:length(predictor)]) {
  boxplot(pat_sat[[col]], main = paste("Boxplot of", col),
          xlab = col, col = "pink", border = "black")
}
```







Summary

```
summary(pat_sat[c("pat_age", "severity", "anxiety")])
```

```
##     pat_age         severity        anxiety
##  Min.   :22.00   Min.   :41.00   Min.   :1.800
##  1st Qu.:31.25   1st Qu.:48.00   1st Qu.:2.100
##  Median :37.50   Median :50.50   Median :2.300
##  Mean   :38.39   Mean   :50.43   Mean   :2.287
##  3rd Qu.:44.75   3rd Qu.:53.00   3rd Qu.:2.475
##  Max.   :55.00   Max.   :62.00   Max.   :2.900
```

1. (b) Scatter Plot Matrix and Correlation Matrix

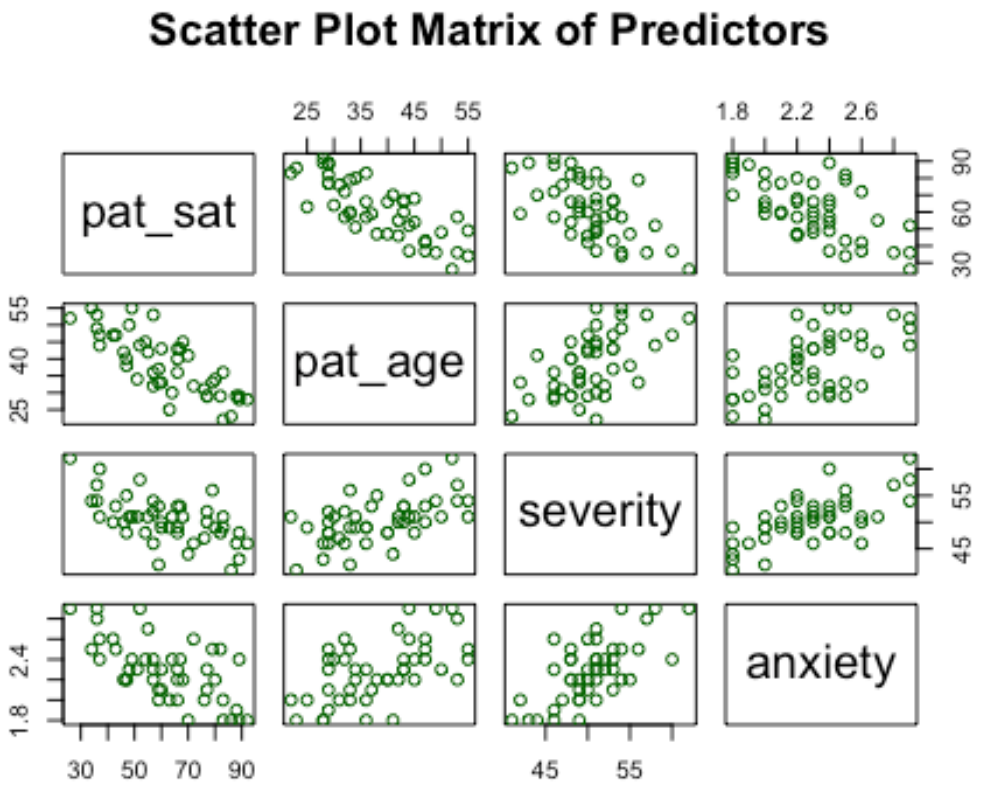
Interpretation:

- patient satisfaction is inversely proportional with all of the predictors (patient age, severity and anxiety) which means the patient satisfaction is decreasing with the increase in each of predictors.
- The patient age has positive linear relationship with the severity and anxiety, which means severity and anxiety is increasing with the increase in age.

- severity and anxiety are is also positively linear relation with each other.

Extreme observations: The relationship between the predictors looks positive linear whereas the relationship between predictors and target is negative linear (inversely proportional)

```
numeric_cols <- sapply(pat_sat, is.numeric)
numeric_data <- pat_sat[, numeric_cols]
pairs(numeric_data, main = "Scatter Plot Matrix of Predictors",
col="darkgreen")
```




```
correlation_matrix <- cor(pat_sat)
print(correlation_matrix)

##             pat_sat    pat_age   severity    anxiety
## pat_sat   1.0000000 -0.7867555 -0.6029417 -0.6445910
## pat_age  -0.7867555  1.0000000  0.5679505  0.5696775
## severity -0.6029417  0.5679505  1.0000000  0.6705287
## anxiety  -0.6445910  0.5696775  0.6705287  1.0000000
```

1. (c) Multiple Linear Regression Model

The estimated regression function based on the multiple linear regression model is:

Y=158.4913−1.1416(pat_age)−0.4420(severity)−13.4702(anxiety)

Interpretation: The coefficient β2 =−0.4420 means that for each one-unit increase in the severity index (X_2), patient satisfaction (Y) is expected to decrease by 0.4420 units, assuming all other variables (age and anxiety) are held constant.

```
multi_reg_model <- lm(pat_sat ~ pat_age + severity + anxiety, data = pat_sat)
summary(multi_reg_model)
```

```
##
## Call:
## lm(formula = pat_sat ~ pat_age + severity + anxiety, data = pat_sat)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -18.3524  -6.4230   0.5196   8.3715  17.1601
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 158.4913    18.1259   8.744 5.26e-11 ***
## pat_age      -1.1416     0.2148  -5.315 3.81e-06 ***
## severity     -0.4420     0.4920  -0.898   0.3741
## anxiety     -13.4702     7.0997  -1.897   0.0647 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 10.06 on 42 degrees of freedom
## Multiple R-squared:  0.6822, Adjusted R-squared:  0.6595
## F-statistic: 30.05 on 3 and 42 DF,  p-value: 1.542e-10
```

1. (d) Significance of overall model

Null Hypothesis (H_0): The model with all predictors (pat_age, severity, anxiety) does not significantly explain patient satisfaction.

Ho:β1=β2=β3=0

Alternative Hypothesis (H_1): At least one of the predictors has a significant relationship with patient satisfaction.

H1:At least one βi != 0

p-value: 1.542e-10

Since the p-value (1.542e-10) is much smaller than the significance level (α = 0.05), we reject the null hypothesis (H_0). The overall model is statistically significant, meaning at least one of the predictors (pat_age, severity, anxiety) is related to patient satisfaction.

1. (e) Confidence Interval

The 90% confidence interval for β1 (the coefficient for pat_age) is [-1.5029, -0.7803].

It suggests that pat_age has a statistically significant negative effect on patient satisfaction at the 90% confidence level. This means we are 90% confident that for every one-year increase in patient age, satisfaction decreases by between 0.78 and 1.50 units, on average, while holding other factors constant.

```
confint(multi_reg_model, level = 0.9)
```

```
##                     5 %        95 %
## (Intercept) 128.004370 188.9781330
## pat_age      -1.502893  -0.7803305
## severity     -1.269467   0.3854587
## anxiety     -25.411454  -1.5288719
```

1. (f) Coefficient of multiple determination

The coefficient of multiple determination value produced by our model is $R^2 = 0.6822$ means that 68.22% of the variation in patient satisfaction (pat_sat) is explained by the predictor variables (pat_age, severity, and anxiety) in the model.

```
# R² value from model summary
summary(multi_reg_model)$r.squared
```

```
## [1] 0.6821943
```

1. (g) Prediction on new data

Interpreting prediction interval:

The predicted patient satisfaction for this new patient is 69.01. Based on a 90% prediction interval, we expect that a new observation of patient satisfaction for similar patients will fall between 48.01 and 90.01 with 90% confidence.

```
new_data <- data.frame(pat_age = 35, severity = 45, anxiety = 2.2)
predicted_value <- predict(multi_reg_model, newdata = new_data, interval =
"prediction")
print(predicted_value)
```

```
##        fit      lwr      upr
## 1 69.01029 48.01224 90.00833
```

1. (h) Forward and backward selection

Yes! Both forward and backward selection produced the same final model. This means both methods agree that “severity” does not contribute significantly to the model and should be removed.

```
pat_sat ~ pat_age + anxiety
```

They produced identical coefficients and AIC values (AIC = 347.603), indicating they agree on the best model.
This model offers a good balance between explanatory power and model simplicity.

```
# Begin with the null model
null_model <- lm(pat_sat ~ 1, data = pat_sat)
summary(null_model)

##
## Call:
## lm(formula = pat_sat ~ 1, data = pat_sat)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -35.565 -13.315  -1.565  15.185  30.435
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   61.565      2.541   24.23   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 17.24 on 45 degrees of freedom

full_model <- lm(pat_sat ~ ., data = pat_sat)

forward <- step(null_model, direction = 'forward',
                scope = formula(full_model), trace = 1)

## Start:  AIC=262.92
## pat_sat ~ 1
##
##            Df Sum of Sq     RSS    AIC
## + pat_age   1    8275.4  5093.9 220.53
## + anxiety   1    5554.9  7814.4 240.21
## + severity  1    4860.3  8509.0 244.13
## <none>                  13369.3 262.92
##
## Step:  AIC=220.53
## pat_sat ~ pat_age
##
##            Df Sum of Sq    RSS    AIC
## + anxiety   1    763.42 4330.5 215.06
## + severity  1    480.92 4613.0 217.97
## <none>                  5093.9 220.53
##
## Step:  AIC=215.06
## pat_sat ~ pat_age + anxiety
##
##            Df Sum of Sq    RSS    AIC
```

```
## <none>                  4330.5 215.06
## + severity  1    81.659 4248.8 216.19

forward$coefficients

## (Intercept)     pat_age     anxiety
##  145.941228   -1.200471  -16.742052

backward <- step(full_model, direction = 'backward',
                 scope = formula(full_model), trace = 1)

## Start:  AIC=216.18
## pat_sat ~ pat_age + severity + anxiety
##
##            Df Sum of Sq    RSS    AIC
## - severity  1     81.66 4330.5 215.06
## <none>                  4248.8 216.19
## - anxiety   1    364.16 4613.0 217.97
## - pat_age   1   2857.55 7106.4 237.84
##
## Step:  AIC=215.06
## pat_sat ~ pat_age + anxiety
##
##           Df Sum of Sq    RSS    AIC
## <none>                 4330.5 215.06
## - anxiety  1     763.4 5093.9 220.53
## - pat_age  1    3483.9 7814.4 240.21

backward$coefficients

## (Intercept)     pat_age     anxiety
##  145.941228   -1.200471  -16.742052

# Compare AIC values
AIC(forward)

## [1] 347.603

AIC(backward)

## [1] 347.603
```

Question Number: 2

2. (a) Correlation between age and muscle

Interpretation: The correlation coefficient (-0.866064) indicates a strong negative correlation between age and muscle mass. This means that as age increases, muscle mass tends to decrease.

```r
muscle_mass <- read.table("muscle_mass.txt", header = TRUE)
# visualizing first five data
head(muscle_mass)
```

```
##   mmass age
## 1   106  43
## 2   106  41
## 3    97  47
## 4   113  46
## 5    96  45
## 6   119  41
```

```r
# calculating correlation
cor(muscle_mass)
```

```
##           mmass       age
## mmass  1.000000 -0.866064
## age   -0.866064  1.000000
```

2.(b) First Order Model

Interpretation: Q. Good Fit?

R^2 =0.7501 close to 1 indicates a good fit, meaning the model explains most of the variance in Y.

```r
first_order_model <- lm(mmass ~ age, data = muscle_mass)
summary(first_order_model)
```

```
##
## Call:
## lm(formula = mmass ~ age, data = muscle_mass)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -16.1368  -6.1968  -0.5969   6.7607  23.4731
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 156.3466     5.5123   28.36   <2e-16 ***
## age          -1.1900     0.0902  -13.19   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8.173 on 58 degrees of freedom
## Multiple R-squared:  0.7501, Adjusted R-squared:  0.7458
## F-statistic: 174.1 on 1 and 58 DF,  p-value: < 2.2e-16
```

```r
plot(muscle_mass$age, muscle_mass$mmass, main = " First Order Regression
Model",
```

```
     xlab = "Age", ylab = "Muscle Mass", pch = 19, col = "blue")
abline(first_order_model, col = "red")
```

2. (c) Second Order Model

```
second_order_model <- lm(mmass ~ age + I(age^2), data = muscle_mass)
summary(second_order_model)
```

```
##
## Call:
## lm(formula = mmass ~ age + I(age^2), data = muscle_mass)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -15.086  -6.154  -1.088   6.220  20.578
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) 207.349608  29.225118   7.095 2.21e-09 ***
## age          -2.964323   1.003031  -2.955  0.00453 **
## I(age^2)      0.014840   0.008357   1.776  0.08109 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## Residual standard error: 8.026 on 57 degrees of freedom
## Multiple R-squared:  0.7632, Adjusted R-squared:  0.7549
## F-statistic: 91.84 on 2 and 57 DF,  p-value: < 2.2e-16
```

2. (d) Plot regression function (a) and (b) together

The second-order model is a better fit because it has a higher R-squared (0.7632 vs 0.7501 in first-order) and a lower residual standard error (8.026 vs 0.8173 in first-order), indicating improved explanatory power and better fit to the data.

```
plot(muscle_mass$age, muscle_mass$mmass,
     main = "First and second Order regression model",
     xlab = "Age", ylab = "Muscle Mass",
     pch = 19, col = "blue")

abline(first_order_model, col = "red", lwd = 2)  # Red line for linear model

age_seq <- seq(min(muscle_mass$age), max(muscle_mass$age), length.out = 100)
mmass_pred_quadratic <- predict(second_order_model, newdata = data.frame(age
= age_seq))
lines(age_seq, mmass_pred_quadratic, col = "green", lwd = 2)  # Green line
for quadratic model

legend("topright", legend = c("Linear Fit", "Quadratic Fit"),
       col = c("red", "green"), lty = 1, lwd = 2)
```

2. (e) Testing regression relation for the model (b)

Interpretation: The overall model is significant since the F-statistic p-value is < 2.2e-16, which is much smaller than 0.05. This means the second-order model provides a statistically significant regression relationship on the basis that there is at least one predictor which is significant.

2. (f)

Interpretation: Since p = 0.08109 is greater alpha=0.05 we fail to reject the null hypothesis. This means the quadratic term does not significantly improve the model at the 5% significance level. Therefore, we can drop the quadratic term and consider using the first-order model instead.

2.(g) Third Order Model

Interpretation: The p-value of the cubic term β111= 0.719 which is very > than alpha=0.05, meaning we fail to reject the null hypothesis. This suggest that the cubic term does not significantly contribute the model.

```r
third_order_model <- lm(mmass ~ age + I(age^2) + I(age^3), data =
muscle_mass)
summary(third_order_model)
```

```
##
## Call:
## lm(formula = mmass ~ age + I(age^2) + I(age^3), data = muscle_mass)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -15.3671  -5.8483  -0.6755   6.1376  20.0637
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.404e+02  1.877e+02   0.748    0.458
## age          5.648e-01  9.822e+00   0.058    0.954
## I(age^2)    -4.559e-02  1.675e-01  -0.272    0.786
## I(age^3)     3.369e-04  9.327e-04   0.361    0.719
##
## Residual standard error: 8.087 on 56 degrees of freedom
## Multiple R-squared:  0.7637, Adjusted R-squared:  0.7511
## F-statistic: 60.34 on 3 and 56 DF,  p-value: < 2.2e-16
```

Question Number 3: Qualitative predictors

```r
# Load the data
cdi <- read.table("cdi.txt", header = TRUE)
cdi <- cdi[, -1]  # Remove the first column

# View structure and unique values of geographic_region
str(cdi)
```

```
## 'data.frame':    440 obs. of  16 variables:
##  $ county                        : chr  "Los_Angeles" "Cook" "Harris"
"San_Diego" ...
##  $ state                         : chr  "CA" "IL" "TX" "CA" ...
##  $ land_area_sq_mi               : int  4060 946 1729 4205 790 71 9204 614
1945 880 ...
##  $ total_population              : int  8863164 5105067 2818199 2498016
2410556 2300664 2122101 2111687 1937094 1852810 ...
##  $ percent_population_18_34      : num  32.1 29.2 31.3 33.5 32.6 28.3 29.2
27.4 27.1 32.6 ...
##  $ percent_population_65_plus    : num  9.7 12.4 7.1 10.9 9.2 12.4 12.5
12.5 13.9 8.2 ...
##  $ number_active_physicians      : int  23677 15153 7553 5905 6062 4861
4320 3823 6274 4718 ...
##  $ number_hospital_beds          : int  27700 21550 12449 6179 6369 8942
6104 9490 8840 6934 ...
##  $ total_serious_crimes          : int  688936 436936 253526 173821 144524
680966 177593 193978 244725 214258 ...
##  $ percent_high_school_graduates : num  70 73.4 74.9 81.9 81.2 63.7 81.5
```

```
70 65 77.1 ...
##  $ percent_bachelors_degrees     : num  22.3 22.8 25.4 25.3 27.8 16.6 22.1
13.7 18.8 26.3 ...
##  $ percent_below_poverty_level   : num  11.6 11.1 12.5 8.1 5.2 19.5 8.8
16.9 14.2 10.4 ...
##  $ percent_unemployment          : num  8 7.2 5.7 6.1 4.8 9.5 4.9 10 8.7
6.1 ...
##  $ per_capita_income             : int  20786 21729 19517 19588 24400
16803 18042 17461 17823 21001 ...
##  $ total_personal_income_millions: int  184230 110928 55003 48931 58818
38658 38287 36872 34525 38911 ...
##  $ geographic_region             : int  4 2 3 4 4 1 4 2 3 3 ...
```

```
unique(cdi$geographic_region)
```

```
## [1] 4 2 3 1
```

```
# Since we have 4 geographical regions, assume:
# 1 = Northeast (Baseline)
# 2 = North Central (X3)
# 3 = South (X4)
# 4 = West (X5)
```

```
## Convert Geographic Region to Factor
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
cdi$geographic_region <-factor(cdi$geographic_region)
```

3.(a) Fit Multiple Linear Regression Model

Regression Equation:

Y = B0 + B1X1 + B2X2 + B3X3 + B4X4 + B5X5 + E

Where:
- **Y** = number_active_physicians (Dependent Variable)
- **X1** = total_population
- **X2** = total_personal_income_millions
- **X3, X4, X5** represent the geographic regions, encoded as dummy variables:

- **X3** = 1 if the region is **North Central**, 0 otherwise
- **X4** = 1 if the region is **South**, 0 otherwise
- **X5** = 1 if the region is **West**, 0 otherwise

The **Northeast** region is the reference category, meaning it is represented when **X3 = X4 = X5 = 0**.

Encoding of Geographic Region Dummy Variables

Geographic.Region	X3...North.Central	X4...South	X5...West
Northeast	0	0	0
North Central	1	0	0
South	0	1	0
West	0	0	1

```
model <- lm(number_active_physicians ~ total_population +
total_personal_income_millions + geographic_region, data = cdi)

## Rename Coefficients
names(model$coefficients) <- gsub("geographic_region2", "X3",
names(model$coefficients))
names(model$coefficients) <- gsub("geographic_region3", "X4",
names(model$coefficients))
names(model$coefficients) <- gsub("geographic_region4", "X5",
names(model$coefficients))
```

3. (b) Coefficients β2 and β3

- B2 (Total Personal Income) = 0.107

Represents the change in the number of active physicians for each additional unit increase in total personal income (in millions).

Interpretation: For every 1 million dollar increase in total personal income, the number of active physicians increases by 0.107, assuming all other factors remain constant.

The p-value (< 0.001) indicates strong statistical significance, meaning the relationship between personal income and physicians is robust.

- B3 (North Central Region) = -3.493

Represents the difference in the number of active physicians between the North Central and the reference category (Northeast).

Interpretation: Compared to the Northeast (baseline region), being in the North Central is associated with 3.493 fewer active physicians, holding total personal income and population constant.

However, the p-value (0.9647) is very high, indicating that this effect is not statistically significant, meaning there is no strong evidence that the number of physicians is truly different in the North Central region compared to the Northeast.

```
summary(model)
```

```
##
## Call:
## lm(formula = number_active_physicians ~ total_population +
total_personal_income_millions +
##     geographic_region, data = cdi)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1866.8  -207.7   -81.5    72.4  3721.7
##
## Coefficients:
##                                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)                   -5.848e+01  5.882e+01  -0.994   0.3207
## total_population               5.515e-04  2.835e-04   1.945   0.0524 .
## total_personal_income_millions  1.070e-01  1.325e-02   8.073  6.8e-15 ***
## X3                            -3.493e+00  7.881e+01  -0.044   0.9647
## X4                             4.220e+01  7.402e+01   0.570   0.5689
## X5                            -1.490e+02  8.683e+01  -1.716   0.0868 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 566.1 on 434 degrees of freedom
## Multiple R-squared:  0.9011, Adjusted R-squared:  0.8999
## F-statistic: 790.7 on 5 and 434 DF,  p-value: < 2.2e-16
```