STAT Assignment HW1

Asha Shah | Prabhath Pasula | Rithwik Reddy Nandyala

2025-04-06

Load necessary libraries

```
library(MASS)
library(ISLR2)
```

```
##
## Attaching package: 'ISLR2'

## The following object is masked from 'package:MASS':
##
##     Boston
```

#____#

PATIENT SATISFACTION ANALYSIS

#____#

Load data from file

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

Initial exploration

```
# View the first 10 rows
head(pat_sat, n=10)
```

```
##    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
## 7       46      42       50     2.2
## 8       54      45       48     2.4
## 9       26      52       62     2.9
## 10      77      29       50     2.1
```

```r
# View the last 10 rows
tail(pat_sat, n=10)
```

```
##    pat_sat pat_age severity anxiety
## 37      82      29       48     2.5
## 38      64      30       51     2.4
## 39      37      47       60     2.4
## 40      42      47       50     2.6
## 41      66      43       53     2.3
## 42      83      22       51     2.0
## 43      37      44       51     2.6
## 44      68      45       51     2.2
## 45      59      37       53     2.1
## 46      92      28       46     1.8
```

```r
# Summary statistics for the dataset
summary(pat_sat)
```

```
##     pat_sat         pat_age         severity        anxiety
##  Min.   :26.00   Min.   :22.00   Min.   :41.00   Min.   :1.800
##  1st Qu.:48.25   1st Qu.:31.25   1st Qu.:48.00   1st Qu.:2.100
##  Median :60.00   Median :37.50   Median :50.50   Median :2.300
##  Mean   :61.57   Mean   :38.39   Mean   :50.43   Mean   :2.287
##  3rd Qu.:76.75   3rd Qu.:44.75   3rd Qu.:53.00   3rd Qu.:2.475
##  Max.   :92.00   Max.   :55.00   Max.   :62.00   Max.   :2.900
```

```r
# Structure of the dataset - to check the types of variables
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 ...
```

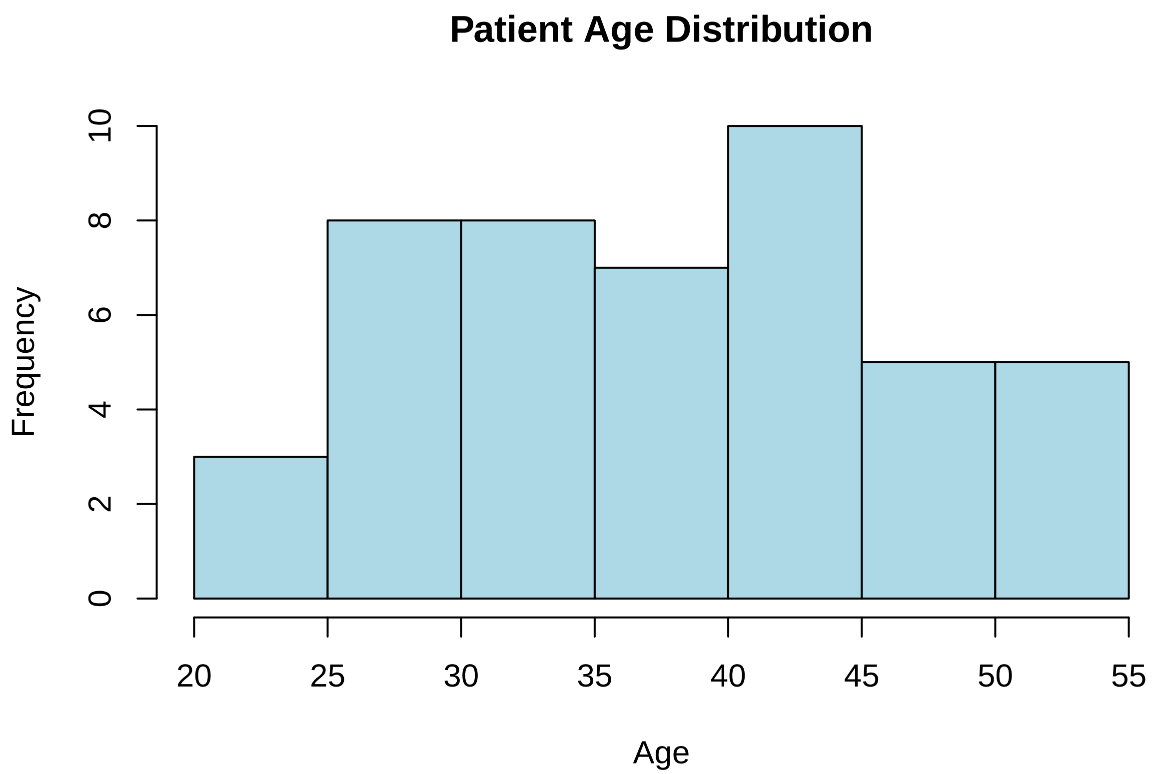
```r
# Dimensions of the dataset - to know the number of rows and columns
dim(pat_sat)
```

```
## [1] 46  4
```

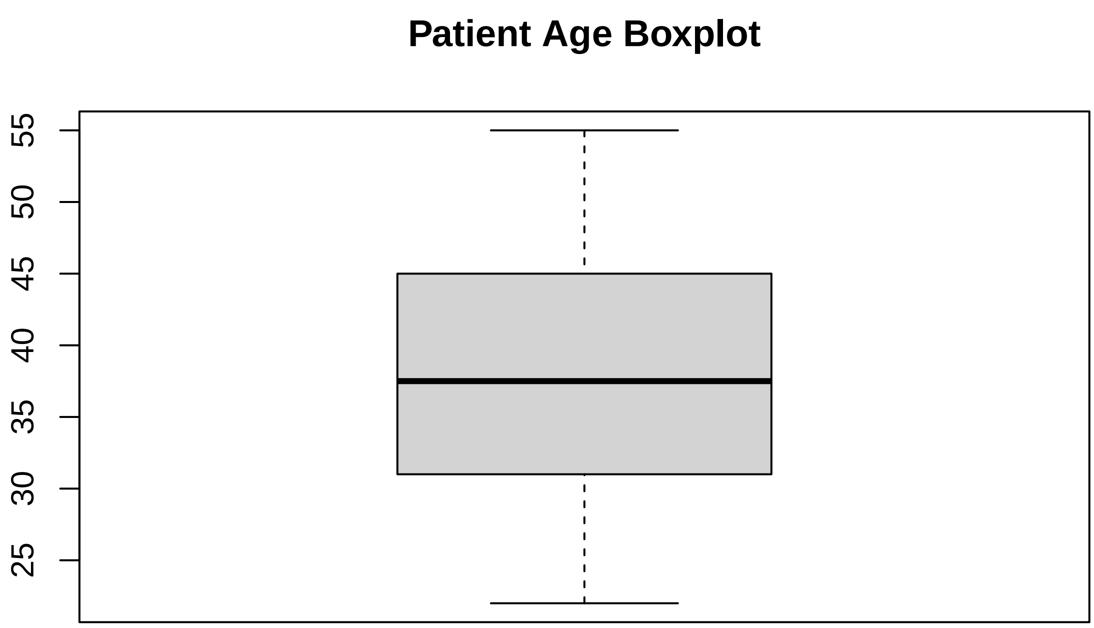
Part 1a: Histograms and Boxplots

```r
### Question: Prepare a histogram and box plot for each of the predictor variables using the hist()
#and boxplot() functions in R.

#Explanation
#Histograms: The histograms provide insight into the shape of each variable's distribution, such as
#whether it is symmetric, skewed, or multimodal. They show where most of the data points lie
#(central tendency), how spread out they are (variability), and whether there are potential extreme
#values.

#Boxplot: The boxplots complement this by visually summarizing the five-number summary (minimum, Q1,
#median, Q3, maximum) and clearly identifying potential outliers. They help assess skewness and
#compare spread between variables.
```

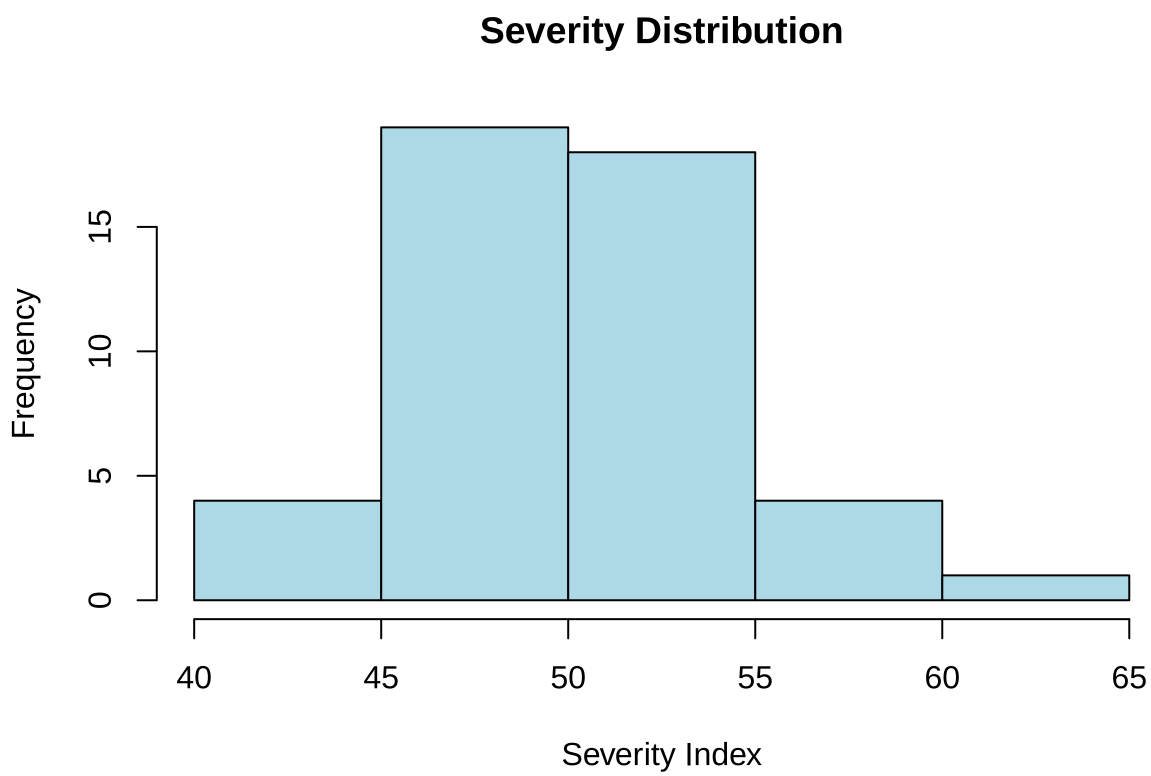
```r
# Histogram for Patient Age
hist(pat_sat$pat_age, col="lightblue", main="Patient Age Distribution", xlab="Age")
```



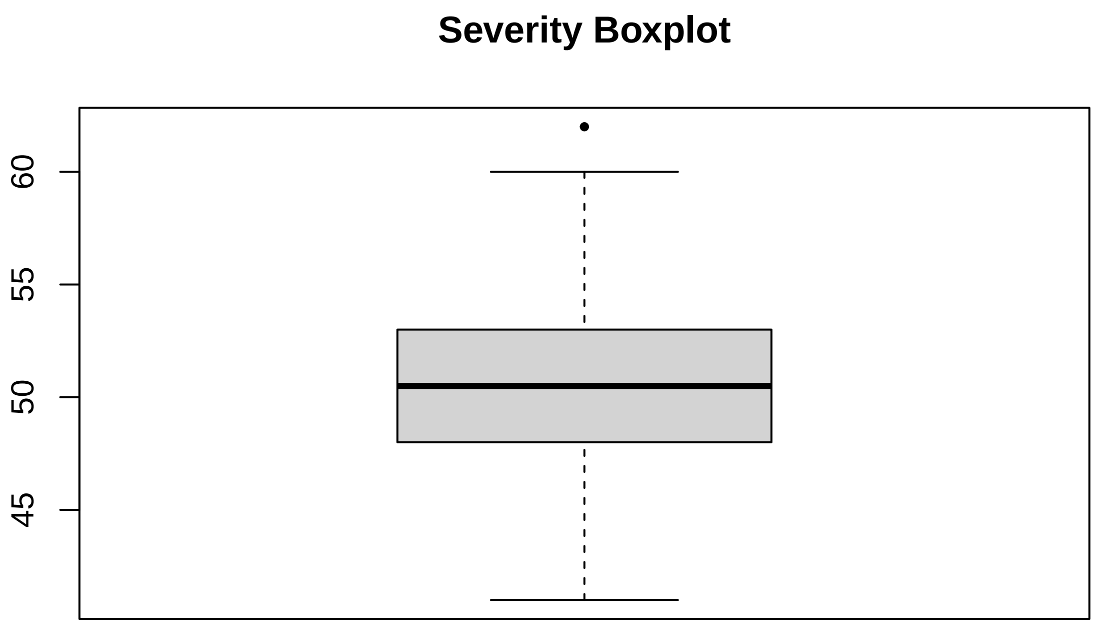

```r
# Boxplot for Patient Age
boxplot(pat_sat$pat_age, col="lightgrey", pch=19, cex=0.5, main="Patient Age Boxplot")
```



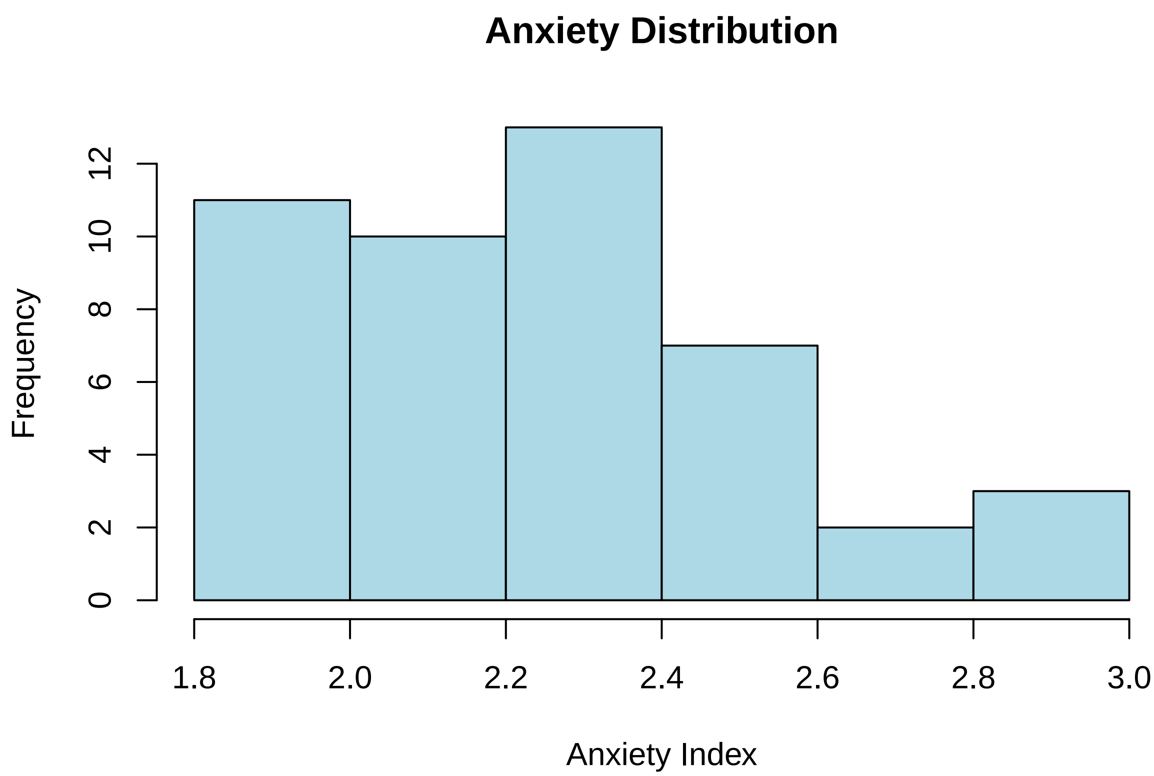

```r
# Histogram for Severity
hist(pat_sat$severity, col="lightblue", main="Severity Distribution", xlab="Severity Index")
```



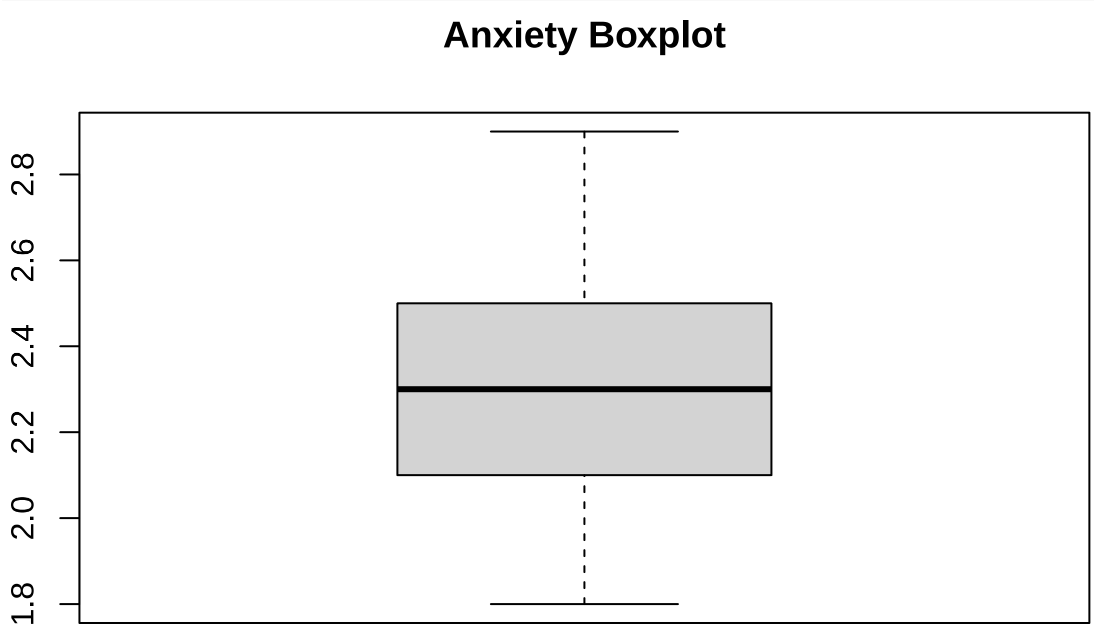

```r
# Boxplot for Severity
boxplot(pat_sat$severity, col="lightgrey", pch=19, cex=0.5, main="Severity Boxplot")
```




```r
# Histogram for Anxiety
hist(pat_sat$anxiety, col="lightblue", main="Anxiety Distribution", xlab="Anxiety Index")
```




```r
# Boxplot for Anxiety
boxplot(pat_sat$anxiety, col="lightgrey", pch=19, cex=0.5, main="Anxiety Boxplot")
```




```r
# Summary of each predictor variable
cat ("Summary of Patient Age: \n")
```

```
## Summary of Patient Age:
```

```r
summary(pat_sat$pat_age)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
```

```
##   22.00   31.25   37.50   38.39   44.75   55.00
```

```r
cat ("Summary of Severrity:\n")
```

```
## Summary of Severrity:
```

```r
summary(pat_sat$severity)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   41.00   48.00   50.50   50.43   53.00   62.00
```

```r
cat ("Summary of Anxiety:\n")
```

```
## Summary of Anxiety:
```

```r
summary(pat_sat$anxiety)
```

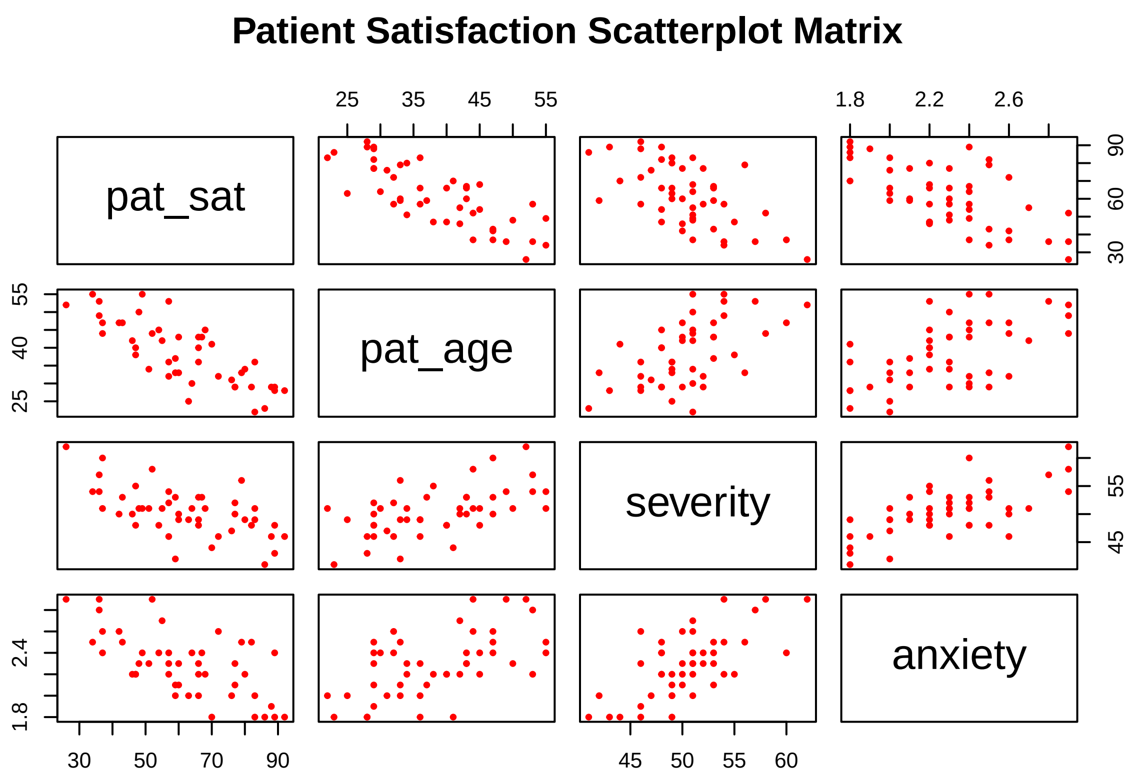
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   1.800   2.100   2.300   2.287   2.475   2.900
```

Observations for Part 1a

```
**Observations for Part 1a:**
Histograms:
- Patient Age: More younger patient than older, skewed slightly right.
- Severity: The histogram shows that most patients have severity levels between 45 and 55,
with the highest frequency around 50. There is a slight skewness towards higher severity
levels with fewer patients having levels above 55
- Anxiety: It shows that most patients have anxiety levels around 2.2 and 2.4. It also shows
a slight skewness towards higher anxiety levels.

Boxplot:
- Patient Age: The median patient age is around 38, with ages ranging from approximately 25 to 55
- Severity: The median severity index is around 50, with most values ranging from approximately 45
to 55, and a single outlier above 60
- Anxiety: The median anxiety index is around 2.3, with most values ranging from approximately 1.8
to 2.6, indicating no outliers
```

Part 1b: Scatterplot and Correlation Matrix

```r
### Question: Obtain the scatter plot matrix and the correlation matrix using the pairs()
#and cor() functions respectively.

# Scatter Plot: The scatter plot matrix shows pairwise scatter plots of all variables. It
#helps us visually assess relationships and correlations between variables, and identify
#any potential outliers.

# Correlation Matrix: The correlation matrix provides numerical values for the strength
#and direction of the relationships between variables. It helps us identify which
#variables are strongly or weakly correlated.

pairs(pat_sat, pch=19, cex=0.5, col="red", main="Patient Satisfaction Scatterplot Matrix")
```




```r
cor_matrix <- cor(pat_sat)
print(cor_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
```

Observations for Part 1b

```
The scatterplot matrix compares pat_sat (patient satisfaction), pat_age (patient age),
severity, and anxiety.

1. pat_sat vs. pat_age:
   - Negative correlation: As age increases, satisfaction decreases.

2. pat_sat vs. severity:
   - Negative correlation: Higher severity leads to lower satisfaction.

3. pat_sat vs. anxiety:
   - Negative correlation: Higher anxiety corresponds to lower satisfaction.

4. pat_age vs. severity:
   - Slight positive correlation: Older patients have higher severity levels.

5. pat_age vs. anxiety:
   - No strong correlation: Anxiety levels do not vary significantly with age.

6. severity vs. anxiety:
```

```
  - Positive correlation: Higher severity is associated with higher anxiety.

Conclusion:

- Negative Correlations:
  - pat_sat decreases with increasing pat_age, severity, and anxiety.

- Positive Correlations:
  - Positive relationship between severity and anxiety.
```

Part 1c: Multiple Linear Regression

```r
### Fit a multiple linear regression model for three predictor variables to the data and
#state the estimated regression function. How is  2 interpreted here?

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

```r
# Explanation: The multiple linear regression model helps us understand the relationship
#between patient satisfaction (Y) and the predictor variables (age, severity, and
#anxiety). The summary provides the estimated coefficients, p-values, and other statistics
#for the model.

anova(pat_model)
```

```
## Analysis of Variance Table
##
## Response: pat_sat
##           Df Sum Sq Mean Sq F value    Pr(>F)
## pat_age    1 8275.4  8275.4 81.8026 2.059e-11 ***
## severity   1  480.9   480.9  4.7539   0.03489 *
## anxiety    1  364.2   364.2  3.5997   0.06468 .
```

```
## Residuals 42 4248.8   101.2
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
# Explanation: The ANOVA table helps us assess the overall significance of the model.

# Estimated regression function
cat("\nEstimated Regression Function:\n")
```

```
##
## Estimated Regression Function:
```

```r
cat("pat_sat =",
    round(coef(pat_model)[1], 2), "+",
    round(coef(pat_model)[2], 2), "(age) +",
    round(coef(pat_model)[3], 2), "(severity) +",
    round(coef(pat_model)[4], 2), "(anxiety)\n\n")
```

```
## pat_sat = 158.49 + -1.14 (age) + -0.44 (severity) + -13.47 (anxiety)
```

```r
# Interpretation of   (severity coefficient)
cat("Interpretation of   (severity coefficient):\n")
```

```
## Interpretation of   (severity coefficient):
```

```r
cat("For each 1-unit increase in severity index,",
    "patient satisfaction decreases by", abs(round(coef(pat_model)[3], 2)),
    "points on average,",
    "holding age and anxiety level constant.\n")
```

```
## For each 1-unit increase in severity index, patient satisfaction decreases by 0.44 points on average
```

```
Estimated Regression Function:
pat_sat = 158.49 + -1.14 (age) + -0.44 (severity) + -13.47 (anxiety)

Interpretation of   (severity coefficient):
For each 1-unit increase in severity index, patient satisfaction decreases by 0.44 points
on average, holding age and anxiety level constant.
```

Part 1d: Model Significance (from summary output)

```r
###Question: Conduct a test to check if the overall model is significant; use   = .05.
#State the null and alternative hypotheses, p-value decision whether to reject H0 or to
#fail to reject H0, and your conclusion (Hint : Use the F-test.).

# Explanation: The F-statistic and p-value in the ANOVA table help us determine if the
#overall model is statistically significant.

anova_results <- anova(pat_model)

# Hypotheses
cat("Null Hypothesis (H0): The overall model is not significant.\n")
```

```
## Null Hypothesis (H0): The overall model is not significant.
```

```r
cat("Alternative Hypothesis (H1): The overall model is significant.\n")
```

```
## Alternative Hypothesis (H1): The overall model is significant.
```

```r
# Extract F-statistic and p-value from the ANOVA results
f_statistic <- anova_results$`F value`[1]
p_value <- anova_results$`Pr(>F)`[1]

# Print the F-statistic and p-value
cat("\nF-statistic:", f_statistic, "\n")
```

```
##
## F-statistic: 81.80263
```

```r
cat("p-value:", p_value, "\n")
```

```
## p-value: 2.059138e-11
```

```r
# Decision Rule and Conclusion
alpha <- 0.05
if (p_value < alpha) {
  cat("\nDecision: Reject the null hypothesis (H0) as p-value: 2.059138e-11 < alpha(0.05)\n")
  cat("Conclusion: There is significant evidence to conclude that the overall model is significant.\n")
} else {
  cat("\nDecision: Fail to reject the null hypothesis (H0).\n")
  cat("Conclusion: There is not enough evidence to conclude that the overall model is significant.\n")
}
```

```
##
## Decision: Reject the null hypothesis (H0) as p-value: 2.059138e-11 < alpha(0.05)
## Conclusion: There is significant evidence to conclude that the overall model is significant.
```

Part 1e: 90% Confidence Interval for 1

```r
###Question: Obtain a 90% confidence interval for  1 using the code below. Interpret your
#results.model <- lm(...) confint(model,level=0.9) #95% confidence intervals for the model
#coefficients

confint(pat_model, level=0.90)[2,]
```

```
##        5 %       95 %
## -1.5028932 -0.7803305
```

```r
# Explanation: The 90% confidence interval for  1 (age) provides a range within which we
#are 90% confident that the true value of the coefficient lies. It helps us understand the
#precision and reliability of the estimate.
```

Observations for Part 1e

```
This means that, with 90% confidence, the true value of  1 lies within this range. In
other words, for every 1-unit increase in pat_age, the predicted pat_sat decreases by
between 0.7803 and 1.5029 units.
```

Part 1f: Coefficient of Determination

```r
### Question: What is the coefficient of multiple determination value produced by your
#model (this is same as R2)?
```

```r
summary(pat_model)$r.squared
```

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

```r
# Explanation: The R-squared value indicates the proportion of the variance in the
#dependent variable (patient satisfaction) that is predictable from the independent
#variables (age, severity, anxiety). A higher R-squared value indicates a better fit of
#the model.
```

Observations for Part 1f

```
R² = 0.6822 means that approximately 68.22% of the variance in patient satisfaction
(pat_sat) can be explained by the predictor variables (patient age, severity of illness,
and anxiety level) in the regression model.

This indicates that model fits the data reasonably well, as it explains a substantial
portion of the variability in patient satisfaction. However, it also means that 31.78% of
the variability in patient satisfaction remains unexplained by the model, suggesting that
other factors not included in the model may influence patient satisfaction.
```

Part 1g: Prediction

```r
### Question: Predict the patient satisfaction for a new patient with X1 = 35, X2 = 45,
#and X3 = 2.2. Also give a 90 percent prediction interval for this new observation.

new_patient <- data.frame(pat_age=35, severity=45, anxiety=2.2)
prediction <- predict(pat_model, newdata=new_patient, interval="prediction", level=0.90)
print(prediction)
```

```
##        fit      lwr      upr
## 1 69.01029 51.50965 86.51092
```

```r
# Explanation: The prediction provides an estimate of patient satisfaction for a new
#patient with specified values for age, severity, and anxiety. The prediction interval
#gives a range within which we expect the true satisfaction score to lie with 90%
#confidence.
```

Observations for Part 1g

```
Predicted patient satisfaction (fit) = 69.01

Additionally, the 90% prediction interval for the predicted patient satisfaction is:
Lower bound (lwr) = 51.51
Upper bound (upr) = 86.51

This means that, with 90% confidence, the true patient satisfaction score for this new
patient would fall between 51.51 and 86.51. The prediction point estimate is 69.01, but
this range accounts for the uncertainty in the model's prediction.
```

Part 1h: Model Selection

```
### Question: Use both forward and backward selection criteria to select a final model.
```

Forward selection

```r
intercept_only <- lm(pat_sat ~ 1, data=pat_sat)
full_model <- lm(pat_sat ~ pat_age + severity + anxiety, data=pat_sat)
forward_model <- step(intercept_only, 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
```

```r
print(forward_model$coefficients)
```

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

```r
# Explanation: Forward selection starts with no predictors and adds predictors one-by-one
#based on some criterion (e.g., AIC) until no more predictors improve the model. This
#helps in identifying the most significant predictors.
```

Backward selection

```r
backward_model <- step(full_model, direction='backward', 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
```

```r
print(backward_model$coefficients)
```

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

```r
# Explanation: Backward selection starts with all predictors and removes the least
#significant predictors one-by-one based on some criterion (e.g., AIC) until no more
#predictors can be removed without worsening the model.
```

Observations for Part 1h

```
Interpretation:

Forward Selection Model:
The model selected towards the end after forward selection has pat_age and anxiety as the
predictors with an intercept term of 145.9412, a coefficient of -1.2005 for pat_age, and
-16.7421 for anxiety.

Backward Selection Model:
The model selected towards the end after backward selection has pat_age and anxiety but
with slightly different coefficients since variable deletion is involved.

Comparison of Models:

Forward Selection:
Model: pat_sat ~ pat_age + anxiety
AIC: 215.06
Coefficients: Intercept = 145.9412, pat_age = -1.2005, anxiety = -16.7421.

Backward Selection:
Model: pat_sat ~ pat_age + anxiety
AIC: 215.06
Coefficients: Intercept = 145.9412, pat_age = -1.2005, anxiety = -16.7421.

Conclusion:
Both forward and backward selection selected the same model: pat_sat ~ pat_age + anxiety,
having the same AIC values (215.06). Since both criteria result in the same model, they
offer the same balance between complexity and explanatory power. Since both models are
identical and yield the same AIC.

This means that pat_age and anxiety are the best predictors of pat_sat in this case, and
adding severity doesn't improve the model sufficiently (as can be seen from the higher AIC
when adding it). The lowest AIC value in this case is 215.06, and since it's the same for
both selection methods, it's your final model.
```

#____#

MUSCLE MASS ANALYSIS (Polynomial regression)

#____#

Load data from file

```
mmass <- read.table("muscle_mass.txt", header=TRUE)
View(mmass)
```

Initial exploration

```
head(mmass, n=10)
```

```
##    mmass age
## 1    106  43
## 2    106  41
## 3     97  47
## 4    113  46
## 5     96  45
## 6    119  41
## 7     92  47
## 8    112  41
## 9     92  48
## 10   102  48
```

```
summary(mmass)
```

```
##      mmass             age
##  Min.   : 52.00   Min.   :41.00
##  1st Qu.: 73.00   1st Qu.:50.25
##  Median : 84.00   Median :60.00
##  Mean   : 84.97   Mean   :59.98
##  3rd Qu.: 97.00   3rd Qu.:70.00
##  Max.   :119.00   Max.   :78.00
```

```
str(mmass)
```

```
## 'data.frame':    60 obs. of  2 variables:
##  $ mmass: int  106 106 97 113 96 119 92 112 92 102 ...
##  $ age  : int  43 41 47 46 45 41 47 41 48 48 ...
```

Part 2a: Correlation

```
### Question: What is the correlation between age and muscle mass measure?
correlation <- cor(mmass$age, mmass$mmass)
print(correlation)
```

```
## [1] -0.866064
```

```r
# Explanation: The correlation coefficient quantifies the strength and direction of the
#linear relationship between age and muscle mass. A negative value would indicate that
#muscle mass decreases as age increases.
```

Observations for Part 2a

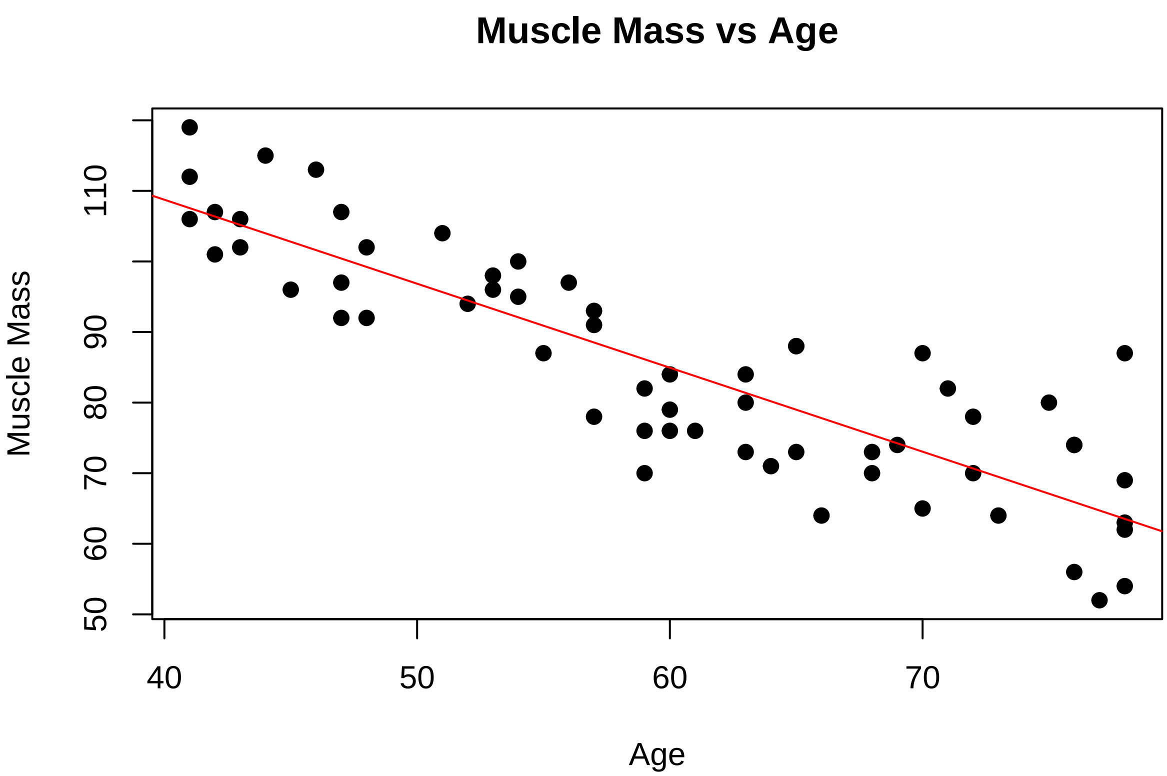
```
A correlation of -0.866 shows that there is a very strong negative relationship between
age and muscle mass in the data. As muscle mass goes down, age goes up, and the
relationship is extremely strong given the value of the correlation is close to -1.
```

Part 2b: First-Order Model

```r
### Question: Fit a first-order regression model to the data and plot the fitted
#regression function and the data.
mmass_model1 <- lm(mmass ~ age, data=mmass)
summary(mmass_model1)
```

```
##
## Call:
## lm(formula = mmass ~ age, data = mmass)
##
## 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 with regression line
plot(mmass$age, mmass$mmass, pch=19, col="black",
     main="Muscle Mass vs Age", xlab="Age", ylab="Muscle Mass")
abline(mmass_model1, col="red")
```




```r
# Explanation: The first-order regression model (linear regression) helps us understand
#the relationship between muscle mass and age. The plot shows the data points and the
#fitted regression line, indicating how muscle mass changes with age.
```

Observations for Part 2b

```
Goodness of Fit
R-squared (R²): 0.7501
This indicates that approximately 75.01% of the variation in muscle mass is explained by
age.

Adjusted R-squared: 0.7458
This adjustment for the number of predictors in the model remains high, and the goodness
of fit is established.

The graph displays the points and the regression line fitted. The line captures the
general trend of the data to decline, showing muscle mass as age increases. The large
R-squared and low p-values for the coefficients support that the regression function "fits
the data very well".
```

Part 2c

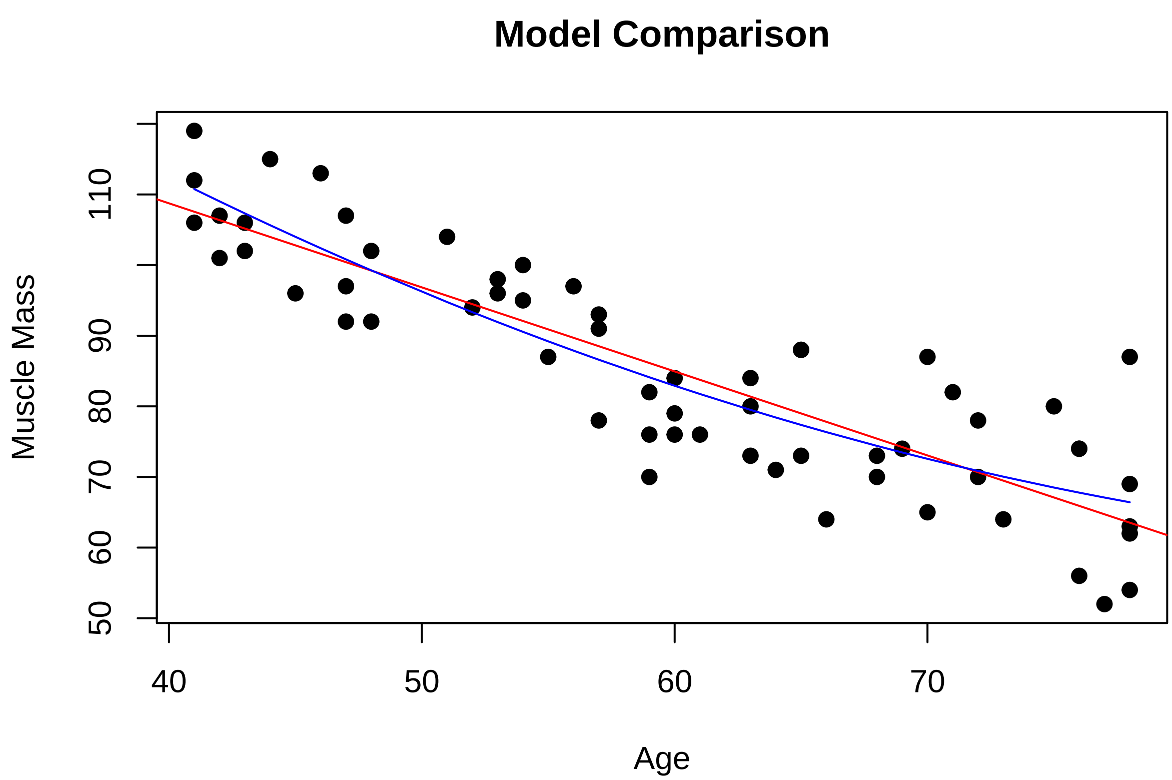
```r
###Question: Fit a second-order regression model Yi =  0 +  1Xi +  11Xi^2 +  i
mmass_model2 <- lm(mmass ~ age + I(age^2), data=mmass)
summary(mmass_model2)
```

```
##
## Call:
```

```
## lm(formula = mmass ~ age + I(age^2), data = mmass)
##
## 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
```

#Part 2d

```r
###Question: Plot the fitted regression functions in a) and b) on the same scatterplot of
#the data using different colors. Which of the regression functions appears to be a better
#fit?

# Plot comparison
plot(mmass$age, mmass$mmass, pch=19, col="black",
     main="Model Comparison", xlab="Age", ylab="Muscle Mass")
abline(mmass_model1, col="red")
lines(sort(mmass$age), predict(mmass_model2, newdata=data.frame(age=sort(mmass$age))), col="blue")
```



```
# Explanation: The second-order regression model includes a quadratic term (age^2) to
#capture any non-linear relationships. The plot compares the first-order (linear) and
#second-order (quadratic) models, showing which model better fits the data.
```

Observations for Part 2d

```
Model Fitting:
model1 fits the first-order (linear) regression model (red)
model2 fits the second-order (quadratic) regression model (blue)

Plotting:
The plot function creates a scatterplot of muscle mass (mmass) vs. age.


Model Summaries:
From the summaries of the models, we can get the R² values and compare the fit:

First-Order Model (Linear):
Estimated Regression Function: mmass=156.35−1.19×age
R²: 0.7501

Second-Order Model (Quadratic):
Estimated Regression Function: mmass= 0+ 1×age+ 2×age^2
R²: 0.7632

Conclusion:
The R² value of the second-order model (R² = 0.7632) is greater than that of the
first-order model (R² = 0.7501), indicating a better fit to capture non-linear
relationships. The comparison plot shows the first-order regression line in red and the
second-order regression line in blue. The second-order model (blue line) better captures
the non-linear trend in the data and is thus a better fit.
```

Part 2e

```
Question: Test whether or not there is a significant regression relation for the model in
b); use  = (Just give the conclusion of the test and report the p-value).
==>
--------------------------------------------------------------
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 ***
--------------------------------------------------------------

p-value: < 2e-16.
Since the p-value is less than the significance level (alpha = 0.05), we reject the null
hypothesis.
```

Part 2f

```
Question: Test whether the quadratic term can be dropped from the regression model; use
=.05. (Hint: This is where you use the p-value for the quadratic term produced in the
summary. Your null hypothesis is H0 :  11 = 0 against the alternative Ha :  11 = 0)
==>
---------------------------------------------------------
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
---------------------------------------------------------

Quadratic Regression (mmass ~ age + I(age^2))
Null hypothesis for age: The coefficient for age is zero, meaning age has no linear effect
on muscle mass.
Null hypothesis for I(age^2): The coefficient for age^2 is zero, meaning there is no
quadratic effect of age on muscle mass.

From the summary output:
p-value for age is 0.00453, which is less than (alpha=0.05), so we reject the null
hypothesis for age.

p-value for I(age^2) is 0.08109, which is greater than 0.05, meaning we fail to reject the
null hypothesis for age^2.

Hence, For the quadratic model, while age remains significant, the quadratic term (age^2)
is not significant at the 5% level. This suggests that the quadratic term might not add
much explanatory power.
```

Part 2g: Third-Order Model

```
###Question: Fit a third-order model and test whether or not  111 = 0 : use   = conclusion
#and p-value.
mmass_model3 <- lm(mmass ~ age + I(age^2) + I(age^3), data=mmass)
summary(mmass_model3)
```

```
##
## Call:
## lm(formula = mmass ~ age + I(age^2) + I(age^3), data = mmass)
##
## 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
```

```
###Explanation: The third-order regression model includes cubic terms to capture even more
#complex relationships. The summary output helps us determine if the cubic term is
#significant.
```

Observations for Part 2g

```
--------------------------------------------------------
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

--------------------------------------------------------
p-value for I(age^3): 0.719
Since the p-value (0.719) is greater than 0.05, we fail to reject the null hypothesis.
Conclusion: The cubic term is not significant at the 0.05 significance level.
```

#____#

CDI DATA ANALYSIS (Qualitative predictors)

#____#

Load data from file

```
cdi <- read.table("cdi.txt", header = TRUE)
View(cdi)  # Characteristic data inspection
```

Initial exploration

```
head(cdi, n=10)
```

```
##    id_number      county state land_area_sq_mi total_population
## 1          1 Los_Angeles    CA            4060          8863164
## 2          2        Cook    IL             946          5105067
## 3          3      Harris    TX            1729          2818199
## 4          4   San_Diego    CA            4205          2498016
## 5          5      Orange    CA             790          2410556
## 6          6       Kings    NY              71          2300664
## 7          7    Maricopa    AZ            9204          2122101
## 8          8       Wayne    MI             614          2111687
## 9          9        Dade    FL            1945          1937094
## 10        10      Dallas    TX             880          1852810
##    percent_population_18_34 percent_population_65_plus number_active_physicians
## 1                      32.1                        9.7                    23677
## 2                      29.2                       12.4                    15153
## 3                      31.3                        7.1                     7553
```

```
## 4                     33.5                       10.9                       5905
## 5                     32.6                        9.2                       6062
## 6                     28.3                       12.4                       4861
## 7                     29.2                       12.5                       4320
## 8                     27.4                       12.5                       3823
## 9                     27.1                       13.9                       6274
## 10                    32.6                        8.2                       4718
##    number_hospital_beds total_serious_crimes percent_high_school_graduates
## 1                 27700               688936                          70.0
## 2                 21550               436936                          73.4
## 3                 12449               253526                          74.9
## 4                  6179               173821                          81.9
## 5                  6369               144524                          81.2
## 6                  8942               680966                          63.7
## 7                  6104               177593                          81.5
## 8                  9490               193978                          70.0
## 9                  8840               244725                          65.0
## 10                 6934               214258                          77.1
##    percent_bachelors_degrees percent_below_poverty_level percent_unemployment
## 1                       22.3                        11.6                  8.0
## 2                       22.8                        11.1                  7.2
## 3                       25.4                        12.5                  5.7
## 4                       25.3                         8.1                  6.1
## 5                       27.8                         5.2                  4.8
## 6                       16.6                        19.5                  9.5
## 7                       22.1                         8.8                  4.9
## 8                       13.7                        16.9                 10.0
## 9                       18.8                        14.2                  8.7
## 10                      26.3                        10.4                  6.1
##    per_capita_income total_personal_income_millions geographic_region
## 1              20786                         184230                 4
## 2              21729                         110928                 2
## 3              19517                          55003                 3
## 4              19588                          48931                 4
## 5              24400                          58818                 4
## 6              16803                          38658                 1
## 7              18042                          38287                 4
## 8              17461                          36872                 2
## 9              17823                          34525                 3
## 10             21001                          38911                 3
```

```
summary(cdi)
```

```
##    id_number        county               state           land_area_sq_mi
##  Min.   :  1.0   Length:440         Length:440         Min.   :   15.0
##  1st Qu.:110.8   Class :character   Class :character   1st Qu.:  451.2
##  Median :220.5   Mode  :character   Mode  :character   Median :  656.5
##  Mean   :220.5                                         Mean   : 1041.4
##  3rd Qu.:330.2                                         3rd Qu.:  946.8
##  Max.   :440.0                                         Max.   :20062.0
##  total_population  percent_population_18_34 percent_population_65_plus
##  Min.   : 100043   Min.   :16.40            Min.   : 3.000
##  1st Qu.: 139027   1st Qu.:26.20            1st Qu.: 9.875
##  Median : 217280   Median :28.10            Median :11.750
##  Mean   : 393011   Mean   :28.57            Mean   :12.170
```

```
##  3rd Qu.: 436064   3rd Qu.:30.02           3rd Qu.:13.625
##  Max.   :8863164   Max.   :49.70           Max.   :33.800
##  number_active_physicians number_hospital_beds total_serious_crimes
##  Min.   :   39.0          Min.   :   92.0      Min.   :   563
##  1st Qu.:  182.8          1st Qu.:  390.8      1st Qu.:  6220
##  Median :  401.0          Median :  755.0      Median : 11820
##  Mean   :  988.0          Mean   : 1458.6      Mean   : 27112
##  3rd Qu.: 1036.0          3rd Qu.: 1575.8      3rd Qu.: 26280
##  Max.   :23677.0          Max.   :27700.0      Max.   :688936
##  percent_high_school_graduates percent_bachelors_degrees
##  Min.   :46.60                 Min.   : 8.10
##  1st Qu.:73.88                 1st Qu.:15.28
##  Median :77.70                 Median :19.70
##  Mean   :77.56                 Mean   :21.08
##  3rd Qu.:82.40                 3rd Qu.:25.32
##  Max.   :92.90                 Max.   :52.30
##  percent_below_poverty_level percent_unemployment per_capita_income
##  Min.   : 1.400              Min.   : 2.200       Min.   : 8899
##  1st Qu.: 5.300              1st Qu.: 5.100       1st Qu.:16118
##  Median : 7.900              Median : 6.200       Median :17759
##  Mean   : 8.721              Mean   : 6.597       Mean   :18561
##  3rd Qu.:10.900              3rd Qu.: 7.500       3rd Qu.:20270
##  Max.   :36.300              Max.   :21.300       Max.   :37541
##  total_personal_income_millions geographic_region
##  Min.   :  1141                 Min.   :1.000
##  1st Qu.:  2311                 1st Qu.:2.000
##  Median :  3857                 Median :3.000
##  Mean   :  7869                 Mean   :2.461
##  3rd Qu.:  8654                 3rd Qu.:3.000
##  Max.   :184230                 Max.   :4.000
```

```r
str(cdi)
```

```
## 'data.frame':    440 obs. of  17 variables:
##  $ id_number                     : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ 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 2111
##  $ 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 2447
##  $ 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 3891
##  $ geographic_region             : int  4 2 3 4 4 1 4 2 3 3 ...
```

Part 3a: Multiple Regression with Qualitative Predictors

```r
# Questions: Fit a multiple linear regression model. Write the regression equation,
#specify what X3,X4and X5 are and how they are encoded.

#Explanation: The multiple linear regression model includes both quantitative and
#qualitative predictors. Converting geographic_region to a factor allows us to include it
#as a categorical variable in the model.

# Convert region to factor (professor's approach for categorical variables)
cdi$geographic_region <- factor(cdi$geographic_region,
                                levels = c(1, 2, 3, 4),
                                labels = c("NE", "NC", "S", "W"))

# Fit model (professor's compact format)
cdi.model <- lm(number_active_physicians ~ total_population +
               total_personal_income_millions + geographic_region,
               data = cdi)

# Model summary (professor always includes both)
summary(cdi.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 ***
## geographic_regionNC            -3.493e+00  7.881e+01  -0.044   0.9647
## geographic_regionS              4.220e+01  7.402e+01   0.570   0.5689
## geographic_regionW             -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
```

```r
anova(cdi.model)
```

```
## Analysis of Variance Table
##
## Response: number_active_physicians
##                                 Df     Sum Sq    Mean Sq   F value    Pr(>F)
## total_population                 1 1243181164 1243181164 3878.9792 < 2.2e-16
## total_personal_income_millions   1   22058054   22058054   68.8256 1.369e-15
## geographic_region                3    1873626     624542    1.9487     0.121
```

```
## Residuals                       434  139093455     320492
##
## total_population                ***
## total_personal_income_millions ***
## geographic_region
## Residuals
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
# Regression equation (formatted output)
cat("\nEstimated Regression Function:\n")
```

```
##
## Estimated Regression Function:
```

```r
cat("physicians =",
    round(coef(cdi.model)[1], 2), "+",
    round(coef(cdi.model)[2], 5), "(population) +",
    round(coef(cdi.model)[3], 2), "(income) +",
    round(coef(cdi.model)[4], 2), "(regionNC) +",
    round(coef(cdi.model)[5], 2), "(regionS) +",
    round(coef(cdi.model)[6], 2), "(regionW)\n")
```

```
## physicians = -58.48 + 0.00055 (population) + 0.11 (income) + -3.49 (regionNC) + 42.2 (regionS) + -14
```

```r
# Encoding explanation
cat("\nGeographic Region Encoding (Reference = NE):\n")
```

```
##
## Geographic Region Encoding (Reference = NE):
```

```r
cat("X3 = regionNC (North Central)\n")
```

```
## X3 = regionNC (North Central)
```

```r
cat("X4 = regionS (South)\n")
```

```
## X4 = regionS (South)
```

```r
cat("X5 = regionW (West)\n")
```

```
## X5 = regionW (West)
```

```r
print(contrasts(cdi$geographic_region))  # Show dummy coding
```

```
##    NC S W
## NE  0 0 0
## NC  1 0 0
## S   0 1 0
## W   0 0 1
```

Observations for Part 3a

```
Estimated Regression Function:
physicians = -58.48 + 0.00055 (population) + 0.11 (income) + -3.49 (regionNC) + 42.2
(regionS) + -149.02 (regionW)

Geographic Region Encoding (Reference = NE):
X3 = regionNC (North Central)
```

```
X4 = regionS (South)
X5 = regionW (West)
   NC S W
NE  0 0 0
NC  1 0 0
S   0 1 0
W   0 0 1

The geographic_region variable is treated as a categorical variable with factor
encoding in the regression model in R. The base category is NE (Northeast).
Three dummy variables are used to encode the rest of the regions:
X3 = regionNC (North Central)
X4 = regionS (South)
X5 = regionW (West)

This means:
For NE: X3 = 0, X4 = 0, X5 = 0
For NC: X3 = 1, X4 = 0, X5 = 0
For S: X3 = 0, X4 = 1, X5 = 0
For W: X3 = 0, X4 = 0, X5 = 1

These dummy variables allow the model to measure the effect of each region relative
to the Northeast. The regression coefficients on these variables show how the
number of practicing physicians in each region differs from the Northeast,
holding all other predictors constant.
```

Part 3b: Coefficient Interpretation

```r
###Question: Briefly explain what the coefficients B2 and B3 in the context of the model.

cat("\nInterpretation of Coefficients:\n")
```

```
##
## Interpretation of Coefficients:
```

```r
cat("  (income): For each $1 million increase in personal income,",
    "we expect", abs(round(coef(cdi.model)[3], 2)),
    "more physicians, holding population and region constant.\n")
```

```
##   (income): For each $1 million increase in personal income, we expect 0.11 more physicians, holding
```

```r
cat("  (regionNC): North Central counties have",
    round(coef(cdi.model)[4], 2),
    "fewer physicians than Northeast counties (reference group),",
    "when controlling for population and income.\n")
```

```
##   (regionNC): North Central counties have -3.49 fewer physicians than Northeast counties (reference g
```

Observations for Part 3b

```
---------------------------------------------------------------------------
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 ***
geographic_regionNC            -3.493e+00  7.881e+01  -0.044   0.9647
geographic_regionS              4.220e+01  7.402e+01   0.570   0.5689
geographic_regionW             -1.490e+02  8.683e+01  -1.716   0.0868 .
-------------------------------------------------------------------------
Interpretation of Coefficients:
B2 (income): For each $1 million increase in personal income, we expect 0.11 more
physicians, holding population and region constant.
B3 (regionNC): North Central counties have -3.49 fewer physicians than Northeast counties
(reference group), when controlling for population and income.
```