Patient Satisfaction Regression Analysis

Ricca D. Callis

Johns Hopkins University

Patient Satisfaction Regression Analysis

This assignment provided students enrolled in a Statistical Models and Regression course (625.661), at Johns Hopkins University, the opportunity to implement the regression modeling techniques taught throughout the semester. This regression analysis utilized Table B.17 'Patient Satisfaction Data' from Montgomery et al. (2012). As a hospital administrator, it is of particular interest to study the patient experience, as it is is a crucial factor to achieving high-quality care and improved patient outcomes. Here, the dataset explored the relationship between patient satisfaction (Y, as a response variable) and patient's age (X_1, in years), severity of illness (X_2), and anxiety level (X_3). Data are reported for 25 randomly selected patients. Analysis was conducted with the intent to both understand patient satisfaction as a function of the attributes listed above, but also to predict patient satisfaction for newly admitted hospital patients.

Introduction

Regulatory changes, labor shortages, declining reimbursement rates, and operating losses are forcing a fundamental change in the way healthcare is delivered. Furthermore, as the availability of healthcare choices expands, the strategic importance of patient confidence, trust, and willingness to not only return, but also recommend friends and family, is vital to a hospital's long-term success. Most patient satisfaction surveys request feedback in a variety of areas including: communication with and education from clinicians, responsiveness of hospital staff and waiting times, cleanliness and quietness of the hospital environment, pain management, communication about medicines, and discharge information. This feedback reveals key indicators regarding the quality of care provided to patients – a benchmark for evaluating hospital performance. Furthermore, patient satisfaction plays a pivotal role in improving patient

outcomes. A patient who has a positive experience at a facility and who trusts her healthcare provider is more likely to comply with treatment recommendations.

The Patient Satisfaction Dataset

As previously mentioned, the cumulative file obtained from Montgomery et al. (2012) includes 25 observations obtained from recently discharged patients (see Table 1).

Satisfaction	Age	Severity	Surgical-Medical	Anxiety
68	55	50	0	2.1
77	46	24	1	2.8
96	30	46	1	3.3
80	35	48	1	4.5
43	59	58	0	2
44	61	60	0	5.1
26	74	65	1	5.5
88	38	42	1	3.2
75	27	42	0	3.1
57	51	50	1	2.4
56	53	38	1	2.2
88	41	30	0	2.1
88	37	31	0	1.9
102	24	34	0	3.1
88	42	30	0	3
70	50	48	1	4.2
82	58	61	1	4.6
43	60	71	1	5.3
46	62	62	0	7.2
56	68	38	0	7.8
59	70	41	1	7
26	79	66	1	6.2
52	63	31	1	4.1
83	39	42	0	3.5
75	49	40	1	2.1

Table 1: Patient Satisfaction Dataset obtained from Montgomery et al. (2012)

The dataset includes one response (or target) variable, Satisfaction, and four potential predictor (or regressor) variables (Age, Severity, Surgical-Medical, Anxiety). The response

variable, Satisfaction, is a subjective patient response measure on an increasing scale. The regressor variable Age indicates the age (in years) of the patient at the time of discharge. The regressor variable Severity is an index measuring the severity of the patient's illness. The regressor variable Surgical-Medical is an indicator of whether the patient is seen for surgery or for other medical purposes (where, 0 = surgical and 1 = medical). The regressor variable Anxiety is an index measuring the patient's anxiety level. For all index variables (Satisfaction, Severity, Anxiety), higher values indicate more satisfaction, increased severity of illness, or more anxiety.

Variable Name	Variable Type	Variable Classification	Description
Satisfaction	Response; Target; Dependent	Quantitative; Continuous; Ratio	A subjective patient response measure on an increasing scale describing the satisfaction of the patient during their hospital visit. Serves as a proxy for how positive the patient's experience was.
Age	Predictor; Regressor; Independent	Quantitative; Continuous; Ratio	Age (in years) of the patient at the time of discharge
Severity	Predictor; Regressor; Independent	Quantitative; Continuous; Ratio	An index measuring the severity of the patient's illness.
Surgical-Medical	Predictor; Regressor; Independent Indicator; Dummy	Qualitative; Categorical; Binary; Dichotomous	An indicator of whether the patient is seen for surgery or for other medical purposes (where, 0 = surgical and 1 = medical)
Anxiety	Predictor; Regressor; Independent	Quantitative; Continuous; Ratio	An index measuring the patient's anxiety level

Table 2: Feature and target descriptions for the Patient Satisfaction Dataset obtained from Montgomery et al., (2012).

Model Building Outline

> In any event, we must always be aware of the fact that statistical models, such as the regression model, never 'emerge' from data. Instead, we simply 'impose' these models and all their attendant assumptions on our data. As the term suggests, a model is only a representation designed to display the basic structure of a more complex set of phenomena. In the end, we must be prepared to justify, on theoretical and empirical grounds, our choice of a particular model to represent our data.
>
> (Allen, 1997)

This analysis followed an iterative process, as outlined in Figure 1 (below). After obtaining the data and performing a basic exploratory data analysis (EDA), a linear regression model is specified and regression assumptions considered a priori. This analysis began by fitting the full model first, initially including all regressor variables. After estimating the parameters (i.e., β coefficients), adequacy of the model is checked via global F-test, individual t-tests for each β coefficient, residual analysis, and adjusted R^2 calculation. Based on these factors, transformations or model re-specifications were considered (then parameter estimation and model adequacy were re-examined). All possible regressions was performed using criteria such as Mallow's C, adjusted R^2, and the PRESS statistic to rank the best subset models. The best models recommended by each criterion were compared and analyzed prior to model validation. To further understand likely error of the model when applied to new data (i.e., model validation), learning curves, cross-fold validation, and decision tree regression were employed.

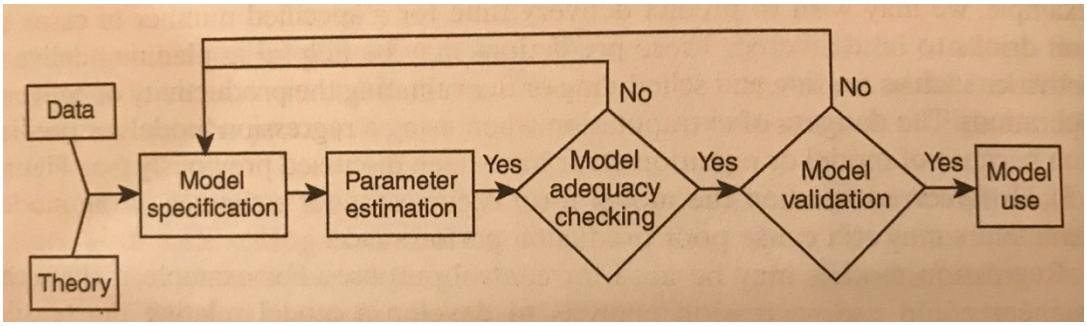


Figure 1: Obtained from Montgomery et al. (2012).

Regression Assumptions

In the regression model, the dependent variable,Y_i, is a function of each independent variable$\left(X_{2i}, X_{3i}, \ldots, X_{kj}\right)$, the parameters $\left(\beta_0, \beta_1, \ldots, \beta_j\right)$ and the value of the *continuous* error term. A regression model will only provide accurate, valid, and reliable results if specific assumptions are met. The specific assumptions for linear regression analysis are explored below (Montgomery et al, 2012).

Linear In Parameters

The dependent variable, *Y,* can be calculated as a linear function of a specific set of independent variables plus an error term. It should be noted that the regression equation must be linear in parameters, but does not have to be linear in the *X*'s.

Random Sample of *n* Observations

The sample consists of n-paired observations that are drawn from the population $\left\{Y_i: X_{2i}, X_{3i}, \ldots, X_{kj}\right\}$. The number of observations is greater than the number of parameters to be estimated (i.e., $n > k$). The predictor X_i is deterministic (nonstochastic, fixed values, not random). All independent variables have nonzero variance (i.e., each independent variable has some variation in value. This ensures that the model examines the relationship between X and Y, rather than the relationship between Y and the error term.

No Multicollinearity

There is not perfect multicollinearity (i.e., there is no exact linear relationship between two or more of the independent variables). A violation in this assumption could result in an infinite number of regression models that fit the observed data equally well. It'd be impossible to isolate the effect of one independent variable, holding all other variables fixed. Furthermore, high correlation among independent variables will cause large standard errors in the partial slope coefficient estimators.

Errors Have Zero Mean

At each set of values for the *n* independent variables, $(X_{2i}, X_{3i}, \dots, X_{kj})$, $E(\varepsilon_i | X_{2i}, X_{3i}, \dots, X_{kj-1}) = 0$. In other words, the mean value of the error term is zero. When this condition is violated, $E(\varepsilon_i | X_{2i}, X_{3i}, \dots, X_{kj-1}) = \mu_i$, where μ_i is nonzero and either remains constant across observations or, more problematically, varies.

Homoskedasticity

Variation about mean does not depend on X_i, i.e. $var(\varepsilon_i) = \sigma^2$ & $Cov(\varepsilon_i, \varepsilon_j) = 0 \; for \; i \neq j$. And, at each set of values for the j-1 independent variables, $(X_{2j}, X_{3j}, \dots, X_{kj-1})$, $var(\varepsilon_i | X_{2j}, X_{3j}, \dots, X_{kj-1}) = \sigma^2$, where σ^2 is a constant (i.e., the conditional variance of the error term is constant). Said differently, homoscedasticity is an assumption that, for each value of an independent variable, the conditional variance of Y_i and ε_i must be identical (i.e., conditional variance of Y_i must equal the constant σ^2).

Errors Are Independent

For each $X_j, Cov\,(X_{kj}, \varepsilon_i) = 0$. In other words, each independent variable is uncorrelated with the error term. This is known as a lack of autocorrelation. This is ensured by the least squares criterion. Since the error term represents the effects of all excluded independent variables

plus the random component of the dependent variable, it must be assumed that, for each set of observations, the net effect of the excluded variables and random behavior of *Y* are uncorrelated.

Errors Are Normally Distributed

At each set of values for the j-1 independent variables, ε_i is normally distributed. $\varepsilon_i \sim N(0, \sigma^2)$. This is key for testing the statistical significance of coefficient estimators and for constructing confidence intervals. It should be noted, however, that with large sample sizes, the Central Limit Theorem ensures that the coefficient estimators are normally distributed regardless of whether the error term is normally distributed.

Exploratory Data Analysis (EDA)

Single Variable EDA

Prior to model selection, each variable was visualized using either a bar chart or histogram of frequencies or densities. Each graph was examined for skewness, kurtosis, and presence outliers. Then descriptive statistics, including Tukey's 5 (minimum, lower-hinge, median, upper-hinge, maximum), were generated for the input data. Tables 3 and 4 summarize the target and feature variables. Relevant R-Code is presented at the end of each section and full code is copied at the end of this write-up.

Features/Target	Mean	Std. Dev.	Minimum	Median	Maximum
Satisfaction	66.72	21.24	26	70	102
Age	50.84	14.81	24	51	79
Severity	45.9	13.03	24	42	71
Surgical-Medical	N/A	N/A	N/A	N/A	N/A
Anxiety	3.93	1.76	1.90	3.30	7.80

Table 3: Data summary (excluding 'Surgical-Medical').

```
> #Target/Response Variable = Satisfaction
> #Predictor Variables = Age, Severity, Surgical-Medical, Anxiety
```

```
> #Create data frame
> data <- data.frame(Satisfaction, Age, Severity, `Surgical-Medical`, Anxiety)

> #Run EDA
> summary(data)
```

Satisfaction	Age	Severity	Surgical.Medical	Anxiety
Min. : 26.00	Min. :24.00	Min. :24.00	Min. :0.00	Min. :1.900
1st Qu.: 52.00	1st Qu.:39.00	1st Qu.:38.00	1st Qu.:0.00	1st Qu.:2.400
Median : 70.00	Median :51.00	Median :42.00	Median :1.00	Median :3.300
Mean : 66.72	Mean :50.84	Mean :45.92	Mean :0.56	Mean :3.932
3rd Qu.: 83.00	3rd Qu.:61.00	3rd Qu.:58.00	3rd Qu.:1.00	3rd Qu.:5.100
Max. :102.00	Max. :79.00	Max. :71.00	Max. :1.00	Max. :7.800

```
> #Get standard deviations
> sapply(data, sd)
```

Satisfaction	Age	Severity	Surgical.Medical	Anxiety
21.2436657	14.8090063	13.0285584	0.5066228	1.7641617

Satisfaction

As previously mentioned, Satisfaction is the target variable for this analysis. Satisfaction scores ranged from 26 to 102. The mean Satisfaction score was 66.72, with a standard deviation of 21.24. The medial score was 70. Considering the difference between the mean and median, a slight skew was expected in the analysis. Skewness calculation yielded -0.307575, indicating a very slight left/negative skew. Kurtosis calculation yielded 2.120311, indicating low kurtosis. See Figures 2 and 3 for count and density histograms, respectively. Initial EDA appears to show Satisfaction scores have a mostly normal range, a slightly higher than expected average, and a skewed distribution in favor of higher patient satisfaction values (a good thing).

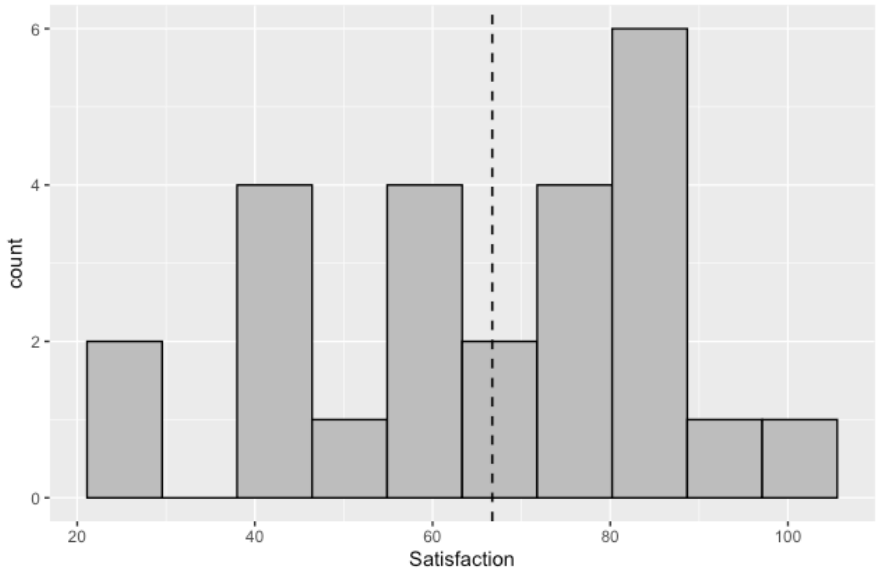


Figure 2: Histogram of Satisfaction counts. Note that the dashed line represents the mean.

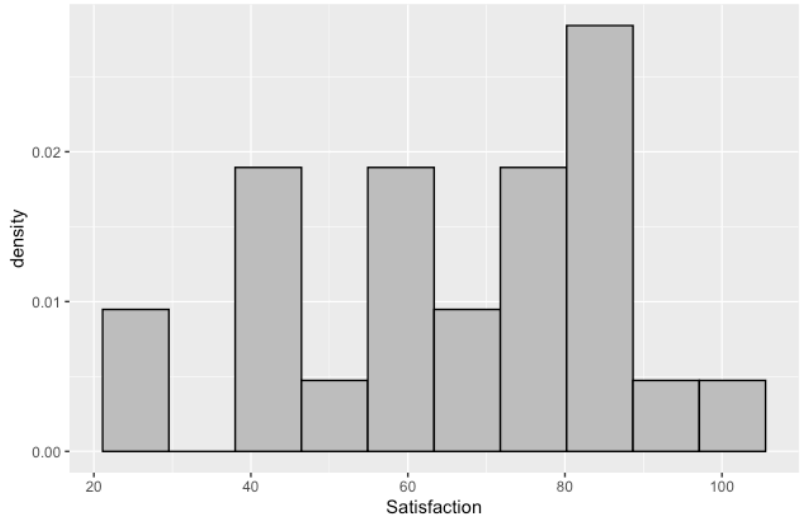


Figure 3: Histogram of Satisfaction counts. Note that the dashed line represents the mean.

> #Satisfaction

>#Tukey's 5 (minimum, lower-hinge, median, upper-hinge, maximum)
> fivenum(data$Satisfaction, na.rm = TRUE)
[1] 26 52 70 83 102
> #min=26 ; lower-hinge=52; median=70; upper-hinge=82; max=102

> #Get interquartile range
> IQR(data$Satisfaction, na.rm = TRUE)
[1] 31

```
> #IQR = 31

> #Skewness
> skewness(data$Satisfaction)
[1] -0.307575
> #-0.307575
> #very slight or moderate left/negative skew

> #Kurtosis
> kurtosis(data$Satisfaction)
[1] 2.120311
>#2.120311
>#kurtosis less than 3 = low kurtosis

> #Histogram for each variable
> #Satisfaction
> ggplot(data=data)+geom_histogram(mapping=aes(x=data$Satisfaction))
`stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
> ggplot(data=data)+geom_histogram(mapping=aes(x=data$Satisfaction), binwidth = 10)
> SatisfactionPlot <-ggplot(data, aes(x=Satisfaction))

> #Density Plot
> #y axis scale = density
> SatisfactionPlot +geom_density()+geom_vline(aes(xintercept=mean(Satisfaction)), linetype =
"dashed", size = 0.6)
> #Change y axis to count instead of density
> SatisfactionPlot +geom_density(aes(y=..count..),
fill="lightgray")+geom_vline(aes(xintercept=mean(Satisfaction)), linetype="dashed", size=0.6,
color="#FC4E07")

> #Histogram plot (counts)
> SatisfactionPlot+geom_histogram(bins=10, color="black",
fill="gray")+geom_vline(aes(xintercept=mean(Satisfaction)),linetype="dashed",size=0.6)

> #Histogram plot (density)
> SatisfactionPlot+geom_histogram(aes(y=..density..), color="black",
fill="white")+geom_density(alpha=0.2,fill="#FF6666")
`stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
> SatisfactionPlot+geom_histogram(bins=10, aes(y=..density..), color="black", fill="gray")

> #Basic frequency polygon
> SatisfactionPlot+geom_freqpoly(bins=10)
```

Age

As previously mentioned, Age is a regressor variable indicating the patient's age, in years, at the time of discharge. Ages ranged from 24 to 79 years old. The mean age was 50.84 with a standard deviation of 14.81. The median age was 51. Considering that the mean and median were close in value, no skewness was anticipated for this variable. Skewness calculation yielded -0.01654419, indicating almost nonexistent skew. Kurtosis calculation yielded 2.160039, indicating low kurtosis. See Figures 4 and 5 for count and density histograms, respectively. Initial EDA appears to show Age observations have a mostly normal range, an expected value as the mean, and no skew or kurtosis in it's distribution.

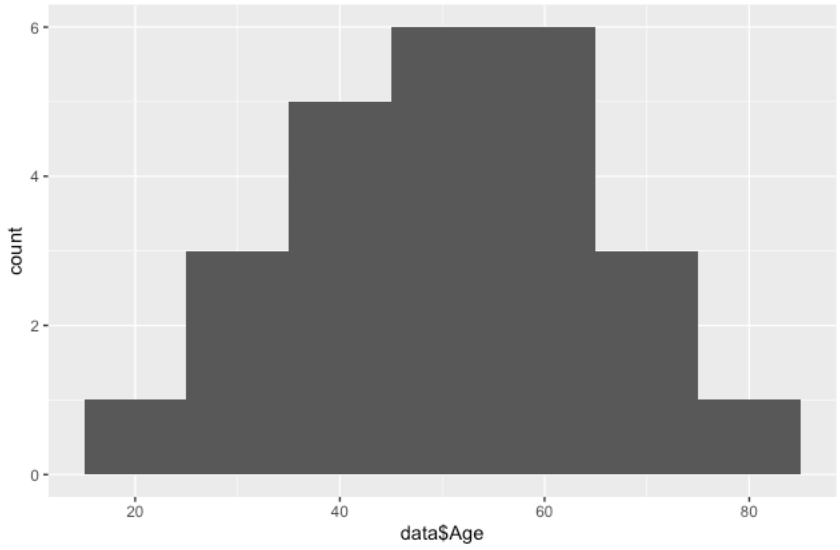


Figure 4: Histogram of Age counts

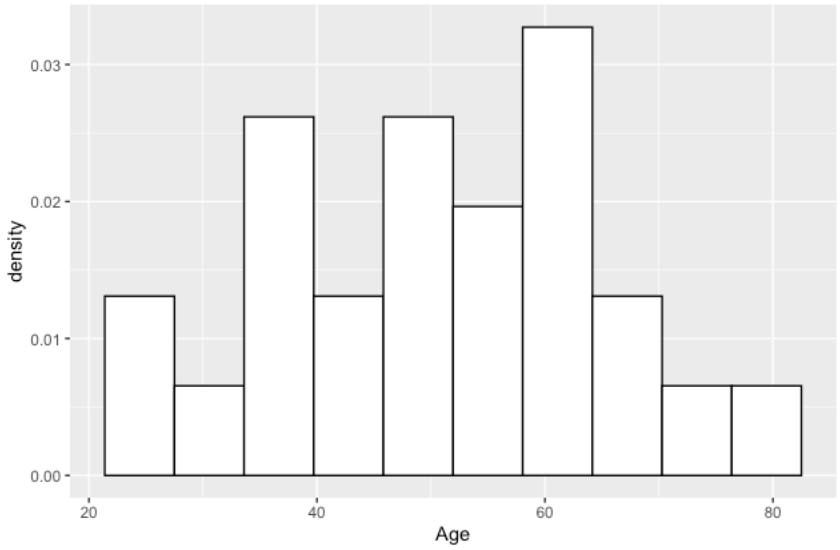


Figure 5: Histogram of Age densities

```
> #Age
> #Get Tukey's 5 (minimum, lower-hinge, median, upper-hinge, maximum)
> fivenum(data$Age, na.rm = TRUE)
[1] 24 39 51 61 79
> #min=24; lower-hinge=39; median=51; upper-hinge=61; max=79

> #Get interquartile range
> IQR(data$Age, na.rm = TRUE)
[1] 22
> #IQR = 22
> #Skewness
> skewness(data$Age)
[1] -0.01654419
> #-0.01654419
> #Almost no skew

> #Kurtosis
> kurtosis(data$Age)
[1] 2.160039
>#2.160039
>Value is less than 3 = low kurtosis

> #Plot Individual Histograms
> ggplot(data=data)+geom_histogram(mapping=aes(x=data$Age))
`stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
> ggplot(data=data)+geom_histogram(mapping=aes(x=data$Age), binwidth = 10)
> AgePlot <-ggplot(data, aes(x=Age))
```

```
> #Density Plot
> #y axis scale = density
> AgePlot +geom_density()+geom_vline(aes(xintercept=mean(Age)), linetype = "dashed", size
= 0.6)
> #Change y axis to count instead of density
> AgePlot +geom_density(aes(y=..count..),
fill="lightgray")+geom_vline(aes(xintercept=mean(Age)), linetype="dashed", size=0.6,
color="#FC4E07")

> #Histogram plot (counts)
> AgePlot+geom_histogram(bins=10, color="black",
fill="gray")+geom_vline(aes(xintercept=mean(Age)),linetype="dashed",size=0.6)

> #Histogram plot (density)
> AgePlot+geom_histogram(bins=10, aes(y=..density..), color="black",
fill="white")+geom_density(alpha=0.2,fill="#FF6666")
> AgePlot+geom_histogram(bins=10, aes(y=..density..), color="black", fill="white")

> #Basic frequency polygon
> AgePlot+geom_freqpoly(bins=10)
```

Severity

As previously mentioned, Severity is a regressor variable indicating the severity of the patient's illness. Severity ranged from 24 to 71. The mean Severity was 45.92 with a standard deviation of 13.03. The median Severity was 42. Considering that the mean and median were close in value, no skewness was anticipated for this variable. Skewness calculation yielded -0.282166, indicating almost nonexistent skew. Kurtosis calculation yielded 2.047968, indicating low kurtosis. See Figures 6 and 7 for count and density histograms, respectively. Initial EDA appears to show Severity observations have a mostly normal range, an expected value as the mean, and no skew or kurtosis in it's distribution. Transformations may be necessary.

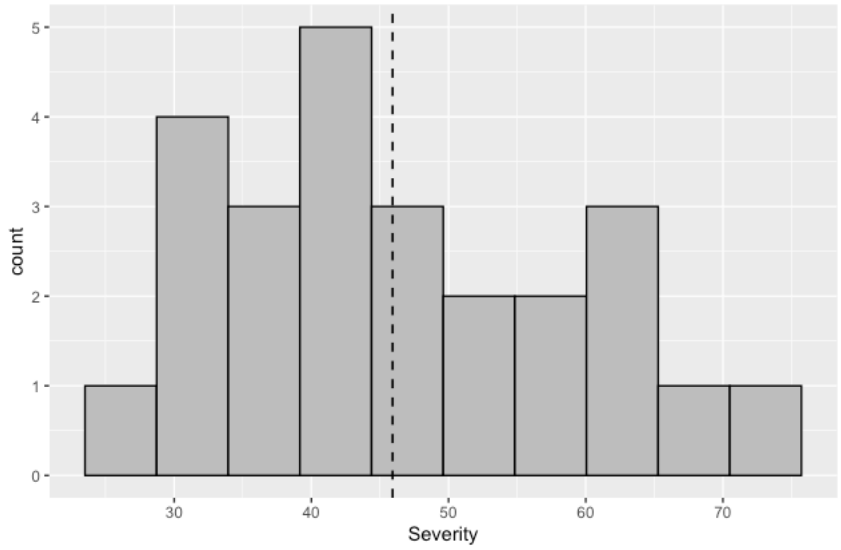


Figure 6: Histogram of Severity counts. Note that the dashed line indicates the mean.

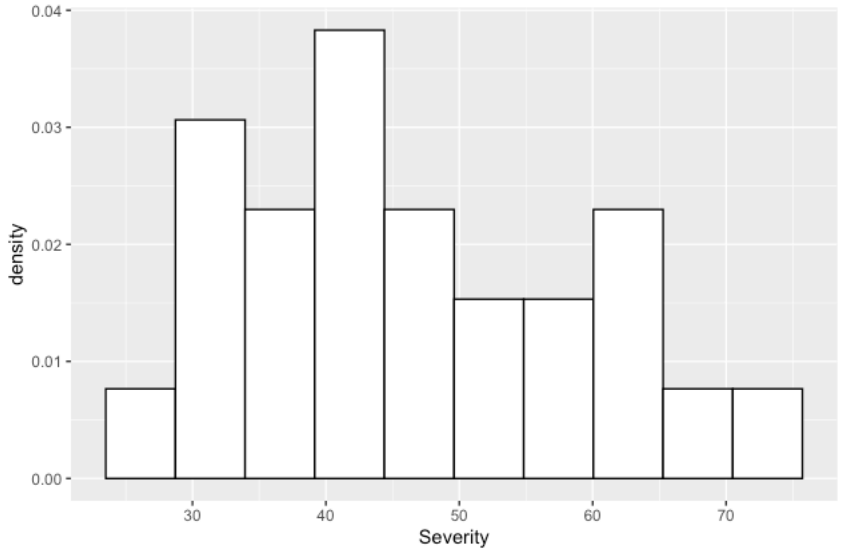


Figure 7: Histogram of Severity density.

```
> #Severity
> #Get Tukey's 5 (minimum, lower-hinge, median, upper-hinge, maximum)
> fivenum(data$Severity, na.rm = TRUE)
[1] 24 38 42 58 71
> #min=24; lower-hinge=38; median=42; upper-hinge=58; maximum=71

> #Get interquartile range
> IQR(data$Severity, na.rm = TRUE)
```

```
[1] 20

> #IQR = 20

> #Skewness
> skewness(data$Severity)
[1] 0.282166
> #[1] 0.282166
> #Almost no skew

> #Kurtosis
> kurtosis(data$Severity)
[1] 2.047968
>#2.047968
>#Value is less than 3 = low kurtosis

> #Plot Individual Histograms
> ggplot(data=data)+geom_histogram(mapping=aes(x=data$Severity))
`stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
> ggplot(data=data)+geom_histogram(mapping=aes(x=data$Severity), binwidth = 5)
> SeverityPlot <-ggplot(data, aes(x=Severity))
> #Density Plot
> #y axis scale = density
> SeverityPlot +geom_density()+geom_vline(aes(xintercept=mean(Severity)), linetype =
"dashed", size = 0.6)
> #Change y axis to count instead of density
> SeverityPlot +geom_density(aes(y=..count..),
fill="lightgray")+geom_vline(aes(xintercept=mean(Severity)), linetype="dashed", size=0.6,
color="#FC4E07")
> #Histogram plot (counts)
> SeverityPlot+geom_histogram(bins=10, color="black",
fill="gray")+geom_vline(aes(xintercept=mean(Severity)),linetype="dashed",size=0.6)
> #Histogram plot (density)
> SeverityPlot+geom_histogram(bins=10, aes(y=..density..), color="black",
fill="white")+geom_density(alpha=0.2,fill="#FF6666")
> SeverityPlot+geom_histogram(bins=10, aes(y=..density..), color="black", fill="white")
> #Basic frequency polygon
> SeverityPlot+geom_freqpoly(bins=10)
```

Surgical-Medical

As previously mentioned, Surgical-Medical is a binary/dichotomous regressor variable indicating whether the patient was seen for surgery (0 = surgery) or for some other medical

reason (1 = medical). In this data set, there were 11 surgical observations and 14 medical observations (see Table 4 and Figure 8). Relatively equal observations of each.

Surgical Counts	Medical Counts
11	14

Table 4: Frequency/Counts for variable 'Surgical-Medical'

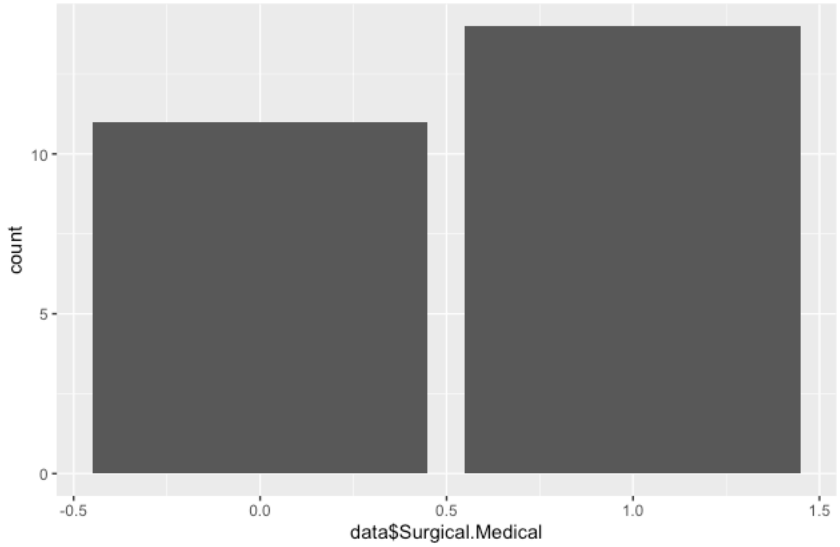


Figure 8: Number of patients seen for surgery (x = 0; i.e., the left-most bar) and number of patients seen for some other medical purpose (x = 1; i.e., the right-most bar).

```
> #Get Surgical-Medical Counts
> table(data$Surgical.Medical)

 0  1
11 14
> #Number of Observations for 0=Surgical: 11
> #Number of Observations for 1=Medical: 14

> #Examine distribution of Surgical-Medical
> ggplot(data = data) + geom_bar(mapping=aes(x=data$Surgical.Medical))
> histogram(`Surgical-Medical`, data=data)
```

Anxiety

As previously mentioned, Anxiety is a regressor variable indicating the amount of anxiety expressed by the patient. Unfortunately, Montgomery et al. (2012) do not cite an outside

source for this dataset, nor do they explain this variable. Thus, it is unclear whether this was a likert-scale, or ordinal subjective response variable. It is suspected that values could have ranged from 0 to 10, but that cannot be confirmed at this time. Descriptive statistics indicate that anxiety ranged from 1.90 to 7.80. The mean Anxiety was 3.93 with a standard deviation of 1.76. The median Anxiety was 3.30. Considering that the mean and median were close in value, no skewness was anticipated for this variable. However, skewness calculation yielded 0.7346762, indicating some skew. Kurtosis calculation yielded 2.454065, indicating a nearly normal shape. See Figures 9 and 10 for count and density histograms, respectively. Initial EDA appears to show Anxiety observations have a mostly normal range, a lower than expected mean (a good thing), right/positive skew (in favor of low anxiety; also a good thing), and a mostly normal distribution shape. Upon further examination of Figure 9 and 10, however, it is anticipated that this variable will require a transformation.

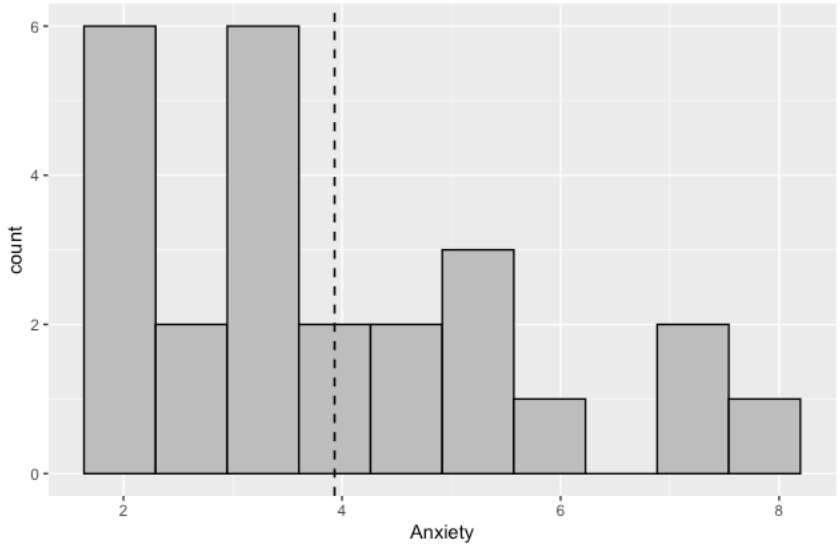


Figure 9: Histogram of Anxiety counts. Note that the dashed line indicates the mean.

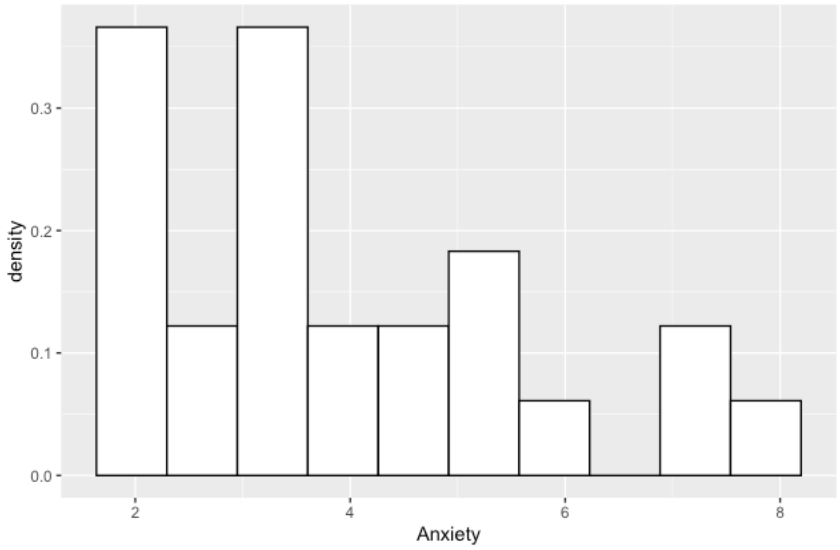


Figure 10: Histogram of Anxiety density.

```
> #Anxiety
< #Get Tukey's 5 (minimum, lower-hinge, median, upper-hinge, maximum)
> fivenum(data$Anxiety, na.rm = TRUE)
[1] 1.9 2.4 3.3 5.1 7.8
> #min=1.9, lower-hinge=2.4, median=3.3, upper-hinge=5.1, max=7.8

> #Get interquartile range
> IQR(data$Anxiety, na.rm = TRUE)
[1] 2.7
> #IQR = 2.7

> #Skewness
> skewness(data$Anxiety)
[1] 0.7346762
> #[1] 0.7346762
> #Value close to positive 1, indicating some skew
> #Kurtosis
> kurtosis(data$Anxiety)
[1] 2.454605
>#2.454065
>#Value near 3, indicating nearly normal shape

> #Anxiety Histogram
> ggplot(data=data)+geom_histogram(mapping=aes(x=data$Anxiety))
`stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
> ggplot(data=data)+geom_histogram(mapping=aes(x=data$Anxiety), binwidth = 1)
```

```
> AnxietyPlot <-ggplot(data, aes(x=Anxiety))

> #Density Plot
> #y axis scale = density
> AnxietyPlot +geom_density()+geom_vline(aes(xintercept=mean(Anxiety)), linetype =
"dashed", size = 0.6)
> #Change y axis to count instead of density
> AnxietyPlot +geom_density(aes(y=..count..),
fill="lightgray")+geom_vline(aes(xintercept=mean(Anxiety)), linetype="dashed", size=0.6,
color="#FC4E07")

> #Histogram plot (counts)
> AnxietyPlot+geom_histogram(bins=10, color="black",
fill="gray")+geom_vline(aes(xintercept=mean(Anxiety)),linetype="dashed",size=0.6)

> #Histogram plot (density)
> AnxietyPlot+geom_histogram(bins=10, aes(y=..density..), color="black",
fill="white")+geom_density(alpha=0.2,fill="#FF6666")
> AnxietyPlot+geom_histogram(bins=10, aes(y=..density..), color="black", fill="white")

> #Basic frequency polygon
> AnxietyPlot+geom_freqpoly(bins=10)
```

Pair-Wise Comparison EDA

Each regressor variable was compared to Satisfaction (the response variable), using both correlations and scatterplots (see Figure 11). Regressors with high correlations to Satisfaction (defined as r > | 0.5|), were expected to make a strong contribution to the regression. Scatterplots were used to assist with curve estimation.

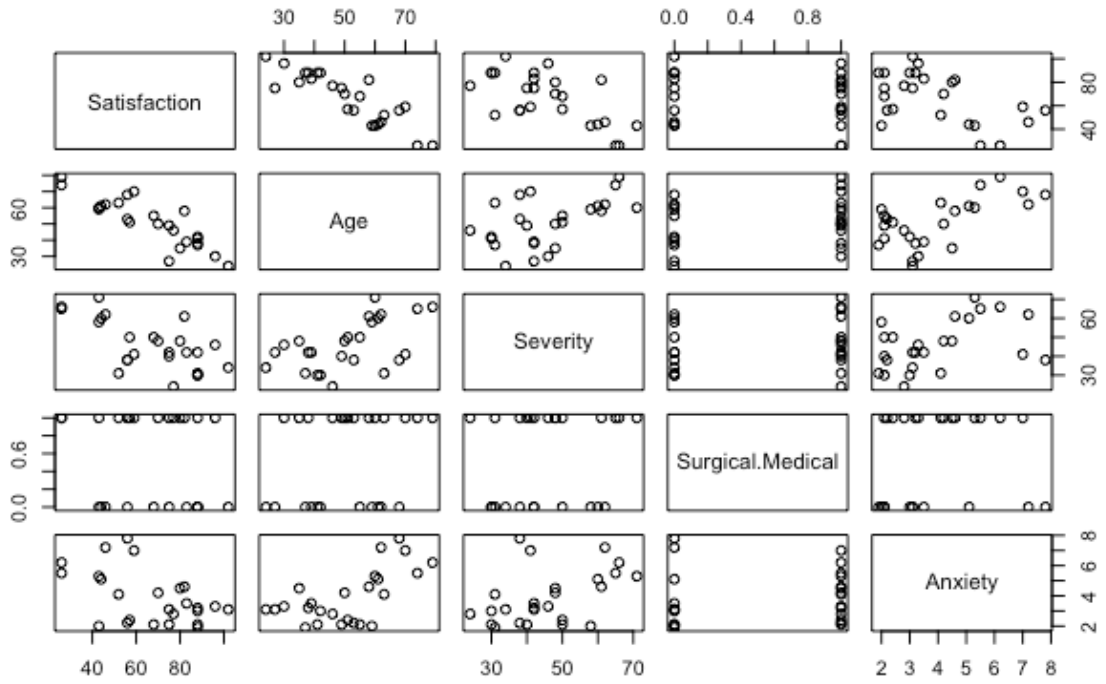


Figure 11: Pair-wise scatter plots

> #Correlations
> pairs(data)
> cor(data)

	Satisfaction	Age	Severity	Surgical.Medical	Anxiety
Satisfaction	1.0000000	-0.8707049	-0.6531434	-0.1822682	-0.5127287
Age	-0.8707049	1.0000000	0.5290246	0.2456932	0.6212453
Severity	-0.6531434	0.5290246	1.0000000	0.1775101	0.4471567
Surgical.Medical	-0.1822682	0.2456932	0.1775101	1.0000000	0.1096486
Anxiety	-0.5127287	0.6212453	0.4471567	0.1096486	1.0000000

> #Age has a strong (negative) correlation with Satisfaction (r = -0.871)
> #Severity has a strong (negative) correlation with Satisfaction (r = -653)
> #Surgical-Medical has a weak correlation with Satisfaction (r = -0.182)
> #Anxiety has strong correlation with Satisfaction (r = -0.513)
> #Where, strong correlation is r >= |0.5|
> #May have a multicollinearity problem:
> #Severity & Age r = 0.529
> #Anxiety & Age r = 0.621

Satisfaction x Age

As seen from the R-code above, Age and Satisfaction have a Pearson Correlation coefficient of -0.87, indicating a strong negative correlation. As age increases, Satisfaction decreases. This linear relationship is made visible by the scatterplot in Figures 11 and 12. It is anticipated that age will remain in the regression model.

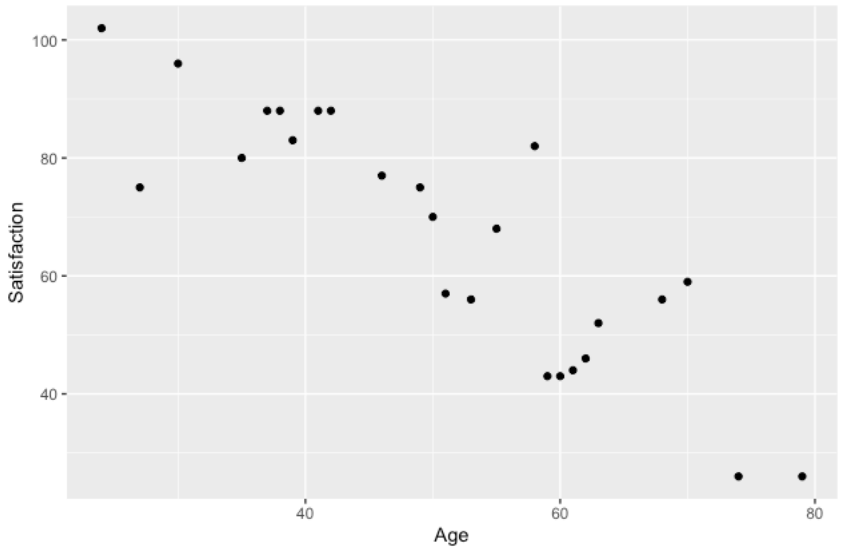


Figure 11: Scatterplot of Age x Satisfaction. A clear linear trend is observed. As age increases, satisfaction decreases

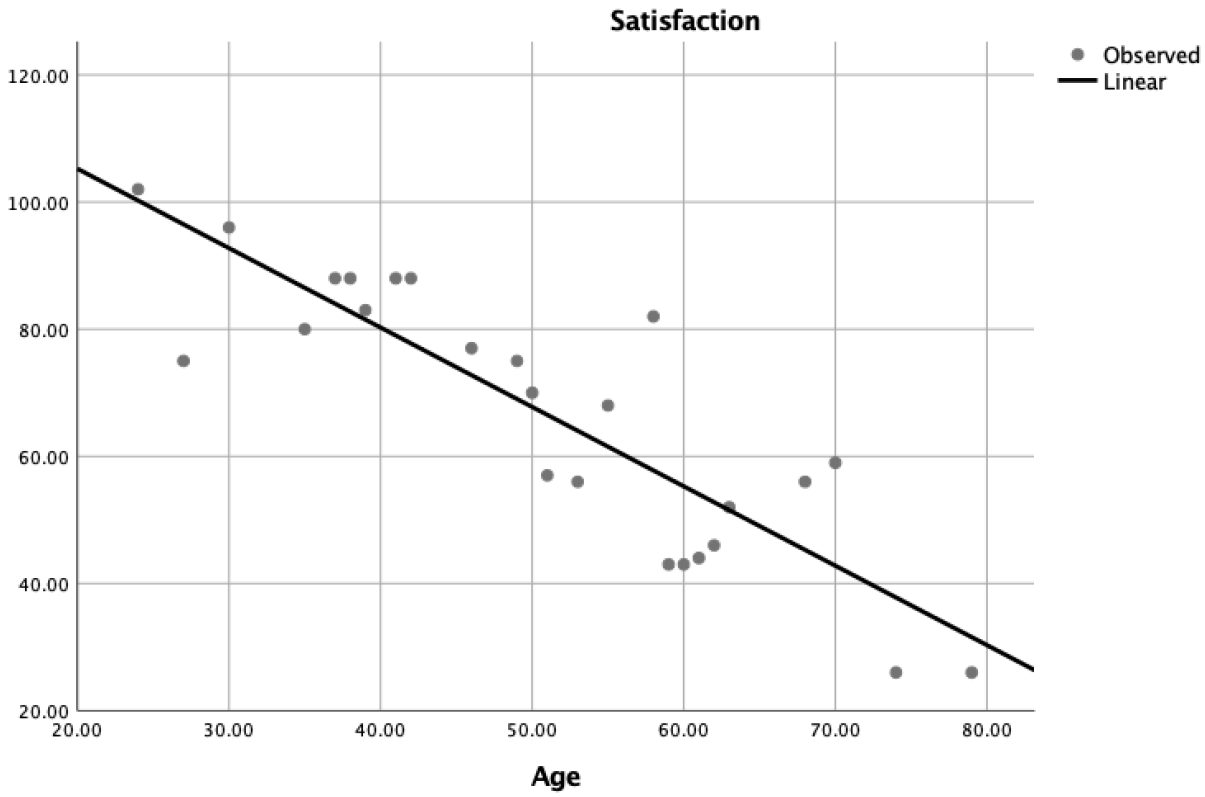


Figure 12: Scatterplot of Age x Satisfaction with linear regression line.

```
> corr.test(Satisfaction, Age)
Call:corr.test(x = Satisfaction, y = Age)
Correlation matrix
[1] -0.87
Sample Size
[1] 25
Probability values  adjusted for multiple tests.
[1] 0
> #r=-0.87 (strong negative)
> corr.test(Satisfaction, Age, method = "spearman")
Call:corr.test(x = Satisfaction, y = Age, method = "spearman")
Correlation matrix
[1] -0.85
Sample Size
[1] 25
Probability values  adjusted for multiple tests.
[1] 0
> #rho = -0.85 (strong negative)

> #Plot each IV against the DV
> ggplot(data=data,aes(x=Age,y=Satisfaction))+geom_point()
```

Satisfaction x Severity

As seen from the R-code, Severity and Satisfaction have a Pearson Correlation coefficient of -0.65, indicating a strong negative correlation. Similarly, Spearman's rho was calculated at -0.6. Thus, as Severity increases, Satisfaction decreases. This relationship is made visible by the scatterplot in Figure 13. However, after visual examination, it is unclear whether this trend is linear, or cubic. It is anticipated that Severity will remain in the regression model. However, a transformation of this variable may be required.

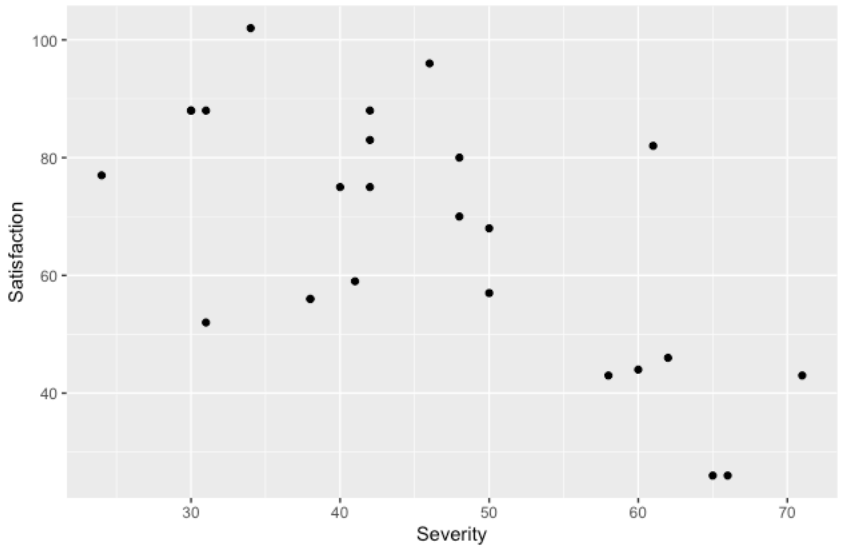


Figure 13: Scatterplot of Age x Satisfaction. As severity of illness increases, satisfaction decreases. It is unclear from this scatterplot whether this is a linear or cubic trend.

Upon further inspection using SPSS' Curve Estimation (see Figure 14 below), cubic regression line appears as a better fit (as indicated by a higher R^2; more on that later). Specifically, a linear regression of Satisfaction as a function of Severity has a coefficient of determination of 0.427, indicating that it accounts for 42.7% of the observed variability. However, the cubic regression has a coefficient of determination of 0.478, indicating that it accounts for 47.8% of the observed variability. Results seem to indicate that a cubic transformation may be necessary.

Model Summary and Parameter Estimates

Dependent Variable: Satisfaction

Equation	Model Summary					Parameter Estimates			
	R Square	F	df1	df2	Sig.	Constant	b1	b2	b3
Linear	.427	17.111	1	23	.000	115.624	-1.065		
Cubic	.478	6.415	3	21	.003	26.295	3.679	-.073	.000

The independent variable is Severity.

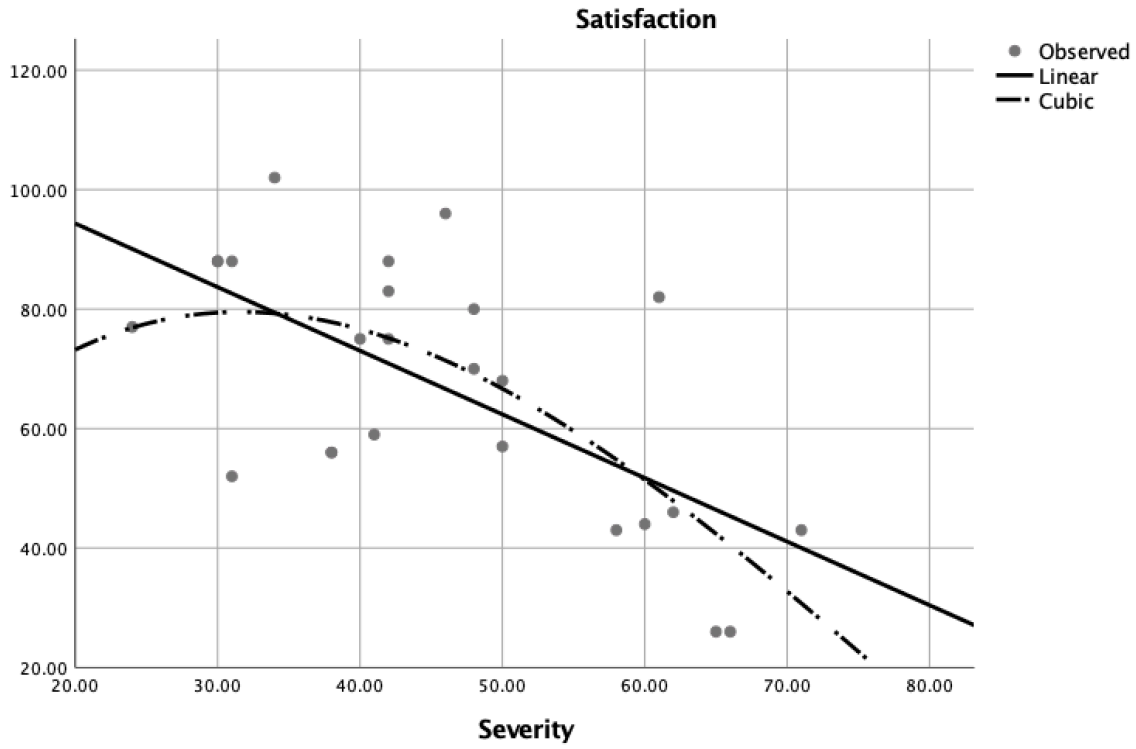


Figure 14: Scatterplot of Severity x Satisfaction. Linear and Cubic regression lines are added for comparison. Model summary indicates a higher R^2 for the cubic model and the linear model.

```
> corr.test(Satisfaction, Severity)
Call:corr.test(x = Satisfaction, y = Severity)
Correlation matrix
[1] -0.65
Sample Size
[1] 25
Probability values  adjusted for multiple tests.
[1] 0
> #r=-0.65 (strong negative)
> corr.test(Satisfaction, Severity, method = "spearman")
Call:corr.test(x = Satisfaction, y = Severity, method = "spearman")
Correlation matrix
```

[1] -0.6
Sample Size
[1] 25
Probability values adjusted for multiple tests.
[1] 0
> #rho = -0.6 (strong negative)

> #Plot each IV against the DV
> ggplot(data=data,aes(x=Severity,y=Satisfaction))+geom_point()

Satisfaction x Surgical-Medical

As seen from the R-code, Surgical-Medical and Satisfaction have a Pearson Correlation coefficient of -0.10, indicating very little correlation. Figure 15 shows a scatterplot of Surgical Patient Satisfaction scores and Figure 16 shows a scatterplot of Medical Patient Satisfaction scores.

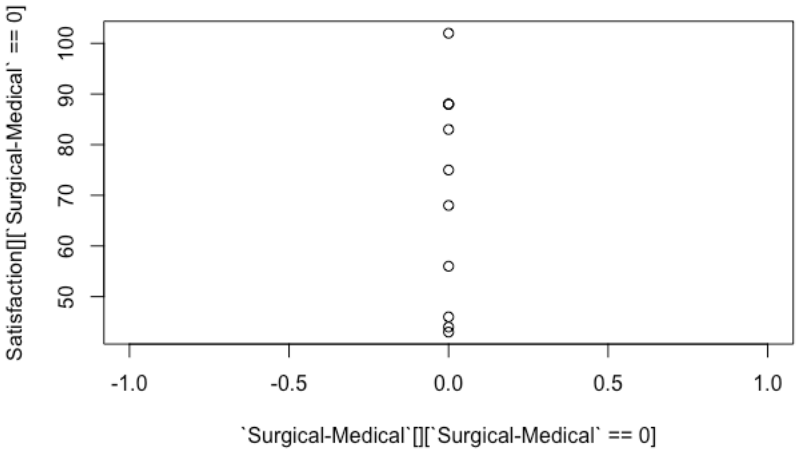


Figure 15: Scatter Plot of Surgical Patient Satisfaction Scores.

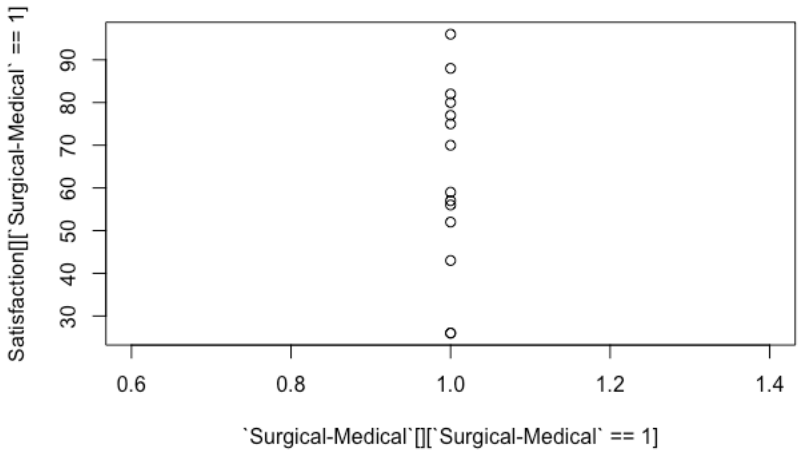


Figure 16: Scatterplot of Medical Patient Satisfaction scores.

```
> corr.test(Satisfaction, `Surgical-Medical`)
Call:corr.test(x = Satisfaction, y = `Surgical-Medical`)
Correlation matrix
[1] -0.18
Sample Size
[1] 25
Probability values  adjusted for multiple tests.
[1] 0.38
> #r=-0.18 (weak negative)
> corr.test(Satisfaction, `Surgical-Medical`, method = "spearman")
Call:corr.test(x = Satisfaction, y = `Surgical-Medical`, method = "spearman")
Correlation matrix
[1] -0.17
Sample Size
[1] 25
Probability values  adjusted for multiple tests.
[1] 0.42
> #rho = -0.17 (weak negative)

> #Plot each IV against the DV
> plot(Satisfaction[][`Surgical-Medical`==0]~`Surgical-Medical`[][`Surgical-Medical`==0])
> plot(Satisfaction[][`Surgical-Medical`==1]~`Surgical-Medical`[][`Surgical-Medical`==1])
```

Satisfaction x Anxiety

As seen from the R-code, Anxiety and Satisfaction have a Pearson Correlation coefficient of -0.51, indicating a strong negative correlation. Spearman's rho, however was calculated at -0.42, indicating a weaker negative correlation. Either way, there's a possible negative correlation, such that as Anxiety increases, Satisfaction decreases. The relationship is made visible by the scatterplot in Figure 17. However, after visual examination, it is unclear whether this trend is linear, or cubic. It is anticipated that Anxiety will remain in the regression model. But, high variability in scores may yield this variable inadequate for the model. However, a transformation of this variable may be required.

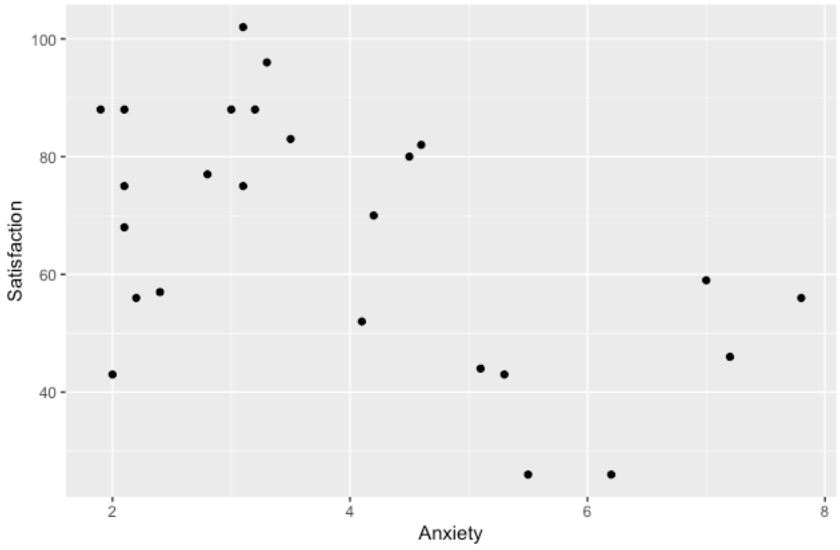


Figure 17: Scatteplot of Satisfaction x Anxiety. A weak correlation is observed, such that as anxiety increases, satisfaction decreases.

Upon further inspection using SPSS' Curve Estimation (see Figure 18 below), cubic regression line appears as a better fit (as indicated by a higher R^2; more on that later). Specifically, a linear regression of Satisfaction as a function of Anxiety has a coefficient of determination of 0.263, indicating that it accounts for only 26.3% of the observed variability (not

very good). However, the cubic regression has a coefficient of determination of 0.546, indicating that it accounts for 54.6% of the observed variability (not great, but better than the linear alternative). Results seem to indicate that a cubic transformation may be necessary.

Model Summary and Parameter Estimates

Dependent Variable: Satisfaction

	Model Summary					Parameter Estimates			
Equation	R Square	F	df1	df2	Sig.	Constant	b1	b2	b3
Linear	.263	8.203	1	23	.009	90.997	-6.174		
Cubic	.546	8.412	3	21	.001	-113.528	153.926	-37.166	2.611

The independent variable is Anxiety.

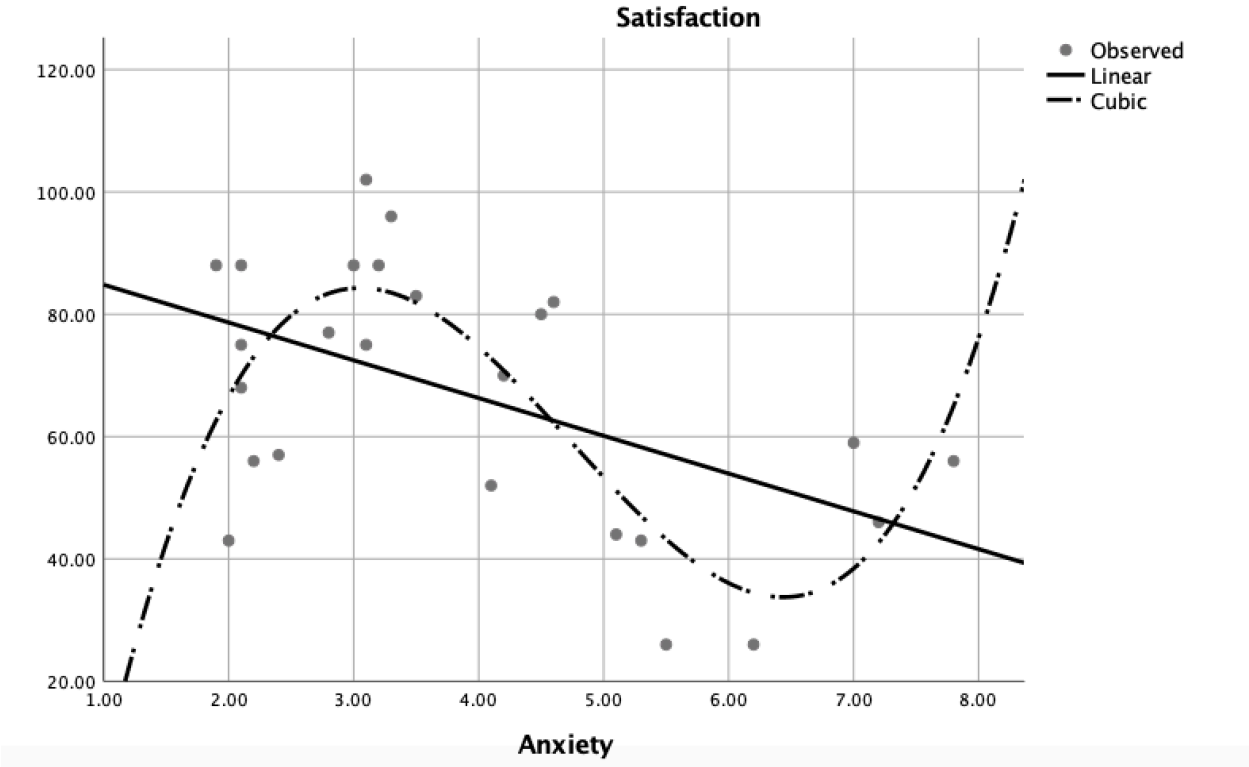


Figure 18: Scatterplot of Anxiety x Satisfaction. Linear and Cubic regression lines are added for comparison. Model summary indicates a higher R^2 for the cubic model and the linear model.

> corr.test(Satisfaction, Anxiety)
Call:corr.test(x = Satisfaction, y = Anxiety)
Correlation matrix
[1] -0.51
Sample Size

[1] 25
Probability values adjusted for multiple tests.
[1] 0.01
> #r=-0.51 (marginally strong negative)
> corr.test(Satisfaction, Anxiety, method = "spearman")
Call:corr.test(x = Satisfaction, y = Anxiety, method = "spearman")
Correlation matrix
[1] -0.42
Sample Size
[1] 25
Probability values adjusted for multiple tests.
[1] 0.04
>#rho = -0.42 (weakly correlated)

> #Plot each IV against the DV
> ggplot(data=data,aes(x=Anxiety,y=Satisfaction))+geom_point()

Model Specification

Modeling Introduction

Given pairs of data $(Y_i, X_{1i}, X_{2i}, X_{3i}, X_{4i}), i = 1, 2, \dots n$, where each i refers to a separate trial, a natural way to represent a linear regression is by fitting a model of the form

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \beta_4 X_{4i} + \varepsilon_i$$

to the observed data. Fitted the data as such serves two purposes. Firstly, it provides a means by which to summarize the observed data such that the relationship between the independent variables $(X_{1i}, X_{2i}, X_{3i}, X_{4i})$ and dependent variable (Y_i) become apparent. Second, it provides a means by which one can predict a value of Y_i for a given value of $X_{1i}, X_{2i}, X_{3i}, X_{4i}$.

In order to build the model at any given point $(X_{1i}, X_{2i}, X_{3i}, X_{4i})$, an estimate of the expected value of the new observation $(\hat{Y}_i)$ or a prediction of $(\hat{Y}_i)$ at the point $(X_{1i}, X_{2i}, X_{3i}, X_{4i})$ is given by:

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_{1i} + \hat{\beta}_2 X_{2i} + \hat{\beta}_3 X_{3i} + \hat{\beta}_4 X_{4i} + \varepsilon_i$$

where $\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2, \hat{\beta}_3, and\ \hat{\beta}_4$ are estimates of the parameters $\beta_0, \beta_1, \beta_2, \beta_3, \beta_4$. These estimated parameters are found using the method of Ordinary Least Squares (OLS; Montgomery et al.,

2012). The OLS method chooses parameters that minimize the sum of the squared residuals (i.e., the difference between the observations and the predicted regression line).

All In Model: Initial Model Proposal

To examine the effect of Age, Severity, Surgical-Medical, and Anxiety on patient Satisfaction scores, a full First-Order Multiple Linear Regression Main Effect model with 3 quantitative independent variables and 1 qualitative independent variable was utilized with the following general form: $Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \beta_4 X_{4i} + \varepsilon_i$

where,

$Y: Patient\ Satisfaction$
$X_1: Age$
$X_2: Severity$
$X_3: Surgical - Medical = \begin{cases} 0\ if\ Surgical \\ 1\ if\ Medical \end{cases}$
$X_4: Anxiety$
$\beta_0: y - intercept\ of\ (k+1)\ dimensonal\ surgace; the\ value\ of\ E(Y)\ when\ X_1 = X_2 = X_3 = X_4 = 0$
$\beta_1: Change\ in\ Satisfaction\ for\ a\ one\ unit\ increase\ in\ Age, when\ all\ other\ variables\ are\ held\ fixed$
$\beta_2: Change\ in\ Satisfaction\ for\ a\ one\ unit\ increase\ in\ Severity, when\ all\ other\ variables\ are\ held\ fixed$
$\beta_3: Change\ in\ Satisfaction\ when\ Surgical - Medical = 1\ , if\ all\ other\ variables\ are\ held\ fixed$
$\beta_4: Change\ in\ Satisfaction\ for\ a\ one\ unit\ increase\ in\ Anxiety, when\ all\ other\ variables\ are\ held\ fixed$

All In Model: Test For Significance of Regression

To test for the significance of the regression, a global F-test was conducted. The test statistic is:

$$F = \frac{\frac{(SS_{yy} - SS_{Res})}{k}}{\frac{SS_{Res}}{[n-(k+1)]}} = \frac{\frac{R^2}{k}}{\frac{(1-R^2)}{[n-(k+2)]}} = \frac{Mean\ Square\ (Model\ or\ MS_{Reg})}{Mean\ Square\ (Error\ or\ MS_{Res})}$$

$$= \frac{Variability\ in\ y\ explained\ by\ the\ model}{Unexplained\ Variability}$$

where,

$n = sample\ size$

$k = number\ of\ terms\ in\ the\ model$

Hypothesis testing was conducted as follows:

$H_0: \beta_1 = \beta_2 = \beta_3 = \beta_4 = 0 \quad (All\ model\ terms\ are\ unimportant\ for\ predicting\ y)$

$H_a: At\ least\ one\ \beta_i \neq 0 \quad (At\ least\ one\ model\ term\ is\ useful\ for\ predicting\ y)$

The rejection region was considered:

- $F_{obtained} > F_{\alpha}$, with k numerator degrees of freedom and $[n - (k + 1)]$ denominator degrees of freedom
- $\alpha > p - value$

The decision rule was:

- Reject H_0 if $\alpha > p - value$
- Reject if test statistic falls in rejection region where $F_{obtained} > F_{critical}$

Here, we choose $\alpha = 0.05$.
Since n = 25 and k = 4,

- numerator df = 4
- denominator df = $[n - (k + 1)] = 25 - (4 + 1) = 25 - 5 = 20$

Then the rejection region for the F-test is: $F_{critical} < F_{0.05,4,20}$

- where, $F_{critical} = 2.866$

From the R-output copied below, we find $F_{obtained} = 22.51$

Compare F_c and $F_{0.05,4,20}$:

$$(F_{critical} = 2.866) < \left(F_{0.05,4,20} = 22.51\right)$$

Conclusion: Because the calculated test statistic falls in the critical region, we reject the null hypothesis and conclude that at least one β is non-zero (i.e., at least one model term is useful for predicting Satisfaction).

We can also compare $\alpha = 0.05$ to the p-value and come to the same conclusion.

$$(P - value = 3.611\text{e} - 07) < \ (\alpha = 0.05)$$

Thus, the All-In Fitted Model is:

$$Satisfaction = 140.1689 - (1.1428\ * Age_i) - (0.4699 * Severity_i)$$

$$+ (2.2259 * SurgicalMedical_i) + (1.2673 * Anxiety_i) + \varepsilon_i$$

A scatterplot of the fitted All-In Model is found in Figure 18:

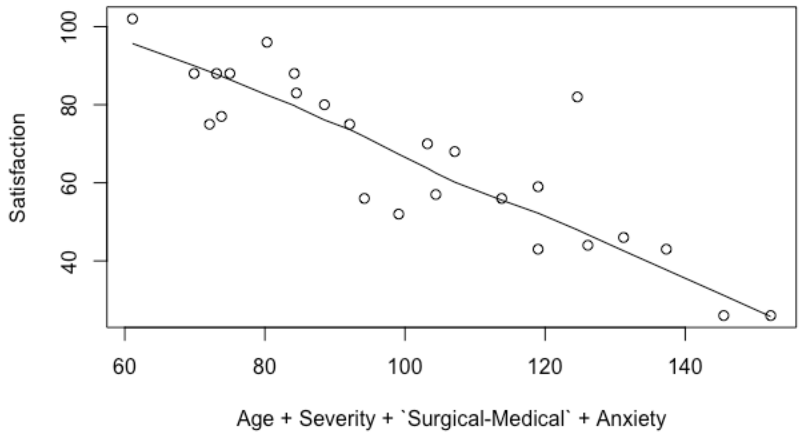


Figure 18: First-Order Linear Main Effect Model of Satisfaction as a function of Age, Severity, Surgical-Medical, and Anxiety

```
> #Start with All-In Model (First Order Linear Multiple Regression, Main Effects)
> AllInModel <-lm(Satisfaction~Age+Severity+`Surgical-Medical`+Anxiety, data = data)

> #Obtain results
> summary(AllInModel)

Call:
lm(formula = Satisfaction ~ Age + Severity + `Surgical-Medical` +
    Anxiety, data = data)

Residuals:
     Min       1Q   Median       3Q      Max
-18.506   -5.096    1.306    4.738   28.722

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)     140.1689    8.3191  16.849 2.77e-13 ***
Age              -1.1428     0.1904  -6.002 7.22e-06 ***
Severity         -0.4699     0.1866  -2.518   0.0204 *
`Surgical-Medical`  2.2259     4.1402   0.538   0.5968
Anxiety          1.2673     1.4922   0.849   0.4058
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 9.921 on 20 degrees of freedom
```

Multiple R-squared: 0.8183, Adjusted R-squared: 0.7819
F-statistic: 22.51 on 4 and 20 DF, p-value: 3.611e-07

```
> #Satisfaction = 140.1689 - (1.1428*Age) - (0.4699*Severity) + (2.2259*Surgical-Medical) +
(1.2673*Anxiety)
> #R^2 = 0.8183, Ra^2 = 0.7819
> #Beta Coefficients for Age & Severity have p<0.05

> #Confirm coefficients
> coefficients(AllInModel) #model coefficients
     (Intercept)              Age         Severity `Surgical-Medical`          Anxiety
     140.1689450       -1.1428491       -0.4698578        2.2258836        1.2672630

> #Get standard deviations
> sapply(data, sd)
      Satisfaction            Age       Severity Surgical.Medical        Anxiety
      21.2436657     14.8090063     13.0285584      0.5066228      1.7641617

> #Confirm with ols package
> ols_regress(AllInModel)
                Model Summary
---------------------------------------------------------------
R                       0.905       RMSE                 9.921
R-Squared               0.818       Coef. Var           14.870
Adj. R-Squared          0.782       MSE                 98.427
Pred R-Squared          0.735       MAE                  6.723
---------------------------------------------------------------
 RMSE: Root Mean Square Error
 MSE: Mean Square Error
 MAE: Mean Absolute Error

                        ANOVA
-------------------------------------------------------------------------------
                Sum of
               Squares      DF    Mean Square       F         Sig.
-------------------------------------------------------------------------------
Regression     8862.507      4       2215.627       22.51     0.0000
Residual       1968.533     20         98.427
Total          10831.040    24
-------------------------------------------------------------------------------
```

Parameter Estimates

model	Beta	Std. Error	Std. Beta	t	Sig	lower	upper
(Intercept)	140.169	8.319	16.849	0.000	122.816	157.522	
Age	-1.143	0.190	-0.797	-6.002	0.000	-1.540	-0.746
Severity	-0.470	0.187	-0.288	-2.518	0.020	-0.859	-0.081
`Surgical-Medical`	2.226	4.140	0.053	0.538	0.597	-6.410	10.862
Anxiety	1.267	1.492	0.105	0.849	0.406	-1.845	4.380

```
> #Remember, Fc = (SSR/k) / (SSE/n-k+1)
> #Alpha = .05
> #Fo = F(1-alpha, df, df)
> #Fo = F(1-.05,4,20)
> qf(0.95,4,20)
[1] 2.866081

> #Plot All In Model
> plot(Age+Severity+`Surgical-Medical`+Anxiety, Satisfaction, data=data)
> #Appears linear
> #Can also plot with ols library
> ols_plot_reg_line(Satisfaction,Age+Severity+`Surgical-Medical`+Anxiety)
> ols_plot_response(AllInModel)
```

All In Model: Calculate R^2 And R^2_{Adj}

The Coefficient of Determination (R^2) serves as another utility measure of the regression model. It indicates the variability in the dependent variable that can be predicted by the independent variable (i.e., the contribution of *x* in predicting *y*).

$$R^2 = \frac{Explained\ Sample\ Variability}{Total\ Sample\ Variability} = 1 - \frac{SS_{Res}}{SS_{yy}} \qquad 0 \leq R^2 \leq 1$$

where,

$R^2 = 0\ implies\ a\ complete\ lack\ of\ fit\ of\ the\ model\ to\ the\ data$

$R^2 = 1\ implies\ a\ perfect\ fit, with\ the\ model\ passing\ through\ every\ data\ point$

As determined from the R-output listed below, $R^2 = 0.818$. Thus, 66.4% of the sample variation in $Y\ (Satisfaction)$ can be explained by (or attributed to) the current model.

The Adjusted Multiple Coefficient of Determination (R_a^2) adjusts for sample size n and the number of β parameters in the model. It should be noted that R_a^2 will always be smaller than R^2 and cannot be forced to 1 by simply adding more independent variables to the model. For some poor-fitting models, R_a^2 may be negative.

$$R_a^2 = \frac{Explained\ Sample\ Variability}{Total\ Sample\ Variability} = 1 - \left[\frac{(n-1)}{n-(k+1)}\right]\frac{SS_{Res}}{SS_{yy}}$$

$$= 1 - \left[\frac{(n-1)}{n-(k+1)}\right](1-R^2)$$

Here,

$$R_a^2 = 1 - \left[\frac{(25-1)}{25-(4+1)}\right](1-0.818) = 1 - \left(\frac{24}{20}\right)(1-0.818) = 0.782$$

Indeed, we can confirm from the output below that $R_a^2 = 0.782$

```
> #R^2
> R2 <-summary(AllInModel)$r.squared
> R2
[1] 0.8182508

> #Confirm with ols package
> ols_regress(AllInModel)
                Model Summary
---------------------------------------------------------------
R                       0.905       RMSE              9.921
R-Squared               0.818       Coef. Var        14.870
Adj. R-Squared          0.782       MSE              98.427
Pred R-Squared          0.735       MAE               6.723
---------------------------------------------------------------
 RMSE: Root Mean Square Error
 MSE: Mean Square Error
 MAE: Mean Absolute Error
```

ANOVA

	Sum of Squares	DF	Mean Square	F	Sig.
Regression	8862.507	4	2215.627	22.51	0.0000
Residual	1968.533	20	98.427		
Total	10831.040	24			

Parameter Estimates

model	Beta	Std. Error	Std. Beta	t	Sig	lower	upper
(Intercept)	140.169	8.319	16.849	0.000	122.816	157.522	
Age	-1.143	0.190	-0.797	-6.002	0.000	-1.540	-0.746
Severity	-0.470	0.187	-0.288	-2.518	0.020	-0.859	-0.081
`Surgical-Medical`	2.226	4.140	0.053	0.538	0.597	-6.410	10.862
Anxiety	1.267	1.492	0.105	0.849	0.406	-1.845	4.380

All In Model: Calculate Confidence Intervals

A confidence interval gives an estimated range of values which is likely to include an unknown population parameter, the estimated range being calculated from a given set of sample data. Confidence interval for each β_i is given by:

$$\left[\hat{\beta}_i - \left(t_{\frac{\alpha}{2},n-2}\right) s\hat{\beta}_i\right] \leq \hat{\beta}_i \leq \left[\hat{\beta}_i + \left(t_{\frac{\alpha}{2},n-2}\right) s\hat{\beta}_i\right]$$

where,

$$s = \frac{s}{\sqrt{SS_{xx}}}$$

Since:

$95\%\ C.I. means\ we\ need\ \alpha = 0.05$

For example, $\hat{\beta}_1$(the parameter for Age): $-1.5400724 \leq \beta_1 \leq -0.74562580$

$\therefore 95\%\ of\ such\ intervals\ will\ include\ the\ true\ value\ of\ the\ slope$

Confidence intervals were calculated for each coefficient and are presented below.

```
> #Confidence Intervals (95%)
> confint(AllInModel)
```

	2.5 %	97.5 %
(Intercept)	122.8155628	157.52232723
Age	-1.5400724	-0.74562580
Severity	-0.8590832	-0.08063232
`Surgical-Medical`	-6.4104312	10.86219832
Anxiety	-1.8454028	4.37992884

```
> #Age: Small Range, does not include 0
> #Severity: Small range, does not include 0
> #Surgical-Medical: Large range, includes 0
> #Anxiety: Large range, includes 0
```

All In Model: Determine Contribution of Each Independent Variable

To determine the individual contribution of each independent variable, individual *t*-tests were used to access individual β-parameters. $\alpha = 0.05$ was selected. Since n = 25, df = 20 and, for each $\hat{\beta}_i$:

$$t = \frac{\hat{\beta}_i}{s_{\beta_i}} = \frac{\hat{\beta}_i - 0}{\frac{s}{\sqrt{SS_{xx}}}}$$

Two-tailed hypothesis testing was conducted:

- $H_0: \beta_i = 0$
- $H_a: \beta_i \neq 0$

The rejection region was considered when: $|t| > t_{\alpha/2}$ or $|t| > t_{0.025}$ which = 2.056

And, the decision rule to reject H_0 was:

- Reject H_0 if $\alpha > p - value$
- Reject H_0 if test statistic falls in rejection region such that $t_{obtained} > t_{\alpha/2}$

For β_1, the output given below shows:

$$|(t_{obtained} = -6.002)| > (t_{0.025} = 2.056) \text{ and } (p - value = 0.00) < (\alpha = 0.05)$$

$\therefore We\ conclude\ that\ \beta_1\ makes\ a\ significant\ contribution\ to\ the\ proposed\ model$

For β_2, the output given above shows:

$|(t_{obtained} = -2.518)| > (t_{0.025} = 2.056)$ and $(p - value = 0.020) < (\alpha = 0.05)$

$\therefore We\ conclude\ that\ \beta_2\ makes\ a\ significant\ contribution\ to\ the\ proposed\ model$

For β_3, the output given above shows:

$|(t_{obtained} = 0.538)| < (t_{0.025} = 2.056)$ and $(p - value = 0.597) > (\alpha = 0.05)$

$\therefore We\ conclude\ that\ \beta_3\ fails\ to\ make\ a\ significant\ contribution\ to\ the\ proposed\ model$

For β_4, the output given above shows:

$|(t_{obtained} = 0.849)| < (t_{0.025} = 2.056)$ and $(p - value = 0.406) > (\alpha = 0.05)$

$\therefore We\ conclude\ that\ \beta_4\ fails\ to\ make\ a\ significant\ contribution\ to\ the\ proposed\ model$

Thus, we see that it is unnecessary to include both $X_3\ and\ X_4$ in the model.

```
> #Confirm with ols package
> ols_regress(AllInModel)
                Model Summary
---------------------------------------------------------------
R                  0.905       RMSE               9.921
R-Squared          0.818       Coef. Var         14.870
Adj. R-Squared     0.782       MSE               98.427
Pred R-Squared     0.735       MAE                6.723
---------------------------------------------------------------
 RMSE: Root Mean Square Error
 MSE: Mean Square Error
 MAE: Mean Absolute Error

                       ANOVA
--------------------------------------------------------------------------------
                Sum of
               Squares        DF    Mean Square      F         Sig.
--------------------------------------------------------------------------------
Regression     8862.507        4     2215.627        22.51     0.0000
Residual       1968.533       20     98.427
Total          10831.040      24
--------------------------------------------------------------------------------

                            Parameter Estimates
-----------------------------------------------------------------------------------------------------------------
 model          Beta       Std. Error      Std. Beta       t         Sig         lower      upper
-----------------------------------------------------------------------------------------------------------------
 (Intercept)    140.169        8.319          16.849       0.000    122.816    157.522
  Age           -1.143         0.190          -0.797      -6.002     0.000     -1.540     -0.746
```

Severity	-0.470	0.187	-0.288	-2.518	0.020	-0.859	-0.081
`Surgical-Medical`	2.226	4.140	0.053	0.538	0.597	-6.410	10.862
Anxiety	1.267	1.492	0.105	0.849	0.406	-1.845	4.380

Residual Analysis

The overall strategy employed for detecting model lack of fit with residuals was:

1. Plot the residuals, $\hat{\varepsilon}$, on the vertical axis against each of the independent variables, $x_1, x_2, \ldots, x_k$, on the horizontal axis
2. Plot the residuals, $\hat{\varepsilon}$, on the vertical axis against the predicted value, $\hat{y}$, on the horizontal axis.
3. In each plot, look for trends, dramatic changes in variability, and/or more than 5% of residuals that lie outside 2s of 0. Any of these patterns indicate a problem with model fit.

A Normal Probability Plot (P-Plot) tests the normality of residuals. A normal P-Plot compares observed cumulative distribution function (CDF) of the standard residual to expected CDF of the normal distribution. If residuals are normally distributed, the P-Plot will plot as a straight line, which is usually determined visually, with an emphasis on the central values rather than extremes. Similarly, a Q-Q Plot compares observed quantile with theoretical quantile of a normal distribution. From Figures 19 and 20, we see a relatively normal distribution, with residuals plotted in an approximately straight line. There are no small or large deviations at either tail of the reference line, indicated the data is not skewed.

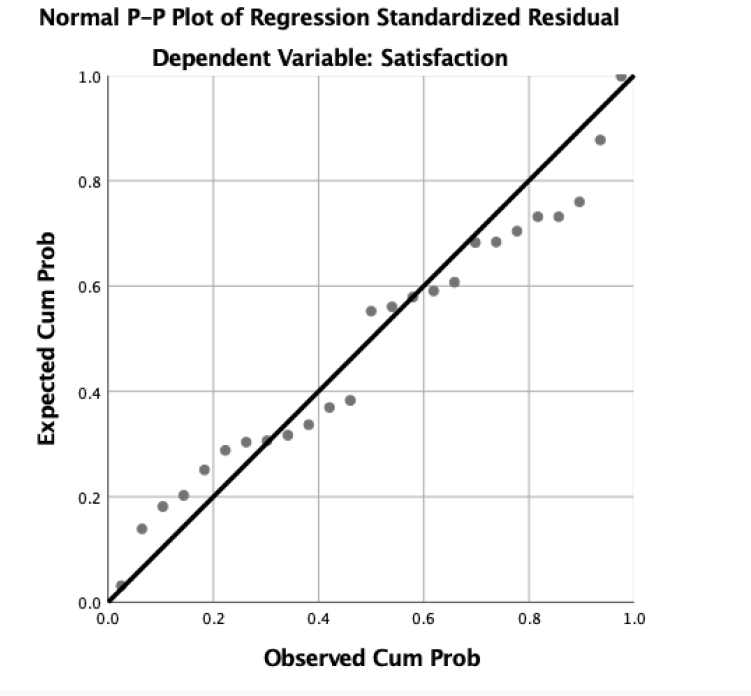


Figure 19: Normal Probability Plot for the All In Model. Assumption of Normality maintained.

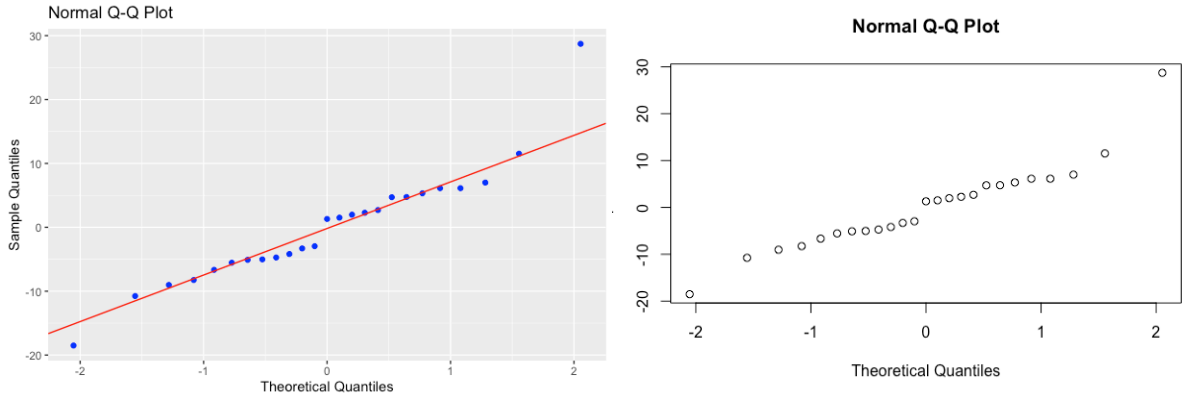


Figure 20: QQ-Plot for the All In Model. Assumption of Normality Obtained. Though, we may have one outlier in the dataset.

We can confirm a relatively normal distribution shape by examining the histogram, as well. Thus, there does not seem to be any problem with the normality assumption. A similar normality is observed in the residual box plot in Figure 22 below.

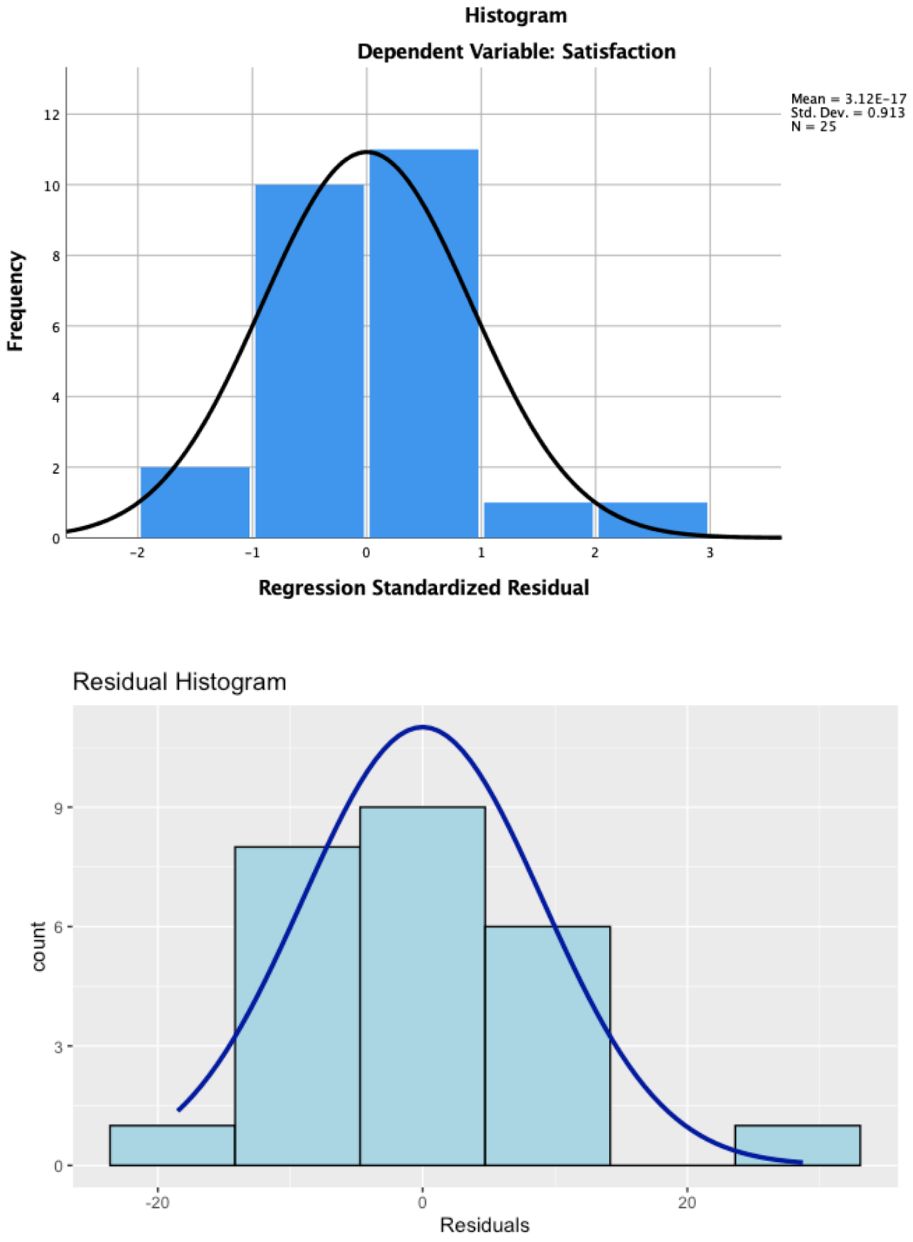


Figure 21: All In Model Residual Histogram indicating a normal distribution.

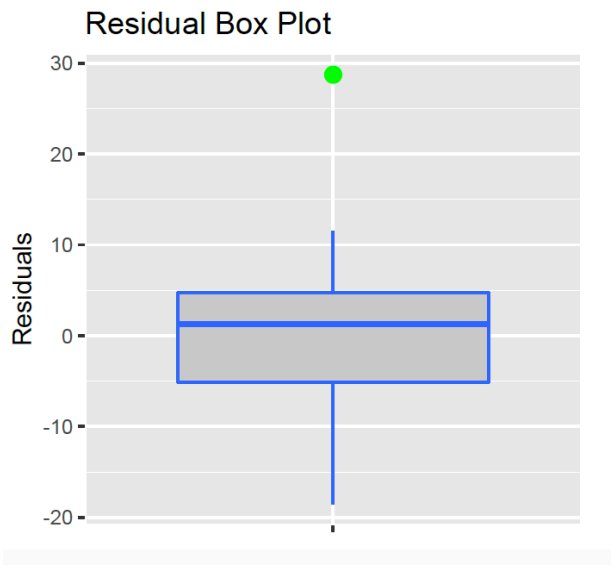
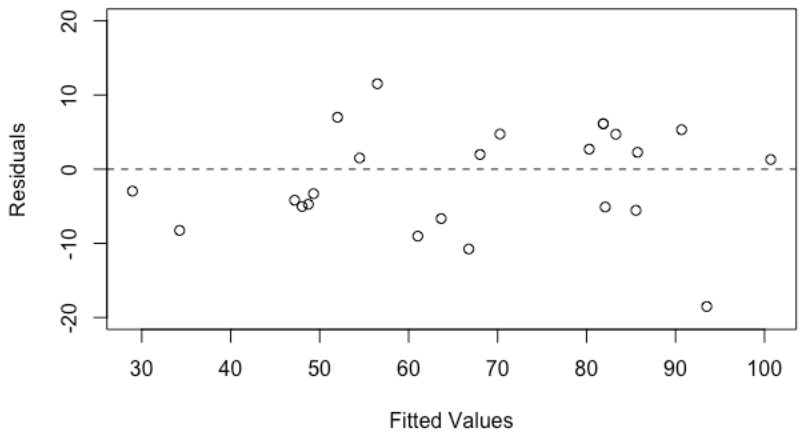


Figure 22: Residual Box Plot for the All In Model

When plotting the residuals vs the predicted response, we examine the shape of the data points. If the residuals can be contained in a horizontal band, then there are no obvious model defects. Figure 21 shows that the data points are mostly contained in a horizontal band. There are, however, a few observations with greater variance.



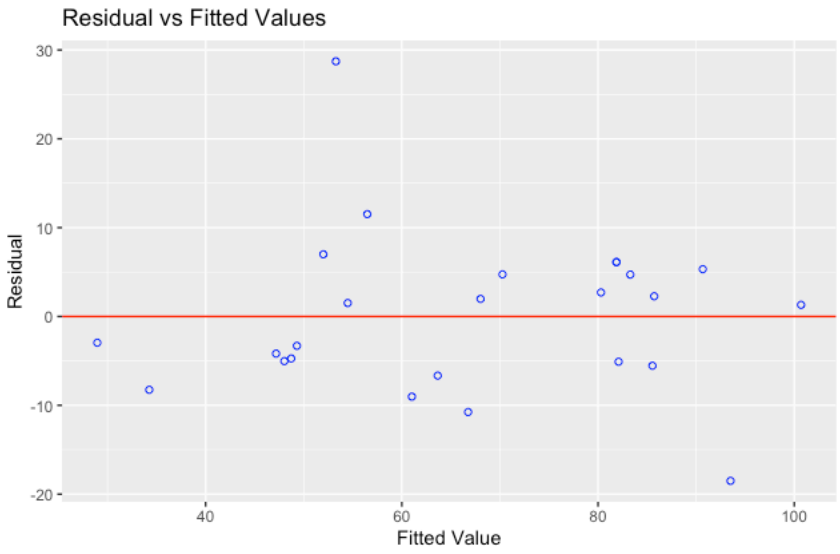


Figure 22: Residuals vs Predicted.

Studentized residuals:

$$r_i = \frac{e_i}{\sqrt{MS_{Res}(1 - h_{ii})}}, \quad i = 1,2,..n$$

Studentized residuals have a constant variance $Var(r_i) = 1$ regardless of the location of x_i, when the model form is correct. Studentized residuals are going to be more effective for detecting outlying Y observations than standardized residuals. Studentized residuals are internally scaled because MS_{Res}

is an internally generated estimate of σ^2 obtained from fitting the model to all n observations.

R-Student (aka studentized deleted residual)

Rather than using MS_{Res} to estimate σ^2, R-Student removes the ith observation, making it an externally studentized residual. R-Student is given by:

$$t_i = \frac{e_i}{\sqrt{S^2_{(i)}(1 - h_{ii})}}, \quad i = 1,2, \ldots, n$$

where,

$$S_{(i)}^2 = \frac{(n-p)MS_{Res} - e_i^2/(1-h_{ii})}{n-p-1}$$

In many situations, t_i will differ little from the studentized residual r_i. However, if the ith observation is influential, then $S_{(i)}^2$ can differ significantly from MS_{Res}, and thus the R-Student statistic will be more sensitive to this point. Generally, if an observation has an externally studentized residual that is larger than 3 (in absolute value) we can call it an outlier. See Figure 24 for Studentized, and Studentized Deleted Residuals. We see that the studentized residual has a minimum of -2.087, a maximum of 3.083, a mean of -0.99 and a standard deviation of 1.002. The R-Student (or studentized deleted residual) has a minimum of -2.301, a maximum of 4.148 (a potential outlier), a mean of 0.026, and a standard deviation of 1.166. Here we see that observations 17 and 9 may be outliers.

Residuals Statistics[a]

	Minimum	Maximum	Mean	Std. Deviation	N
Predicted Value	28.9562	100.6939	66.7200	19.21643	25
Std. Predicted Value	-1.965	1.768	.000	1.000	25
Standard Error of Predicted Value	2.785	6.576	4.359	.843	25
Adjusted Predicted Value	29.7576	100.3194	66.9052	19.30933	25
Residual	-18.50651	28.72233	.00000	9.05661	25
Std. Residual	-1.865	2.895	.000	.913	25
Stud. Residual	-2.087	3.083	-.009	1.002	25
Deleted Residual	-23.17436	32.56689	-.18515	10.93280	25
Stud. Deleted Residual	-2.301	4.148	.026	1.166	25
Mahal. Distance	.931	9.583	3.840	1.897	25
Cook's Distance	.001	.254	.041	.064	25
Centered Leverage Value	.039	.399	.160	.079	25

a. Dependent Variable: Satisfaction

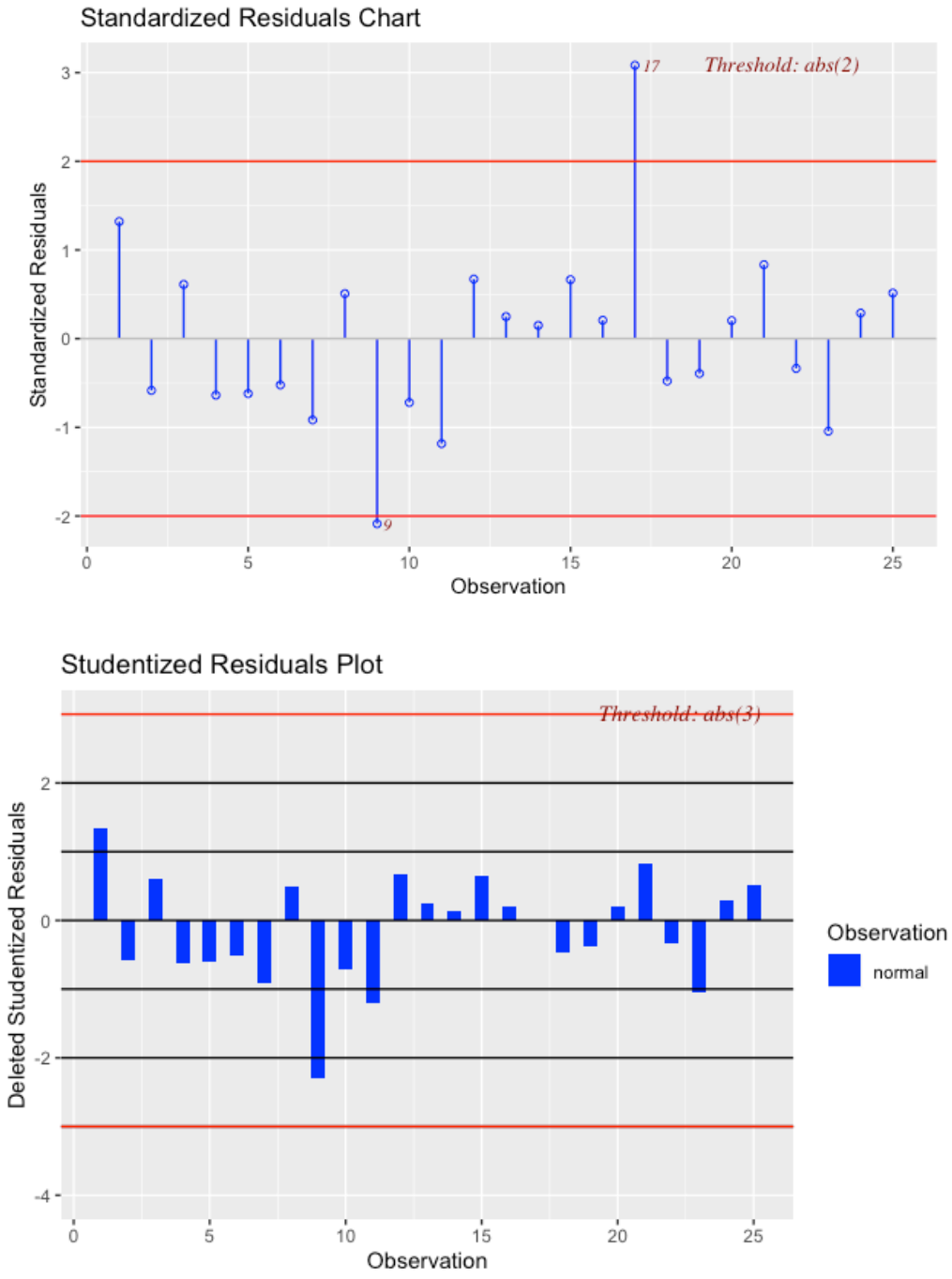


Figure 24: Studentized and Studentized Deleted Residuals for the All In Model.

> #Obtain Fitted Values
> fitted(AllInModel) #predicted values
1 2 3 4 5 6 7 8 9 10
56.48061 82.09552 90.67787 85.54462 48.02362 48.72672 34.25319 83.28778 93.50651
63.65807

```
      11       12       13       14       15       16       17       18       19       20
 66.75721  81.87765  85.72574 100.69392  81.87534  68.02170  53.27767  47.18047  49.30541
54.48526
      21       22       23       24       25
 52.00206  28.95617  61.02552  80.29922  70.26216

> #Regression Diagnostics Battery
> ols_plot_diagnostics(AllInModel)

> #Obtain residuals
> residuals(AllInModel) #residuals
       1        2        3        4        5        6        7        8        9
 11.519393  -5.095519   5.322134  -5.544620  -5.023622  -4.726724  -8.253185   4.712223 -
18.506508
       10       11       12       13       14       15       16       17       18
 -6.658066 -10.757209   6.122350   2.274264   1.306083   6.124662   1.978296  28.722334  -
4.180474
       19       20       21       22       23       24       25
 -3.305412   1.514739   6.997937  -2.956166  -9.025522   2.700776   4.737837

> e <-residuals(AllInModel)
> e
       1        2        3        4        5        6        7        8        9
 11.519393  -5.095519   5.322134  -5.544620  -5.023622  -4.726724  -8.253185   4.712223 -
18.506508
       10       11       12       13       14       15       16       17       18
 -6.658066 -10.757209   6.122350   2.274264   1.306083   6.124662   1.978296  28.722334  -
4.180474
       19       20       21       22       23       24       25
 -3.305412   1.514739   6.997937  -2.956166  -9.025522   2.700776   4.737837

> boxplot(e, ylab = "Residuals")
> #Plot residuals against Y_hat
> yhat <- fitted(AllInModel)
> plot(yhat, e, xlab = "Fitted Values", ylab = "Residuals", ylim = c(-20,20))
> abline(h=0, lty = 2)

> #Plot residuals against each of the predictor variables
> #Age Residuals
> plot(Age,e, xlab="Age", ylab = "Residuals", ylim = c(-20,20))
> abline(h=0, lty=2)

> #Severity Residuals
> plot(Severity, e, xlab = "Severity", ylab = "Residuals", ylim = c(-20,20))
> abline(h=0, lty=2)
```

```
> #Surgical-Medical Residuals
> plot(`Surgical-Medical`, e, xlab = "Surgical-Medical", ylab = "Residuals", ylim = c(-20,20))
> abline(h=0,lty=2)

> #Anxiety
> plot(Anxiety,e, xlab = "Anxiety", ylab = "Residuals", ylim = c(-20,20))
> abline(h=0, lty=2)

> #3d Scatterplot
> scatterplot3d(Age, Severity, e, xlab = "Age", ylab = "Severity", zlab = "Residuals")
> scatterplot3d(Age,`Surgical-Medical`, e, xlab = "Age", ylab = "Surgical-Medical", zlab =
"Residuals")
> scatterplot3d(Age,Anxiety, e, xlab = "Age", ylab = "Anxiety", zlab = "Residuals")
> scatterplot3d(Severity, `Surgical-Medical`, e, xlab = "Severity", ylab = "Surgical-Medical",
zlab = "Residuals")
> scatterplot3d(Severity, Anxiety, e, xlab = "Severity", ylab = "Anxiety", zlab = "Residuals")
> scatterplot3d(`Surgical-Medical`, Anxiety, e, xlab = "Surgical-Medical", ylab = "Anxiety",
zlab = "Residuals")

> #More Residuals
> summary(AllInModel)$residuals
       1        2        3        4        5        6        7        8        9
 11.519393  -5.095519   5.322134  -5.544620  -5.023622  -4.726724  -8.253185   4.712223 -
18.506508
      10       11       12       13       14       15       16       17       18
 -6.658066 -10.757209   6.122350   2.274264   1.306083   6.124662   1.978296  28.722334  -
4.180474
      19       20       21       22       23       24       25
 -3.305412   1.514739   6.997937  -2.956166  -9.025522   2.700776   4.737837

> ols_plot_resid_stud(AllInModel)

> #Residual Normality Test
> ols_test_normality(AllInModel)
-----------------------------------------------
      Test                  Statistic      pvalue
-----------------------------------------------
Shapiro-Wilk                0.9203       0.0520
Kolmogorov-Smirnov          0.1399       0.6616
Cramer-von Mises            2.0933       0.0000
Anderson-Darling            0.5868       0.1151
-----------------------------------------------

> #Correlation Between Observed Residuals and Expected Residuals Under Normality
> ols_test_correlation(AllInModel)
[1] 0.9483583
```

```
> #Correlation = 0.9483583

> #Residual vs Fitted Values Plot
> ols_plot_resid_fit(AllInModel)

> #Residual Histogram
> ols_plot_resid_hist(AllInModel)

> #Studentized Residuals vs Leverage Plot
> ols_plot_resid_lev(AllInModel)

> #Deleted Studentized residual vs predicted values
> ols_plot_resid_stud_fit(AllInModel)

> #Residuals
> r<-resid(AllInModel)
> r
       1        2        3        4        5        6        7        8        9
 11.519393  -5.095519   5.322134  -5.544620  -5.023622  -4.726724  -8.253185   4.712223 -
18.506508
      10       11       12       13       14       15       16       17       18
 -6.658066 -10.757209   6.122350   2.274264   1.306083   6.124662   1.978296  28.722334  -
4.180474
      19       20       21       22       23       24       25
 -3.305412   1.514739   6.997937  -2.956166  -9.025522   2.700776   4.737837

> #QQ Plot
> qqnorm(resid(AllInModel))
> ols_plot_resid_qq(AllInModel)

> #Predictively adjusted residuals
> (pr<-resid(AllInModel)/(1-lm.influence(AllInModel)$hat))
       1        2        3        4        5        6        7        8        9
 14.904925  -6.587249   6.912657  -7.207375  -7.521660  -5.676004 -10.015773   5.388592 -
23.174362
      10       11       12       13       14       15       16       17       18
 -7.668690 -12.819820   7.261376   2.675657   1.680566   7.109580   2.147497  32.566885  -
5.374039
      19       20       21       22       23       24       25
 -4.601232   2.701501   9.777218  -3.757628 -11.872648   3.025197   5.496027

> #Construct residual vs predicted response
> plot(predict(AllInModel), resid(AllInModel))
> #standardized residuals
> ols_plot_resid_stand(AllInModel)
```

```
> #Normal Probability Plot of Residuals
> n <-length(e)
> MSE <-sum(e^2)/(n-4)
> RankofRes <-rank(e)
> Zscore <- qnorm((RankofRes-0.375)/(n+0.25))
> ExpRes <- Zscore * sqrt(MSE)
> plot(ExpRes, e, xlab = "Expected Score", ylab = "Residuals")
> abline(a = 0, b = 1)

> #Verify Assumption of Constant Variance (Breusch-Pagan test)
> SSE <- sum(e^2)
> SSE
[1] 1968.533
> #1968.533
> n <-length(e)
> reg2 <-lm (e^2~Age+Severity+`Surgical-Medical`+Anxiety, data = data)
> y2hat<-fitted(reg2)
> SSR2 <-sum((y2hat-mean(y2hat))^2)
> SSR2
[1] 43186.72
> #43186.72
> chiBP <-(SSR2/2)/(SSE/n)^2
> chiBP
[1] 3.482692
> #3.482692
> chiTAB<-qchisq(0.99,4)
> chiTAB
[1] 13.2767
> #13.2767
> chiTab <-qchisq(0.95,4)
> chiTab
[1] 9.487729
```

PRESS

R^2, also known as coefficient of determination, is a popular measure of quality of fit in regression. However, it does not offer any significant insights into how well our regression model can predict future values. Instead, the PRESS statistic (the predicted residual sum of squares) can be used as a measure of predictive power. The PRESS statistic can be computed in the leave-one-out cross validation process, by adding the square of the residuals for the case that

is left out. As a reminder, in the leave-one-out cross validation, one case of the data set is used as the testing set and the remaining are used as the testing set. We iterate this process, until all cases have served as the testing set. Or, we can calculate it in the following manner:

```
> #Cross-validated residuals
> #Regular RSS is
> sum(r^2)
[1] 1968.533

> #PRESS is
> sum(pr^2)
[1] 2869.485
> #2869.485
> #2869.485
> #Note PRESS is bigger because predicting is harder than fitting
```

All In Model: Search For High-Leverage Or Overly Influential Observations

Outliers can be found by examining standardized residuals, leverage (how far an observation deviates from the mean of that variable), Cook's Distance (which combines leverage and residuals), and DfBeta (which is a more specific measure of influence).

Standardized Residuals: Outliers have values greater than 2 and less than -2. From Figure 25, we see that observations 17 and 9 may be outliers.

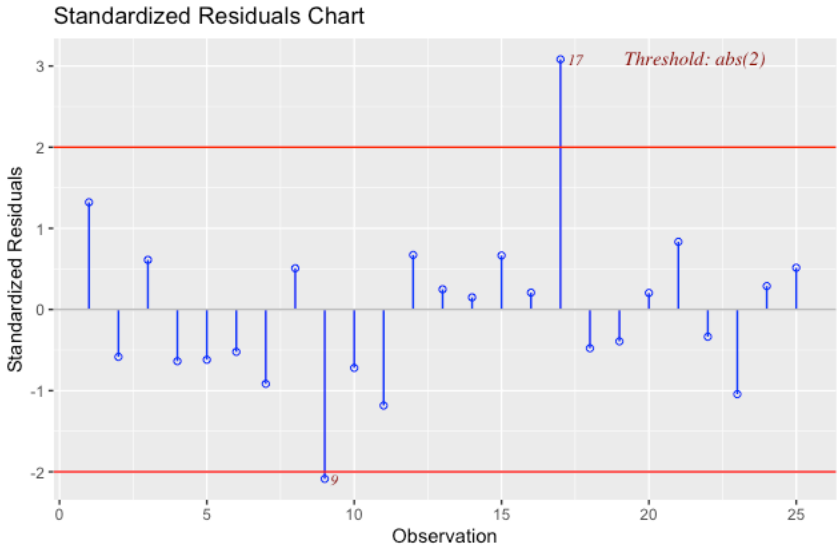


Figure 25: Standardized Residuals for the All In Model. Outliers observed at observation 17 and 9.

Leverage indicates how far an observation deviates from the mean of that variable. It is calculated by:

$$Leverage = \frac{2k+2}{n}, \quad where\ k = \#\ predictors, n = \#\ observations$$

For this data set, the leverage threshold is 0.4. After sorting the leverage threshold for each value and by visually examining Figure 26, we see that observations 17 and 9 are outliers and observation 20 has an abnormally high/strong leverage.

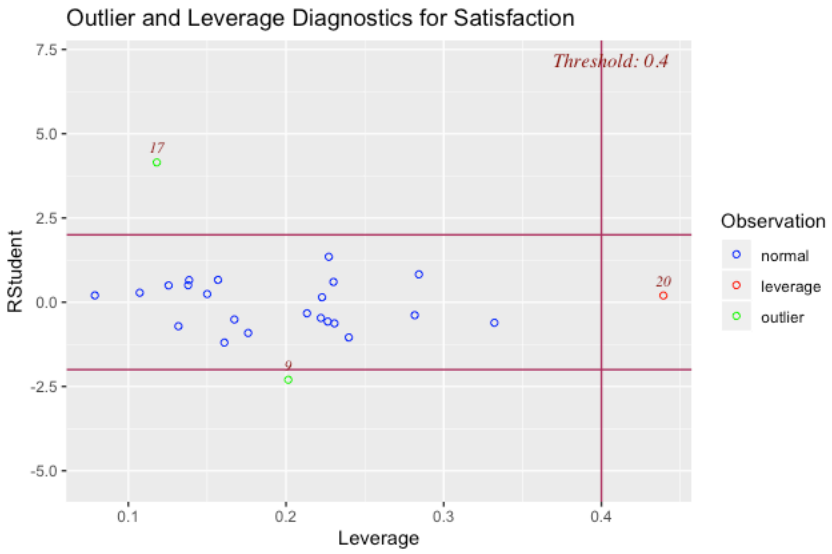
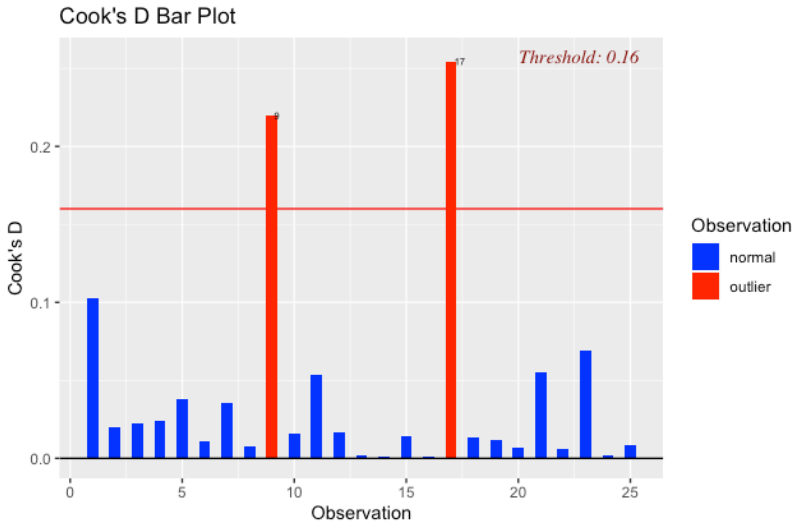


Figure 26: Leverage Diagnostic for the All In Model

Cook's Distance combines leverage and residuals. The general rule of thumb is the higher the value, the better. The lowest possible value is 0 and the conventional cut off is calculated by 4/n (which is 0.16 for this data set). After sorting each observation's Cook's Distance and analyzing the diagnostic plots below in Figure 27, we find two observations below the cut off (9 and 17).



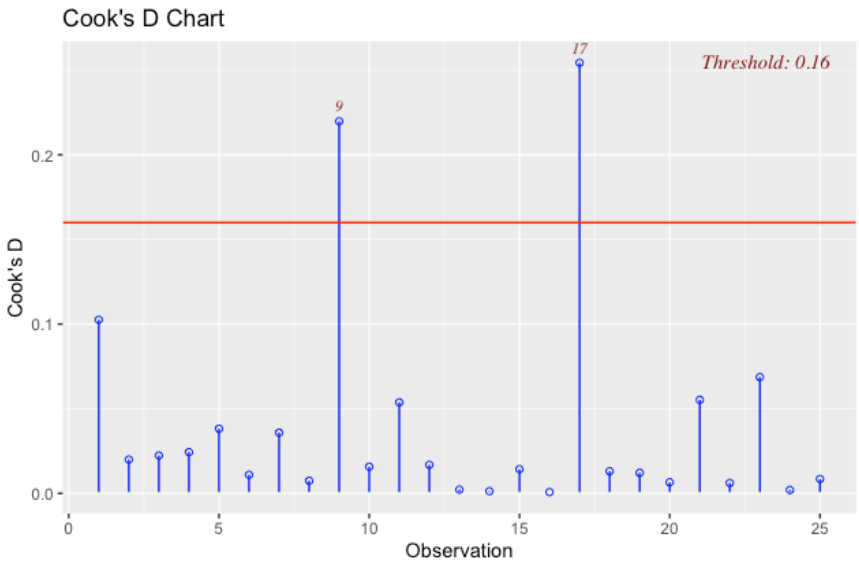
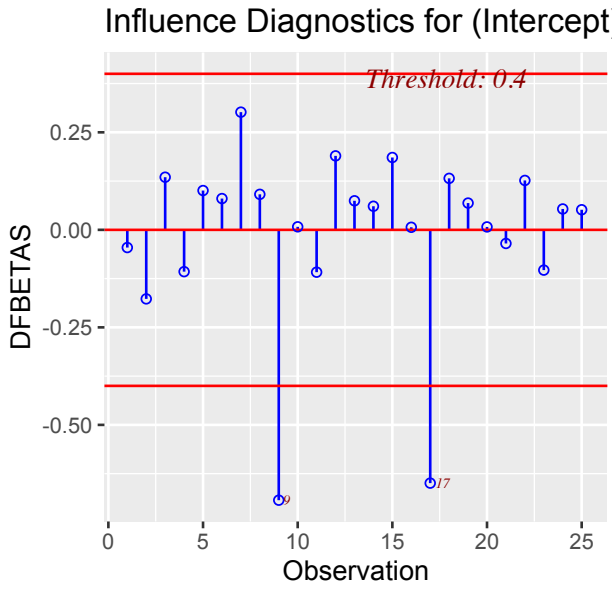
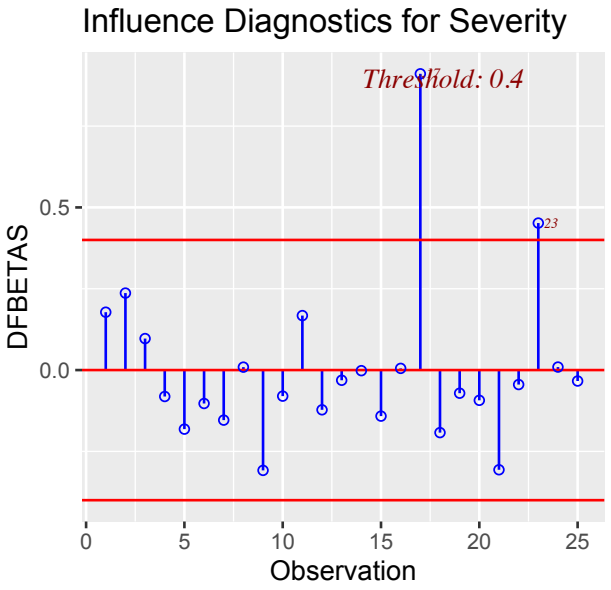
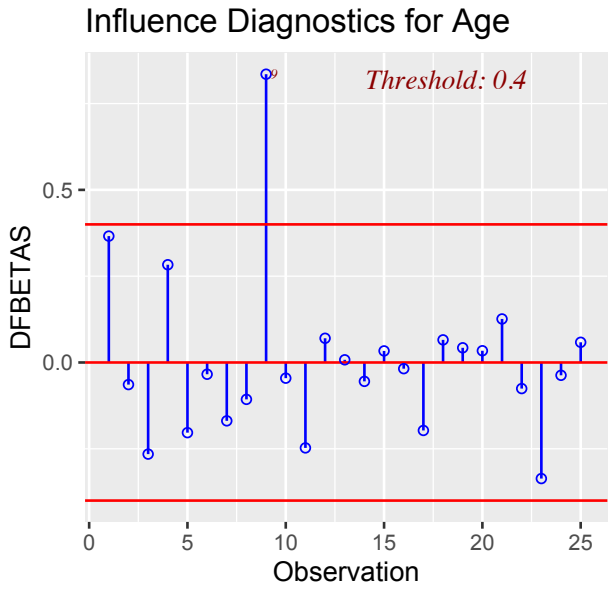


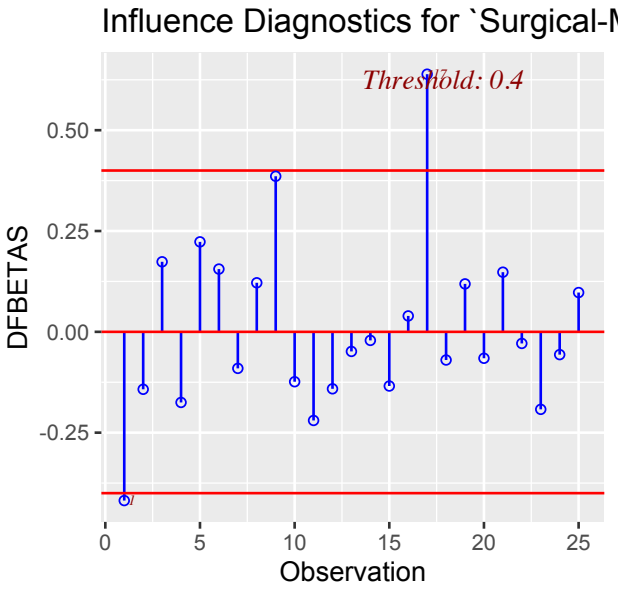
Figure 27: Cook's Distance Bar Chart for the All In Model

DfBETA is a more specific measure of influence. It measures the difference in each parameter estimate with and without the influential point. There is a DFBETA for each data point (i.e., if there are n observations and k variables, there will be n∗kn∗k DFBETAs). In general, large values of DFBETAS indicate observations that are influential in estimating a given parameter. Belsley, Kuh, and Welsch recommend 2 as a general cutoff value to indicate influential observations and $\frac{2}{\sqrt{n}}$ as a size-adjusted cutoff. Thus, we must sort the DfBeta observations and check that each observation passes the threshold of 0.4. From Figure 28, we see observations 9 and 17 as outliers.


Influence Diagnostics for (Intercept)
Threshold: 0.4
DFBETAS
0.25
0.00
-0.25
-0.50
9
17
0
5
10
15
20
25
Observation


Influence Diagnostics for Severity
Threshold: 0.4
DFBETAS
0.5
0.0
23
0
5
10
15
20
25
Observation


Influence Diagnostics for Age
Threshold: 0.4
DFBETAS
0.5
0.0
9
0
5
10
15
20
25
Observation


Influence Diagnostics for `Surgical-M
Threshold: 0.4
DFBETAS
0.50
0.25
0.00
-0.25
17
1
0
5
10
15
20
25
Observation

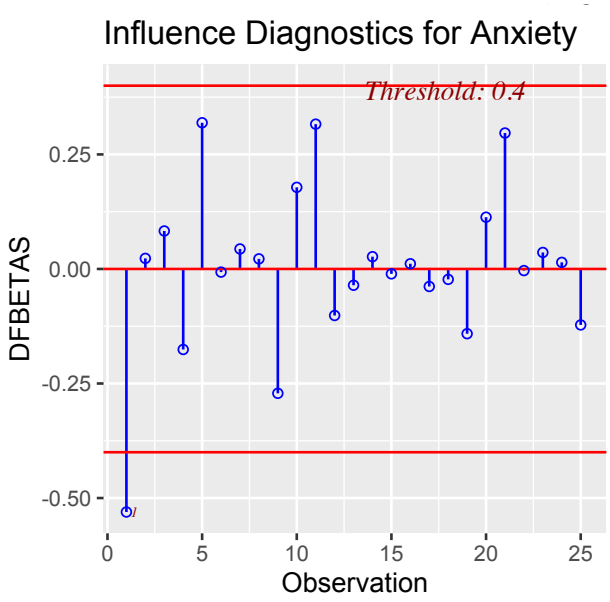


Figure 28: DfBeta Diagnostics for the All In Model

Our final measure of influence is the Variance Indicator Factor (VIF), which is calculated as:

$$VIF = \frac{1}{1 - R^2}$$

VIF can be used to detect collinearity, which causes instability in parameter estimation in regression-type models. The VIF is based on the square of the multiple correlation coefficient resulting from regressing a predictor variable against all other predictor variables. If a variable has a strong linear relationship with at least one other variable, the correlation coefficient would be close to 1, and VIF for that variable would be large. A VIF greater than 10 is a signal that the model has a collinearity problem. We will use a general VIF cut off VIF > 5. From the R-output copied below, no variables have a VIF > 5, indicating that multicollinearity is not a problem for this model.

```
> vif(AllInModel)
      Age          Severity      `Surgical-Medical`      Anxiety
 1.939128          1.441055          1.072782          1.689768
```

```
> #Obtain regression diagnostics for each observation: y_hat, coefficients, sigma, weighted
residual
> influence(AllInModel) #regression diagnostics
$hat
        1         2         3         4         5         6         7         8         9
0.22714184 0.22645722 0.23008849 0.23070193 0.33211257 0.16724449 0.17598122
0.12551879 0.20142321
       10        11        12        13        14        15        16        17        18
0.13178567 0.16089235 0.15686095 0.15001674 0.22283152 0.13853397 0.07879001
0.11805092 0.22209843
       19        20        21        22        23        24        25
0.28162471 0.43929726 0.28426093 0.21328924 0.23980549 0.10723961 0.13795244

$coefficients
   (Intercept)         Age      Severity `Surgical-Medical`      Anxiety
1  -0.37019698  0.068330281  0.0325387321       -1.69801638 -0.775891888
2  -1.49760553 -0.012400805  0.0449191070       -0.59989410  0.035052846
3   1.14095490 -0.051364678  0.0183661487        0.73081332  0.125712776
4  -0.90551122  0.054753502 -0.0153985336       -0.73606458 -0.266316133
5   0.85166278 -0.039334153 -0.0344237716        0.93869802  0.483796499
6   0.67967784 -0.006606940 -0.0195454458        0.65665743 -0.010432985
7   2.52055681 -0.032328436 -0.0288300692       -0.37659991  0.065402516
8   0.77197680 -0.020736802  0.0017123491        0.51321255  0.033518776
9  -5.23130571  0.144286161 -0.0522132820        1.44919993 -0.367521684
10  0.06524995 -0.008756506 -0.0151394819       -0.51857960  0.269275892
11 -0.89307173 -0.046616775  0.0309152822       -0.90019921  0.466621972
12  1.60000799  0.013547805 -0.0231212559       -0.59318428 -0.153716198
13  0.63463698  0.001509556 -0.0059922881       -0.20596233 -0.054387529
14  0.51581349 -0.010681152 -0.0003544754       -0.09010122  0.040990109
15  1.56559490  0.006475130 -0.0268005915       -0.56429848 -0.016204728
16  0.05610627 -0.003499209  0.0009912929        0.16735061  0.017410198
17 -4.01467501 -0.027868634  0.1263087224        1.96596393 -0.042432487
18  1.12026045  0.012724683 -0.0366043266       -0.29493276 -0.034616446
19  0.58359324  0.008279495 -0.0135618189        0.50303909 -0.215543035
20  0.06372184  0.006634618 -0.0177838924       -0.27802732  0.172778695
21 -0.29285216  0.024203162 -0.0576452146        0.61729270  0.446296151
22  1.07803495 -0.014724589 -0.0084992492       -0.12179027 -0.005405363
23 -0.85667310 -0.063879573  0.0841278109       -0.79377990  0.053604358
24  0.45574278 -0.007240035  0.0017652586       -0.23946857  0.021971064
25  0.43693262  0.011353401 -0.0064012694        0.41120442 -0.185815527

$sigma
       1        2        3        4        5        6        7        8        9       10
 9.724729 10.091599 10.083187 10.074905 10.080587 10.109151  9.962748 10.112890
9.001918 10.045879
      11       12       13       14       15       16       17       18       19       20
```

```
 9.815742 10.063158 10.163007 10.173075 10.065545 10.167762  7.373978 10.120502
10.139354 10.168166
      21       22       23       24       25
10.000295 10.149992  9.897835 10.157606 10.111206

$wt.res
        1         2         3         4         5         6         7         8         9
 11.519393  -5.095519   5.322134  -5.544620  -5.023622  -4.726724  -8.253185   4.712223 -
18.506508
       10        11        12        13        14        15        16        17        18
 -6.658066 -10.757209   6.122350   2.274264   1.306083   6.124662   1.978296  28.722334  -
4.180474
       19        20        21        22        23        24        25
 -3.305412   1.514739   6.997937  -2.956166  -9.025522   2.700776   4.737837

> #Diagnostic plots
> layout(matrix(c(1,2,3,4),2,2)) #optional 4 graphs/page

> plot(AllInModel)

> #Detect Influence with Leverage
> #The observed value of y_i is influential if h_i > [2(k+1)]/n
> #Where h_i = leverage for the ith obervation
> #k = # of betas in the model (excluding b_0)
> # [2(k+1)]/n = [2(5+1)/25] = 12/25 = 0.48
> ols_leverage(AllInModel)
 [1] 0.22714184 0.22645722 0.23008849 0.23070193 0.33211257 0.16724449 0.17598122
0.12551879
 [9] 0.20142321 0.13178567 0.16089235 0.15686095 0.15001674 0.22283152 0.13853397
0.07879001
[17] 0.11805092 0.22209843 0.28162471 0.43929726 0.28426093 0.21328924 0.23980549
0.10723961
[25] 0.13795244
> #Check for Collinearity
> ols_coll_diag(AllInModel)
> rmatrix <- rcorr(as.matrix(data)) #can be pearson or spearman
> rmatrix
                Satisfaction   Age Severity Surgical.Medical Anxiety
Satisfaction            1.00 -0.87    -0.65           -0.18   -0.51
Age                    -0.87  1.00     0.53            0.25    0.62
Severity               -0.65  0.53     1.00            0.18    0.45
Surgical.Medical       -0.18  0.25     0.18            1.00    0.11
Anxiety                -0.51  0.62     0.45            0.11    1.00

n= 25
```

```
P
                Satisfaction Age    Severity Surgical.Medical Anxiety
Satisfaction                0.0000 0.0004   0.3832          0.0088
Age             0.0000              0.0065   0.2365          0.0009
Severity        0.0004      0.0065          0.3959          0.0250
Surgical.Medical 0.3832      0.2365 0.3959                   0.6018
Anxiety         0.0088      0.0009 0.0250   0.6018
> vcov(AllInModel) # covariance matrix for model parameters
                (Intercept)        Age    Severity `Surgical-Medical`     Anxiety
(Intercept)        69.2077733 -0.63906056 -0.75928064       -1.15763609  0.69521211
Age               -0.6390606  0.03626236 -0.01200416       -0.15037085 -0.14473026
Severity          -0.7592806 -0.01200416  0.03481682       -0.05336886 -0.05069426
`Surgical-Medical`  -1.1576361 -0.15037085 -0.05336886       17.14129126  0.42066763
Anxiety            0.6952121 -0.14473026 -0.05069426        0.42066763  2.22664821
> #Cook's Distance: Combines leverage and residuals
> #Higher value, the better
> #Lowest Value = 0
> #Conventional Cut off is 4/n
> ols_plot_cooksd_bar(AllInModel)
> ols_plot_cooksd_chart(AllInModel)
> cooks.distance(AllInModel)
         1          2          3          4          5          6          7
0.1025354692 0.0199669511 0.0223409865 0.0243513430 0.0381794931 0.0109485009
0.0358717623
         8          9         10         11         12         13         14
0.0074058822 0.2198074931 0.0157480910 0.0537299748 0.0168062211 0.0021823176
0.0012788070
        15         16         17         18         19         20         21
0.0142285827 0.0007383353 0.2544129532 0.0130336058 0.0121153642 0.0065145740
0.0552160322
        22         23         24         25
0.0061194701 0.0686865463 0.0019942522 0.0084672912
> #dfbetas:measures the difference in each parameter estimate with and without the influential
point
> ols_plot_dfbetas(AllInModel)
> dfbeta(AllInModel)
  (Intercept)         Age      Severity `Surgical-Medical`      Anxiety
1  -0.37019698  0.068330281  0.0325387321       -1.69801638 -0.775891888
2  -1.49760553 -0.012400805  0.0449191070       -0.59989410  0.035052846
3   1.14095490 -0.051364678  0.0183661487        0.73081332  0.125712776
4  -0.90551122  0.054753502 -0.0153985336       -0.73606458 -0.266316133
5   0.85166278 -0.039334153 -0.0344237716        0.93869802  0.483796499
6   0.67967784 -0.006606940 -0.0195454458        0.65665743 -0.010432985
7   2.52055681 -0.032328436 -0.0288300692       -0.37659991  0.065402516
8   0.77197680 -0.020736802  0.0017123491        0.51321255  0.033518776
```

9	-5.23130571	0.144286161	-0.0522132820	1.44919993	-0.367521684
10	0.06524995	-0.008756506	-0.0151394819	-0.51857960	0.269275892
11	-0.89307173	-0.046616775	0.0309152822	-0.90019921	0.466621972
12	1.60000799	0.013547805	-0.0231212559	-0.59318428	-0.153716198
13	0.63463698	0.001509556	-0.0059922881	-0.20596233	-0.054387529
14	0.51581349	-0.010681152	-0.0003544754	-0.09010122	0.040990109
15	1.56559490	0.006475130	-0.0268005915	-0.56429848	-0.016204728
16	0.05610627	-0.003499209	0.0009912929	0.16735061	0.017410198
17	-4.01467501	-0.027868634	0.1263087224	1.96596393	-0.042432487
18	1.12026045	0.012724683	-0.0366043266	-0.29493276	-0.034616446
19	0.58359324	0.008279495	-0.0135618189	0.50303909	-0.215543035
20	0.06372184	0.006634618	-0.0177838924	-0.27802732	0.172778695
21	-0.29285216	0.024203162	-0.0576452146	0.61729270	0.446296151
22	1.07803495	-0.014724589	-0.0084992492	-0.12179027	-0.005405363
23	-0.85667310	-0.063879573	0.0841278109	-0.79377990	0.053604358
24	0.45574278	-0.007240035	0.0017652586	-0.23946857	0.021971064
25	0.43693262	0.011353401	-0.0064012694	0.41120442	-0.185815527

All In Model: Test Lack of Fit

Finally, a Lack of Fit test was conducted and testing using the following hypothesis testing

H_0: *The relationship assumed in the model is reasonable; i.e., there is no lack of fit in the model*

H_a: *The relationship assumed in the model is not reasonable; i.e., there is a lack of fit in the model*

As before, rejection criteria was:

- $F_{obtained} > F_{critical}$
- $\alpha > p - value$

From the R-output listed below, we fail to reject the null hypothesis and conclude that the model is reasonable (i.e., there is no lack of fit).

```
> #lack of fit
> ols_test_f(AllInModel)

 F Test for Heteroskedasticity
 -----------------------------
```

Ho: Variance is homogenous
Ha: Variance is not homogenous

Variables: fitted values of Satisfaction

Test Summary

Num DF = 1
Den DF = 23
F = 0.09582794
Prob > F = 0.7596821

Apply Transformations & Re-Specify Model

Although the All In Model maintained all regression assumptions, had a relatively high adjusted R^2, and had no lack of fit, a transformation was applied to the Severity variable in order to determine whether that would improve the model. R_a^2 was used as the primary metric for initial model comparisons.

First-Order Main Effect Model with Severity:

- $y = \beta_0 + \beta_1 Age + \beta_2 Severity + \beta_3 SurgicalMedical + \beta_4 Anxiety$
- Overall model: $F_{obtained}$ = 22.57, p = 3.611e-07 (significant)
- Main Effect of Age: $t_{obtained}$ = -6.002, p = 0.000 (significant)
- Main Effect of Severity: $t_{obtained}$ = -2.518, p = 0.02 (significant)
- No significant main effect of Surgical-Medical or Anxiety
- $R_a^2 = 0.7819$

Polynomial Main Effect Model with $Severity^3$:

- $y = \beta_0 + \beta_1 Age + \beta_2 Severity^3 + \beta_3 SurgicalMedical + \beta_4 Anxiety$
- Overall model: $F_{obtained}$ = 22.86, p = 3.195e-07 (significant)
- Main Effect of Age: $t_{obtained}$ = -5.751, p = 1.26e-05 (significant)

- Main Effect of Severity^3: $t_{obtained}$ = -2.582, p = 0.0178 (significant)
- No significant main effect of Surgical-Medical or Anxiety
- $R_a^2 = 0.7846$ (improvement)

The increased R_a^2 , significant global F-test and significant main effects for Age and Severity^3 suggest that Severity^3 should be included in the model.

```
> data$SeverityCubed <-data$Severity*data$Severity*data$Severity
> cor.test(Satisfaction, data$SeverityCubed)

        Pearson's product-moment correlation

data:  Satisfaction and data$SeverityCubed
t = -4.5383, df = 23, p-value = 0.0001471
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 -0.8512704 -0.4011489
sample estimates:
      cor
-0.6873331

> #-0.6873331  (improvement)
> corr.test(Age, data$SeverityCubed)
Call:corr.test(x = Age, y = data$SeverityCubed)
Correlation matrix
[1] 0.57
Sample Size
[1] 25
Probability values  adjusted for multiple tests.
[1] 0

 To see confidence intervals of the correlations, print with the short=FALSE option
> #-0.6873331  (improvement)
> corr.test(Age, data$SeverityCubed)
Call:corr.test(x = Age, y = data$SeverityCubed)
Correlation matrix
[1] 0.57
Sample Size
[1] 25
Probability values  adjusted for multiple tests.
[1] 0

 To see confidence intervals of the correlations, print with the short=FALSE option
```

```
> #Age & Severity r = 0.529

> #Create model
> modelSeverityCubed <-lm(Satisfaction~Age+SeverityCubed+`Surgical-Medical`+Anxiety,
data = data)
> #Obtain Result
> summary(modelSeverityCubed)

Call:
lm(formula = Satisfaction ~ Age + SeverityCubed + `Surgical-Medical` +
    Anxiety, data = data)

Residuals:
      Min       1Q   Median       3Q      Max
-19.0019  -5.9868   0.4121   3.8628  28.6010

Coefficients:
                  Estimate   Std. Error  t value  Pr(>|t|)
(Intercept)     1.252e+02  7.443e+00   16.814   2.88e-13 ***
Age            -1.111e+00  1.932e-01   -5.751   1.26e-05 ***
SeverityCubed  -6.693e-05  2.591e-05   -2.583   0.0178 *
`Surgical-Medical`  2.237e+00  4.114e+00   0.544   0.5927
Anxiety         1.230e+00  1.479e+00    0.831   0.4157
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 9.859 on 20 degrees of freedom
Multiple R-squared:  0.8205,  Adjusted R-squared:  0.7846
F-statistic: 22.86 on 4 and 20 DF,  p-value: 3.195e-07
```

Further Model Comparisons

Models were further compared by running an anova on the All In Model and the model with Severity^3. A Lack of Fit test on the transformed model with Severity^3 was also conduced.

Results indicate a positive contribution of Severity^3 within the model. R-Code copied below.

```
> #Lack of Fit
> anova(modelSeverityCubed, AllInModel)
Analysis of Variance Table

Model 1: Satisfaction ~ Age + SeverityCubed + `Surgical-Medical` + Anxiety
Model 2: Satisfaction ~ Age + Severity + `Surgical-Medical` + Anxiety
  Res.Df  RSS Df  Sum of Sq  F  Pr(>F)
1     20  1944.1
```

```
        2                20              1968.5          0        -24.424

> ols_test_f(modelSeverityCubed)

 F Test for Heteroskedasticity
 -----------------------------
 Ho: Variance is homogenous
 Ha: Variance is not homogenous

 Variables: fitted values of Satisfaction

      Test Summary
 --------------------------
 Num DF     =    1
 Den DF     =    23
 F          =    0.01827019
 Prob > F   =    0.8936556
```

Apply Transformations & Re-Specify Model

Although the All In Model maintained all regression assumptions, had a relatively high adjusted R^2, and had no lack of fit, a transformation was applied to the Anxiety variable in order to determine whether that would improve the model. R_a^2 was used as the primary metric for initial model comparisons.

First-Order Main Effect Model with Anxiety:

- $y = \beta_0 + \beta_1 Age + \beta_2 Severity + \beta_3 SurgicalMedical + \beta_4 Anxiety$
- Overall model: $F_{obtained}$ = 22.57, p = 3.611e-07 (significant)
- Main Effect of Age: $t_{obtained}$ = -6.002, p = 0.000 (significant)
- Main Effect of Severity: $t_{obtained}$ = -2.518, p = 0.02 (significant)
- No significant main effect of Surgical-Medical or Anxiety
- $R_a^2 = 0.7819$

Polynomial Main Effect Model with Anxiety^3:

- $y = \beta_0 + \beta_1 Age + \beta_2 Severity + \beta_3 SurgicalMedical + \beta_4 Anxiety^3$

- Overall model: $F_{obtained} = 22.57$, p = 3.53e-07 (significant)
- Main Effect of Age: $t_{obtained} = -5.718$, p = 1.35e-05 (significant)
- Main Effect of Severity: $t_{obtained} = -2.410$, p = 0.0257 (significant)
- No significant main effect of Surgical-Medical or Anxiety[3]
- $R_a^2 = 0.7824$ (decrease)

Although the R_a^2 did not increase by replacing Anxiety with Anxiety[3], the global F-test remaining significant and the model still had significant main effects for Age and Severity.

```
> #Try Anxiety Cubed
> corr.test(Satisfaction, Anxiety)
Call:corr.test(x = Satisfaction, y = Anxiety)
Correlation matrix
[1] -0.51
Sample Size
[1] 25
Probability values  adjusted for multiple tests.
[1] 0.01

 To see confidence intervals of the correlations, print with the short=FALSE option
> #[1] -0.51

> cor.test(Satisfaction, data$AnxietyCubed)
> #-0.4908339 (not an improvement)
> cor.test(Age, data$AnxietyCubed)
> #Anxiety & Age r = 0.621
> data$AnxietyCubed <-data$Anxiety*data$Anxiety*data$Anxiety

> modelAnxietyCubed <-lm(Satisfaction~Age+Severity+`Surgical-Medical`+AnxietyCubed,
data=data)

> #Obtain Result
> summary(modelAnxietyCubed)

Call:
lm(formula = Satisfaction ~ Age + Severity + `Surgical-Medical` +
    AnxietyCubed, data = data)

Residuals:
      Min       1Q   Median       3Q      Max
-18.3794  -  5.4093   0.5531   4.5162  29.0722
```

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	142.90457	9.03144	15.823	8.95e-13 ***
Age	-1.16760	0.20420	-5.718	1.35e-05 ***
Severity	-0.44161	0.18326	-2.410	0.0257 *
`Surgical-Medical`	2.89325	4.25325	0.680	0.5041
AnxietyCubed	0.01838	0.02094	0.877	0.3907

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 9.909 on 20 degrees of freedom
Multiple R-squared: 0.8187, Adjusted R-squared: 0.7824
F-statistic: 22.57 on 4 and 20 DF, p-value: 3.53e-07

```
> #Ra^2=0.7842
> #Compare to AllInModel: Ra^2 = 0.7819
> #Improvement
```

Further Model Comparisons

Models were further compared by running an anova on the All In Model and the model with Anxiety^3. A Lack of Fit test on the transformed model with Anxiety^3 was also conduced. It is still unclear whether Anxiety^3 should be included in the model. One further analysis will be conducted before attempting All Possible Regressors. R-Code copied below.

```
> #Lack of Fit
> anova(modelAnxietyCubed, AllInModel)
Analysis of Variance Table

Model 1: Satisfaction ~ Age + Severity + `Surgical-Medical` + AnxietyCubed
Model 2: Satisfaction ~ Age + Severity + `Surgical-Medical` + Anxiety
  Res.Df    RSS Df Sum of Sq F Pr(>F)
1     20 1963.9
2     20 1968.5  0   -4.5999
> ols_test_f(modelAnxietyCubed)

 F Test for Heteroskedasticity
 -----------------------------
 Ho: Variance is homogenous
```

Ha: Variance is not homogenous

Variables: fitted values of Satisfaction

Test Summary

Num DF = 1
Den DF = 23
F = 0.09964379
Prob > F = 0.7551041

Apply Transformations & Re-Specify Model

This final polynomial comparison sought to compare whether inclusion of both Severity^3 and Anxiety^3 improved the model over that established in the All In Model. R_a^2 was used as the primary metric for initial model comparisons.

First-Order Main Effect Model with Anxiety:

- $y = \beta_0 + \beta_1 Age + \beta_2 Severity + \beta_3 SurgicalMedical + \beta_4 Anxiety$
- Overall model: $F_{obtained}$ = 22.57, p = 3.611e-07 (significant)
- Main Effect of Age: $t_{obtained}$ = -6.002, p = 0.000 (significant)
- Main Effect of Severity: $t_{obtained}$ = -2.518, p = 0.02 (significant)
- No significant main effect of Surgical-Medical or Anxiety
- $R_a^2 = 0.7819$

Polynomial Main Effect Model with Severity^3 and Anxiety^3:

- $y = \beta_0 + \beta_1 Age + \beta_2 Severity^3 + \beta_3 SurgicalMedical + \beta_4 Anxiety^3$
- Overall model: $F_{obtained}$ = 22.79, p = 3.275e-07 (significant)
- Main Effect of Age: $t_{obtained}$ = -5.405 p = 2,73e-05 (significant)
- Main Effect of Severity^3: $t_{obtained}$ = -2.450, p = 0.0236 (significant)
- No significant main effect of Surgical-Medical or Anxiety^3

- $R_a^2 = 0.7841$ (very minor decrease)

Although the R_a^2 did not increase by replacing Anxiety with Anxiety^3 and Severity with Severity^3, the global F-test remaining significant and the model still had significant main effects for Age and Severity^3.

```
> #Try Anxiety Cubed AND Severity Cubed
> modelAnxietyAndSeverityCubed <-lm(Satisfaction~Age+SeverityCubed+`Surgical-
Medical`+AnxietyCubed,data=data)

> #Obtain Result
> summary(modelAnxietyAndSeverityCubed)

Call:
lm(formula = Satisfaction ~ Age + SeverityCubed + `Surgical-Medical` +
    AnxietyCubed, data = data)

Residuals:
      Min       1Q   Median       3Q      Max
-18.8154  -6.3216  -0.0471   3.7581  28.8957

Coefficients:
                 Estimate   Std. Error  t value  Pr(>|t|)
(Intercept)     1.285e+02  8.253e+00    15.571  1.21e-12 ***
Age            -1.132e+00  2.093e-01    -5.405  2.73e-05 ***
SeverityCubed  -6.269e-05  2.558e-05    -2.450    0.0236 *
`Surgical-Medical`  2.824e+00  4.235e+00  0.667    0.5126
AnxietyCubed    1.668e-02  2.087e-02    0.799    0.4336
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 9.872 on 20 degrees of freedom
Multiple R-squared:  0.8201,  Adjusted R-squared:  0.7841
F-statistic: 22.79 on 4 and 20 DF,  p-value: 3.275e-07

> #Ra^2 = 0.7841
> #Compare to AllInModel: Ra^2 = 0.7819
> #Improvment
> #AnxietyCubed Nonsignificant
```

Further Model Comparisons

Models were further compared by running an anova on the All In Model and the model with Anxiety^3 and Severity^3. A Lack of Fit test on the transformed model with Anxiety^3 and Severity^3 was also conduced. It is still unclear whether both Anxiety^3 and Severity^3 should be included in the model. Thus, the next step will involve All Regression analysis. R-Code copied below.

```
> #Lack of Fit
> anova(modelAnxietyAndSeverityCubed, AllInModel)
Analysis of Variance Table

Model 1: Satisfaction ~ Age + SeverityCubed + `Surgical-Medical` + AnxietyCubed
Model 2: Satisfaction ~ Age + Severity + `Surgical-Medical` + Anxiety
  Res.Df RSS Df Sum of Sq F Pr(>F)
1     20          1949.0
2     20          1968.5 0 -19.513
> ols_test_f(modelAnxietyCubed)

 F Test for Heteroskedasticity
 -----------------------------
 Ho: Variance is homogenous
 Ha: Variance is not homogenous

 Variables: fitted values of Satisfaction

      Test Summary
 --------------------------
 Num DF     =    1
 Den DF     =    23
 F          =    0.09964379
 Prob > F   =    0.7551041
```

Final Model Specification

Use All Possible Regressions Method to Find "Best Subset"

All Possible Regressions method was run to find the "Best Subset". All Possible Regressions was performed using criteria such as Mallow's C, adjusted R^2, and the PRESS

statistic to rank the best subset models. The best models recommended by each criterion were compared and analyzed prior to model validation. The following models had the highest adjusted R^2 and best Mallow's Cp:

- Satisfaction ~ Age + Severity^3 + Surgical-Medical + Anxiety^3
  - $R_a^2 = 0.820$
  - Mallow's Cp = 5
- Satisfaction ~ Age + Severity^3
  - $R_a^2 = 0.795$
  - Mallow's Cp = 1.88
- Satisfaction ~ Age + Severity^3 + Anxiety^3
  - $R_a^2 = 0.790$
  - Mallow's Cp = 3.44

```
> #Test Fit of modelAnxietyAndSeverityCubed
> ols_step_all_possible(modelAnxietyAndSeverityCubed)
# A tibble: 15 x 6
   Index N Predictors                                  `R-Square` `Adj. R-Square` `Mallow's Cp`
   <int> <int> <chr>                                       <dbl>          <dbl>          <dbl>
 1     1     1 Age                                         0.758          0.748           5.88
 2     2     1 SeverityCubed                               0.472          0.449          37.6
 3     3     1 AnxietyCubed                                0.241          0.208          63.4
 4     4     1 `Surgical-Medical`                          0.0332        -0.00881        86.5
 5     5     2 Age SeverityCubed                           0.812          0.795           1.88
 6     6     2 Age AnxietyCubed                            0.763          0.742           7.30
 7     7     2 Age `Surgical-Medical`                      0.759          0.737           7.76
 8     8     2 SeverityCubed AnxietyCubed                  0.550          0.509          31.0
 9     9     2 SeverityCubed `Surgical-Medical`            0.475          0.428          39.3
10    10     2 `Surgical-Medical` AnxietyCubed             0.280          0.214          61.1
11    11     3 Age SeverityCubed AnxietyCubed              0.816          0.790           3.44
12    12     3 Age SeverityCubed `Surgical-Medical`        0.814          0.788           3.64
13    13     3 Age `Surgical-Medical` AnxietyCubed         0.766          0.733           9.00
14    14     3 SeverityCubed `Surgical-Medical` AnxietyCubed  0.557          0.494          32.2
15    15     4 Age SeverityCubed `Surgical-Medical` Anxiety…  0.820          0.784           5
```

The leaps() function was also utilized for variable selection because it performs an exhaustive search for the best subset of the variables in X for predicting Y in a linear regression. Results in Figure 29 illustrate the optimal number of variables yielding:

- The highest R^2 : 5 variables
- The highest R_a^2: 3 variables

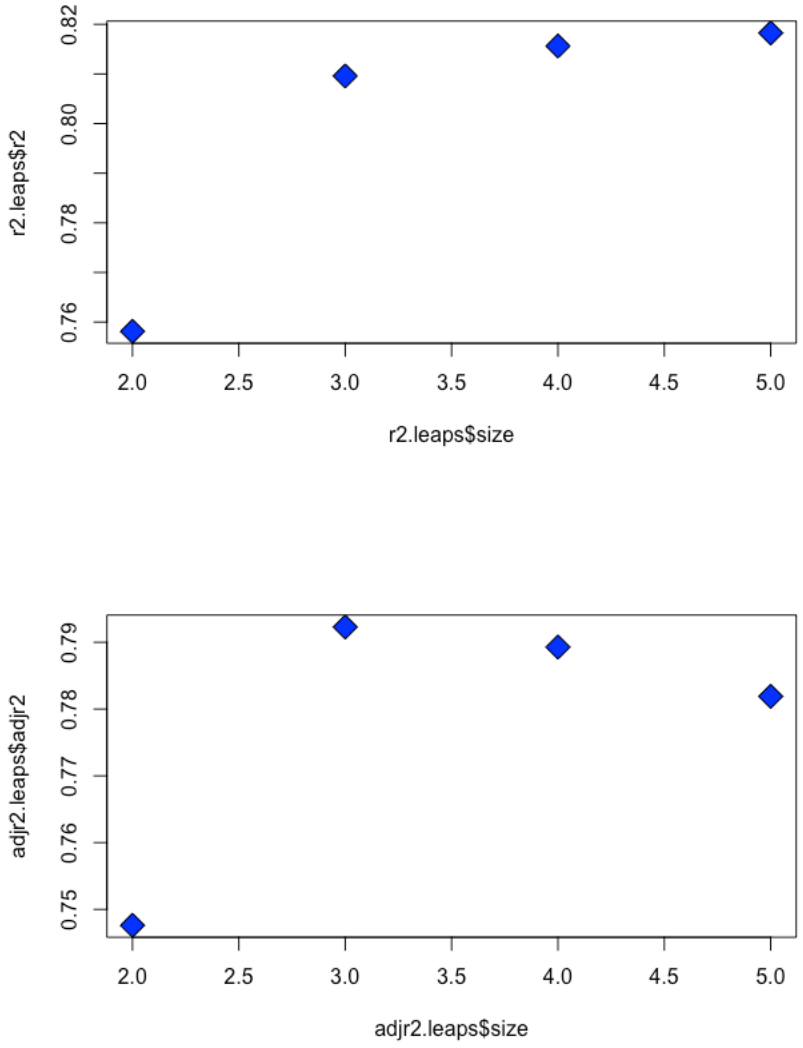


Figure 29: Leap() identifying optimal number of variables to increase R^2 and adjusted R^2

```
> #Perform All Possible Regressions (for Variable Selection)
> fitAllWithCubes = regsubsets(Satisfaction~.,data=data, nbest =1, nvmax = NULL, force.in =
NULL, force.out = NULL, method = 'exhaustive')

> summary(fitAllWithCubes)
Subset selection object
Call: regsubsets.formula(Satisfaction ~ ., data = data, nbest = 1,
    nvmax = NULL, force.in = NULL, force.out = NULL, method = "exhaustive")

                  6 Variables  (and intercept)
                  Forced in    Forced out
Age               FALSE        FALSE
Severity          FALSE        FALSE
Surgical.Medical  FALSE        FALSE
Anxiety           FALSE        FALSE
SeverityCubed     FALSE        FALSE
AnxietyCubed      FALSE        FALSE
1 subsets of each size up to 6

Selection Algorithm: exhaustive
         Age Severity Surgical.Medical Anxiety SeverityCubed AnxietyCubed
1  ( 1 ) "*" " "      " "              " "      " "           " "
2  ( 1 ) "*" " "      " "              " "      "*"           " "
3  ( 1 ) "*" " "      " "              "*"      "*"           " "
4  ( 1 ) "*" " "      "*"              "*"      "*"           " "
5  ( 1 ) "*" "*"      "*"              "*"      "*"           " "
6  ( 1 ) "*" "*"      "*"              "*"      "*"           "*"

> fitAllWithCubesOutput = summary(fitAllWithCubes)
> names(fitAllWithCubesOutput)
[1] "which"  "rsq"    "rss"    "adjr2"  "cp"     "bic"    "outmat" "obj"

> as.data.frame(fitAllWithCubesOutput$outmat)
         Age Severity Surgical.Medical Anxiety SeverityCubed AnxietyCubed
1  ( 1 )  *
2  ( 1 )  *                                    *
3  ( 1 )  *                             *      *
4  ( 1 )  *                          *  *      *
5  ( 1 )  *        *                 *  *      *
6  ( 1 )  *        *                 *  *      *             *

> View(fitAllWithCubesOutput)
> #Which model has the highest R^2?
> which.max(fitAllWithCubesOutput$rsq)
[1] 6
```

```
> #Model 6 = Age, Severity, SurgicalMedical, Anxiety, SeverityCubed, AnxietyCubed
> #See which model
> #Variables marked with TRUE are the chosen ones
> fitAllWithCubesOutput$which[6,]
      (Intercept)             Age        Severity  Surgical.Medical         Anxiety
             TRUE            TRUE            TRUE            TRUE            TRUE

   SeverityCubed     AnxietyCubed
             TRUE            TRUE

> #Plot Best R^2
> par(mfrow = c(2,2))
> plot(fitAllWithCubesOutput$rsq, xlab = "Number of Variables", ylab = "R^2", type = "b")
> best.r2=which.max(fitAllWithCubesOutput$rsp)
> points(best.r2,fitAllWithCubesOutput$rsq[best.r2], col = "red", cex = 2, pch =20)

> #Which model has the highest Ra^2?
> which.max(fitAllWithCubesOutput$adjr2)
[1] 2
> #The model with 2 variables (Age, SeverityCubed) has the highest Ra^2
> #See which model
> #Variables marked with TRUE are the chosen ones
> fitAllWithCubesOutput$which[6,]
      (Intercept)             Age        Severity  Surgical.Medical         Anxiety
             TRUE            TRUE            TRUE            TRUE            TRUE

   SeverityCubed     AnxietyCubed
             TRUE            TRUE

> #Plot Best Ra^2
> plot(fitAllWithCubesOutput$adjr2, xlab = "Number of Variables", ylab = "Adjusted R^2",
type = "b")
> best.adjr2=which.max(fitAllWithCubesOutput$adjr2)
> points(best.adjr2,fitAllWithCubesOutput$adjr2[best.adjr2], col = "red", cex = 2, pch =20)
```

Identify Best Subset & Re-Run Regression

After All Regression analysis, the Best Subset was established by selecting the subset with the highest R_a^2, which also satisfied the principle of parsimony. A polynomial Multiple Linear Regression model was fit with 2 quantitative independent variables to examining the effect of Age and Severity^3 on patient Satisfaction scores. The following is the general model form:

$$Satisfaction = \beta_0 + \beta_1 Age_i + \beta_2 Severity_i^3$$

where,

$Y: Patient\ Satisfaction$
$X_1: Age\ in\ years$
$X_2: Severity^3$
$\beta_0: y - intercept\ of\ (k+1)\ dimensonal\ surgace; the\ value\ of\ E(Y)\ when\ X_1 = X_2 = X_3 = X_4 = 0$
$\beta_1: Change\ in\ Satisfaction\ for\ a\ one\ unit\ increase\ in\ Age, when\ all\ other\ variables\ are\ held\ fixed$
$\beta_2: Change\ in\ Satisfaction^3\ for\ a\ one\ unit\ increase\ in\ Severity, when\ all\ other\ variables\ are\ held\ fixed$

Test For Significance of Regression

A global F-test was conducted to test for the overall significance of the revised regression. Results are significant for $F_{obtained} = 47.56\ and\ p = 1.027e-08$. Also see Figure 30.

```
#Do regression with the best model
> revisedPolynomialModel <- lm(Satisfaction~Age+SeverityCubed, data=data)
> summary(revisedPolynomialModel)

Call:
lm(formula = Satisfaction ~ Age + SeverityCubed, data = data)

Residuals:
     Min       1Q   Median       3Q      Max
-18.845   -6.366    1.367    5.349   29.268

Coefficients:
              Estimate   Std. Error  t value  Pr(>|t|)
(Intercept)  1.260e+02  7.210e+00   17.470   2.21e-14 ***
Age          -1.018e+00  1.613e-01   -6.308   2.39e-06 ***
SeverityCubed -6.256e-05  2.487e-05   -2.516   0.0197 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 9.616 on 22 degrees of freedom
Multiple R-squared:  0.8122,  Adjusted R-squared:  0.7951
F-statistic: 47.56 on 2 and 22 DF,  p-value: 1.027e-08
> #p = 1.027e-08 (significant); Ra^2 = 0.7951 (higher);both significant t-values
> #Satisfaction = 125.957 - (1.018*Age) - (6.256002e-05*Severity)
```

```
> #Confirm coefficients
> coefficients(revisedPolynomialModel) #model coefficients
  (Intercept)           Age                SeverityCubed
 1.259573e+02 -1.017675e+00 -6.256002e-05


> #Confirm with ols package
> ols_regress(revisedPolynomialModel)
                Model Summary
-----------------------------------------------------------------
R                       0.901       RMSE                 9.616
R-Squared               0.812       Coef. Var           14.413
Adj. R-Squared          0.795       MSE                 92.477
Pred R-Squared          0.761       MAE                  6.513
-----------------------------------------------------------------
 RMSE: Root Mean Square Error
 MSE: Mean Square Error
 MAE: Mean Absolute Error

                        ANOVA
-----------------------------------------------------------------------------------
                Sum of
                Squares         DF    Mean Square      F         Sig.
-----------------------------------------------------------------------------------
Regression      8796.548        2     4398.274         47.561    0.0000
Residual        2034.492        22    92.477
Total           10831.040       24
-----------------------------------------------------------------------------------

                              Parameter Estimates
-----------------------------------------------------------------------------------------------------------------
 model          Beta            Std. Error      Std. Beta       t         Sig       lower      upper
-----------------------------------------------------------------------------------------------------------------
 (Intercept)    125.957 7.210   17.470          0.000                               111.005    140.910
 Age            -1.018          0.161           -0.709          -6.308    0.000     -1.352     -0.683
SeverityCubed   0.000           0.000           -0.283          -2.516    0.020     0.000      0.000
-----------------------------------------------------------------------------------------------------------------
```

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	8796.548	2	4398.274	47.561	.000[b]
	Residual	2034.492	22	92.477		
	Total	10831.040	24			

a. Dependent Variable: Satisfaction

b. Predictors: (Constant), SeverityCubed, Age

Figure 30: Revised global F-test for significance of regression.

Thus, the final fitted model is:

$$Satisfaction = 125.957 + (1.018 * Age) + (6.256002e - 05 * Severity_i^3)$$

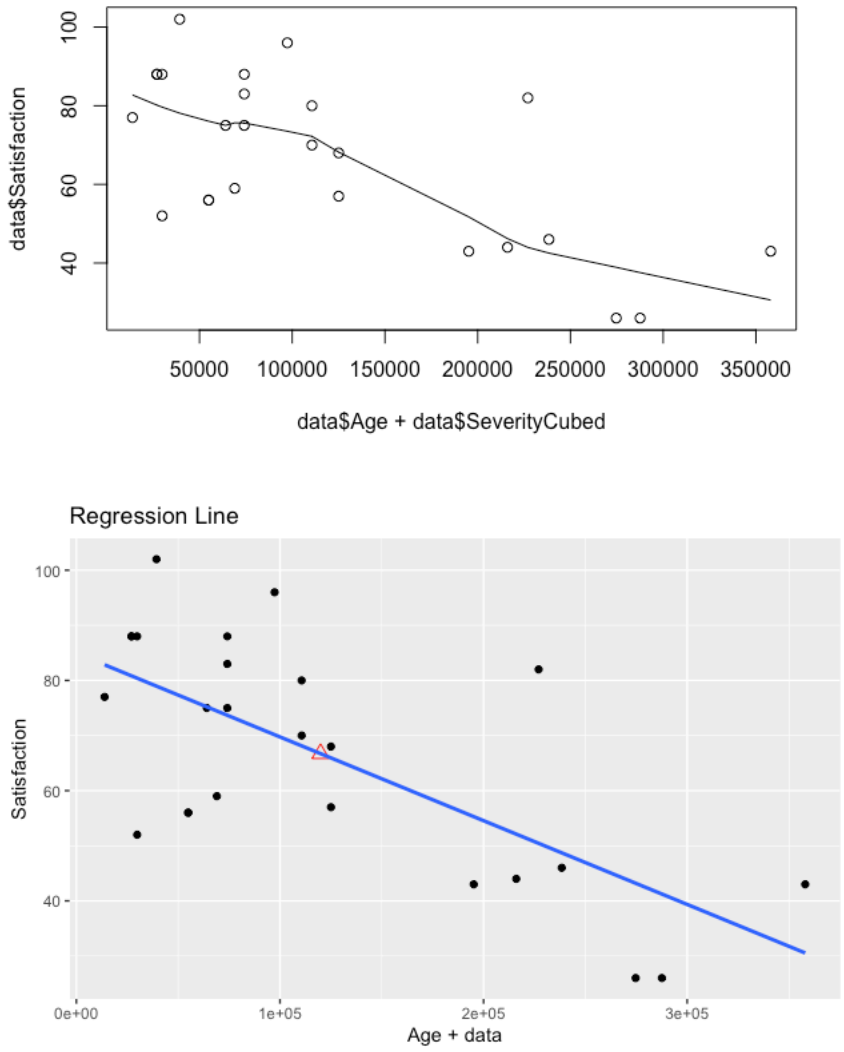


Figure 31: Final Linear Polynomial Model, Satisfaction ~ Age + Severity^3

```
> #Plot New Model
> plot(data$Age+data$SeverityCubed, data$Satisfaction, data=data)
> lines(lowess(data$Age+data$SeverityCubed, Satisfaction))
> #Appears linear
> #Can also plot with ols library
> ols_plot_reg_line(data$Satisfaction,data$Age+data$SeverityCubed)
```

Calculate R^2 And R^2_{Adj} For This Model

Adjusted R^2 for this model is 0.795, indicating that the model accounts for 79.5% of the observed variability.

Determine Contribution of Each Independent Variable

All variables included in the model had significant t-tests. Specifically, for Age: $t_{obtained} = -6.308, p = 2.39e-06$. And for Severity^3, $t_{obtained} = -2.516, p = 0.0197$. See Figure 31 below.

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients			95.0% Confidence Interval for B		Correlations			Collinearity Statistics	
		B	Std. Error	Beta	t	Sig.	Lower Bound	Upper Bound	Zero-order	Partial	Part	Tolerance	VIF
1	(Constant)	125.957	7.210		17.470	.000	111.005	140.910					
	Age	-1.018	.161	-.709	-6.308	.000	-1.352	-.683	-.871	-.802	-.583	.675	1.481
	SeverityCubed	-6.256E-5	.000	-.283	-2.516	.020	.000	.000	-.687	-.473	-.232	.675	1.481

a. Dependent Variable: Satisfaction

Figure 32: Revised t-tests to determine individual contributions of variables.

Calculate Confidence Intervals

Confidence intervals were calculated for each coefficient and are presented below.

```
> #Confidence Intervals
> confint(revisedPolynomialModel)
                      2.5 %          97.5 %
(Intercept)    1.110046e+02   1.409100e+02
Age           -1.352257e+00  -6.830922e-01
SeverityCubed -1.141337e-04  -1.098640e-05

> #Age: Small Range, does not include 0
> #Severity: Small range, does not include 0
> #Severity^3: Small range, does not include 0
```

Response Variable Diagnostics

Response variable diagnostics indicate a slight left/negative skew (a good thing) in patient satisfaction scores (Figure 33).

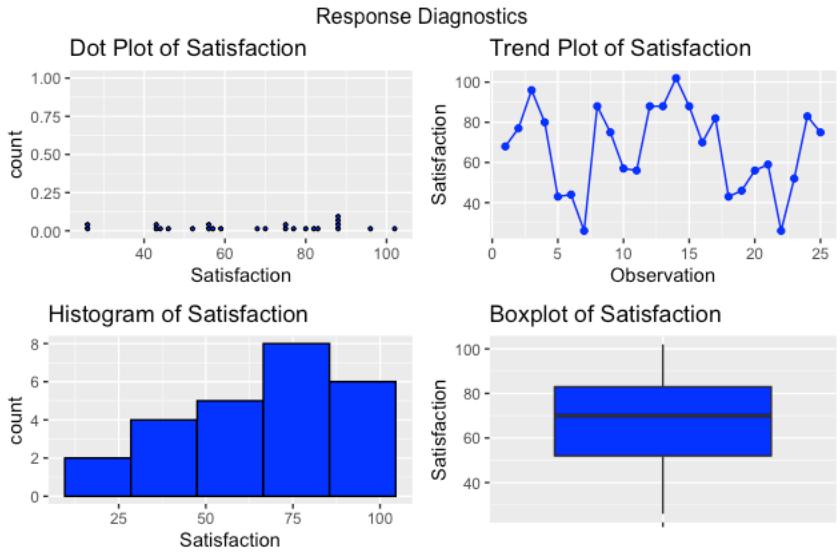


Figure 33: Response Variable Diagnostics

Residual Analysis

A 3d scatterplots was created to visualize the relationship of each variable's residuals (Figure 34).

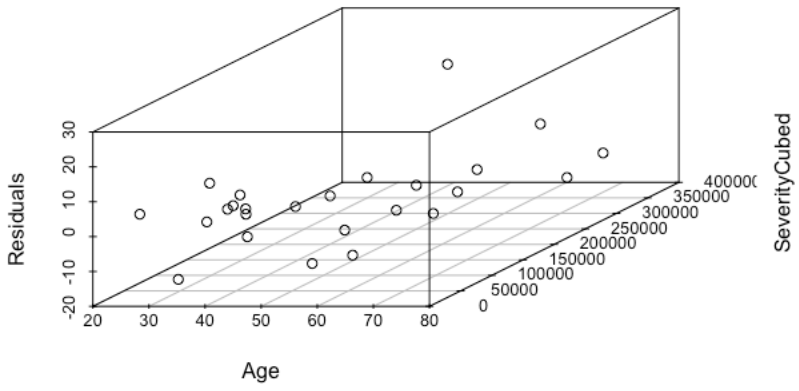
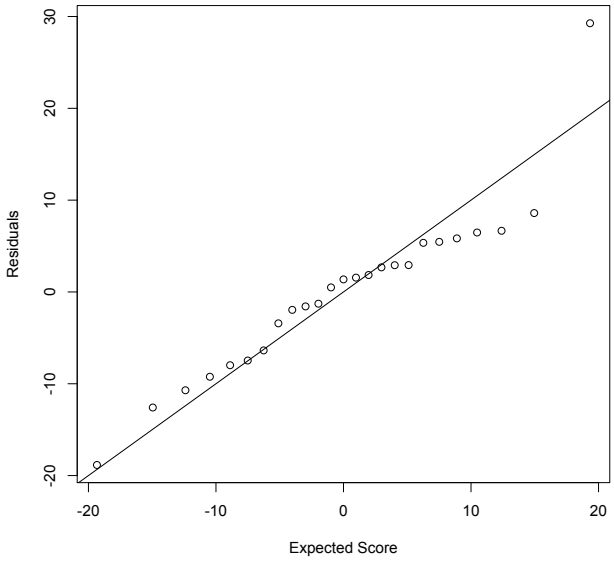


Figure 34: 3d Residual Scatterplot

A Normal Probability Plot (P-Plot) tests the normality of residuals. A normal P-Plot compares observed cumulative distribution function (CDF) of the standard residual to expected CDF of the normal distribution. If residuals are normally distributed, the P-Plot will plot as a straight line, which is usually determined visually, with an emphasis on the central values rather than extremes. Similarly, a Q-Q Plot compares observed quantile with theoretical quantile of a normal distribution. From Figures 35 we see a relatively normal distribution, with residuals plotted in an approximately straight line. There are no small or large deviations at either tail of the reference line, indicated the data is not skewed.



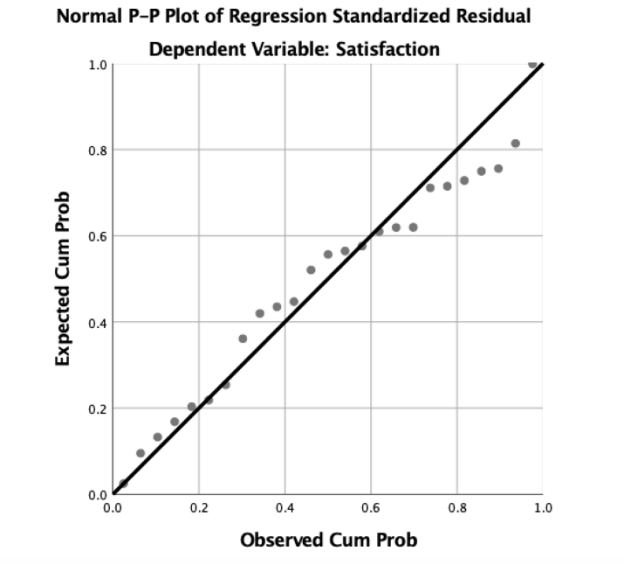
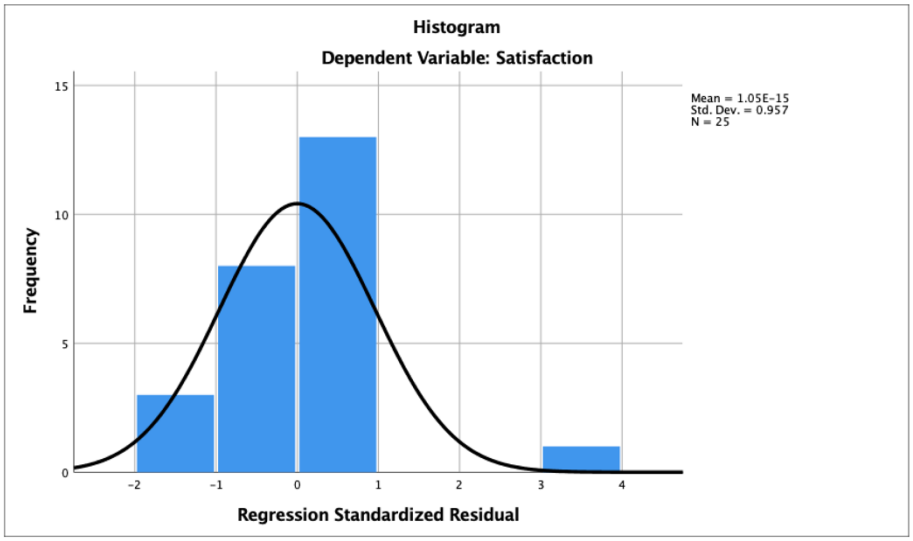


Figure 35: Revised Model Normal P-P Plot.

We can confirm a relatively normal distribution shape by examining the histogram, as well (Figure 36). Thus, there does not seem to be any problem with the normality assumption. A similar normality is observed in the residual box plot in Figure 37 below.



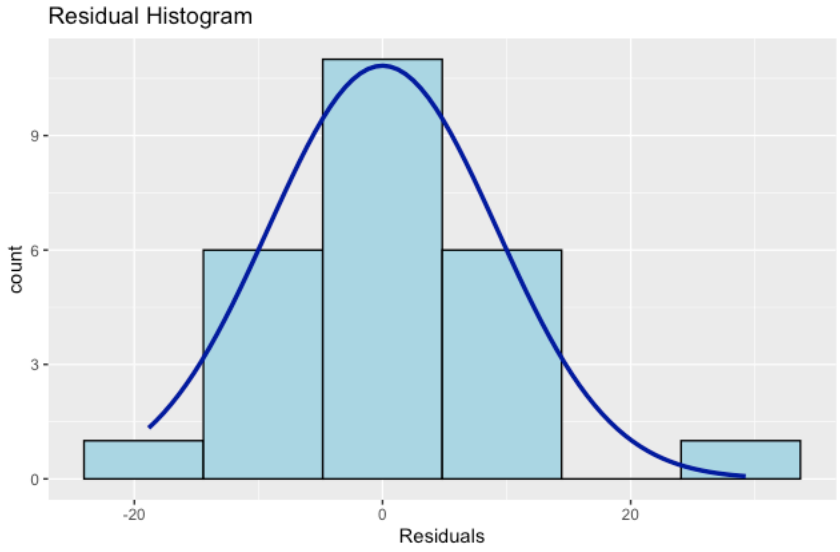


Figure 36: Revised Model Residual Histogram

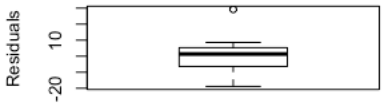


Figure 37: Residual Boxplot

When plotting the residuals vs the predicted response, we examine the shape of the data points. If the residuals can be contained in a horizontal band, then there are no obvious model defects. Figure 38 shows that the data points are mostly contained in a horizontal band. There are, however, a few observations with greater variance.

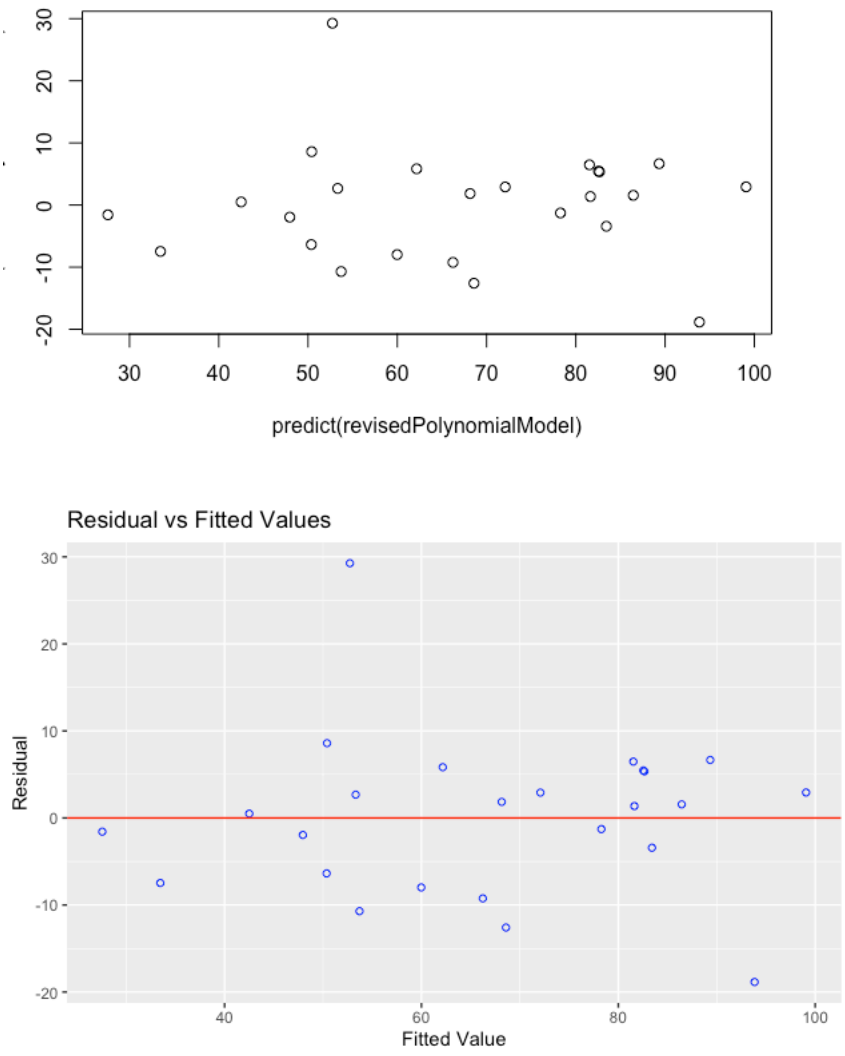


Figure 38: Residuals v Predicted

As previously mentioned, studentized residuals have a constant variance $Var(r_i) = 1$ regardless of the location of x_i, when the model form is correct. Thus, Studentized residuals are going to be more effective for detecting outlying Y observations than standardized residuals.

Rather than using MS_{Res} to estimate σ^2, R-Student removes the ith observation, making it an externally studentized residual. In many situations, t_i will differ little from the studentized residual r_i. However, if the ith observation is influential, then $S^2_{(i)}$ can differ significantly from

MS_{Res}, and thus the R-Student statistic will be more sensitive to this point. See Figure 39 for Studentized, and Studentized Deleted Residuals. We see that the studentized residual has a minimum of -2.138, a maximum of 3.196, a mean of -0.002 and a standard deviation of 1,016. The R-Student (or studentized deleted residual) has a minimum of -2.347, a maximum of 4.266 (a potential outlier), a mean of 0.030, and a standard deviation of 1.184. Here we see that observations 17 and 9 may be outliers.

Residuals Statistics[a]

	Minimum	Maximum	Mean	Std. Deviation	N
Predicted Value	27.5752	99.0743	66.7200	19.14479	25
Std. Predicted Value	-2.045	1.690	.000	1.000	25
Standard Error of Predicted Value	1.927	5.564	3.202	.939	25
Adjusted Predicted Value	28.0134	98.4355	66.7527	19.23908	25
Residual	-18.84514	29.26776	.00000	9.20709	25
Std. Residual	-1.960	3.043	.000	.957	25
Stud. Residual	-2.138	3.196	-.002	1.016	25
Deleted Residual	-22.43463	32.27673	-.03270	10.37532	25
Stud. Deleted Residual	-2.347	4.266	.030	1.184	25
Mahal. Distance	.003	7.073	1.920	1.729	25
Cook's Distance	.001	.350	.042	.087	25
Centered Leverage Value	.000	.295	.080	.072	25

a. Dependent Variable: Satisfaction

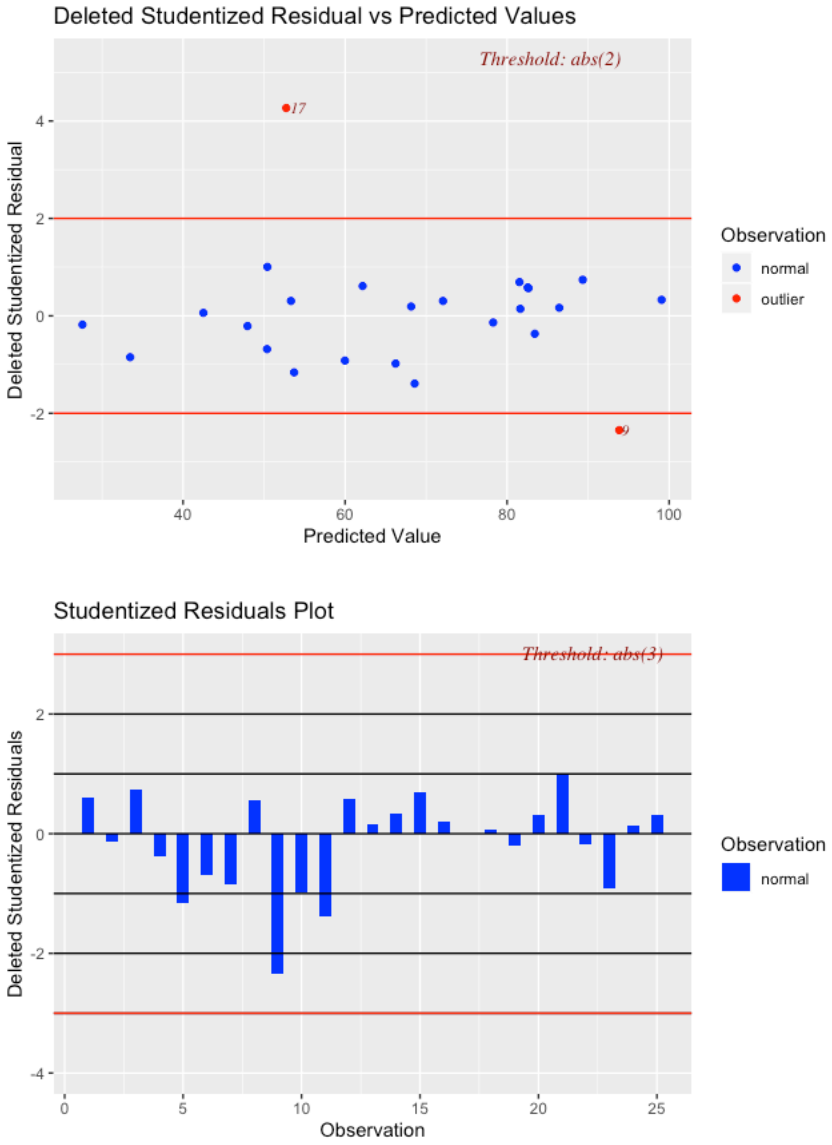


Figure 39: Studentized and R-Student Residuals

> #3d Scatterplot
> scatterplot3d(data$Age, data$SeverityCubed, e, xlab = "Age", ylab = "SeverityCubed", zlab = "Residuals")

> #Obtain Fitted Values
> fitted(revisedPolynomialModel) #predicted values

1	2	3	4	5	6	7	8	9	10	11
62.16520	78.27944	89.33772	83.42005	53.70829	50.36618	33.46883	82.65072	93.84514	66.23589	68.58775

12	13	14	15	16	17	18	19	20	21	22

```
82.54352 86.43962 99.07425 81.52585 68.15493 52.73224 42.50590 47.95167 53.32263
50.40838 27.57525
      23       24       25
59.98008 81.63305 72.08740

> #Obtain regression diagnostics for each observation: y_hat, coefficients, sigma, weighted
residual

> influence(revisedPolynomialModel) #regression diagnostics
$hat
         1          2          3          4          5          6          7          8          9
0.04399040 0.09640421 0.14240896 0.10392864 0.06623501 0.08254638 0.17384964
0.07134278 0.15999756
        10         11         12         13         14         15         16         17         18
0.04014293 0.07650469 0.07972445 0.08650664 0.17920640 0.07906022 0.04038831
0.09322414 0.33471104
        19         20         21         22         23         24         25
0.10350983 0.20629101 0.20895862 0.21761819 0.19005032 0.06666072 0.05673890

$coefficients
   (Intercept)          Age SeverityCubed
1  -0.02962253  6.370878e-03 -4.183164e-07
2  -0.06595339 -1.784502e-03  8.346073e-07
3   2.06766380 -4.123847e-02  2.833565e-06
4  -0.82240626  1.614061e-02 -1.258514e-06
5   0.20960932 -4.996452e-03 -3.456438e-06
6   0.21821306 -3.344660e-03 -2.717481e-06
7   1.37605762 -2.432844e-02 -4.178226e-06
8   0.94930439 -1.429552e-02  6.581503e-08
9  -6.49696348  1.251351e-01 -6.359579e-06
10 -0.38994195  7.887861e-04 -2.923815e-07
11  0.19254336 -3.019577e-02  6.652406e-06
12  0.64810635 -2.803570e-03 -2.239244e-06
13  0.26640128 -2.848660e-03 -4.442468e-07
14  1.09707885 -1.982547e-02  4.457491e-07
15  0.68867565 -1.345447e-03 -2.828824e-06
16  0.08710608 -1.365181e-05 -7.928235e-08
17  0.24502887 -2.046117e-02  1.740538e-05
18  0.03310504 -2.457603e-03  1.014034e-06
19  0.07100523 -4.599203e-04 -1.123797e-06
20 -0.62167681  2.171392e-02 -2.897684e-06
21 -2.17856715  7.225468e-02 -8.846779e-06
22  0.40885659 -7.610658e-03 -8.548559e-07
23  1.36954752 -5.566829e-02  8.897767e-06
24  0.22475288 -3.222535e-03 -1.948728e-08
25  0.10934898  2.667026e-03 -1.013051e-06
```

```
$sigma
      1       2       3       4       5       6       7       8       9      10      11
9.756270 9.838410 9.716789 9.811167 9.541117 9.735349 9.678080 9.767973 8.760595
9.625420 9.418609
     12      13      14      15      16      17      18      19      20      21      22
9.764220 9.836344 9.817534 9.732076 9.834208 7.203916 9.841906 9.832510 9.820922
9.614417 9.835118
     23      24      25
9.650729 9.837949 9.821014

$wt.res
        1         2         3         4         5         6         7         8
  5.8348049  -1.2794403   6.6622777  -3.4200533 -10.7082881  -6.3661850  -7.4688329
5.3492796
        9        10        11        12        13        14        15        16
-18.8451418  -9.2358938 -12.5877538   5.4564772   1.5603835   2.9257464   6.4741519
1.8450667
       17        18        19        20        21        22        23        24
 29.2677626   0.4940959  -1.9516701   2.6773662   8.5916213  -1.5752495  -7.9800751
1.3669543
       25
  2.9125954

> #Diagnostic plots
> layout(matrix(c(1,2,3,4),2,2)) #optional 4 graphs/page
> plot(revisedPolynomialModel)

> #Regression Diagnostics Battery
> ols_plot_diagnostics(revisedPolynomialModel)

> #Obtain residuals
> residuals(revisedPolynomialModel) #residuals
        1         2         3         4         5         6         7         8
  5.8348049  -1.2794403   6.6622777  -3.4200533 -10.7082881  -6.3661850  -7.4688329
5.3492796
        9        10        11        12        13        14        15        16
-18.8451418  -9.2358938 -12.5877538   5.4564772   1.5603835   2.9257464   6.4741519
1.8450667
       17        18        19        20        21        22        23        24
 29.2677626   0.4940959  -1.9516701   2.6773662   8.5916213  -1.5752495  -7.9800751
1.3669543
       25
  2.9125954
> e <-residuals(revisedPolynomialModel)
> e
```

```
        1         2         3         4         5         6         7         8
 5.8348049 -1.2794403  6.6622777 -3.4200533 -10.7082881 -6.3661850 -7.4688329
5.3492796
        9        10        11        12        13        14        15        16
-18.8451418 -9.2358938 -12.5877538  5.4564772  1.5603835  2.9257464  6.4741519
1.8450667
       17        18        19        20        21        22        23        24
 29.2677626  0.4940959 -1.9516701  2.6773662  8.5916213 -1.5752495 -7.9800751
1.3669543
       25
 2.9125954

> boxplot(e, ylab = "Residuals")

> #Plot residuals against Y_hat
> yhat <- fitted(revisedPolynomialModel)
> plot(yhat, e, xlab = "Fitted Values", ylab = "Residuals", ylim = c(-20,20))
> abline(h=0, lty = 2)

> #Plot residuals against each of the predictor variables
> #Age Residuals
> plot(Age,e, xlab="Age", ylab = "Residuals", ylim = c(-20,20))
> abline(h=0, lty=2)
> #Severity Residuals
> plot(data$SeverityCubed, e, xlab = "SeverityCubed", ylab = "Residuals", ylim = c(-20,20))
> abline(h=0, lty=2)
> #3d Scatterplot
> scatterplot3d(data$Age, data$SeverityCubed, e, xlab = "Age", ylab = "SeverityCubed", zlab =
"Residuals")

> #More Residuals
> summary(revisedPolynomialModel)$residuals
        1         2         3         4         5         6         7         8
 5.8348049 -1.2794403  6.6622777 -3.4200533 -10.7082881 -6.3661850 -7.4688329
5.3492796
        9        10        11        12        13        14        15        16
-18.8451418 -9.2358938 -12.5877538  5.4564772  1.5603835  2.9257464  6.4741519
1.8450667
       17        18        19        20        21        22        23        24
 29.2677626  0.4940959 -1.9516701  2.6773662  8.5916213 -1.5752495 -7.9800751
1.3669543
       25
 2.9125954
> ols_plot_resid_stud(revisedPolynomialModel)

> #Residual Normality Test
```

```
> ols_test_normality(revisedPolynomialModel)
-----------------------------------------------
      Test             Statistic       pvalue
-----------------------------------------------
Shapiro-Wilk              0.9145         0.0384
Kolmogorov-Smirnov        0.1547         0.5374
Cramer-von Mises          1.8829         0.0000
Anderson-Darling          0.6239         0.0925
-----------------------------------------------
> #Correlation Between Observed Residuals and Expected Residuals Under Normality
> ols_test_correlation(revisedPolynomialModel)
[1] 0.9452152
> #Correlation = 0.9452152

> #Residual vs Fitted Values Plot
> ols_plot_resid_fit(revisedPolynomialModel)

> #Residual Histogram
> ols_plot_resid_hist(revisedPolynomialModel)


> #Deleted Studentized residual vs predicted values
> ols_plot_resid_stud_fit(revisedPolynomialModel)

> #Residuals
> r<-resid(revisedPolynomialModel)
> r
        1         2         3         4         5         6         7         8
 5.8348049 -1.2794403  6.6622777 -3.4200533 -10.7082881 -6.3661850 -7.4688329
5.3492796
        9        10        11        12        13        14        15        16
-18.8451418 -9.2358938 -12.5877538  5.4564772  1.5603835  2.9257464  6.4741519
1.8450667
       17        18        19        20        21        22        23        24
 29.2677626  0.4940959 -1.9516701  2.6773662  8.5916213 -1.5752495 -7.9800751
1.3669543
       25
 2.9125954

> #QQ Plot
> qqnorm(resid(revisedPolynomialModel))
> ols_plot_resid_qq(revisedPolynomialModel)

> #Predictively adjusted residuals
> (pr<-resid(revisedPolynomialModel)/(1-lm.influence(revisedPolynomialModel)$hat))
        1         2         3         4         5         6         7         8
```

```
  6.1032911  -1.4159432   7.7685953  -3.8167198 -11.4678620  -6.9389720  -9.0405249
5.7602305
         9         10         11         12         13         14         15         16
-22.4346274  -9.6221553 -13.6305552   5.9291776   1.7081498   3.5645336   7.0299405
1.9227222
        17         18         19         20         21         22         23         24
 32.2767332   0.7426786  -2.1770123   3.3732341  10.8611528  -2.0134025  -9.8525566
1.4645845
        25
  3.0877934

> #Construct residual vs predicted response
> plot(predict(revisedPolynomialModel), resid(revisedPolynomialModel))


> #Normal Probability Plot of Residuals
> n <-length(e)
> MSE <-sum(e^2)/(n-4)
> RankofRes <-rank(e)
> Zscore <- qnorm((RankofRes-0.375)/(n+0.25))
> ExpRes <- Zscore * sqrt(MSE)
> plot(ExpRes, e, xlab = "Expected Score", ylab = "Residuals")
> abline(a = 0, b = 1)
```

PRESS

R^2, also known as coefficient of determination, is a popular measure of quality of fit in regression. However, it does not offer any significant insights into how well our regression model can predict future values. Instead, the PRESS statistic (the predicted residual sum of squares) can be used as a measure of predictive power. The PRESS statistic can be computed in the leave-one-out cross validation process, by adding the square of the residuals for the case that is left out. As a reminder, in the leave-one-out cross validation, one case of the data set is used as the testing set and the remaining are used as the testing set. We iterate this process, until all cases have served as the testing set. Or, we can calculate it in the following manner:

```
> #Cross-validated residuals
> #Regular RSS is
> sum(r^2)
```

```
[1] 2034.492
> #2034.492
> #PRESS is
> sum(pr^2)
[1] 2583.56
> #2583.56
> #2583.56
> #Note PRESS is bigger because predicting is harder than fitting
> #Another way to calculate the PRESS statistic
> PRESS(revisedPolynomialModel)
[1] 2583.56
```

Search For High-Leverage Or Overly Influential Observations

Outliers can be found by examining standardized residuals, leverage (how far an observation deviates from the mean of that variable), Cook's Distance (which combines leverage and residuals), and DfBeta (which is a more specific measure of influence).

Standardized Residuals: Outliers have values greater than 2 and less than -2. Results indicate that observations 17 and 9 may be outliers.

Leverage indicates how far an observation deviates from the mean of that variable. It is calculated by:

$$Leverage = \frac{2k+2}{n}, \quad where\ k = \#\ predictors, n = \#\ observations$$

For this data set, the leverage threshold is 0.24. After sorting the leverage threshold for each value and by visually examining Figure 40, we see that observations 17 and 9 are outliers and observation 18 has an abnormally high/strong leverage.

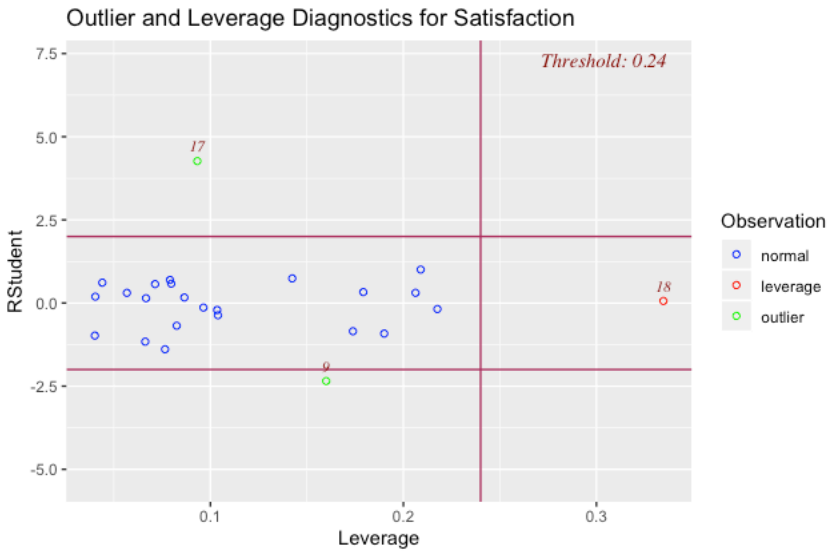


Figure 40: R Student and Leverage.

Cook's Distance combines leverage and residuals. The general rule of thumb is the higher the value, the better. The lowest possible value is 0 and the conventional cut off is calculated by 4/n (which is 0.16 for this data set). After sorting each observation's Cook's Distance and analyzing the diagnostic plots below in Figure 41, we find two observations below the cut off (9 and 17).

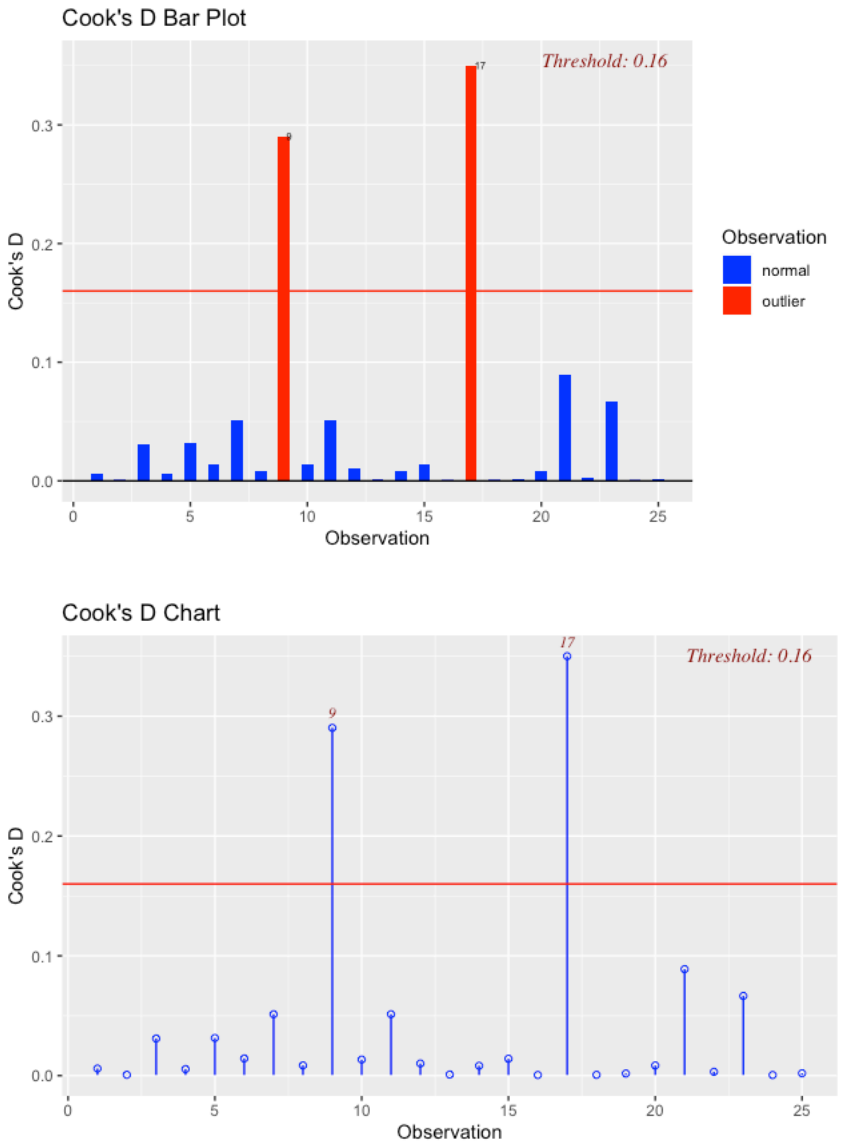


Figure 41: Cook's Distance

DfBETA is a more specific measure of influence. It measures the difference in each parameter estimate with and without the influential point. There is a DFBETA for each data point (i.e., if there are n observations and k variables, there will be n∗kn∗k DFBETAs). In general, large values of DFBETAS indicate observations that are influential in estimating a given parameter. Belsley, Kuh, and Welsch recommend 2 as a general cutoff value to indicate influential observations and $\frac{2}{\sqrt{n}}$ as a size-adjusted cutoff. Thus, we must sort the DfBeta

observations and check that each observation passes the threshold of 0.4. From Figure 42, we see observations 9 and 17 as outliers.

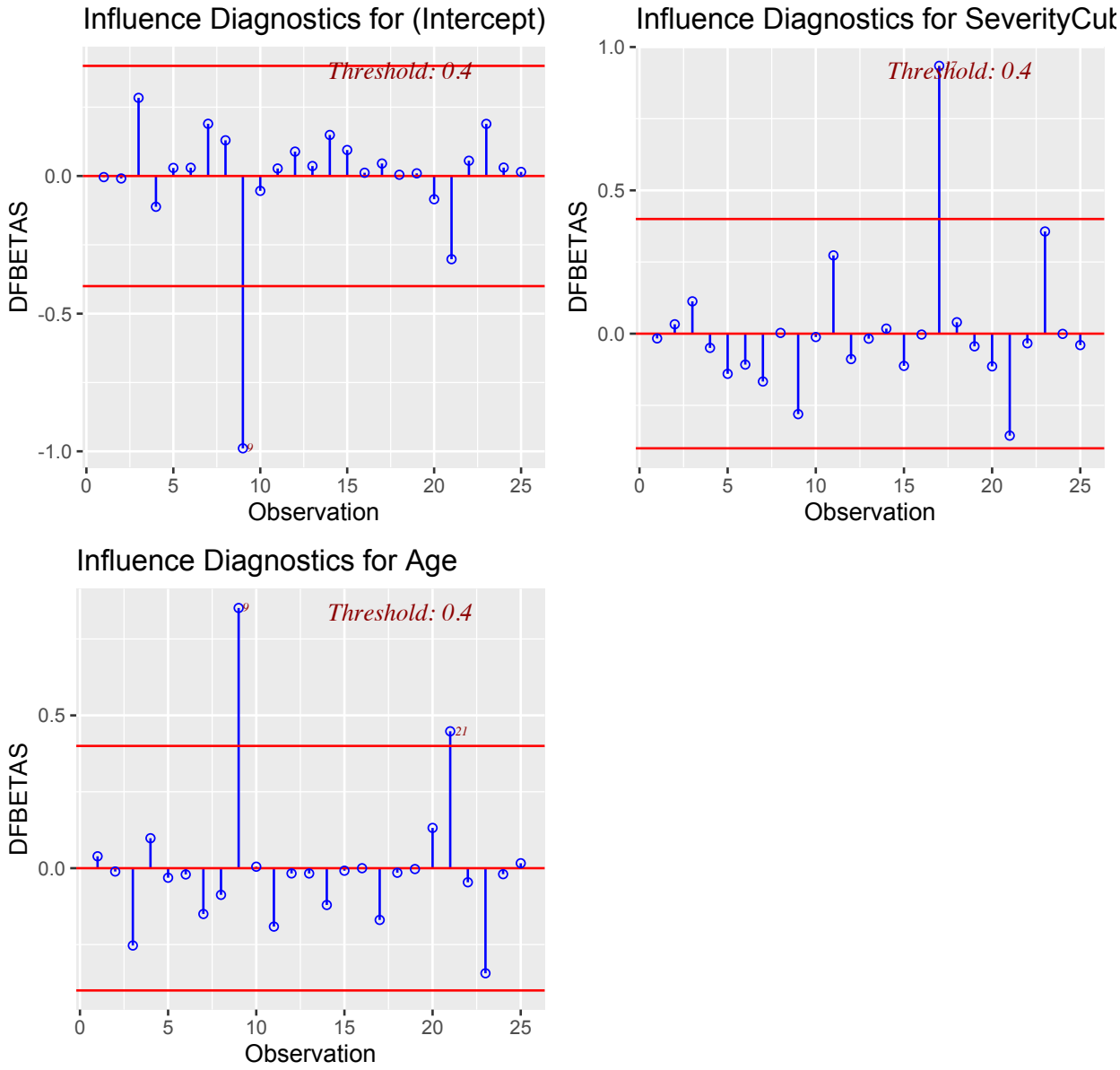


Figure 42: DfBetas

VIF can be used to detect collinearity, which causes instability in parameter estimation in regression-type models. The VIF is based on the square of the multiple correlation coefficient resulting from regressing a predictor variable against all other predictor variables. If a variable

has a strong linear relationship with at least one other variable, the correlation coefficient would be close to 1, and VIF for that variable would be large. A VIF greater than 10 is a signal that the model has a collinearity problem. We will use a general VIF cut off VIF > 5. From the R-output copied below, no variables have a VIF > 5, indicating that multicollinearity is not a problem for this model.

```
> # VIF = (1/1-R^2). VIF > 5 indicates associated regression coefficients are poorly estimated
b/c multicollinearity
> ols_vif_tol(revisedPolynomialModel)
# A tibble: 2 x 3
  Variables       Tolerance   VIF
  <chr>           <dbl>       <dbl>
1 Age             0.675        1.48
2 SeverityCubed   0.675        1.48

> #standardized residuals
> ols_plot_resid_stand(revisedPolynomialModel)


> #Studentized Residuals vs Leverage Plot
> ols_plot_resid_lev(revisedPolynomialModel)


> #Detect Influence with Leverage
> #The observed value of y_i is influential if h_i > [2(k+1)]/n
> #Where h_i = leverage for the ith obervation
> #k = # of betas in the model (excluding b_0)
> # [2(2+1)]/n = [2(3)/25] = 6/25 = 0.5
> ols_leverage(revisedPolynomialModel)
 [1] 0.04399040 0.09640421 0.14240896 0.10392864 0.06623501 0.08254638 0.17384964
0.07134278
 [9] 0.15999756 0.04014293 0.07650469 0.07972445 0.08650664 0.17920640 0.07906022
0.04038831
[17] 0.09322414 0.33471104 0.10350983 0.20629101 0.20895862 0.21761819 0.19005032
0.06666072
[25] 0.05673890
```

```
> #Check for Collinearity
> ols_coll_diag(revisedPolynomialModel)
Tolerance and Variance Inflation Factor
---------------------------------------
# A tibble: 2 x 3
  Variables      Tolerance      VIF
  <chr>          <dbl>          <dbl>
1 Age            0.675          1.48
2 SeverityCubed  0.675          1.48


Eigenvalue and Condition Index
------------------------------
  Eigenvalue Condition Index   intercept         Age SeverityCubed
1 2.73601707        1.000000 0.008782298 0.006574019    0.02903965
2 0.23364281        3.422025 0.083889137 0.013358569    0.73089205
3 0.03034012        9.496219 0.907328565 0.980067412    0.24006830

> vcov(revisedPolynomialModel) #covariance matrix for model parameters
             (Intercept)          Age SeverityCubed
(Intercept)    5.198467e+01 -1.049127e+00  4.214764e-05
Age           -1.049127e+00  2.602807e-02 -2.287087e-06
SeverityCubed  4.214764e-05 -2.287087e-06  6.184307e-10

> #Cook's Distance: Combines leverage and residuals
> #Higher value, the better
> #Lowest Value = 0
> #Conventional Cut off is 4/n
> ols_plot_cooksd_bar(revisedPolynomialModel)
> ols_plot_cooksd_chart(revisedPolynomialModel)
> cooks.distance(revisedPolynomialModel)
          1           2           3           4           5           6           7
0.0059065174 0.0006966796 0.0309790221 0.0054570912 0.0313977056 0.0143262892
0.0512160977
          8           9          10          11          12          13          14
0.0085324775 0.2902661979 0.0133967428 0.0512342704 0.0101024297 0.0009098017
0.0082073762
         15          16          17          18          19          20          21
0.0140833734 0.0005381882 0.3500683950 0.0006654528 0.0017682707 0.0084609409
0.0888500179
         22          23          24          25
0.0031798145 0.0664985038 0.0005153998 0.0019499468

> #dfbetas:measures the difference in each parameter estimate with and without the influential
point
> ols_plot_dfbetas(revisedPolynomialModel)
```

```
> dfbeta(revisedPolynomialModel)
   (Intercept)          Age SeverityCubed
1  -0.02962253  6.370878e-03 -4.183164e-07
2  -0.06595339 -1.784502e-03  8.346073e-07
3   2.06766380 -4.123847e-02  2.833565e-06
4  -0.82240626  1.614061e-02 -1.258514e-06
5   0.20960932 -4.996452e-03 -3.456438e-06
6   0.21821306 -3.344660e-03 -2.717481e-06
7   1.37605762 -2.432844e-02 -4.178226e-06
8   0.94930439 -1.429552e-02  6.581503e-08
9  -6.49696348  1.251351e-01 -6.359579e-06
10 -0.38994195  7.887861e-04 -2.923815e-07
11  0.19254336 -3.019577e-02  6.652406e-06
12  0.64810635 -2.803570e-03 -2.239244e-06
13  0.26640128 -2.848660e-03 -4.442468e-07
14  1.09707885 -1.982547e-02  4.457491e-07
15  0.68867565 -1.345447e-03 -2.828824e-06
16  0.08710608 -1.365181e-05 -7.928235e-08
17  0.24502887 -2.046117e-02  1.740538e-05
18  0.03310504 -2.457603e-03  1.014034e-06
19  0.07100523 -4.599203e-04 -1.123797e-06
20 -0.62167681  2.171392e-02 -2.897684e-06
21 -2.17856715  7.225468e-02 -8.846779e-06
22  0.40885659 -7.610658e-03 -8.548559e-07
23  1.36954752 -5.566829e-02  8.897767e-06
24  0.22475288 -3.222535e-03 -1.948728e-08
25  0.10934898  2.667026e-03 -1.013051e-06
```

Test Lack of Fit

Finally, a Lack of Fit test was conducted and testing using the following hypothesis testing

H_0: $The\ relationship\ assumed\ in\ the\ model\ is\ reasonable; i.e., there\ is\ no\ lack\ of\ fit\ in\ the\ model$

H_a: $The\ relationship\ assumed\ in\ the\ model\ is\ not\ reasonable; i.e., there\ is\ a\ lack\ of\ fit\ in\ the\ model$

As before, rejection criteria was:

- $F_{obtained} > F_{critical}$
- $\alpha > p - value$

From the R-output listed below, we fail to reject the null hypothesis and conclude that the model is reasonable (i.e., there is no lack of fit).

```
> #lack of fit
> ols_test_f(revisedPolynomialModel)

 F Test for Heteroskedasticity
 -----------------------------
 Ho: Variance is homogenous
 Ha: Variance is not homogenous

 Variables: fitted values of Satisfaction

     Test Summary
 -------------------------
 Num DF     =    1
 Den DF     =    23
 F          =    0.0859901
 Prob > F   =    0.7719677
```

References

Allen, M.P. (1997). Understanding Regression Analysis. New York: Plenum Press.

Montgomery, D.C., Peck, E.A., and Vining, C.G. (2012) Introduction to Linear Regression Analysis. 5th Edition, John Wiley and Sons.

Full R-Code:

```
#Course Project
#Data from Table B.17 page 570 of Textbook - Intro to Linear Regression Analysis
        (Montogomery, et al 2012)
#Standard libraries
library("ggplot2")
library(readxl)
library (MPV)
library (olsrr)
library("ggpubr")
library("Hmisc")
library(caret)
library (lattice)
library(leaps)
library(MASS)
library(alr3)
library(rms)
library(ppcor)
library(ggthemes)
library(data.table)
library(ISLR)
library(tibble)
library(aod)
library(tidyverse)
library(modelr)
library(broom)
library(WVPlots)
library(lmtest)
library(drc)
library(nlme)
library(aomisc)
library(scatterplot3d)
library(dplyr)
#install.packages("pastecs")
library(pastecs)
#install.packages("psych")
library(psych)
#install.packages("moments")
library(moments)

#Import Data
data_table_B17 <- read_excel("Desktop/JHU/Data Science/Statistical Models and
        Regression/Course Project/data-table-B17.xls")
View(data_table_B17)
attach(data_table_B17)
```

```
#Create data frame
#Target/Response Variable = Satisfaction
#Predictor Variables = Age, Severity, Surgical-Medical, Anxiety
data <- data.frame(Satisfaction, Age, Severity, `Surgical-Medical`, Anxiety)

#Run EDA
summary(data)
#Get standard deviations
sapply(data, sd)

#Get n, nmiss, unique, mean, 5, 10, 25, 50, 75, 90, 95th percentiles
#5 lowest & 5 highest scores
describe(data)

#nbr.val, nbr.null, nbr.na, min max, range, sum
#median, mean, SE.mean, CI.mean, var, st.dev, coef.var
stat.desc(data)

#n, mean, sd, median, trimmed, mad, min, max, range, skew, kurtosis, se
describeBy(data, group=NULL)

#Get Tukey's 5 (minimum, lower-hinge, median, upper-hinge, maximum)

#Satisfaction
fivenum(data$Satisfaction, na.rm = TRUE)
#[1]  26  52  70  83 102
#Get interquartile range
IQR(data$Satisfaction, na.rm = TRUE)
#IQR = 31
#Skewness
skewness(data$Satisfaction)
#-0.307575
#very slight or moderate left/negative skew
#Kurtosis
kurtosis(data$Satisfaction)
#2.120311
#value is less than 3 = low kurtosis


#Age
#Get Tukey's 5 (minimum, lower-hinge, median, upper-hinge, maximum)
fivenum(data$Age, na.rm = TRUE)
#[1]24 39 51 61 79
#Get interquartile range
IQR(data$Age, na.rm = TRUE)
```

```
#IQR = 22
#Skewness
skewness(data$Age)
#-0.01654419
#Almost no skew
#Kurtosis
kurtosis(data$Age)
#2.160039
#value is less than 3 = low kurtosis

#Severity
fivenum(data$Severity, na.rm = TRUE)
#[1] 24 38 42 58 71
#Get interquartile range
IQR(data$Severity, na.rm = TRUE)
#IQR = 20
#Skewness
skewness(data$Severity)
#[1] 0.282166
#Almost no skew
#Kurtosis
kurtosis(data$Severity)
#2.047968
#value is less than 3 = low kurtosis

#Anxiety
fivenum(data$Anxiety, na.rm = TRUE)
#[1] 1.9 2.4 3.3 5.1 7.8
#Get interquartile range
IQR(data$Anxiety, na.rm = TRUE)
#IQR = 2.7
#Skewness
skewness(data$Anxiety)
#[1] 0.7346762
#Value is close to positive 1, indicating some skew
#Kurtosis
kurtosis(data$Anxiety)
#2.454605
#value is close to 3, indicating nearly normal shape

#Get Surgical-Medical Counts
table(data$Surgical.Medical)
#Number of Observations for 0=Surgical: 11
#Number of Observations for 1=Medical: 14
#Examine distribution of Surgical-Medical
ggplot(data = data) + geom_bar(mapping=aes(x=data$Surgical.Medical))
```

```
histogram(`Surgical-Medical`, data=data)

#Histogram for each variable
#Satisfaction
ggplot(data=data)+geom_histogram(mapping=aes(x=data$Satisfaction))
ggplot(data=data)+geom_histogram(mapping=aes(x=data$Satisfaction), binwidth = 10)
SatisfactionPlot <-ggplot(data, aes(x=Satisfaction))
#Density Plot
#y axis scale = density
SatisfactionPlot +geom_density()+geom_vline(aes(xintercept=mean(Satisfaction)), linetype =
        "dashed", size = 0.6)
#Change y axis to count instead of density
SatisfactionPlot +geom_density(aes(y=..count..),
        fill="lightgray")+geom_vline(aes(xintercept=mean(Satisfaction)), linetype="dashed",
        size=0.6, color="#FC4E07")
#Histogram plot (counts)
SatisfactionPlot+geom_histogram(bins=10, color="black",
        fill="gray")+geom_vline(aes(xintercept=mean(Satisfaction)),linetype="dashed",size=0.6)
#Histogram plot (density)
SatisfactionPlot+geom_histogram(aes(y=..density..), color="black",
        fill="white")+geom_density(alpha=0.2,fill="#FF6666")
SatisfactionPlot+geom_histogram(bins=10, aes(y=..density..), color="black", fill="gray")
#Basic frequency polygon
SatisfactionPlot+geom_freqpoly(bins=10)

#Age
ggplot(data=data)+geom_histogram(mapping=aes(x=data$Age))
ggplot(data=data)+geom_histogram(mapping=aes(x=data$Age), binwidth = 10)
AgePlot <-ggplot(data, aes(x=Age))
#Density Plot
#y axis scale = density
AgePlot +geom_density()+geom_vline(aes(xintercept=mean(Age)), linetype = "dashed", size =
        0.6)
#Change y axis to count instead of density
AgePlot +geom_density(aes(y=..count..),
        fill="lightgray")+geom_vline(aes(xintercept=mean(Age)), linetype="dashed", size=0.6,
        color="#FC4E07")
#Histogram plot (counts)
AgePlot+geom_histogram(bins=10, color="black",
        fill="gray")+geom_vline(aes(xintercept=mean(Age)),linetype="dashed",size=0.6)
#Histogram plot (density)
AgePlot+geom_histogram(bins=10, aes(y=..density..), color="black",
        fill="white")+geom_density(alpha=0.2,fill="#FF6666")
AgePlot+geom_histogram(bins=10, aes(y=..density..), color="black", fill="white")
#Basic frequency polygon
AgePlot+geom_freqpoly(bins=10)
```

```
#Severity
ggplot(data=data)+geom_histogram(mapping=aes(x=data$Severity))
ggplot(data=data)+geom_histogram(mapping=aes(x=data$Severity), binwidth = 5)
SeverityPlot <-ggplot(data, aes(x=Severity))
#Density Plot
#y axis scale = density
SeverityPlot +geom_density()+geom_vline(aes(xintercept=mean(Severity)), linetype = "dashed",
        size = 0.6)
#Change y axis to count instead of density
SeverityPlot +geom_density(aes(y=..count..),
        fill="lightgray")+geom_vline(aes(xintercept=mean(Severity)), linetype="dashed",
        size=0.6, color="#FC4E07")
#Histogram plot (counts)
SeverityPlot+geom_histogram(bins=10, color="black",
        fill="gray")+geom_vline(aes(xintercept=mean(Severity)),linetype="dashed",size=0.6)
#Histogram plot (density)
SeverityPlot+geom_histogram(bins=10, aes(y=..density..), color="black",
        fill="white")+geom_density(alpha=0.2,fill="#FF6666")
SeverityPlot+geom_histogram(bins=10, aes(y=..density..), color="black", fill="white")
#Basic frequency polygon
SeverityPlot+geom_freqpoly(bins=10)

#Anxiety
ggplot(data=data)+geom_histogram(mapping=aes(x=data$Anxiety))
ggplot(data=data)+geom_histogram(mapping=aes(x=data$Anxiety), binwidth = 1)
AnxietyPlot <-ggplot(data, aes(x=Anxiety))
#Density Plot
#y axis scale = density
AnxietyPlot +geom_density()+geom_vline(aes(xintercept=mean(Anxiety)), linetype = "dashed",
        size = 0.6)
#Change y axis to count instead of density
AnxietyPlot +geom_density(aes(y=..count..),
        fill="lightgray")+geom_vline(aes(xintercept=mean(Anxiety)), linetype="dashed",
        size=0.6, color="#FC4E07")
#Histogram plot (counts)
AnxietyPlot+geom_histogram(bins=10, color="black",
        fill="gray")+geom_vline(aes(xintercept=mean(Anxiety)),linetype="dashed",size=0.6)
#Histogram plot (density)
AnxietyPlot+geom_histogram(bins=10, aes(y=..density..), color="black",
        fill="white")+geom_density(alpha=0.2,fill="#FF6666")
AnxietyPlot+geom_histogram(bins=10, aes(y=..density..), color="black", fill="white")
#Basic frequency polygon
AnxietyPlot+geom_freqpoly(bins=10)
```

```
#Plot each IV against the DV
ggplot(data=data,aes(x=Age,y=Satisfaction))+geom_point()
ggplot(data=data,aes(x=Severity,y=Satisfaction))+geom_point()
plot(Satisfaction[][`Surgical-Medical`==0]~`Surgical-Medical`[][`Surgical-Medical`==0])
plot(Satisfaction[][`Surgical-Medical`==1]~`Surgical-Medical`[][`Surgical-Medical`==1])
ggplot(data=data,aes(x=Anxiety,y=Satisfaction))+geom_point()

#Correlations
pairs(data)
cor(data)
#Age has a strong (negative) correlation with Satisfaction (r = -0.871)
#Severity has a strong (negative) correlation with Satisfaction (r = -653)
#Surgical-Medical has a weak correlation with Satisfaction (r = -0.182)
#Anxiety has strong correlation with Satisfaction (r = -0.513)
#Where, strong correlation is r >= |0.5|
#May have a multicollinearity problem:
#Severity & Age r = 0.529
#Anxiety & Age r = 0.621
corr.test(Satisfaction, Age)
#r=-0.87 (strong negative)
corr.test(Satisfaction, Age, method = "spearman")
#rho = -0.85 (strong negative)
corr.test(Satisfaction, Severity)
#r=-0.65 (strong negative)
corr.test(Satisfaction, Severity, method = "spearman")
#rho = -0.6 (strong negative)
corr.test(Satisfaction, `Surgical-Medical`)
#r=-0.18 (weak negative)
corr.test(Satisfaction, `Surgical-Medical`, method = "spearman")
#rho = -0.17 (weak negative)
corr.test(Satisfaction, Anxiety)
#r=-0.51 (strong negative)
corr.test(Satisfaction, Anxiety, method = "spearman")
#rho = -0.42 (medium negative)

#Start with All-In Model (First Order Linear Multiple Regression, Main Effects)
AllInModel <-lm(Satisfaction~Age+Severity+`Surgical-Medical`+Anxiety, data = data)

#Obtain results
summary(AllInModel)
#Satisfaction = 140.1689 - (1.1428*Age) - (0.4699*Severity) + (2.2259*Surgical-Medical) +
        (1.2673*Anxiety)
#R^2 = 0.8183, Ra^2 = 0.7819
#Beta Coefficients for Age & Severity have p<0.05
#Confirm coefficients
coefficients(AllInModel) #model coefficients
```

```
#Get standard deviations
sapply(data, sd)
#Confirm with ols package
ols_regress(AllInModel)
#Remember, Fc = (SSR/k) / (SSE/n-k+1)
#Alpha = .05
#Fo = F(1-alpha, df, df)
#Fo = F(1-.05,4,20)
qf(0.95,4,20)
#Fo = 2.866081

#Confidence Intervals (95%)
confint(AllInModel)
#Age: Small Range, does not include 0
#Severity: Small range, does not include 0
#Surgical-Medical: Large range, includes 0
#Anxiety: Large range, includes 0
#Another way to calculate confidence intervals:
b <- summary(AllInModel)$coef[,1]
b
seb <-summary(AllInModel)$coef[,2]
seb
alpha <- 0.05
tval <- qt(1-alpha/8,20)
lower.lim <-b-tval*seb
lower.lim
upper.lim<-b+tval*seb
upper.lim

#R^2
R2 <-summary(AllInModel)$r.squared
R2
#R^2 = 0.8182508

#Plot All In Model
plot(Age+Severity+`Surgical-Medical`+Anxiety, Satisfaction, data=data)
lines(lowess(Age+Severity+`Surgical-Medical`+Anxiety, Satisfaction))
#Appears linear
#Can also plot with ols library
ols_plot_reg_line(Satisfaction,Age+Severity+`Surgical-Medical`+Anxiety)
ols_plot_response(AllInModel)

#Obtain Fitted Values
fitted(AllInModel) #predicted values
```

```
#Obtain regression diagnostics for each observation: y_hat, coefficients, sigma, weighted
        residual
influence(AllInModel) #regression diagnostics
#Diagnostic plots
layout(matrix(c(1,2,3,4),2,2)) #optional 4 graphs/page
plot(AllInModel)
#Regression Diagnostics Battery
ols_plot_diagnostics(AllInModel)

#Obtain residuals
residuals(AllInModel) #residuals
e <-residuals(AllInModel)
e
boxplot(e, ylab = "Residuals")

#Plot residuals against Y_hat
yhat <- fitted(AllInModel)
plot(yhat, e, xlab = "Fitted Values", ylab = "Residuals", ylim = c(-20,20))
abline(h=0, lty = 2)

#Plot residuals against each of the predictor variables
#Age Residuals
plot(Age,e, xlab="Age", ylab = "Residuals", ylim = c(-20,20))
abline(h=0, lty=2)
#Severity Residuals
plot(Severity, e, xlab = "Severity", ylab = "Residuals", ylim = c(-20,20))
abline(h=0, lty=2)
#Surgical-Medical Residuals
plot(`Surgical-Medical`, e, xlab = "Surgical-Medical", ylab = "Residuals", ylim = c(-20,20))
abline(h=0,lty=2)
#Anxiety
plot(Anxiety,e, xlab = "Anxiety", ylab = "Residuals", ylim = c(-20,20))
abline(h=0, lty=2)

#3d Scatterplot
scatterplot3d(Age, Severity, e, xlab = "Age", ylab = "Severity", zlab = "Residuals")
scatterplot3d(Age,`Surgical-Medical`, e, xlab = "Age", ylab = "Surgical-Medical", zlab =
        "Residuals")
scatterplot3d(Age,Anxiety, e, xlab = "Age", ylab = "Anxiety", zlab = "Residuals")
scatterplot3d(Severity, `Surgical-Medical`, e, xlab = "Severity", ylab = "Surgical-Medical", zlab
        = "Residuals")
scatterplot3d(Severity, Anxiety, e, xlab = "Severity", ylab = "Anxiety", zlab = "Residuals")
scatterplot3d(`Surgical-Medical`, Anxiety, e, xlab = "Surgical-Medical", ylab = "Anxiety", zlab
        = "Residuals")

#More Residuals
```

```
summary(AllInModel)$residuals
ols_plot_resid_stud(AllInModel)
#Residual Normality Test
ols_test_normality(AllInModel)
#Correlation Between Observed Residuals and Expected Residuals Under Normality
ols_test_correlation(AllInModel)
#Correlation = 0.9483583
#Residual vs Fitted Values Plot
ols_plot_resid_fit(AllInModel)
#Residual Histogram
ols_plot_resid_hist(AllInModel)
#Studentized Residuals vs Leverage Plot
ols_plot_resid_lev(AllInModel)
#Deleted Studentized residual vs predicted values
ols_plot_resid_stud_fit(AllInModel)
#Residuals
r<-resid(AllInModel)
r
#QQ Plot
qqnorm(resid(AllInModel))
ols_plot_resid_qq(AllInModel)
#Predictively adjusted residuals
(pr<-resid(AllInModel)/(1-lm.influence(AllInModel)$hat))
#Construct residual vs predicted response
plot(predict(AllInModel), resid(AllInModel))
#Cross-validated residuals
#Regular RSS is
sum(r^2)
#PRESS is
sum(pr^2)
#2869.485
#Note PRESS is bigger because predicting is harder than fitting
#Another way to calculate the PRESS statistic
PRESS<-function(AllInModel){
  pr<-residuals(AllInModel)/(1-lm.influence(AllInModel$hat)
                          sum(pr^2)
                          }
PRESS(AllInModel)
#Same result: 2869.485

#standardized residuals
ols_plot_resid_stand(AllInModel)

#Normal Probability Plot of Residuals
n <-length(e)
MSE <-sum(e^2)/(n-4)
```

```
RankofRes <-rank(e)
Zscore <- qnorm((RankofRes-0.375)/(n+0.25))
ExpRes <- Zscore * sqrt(MSE)
plot(ExpRes, e, xlab = "Expected Score", ylab = "Residuals")
abline(a = 0, b = 1)

#Detect Influence with Leverage
#The observed value of y_i is influential if h_i > [2(k+1)]/n
#Where h_i = leverage for the ith obervation
#k = # of betas in the model (excluding b_0)
# [2(k+1)]/n = [2(5+1)/25] = 12/25 = 0.48
ols_leverage(AllInModel)

# VIF = (1/1-R^2). VIF > 5 indicates associated regression coefficients are poorly estimated b/c
        multicollinearity
ols_vif_tol(AllInModel)
#Result: All VIFs < 5 ?
vif(AllInModel)

#Check for Collinearity
ols_coll_diag(AllInModel)
rmatrix <- rcorr(as.matrix(data)) #can be pearson or spearman
rmatrix
vcov(AllInModel) # covariance matrix for model parameters

#Cook's Distance: Combines leverage and residuals
#Higher value, the better
#Lowest Value = 0
#Conventional Cut off is 4/n
ols_plot_cooksd_bar(AllInModel)
ols_plot_cooksd_chart(AllInModel)
cooks.distance(AllInModel)

#dfbetas:measures the difference in each parameter estimate with and without the influential
        point
ols_plot_dfbetas(AllInModel)
dfbeta(AllInModel)

#lack of fit
ols_test_f(AllInModel)
#Fail to reject H0; p = 0.7596821


#Another tyoe of Lack of Fit Testing
#Pure Error Analysis of Variance: For a linear model object, finds the sum of squares for lack of
        fit and the sum of
```

```
#squares for pure error. These are added to the standard anova table to give a test for lack of fit.
        If there is no
#pure error, then the regular anova table is returned.
#For regression models with one predictor, say y ~ x, this method fits y ~ x + factor(x) and prints
        the anova table.
#With more than one predictor, it computes a random linear combination L of the terms in the
        mean function and then
#gives the anova table for update(mod, ~.+factor(L)).
pureErrorAnova(AllInModel) #include pure error
anova(AllInModel)

#Verify Assumption of Constant Variance (Breusch-Pagan test)
SSE <- sum(e^2)
SSE
#1968.533
n <-length(e)
reg2 <-lm (e^2~Age+Severity+`Surgical-Medical`+Anxiety, data = data)
y2hat<-fitted(reg2)
SSR2 <-sum((y2hat-mean(y2hat))^2)
SSR2
#43186.72
chiBP <-(SSR2/2)/(SSE/n)^2
chiBP
#3.482692
chiTAB<-qchisq(0.99,4)
chiTAB
#13.2767
chiTab <-qchisq(0.95,4)
chiTab
#9.4877729

#Predictions
#Interval Estimate of the mean satisfaction when X_h1 = #, X_h2 = #, X_h3 = #, X_h4 = #
#95% Confidence Coefficient
X <- cbind(1, Age, Severity, `Surgical-Medical`, Anxiety)
Y <-matrix(Satisfaction, ncol=1)
XX <-t(X) %*% X
XY <-t(X) %*% Y
b <-solve(XX) %*% XY
Xh <-matrix(c(1,35,45,1, 2.2), nrow=1) #given values of x
EYh <- Xh %*%b
EYh
#84.03949
VarEYh <- (Xh %*% solve(XX) %*% t(Xh)) * MSE
VarEYh
#15.34136
```

```
alpha <- 0.05
tval <- qt(1-alpha/3,20)
tval
#2.285497
c(EYh - tval *sqrt(VarEYh), EYh+tval*sqrt(VarEYh))
#75.08764 92.99133

#Obtain a prediction interval for a new patient's satisfaction when X_h1 = #, X_h2 = #, X_h3 =
        #, X_h4=#
pred <-Xh %*% b
pred
#84.03949
VarPred <- (1+(Xh %*% solve(XX) %*% t(Xh))) *MSE
VarPred
#109.081
c(pred-tval*sqrt(VarPred), pred+tval*sqrt(VarPred))
#60.16933 107.90964

#TRY ELIMINATING VARIABLE SURGICAL MEDICAL
#Remember, for AllInModel:
#R^2 = 0.818
#Ra^2 = 0.782
#MSE = 98.427

#No Surgical-Medical Variable (Model is First Order Linear Multiple Regression, MEs only)
modelNoSurgMed <-lm(Satisfaction~Age+Severity+Anxiety, data = data)

#Obtain Results
summary(modelNoSurgMed)
#Satisfaction = 140.3193 - (1.1233*Age) - (0.4629*Severity) + (1.2126*Anxiety)
#Fo < .05 (significant model)
#Anxiety coefficient nonsignificant
#R^2 = 0.816 (Decreased)
#Ra^2 = 0.789 (Increased)
#MSE = 95.094 (Decreased)
#Confirm with ols
ols_regress(modelNoSurgMed)

#TRY ELIMINATING ANXIETY (RE-INSERT SURGICAL MEDICAL)
#No Anxiety Variable (Model is First Order Linear Multiple Regression, MEs only)
modelNoAnxiety <-lm(Satisfaction~Age+Severity+`Surgical-Medical`, data = data)

#Obtain Results
summary(modelNoAnxiety)
#Satisfaction = 139.7722 - (1.0605*Age) - (0.4410*Severity) + (1.9865*SurgicalMedical)
#Fo < .05 (significant model)
```

```
#SurgicalMedical coefficient nonsignificant
#R^2 = 0.812 (decreased)
#Ra^2 = 0.740 (decreased)
#MSE = 97.120 (Increased)
#Confirm with ols
ols_regress(modelNoAnxiety)

#TRY ELIMINATING SURGICAL MEDICAL & ANXIETY
#No SurgMed No Anxiety Variable (Model is First Order Linear Multiple Regression, MEs
        only)
modelNoSurgMedNoAnxiety <-lm(Satisfaction~Age+Severity, data = data)

#Obtain Results
summary(modelNoSurgMedNoAnxiety)
#Satisfaction = 139.9233 - (1.0462*Age) - (0.4359*Severity)
#Fo < .05 (significant model)
#Both coefficients significant
#R^2 = 0.810 (decreased)
#Ra^2 = 0.792 (Increased)
#MSE = 93.740 (Decreased)
#Confirm with ols
ols_regress(modelNoSurgMedNoAnxiety)

#Test Fit of AllInModel
ols_step_all_possible(AllInModel)
#Best Models:
#Satisfaction ~ Age+Severity
#Ra^2 = 0.792
#Mallow's Cp = 1.95
#Satisfaction ~ Age+Severity+Anxiety
#Ra^2 = 0.789
#Mallow's Cp = 3.29


#Perform All Possible Regressions (for Variable Selection)
#Use leaps() b/c it performs an exhaustive search for the best subsets of the variables in x for
        predicting y in a linear regression
fitAll = regsubsets(Satisfaction~.,data=data, nbest =1, nvmax = NULL, force.in = NULL,
        force.out = NULL, method = 'exhaustive')
summary(fitAll)
fitAllOutput = summary(fitAll)
names(fitAllOutput)
as.data.frame(fitAllOutput$outmat)
View(fitAllOutput)
#Which model has the highest R^2?
which.max(fitAllOutput$rsq)
```

```
#The model with all 4 variables has the highest R^2
#See which model
#Variables marked with TRUE are the chosen ones
fitAllOutput$which[4,]
#Age + Severity + Surgical-Medical + Anxiety

#Plot Best R^2
par(mfrow = c(2,2))
plot(fitAllOutput$rsq, xlab = "Number of Variables", ylab = "R^2", type = "b")
best.r2=which.max(fitAllOutput$rsp)
points(best.r2,fitAllOutput$rsq[best.r2], col = "red", cex = 2, pch =20)

#Which model has the highest Ra^2?
which.max(fitAllOutput$adjr2)
#The model with 2 variables has the highest Ra^2
#See which model
#Variables marked with TRUE are the chosen ones
fitAllOutput$which[4,]
#Variables Age, Severity, Surgical-Medical, & Anxiety are TRUE

#Plot Best Ra^2
plot(fitAllOutput$adjr2, xlab = "Number of Variables", ylab = "Adjusted R^2", type = "b")
best.adjr2=which.max(fitAllOutput$adjr2)
points(best.adjr2,fitAllOutput$adjr2[best.adjr2], col = "red", cex = 2, pch =20)

#Do regression with the best model
best.model <- lm(Satisfaction~Age+Severity, data=data)
summary(best.model)
#p = 1.93e-08; Ra^2 = 0.7923;both significant t-values
#Satisfaction = 139.9233 - (1.0462*Age) - (0.4359*Severity)

#Another way to run leaps()
#Take a design matrix as an argument; throw away the intercept column
xs <- model.matrix(AllInModel)[,-1]
#Look at R^2 of all subsets
r2.leaps <-leaps(xs,data$Satisfaction, nbest=1, method='r2') #nbest = 1 (1 best model for each
        number of predictors)
plot(r2.leaps$size, r2.leaps$r2, pch=23, bg='blue', cex=2)
#Which variable has the highest R^2?
best.model.r2 <-r2.leaps$which[which((r2.leaps$r2 == max(r2.leaps$r2))),]
print(best.model.r2)
#All variables fit

#Adjusted R^2 of all subsets
adjr2.leaps <-leaps(xs,data$Satisfaction, nbest =1, method='adjr2') #nbest = 1 (1 best model for
        each number of predictors)
```

```
plot(adjr2.leaps$size, adjr2.leaps$adjr2, pch=23, bg = 'blue', cex=2)
best.model.adjr2 <-adjr2.leaps$which[which((adjr2.leaps$adjr2 == max(adjr2.leaps$adjr2))),]
print(best.model.adjr2)
#Columns 1 (Age) and 2 (Severity) only

#Validate RMSE
set.seed(12022019)
numVars = ncol(data)-1
trnIdx = sample(c(TRUE, FALSE), nrow(data), rep = TRUE)
tstIdx = (!trnIdx)
fitAllValidateRMSE = regsubsets(Satisfaction~., data=data[trnIdx,], nvmax = numVars)
testMat = model.matrix(Satisfaction~., data=data[tstIdx,])

testError = rep(0, times = numVars)
for (i in seq_along (testError)){
  coefs = coef(fitAllValidateRMSE, id = i)
  pred = testMat[,names(coefs)] %*% coefs
  testError[i] <- sqrt(mean((data$Satisfaction[tstIdx]-pred)^2))
}
testError
#8.904410 6.679043 7.010029 7.001217

#Plot
plot(testError, type = 'b', ylab = "Test Set RMSE", xlab = "Number of Predictors")

#Which variable has the lowest error?
which.min(testError)
#2 variable model has the lowest error

#Determine which 2 variables it is
coef(fitAllValidateRMSE, which.min(testError))
#Age (-1.0260027)
#Severity (-0.3419519)

#Another way to run all possible regressions using olsr:
ols_step_all_possible(AllInModel)
ols_step_all_possible_betas(AllInModel)
ols_step_best_subset(AllInModel)
#Same results as when using leap()

#Try Severity Cubed
corr.test(Satisfaction, Severity)
#[1] -0.65
data$SeverityCubed <-data$Severity*data$Severity*data$Severity
cor.test(Satisfaction, data$SeverityCubed)
#-0.6873331  (improvement)
```

```
corr.test(Age, data$SeverityCubed)
#Age & Severity r = 0.529
#Create model
modelSeverityCubed <-lm(Satisfaction~Age+SeverityCubed+`Surgical-Medical`+Anxiety, data
        = data)
#Obtain Result
summary(modelSeverityCubed)
#Ra^2 = 0.7846
#Compare to AllInModel: Ra^2 = 0.7819
#Improvment
#Lack of Fit
anova(modelSeverityCubed, AllInModel)

ols_test_f(modelSeverityCubed)
#p = 0.8936556

#Try Anxiety Cubed
corr.test(Satisfaction, Anxiety)
#[1] -0.51
cor.test(Satisfaction, data$AnxietyCubed)
#-0.4908339 (not an improvement)
cor.test(Age, data$AnxietyCubed)
#
#Anxiety & Age r = 0.621
data$AnxietyCubed <-data$Anxiety*data$Anxiety*data$Anxiety
modelAnxietyCubed <-lm(Satisfaction~Age+Severity+`Surgical-Medical`+AnxietyCubed,
        data=data)
#Obtain Result
summary(modelAnxietyCubed)
#Ra^2=0.7842
#Compare to AllInModel: Ra^2 = 0.7819
#Improvment
#Improvment
#Lack of Fit
anova(modelAnxietyCubed, AllInModel)

ols_test_f(modelAnxietyCubed)
#p = 0.8936556

#Try Anxiety Cubed AND Severity Cubed
modelAnxietyAndSeverityCubed <-lm(Satisfaction~Age+SeverityCubed+`Surgical-
        Medical`+AnxietyCubed,data=data)
#Obtain Result
summary(modelAnxietyAndSeverityCubed)
#Ra^2 = 0.7841
#Compare to AllInModel: Ra^2 = 0.7819
```

```
#Improvment
#AnxietyCubed Nonsignificant
#Lack of Fit
anova(modelAnxietyAndSeverityCubed, AllInModel)

ols_test_f(modelAnxietyCubed)
#p = 0.7551041

#Confirm coefficients
coefficients(modelAnxietyAndSeverityCubed) #model coefficients
#Intercept: 1.285138e+02
#Age: -1.131516e+00
#SeverityCubed: -6.268721e-05
#SurgicalMedical: 2.823530e+00
#AnxietyCubed: 1.668141e-02
#Confirm with ols package
ols_regress(modelAnxietyAndSeverityCubed)
#Remember, Fc = (SSR/k) / (SSE/n-k+1)
#Alpha = .05
#Fo = F(1-alpha, df, df)
#Fo = F(1-.05,4,20)
qf(0.95,4,20)
#Fo = 2.866081

#Confidence Intervals (95%)
confint(modelAnxietyAndSeverityCubed)
#Age: Small Range, does not include 0
#Severity: Small range, does not include 0
#Surgical-Medical: Large range, includes 0
#Anxiety: Large range, includes 0

#Test Fit of modelAnxietyAndSeverityCubed
ols_step_all_possible(modelAnxietyAndSeverityCubed)
#Best Models:
#Satisfaction ~ Age+SeverityCubed
#Ra^2 = 0.795
#Mallow's Cp = 1.88
#Satisfaction ~ Age+SeverityCubed+AnxietyCubed
#Ra^2 = 0.788
#Mallow's Cp = 3.64
#AllVariables
#Ra^2 = 0.784
#Mallow's Cp 5

#Perform All Possible Regressions (for Variable Selection)
```

```
#Use leaps() b/c it performs an exhaustive search for the best subsets of the variables in x for
        predicting y in a linear regression
fitAllWithCubes = regsubsets(Satisfaction~.,data=data, nbest =1, nvmax = NULL, force.in =
        NULL, force.out = NULL, method = 'exhaustive')
summary(fitAllWithCubes)
fitAllWithCubesOutput = summary(fitAllWithCubes)
names(fitAllWithCubesOutput)
as.data.frame(fitAllWithCubesOutput$outmat)
View(fitAllWithCubesOutput)
#Which model has the highest R^2?
which.max(fitAllWithCubesOutput$rsq)
#The model with all variables has the highest R^2
#See which model
#Variables marked with TRUE are the chosen ones
fitAllWithCubesOutput$which[6,]
#Age + Severity + Surgical-Medical + Anxiety

#Plot Best R^2
par(mfrow = c(2,2))
plot(fitAllWithCubesOutput$rsq, xlab = "Number of Variables", ylab = "R^2", type = "b")
best.r2=which.max(fitAllWithCubesOutput$rsp)
points(best.r2,fitAllWithCubesOutput$rsq[best.r2], col = "red", cex = 2, pch =20)

#Which model has the highest Ra^2?
which.max(fitAllWithCubesOutput$adjr2)
#The model with 2 variables (Age, SeverityCubed) has the highest Ra^2
#See which model
#Variables marked with TRUE are the chosen ones
fitAllWithCubesOutput$which[6,]
#Variables Age, Severity, Surgical-Medical, Anxiety, AnxietyCubed, SeverityCubed are TRUE

#Plot Best Ra^2
plot(fitAllWithCubesOutput$adjr2, xlab = "Number of Variables", ylab = "Adjusted R^2", type
        = "b")
best.adjr2=which.max(fitAllWithCubesOutput$adjr2)
points(best.adjr2,fitAllWithCubesOutput$adjr2[best.adjr2], col = "red", cex = 2, pch =20)

#Do regression with the best model
revisedPolynomialModel <- lm(Satisfaction~Age+SeverityCubed, data=data)
summary(revisedPolynomialModel)
#p = 1.027e-08 (significant); Ra^2 = 0.7951 (higher);both significant t-values
#Satisfaction = 125.957 - (1.018*Age) - (6.256002e-05*Severity)
#Confirm coefficients
coefficients(revisedPolynomialModel) #model coefficients
#Confirm with ols package
ols_regress(revisedPolynomialModel)
```

```
#Confidence Intervals
confint(revisedPolynomialModel)

#Previous Best Model:
#p = 1.93e-08; Ra^2 = 0.7923;both significant t-values
#Satisfaction = 139.9233 - (1.0462*Age) - (0.4359*Severity)

#REDO ALL RESIDUAL ANALYSIS
View(data)

#Plot New Model
plot(data$Age+data$SeverityCubed, data$Satisfaction, data=data)
lines(lowess(data$Age+data$SeverityCubed, Satisfaction))
#Appears linear
#Can also plot with ols library
ols_plot_reg_line(data$Satisfaction,data$Age+data$SeverityCubed)
ols_plot_response(revisedPolynomialModel)

#Obtain Fitted Values
fitted(revisedPolynomialModel) #predicted values

#Obtain regression diagnostics for each observation: y_hat, coefficients, sigma, weighted
        residual
influence(revisedPolynomialModel) #regression diagnostics
#Diagnostic plots
layout(matrix(c(1,2,3,4),2,2)) #optional 4 graphs/page
plot(revisedPolynomialModel)
#Regression Diagnostics Battery
ols_plot_diagnostics(revisedPolynomialModel)

#Obtain residuals
residuals(revisedPolynomialModel) #residuals
e <-residuals(revisedPolynomialModel)
e
boxplot(e, ylab = "Residuals")

#Plot residuals against Y_hat
yhat <- fitted(revisedPolynomialModel)
plot(yhat, e, xlab = "Fitted Values", ylab = "Residuals", ylim = c(-20,20))
abline(h=0, lty = 2)

#Plot residuals against each of the predictor variables
#Age Residuals
plot(Age,e, xlab="Age", ylab = "Residuals", ylim = c(-20,20))
abline(h=0, lty=2)
#Severity Residuals
```

```
plot(data$SeverityCubed, e, xlab = "SeverityCubed", ylab = "Residuals", ylim = c(-20,20))
abline(h=0, lty=2)

#3d Scatterplot
scatterplot3d(data$Age, data$SeverityCubed, e, xlab = "Age", ylab = "SeverityCubed", zlab =
        "Residuals")

#More Residuals
summary(revisedPolynomialModel)$residuals
ols_plot_resid_stud(revisedPolynomialModel)
#Residual Normality Test
ols_test_normality(revisedPolynomialModel)
#Correlation Between Observed Residuals and Expected Residuals Under Normality
ols_test_correlation(revisedPolynomialModel)
#Correlation = 0.9452152
#Residual vs Fitted Values Plot
ols_plot_resid_fit(revisedPolynomialModel)
#Residual Histogram
ols_plot_resid_hist(revisedPolynomialModel)
#Studentized Residuals vs Leverage Plot
ols_plot_resid_lev(revisedPolynomialModel)
#Deleted Studentized residual vs predicted values
ols_plot_resid_stud_fit(revisedPolynomialModel)
#Residuals
r<-resid(revisedPolynomialModel)
r
#QQ Plot
qqnorm(resid(revisedPolynomialModel))
ols_plot_resid_qq(revisedPolynomialModel)
#Predictively adjusted residuals
(pr<-resid(revisedPolynomialModel)/(1-lm.influence(revisedPolynomialModel)$hat))
#Construct residual vs predicted response
plot(predict(revisedPolynomialModel), resid(revisedPolynomialModel))
#Cross-validated residuals
#Regular RSS is
sum(r^2)
#2034.492
#PRESS is
sum(pr^2)
#2583.56
#Note PRESS is bigger because predicting is harder than fitting
#Another way to calculate the PRESS statistic
PRESS<-function(revisedPolynomialModel){pr<-residuals(revisedPolynomialModel)/(1-
        lm.influence(revisedPolynomialModel$hat)
                sum(pr^2)}
PRESS(revisedPolynomialModel)
```

```
#Same result: 2583.56

#standardized residuals
ols_plot_resid_stand(revisedPolynomialModel)

#Normal Probability Plot of Residuals
n <-length(e)
MSE <-sum(e^2)/(n-4)
RankofRes <-rank(e)
Zscore <- qnorm((RankofRes-0.375)/(n+0.25))
ExpRes <- Zscore * sqrt(MSE)
plot(ExpRes, e, xlab = "Expected Score", ylab = "Residuals")
abline(a = 0, b = 1)

#Detect Influence with Leverage
#The observed value of y_i is influential if h_i > [2(k+1)]/n
#Where h_i = leverage for the ith obervation
#k = # of betas in the model (excluding b_0)
# [2(2+1)]/n = [2(3)/25] = 6/25 = 0.5
ols_leverage(revisedPolynomialModel)

# VIF = (1/1-R^2). VIF > 5 indicates associated regression coefficients are poorly estimated b/c
        multicollinearity
ols_vif_tol(revisedPolynomialModel)
#Result: All VIFs < 5
#Age Tolderance = 0.675
#Age VIF = 1.48
#SeverityCubed Tolerance = 0.675
#SeverityCubed VIF = 1.48

#Check for Collinearity
ols_coll_diag(revisedPolynomialModel)
vcov(revisedPolynomialModel) #covariance matrix for model parameters

#Cook's Distance: Combines leverage and residuals
#Higher value, the better
#Lowest Value = 0
#Conventional Cut off is 4/n
ols_plot_cooksd_bar(revisedPolynomialModel)
ols_plot_cooksd_chart(revisedPolynomialModel)
cooks.distance(revisedPolynomialModel)

#dfbetas:measures the difference in each parameter estimate with and without the influential
        point
ols_plot_dfbetas(revisedPolynomialModel)
dfbeta(revisedPolynomialModel)
```

```
#lack of fit
ols_test_f(revisedPolynomialModel)
#Fail to reject H0; p = 0.7719677
```