Stat 21 Class 20 Practice Problem

Stat 021 F 2020 Swarthmore College

Probem 1

(This is Problem 3.17 in your textbook).

Consider the following patient satisfaction data. Perform a thorough analysis of these data where we will consider the response variable `Satisfaction` and the predictor variables `Age`, `Severity`, and `Anxiety`.

```
library('tidyverse')
patient_satisfaction <- read_table2(
  url("http://www.swarthmore.edu/NatSci/sthornt1/DataFiles/patient_satisfaction.txt")) %>%
  select(-c(SurgicalMedical,X6))
  ## This last step just gets ride of two columns we aren't going to use in our analysis.
head(patient_satisfaction)
```

```
## # A tibble: 6 x 4
##   Satisfaction   Age Severity Anxiety
##          <dbl> <dbl>    <dbl>   <dbl>
## 1           68    55       50     2.1
## 2           77    46       24     2.8
## 3           96    30       46     3.3
## 4           80    35       48     4.5
## 5           43    59       58     2
## 6           44    61       60     5.1
```

Step 1: Make sure we understand the variables and that they're in the correct form.

We were provided with the data set but not much context for this data therefore, before any analysis of the relationships among these variables, we need to first make sure that we understand each of variables in our data set.

First, ask yourself what are the meanings of the numeric values for `Satisfaction`, `Severity`, and `Anxiety`? Are they all measured on the same scale? If not, is that something we can fix? Also, is there a good reason to think that `Severity`, `Anxiety`, and `Age` are related to `Satisfaction`? (This seems obvious but you'd be suprised how often this question is ignored and instead researchers jump straight to the individual t-test p-values.)

A quick way to get a general picture of quantitative variables is to look at a summary for each of the variables.

```
patient_satisfaction %>% summary
```

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

From the summary we can see that there appears to be a mistake in the data since there is one observation where `Satifaction` is higher than 100. Because this outlier is likely due to a mistake in transcribing the data and because we can't reach out to whoever transcribed the data to check, we are going to eliminate this data point from our analysis using the following code.

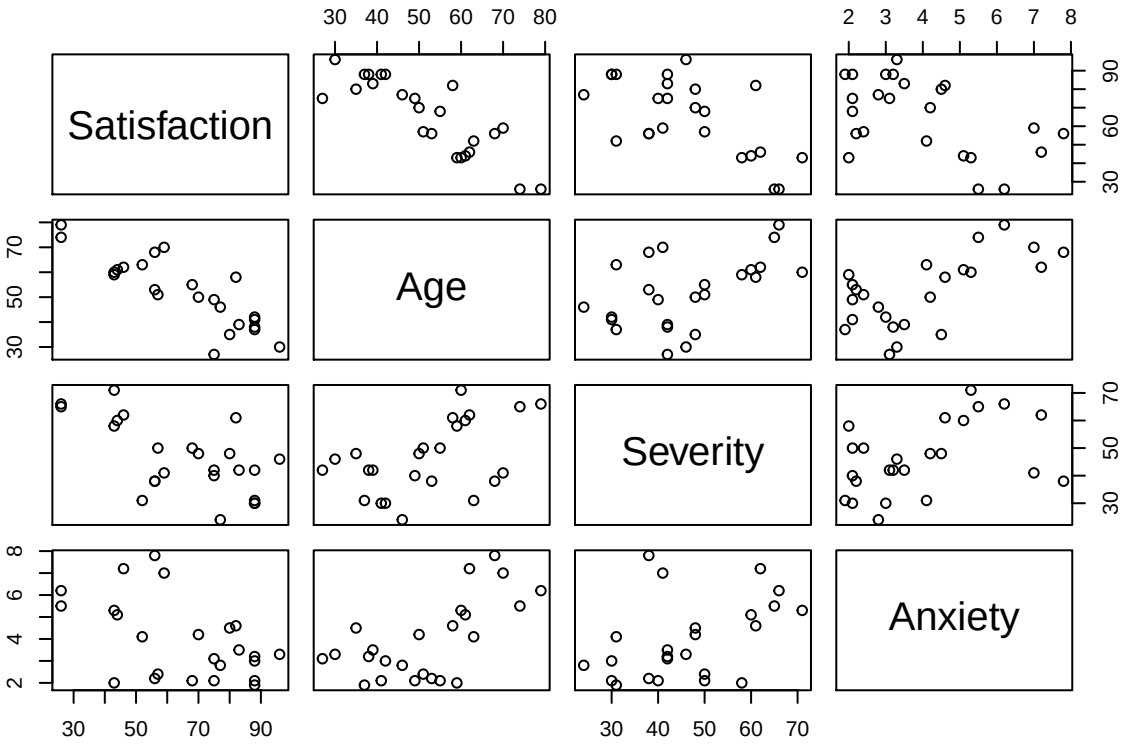
```
patient_satisfaction <- patient_satisfaction %>% filter(Satisfaction<=100)
```

Also, make sure that any categorical predictor variables are of variable class "factor" in R. (This does not apply in this problem since all predictors are quantitative.)

Step 2: Visualize the data.

Now that we are certain we understand each of the variables in our data set and have gotten rid of any severly problematic data points (i.e. data points that appear to be a result of human error or data points with missing entries), we move to the next step of our analysis and visualize the data. Since we are building a MLR model with only quantitative predictors, we can create a matrix scatterplot for all the variables at once.

```
patient_satisfaction %>% pairs
```



If we don't see a linear relationship between each of the pairwise plots of the predictors and the response that is not a problem! We are no longer interested in pairwise linear relationship, rather we want to explore a multidimensional linear relationship among several predictors grouped together and the response. We do want to look out for any noteworthy trends, get a sense of the range of predictor values we observe (to understand how to look out for extrapolation), and see if we notice any pairwise linear trends among the predictor variables (as evidence of multicollinearity).

Step 3: Fit the estimated regression model.

Now that we have a better understanding of our data set, we can go ahead and fit the estimated regression model. At this step we won't yet consider any of the individual t-test results in the model summary. For now, let's just get estimates for all of the model parameters and consider the model fit (for this we can look at both the adjusted R-squared value and the p-value for the F-test of overall fit).

```r
mod1 <- lm(Satisfaction ~ Age + Severity + Anxiety, data=patient_satisfaction)
summary(mod1)
```

```
##
## Call:
## lm(formula = Satisfaction ~ Age + Severity + Anxiety, data = patient_satisfaction)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -19.163  -5.899  -1.201   5.306  29.496
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 140.0820     9.0378  15.500 1.31e-12 ***
## Age          -1.1180     0.2032  -5.501 2.20e-05 ***
## Severity     -0.4626     0.1875  -2.467   0.0228 *
## Anxiety       1.1925     1.5267   0.781   0.4439
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 9.991 on 20 degrees of freedom
## Multiple R-squared:  0.7906, Adjusted R-squared:  0.7592
## F-statistic: 25.17 on 3 and 20 DF,  p-value: 5.396e-07
```

Based on the output above, the fitted regression equation is

$$\hat{y} = 140 - 1.1x_1 - 0.5x_2 + 1.2x_3, \quad \hat{\sigma} = 10.$$

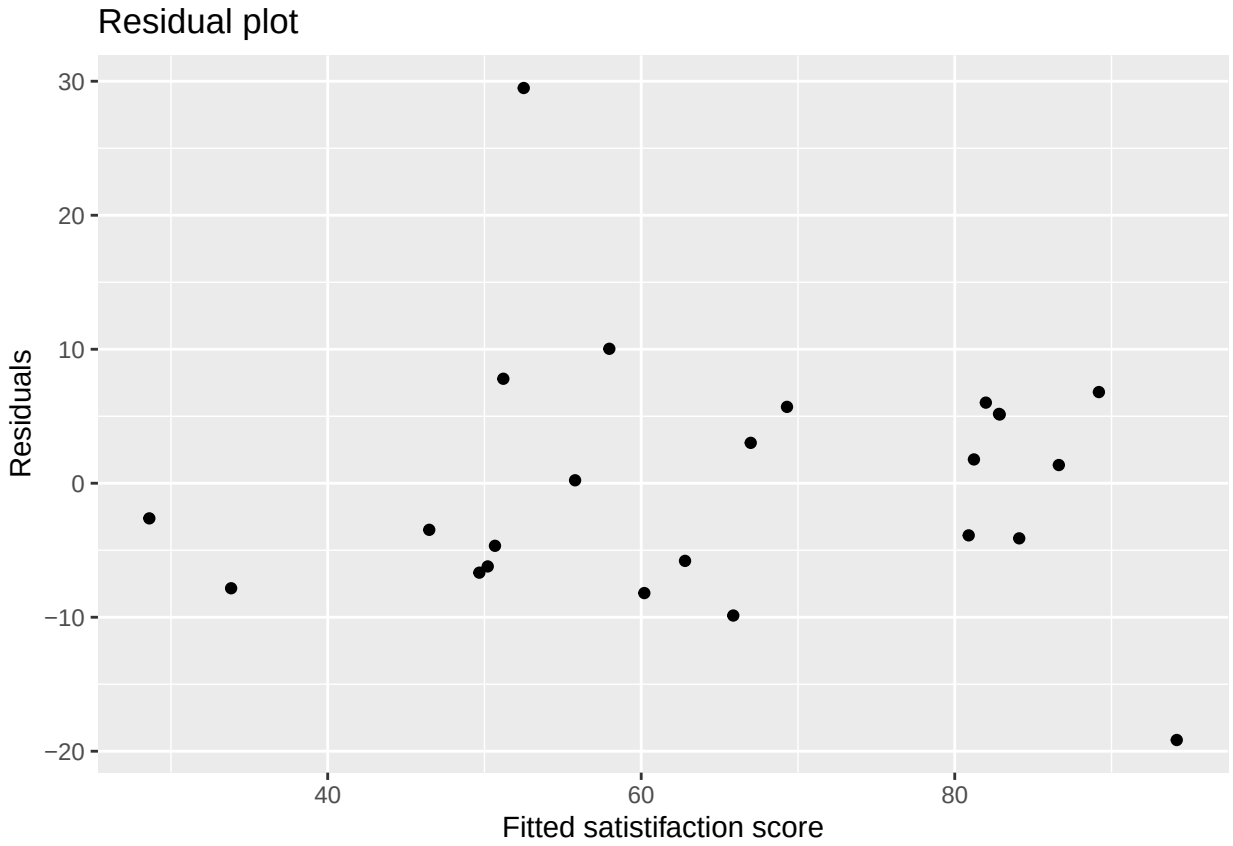
At this step it is a good idea to stop and think about the interpretation of each of the model parameters. Think about questions such as: What are the units of $\hat{\sigma}$? Do the signs of each of the $\hat{\beta}$ coefficients make sense in this context? If not, how might we address that? Which predictor has the largest effect size? Can we tell or do we need to transform the data? (Again, for now answer these questions without considering the results of the individual t-tests.)

Step 4: Residual analysis and checking the necessary assumptions.

While all of the previous steps are important, this is the step that informs us about how useful this MLR model will be to us. First and foremost, think about the independent errors assumption. If the random error is independent, then the values of our response are independent. Do we have any reason to suspect that the satisfaction scores for one patient in the study are related to the satisfaction scores of another patient in the study? Maybe, if the patients reside in the same household or neighborhood. It's more likely however that this data pertains to an arbitrary set of patients who only share their illness and access to the doctor in common. So there's no strong reason to suspect that the satisfaction scores of these patients are dependent. Also worth mentioning is that, although the patients are probably not randomly assigned (they are seeing the doctor for a particular reason not because the experimenter told them to), they may still be representative of the population provided the researcher took care to collect data for a diverse group of patients. (Also - what is the population of interest?)

Now, let's take a look at the residual plot based on our estimated regression model from Step 3.

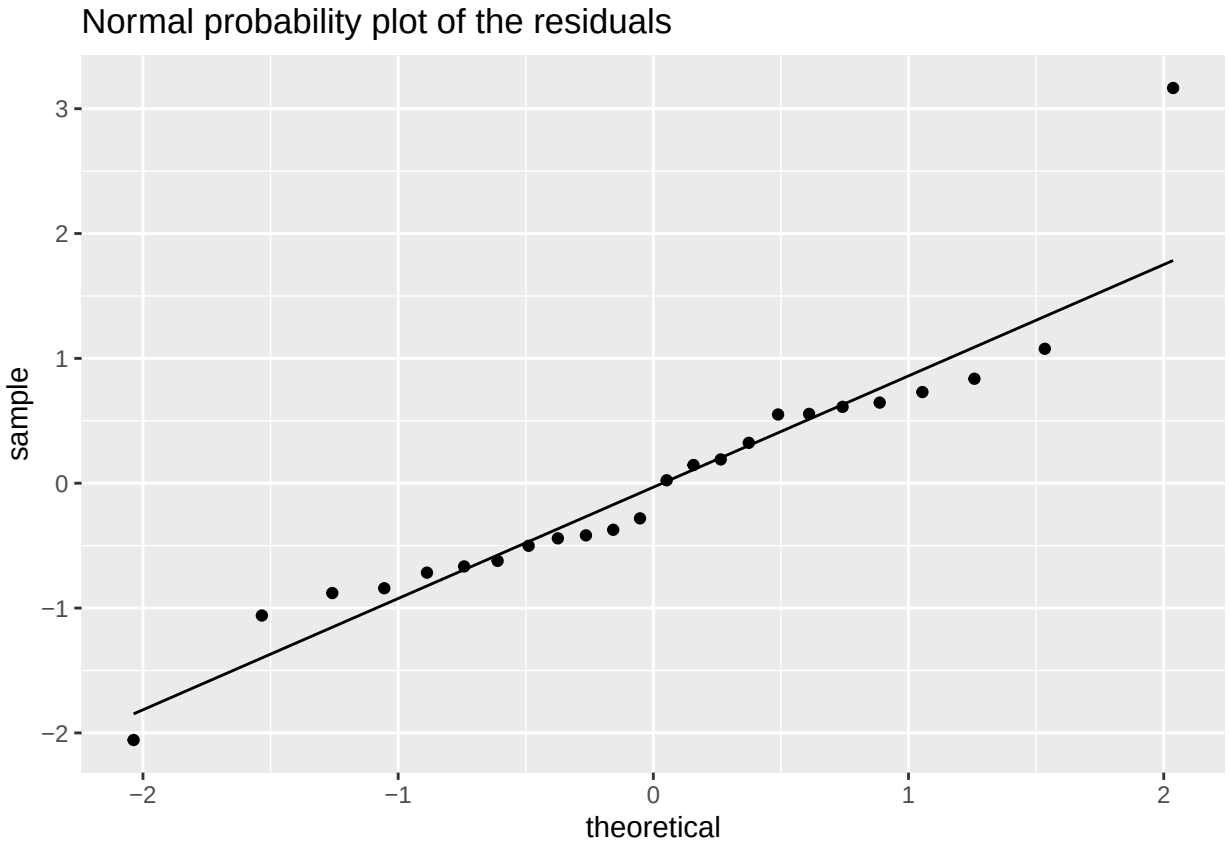
```r
satifaction2 <- patient_satisfaction %>% mutate(res = residuals(mod1),
                                                fit = fitted.values(mod1))
ggplot(satifaction2, aes(x=fit, y=res)) +
  geom_point() +
  labs(title="Residual plot", x="Fitted satistifaction score", y="Residuals")
```



The things we want to look out for here are: evidence of heteroscedasticity and evidence of a non-linear relationship. The first will better inform us as to whether or not we can assume the random noise has a constant (unchanging) variance. The second will help us determine if a linear model is an appropriate fit for this data or if we should try transforming the data (the predictors and/or the response) or perhaps if we should try fitting another model altogether.

If we want to examine the relative significance of each of the predictor variables using the individual t-tests or if we want to calculate a confidence interval for the average response or a prediction interval for an unobserved response, then we also need to check if the assumption that the random error is normally distributed is reasonable. We can do this by looking at a normal probably plot of the standardized residuals to see how well the quantiles of the residuals line up with the quantiles of a standard normal distribution.

```
ggplot(satifaction2, aes(sample=scale(res))) +
  ## I used the scale function here instead of typing out: (res - mean(res))/sd(res)
  stat_qq() +
  stat_qq_line() +
  labs(title="Normal probability plot of the residuals")
```



Additionally we need to consider any evidence of multicollinearity among the predictors. If there was evidence of multicollinearity in Step 2 then we should repeat Step 3 and 4 with fewer predictor variables to try to eliminate any redundant information contained in the predictors. This is not necessary however if we do not want to consider any inferential questions.

Step 5: Evaluate the model and draw inferential conclusions

Proceed by either making necessary adjustments (such as variable transformation or collecting more data or including interaction terms) and then repeat Steps 3 and 4, or proceed to carefully analyse the regression model. Now you can consider the relative *statistical significance* of each of the predictor variables. Remember that these p-values correspond to testing whether or not the β coefficient for a particular predictor is equal to zero or not. This does NOT mean that we are testing whether or not each predictor is a significant predictor of the response in and of itself, rather, it means we are testing whether or not each predictor is significant given that all the other predictors are included in the model.

For more practice: Read the analyses of this data set as outlined in Sections 2.7 and 3.6 of your textbook. What are some differences between the analyses in your book and the one we did her?