Homework #1

Dustin Leatherman

January 12, 2019

```
knitr::opts_chunk$set(echo = TRUE)
library(ggplot2)
library(kableExtra)
library(dplyr)
library(grid)
library(gridExtra)
library(tidyr)
library(broom)

# Create residual plots for a dataframe containing an x and y variable.
# returns the created model in the event the caller wishes to do further analysis
plotSummary <- function (df) {
  model <- lm(y~x, data=df)

  # Scatterplot with Regression Model
  df.splot <-
    ggplot(model, aes(x = x, y = y)) +
      geom_point() +
      geom_smooth(method = "lm") +
      xlab("X") +
      ylab("Y") +
      ggtitle("Scatterplot with Regression") +
        theme(plot.title = element_text(hjust = 0.5))

  # Residual vs fitted values
  model.resfit <-
    ggplot(model, aes(x = .fitted, y = .resid)) +
      geom_point() +
      geom_hline(yintercept = 0, lty = 2) +
      geom_smooth() +
      ggtitle("Residual vs Fitted") +
        theme(plot.title = element_text(hjust = 0.5)) +
      xlab("Fitted Values") +
      ylab("Residuals")

  # Residual vs predictor (x)
  model.predict <- ggplot(model, aes(x = x, y = .resid)) +
    geom_point() +
    geom_hline(yintercept = 0, lty = 2) +
    geom_smooth() +
    ggtitle("Residual vs Predictor (X)") +
      theme(plot.title = element_text(hjust = 0.5)) +
    xlab("X") +
    ylab("Residuals")

  # Normality plot against residuals
  model.qq <- ggplot(aes(sample = .resid), data = model) +
```

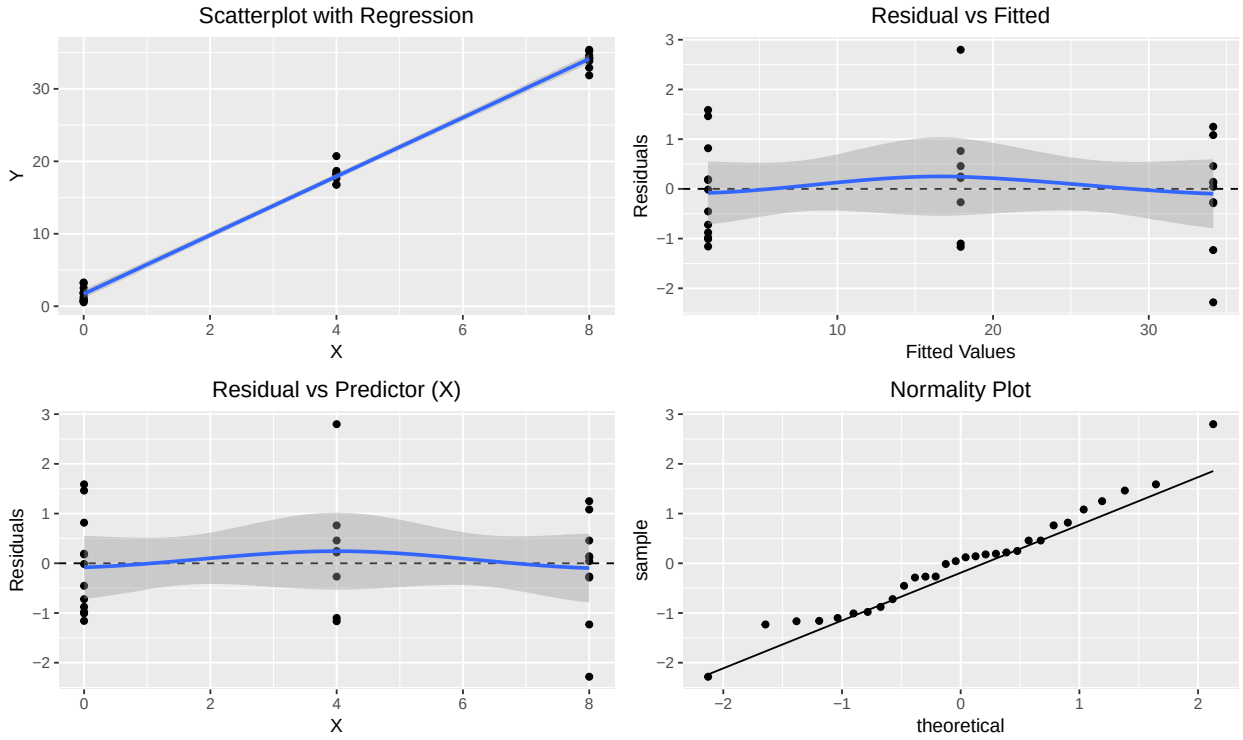
```
    stat_qq() +
    stat_qq_line() +
    ggtitle("Normality Plot") +
      theme(plot.title = element_text(hjust = 0.5))

  grid.arrange(df.splot, model.resfit, model.predict, model.qq,
               widths = c(1,1),
               ncol = 2)
  return (model)
}
```

Question 1

Introduction

Each dataset consists of an explanatory variable (x) and a response variable (y). There are no particular references attached to these datasets so each can be considered depersonalized. With depersonalized datasets, the analyst is able to extract quantify the relationship between the explanatory variable(s) and the response variable(s) through statistical analysis but remain unaware of the implications of their analysis. This type of dataset is common in industries that regularly deal with sensitive data such as the Financial or Healthcare industries.

Dataset #1



##

```
##  Shapiro-Wilk normality test
##
## data:  model1$residuals
## W = 0.97412, p-value = 0.6567
```

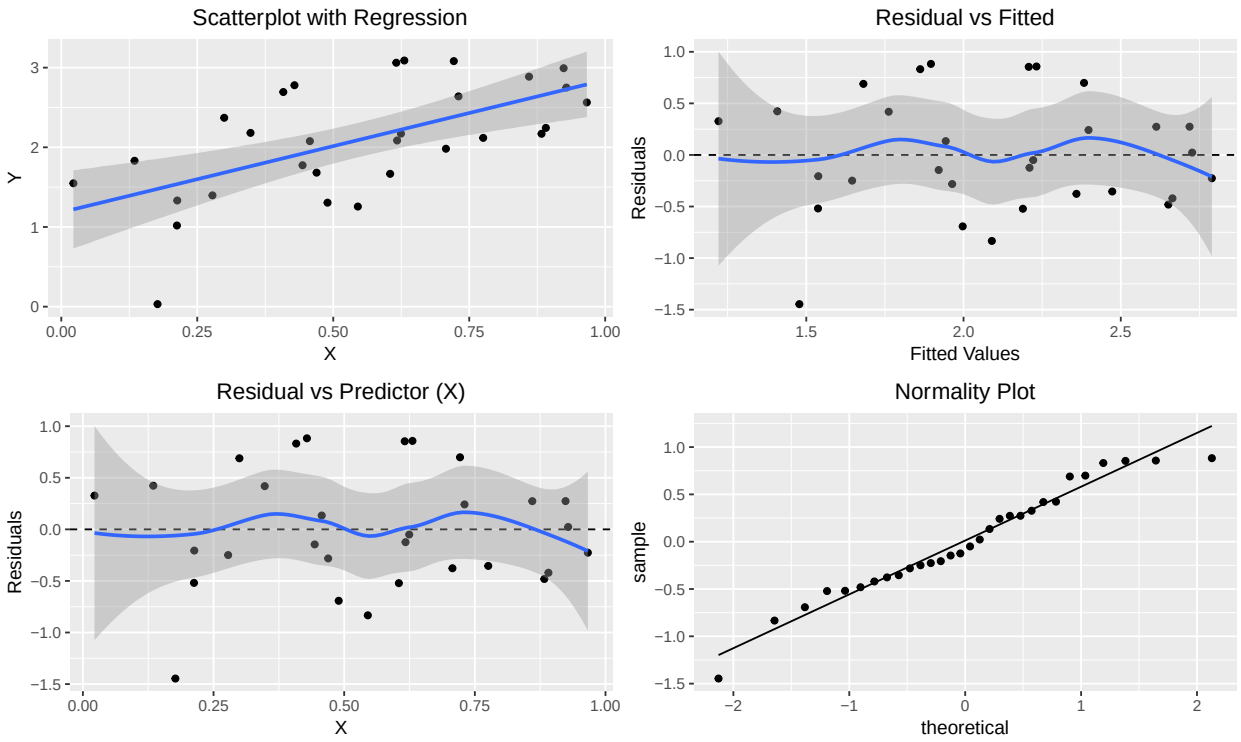
```
##                2.5 %   97.5 %
## (Intercept) 1.110900 2.285148
## x           3.939661 4.171742
```

```
##
## Call:
## lm(formula = y ~ x, data = df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.28248 -0.83863  0.08136  0.45802  2.79821
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.69802    0.28662   5.924 2.24e-06 ***
## x            4.05570    0.05665  71.593  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.06 on 28 degrees of freedom
## Multiple R-squared:  0.9946, Adjusted R-squared:  0.9944
## F-statistic:  5126 on 1 and 28 DF,  p-value: < 2.2e-16
```

The first dataset clearly has a categorical explanatory variable since each variable is associated with more than one response. The Normality Plot for this model appears suspect but the p-value of the Shapiro-Wilk test indicates that there is not enough evidence to suggest that it is non-normal. The residual plots appear symmetric about the gression line and the spread is equal. The line at y = 0 remains within the confidence band which indicates the spread of residuals is fairly even around the regression line. The R^2 value is close to 1 which indicates that 99.44% of the data is explained by this model. This provides a level of confidence that this model is accurate.

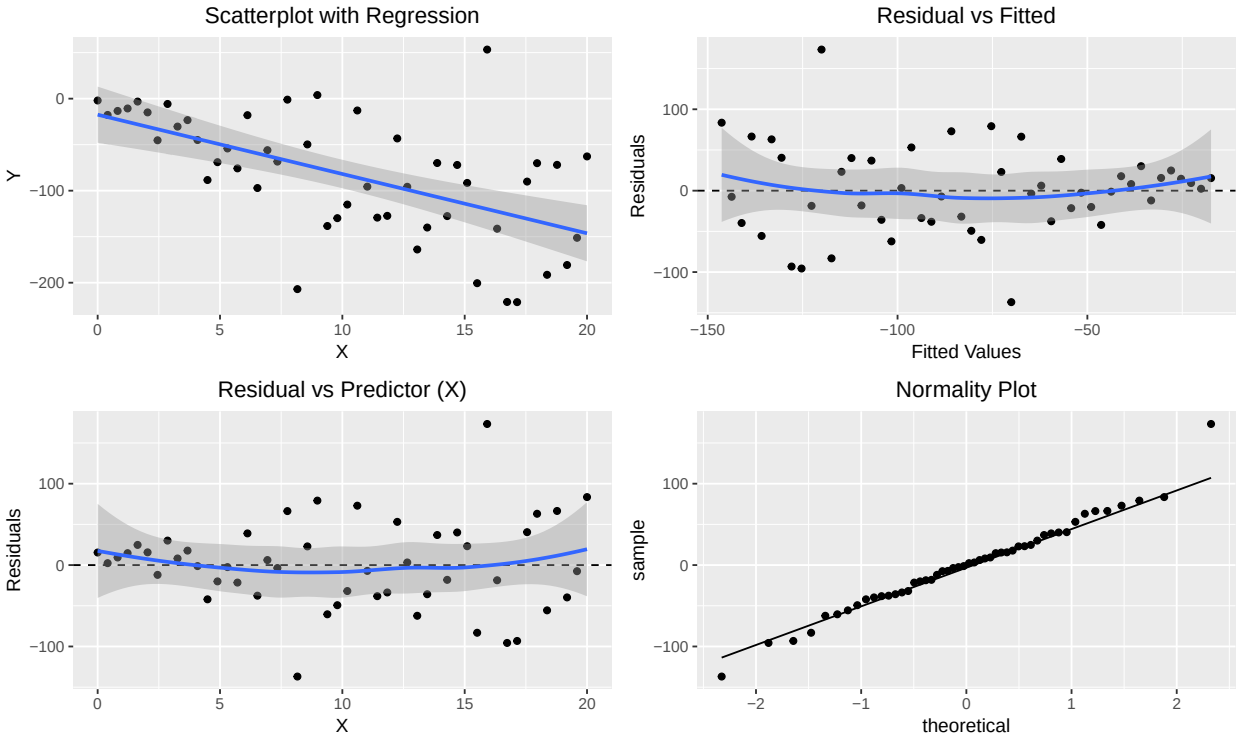
There is convincing evidence that there is a relationship between the explanatory and response variable (p-value $<$ 2e-16). It is estimated that for each increase in the explanatory variable, there is an associated increase in average response value by 4.0557 units. With 95% confidence, for every increase in the explanatory variable, there is an associated increase in the mean response variable by between 3.94 and 4.172 units.

Dataset #2



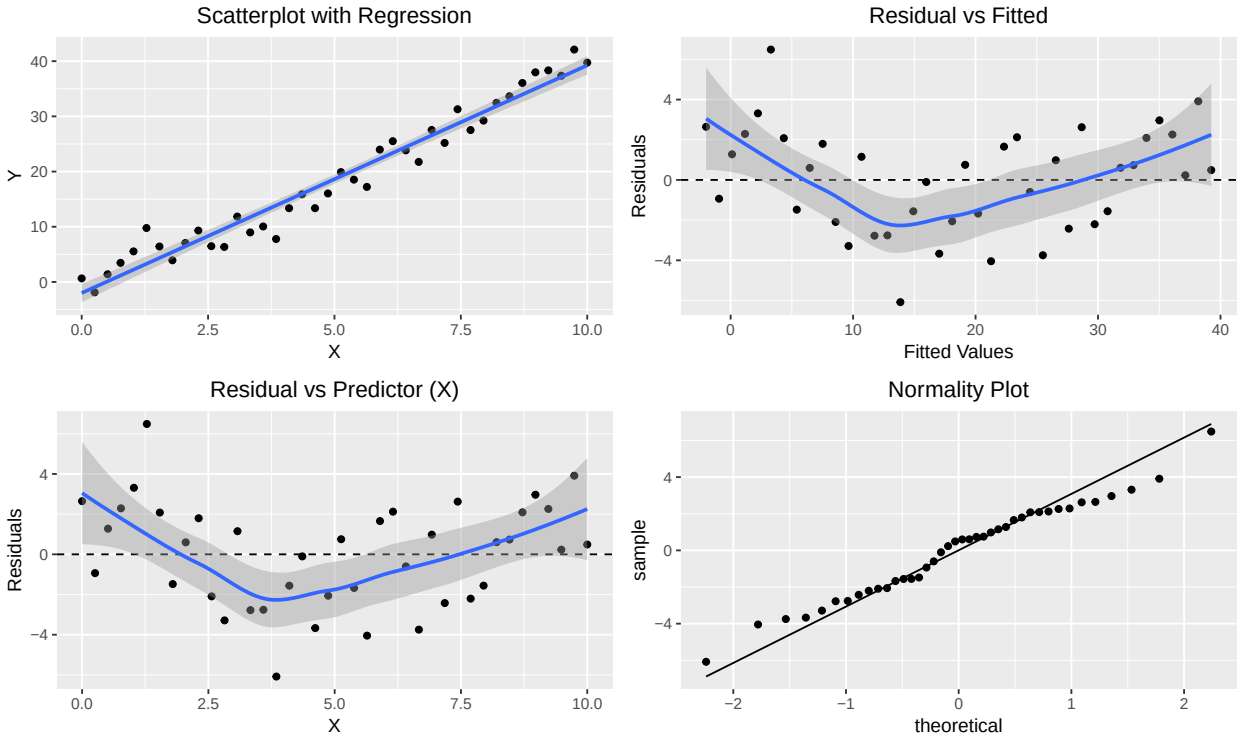
The residul plots are not quite symmetric about the regression line. There is fanning in the tail ends of the confidence bands in the residual plots and the stronger deviation from normality in the tail ends of the normality plot which indicate that the spread of the tails are wider than the other parts of the model. This indicates that this model does not work well at extreme values due to lack of data for extreme values of X. Other than the tails, the normality plot indicates that the data is normal. The uneven spread violates the equal variance assumption and the lack of symmetry violates the Linearity Assumption, both of which are required for Linear Regression to be accurate. As such, a Simple Linear Regression model is not appropriate.

Dataset #3



The residual plots are not symmetric about the regression line indicating the data is not linear. There is also uneven spread about the regression line which indicates that the variances are not equal. The normality plot indicates that this data is normal barring the most extreme values on either tail. The response value contains negative values which means it is not a candidate for a log transformation. Because of the violation of both A Simple Linear Regression Model is not appropriate.

Dataset #4



statistic	p.value	method
0.9817055	0.7524294	Shapiro-Wilk normality test

The residual plots are not symmetric about the regression line thus violating the Linearity assumption. This is readily apparent because the regression line lies outside the confidence interval at times in the residual plots. The spread seems to be fairly consistent so the Equal Spread assumption is met. The normality plot appears to be close for values near the median but drift closer to the tails. The Shapiro-Wilk test indicates that there is not enough evidence to indicate that the data is non-normal so it can be said that the Normality assumption is met. Because the data is not linear, a Simple Linear Regression model is not appropriate.

Question #2

a. Culturally transmitted song exchange between humpback whales (Megaptera novaeangliae) in the southeast Atlantic and southwest Indian Ocean basins

b. The article's headline vaguely implies a generalization about songs and the culture of humpback whales in the Southern Hemisphere.

c. The headline is accurate and I personally appreciate that the headline doesn't give it all away. The study followed pitches of humback whales trying to find similarities between various groups throughout the Southern Oceans. The hypothesis is that groups of humpback whales pass on songs that they have heard to other whales. The conclusion suggests how that these different groups of humpback whales intermingle since they have heard similar songs around the southern oceans.