1. In last week’s homework you used regression to analyze the Sparrows data. The goal was to examine whether a linear model was appropriate for using a bird’s wing length to predict the weight of the bird. Continuing this analysis, create the plots listed in parts (a) to (d), and for each plot explain the following:
   i. Which condition(s) for linear regression can you check with this plot?
   ii. Based on the plot is it reasonable to assume that the condition(s) is (or are) satisfied? Explain your reasoning.

Lecture 3 covered the conditions that can be checked with various plots, including assumptions and questions. Rather than writing them out for each of the parts a to d, they are listed here and will be referenced by letter/number below. Note that all of the assumptions are about the population, but we check them using plots of sample data because we don’t have the population. Here is the list:
   A1: There is a linear relationship between X and Y
   A2: The error terms have constant variance.
   [A3, independence of the errors, cannot be checked with our plots]
   A4: The error terms are normally distributed.
   Q1: Are there major outliers?
   Q2: Are there other possible predictors that should be included in the model?

a. Scatter plot of residuals versus fitted values (See plot 1 at end.)
   i. The conditions that can be checked with this plot are A1, A2, Q1
   ii. A1 and A2 appear to be satisfied, and there are no major outliers (Q1).

b. Q-Q plot (normal probability plot) of the standardized residuals. (See plot 2 at end.)
   i. A4 can be checked with this plot.
   ii. It is reasonable to assume that the normality condition is satisfied, because the points are close to a straight line.

c. Stemplot of the standardized residuals. (See plot 3 at end.)
   i. A4 can be checked and Q1 can be answered.
   ii. Normality appears reasonable, and there are no major outliers. The cases with standardized residuals above 2.0 and below -2.0 could be investigated, but we expect about 5% of the data to fall beyond 2.0 and -2.0, so with 116 data points we would expect at least 5 or 6 to be beyond 2 and -2, and there are 5 of them here.

d. Scatter plot of Y = Weight vs. X = Wing Length using a different color for each treatment category (control, reduced, enlarged) with a separate regression line for each treatment category also shown in the plot. This plot should be created using the ggplot2 package, covered on Fri, Oct 13. (Note that it’s okay if you need to print your plot in black and white; we will be able to see whether you did it correctly.) (See plot 4 at end.)
   i. Any scatterplot of X vs Y allows us to check A1 and A2, but in addition this one allows us to check whether Treatment Group could help by being added as an explanatory variable, addressing Q2.
   ii. It does appear that adding Treatment Group might help because they all have different lines.
2. Carry out the 5 steps of a hypothesis test for the slope to determine if the data provide convincing evidence of a linear relationship between weight and wing length in the population of sparrows similar to these. You may refer to your plots from Exercise 1 for Step 2.

The 5 steps, underlined below, were shown in Lecture 4.

Step 1: Determine the null and alternative hypotheses. The hypotheses are \( H_0: \beta_1 = 0 \) versus \( H_a: \beta_1 \neq 0 \).

Step 2: Verify necessary data conditions, and if met, summarize the data into a test statistic. Checking conditions was done in Exercise 1. Conditions appear to be satisfied. The test statistic is part of the R output. The appropriate part of the R output is shown below. The test statistic is \( t = 13.463 \). You could use the \( F \) statistic instead; from output not shown, \( F = 181.3 \).

| Coefficients:   | Estimate | Std. Error | t value | Pr(>|t|) |
|-----------------|----------|------------|---------|----------|
| (Intercept)     | 1.36549  | 0.95731    | 1.426   | 0.156    |
| WingLength      | 0.46740  | 0.03472    | 13.463  | <2e-16   *** |

Step 3: Assuming the null hypothesis is true, find the p-value. The p-value is given in the R output above as \(<2e-16\). This is essentially 0.

Step 4: Decide whether or not the result is statistically significant based on the p-value. Since the p-value is essentially 0, which is less than any reasonable significance level, reject the null hypothesis and conclude that \( \beta_1 \neq 0 \). In other words, the result is statistically significant.

Step 5: Report the conclusion in the context of the situation. There is a statistically significant linear relationship between wing length and weight for sparrows like the ones in this study. In other words, we conclude that there is a linear relationship between wing length and weight in the population of sparrows similar to these.

3. For this problem, you will use the highway sign data and the Rossman/Chance applet for guessing and viewing a regression line, both linked to the class website (and provided below as well).

a. Copy and paste the highway sign data into the data box at the applet website, removing the data that is there when you open the applet. Check the box “Show regression line” and write down the equation it provides, in the form shown.

Here is how it looks. \( \text{response}^\wedge = 576.68 + -3.01 \times \text{explanatory} \)

b. Remove the point that looked like a slight outlier when the example was discussed in class, with a standardized residual of 2.3. (You can remove it by just deleting it in the data box. Make sure you click on “Use Data” after you remove it, or it will use the original data.) Now check the box “Show movable line.” Move the line until you think you have found the right place for the regression line. Write down the equation for the line you have placed. (This answer will differ for each student, but must be plausible to get credit.) Now check the box “Show regression line” and write the equation it gives. How well did you do in guessing where the line should go?

http://www.ics.uci.edu/~jutts/110/HighwaySign.txt
http://www.rossmanchance.com/applets/RegShuffle.htm

The point to remove is the one with age of 75 and distance of 460. The estimated equation will differ for each student. The actual equation is: \( \text{response}^\wedge = 583.23 + -3.21 \times \text{explanatory} \), so you should have gotten something close to that.

4. Do Exercise 1.26 in the textbook with the following modifications. Do parts (a) and (b) as stated.

a. The scatterplot (shown at the end as Plot 5) indicates a linear trend, with price increasing as number of pages increases.
Part (c) asks for “relevant residual plots.” For that part, create and discuss the following plots:

(i) Residuals vs X. (The plot of residuals vs $X = \text{pages}$ is shown at the end as Plot 6.)

   It basically looks okay except for possible increasing variance.

(ii) Normal probability plot of residuals (Shown at the end as plot 7.)

   The normal probability plot shows some problems at the two ends, indicating that the “tails” of
   the residual distribution may be too wide.

(iii) Stemplot of residuals (Shown at the end as plot 8.)

   The stemplot shows some possible outliers, especially the one at the upper end, which will show
   as either 72 or 73, depending on how you did the plot.

Part (d) [NOT in book] If you did the plots correctly, you should identify an outlier with a residual
of $72. From the list of data shown in the book (or by examining the TextPrices data set some
other way), identify which book (Pages and Price) the outlier is associated with. (Hint: You
should be able to figure this out from one of the plots and the list of data.) Discuss which of the three
reasons for outliers given in class you think might apply in this case. If you don’t think it’s clear
which reason applies, discuss all of the reasons that could apply, and why that reason might apply.

The residual is for a book with 400 pages and a price of $128.50. What counts for credit on this
part is your reasoning, because you could reasonably argue for any of the 3 reasons. You only
need to cover one of them. Here are some arguments:

1. It could be a mistake. Perhaps the price was $28.50, not $128.50.
2. It could belong to a different population. Perhaps it’s a science book with lots of glossy color
   photos, thus making it very expensive, but the other books are all standard either black and white
   or four color (a common format for textbooks).
3. It could (and probably does) just represent natural variability in textbook prices due to things
   like quantity sold, quality of the pages, and so on.

PLOTS AND R CODE

Before creating any of the plots, here is the R code and partial results for running the model:

```r
> library(Stat2Data)
> data(Sparrows)
> HW2Mod <- lm(Weight ~ WingLength, data = Sparrows)
> summary(HW2Mod)

Call:
  lm(formula = Weight ~ WingLength, data = Sparrows)

Residuals:
    Min      1Q  Median      3Q     Max
-3.5440 -0.9935  0.0809  1.0559  3.4168

Coefficients:
                Estimate Std. Error t value  Pr(>|t|)
(Intercept) 1.36549    0.95731   1.426   0.156
WingLength  0.46740    0.03472  13.463 <2e-16 ***
```
Residual standard error: 1.4 on 114 degrees of freedom
Multiple R-squared: 0.6139, Adjusted R-squared: 0.6105
F-statistic: 181.3 on 1 and 114 DF, p-value: < 2.2e-16

Plot 1, Exercise 1a.

\[
\begin{align*}
\text{> plot(resid(HW2Mod) ~ fitted(HW2Mod))} \\
\text{> abline(h=0)}
\end{align*}
\]

![Plot 1](image1)

Plot 2, Exercise 1b.

\[
\begin{align*}
\text{> qqnorm(rstandard(HW2Mod))} \\
\text{> qqline(rstandard(HW2Mod))}
\end{align*}
\]

![Plot 2](image2)

Plot 3, Exercise 1c

\[
\begin{align*}
\text{> stem(rstandard(HW2Mod))}
\end{align*}
\]

The decimal point is at the |

-2  |  7
-2  |  32
-1  |  988665
-1  |  443222100
-0  |  9999888777776666655
-0  |  444333322111111000
  0  |  011122222233334444
  0  |  55555666678888889999
  1  |  000112344
  1  |  5666688
  2  |  3
  2  |  5
Plot 4, Exercise 1d

```r
> ggplot(Sparrows, aes(x = WingLength, y = Weight, color = Treatment)) +
  + geom_point() + geom_smooth(method = "lm", formula = y ~ x, se = F) +
  + ggtitle("Scatterplot: Weight versus Wing Length") +
  + xlab("Wing Length (in mm)") +
  + ylab("Weight (in grams") +
  + scale_color_discrete("Treatment")
```

Plot 5, Exercise 4a (Note that you do not need to include the line.)

```r
> PriceModel <- lm(Price~Pages, data=TextPrices)
> plot(TextPrices$Pages, TextPrices$Price)
> abline(PriceModel)
```
Plot 6, Exercise 4c (i)

```r
> plot(TextPrices$Pages, PriceModel$residuals)
> abline(h=0)
```

Plot 7, Exercise 4c (ii)

```r
> qqnorm(resid(PriceModel))
> qqline(resid(PriceModel))
```

Plot 8, Exercise 4c (iii)

```r
> stem(resid(PriceModel), scale=2)
The decimal point is 1 digit(s) to the right of the |
```

-6 53
-5
-4
-3 333
-2 2
-1 94
0 87522110
1 126
2 00077
3 35
4 12
5 2
6
7 3