B27
Hector G. Flores Rodriguez

Stats 110 - HW 2
October 18, 2017

1. Continue analysis of the Sparrows data

We'll first load the data then take a peak at it

```
sparrows <- read.csv("./data/Sparrows.csv")
head(sparrows)
```

```
##   Treatment Weight WingLength
## 1   control   14.9       29.0
## 2   control   15.0       31.0
## 3   control   14.3       25.0
## 4   control   17.0       29.0
## 5   control   16.0       30.0
## 6   control   16.2       31.5
```

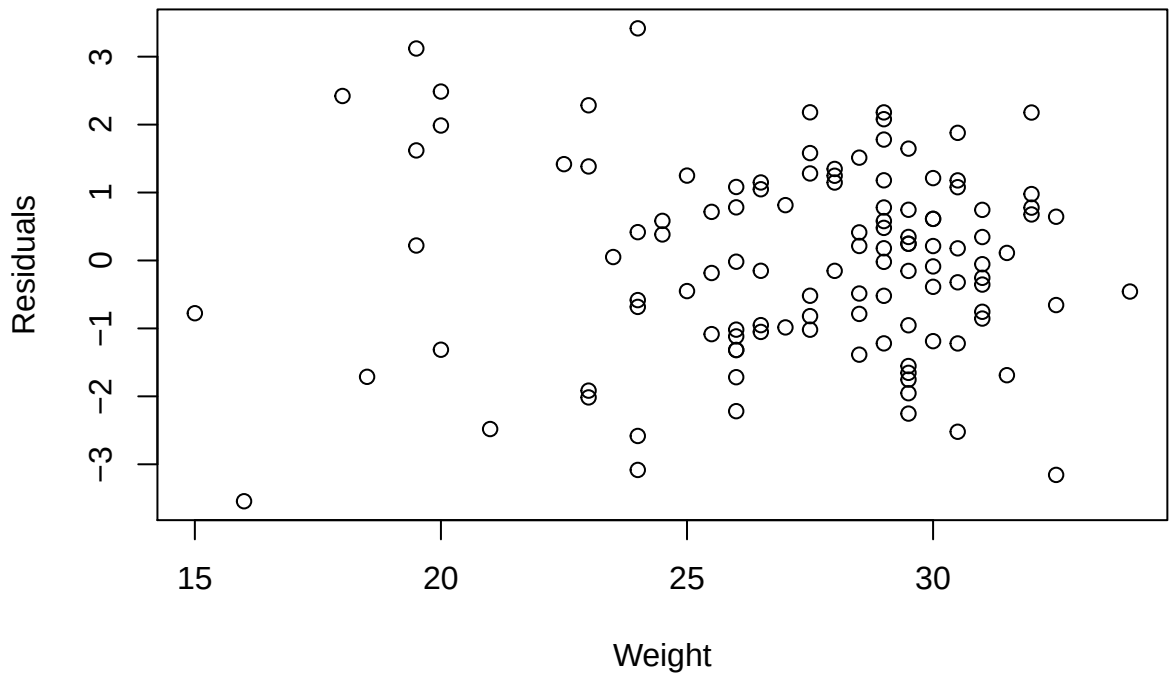
Now let's fit the data

```
sparrows_lm <-lm("Weight ~ WingLength", data=sparrows)
print(sparrows_lm)
```

```
##
## Call:
## lm(formula = "Weight ~ WingLength", data = sparrows)
##
## Coefficients:
## (Intercept)   WingLength
##      1.3655       0.4674
```

1-(a). Create a scatter plot of residuals vs fitted values

```
sparrows_res = resid(sparrows_lm)
plot(sparrows$WingLength, sparrows_res, ylab="Residuals", xlab="Weight")
```

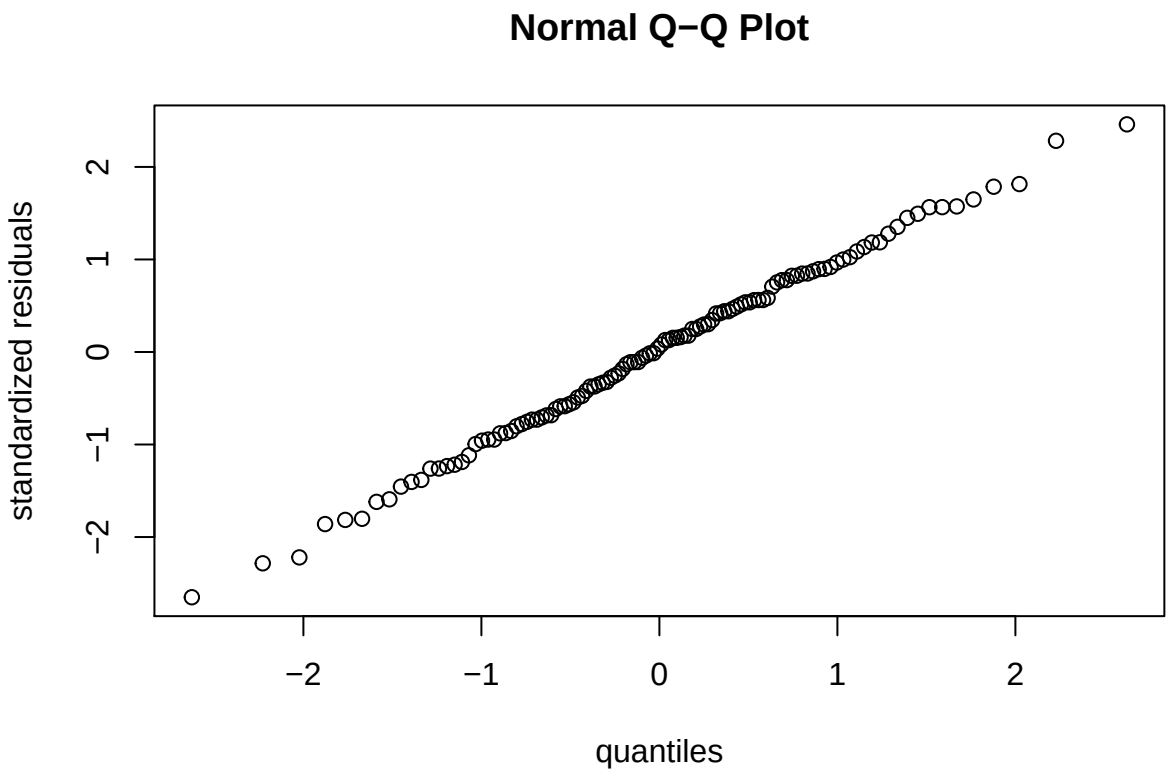


(i). The scatter plot of the residuals vs fitted values is usefult for checking if the variance is constant and for linearity.

(ii). Based on the plot, it is safe to assume constant variance and linearity.

1-(b). Q-Q plot of the standardized residuals

```
stdres <- rstandard(sparrows_lm)
qqnorm(stdres, ylab="standardized residuals", xlab="quantiles")
```



(i). The Q-Q plot is common way of checking for normality.

(ii). The tails do no look so "heavy" and there does not seem to be much of skewness. It is reasonable to assume normality here.

1-(c). Stemplot of the standardized residuals

```
print(stem(stdres))
```

```
##
##   The decimal point is at the |
##
##   -2 | 7
##   -2 | 32
##   -1 | 988665
##   -1 | 4433222100
##   -0 | 99999888777776666555
##   -0 | 444333332211111000
##    0 | 0111222222233334444
##    0 | 555556666788888889999
##    1 | 0001122344
##    1 | 5666688
##    2 | 3
##    2 | 5
##
## NULL
```

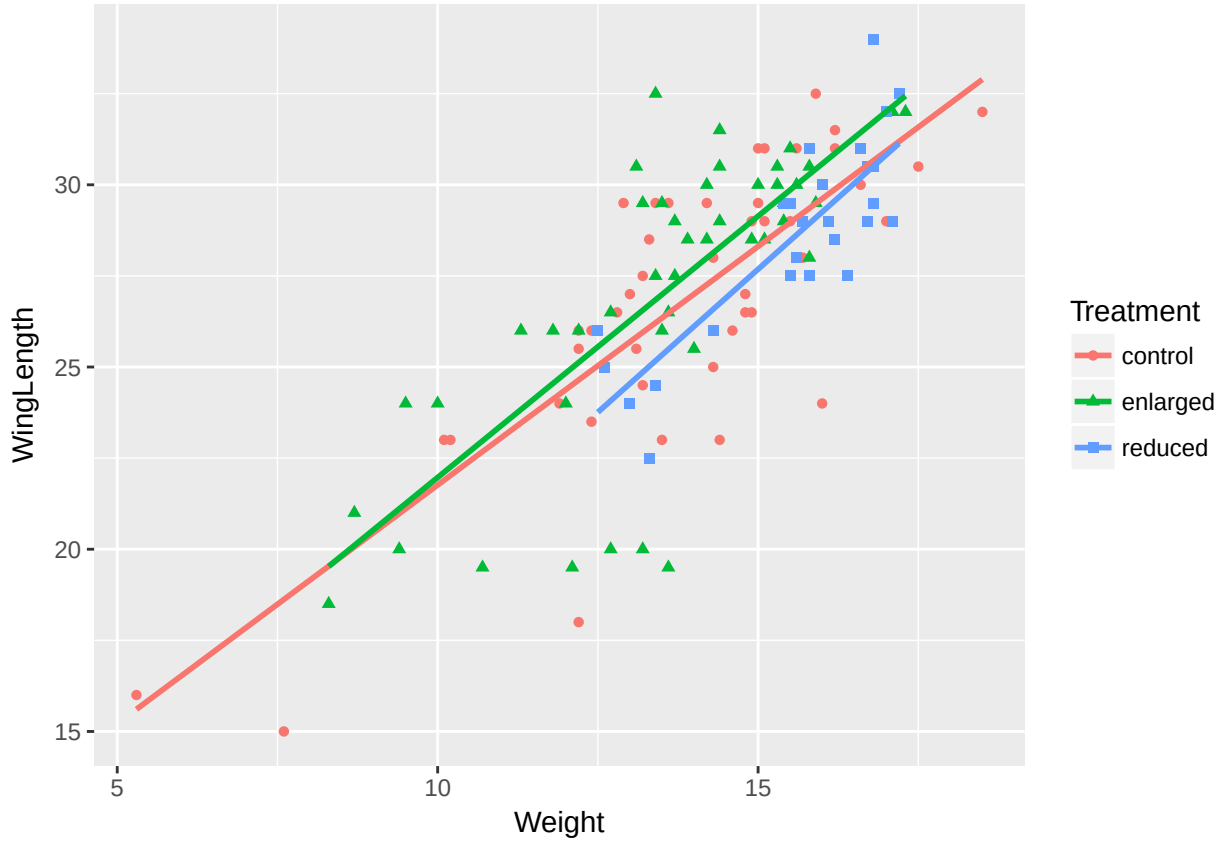
(i). The stemplot is used to check normality assumptions.

(ii). It is reasonable to assume normality.

1-(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

```
# import ggplot2 library
library(ggplot2)

ggplot(sparrows, aes(Weight, WingLength, color=Treatment, shape=Treatment)) +
  geom_point() +
  geom_smooth(method=lm, se=FALSE)
```



2. Carry out the 5 steps of a hypothesis test for the slope to determine if the data provided convincing evidence of a linear relationship between weight and wing length in the population of sparrows similar to these.

(1). Determine the null and alternative hypotheses.

To test for a linear relationship we want our null hypothesis to be such that the slope, β_1, is equal to 0.

Hence,

$$H_0 : \beta_1 = 0$$
$$H_1 : \beta_1 \neq 0$$

(2). Verfiy necessary data conditions, and if met, summarize the data into an appropriate test statistic.

From the previous problem, it is reasonable to assume that our model assumptions are valid. Therefore, we will continue with a t-test.

```
summary(sparrows_lm)
```

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

We can see from the summary of our model that the t-statisic is:

$t = 1.426$

(3). Assuming the null hypothesis is true, find the p-value

From the summary we can see that $p < 2e^{-16}$

(4). Decide whether or not the result is statistically significant based on p-value.

Let us set the significance level to $\alpha = 0.05$.

Since $p << \alpha$, it is reasonable to assume that the result is statistically significant. Therefore, we reject the null hypothesis in favor of the alternative.

(5). Report the conclusion in the context of the situtation

Hence, there is significant evidence that there is a linear relationship between weight and wing length in the population of sparrows.

3. Use the hightway sign data and the Rossman/Chance applet for guessing and viewing a regression line.

(a). Copy and paste the highway sign data into the data box at the applet website. Check the box "Show regression line" and write down the equation

Responseˆ = 576.68 + -3.01 x Explanatory

(b). Remove the slight outlier, guess the regression line and write down the equation of the guess line. Write down the equation of the actual regression

Guess: Responseˆ = 599.36 + -3.67 x Explanatory

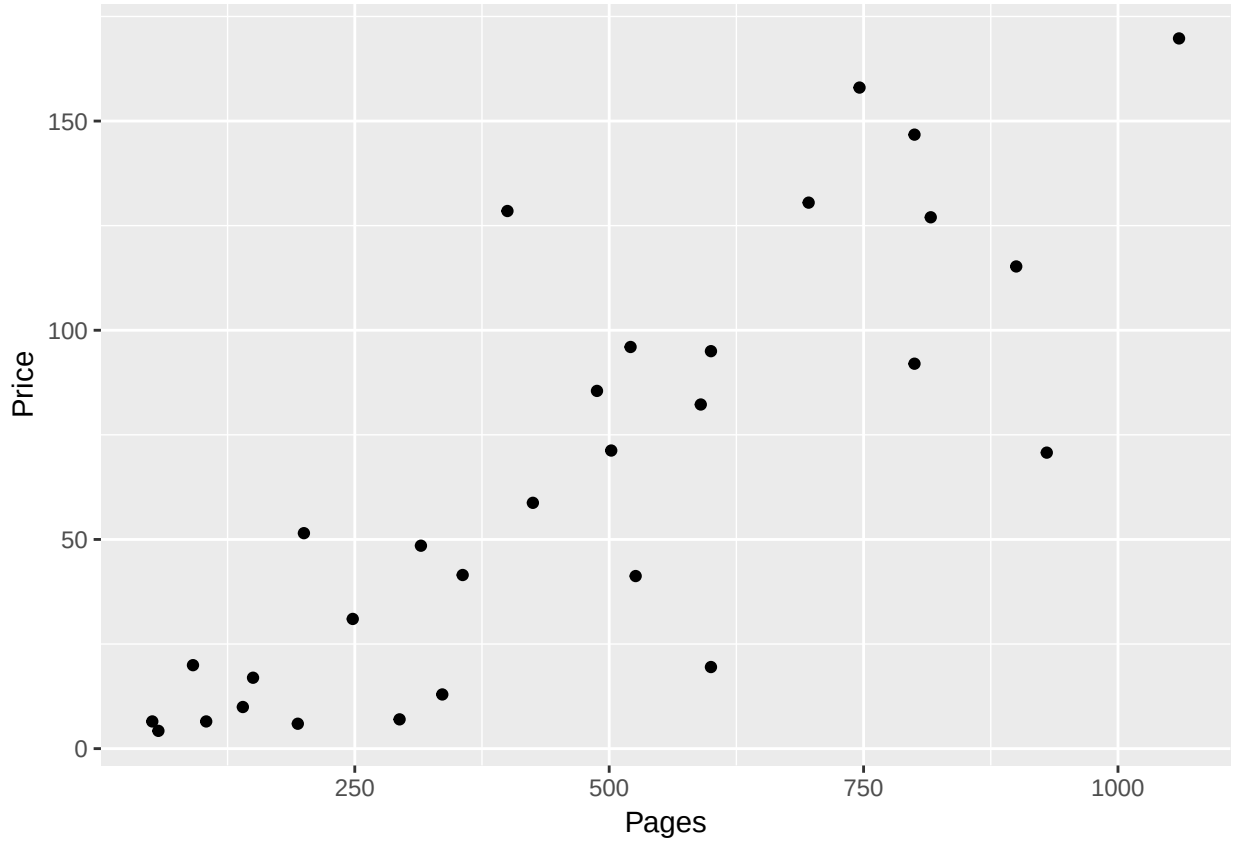
True: Responseˆ = 583.23 + -3.21 x Explanatory

The guess was fairly accurate.

4. Textbook exercise 1.26.

(a). Produce the relevant scatter plot.

```
textprices = read.csv("./data/TextPrices.csv")
ggplot(data=textprices, aes(Pages, Price)) + geom_point()
```



There appears to be a positive linear relationship between pages and price.

(b). Determine the equation of the regression line for prediction.

```
textprices_lm = lm("Price ~ Pages", data=textprices)
print(textprices_lm)
```

```
##
## Call:
## lm(formula = "Price ~ Pages", data = textprices)
##
## Coefficients:
## (Intercept)        Pages
##     -3.4223       0.1473
```

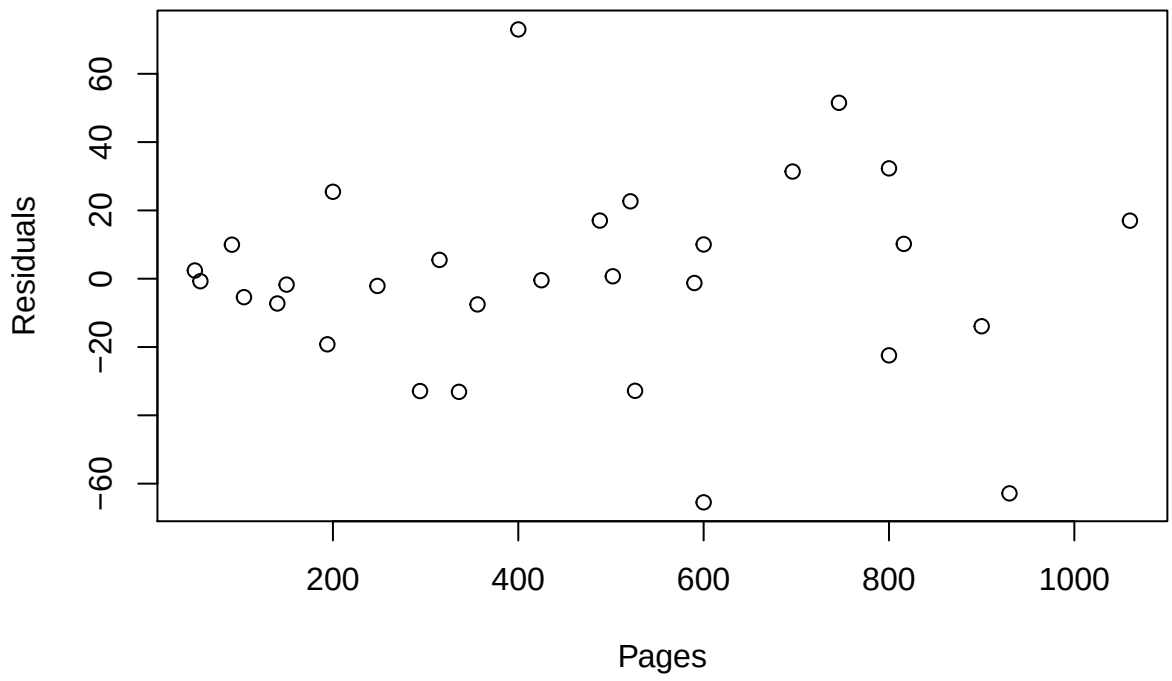
The equation of the regression line is:

$$\hat{y} = -3.4223 + 0.1437\mathbf{x}$$

(c). Create and discuss the following plots:

(i). Residuals vs X

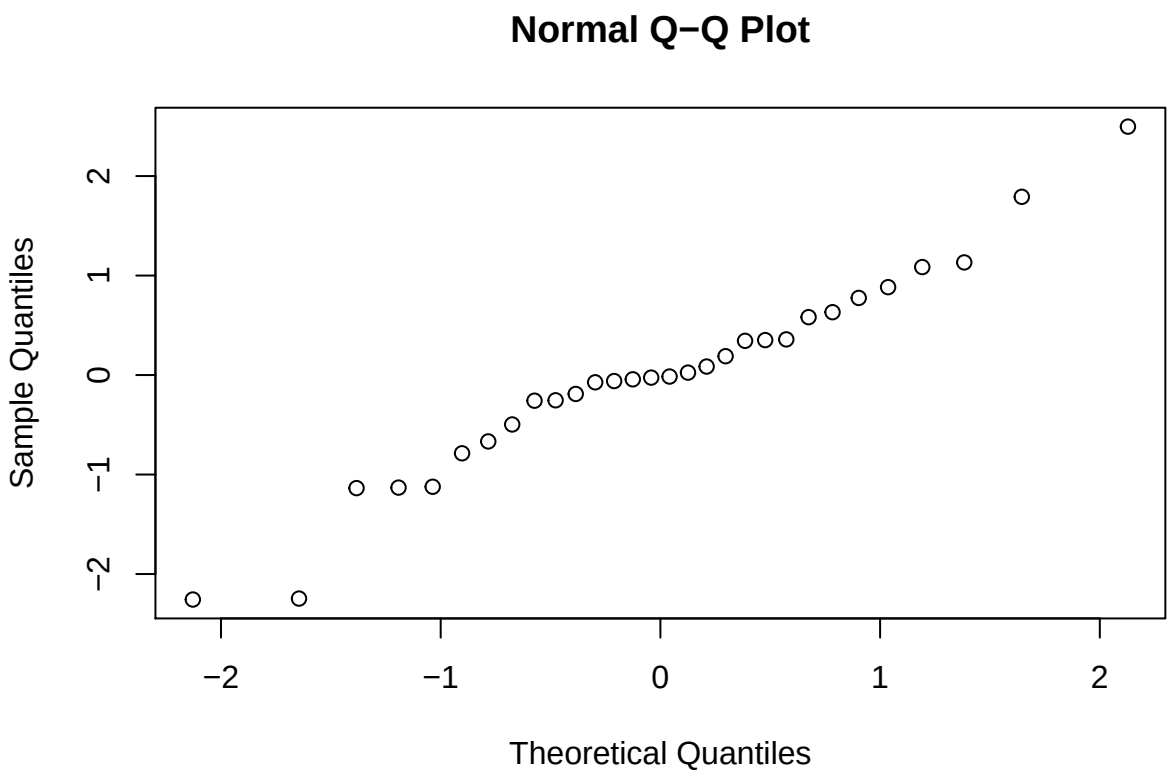
```
textprices_res = resid(textprices_lm)
plot(textprices$Pages, textprices_res, ylab="Residuals", xlab="Pages")
```



From the residual plot, there seems to be an outlier around the 400 pages mark. For the most part, it is reaosonable to assume constant variance and linearity.

(ii). Q-Q plot of residuals

```
tp_stdres = rstandard(textprices_lm)
qqnorm(tp_stdres)
```



The Q-Q plot looks slightly skewed with "heavy" lower tail.

(iii). Stemplot of residuals

```
print(stem(tp_stdres))
```

```
##
##   The decimal point is at the |
##
##   -2 | 32
##   -1 |
##   -1 | 111
##   -0 | 875
##   -0 | 33211000
##    0 | 012344
##    0 | 6689
##    1 | 11
##    1 | 8
##    2 |
##    2 | 5
##
## NULL
```

(d). There's an outlier with a residual of $72. Identify which book (Pages and Price) the outlier is associated with. Discuss which of the three reasons for outliers given in class you think might apply in this case.

Book #4 with 400 pages and a price of $128.50 is the oulier. Most likely the outlier is a result of natural variability since the books were randomly sampled and course textbooks can vary in range depending on the type of course and the edition of the book.