Benji_HW1

January 30, 2018

1 Benji Wagner

1a. Derive the formulas for $\hat{\beta_0}$ and $\hat{\beta_1}$ in a simple linear regression model.

$$Q = \sum_{i=1}^{n} \epsilon_i^2$$

$$\epsilon_i = Y_i - \beta_0 - \beta_1 X_i$$

First we will solve for $\hat{\beta_0}$:

$$\frac{\partial Q}{\partial \beta_0} = 0$$

$$\frac{\partial Q}{\partial \beta_0} = \sum_{i=1}^{n}[\frac{\partial}{\partial \beta_0}\epsilon_i^2] = \sum_{i=1}^{n}[2\epsilon_i \frac{\partial \epsilon_i}{\partial \beta_0}] = 2\sum_{i=1}^{n}\epsilon_i(-1) = 0\sum_{i=1}^{n}\epsilon_i = 0$$

$$\sum_{i=1}^{n}(Y_i - B_0 - B_1 X_i) = 0$$

$$\sum_{i=1}^{n} Y_i - n\beta_0 - \beta_1 \sum_{i=1}^{n} X_i = 0$$

$$n\beta_0 = \sum_{i=1}^{n} Y_i - \beta_1 \sum_{i=1}^{n} X_i$$

$$\beta_0 = \frac{1}{n}(\sum_{i=1}^{n} Y_i - \beta_1 \sum_{i=1}^{n} X_i)$$

$$\beta_0 = \overline{Y} - \beta_1 \overline{X}$$

Now we solve for $\hat{\beta_1}$:

$$\frac{\partial Q}{\partial \beta_1} = 0$$

$$\frac{\partial Q}{\partial \beta_1} = \sum_{i=1}^{n}[\frac{\partial}{\partial \beta_1}\epsilon_i^2] = \sum_{i=1}^{n}[2\epsilon_i \frac{\partial \epsilon_i}{\partial \beta_1}] = 2\sum_{i=1}^{n}\epsilon_i(-X_i) = 0$$

$$\sum_{i=1}^{n} \epsilon_i = 0$$

$$\sum_{i=1}^{n} \epsilon_i X_i = 0 \sum_{i=1}^{n} (Y_i X_i - B_0 X_i - B_1 X_i^2) = 0$$

Plugging in β_0 we get:

$$\sum_{i=1}^{n} Y_i X_i - \frac{1}{n}(\sum_{i=1}^{n} Y_i - \beta_1 \sum_{i=1}^{n} X_i)(\sum_{i=1}^{n} X_i) - \beta_1 \sum_{i=1}^{n} X_i^2 = 0$$

$$\sum_{i=1}^{n} Y_i X_i - \frac{1}{n}(\sum_{i=1}^{n} X_i)(\sum_{i=1}^{n} X_i) + \frac{1}{n}\beta_1(\sum_{i=1}^{n} X_i)(\sum_{i=1}^{n} X_i) - \beta_1 \sum_{i=1}^{n} X_i^2 = 0$$

$$\beta_1 = \frac{\sum_{i=1}^{n} Y_i X_i - \frac{1}{n}(\sum_{i=1}^{n} X_i)(\sum_{i=1}^{n} Y_i)}{\sum_{i=1}^{n} X_i X_i - \frac{1}{n}(\sum_{i=1}^{n} X_i)(\sum_{i=1}^{n} X_i)}$$

$$\beta_1 = \frac{\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})}{\sum_{i=1}^{n}(X_i - \overline{X})^2}$$

Thus:

$$\hat{\beta_0} = \overline{Y} - \beta_1 \overline{X}$$

$$\hat{\beta_1} = \frac{XY}{XX}$$

1b. Derive the matrix forms of $\hat{\beta_0}$ and $\hat{\beta_1}$

Let X =

$$\begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ 1 & x_3 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix}$$

Let Y =

$$\begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ \vdots \\ y_n \end{bmatrix}$$

$X^T X =$

$$\begin{bmatrix} 1 & 1 & 1 & \dots & 1 \\ x_1 & x_2 & x_3 & \dots & x_n \end{bmatrix} \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ 1 & x_3 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix} = \begin{bmatrix} n & \sum_{i=1}^{n} x_i \\ \sum_{i=1}^{n} x_i & \sum_{i=1}^{n} x_i^2 \end{bmatrix}$$

$X^TY =$

$$\begin{bmatrix} 1 & 1 & 1 & \dots & 1 \\ x_1 & x_2 & x_3 & \dots & x_n \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} \sum_{i=1}^n y_i \\ \sum_{i=1}^n x_i y_i \end{bmatrix}$$

We know Y = X β + ϵ in matrix form. This implies that $\hat{\beta} = (X^TX)^{-1}X^TY$

$$\begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} = \begin{bmatrix} n & \sum_{i=1}^n x_i \\ \sum_{i=1}^n x_i & \sum_{i=1}^n x_i^2 \end{bmatrix}^{-1} \begin{bmatrix} \sum_{i=1}^n y_i \\ \sum_{i=1}^n x_i y_i \end{bmatrix}$$

2. Show that:

$$\sum_{i=1}^n (\hat{Y_i} - \overline{Y})(Y_i - \hat{Y_i}) = 0$$

Let $(Y_i - \hat{Y_i}) = \epsilon_i$ and $\hat{Y_i} = \hat{\beta_0} + \hat{\beta_1} X_i$
Then:

$$\sum_{i=1}^n (\hat{Y_i} - \overline{Y})(Y_i - \hat{Y_i}) = \sum_{i=1}^n (\hat{Y_i} - \overline{Y})\epsilon_i$$

$$\sum_{i=1}^n (\hat{Y_i} - \overline{Y})(Y_i - \hat{Y_i}) = \sum_{i=1}^n (\hat{\beta_0} + \hat{\beta_1}\overline{X_i} - \overline{Y})\epsilon_i$$

$$\sum_{i=1}^n (\hat{Y_i} - \overline{Y})(Y_i - \hat{Y_i}) = \hat{\beta_0}\sum_{i=1}^n \epsilon_i + \hat{\beta_1}\sum_{i=1}^n \overline{X_i}\epsilon_i - \overline{Y}\sum_{i=1}^n \epsilon_i$$

Since

$$\sum_{i=1}^n \epsilon_i = 0$$

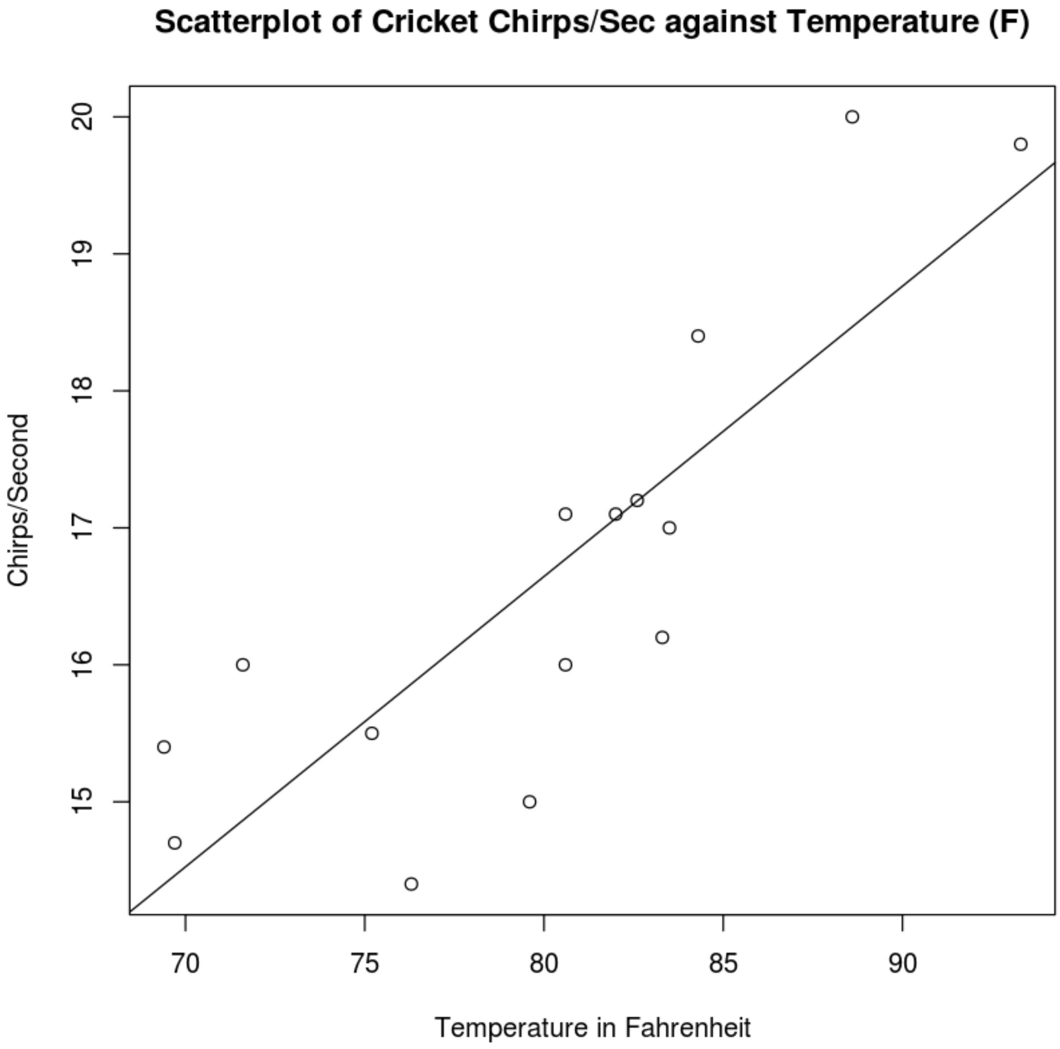
$$\sum_{i=1}^n (\hat{Y_i} - \overline{Y})(Y_i - \hat{Y_i}) = 0$$

3. Consider crickets chirp frequency data presented in cricket data.csv posted on Sakai. There are n=15 bivariate measurements on striped ground crickets, where Y = chirps per seconds and X= temperture in Fahrenheit:

```
In [1]: crickets <- read.csv(file = 'cricket_data.csv', header = TRUE, sep = ',')
```

a. Obtain a scatterplot of these measurements.

```
In [2]: plot(x = crickets$x, y = crickets$y,
             main = "Scatterplot of Cricket Chirps/Sec against Temperature (F)",
             xlab = "Temperature in Fahrenheit", ylab = "Chirps/Second")
        fit <- lm(crickets$y ~ crickets$x)
        abline(fit)
```



b. Specify the simple linear regression model for these data. Identify all parameters in the model, providing interpretation of each.

```
In [3]: summary(lm(crickets$y ~ crickets$x))


Call:
lm(formula = crickets$y ~ crickets$x)

Residuals:
     Min       1Q   Median       3Q      Max
-1.56009 -0.57930  0.03129  0.59020  1.53259
```

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.30914    3.10858  -0.099 0.922299
crickets$x   0.21193    0.03871   5.475 0.000107 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 0.9715 on 13 degrees of freedom
Multiple R-squared:  0.6975,Adjusted R-squared:  0.6742
F-statistic: 29.97 on 1 and 13 DF,  p-value: 0.0001067
```

Looking at the summary, $\hat{\beta_0} = -0.30914$ and $\hat{\beta_1} = 0.21193$. This can be interpreted as for every 1-degree Fahrenheit increase in temperature, assuming everything else is held constant, the chirps per second will increase by 0.21193. Additionally, the $\hat{\beta_0}$ intercept is just a centering constant since it is not interpretable within the context of our data.

c. Explain how the interpretation (and the estimate) of the slope parameter changes if temperature is expressed in Celsius.

If temperature were to be expressed in Celsius, the interpretation would not change, but the slope parameter would change to accommodate the change to Celsius.

d. Estimate the mean chirp frequency among crickets in a temperature of 80 degrees F. Estimate the standard deviation among chirp frequency measurements made at this fixed temperature.

Using:

$$\hat{Y_i} = \hat{\beta_0} + \hat{\beta_1} X_i \hat{\sigma^2} = S^2_{y.x} = \frac{1}{n-2}\sum_{i=1}^{n}(Y_i - \hat{Y_i})^2$$

```
In [35]: (beta_0 <- lm(crickets$y ~ crickets$x)$coefficients[1])
         (beta_1 <- lm(crickets$y ~ crickets$x)$coefficients[2])
         s_yx <- sqrt(sum((crickets$y - predict(fit))^2) / (length(crickets$y) - 2))
```

(Intercept): -0.309144391026669
crickets\$x: 0.211925009049975

```
In [36]: cat("Estimated mean chirp frequency: ", beta_0 + beta_1*80, "Chirps/Second")
         cat("\nEstimated standard deviation among chirp frequencies: ", s_yx)
```

```
Estimated mean chirp frequency:  16.64486 Chirps/Second
Estimated standard deviation among chirp frequencies:  0.9715177
```

e. Estimate the mean chirp frequency among crickets in a temperature of 105 degrees F.

```
In [6]: cat("Estimated mean chirp frequency: ", beta_0 + beta_1*105, "Chirps/Second")
```

```
Estimated mean chirp frequency:  21.94298 Chirps/Second
```

If we extrapolated the model, the estimated mean chirp frequency would be ~22 chirps/second. However, since our measured temperatures do not extend to 105 degrees Fahrenheit, we cannot accurately estimate the mean chirp frequency at 105 degrees Fahrenheit.

f. Report the sum of squares deviations between the fitted values and the average chirp frequency $\overline{Y}$

Using:

$$SSR = \sum_{i=1}^{n} (\hat{Y}_i - \overline{Y})^2$$

```
In [7]: sum((predict(fit) - mean(crickets$y))^2)
```

```
28.2873263579725
```

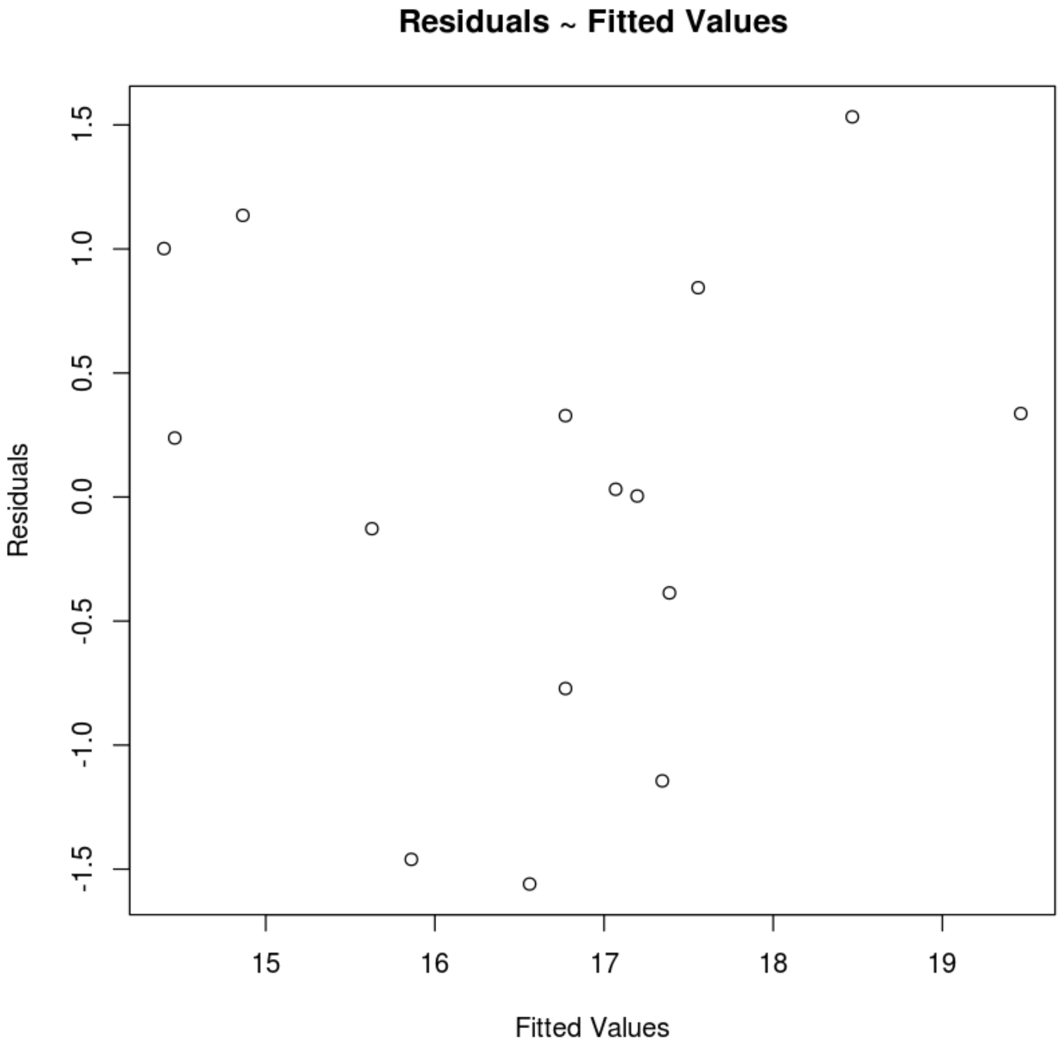
g. What proportion of variance in chirp frequencies is explained by the linear regression model?

```
In [8]: cat("The proportion of variance in chirp frequencies explained by the linear regression
            cor(crickets$x, crickets$y)^2)
```

```
The proportion of variance in chirp frequencies explained by the linear regression model is:
 0.6974651
```

h. Obtain a plot of the residuals against the fitted values

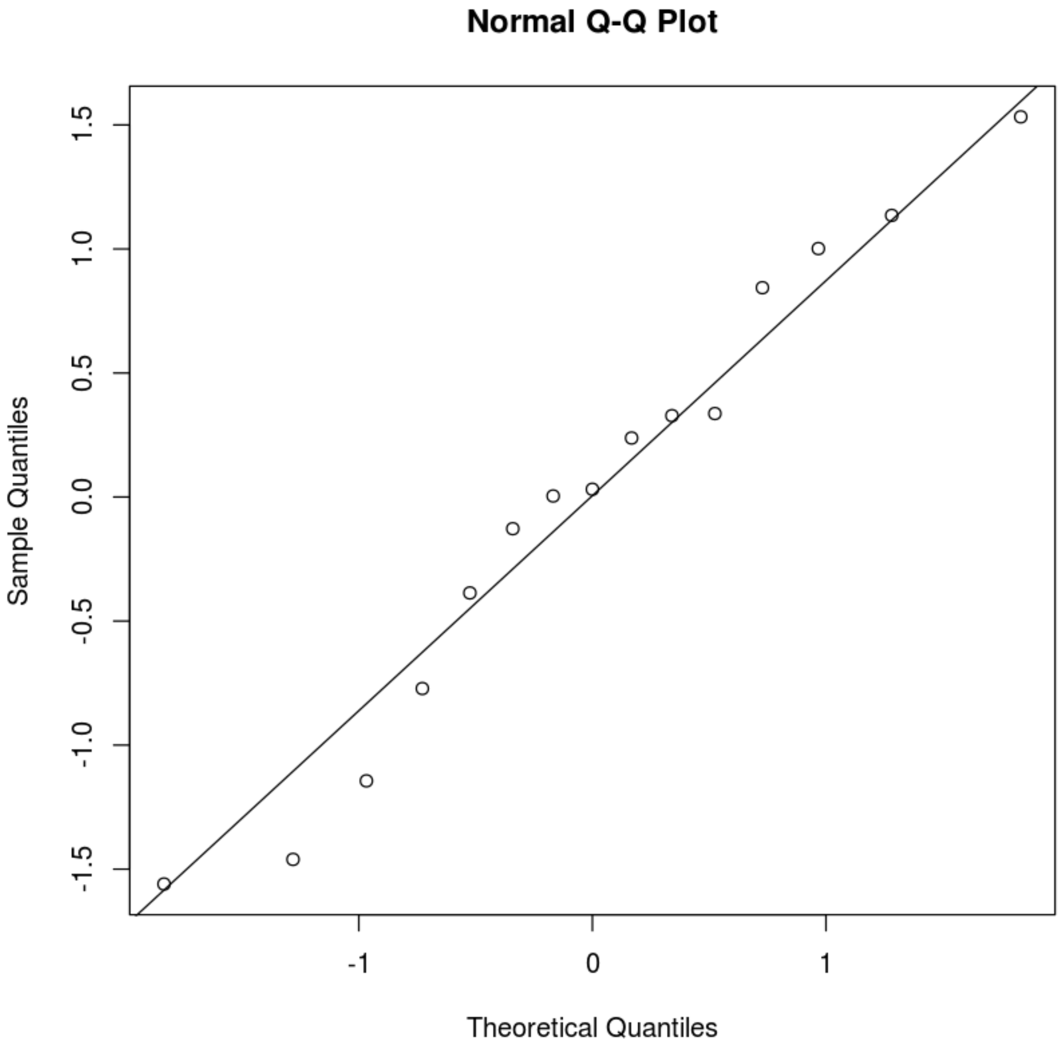
```
In [9]: plot(crickets$y - predict(fit) ~ predict(fit),
             xlab = "Fitted Values",
             ylab = "Residuals",
             main = "Residuals ~ Fitted Values")
```



i. Obtain a plot of the ordered residuals against the corresponding quantiles from the standard normal distribution.

```
In [10]: ordered_residuals <-(crickets$y - predict(fit))[order(crickets$y - predict(fit))]
```

```
In [11]: qqnorm(ordered_residuals)
         qqline(ordered_residuals)
```

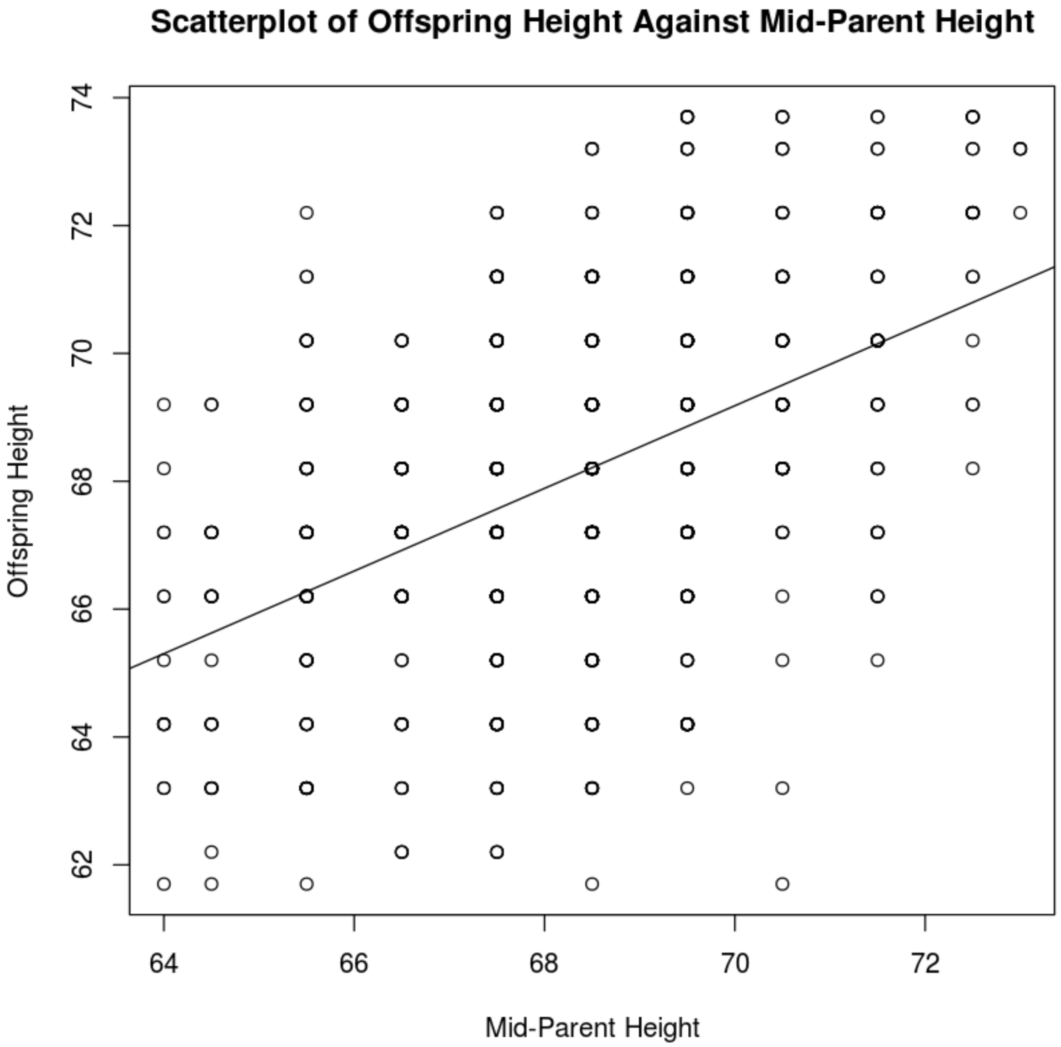


4. Stigler, History of Statistics pg. 285 gives Galton's famous data on heights of sons (Y in inches) and average parents height (X in inches) scaled to represent a male height (essentially sons' heights versus fathers' heights). Data are given in parents offsprings.csv on Sakai. Consider a statistical model for these data, randomly sampled from some population of interest. In particular, choose a model which accounts for the apparent linear dependence of the mean height of sons on midparent height X.

```
In [12]: heights <- read.csv(file = 'parents_offsprings.csv', header = TRUE)
```

A. Obtain a scatterplot of these data.

```
In [13]: plot(x = heights$midparent_height, y = heights$offspring_height,
              xlab = "Mid-Parent Height", ylab = "Offspring Height",
              main = "Scatterplot of Offspring Height Against Mid-Parent Height")
         abline(lm(heights$offspring_height ~ heights$midparent_height))
```



```
In [14]: beta_0 <- unname(lm(heights$offspring_height ~ heights$midparent_height)$coefficients[1
         beta_1 <- unname(lm(heights$offspring_height ~ heights$midparent_height)$coefficients[2

In [15]: summary(lm(heights$offspring_height ~ heights$midparent_height))
```

```

Call:
lm(formula = heights$offspring_height ~ heights$midparent_height)

Residuals:
    Min      1Q  Median      3Q     Max
-7.8050 -1.3661  0.0487  1.6339  5.9264
```

```
Coefficients:
                         Estimate Std. Error t value Pr(>|t|)
(Intercept)              23.94153    2.81088   8.517   <2e-16 ***
heights$midparent_height  0.64629    0.04114  15.711   <2e-16 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 2.239 on 926 degrees of freedom
Multiple R-squared:  0.2105,Adjusted R-squared:  0.2096
F-statistic: 246.8 on 1 and 926 DF,  p-value: < 2.2e-16
```

B. What is the meaning, in words, of β_1?

Looking at the summary, $\hat{\beta_1} = 0.64629$. This can be interpreted as when you hold everything else constant, for every 1-inch increase of midparent height, the estimated offspring height increases by 0.64629 inches.

C. What is the observed value of $\hat{\beta_1}$?

```
In [16]: cat("The observed value of beta_1 is: ", beta_1)
```

```
The observed value of beta_1 is:  0.6462906
```

D. How much does $\hat{\beta_1}$ vary about β_1 from sample to sample? (Provide an estimate of the standard error, as well as an expression indicating how it was computed)

Using:

$$\widehat{\text{var}(\hat{\beta_1})} = \frac{S^2_{y.x}}{\sum_{i=1}^{n}(X_i - \overline{X})^2}$$

and

$$\sigma^2 = S^2_{y.x} = \frac{1}{n-2}\sum_{i=1}^{n}(Y_i - \hat{Y_i})^2$$

```
In [17]: residuals <- lm(heights$offspring_height~heights$midparent_height)$residuals
         var_b1 <- (sum(residuals^2)/(length(heights$midparent_height) - 2))/
                     sum((heights$midparent_height - mean(heights$midparent_height))^2)
         cat("The estimate of the standard error is: ", var_b1^0.5)
```

```
The estimate of the standard error is:  0.04113588
```

E. What is a region of plausible values for β_1 suggested by the data?

Using a 95% confidence interval for β_1, we can use the following formula:

$$\hat{\beta_1} \pm t_{0.975,n-2} SE(bet\hat{a}_1)$$

```
In [18]: cat("The region of plausible values for beta_1 suggested by the data is : (",
             beta_1 - qt(p = 0.975, df = length(heights$midparent_height) - 2) * var_b1^0.5,
             ", ",
             beta_1 + qt(p = 0.975, df = length(heights$midparent_height) - 2) * var_b1^0.5,
             ")"
             )
```

```
The region of plausible values for beta_1 suggested by the data is : ( 0.5655602 ,  0.7270209 )
```

F. What is the line that best fits these data, using criterion that smallest sum of squared residuals is "best?"

The line that fits best is the one we used above in the linear model: $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i = 23.94153 + 0.64629X_i$

G. How much of the observed variation in the heights of sons (the y-axis) is explained by this "best" line?

We can measure the observed variation in the heights of sons explained by the best-fit line using the R^2 value:

$$R^2 = 1 - \frac{SSE}{SST}$$

```
In [19]: SSR <- sum((predict(lm(heights$offspring_height ~ heights$midparent_height)) -
                   mean(heights$offspring_height))^2)
         SSE <- sum((heights$offspring_height -
                   predict(lm(heights$offspring_height ~ heights$midparent_height)))^2)
         SST <- sum((heights$offspring_height - mean(heights$offspring_height))^2)
```

```
In [20]: cat("Check for equivalency:\nSSR + SST = ", SSR + SSE, "\nSST = ", SST)
```

```
Check for equivalency:
SSR + SST =  5877.207
SST =  5877.207
```

R^2 value is:

```
In [21]: 1 - SSE/SST
```

0.210462910561639

H. What is the estimated average height of sons whose midparent height is x = 68?

```
In [22]: cat("The estimated average height of sons whose midparent height is 68 is: ",
             beta_0 + beta_1*68)
```

```
The estimated average height of sons whose midparent height is 68 is:  67.88929
```

I. Is this the true average height in the whole population of sons whose midparent height is x = 68?

No, the *true* average height in the whole population of sons whose midparent height is 68 inches is what we are trying to approximate with our model, but the true average is unknown.

J. Under the model, what is the true average height of sons with midparent height is x = 68?
$Y = \beta_0 + \beta_1(68)$

K. What is the estimated standard deviation among the population of sons whose midparent height is x = 68? Would you call this standard deviation a ”standard error”?

Since:

$$MSE = \frac{SSE}{n-2}$$

```
In [23]: MSE <- SSE/(length(heights$midparent_height)-2)
```

```
In [24]: cat("The estimated standard deviation is: ", sqrt(MSE))
```

```
The estimated standard deviation is:  2.238547
```

I would not call this standard deviation a "standard error" since it’s being used to estimate σ^2, which is the population parameter and not the sample statistic.

L. What is the estimated standard deviation among the population of sons whose midparent height is x = 72? Bigger, smaller, or the same that for x = 68? Is your answer obviously supported or refuted by inspection of the scatterplot?

```
In [25]: cat("The estimated standard deviation is: ", sqrt(MSE))
```

```
The estimated standard deviation is:  2.238547
```

The estimated standard deviation is the same for x = 68 due to homoskedasticity. My answer is supported by my scatterplot.

M. What is the estimated standard error of the estimated average for sons with midparent height x = 68, $\hat{\mu}(68) = \hat{\beta_0} + \hat{\beta_1}(68)$? Provide an expression for this standard error.

$$SE(\hat{\mu}(68)) = \sqrt{MSE(\frac{1}{n} + \frac{(68 - \overline{X})^2}{\sum(x_i - \overline{x})^2})}$$

```
In [26]: xs <- heights$midparent_height
         ys <- heights$offspring_height
         n <- length(xs)
```

```
In [27]: cat("The estimated standard error of the estimated average
             height for sons w/ midparent height x = 68 is:\n",
             sqrt((((68 - mean(xs))^2/sum((xs - mean(xs))^2)) + (1/n))*MSE))
```

```
The estimated standard error of the estimated average
    height for sons w/ midparent height x = 68 is:
 0.07456949
```

N. Is the estimated standard error of $\hat{\mu}(72)$ bigger, smaller, or the same as that for $\hat{\mu}(68)$?

```
In [28]: cat("The estimated standard error of the estimated average height
            for sons w/ midparent height x = 72 is:\n",
            sqrt((((72 - mean(xs))^2/sum((xs - mean(xs))^2)) + (1/n))*MSE))
```

```
The estimated standard error of the estimated average height
    for sons w/ midparent height x = 72 is:
 0.1687102
```

$\hat{SE}(\hat{\mu}(72)) > \hat{SE}(\hat{\mu}(68))$

O. Is the observed linear association between son's height and midparent height strong? Report a test statistic.

Using:

$$r = \hat{\beta}_1 \frac{S_x}{S_y}$$

```
In [29]: r <- beta_1*sqrt(sum((xs - mean(xs))^2)/sum((ys - mean(ys))^2))
```

```
In [30]: cat("The observed linear association can be measured using
            Pearson's correlation coefficient r: ", r,
           "\nThis is moderately strong and positive.")
```

```
The observed linear association can be measured using
    Pearson's correlation coefficient r:  0.4587624
This is moderately strong and positive.
```

P. What quantity can you use to describe or characterize the linear association between son's height and midparent height in the whole population? Is this a parameter or a statistic?

You could use ρ, which is a parameter of linear association for a population.

Q. Let Y denote the height of a male randomly sampled from this population and X his midparent height. Is it true that the population correlation coefficient ρ satisfies:

$$\rho = E[(\frac{Y - \mu_Y}{\sigma_Y}) \times (\frac{X - \mu_X}{\sigma_X})]$$

This is true!

R. Define ţY , Y , ţX, X, . Parameters or statistics?

These are all population parameters.

μ_Y is the mean height of an offspring from the population.

μ_X is the mean midparent height from the population.

σ_Y is the standard deviation of offspring heights for the population.

σ_X is the standard deviation of midparent heights for the population.

ρ is the population correlation coefficient between X and Y.

S. What are the plausible values for ρ suggested by the data?
Using a 95% confidence interval:

```
In [31]: lower_bound <- (((1+r)/(1-r))*exp(-2*1.96/sqrt(n-3)) - 1)/
             (((1+r)/(1-r))*exp(-2*1.96/sqrt(n-3)) + 1)
```

```
In [32]: upper_bound <- (((1+r)/(1-r))*exp(2*1.96/sqrt(n-3)) - 1)/
             (((1+r)/(1-r))*exp(2*1.96/sqrt(n-3)) + 1)
```

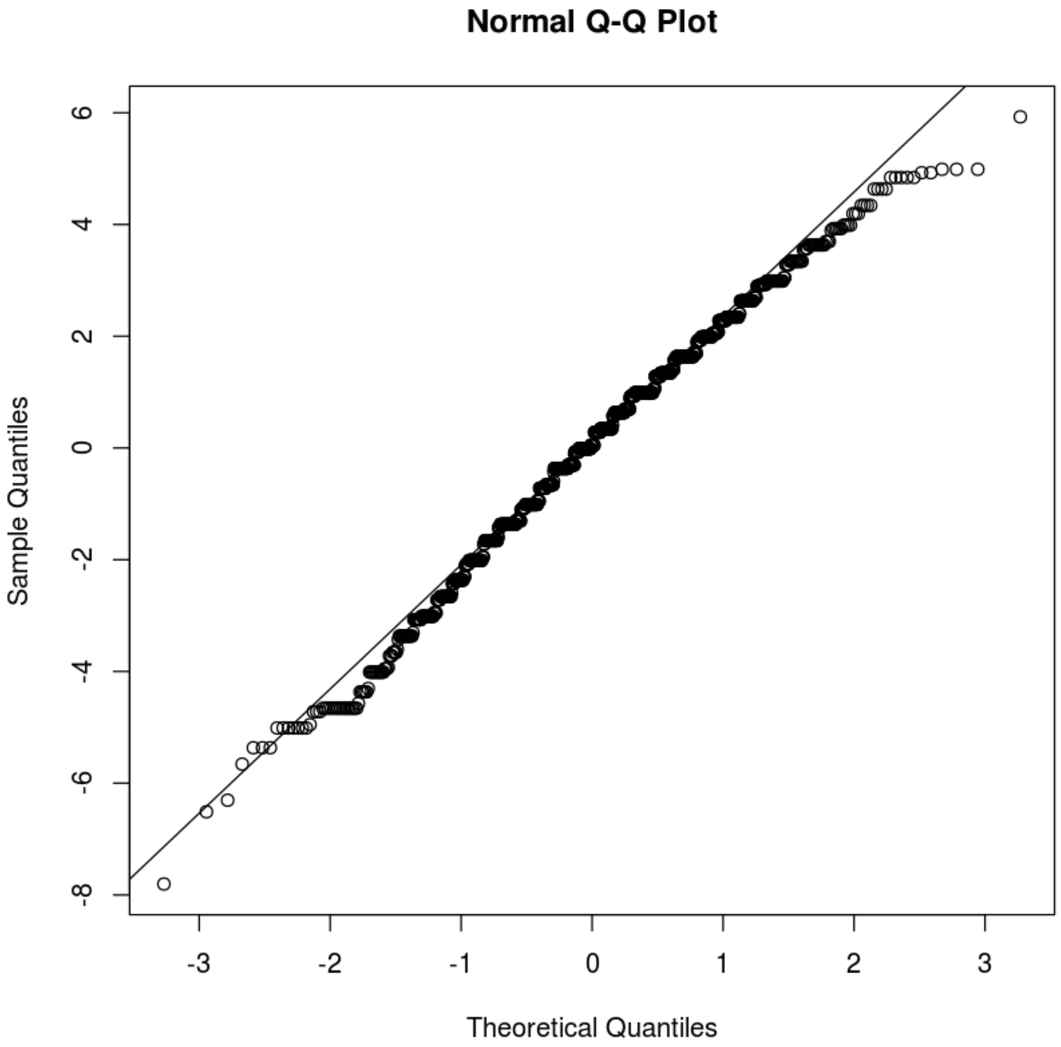
The plausible values for ρ are:

```
In [33]: cat("( ", lower_bound, ", ", upper_bound, ")")
```

```
(  0.4064057 ,  0.5081162 )
```

T. Are the residuals $e_1, e_2, \ldots, e_{928} \overset{iid}{\sim} N(0, \sigma^2)$ a reasonable assumption?
Yes. Looking at the QQNorm Plot, this appears to be a reasonable assumption since there aren't any heavy-tails.

```
In [34]: qqnorm(y = residuals)
         qqline(y = residuals)
```


Normal Q-Q Plot
Sample Quantiles
Theoretical Quantiles
6
4
2
0
-2
-4
-6
-8
-3
-2
-1
0
1
2
3