BIOS705_HW01

January 30, 2018

1 BIOS705 HW01

1.0.1 Author: Kuei-Yueh Ko


```
In [2]: # Setup Environment
        #workdir <- ""
        pathData <- "./Data"
```

```
In [3]: # Read in data
        dat_cricket <- read.csv(
            file.path(pathData, "cricket_data.csv"),
            header = TRUE,
            stringsAsFactors = FALSE)

        dat_parents_offspring <- read.csv(
            file.path(pathData, "parents_offsprings.csv"),
            header = TRUE,
            stringsAsFactors = FALSE)
```

2 Question 01

(a) (Slide 11) Derive the formulas for $\hat{\beta}_0$ and $\hat{\beta}_1$

See the scanned images for Q1 part (a)

2.1 (b) (Slide 48) As in part (a) above, derive the same formulas (ie equivalency to formulas on Slide 11).

See the scanned images for Q1 part (b)

3 Question 02

Show that $\sum(\hat{Y} - \bar{Y})(Y_i - \hat{Y}_i) = 0$

See the scanned image for Q2

BIOS 705 HW 1 Kuei-Yueh Ko

Q1

(a) (Slide 11) $\hat{\beta}_0$ $\hat{\beta}_1$ Derivation

(1) $Q = \sum_i \varepsilon_i^2$

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

$$\frac{\partial Q}{\partial \beta_0} = \sum_i \left[\frac{\partial}{\partial \beta_0} \left(\varepsilon_i^2 \right) \right] = \sum_i 2\varepsilon_i \cdot \frac{\partial \varepsilon_i}{\partial \beta_0} = 2\sum_i \varepsilon_i (-1) = -2\sum_i \varepsilon_i$$

$$\frac{\partial Q}{\partial \beta_1} = \sum_i \left[\frac{\partial}{\partial \beta_1} \left(\varepsilon_i^2 \right) \right] = \sum_i 2\varepsilon_i \cdot \frac{\partial \varepsilon_i}{\partial \beta_1} = 2\sum_i \varepsilon_i (-X_i) = -2\sum_i \varepsilon_i X_i$$

(2) let $\frac{\partial Q}{\partial \beta_0} = 0$, $\frac{\partial Q}{\partial \beta_1} = 0$

$$\Rightarrow \begin{cases} -2\sum_i \varepsilon_i = 0 \\ -2\sum_i \varepsilon_i X_i = 0 \end{cases} \Rightarrow \begin{cases} \sum_i \varepsilon_i = 0 \\ \sum_i \varepsilon_i X_i = 0 \end{cases}$$

$$\Rightarrow \begin{cases} \sum_i (Y_i - \beta_0 - \beta_1 X_i) = 0 \\ \sum_i (X_i Y_i - \beta_0 X_i - \beta_1 X_i^2) = 0 \end{cases} \Rightarrow \begin{cases} \left(\sum_i Y_i\right) - n\beta_0 - \beta_1 \left(\sum_i X_i\right) = 0 \quad \cdots ① \\ \left(\sum_i X_i Y_i\right) - \beta_0 \left(\sum_i X_i\right) - \beta_1 \left(\sum X_i^2\right) \quad \cdots ② \end{cases}$$

(3) ① $\Rightarrow \beta_0 = \frac{1}{n}\left(\sum Y_i - \beta_1 \sum X_i\right)$, substitute β_0 in ②

② $\Rightarrow \left(\sum_i X_i Y_i\right) - \frac{1}{n}\left(\sum Y_i - \beta_1 \sum X_i\right)\left(\sum_i X_i\right) - \beta_1 \left(\sum X_i^2\right) = 0$

$\Rightarrow \left(\sum_i X_i Y_i\right) - \frac{1}{n}\left(\sum X_i\right)\left(\sum Y_i\right) + \frac{1}{n}\beta_1 \left(\sum X_i\right)\left(\sum X_i\right) - \beta_1 \left(\sum X_i^2\right) = 0$

$$\Rightarrow \beta_1 = \frac{\sum X_i Y_i - \frac{1}{n}\left(\sum X_i\right)\left(\sum Y_i\right)}{\sum X_i X_i - \frac{1}{n}\left(\sum X_i\right)\left(\sum X_i\right)} = \frac{\sum (X_i - \bar{X})(Y_i - \bar{Y})}{\sum (X_i - \bar{X})(X_i - \bar{X})}$$

where $\bar{X} = \frac{1}{n}\sum X_i$, $\bar{Y} = \frac{1}{n}\sum Y_i$

(4) $\hat{\beta}_0 = \frac{1}{n}\left(\sum Y_i - \beta_1 \sum X_i\right) = \bar{Y} - \beta_1 \bar{X}$

$$\hat{\beta}_1 = \frac{\sum (X_i - \bar{X})(Y_i - \bar{Y})}{\sum (X_i - \bar{X})(X_i - \bar{X})}$$

※

BIOS705 HW01

Q1

(b) (Slide 48) $\hat{\beta} = (X^TX)^{-1}(X^TY)$

(1) $X = \begin{bmatrix} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix} \quad Y = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix}$

$$X^TX = \begin{bmatrix} 1 & \cdots & 1 \\ x_1 & \cdots & x_n \end{bmatrix} \begin{bmatrix} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix} = \begin{bmatrix} n & \sum x_i \\ \sum x_i & \sum x_i^2 \end{bmatrix}$$

$$(X^TX)^{-1} = \frac{1}{n\sum x_i^2 - (\sum x_i)(\sum x_i)} \begin{bmatrix} \sum x_i^2 & -\sum x_i \\ -\sum x_i & n \end{bmatrix}$$

$$(X^TY) = \begin{bmatrix} 1 & \cdots & 1 \\ x_1 & \cdots & x_n \end{bmatrix} \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} \sum y_i \\ \sum x_i y_i \end{bmatrix}$$

$$(X^TX)^{-1}(X^TY) = \frac{1}{n\sum x_i^2 - (\sum x_i)(\sum x_i)} \begin{bmatrix} \sum x_i^2 \sum y_i - \sum x_i \sum x_i y_i \\ -\sum x_i \sum y_i + n \sum x_i y_i \end{bmatrix}$$

$$\Rightarrow \begin{cases} \hat{\beta}_0 = \dfrac{\sum x_i^2 \sum y_i - \sum x_i \sum x_i y_i}{n\sum x_i^2 - \sum x_i \sum x_i} \\ \hat{\beta}_1 = \dfrac{n\sum x_i y_i - \sum x_i \sum y_i}{n\sum x_i^2 - \sum x_i \sum x_i} = \dfrac{\sum x_i y_i - \frac{1}{n}\sum x_i \sum y_i}{\sum x_i x_i - \frac{1}{n}\sum x_i \sum x_i} \text{ ※} \end{cases}$$

(2) The $\hat{\beta}_0$ can be derived from $X^TX\hat{\beta} = X^TY$

$$\Rightarrow \begin{bmatrix} n & \sum x_i \\ \sum x_i & \sum x_i^2 \end{bmatrix} \begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \end{bmatrix} = \begin{bmatrix} \sum y_i \\ \sum x_i y_i \end{bmatrix}$$

$$\Rightarrow \begin{cases} n\hat{\beta}_0 + \hat{\beta}_1 \sum x_i = \sum y_i \longrightarrow \frac{1}{n}\sum y_i = \hat{\beta}_1 \frac{1}{n}\sum x_i + \hat{\beta}_0 \\ \hat{\beta}_0 \sum x_i + \hat{\beta}_1 \sum x_i^2 = \sum x_i y_i \end{cases}$$

$$\Rightarrow \hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X} \text{ ※}$$

BIOS 705 HW1

Q2 Show that $\sum_{i=1}^{n} (\hat{Y}_i - \bar{Y})(Y_i - \hat{Y}_i) = 0$

(1) $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i$ $\qquad \bar{Y} = \frac{1}{n}\sum Y_i$

$Y_i = \hat{\beta}_0 + \hat{\beta}_1 X_i + \varepsilon_i$ $\qquad \bar{Y} = \hat{\beta}_0 + \hat{\beta}_1 \bar{X}$

(2) from Q1, we know that

$Q = \sum_i \varepsilon_i^2$, $\frac{\partial Q}{\partial \beta_0} = 0$, $\frac{\partial Q}{\partial \beta_1}$

$$\Rightarrow \begin{cases} \frac{\partial Q}{\partial \beta_0} = 0 \Rightarrow \sum \varepsilon_i = 0 \\ \frac{\partial Q}{\partial \beta_1} = 0 \Rightarrow \sum \varepsilon_i x_i = 0 \end{cases}$$

(3)
$$\begin{aligned} \sum_i (\hat{Y}_i - \bar{Y})(Y_i - \hat{Y}_i) &= \sum_{i=1}^{n} \left[(\hat{\beta}_0 + \hat{\beta}_1 X_i) - (\hat{\beta}_0 + \hat{\beta}_1 \bar{X}) \right] \varepsilon_i \\ &= \sum_i \hat{\beta}_1 (X_i - \bar{X}) \varepsilon_i \\ &= \hat{\beta}_1 \sum X_i \varepsilon_i - \hat{\beta}_1 \bar{X} \sum \varepsilon_i \\ &= 0 \end{aligned}$$

4 Question 03

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 °F**:

(a) Obtain a scatter plot of these meansurements

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

(c) Expalin how the interpretation (and the estimate) of the slope parameter changes if temperature is expressed in degree celsius

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

(e) Estimate the mean chirp frequency among crickets in a temperature of 105F

(f) Report the sum of squares deviations between the tted values and the average chirp frequency $\bar{Y}$

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

(h) Obtain a plot of the residuals against the tted values.

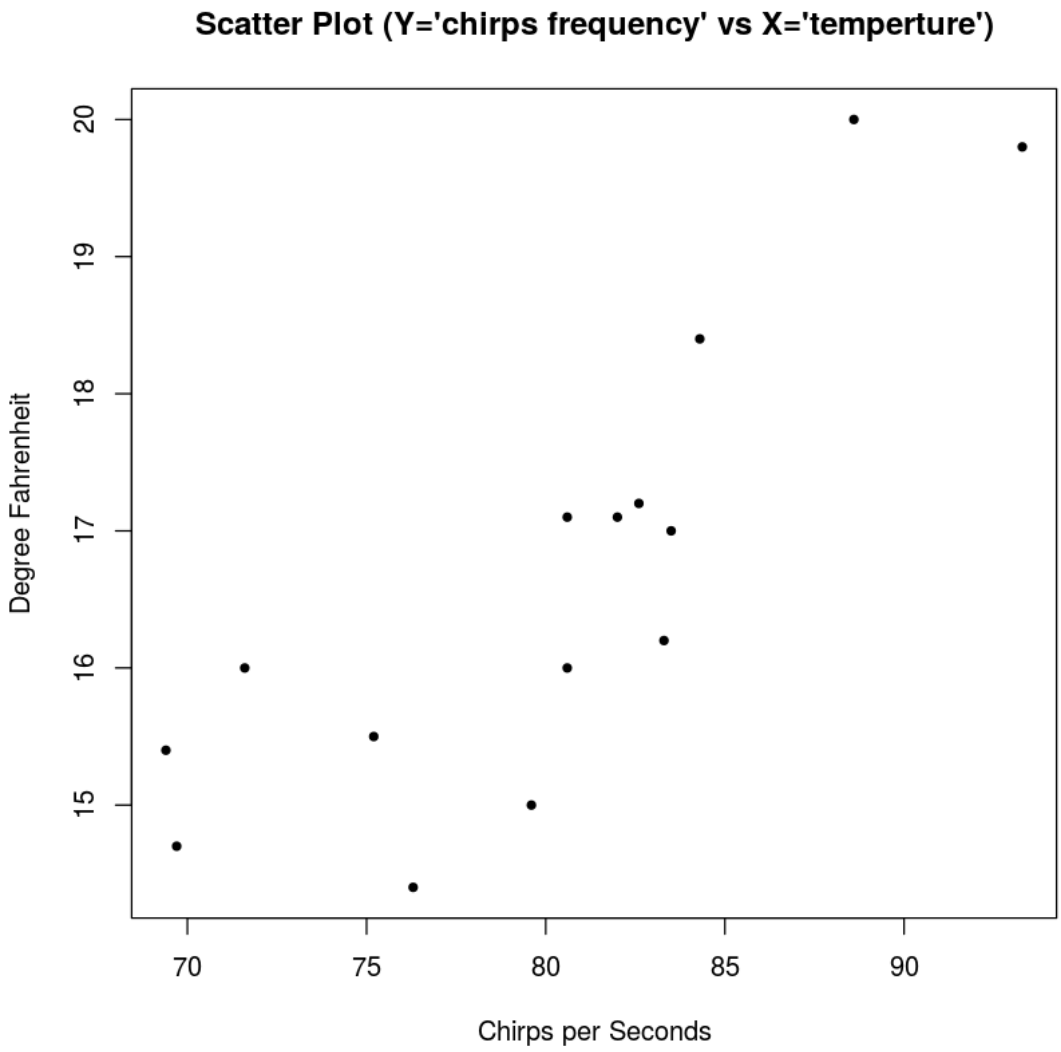
(i) Obtain a plot of the ordered residuals against the corresponding quantiles from the standard normal distribution.

```
In [35]: # initialization
         dat <- dat_cricket
         n     <- nrow(dat)
         x     <- dat$x
         y     <- dat$y
         head(dat)
```

y	x
20.0	88.6
16.0	71.6
19.8	93.3
18.4	84.3
17.1	80.6
15.5	75.2

4.0.1 Q3(a) Obtain a scatter plot of these meansurements

```
In [36]: plot(x, y, pch=20,
              main = "Scatter Plot (Y='chirps frequency' vs X='temperture')",
              xlab = "Chirps per Seconds",
              ylab = "Degree Fahrenheit")
```



4.0.2 Q3(b) Specify the simple linear regression model for these data. Identify all parameters in the model, providing interpretation of each.

```
In [37]: # calculation
         xy    <- crossprod(x, y) ; xy <- drop(xy) # sum(x * y)
         xx    <- crossprod(x, x) ; xx <- drop(xx) # sum(x * x)
         sum_x <- sum(x)
         sum_y <- sum(y)
         xbar  <- mean(x)
         ybar  <- mean(y)

         # Parameters
         beta_1 <- (xy - 1/n * sum_x * sum_y) / (xx - 1/n * sum_x * sum_x)
```

```
beta_0 <- 1/n * (sum_y - beta_1 * sum_x)

# another way to calculate beta 1
# sum((x - xbar) * (y - ybar)) / sum((x - xbar) * (x - xbar))

# prediction
yhat <- beta_0 + beta_1 * x

# MSE
sig2 <- 1/(n-2) * sum((y - yhat)^2)

cat("",
    "Beta 0 hat (intercept)  = ", beta_0, "\n",
    "Beta 1 hat (slope)      = ", beta_1, "\n",
    "Estimate Error Var = ", sig2, "\n",
    "Estimate Error std = ", sig2**0.5)
```

```
Beta 0 hat (intercept)  =  -0.3091444
Beta 1 hat (slope)      =  0.211925
Estimate Error Var =  0.9438467
Estimate Error std =  0.9715177
```

4.0.3 Q3(c) Expalin how the interpretation (and the estimate) of the slope parameter changes if temperature is expressed in °C

$$F = C * \frac{9}{5} + 32$$

$\hat{Y}$
$= \hat{\beta_0} + \hat{\beta_1} * \text{X(F)}$
$= \hat{\beta_0} + \hat{\beta_1} * (\text{X(C)} * \frac{9}{5} + 32)$
$= (\hat{\beta_0} + \hat{\beta_1} * 32) + (\hat{\beta_1} * \frac{9}{5}) * \text{X(C)}$

based on the equation, we can estimate the new $\hat{\beta_1}$ as
$\hat{\beta_{new}} = \frac{9}{5} * \hat{\beta_1}$

```
In [43]: beta_1 * 9 / 5
```

0.381465016289957

The new β_1 estimates the average increase of chirps frequency when you increase temperature one unit, given all other factors fixed.

4.0.4 Q3(d) Estimate the mean chirp frequency among crickets in a temperature of 80°F. Estimate the standard deviation among chirp frequency measurements made at this xed temperature.

```
In [38]: # Estimate mean chirp frequency among crickets in a temperature of 80F
         beta_0 + beta_1 * 80
```

16.6448563329713

```
In [39]: # Estimate the standard deviation among chirp frequency measurements made at this xed t
         (sig2 * (1 + 1/n + (80 - xbar)^2 / sum((x-xbar)^2)))^0.5
```

1.00338038689305

4.0.5 Q3(e) Estimate the mean chirp frequency among crickets in a temperature of 105F

Since 105F is outside the range of data, we are not sure whether the linear model is applicable in a temperature of 105F

However, if we assume that the linear model still holds for data point at x = 105, then we can estimate the mean chirp frequency conditioned on x=105 as follow:

```
In [40]: # Estimate the mean chirp frequency among crickets in a temperature of 105F
         beta_0 + beta_1 * 105
```

21.9429815592207

4.0.6 Q3(f) Report the sum of squares deviations between the tted values and the average chirp frequency $\bar{Y}$

$$\sum_i (\hat{Y} - \bar{Y})^2$$

```
In [41]: sum((yhat - ybar)^2)
```

28.2873263579728

4.0.7 Q3(g) What proportion of variance in chirp frequencies is explained by the linear regression model?.

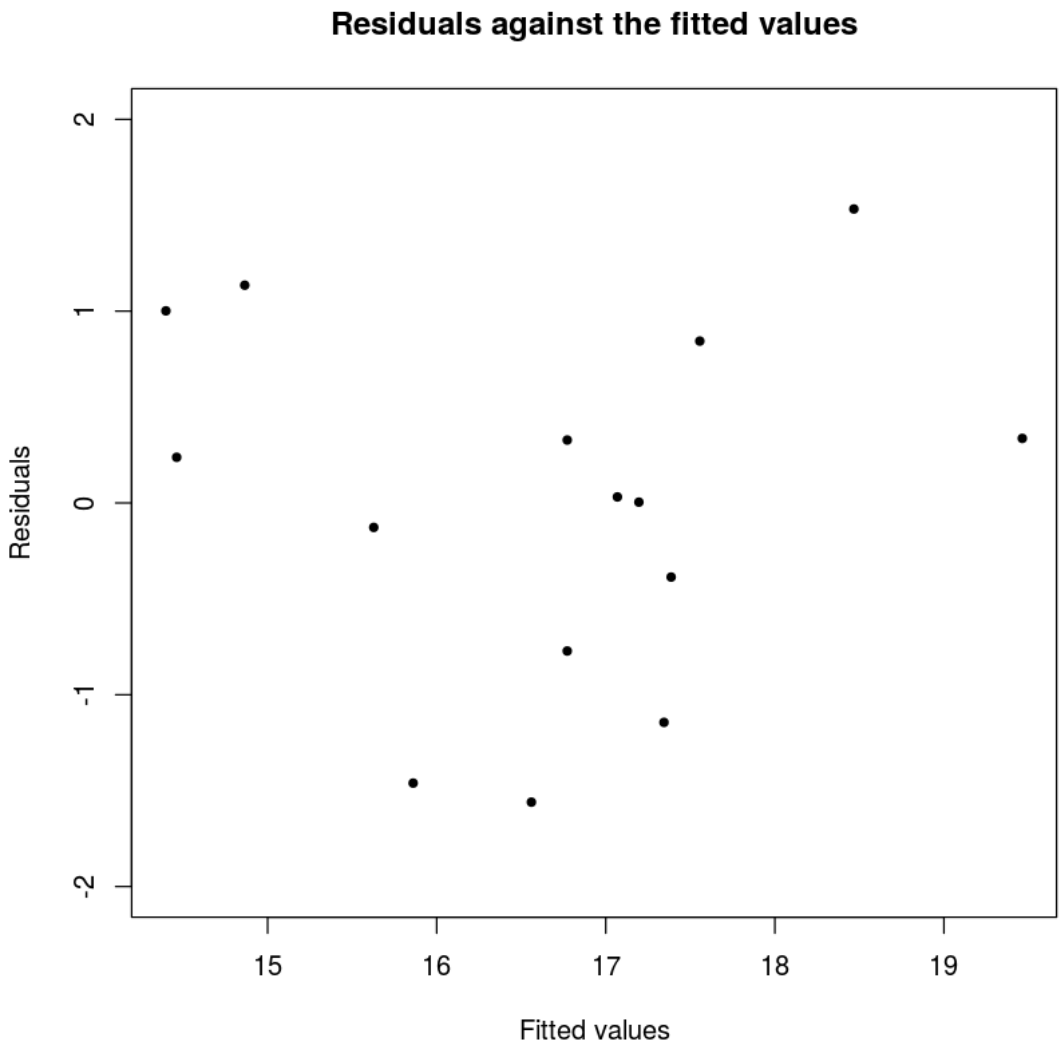
$$R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST}$$

```
In [42]: SST <- sum((y    - ybar)^2)
         SSR <- sum((yhat - ybar)^2)
         SSE <- sum((y    - yhat)^2)

         cat("",
             "SST      =", SST, "\n",
             "SSR      =", SSR, "\n",
             "SSE      =", SSE, "\n",
             "R2       =", 1 - SSE/SST)
             #"R2 (adj) =", 1 - (SSE/(n-2)) / (SST/(n-1)), "\n")
```

```
 SST      = 40.55733
 SSR      = 28.28733
 SSE      = 12.27001
 R2       = 0.6974651
```

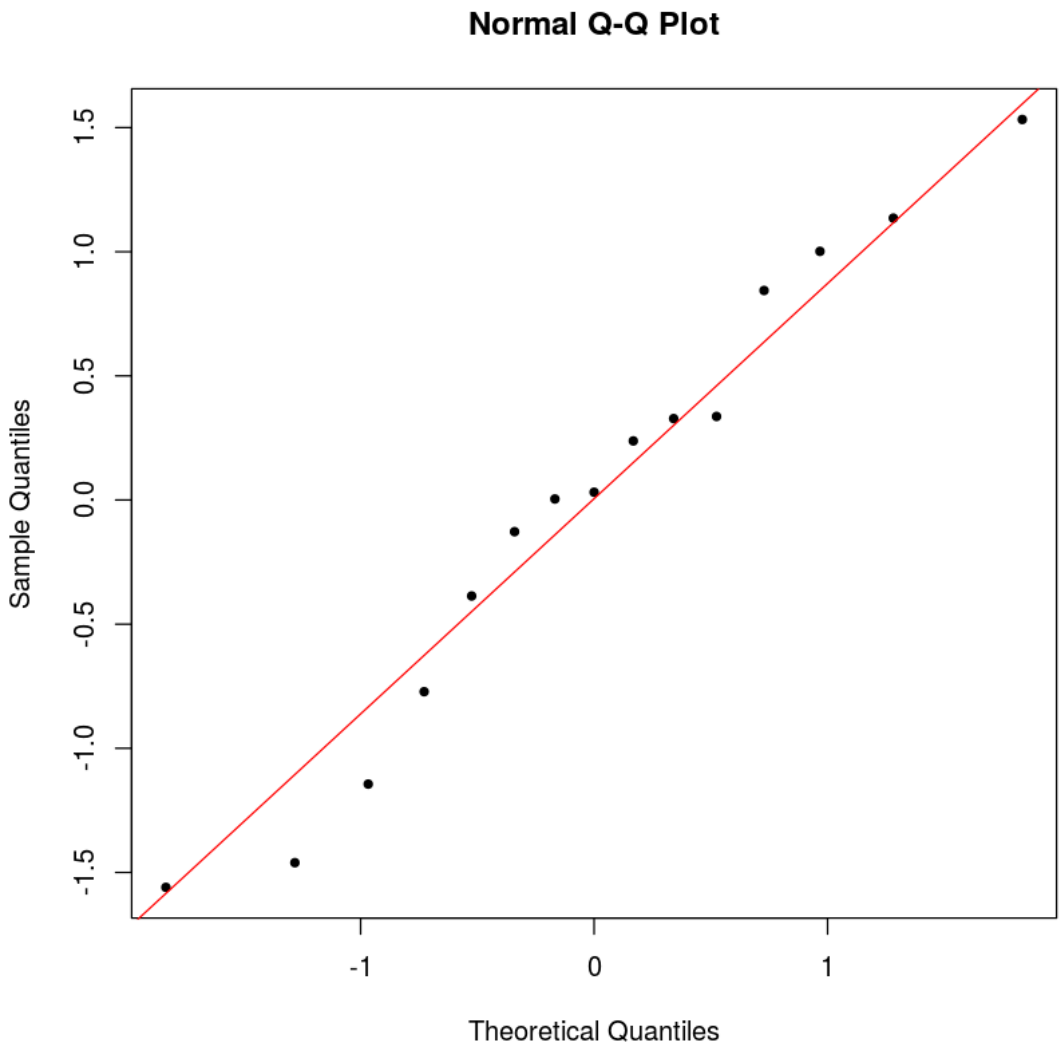
4.0.8 Q3(h) Obtain a plot of the residuals against the tted values.

```
In [11]: plot(yhat, y - yhat, pch=20,
              main = "Residuals against the tted values",
              xlab = "Fitted values",
              ylab = "Residuals",
              ylim = c(-2, 2))
```



4.0.9 Q3(i) Obtain a plot of the ordered residuals against the corresponding quantiles from the standard normal distribution.

```
In [12]: qqnorm(y - yhat, pch = 20)
         qqline(y - yhat, col = "red")
```



5 Question 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

1) Obtain a scatterplot of these data.

2) What is the meaning, in words, of β_1?

3) What is the observed value of $\hat{\beta}_1$?

4) 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)

5) What is a region of plausible values for β_1 suggested by the data?

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

7) How much of the observed vaiation is the heights of sons (the y-axis) is explained by this "best" line?

8) What is the estimated average height of sons whose midparent height is x = 68?

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

10) Under the model, what is the true average height of sons with midparent height is x = 68?

11) 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"?

12) 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?

13) What is th estimated standard error of the estimated average for sons with mid-parent height x = 68, $\hat{\mu}(68) = \hat{\beta}_0 + 68\hat{\beta}_1$? Provide an expression for this standard error.

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

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

16) 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?

17) 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

18) Define $\mu_Y, \sigma_Y, \mu_X, \sigma_X, \rho$. Parameters or statistics?

19) What are the plausible values for ρ suggested by the data?

20) Are the residuals $e_1, e_2, ..., e_{928} \overset{iid}{\sim} N(\theta, \sigma_2)$ a reasonable assumption?

```
In [47]: # initialization
         dat <- dat_parents_offspring
         n     <- nrow(dat)
         x     <- dat$midparent_height
         y     <- dat$offspring_height
         head(dat, 3)
```

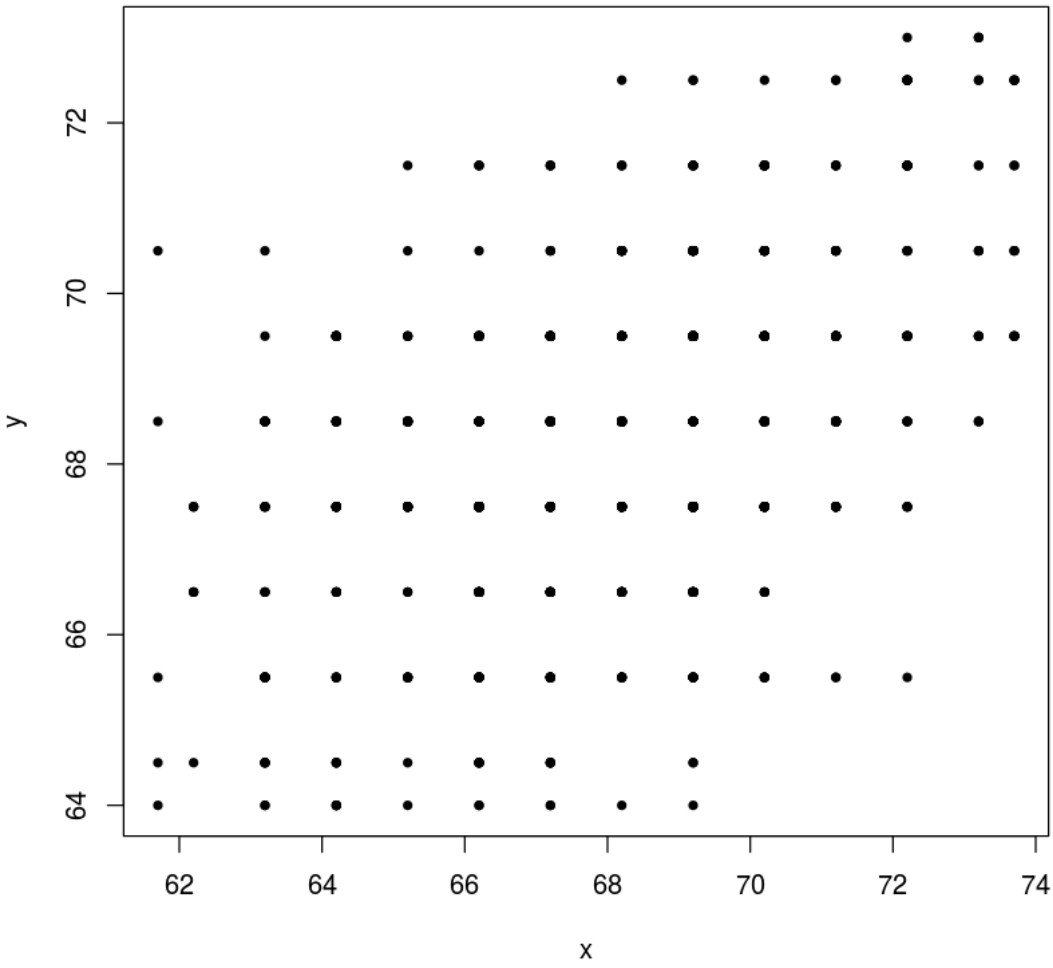
midparent_height	offspring_height
69.5	70.2
67.5	70.2
68.5	67.2

In [48]:
```
# calculate the parameters
xy    <- crossprod(x, y) ; xy <- drop(xy) # sum(x * y)
xx    <- crossprod(x, x) ; xx <- drop(xx) # sum(x * x)
sum_x <- sum(x)
sum_y <- sum(y)
xbar  <- mean(x)
ybar  <- mean(y)

# beta
beta_1 <- (xy - 1/n * sum_x * sum_y) / (xx - 1/n * sum_x * sum_x)
beta_0 <- 1/n * (sum_y - beta_1 * sum_x)

# MSE
yhat <- beta_0 + beta_1 * x
sig2 <- 1/(n-2) * sum((y - yhat)^2)

cat("",
    "Beta 0 hat (intercept)  = ", beta_0, "\n",
    "Beta 1 hat (slope)      = ", beta_1, "\n",
    "Estimate Error Var = ", sig2, "\n",
    "Estimate Error std = ", sig2**0.5)
```

```
Beta 0 hat (intercept)  =  23.94153
Beta 1 hat (slope)      =  0.6462906
Estimate Error Var =  5.011094
Estimate Error std =  2.238547
```

5.0.1 Q4 (1) Obtain a scatterplot of these data.

In [38]:
```
# heights of sons (Y in inches) and average parents height (X in inches)
plot(x, y, pch = 20)
```



5.0.2 Q4 (2) What is the meaning, in words, of β_1?

There are two ways to intepret β_1
- Average change in y, the height of offspring, as you change one unit of x, the height of parents. - When you have two populations with a unit difference in the heights of parents, you will expected to get a β_1 difference of the height of offspring

5.0.3 Q4 (3) What is the observed value of $\hat{\beta}_1$?

```
In [7]: cat("Beta 1 hat (slope) = ", beta_1)
```

```
Beta 1 hat (slope) =  0.6462906
```

5.0.4 Q4 (4) 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)

$$\text{var}(\hat{\beta}_1) = \frac{\sigma^2}{\sum(x_i - \bar{x})^2}$$

$$\widehat{\text{var}(\hat{\beta}_1)} = \frac{s^2_{y.x}}{\sum(x_i - \bar{x})^2}$$

```
In [8]: sig2 <- 1/(n-2) * sum((y - yhat)^2)
        var_beta_1 <- sig2 / sum((x - xbar)^2)
        cat("Standard Error of Statistic 'beta 1 hat' =", var_beta_1^0.5)
```

```
Standard Error of Statistic 'beta 1 hat' = 0.04113588
```

5.0.5 Q4 (5) What is a region of plausible values for β_1 suggested by the data?

95% Confidence interval of $\hat{\beta}_1$

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

```
In [9]: qt(0.975, df = n - 2)
```

1.96252912737767

```
In [19]: t_975 <- qt(0.975, df = n - 2)
         cat(
             "(",
                 beta_1 - t_975 * var_beta_1^0.5,
             ",",
                 beta_1 + t_975 * var_beta_1^0.5,
             ")")
```

```
( -0.9993268 , 1.650622 )
```

5.0.6 Q4 (6) What is the line that best fits these data, using criterion that smallest sum of squared residuals is "best?"

The line that best fits these data is $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 * x$.

By minimizing the sum of squared residuls, the calculation of $\hat{\beta}_0$ and $\hat{\beta}_1$ can be derived to the following formula:

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 * \bar{x}$$

$$\hat{\beta}_1 = \sum \frac{(y_i - \bar{y})(x_i - \bar{x})}{(x_i - \bar{x})(x_i - \bar{x})}$$

The values are of $\hat{\beta}_0$ and $\hat{\beta}_1$ is shown as follow

```
In [20]: cat("",
              "Beta 0 hat (intercept)  = ", beta_0, "\n",
              "Beta 1 hat (slope)      = ", beta_1)
```

```
 Beta 0 hat (intercept)  =  46.13535
 Beta 1 hat (slope)      =  0.3256475
```

5.0.7 Q4 (7) How much of the observed vaiation is the heights of sons (the y-axis) is explained by this "best" line?

The proportion of observed variation explained by regression:

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

```
In [10]: SST <- sum((y    - ybar)^2)
         SSR <- sum((yhat - ybar)^2)
         SSE <- sum((y    - yhat)^2)

         cat("",
             "SST      =", SST, "\n",
             "SSR      =", SSR, "\n",
             "SSE      =", SSE, "\n",
             "R2       =", 1 - SSE/SST)
             #"R2 (adj) =", 1 - (SSE/(n-2)) / (SST/(n-1)), "\n")
```

```
 SST      = 5877.207
 SSR      = 1236.934
 SSE      = 4640.273
 R2       = 0.2104629
```

5.0.8 Q4 (8) What is the estimated average height of sons whose midparent height is x = 68?

$$\hat{\mu_x} = \hat{\beta_0} + \hat{\beta_1} * x$$

```
In [11]: beta_0 + beta_1 * 68
```

67.8892897559802

5.0.9 Q4 (9) Is this the true average height in the whole population of sons whose midparent height is x = 68?

$$\mu_x = E[Y|X = x]$$
$$\hat{\mu_x} = \hat{\beta_0} + \hat{\beta_1} * x$$

No, the value 68.279 is an estimated average. The true average height E[Y | X=68] is a parameter of the whole population of sons given midparent height equal to 68.

5.0.10 Q4 (10) Under the model, what is the true average height of sons with midparent height is x = 68?

Based on the model,

$$Y_i = \beta_0 + \beta_1 * X_i + \epsilon_i$$

where ϵ follows the standard normal distribution $N(0, \sigma^2)$

Given β_0, β_1, and X_i as constants, the distribution of Y is

$$Y_i \sim N(\beta_0 + \beta_1 * X_i, \sigma^2)$$

Therefore, under the model, the true average height of sons with midparent height is x = 68 is

$$\beta_0 + \beta_1 * 68$$

5.0.11 Q4 (11) 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 Y follows the distribution

$$Y_i \sim N(\beta_0 + \beta_1 * X_i, \sigma^2)$$

we know that the esitmated standard deviatiton among the population of sons whose midparent height is x = 68 is σ

we can estimate the σ^2 by $\hat{\sigma^2} = s^2_{y.x}$

$$s^2_{y.x} = \frac{1}{n-2}\sum_i (Y_i - \hat{Y})^2$$

Note that $s^2_{y.x}$ is the MSE of the data

The MSE is not standard error because MSE is not the standard deviation of a statistics, but of the population of sons is conditioned on x = 68 (such standard deviation is used in the prediction invertal).

```
In [12]: sig2^0.5
```

2.23854719318213

5.0.12 Q4 (12) 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?

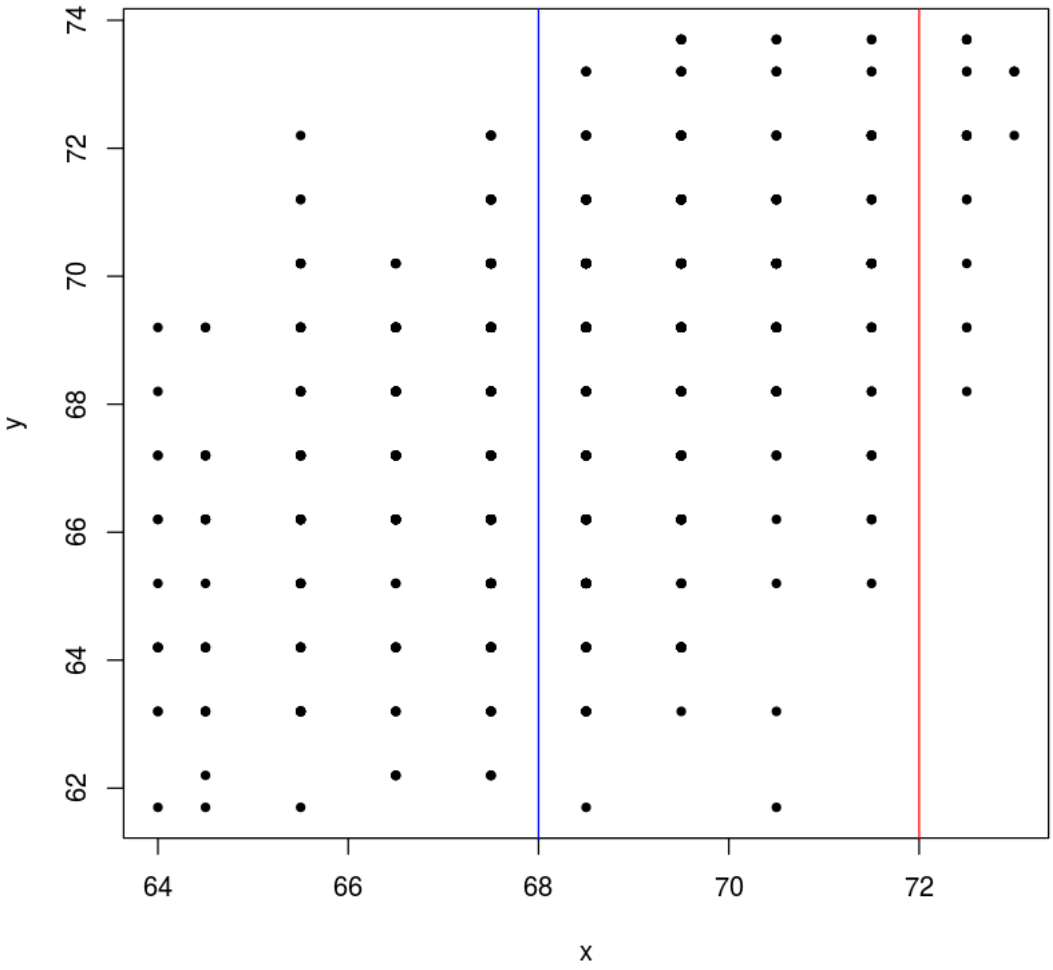
Since Y follows the distribution

$$Y_i \sim N(\beta_0 + \beta_1 * X_i, \sigma^2)$$

Since the it turns out that the the estimated standard deviation among the population of sons whose midparent height is x = 72 is the same as that for x = 68, which is equal to $s_{y.x}$.

No, from the scatterplot below, the variation of y conditioned on x = 68 is larger than that on x = 72.

```
In [26]: plot(x, y, pch = 20)
         abline(v = 68, col = "blue")
         abline(v = 72, col = "red")
```



5.0.13 Q4 (13) What is the estimated standard error of the estimated average for sons with mid-parent height x = 68, $\hat{\mu}(68) = \hat{\beta}_0 + 68\hat{\beta}_1$? Provide an expression for this standard error.

$$\text{var}(\hat{\mu_x}) = \sqrt{\sigma^2 * [\frac{1}{n} + \frac{(x - \bar{x})^2}{\sum(x_i - \bar{x})^2}]}$$

```
In [28]: (sig2 * (1/n + (68 - xbar)^2 / sum((x-xbar)^2)))^0.5
```

1.70008651544676

5.0.14 Q4 (14) Is the estimated standard error of $\hat{\mu}(72)$ bigger, smaller, or the same as that for $\hat{\mu}(68)$

$$\text{var}(\hat{\mu_x}) = \sqrt{\sigma^2 * [\frac{1}{n} + \frac{(x - \bar{x})^2}{\sum(x_i - \bar{x})^2}]}$$

```
In [22]: cat("",
             "x = 68 ---", (sig2 * (1/n + (68 - xbar)^2 / sum((x - xbar)^2)))^0.5, "\n",
             "x = 72 ---", (sig2 * (1/n + (72 - xbar)^2 / sum((x - xbar)^2)))^0.5)
```

```
 x = 68 --- 0.07456949
 x = 72 --- 0.1687102
```

Yes, the estimated standard error of $\hat{\mu}(72)$ bigger than that for $\hat{\mu}(68)$

5.0.15 Q4 (15) Is the observed linear association between son's height and midparent height strong? Report a test statistic.

```
In [49]: 1 / (n - 1) * sum((y - ybar) * (x - xbar)) / sd(x) / sd(y)
```

0.458762368292822

```
In [24]: # Note: the result is the same as the value calculated by the built-in function in R
         cor(x, y)
```

0.458762368292822

```
In [50]: r <- 1 / (n - 1) * sum((y - ybar) * (x - xbar)) / sd(x) / sd(y)
         (r^2)^0.5
```

0.458762368292822

Since the ρ range from -1 to 1, we can square it and then take a square root of it to make it range from 0 to 1. It turns out the result is 0.458, which is not a strong linear association.

5.0.16 Q4 (16) 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?

Parameters

$$\rho = \frac{E[(Y - \mu_Y)(X - \mu_X)]}{\sigma_x \sigma_y}$$

Statistics

$$\hat{\rho} = \frac{\sum(Y_i - \bar{Y})(X_i - \bar{X})}{s_x s_y}$$

5.0.17 Q4 (17) 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

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

Yes, it is. The correlation coefficient ρ is the standardized covariance.

$$\text{Cov(X, Y)} = E[(Y - \mu_y)(X - \mu_x)]$$

$$\rho = \text{Cor(X, Y)} = \frac{\text{Cov(X, Y)}}{\sigma_Y \sigma_X}$$

5.0.18 Q4 (18) Define $\mu_Y, \sigma_Y, \mu_X, \sigma_X, \rho$. Parameters or statistics?

$\mu_Y, \sigma_Y, \mu_X, \sigma_X, \rho$ are all parameters

- μ_Y
  - The mean of the variable Y in population
- σ_Y
  - The variance of the variable Y in population
- μ_X
  - The mean of the variable X in population
- σ_X
  - The variance of the variable X in population
- ρ
  - The correlation between the random variable X and Y

5.0.19 Q4 (19) What are the plausible values for ρ suggested by the data?

$$(\frac{\frac{1+r}{1-r}e^{\frac{-2z}{\sqrt{n-3}}} - 1}{\frac{1+r}{1-r}e^{\frac{-2z}{\sqrt{n-3}}} + 1}, \frac{\frac{1+r}{1-r}e^{\frac{2z}{\sqrt{n-3}}} - 1}{\frac{1+r}{1-r}e^{\frac{2z}{\sqrt{n-3}}} + 1})$$

In [33]:
```
r <- beta_1 * sd(x) / sd(y) # cor(x, y)
r
```

0.458762368293266

Calculate the 95% CI of correlation

In [31]:
```
tmp1 <- (1 + r) / (1 - r)
tmp2 <- exp(2 * qnorm(0.975) / (n - 3)^0.5)

cat("(",
    (tmp1 * 1/tmp2 - 1) / (tmp1 * 1/tmp2 + 1)
    ,",",
    (tmp1 * tmp2 - 1) / (tmp1 * tmp2 + 1)
    ,")")
```

```
( 0.4064067 , 0.5081153 )
```

5.0.20 Q4 (20) Are the residuals $e_1, e_2, ..., e_{928} \overset{iid}{\sim} N(\theta, \sigma_2)$ a reasonable assumption?

Yes, from the q-q plot, since the data points forms in a line, it is reasonable to make such assumption.

```
In [25]: qqnorm(y - yhat, pch = 20)
         qqline(y - yhat, col = "red")
```