Assignment 2

Rin Meng
Student ID: 51940633

October 27, 2024

1. Given the summary output, it is true that:

   (a)
   $$\text{F-Statistic} = 52.77$$
   $$\text{MSE} = (\text{Residual Standard Error})^2 = 2.792^2 = 7.80$$

   (b) To build the anova table, we need to do calculations as follows:
   Finding the degrees of freedom:
   $$DF_{Reg} = 1$$
   $$DF_{Error} = 7$$
   $$DF_{Total} = 8$$
   $$SS_{Total} = SS_{Reg} + SS_{Error}$$
   Calculate the Mean Squares Reg and Error:
   $$F = \frac{MS_{Reg}}{MS_{Error}} = 52.77$$
   $$MS_{Error} = MSE = 7.80$$
   $$MS_{Reg} = F \cdot MS_{Error} = 52.77 \cdot 7.80 = 411.34$$
   Calculate the Sum Squares Reg and Error:
   $$MS_{Reg} = \frac{SS_{Reg}}{DF_{Reg}} \Leftrightarrow SS_{Reg} = MS_{Reg} \cdot DF_{Reg}$$

$$MS_{Error} = \frac{SS_{Error}}{DF_{Error}} \Leftrightarrow SS_{Error} = MS_{Error} \cdot DF_{Error}$$
$$SS_{Total} = SS_{Reg} + SS_{Error}$$
$$SS_{Reg} = MS_{Reg} \cdot DF_{Reg} = 411.34 \cdot 1 = 411.34$$
$$SS_{Error} = MS_{Error} \cdot DF_{Error} = 7.80 \cdot 7 = 54.60$$
$$SS_{Total} = SS_{Reg} + SS_{Error} = 411.34 + 54.60 = 465.94$$

Source	DF	SS	MS	F
Reg.	1	411.34	411.34	52.77
Error	7	54.60	7.80	
Total	8	465.94		

(c) The anova function returns:

```
Analysis of Variance Table

Response: R
        Df Sum Sq Mean Sq F value    Pr(>F)
W          1 411.42  411.42  52.767 0.0001679 ***
Residuals  7  54.58    7.80
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01
‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Our hand calculations are fairly consistent with the output from the anova function in R, minus a few rounding errors.

(d) Calculating the $\sqrt{F}$ value:

$$\sqrt{F} = \sqrt{52.77} = 7.27$$

The t value for $\hat{\beta}_1$ is 7.264, which is very close to the $\sqrt{F}$ value. This is expected because, for simple linear regression with one predictor, the square of the t-value for the slope is equal to the F-statistic

$$\sqrt{F} = t \Leftrightarrow F = t^2$$

2. Given

(a) - **First task time**: ϵ is exponentially distributed with mean $\frac{1}{\lambda}$, so $E[\epsilon] = \frac{1}{\lambda}$.

- **Second task time**: Proportional to x, with a proportionality constant β. So the time required should be βx.
- **Total time**: sum of times it takes to complete the two tasks impying that the total time is $y = \beta x + \epsilon$.

Then the final linear model would be

$$y = \beta x + \epsilon$$

(b) To derive the maximum likelihood estimator for β and λ, we need to find the pdf of the exponential distribution.

$$f(\epsilon) = \lambda e^{-\lambda\epsilon} \text{ for } x \geq 0$$

$$\Rightarrow f(y) = \lambda e^{-\lambda(y-\beta x)} \text{ for } y \geq \beta x$$

Deriving maximum likelihood estimator for β:

$$\begin{aligned}
L(\beta, \lambda) &= \prod_{i=1}^{n} \lambda e^{-\lambda(y_i - \beta x_i)} \\
\log(\beta, \lambda) = \ell(\beta, \lambda) &= \sum_{i=1}^{n} \log(\lambda e^{-\lambda(y_i - \beta x_i)}) \\
\frac{\partial \ell(\beta, \lambda)}{\partial \beta} &= \sum_{i=1}^{n} \frac{\partial}{\partial \beta} \left(\log(\lambda e^{-\lambda(y_i - \beta x_i)})\right) \\
\frac{\partial \ell(\beta, \lambda)}{\partial \beta} &= \sum_{i=1}^{n} \frac{\partial}{\partial \beta} \left(\log \lambda - \lambda(y_i - \beta x_i)\right) \\
\frac{\partial \ell(\beta, \lambda)}{\partial \beta} &= \sum_{i=1}^{n} -x_i\lambda + x_i\lambda\beta \\
\frac{\partial \ell(\beta, \lambda)}{\partial \beta} &= \sum_{i=1}^{n} -x_i\lambda + x_i\lambda\beta = 0 \\
\sum_{i=1}^{n} x_i\lambda &= \sum_{i=1}^{n} x_i\lambda\beta \\
\beta &= 1
\end{aligned}$$

Deriving maximum likelihood estimator for λ:

$$L(\beta, \lambda) = \prod_{i=1}^{n} \lambda e^{-\lambda(y_i - \beta x_i)}$$

$$\log L(\beta, \lambda) = \ell(\beta, \lambda) = \sum_{i=1}^{n} \log \lambda e^{-\lambda(y_i - \beta x_i)}$$

$$\frac{\partial \ell(\beta, \lambda)}{\partial \lambda} = \sum_{i=1}^{n} \frac{\partial}{\partial \lambda} \left(\log \lambda e^{-\lambda(y_i - \beta x_i)}\right)$$

$$\frac{\partial \ell(\beta, \lambda)}{\partial \lambda} = \sum_{i=1}^{n} \frac{\partial}{\partial \lambda} \left(\log \lambda - \lambda(y_i - \beta x_i)\right)$$

$$\frac{\partial \ell(\beta, \lambda)}{\partial \lambda} = \sum_{i=1}^{n} \frac{1}{\lambda} - (y_i - \beta x_i)$$

$$\frac{\partial \ell(\beta, \lambda)}{\partial \lambda} = \sum_{i=1}^{n} \frac{1}{\lambda} - (y_i - \beta x_i) = 0$$

$$\sum_{i=1}^{n} \frac{1}{\lambda} = \sum_{i=1}^{n} (y_i - \beta x_i)$$

$$\frac{n}{\lambda} = \sum_{i=1}^{n} (y_i - \beta x_i)$$

$$\lambda = \frac{n}{\sum_{i=1}^{n} (y_i - \beta x_i)}$$

$\therefore$ The maximum likelihood estimator for β is 1 and λ is $\frac{n}{\sum_{i=1}^{n}(y_i - \beta x_i)}$.

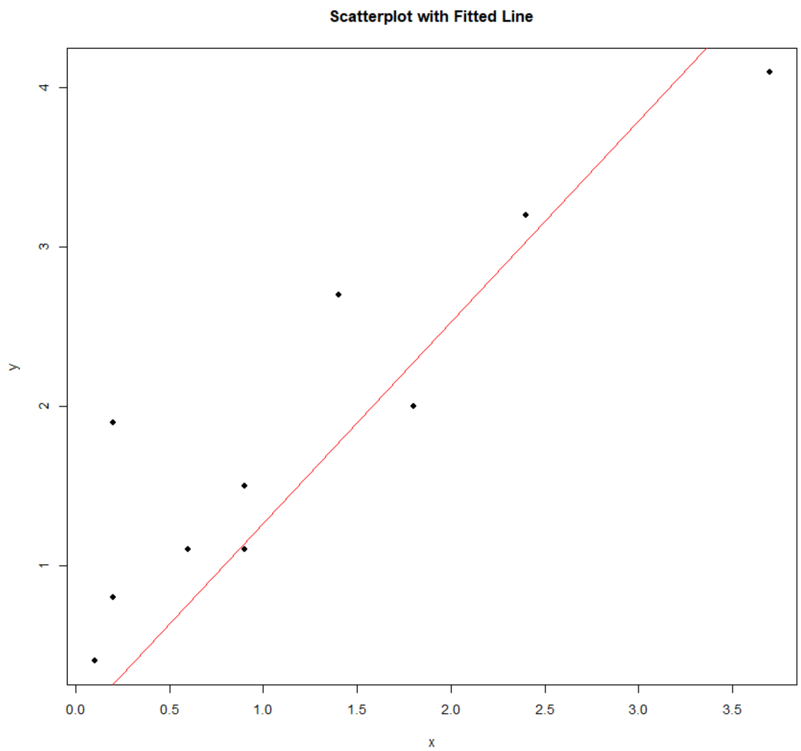
(c) Summary function returns:

```
Call:
lm(formula = y ~ 0 + x, data = data)

Residuals:
     Min       1Q   Median       3Q      Max
-0.58177  0.01169  0.30713  0.50050  1.64693

Coefficients:
  Estimate Std. Error t value Pr(>|t|)
```

```
x   1.2653     0.1389   9.111 7.72e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01
‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error:
0.7179 on 9 degrees of freedom
Multiple R-squared:  0.9022,
Adjusted R-squared:  0.8913
F-statistic: 83 on 1 and 9 DF,
p-value: 7.725e-06
```

(d) To plot the scatter plot, we can run this code:

```
plot(data$x, data$y,
main = "Scatterplot with Fitted Line",
xlab = "x", ylab = "y", pch = 19)
abline(model, col = "red")
```



(e) We can calculate the residuals using the equation $e_i = y_i - \hat{y}_i$ or the function

```
residuals(model)
```

Which will return

```
          1           2           3           4           5
-0.27761976  0.27346557  1.64693114  0.92851796  0.16317365
          6           7           8           9          10
-0.03880988  0.36119012  0.34079341  0.54693114 -0.58177395
```

and the mean of the residuals is 0.336.

Comments:

i. The residuals are not centered around 0 since the mean is 0.336, that means the model may not be a good fit.

ii. The residuals does not seem to be symmetrically distributed around 0 (there are more positives than negatives), therefore, the assumption of normality is not satisfied.

(f) The OLS being used to fit would be biased in this case because:

i. The error term ϵ is not centered at zero (exponential distribution is always positive)

ii. Key assumtion that $E[\epsilon] = 0$ but it's actually $E[\epsilon] = \frac{1}{\lambda}$

iii. For all ϵ, it would be expoentially distributed

For the model, we can express the expected value of y,

$$E[y] = E[\beta x + \epsilon] = \beta x + E[\epsilon] = \beta x + \frac{1}{\lambda}$$

Here, we reap what we sow. When we apply OLS to the model, we will be making an assumtion that $E[\epsilon] = 0$, and we can see that OLS will be estimating β biased towards $\frac{1}{\lambda}$ which would be $E[\epsilon] = 0.336$.

(g) For the least-squares estimate of β, to be unbiased, assuming the regression through the origin model, the condition required should be that

i. The error term ϵ is centered at 0

ii. $Cov(x, \epsilon) = 0$

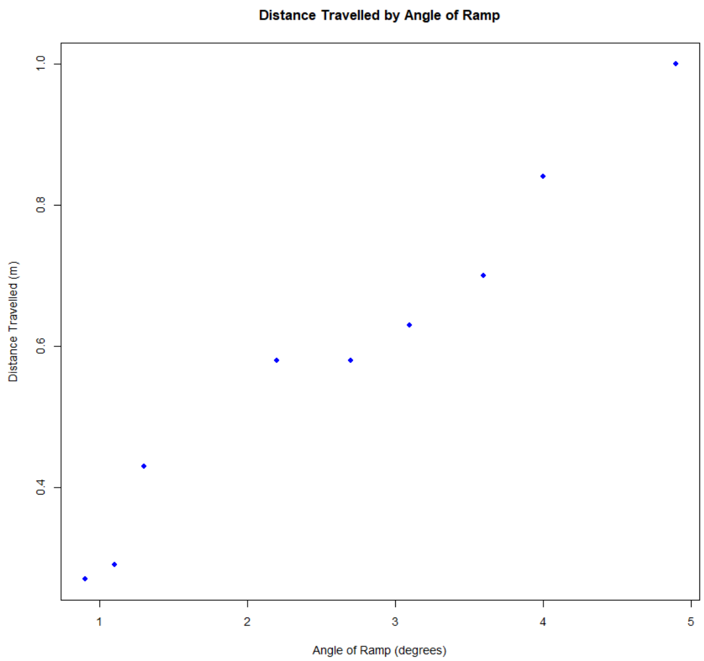
iii. The error term ϵ is normally distributed

Withholding these conditions, the OLS estimate of β will be unbiased.

3. Ploting the data, we can run this code:

(a)
```
# Define the vectors for angle and distance
angle <- c(1.3, 4.0, 2.7, 2.2, 3.6,
             4.9, 0.9, 1.1, 3.1)
distance <- c(0.43, 0.84, 0.58, 0.58,
            0.70, 1.00, 0.27, 0.29, 0.63)
plot(angle, distance,
        xlab = "Angle of Ramp (degrees)",
        ylab = "Distance Travelled (m)",
        main = "Distance Travelled by Angle of Ramp",
        pch = 19, col = "blue")
```

Which will return



Where angle is our predictor y (angle), and distance is our response x (distance).

Is a linear model reasonable?

i. On physical grounds, we know that a steep ramp angle could result in higher gravitational component acting along the ramp, so we would expect some increase in the distance travelled, but should not be perfectly linear becauuse of other factors like friction, air resistance, etc.

ii. On statistical grounds, the since the points are not perfectly linear (noticable curve and dispersion), we can say that the relationship between the angle of the ramp and the distance travelled is not linear, therefore suggesting that a linear model may not be the best fit.

(b) Now we compute, the slope and intercept for the linear model, i.e S_{xx}, S_{xy}, $\bar{y}$, $\bar{x}$. First we recall that

i. $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$

ii. $\bar{y} = \frac{1}{n} \sum_{i=1}^{n} y_i$

iii. $S_{xx} = \sum_{i=1}^{n} (x_i - \bar{x})^2$

iv. $S_{xy} = \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})$

v. $\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}}$

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

Now we can calculate the values:

$$\bar{x} = \frac{23.8}{9} = 2.64$$

$$\bar{y} = \frac{5.32}{9} = 0.59$$

Using `Sxx <- sum((angle - sqr(mean(angle))))`,

$$S_{xx} = 15.48$$

Using `Sxy <- sum((angle - mean(angle))*(distance - ybar))` where, `ybar <- mean(distance)`,

$$S_{xy} = 2.62$$

Using `beta1 <- Sxy / Sxx`,

$$\hat{\beta}_1 = 0.169$$

Using `beta0 <- ybar - beta1 * xbar`,

$$\hat{\beta}_0 = 0.144$$

(c) Now let us provide a 95% confidence interval for the slope and the intercept parameters.

i. The confidence interval for the slope β_1 is given by

$$\hat{\beta}_1 \pm t_{\alpha/2,7} \times SE(\hat{\beta}_1)$$

$t_{\alpha/2,7} = 2.364 =$ `tval <- 2.364`
Using `sebeta1 <- sqrt(MSE / Sxx)`,

$$SE(\hat{\beta}_1) = 0.0123$$

Using `cibeta1 <- c(beta1 - tval * sebeta1, beta1 + tval * sebeta1)`

$$\Rightarrow 0.140 \leq \hat{\beta}_1 \leq 0.198$$

ii. The confidence interval for the intercept β_0 is given by

$$\hat{\beta}_0 \pm t \times SE(\hat{\beta}_0)$$

$t_{\alpha/2,7} = 2.364 =$ `tval <- 2.364`
Using `sebeta0 <- sqrt(MSE * (1/9 + xbar^2 / Sxx)`,

$$SE(\hat{\beta}_0) = 0.043$$

Using `cibeta0 <- c(beta0 - tval * sebeta0, beta0 + tval * sebeta0)`

$$\Rightarrow 0.058 \leq \hat{\beta}_0 \leq 0.230$$

(d) ANOVA Table of whether or not the distance depends on the angle of the ramp, we first need to calculate:

i. $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i =$ `yhat <- beta0 + beta1 * angle`
ii. $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i =$ `ybar <- mean(distance)`
iii. $SST = \sum(y_i - \bar{y})^2 =$ `SST <- sum((distance - ybar)^2)`
iv. $SSE = \sum(y_i - \hat{y}_i)^2 =$ `SSE <- sum((distance - yhat)^2)`
v. $SSR = \sum(\hat{y}_i - \bar{y})^2 =$ `SSR <- SST - SSE`
vi. $DF_R = 1 =$ `dfR <- 1`
vii. $DF_E = n - 2 = 9 - 2 = 7 =$ `dfE <- 7`
viii. $DF_T = n - 1 = 9 =$ `dfT <- 9`
ix. $MSR = \frac{SSR}{DF_R} = \frac{SSR}{1} =$ `MSR <- SSR / dfR`
x. $MSE = \frac{SSE}{DF_E} = \frac{SSE}{n-2} =$ `MSE <- SSE / dfE`
xi. $F = \frac{MSR}{MSE} =$ `F = MSR / MSE`

Using `SST = sum((distance - mean(distance))^2)`

$$SST = 0.462$$

Using `SSE = sum((distance - yhat)^2)`

$$SSE = 0.0166$$

Using `SSR = SST - SSE`

$$SSR = 0.446$$

Using `MSR <- SSR / dfR`

$$MSR = 0.446$$

Using `MSE <- SSE / dfE`

$$MSE = 0.0024$$

Using `F = MSR / MSE`

$$F = 188.6$$

Source	DF	SS	MS	F
Reg.	1	0.446	0.446	188.6
Error	7	0.0166	0.0024	
Total	8	0.462		

On the F-table of critical values for $\alpha = 0.05$,

$$F_{\alpha,DF_1,DF_2} = F_{\alpha,DF_R,DF_E} = F_{0.05,1,7} = 5.59$$

So then, since $F = 188.6 > 5.59$, we reject the null hypothesis that the angle of the ramp significantly affects the distance travelled.

(e) To find a 95% confidence interval for the expected distance at an angle of 2.5 degrees, we have to calculate:

i. $x_0 = 2.5$ `x0 <- 2.5`

ii. $n = 9$

iii. $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i =$ `xbar <- mean(angle)`

iv. $S_{xx} = \sum_{i=1}^{n}(x_i - \bar{x})^2 =$ `Sxx <- sum((angle - xbar)^2)`

v. $s^2 = \frac{SSE}{n-2} =$ `ssquared <- SSE / dfE`

vi. $yhat0 = \beta_0 + \beta_1 x_0 =$ `yhat0 <- beta0 + beta1 * x0`

vii. $t_{\alpha/2,n-2} = t_{0.025,7} = 2.364 =$ `tval <- 2.364`

viii. $\text{SE}(\hat{y}_0) = \sqrt{s^2\left(\frac{1}{n} + \frac{(x_0-\bar{x})^2}{S_{xx}}\right)}$
$=$ `semr <- sqrt(ssquared * (1/9 + (x0 - xbar)^2 / Sxx)`

ix. $\text{CI} = \hat{y}_0 \pm t_{\alpha/2,n-2} \times \sqrt{s^2\left(\frac{1}{n} + \frac{(x_0-\bar{x})^2}{S_{xx}}\right)}$
$=$ `cimr <- c(yhat0 - tval * semr, yhat0 + tval * semr)`

Using `x0 <- 2.5`

$$x_0 = 2.5$$

Using `xbar <- mean(angle)`

$$\bar{x} = 2.64$$

Using `Sxx <- sum((angle - xbar)^2)`

$$S_{xx} = 15.48$$

Using `ssquared <- SSE / dfE`

$$s^2 = 0.0024$$

Using `y0 <- beta0 + beta1 * x0`

$$y_0 = 0.56$$

Using the t-table, $t_{\alpha/2,n-2} = t_{0.025,7} = 2.364$ Using `semr <- sqrt(ssquared * (1/9 + (x0 - xbar)^2 / Sxx)`

$$SE(\hat{y}_0) = 0.0512$$

Using `cimr <- c(yhat0 - tval * semr, yhat0 + tval * semr)`

$$CI = 0.528 \leq \hat{y}_0 \leq 0.605$$

Now calculating the 95% prediction interval for the distance at an angle of 2.5 degrees, we have:

$$\hat{y}_0 \pm t_{\alpha/2,n-2} \times \sqrt{s^2 \left(1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}\right)}$$

Where $1 + \frac{1}{n} + \frac{(x_0-\bar{x})^2}{S_{xx}}$ is the variability in a new single observation, it also makes the prediction interval wider than the confidence interval. Using `sepi <- sqrt(ssquared * (1 + 1/n + (x0 - xbar)^2 / Sxx))`

$$SE(\hat{y}_0) = 0.0513$$

Using `piso <- c(yhat0 - tval * sepi, yhat0 + tval * sepi)`

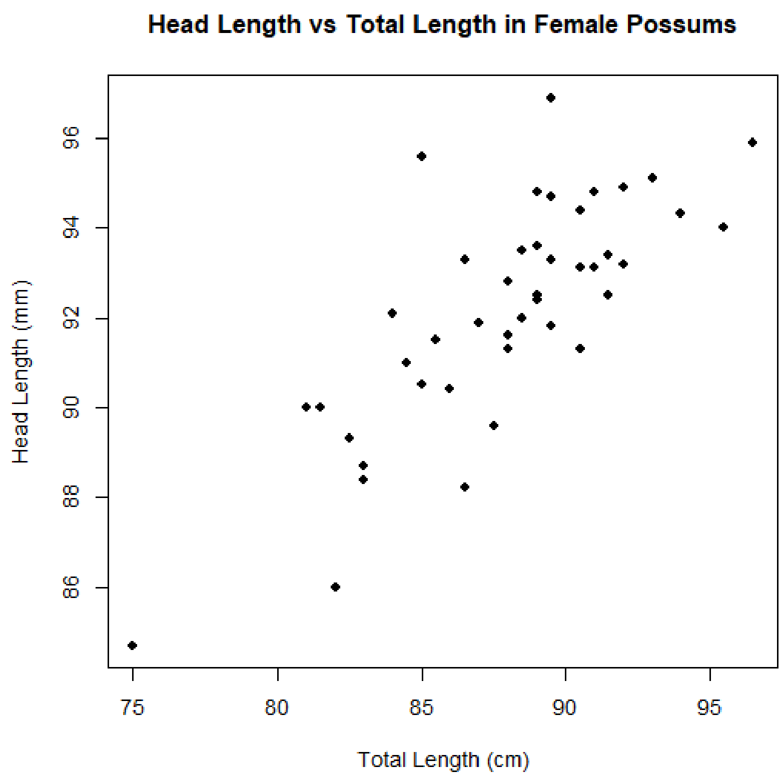
$$PI = 0.445 \leq \hat{y}_0 \leq 0.688$$

4. Using R for this problem, we can run the following code:

(a) Code:

```
library(DAAG)
data("fossum", package = "DAAG")
plot(fossum$totlngth, fossum$hdlngth,
xlab = "Total Length (cm)",
ylab = "Head Length (mm)",
main = "Head Length vs Total Length in Female Possums",
pch = 16, col = "blue")
```

Output:



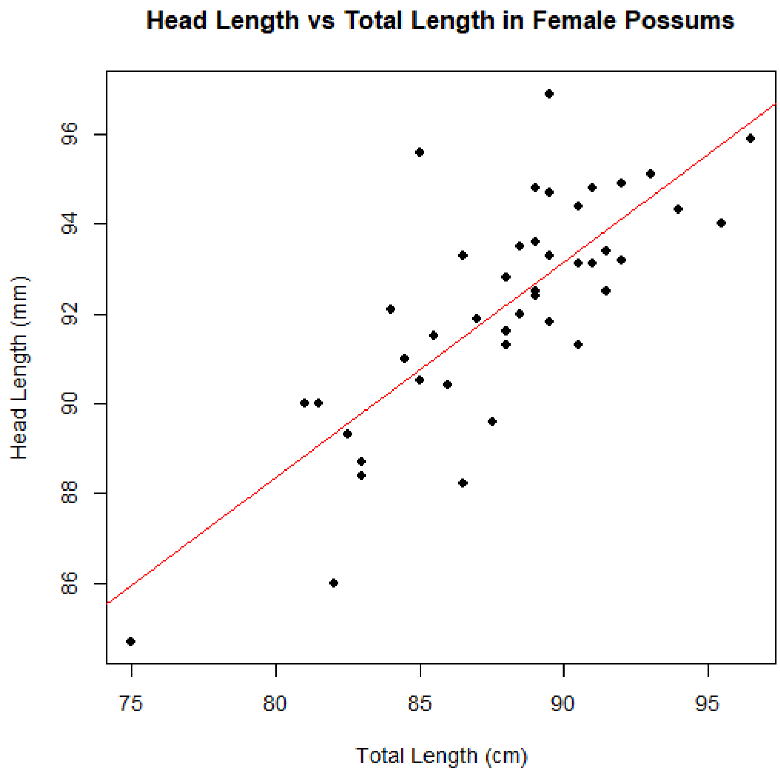
(b) Code:

```
model <- lm(fossum$hdlngth ~ fossum$totlngth)
```

```
summary(model)
```

Output:

```
Call:
lm(formula = hdlngth ~ totlngth, data = fossum)

Residuals:
    Min      1Q  Median      3Q     Max
-3.3149 -0.9030 -0.2548  0.9259  4.8458

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 49.97455    5.30325   9.423 8.14e-12 ***
totlngth     0.47976    0.06026   7.961 7.50e-10 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01
'*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.633 on 41 degrees of freedom
Multiple R-squared:  0.6072,
Adjusted R-squared:  0.5976
F-statistic: 63.38 on 1 and 41 DF,
p-value: 7.501e-10
```



(c) Code:

```
anova(model)
```

```
Analysis of Variance Table

Response: hdlngth
            Df Sum Sq Mean Sq F value    Pr(>F)
totlngth     1 169.09 169.089  63.383 7.501e-10 ***
Residuals 41 109.38   2.668
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01
‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

(d) Code:

```
# F-statistic from ANOVA table
F_statistic <- 63.38

# degrees of freedom
dfR <- 1  # Regression
dfE <- 41  # Error

# critical f-value for a two-tailed test at alpha = 0.05
alpha <- 0.05
F_critical <- qf(1 - alpha, dfR, dfE)

# output
cat("Calculated F-statistic:", F_statistic, "\n")
cat("Critical F-value (two-tailed):", F_critical, "\n")

# conclusion
if (F_statistic > F_critical) {
    cat("Reject the null hypothesis:
There is a significant relationship
between hdlngth and totlngth.\n")
} else {
    cat("Fail to reject the null hypothesis:
There is no significant relationship
between hdlngth and totlngth.\n")
}
```

Output:

```
Calculated F-statistic: 63.38
Critical F-value (two-tailed): 4.078546
Reject the null hypothesis:
There is a significant relationship
between hdlngth and totlngth.
```

(e) Code:

```
new_data <- data.frame(totlngth = 85)
predict(model, new_data, interval = "prediction", level = 0.95)
```

Output:

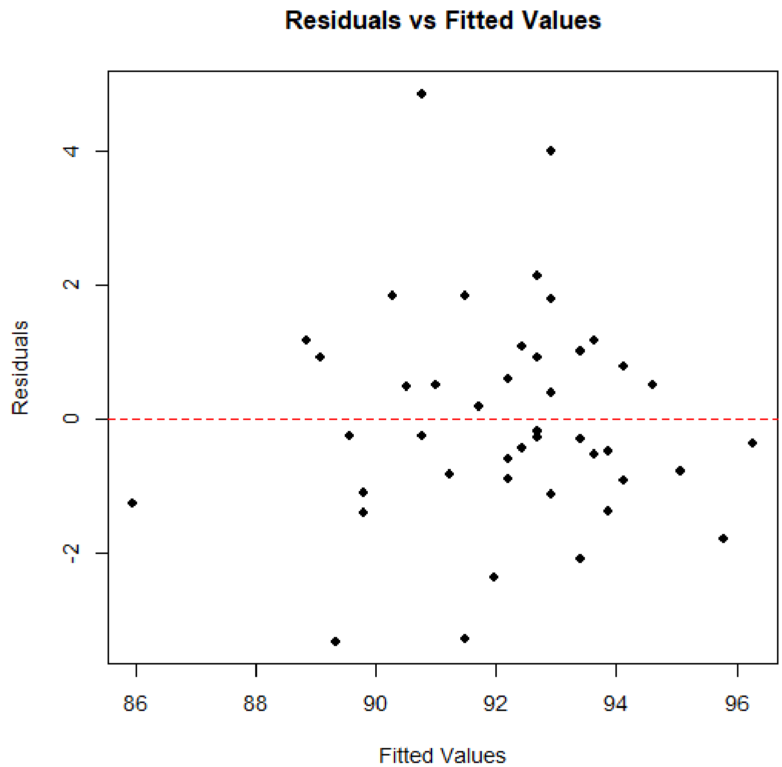
```
fit      lwr      upr
1 90.75419 87.39878 94.10959
```

$$\Rightarrow 87.40 \le \hat{y}_0 \le 94.11$$

(f) Code:

```
plot(fitted(model), resid(model),
xlab = "Fitted Values",
ylab = "Residuals",
main = "Residuals vs Fitted Values",
pch = 16, col = "black")
abline(h = 0, col = "red", lty = 2)
```

Output:



They look fairly symmetrically distributed around 0, but there are some outliers, so we can assume that the residuals are normally distributed.

End of Assignment 2.