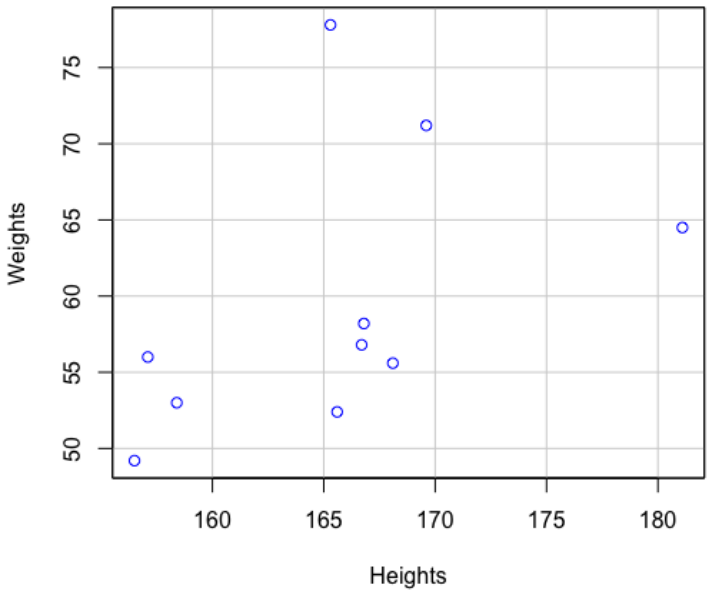
Exercise 1: Do problem 2.1. Complete 2.1.1, 2.1.2, and only do the first two sentences of 2.1.3.

2.1.1 Draw a scatterplot of *wt* on the vertical axis versus *ht* on the horizontal axis. On the basis of this plot, does a simple linear regression model make sense for these data? Why or why not?

Figure 1: Height vs Weight Scatter Plot



Code:

```
> df <- Htwt
> scatterplot(df$ht, df$wt, regLine = FALSE,
              boxplots = FALSE, smooth = FALSE,
              xlab = 'Heights', ylab = 'Weights')
```

The data does not look ideal for a linear regression model. You can see that as the height increase we experience more variance in the weight values. A linear regression model assumes constant variance.

2.1.2 Show that $\bar{x} = 165.52$, and $\bar{y} = 59.37$, $SXX = 472.08$, $SYY = 731.96$, and $SXY = 274.79$. Compute the estimates of the slope and intercept for the regression of Y on X. Draw the fitted line on you scatterplot.

Solution:
Calculating $\bar{x}$ and $\bar{y}$ just applying the mean formula on the data. Computing SXX, SYY, and SXY involves using the following formula,

$$SXX = \sum (x_i - \bar{x})^2,$$

$$SYY = \sum (y_i - \bar{y})^2,$$

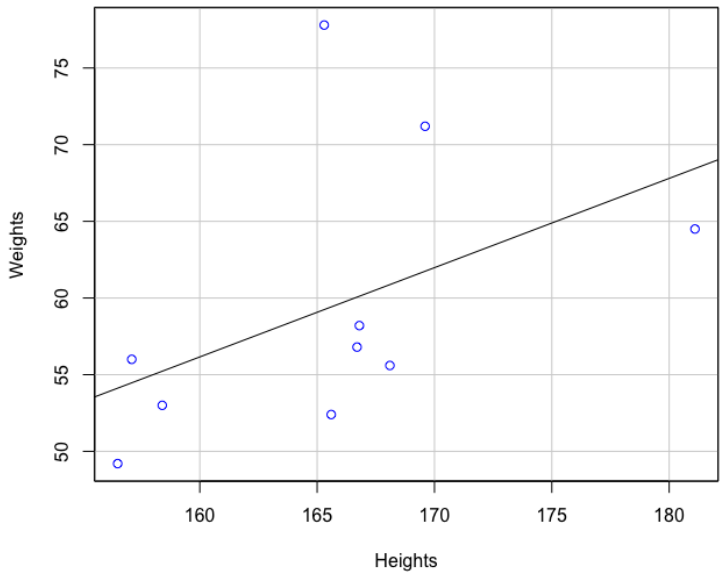
$$SXY = \sum (x_i - \bar{x})(y_i - \bar{y})^2.$$

Given our values we can finally solve for $\hat{\beta}_0$ and $\hat{\beta}_1$

$$\hat{\beta}_1 = \frac{SXY}{SXX} = \frac{274.786}{472.076} = 0.58208$$

$$\hat{\beta}_2 = \bar{y} - \bar{x}\hat{\beta}_1 = 59.37 - 165.52(0.58208) = -36.97588$$

Figure 2: Height vs Weight Scatter Plot with Regression



Code:

```
> x_bar = mean(df$ht)
    [1] 165.52
> y_bar = mean(df$wt)
    [1] 59.47


> SXX = sum((df$ht - x_bar)^2)
    [1] 472.076
> SYY = sum((df$wt - y_bar)^2)
    [1] 731.961
> SXY = sum((df$ht - x_bar)*(df$wt - y_bar))
    [1] 274.786



> b_1 = SXY/SXX
> b_0 = y_bar - x_bar*(b_1)
> scatterplot(df$ht, df$wt, regLine = FALSE,
              boxplots = FALSE, smooth = FALSE,
              xlab = 'Heights', ylab = 'Weights')
> abline(b_0, b_1)
```

2.1.3 Obtain the estimate of σ^2 and find the estimated standard errors of $\hat{\beta}_0$ and $\hat{\beta}_1$. Also find the estimated covariance between $\hat{\beta}_0$ and $\hat{\beta}_1$.

Solution:
Recall that our estimator for σ^2 is computed by,

$$\hat{\sigma}^2 = \frac{RSS}{n-2} = \frac{\sum(y_i - \hat{y}_i)^2}{n-2} = 71.5142$$

Now recall the formulas for computing the standard errors of $\hat{\beta}_0$ and $\hat{\beta}_1$,

$$SE(\hat{\beta}_0) = \sqrt{\hat{\sigma}^2\left(\frac{1}{n} + \frac{\bar{x}^2}{SXX}\right)} = \sqrt{71.5142\left(\frac{1}{10} + \frac{165.52^2}{472.076}\right)} \approx 64.478$$

$$SE(\hat{\beta}_1) = \sqrt{\hat{\sigma}^2\frac{1}{SXX}} = \sqrt{71.5142\frac{1}{472.076}} \approx 0.389$$

Solving for the estimated covariance,

$$Cov(\hat{\beta}_0, \hat{\beta}_1) = -\hat{\sigma}^2\frac{\bar{x}}{SXX} = -71.5142\frac{165.52}{472.076} \approx 25.0744$$

Exercise 2: For the data set in problem 2.1, do the following,

a. Write out the simple linear regression model, including the mean and the variance function.

Solution:
From the previous problem we have already computed the simple linear regression model,

$$y = \hat{\beta}_0 + \hat{\beta}_1 x = -36.976 + 0.582x$$

We also know that the expected value and mean functions,

$$E(Y) = -36.976 + 0.582x,$$

$$V(E(Y)) = \hat{\sigma}^2 \left(\frac{1}{n} + \frac{(x - \bar{x})^2}{SXX} \right) = 71.5142 \left(\frac{1}{10} \frac{(x - 165.52)^2}{472.08} \right).$$

b. Interpret the intercept and slope estimates you obtained above.

Solution:
The intercept of our linear regression model suggests that on average when someone is 0 centimeters then they weigh approximately negative 36 kilos. The slope of our linear regression model suggests that for every 1 centimeter of height we can expect the weight to increase by approximately 0.6 kilos.

c. What does it mean when we call the fitted model the "best" in other words what is the idea behind OLS?

Solution:
The fitted (OLS)model is created by minimizing the residual sum of squares(RSS) function.

$$RSS(\hat{\beta}_0, \hat{\beta}_1) = \sum_{i=1}^{n} (y_i - \hat{\beta}_0 + \hat{\beta}_1 x_i)^2$$

The idea behind this is, if the residual is a good estimate for the error or inaccuracy in the model, it makes sense to create a model by minimizing the residual. Our (OLS) model give the smallest residual or highest accuracy, so it is the 'best' fit for the data.

d. What does it mean to say that the OLS estimators are linear?

Solution:
We say that the OLS estimators are linear because they are produced via a linear combination of y_i. More accurately we would say that the OLS estimators are linear with respect to the response variable.

e. Using the fitted model, predict the weight of a person from this population who is 171.0 cm tall.

Solution:

$$-36.976 + 0.582(171.0cm) = 62.546kg$$

Exercise 3: For the data set brains in the alr4 library, do the following,

a. Fit the simple linear regression model using log(BrainWT) as the response and log(BodyWT) as the predictor. Report the estimated slope and residual standard error of the fitted model.

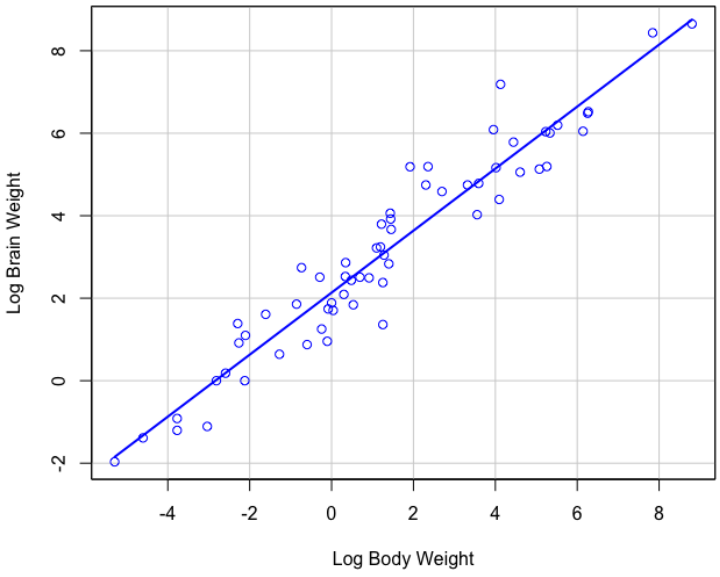
Solution:
Fitting the simple linear regression using r we get the following,

$$\hat{\beta}_1 = 0.75169$$

$$\hat{\beta}_0 = 2.13479$$

$$RSE = 0.6943$$

Figure 3: Log Body-Weight vs Log Brain-Weight Scatter Plot with Regression



Code:

```
> summary(lm(log(df$BrainWt)~log(df$BodyWt)))

Call:
lm(formula = log(df$BrainWt) ~ log(df$BodyWt))

Residuals:
```

```
     Min       1Q   Median       3Q      Max
-1.71550 -0.49228 -0.06162  0.43598  1.94833

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)      2.13479    0.09604   22.23   <2e-16 ***
log(df$BodyWt)   0.75169    0.02846   26.41   <2e-16 ***
---
Residual standard error: 0.6943 on 60 degrees of freedom
Multiple R-squared:  0.9208,    Adjusted R-squared:  0.9195
F-statistic: 697.4 on 1 and 60 DF,  p-value: < 2.2e-16
```

b. What do you know about the point defined by the mean of $log(BodyWT)$ and the mean of $log(BrainWT)$ related to the fitted model.

Solution:
We know, by how $\hat{\beta}_0$ is defined that the mean pair $(\bar{x}, \bar{y})$ lies on the fitted line.

c. If the errors are normally distributed in the population, what are the distributions of the slope estimator and the error variance estimator.

Solution:
As discussed in the lecture, if e_i $N(0, \sigma^2)$ for all i we know that the

$$Y_i \; N(\hat{\beta}_0 + \hat{\beta}_1 x_i, \sigma^2)$$

Since slope estimator and the error variance estimator are a linear combination with respect to y_i then they must also be normally distributed.

d. What is the fitted value for Raccoons? What is the residual?

Solution:
Computed the fitted sample, using the linear model estimators from r.

$$\hat{y}_{raccoon} = 2.1347883 + 0.7516861 * (log(4.288)) = 3.229108$$

Computing the residual, with the given data.

$$e_{raccoon} = |3.229108 - log(39.201)| = 0.439594$$

e. If a new animal species was sampled, which had a $log(BodyWT)$ of 1.46, would the variance of its prediction error match the variance of the fitted value for the racoons? Why or why not?

Solution:
We know that the variance for the fitted value is going to be smaller that the variance of the prediction error. This is true by definition and is shown below, where the variance of the fitted value is given by,

$$V(E(Y)) = V(\hat{y}) = \sigma^2 \left(\frac{1}{n} + \frac{(x - \bar{x})^2}{SXX} \right)$$

and the variance of the prediction error is given by,

$$V(\hat{y} - y^*) = \sigma^2 \left(1 + \frac{1}{n} + \frac{(x - \bar{x})^2}{SXX} \right)$$

This makes sense, as its harder to predict a parameter then it is to estimate one.