Exercise: simple linear regression: body weight and plasma volume. Example data contain the body weight (kg) and plasma volume (literes) for eight healthy men.

1 Estimating model coefficients

weight [kg]	plasma [l]	$x_i - \overline{x}$	$y_i - \overline{y}$	$(x_i - \overline{x})(y_i - \overline{y})$	$(x_i - \overline{x})^2$	$(y_i - \overline{y})^2$	x^2
58.00	2.75						
70.00	2.86						
74.00	3.37						
63.50	2.76						
62.00	2.62						
70.50	3.49						
71.00	3.05						
66.00	3.12						

1. Calculate:

$\overline{y} = \frac{1}{n}\sum_{i=1}^{n} y_i =$

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

2. Fill in columns 3rd to 6th (leave the last 2 columns for now)

3. Calculate $\hat{\beta}_1$:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^{n}(x_i - \overline{x})^2} =$$

4. Calculate $\hat{\beta}_0$:
$\hat{\beta}_0 = \overline{y} - \hat{\beta}_1 \overline{x} =$

5. Write equation for the best-fitting straight line:

2 Accuracy of the coefficient estimates

1. Fill in the remainig columns in the table above

2. Calculate s

$$s = \sqrt{\left[\frac{\sum_{i=1}^{n}(y_i - \overline{y})^2 - \overline{\beta}_1 \sum_{i=1}^{n}(x_i - \overline{x})^2}{n-2}\right]} =$$

3. Calculate $s.e(\hat{\beta}_0) = s * \sqrt{\left[\frac{1}{n} + \frac{x_i^2}{\sum_{i=1}^{n}(x_i - \overline{x})^2}\right]} =$

4. Calculate $s.e(\hat{\beta}_1) = \frac{s}{\sqrt{\sum_{i=1}^{n}(x_i - \overline{x})^2}} =$

5. Have a look at Figure 3.3 in *An Introduction to Statistical Learning* and answer questions

   - What do 10 light blue lines represent on the plot (right)?
   - What is an `unbiased estimator`?
   - Have we underestimated or overestimated β_1?

3 Hypothesis testing

Is there an association between body weight and plasma volume?

1. Write down the null hypothesis and alternative hypothesis

2. Calculate t-statistics for $\hat{\beta}_1$

$t = \frac{\hat{\beta}_1 - 0}{s.e.(\hat{\beta}_1)} =$

3. Use t distribution table containing critical values of the t distribution, to check if whether the p-value for our calculated t-statistics is lower than 5% threshold? Is it lower than 1% threshold?

4. Can we reject the null hypothesis? Is there an association between body weight and plasma volume.

4 Prediction

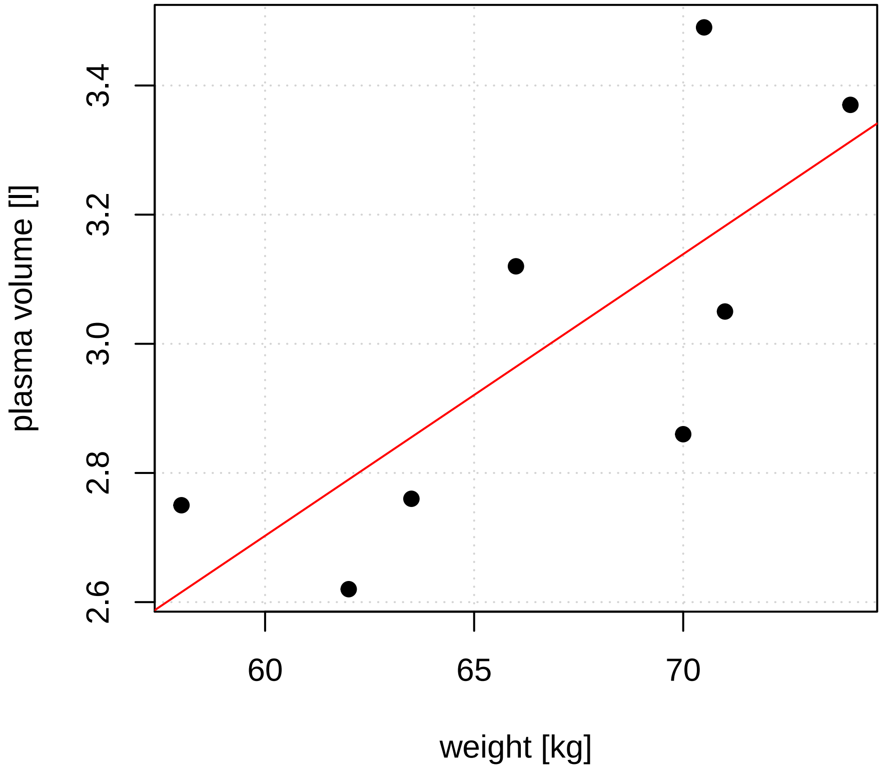


Figure 1: Body weight vs. plasma volume

1. Given Figure 1 predict ‘plasma volume‘ for weight values of 60, 65, and 70 kg
2. Calculate predicted values of ‘plasma volume‘ for weight values of 60, 65 and 70 kg using the equation $y'_i = \hat{\beta}_0 + \hat{\beta}_1 x'_i$

$y'_{60} =$

$y'_{65} =$

$y'_{70} =$

3. Calculate standard error for the predicted ‘plasma volume‘ for weight value of 60 kg

$s.e.(y'_i) = s\sqrt{[1 + \frac{1}{n} + \frac{(x_i - \overline{x_i})^2}{\sum_{i=1}^{n}(x_i - \overline{x_i})^2}]} =$

5 Asesssing the Accuracy of the Model & Correlation

1. Using Given Figure 1 try to calculate (estimate) the RSE. We will check which group gets results closeset to the computed ones.

2. Using lecture and this pen-and-paper docs, calculate R^2, i.e. do not use computer to calculated. Hint: most of the values have been reported / calculated before. It is ok to use mobiles for adding and dividing things up.

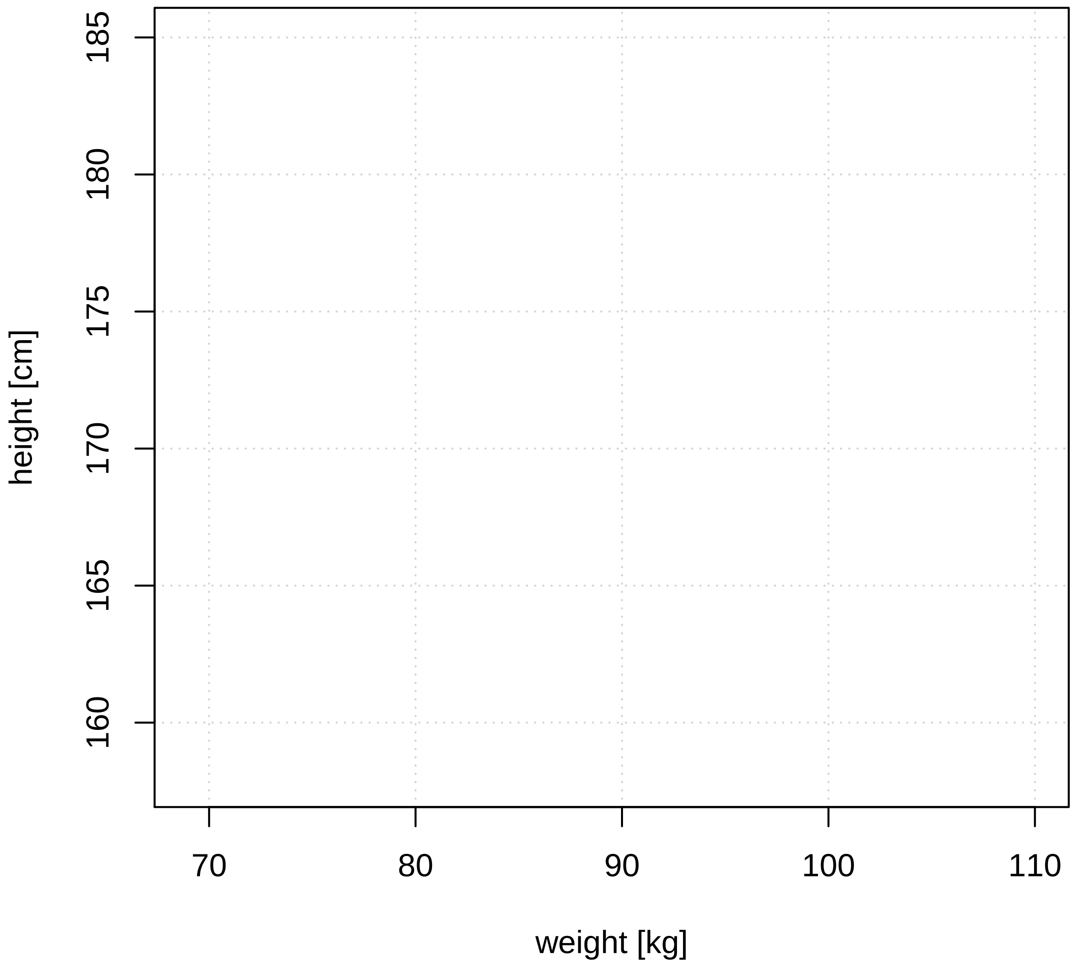
$$R^2 = \frac{TSS-RSS}{TSS} = 1 - \frac{RSS}{TSS} =$$

3. Calculate correlation

$$Cor(X,Y) = \frac{\sum_{i=1}^{n}(x_i-\overline{x})(y_i-\overline{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\overline{x})^2}\sqrt{\sum_{i=1}^{n}(y_i-\overline{y})^2}} = s$$

6 Extra dataset to practise more

weight [kg]	height [cm]	$x_i - \overline{x}$	$y_i - \overline{y}$	$(x_i - \overline{x})(y_i - \overline{y})$	$(x_i - \overline{x})^2$	$(y_i - \overline{y})^2$	x^2
110.00	182.00						
74.00	170.00						
96.00	185.00						
100.00	178.00						
94.00	172.00						
69.00	168.00						
83.00	170.00						
76.00	170.00						
80.00	168.00						
71.00	158.00						


height [cm]
185
180
175
170
165
160
70
80
90
100
110
weight [kg]