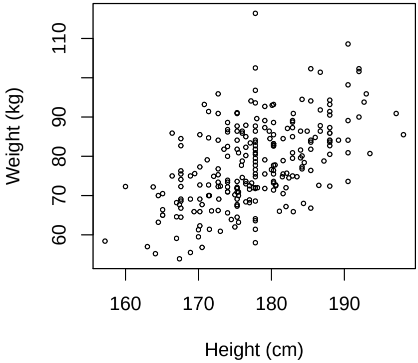
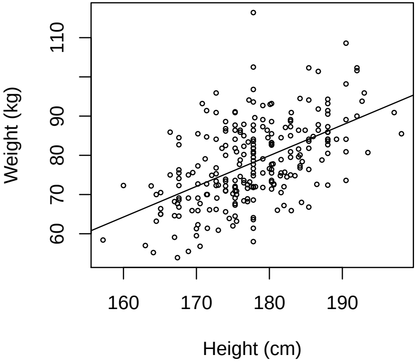
Lecture 2
Least Squares Regression Line
STAT 432, Spring 2021

A scatterplot of weight (Y) versus height (X) for 247 physically active men.



One way to describe the relationship between the two variables is with a straight line. The points do not fall directly on the line, so there is some variability in the points around the line.



Simple Linear Regression Model

Let $\{(x_i, y_i) : i = 1, \cdots, n\}$ be a collection of n data points. A **simple linear regression model** expressing the relationship between y_i and x_i is given by:

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i$$

- y_i response variable
- x_i explanatory variable
- β_0 intercept parameter
- β_1 slope parameter
- ϵ_i is the random error term; assume $\epsilon_i \sim N(0, \sigma^2)$

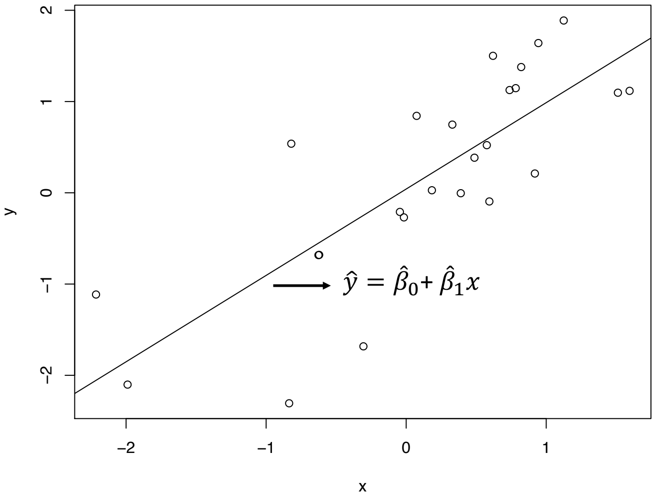
Remark: y_i is also sometimes called the **dependent** variable, and x_i the **independent** or **predictor** variable. Notation and terminology may vary depending on the textbook and context.

Fitted Values and Residuals

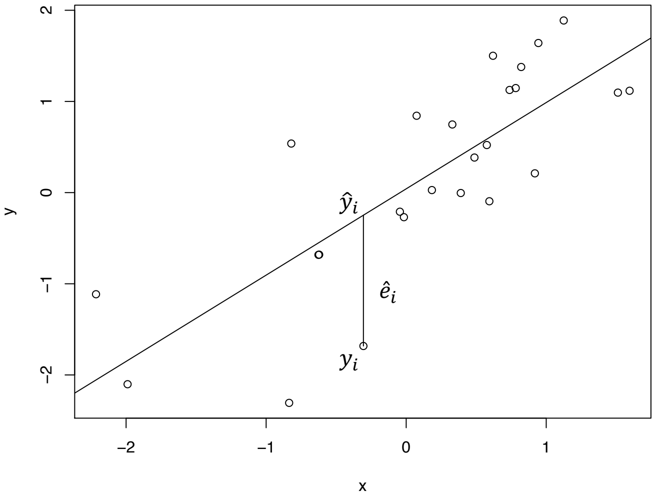
- The line that we we estimate, or fit to the data in the scatterplot, is written as
$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$$
where $\hat{\beta}_0$ and $\hat{\beta}_1$ denote the estimates of the unkown regression parameters β_0 and β_1.

- The fitted (or predicted) value for the i^{th} observation (x_i, y_i):
$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$$

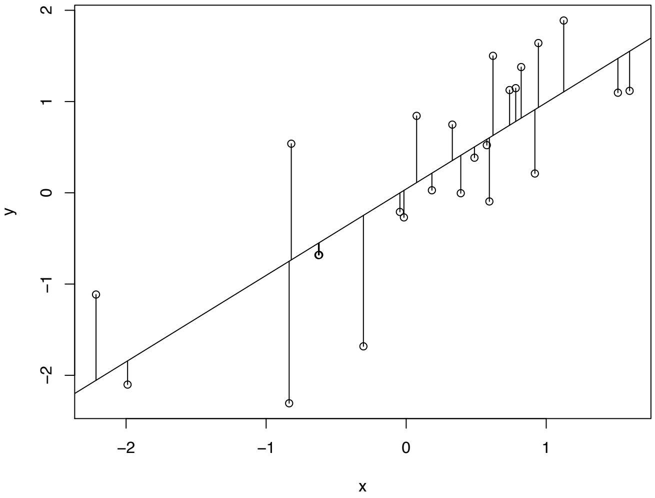
- The **residual** for the i^{th} observation is the difference between the observed value (y_i) and the predicted value $(\hat{y}_i)$:
$$\hat{e}_i = y_i - \hat{y}_i = y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i)$$



$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$
y
x
2
1
0
−1
−2
−2
−1
0
1



$\hat{y}_i$
$\hat{e}_i$
y_i
y
x
2
1
0
−1
−2
−2
−1
0
1



y
2
1
0
−1
−2
−2
−1
0
1
x

Sum of Squared Residuals

- Intuitively, a line that fits the data well has small residuals.
- The **least squares line** minimizes the **sum of squared residuals**:

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

- That is, out of all possible lines we could draw on the scatterplot, the least squares line is the "best fit" since it has the smallest sum of squared residuals.

Least Squares Estimation

Formally, the estimates $\hat{\beta}_0$ and $\hat{\beta}_1$ of the intercept and slope are found by using calculus to minimize the residual sum of the squares (RSS):

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

To minimize set the partial derivatives equal to zero:

$$\frac{\partial RSS}{\partial \hat{\beta}_0} = -2 \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0$$

$$\frac{\partial RSS}{\partial \hat{\beta}_1} = -2 \sum_{i=1}^{n} x_i (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0$$

Least Squares Estimation

Using some algebraic manipulation we can solve these two equations to obtain the least squares estimates of the intercept and slope:

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

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

Note that the equation for the intercept guarantees the least squares line passes through $(\bar{x}, \bar{y})$.

Interpretation

- **Slope**: an increase in the explanatory variable (x) by one unit is associated with a change of $\hat{\beta}_1$ in the predicted response ($\hat{y}$).
- **Intercept**: the prediction for the response variable ($\hat{y}$) when the value for the explanatory variable is zero ($x = 0$). It may not make sense to try to interpret the intercept depending on the application.