Lecture 14:
Simple Linear Regression
STAT 630, Fall 2021

Scatterplots

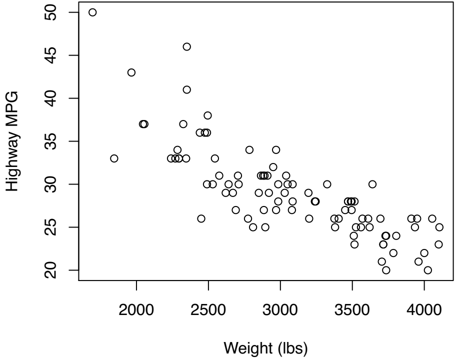
- A scatterplot a graphical display used to study the relationship between two variables x and y.
- Data displayed on a scatterplot are collected in pairs:

$$(x_1, y_1), (x_2, y_2), \cdots, (x_n, y_n)$$

  where n denotes the total number of cases or pairs.
- A scatterplot provides insight into how two variables are related.

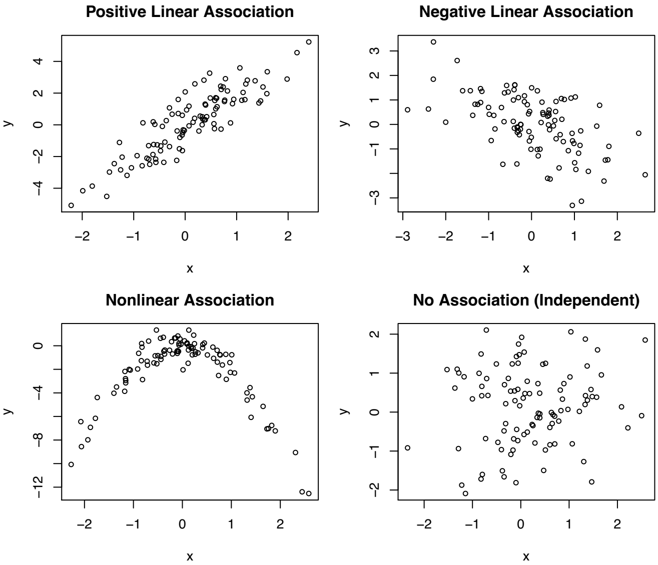
Example

```
> library(MASS)
> plot(Cars93$Weight, Cars93$MPG.highway,
    xlab = "Weight (lbs)", ylab = "Highway MPG")
```



Types of Relationships between Variables

- Two variables are said to be **associated** if the scatterplot shows a discernible pattern or trend.
- An association is **positive** if y increases as x increases.
- An association is **negative** if y decreases as x increases.
- An association is **linear** if the scatterplot between x and y has a linear trend; otherwise, the association is called **nonlinear**.


Positive Linear Association
Negative Linear Association
Nonlinear Association
No Association (Independent)
x
y
−4 −2 0 2 4
−2 −1 0 1 2
−3 −1 0 1 2 3
−3 −2 −1 0 1 2
−12 −8 −4 0
−2 −1 0 1 2
−2 −1 0 1 2
−2 −1 0 1 2

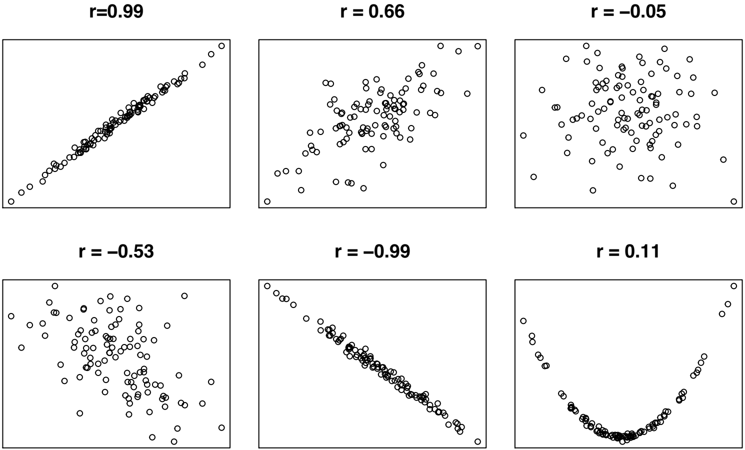
Correlation Coefficient

The **correlation coefficient**, denoted by r, is a number between -1 and 1 that describes the strength of the linear association between two numerical variables.

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}} = \frac{1}{n-1}\sum_{i=1}^{n}\left(\frac{x_i - \bar{x}}{s_x}\right)\left(\frac{y_i - \bar{y}}{s_y}\right)$$

- $\bar{x}$ and $\bar{y}$ are the sample means
- s_x and s_y are the sample standard deviations

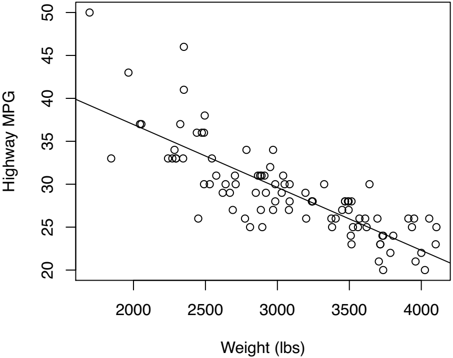
Correlation Coefficient



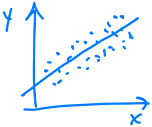
Correlation Coefficient

- $r \approx 1$ when there is a strong positive linear association between the variables.
- $r \approx -1$ when there is a strong negative linear association between the variables.
- $r \approx 0$ when there is no relationship between the variables (i.e., independent).
- The correlation coefficient is only useful for evaluating the linear association between two variables. It is not a useful measure for nonlinear relationships.

Simple linear regression is a method for fitting a straight line to data that show a linear trend when displayed on a scatterplot. It is a useful tool for making predictions for a quantitative response variable.



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 (random)
- x_i explanatory variable (non-random)
- β_0 intercept parameter (non-random)
- β_1 slope parameter (non-random)
- ϵ_i is the random error term, $\epsilon_i \sim N(0, \sigma)$

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

Fitted Values and Residuals

$\hat{\beta}_0, \hat{\beta}_1$ are estimates of parameters β_0, β_1

The line that we ~~we~~ estimate, or fit to the data in the scatterplot, is written as

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

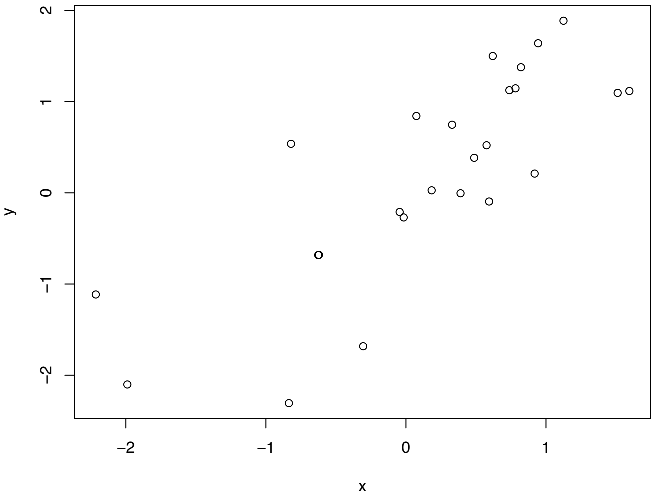
The fitted (or predicted) value for the i^{th} observation (x_i, y_i):

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$$

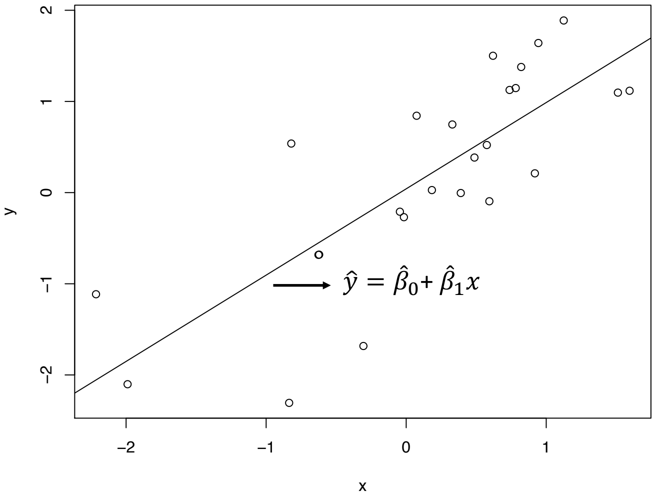
The **residual** for the i^{th} observation is the difference between the observed value (y_i) and the predicted value $(\hat{y}_i)$:

$$\hat{e}_i = y_i - \hat{y}_i = y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i)$$

residual = observed − predicted

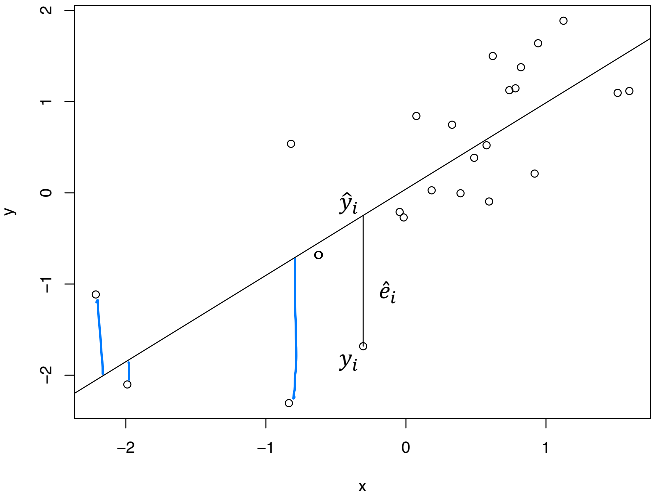


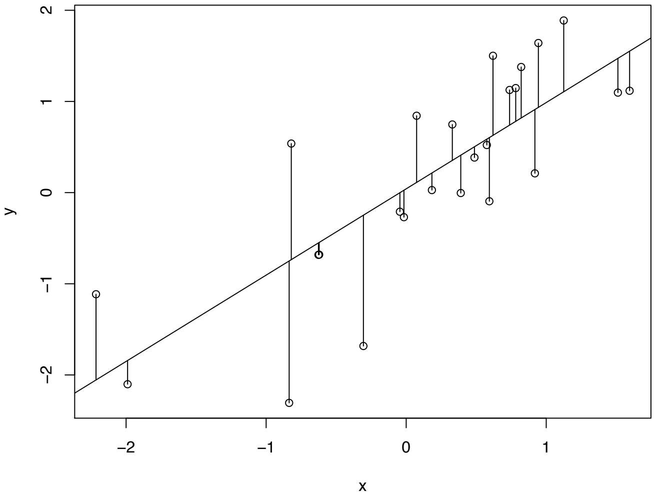
y
x
2
1
0
−1
−2
−2
−1
0
1



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

$\hat{e}_i = y_i - \hat{y}_i$





y
x
2
1
0
−1
−2
−2
−1
0
1

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.

$RSS = \hat{e}_1^2 + \hat{e}_2^2 + \dots + \hat{e}_n^2$

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 sum of squared residuals:

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

To minimize set the partial derivatives equal to zero:

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

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

Find $\hat{\beta}_0, \hat{\beta}_1$ that minimize

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

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

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

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 $(\bar{x}, \bar{y})$.

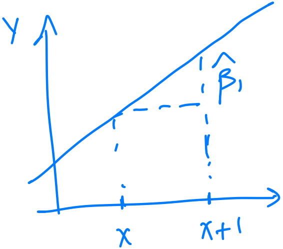
$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} \Rightarrow \bar{y} = \hat{\beta}_0 + \hat{\beta}_1 \bar{x}$$

Interpretation

least squares line: $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$

- **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.

Slope



Intercept

Set $x = 0$

$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1(0) = \hat{\beta}_0$

Coefficient of Determination

The **coefficient of determination** (R^2) is a measure of how well the linear regression model fits the data.

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

- $TSS = \sum_{i=1}^{n}(y_i - \bar{y})^2$ is the total sum of squares (total variability in the response variable)
- $RSS = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$ is the residual sum of squares (unexplained variability)

Coefficient of Determination

- R^2 can be interpreted as the proportion of variability in the response variable y that is explained by x.
- $0 \leq R^2 \leq 1$; the closer R^2 is to 1, the better the linear regression model fits the data.
- R^2 can be computed as the correlation coefficient r squared.
- R^2 is arguably one of the most commonly misused statistics. Always look at a scatterplot of your data first, and check whether fitting a line makes sense and for any outliers.