2022 Winter School

Introduction to linear models

Kathryn Kemper

Introduction to linear models: Outline

- Definition, terminology
- LS estimation of regression parameters
- Diagnostics

Simple linear regression

In most cases,

linear model ≈ linear regression

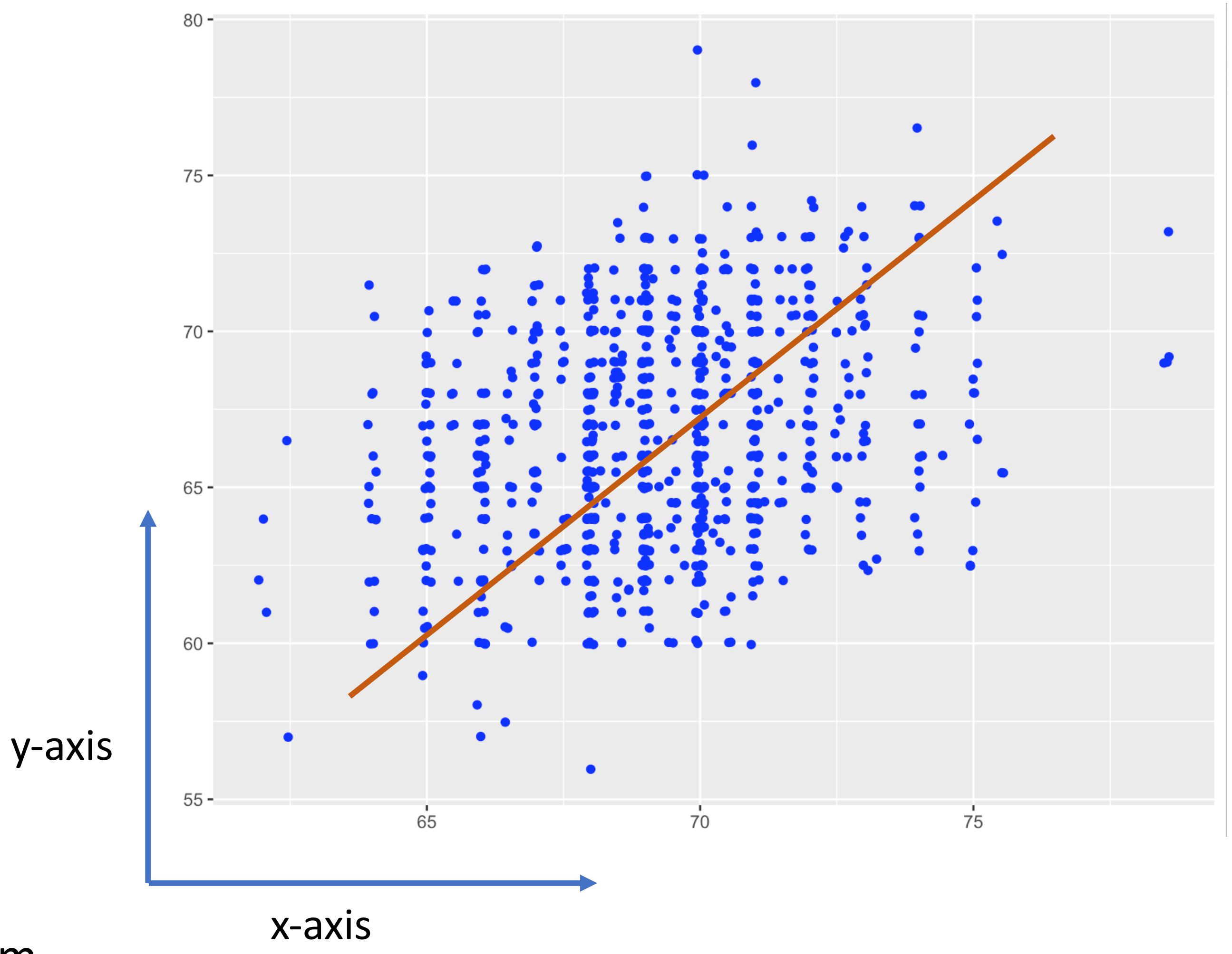
- describes the relationship between two variables
- We want to find ’the best’ line to describe the relationship, i.e.

$$y = a + bx$$

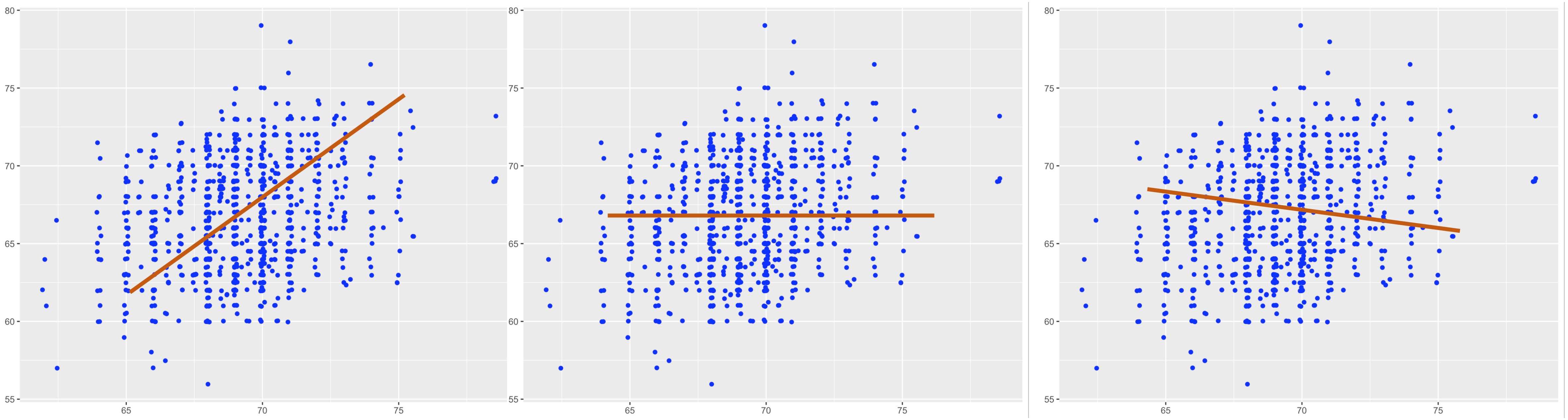
intercept slope

Today we will:

- show how to obtain ‘best fitting’ line using OLS (ordinary least squares)
- review the metrics that describe ‘model fit’
- generalize the the basic model to matrix form



How to find the ‘best’ line to describe the data?



Simple linear regression

Independent (response) variable

intercept

slope

dependent (predictor) variable

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

residual

$$i = 1 \ \dots \ n$$

$$y_1 = \beta_0 + \beta_1 x_1 + \varepsilon_1$$
$$y_2 = \beta_0 + \beta_1 x_2 + \varepsilon_2$$
$$....$$
$$y_n = \beta_0 + \beta_1 x_n + \varepsilon_n$$

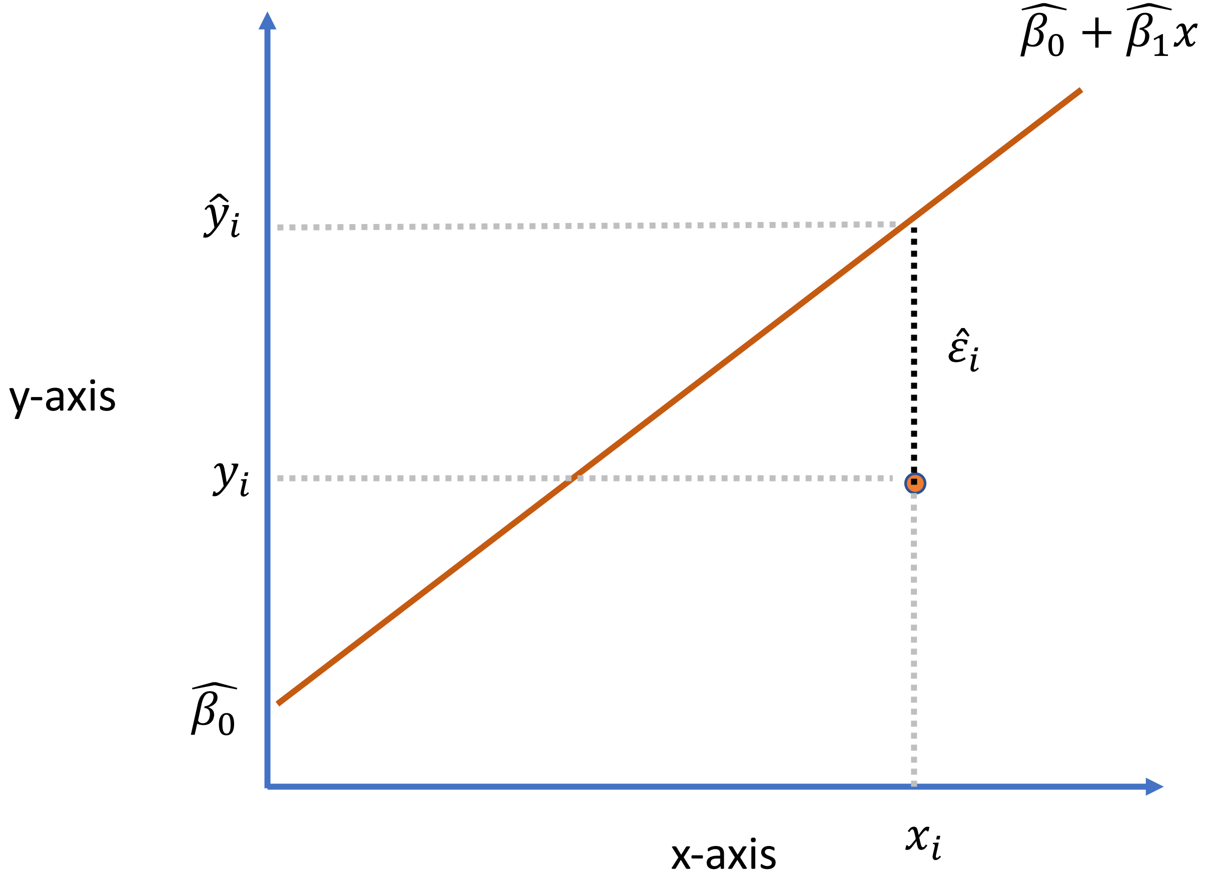
$\beta_0 \ and \ \beta_1$ are unknown population parameters
$\widehat{\beta_0} \ and \ \widehat{\beta_1}$ (ie 'beta-hat') are the population estimates

The 'predicted' value of y (ie y-hat) is:

$$\hat{y}_i = \widehat{\beta_0} + \widehat{\beta_1} x_i$$

The residual is:

$$\hat{\varepsilon}_i = y - \hat{y}_i$$



How good is the regression?

SSQ in y:

$$\sum (y - \bar{y})^2$$

SSQ explained by the regression:

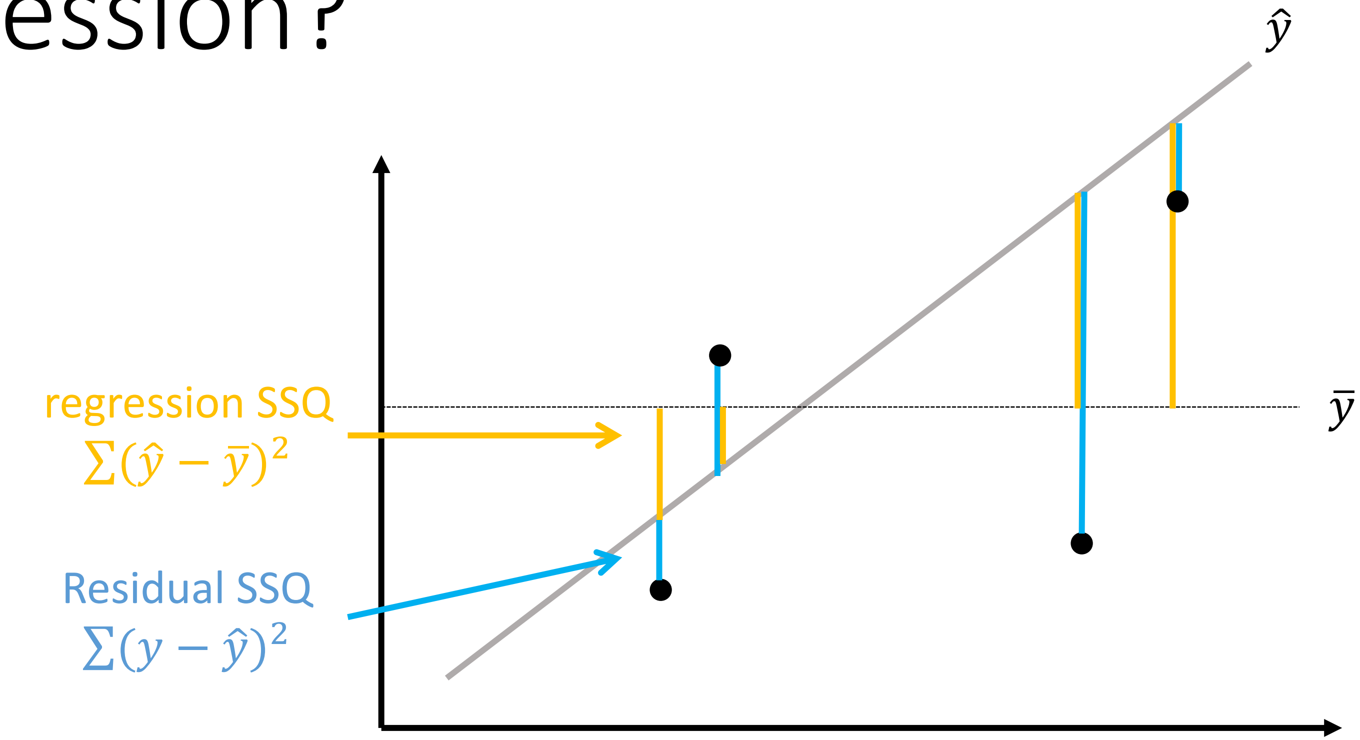
$$\sum (\hat{y} - \bar{y})^2$$

SSQ residuals:

$$\sum (y - \hat{y})^2$$

Thus,
Total SSQ = regression SSQ + residual SSQ

$$\Sigma(y - \bar{y})^2 = \Sigma(\hat{y} - \bar{y})^2 + \Sigma(y - \hat{y})^2$$



regression SSQ
$\Sigma(\hat{y} - \bar{y})^2$

Residual SSQ
$\Sigma(y - \hat{y})^2$

R^2 = variance explained by the regression

$$= \frac{regression\ SSQ}{total\ SSQ} = \frac{\Sigma(\hat{y} - \bar{y})^2}{\Sigma(y - \bar{y})^2}$$

$$= 1 - \frac{\Sigma(y - \hat{y})^2}{\Sigma(y - \bar{y})^2}$$

How good is the regression?

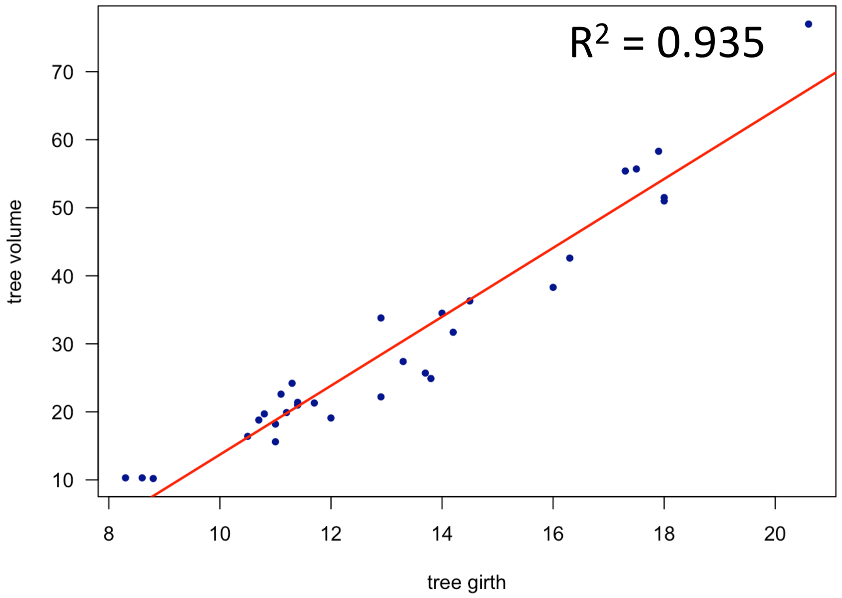
- R^2 = variance explained by the regression

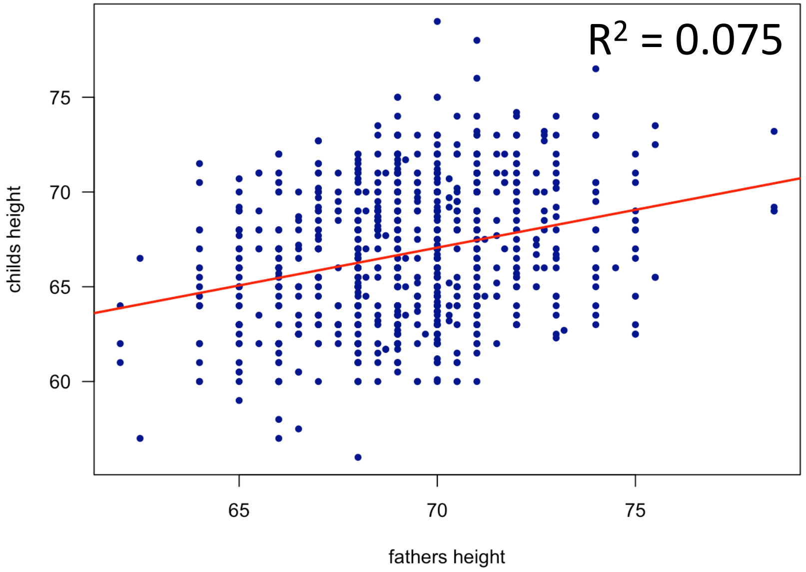
$$R^2 = \frac{regression\ SSQ}{total\ SSQ} = \frac{\sum(\hat{y} - \bar{y})^2}{\sum(y - \bar{y})^2}$$

- A value that ranges from
  - 0 (regression explains no variation)
  - 1 (perfect fit)

➢"93.5% of the variation in tree volume can be explained by tree girth"

➢"7.5% of the variation in a height of children can be explained by their father's height"



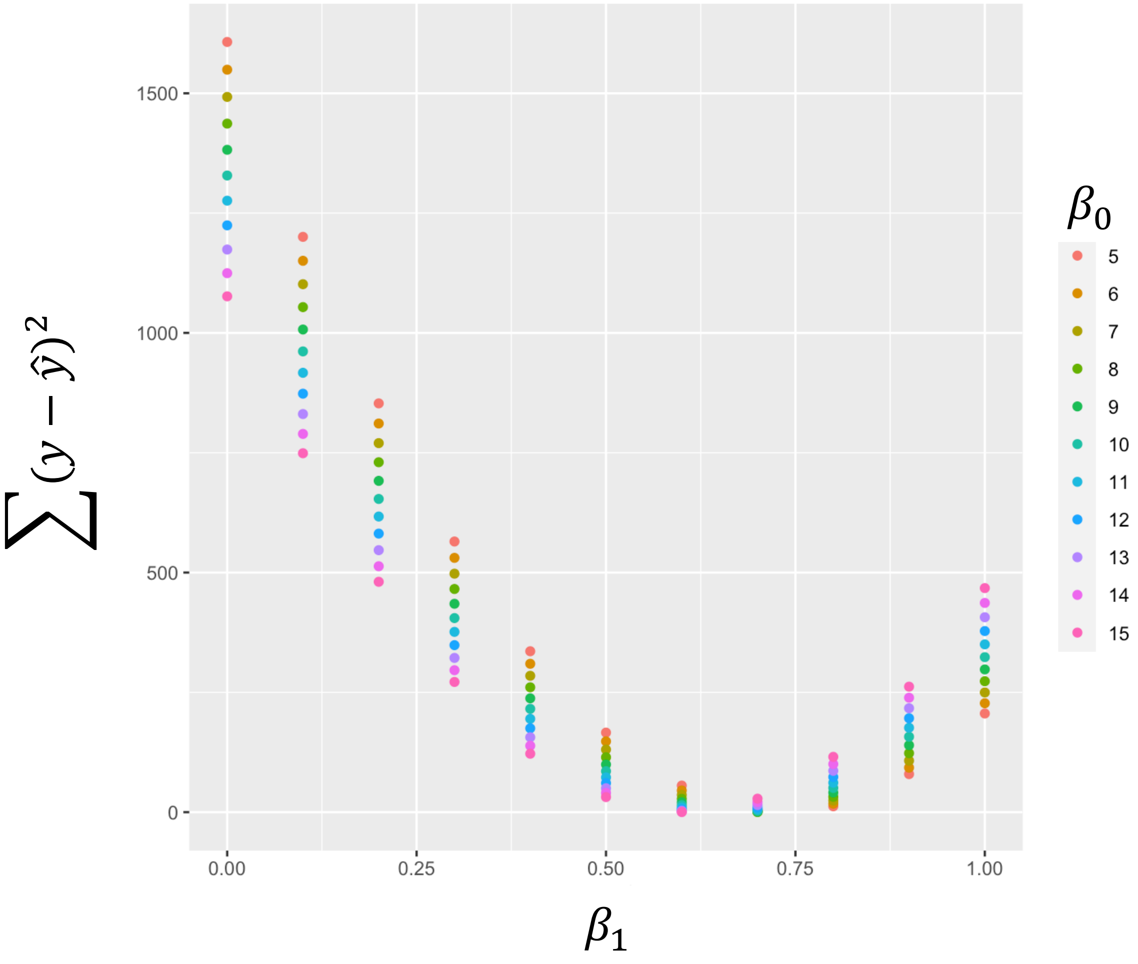


How do we determine β_0 and β_1?

We can use a grid-search,

x	y
76.0	61.2
72.6	57.9
74.6	59.2
75.8	60.6
74.5	62.0
74.9	58.7
74.4	59.1
75.7	59.5
73.4	60.1
75.5	62.3

1. take our data
2. guess values β_0 and β_1
3. calculate $\hat{y}$
4. calculate SSQ
5. chose model with 'best fit'



Not an ideal approach! -> do not do this
'best fit' minimizes SSQ residuals
(or maximizes R^2)

How do we determine β_0 and β_1?

- Briefly, take partial derivatives of $\sum(y-\hat{y})^2$ (w.r.t. β_0 and then β_1), set to zero and solve.

- Result,
  - slope:

$$\widehat{\beta_1} = \frac{\sum(y-\bar{y})(x-\bar{x})}{\sum(x-\bar{x})^2} = \frac{SSQ_{xy}}{SSQ_x}$$

  - intercept:

$$\widehat{\beta_0} = \bar{y} - \widehat{\beta_1}\bar{x}$$

Simple linear regression

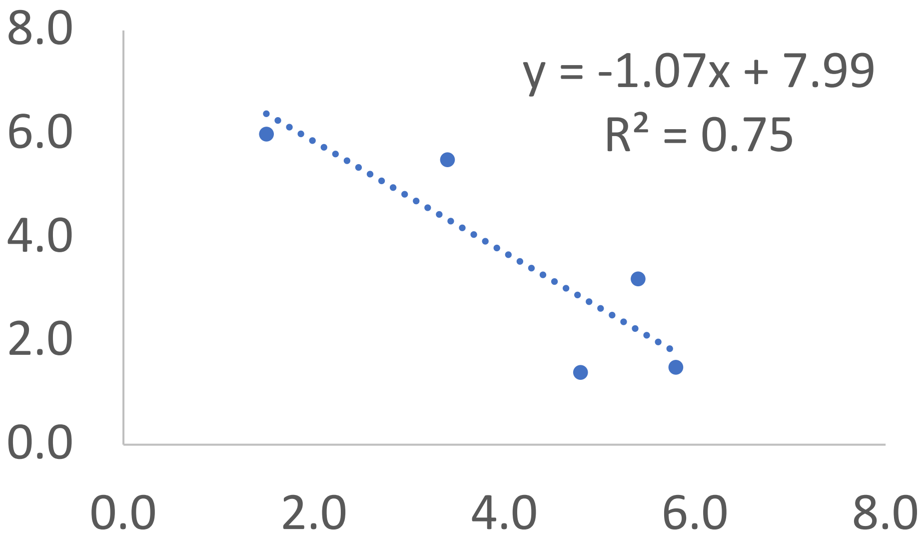
Recall: $\sum(y-\bar{y})(x-\bar{x}) = n\sum xy - \sum x \sum y$
$\sum(x-\bar{x})^2 = n\sum x^2 - (\sum x)2$

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$
$$i = 1 \ \ldots n$$

	y	x	xy	x^2
	6.0	1.5	9.0	2.3
	1.4	4.8	6.7	23.0
	1.5	5.8	8.7	33.6
	5.5	3.4	18.7	11.6
	3.2	5.4	17.3	29.2
$\sum$	17.6	20.9	60.4	99.6

$$\widehat{\beta_1} = \frac{5x60.4 - 17.6x20.9}{5x99.6 - (20.9)^2} = -1.07$$

$$\widehat{\beta_0} = \frac{17.6}{5} + \frac{1.07x20.9}{5} = 7.99$$



$$R^2 = \frac{regression\ SSQ}{total\ SSQ}$$

Variance = SSQ / ‘n’

Hypothesis testing for ‘overall fit’

- H_0: All regression coefficients = 0
- Use an F-test to determine the support for H_0

$$F = \frac{variance\ explained\ by\ regression}{variance\ not\ explained\ by\ regression}$$

Number of parameters in our model
(i.e. = 2; β_0 and β_1)

$$F = \frac{SSQ_{reg}/(p_{reg} - 1)}{SSQ_{\varepsilon}/(n - p_{reg})}$$

Always = 1 as we compared regression to a model with only intercept (or mean β_0)

Sample size

Scaler form:

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$
$$i = 1 \;\ldots n$$

$$\begin{gathered} y_1 = \beta_0 + \beta_1 x_1 + \varepsilon_1 \\ y_2 = \beta_0 + \beta_1 x_2 + \varepsilon_2 \\ \ldots. \\ y_n = \beta_0 + \beta_1 x_n + \varepsilon_n \end{gathered}$$

Matrix form:

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

Why? Convenient & generalizable

Matrix form:

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

$\boldsymbol{Y}$ is a n x 1 column vector of observations
$\boldsymbol{X}$ is a n x 2 'design' matrix
$\boldsymbol{\beta}$ is a 2 x 1 column vector of parameters
$\boldsymbol{\varepsilon}$ is a n x 1 column vector of residuals
where n is the number of observations

Quick check:
$\boldsymbol{X}\boldsymbol{\beta}$, (n x 2) X (2 x 1) = (n x 1) matrix

$$\boldsymbol{Y} = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

$$= \begin{bmatrix} \beta_0 + \beta_1 x_1 \\ \beta_0 + \beta_1 x_2 \\ \vdots \\ \beta_0 + \beta_1 x_n \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix} = \begin{bmatrix} \beta_0 + \beta_1 x_1 + \varepsilon_1 \\ \beta_0 + \beta_1 x_2 + \varepsilon_2 \\ \vdots \\ \beta_0 + \beta_1 x_n + \varepsilon_n \end{bmatrix}$$

Estimating parameters in matrix form

$$\boldsymbol{Y} = \boldsymbol{X\beta} + \boldsymbol{\varepsilon}$$

We want to minimize SSQ residuals,

residuals: $\boldsymbol{\varepsilon} = \boldsymbol{Y} - \boldsymbol{X\beta}$

$$\sum \boldsymbol{\varepsilon}^2 = \boldsymbol{\varepsilon}'\boldsymbol{\varepsilon} = (\boldsymbol{Y} - \boldsymbol{X\beta})'(\boldsymbol{Y} - \boldsymbol{X\beta})$$

Like previously, take derivatives w.r.t. $\boldsymbol{\beta}$, set to zero and solve.

Final result:

$$\boldsymbol{\beta} = [\boldsymbol{X'X}]^{-1}\boldsymbol{X'Y}$$

Hat matrix for prediction

$$\boldsymbol{Y} = \boldsymbol{X\beta} + \boldsymbol{\varepsilon}$$

- Parameter estimates: $\boldsymbol{\beta} = [\boldsymbol{X}'\boldsymbol{X}]^{-\mathbf{1}}\boldsymbol{X}'\boldsymbol{Y}$
- Predicted values:

$$\widehat{\boldsymbol{Y}} = \boldsymbol{X\beta}$$

$$\widehat{\boldsymbol{Y}} = \boldsymbol{X}[\boldsymbol{X}'\boldsymbol{X}]^{-\mathbf{1}}\boldsymbol{X}'\boldsymbol{Y}$$

$$\widehat{\boldsymbol{Y}} = \boldsymbol{HY},$$

where $\mathbf{H} = \boldsymbol{X}[\boldsymbol{X}'\boldsymbol{X}]^{-\mathbf{1}}\boldsymbol{X}'$

$\mathbf{H}$ is called the ‘hat matrix’ because it turns $\boldsymbol{Y}$ into $\widehat{\boldsymbol{Y}}$

Other scalar forms for estimating β_1

$$\widehat{\beta_1} = \frac{SSQ_{xy}}{SSQ_x} = \frac{\sum(y-\bar{y})(x-\bar{x})}{\sum(x-\bar{x})^2}$$

$$\widehat{\beta_1} = \frac{cov(x,y)}{var(x)}$$

correlation

$$\widehat{\beta_1} = r\,\frac{s_y}{s_x}$$

SD of x and y

$$= \frac{SSQ_{xy}}{\sqrt{SSQ_x SSQ_y}}\sqrt{\frac{SSQ_y}{n-1}}\sqrt{\frac{n-1}{SSQ_x}} = \frac{SSQ_{xy}}{SSQ_x}$$

NB:

1. Variance = SSQ / ‘n’
2. R^2 (variance explained by model) = r^2 (sq. correlation) in SLR

Estimation of effects for discrete variables

- So far:

$$\boldsymbol{Y} = \boldsymbol{X\beta} + \boldsymbol{\varepsilon}$$

$\boldsymbol{Y}$ is a n x 1 column vector of observations
$\boldsymbol{X}$ is a n x **p** ‘design’ matrix
$\boldsymbol{\beta}$ is a **p** x 1 column vector of parameters
$\boldsymbol{\varepsilon}$ is a n x 1 column vector of residuals
where n is the number of observations, &
p is the number of parameters

- The framework can also be used to estimate the effect of discrete factors (or levels)

Estimation of effects for discrete variables

$$\boldsymbol{Y} = \boldsymbol{X\beta} + \boldsymbol{\varepsilon}$$
$$\boldsymbol{\beta} = [\boldsymbol{X}'\boldsymbol{X}]^{-1}\boldsymbol{X}'\boldsymbol{Y}$$

Example: We have measured weight and SNP genotypes (AA, AB, BB) for 7 people, 2 with AA genotype, 2 with AB and 3 BB. What is the mean effect for each genotype?

Need to a new 'design matrix' $\boldsymbol{X}$*:*

$$\boldsymbol{X} = \begin{bmatrix}1 & 0 & 0\\ 1 & 0 & 0\\ 0 & 1 & 0\\ 0 & 1 & 0\\ 0 & 0 & 1\\ 0 & 0 & 1\\ 0 & 0 & 1\end{bmatrix}\textit{, i.e. } \boldsymbol{X} \textit{ is now a 7 x 3 matrix, then } \boldsymbol{\beta} \textit{ becomes a 3x1 matrix, } \boldsymbol{\beta} = \begin{bmatrix}\beta_0\\ \beta_1\\ \beta_2\end{bmatrix} \begin{array}{l} \longleftarrow \text{mean AA}\\ \longleftarrow \text{mean AB}\\ \longleftarrow \text{mean BB}\end{array}$$

AA genotype
AB genotype
BB genotype

Estimation of effects for discrete variables

$$\boldsymbol{Y} = \boldsymbol{X\beta} + \boldsymbol{\varepsilon}$$

$$\boldsymbol{\beta} = [\boldsymbol{X'X}]^{-1}\boldsymbol{X'Y}$$

Example:

(7x1) (7x3) (3x3) (3x1)

$$\boldsymbol{Y} = \begin{bmatrix} 45 \\ 52 \\ 63 \\ 46 \\ 54 \\ 65 \\ 70 \end{bmatrix} \qquad \boldsymbol{X} = \begin{bmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \end{bmatrix}$$

$$\boldsymbol{X'X} = \begin{bmatrix} 2 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 3 \end{bmatrix}$$

N (AA geno)

N (AB geno)

N (BB geno)

$$\boldsymbol{X'Y} = \begin{bmatrix} 97 \\ 109 \\ 189 \end{bmatrix} \begin{matrix} \leftarrow \sum(AA\ geno) \\ \leftarrow \sum(AB geno) \\ \leftarrow \sum(BB geno) \end{matrix}$$

then $\boldsymbol{X'Y}$ ‘divided by’ $\boldsymbol{X'X}$ will be equal to the average per group....

$$[\boldsymbol{X'X}]^{-1}\boldsymbol{X'Y} = \boldsymbol{\beta} = \begin{bmatrix} \mathbf{48.5} \\ \mathbf{54.5} \\ \mathbf{63.0} \end{bmatrix} \begin{matrix} \leftarrow \text{mean AA} \\ \leftarrow \text{mean AB} \\ \leftarrow \text{mean BB} \end{matrix}$$

Setting up the design matrix

$$\boldsymbol{X_1} = \begin{bmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \end{bmatrix}$$

Rank = number of independent rows of a matrix

- If $\boldsymbol{X}$ is a p x n matrix, then $\boldsymbol{X}'\boldsymbol{X}$ is p x p
- $\boldsymbol{X}$ must be 'full rank' for $[\boldsymbol{X}'\boldsymbol{X}]^{-\mathbf{1}}$ to exist
- If $[\boldsymbol{X}'\boldsymbol{X}]^{-\mathbf{1}}$ exists, then there is a unique $\widehat{\boldsymbol{\beta}}$

- Previously we estimated a mean for each genotype using $\boldsymbol{X_1}$ (above), equally we could use $\boldsymbol{X_2}$ to estimate a mean for AA genotypes and deviations for AB and BB genotype classes.

- However, we cannot estimate an overall mean, and 3 genotypes deviations as we only have 3 groups. Therefore the 4th number is a linear combination of the 3 others

$$\boldsymbol{X_2} = \begin{bmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \\ 1 & 0 & 1 \\ 1 & 0 & 1 \end{bmatrix}$$

Estimation of effects for discrete variables

$$\boldsymbol{Y} = \boldsymbol{X\beta} + \boldsymbol{\varepsilon}$$
$$\boldsymbol{\beta} = [\boldsymbol{X'X}]^{-1}\boldsymbol{X'Y}$$

$$\boldsymbol{Y} = \begin{bmatrix} 45 \\ 52 \\ 63 \\ 46 \\ 54 \\ 65 \\ 70 \end{bmatrix}$$

OPTION 1:

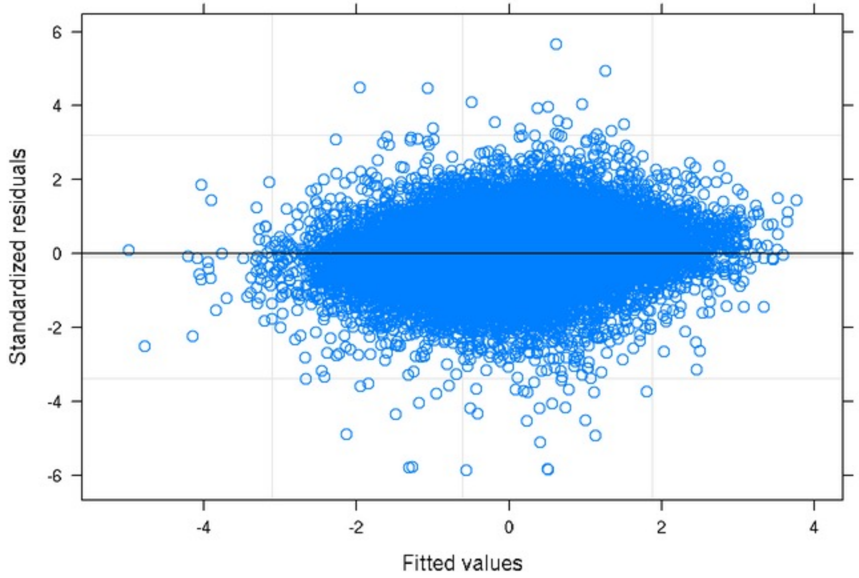
$$\boldsymbol{X_1} = \begin{bmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \end{bmatrix} \quad \boldsymbol{\widehat{\beta}} = \begin{bmatrix} \boldsymbol{48.5} \\ \boldsymbol{54.5} \\ \boldsymbol{63.0} \end{bmatrix}$$

OPTION 2:

$$\boldsymbol{X_2} = \begin{bmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \\ 1 & 0 & 1 \\ 1 & 0 & 1 \end{bmatrix} \quad \boldsymbol{\widehat{\beta}} = \begin{bmatrix} \boldsymbol{48.5} \\ \boldsymbol{6.0} \\ \boldsymbol{14.5} \end{bmatrix}$$

Model diagnostics

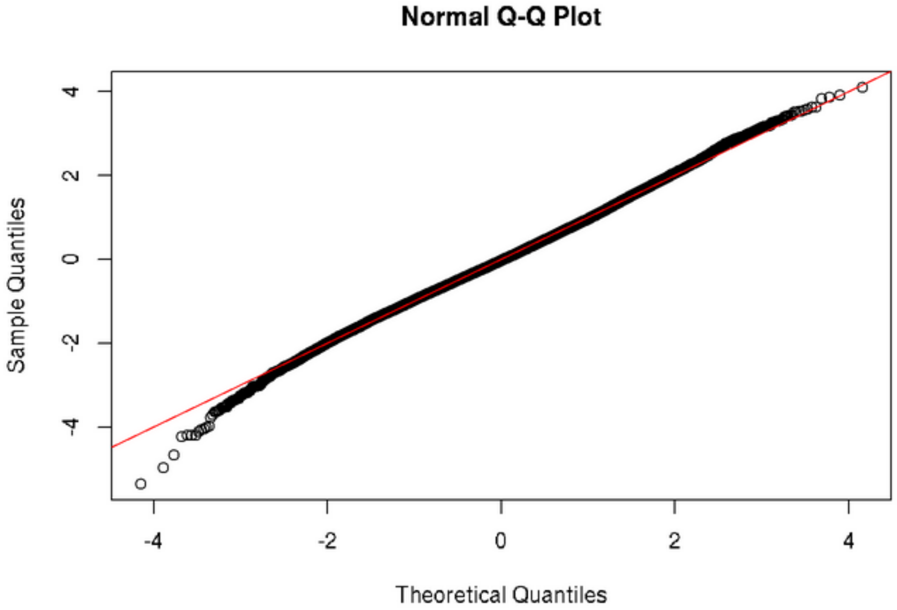
Hypothesis testing in OLS (ordinary least squares) assumes heteroscedastic, uncorrelated residuals, i.e. $\boldsymbol{\varepsilon} \sim MVN(0, \boldsymbol{I}\sigma_e^2)$



It’s all about the residuals!

e.g. plot residuals on y or $\hat{\boldsymbol{y}}$

- Should look ‘stary night’
- Screen for outliers
- test for normality, Q-Q plot or Wilk-Shapiro test



If $\boldsymbol{\varepsilon} \sim MVN(0, \boldsymbol{I}\sigma_e^2)$, then $\hat{\boldsymbol{\beta}} \sim MVN(\boldsymbol{\beta}, [\boldsymbol{X}'\boldsymbol{X}]^{-1}\sigma_e^2)$

basics – practical 3

- Work through “basicsPrac3.pdf”
    - Q1: variance, correlation & covariance
    - Q2: SLR using matrix notation
    - Q3: model fit
    - Q4: variance of predictor (as example of variance properties)
    - Q5: QQ-plot
- Software: R, you need to change the directory for the data!