

G&G 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 \approx linear regression

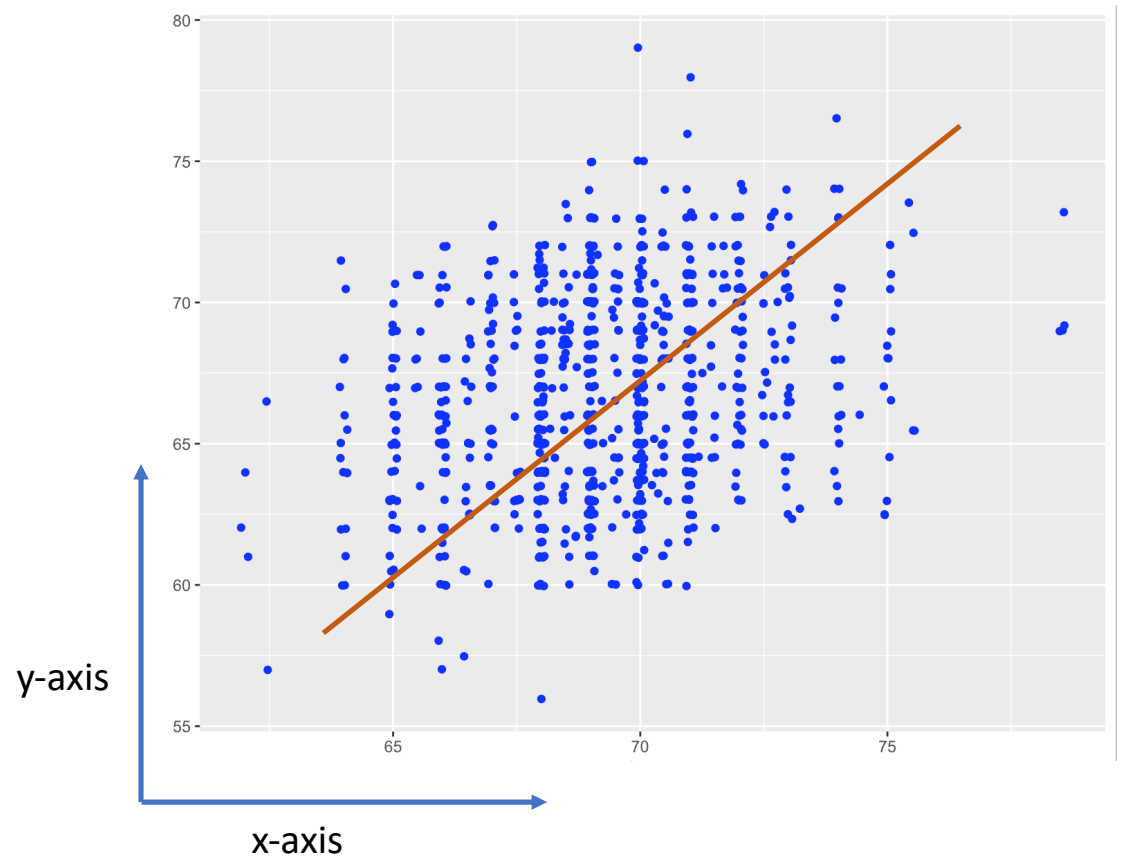
- describes the relationship between two variables
- We want to find 'the best' line to describe the relationship, i.e.

$$y = a + bx$$

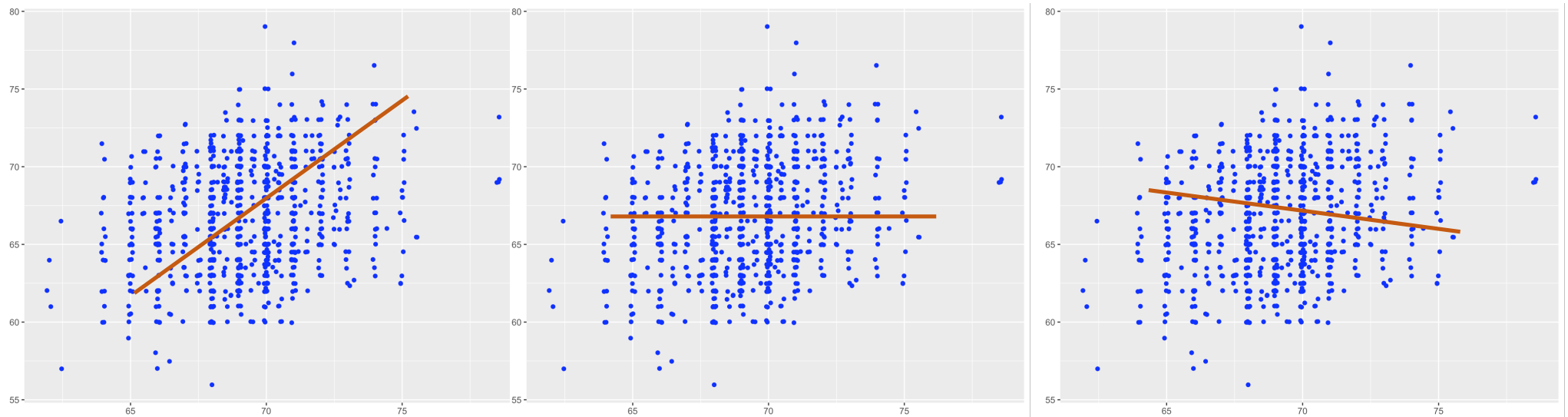
intercept slope

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

dependent (response) variable

intercept

slope

independent (predictor) variable

error

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

$$y_1 = \beta_0 + \beta_1 x_1 + \varepsilon_1$$

$$y_2 = \beta_0 + \beta_1 x_2 + \varepsilon_2$$

....

$$y_n = \beta_0 + \beta_1 x_n + \varepsilon_n$$

β_0 and β_1 are unknown population parameters

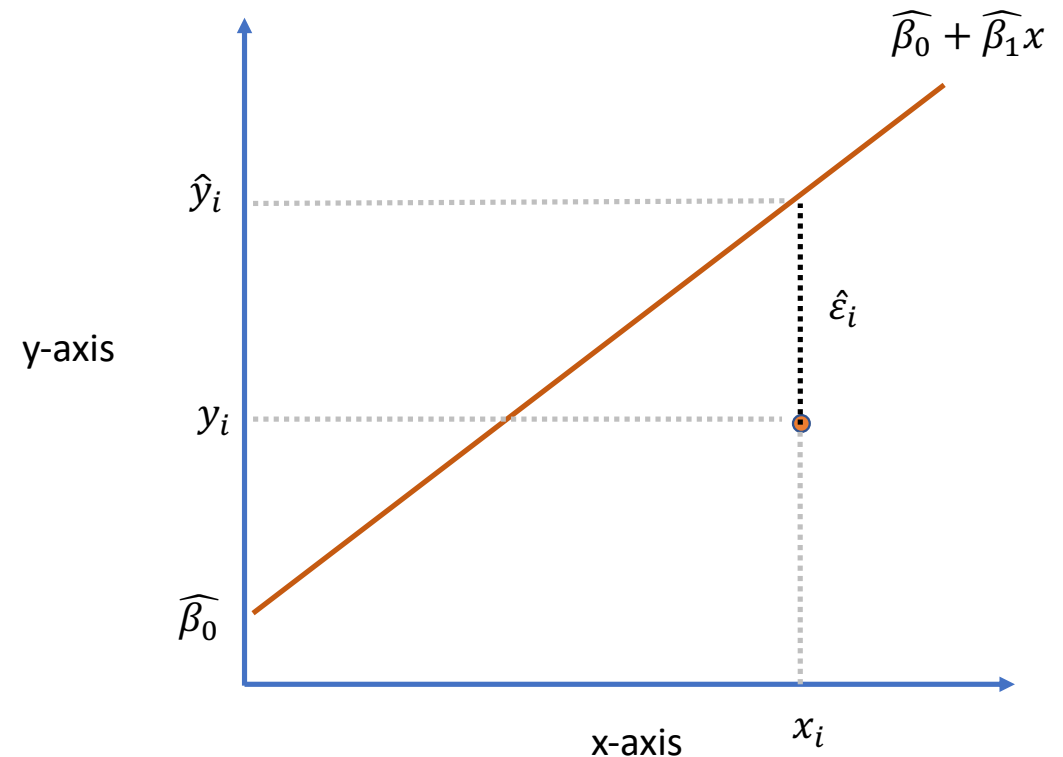
$\widehat{\beta}_0$ and $\widehat{\beta}_1$ (ie 'beta-hat') are the population estimates

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

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

The residual (an estimate of the error) 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:

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

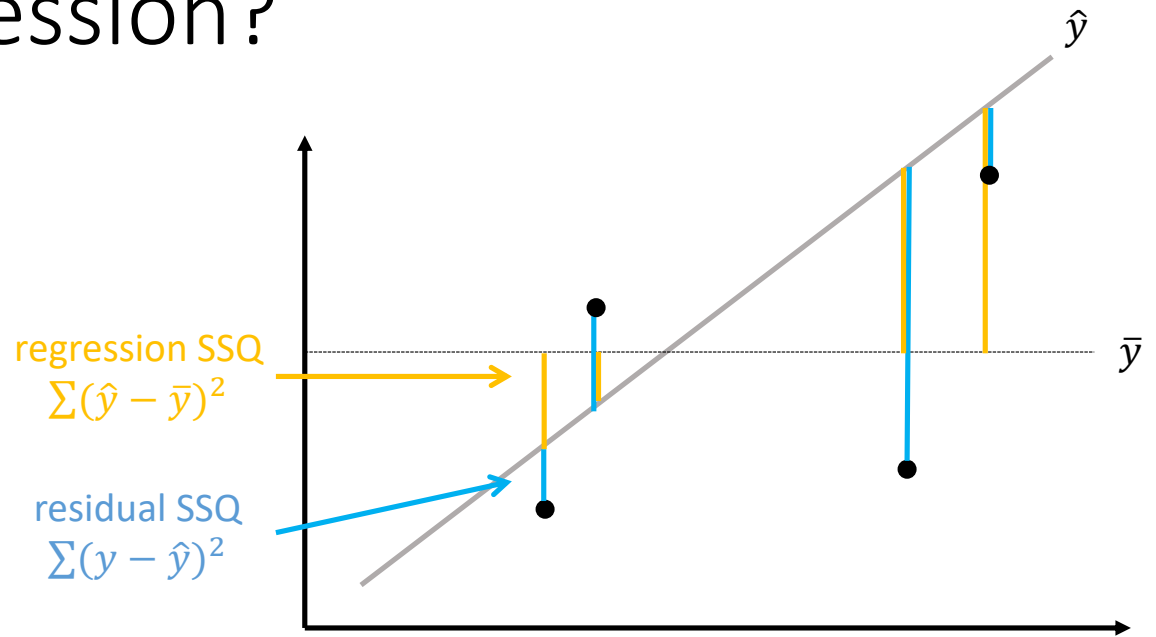
residual SSQ:

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

Thus,

Total SSQ = regression SSQ + residual SSQ

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



R^2 = variance explained by the regression

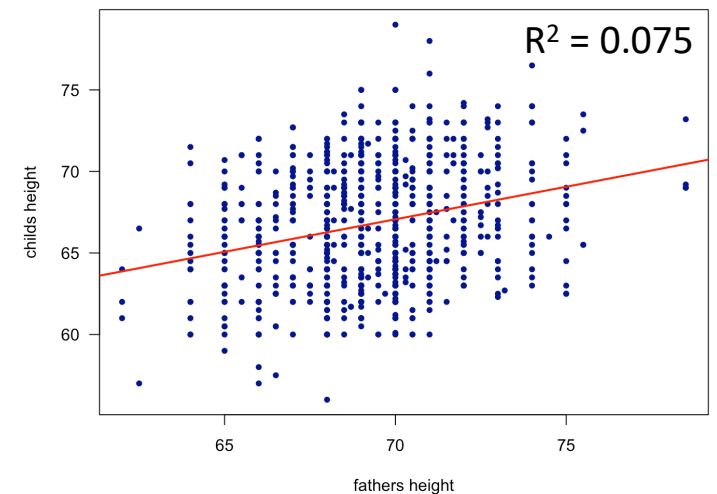
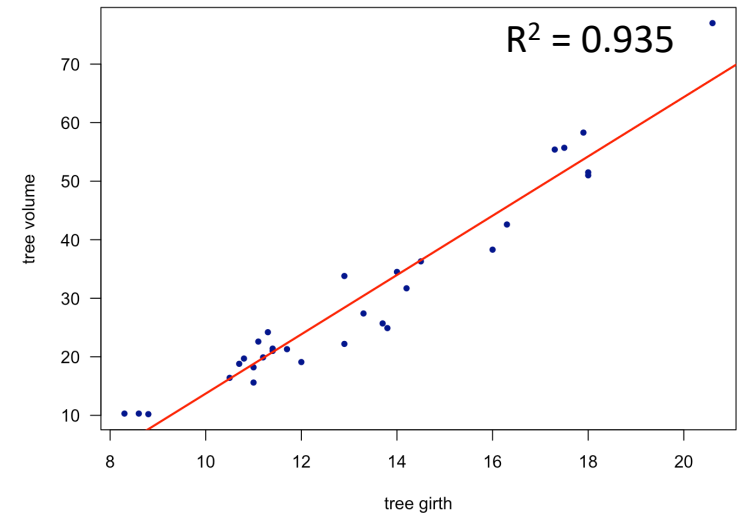
$$\begin{aligned} &= \frac{\text{regression SSQ}}{\text{total SSQ}} = \frac{\sum (\hat{y} - \bar{y})^2}{\sum (y - \bar{y})^2} \\ &= 1 - \frac{\sum (y - \hat{y})^2}{\sum (y - \bar{y})^2} \end{aligned}$$

How good is the regression?

- R^2 = variance explained by the regression

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

- 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”



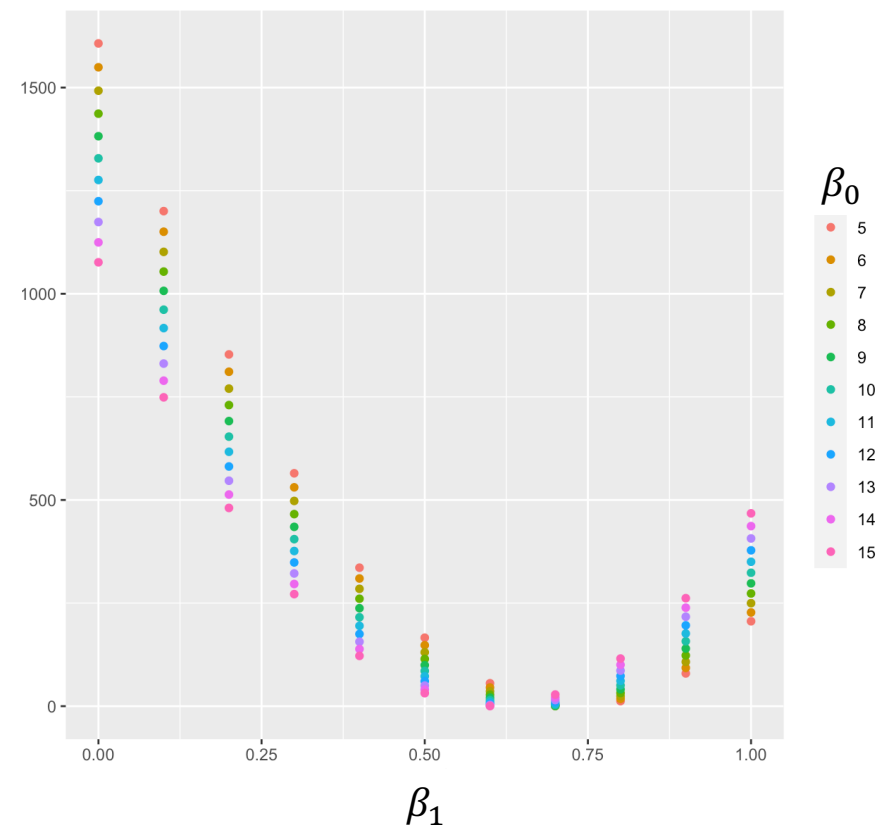
How do we determine β_0 and β_1 ?

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'

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



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:

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

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

Simple linear regression

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

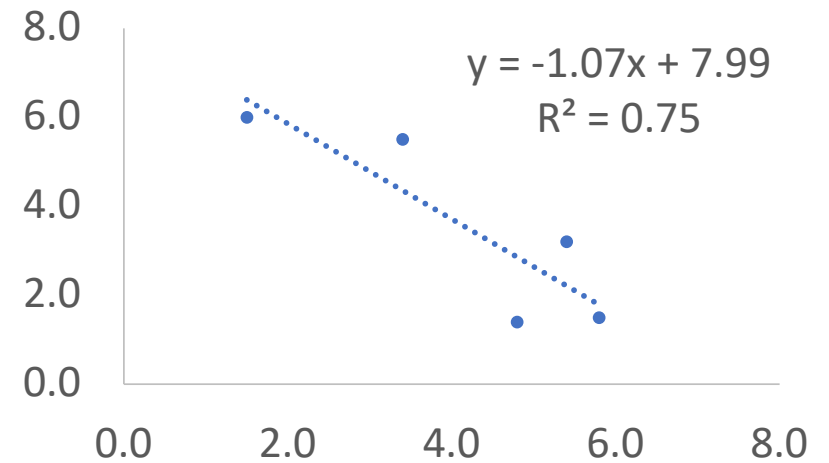
$$i = 1 \dots n$$

| 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 |
| Σ 17.6 | 20.9 | 60.4 | 99.6 |

$$\widehat{\beta}_1 = \frac{5 \times 60.4 - 17.6 \times 20.9}{5 \times 99.6 - (20.9)^2} = -1.07$$

$$\widehat{\beta}_0 = \frac{17.6}{5} + \frac{1.07 \times 20.9}{5} = 7.99$$

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



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

$$\text{Variance} = \text{SSQ} / 'n'$$

Hypothesis testing for 'overall fit'

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

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

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

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

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

Sample size

Scalar form:

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

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

Matrix form:

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

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

Why? Convenient & generalizable

Matrix form:

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

\mathbf{Y} is a $n \times 1$ column vector of observations

\mathbf{X} is a $n \times 2$ 'design' matrix

$\boldsymbol{\beta}$ is a 2×1 column vector of parameters

$\boldsymbol{\varepsilon}$ is a $n \times 1$ column vector of errors

where n is the number of observations

Quick check:

$\mathbf{X}\boldsymbol{\beta}$, $(n \times 2) \times (2 \times 1) = (n \times 1)$ matrix

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

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

Estimating parameters in matrix form

$$Y = X\beta + \varepsilon$$

We want to minimize residual SSQ,

$$\begin{aligned} \text{residuals: } \hat{\varepsilon} &= Y - X\hat{\beta} \\ \sum \hat{\varepsilon}^2 &= \hat{\varepsilon}'\hat{\varepsilon} = (Y - X\hat{\beta})'(Y - X\hat{\beta}) \end{aligned}$$

Like before, take derivatives w.r.t. β , set to zero and solve.

Final result:

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

Hat matrix for prediction

$$Y = X\beta + \varepsilon$$

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

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

$$\hat{Y} = HY,$$

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

H is called the 'hat matrix' because it turns Y into \hat{Y}

Estimation of effects for discrete variables

- So far:

$$Y = X\beta + \varepsilon$$

Y is a $n \times 1$ column vector of observations

X is a $n \times p$ 'design' matrix

β is a $p \times 1$ column vector of parameters

ε is a $n \times 1$ column vector of errors

where n is the number of observations, &

p is the number of parameters

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

Estimation of effects for discrete variables

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

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' X :

$$X = \begin{bmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \end{bmatrix}$$

, i.e. X is now a 7 x 3 matrix, then β becomes a 3x1 matrix,

$$\beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{bmatrix}$$

← mean AA
← mean AB
← mean BB

AA genotype
AB genotype
BB genotype

Estimation of effects for discrete variables

$$Y = X\beta + \varepsilon$$

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

Example:

$$\begin{array}{ccc}
 \begin{array}{c} (7 \times 1) \\ Y = \begin{bmatrix} 45 \\ 52 \\ 63 \\ 46 \\ 54 \\ 65 \\ 70 \end{bmatrix} \end{array} &
 \begin{array}{c} (7 \times 3) \\ X = \begin{bmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \end{bmatrix} \end{array} &
 \begin{array}{c} (3 \times 3) \\ X'X = \begin{bmatrix} 2 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 3 \end{bmatrix} \end{array} &
 \begin{array}{c} (3 \times 1) \\ X'Y = \begin{bmatrix} 97 \\ 109 \\ 189 \end{bmatrix} \end{array}
 \end{array}$$

$\leftarrow \sum(AA \text{ geno})$
 $\leftarrow \sum(AB \text{ geno})$
 $\leftarrow \sum(BB \text{ geno})$

N (AA geno) \leftarrow
 N (AB geno) \leftarrow
 N (BB geno) \leftarrow

then $X'Y$ 'divided by' $X'X$ will be equal to the average per group....

$$[X'X]^{-1}X'Y = \hat{\beta} = \begin{bmatrix} 48.5 \\ 54.5 \\ 63.0 \end{bmatrix}$$

\leftarrow mean AA
 \leftarrow mean AB
 \leftarrow mean BB

Setting up the design matrix

Rank = number of independent rows of a matrix

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

- Previously we estimated a mean for each genotype using \mathbf{X}_1 (above), equally we could use \mathbf{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

$$\mathbf{X}_1 = \begin{bmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \end{bmatrix}$$

$$\mathbf{X}_2 = \begin{bmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \\ 1 & 0 & 1 \\ 1 & 0 & 1 \end{bmatrix}$$

Estimation of effects for discrete variables

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

OPTION 1:

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

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

OPTION 2:

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

Model diagnostics

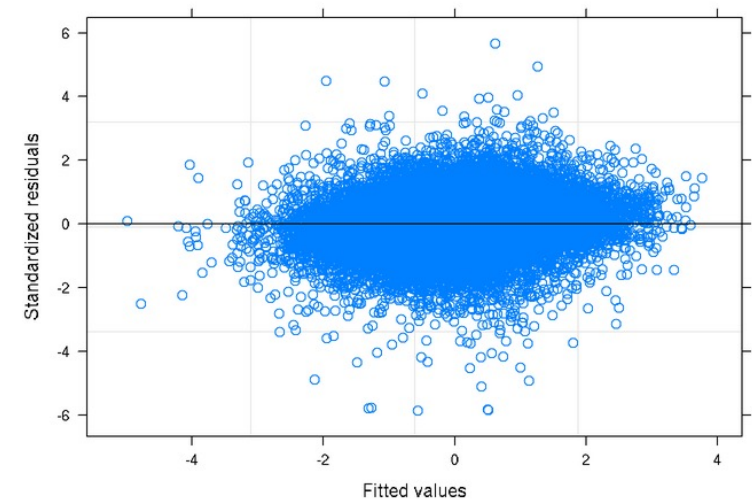
Hypothesis testing in OLS (ordinary least squares) assumes heteroscedastic, uncorrelated errors, i.e. $\boldsymbol{\varepsilon} \sim MVN(0, \mathbf{I}\sigma_e^2)$

It's all about the residuals!

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

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

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



Normal Q-Q Plot

