SM-4331 Advanced Statistics
Chapter 5 (Linear Regression)

Dr Haziq Jamil
FOS M1.09

Universiti Brunei Darussalam

Semester II 2019/20

Outline

Introduction

Regression analysis is one of the most frequently used statistical techniques. It aims to build up an explicit relationship between a response variable and one or more explanatory variables.

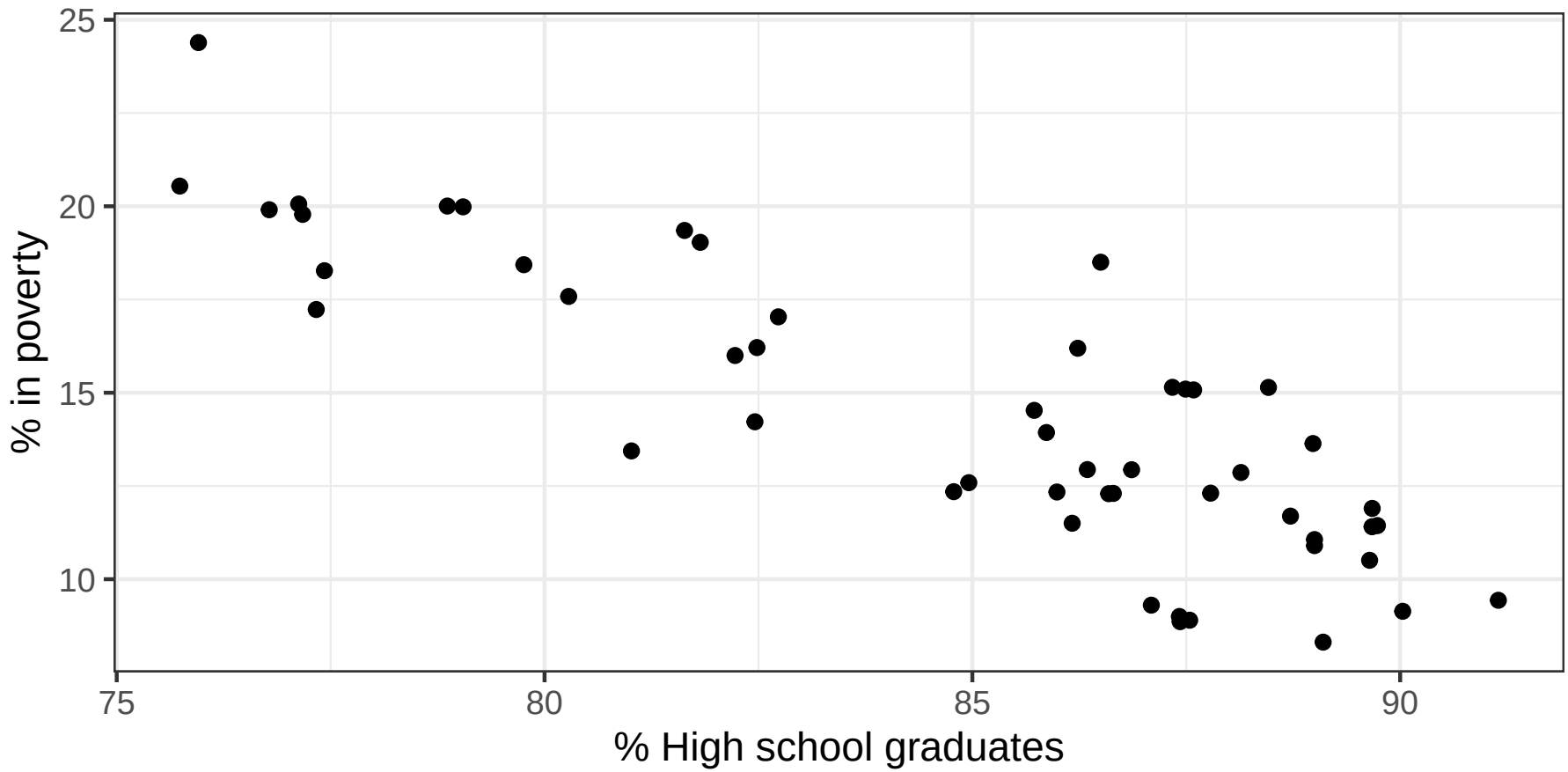
- **Response variables** are often denoted y. Alternative names are *dependent variables* or *output variables*.
- **Explanatory variables** are often denoted $x_1, \ldots, x_p$. Alternative names are *independent variables*, *covariates*, *regressors* or *input variables*.

GOAL: 1) To understand how y depends on $x_1, \ldots, x_p$ linearly; 2) To predict or control unobserved y based on observed $x_1, \ldots, x_p$.

Linear associations

Motivating example

How does the rate of those living under the poverty line in 50 US states plus the federal district vary with the prevalence of high school graduates?



Quantifying linear associations

Correlation describes the strength of the linear association between two variables x and y. The formula is given by

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

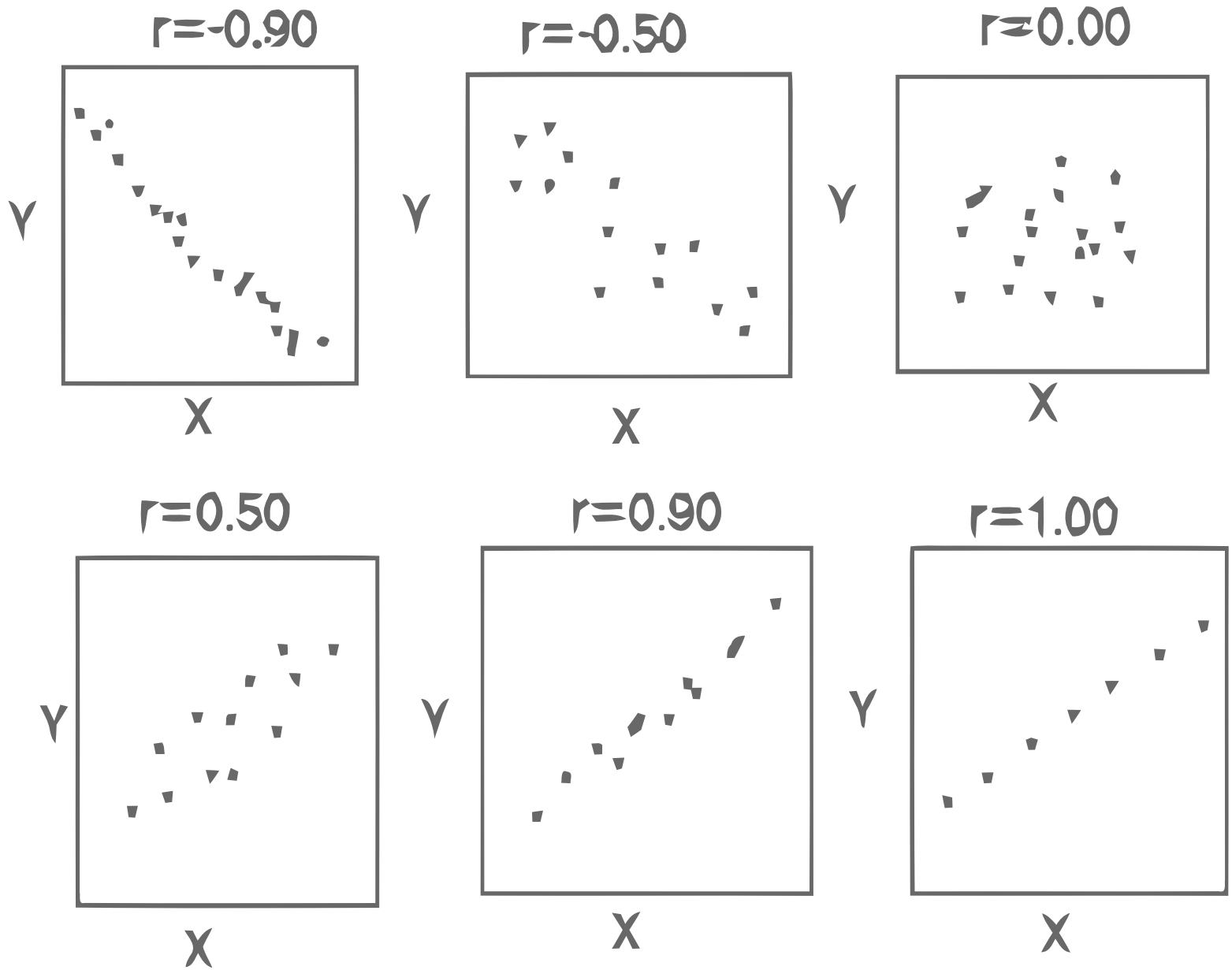
This relates to the the other formula for correlation involving random variables, i.e.

$$\rho = \frac{\text{Cov}(X, Y)}{\sqrt{\text{Var}(X)}\sqrt{\text{Var}(Y)}}$$

As we know, r or ρ takes values between -1 and 1.

- A value of zero indicates no linear association.
- On the other hand, a value of one indicates a perfect linear association.
- In a similar manner, a value of negative one indicates a perfect linear association in the opposite direction.

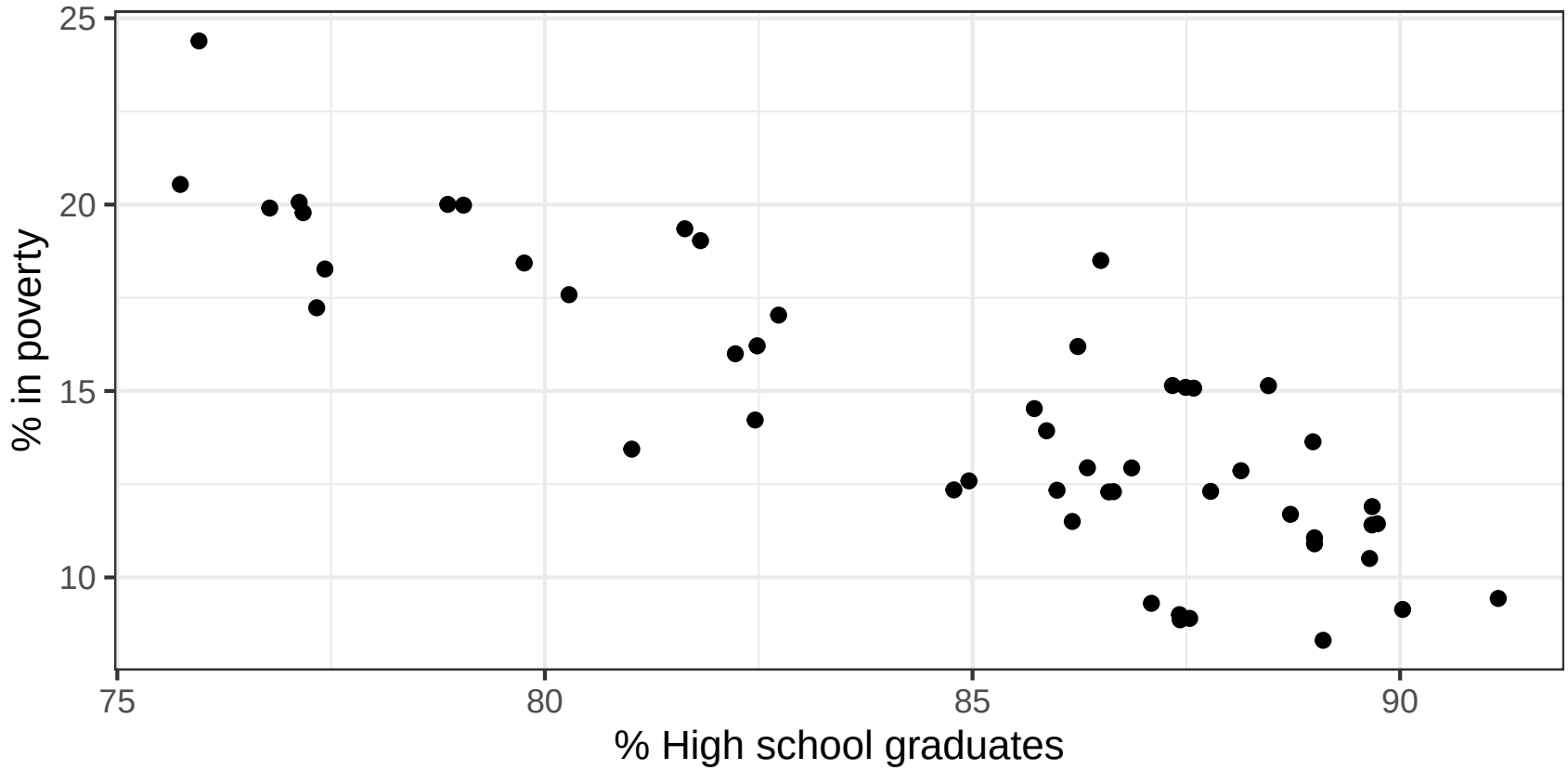
Quantifying linear associations (cont.)



Quantifying linear associations (cont.)

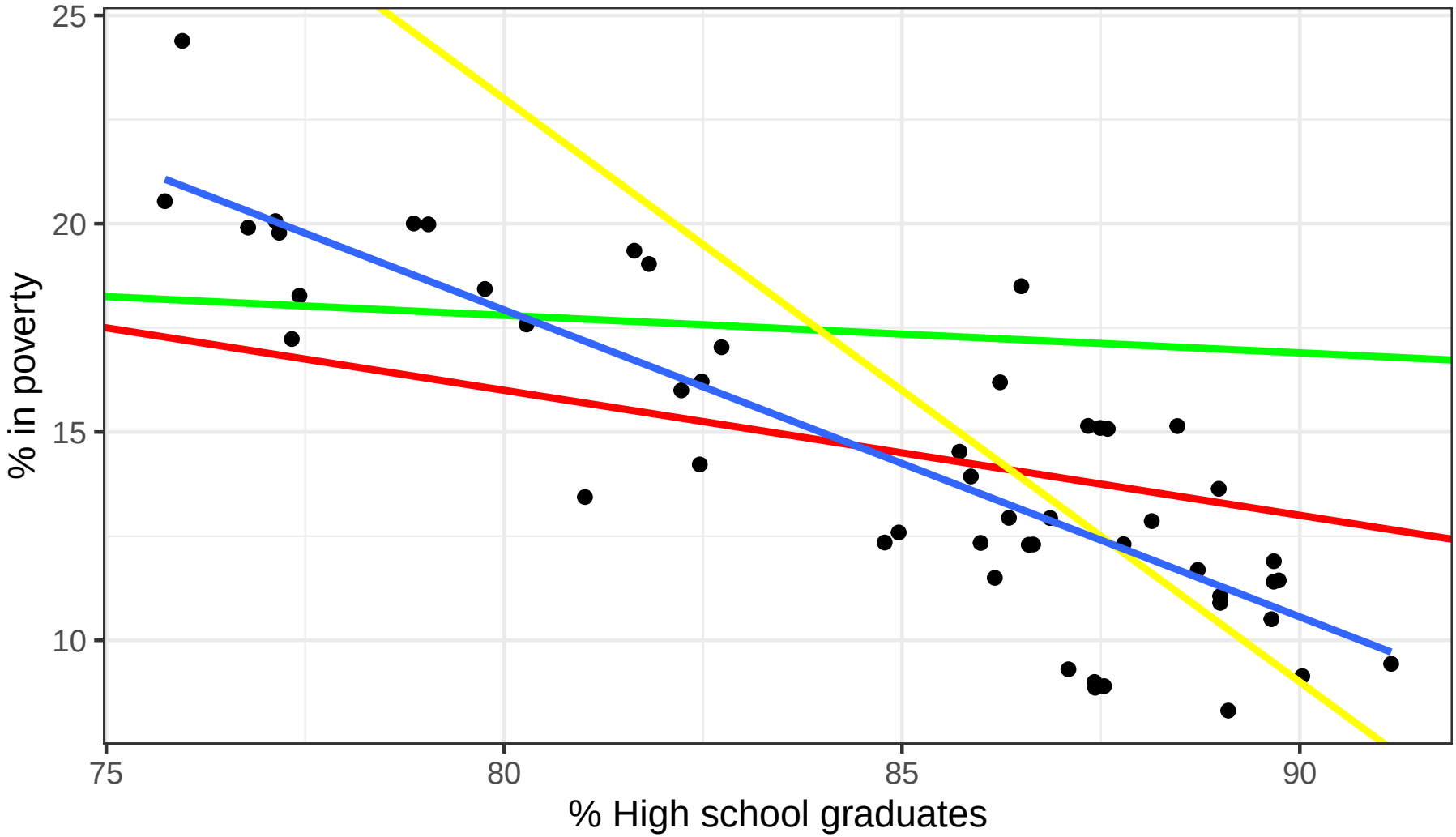
Question

What do you think the correlation between x and y in the data is?



Eyeballing the line

Linear means, roughly speaking, we can pass a straight line through the data points. Which line passes "the best" through all the data points?



Equation of a line

The equation of a line in two dimensions is

$$y = \beta_0 + \beta_1 x.$$

It is parameterised by

- an intercept (or a constant) β_0;
- a slope (or a gradient) β_1.

In the previous example, the y variable is '% in poverty', whereas the x variable is '% high school graduates'. Intuitively, we understand that

- The intercept β_0 gives the poverty rate in states where there is zero high school graduates.
- The slope β_1 gives the *rate of change* of poverty with respect to high school graduates.

Equation of a line (cont.)

Remark

The equation of a plane in $\mathbb{R}^3$ is given by

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2$$

which is a regression model with two covariates.

By extension, the equation of a hyperplane in $\mathbb{R}^{p+1}$ is given by

$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p$$

which is essentially the linear regression model using p covariates. In every case, the relationship between y and $x_1, \ldots, x_p$ is linear.

Imperfections

In most practical situations involving data, we are not able to fit a straight line that passes through all the data points. In a sense, we have these "errors", which we can define by

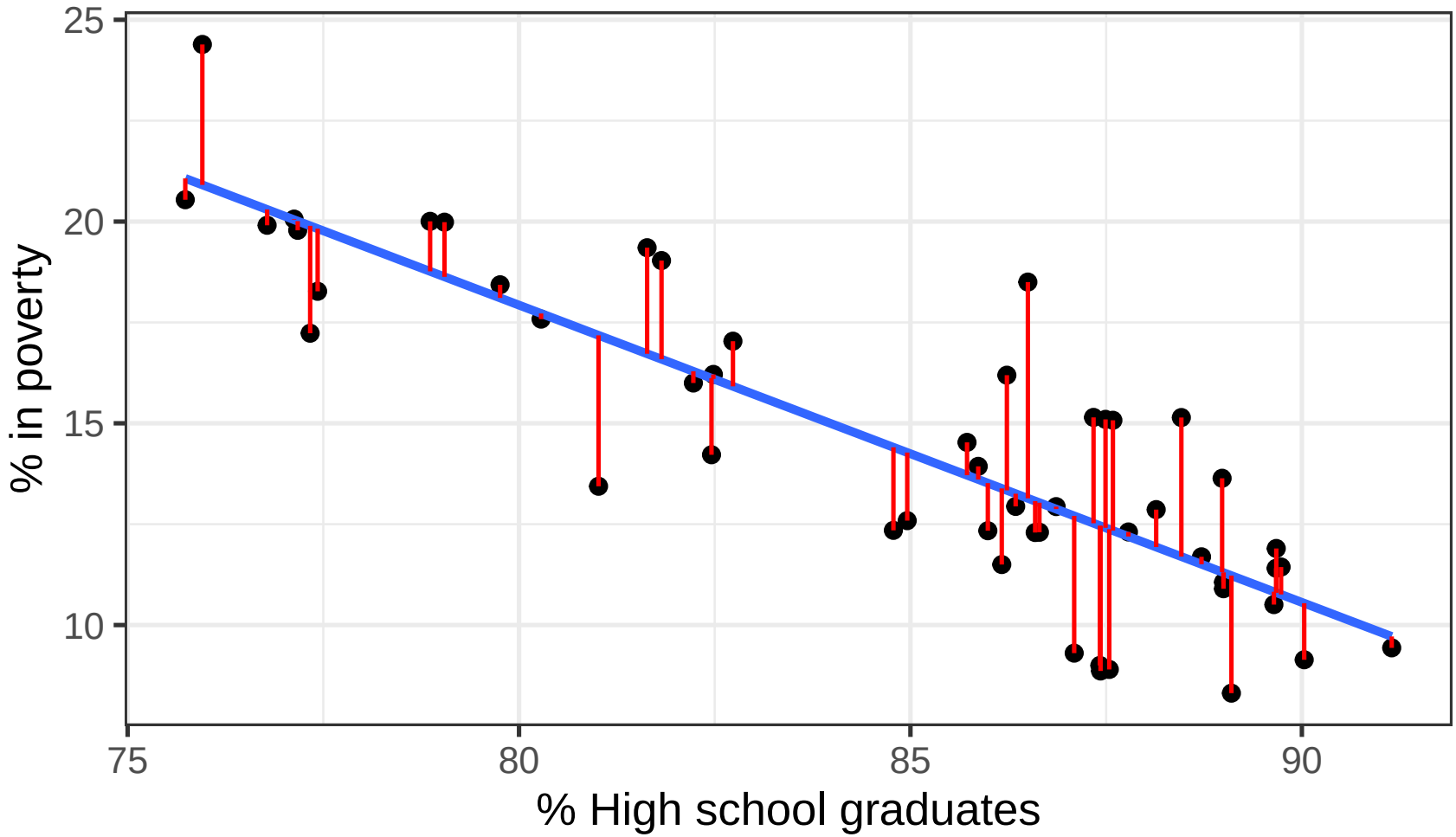
$$\epsilon = y - (\beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p)$$

Errors are a result of various things, such as

- Measurement error.
- Random fluctuations in data collection.
- Deviation of observations from hypothesised linear model.
- Etc.

Optimising the line

It seems that the "best fit" line is the one that minimises all of these errors.



In the next section we describe the technique of "least squares regression".

Linear model

The linear regression model assumes the following relationship between y and x:

$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p + \epsilon.$$

The regressors may be

- Quantitative inputs
- Transformations of quantitative inputs, such as logs, square-root, etc.
- Numeric or "dummy" coding of the levels if qualitative variables.
- Interactions between variables, e.g. $x_3 = x_1 x_2$

Remark

- Indeed, the capacity of the model is large (one can fit as many variables as one likes for many purposes).
- The model is **linear** with respect to the unknown coefficients $\beta_0, \beta_1, \ldots, \beta_p$.

Linear model (cont.)

Example 1

Continuing the motivating example, suppose we fit a linear regression model with $p = 4$.

1. Variable x_1 is the percentage high school graduates.
2. Variable $x_2 = p/(1-p)$, where p is the average state percentage of home ownership. x_2 is then the odds of home ownership.
3. Variable x_3 is defined to be the state's "party majority" (either Republican or Democrat). This qualitative variable can be converted to a "dummy" variable as follows:

$$x_3 = \begin{cases} 1 & \text{if Republican} \\ 0 & \text{if Democrat} \end{cases}$$

Linear model (cont.)

Example 1

4. Variable $x_4 = x_1 x_3$ is the *interaction effect* between state party majority and high school graduates. Effectively,

$$x_4 = \begin{cases} x_1 & \text{if Republican} \\ 0 & \text{if Democrat} \end{cases}$$

The dummy variables and interaction has an interesting effect on the regression line $y = \beta_0 + \sum_{k=1}^{4} \beta_k x_k$:

$$\begin{aligned} y &= \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_1 x_3 \\ &= \begin{cases} \overbrace{(\beta_0 + \beta_3)}^{\tilde{\beta}_0} + \overbrace{(\beta_1 + \beta_4)}^{\tilde{\beta}_1} x_1 + \beta_2 x_2 & \text{if Republican} \\ \beta_0 + \beta_1 x_1 + \beta_2 x_2 & \text{if Democrat} \end{cases} \end{aligned}$$

Linear model (cont.)

Remark

- A qualitative or categorical variable may have more than two distinct outcomes, e.g. faculty $x = \{\text{FOS}, \text{FASS}, \text{FIT}, \text{Others}\}$. If this is the case, we create $m - 1$ dummy variables, where m is the number of distinct outcomes.

$$x_1 = \begin{cases} 1 & \text{if FOS} \\ 0 & \text{otherwise} \end{cases} \quad x_2 = \begin{cases} 1 & \text{if FASS} \\ 0 & \text{otherwise} \end{cases} \quad x_3 = \begin{cases} 1 & \text{if FIT} \\ 0 & \text{otherwise} \end{cases}$$

  Note that the remaining dummy variable is not needed, because if all else is zero, what remains must be the last group by elimination.

- Interactions need not be between continuous and dummy variables only. E.g. strength of metal depending on interaction between temperature and pressure.
- One is not limited to interactions between two variables only.

Data

In practice, we typically have n observations $\{(y_i, x_{i1}, \ldots, x_{ip})\}_{i=1}^{n}$. Each set of observation should follow the line given by the equation

$$y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} + \epsilon_i.$$

Thus in matrix and vector form, we may write

$$\overbrace{\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}}^{\mathbf{y}\in\mathbb{R}^n} = \overbrace{\begin{pmatrix} 1 & x_{11} & \cdots & x_{1p} \\ 1 & x_{21} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \cdots & x_{np} \end{pmatrix}}^{\mathbf{X}\in\mathbb{R}^{n\times(p+1)}} \overbrace{\begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{pmatrix}}^{\boldsymbol{\beta}\in\mathbb{R}^{p+1}} + \overbrace{\begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{pmatrix}}^{\boldsymbol{\epsilon}\in\mathbb{R}^n}$$

i.e. $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$.

Question

Why is there a column of 1's in the X matrix?

Assumptions

- Note that ϵ_i represents the (random) noise in the ith observation.
  - $\mathrm{E}(\epsilon_i) = 0$ and $\mathrm{Var}(\epsilon_i) = \sigma^2$.
  - $\mathrm{Cov}(\epsilon_i, \epsilon_j) = 0, \forall i \neq j$.
  - Sometimes, for convenience, we also assume normality, i.e. $\boldsymbol{\epsilon} \sim \mathrm{N}_n(\mathbf{0}, \sigma^2 \mathbf{I}_n)$.
- The explanatory variables are treated as deterministic variables (i.e. non-random) for technical convenience.
- The regression parameters $\Theta = \{\boldsymbol{\beta}, \sigma^2\}$ are treated as fixed but unknown parameters to be estimated.

From the above, it's easy to see that

$$\mathrm{E}(\mathbf{y}) = \mathbf{X}\boldsymbol{\beta}, \quad \mathrm{Var}(\mathbf{y}) = \sigma^2 \mathbf{I}_n$$

and if $\boldsymbol{\epsilon}$ is assumed to be normally distributed, then so is $\mathbf{y}$.

Remark

The observations $y_1, \ldots, y_n$ are assumed to be uncorrelated (or even independent) but not identically distributed.

Least squares estimation–no intercept

Consider firstly the simple linear regression through the origin (i.e. zero intercept)

$$y_i = x_i\beta + \epsilon_i, \quad i = 1, \ldots, n$$

The least squares estimator $\hat{\beta}$ for β is defined to be the minimiser of

$$\sum_{i=1}^{n} \overbrace{(y_i - x_i\beta)}^{\text{“errors”}}{}^2 = \|\mathbf{y} - \mathbf{x}\beta\|^2 = (\mathbf{y} - \mathbf{x}\beta)^\top(\mathbf{y} - \mathbf{x}\beta),$$

where $\mathbf{y} = (y_1, \ldots, y_n)^\top$ and $\mathbf{x} = (x_1, \ldots, x_n)^\top$.

Recall

$\|\mathbf{a}\| = \|(a_1, \ldots, a_n)^\top\| := \sqrt{a_1^2 + a_2^2 + \cdots + a_n^2} \in \mathbb{R}$ is the called the 2-norm.

Least squares estimation–no intercept (cont.)

Trick: Add and subtract $\mathbf{x}\hat{\beta}$ into the 2-norm

$$\begin{aligned}\|\mathbf{y}-\mathbf{x}\beta\|^2 &= \|\mathbf{y}-\mathbf{x}\hat{\beta}+\mathbf{x}\hat{\beta}-\mathbf{x}\beta\|^2 \\ &= \|\mathbf{y}-\mathbf{x}\hat{\beta}\|^2+\|\mathbf{x}\hat{\beta}-\mathbf{x}\beta\|^2+2(\mathbf{x}\hat{\beta}-\mathbf{x}\beta)^\top(\mathbf{y}-\mathbf{x}\hat{\beta}) \\ &= \|\mathbf{y}-\mathbf{x}\hat{\beta}\|^2+\|(\hat{\beta}-\beta)\mathbf{x}\|^2+2(\hat{\beta}-\beta)\mathbf{x}^\top(\mathbf{y}-\mathbf{x}\hat{\beta})\end{aligned}$$

Now, if we choose a value $\hat{\beta}$ such that $\mathbf{x}^\top(\mathbf{y}-\mathbf{x}\hat{\beta})=0$, then we are left with

$$\begin{aligned}\|\mathbf{y}-\mathbf{x}\beta\|^2 &= \|\mathbf{y}-\mathbf{x}\hat{\beta}\|^2+\|(\hat{\beta}-\beta)\mathbf{x}\|^2 \\ &\geq \|\mathbf{y}-\mathbf{x}\hat{\beta}\|^2\end{aligned}$$

for all β.

Least squares estimation–no intercept (cont.)

Hence the LSE $\hat{\beta}$ is the solution of the equation

$$\mathbf{x}^\top(\mathbf{y} - \mathbf{x}\hat{\beta}) = 0$$
$$\mathbf{x}^\top\mathbf{y} - \mathbf{x}^\top\mathbf{x}\hat{\beta} = 0$$
$$\Rightarrow \hat{\beta} = \frac{\mathbf{x}^\top\mathbf{y}}{\mathbf{x}^\top\mathbf{x}}$$

In other words,

$$\hat{\beta} = \frac{\sum_{i=1}^{n} x_i y_i}{\sum_{i=1}^{n} x_i^2}$$

Least squares estimation–simple regression

Next, consider the simple linear regression

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i, \quad i = 1, \ldots, n$$

The principle is the same, i.e. the LSE $\hat{\beta}_0$ and $\hat{\beta}_1$ minimises

$$\begin{aligned}
&\sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2 \\
&\quad = \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i + \hat{\beta}_0 - \hat{\beta}_1 x_i - \beta_0 - \beta_1 x_i)^2 \\
&\quad = \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2 + \sum_{i=1}^{n} (\hat{\beta}_0 - \hat{\beta}_1 x_i - \beta_0 - \beta_1 x_i)^2 \\
&\qquad + 2 \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)(\hat{\beta}_0 - \hat{\beta}_1 x_i - \beta_0 - \beta_1 x_i)
\end{aligned}$$

Least squares estimation–simple regression (cont.)

As before, if we choose the third term in the expansion above (i.e. the crossproduct term) to be zero, then we see that

$$\begin{aligned}
&\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)^2 \\
&\quad = \sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2 + \sum_{i=1}^{n}(\hat{\beta}_0 - \hat{\beta}_1 x_i - \beta_0 - \beta_1 x_i)^2 \\
&\quad \geq \sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2
\end{aligned}$$

for all values of β_0 and β_1.

Least squares estimation–simple regression (cont.)

Inspecting the crossproduct term further,

$$
\begin{aligned}
&2\sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)(\hat{\beta}_0 - \hat{\beta}_1 x_i - \beta_0 - \beta_1 x_i) \\
&\quad = 2\sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)\big(\hat{\beta}_0 - \beta_0 + (\hat{\beta}_1 - \beta_1)x_i\big) \\
&\quad = 2(\hat{\beta}_0 - \beta_0)\sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) + 2(\hat{\beta}_1 - \beta_1)\sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)x_i.
\end{aligned}
$$

For this crossproduct term to be zero, both terms need to be zero. Specifically, $\hat{\beta}_0$ and $\hat{\beta}_1$ need to be chosen such that

$$
\sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0 \quad \text{and} \quad \sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)x_i = 0
$$

Least squares estimation–simple regression (cont.)

Solving the two simultaneous equations in two unknowns, we can obtain the following results:

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

and

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

You will prove this in one of the questions in Exercise 5.

Least squares estimation

Now back to the general model

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

where $\mathbf{X}$ is an $n \times (p+1)$ matrix. The LSE $\hat{\beta}$ is defined to minimise

$$\begin{aligned}
&\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_{i1} - \cdots - \beta_p x_{ip})^2 \\
&\quad = \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|^2 \\
&\quad = \|\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}} + \mathbf{X}\hat{\boldsymbol{\beta}} - \mathbf{X}\boldsymbol{\beta}\|^2 \\
&\quad = \|\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}\|^2 + \|\mathbf{X}\hat{\boldsymbol{\beta}} - \mathbf{X}\boldsymbol{\beta}\|^2 + 2(\mathbf{X}\hat{\boldsymbol{\beta}} - \mathbf{X}\boldsymbol{\beta})^\top(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})
\end{aligned}$$

As usual, choosing the $\hat{\boldsymbol{\beta}}$ such that the crossproduct term disappears, we get that

$$\|\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}\|^2 + \|\mathbf{X}\hat{\boldsymbol{\beta}} - \mathbf{X}\boldsymbol{\beta}\|^2 \geq \|\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}\|^2$$

Least squares estimation (cont.)

That is, the LSE solution satisfies

$$\begin{aligned} 2(\mathbf{X}\hat{\boldsymbol{\beta}} - \mathbf{X}\boldsymbol{\beta})^\top(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}) &= 0 \\ 2(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})^\top\mathbf{X}^\top(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}) &= 0 \\ \Rightarrow \mathbf{X}^\top(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}) &= 0 \\ \mathbf{X}^\top\mathbf{y} &= \mathbf{X}^\top\mathbf{X}\hat{\boldsymbol{\beta}} \\ \Rightarrow \hat{\boldsymbol{\beta}} = (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{y} & \end{aligned}$$

Remark

We did not require the normality assumption for the errors. It turns out that using the maximum likelihood method we get the exact same solution. This formula is very important for every statistician to know!

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{y}$$

Least squares estimation (cont.)

Example 2

For the motivating example (poverty data set), we have the following statistics:

$$\bar{x} = 84.77, \quad \bar{y} = 14.41, \quad \sum_{i=1}^{51}(x_i - \bar{x})(y_i - \bar{y}) = -718.49,$$

$$\sum_{i=1}^{51}(x_i - \bar{x})^2 = 975.38, \quad \sum_{i=1}^{51}(y_i - \bar{y}) = 738.78.$$

Least squares estimation (cont.)

Example 2

The LSE $\hat{\beta}_1$ is given by

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{-718.49}{975.38} = -0.73662$$

while the LSE $\hat{\beta}_0$ is given by

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1\bar{x} = 14.41 - (-0.73662)84.77 = 76.85835$$

The equation of the best fit line is therefore

$$y = 76.9 - 0.737x.$$

Maximum likelihood estimation

Suppose we assume that $\boldsymbol{\epsilon} \sim \mathrm{N}_n(\mathbf{0}, \mathbf{I}_n)$. Then

$$\mathbf{y} \sim \mathrm{N}_n(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I}_n),$$

and the likelihood function for the parameters is given by

$$L(\boldsymbol{\beta}, \sigma^2) = \frac{1}{(\sigma\sqrt{2\pi})^n} \exp\left\{-\frac{1}{2\sigma^2}\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|^2\right\}.$$

Since we proved that the LSE $\hat{\beta}$ satisfies $\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|^2 \geq \|\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}\|^2$ for all values of $\boldsymbol{\beta}$, it must be that case that the exponent is maximised at $\hat{\boldsymbol{\beta}}$, so

$$L(\hat{\boldsymbol{\beta}}, \sigma^2) \geq L(\boldsymbol{\beta}, \sigma^2)$$

which implies that $\hat{\boldsymbol{\beta}}$ is also the ML estimate.

Expectation and variance of $\hat{\boldsymbol{\beta}}$

Since $\mathrm{E}(\mathbf{y}) = X\boldsymbol{\beta}$, $\mathrm{Var}(\mathbf{y}) = \sigma^2\mathbf{I}_n$, and the $\mathbf{X}$ matrix is fixed,

- **Expectation of $\hat{\boldsymbol{\beta}}$.**

$$\begin{aligned}\mathrm{E}(\hat{\boldsymbol{\beta}}) &= \mathrm{E}[(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{y}] \\ &= (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\,\mathrm{E}(\mathbf{y}) \\ &= \cancel{(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{X}}\,\boldsymbol{\beta}\end{aligned}$$

- **Variance of $\hat{\boldsymbol{\beta}}$.**

$$\begin{aligned}\mathrm{Var}(\hat{\boldsymbol{\beta}}) &= \mathrm{Var}[(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{y}] \\ &= (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\,\mathrm{Var}(\mathbf{y})\mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1} \\ &= (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\,\sigma^2\mathbf{I}_n\,\mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1} \\ &= \sigma^2(\mathbf{X}^\top\mathbf{X})^{-1}\end{aligned}$$

Expectation and variance of $\hat{\boldsymbol{\beta}}$ (cont.)

Example 3

In the special case of $p = 1$ (i.e. the simple linear regression), we have that

$$\begin{aligned}
\sigma^2(\mathbf{X}^\top\mathbf{X})^{-1} &= \sigma^2 \left(\begin{pmatrix} 1 & \dots & 1 \\ x_1 & \dots & x_n \end{pmatrix} \begin{pmatrix} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix} \right)^{-1} \\
&= \sigma^2 \begin{pmatrix} n & \sum_i x_i \\ \sum_i x_i & \sum_i x_i^2 \end{pmatrix}^{-1} \\
&= \frac{\sigma^2}{n\sum_i x_i^2 - n^2\bar{x}^2} \begin{pmatrix} \sum_i x_i^2 & -n\bar{x} \\ -n\bar{x} & n \end{pmatrix} \\
&= \frac{\sigma^2}{n\sum_i (x_i - \bar{x})^2} \begin{pmatrix} \sum_i x_i^2 & -n\bar{x} \\ -n\bar{x} & n \end{pmatrix}
\end{aligned}$$

Estimating σ^2

Under the assumption $\boldsymbol{\epsilon} \sim \mathrm{N}_n(\mathbf{0}, \sigma^2 \mathbf{I}_n)$, the MLE for σ^2 is $\|\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}\|^2/n$. But in practice we use the following estimator:

$$\hat{\sigma}^2 = \frac{1}{n-p-1}\|\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}\|^2$$

which is unbiased, i.e. $\mathrm{E}(\hat{\sigma}^2) = \sigma^2$. The proof for this is omitted, but can be obtained by doing a Google search.

Fitted and residual values

Remember when we said that all of the data points are not expected to fall on the regression line?

Definition 4 (Fitted values)

The points that are meant to be on the regression line, called the **fitted values**, are given by

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \cdots + \hat{\beta}_p x_{ip}, i = 1, \ldots, n.$$

Definition 5 (Residuals)

The difference between the observed value and the fitted value is called the residual

$$\hat{\epsilon}_i = y_i - \hat{y}_i, i = 1, \ldots, n.$$

Initially we called residuals "errors", which is actually an alternative name for it.

Fitted and residual values (cont.)

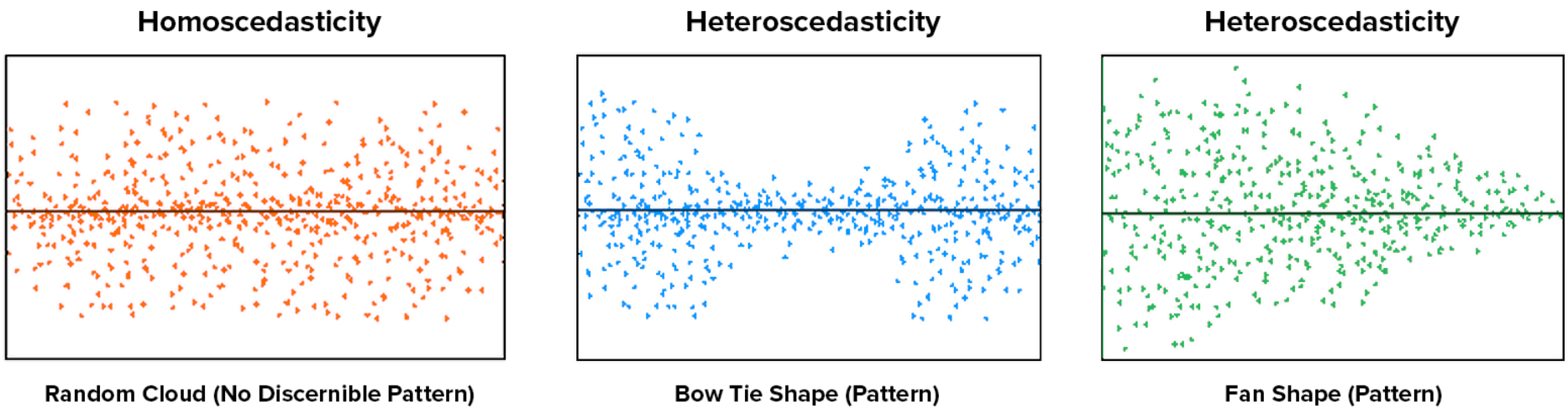
If the model is correct, the residuals should behave like random noise. One can prove that

$$\mathrm{E}(\hat{\boldsymbol{\epsilon}}) = 0$$

and

$$\mathrm{Var}(\hat{\boldsymbol{\epsilon}}) = \sigma^2\left(\mathbf{I}_n - \mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\right).$$

An effective way of checking is to plot $\hat{\epsilon}_i$ against y_i, or against x_{ij} for $j = 1, \ldots, p$. Ideally we want a 'random noise' pattern to emerge.



Otherwise, it may be indicative of variable transformation being required.

Fitted and residual values (cont.)

Another important property of the residuals is that

$$\sum_{i=1}^{n} \hat{\epsilon}_i = \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \cdots - \hat{\beta}_p x_{ip}) = \sum_{i=1}^{n} (y_i - \mathbf{x}_i^\top \hat{\boldsymbol{\beta}}) = 0$$

This comes from the *first order condition* of minimising the sum of squared residuals,

$$\hat{\boldsymbol{\beta}} := \arg\min_{\boldsymbol{\beta}} \sum_{i=1}^{n} (y_i - \mathbf{x}_i^\top \boldsymbol{\beta})^2.$$

This implies that the following minimising condition must be satisfied:

$$\frac{\partial}{\partial \boldsymbol{\beta}} \sum_{i=1}^{n} (y_i - \mathbf{x}_i^\top \boldsymbol{\beta})^2 \Big|_{\boldsymbol{\beta}=\hat{\boldsymbol{\beta}}} = 2 \sum_{i=1}^{n} (y_i - \mathbf{x}_i^\top \hat{\boldsymbol{\beta}}) = 0.$$

Predicted values

Besides fitted values, we can use the model to obtain **predicted values** for observations that we have not seen before.

Let $\mathbf{x}_{new} = \left(1, x_{new,1}, \ldots, x_{new,p}\right)^\top$. According to the model, $y_{new} = \mathbf{x}_{new}^\top \boldsymbol{\beta} + \epsilon$, and

$$\mathrm{E}(y_{new}) = \mathbf{x}_{new}^\top \boldsymbol{\beta}$$

which is estimated by $\mathbf{x}^{(new)\top} \hat{\boldsymbol{\beta}}$. We use this value as a point estimate of the *predicted value* given this new set of values.

Remark

There is variability in obtaining the predicted value, since $\hat{\boldsymbol{\beta}}$ itself is random (and also ϵ). This implies that the predicted value is also random, i.e. it has a mean and variance and a distribution. We can thusly build confidence sets for the predicted value (see the remark beginning on slide 78).

Gauss-Markov Theorem

The LSE $\hat{\boldsymbol{\beta}}$ is the best linear unbiased estimator (BLUE) for $\boldsymbol{\beta}$. In other words, $\hat{\boldsymbol{\beta}}$ has the minimum variance among all the linear unbiased estimators for $\boldsymbol{\beta}$.

Theorem 6 (Gauss-Markov Theorem)

For the regression model $\mathbf{y} = \mathbf{X} + \boldsymbol{\epsilon}$, where $\mathrm{E}(\boldsymbol{\epsilon}) = 0$ and $\mathrm{Var}(\boldsymbol{\epsilon}) = \sigma^2 \mathbf{I}_n$, $\hat{\boldsymbol{\beta}} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y}$ is the BLUE of $\boldsymbol{\beta}$ in the sense that

$$\mathrm{Var}(\mathbf{B}\mathbf{y}) - \mathrm{Var}(\hat{\boldsymbol{\beta}}) \geq 0$$

is a positive semi-definite matrix for any $p \times n$ constant matrix $\mathbf{B}$ for which $\mathrm{E}(\mathbf{B}\mathbf{y}) = \boldsymbol{\beta}$.

Remark

For $k = 1, \ldots, p$, $\hat{\beta}_k$ is the MVUE (linear) estimator for β_k.

Wald tests

When we do not make any distributional assumptions on the errors, then we don't know what the distribution of $\hat{\beta}$ is. However can rely on the asymptotic normality of $\hat{\beta}$ (being an MLE):

$$\frac{\hat{\beta} - \beta}{\mathrm{SE}(\hat{\beta})} \xrightarrow{\mathrm{D}} \mathrm{N}(0, 1)$$

as $n \to \infty$. From this, we can construct Wald test statistics to test the hypothesis $H_0 : \beta = b$ against the alternative $H_1 : \beta \neq b$. For example,

- $H_0 : \beta_1 = 0$ v.s. $H_1 : \beta_1 \neq 0$ "*is the slope significant?*"
- $H_0 : \beta_1 = 0$ v.s. $H_1 : \beta_1 > 0$ "*is the slope positive?*"
- $H_0 : \beta_1 = a + c\beta_0$ v.s. $H_1 : \beta_1 \neq a + c\beta_0$ "*is the slope a linear combination of the intercept?*"

Normal linear regression models

On the other hand, assuming normality of the errors, i.e. $\boldsymbol{\epsilon} \sim \mathrm{N}_n(\mathbf{0}, \sigma^2 \mathbf{I}_n)$, allows us to derive the test statistics for $\hat{\boldsymbol{\beta}}$ and $\hat{\sigma}^2$ exactly.

Theorem 7

If $\boldsymbol{\epsilon} \sim \mathrm{N}_n(\mathbf{0}, \sigma^2 \mathbf{I}_n)$, then

- $\hat{\boldsymbol{\beta}} \sim \mathrm{N}\left(\boldsymbol{\beta}, \sigma^2 (\mathbf{X}^\top \mathbf{X})^{-1}\right)$;
- $(n - p - 1)\hat{\sigma}^2 / \sigma^2 \sim \chi^2_{n-p-1}$; and
- $\hat{\boldsymbol{\beta}}$ and $\hat{\sigma}^2$ are independent.

The proof of this theorem requires a little more attention, so for brevity, is omitted. However, many proofs of this theorem exists and can be found easily in other sources.

t-tests for β_j

By Theorem 7, we clearly see that $\hat{\beta}_k \sim \mathrm{N}(\beta_k, \sigma^2 v_{kk})$, where v_{kk} is the $(k+1, k+1)$th entry of the matrix $(\mathbf{X}^\top \mathbf{X})^{-1}$. Therefore,

$$T = \frac{\overbrace{(\hat{\beta}_k - \beta_k)/\sigma\sqrt{v_{kk}}}^{\sim \mathrm{N}(0,1)}}{\sqrt{\frac{\overbrace{(n-p-1)\hat{\sigma}^2/\sigma^2}^{\sim \chi^2_{n-p-1}}}{n-p-1}}} = \frac{\hat{\beta}_k - \beta_k}{\underbrace{\hat{\sigma}\sqrt{v_{kk}}}_{\mathrm{SE}(\hat{\beta}_k)}} \sim t_{n-p-1}$$

Thus, we reject the null hypothesis $H_0 : \beta_k = b$ for large values as compared to the t_{n-p-1} distribution.

Remark

When n is large, then t-test $\approx$ Wald test.

F-tests for β

An important F-test for regression analysis is the so-called "overall F-test" for the regression coefficients. The hypothesis to be tested is

$$H_0 : \beta_1 = \cdots = \beta_p = 0 \quad \text{v.s.} \quad H_1 : \text{not all } \beta_j \text{ are zero}$$

Under the null hypothesis,

$$y_i \sim \mathrm{N}(\overbrace{\beta_0}^{\mu}, \sigma^2), i = 1, \ldots, n.$$

As we know, the MLE for β_0 is $\bar{y}$. Since each fitted value is going to be the same ($\bar{y}$), the **total sum of squares (SS)** is given by

$$\sum_{i=1}^{n} (y_i - \bar{y})^2$$

F-tests for $\boldsymbol{\beta}$ (cont.)

We introduce the decomposition:

$$\overbrace{\sum_{i=1}^{n}(y_i - \bar{y})^2}^{\text{Total SS}} = \overbrace{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}^{\text{Reg SS}} + \overbrace{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}^{\text{Resid SS}}$$

where $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \cdots + \hat{\beta}_p x_{ip}$ are the fitted values under the unrestricted model (H_1).

- The term of the LHS (Total SS) is the total variation in the data $\{y_i\}$.
- The residual sum of squares tells us the total deviation of our fitted values from the observed data y.
- The regression sum of squares tells us how much our fitted values differ from the mean value of the observed data y.

F-tests for β (cont.)

Notice that

$$\sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 + \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$
$$\geq \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

or to put it simply, $\tilde{F} = \text{Total SS}/\text{Resid SS} \geq 1$. The statistical question here is to determine whether or not the statistic $\tilde{F}$ is greater than 1 or not.

Remark

Note that this implies that the sum of residuals from the null model (intercept only) is always going to be larger or equal than the sum of residuals from an unrestricted model. Generally speaking, the more variables you add to the model, the smaller its sum of residuals are.

F-tests for β (cont.)

Lemma 8

For the normal linear regression model,

- $\frac{1}{\sigma^2}\text{Total SS} \sim \chi^2_{n-1}$
- $\frac{1}{\sigma^2}\text{Reg SS} \sim \chi^2_p$
- $\frac{1}{\sigma^2}\text{Resid SS} \sim \chi^2_{n-p-1} \Rightarrow (n-p-1)\hat{\sigma}^2/\sigma^2 \sim \chi^2_{n-p-1}$

Using the above lemma, we can build the test statistic

$$\tilde{F} = \frac{\text{Total SS}}{\text{Resid SS}} = \frac{\text{Resid SS} + \text{Reg SS}}{\text{Total SS}}$$
$$= 1 + \underbrace{\frac{\frac{1}{\sigma^2}\text{Reg SS}/p}{\frac{1}{\sigma^2}\text{Resid SS}/(n-p-1)}}_{F \sim F_{p,n-p-1}}$$

which suggests comparing this test statistic against critical values from the $F_{p,n-p-1}$ distribution in order to reject H_0.

The ANOVA table for regression

This gives rise to the following ANOVA table.

Source	SS	d.f.	MSS	F-statistic
Regressors	$\sum_i(\hat{y}_i - \bar{y})^2$	p	$\frac{\sum_i(\hat{y}_i-\bar{y})^2}{p}$	$\frac{\sum_i(\hat{y}_i-\bar{y})^2/p}{\sum_{i,j}(X_{ij}-\bar{X}_j)^2/(n-p-1)}$
Residuals	$\sum_i(y_i - \hat{y}_i)^2$	$n-p-1$	$\frac{\sum_i(y_i-\hat{y}_i)^2}{n-p-1}$	
Total	$\sum_i(y_i - \bar{y})^2$	$n-1$		

The above ANOVA table and its corresponding F-statistic is used for the overall F-test for the regression coefficients.

F-tests for $\boldsymbol{\beta}$ (cont.)

Remark

More generally, one uses the F-test to test any kind of restrictions on the coefficients $\boldsymbol{\beta}$. This can be motivated by two ways:

- Deducing that the test statistic for restrictions on the $\boldsymbol{\beta}$, for e.g. $H_0 : \beta_1 = 0, \beta_2 = 4, \beta_3 = \beta_4$ is a ratio of χ^2 variates, thus has an F-distribution.
- The *restricted LSE* $\tilde{\boldsymbol{\beta}}$ can be obtained, and one can show that the scaled difference between the sum of squares of the restricted and unrestricted model follows a χ^2-distribution, i.e.

$$\left(\text{Resid SS}(\tilde{\boldsymbol{\beta}}) - \text{Resid SS}(\hat{\boldsymbol{\beta}})\right)/\sigma^2 \sim \chi^2_k$$

  where k is the number of restriction imposed, and $\hat{\boldsymbol{\beta}}$ is the LSE of the unrestricted model. From this fact, an F-statistic can be built as well.

In addition, one can also use the LRT to compare two nested models. Proving all the facts above are beyond the scope of this module.

The ANOVA table for regression (cont.)

Example 9

Let's revisit the ANOVA problem, but formulated slightly differently. Consider the regression model $y_{ij} = \beta_0 + \sum_{k=1}^{m} \beta_k x_{ik}$, where $y \in \mathbb{R}$ and x_j is a dummy variable to indicate which group the observation belongs to, i.e. $x_{ik} = 1$ if the observation y_{ij} belongs to group $k \in \{1, \ldots, m\}$, and 0 otherwise.

Technically we need one constraint on the regression model (otherwise it is not identifiable)–think of this way: there are information from m groups in total, but we have to estimate $m + 1$ parameters.

We usually set $\beta_m = 0$ (cornerpoint constraint). Alternatively we can also set any of the other β_j to be zero, or the "grand mean" β_0 to be zero.

The ANOVA table for regression (cont.)

Example 9

There are n_j observations in each group $j = 1, \ldots, m$, and the total number of observations is $n = n_1 + \cdots + n_m$. The data looks like this

Group 1	Group 2	$\cdots$	Group m
y_{11}	y_{12}	$\cdots$	y_{1m}
y_{21}	y_{12}	$\cdots$	y_{1m}
$\vdots$	$\vdots$	$\vdots$	$\vdots$
$y_{n_1 1}$	$y_{n_2 2}$	$\cdots$	$y_{n_m m}$

We can even estimate the sample group means $\bar{y}_j = \sum_{i=1}^{n_j} y_{ij}/n_j$

Recall that in an ANOVA setting, we assume $y_{ij} \sim \mathrm{N}(\mu_j, \sigma^2)$, and we're interested in testing whether all of these means $\mu_1, \ldots, \mu_m$ are equivalent or not.

The ANOVA table for regression (cont.)

Example 9

Under the normal linear regression setting, we have $y_{ij} \sim \mathrm{N}\left(\mathrm{E}(y_{ij}), \sigma^2\right)$, where

$$\mathrm{E}(y_{ij}) = \begin{cases} \mu_1 := \beta_0 + \beta_1 & \text{if all except } x_{i1} = 1 \\ \mu_2 := \beta_0 + \beta_2 & \text{if all except } x_{i2} = 1 \\ \vdots & \\ \mu_m := \beta_0 & \text{if all } x_{ij} = 0 \end{cases}$$

So the regression model is estimating m different means (one from each group). Specifically, the mean of group j is $\beta_0 + \beta_j$ for $j = 1, \ldots, m-1$, and the mean of group $j = m$ is β_0 (since we constrained $\beta_m = 0$).

The ANOVA table for regression (cont.)

Example 9

The ANOVA null hypothesis under the regression setting is written

$$H_0 : \beta_0 + \beta_1 = \cdots = \beta_0 + \beta_{m-1} = \beta_0$$

which is tested against the alternative hypothesis that not all means are equal. Subtracting β_0 from all sides of the equality we get

$$H_0 : \beta_1 = \cdots = \beta_{m-1} = 0$$

which is the overall F-test null hypothesis for the regression coefficients. This F-test statistic has distribution $F_{m-1,n-m}$ is **identical** to the ANOVA F-test statistic that we saw in Chapter 3.

The ANOVA table for regression (cont.)

Remark

We have seen how the classical ANOVA test can be reparameterised into a regression model. Interestingly,

- We can add interaction effects to the ANOVA regression model to do two-way ANOVA, three-way ANOVA, etc. (multiway ANOVA).
- We can add additional covariates to the model as well, e.g. controlling for age and sex in randomised clinical trials.

Remark

When we are testing the means of two groups ($m = 2$), of which there are n_1 and n_2 samples in each group. Then under the regression setting we use a test statistic $T^2 \sim F_{1,n_1+n_2-2}$, which implies that $T \sim t_{n-2}$. This is exactly the test statistic for the two-sample t-test.

Coefficient of determination

Definition 10 (Coefficient of determination)

For a linear regression model, the coefficient of determination, denoted R^2, is given by

$$R^2 = \frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} = \frac{\text{Reg SS}}{\text{Total SS}} \in [0, 1]$$

It measures the amount of *model agreement*, i.e. the proportion in which the total variation in $\{y_i\}$ is explained by all the regressors.

Another way of writing R^2 is

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} = 1 - \frac{\text{Resid SS}}{\text{Total SS}}.$$

From this we can clearly see that the smaller the residuals are, the better the fit of the model, and hence R^2 is closer to 1.

Coefficient of determination (cont.)

Remark

- Other names for R^2 include regression correlation coefficient.
- We don't know the distribution of R^2 in general, so often times the *Adjusted* R^2 is used instead

$$\tilde{R}^2 = 1 - \overbrace{\frac{\text{Resid SS}/(n-p-1)}{\text{Total SS}/(n-1)}}^{F^{-1}} = 1 - \frac{n-1}{n-p-1}(1-R^2)$$

  and the term denoted by X can be shown to have an F-distribution.
- For large n, $\tilde{R}^2 \approx R^2$.

Model selection criteria

When we are building a model for inference, then it is important that the β coefficients included in the model are significant (using Wald tests, t-tests, and F-tests).

We can also look at some model selection criteria to choose between two or more competing models. In addition to

- Residual standard error $\hat{\sigma}$ (lower is better)
- Coefficient of determination R^2 (higher is better)

here are several other model selection criteria

- Max. log-likelihood value $\hat{l} := l(\hat{\beta}, \hat{\sigma}^2)$ (higher is better)
- Akaike information criterion AIC $= -2\hat{l} + 2p$ (lower is better)
- Bayesian information criterion AIC $= -2\hat{l} + p \log n$ (lower is better)
- Mallow's $C_p = (\text{Resid SS} + 2p\hat{\sigma}^2)/n$ (lower is better)

Variable selection

In a regression setting, model selection is essentially synonymous with variable selection. That is, from a set of p predictors $\{x_1, \ldots, x_p\}$, the task is to select a subset of the variables that "best fits" the data (perhaps using one of the criterion mentioned above).

Remark

A model with fewer predictors improves the interpretability, and practically speaking fewer variables need to be collected during collection stage. More importantly, it leads to more stable statistical inference.

If each variable can be selected or not, there are a total of 2^p possible models. This gets large exponentially as p grows, so doing pairwise comparison of all models is not feasible! How to do it in practice then?

Stepwise regression

The most common method is a stepwise addition and deletion scheme.

Step 1. Start with an initial model (e.g. intercept-only model).

Step 2.

(a) Specify one variable such that by adding it to the model, the decrease of the Resid SS is maximised among all candidate variables.
(b) Only add the variable specified in 2.(a) to the model if the value of the AIC (or Mallow's C_p) decreases.

Step 3.

(a) Specify one variable among all the variables in the model, such that by deleting it from the model, the increase of the Resid SS is minimised.
(b) Only delete the variable specified in 3.(a) if the value of the AIC (or Mallow's C_p) decreases.
(c) Repeat 3.(a) and 3.(b) such that no more variables can be deleted.

Step 4. Repeat Steps 2 and 3 above until no more variables can be added to or deleted from the model.

Stepwise regression

Remark

- Luckily, this is very easily done using the `step()` function in base R. It is implemented using the AIC.
- Starting with the null model and adding in variables is called *forward selection*. Starting with the full model and deleting variables is called *backwards elimination*.

Inspect the data set

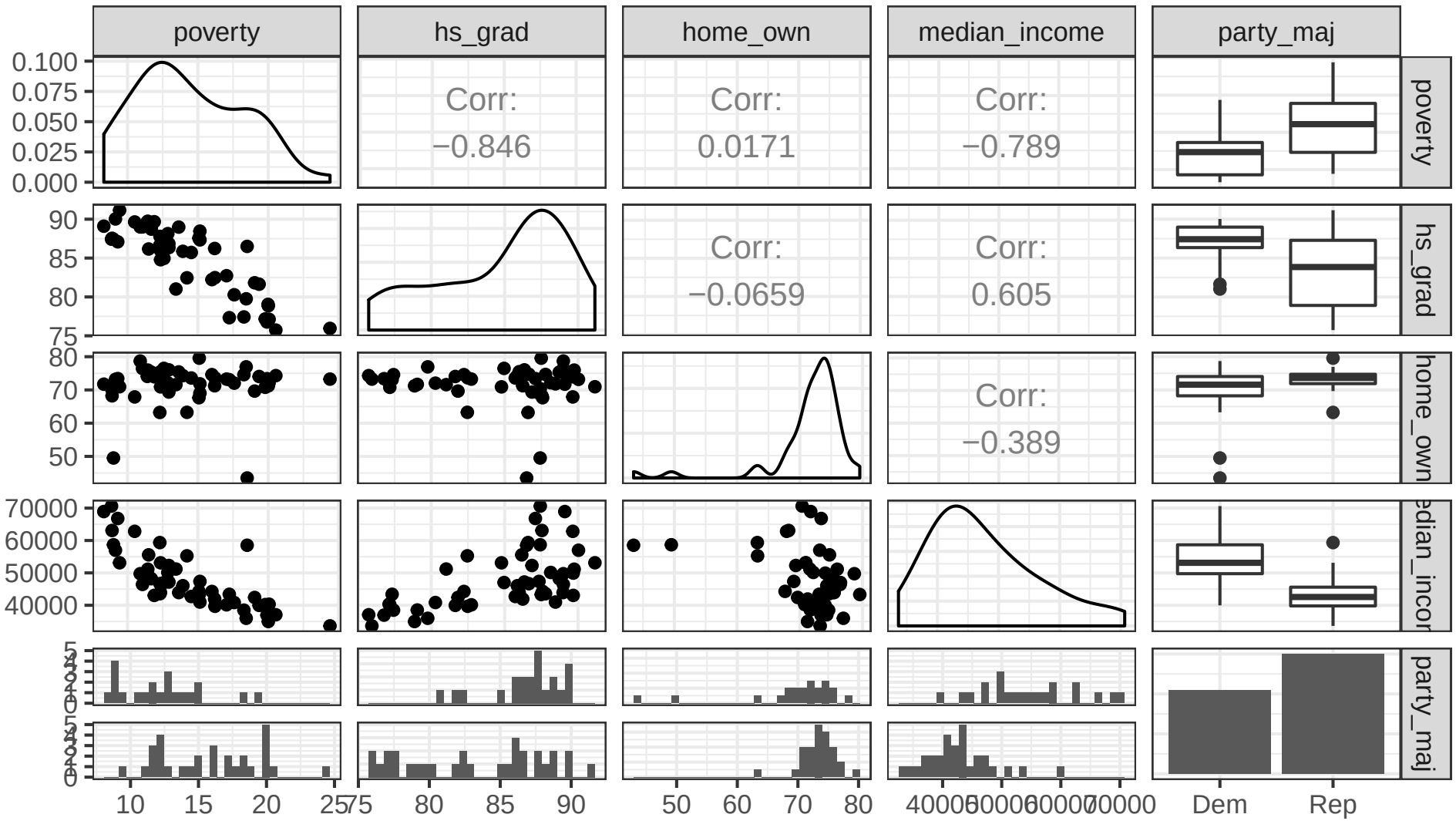
Back to poverty data set: It contains $n = 51$ observations and $p = 4$ predictors (% high school graduates, % home ownership, median income in US$ per year, and party majority).

```
poverty
```

```
## # A tibble: 51 x 6
##    state                poverty hs_grad home_own median_income party_maj
##    <chr>                  <dbl>   <dbl>    <dbl>         <dbl> <chr>
##  1 Alabama                19.9     76.8     73.4         36963. Rep
##  2 Alaska                 12.3     86.6     63.2         59351. Rep
##  3 Arizona                19.0     81.8     69.7         42418. Rep
##  4 Arkansas               20.0     78.9     71.2         34983. Rep
##  5 California             14.2     82.5     63.3         55266. Dem
##  6 Colorado               12.9     88.1     72.1         50136. Dem
##  7 Connecticut             8.31    89.1     71.7         68935. Dem
##  8 Delaware               11.5     86.2     74.7         55568. Dem
##  9 District of Columbia   18.5     86.5     43.5         58526  Dem
## 10 Florida                16.0     82.2     74.6         44269. Rep
## # ... with 41 more rows
```

Inspect the data set (cont.)

```
GGally::ggpairs(poverty[, -1]) + theme_bw()
```



Inspect the data set (cont.)

As part of an exploratory data analysis, we look for

- Inspect distribution of data and look for outliers. Notice that the *scale* of the `median_income` variable is in the thousands, the only one different to the rest.
- Pairwise patterns between variables (including response variables)–any linear/non-linear relationships? Based on what is seen, a linear model seems reasonable.
- Identify strong relationships (correlations) between variables. We see strong correlations between `poverty` and covariates `hs_grad` (-0.846) and `median_income` (-0.789), but not with `home_own` (0.0171).
- Ideally no extremely strong relationship between covariates, otherwise we face the issue of *collinearity*. In this case it looks OK (maximum correlation of 0.605 between covariates).

Remark

Correlations are calculated only among continuous variables.

Collinearity

This is a huge problem for linear regression. Consider the linear regression model $y = \beta_0 + \beta_1 x + \beta_2 z + \epsilon$. Suppose that the variables x and z are perfectly correlated, i.e. x and z are linearly related:

$$z = a + cx \Rightarrow \text{Cov}(z, x) = \text{Cov}(a + cx, x) = c\,\text{Var}(x)$$
$$\Rightarrow \rho(z, x) = \frac{c\,\text{Var}(x)}{\sqrt{c^2\,\text{Var}(x)\,\text{Var}(x)}} = 1.$$

This then implies that in the regression model,

$$\begin{aligned} y &= \beta_0 + \beta_1 x + \beta_2 z + \epsilon \\ &= \beta_0 + \beta_1 x + \beta_2 (a + cx) + \epsilon \\ &= \underbrace{(\beta_0 + a\beta_2)}_{\tilde{\beta}_0} + \underbrace{(\beta_1 + c\beta_2)}_{\tilde{\beta}_1} x + \epsilon \end{aligned}$$

This seems to suggest that the variable z is **redundant** in the model.

Model fitting

The model to be fitted is the following:

$$\begin{aligned}\texttt{poverty} = \beta_0 + \beta_1 \cdot \texttt{hs\_grad} + \beta_2 \cdot \texttt{home\_own} \\ + \beta_3 \cdot \texttt{median\_income}/1000 + \beta_4 \cdot \texttt{party\_maj} + \epsilon \\ \epsilon \sim \mathrm{N}(0, \sigma^2)\end{aligned}$$

Notes:

- Decided to scale the `median_income` variable by dividing by 1000.
- `party_maj` is a dummy variable. It equals 1 if 'Republican', and 0 if 'Democrat'. In this case, the label 'Democrat' is called the *baseline*.

Model fitting (cont.)

```
mod <- lm(formula = poverty ~ hs_grad + home_own + I(median_income/1000) +
            party_maj, data = poverty)
summary(mod)
```

```
## Call:
## lm(formula = poverty ~ hs_grad + home_own + I(median_income/1000) +
##     party_maj, data = poverty)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.8794 -0.8104 -0.0762  0.8412  3.2827
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)           76.07336    4.58210  16.602  < 2e-16 ***
## hs_grad               -0.45345    0.05599  -8.098 2.12e-10 ***
## home_own              -0.14614    0.03538  -4.130 0.000151 ***
## I(median_income/1000) -0.26101    0.03427  -7.616 1.10e-09 ***
## party_majRep          -0.51100    0.51347  -0.995 0.324857
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.354 on 46 degrees of freedom
## Multiple R-squared:  0.8859, Adjusted R-squared:  0.876
## F-statistic: 89.28 on 4 and 46 DF,  p-value: < 2.2e-16
```

Model fitting (cont.)

```r
fitted(mod)  # obtain fitted values
```

```
##         1         2         3         4         5         6         7         8
## 20.373513 11.565787 17.210837 20.261402 15.012070 12.477839  7.194859 11.576235
##         9        10        11        12        13        14        15        16
## 15.217284 15.819316 19.729882 13.879387 14.791221 13.493640 13.587406 11.915343
##        17        18        19        20        21        22        23        24
## 13.474400 20.683055 19.524389 13.240073  8.420077  9.104415 12.953813 11.231420
##        25        26        27        28        29        30        31        32
## 21.637436 17.040754 14.251766 12.859763 13.400985  9.679679  7.710733 17.805235
##        33        34        35        36        37        38        39        40
## 12.902249 17.969372 14.065428 13.752141 16.863705 14.909914 13.171882  9.942183
##        41        42        43        44        45        46        47        48
## 19.186426 15.105553 19.505661 18.487914 10.424302 11.556045 15.513875 14.018749
##        49        50        51
## 18.759122 11.743215  9.996266
```

Model fitting (cont.)

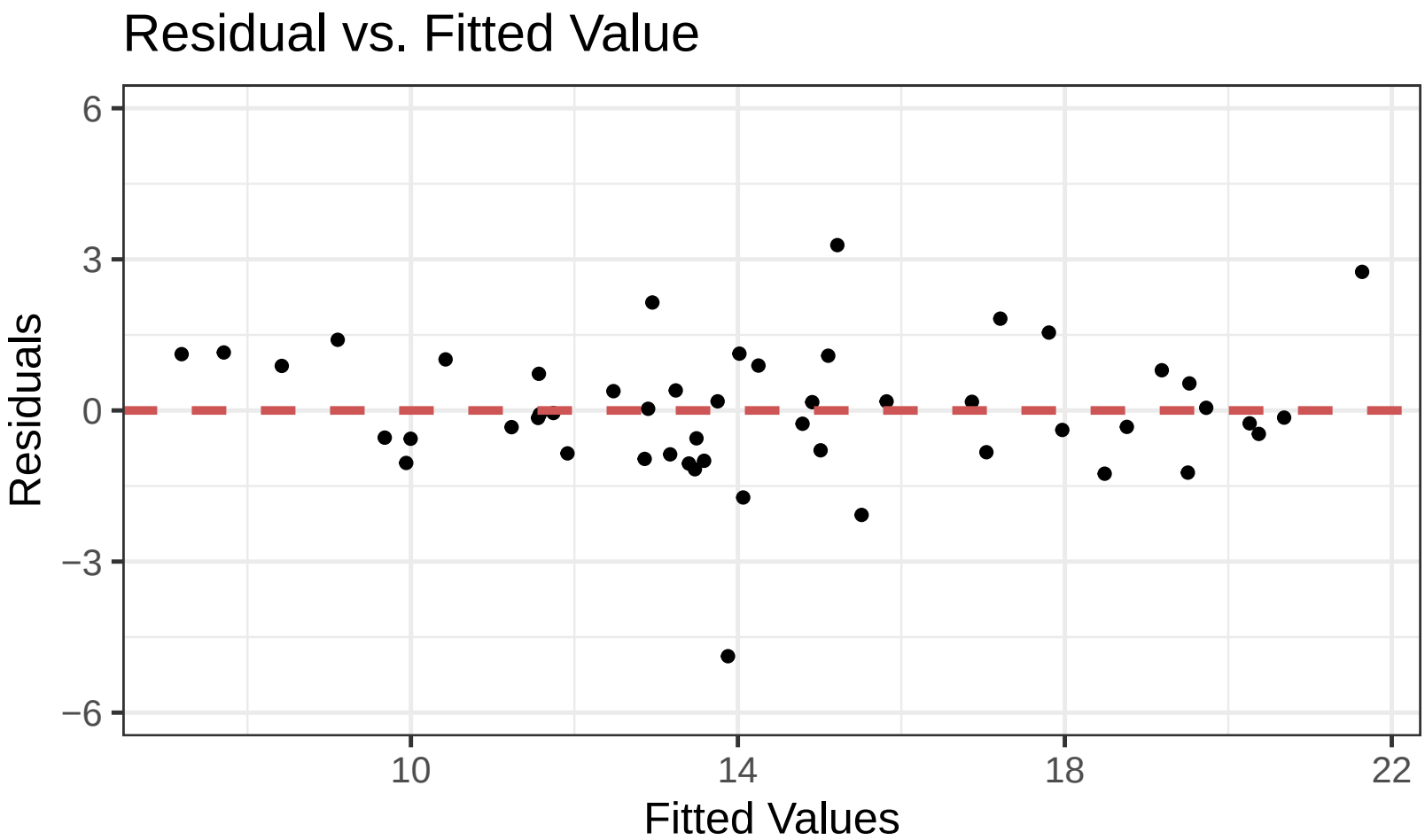
```r
resid(mod)  # obtain residuals
```

```
##           1           2           3           4           5           6
## -0.46605065  0.72731679  1.82249609 -0.25606877 -0.79138008  0.38309895
##           7           8           9          10          11          12
##  1.11764146 -0.07623466  3.28271639  0.17919156  0.05250815 -4.87938690
##          13          14          15          16          17          18
## -0.26394784 -0.55344392 -0.99936285 -0.85271642 -1.16773370 -0.14222118
##          19          20          21          22          23          24
##  0.53654811  0.39742676  0.88408973  1.40272776  2.14377701 -0.33141953
##          25          26          27          28          29          30
##  2.75158873 -0.82944933  0.89109102 -0.96191335 -1.05392611 -0.53967877
##          31          32          33          34          35          36
##  1.15117164  1.54627966  0.03323485 -0.38737151 -1.72769199  0.18081322
##          37          38          39          40          41          42
##  0.17136036  0.16508568 -0.87188216 -1.04218336  0.79835675  1.08687105
##          43          44          45          46          47          48
## -1.23408238 -1.25563045  1.01362932 -0.14890176 -2.07432310  1.12740532
##          49          50          51
## -0.32639451 -0.05154804 -0.56148303
```

Model diagnostics

Ideally, a random noise pattern should be seen.

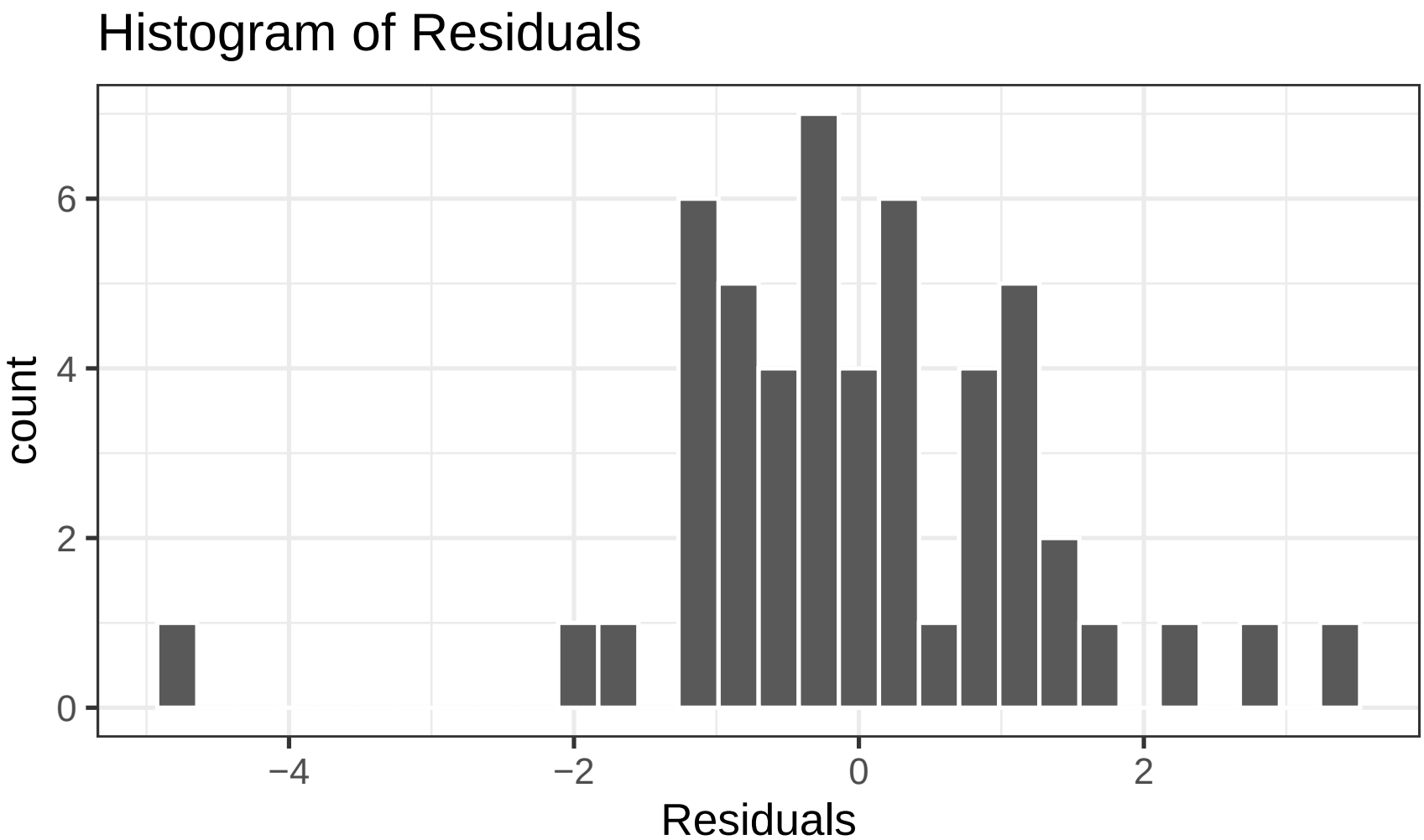
```
diag.plots <- lindia::gg_diagnose(mod, theme = theme_bw(), plot.all = FALSE)
diag.plots$res_fitted
```



Model diagnostics (cont.)

Check the distribution of the residuals. Does it look like a bell curve?

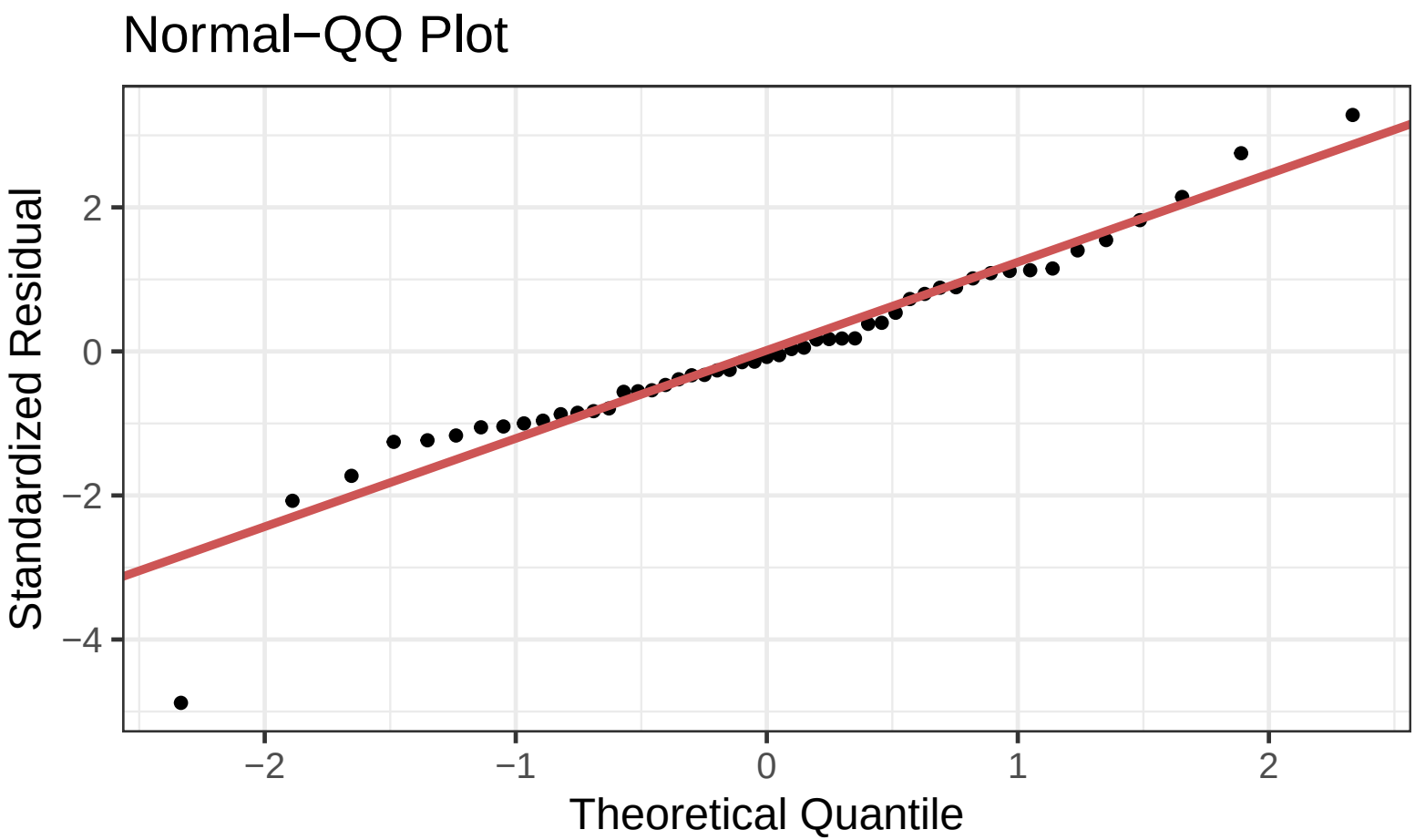
```
diag.plots$residual_hist
```



Model diagnostics (cont.)

A QQ-plot determines if the residuals follow a normal distribution. Everything is fine, except for one outlier on the left hand side of the plot. May want to follow up with this data point–any errors (measurement, collection, sampling, etc.)?

```
diag.plots$qqplot
```



Model diagnostics (cont.)

Remarks on model diagnostics:

- Normality assumption is satisfied satisfactorily (look at histogram, QQ-plot).
- Constant variance assumption is satisfied (look at residual plot pattern).
- $R^2 = 0.88$ indicates a strong linear association of the predictors on the response variable.

Manual calculations

```
X <- model.matrix(mod); head(X)  # n by (p + 1) matrix
```

```
##   (Intercept)  hs_grad home_own I(median_income/1000) party_majRep
## 1           1 76.78209 73.38657                36.96294           1
## 2           1 86.59310 63.22759                59.35103           1
## 3           1 81.82000 69.65333                42.41793           1
## 4           1 78.86400 71.23067                34.98271           1
## 5           1 82.45862 63.26724                55.26572           0
## 6           1 88.13906 72.14531                50.13584           0
```

```
y <- poverty$poverty; head(y)  # n by 1 vector
```

```
## [1] 19.90746 12.29310 19.03333 20.00533 14.22069 12.86094
```

Manual calculations (cont.)

```
XtX <- t(X) %*% X  # (p + 1) by (p + 1) matrix
colnames(XtX) <- rownames(XtX) <- NULL
XtX
```

```
##           [,1]       [,2]       [,3]       [,4]     [,5]
## [1,]   51.000   4323.498   3657.123   2430.910   30.000
## [2,] 4323.498 367497.623 309942.167 207273.903 2498.765
## [3,] 3657.123 309942.167 264098.108 173256.517 2202.333
## [4,] 2430.910 207273.903 173256.517 119871.817 1284.466
## [5,]   30.000   2498.765   2202.333   1284.466   30.000
```

```
Xty <- t(X) %*% y; head(Xty)  # (p + 1) by 1 vector
```

```
##                                [,1]
## (Intercept)                734.9980
## hs_grad                  61590.5786
## home_own                 52725.4769
## I(median_income/1000)    33676.2834
## party_majRep               476.7111
```

Manual calculations (cont.)

```r
as.numeric(beta <- solve(XtX, Xty))  # regression coefficients
```

```
## [1] 76.0733587 -0.4534462 -0.1461374 -0.2610123 -0.5109967
```

```r
y.hat <- as.numeric(X %*% beta.hat); head(y.hat)  # fitted values
```

```
## [1] 20.37351 11.56579 17.21084 20.26140 15.01207 12.47784
```

```r
eps.hat <- y - y.hat; head(eps.hat)  # residuals
```

```
## [1] -0.4660507  0.7273168  1.8224961 -0.2560688 -0.7913801  0.3830990
```

```r
(sigma.hat <- sqrt(sum(eps.hat ^ 2) / (51 - 4 - 1)))  # residual SE
```

```
## [1] 1.353753
```

Manual calculations (cont.)

```
(var.beta.hat <- sigma.hat ^ 2 * solve(XtX))  # Estimate of Var(beta.hat)
```

```
##               [,1]          [,2]          [,3]          [,4]         [,5]
## [1,] 20.995620883 -0.1849707465 -0.0674556619 -0.0029825145 -0.509316515
## [2,] -0.184970747  0.0031353530 -0.0004649709 -0.0010121404  0.001289793
## [3,] -0.067455662 -0.0004649709  0.0012520320  0.0003869122 -0.002294815
## [4,] -0.002982514 -0.0010121404  0.0003869122  0.0011746260  0.008589991
## [5,] -0.509316515  0.0012897933 -0.0022948145  0.0085899906  0.263655017
```

```
(se.beta.hat <- sqrt(diag(var.beta.hat)))  # SE beta.hat
```

```
## [1] 4.58209787 0.05599422 0.03538406 0.03427282 0.51347348
```

```
as.numeric(beta.hat / se.beta.hat)  # test statistic value
```

```
## [1] 16.6022990 -8.0980881 -4.1300351 -7.6157228 -0.9951764
```

Manual calculations (cont.)

```
(total.SS <- sum((y - mean(y)) ^ 2))  # Total SS
```

```
## [1] 738.7815
```

```
(resid.SS <- sum(eps.hat ^ 2))  # Resid SS
```

```
## [1] 84.30172
```

```
(reg.SS <- total.SS - resid.SS)  # Reg SS
```

```
## [1] 654.4798
```

```
(reg.SS / 4) / (resid.SS / (51 - 4 - 1))  # F-statistic
```

```
## [1] 89.28071
```

```
1 - resid.SS / total.SS  # R^2 value
```

```
## [1] 0.8858909
```

Inference and interpretation

Are poverty rates in US states dependent on the party that is in majority? We would like to test

$$H_0 : \beta_4 = 0 \quad \text{v.s.} \quad H_1 : \beta_4 \neq 0$$

We can conduct a t-test: Under the null hypothesis, the test statistic

$$T = \frac{\hat{\beta}_4 - 0}{\text{SE}(\hat{\beta}_4)} \sim t_{51-1}.$$

The observed value for T is $T = -0.995$, and the p-value associated with this is $p = 0.325$. Thus we do not have sufficient evidence to reject H_0.

Inference and interpretation (cont.)

Out of all the coefficients, β_4 is the only one which is not significant (look at the summary output). We can then proceed to make interpretations such as:

- The expected poverty rate in a state with zero high school graduates, home ownership and median income is

$$\beta_0 + \beta_1 \cdot 0 + \beta_2 \cdot 0 + \beta_3 \cdot 0 = \beta_0$$

  which is estimated by $\hat{\beta}_0 = 76.1\%$. Admittedly not very informative–no such state exists!

- The average percentages of high school graduates and home ownership, and the average annual median income are

$$\bar{x}_1 = 84.8, \quad \bar{x}_2 = 71.7 \quad \bar{x}_3 = 47.7$$

  The expected poverty rate in an 'average' state is therefore estimated to be

$$\hat{\beta}_0 + \hat{\beta}_1 \cdot \bar{x}_1 + \hat{\beta}_2 \cdot \bar{x}_3 + \hat{\beta}_3 \cdot \bar{x}_3 = 14.7\% \in [11.9\%, 17.5\%]^1$$

[1] See remark at the end of the slides

Inference and interpretation (cont.)

- By how much does poverty rate change given different percentages of high school graduates in a state? Let

$$\mathrm{E}(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3.$$

Keeping all other variables constant in value, the expected outcome of an increase in high school graduates by one percentage point is

$$\mathrm{E}(y') = \beta_0 + \beta_1 (x_1 + 1) + \beta_2 x_2 + \beta_3 x_3.$$

Taking differences between $\mathrm{E}(y')$ and $\mathrm{E}(y)$, we get

$$\begin{aligned} \mathrm{E}(y') - \mathrm{E}(y) &= \cancel{\beta_0} + \beta_1 (x_1 + 1) + \cancel{\beta_2 x_2} + \cancel{\beta_3 x_3} \\ &\quad - \left[\cancel{\beta_0} + \beta_1 x_1 + \cancel{\beta_2 x_2} + \cancel{\beta_3 x_3}\right] \\ &= \cancel{\beta_1 x_1} + \beta_1 - \cancel{\beta_1 x_1} = \beta_1 \end{aligned}$$

Therefore, it is estimated that poverty rates decline by $|\hat{\beta}_1| = 0.45$ percentage points for every 1% increase in high school graduates, with everything else remaining constant.

Inference and interpretation (cont.)

- In a similar manner we can say that
  - The expected poverty rates decline by $|\hat{\beta}_2| = 0.14$ percentage points for every 1% increase in home ownership in a state;
  - The expected poverty rates decline by $|\hat{\beta}_2| = 0.26$ percentage points for every US$1,000 increase in annual median income;

  keeping all other variables constant. Of course, keep in mind the **sign** of the coefficients and ensure they make sense. In our case, it makes sense that improvements in high school graduates, home ownership rates and median income would see a decrease in poverty rates (negative coefficients).
- A test of $H_0 : \beta_1 = \beta_2 = \beta_3 = \beta_4 = 0$ is rejected at the 0.01 level ($F = 89.26$ on 4 and 46 d.f., $p < 2.2e^{-16}$), and we conclude that not all regression coefficients are zero.

Inference and interpretation (cont.)

Remark

The expected poverty rate in an 'average' state is actually a **prediction**–since we never observe a state with values exactly equal to the average. Under the normal model, we expect that $y_{new} = \mathbf{x}_{new}^\top \boldsymbol{\beta} + \epsilon$ to have distribution

$$y_{new} \sim \mathrm{N}(\mathbf{x}_{new}^\top \boldsymbol{\beta}, \sigma^2)$$

The predicted value for y_{new} is $\hat{y}_{new} = \mathbf{x}_{new}^\top \hat{\boldsymbol{\beta}} + \epsilon$. This has distribution given by

$$\hat{y}_{new} \sim \mathrm{N}\left(\mathbf{x}_{new}^\top \boldsymbol{\beta}, \sigma^2(\mathbf{x}_{new}^\top (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{x}_{new} + 1)\right).$$

Using the estimated value $\hat{\sigma}$, the distribution of $\hat{y}_{new}$ follows a t-distribution

$$\frac{\hat{y}_{new} - \mathbf{x}_{new}^\top \boldsymbol{\beta}}{\hat{\sigma}\sqrt{\mathbf{x}_{new}^\top (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{x}_{new} + 1}} \sim t_{n-p},$$

for which a 95% confidence interval can be constructed.

Remark on predictive intervals

Remark

Even though technically y_{new} is random, a point estimate for it is its mean value which is $\mathrm{E}(y_{new}) = x_{new}^\top \boldsymbol{\beta}$. This in turn is estimated by $x_{new}^\top \hat{\boldsymbol{\beta}}$.

There is a subtle difference between the distribution of the estimate for $\mathrm{E}(y_{new})$ and y_{new} itself. As for $\widehat{\mathrm{E}(y_{new})} = x_{new}^\top \hat{\boldsymbol{\beta}}$, the only variability comes from the estimator, thus

$$\widehat{\mathrm{E}(y_{new})} \sim \mathrm{N}\left(\mathbf{x}_{new}^\top \boldsymbol{\beta}, \sigma^2 \mathbf{x}_{new}^\top (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{x}_{new}\right)$$

One is also able to show that

$$\frac{\widehat{\mathrm{E}(y_{new})} - \mathbf{x}_{new}^\top \boldsymbol{\beta}}{\hat{\sigma}\sqrt{\mathbf{x}_{new}^\top (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{x}_{new}}} \sim t_{n-p},$$

Remark on predictive intervals (cont.)

Remark

- The interval based on the first distribution is called the **prediction interval** or the *wide interval*.
- The interval based on the second distribution is called the **confidence interval** or the *narrow interval*.
- Predicting a new point has more uncertainty hence a wider interval. Predicting an *average* of a new point has less uncertainty, hence a narrower interval.
- Both intervals are centred around the point estimate $x_{new}^{\top}\hat{\boldsymbol{\beta}}$ and has the same distribution. The only difference is the standard error.
- We could turn to Wald-based intervals instead of the exact t-based intervals, and in fact for large n this would be approximately the same.

Remark on predictive intervals (cont.)

```
newx <- data.frame(
  hs_grad = tmp[2],
  home_own = tmp[3],
  median_income = tmp[4] * 1000,
  party_maj = "Dem"
)
predict(mod, newx, interval = "confidence", level = 0.95)   # narrow
```

```
##              fit      lwr      upr
## hs_grad 14.71231 13.99451 15.43011
```

```
predict(mod, newx, interval = "prediction", level = 0.95)   # wider
```

```
##              fit      lwr      upr
## hs_grad 14.71231 11.89439 17.53023
```

References I

Casella, G. and R. L. Berger (2002). *Statistical Inference*. 2nd ed. Pacific Grove, CA: Duxbury. ISBN: 978-0-534-24312-8.

Faraway, J. J. (2014). *Linear models with R*. CRC press.

Pawitan, Y. (2001). *In All Likelihood*. Statistical Modelling and Inference Using Likelihood. Oxford University Press. ISBN: 978-0-19-850765-9.

Jamil, H. (Oct. 2018). "Regression modelling using priors depending on Fisher information covariance kernels (I-priors)". PhD thesis. London School of Economics and Political Science.

Wasserman, L. (2013). *All of statistics: a concise course in statistical inference*. Springer Science & Business Media.