UCL Doctoral School: Research Methods for Quantitative Data

Linear Regression

Dr Peter Martin, Department of Applied Health Research

peter.martin@ucl.ac.uk

April 2018

Overview

- Types of relationships
- Simple linear regression
- Multiple linear regression
- Regression diagnostics and further topics

Learning Objectives

By the end of this session you should be able to:

- Describe associations between numerical variables, using graphs and/or statistics
- Understand the principles of ordinary least squares regression, and how to use it to model linear bivariate and multivariate relationships
- Understand results from regression models reported in academic research publications
- Estimate a linear regression model and interpret the results (using SPSS, STATA, or R)
- Understand the importance of model assumptions and assess the fit of a linear regression model
- Compare nested models with one another using hypothesis tests and other statistics

A note about these slides

These slides contain more information than can be covered in a one-hour lecture, and more than students are expected to fully understand at the end of the lecture. For example, students are not expected to follow all calculations and appreciate the details of all equations given in these slides.

The additional material and technical details that these slides provide are intended as a guide towards further study.

Numerical and categorical variables

Numerical variables are variables whose values are numbers that have an interpretable meaning as numbers.

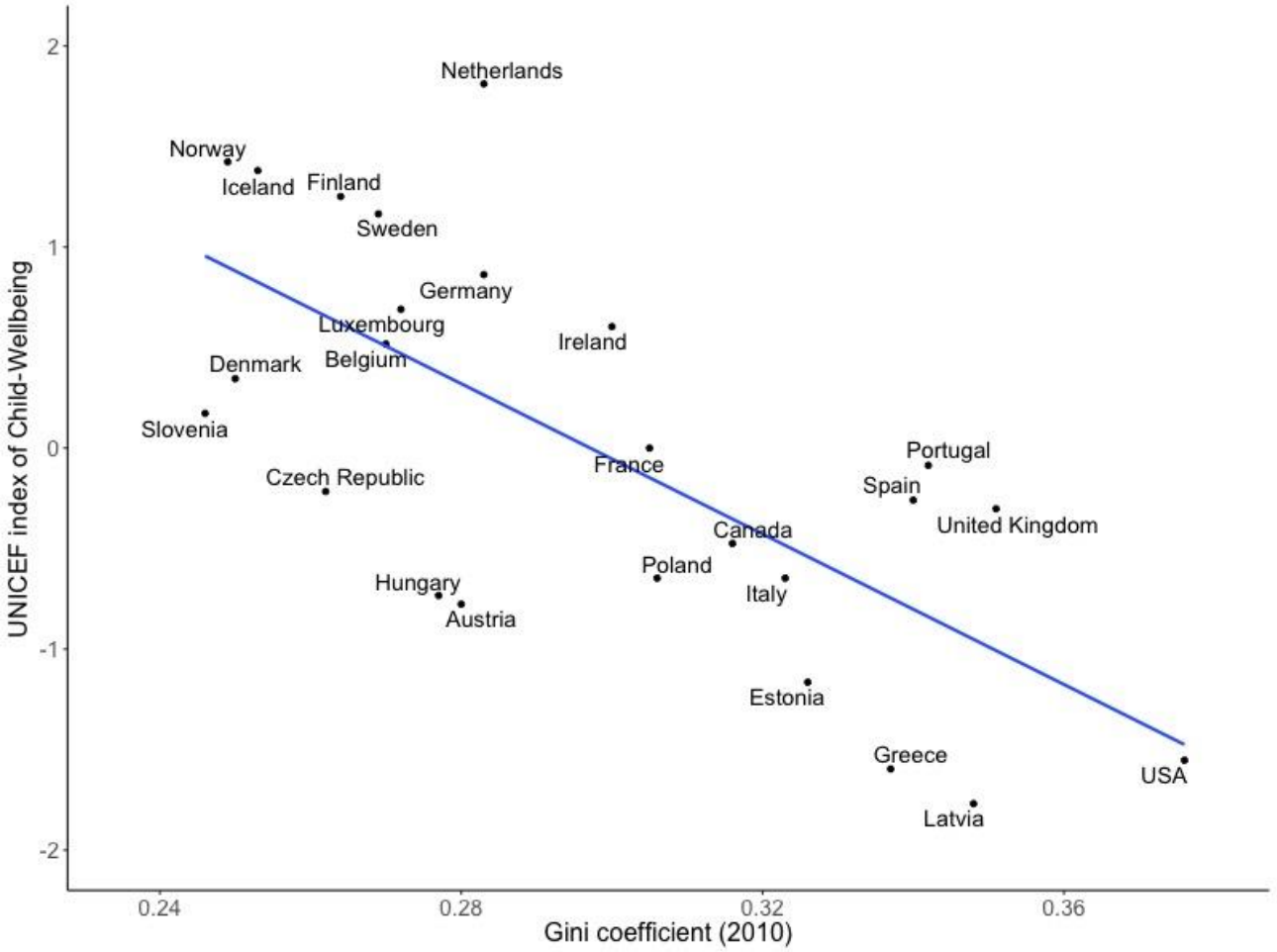
Examples of numerical variables:

- Age
- Height
- GDP
- Gini coefficient
- Number of births per year

Examples of categorical (non-numeric) variables:

- Highest qualification (“Below A-level”, “A-level”, “Degree or equivalent”)
- Gender (“male”, “female”, “non-binary”)

A linear association: child wellbeing and economic inequality (2014)

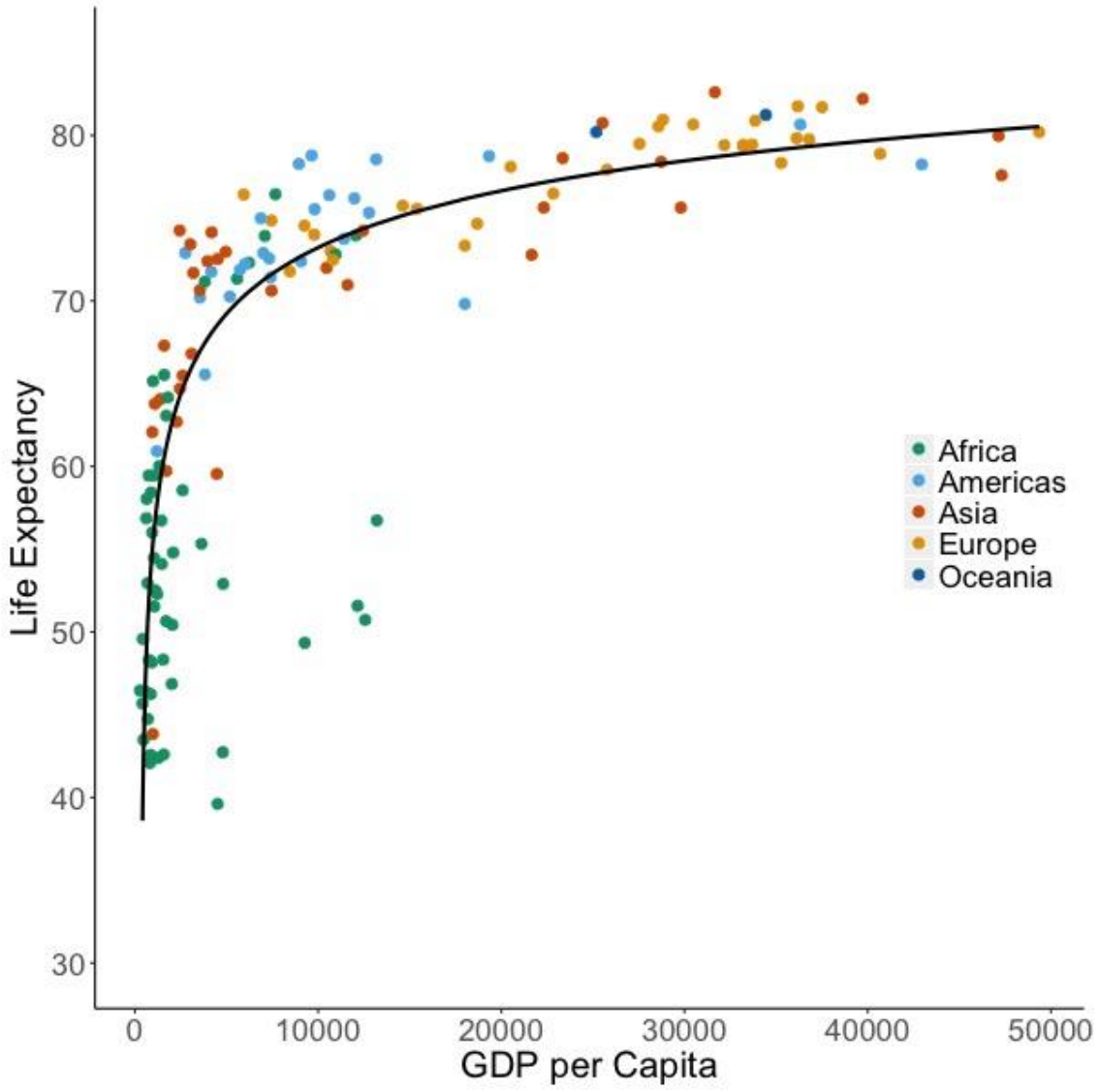


Pearson's r = —0.70

Gini coefficient: higher coefficient indicates more income inequality. **UNICEF index of child wellbeing**: higher number indicates better child wellbeing averaged over four dimensions: health, education, housing and environment, and behaviours).

Sources: Gini coefficient: OECD (http://www.oecd.org/social/income-distribution-database.htm). Child-wellbeing: Martorano et al (2014, Table 15). The figure was inspired Figure 1 in Pickett & Wilkinson (2007), but is based on more recent data.

A non-linear association GDP per capita and life expectancy (2007)



Source: Data from the Gapminder Foundation (Bryan, 2015). See www.gapminder.org .

Descriptive statistics: correlation coefficient

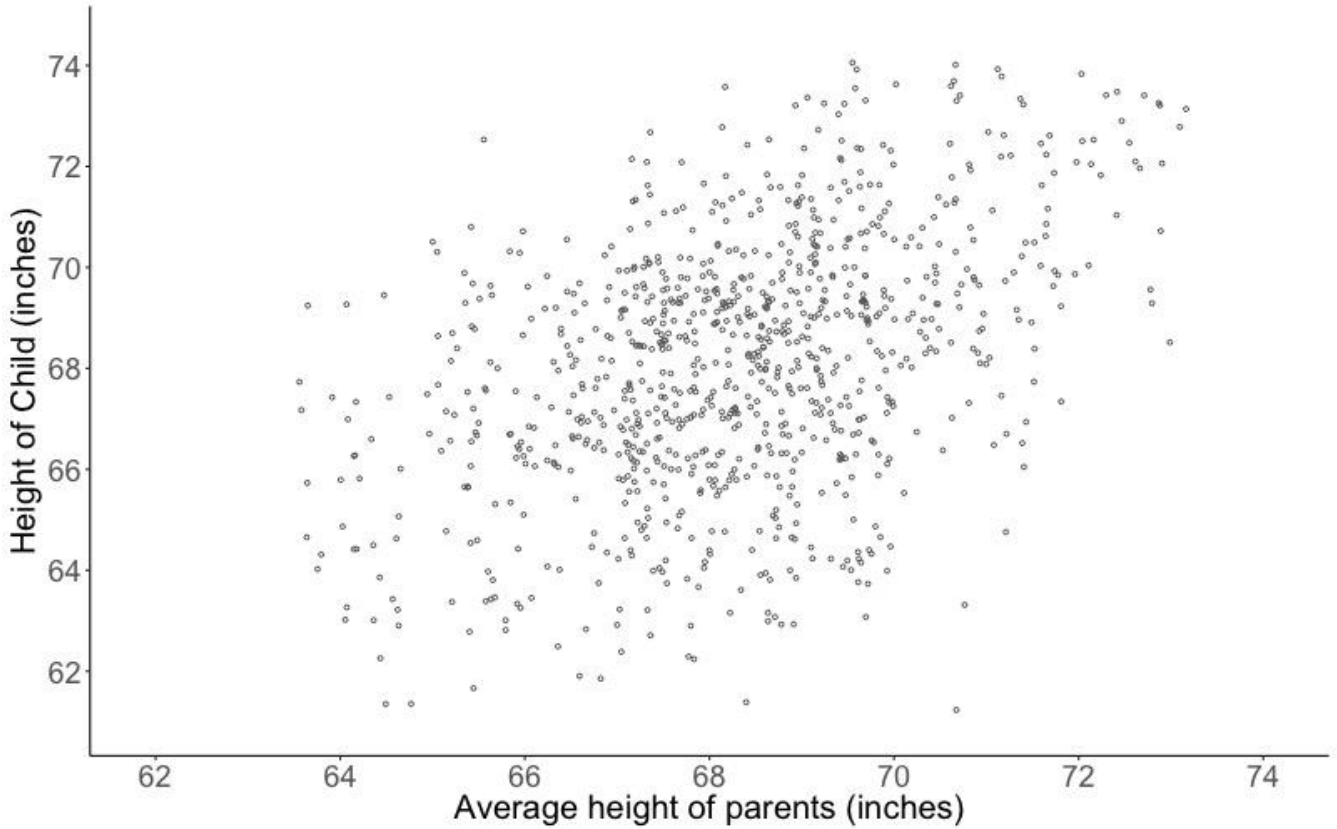
The strength and direction of a linear association between two numerical variables can be described by Pearson's correlation coefficient ("Pearson's r"). For example, the observed correlation between Child Wellbeing and Gini coefficient is r = - 0.70.

Mathematically speaking, *Pearson's r* can be calculated for non-linear relationships also, but it will fail to adequately describe that relationship. Calculating Pearson's r for a non-linear association is missing the point.

As we shall see, the same goes for linear regression.

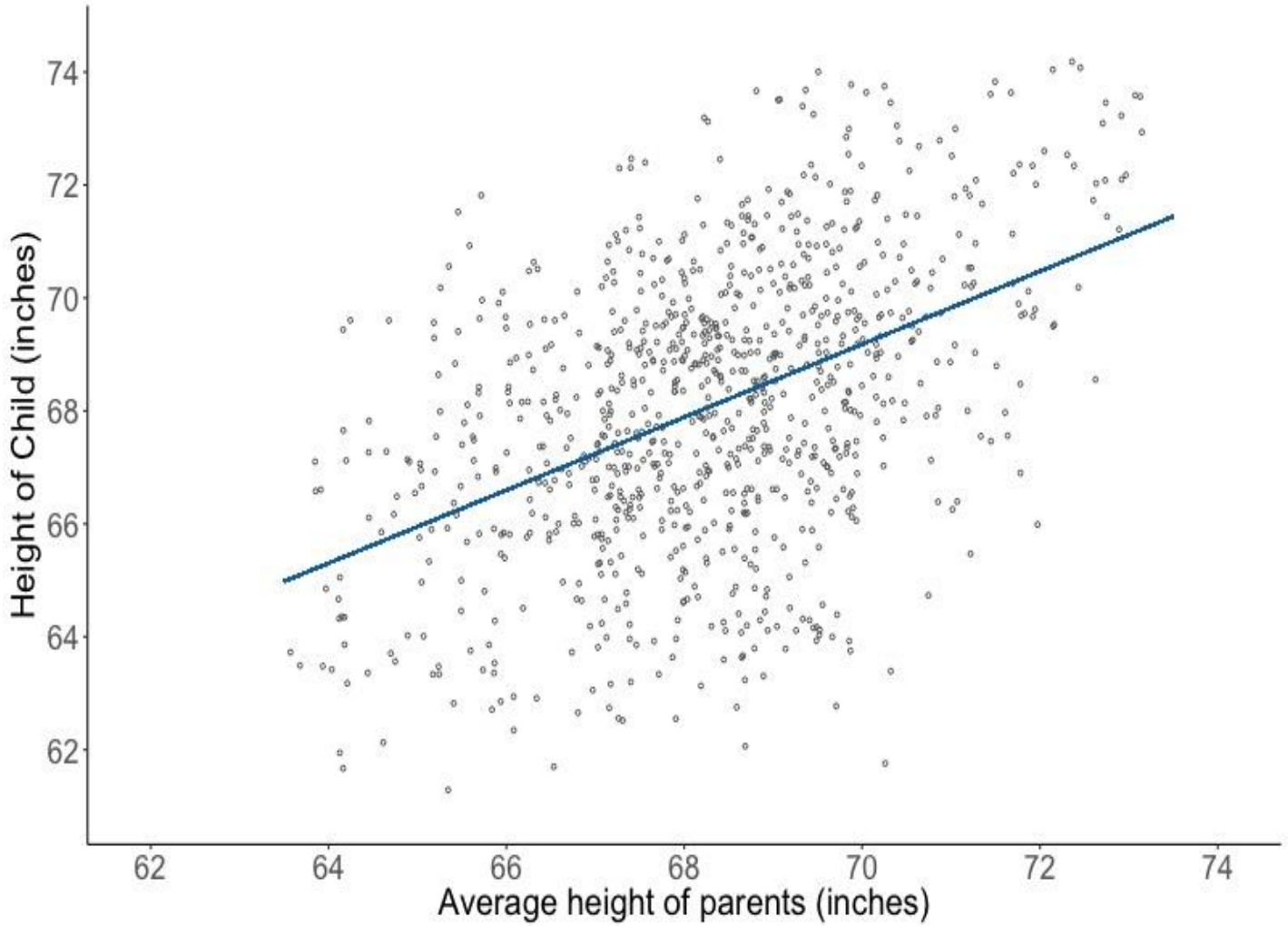
Simple Linear Regression:

Francis Galton's analysis of the heights of parents and their children

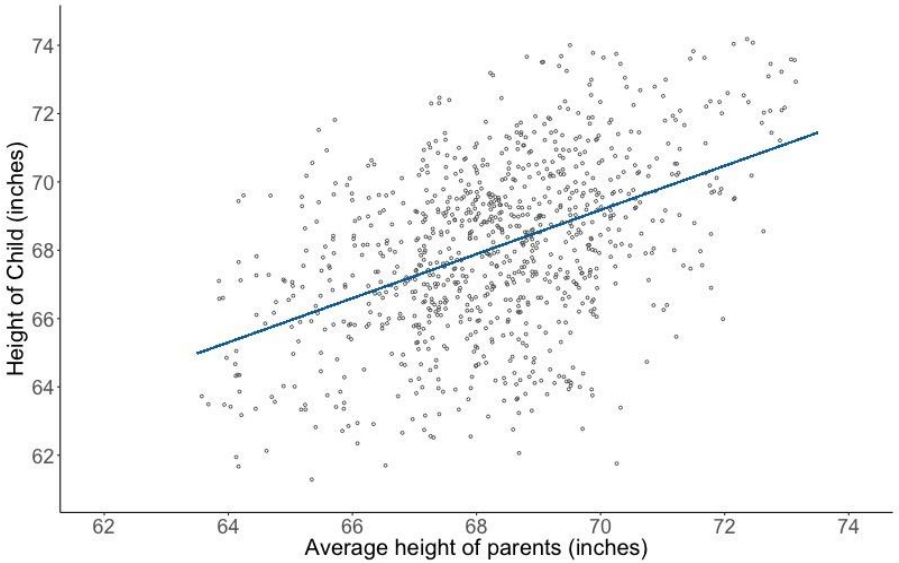


Notes: Data are taken from Galton (1886) via the "psych" package for R (Revelle, 2015)*. Original data have been jittered to make the scatterplot more easily interpretable. Regression estimates shown here therefore differ slightly from Galton's historical result, which were based on discretized data. Parents' height is the average of the mother's and father's height. Female heights (both mothers and daughters) were adjusted to make the comparable to male heights.*

Galton's data with fitted regression line



Galton's data with fitted regression line



Simple linear regression is a statistical model that aims to represent the relationship between two variables, *Y* and *X*, by a straight line. This is called the 'line of best fit' or 'line of best prediction'. We use the variable *X* to predict *Y*.

- *X* is called the **independent variable**, **predictor**, or **exposure**.
- *Y* is called the **dependent variable**, **outcome**, or **response**.

Equation of a regression line

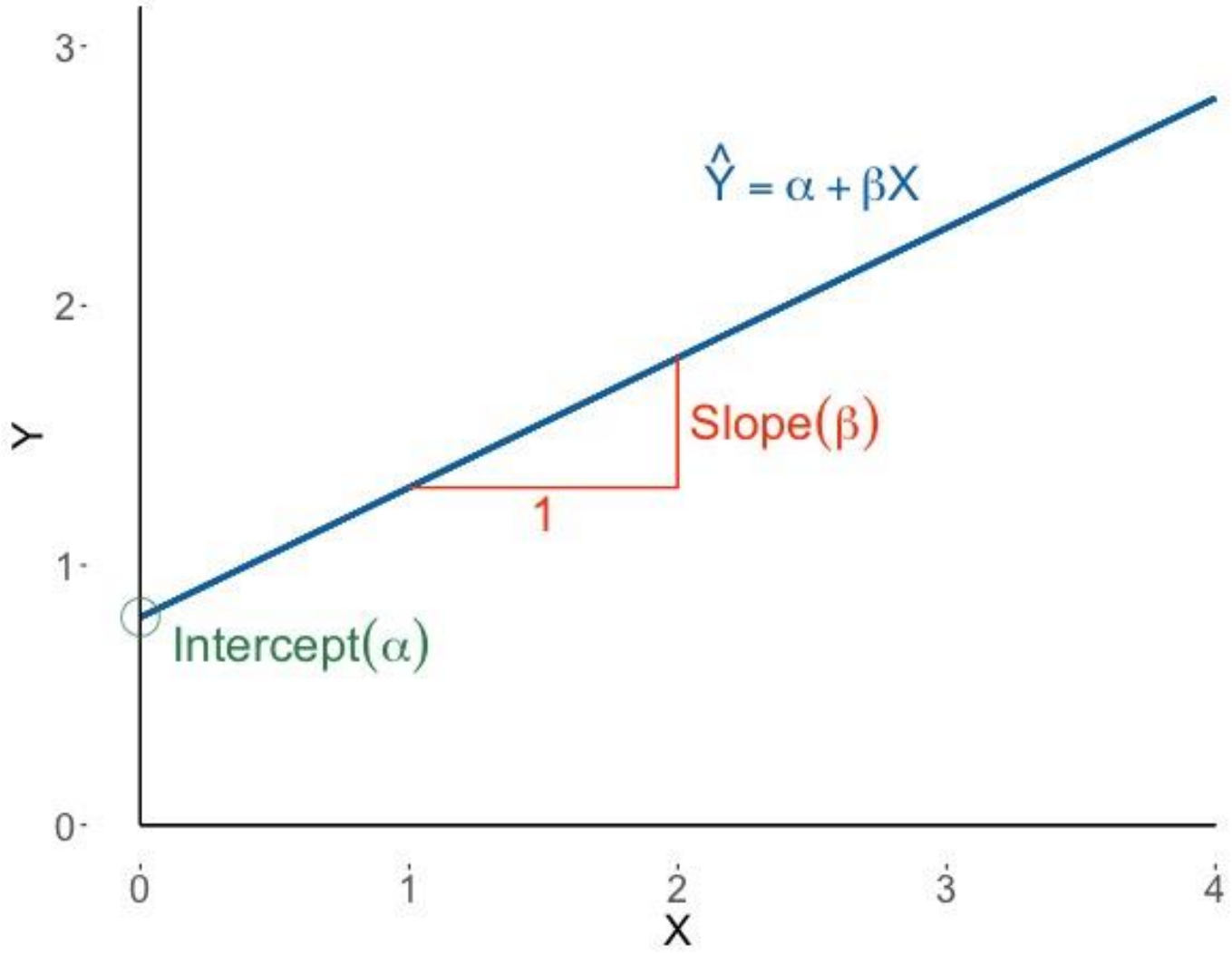
The mathematical representation of a straight line is a linear equation of the form:

$$\hat{Y} = \alpha + \beta X ,$$

where:

- $\hat{Y}$ is the predicted value of the outcome *Y* – in our case, the predicted height of the child (read $\hat{Y}$ as "Y-hat"; the hat indicates a predicted value of *Y*, in contrast to the actual value of a particular person).
- X is the value of the predictor – here, the parents' height
- α is the **intercept** of the regression line (the value of $\hat{Y}$ when X = 0)
- β is the **slope** of the regression line – this is the predicted difference in *Y* for a one-unit difference in X (in our case: the predicted height difference between two children whose parents' heights differ by one inch).

An illustration of the regression line, its intercept and slope



Galton's regression line

The equation of Galton's regression line shown on slide 10 is:

$$\hat{Y} = 23.942 + 0.646X$$

If it helps, you may write this equation as:

$$Predicted\ Child'sHeight = 23.942 + 0.646 \times Parents'Height$$

Interpreting regression coefficients

The intercept (α) and the slope (β) of the regression equation are referred to as the **coefficients**.

The **intercept** is the predicted value of *Y* at the point when *X* is zero. In our example, the intercept is equal to 23.942. Formally, this means that the predicted height of a person whose parents have zero height is 23.942 inches. As a prediction, this obviously does not make sense, because parents of zero height don't exist. The intercept is of scientific interest only when *X* = 0 is a meaningful data point.

The **slope** determines by how much the line rises in the Y-direction for a one-unit step in the X-direction. In our example, the slope is equal to 0.646. This means that a one inch difference in parents' height is associated with a 0.646-inch difference in the height of the children. If the Joneses are one inch taller than the Smiths, the Joneses' children are predicted to be taller than the Smith's children by 0.646 inches on average.

Deriving predictions from the regression line

We can use this equation to derive a predicted height for a child, if we are given the parents' height. For example, take a child whose parents' height is 64.5 inches. Plugging that number into the regression equation, we get:

$$\hat{Y} = 23.942 + 0.646 \times 64.5$$
$$= 65.6$$

A child of parents with height 64.5 inches is predicted to be 65.6 inches tall. The "hat" over Y indicates that this result is a prediction, not the actual height of the child. This is important because the prediction is not perfect: not every child is going to have exactly the height predicted by the regression equation. The aim of the regression equation is to be right *on average*, not necessarily for every individual case.

Regression as a statistical model

The linear regression model looks like this:

$$Y_i = \alpha + \beta X_i + \varepsilon_i \, ,$$

where

- Y_i is the Y value of the i^{th} case
- X_i is the X value of the i^{th} case
- α and β are the intercept and the slope, as before
- ε_i is the called the **error**: the difference between the observed Y and the predicted value $\hat{Y}$.

Errors of prediction

The errors are the difference between the observed values Y_i and the predicted values $\widehat{Y}_i$. To see this, rearrange the regression equation from the previous slide:

$$\begin{aligned} \varepsilon_i &= Y_i - (\alpha + \beta X_i) \\ &= Y_i - \widehat{Y}_i \end{aligned}$$

Taking account of errors is what distinguishes a statistical model from a mathematical one (where relationships are assumed to be without errors).

Errors versus residuals

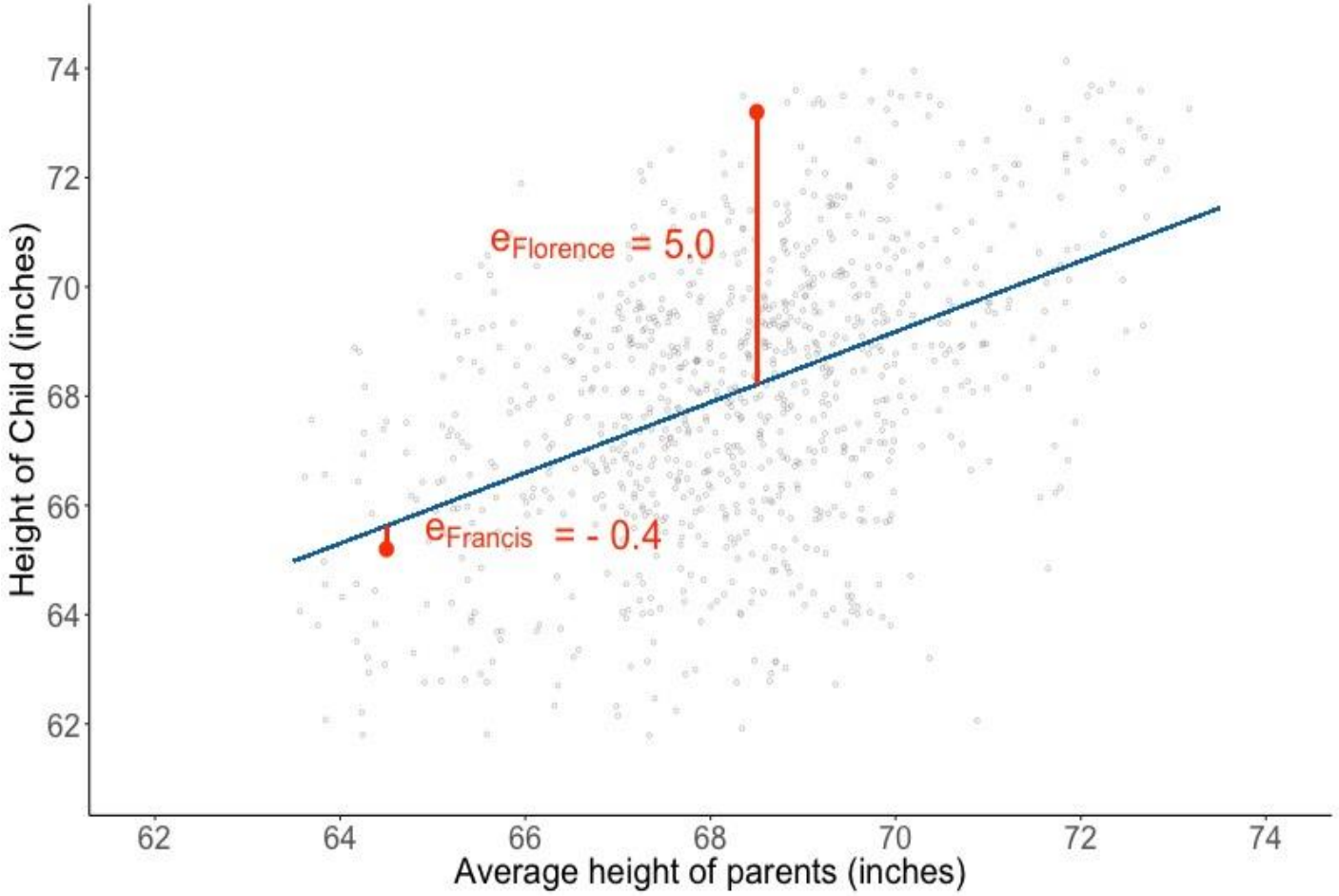
In practice, the errors are unobserved. (We don't know what they are.) This is because in general we do not know the true regression line. All we can do is estimate the regression line from a set of data.

The observed differences between the predicted values of Y and the actual values of Y in regression based on a particular sample are called the residuals. These can be seen as estimates of the errors:

$$e_i = Y_i - \hat{Y}_i$$

Two example residuals in Galton's data

Imagine two children, Francis and Florence…



Francis' parents are 64.5 inches tall. From this, the regression equation predicts Francis' height to be 65.6 inches. But Francis is only 65.2 inches tall. Between Francis' actual height and the prediction, there is a difference of 0.4 inches.

So Francis' residual is:

$$\begin{aligned} e_{Francis} &= Y_{Francis} - \hat{Y}_{Francis} \\ &= 65.2\ inches - 65.6\ inches \\ &= -0.4\ inches \end{aligned}$$

Francis' residual is a negative number, because Francis is shorter than predicted.

Now consider Florence. Her parents' height is 68.5 inches, and consequently her predicted height is 68.2 inches. But Florence is in fact 73.2 inches tall. Because she is taller than predicted, her residual is a positive number:

$$\begin{aligned} e_{Florence} &= Y_{Florence} - \hat{Y}_{Florence} \\ &= 73.2\ inches - 68.2\ inches \\ &= 5.0\ inches \end{aligned}$$

Estimating a regression line: the least squares procedure

The line of best fit is found by minimizing the sum of the squared residuals. That is, from all possible regression lines that could be drawn, the line of best fit is the one that has the smallest sum of squared residuals.

Because of this, linear regression is also called **ordinary least-squares** (OLS) regression.

It can be shown mathematically that for any given set of points, there is exactly one best regression line.

Typical regression results table

	Coefficient estimate	Std. Error	99 % Confidence Interval for the coefficient Lower bound	Upper bound	t-test of H_0: $\beta = 0$ t	df	p
Intercept	23.942	2.811					
Parents' Mean Height	0.646	0.041	0.540	0.752	15.711	926	.000

This is a typical output table for a simple linear regression analysis. We will look at formal definitions of confidence intervals and t-tests below.

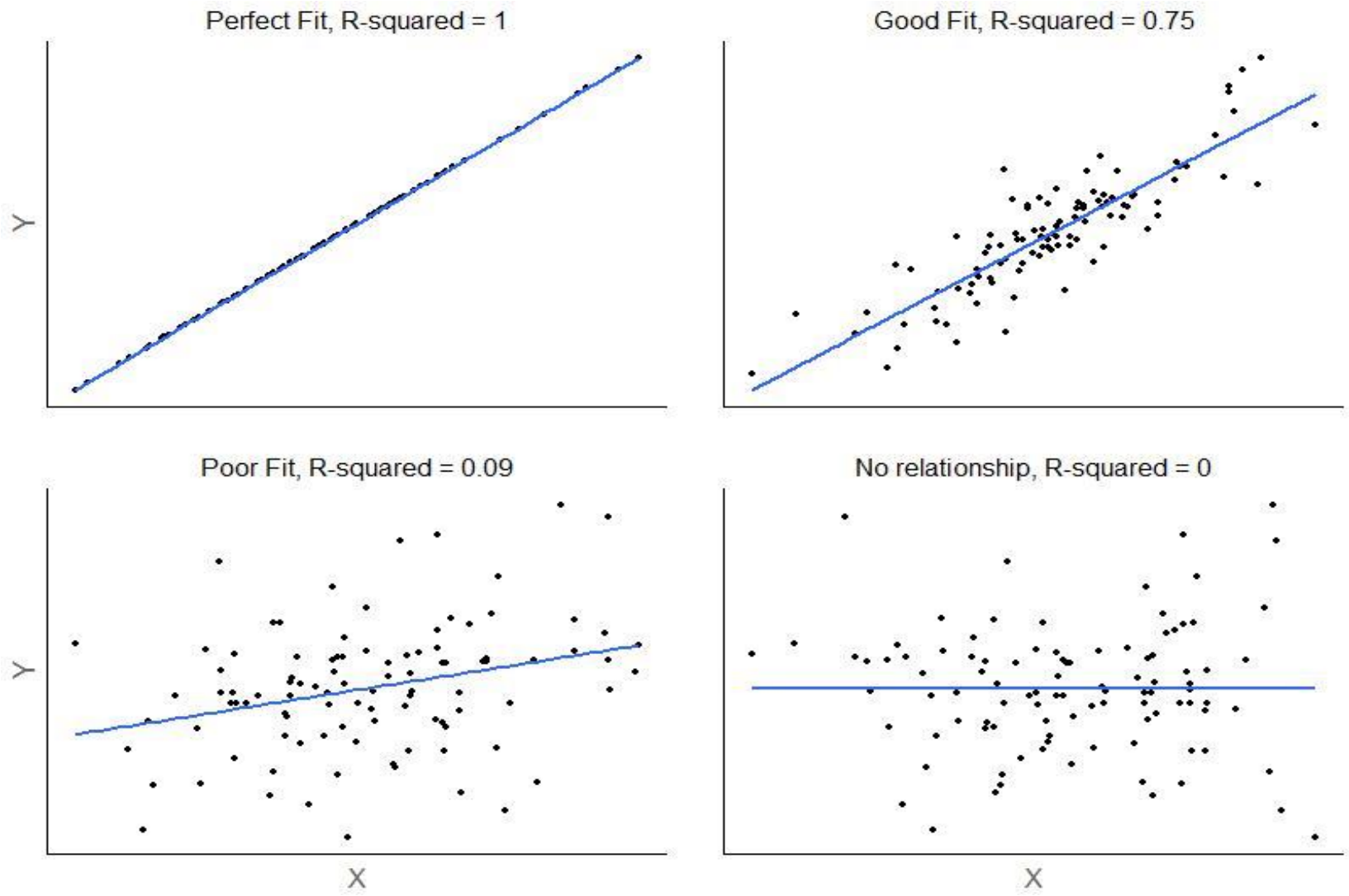
Assessment of Model Fit: the coefficient of determination (R^2)

We can measure how well X predicts Y. This is an aspect of model fit. In a well-fitting model, all predicted values are close to the observed values, and there is no non-linear pattern in the data.

The R^2 statistic (also called "coefficient of determination") is often used to describe model fit in a linear regression. This measures the proportional reduction in error that we achieve by using X to predict Y, compared to a situation where we had no information about X.

For example, the R^2 for Galton's regression is $R^2 = 0.21$. This is often interpreted to mean that X 'accounts for' 21 % of the variation in Y. In this case, we might say that 21 % of the variation in heights is 'accounted for' if we know the heights of people's parents. The remaining 79% are presumably due to other factors (including, for example, environmental influences during the growth period, measurement error when measuring the variables, etc.).

Note that in simple linear regression, R^2 is actually the square of Pearson's r.



R^2 is a number between 0 and 1.

Regression and Causality

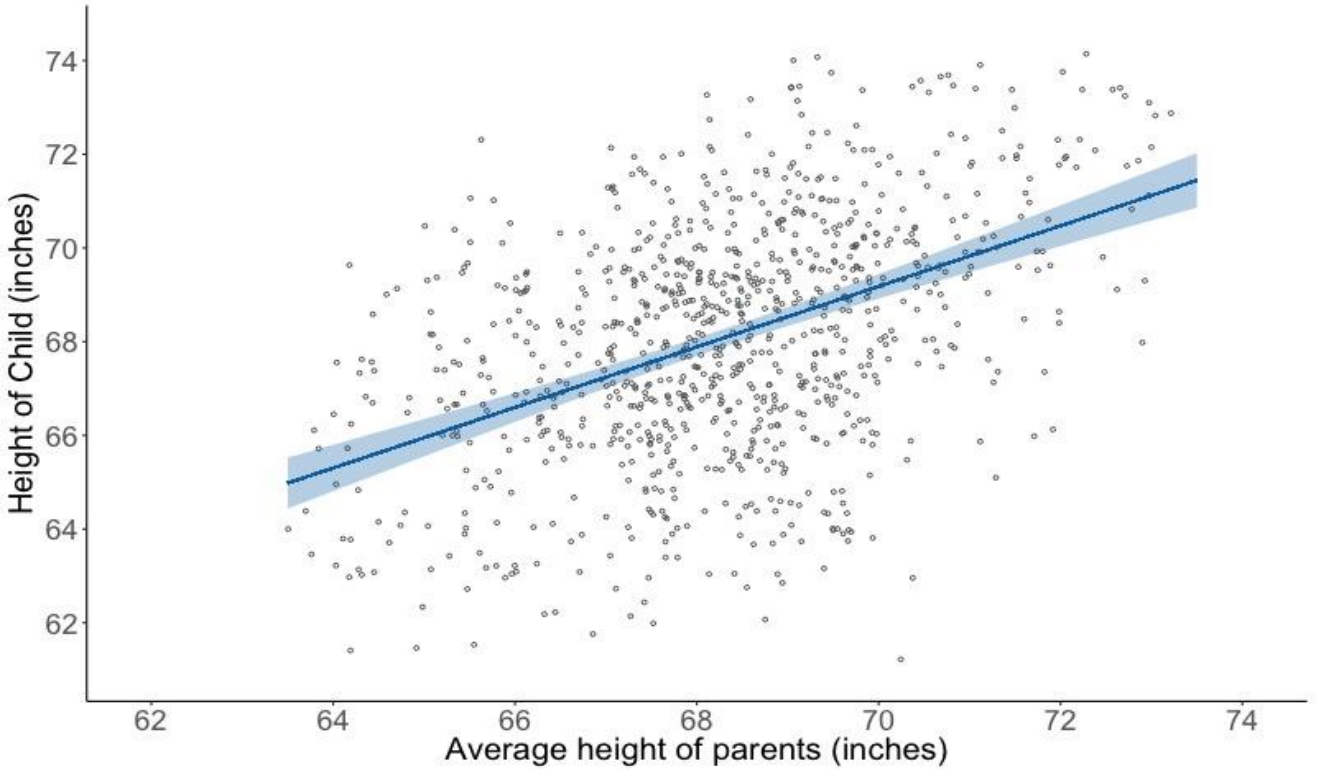
Linear regression is a directional model. We predict an outcome (*Y*) by a predictor (*X*). Estimates of coefficients will not in general be the same if we instead predict *X* by *Y*.

Linear regression cannot prove causality, however. Our ability to predict *Y* from *X* does not imply that *X* causes *Y*. However, if we have theoretical grounds to believe that *X* plays a part in causing *Y*, then the observation that we can predict *Y* by *X* to some extent may be interpreted as support for our theory (although not proof).

How strong or weak the evidence for causality is depends on the research design. If *X* values are the result of randomization (e.g. randomly assign doses of medicine to participants), then the argument for causality is stronger. In observational data, we may consider logical relationships: parents by definition were causally responsible for their children's existence (not the other way around), so the relationship between parents and children's heights may be seen as evidence for a hereditary component of height. However, we might consider that parents and children often may have similar life circumstances in their growth periods, and that this may (partially) explain the association.

Confidence region for the regression line: mean prediction

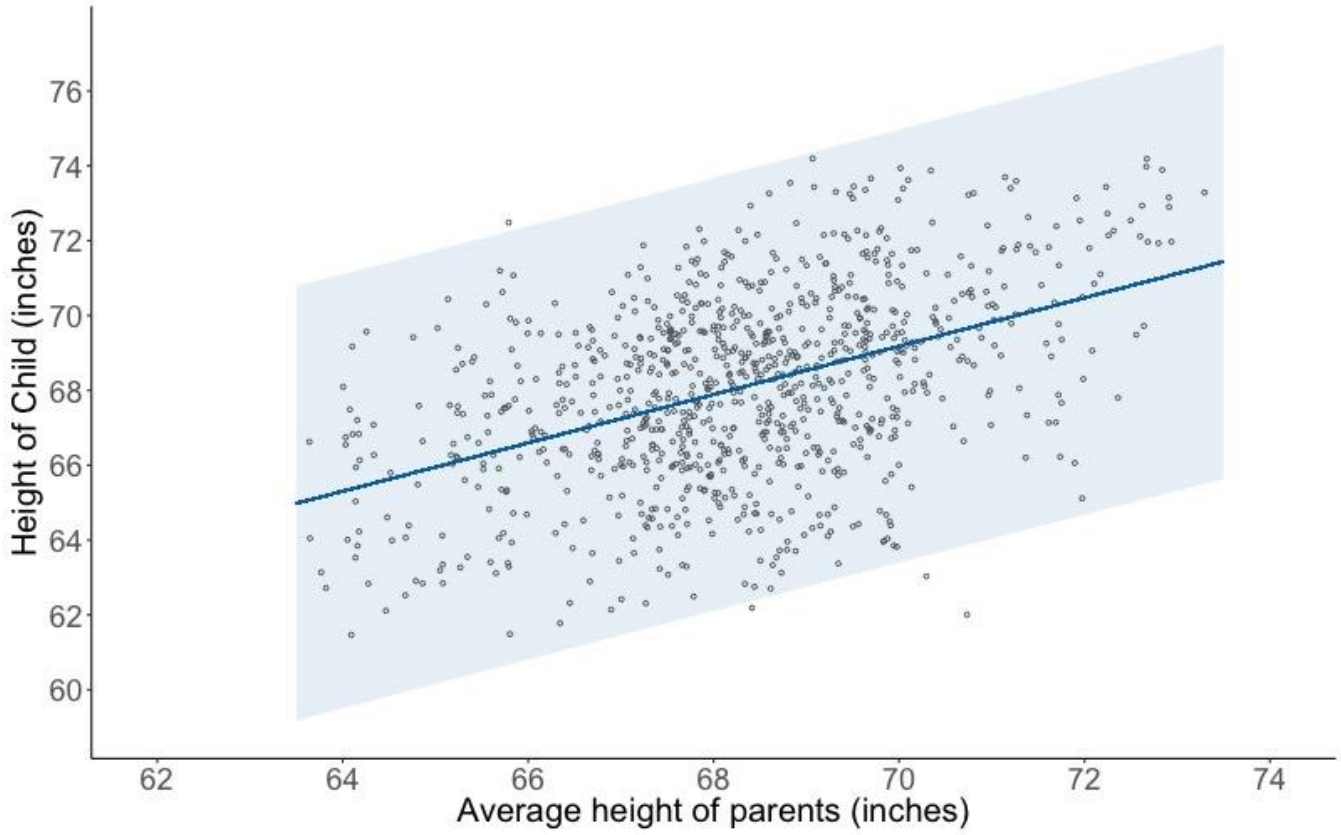
Mean prediction means in our example: the predicted <u>average</u> height of all children whose parents are X inches tall), with a 95 % confidence region.



The shaded area is a 95 % confidence region for the regression line, i.e. we are 95 % 'confident' that the shaded region contains the true regression line.

Confidence ranges for individual prediction

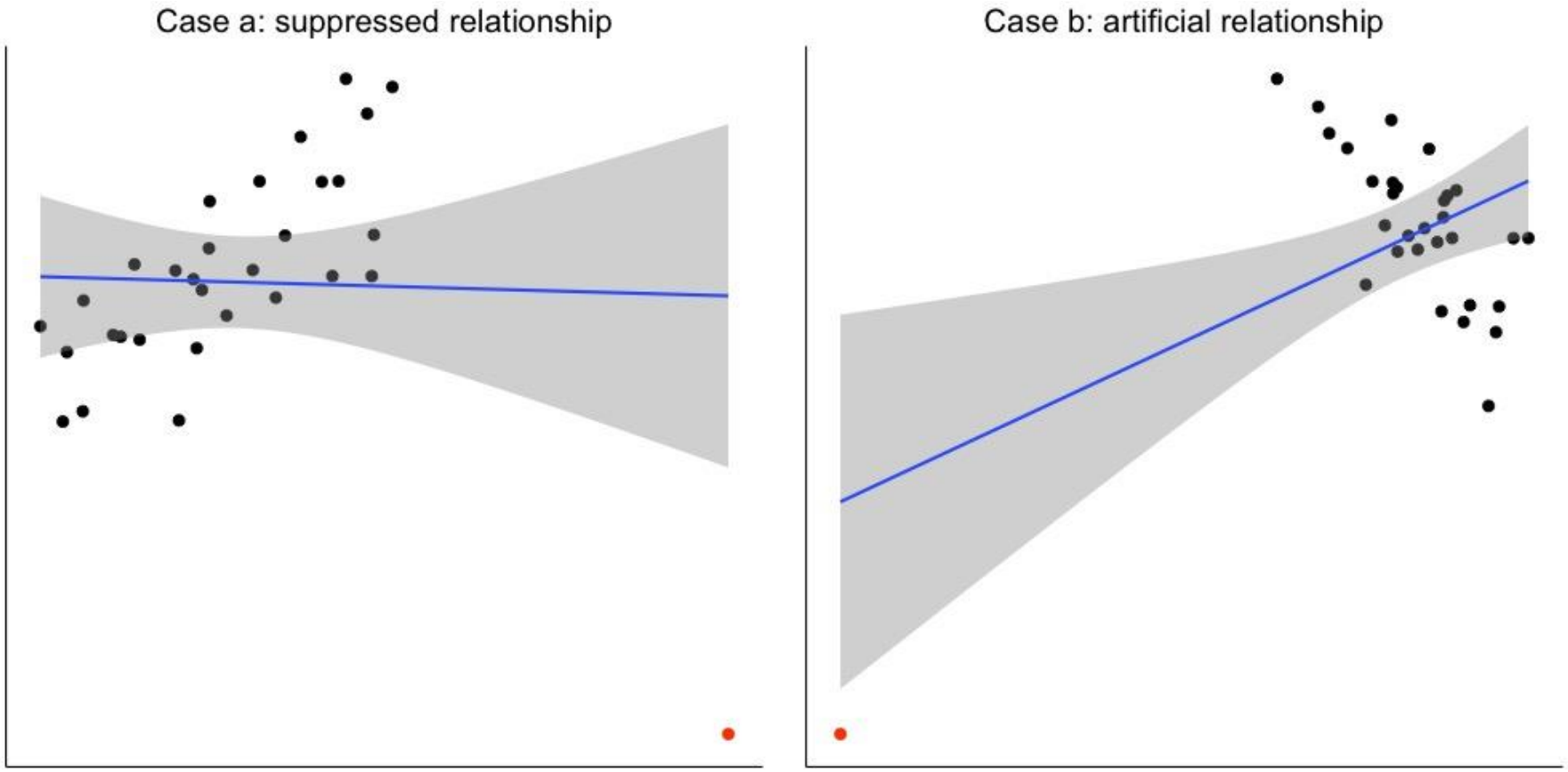
Individual **prediction** in our example: the predicted height of an individual child whose parents are *X* inches tall, with a 95% confidence range.



Note that the prediction of an individual value has a much larger margin of error than the prediction of a population average.

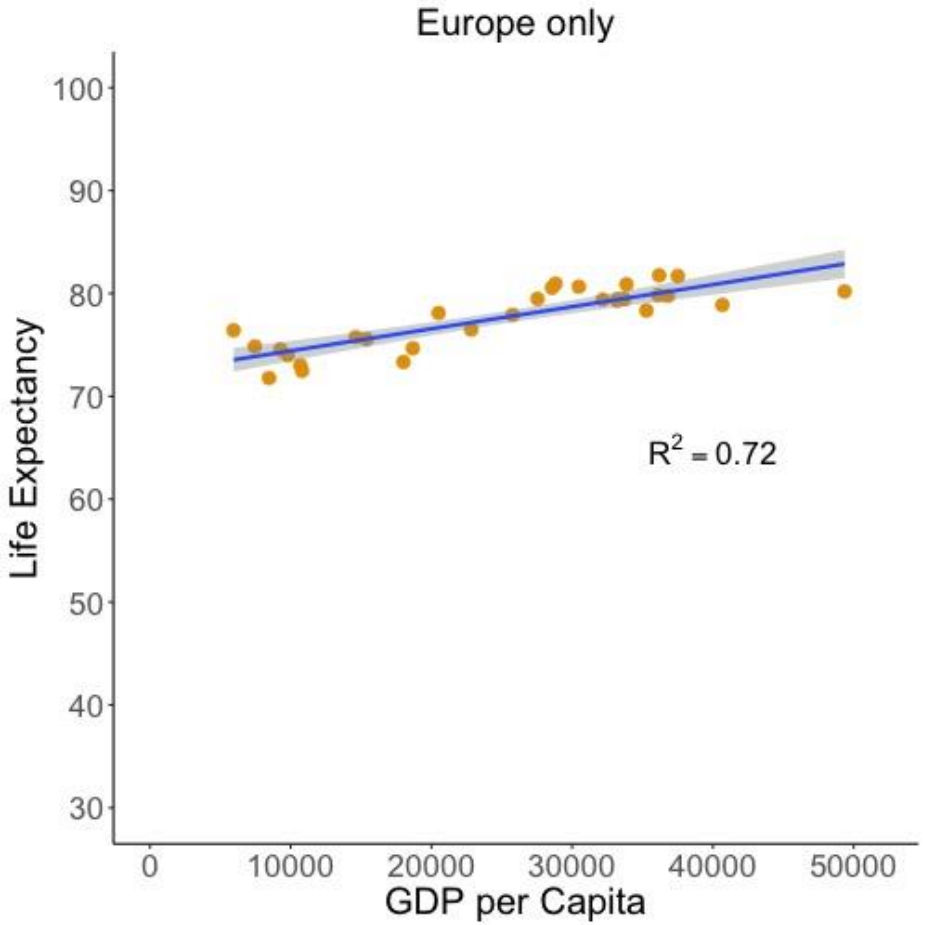
Regression in practice: things that can go wrong

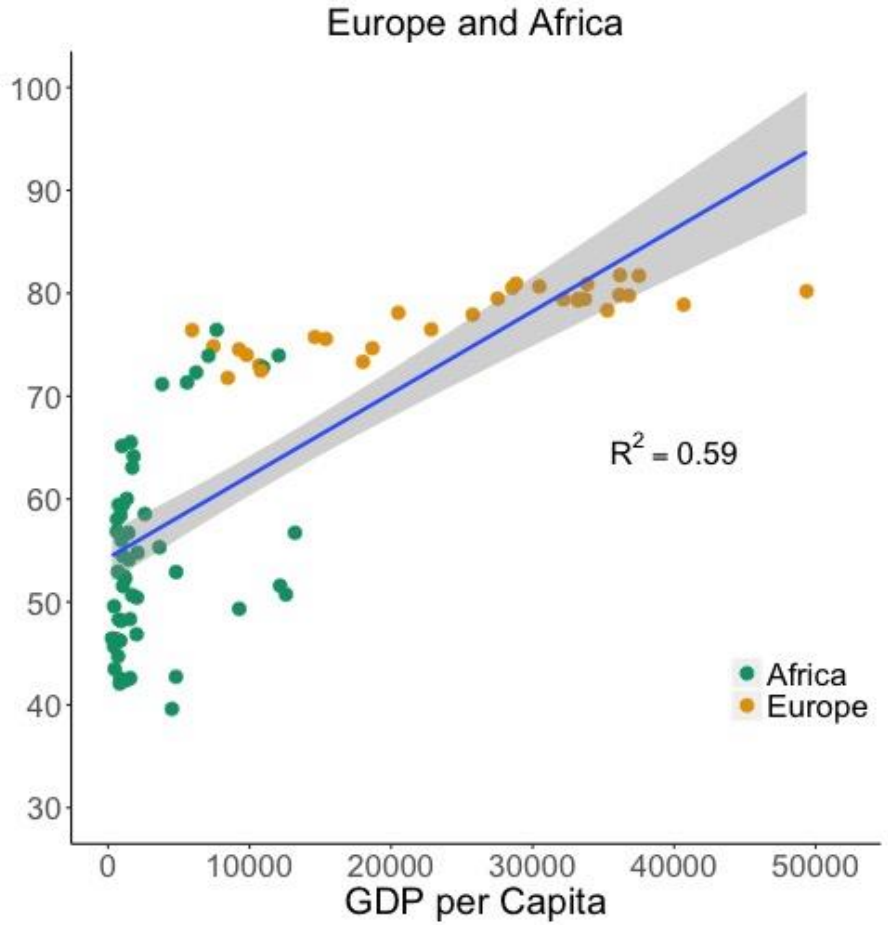
Misleading regression lines resulting from influential observations



The influential observation (red dot) leads to a misleading regression line that does a poor job at representing the data. The shaded areas give 95 % confidence regions for the regression lines.

Linear regression may be appropriate for one subset of a data set, but inappropriate for another

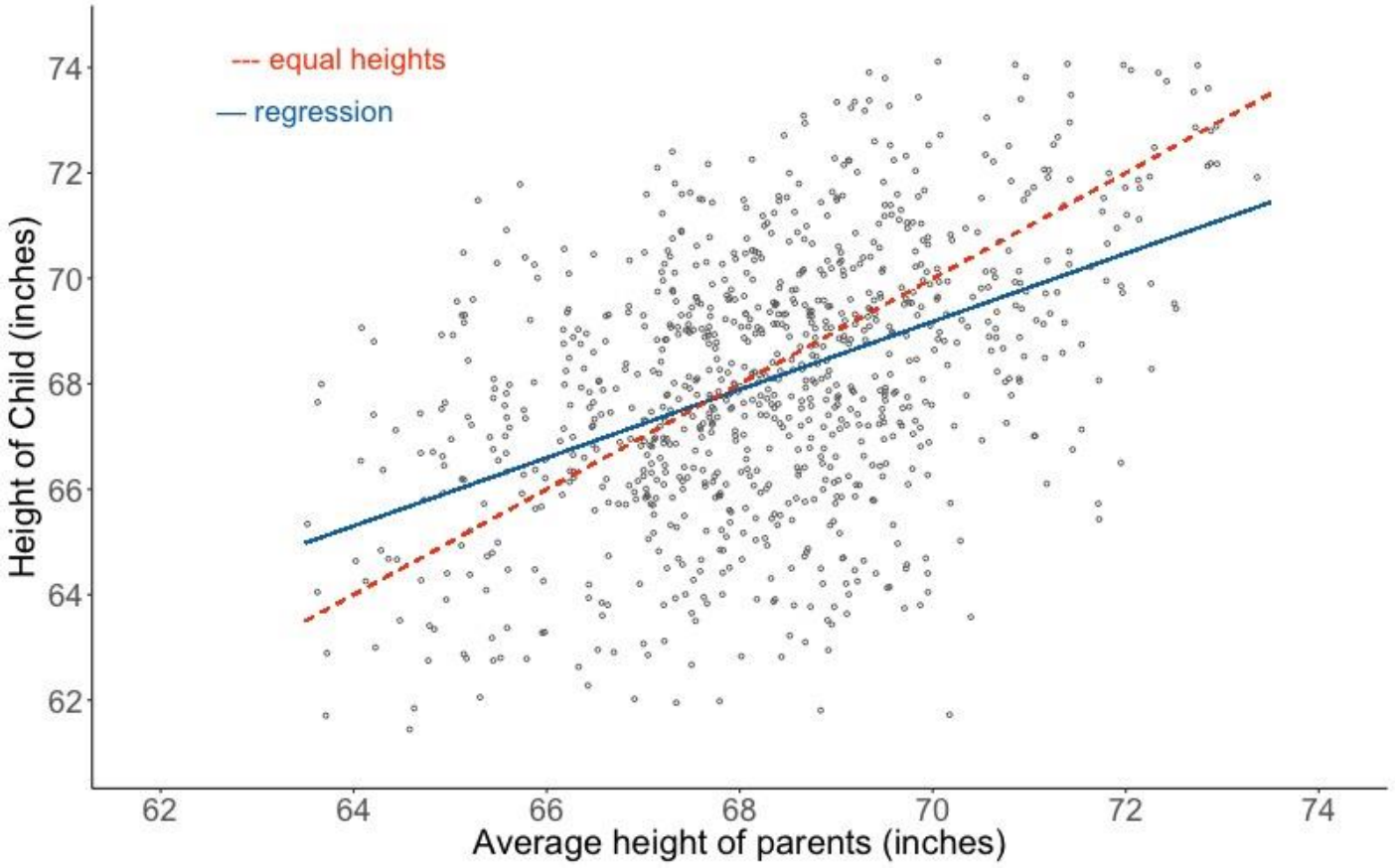




Left panel: Linear model for European countries only.

Right panel: Linear model fitted to both European and African countries, both continents as a single group.

Historical note: why regression is called regression



Galton's discovery of "regression to the mean"

Galton noted that the predicted height of the child is not, in general, the average height of the parents. Instead: "When the [parents] are taller than average, their Children tend to be shorter than they", and when parents "are shorter than average, their Children tend to be taller than they"(Galton, 1886, plate IX).

The children of parents with non-average height tend to "regress" to the average height of the population. Galton called this "filial regression to mediocrity" (Galton, 1886, p. 246). He uses the word 'regress' in the sense of 'return towards': his idea is that a child of parents whose heights deviate from the population mean has a tendency to "return" some way towards that population mean. Today, we call this phenomenon *regression to the mean*.

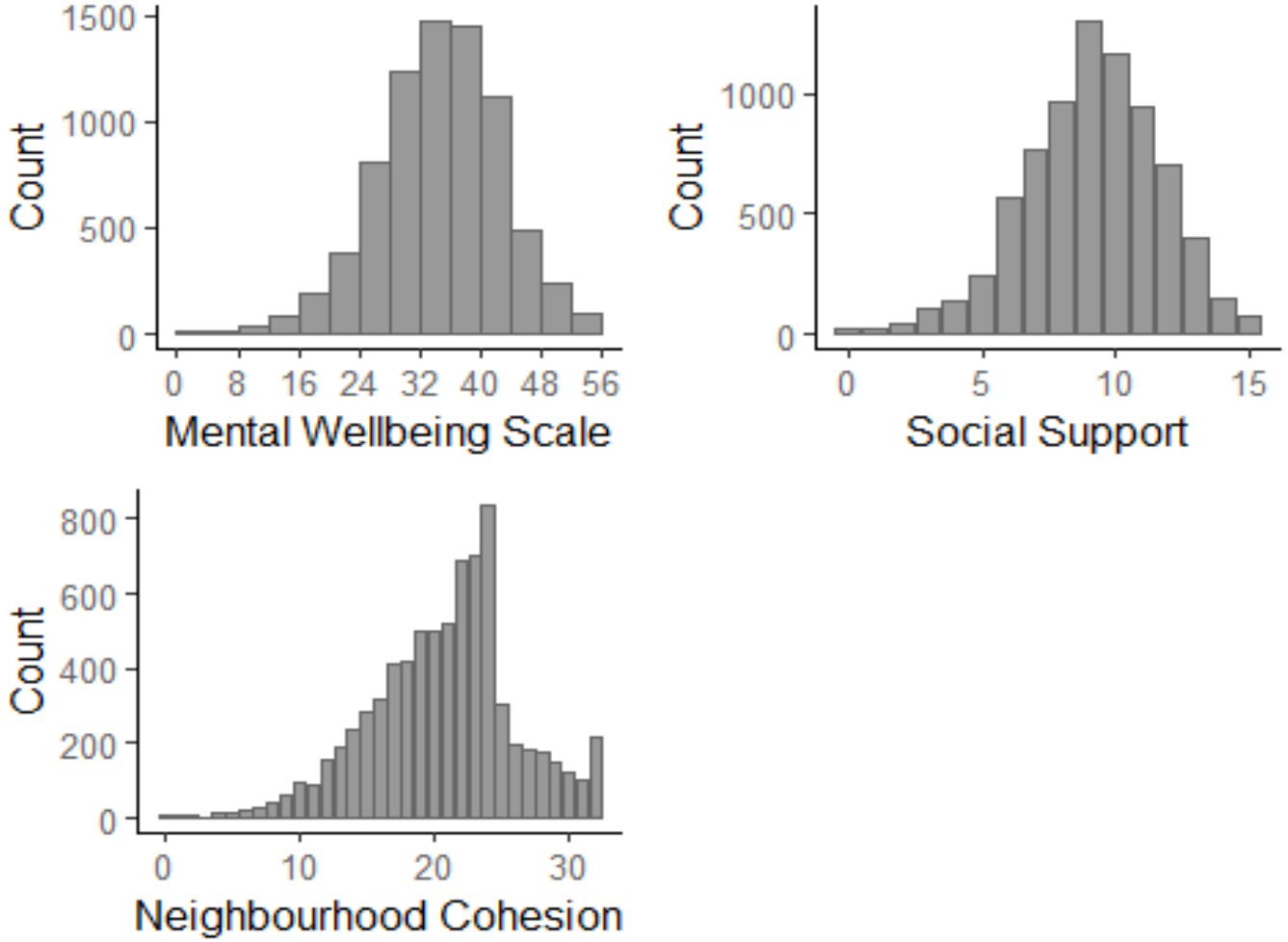
This is why the statistical model that is the subject of this lecture bears the name "regression".

Multiple linear regression: motivating example

Elliott et al (2014) used data from the National Child Development Study (NCDS) to investigate predictors of mental wellbeing. They were particularly interested in the association of Neighbourhood Cohesion (NHC) with Mental Wellbeing (MWB). The data were collected from a survey when the study participants were about 50 years old. NHC and MWB were measured at the same time, so it is difficult to infer causality.

Elliott et al (2014) were concerned about possible confounder variables, such as the availability of social support, the presence of a longstanding limiting illness, and others, which may be related both to NHC and MWB.

Descriptive statistics



Note: Data from the National Child Development Study (2008). See: http://www.cls.ioe.ac.uk/page.aspx?&sitesectionid=724&sitesectiontitle=National+Child+Development+Study

Multiple linear regression: the model

Multiple linear regression is a model of the form:

$$Y_i = \alpha + \beta_1 X_{1i} + \beta_2 X_{2i} + \ldots + \beta_p X_{pi} + \epsilon_i$$

The variables X_1, X_2, etc. may represent numeric predictors, dummy variables (coded 0 or 1, such as Limiting illness), or interactions between predictors. We will use the letter *p* to denote the number of predictors. In principle, there is no limit to the number of predictors in a multiple linear regression model.

Predicted values of the outcome are given by:

$$\widehat{Y_i} = \alpha + \beta_1 X_{1i} + \beta_2 X_{2i} + \ldots + \beta_p X_{pi}$$

Multiple linear regression: the model (2)

$$Y_i = \alpha + \beta_1 X_{1i} + \beta_2 X_{2i} + \ldots + \beta_p X_{pi} + \epsilon_i$$

The coefficient α is called the **intercept** (or, alternatively, the **constant**). It represents the predicted value of the outcome *Y* for a case whose value on all predictors is zero. The coefficients β_1, β_2 , etc. are called the **slopes**. In multiple regression the slope represents the effect of a predictor **when all other predictors are taken into account**.

You will find researchers acknowledging this interpretation using any of the following formulations. They might say that a coefficient for X_1 in a multiple regression with additional predictor X_2 represents ...

- ... the effect of X_1 controlling for X_2 (social science)
- ... the effect of X_1 adjusting for X_2 (medical and epidemiological research)
- ... the effect of X_1 holding X_2 constant (statistical textbooks)
- ... the effect of X_1 in the presence of X_2 (statistical textbooks).

Simple linear regression (no control variable)

We may estimate a simple linear regression of Mental Wellbeing (MWB) on Neighbourhood Cohesion (NHC).

The model is:

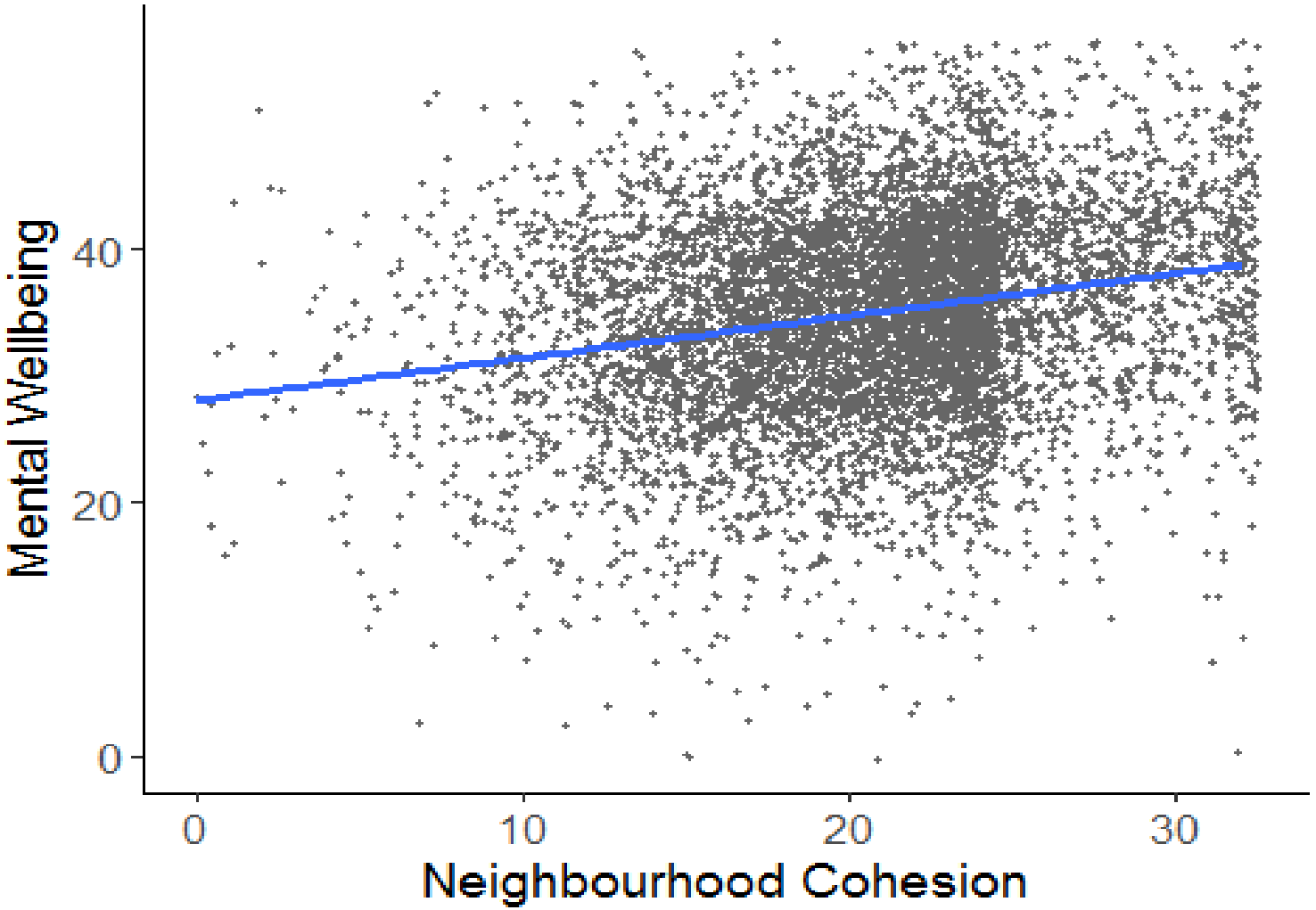
$$MWB_i = \alpha + \beta NHC_i + \varepsilon_i \ ,$$

Estimates may be displayed in a table:

			99 % C.I.	
	Estimate	Std. Error	Lower	Upper
Intercept	28.072	0.361		
Neighbourhood Cohesion	0.334	0.017	0.291	0.378

R^2 = 0.049

Mental Wellbeing and Neighbourhood Cohesion: Regression line



Note: N = 7603

A model with two predictors

The model is:

$$MWB_i = \alpha + \beta_1 NHC_i + \beta_2 SUPP_i + \varepsilon_i \,,$$

Estimates may be displayed in a table:

			99 % C.I.	
	Estimate	Std. Error	Lower	Upper
Intercept	24.124	0.428		
Neighbourhood Cohesion	0.256	0.017	0.212	0.300
Social Support	0.611	0.036	0.517	0.705

R^2 = 0.083, N = 7603

Interpretation of slope coefficients in multiple regression

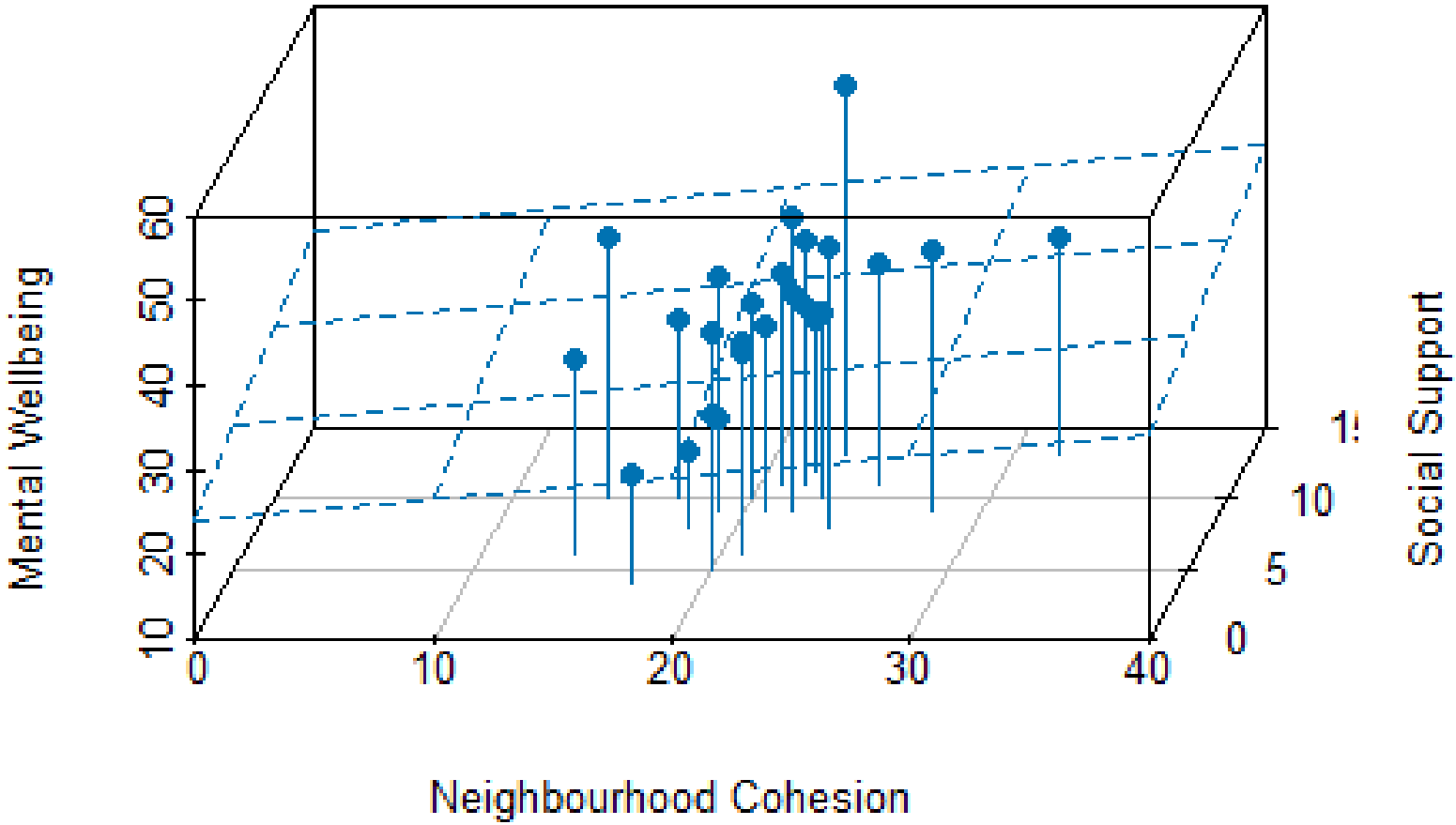
Thus the estimated regression equation is:

$$\widehat{MWB}_i = 24.124 + 0.256 \times NHC_i + 0.611 \times SUPP_i ,$$

Interpretation:

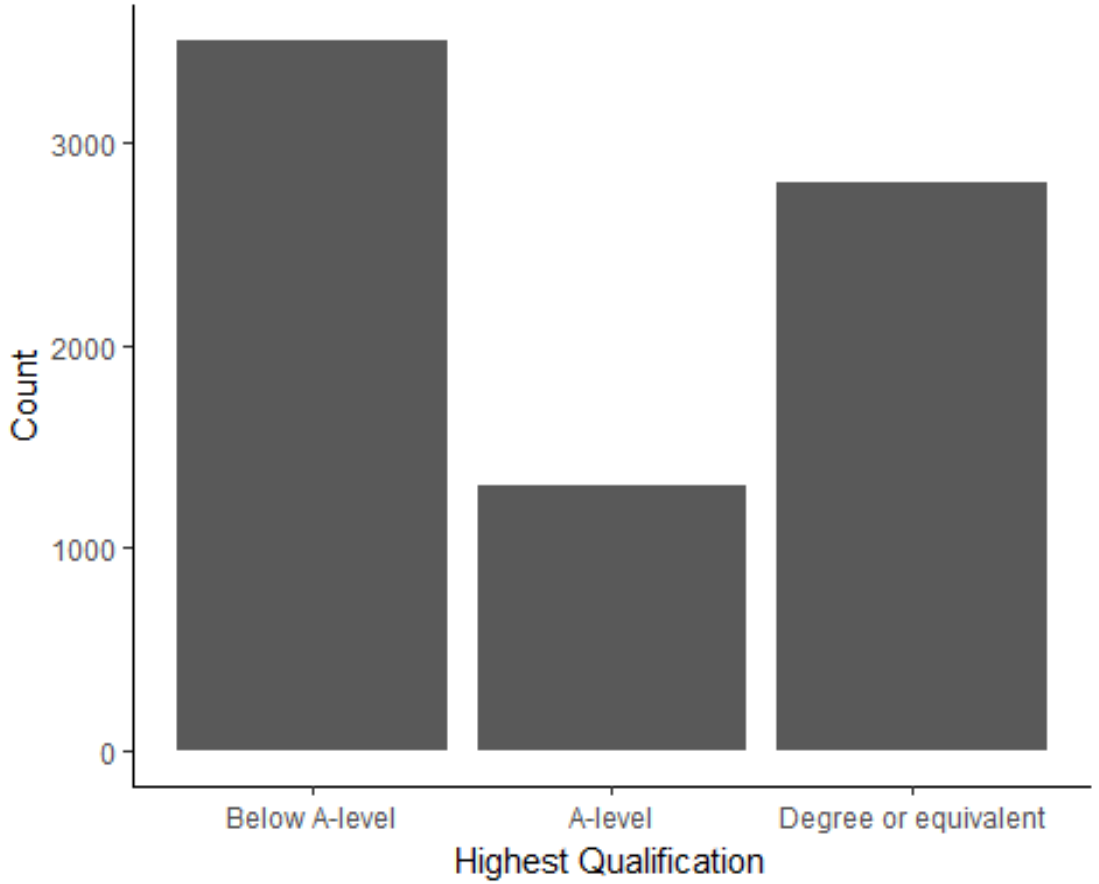
- We estimate that a one-point difference in Neighbourhood Cohesion is associated with a 0.256-point difference in Mental Wellbeing, keeping Social Support constant.
- We estimate that a one-point difference in Social Support is associated with a 0.611-point difference in Mental Wellbeing, keeping Neighbourhood Cohesion constant.

Visualization of multiple regression with two predictors



Accommodating categorical variables

Linear regression assumes that predictor variables are numerical. Consider a categorical variable such as “Highest Qualification”.



Dummy variables

Categorical variables can be used as predictors in a linear regression by transformation into dummy variables. A dummy variable is a predictor that can assume exactly two possible values: zero, or one. It turns out that we can represent a categorical variable with *c* categories by *c - 1* dummy variables.

For example, Highest Qualification has three categories. We can represent these by two dummy variables.

A scheme to represent Highest Qualification by two dummy variables

		Dummy variables	
		"A-level"	"Degree"
Highest Qualification	Below A-level	0	0
	A-level or equivalent	1	0
	Degree or equivalent	0	1

Regression model with dummy variables

The following model posits that Mental Wellbeing (MWB) is predicted by Neighbourhood Cohesion, Social Support, and Highest Qualification, where Highest Qualification is represented by the dummy variables “A-level” and “Degree”, with the reference group “below A-level”:

$$MWB_i = \alpha + \beta_1 NHC_i + \beta_2 SUPP_i + \beta_3 Alevel_i + \beta_4 Degree_i + \varepsilon_i$$

The dummy coefficients represent the predicted difference between people in the relevant dummy category and the reference group, keeping all other variables in the model constant.

Results from the estimation are shown on the next slide.

Estimates from a regression predicting Mental Wellbeing

	Estimate	Std. Error	99 % C.I. Lower	99 % C.I. Upper
Intercept	23.466	0.447		
Neighbourhood Cohesion	0.253	0.017	0.209	0.297
Social Support	0.582	0.036	0.489	0.675
Highest Qualification (ref: below A-level)				
A-level	1.141	0.250	0.497	1.786
Degree	2.127	0.196	1.622	2.633

- The predicted difference in Mental Wellbeing between those with “A-level” qualifications and the reference group is 1.141, keeping Neighbourhood Cohesion and Social Support constant.
- The predicted difference between those with a degree and the reference group is 2.127, keeping the other variables constant.
- This implies that the predicted difference between those with a degree and those with A-levels is 2.127 – 1.141 = 0.986.

Estimated regression equation with dummy variables

$$\widehat{MWB}_i = 23.466 + 0.253 \times NHC_i + 0.582 \times SUPP_i + 1.141 \times Alevel_i + 2.127 \times Degree_i$$

From this, we can derive estimated regression equations for all three qualification groups.

Interpreting predictive equations with dummy coefficients

For the reference group, the regression equation is:

$$\widehat{MWB}_i = 23.466 + 0.253 \times NHC_i + 0.582 \times SUPP_i + \mathbf{1.141} \times \mathbf{0} + \mathbf{2.127} \times \mathbf{0}$$

$$\widehat{MWB}_i = \mathbf{23.466} + 0.253 \times NHC_i + 0.582 \times SUPP_i$$

For those with A-levels, the regression equation is:

$$\widehat{MWB}_i = 23.466 + 0.253 \times NHC_i + 0.582 \times SUPP_i + \mathbf{1.141} \times \mathbf{1} + \mathbf{2.127} \times \mathbf{0}$$

$$\widehat{MWB}_i = (\mathbf{23.466} + \mathbf{1.141}) + 0.253 \times NHC_i + 0.582 \times SUPP_i$$

$$\widehat{MWB}_i = \mathbf{24.607} + 0.253 \times NHC_i + 0.582 \times SUPP_i$$

For those with a degree, the regression equation is:

$$\widehat{MWB}_i = 23.466 + 0.253 \times NHC_i + 0.582 \times SUPP_i + \mathbf{1.141} \times \mathbf{0} + \mathbf{2.127} \times \mathbf{1}$$

$$\widehat{MWB}_i = (\mathbf{23.466} + \mathbf{2.127}) + 0.253 \times NHC_i + 0.582 \times SUPP_i$$

$$\widehat{MWB}_i = \mathbf{25.593} + 0.253 \times NHC_i + 0.582 \times SUPP_i$$

Regression models and statistical inference

Statistical inference is the art of drawing conclusions from a random sample to a population, or from a particular data set to a hypothesized process.

Hypothesis tests, confidence intervals, and confidence ranges for prediction are all examples of inferential statistics.

Inferences about regression models are generally valid only under the condition that certain assumptions about the data are satisfied. We will introduce the inferential procedures first, and finally deal with the assumptions that must be met for these inferences to be valid.

Confidence interval of a single coefficient

If all regression assumptions are satisfied, the estimate of a slope coefficient follows a t-distribution with *n-p-1* degrees of freedom, where *n* is the sample size and *p* is the number of predictors in the regression model.

A 99 % confidence interval for a coefficient is computed as:

$$CI_{.99} = \hat{\beta} \pm s_{\hat{\beta}} \times t_{n-p-1,.995} \text{ ,}$$

where:

- $\hat{\beta}$ is a regression coefficient estimate (subscript omitted)
- $s_{\hat{\beta}}$ is the estimated standard error of $\hat{\beta}$
- $t_{n-p,.995}$ is the 99.5th percentile of the t-distribution with *n-p-1* degrees of freedom

Confidence interval: example

	Estimate	Std. Error	99 % C.I. Lower	99 % C.I. Upper
Intercept	23.466	0.447		
Neighbourhood Cohesion	**0.253**	**0.017**	**0.209**	**0.297**
Social Support	0.582	0.036	0.489	0.675
Highest Qualification (ref: below A-level)				
A-level	1.141	0.250	0.497	1.786
Degree	2.127	0.196	1.622	2.633

99 % Confidence interval for the coefficient of Neighbourhood Cohesion:

$$
\begin{aligned}
CI_{.99} &= 0.253 \pm 0.017 \times t_{n-p-1,.995} \\
&= 0.253 \pm 0.017 \times t_{7603-4-1,.995} \\
&= 0.253 \pm 0.017 \times 2.576 \\
&= (0.209; 0.297)
\end{aligned}
$$

Confidence interval: interpretation

	Estimate	Std. Error	99 % C.I. Lower	99 % C.I. Upper
Intercept	23.466	0.447		
Neighbourhood Cohesion	**0.253**	**0.017**	**0.209**	**0.297**
Social Support	0.582	0.036	0.489	0.675
Highest Qualification (ref: below A-level)				
A-level	1.141	0.250	0.497	1.786
Degree	2.127	0.196	1.622	2.633

Interpretation

We are 99 % “confident” that the interval from 0.209 to 0.297 contains the (true) coefficient for Neighbourhood Cohesion, controlling for Social Support and Highest Qualification.

The confidence interval puts a probable “margin of error” around our estimate of a regression coefficient.

T-test for a single coefficient

We can test hypotheses about a single coefficient using a t-test of the form:

$$t_{obs} = \frac{\hat{\beta} - \beta_0}{s_{\hat{\beta}}}$$

where

- $\hat{\beta}$ and $s_{\hat{\beta}}$ are defined as above, and
- β_0 is a the null hypothesis value of the coefficient (the value that we wish to test)

Often, β_0 is set to zero so that

$$t_{obs} = \frac{\hat{\beta}}{s_{\hat{\beta}}}$$

A p-value can be calculated comparing t_{obs} to the t-distribution with *n-p-1* degrees of freedom (the same distribution as used for the calculation of confidence intervals).

Typical linear regression results table displayed in software

	Estimate	Std. Error	t	p
Intercept	23.466	0.447		
Neighbourhood Cohesion	0.253	0.017	14.816	.000
Social Support	0.582	0.036	16.075	.000
Highest Qualification (ref: below A-level)				
A-level	1.141	0.250	4.562	.000
Degree	2.127	0.196	10.844	.000

Note, however, that the p-values displayed in such a table are typically not adjusted for multiple tests. So it is **not** good practice to make multiple decisions about the 'statistical significance' of covariates on the basis of p-values from the typical software output table.

It may be preferable to generate specific hypotheses in advance of seeing the data, and use systematic model comparisons to investigate which of two or more models is most compatible with the data.

Analysis of Variance Table

	Sum of Squares	Degrees of freedom (df)	Mean Square (MS)
Regression	$SS_{Reg} = \sum(\hat{Y}_i - \overline{Y})^2$	*p*	$MS_{Reg} = SS_{Reg} / p$
Residual	$SS_{Res} = \sum(Y_i - \hat{Y}_i)^2$	*n - p - 1*	$MS_{Res} = SS_{Res} / (n - p - 1)$
Total	$SS_{Tot} = \sum(Y_i - \overline{Y})^2$	*n - 1*	$MS_{Tot} = SS_{Tot} / (n - 1)$

Notes: p: number of predictors in the model; n: sample size. SS: sum of squares; MS: mean square.

The analysis of variance table decomposes the total variation of the outcome (**SS_{Tot}**) into:

- **SS_{Reg}**: the variation 'accounted for' by the predictors
- **SS_{Res}**: the residual variation not accounted for by the predictors

$$SS_{Tot} = SS_{Reg} + SS_{Res}$$

The ANOVA table, R^2 statistic and adjusted R^2 statistic

R^2 statistic: the proportion of outcome variance the model 'accounts for' in the data set at hand:

$$R^2 = 1 - \frac{SS_{Residual}}{SS_{Total}}$$

Adjusted R^2: better than (unadjusted) R^2 as an estimate of how much of the outcome variance the model can 'account for' in the population (or in a new data set):

$$Adj.R^2 = 1 - \frac{MS_{Residual}}{MS_{Total}}$$

Of the two, the adjusted R^2 is the better statistic to use for model comparisons. The (unadjusted) R^2 will always tend to favour the larger model, but this is not the case for the adjusted R^2.

Model comparison

We can compare different models with one another. As an example, consider the question whether Highest Qualification improves the prediction of Mental Wellbeing, in the presence of the variables Neighbourhood Cohesion and Social Support. Note that this question cannot be answered directly via a t-test, since the t-tests addresses hypotheses about a single coefficient, while the two models stated below differ with respect to two coefficients.

Model 1:

$$MWB_i = \alpha + \beta_1 NHC_i + \beta_2 SUPP_i + \varepsilon_i$$

Model 2:

$$MWB_i = \alpha + \beta_1 NHC_i + \beta_2 SUPP_i + \beta_3 Alevel_i + \beta_4 Degree_i + \varepsilon_i$$

Model comparison using adjusted R^2

	Model 1	Model 2
Adjusted R^2	0.083	0.132

Model 2 has the higher adjusted R^2-value. Model 2 'accounts for' about 13 % of the variance of Mental Wellbeing, while Model 1 accounts for about 8 %. This constitutes evidence that Model 2 results in a better fit to the data than Model 1. In our example, this means that we have evidence that Highest Qualification is a predictor of Mental Wellbeing, once Neighbourhood Cohesion and Social Support have been taken into account.

Model comparison using statistical significance tests:

Nested models

We can also use statistical significance tests to investigate hypotheses about models. We can compare two models using significance tests when the two models are nested. Formally, two models are nested if we can turn the larger model into the smaller one by imposing constraints on some of the parameters of the larger model. For example, looking at the models on the previous slide: model 1 is nested within model 2, because we can turn model 2 into model 1 by setting $\beta_3 = 0$ and $\beta_4 = 0$. Put differently: if we assume that the coefficients of the two variables contained in Model 2 but not in Model 1 are zero, then the two models are identical.

If models are nested, we can use either significance test or adjusted R^2 to compare them. For non-nested models, significance tests are not appropriate.

Nested and non-nested models

Consider the following three models. Which pairs of models are nested?

$$Model\ A\text{:}\ \ Y_i = \alpha + \beta_1 X_{1i} + \epsilon_i$$

$$Model\ B\text{:}\ \ Y_i = \alpha + \beta_1 X_{1i} + \beta_2 X_{2i} + \epsilon_i$$

$$Model\ C\text{:}\ \ Y_i = \alpha + \beta_1 X_{1i} + \beta_3 X_{3i} + \ \beta_4 X_{4i} + \epsilon_i$$

- Model A is nested within Model B, because if we set the parameter β_2 in Model B to zero, the two models are identical.
- Similarly, Model A is also nested within Model C, because we can turn Model C into Model A by setting $\beta_3 \ = \ \beta_4 \ = \ 0$.
- However, models B & C are non-nested, because each contains parameters that the other does not have: Model C does not contain a parameter for variable X_2, while Model B does not contain parameters for X_3 and X_4.

Nested model comparison using F-tests

We can compare nested models using a statistical significance test called the F-test.

Null hypothesis: The larger model makes no improvement in predicting the outcome relative to the smaller model.

Alternative hypothesis: The larger model accounts for a larger proportion of the outcome variance than the smaller model.

F-statistic: formal definition

The test statistic of the F-test is called F and is computed as follows:

$$F = \frac{(SS_{Res,S} - SS_{Res,L})/(df_{Res,S} - df_{Res,L})}{MS_{Res,L}},$$

where

- $SS_{Res,L}$, $df_{Res,L}$, and $MS_{Res,L}$ are the residual sum of squares, degrees of freedom and mean square of the larger model, respectively, while
- $SS_{Res,S}$, $df_{Res,S}$ are the residual sum of squares and degrees of freedom of the smaller model, respectively.

Expressed in words, the equation above states that F is computed as:

$$F = \frac{Difference\ in\ regression\ SS\ between\ the\ two\ models\ /\ Difference\ in\ df}{Residual\ MS\ for\ larger\ model}$$

ANOVA table of model comparison

	SS_{Res}	Difference in SS_{Res}	df_{Res}	Difference in df_{Res}	F	df	p-value
Model 1	458604.0		7600				
Model 2	451572.4	7031.6	7598	5	59.155	2, 7598	.000

The F-statistic is calculated from the numbers in the table as follows:

$$F = \frac{(458604.0 - 451572.4)/(7600 - 7598)}{451572.4/7598} = \frac{7031.6 \,/\, 2}{59.433} = 59.155$$

In this example, the p-value is small. So we can consider this as evidence that Model 2 improves the prediction of Mental Wellbeing compared to Model 1.

Model assumptions and regression diagnostics

A regression is a type of statistical model. All statistical models make assumptions about the data – or more accurately, about the processes that have generated the data. The next few slides list the assumptions underlying multiple linear regression. All but one (absence of collinearity) apply to simple linear regression also. Some assumptions underlie the regression procedure as a whole. Others become important only when we wish to use statistical inference (confidence intervals, prediction, and hypothesis tests).

Assumptions of linear regression (1)

Linearity: The relationship between the predictor and the outcome is linear. If this assumption is not met, the linear model misrepresents the true shape of the relationship between the predictor and the outcome.

Normality of errors: The errors follow a normal distribution. If the errors are not normally distributed, hypothesis tests and confidence intervals around model coefficients and model predictions may not be correct.

Homoscedasticity of errors. The variance of the errors around the regression line is the same at every point of the regression line. "Homoscedasticity" is a compound word made of classical Greek roots meaning "same variance". The opposite of homoscedasticity is heteroscedasticity ("different variance"). If the errors are heteroscedastic, hypothesis tests and confidence intervals around model coefficients and model predictions may not be correct.

Assumptions of linear regression (2)

Independence of errors: The errors are independent of one another. If there is dependency between some or all errors, hypothesis tests and confidence intervals around model coefficients and model predictions may not be correct. Dependency occurs, for example, when participants occur in natural groups, such as children being clustered in schools. For such situations, multilevel models (also called mixed-effects models) may be employed.

Randomness of errors: **The errors are the result of a random process.** A random process may be built into the design of our study, such as when we employ random sampling to select survey respondents. But we may also assume randomness in social or natural processes. For example, Galton's data on heights were not from a random sample of child-parent pairs. Nonetheless, our model assumed variations of children's heights around the model predicted value were the result of a random process. Given our knowledge of genetics today (which Galton himself did not have), we might argue that the selection of parent genes that are passed on to their children is such a random process. If the errors are not in fact random, then our inferences to a population or process might be illusory.

Assumptions of linear regression (3)

The predictor is measured without error. The linear regression model has no error term for the predictor variable, and thus implicitly assumes that the predictor variable has no measurement error. This assumption cannot be tested within the framework of linear regression itself. Structural equation models are one way to analyse data where we take into account measurement errors in both predictor and outcome variables.

Absence of extremely influential observations: There are no extreme outliers in the data that distort the observed relationship between the predictor and the outcome. If this assumption is not met, coefficient estimates may be misleading.

Assumptions of linear regression (4)

Absence of collinearity. Multicollinearity occurs when one or more predictor variables can be perfectly or almost perfectly predicted from one or more other predictors.

When multicollinearity is perfect, the model estimation will break down.

When predictors are very highly correlated, even though not perfectly collinear, the model estimation may be mathematically possible, but the results might be unstable. "Unstable" here means that the coefficient estimates would be liable to change drastically in response to small changes in the data.

Errors, residuals, and standardized residuals

The assumptions of normality and homoscedasticity concern the errors. Although we do not observe the errors directly, we do observe the residuals from a regression model. The residuals, however, by definition are correlated with one another, and are heteroscedastic (they have the largest variance near the centre of the regression line, and smaller variance the more you move away from the centre).

The solution is to calculate **standardized residuals**. These can be used to investigate the plausibility of the assumptions of

- Linearity
- Homoscedasticity
- Normality.

Standardized residuals

If all regression assumptions are satisfied, then the standardized residuals should follow a standard normal distribution (with mean 0 and standard deviation 1).

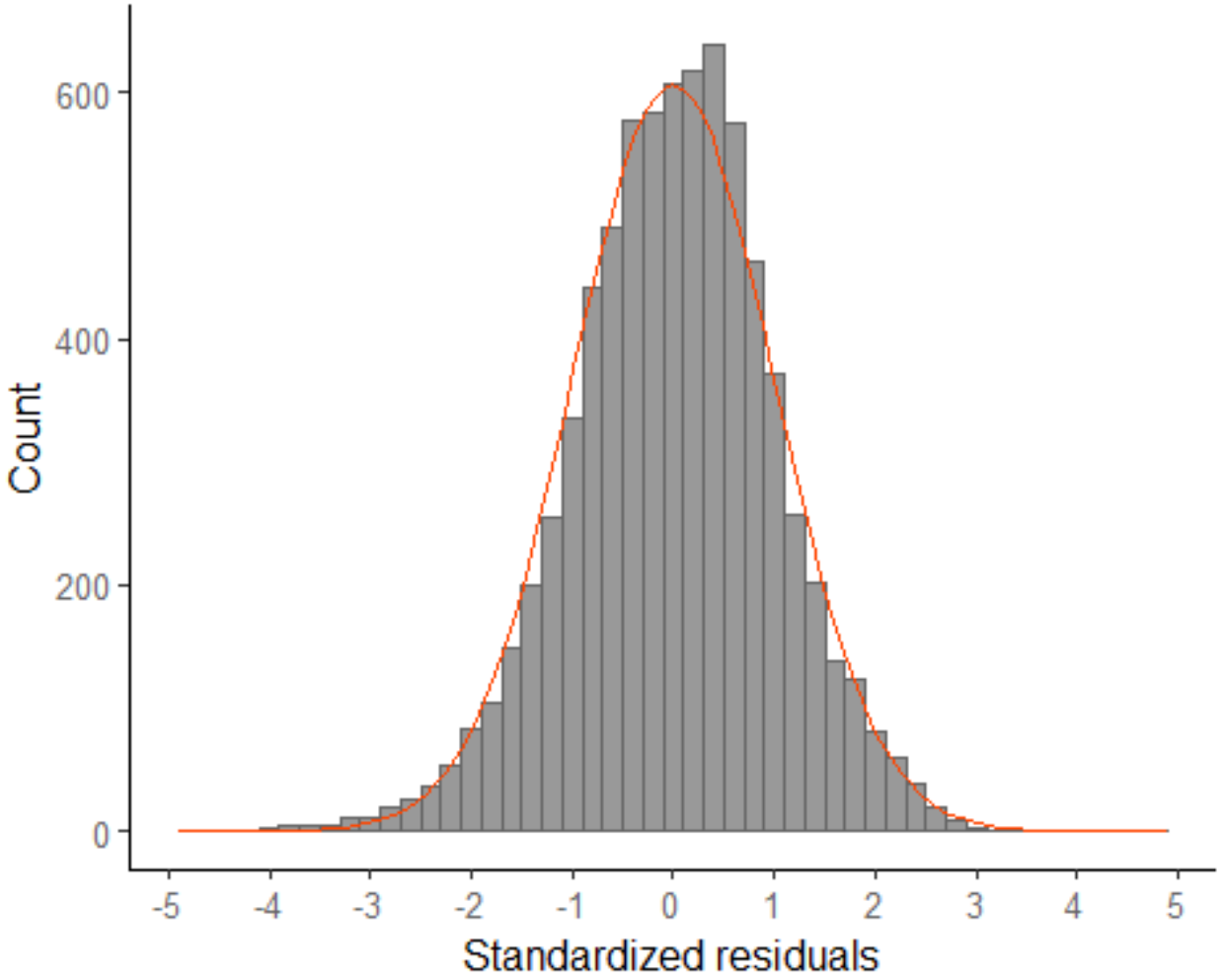
Standardized residuals are calculated as follows:

$$stdres_i = \frac{e_i}{s\sqrt{1 - h_{ii}}}$$

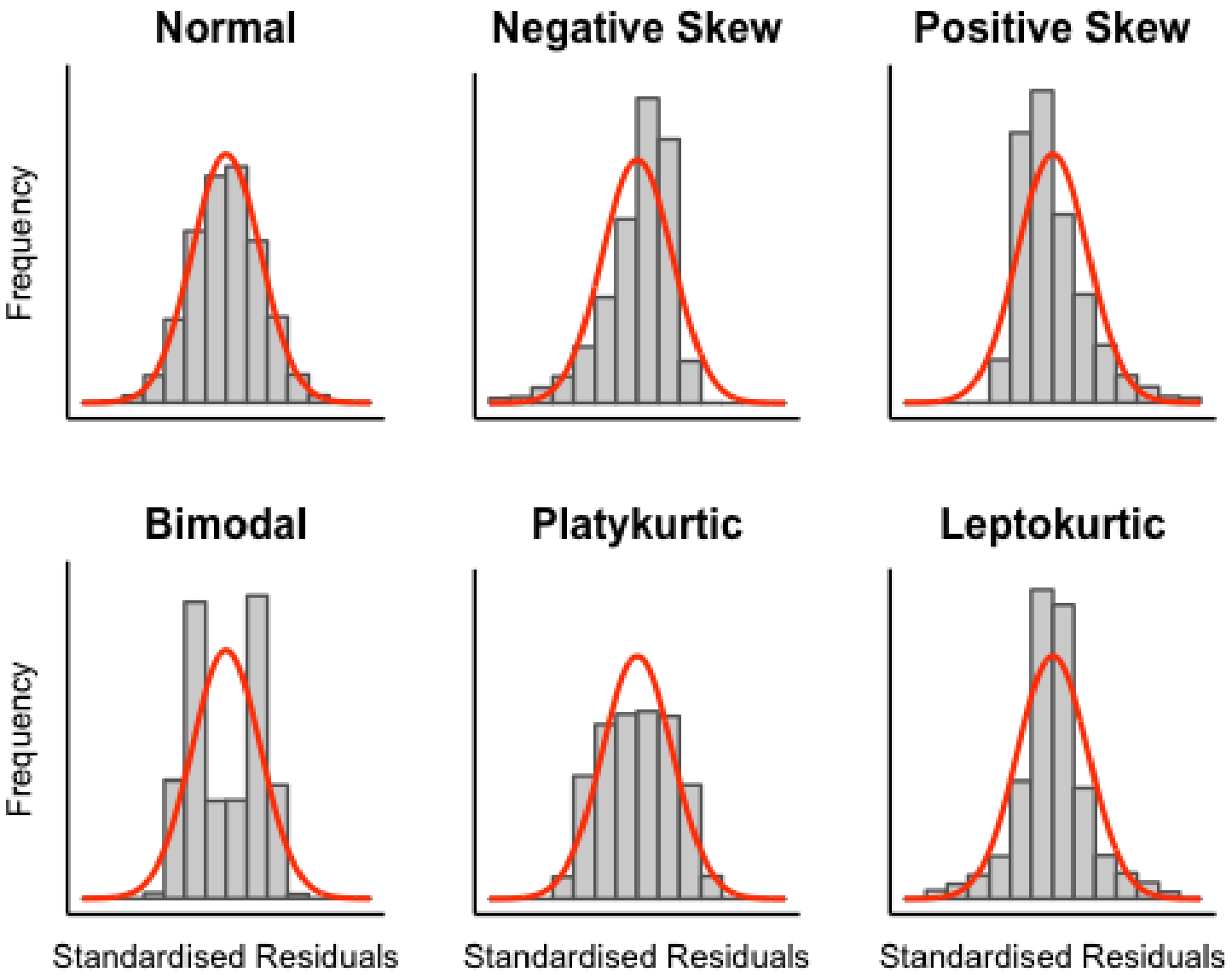
where

- $stdres_i$ is the standardized residual of the i^{th} observation
- e_i is the (unstandardized) residual of the i^{th} observation
- *s* is the *standard error of the estimate* – this is an estimate of the standard deviation of the residuals (but it's calculated with denominator equal to df_{Res}, rather than $n - 1$ as otherwise usual for standard deviations)
- h_{ii} is the *leverage* of the i^{th} observation. The leverage is a number between 1/n and 1 that indicates how untypical the predictor values of the i^{th} observation are.

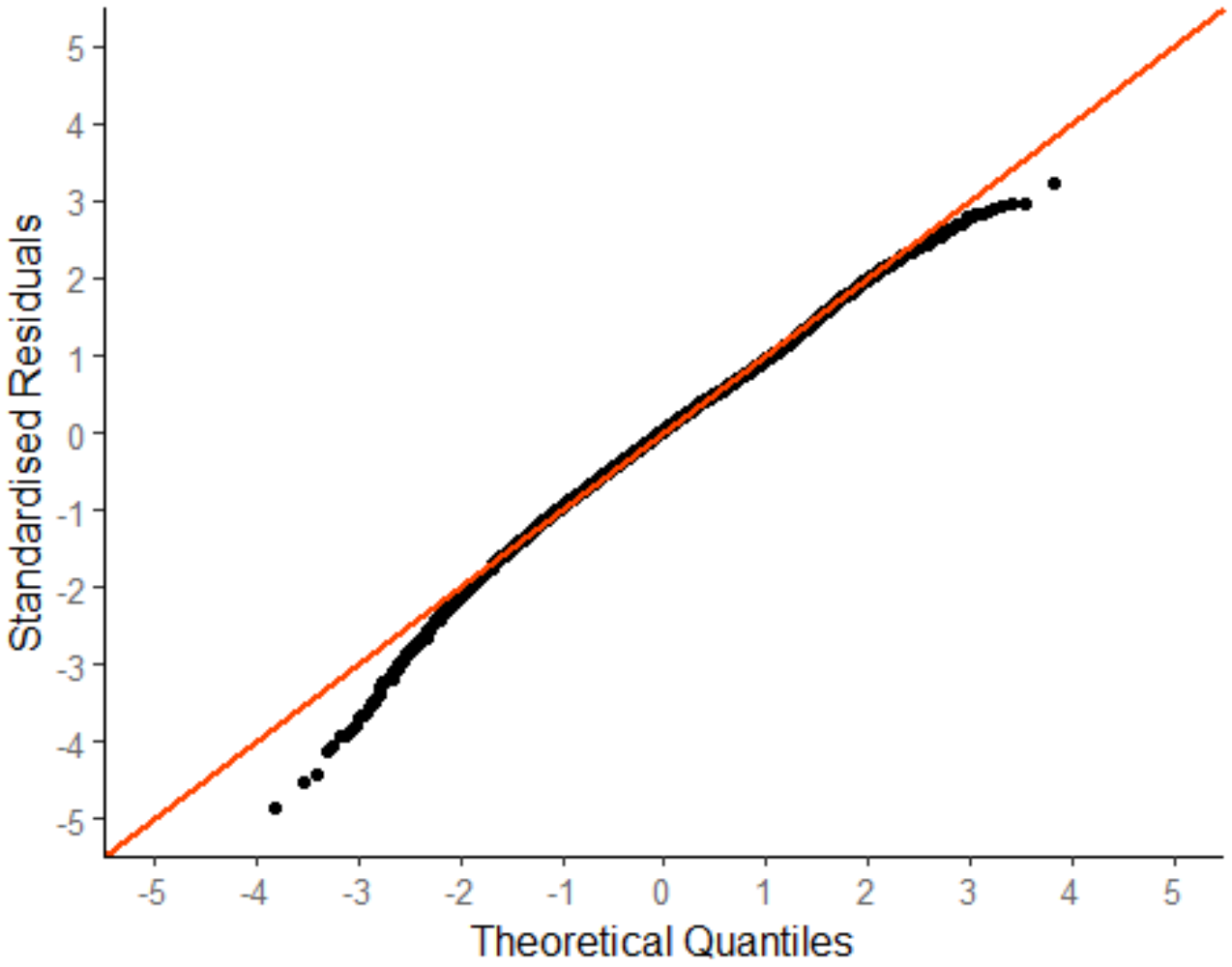
Regression diagnostics: Histogram to assess normality



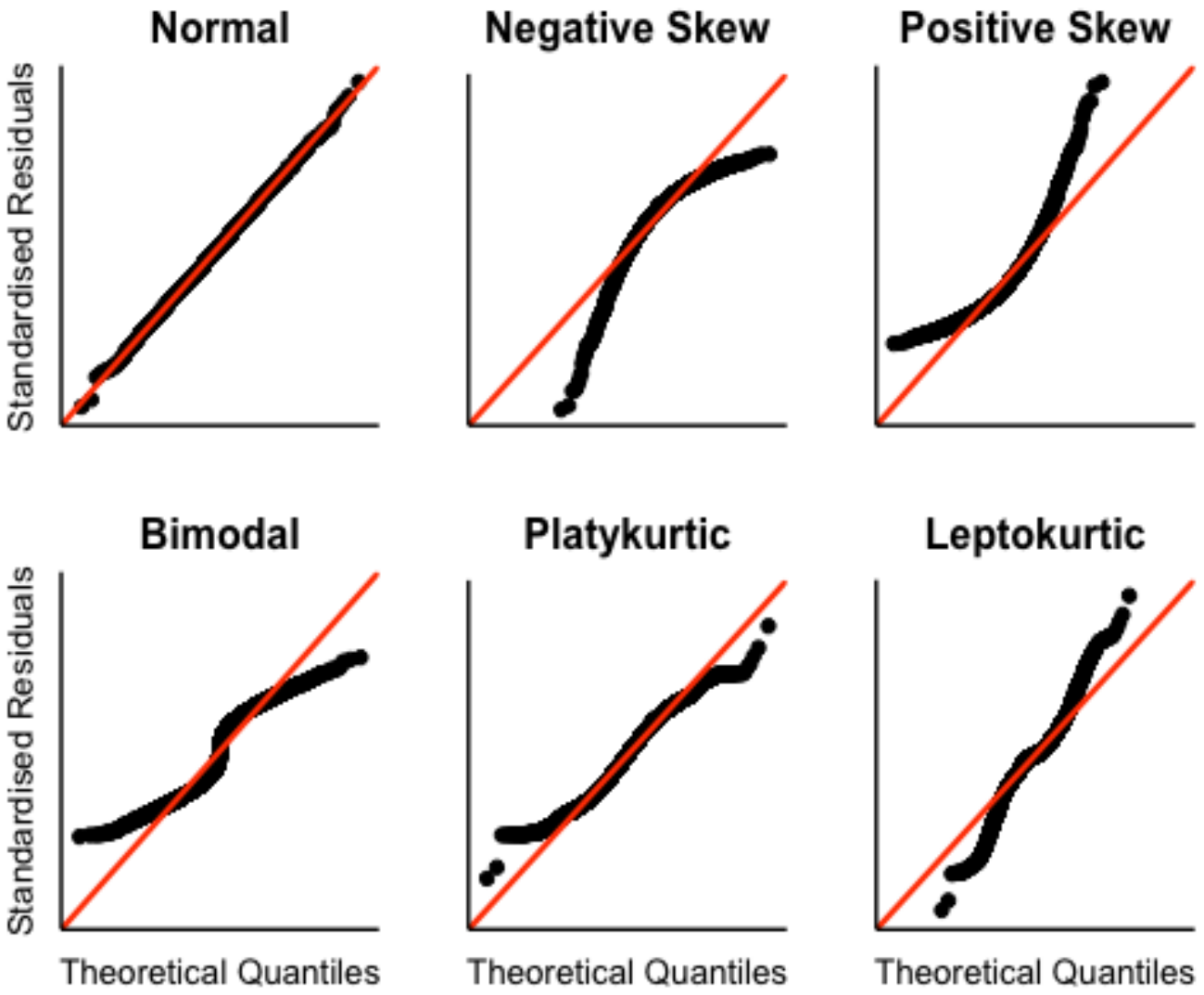
Evaluating histograms



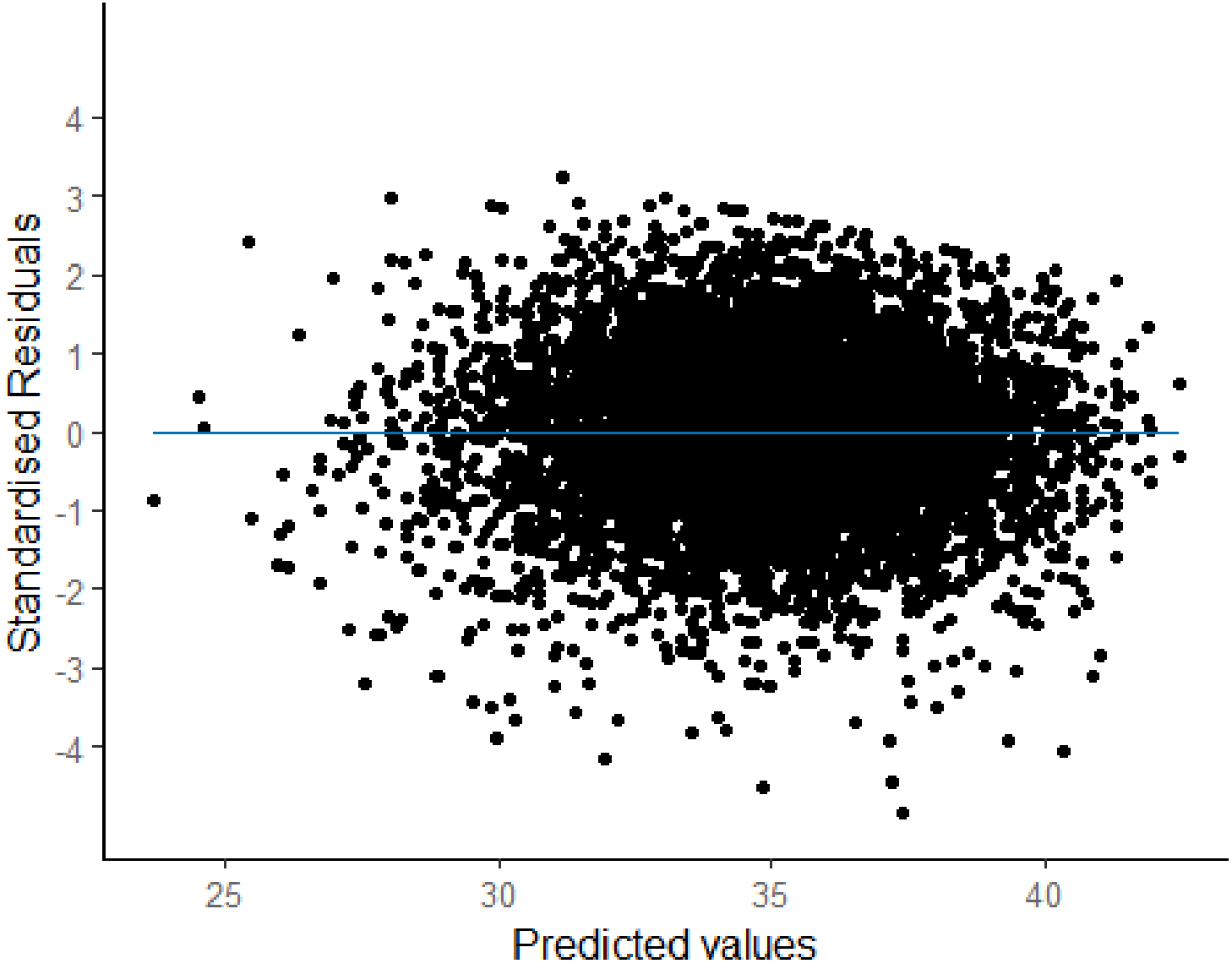
Normal quantile-quantile plots (normal q-q plots)



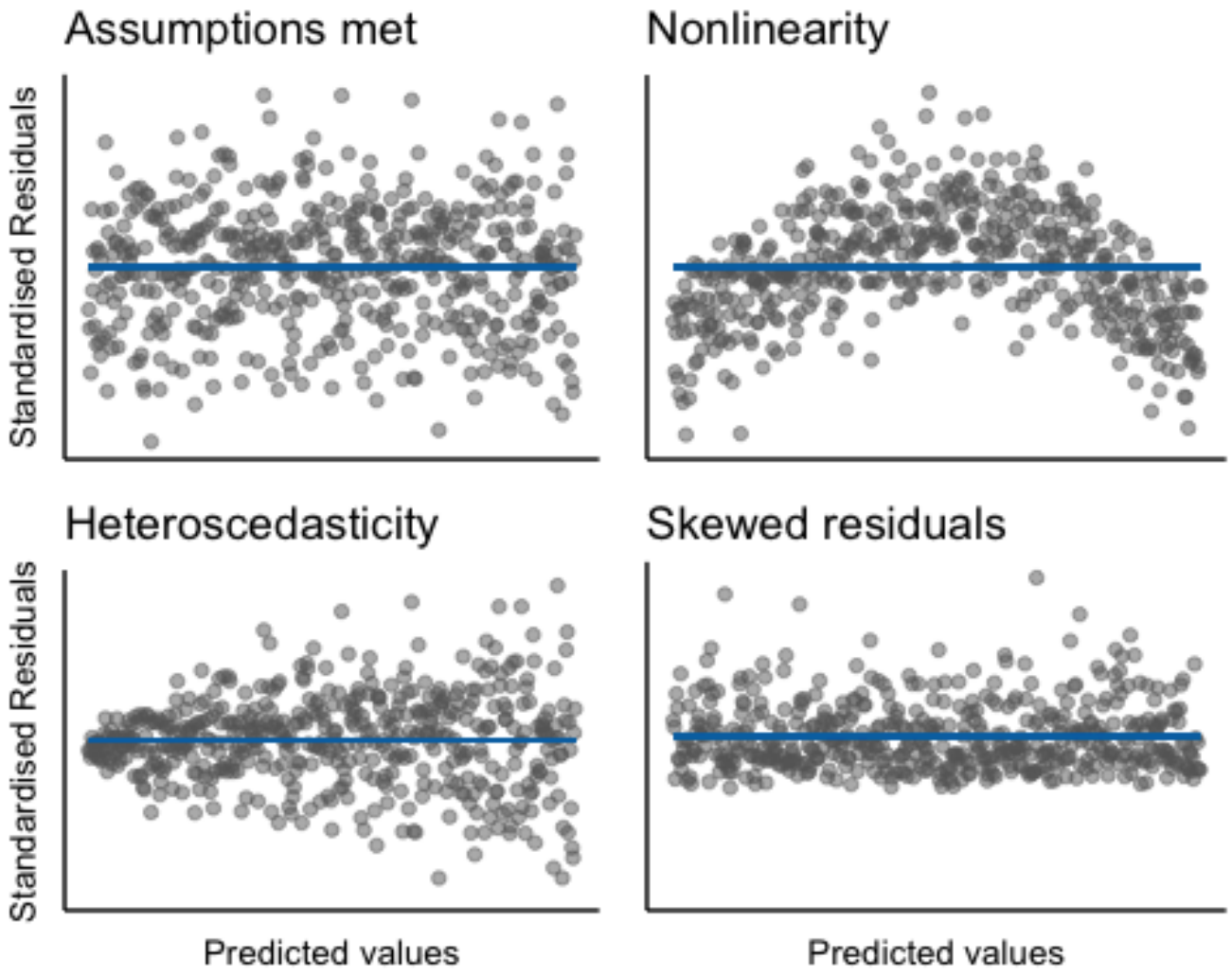
Assessing normal q-q plots



Assessing homoscedasticity: spread-level plot



Assessing spread-level plots



What to do if assumptions are not satisfied?

If you judge that one or several assumptions are not met, you should not trust your model. In particular, significance tests and confidence intervals, although your software will conduct and print them without giving you a warning, may be incorrect and conclusions drawn from them may be misleading.

Remedies depend on which assumption is violated. This is a big topic and beyond the scope of this lecture. One frequent strategy is given on the next slide.

Variable transformations

Non-normality and non-linearity can sometimes be remedied by transforming predictors and/or outcome variables by a mathematical function. For example, consider slide 7. The non-linear relationship between GDP and life expectancy. But it turns out that if we plot Life expectancy against log(GDP), the relationship is approximately linear. Transformations are an important topic, but are beyond the scope of this lecture.

Summary and final comments

Linear regression allows us to model linear relationships between a numeric outcome, or dependent variable, and one or several numeric predictors, or independent variables. Categorical predictors can be accommodated via the use of dummy variables.

Your statistical software of choice will carry out an ordinary least squares regression, calculate coefficient estimates and other statistics, and conduct hypothesis tests about coefficients. However, inferences about models are only valid if model assumptions (such as linearity, normality of errors, and homoscedasticity of errors) are met. The software will not usually give a warning in cases where assumptions are not met, or when the model does not make sense. It is your responsibility to investigate model assumptions and ensure that the results you report are based on a plausible model.

Further reading

Good short introduction to correlation and regression:

Bewick V, Cheek L & Ball J (2003) Statistics Review 7: Correlation and regression. Critical Care 7: 451 – 459. Available online: http://ccforum.com/content/7/6/451 .

Thorough textbook on multivariate data analysis, including regression:

Tabachnik BG & Fidell LS (2013) Using multivariate statistics. 6th edition. Pearson.

Excellent statistics textbook, including good chapters on regression:

Howell DC (2013) Statistical methods for psychology. 8th edition. Cengage.

For a non-mathematical explanation of correlation, consider:

Rowntree, D (2000) Statistics without tears. A primer for non-mathematicians. Chapter 8: Analyzing relationships. Allyn & Bacon.

Learning resources from the UCLA Institute of Digital Research and Education

Regression with STATA

Chapter 1:

https://stats.idre.ucla.edu/stata/webbooks/reg/chapter1/regressionwith-statachapter-1-simple-and-multiple-regression/

Chapter 2:

https://stats.idre.ucla.edu/stata/webbooks/reg/chapter2/stata-webbooksregressionwith-statachapter-2-regression-diagnostics/

Regression with SPSS

Chapter 1:

https://stats.idre.ucla.edu/spss/seminars/introduction-to-regression-with-spss/introreg-lesson1/

Chapter 2:

https://stats.idre.ucla.edu/spss/webbooks/reg/chapter2/spss-webbooksregressionwith-spsschapter-2-regression-diagnostics/

References

Bryan, J. (2015). gapminder: Data from Gapminder. R package. Retrieved from https://cran.r-project.org/package=gapminder

Elliott, J., Gale, C. R., Parsons, S., & Kuh, D. (2014). Neighbourhood cohesion and mental wellbeing among older adults: A mixed methods approach. *Social Science and Medicine*, *107*, 44–51. https://doi.org/10.1016/j.socscimed.2014.02.027

Galton, F. (1886). Regression towards mediocrity in hereditary stature. *The Journal of the Anthropological Institute of Great Britain and Ireland*, *15*, 246–263.

Martorano, B., Natali, L., De Neubourg, C., & Bradshaw, J. (2014). *Child well-being in advanced economies in the late 2000s*. *UNICEF Office of Research Working Paper WP-2013-01* (Vol. 118). Florence. https://doi.org/10.1007/s11205-013-0402-z

Pickett, K. E., & Wilkinson, R. G. (2007). Child wellbeing and income inequality in rich societies: ecological cross sectional study. *British Medical Journal*, *335*, 1080–1084. https://doi.org/10.1136/bmj.39377.580162.55

Revelle, W. (2015). psych: Procedures for Personality and Psychological Research. Evanston, Illinois: Northwestern University. Retrieved from http://cran.r-project.org/package=psych