Applied Statistics

Chiara Seghieri,
Laboratorio MeS
Istituto di Management
Scuola Superiore Sant'Anna
c.seghieri@santannapisa.it

11-12 Feb 2021

Introduction

- We will model the relationship between a set of variables x_s and a single variable y.
- The motivation for using the technique:
  - Analyze the specific relationships between the variables x and the y.
  - Forecast the value of y from the values of the variables x_1, x_2,…$x_{k...}$

Regression model

Relation between variables where changes in some variables may "explain" changes in other variables.

Explanatory variables (X_1, X_2, X_3,…) are termed the **independent** variables and the variable to be explained is termed the **dependent** variable (Y).

We can describe how variables are related using a mathematical function. The function along with other assumptions is called a model. Regression model estimates the nature of the relationship between the independent and dependent variables.

Size of the relationship.
Strength of the relationship.
Statistical significance of the relationship.

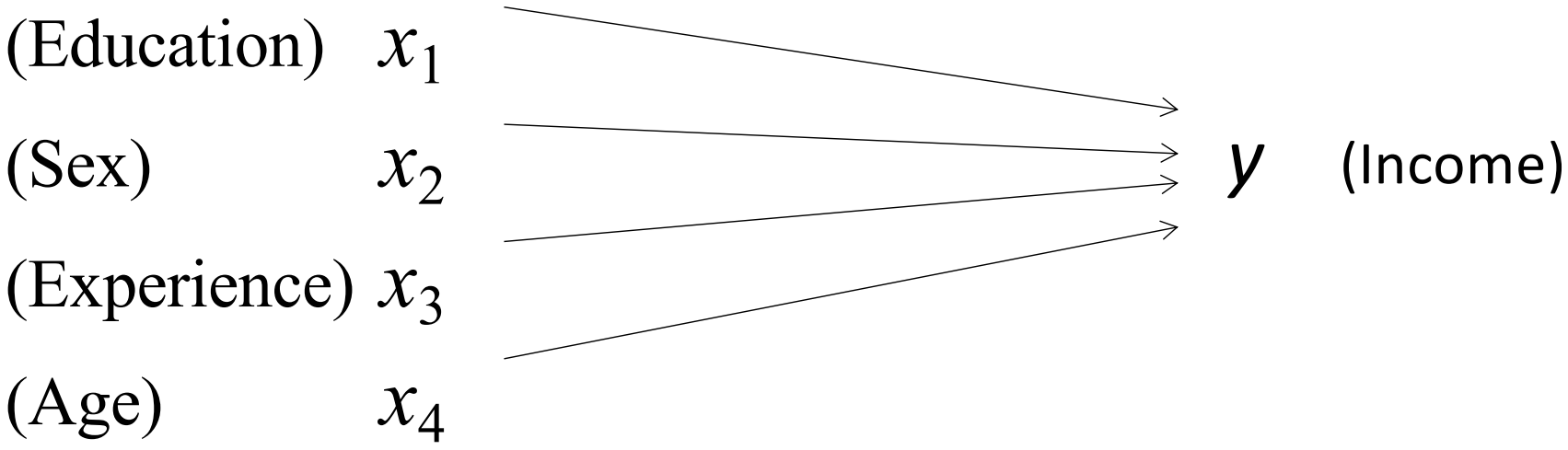
Simple and multivariate models

Bivariate or simple regression model

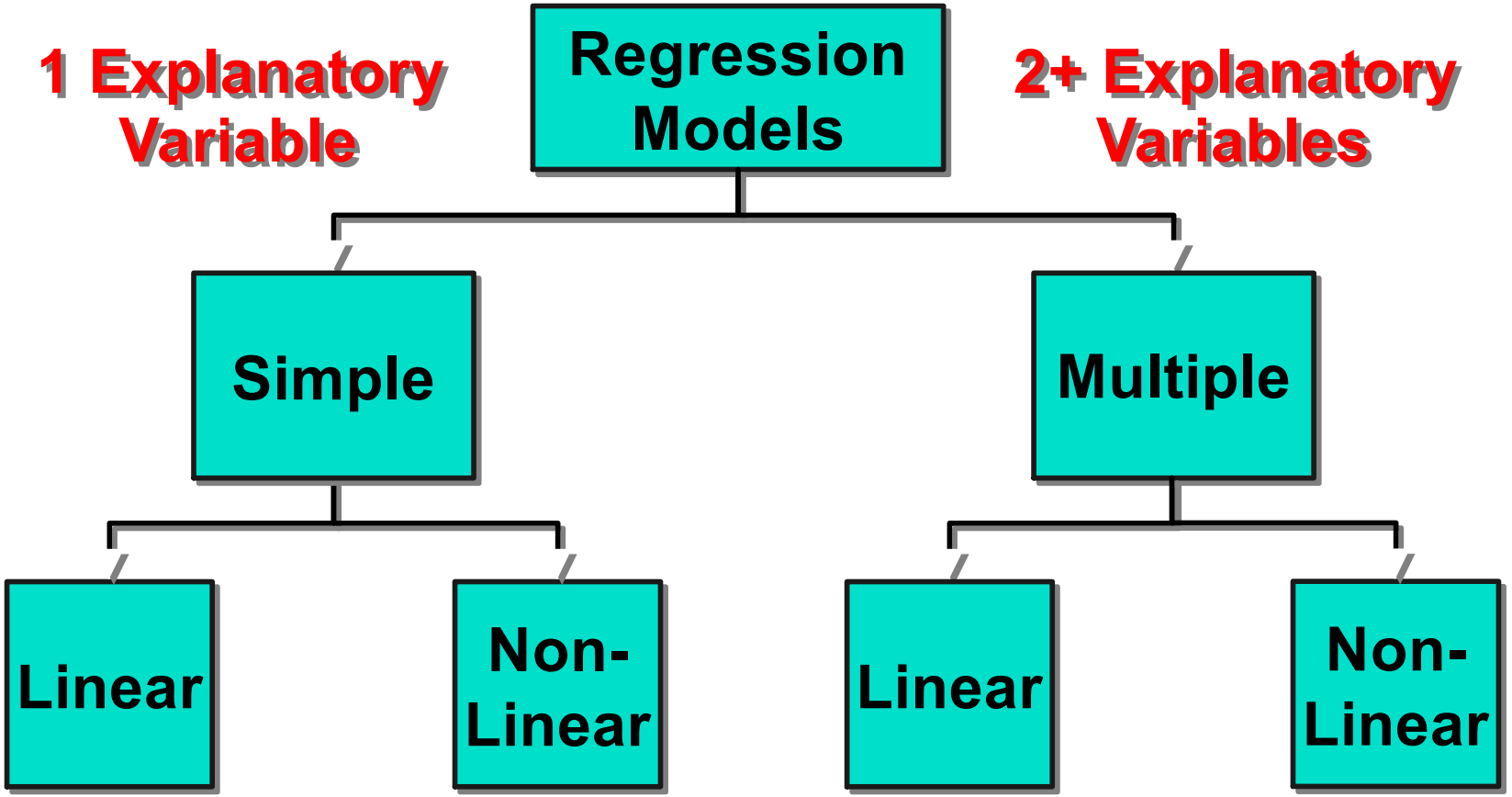
(Education) x ⟶ y

(Income)

Multivariate or multiple regression model



Types of Regression Models

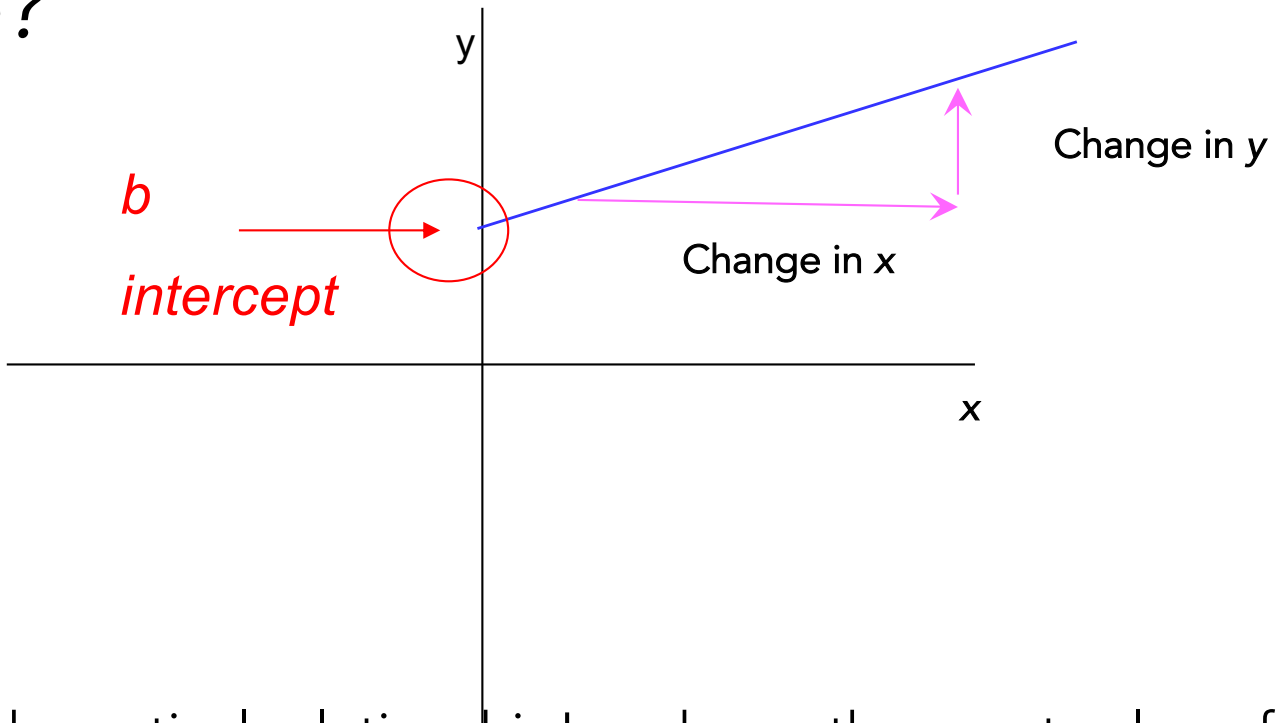


Simple Linear Regression Model

What is "Linear"?

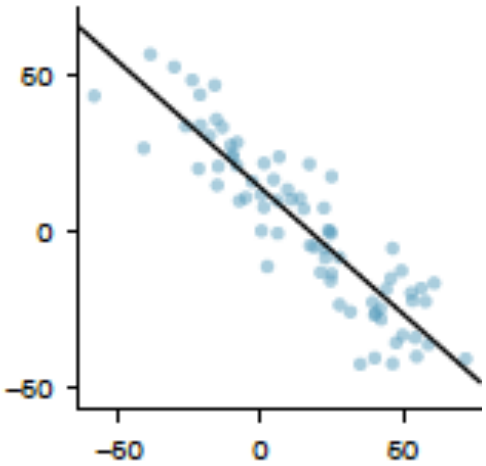
- Remember this:
- *Y=mx+b?*

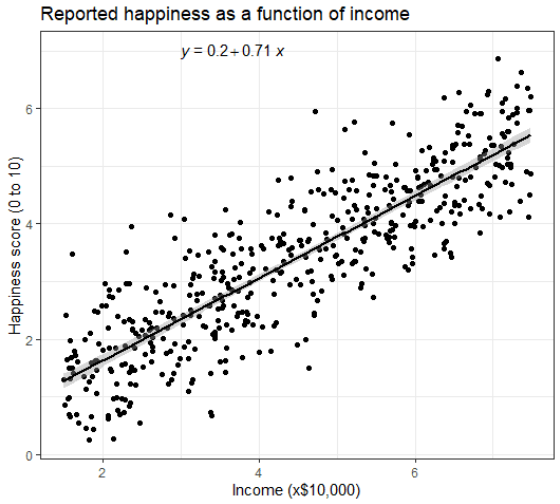
m=slope=change in y/change in x



It is a deterministic mathematical relationship! we know the exact value of y just by knowing the value of x. This is unrealistic in almost any natural process!

Generally social & real world data do not fall on a straight line. For example, if we took family income (x), this value would provide some useful information about food expenditures of a family (y). However, the prediction would be far from perfect, since other factors play a role in deciding the level of expenditures. It's more common for data to appear as a cloud of points.





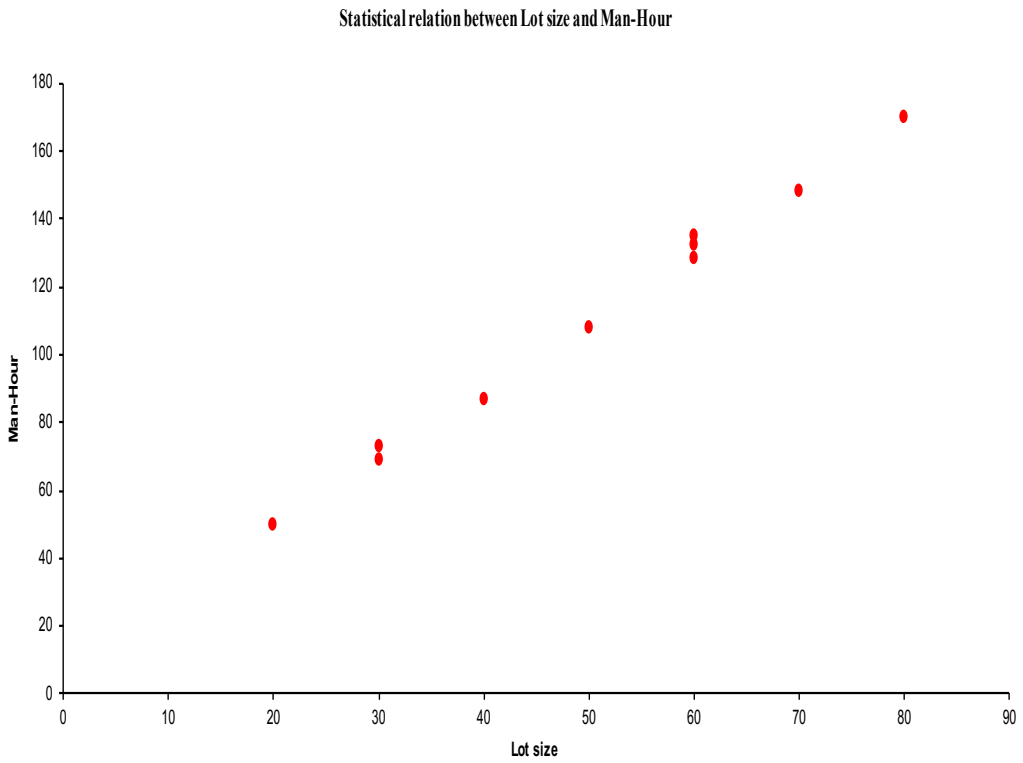
Linear regression is the statistical method for fitting a line to data where the relationship between two variables, x and y, can be modelled by a straight line with some error.

- The primary goal of regression analysis is to use current information about a phenomenon to predict its future behavior.
- Current information is usually in the form of a set of data.
- In a simple case, when the data form a set of pairs of numbers, we may interpret them as representing the observed values of an independent (or predictor) variable X and a dependent (or response) variable Y.

lot size	Man-hours
30	73
20	50
60	128
80	170
40	87
50	108
60	135
30	69
70	148
60	132

In the table on the right the response variable Y represents the man-hours of labor for manufacturing a certain product in lots (X) that vary in size as demand fluctuates.

- We are looking for functional relation
$$y = f(x)$$
between the dependent variable y and the predictor variable x.



Notice that in the graph for some values of X (X=30 and X=60) there correspond more than one value of Y. This is not an ordinary functional relationship between X and Y, where to each value of X a unique value of Y must correspond.

Question: What is a plausible way of thinking about this situation that would still lead to a model y=F(X) in which we would have a unique value y for each value of X?

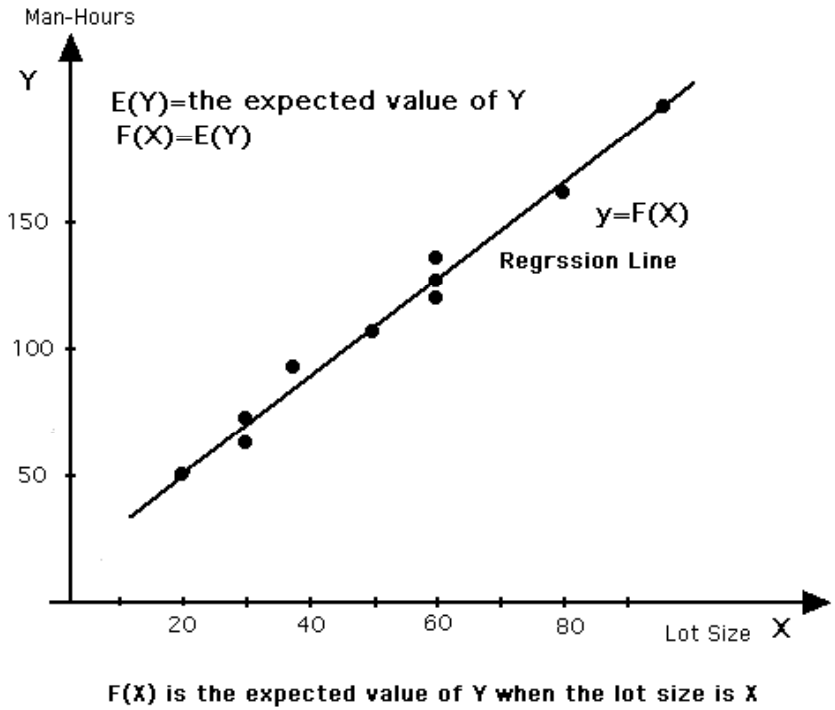
Regression Function

The idea behind this is: when we have several values of y observed for one value of X we take the average of these values of y to assign to x.

- The statement that the relation between X and Y is statistical should be interpreted as providing the following guidelines:

1. **Regard Y as a random variable.**
2. **For each X (fixed variable),** take *f* (*x*) to be the expected value (i.e., mean value) of y.
3. Given that *E* (*Y*) denotes the expected value of Y, call the equation of the regression function.

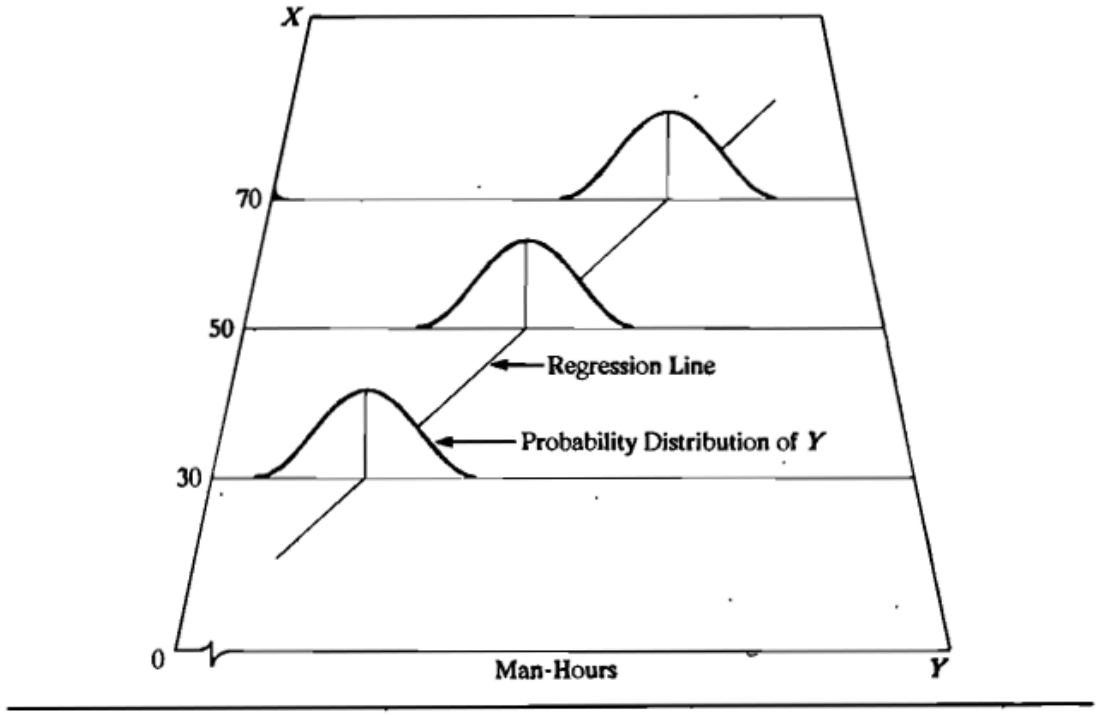
$$E(Y) = f(x)$$



F(X) is the expected value of Y when the lot size is X

So the structural model says that for each value of x the population mean of Y (over all of the subjects who have that particular value "x" for their explanatory variable) can be calculated using the simple linear expression b0 + b1x. Of course we cannot make the calculation exactly, in practice, because the two parameters are unknown therefore we make estimates of the parameters.

Representation of Linear Regression Model

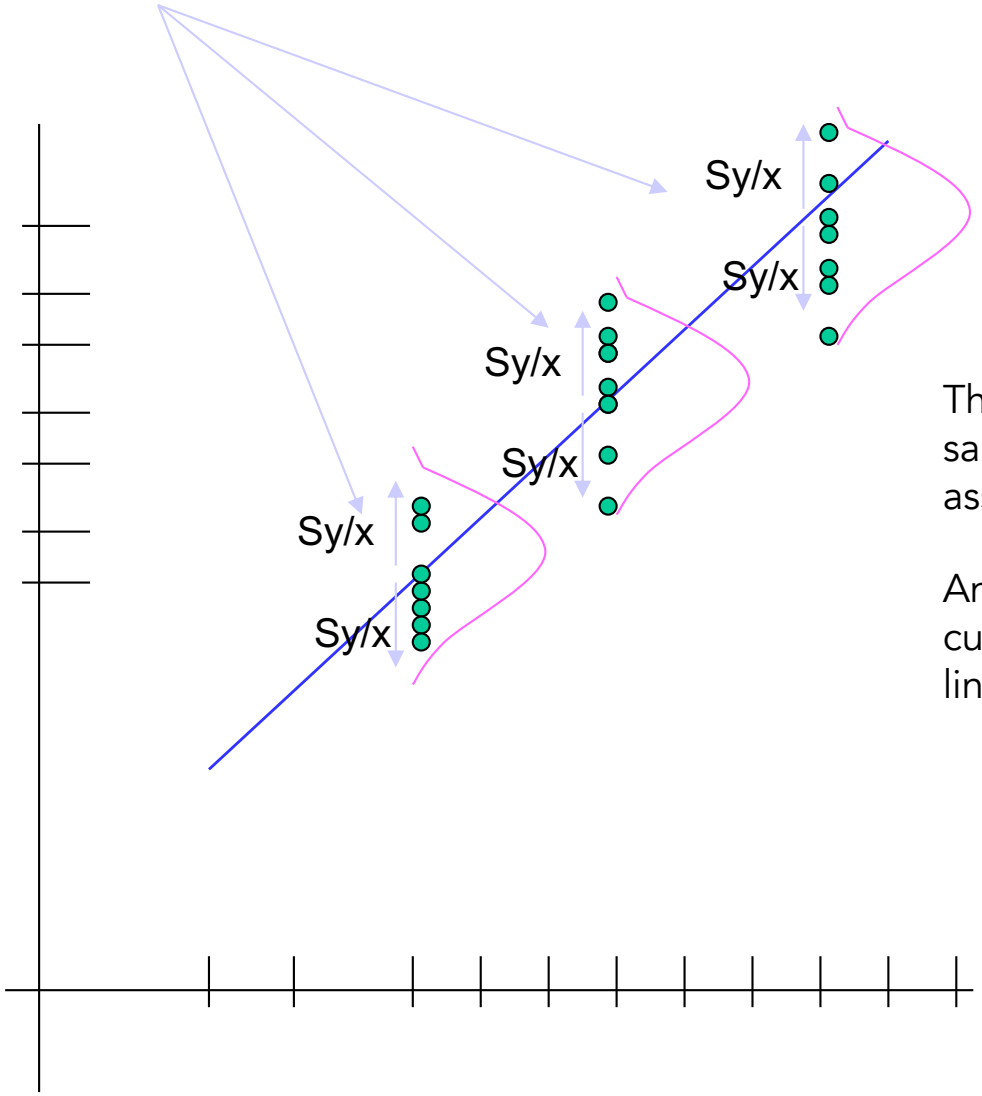


For each X, there is a probability distribution of Y. The figure shows a probability distribution for X= 30, X=50, and X=70

The number of man-hours Y is then viewed as a random selection from this probability distribution. The means of the probability distributions have a systematic relation to the level of X. This systematic relationship is called the **regression function of Y on X.** Here the regression function is linear. This implies that the mean number of man-hours varies linearly with values of X. The straight line, is the line connecting the average of the y values for each level of the independent variable, x. The actual y values for each level of x are normally distributed around the mean of y. In addition, the distribution of y is the same for each value of x, i.e. the variability is the same.

we have several values of y observed for one value of X we take the average of these values of y to assign to x.

The standard error of Y given X is the average variability around the regression line at any given value of X. It is assumed to be equal at all values of X.



The fact that the three Normal curves have the same spreads represents the equal variance assumption.

And the fact that the means of the Normal curves fall along a straight line represents the linearity assumption.

Notation

n observations (sample size).

The variable Y is the response variable, and $y_1, y_2, \ldots, y_n$ are the observed values of the response, Y.
The variable X is the predictor variable and $x_1, x_2, \ldots x_n$ are observed values of the predictor, X.

The observations are considered as coordinates, (xi,yi), for i=1,…,n.
As we saw before, the points, (x1,y1),…,(xn,yn), may not fall exactly on a line. There is some error we must consider.

The general form of the simple linear regression model is:

$$Y=\boldsymbol{\beta}_0+\boldsymbol{\beta}_1 X+\epsilon$$

For an individual observation,

$$y_i=\boldsymbol{\beta}_0+\boldsymbol{\beta}_1 x_i+\epsilon_i$$

where,
$\boldsymbol{\beta}_0$ is the population y-intercept,
$\boldsymbol{\beta}_1$ is the population slope, and
ϵ_i is the error or deviation of y_i from the line, $\boldsymbol{\beta}_0+\boldsymbol{\beta}_1 x_i$.

To make inferences about these unknown population parameters, we must find an estimate for them. There are different ways to estimate the parameters from the sample. In this class, we will present the least squares method.

Once the parameters are estimated, we have the **least square regression equation line (or the estimated regression line)**. We can also use the least squares regression line to estimate the errors, called residuals.
$\widehat{\epsilon_i}= y_i - \hat{y}_i$.

The Simple Linear Regression

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

$$\varepsilon_i \sim N(o, \sigma^2) \quad COV\left(\varepsilon_i \varepsilon_j\right) = 0$$

that means: any variations in Y that are not explained by the X's are independent and identically normally distributed.

The expected value of *Y* at each level of *x* is:

$$E(Y) = \beta_0 + \beta_1 X$$

β_1 is the coefficients that describe the size of the effect the independent variable is having on your dependent variable *Y*, and β_0 is the value *Y* is predicted to have when all the independent variables are equal to zero.

Assumptions

Main assumptions:

- Linearity: the response can be written as a linear combination of the predictors.
- Independence: the errors are independent.
- Normality: the distribution of the errors should follow a normal distribution.
- Equal Variance: the error variance is the same at any set of predictor values.

The linearity assumption is encoded as

$$\boldsymbol{\beta}_0+\boldsymbol{\beta}_1 x_{i1}+\boldsymbol{\beta}_2 x_{i2}+\cdots+\boldsymbol{\beta}_{p-1} x_{i(p-1)}$$

while the remaining three, are all encoded in

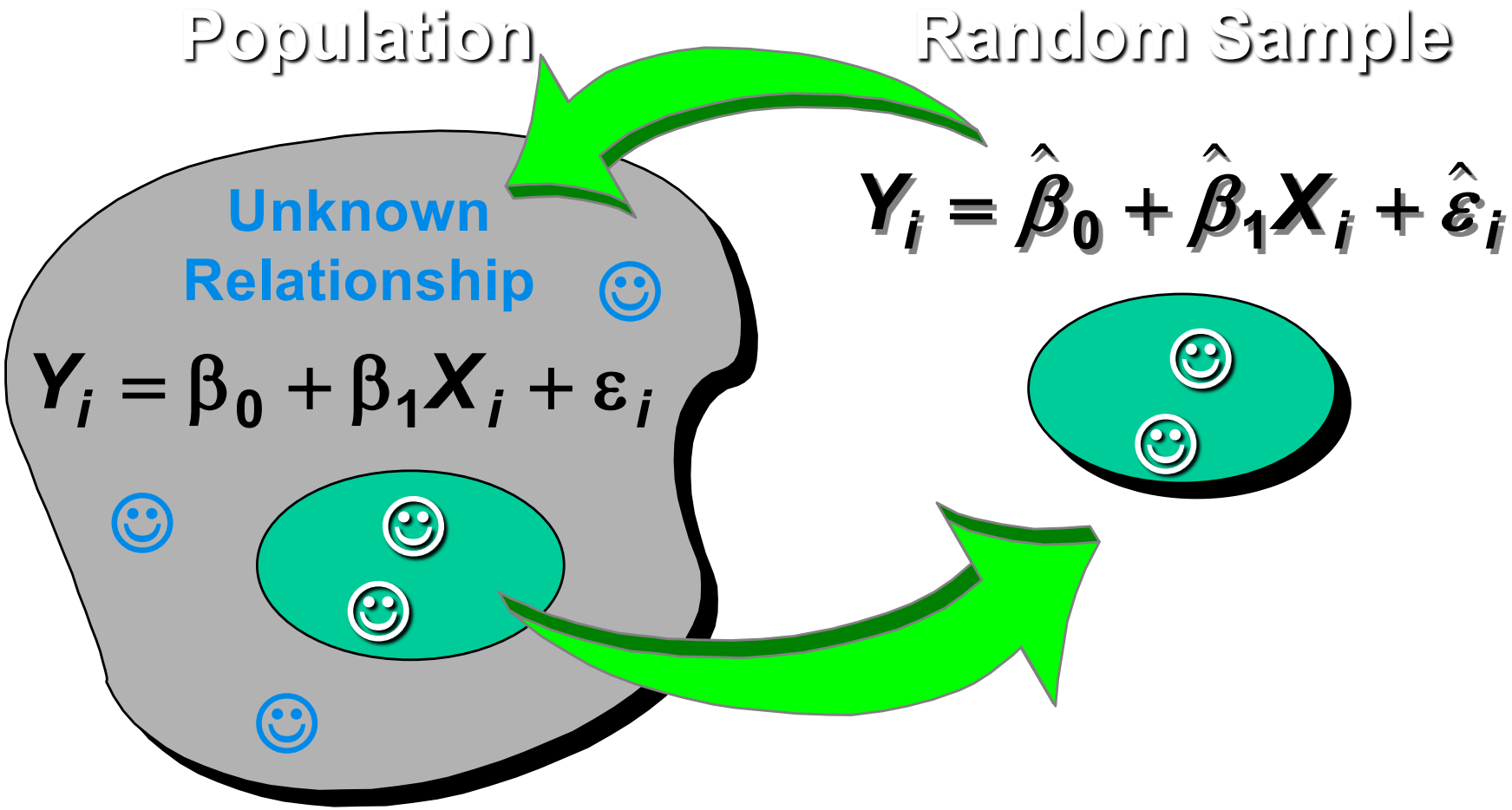
$\epsilon i \sim N(0,\boldsymbol{\sigma}^2)$,since the ϵi are iid normal random variables with constant variance

The linear model

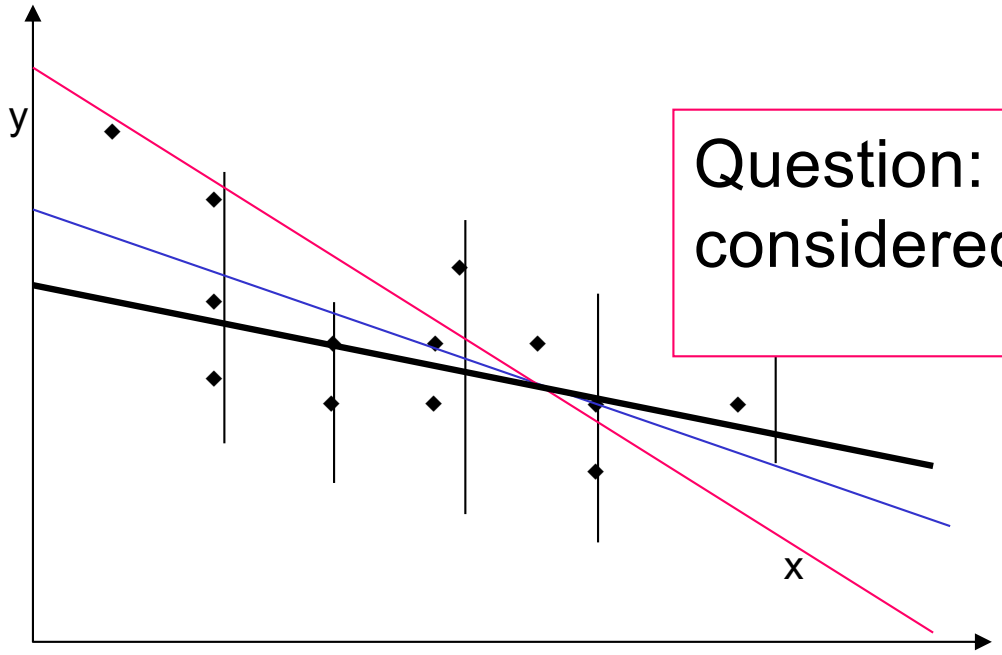
By knowing this equation we can estimate values of y for a given value of x through **the estimation of the coefficients β_0, and β_1**

- Since the estimates are made based on the sample and not the entire population, the estimate will not be perfect, there will be residuals or errors

Population & Sample Regression Models



Estimating the Coefficients



Question: What should be considered a good line?

The Least Squares (Regression) Line

A good line is one that minimizes the sum of squared differences between the points and the line.

It says that we should choose as the best-fit line, that line which minimizes the sum of the squared residuals, where the residuals are the vertical distances from individual points to the best-fit "regression" line.

Least Squares

'Best Fit' Means Difference Between Actual Y Values & Predicted Y Values is a Minimum. *In particular, we use* square errors

$$\sum_{i=1}^{n} \left(Y_i - \hat{Y}_i \right)^2 = \sum_{i=1}^{n} \hat{\varepsilon}_i^2$$

Coefficient Equations

LS minimize:

$$\sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} \left(y_i - \beta_0 - \beta_1 x_i \right)^2$$

- Sample slope

$$\hat{\beta}_1 = \frac{SS_{xy}}{SS_{xx}} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2}$$

- Sample Y - intercept

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

Best Linear Unbiased Estimate (BLUE)

If the following assumptions are met:

- The Model is
  - Linear
  - Additive
- The regression error term is
  - normally distributed
  - has an expected value of 0
  - errors are independent
  - homoscedasticity

Characteristics of OLS if sample is probability sample

- Unbiased
- Efficient
- Consistent

The Three Desirable Characteristics

- Unbiased:
- $E(\hat{\beta})=\beta$
  - On the average we are on target
- Efficient
  - Standard error will be minimum
- Consistent
  - As N increases the standard error decreases and closes in on the population value

Remember: Efficiency refers to how stable a statistic is from one sample to the next. A more efficient statistic has less variability from sample to sample; it is therefore, on average, more precise.

Interpretation of Coefficients

Interpretation of Coefficients

1. Slope ($\hat{\beta}_1$)
   - Estimated change (increase or decrease) of *Y* for Each 1 Unit Increase in *X*
     - If $\hat{\beta}_1 = 2$, then on average increase by 2 for Each 1 Unit Increase in *X*

Interpretation of Coefficients

2. Y-Intercept ($\hat{\beta}_0$)
   - Average Value of *Y* When *X* = 0 (when it makes sense that X=0)
     - *if* $\hat{\beta}_0$ = 4, then Average *Y* Is Expected to Be 4 When *X* Is 0

Testing the Slope

- We can draw inference by testing:

  H_0: $\beta_1 = 0$

  H_1: $\beta_1 \neq 0$ (or < 0, or > 0)

The model in STATA

Sample: 20 cities in US; Y=homicide rate, X=% of families below the poverti line

reg homic poor

Source	SS	df	MS
Model	181.370325	1	181.370325
Residual	531.573154	18	29.5318419
Total	712.943479	19	37.523341

Number of obs	= 20
F(1, 18)	= 6.14
Prob > F	= 0.0233
R-squared	= 0.2544
Adj R-squared	= 0.2130
Root MSE	= 5.4343

homic	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
poor	.9438495	.3808596	2.48	0.023	.1436932	1.744006
_cons	-.8151891	3.344025	-0.24	0.810	-7.840726	6.210348

The regression model is:

Homicide rate = - 0.82 + 0.94 (poor families)

Interpretation and significance of the coefficients [1]

- The average city homicide rates rise by 0.94 with each 1-point increase in the percentage of families below poverty

- The constant estimate implies that the average homicide rate should equal –0.8 in cities with 0 percent below poverty.

That interpretation makes no sense, because we have no cities without poverty. Despite the constant term is important for providing simply interpretation of the regression output, the regression line may yield unreasonable results when projected beyond the X range of the data.

Interpretation and significance of the coefficients [2]

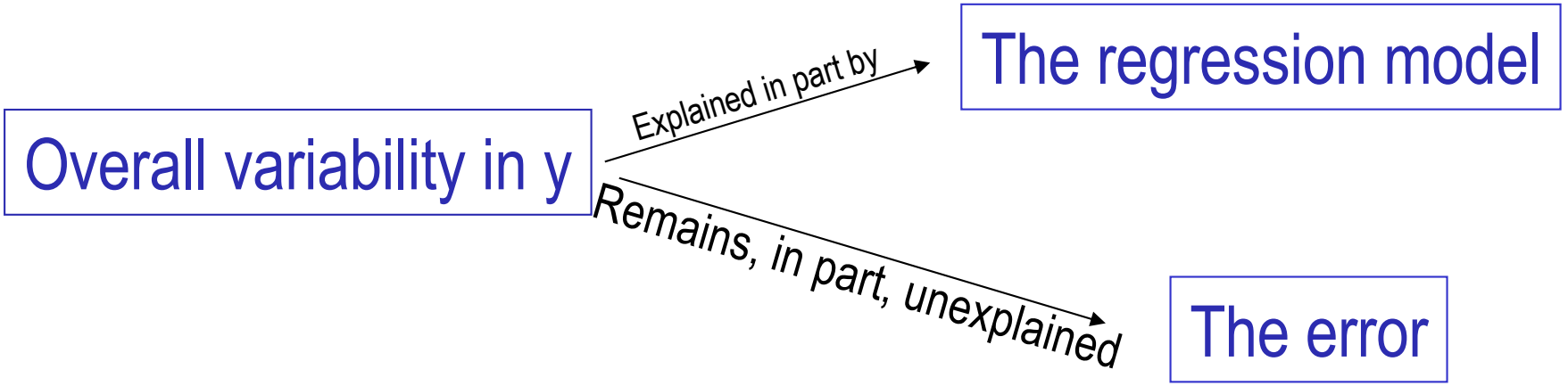
- t test: it verifies the significance of each single parameter estimate. It is based on the two hypotheses:

$$H0: \beta=0 \text{ versus } H1: \beta\neq0$$

→ each coefficient is significantly different from 0.

- P>|t| is the P-value, the null hypothesis is rejected if the P-value is lower than the chosen size (5%). In doing so, we make an error less than 5 times over 100. In this case, the coefficient of β is statistically significant in explaining the city homicide rates (P-value of 'poor' 0.023 < 0.05)

The fit of the model



The fit of the model

- A measure of the goodness of the regression model is the R-squared or the coefficient of determination of the regression: *R2*
- R2 measures the fraction of the variance of Y that is explained by X; it is unitless and ranges between zero (no fit) and one (perfect fit).
- R-squared = 0.2544 means that the regression model explains about 25% of the variation in the Y.
- Prob > F = 0.0233 : p-value of the model. It tests the overall significance of the model, whether R2 is different from 0. (p-value lower than 0.05 shows a statistically significant relationship between X and Y)

Multiple Linear Regression

More than one predictor…

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \ldots + \beta_k x_k + \varepsilon$$

Additive (Effect) Assumption: The **expected change in y** per unit increment in x_j is constant and does not depend on the value of any other predictor. This change in y is equal to b_j.

That is, the amount of change in the outcome variable that would be expected per one unit change of the predictor, if all other variables in the model were held constant.

Standardized Regression Coefficients

- Regression slopes depends on the units of the independent variables
- How do you compare how "strong" the effects of two variables if they have totally different units?
- Example: Education, health status, income
  - Education measured in years, b = 2.5
  - Health status measured on 1-5 scale, b = .18
  - Which is a "bigger" effect? Units aren't comparable

"standardized" coefficients

Standardized Regression Coefficients

Standardized Coefficients called "Betas" or Beta Weights" (is equivalent to Z-scoring all independent variables before doing the regression)

$$\beta_j^* = \left(\frac{s_{X_j}}{s_Y} \right) b_j$$

The unit is standard deviations and Betas indicate the effect a 1 standard deviation change in X_j on Y (an increase of 1 standard deviation in X results in a b standard deviation increase in Y)

Example:

Sample of 20 HHs, food consumption (*Y*) HH income (X_1).
The estimated model is

Y= expenditures * 1000 euros

X_1=HH income * 1000 euros

X_2=HH size

$$\hat{Y} = -1{,}118 + 0{,}148\,X_1 + 0{,}793X_2$$

b1 = 0,148: on average, consumption expend. increase, of **148** Euros each year for an increase of **1000 Euros of the income**, holding X2 fixed

b_2 = 0,793: on average, consumption expend. increase of **793** Euros yearly for an additional component in the HH, holding X1 fixed

Standardized coefficients

$$\hat{Y} = 0.761 X_1 + 0.272 X_2$$

Which variable is contributing more to explain the food expenditures?

How to make a prediction:

Etsimate Y for a family with HH income 90000 € and HHsize = 5

$$\hat{Y} = -1.118 + 0.148(X1) + 0.793(X2)$$
$$= -1.118 + 0.148 \times 90 + 0.793 \times 5$$
$$= 16.167$$

Predicted expenditure 16167 Euro

BE CAREFUL: HH income is in €*1000, therefore X1= 90

Dummy Variables

"Dummy" = a dichotomous variables coded to indicate the presence (1) or absence (0) of something.

First, create a separate dummy variable for **all** categories

- Ex: Gender – make female & male variables
  - DFEMALE: coded as 1 for all women, zero for men
  - DMALE: coded as 1 for all men

Then: Include **all but one** dummy variables into a multiple regression model

- If two dummies, include 1; If 5 dummies, include 4.

Dummy Variables: Interpretation

Example:

$$Y_i = a + b_1 AGE_i + b_2 DFEMALE_i + e_i$$

- What if the case *i* is a male?
- DFEMALE is 0 in case of male, so males are modeled as:

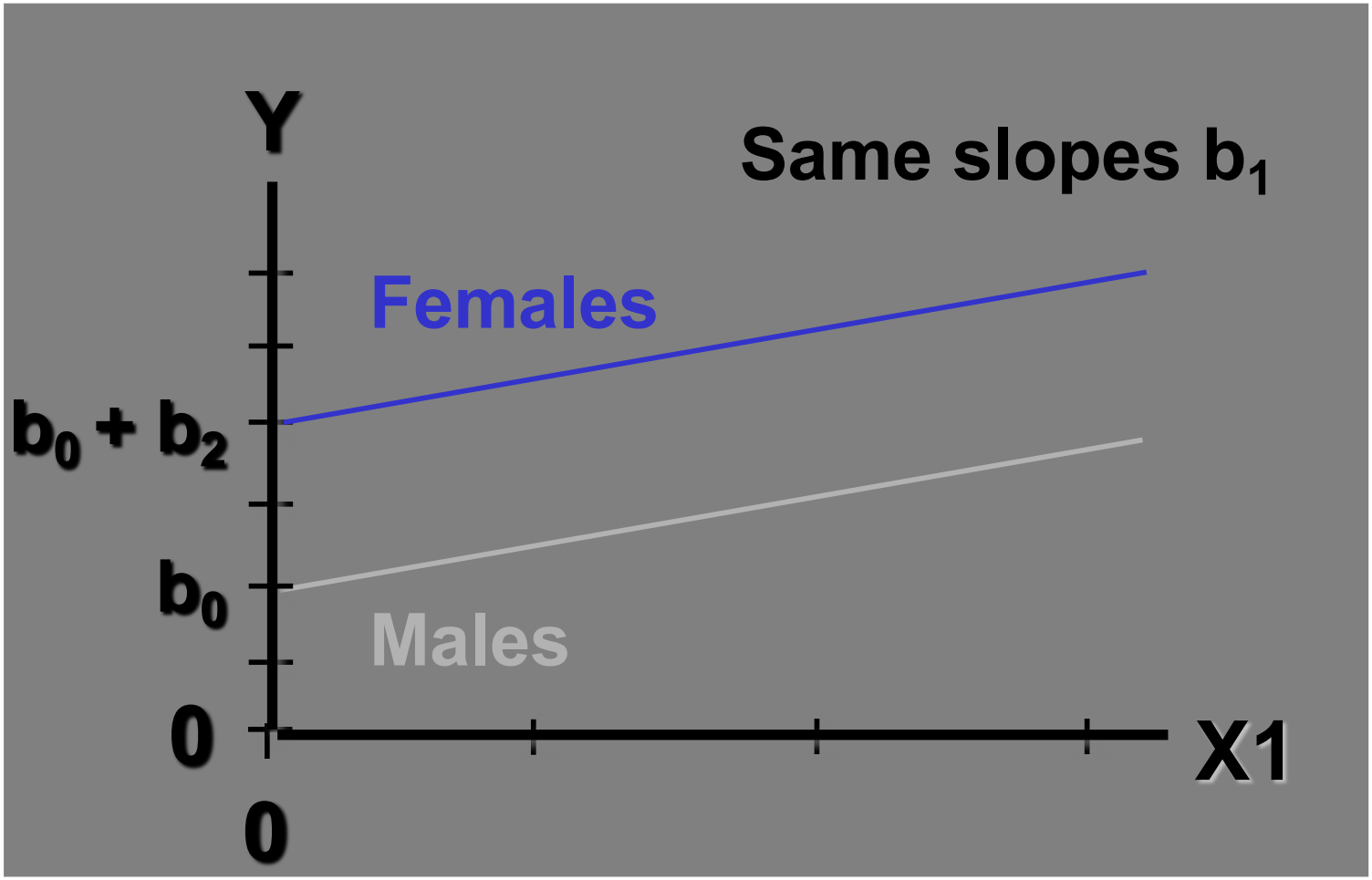
$a + b_1 AGE + e.$

Dummy Variables: Interpretation

$$Y_i = a + b_1 AGE_i + b_2 DFEMALE_i + e_i$$

- What if the case *i* is a female?
- DFEMALE=1 and so females are modeled using a different regression line: $(a+b_2) + b_1$AGE + e

  – Thus, the coefficient of b_2 reflects difference in the **constant** for women.



a different constant generates a different line, either higher or lower. A positive coefficient (b) indicates that women are consistently higher compared to men (on dep. var.). A negative coefficient indicated women are lower

Dummy Variables: Interpretation

A positive coefficient (b) indicates that women are consistently higher compared to men (on dep. var.)

- A negative coefficient indicated women are lower

- Example: If DFEMALE coeff = 1.2:

  "Women are on average 1.2 points higher than men".

Dummy Variables

- What if you want to compare more than 2 groups?
- Example: Race
  - Coded 1=white, 2=black, 3=other
- Make 3 dummy variables and then, include **two** of the three variables in the multiple regression model.
- The contrast is **always** with the category that was **left out** of the equation
  - If DFEMALE is included, the contrast is with males
  - If DBLACK, DOTHER are included, coefficients reflect difference in constant compared to whites.

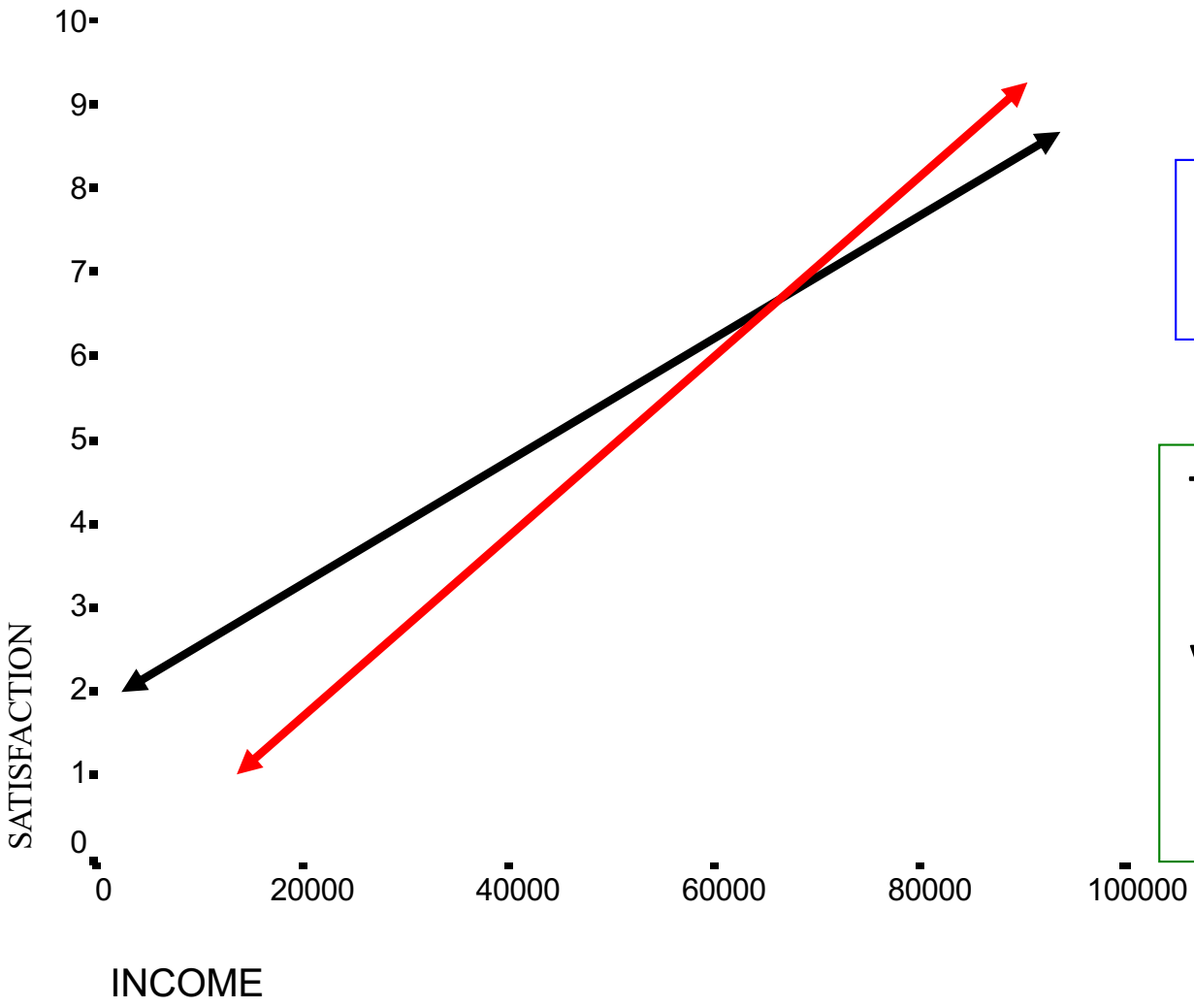
Interactions

What if a variable has a different slope for two different sub-groups in your data?

- Example: Income and Satisfaction with life (Y) – gender
  - Perhaps for men an extra euro increases their satisfaction a lot
  - Whereas for women each euro has a smaller effect on satisfaction (compared to men)

The slope of a variable (income) might differ across groups

Interaction



the slope for men and women differs.

The effect of income on satisfaction (X1 on Y) varies with gender (X2). This is called an "interaction effect"

Interaction: example

- Examples of interaction:
  - Effect of education on income may interact with type of school attended (public vs. private)
    - Private schooling has bigger effect on income
  - Effect of aspirations on educational attainment interacts with poverty
    - Aspirations matter less if you don't have money to pay for college

Interaction terms are needed whenever there is reason to believe that the effect of one independent variable depends on the value of another independent variable

Interaction effect:

- Interaction effects: differences in the relationship (slope) between two variables for each category of a **third variable**
- Option #1: Analyze each group separately (stratify)
    - Look for different slope in each group
- Option #2: Multiply the two variables of interest: (DFEMALE, INCOME) to create a new variable
  - Called: DFEMALE*INCOME
  - Add that variable to the multiple regression model.

Interaction Term in the regression

Example, Y is satisfaction

$$Y_i = a + b_1 INCOME_i + b_2 DFEM * INC_i + e_i$$

if the case *i* is male:

DFEMALE is 0, so b_2(DFEM*INC)=0 and males are modeled using the regression equation:

a + b_1X + e.

$$Y_i = a + b_1 INCOME_i + b_2 DFEM * INC_i + e_i$$

if the case i is female

DFEMALE is 1, so b_2(DFEM*INC) becomes b_2*INCOME, which is added to b_1

Females are then modeled using a different regression line: a + (b_1+b_2) X + e

- Thus, the coefficient of b_2 reflects difference in the **slope** of INCOME for women.

Interpreting Interaction Terms

- Interpreting interaction terms:
- A positive b for DFEMALE*INCOME indicates the slope for income is higher for women vs. men
  - A negative effect indicates the slope is lower
  - Size of coefficient indicates actual difference in slope
- Example: DFEMALE*INCOME, Coefficient = -.58 indicates that the slope of satisfaction and income is .58 points lower for females than for males

Continuous Interaction

- Two continuous variables can also interact
- Example: Effect of education and income on subjective well being

- Multiply Education and Income to create the interaction term "EDUCATION*INCOME"
    - And add it to the model.

Interpreting Interaction

Example: EDUCATION*INCOME: Coefficient = 2.0:

- For each unit change in education, the slope of income – subj wellbeing increases by 2
  - Note: coefficient is symmetrical: For each unit change in income, education slope increases by 2
- Dummy interactions effectively estimate 2 slopes: one for each group. Continuous interactions result in many slopes: Each value of education*income yields a different slope.

Dummy Interactions

- It is also possible to construct interaction terms based on two dummy variables
  - Instead of a "slope" interaction, dummy interactions show difference in **constants**
    - Constant differs across values of a third variable
  - Example: Effect of race on health varies by gender
    - Black have a worse health; but the difference is much larger for black males.

Dummy Interactions

- Strategy for dummy interaction is the same: Multiply both variables
  - Example: Multiply DBLACK, DMALE to create DBLACK*DMALE
    - Then, include all 3 variables in the model
  - Effect of DBLACK*DMALE reflects difference in constant (level) for black males, compared to white males and black females
    - You would observe a negative coefficient, indicating that black males have a worse health than black females or white males.

Interaction Terms: Remarks

If you make an interaction you should also include the component variables in the model:

- In general a model with "DFEMALE * INCOME" should also include DFEMALE and INCOME

Sometimes interaction terms are highly correlated with its components

- That can cause problems of multicollinearity

Interaction Terms:

Make sure you have enough cases in each group for your interaction terms

- Interaction terms involve estimating slopes for sub-groups (e.g., black females vs black males).
  - If you there are hardly any black females in the dataset, you can have problems

Conditions for Regression Inference

- The linear regression model, which is the basis for inference, imposes several conditions.
- We should verify these conditions before proceeding with inference.
- The conditions concern the population, but we can observe only our sample.

Regression Diagnostics

- The three conditions required for the validity of the regression analysis are:
  - the error variable is normally distributed.
  - the error variance is constant for all values of x.
  - The errors are independent of each other.
- How can we diagnose violations of these conditions?

Residuals

- The difference between the observed value y_i and the corresponding fitted $\hat{y}_i$ value.
- A residual is the deviation of an outcome from the predicated mean value for all subjects with the same value for the explanatory variable.
- Residuals are highly useful for studying whether a given regression model is appropriate for the data at hand.

Regression Diagnostics

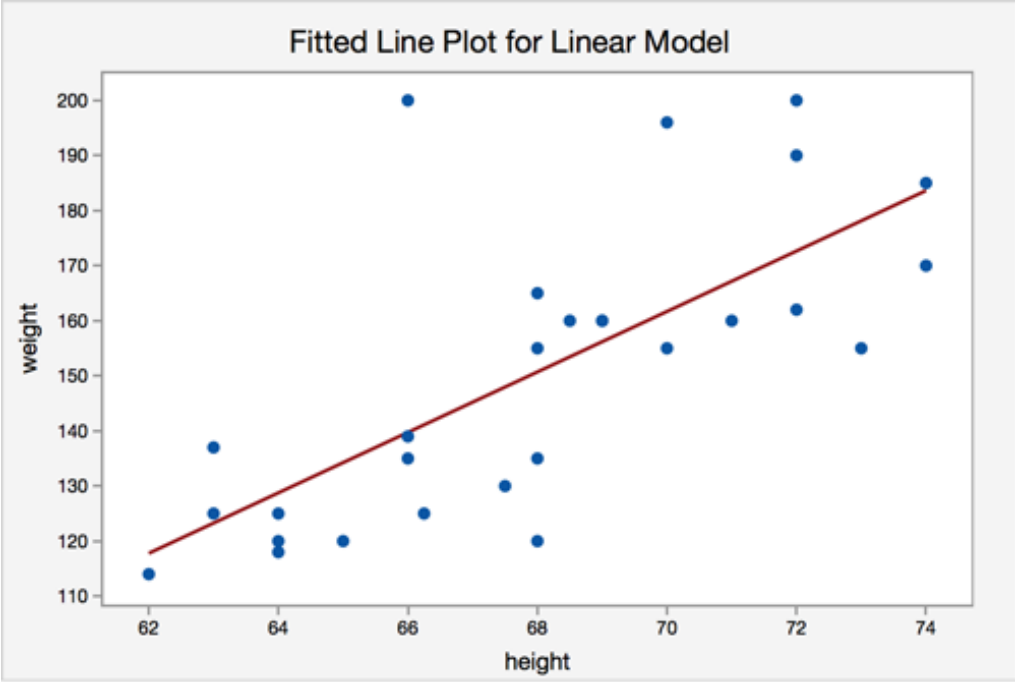
How can we diagnose violations of these conditions?

➔ **Residual Analysis**, that is, examine the ***differences*** between the actual data points and those predicted by the linear equation.

➔ A plot of all residuals on the y-axis vs. the predicted values on the x-axis, called a residual vs. fit plot, is a good way to check the linearity and equal variance assumptions.

➔ A quantile-normal plot of all of the residuals is a good way to check the Normality assumption.

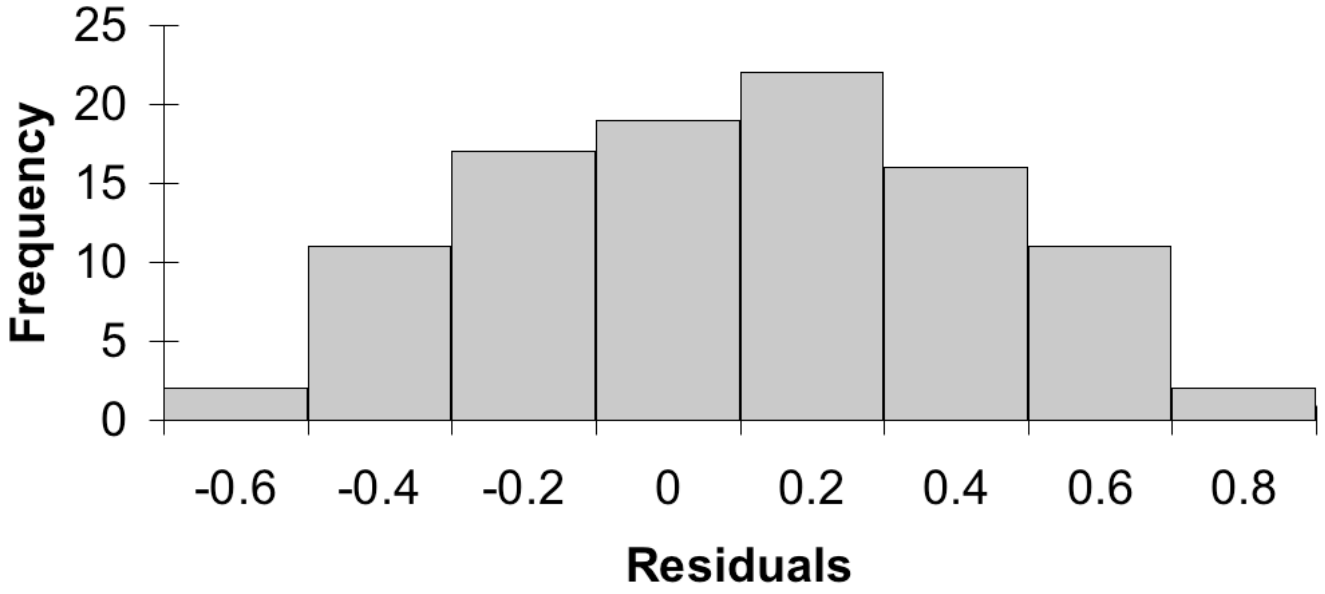
Don’t forget to also check for the Assumption of Linearity
- The relationship between height and weight must be linear.



The scatterplot shows that, in general, as height increases, weight increases. There does not appear to be any clear violation that the relationship is not linear. Also, for a quick assessment of nonlinearity: plot of observed versus predicted values or a plot of residuals versus predicted values. The points should be symmetrically distributed around a diagonal line in the former plot or around horizontal line in the latter plot.

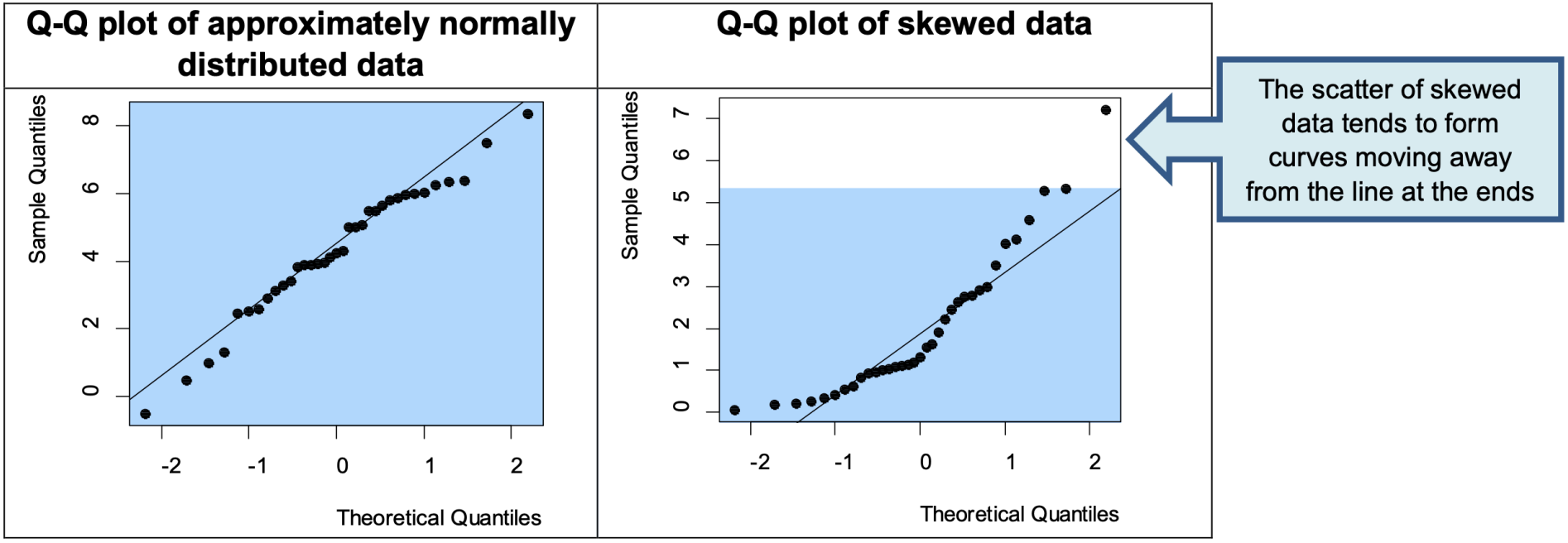
Nonnormality

We can take the residuals and put them into a histogram to visually check for normality…



…we're looking for a bell shaped histogram with the mean close to zero.

The **Q-Q plot** is an alternative graphical method of assessing normality to the histogram and is easier to use when there are small sample sizes. It compares the observed quantile with the theoretical quantile of a normal distribution. The scatter compares the data to a perfect normal distribution. The scatter should lie as close to the line as possible with no obvious pattern coming away from the line for the data to be considered normally distributed.



There are also specific test for which could be used in conjunction with either a histogram or a Q-Q plot.

The Kolmogorov-Smirnov test and the Shapiro-Wilk's W test whether the underlying distribution is normal. Both tests are sensitive to outliers and are influenced by sample size:

- For smaller samples, non-normality is less likely to be detected but the Shapiro-Wilk test should be preferred as it is generally more sensitive
- For larger samples (i.e. more than one hundred), the normality tests are conservative and the assumption of normality might be rejected too easily.

Nonnormality

The Shapiro-Wilk test for normality. It answers the question: is there enough evidence for non-normality to overthrow the null hypothesis (the null hypothesis is that the distribution of the residuals is normal). In stata the command is swilk.

```
swilk e
```

Shapiro-Wilk W test for normal data

Variable	Obs	W	V	z	Prob>z
e	50	0.95566	2.085	1.567	0.05855

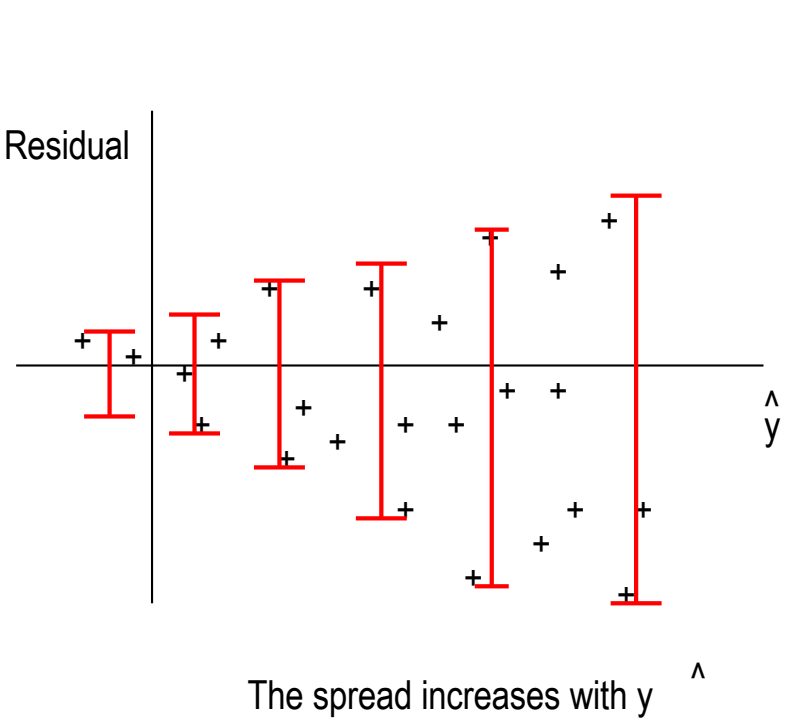
Regression Inference is robust against moderate lack of Normality. On the other hand, outliers and influential observations can invalidate the results of inference for regression. What to do?

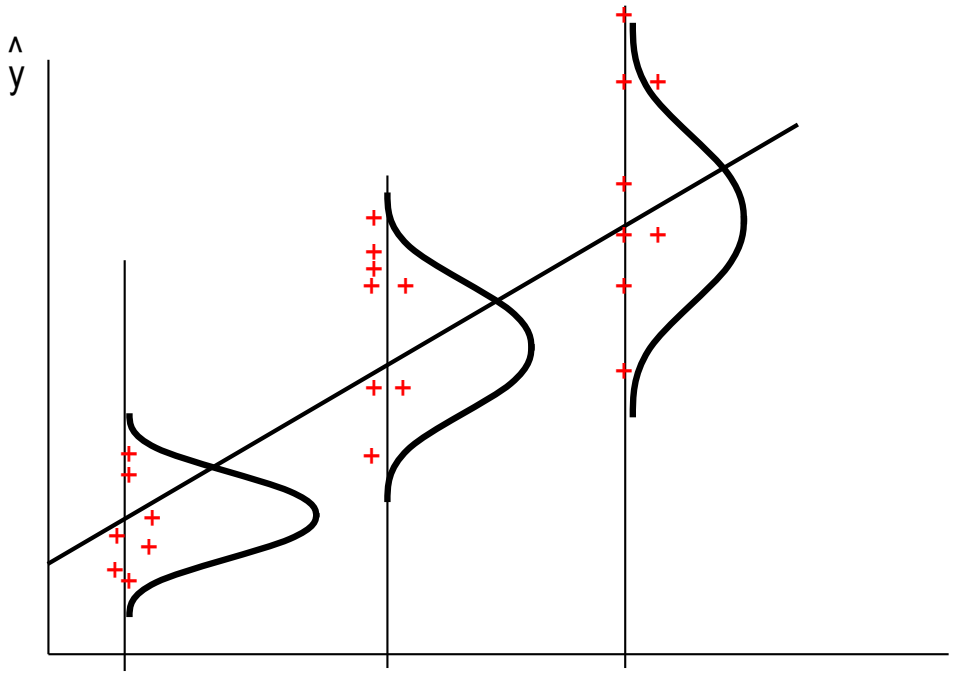
Transform the dependent variable (repeating the normality checks on the transformed data): Common transformations include taking the log or square root of the dependent variable

Use non-parametric methods.

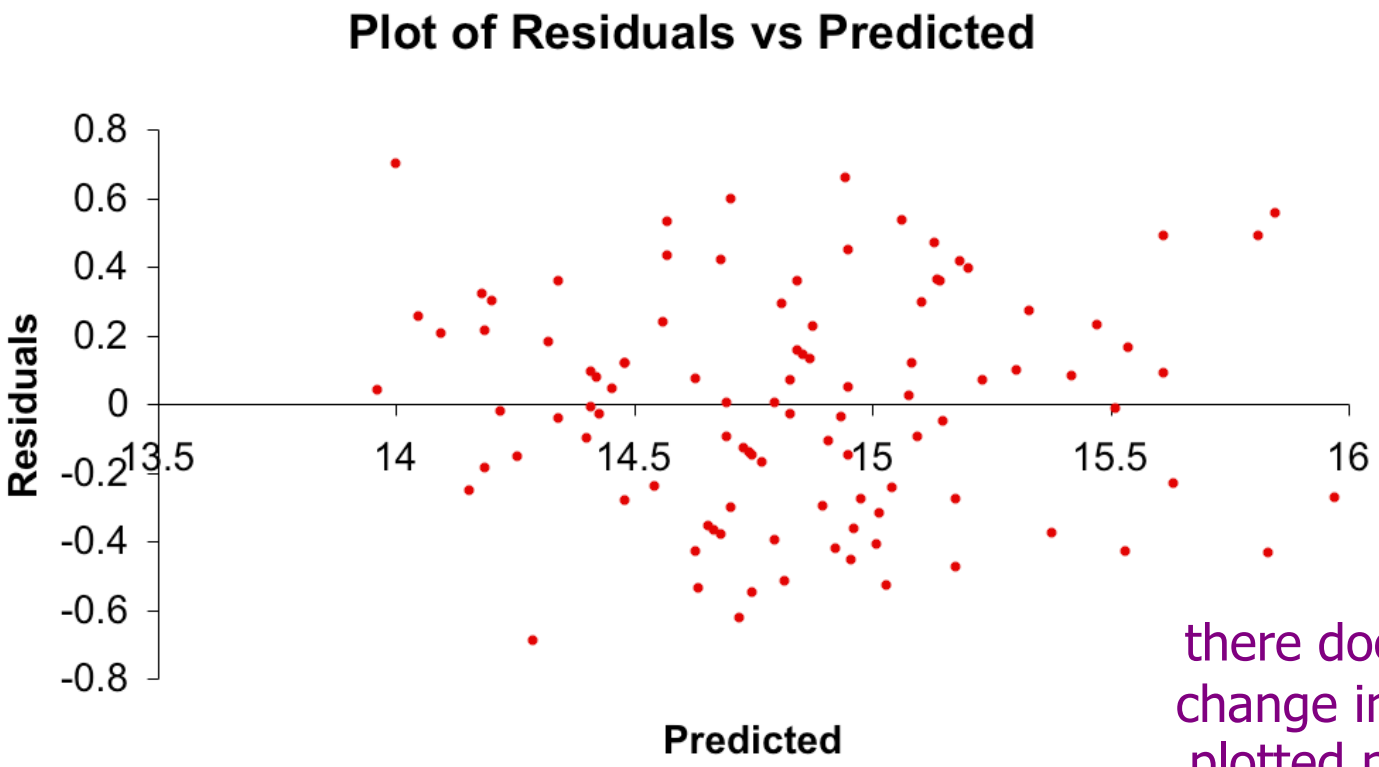
Heteroscedasticity

- When the requirement of a constant variance is violated we have a condition of heteroscedasticity. Heteroscedasticity results in biased standard errors.
- Diagnose heteroscedasticity by plotting the residual against the predicted y.





Heteroscedasticity



there doesn't appear to be a change in the ***spread*** of the plotted points, therefore no ***heteroscedasticity***

Heteroscedasticity

Another way to test for heteorscedasticity is the Breusch-Pagan test. The null hypothesis is that residuals are homoskedastic.

In stata the command is estat hettest (after the regression)

```
. estat hettest

Breusch-Pagan / Cook-Weisberg test for heteroskedasticity
         Ho: Constant variance
         Variables: fitted values of csat

         chi2(1)      =     2.72
         Prob > chi2  =   0.0993
```

If the test statistic is significant, then there is unspecified heteroscedasticity, which you can correct by estimating with the **robust** option to the **regress** command and/or you may use weighted least squares instead of OLS. You may use both **WLS** and **,robust** in the same model.

According to Berry and Feldman (1985) and Tabachnick and Fidell (1996) slight heteroscedasticity has little effect on significance tests; however, when heteroscedasticity is marked it can lead to serious distortion of findings and seriously weaken the analysis thus increasing the possibility of a Type I error.

Non Independence of Error Variables

- **A time series** is constituted if data were collected over time.
- Examining the residuals over time, no pattern should be observed if the errors are independent.
- When a pattern is detected, the errors are said to be autocorrelated.
- Autocorrelation can be detected by graphing the residuals against time.

Additionally, there are issues that can arise during the analysis that are of great concern to data analysts:

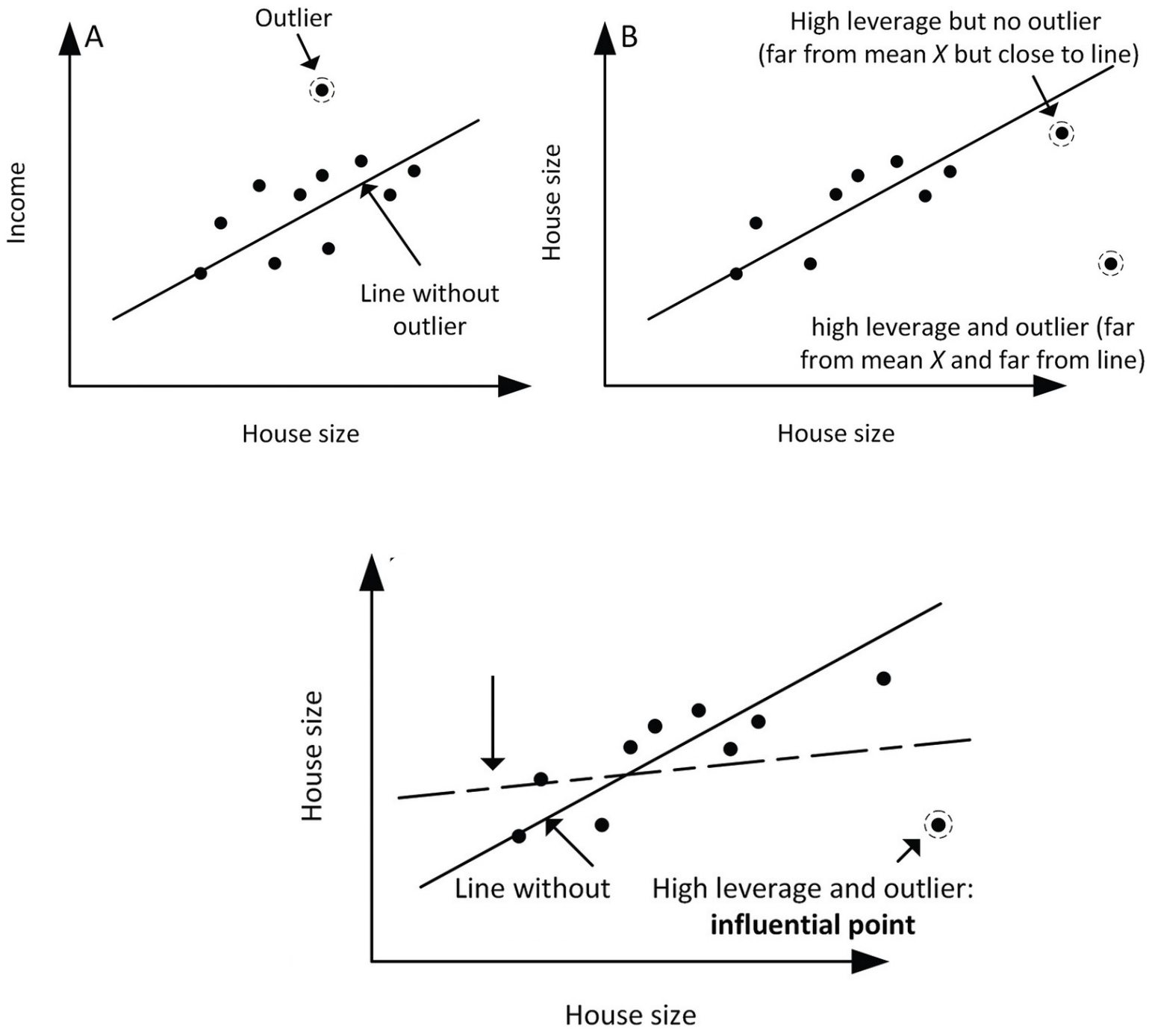
Model specification – the model should be properly specified (including all relevant variables, and excluding irrelevant variables)

- **Multicollinearity** – predictors that are highly related to each other and both predictive of your outcome, can cause problems in estimating the regression coefficients.
- **Unusual and Influential Data**
  - **Outliers**: observations with large residuals (the deviation of the predicted score from the actual score).
  - **Leverage**: measures the extent to which the predictor differs from the mean of the predictor.
  - **Influence**: observations that have high leverage and are extreme outliers, changes coefficient estimates drastically if not included.

Multicollinearity

vif - *variance inflation factor*, a measure of potential multicollinearity. The **vif** command computes a vif for each variable and for the overall regression. (There is no hard and fast rule about acceptable vif's). Any value of VIF over 10 is worrisome

The vif score may be grossly inflated if you use categorical variables, interactions, or exponents. The vif command is rarely used in practice, perhaps because experienced statisticians can recognize multicollinearity in other ways.


A
Outlier
Income
Line without
outlier
House size
B
High leverage but no outlier
(far from mean X but close to line)
House size
high leverage and outlier (far
from mean X and far from line)
House size
House size
Line without
High leverage and outlier:
influential point
House size

Outliers

- in linear regression, an outlier is an observation with large residual. In other words, it is an observation whose dependent-variable value is unusual given its values on the predictor variables. An outlier may indicate a sample peculiarity or may indicate a data entry error or other problem.
- To identify outliers, we will look for observations with large residuals

Outlier (univariate or bivariate) removal is straightforward in most statistical software. However, it is not always desirable to remove outliers. In this case transformations (e.g., square root, log, or inverse), can improve normality, but complicate the interpretation of the results, and should be used deliberately and in an informed manner. A full treatment of transformations is discussed in many popular statistical textbooks.

We can then look at the **standardized residual** for each observation, we can use this fact to identify “large” residuals. For example, values more extreme than 2 may be a problem .

Leverage: A leverage point is defined as an observation that has a value of x that is far away from the mean of x. Leverage is a measure of how far an observation deviates from the mean of that variable.
These leverage points can have an effect on the estimate of regression coefficients.

Leverage - for measuring "unusualness" of x's: A standardized version of the distance to the mean of the predictor for each individual predictor point.
Generally, a point with leverage greater than (2k+2)/n should be carefully examined. Here k is the number of predictors and n is the number of observations

Influence: An observation is said to be influential if removing the observation substantially changes the estimate of coefficients. Influence can be thought of as the product of leverage and outlierness. Thus, influential points have a large influence on the fit of the model. One method to find influential points is to compare the fit of the model with and without each observation.

As our data point of interest has both high leverage and discrepancy, it should also have high influence

A common measure of influence is Cook's Distance, a measure, for each observation, of the extent of change in model estimates when that particular observation is omitted.

Any observation that has Cook's distance close to 1 or more, or that is substantially larger than other Cook's distances (highly influential data points), requires investigation.

General guidelines for regression modelling

1.Make sure all relevant predictors are included. These are based on your research question, theory and knowledge on the topic.
2.Combine those predictors that tend to measure the same thing (i.e. as an index).
3.Consider the possibility of adding interactions
4.Strategy to keep or drop variables:

Predictor not significant and has the expected sign -> Keep it
Predictor not significant and does not have the expected sign -> Drop it
Predictor is significant and has the expected sign -> Keep it
Predictor is significant but does not have the expected sign -> Review, you may need more variables, it may be interacting with another variable in the model or there may be an error in the data.

Gelman, Andrew, Jennifer Hill, *Data Analysis Using Regression and Multilevel/Hierarchical Models*, 2007

What are multilevel data and multilevel analysis?

What are multilevel data?

- Multilevel data are data where observations are clustered in units
- Observations within the same unit may be more similar than observations in separate units, on average
  - What effect does this have on estimation and statistical inference?

Examples of multilevel data with contextual clustering

- Observations of students, clustered within schools
- Observations of siblings, clustered within families
- Observations of individuals, clustered within countries, states, or neighborhoods
- Patients within hospitals
- Repeated test scores, clustered within students (multilevel data with intra-person clustering)

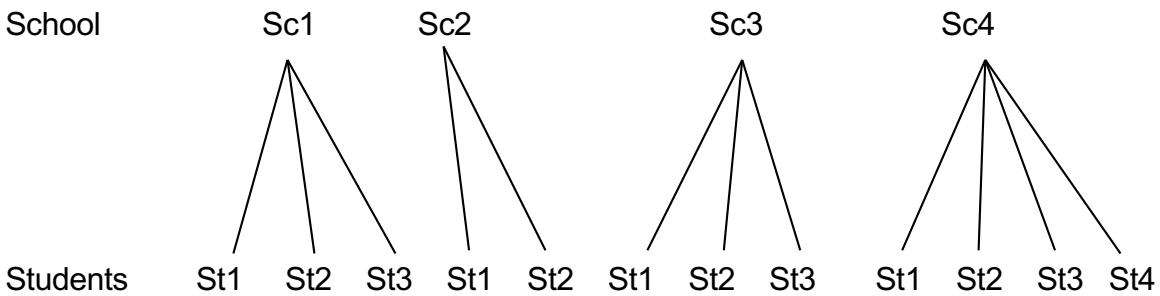
Nested Data

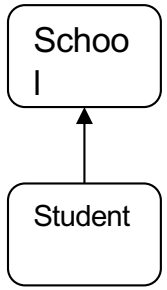
- Data nested within a group tend to be more alike than data from individuals selected at random.

- Nature of group dynamics will tend to exert an effect on individuals.

Two-level hierarchical structures

Students within schools





Students within a school are more alike than a random sample of students. This is the ‘clustering’ effect of schools.

Why Multilevel Modeling
vs.
Traditional Approaches?

Traditional Approaches – 1-Level

1. Individual level analysis (ignore group)

2. Group level analysis (aggregate data and ignore individuals)

Problems with Traditional Approaches

1. Individual level analysis (ignore group)

   Violation of independence of data assumption leading to misestimated standard errors (standard errors are smaller than they should be).

Problems with Traditional Approaches

1. Group level analysis

   (aggregate data and ignore individuals)

   Aggregation bias = the meaning of a variable at Level-1 (e.g., individual level SES) may not be the same as the meaning at Level-2 (e.g., school level SES)

Multilevel Approach

- 2 or more levels can be considered simultaneously
- Can analyze within- and between-group variability

Multilevel regression models

- Also called
  - Hierarchical Linear Models
  - Mixed Models
  - Multilevel Models
  - Growth Models
  - Slopes-as-Outcomes Models

Multilevel Regression Models

- A form of regression models
- Used to answer questions about the relationship of context to individual outcomes
- Used to estimate both within-unit and between-unit relationships (and cross-level interactions)
  - e.g., within- vs. between-school relationships between SES and achievement

Data frame for student within school example

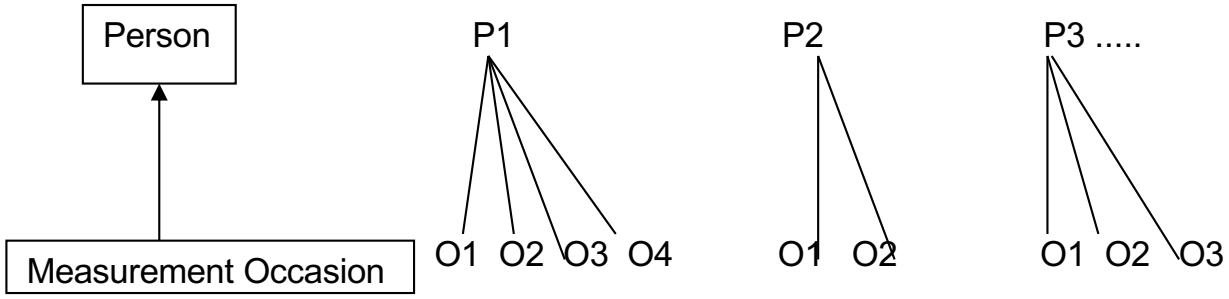
Classifications or levels		**Response**	**Explanatory variables**		
Student i	*School* j	*Student Exam* $score_{ij}$	*Student previous Examination* $score_{ij}$	*Student* $gender_{ij}$	*School* $type_j$
1	1	75	56	M	State
2	1	71	45	M	State
3	1	91	72	F	State
1	2	68	49	F	Private
2	2	37	36	M	Private
3	2	67	56	M	Private
1	3	82	76	F	State

1 Do Males make greater progress than Females?
2 *Does the gender gap vary across schools?
3 *What is the between-school variation in student's progress?

4 * Are students in public schools less variable in their progress?

* Requires multilevel model to answer

Classification, unit diagrams and data frames for **repeated measures structures**.



Person	*H-Occ1*	*H-Occ2*	*H-Occ3*	*Age-Occ1*	*Age-Occ2*	*Age-Occ3*	*Gender*
1	75	85	95	5	6	7	F
2	82	91	*	7	8	*	M
3	88	93	96	5	6	7	F

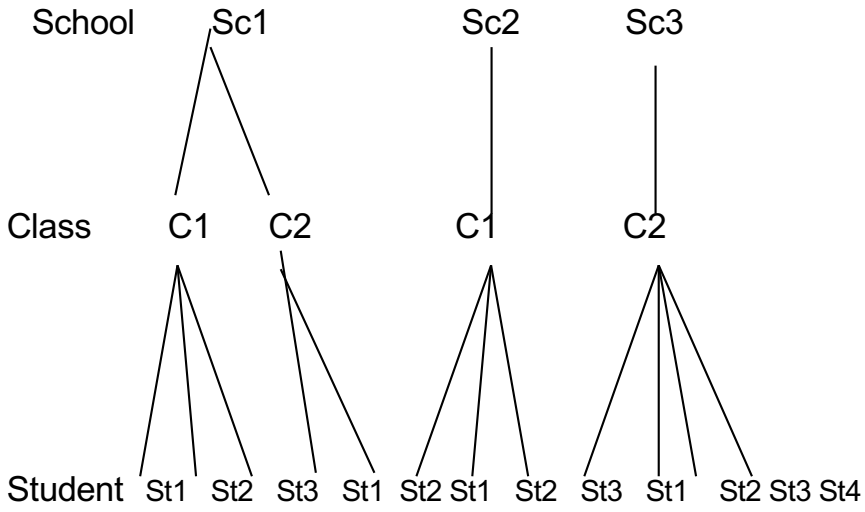
Wide form 1 row per individual

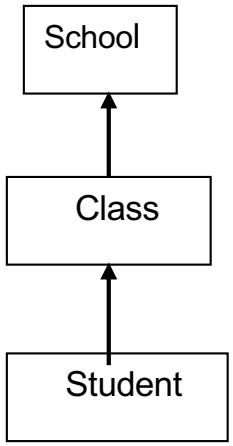
Long form 1 row per occasion(required by *MLwiN*)

Classifications or levels		**Response**	**Explanatory variables**	
Occasion I	*Person* J	$Height_{ij}$	Age_{ij}	$Gender_j$
1	1	75	5	F
2	1	85	6	F
3	1	95	7	F
1	2	82	7	M
2	2	91	8	M
1	3	88	5	F
2	3	93	6	F
3	3	96	7	F

Three level structures

Students:classes:schools





MLM allow a different number of students in each class and a different number of classes in each school.

The multilevel model:

$$y_{ij} = \beta_0 + \beta_1 x_{1ij} + (u_{0j} + e_{ij})$$

- **Fixed part** $\beta_0 + \beta_1$
- **Random part (Level 2)** $u_{0j} \sim N(0, \sigma^2_{u0})$
- **Random part (Level 1)** $e_{0ij} \sim N(0, \sigma^2_{e0})$