Applied Statistics
Lecture 4

Prof.ssa Chiara Seghieri, Dott.ssa Costanza Tortù

Laboratorio di Management e Sanità, Istituto di Management, L'EMbeDS
Scuola Superiore Sant'Anna, Pisa
c.seghieri@santannapisa.it
c.tortu@santannapisa.it

Outline

1. Motivation and Intuition
2. The Linear Regression Model
3. Estimation of the Coefficients (OLS vs MLE)
4. Interpretation of coefficients
5. Regression Diagnostics
6. Goodness of fit
7. Multiple regression models: interpretation of coefficients
8. Correlation vs causation

1) Motivation and intuition

Motivation

- We will model the relationship between a set of variables x_s and a single variable y.

Examples: determinants of income

- 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. This function is called a **mode**l.

$Y=f(x_{1,}\, x_{2, \ldots,}\, x_s)$

N.B Some of these variables may be either *unobservable* or *unimpactful on y.*

To sum up

Regression model estimates the nature of the relationship between the independent and dependent variables.

Specifically, it allows researchers to understand

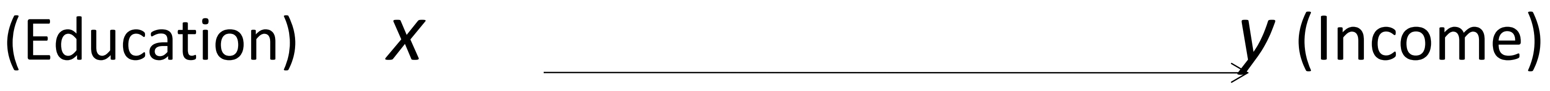
A Size of the relationship.

B Strength of the relationship.

C Statistical significance of the relationship.

Simple and multivariate models

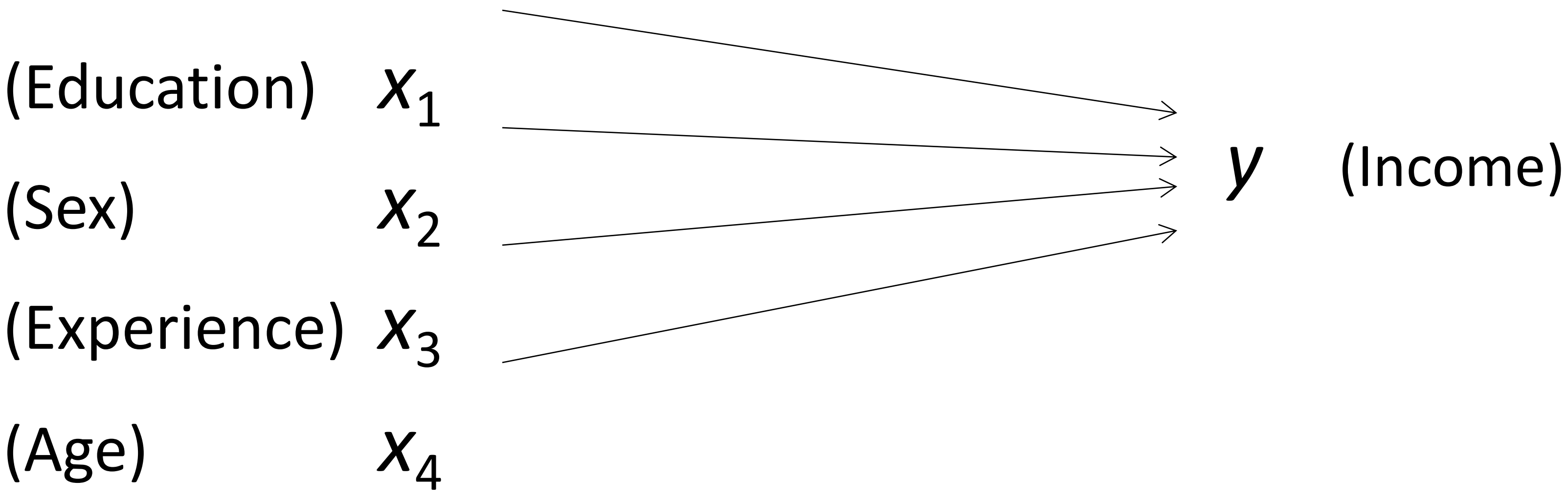
Bivariate or simple regression model



(Education) x

y (Income)

Multivariate or multiple regression model



(Education) x_1

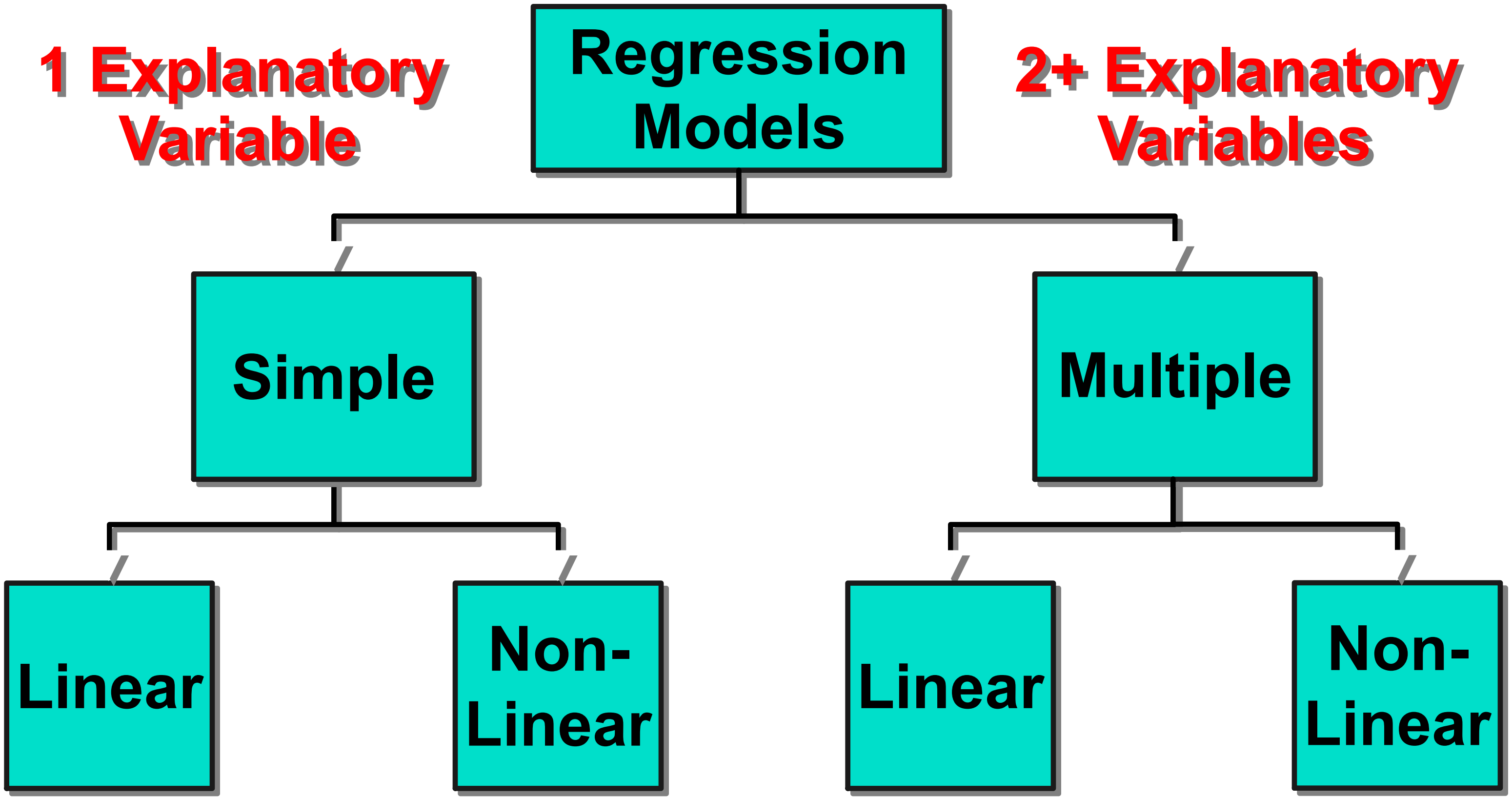
(Sex) x_2

(Experience) x_3

(Age) x_4

y (Income)

Types of Regression Models



1 Explanatory Variable

Regression Models

2+ Explanatory Variables

Simple

Multiple

Linear

Non-Linear

Linear

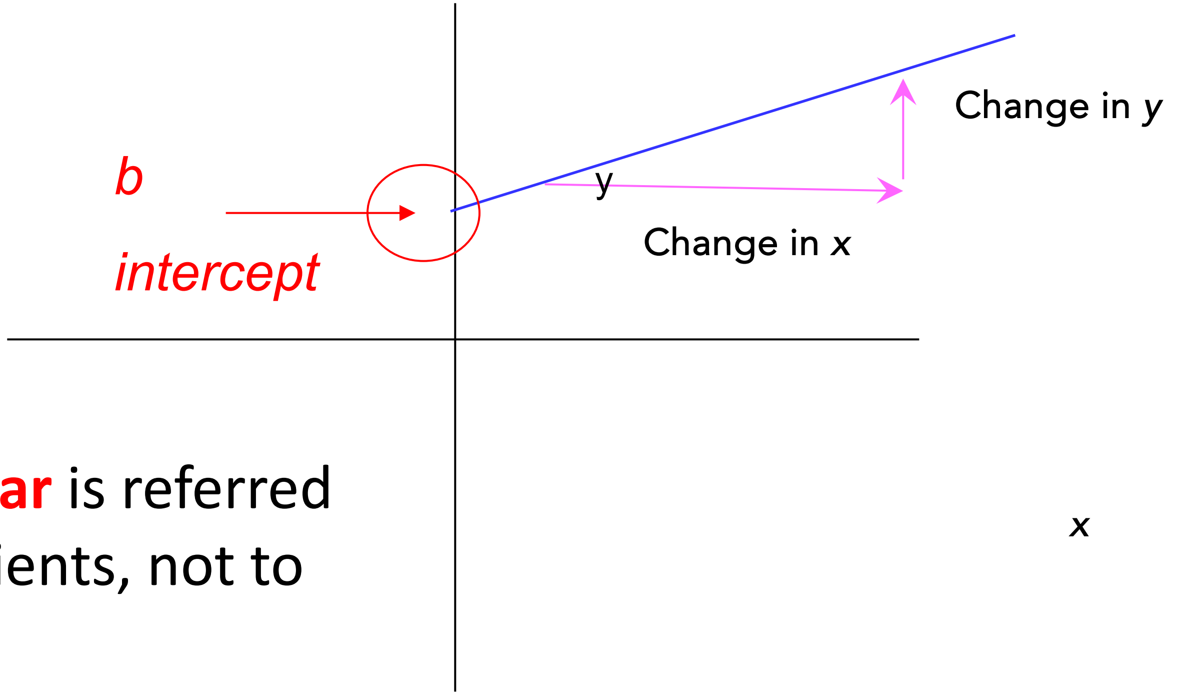
Non-Linear

2) Simple Linear Regression Model

What is “Linear”?

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

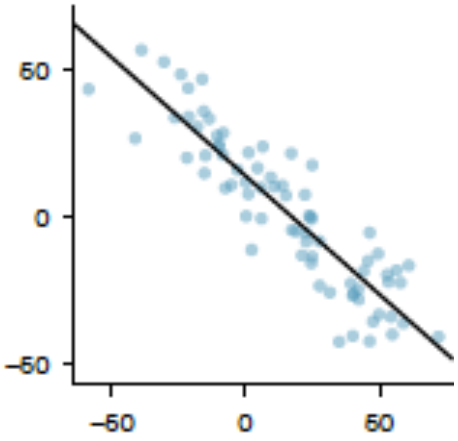
m=slope=change in y/change in x

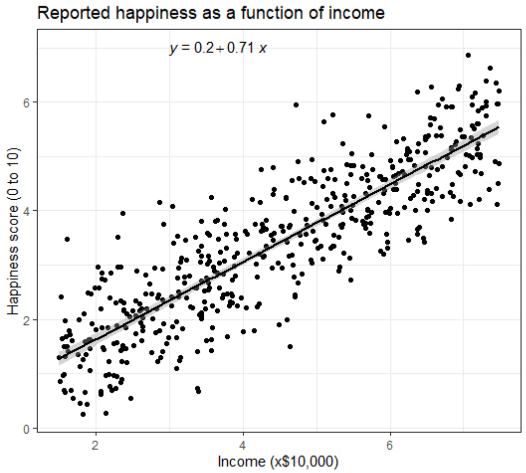


The term **linear** is referred to the coefficients, not to the x,

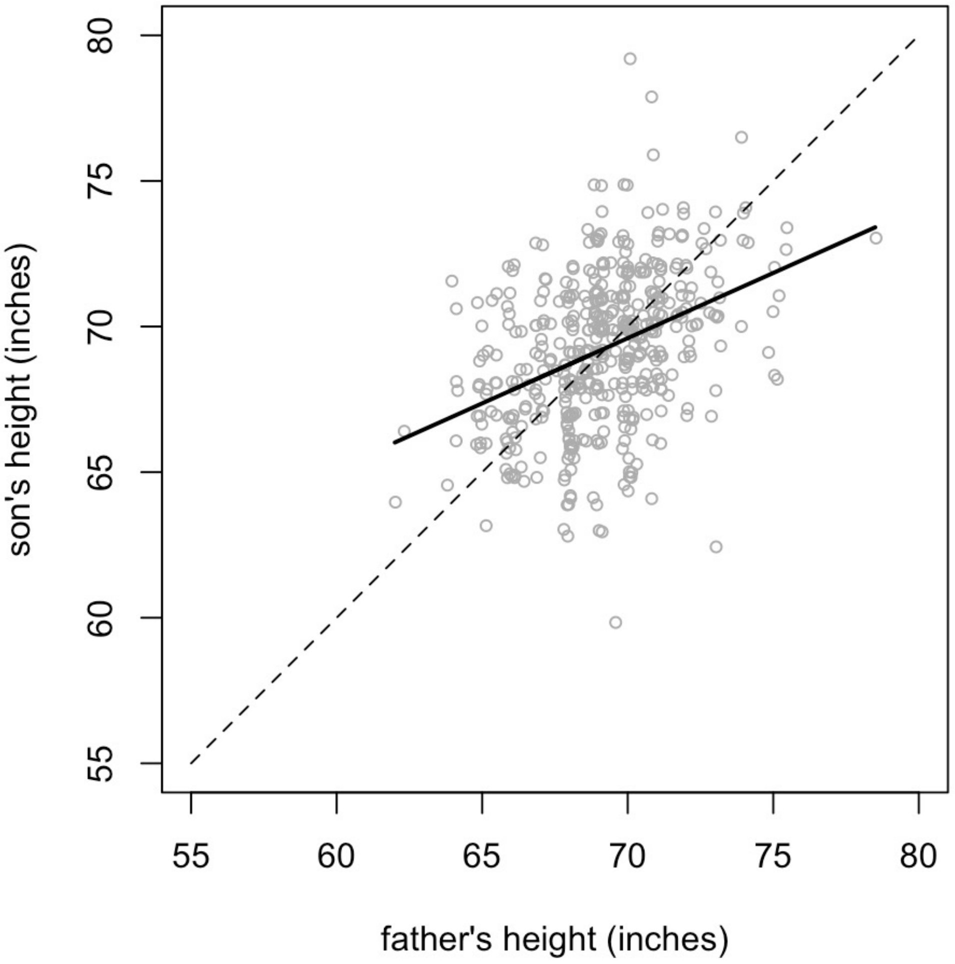
It is a deterministic mathematical relationship! we know the exact value of y just by knowing the value of x. This is unrealistic in almost any natural process!

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.





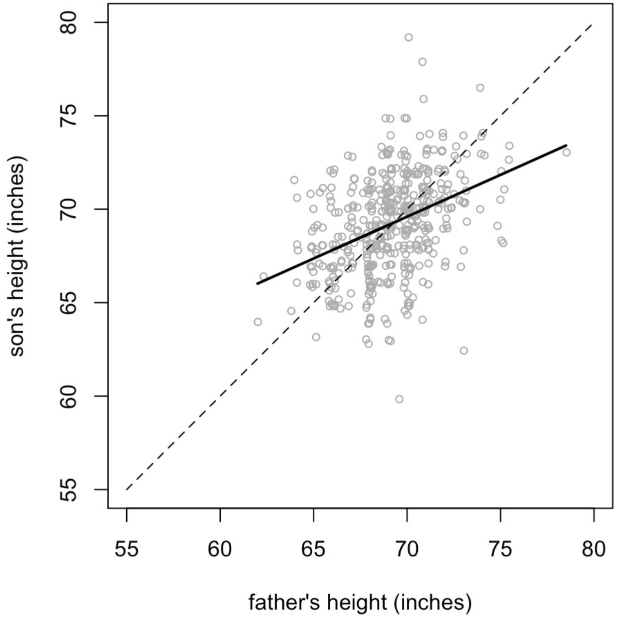
Linear regression is the statistical method for fitting a line to data where the relationship between two variables, x and y, can be modelled by a straight line with some error.


80
75
70
65
60
55
son's height (inches)
55
60
65
70
75
80
father's height (inches)

We are looking for functional relation

between the dependent variable y and the predictor variable x.

$$y = f(x)$$



In the graph for some values of X 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

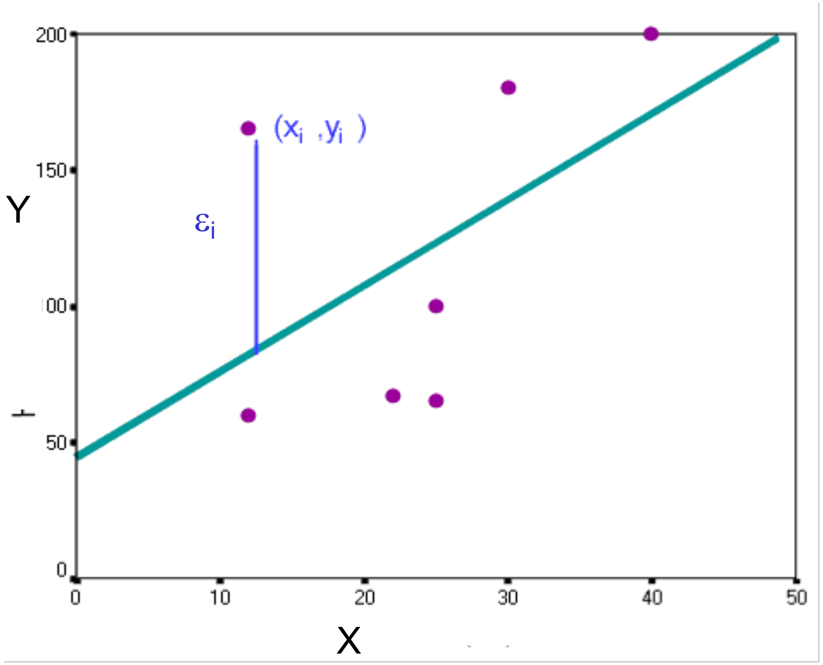
The idea behind this is: when we have several values of y observed for one value of X we take the average of these values of y to assign to x.

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 the regression function.

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



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 $b_0 + b_1x$.

Of course we cannot make the calculation exactly, in practice, because the two parameters are unknown therefore we make estimates of the parameters.

Notation

n observations (sample size).

The variable **Y** is the response variable, and y_1,y_2,…,y_n are the observed values of the **response** Y.

The variable **X** is the predictor variable and x_1,x_2,...x_n are observed values of the **predictor**, X.

The observations are considered as coordinates, (x_i,y_i), for i=1,…,n.

The points, (x_1,y_1),…,(x_n,y_n), may not fall exactly on a line.
There is some **error** we must consider.

The general form of the simple linear regression model is:

$$Y=\beta_0+\beta_1X+\varepsilon$$

For an individual observation: $y_i=\beta_0+\beta_1x_i+\varepsilon_i$

Where:

β_0 is the population **intercept**,
β_1 is the population **slope**, and
ε_i is the **error** or deviation of y_i from the line, $\beta_0+\beta_1x_i$.

To make inference 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.

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

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

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:

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

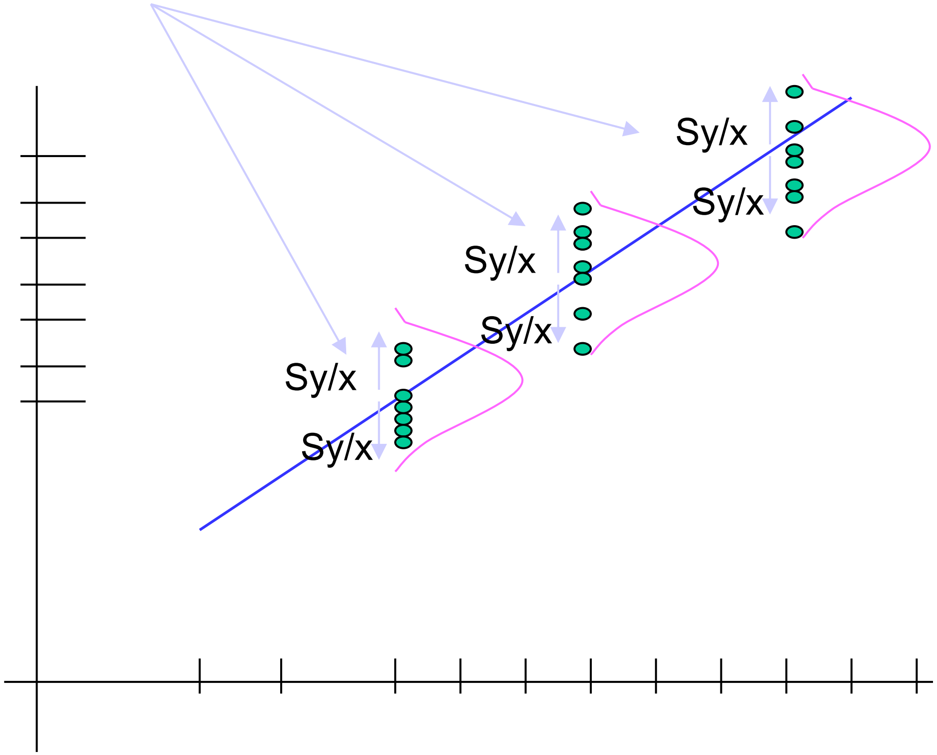
β_1 is the coefficients that describe the size of the effect the independent variable is having on your dependent variable *Y*, and β_0 is the value *Y* is predicted to have when all the independent variables are equal to zero. The "fixed-x" assumption is that the explanatory variable is measured without error.

Assumptions of the Model (1/2)

Linear regression assumes that…

The relationship between X and Y is linear.

The error model underlying a linear regression analysis includes the assumptions of fixed-x, Normality, equal spread, and independent errors.



Assumptions of the Model (1/2)

Linear: does not imply that only linear relationships can be studied. It says that the beta’s must not be in a transformed form. It is OK to transform x or Y , and that allows many non-linear relationships to be represented on a new scale that makes the relationship linear.

Normality and equal variance: if we could collect many subjects with that x value, their distribution around the population mean is Normal with a spread that is the same value for each value of x (and corresponding population

Assumptions of the Model (2/2)

Independent error: the error (deviation of the true outcome value from the population mean of the outcome for a given x value) for one observational unit is not predictable from knowledge of the error for another observational unit.

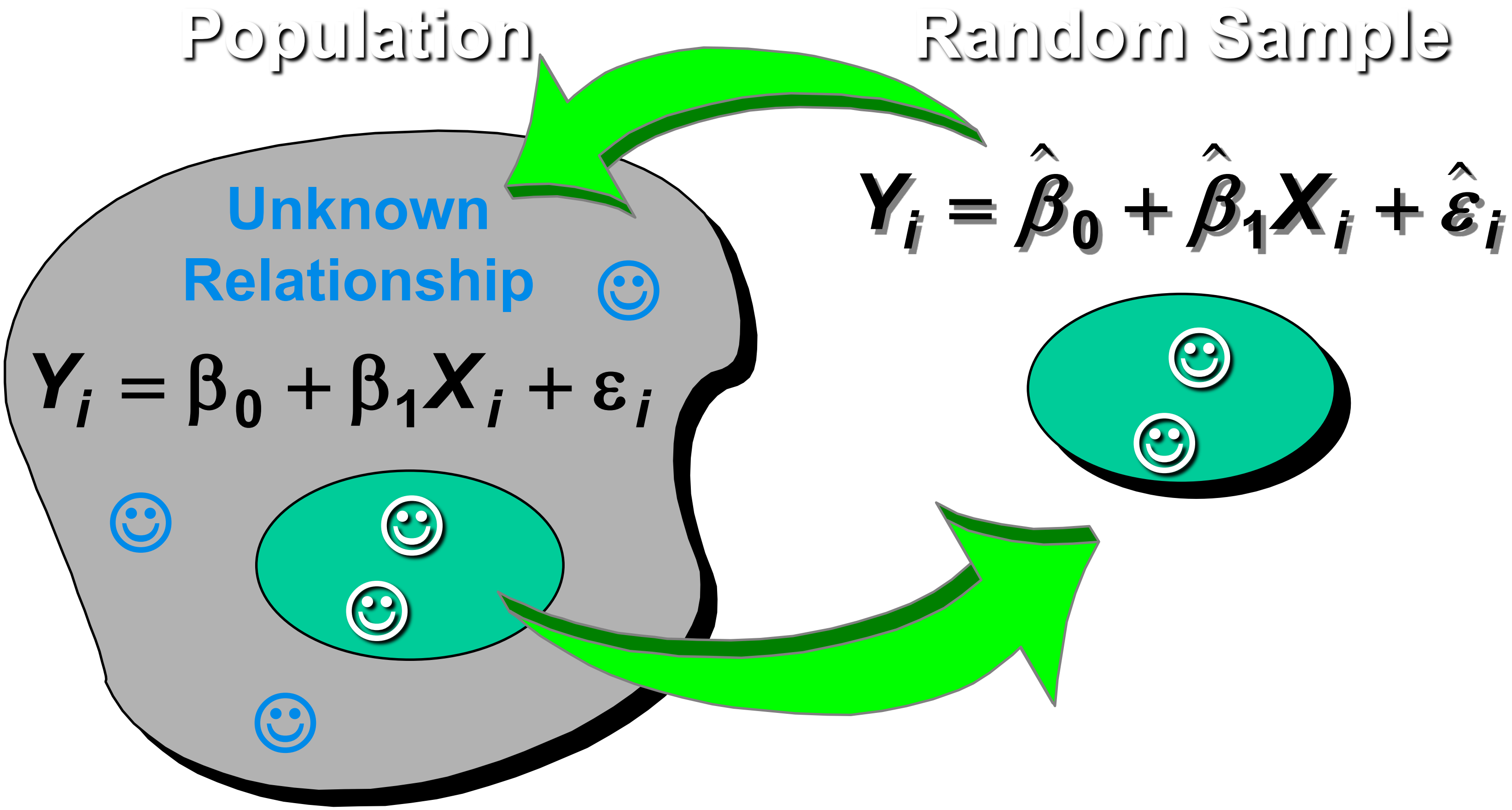
3) Estimation of Coefficients (OLS vs MLE)

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



Population

Random Sample

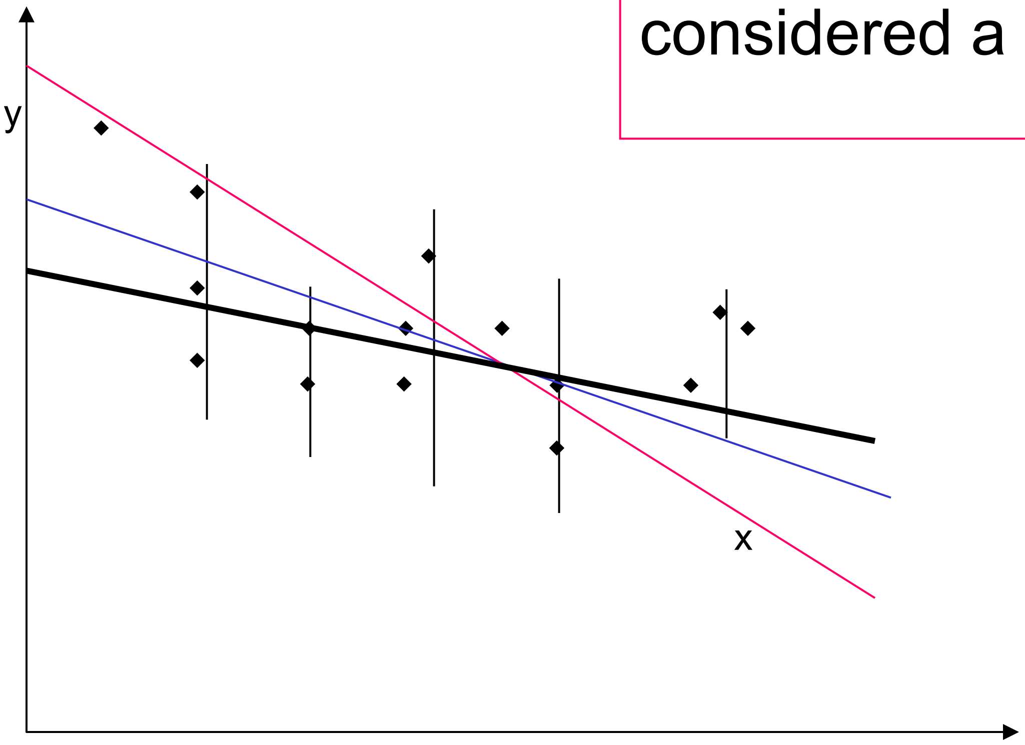
Unknown Relationship

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$

$$Y_i = \hat{\beta}_0 + \hat{\beta}_1 X_i + \hat{\varepsilon}_i$$

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**.

In practice, we don’t try every possible line. Instead we use calculus to find the values of β0 and β1 that give the minimum sum of squared residuals.

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: 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

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

- Sample Y - intercept

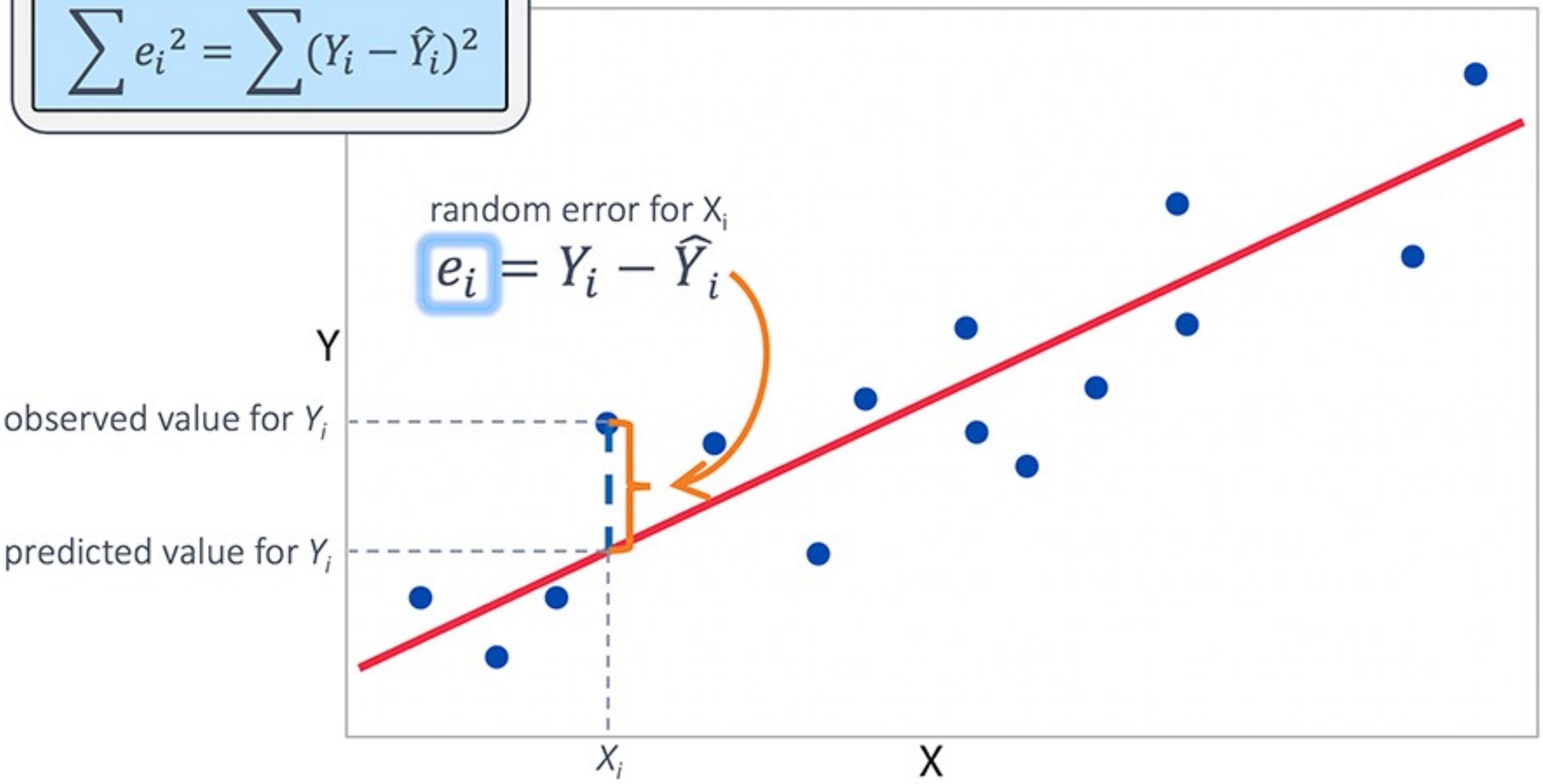
$$\hat{\beta}_0 = \overline{y} - \hat{\beta}_1 \overline{x}$$

Regression line always goes through the point:

OLS: graphical intuition

Method of Least Squares

$$\sum e_i{}^2 = \sum (Y_i - \hat{Y}_i)^2$$



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

a

- 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

4) 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 (β_0)
   - Average Value of *Y* When *X* = 0 (when it makes ^ sense that X=0)
     - *if* β_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)

  (Thanks to the normality assumption!!)

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              Number of obs =      20
-------------+------------------------------           F(  1,    18) =    6.14
       Model |  181.370325     1  181.370325           Prob > F      =  0.0233
    Residual |  531.573154    18  29.5318419           R-squared     =  0.2544
-------------+------------------------------           Adj R-squared =  0.2130
       Total |  712.943479    19   37.523341           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

•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

- 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, i.e. the estimated probability of a Type I error associated to the test statistic: the null 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' $0.023 < 0.05$)

Together with the parameter estimates also **standard errors** are reported.

SE are estimated standard deviations of the corresponding sampling distributions and gives an idea of the scale of the variability of the estimate of the coefficient around the true, unknown value if we repeat the whole experiment many times.

5) Regression diagnostics

Assumptions for Simple Linear Regression

1.Linearity: The relationship between X and Y must be linear.Check this assumption by examining a scatterplot of x and y.

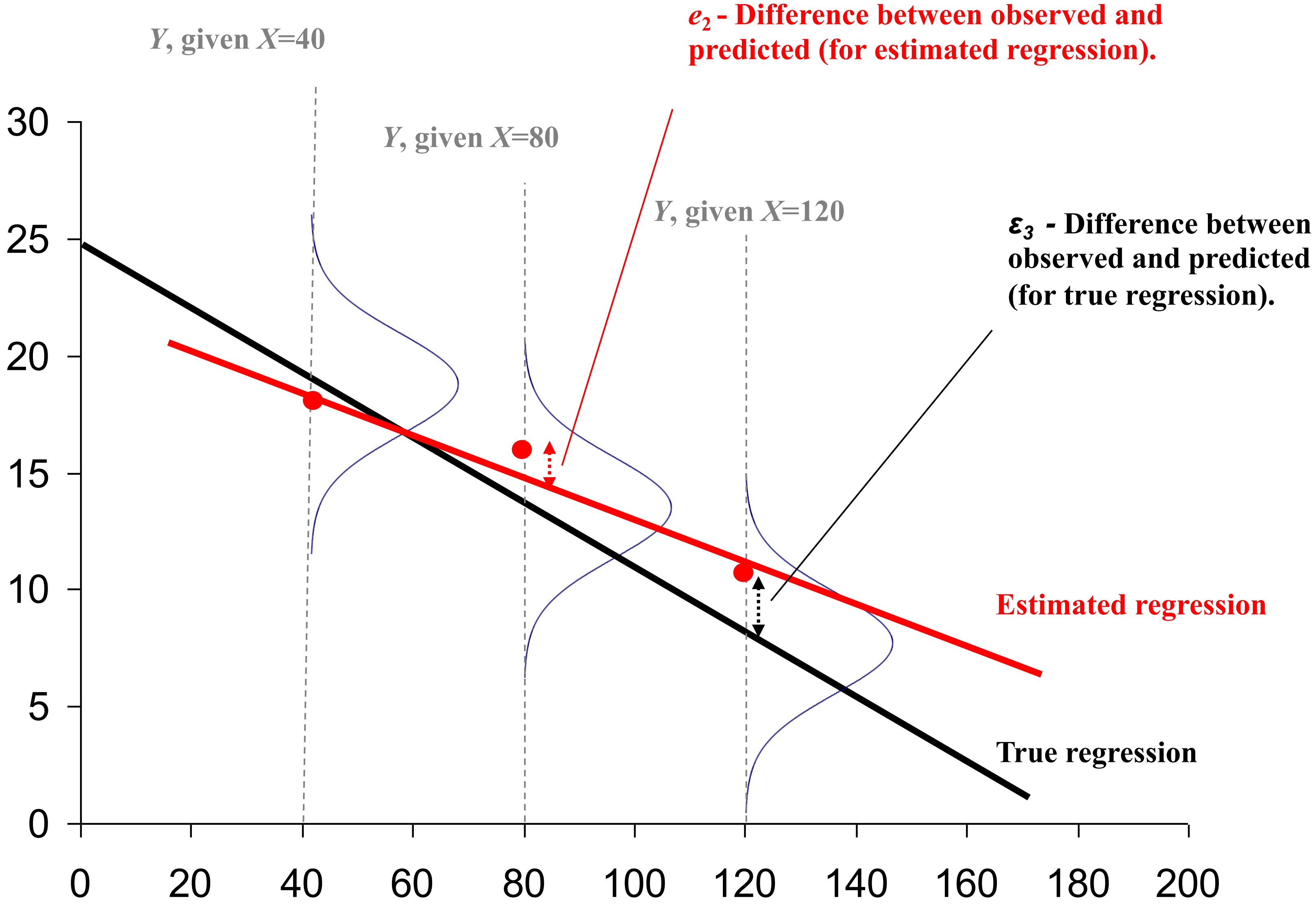
2.Independence of errors: There is not a relationship between the residuals and the Y variable; in other words, Y is independent of errors. Check this assumption by examining a scatterplot of “residuals versus fits”; the correlation should be approximately 0. In other words, there should not look like there is a relationship.

3.Normality of errors: The residuals must be approximately normally distributed. Check this assumption by examining a normal probability plot; the observations should be near the line. You can also examine a histogram of the residuals; it should be approximately normally distributed.

4.Equal variances: The variance of the residuals is the same for all values of X. Check this assumption by examining the scatterplot of “residuals versus fits”; the variance of the residuals should be the same across all values of the *x*-axis. If the plot shows a pattern (e.g., bowtie or megaphone shape), then variances are not consistent, and this assumption has not been met.

Residual

- The difference between the observed value y_i and the corresponding fitted value. $\hat{y}_i$
- A residual is the deviation of an outcome from the predicted 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.



e_2 - Difference between observed and predicted (for estimated regression).

ε_3 - Difference between observed and predicted (for true regression).

Conditions for Regression Inference

- The simple 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?

Regression Diagnostics

How can we diagnose violations of these conditions?

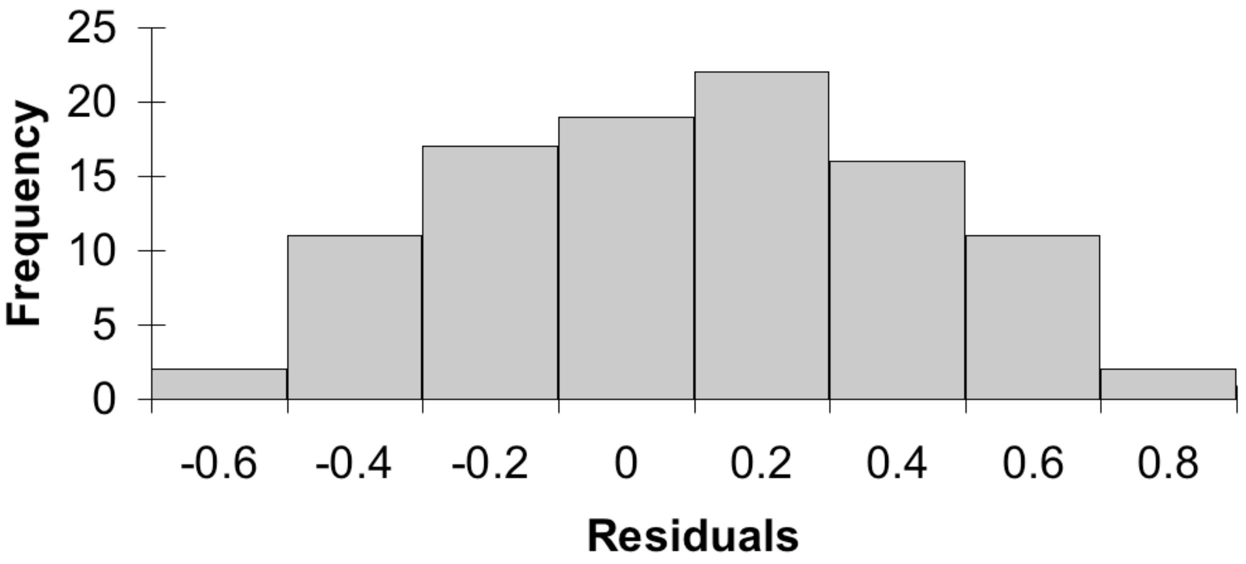
➔ **Residual Analysis**, that is, examine the ***differences*** between the actual data points and those predicted by the linear equation.

➔ 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.

Nonnormality

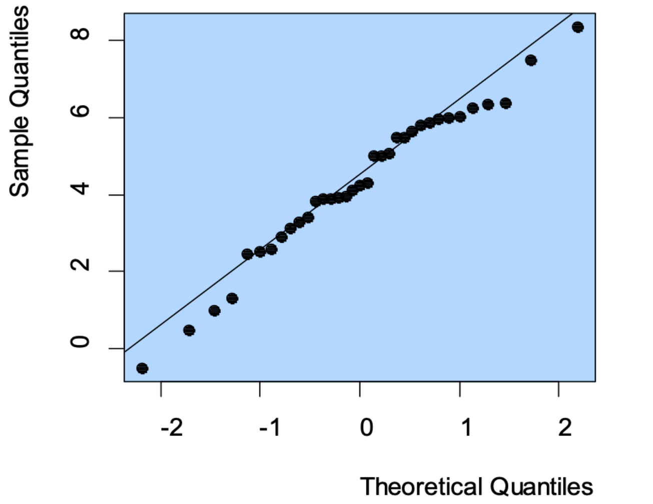
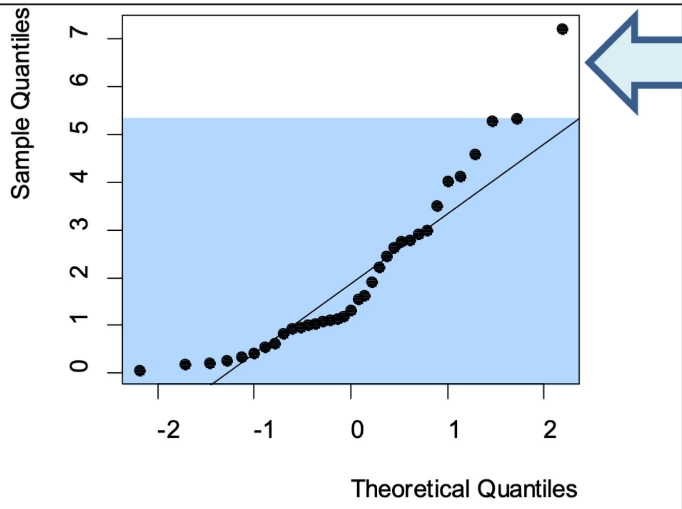
We can take the residuals and put them into a histogram to visually check for normality…



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

Nonnormality

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.

Q-Q plot of approximately normally distributed data	Q-Q plot of skewed data
	

The scatter of skewed data tends to form curves moving away from the line at the ends

Nonnormality

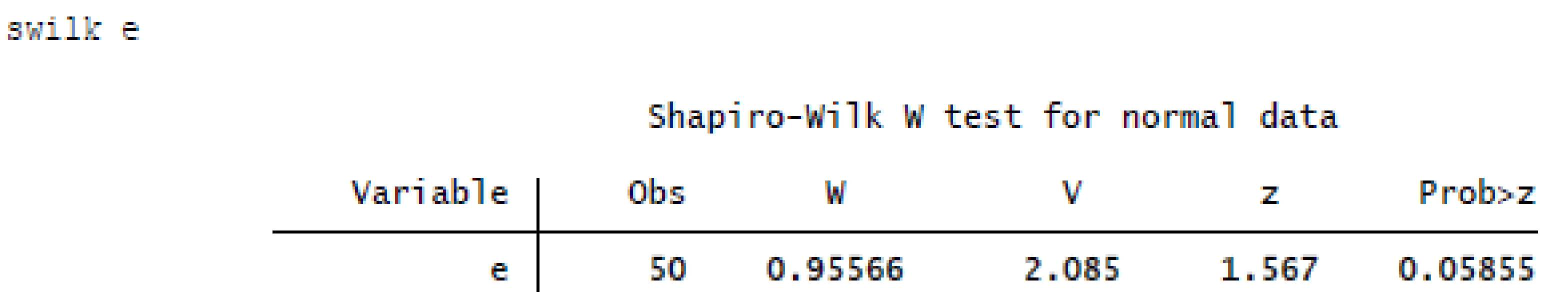
There are also specific test for which could be used in conjunction with either a histogram or a Q-Q plot.

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
```

Nonnormality

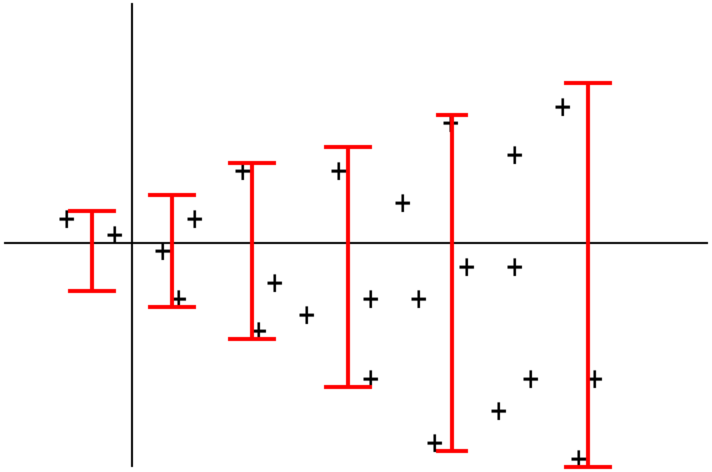
Regression Inference is robust against moderate lack of Normality. On the other hand, outliers and influential observations can invalidate the results of inference for regression. What to do?

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

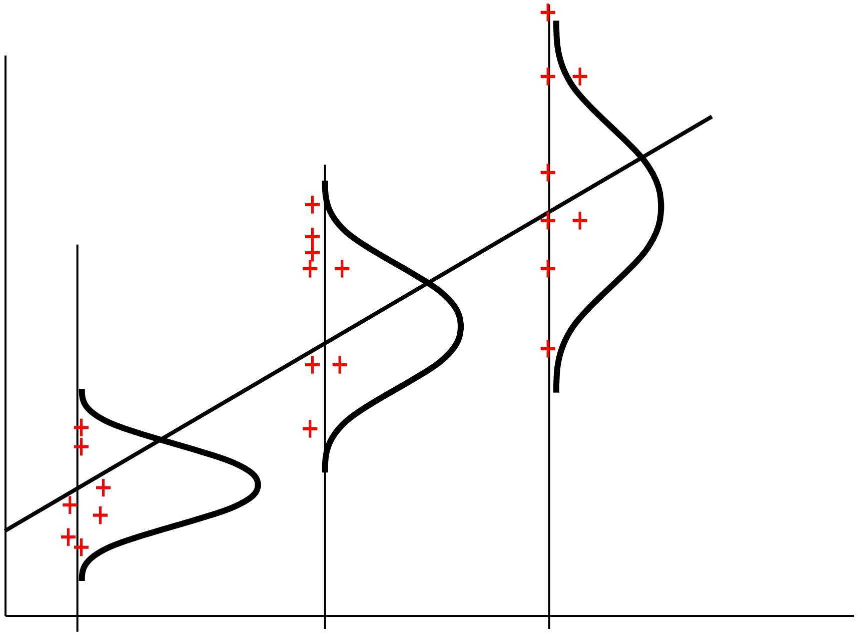
Use non-parametric methods.

Heteroscedasticity

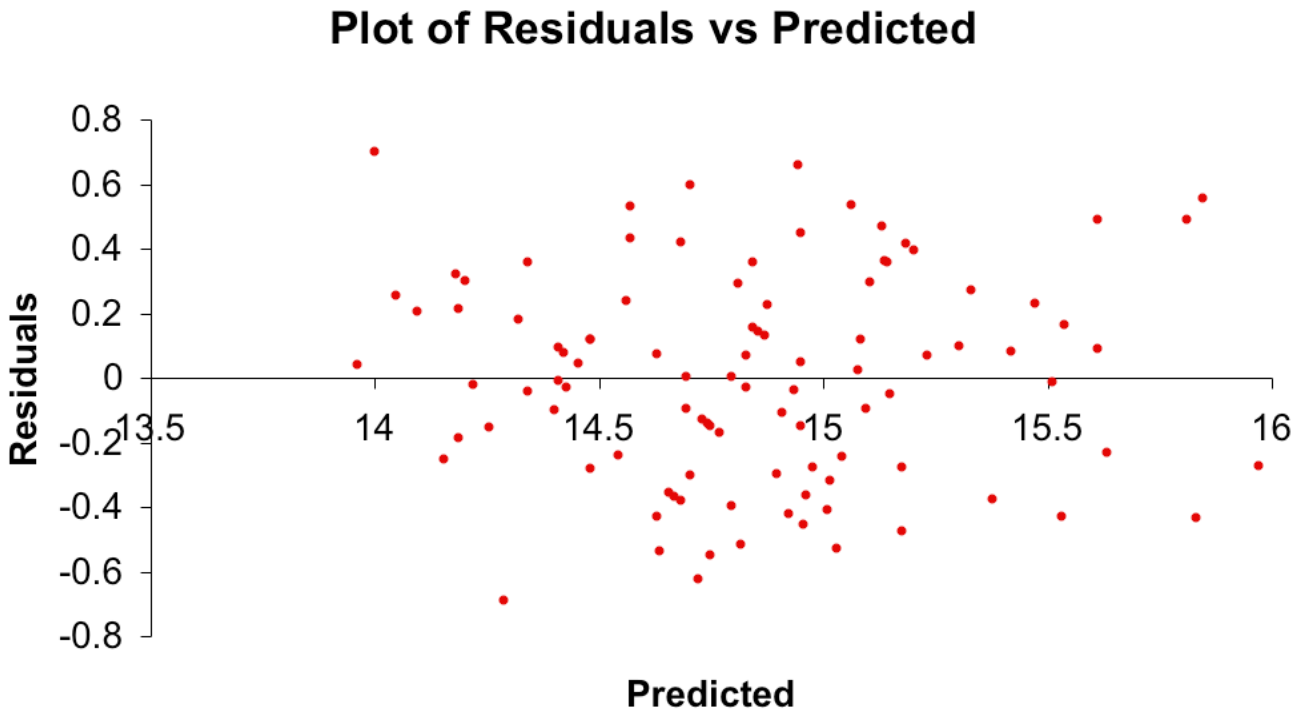
- When the requirement of a constant variance is violated we have a condition of heteroscedasticity. Heteroscedasticity results in biased standard errors.
- Diagnose heteroscedasticity by plotting the residual against the predicted y.



The spread increases with $\hat{y}$

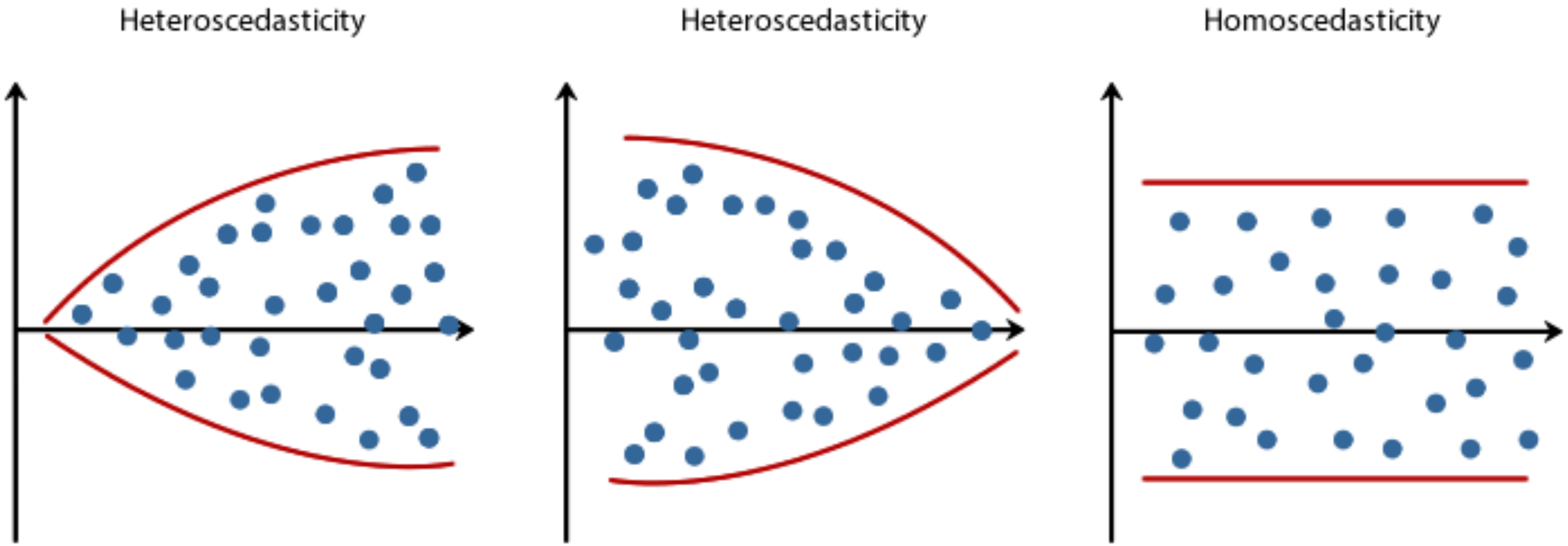


Heteroscedasticity



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

Heteroscedasticity



Heteroscedasticity

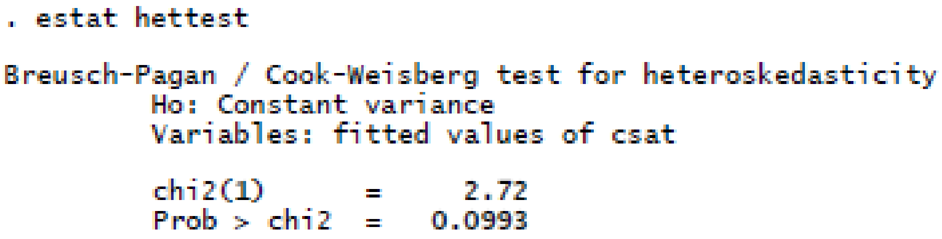
Heteroscedasticity

Homoscedasticity

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
```

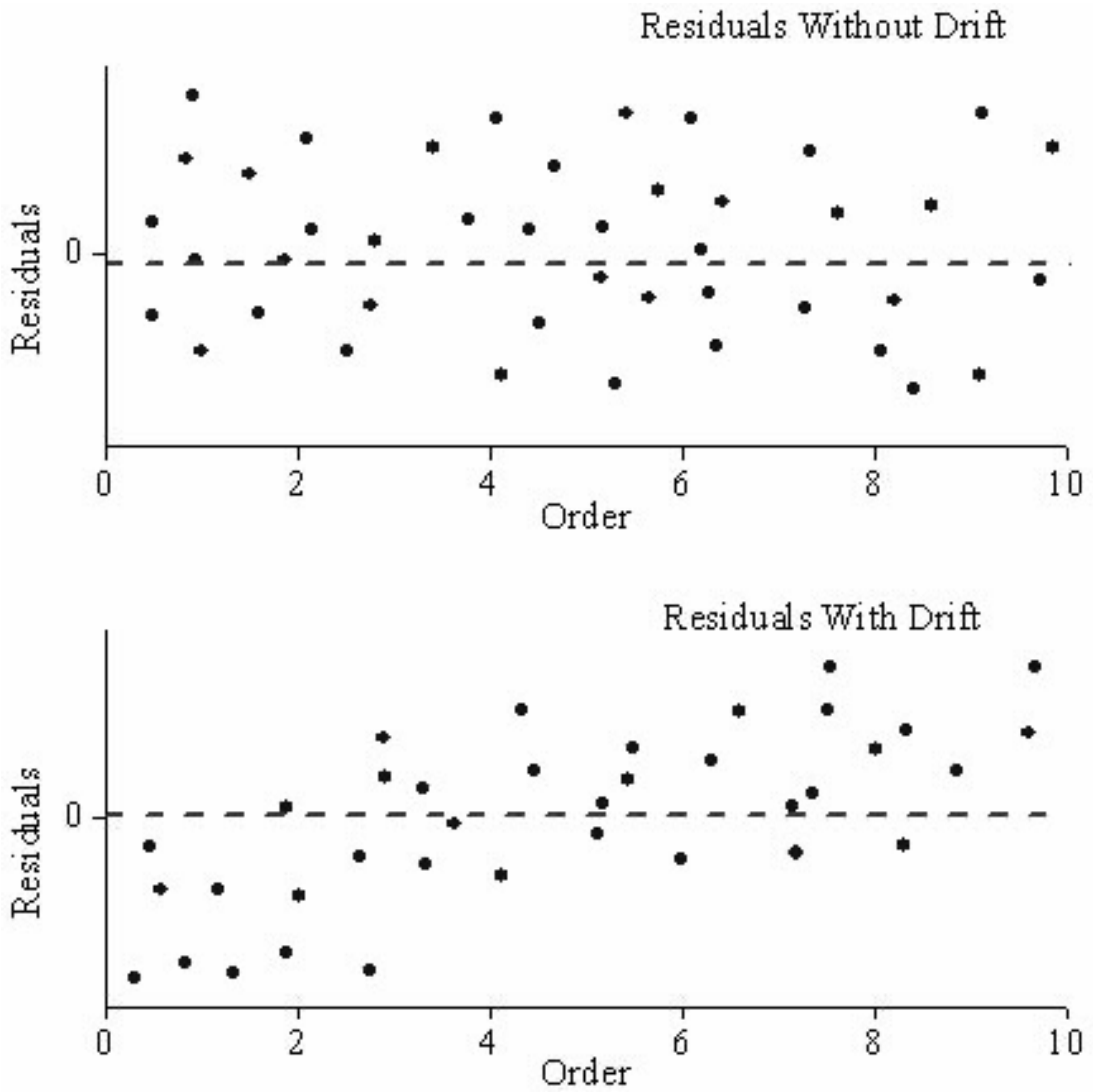
If the test statistic is significant, then there is unspecified heteroscedasticity, which you can correct by estimating with the **robust** option to the **regress** command and/or you may use weighted least squares instead of OLS. You may use both **WLS** and **robust** in the same model.

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 Errors

- **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.

Non-Independence of Errors

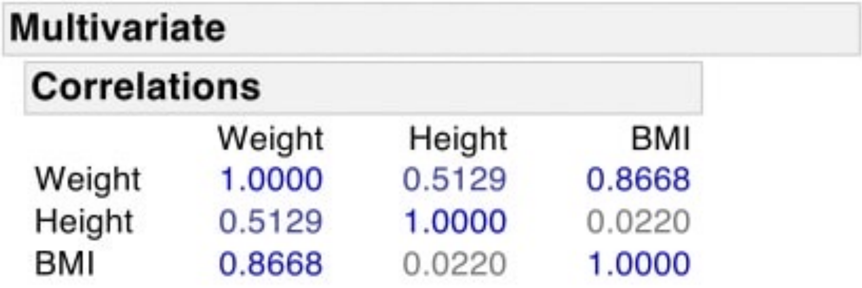


Issues in model specification

Additionally, there are issues that can arise during the analysis that, while strictly speaking are not assumptions of regression, are none the less, 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

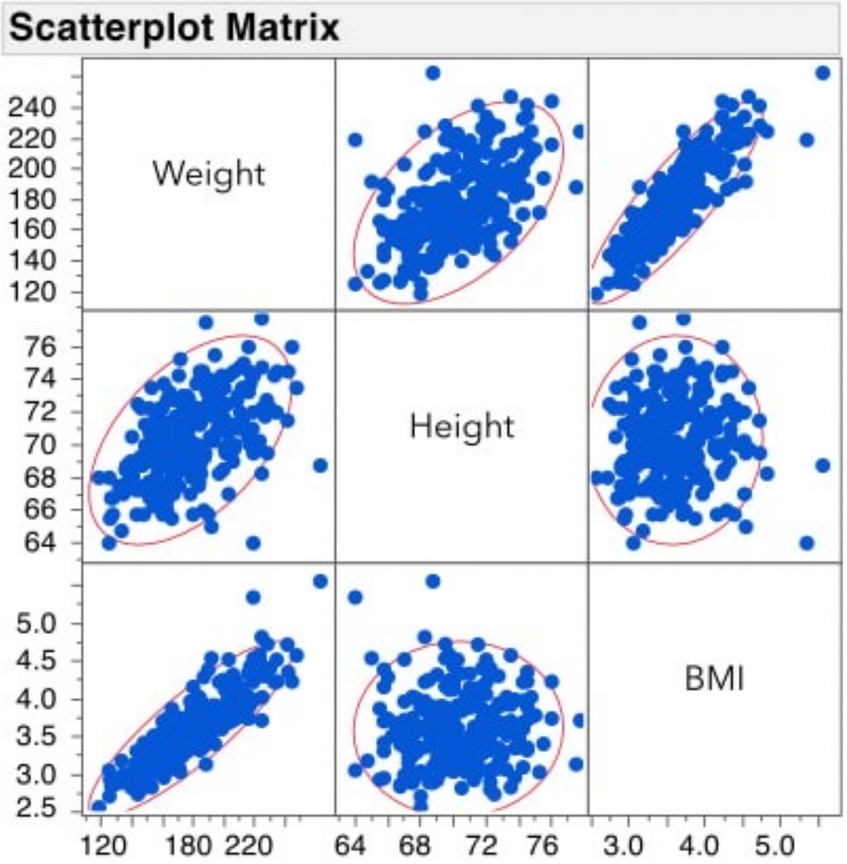
Issues in model specification



Multivariate

Correlations

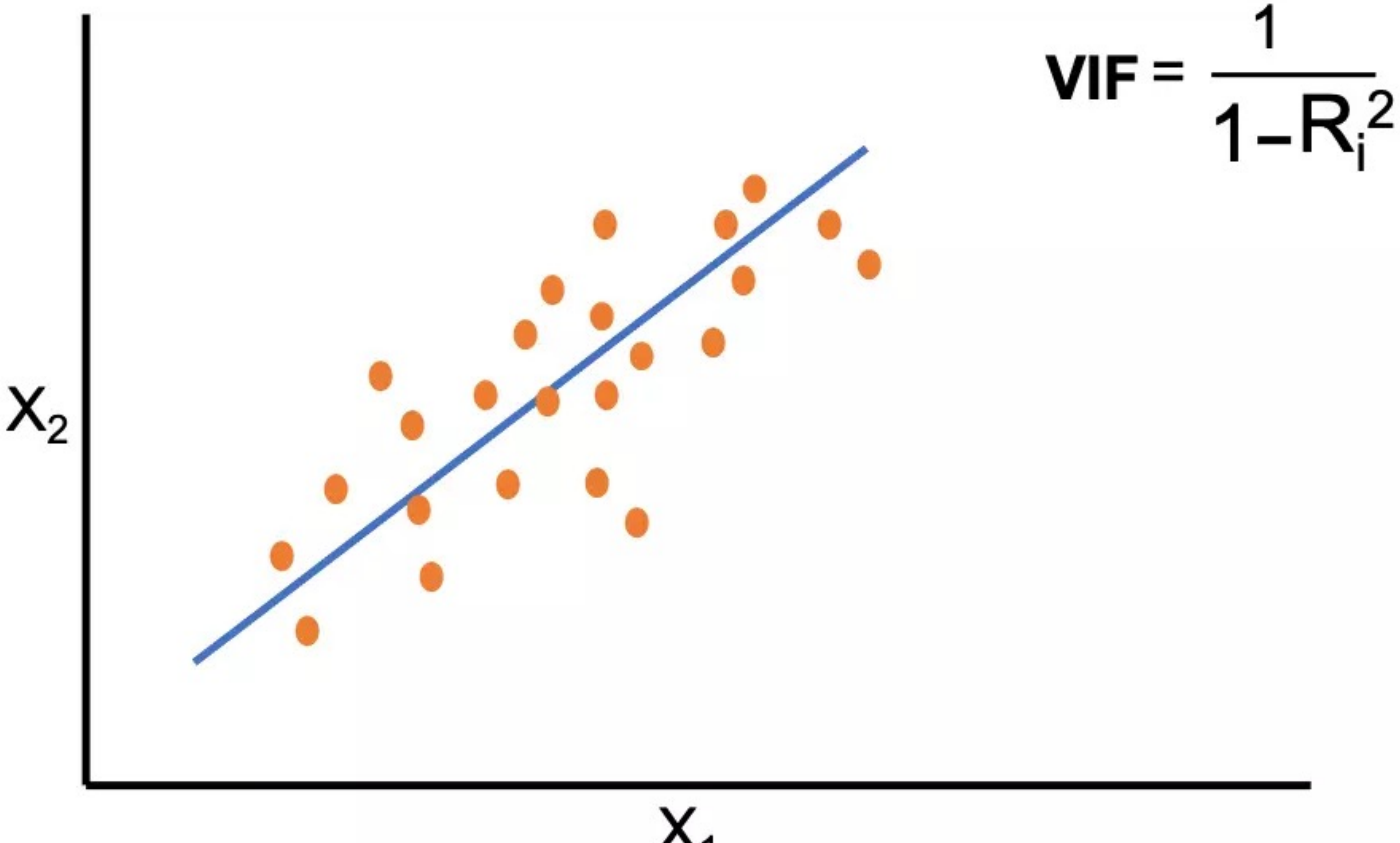
	Weight	Height	BMI
Weight	1.0000	0.5129	0.8668
Height	0.5129	1.0000	0.0220
BMI	0.8668	0.0220	1.0000



Scatterplot Matrix

Issues in model specification

vif - ***variance inflation factor,*** a measure of potential multicollinearity.



Outliers

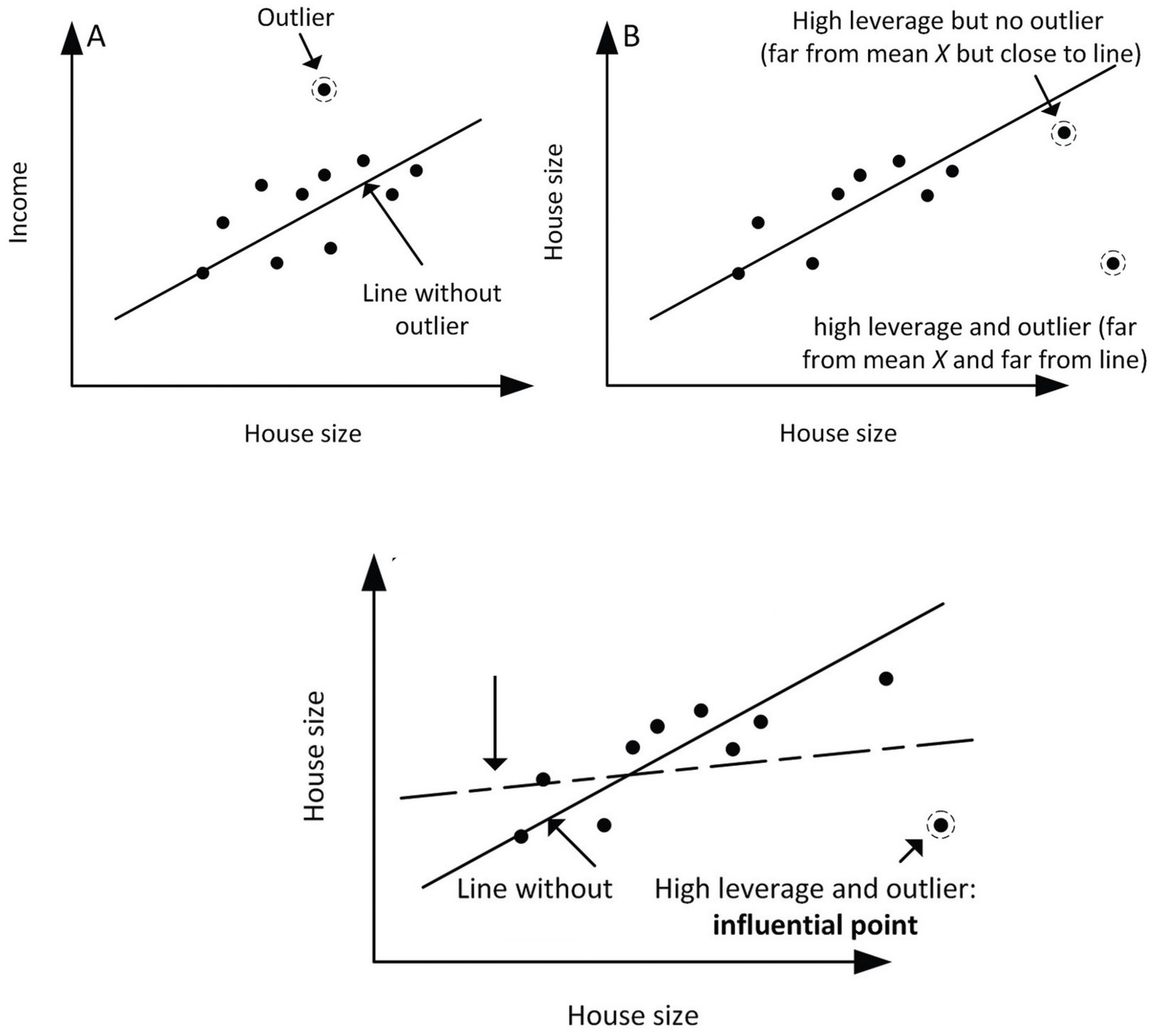
- An outlier is an observation that is unusually small or large.
- Several possibilities need to be investigated when an outlier is observed:
  - There was an error in recording the value.
  - The point does not belong in the sample.
  - The observation is valid.
- Identify outliers from the scatter diagram.

If a single observation (or small group of observations) substantially changes your results, you would want to know about this and investigate further. There are three ways that an observation can be unusual.

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.

Leverage: An observation with an extreme value on a predictor variable is called a point with high leverage. 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.

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.


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

Outlier removal is straightforward in most statistical software. However, it is not always desirable to remove outliers.

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. These leverage points can have an effect on the estimate of regression coefficients. A leverage point will inflate the strength of the regression relationship by both the statistical significance (reducing the ***p*-value** to increase the chance of a significant relationship) and the practical significance (increasing ***r*-square**).

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

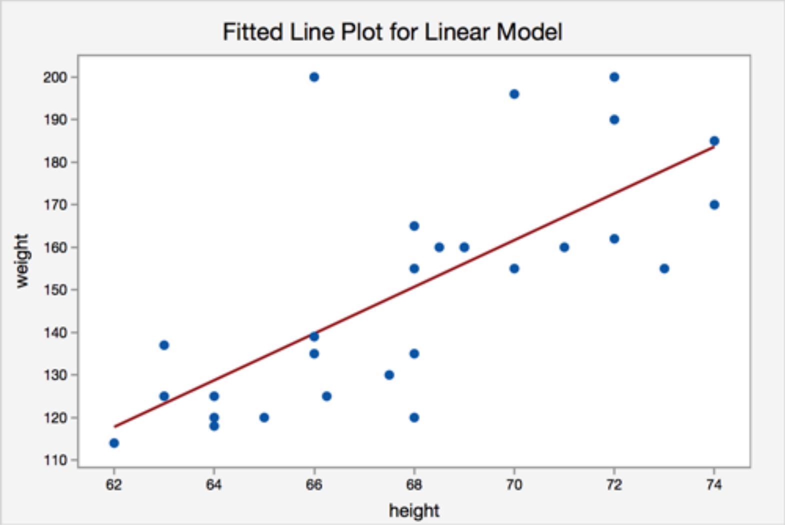
Influence: An observation is said to be influential if removing the observation substantially changes the estimate of coefficients. Influence can be thought of as the product of leverage and outlierness. Thus, influential points have a large influence on the fit of the model. One method to find influential 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.

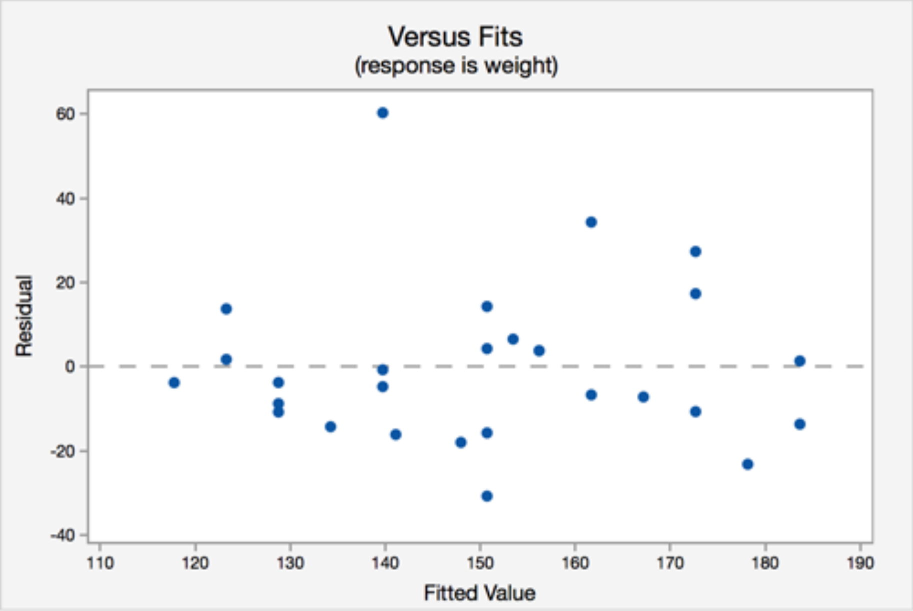
Any observation that has Cook's distance close to 1 or more, or that is substantially larger than other Cook's distances (highly influential data points), requires investigation.

Assumption 1: 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.

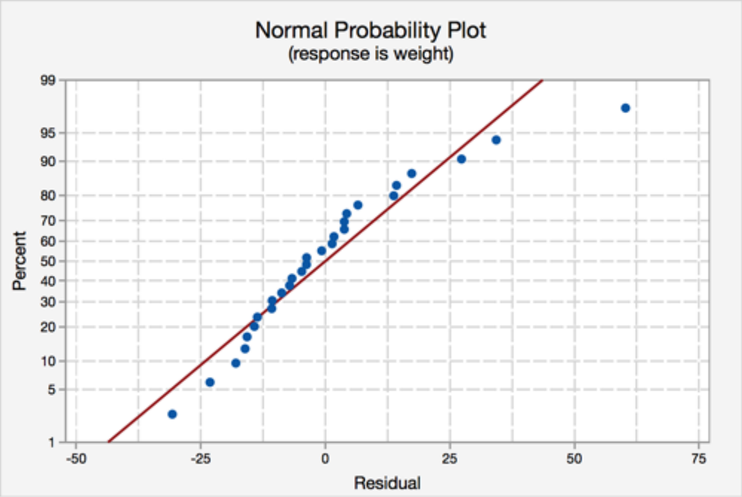
Assumption 2: Independence of errors - There is not a relationship between the residuals and weight.

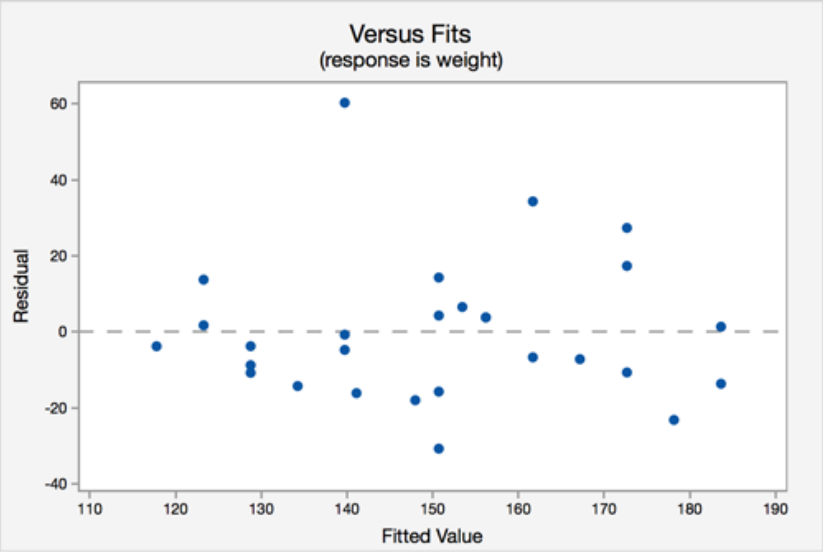


In the residuals versus fits plot, the points seem randomly scattered, and it does not appear that there is a relationship.

Assumption 3: Normality of errors - The residuals must be approximately normally distributed

Assumption 4: Equal Variances - The variance of the residuals is the same for all values of X.





Most of the data points fall close to the line, but there does appear to be a slight curving. There is one data point that stands out.

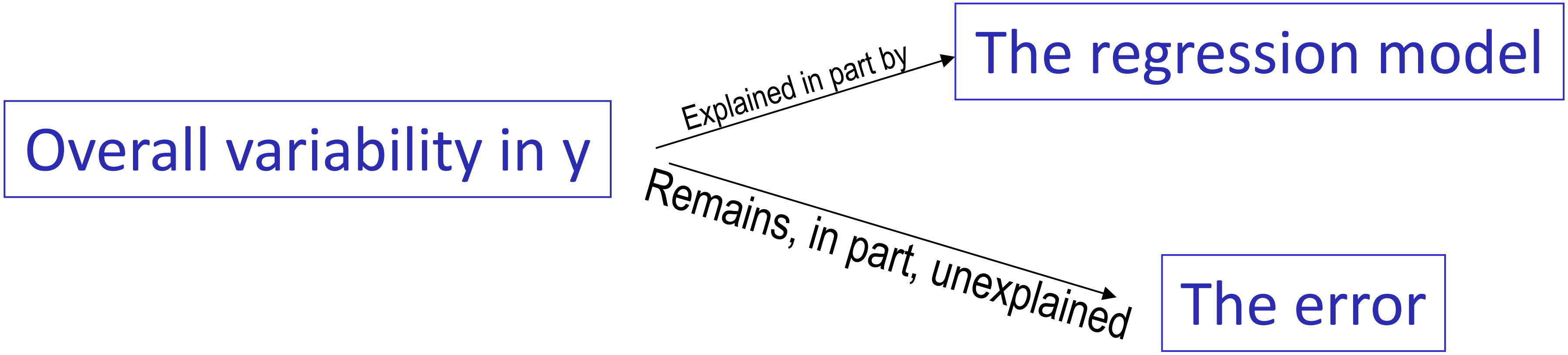
In this plot, there does not seem to be a pattern.

Procedure for Regression Diagnostics

- Develop a model that has a theoretical basis.
- Gather data for the two variables in the model.
- Draw the scatter diagram to determine whether a linear model appears to be appropriate.
- Determine the regression equation.
- Check the required conditions for the errors.
- Check the existence of outliers and influential observations
- Assess the model fit.
- If the model fits the data, use the regression equation.

5) Goodness of fit

The fit of the model



Overall variability in y

Explained in part by

The regression model

Remains, in part, unexplained

The error

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

Total Sum of Square

Regression Sum of Square

Error Sum of Square

The fit of the model

- RSE (Residual Standard Error): standard error of the residuals

Limitation: it is an absolute measure that strictly depends on the magnitude of Y

- R2 (Coefficient of determination) 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).
- 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)

- R2 (adjusted): it takes into account the number of Predictors

7) Multiple regression models: interpretation of coefficients

Multiple Linear Regression

More than one predictor…

$$y = b_0 + b_1x_1 + b_2x_2 + b_3x_3 + \ldots + b_kx_k + e$$

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 $\mathbf{b_j}$.

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

$$\hat{y}_i = -0.412 + 0.184\,x_{1i} \qquad (i = 1,\ 2,\ \ldots,\ 20)$$

Y= expenditures * 1000 euros
X_1=HH income * 1000 euros

Now we include HH size (***X_2***)

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

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

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

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

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

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

Example:

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

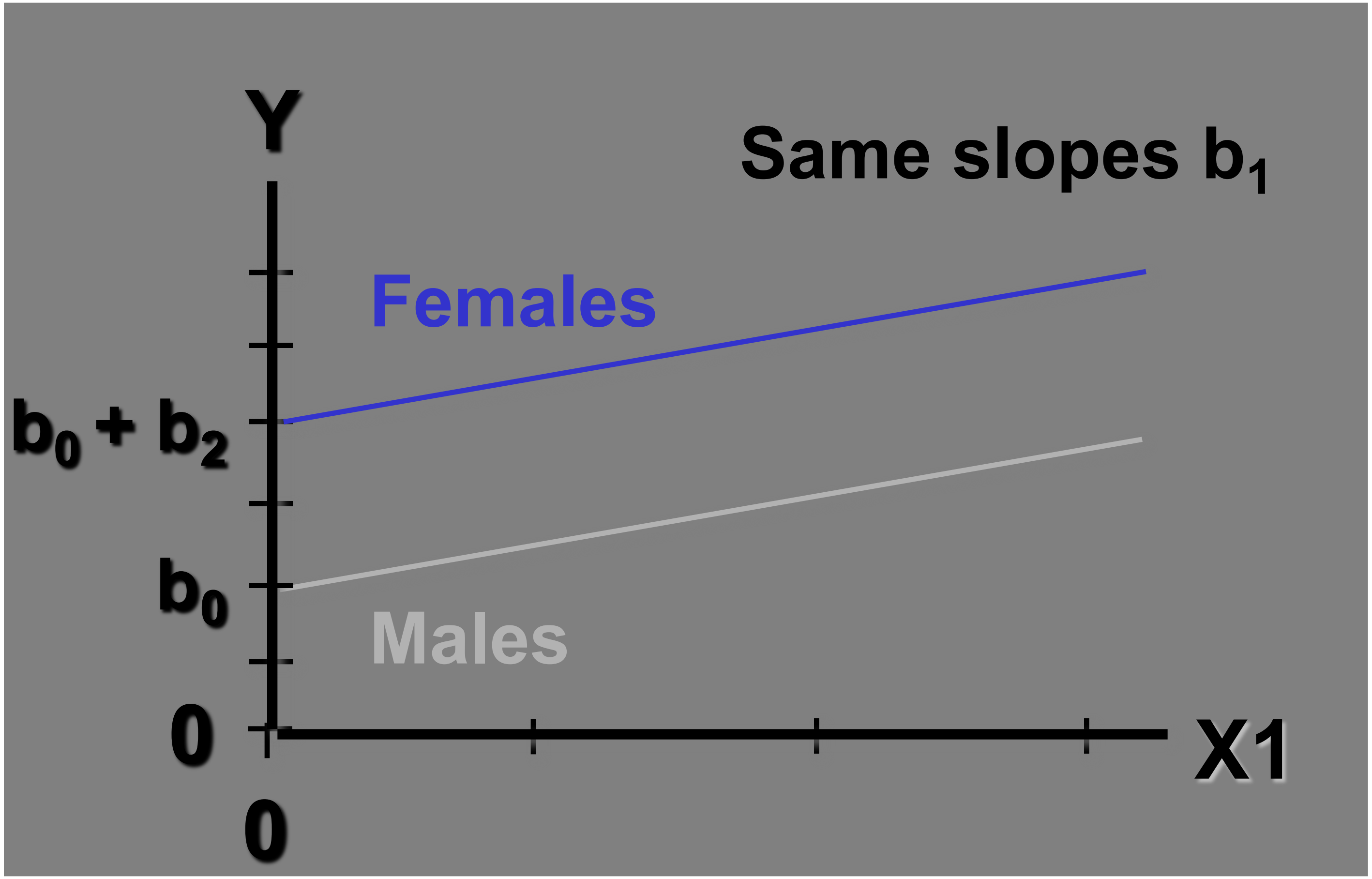
- What if the case *i* is a male?
- DFEMALE is 0 in case of male, so males are modeled as:

a + b_1AGE + e.

Dummy Variables

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

- 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

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.

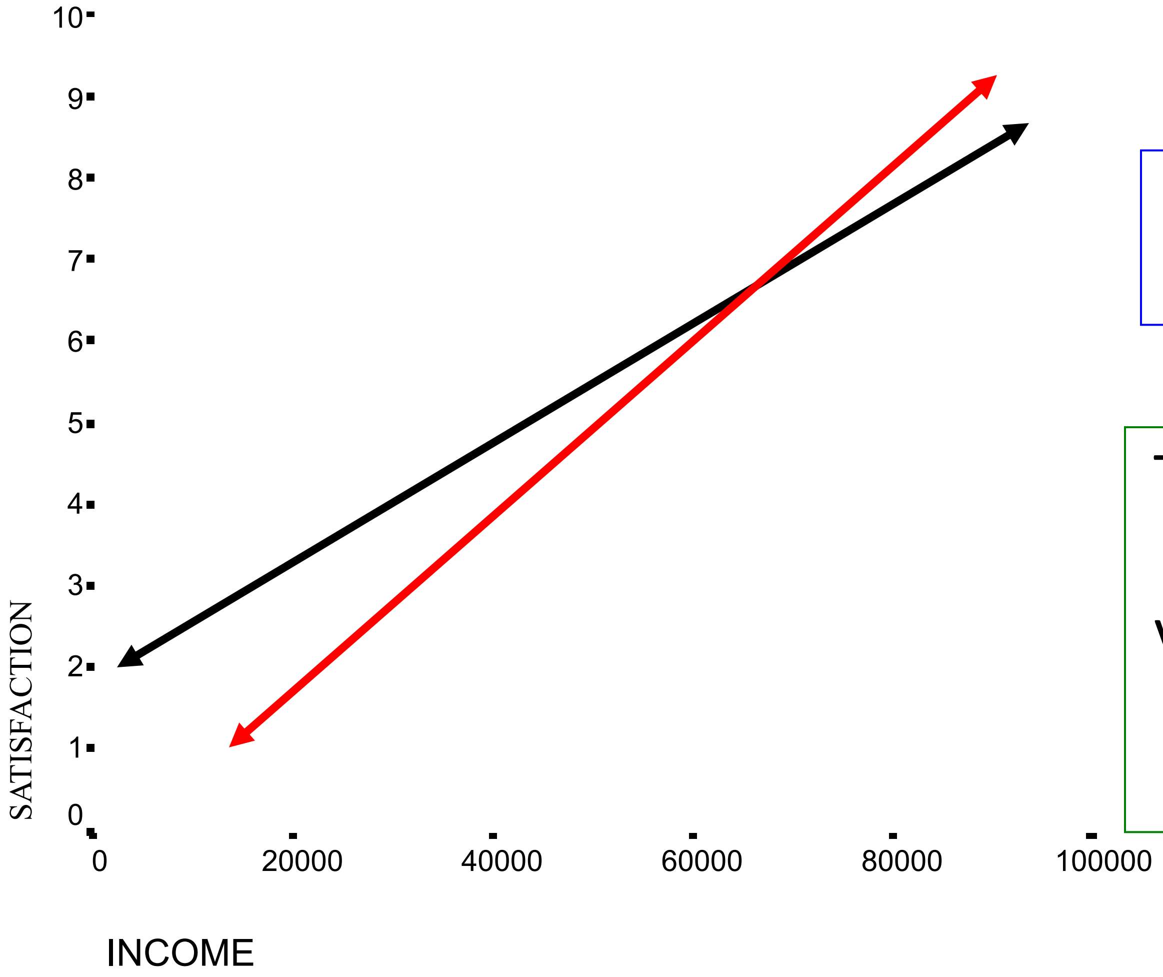
Interactions

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

- Example: Income and Satisfaction with life – 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

Interactions



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"

Interactions

- 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

Interactions

- 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.

Interactions

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.

Interactions

- 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

Interactions: continuous variables

- 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.

Interactions: continuous variables

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.

Interactions: dummy variables

- 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.

Interactions: dummy variables

- 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.

Interactions: final 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

Interactions: final remarks

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

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:

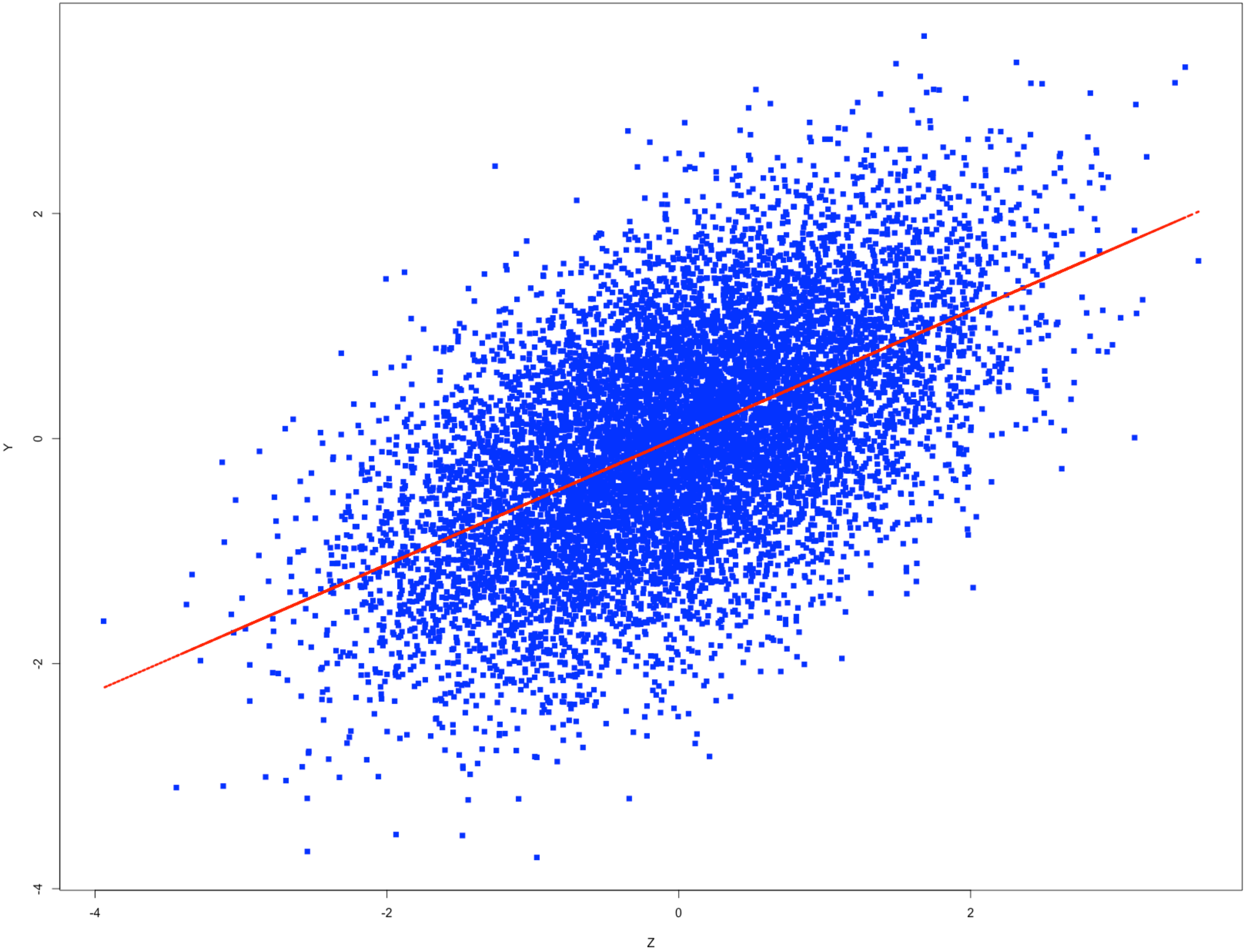
Predictor not significant and has the expected sign -> Keep it
Predictor not significant and does not have the expected sign -> Drop it
Predictor is significant and has the expected sign -> Keep it
Predictor is significant but does not have the expected sign -> Review, you may need more variables, it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

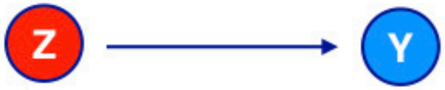
Gelman, Andrew, Jennifer Hill, *Data Analysis Using Regression and Multilevel/Hierarchical Models*, 2007

“Essentially all models are wrong, but some are useful.”

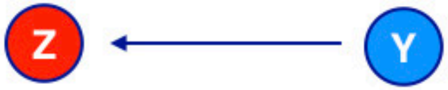
George Box, 1976

8) Correlation vs Causation

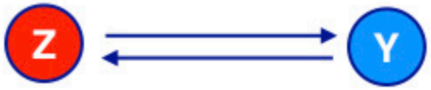

Y
2
0
-2
-4
-4
-2
0
2
Z



Z causes Y

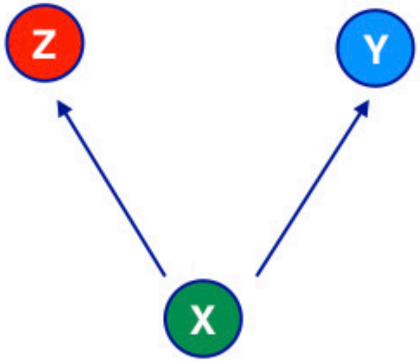


Y causes Z



Z causes Y,
Y causes Z

No causal relationship
between Z and Y



Both Z and Y are affected by
a third factor X
(*confounding* variable)

Correlation does not mean causation!