Linear Regression

Prof.ssa Chiara Seghieri

Laboratorio di Management e Sanità, Istituto di Management, L'EMbeDS
Scuola Superiore Sant'Anna, Pisa
c.seghieri@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, determinants of stock index,…

- The main motivations for using the technique:
  - Analyze the specific relationships between the variables X_s and the Y.
  - Predict the "future" value of Y from the values of the variables X_s

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.*

Notation

N: population size, number of observations in the population
n: sample size, number of observations in the sample
p: number of independent variables
x_{ij}: value of the *j* variable for the observation *i*, where *i*=1,2,…,n and *j*=1,2,…,p

$$X = \begin{pmatrix} x_{11} & \cdots & x_{1p} \\ \vdots & \ddots & \vdots \\ x_{n1} & \cdots & x_{np} \end{pmatrix}$$

yi: value of the dependent variable for the observation *i*, where *i*=1,2,…,n

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 $\longrightarrow$ y (Income)

Multivariate or multiple regression model

(Education) x_1

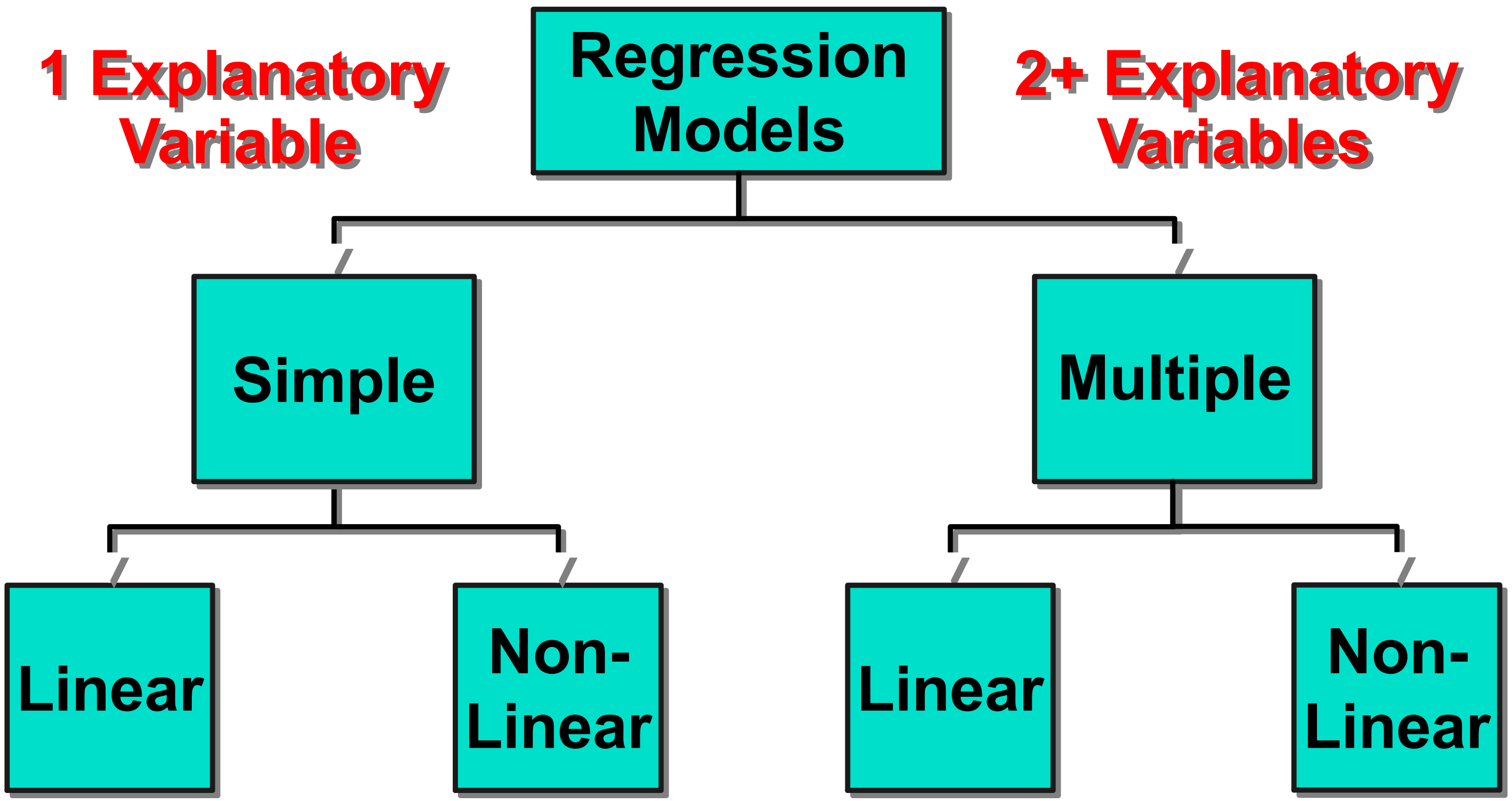
(Sex) x_2

(Experience) x_3

(Age) x_4

y (Income)

Types of Regression Models

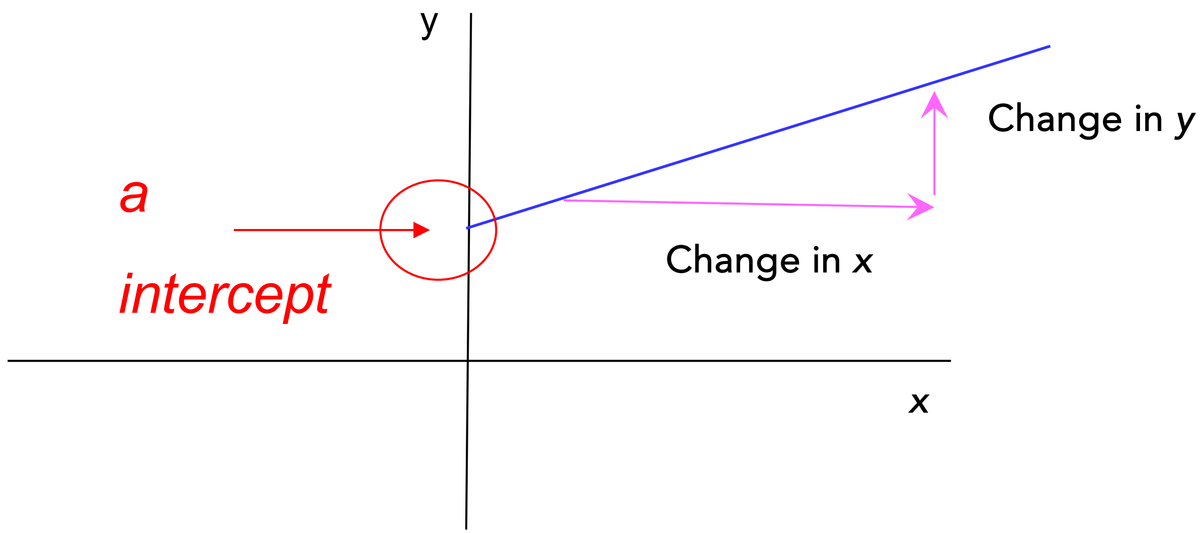


2) Simple Linear Regression Model

Let's start from an analysis for a quantitative response and a single quantitative explanatory variable.

What is “Linear”?

- Remember this: *y=a+bx?*

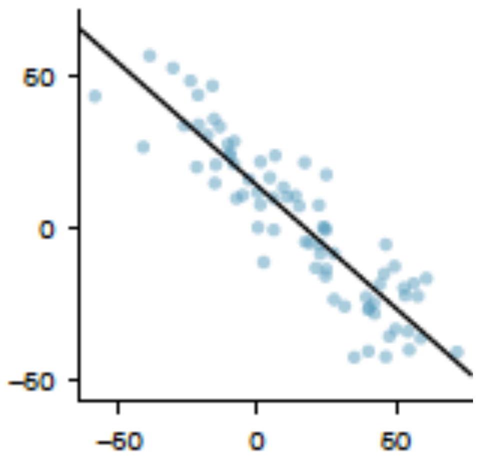


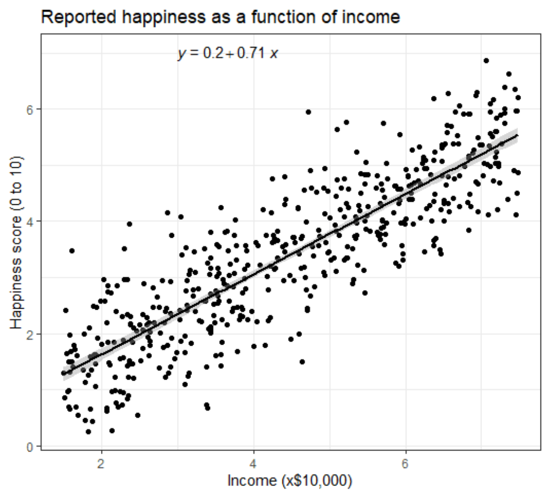
b=slope=change in y/change in x= $(y_2-y_1)/(x_2-x_1)$

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.

The relationship between the response Y and the predictor X can be written in general form as:

$$Y=f(X)+\varepsilon$$

f is some fixed but unknown function of X, is the systematic information that X provides about Y,

ε is the random error term which is independent of X and has mean zero, it cannot be predicted by using X and might contain unmeasured variables that are useful in predicting Y and/or unmeasurable variation (i.e. subjective feeling).

The idea is to estimate f.

Broadly speaking, there can be **parametric** or **non-parametric** methods. The first start making an assumption about the functional form (or shape) of f, the latter do not make any assumption about the form of f.

If f is approximated by a linear function, the general form of the simple linear regression (population regression model) model is:

$$Y=\beta_0+\beta_1 X+\varepsilon$$

For an individual observation: $y_i=\beta_0+\beta_1 x_i+\varepsilon_i$

Where:

β_0 is the population **intercept**, expected value of Y when X=0

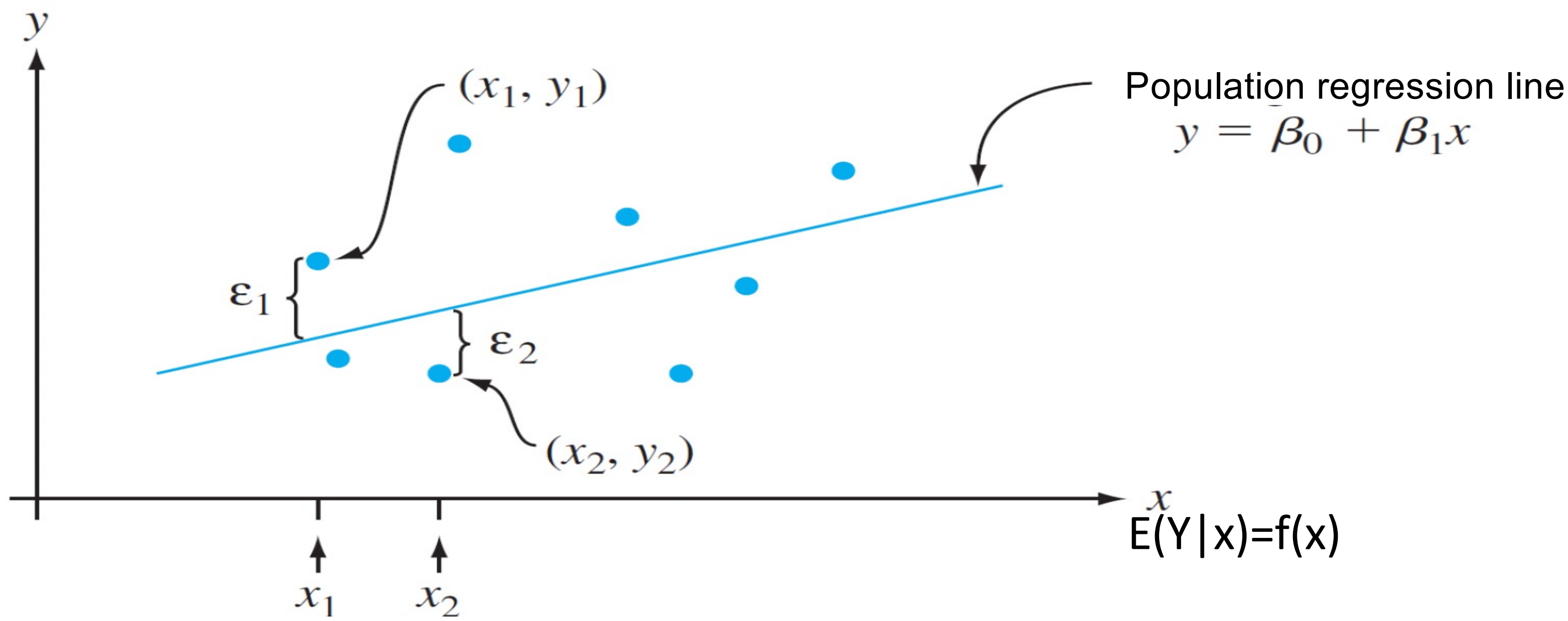
β_1 is the population **slope**, average change in Y associated with one unit increase in X

ε is the random **error** term, independent of X with mean zero, what we miss with this simple model.

The structural model says that the explanatory and outcome variables are linearly related such that the population mean of the outcome for any x value is $\beta_0 + \beta_1 X$.

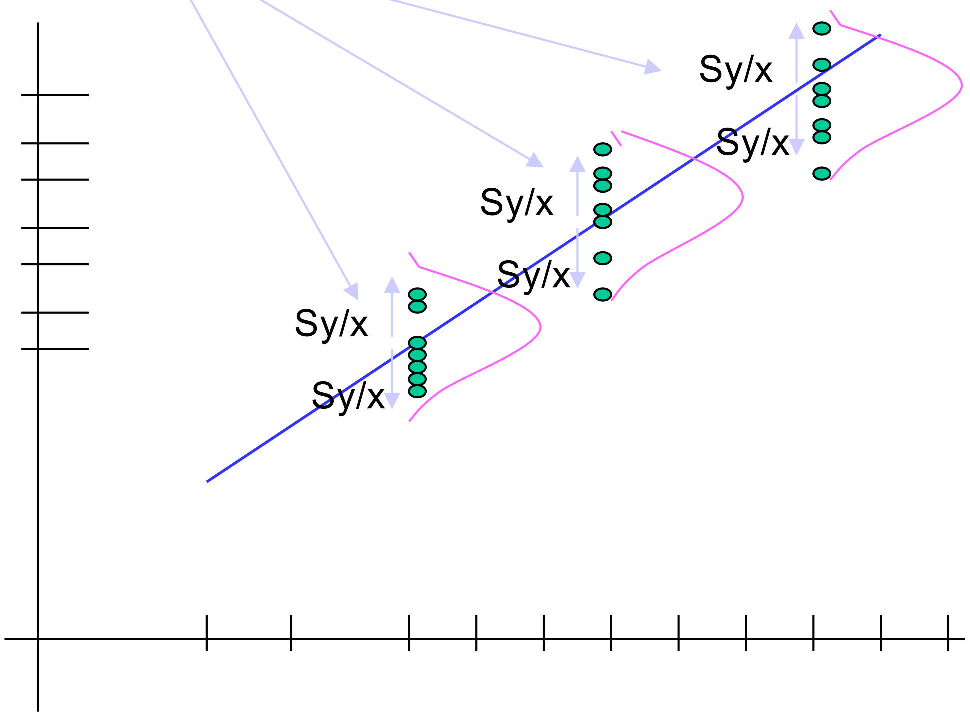
That is: for each value of X=x the population mean of Y (over all of the subjects who have that particular value "x" for their explanatory variable) is:

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


y
(x_1, y_1)
Population regression line
$y = \beta_0 + \beta_1 x$
ε_1
ε_2
(x_2, y_2)
x
E(Y|x)=f(x)
x_1
x_2

The error model that we use is that for each particular x, if we have or could collect many subjects with that x value, their distribution around the population mean is Gaussian with a spread σ^2, that is the same value for each value of x (and corresponding population mean of y).

The error model underlying a linear regression analysis includes the assumptions: Normality with mean zero, equal spread, and independent errors.



Of course, the value of σ^2 is an unknown parameter, and we can make an estimate of it from the data.

Normal curves represent the Normally distributed outcomes (Y values) at each of fixed x values.

The curves have the same spreads represents the equal variance assumption.

The means of the Normal curves fall along a straight line represents the linearity assumption.

Assumptions of the Model

- Linear: 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.

- Same spread around the regression line

- Independent errors: 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.

For example, in predicting time to reach a finish line from age, knowing that the first subject took 4 seconds longer than the mean of all possible subjects with the same age should not tell us anything about how far the next subject's time should be above or below the mean.

Note that the model assumes a Normal distribution of the outcome for each value of the explanatory variable.
(It is equivalent to say that all of the errors are Normally distributed.)
Implicitly this indicates that the outcome should be a continuous quantitative variable.

Fortunately, regression is quite robust to deviations from the Normality assumption, and it is OK to use discrete or continuous outcomes that have at least a moderate number of different values, e.g., 10 or more.

To make inference about the 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 (OLS)**.

3) Estimation of Coefficients: OLS method

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**

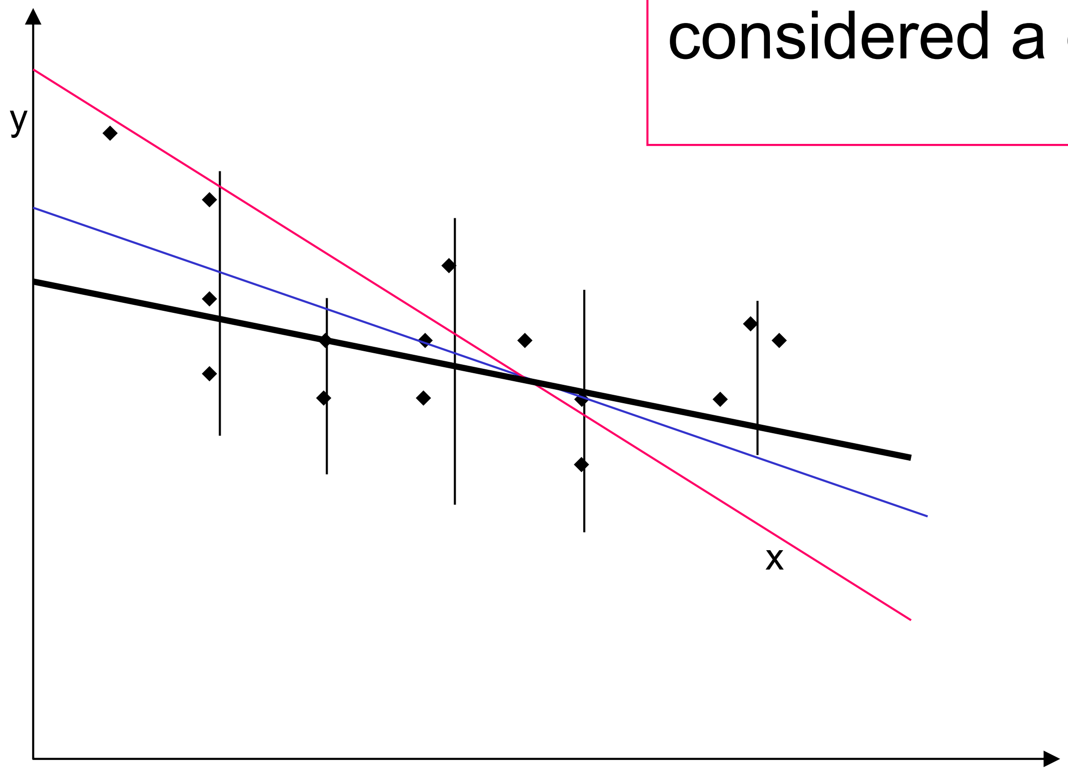
The values of β_0, β_1, and σ^2 will almost never be known to an investigator.

Instead, sample data consists of n observed pairs (x_1, y_1), … , (x_n, y_n), from which the model parameters and the true regression line itself can be estimated.

- 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

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. We use calculus to find the values of β_0 and β_1 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.

The sample data are only one of the possible determinations, that is, the one that was "extracted"

-As the sample and, therefore, the available data change, the estimated regression line will also change.

Least squares: Coefficient Equations

LS minimize:

$$\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = \sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)^2$$

- Sample slope

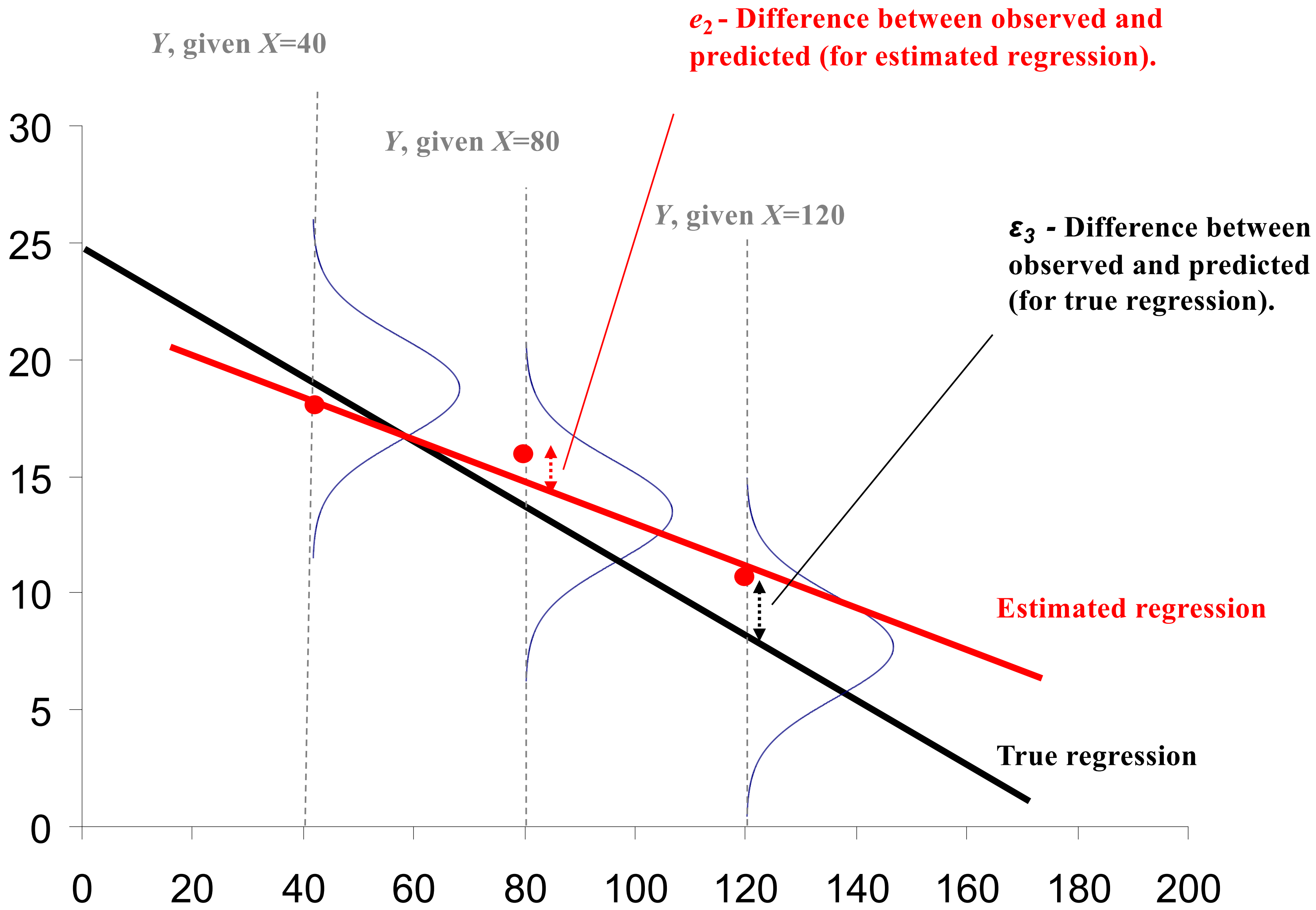
$$\hat{\beta}_1 = \frac{SS_{xy}}{SS_{xx}} = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})^2}$$

- Sample Y - intercept

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

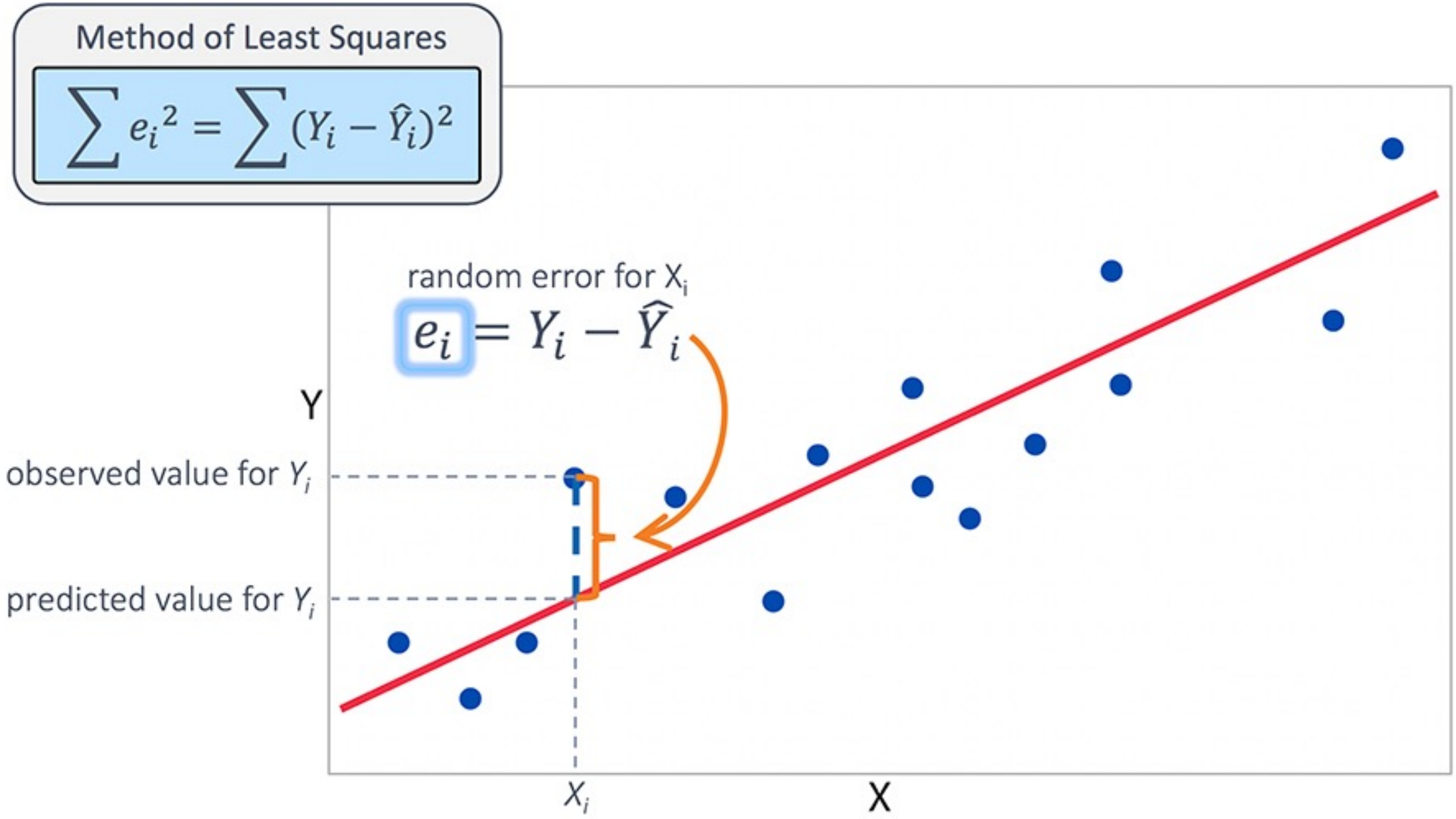
Residual

- 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).
Y, given X=40
Y, given X=80
Y, given X=120
ε_3 - Difference between observed and predicted (for true regression).
Estimated regression
True regression
30
25
20
15
10
5
0
0
20
40
60
80
100
120
140
160
180
200

OLS: graphical intuition



from population to sample regression line

- The error term used in the true population regression line, becomes the residual in the sample regression line
- The coefficients $\widehat{\beta_0}$and $\widehat{\beta_1}$, are estimators of β_0 and β_1
- Compliance with assumptions, allows us to say that the OLS estimator b, is the best correct and linear estimator of β.
- We thus say that $\widehat{\beta}$ is the BLUE (Best Linear Unbiased Estimator) estimator.

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

Note that:

It is sufficient for the unbiasedness of the OLS estimator that the error terms have zero mean and are independent of all explanatory variables, even in the presence of autocorrelation and heteroschedasticity.

In the presence of autocorrelation and heteroschedasticity the OLS estimator can still be correct and consistent, but only relatively efficient (it is no longer BLUE).

- In these cases, the OLS estimator, although correct, is not the best

- Two possibilities open up at this point:

1] One can derive a new estimator (GLS or weighted least squares) that is BLUE

2] One can continue to use the OLS estimator, correcting the standard errors to admit the possibility of heteroschedasticity and/or autocorrelation

….otherwise

- Finally, remember that in many cases the presence of heteroschedasticity and/or autocorrelation, indicates incorrect specification of the model.

- Therefore, one can intervene in another way, namely reconsidering the model.

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 ($\hat{\beta}_0$)
   - 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

How close are the estimates of the parameters to the true values?

$$SE(\hat{\beta}_0)^2 = \sigma^2 \left[\frac{1}{n} + \frac{\bar{x}^2}{\sum_{1=1}^{n}(x_i - \bar{x})^2}\right]$$

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

Where σ^2=Var(ε)

σ^2 is not known and but can be estimated from the data. The estimate is known as residual standard error:

$$\hat{\sigma}^2 = \frac{1}{n-2}\sum_{1=1}^{n}(y_i - \hat{y}_1)^2$$

SE can be used in confidence interval (CI) formulas and hypothesis testing procedures:

$$\hat{\beta}_1 \pm 2 * SE(\hat{\beta}_1)$$

(Thanks to the normality assumption of the errors!!)

Testing the Slope

We can draw inference by testing:

H_0: $\beta_1 = 0$

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

We need to determine whether the estimate is sufficienty far from zero.
In practice we compute the t statistic:

$$t = \frac{\hat{\beta}_1 - 0}{SE(\hat{\beta}_1)}$$

Which measures the number of standard deviations that $\hat{\beta}_1$ is away from zero
If there is no relationship between Y and X then we expect that it has a t-distribution with n-2 df

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 <u>lower</u> than the chosen size (5%). **A small p-value indicates that it is unlikely to observe such a substantial association between the predictor and the response due to chance, in the absence of any real association between the predictor and the response**. In this case, the coefficient of β is statistically significant in explaining the city homicide rates.

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.

4) 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

Due to th epresence of the error term in the regression equation we would not be able to perfectly predict Y from X, if the sum of square of the error is quite large then the model does not fit the data well

The model in STATA

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
```

Model: Model Sum of Squares (MSS)
Residual: Residual Sum of Squares (RSS)
Total: Total Sum of Squares (TSS)

Root MSE: square root of the Average Residual Sum of Squares=Residual Standard Error: homicides rates in each city deviate from the true regression line of about 5 unit on average.

The fit of the model

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

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

- R^2 (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). If it is near to zero this might occur because the linear model is wrong or the error variance σ^2 is high or both.
- F test in the regression output, 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)

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.

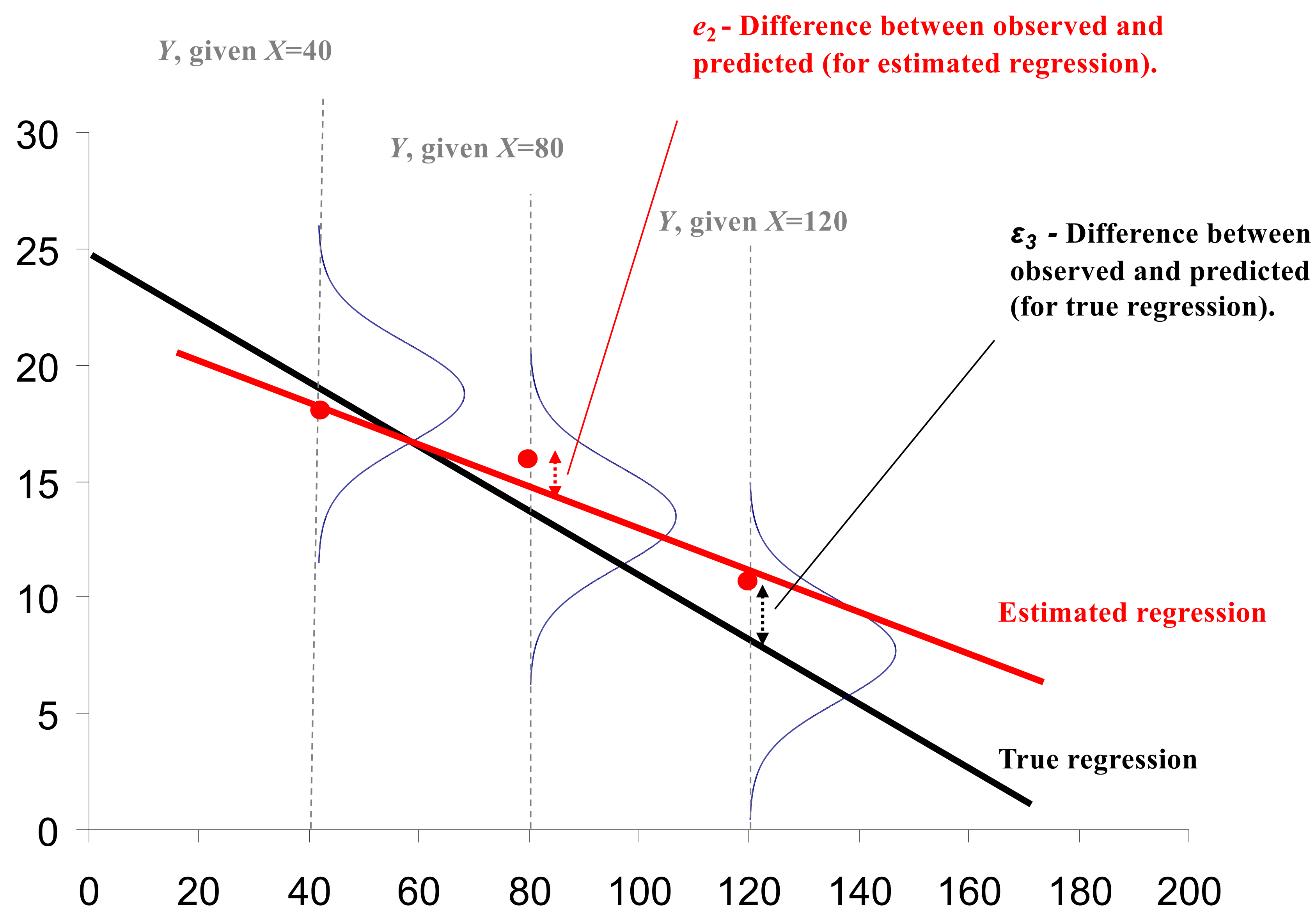
2. **Independence of errors**: pay attention to the type of data, eg panel, time series.

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).
Y, given X=40
Y, given X=80
Y, given X=120
ε_3 - Difference between observed and predicted (for true regression).
Estimated regression
True regression
30
25
20
15
10
5
0
0
20
40
60
80
100
120
140
160
180
200

Regression Diagnostics

How can we diagnose violations of these conditions?

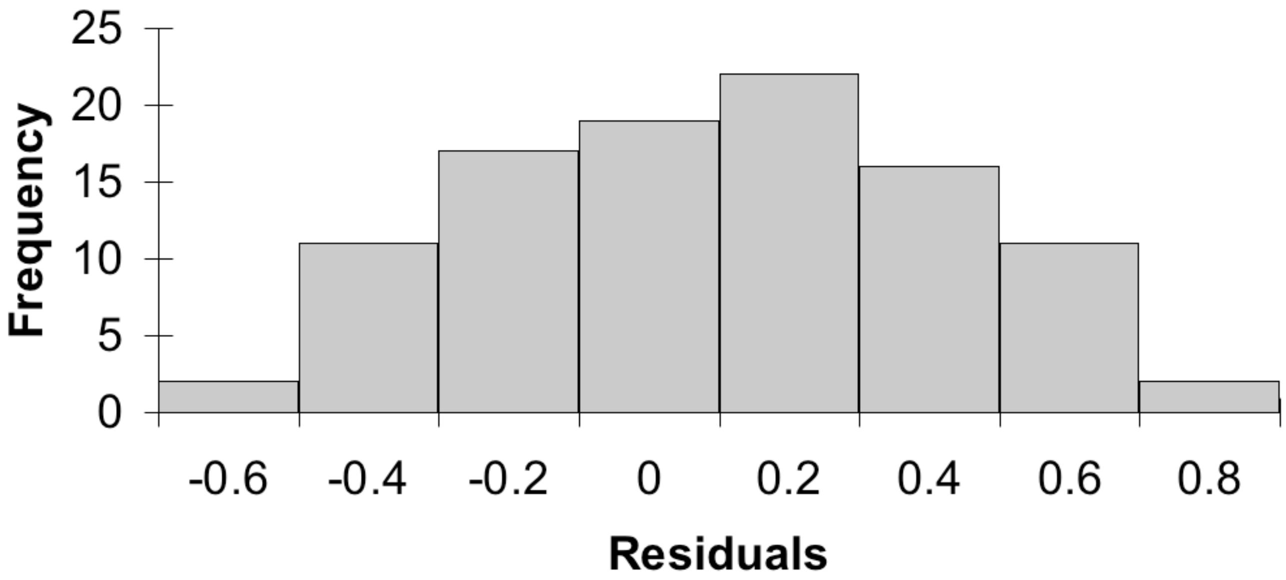
➔ **Residual Analysis**, that is, examine the ***differences*** between the actual data points and those predicted by the linear equation.

➔ 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

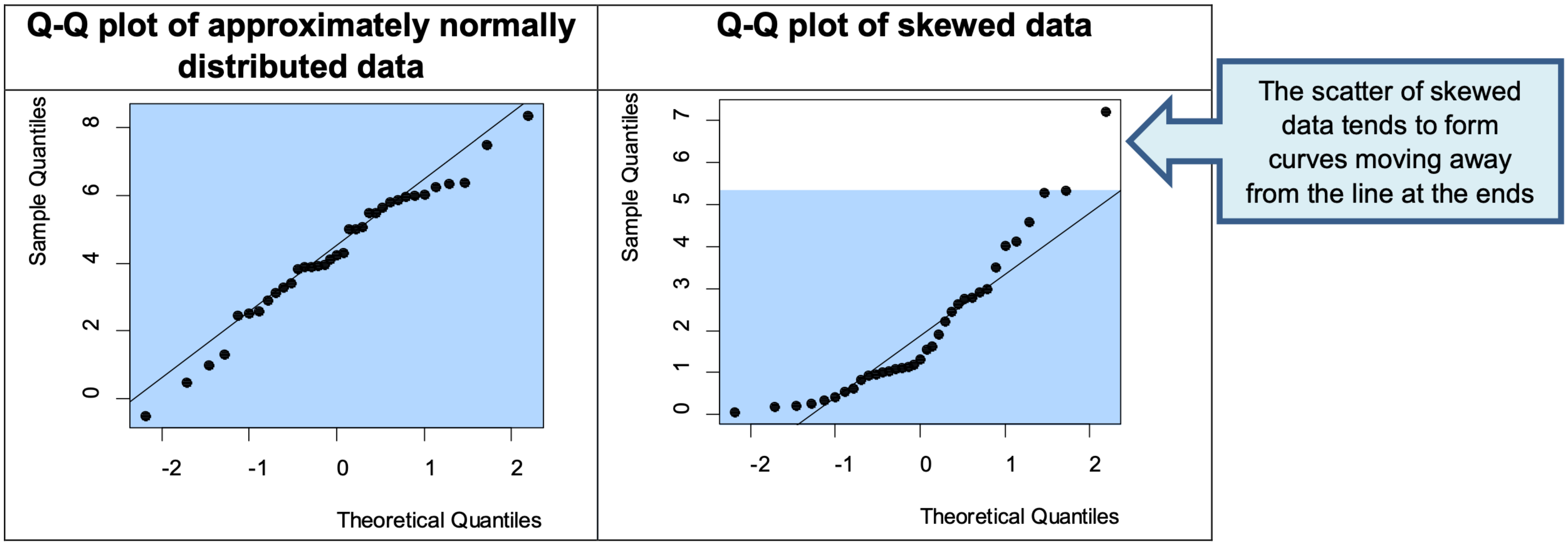
We can take the residuals and put them into a histogram to visually check for normality…



…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.



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

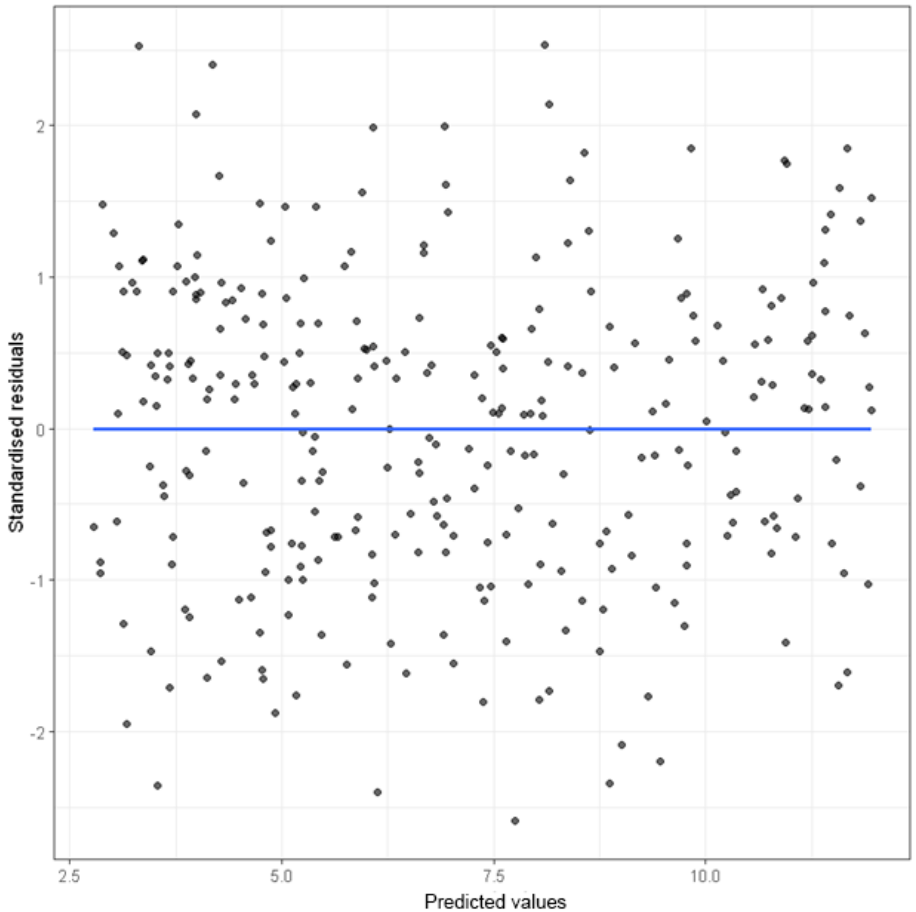
Regression Inference is robust against moderate lack of Normality. On the other hand, outliers and influential observations can invalidate the results of inference for regression. What to do?

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

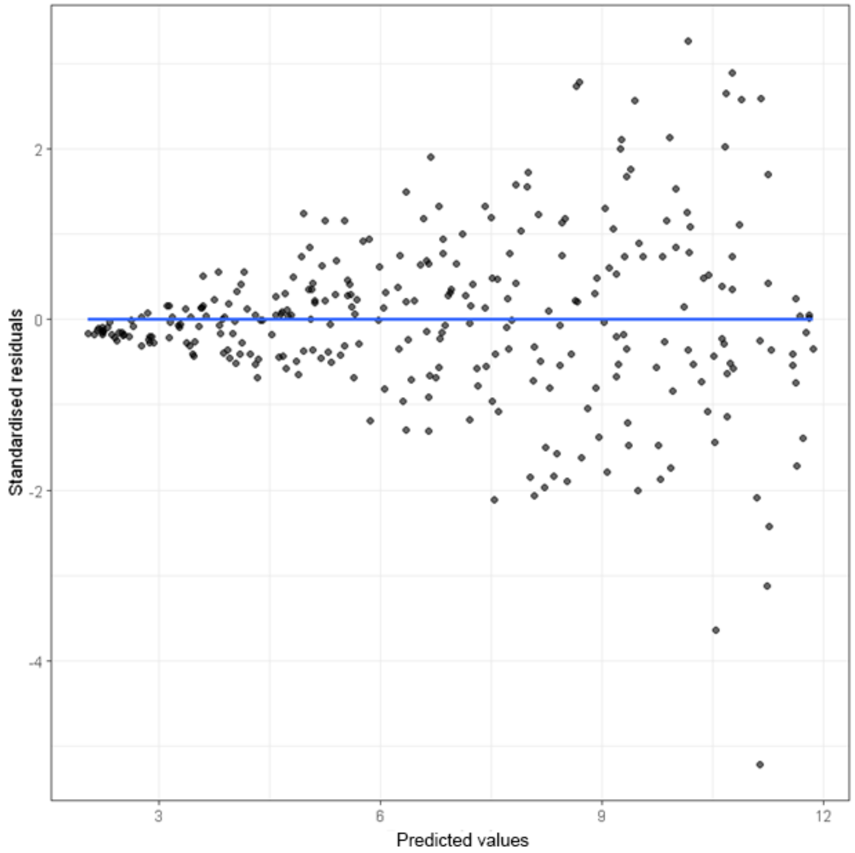
Use non-parametric methods.

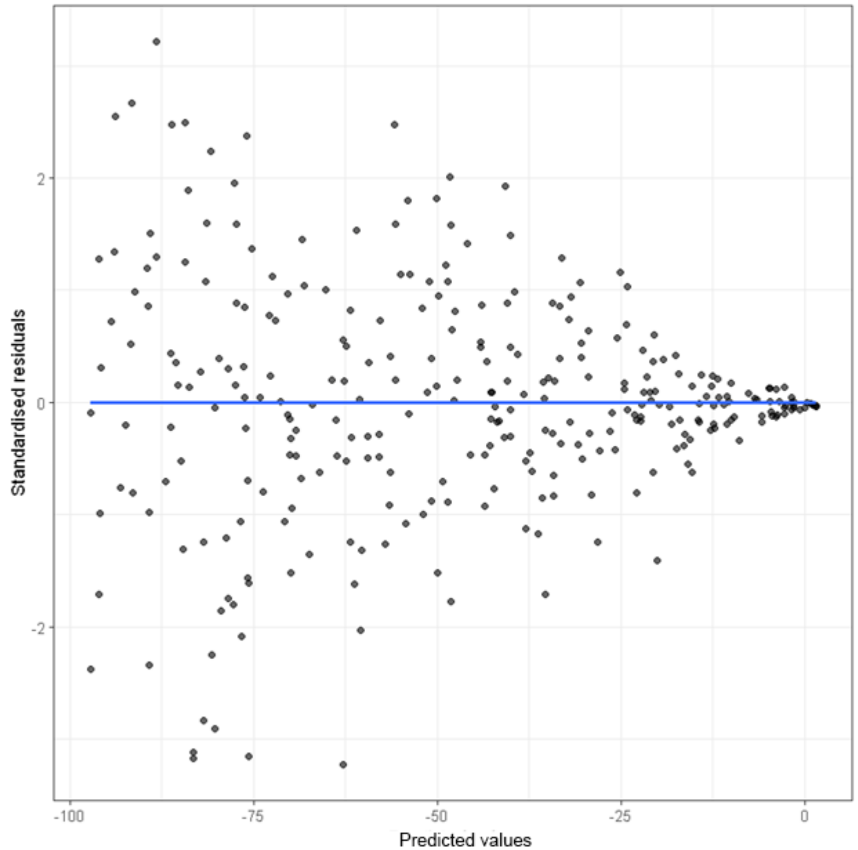
The plot of residuals versus predicted values is useful for checking the assumption of **linearity** and **homoscedasticity. Often, we consider s**tandardized residual: residual divided by an estimate of its standard deviation. They quantify how large the residuals are in standard deviation units.

When both the assumption of linearity and homoscedasticity are met, will be randomly scattered.

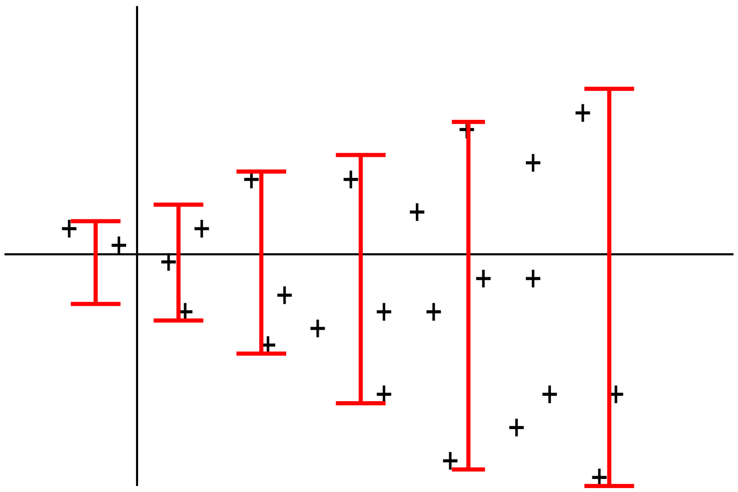


When the homoscedasticity assumption is violated, the “spread” of the points across predicted values are not the same. Heteroscedasticity results in biased standard errors. The following are two plots that indicate a violation of this assumption.

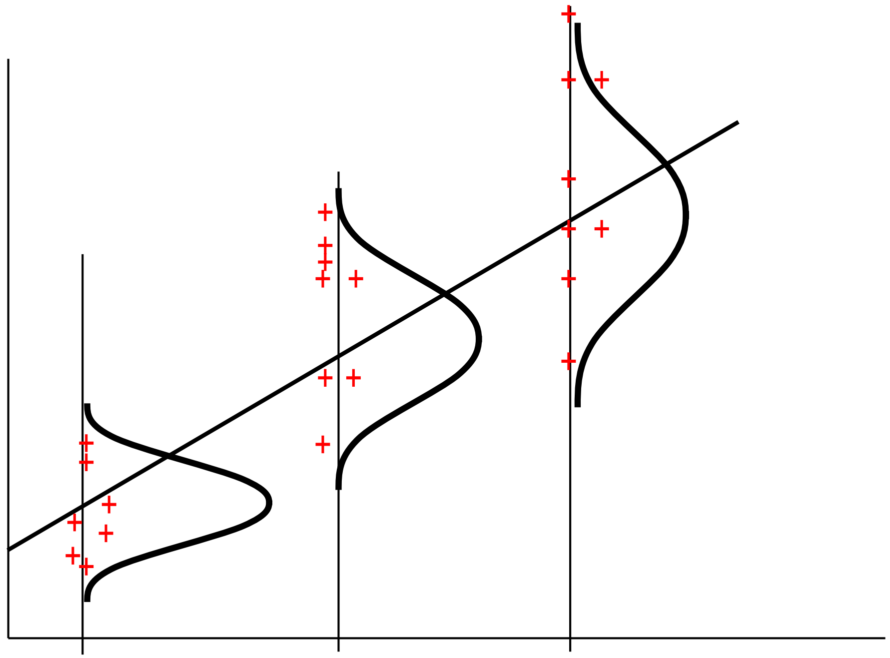




Heteroscedasticity



The spread increases with $\hat{y}$



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)

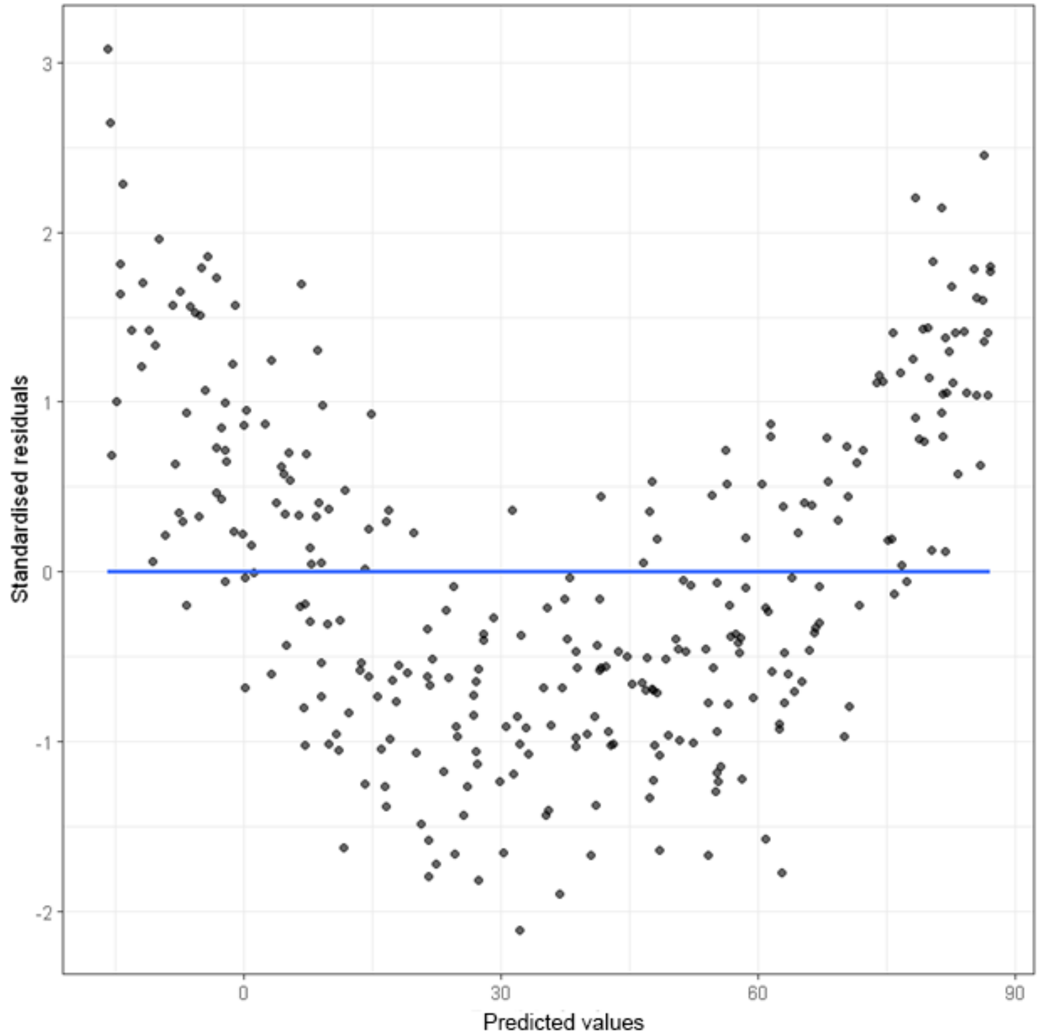
```
. estat hettest

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 linearity

When the linearity assumption is violated, the points in the residual plot will not be randomly scattered. Instead, the points will often show some “curvature”.



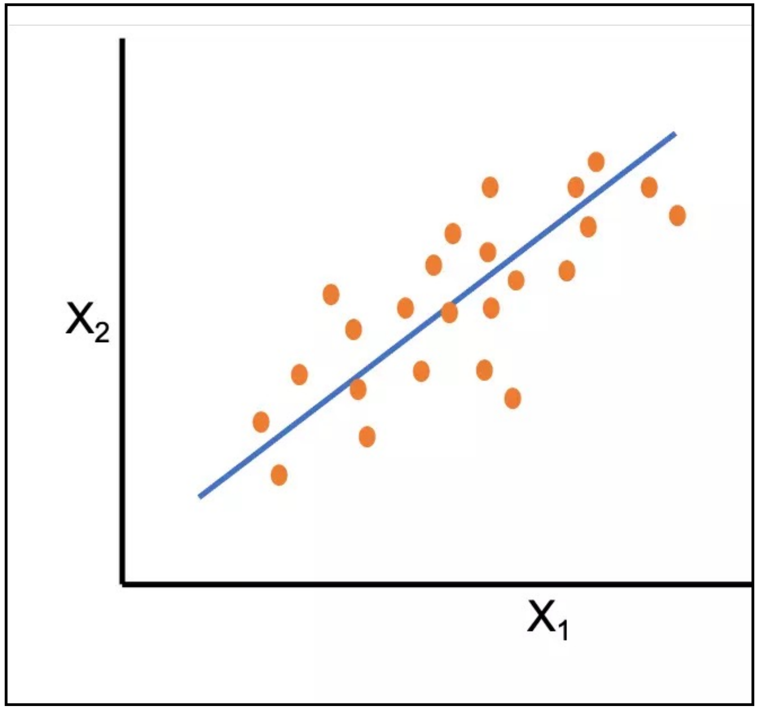
Issues in model specification

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

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



the variance inflation factor for the j^{th} predictor is:

$$VIF_j = \frac{1}{1 - R_j^2}$$

where R_j^2 is the R^2-value obtained by regressing the j^{th} predictor on the remaining predictors

A VIF of 1 means that there is no correlation among the j^{th} predictor and the remaining predictor variables, and hence the variance of b_j is not inflated at all. The general rule of thumb is that VIFs exceeding 4 warrant further investigation, while VIFs exceeding 10 are signs of serious multicollinearity requiring correction.

Influential Points

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.

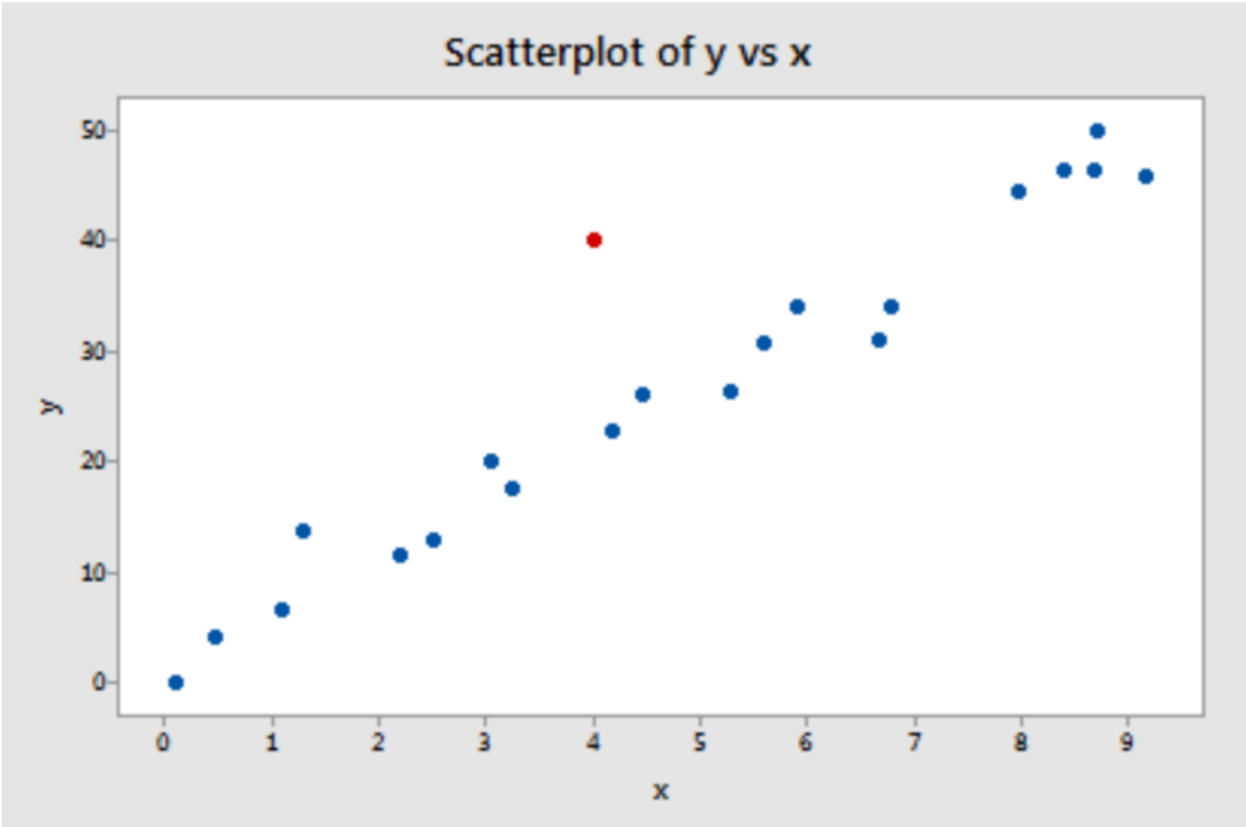
Outliers: In linear regression, an outlier is an observation with large residual. In other words, it is an observation whose dependent-variable value is unusual given its values on the predictor variables. An outlier may indicate a sample peculiarity or may indicate a data entry error or other problem.

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.

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 but very extreme value.
- Identify outliers from the scatter diagram.


Scatterplot of y vs x
y
x
0
10
20
30
40
50
0
1
2
3
4
5
6
7
8
9

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 .

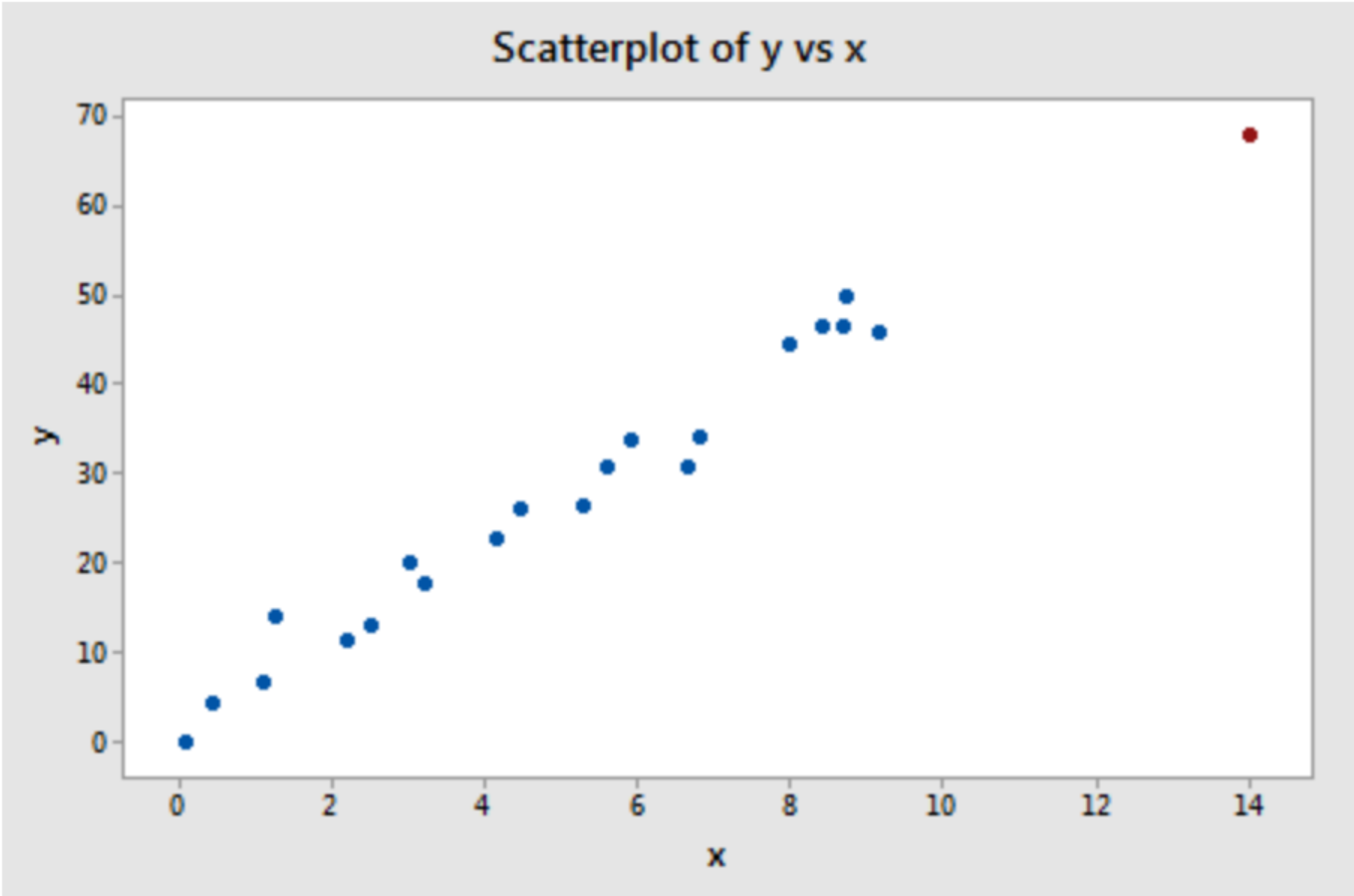
We therefore should pay attention to residuals that exceed +2 or -2, and get even more concerned about residuals that exceed +2.5 or -2.5 and seriously concerned with residuals that exceed +3 or -3.

Not every outlier or high-leverage data point strongly influences the regression analysis. The researcher should always determine if the regression analysis is highly influenced by one or more data points.

Of course, the easy situation occurs for simple linear regression, when we can rely on simple scatter plots to elucidate matters. In the multiple regression situation, we have to rely on various measures to help us determine whether a data point is an outlier, high leverage, or both. Once we've identified such points we then need to see if the points are actually influential.

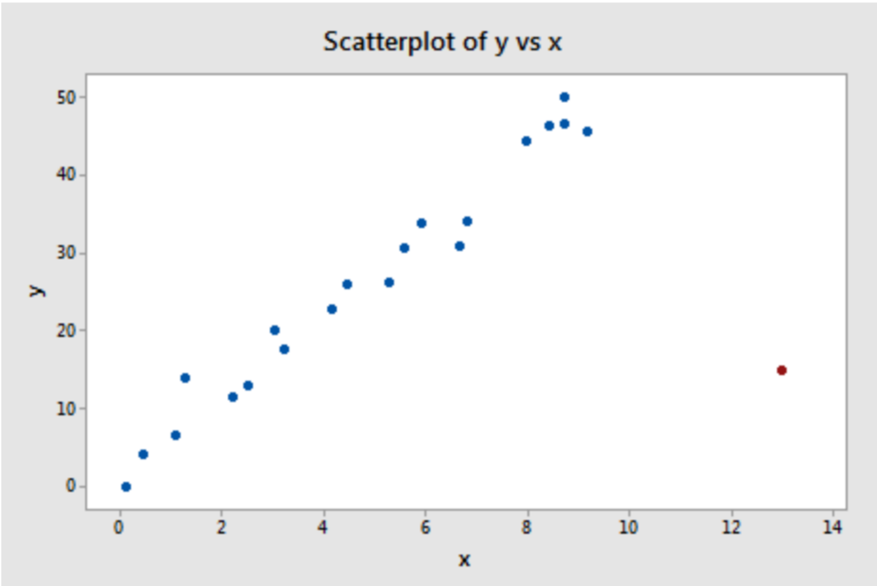
Leverage: A leverage point is defined as an observation that has a value of x that is far away from the mean of x. These leverage points can have an effect on the estimate of regression coefficients. A leverage point will inflate the strength of the regression relationship by both the statistical significance (reducing the ***p*-value** to increase the chance of a significant relationship) and the practical significance (increasing ***r*-square**).

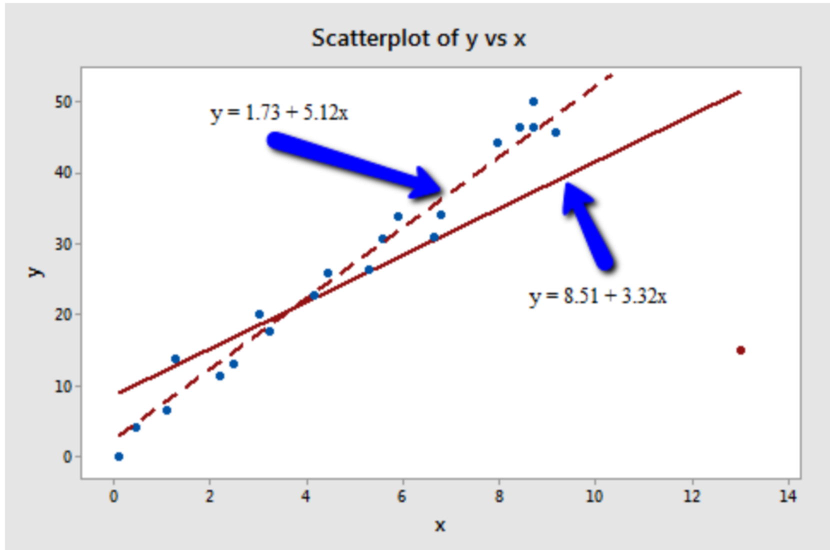
Leverage - for measuring “unusualness” of x’s:
Generally, a point with leverage greater than (2k+2)/n should be carefully examined. Here k is the number of predictors and n is the number of observations


Scatterplot of y vs x
y
70
60
50
40
30
20
10
0
x
0
2
4
6
8
10
12
14

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.


Scatterplot of y vs x
y
x
0
10
20
30
40
50
0
2
4
6
8
10
12
14


Scatterplot of y vs x
y = 1.73 + 5.12x
y = 8.51 + 3.32x
y
x
0
10
20
30
40
50
0
2
4
6
8
10
12
14

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.

7) Multiple regression models: interpretation of coefficients

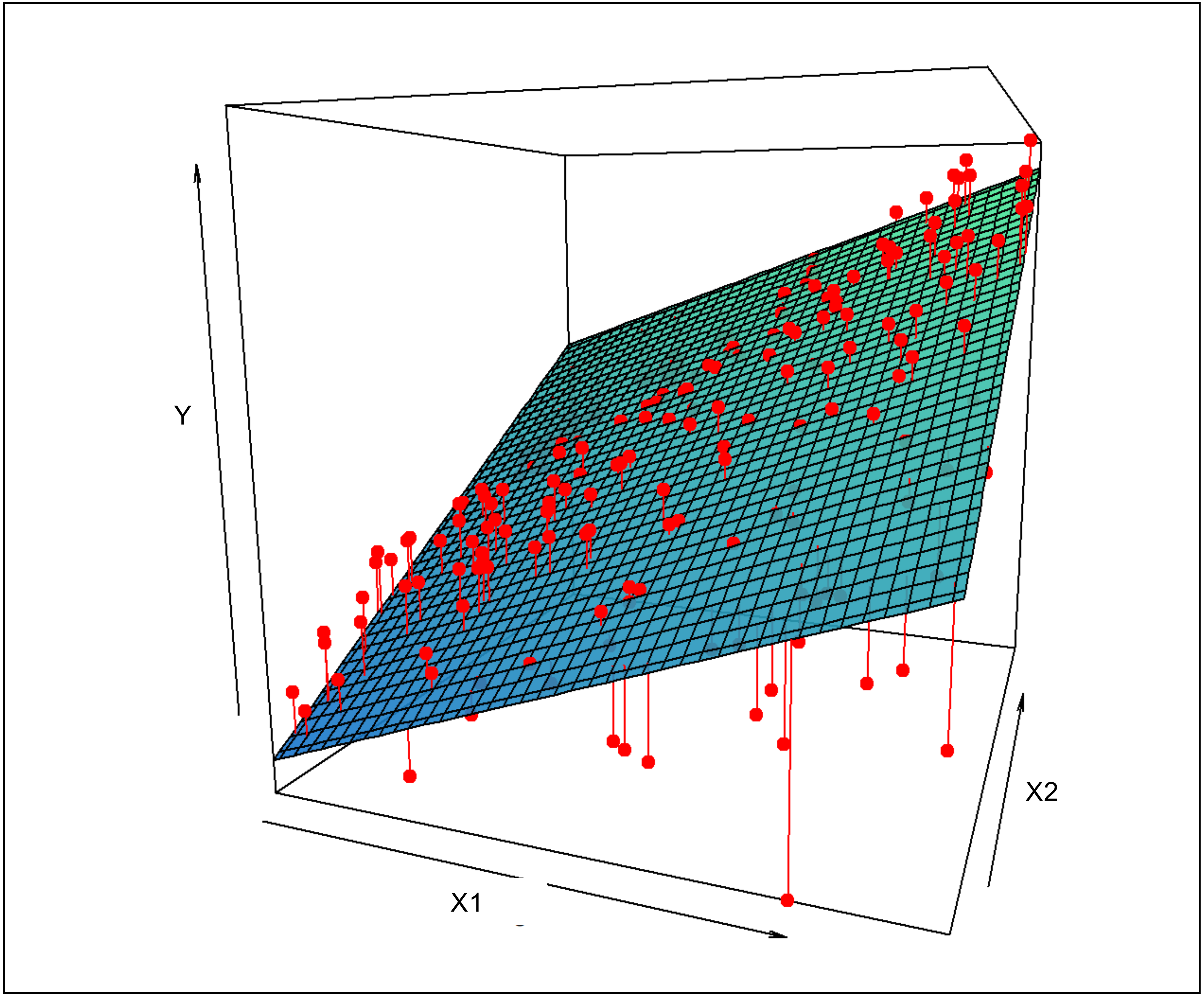
Multiple Linear Regression

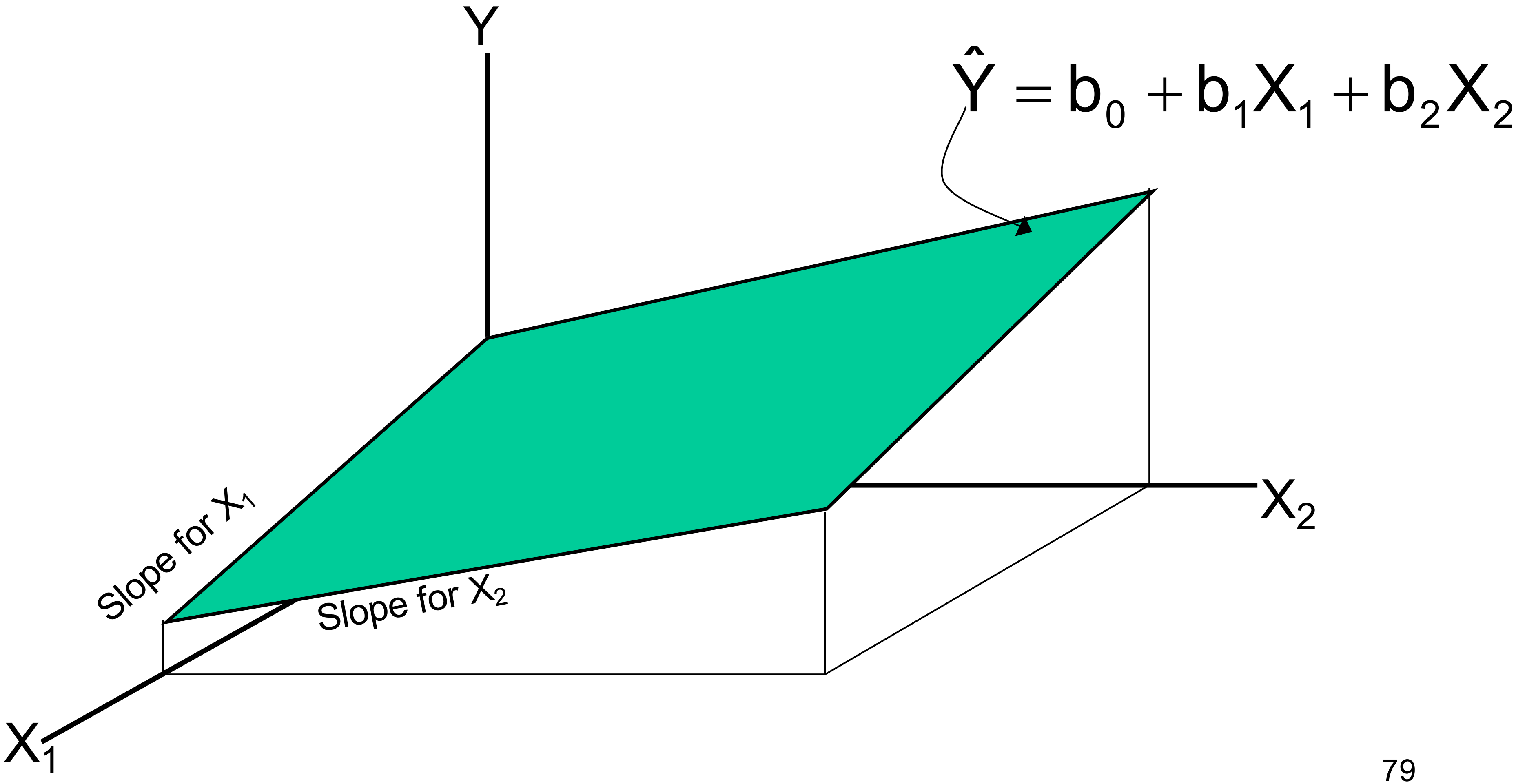
More than one predictor…

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

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.


Y
X1
X2


Y
$\hat{Y} = b_0 + b_1X_1 + b_2X_2$
Slope for X_1
Slope for X_2
X_2
X_1

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***)

$$\hat{Y} = -1.11 + 0.148X_1 + 0.793X_2$$

on average, consumption expend. increase, of **148** Euros each year for an increase of **1000 Euros of the income**, holding X_2 fixed

on average, consumption expend. increase of **793** Euros yearly for an additional component in the HH, holding X_1 fixed

Standardized coefficients

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

Which variable is contributing more to explain the food expenditures?

How to make a prediction:

Estimate 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 16.167 Euro

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

Advertising Dataset: sales (in thousands of units) for a product in function of advertising budget (in thousands of dollars) for TV, radio and newspaper

```
. reg sales TV

      Source |       SS           df       MS      Number of obs   =       200
-------------+----------------------------------   F(1, 198)       =    312.14
       Model |  3314.61817         1  3314.61817   Prob > F        =    0.0000
    Residual |  2102.53058       198  10.6188413   R-squared       =    0.6119
-------------+----------------------------------   Adj R-squared   =    0.6099
       Total |  5417.14875       199   27.221853   Root MSE        =    3.2587

------------------------------------------------------------------------------
       sales |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          TV |   .0475366   .0026906    17.67   0.000     .0422307    .0528426
       _cons |   7.032594   .4578429    15.36   0.000     6.129719    7.935468
------------------------------------------------------------------------------
```

Dataset from "An Introduction to Statistical Learning, with applications in R" (Springer, 2013)

```
. reg sales radio

      Source |       SS           df       MS      Number of obs   =       200
-------------+----------------------------------   F(1, 198)       =     98.42
       Model |   1798.6692         1   1798.6692   Prob > F        =    0.0000
    Residual |  3618.47955       198  18.2751492   R-squared       =    0.3320
-------------+----------------------------------   Adj R-squared   =    0.3287
       Total |  5417.14875       199   27.221853   Root MSE        =    4.2749

------------------------------------------------------------------------------
       sales |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       radio |   .2024958   .0204113     9.92   0.000     .1622443    .2427472
       _cons |   9.311638   .5629005    16.54   0.000     8.201588    10.42169
------------------------------------------------------------------------------
```

```
. reg sales newspaper

      Source |       SS           df       MS      Number of obs   =       200
-------------+----------------------------------   F(1, 198)       =     10.89
       Model |  282.344206         1  282.344206   Prob > F        =    0.0011
    Residual |  5134.80454       198  25.9333563   R-squared       =    0.0521
-------------+----------------------------------   Adj R-squared   =    0.0473
       Total |  5417.14875       199   27.221853   Root MSE        =    5.0925

------------------------------------------------------------------------------
       sales |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
   newspaper |   .0546931   .0165757     3.30   0.001     .0220055    .0873807
       _cons |   12.35141   .6214202    19.88   0.000     11.12596    13.57686
------------------------------------------------------------------------------
```

```
. reg sales TV radio newspaper

      Source |       SS           df       MS      Number of obs   =       200
-------------+----------------------------------   F(3, 196)       =    570.27
       Model |  4860.32349         3  1620.10783   Prob > F        =    0.0000
    Residual |  556.825263       196  2.84094522   R-squared       =    0.8972
-------------+----------------------------------   Adj R-squared   =    0.8956
       Total |  5417.14875       199   27.221853   Root MSE        =    1.6855

------------------------------------------------------------------------------
       sales |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          TV |   .0457646   .0013949    32.81   0.000     .0430137    .0485156
       radio |     .18853   .0086112    21.89   0.000     .1715474    .2055126
   newspaper |  -.0010375    .005871    -0.18   0.860     -.012616     .010541
       _cons |   2.938889   .3119082     9.42   0.000     2.323762    3.554016
------------------------------------------------------------------------------
```

F statistic:

H_0: $\beta_1=\beta_2=\ldots=\beta_p=0$

H_1: at least one β_j is non-zero

The simple and multiple regression coefficients can be quite different (see the case of the estimates for newspaper).
This difference stems from the fact that in the simple regression case, the slope term represents the average effect of a one unite increase in newspaper advertising, ignoring other predictors such as TV and radio.

In contrast, in the multiple regression setting, the coefficient for newspaper represents the average effect of increasing while holding TV and radio fixed.
We know that the correlation between radio and newspaper is 0.35. This reveals a tendency to spend more on newspaper advertising in markets where more is spent on radio advertising.
Now suppose that the multiple regression is correct and newspaper advertising has no direct impact on sales, but radio advertising does increase sales. Then in markets where we spend more on radio our sales will tend to be higher, and as our correlation matrix shows, we also tend to spend more on newspaper advertising in those same markets. Hence, in a simple linear regression which only examines sales versus newspaper, we will observe that higher values of newspaper tend to be associated with higher values of sales, even though newspaper advertising does not actually affect sales. So, newspaper sales are a surrogate for radio advertising; newspaper gets “credit” for the effect of radio on sales.

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
  - FEMALE: coded as 1 for all women, zero for men
  - MALE: coded as 1 for all men, zero for women

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 index measuring satisfaction with life

X_1=income, X_2=Female

$$Y = \beta_0 + \beta_1 INCOME + \beta_2 FEMALE + \varepsilon$$

We run the OLS and obtain the regression equation:

$$\hat{Y} = \widehat{\beta_0} + \widehat{\beta_1} INCOME + \widehat{\beta_2} FEMALE$$

- What if the case *for* a male?

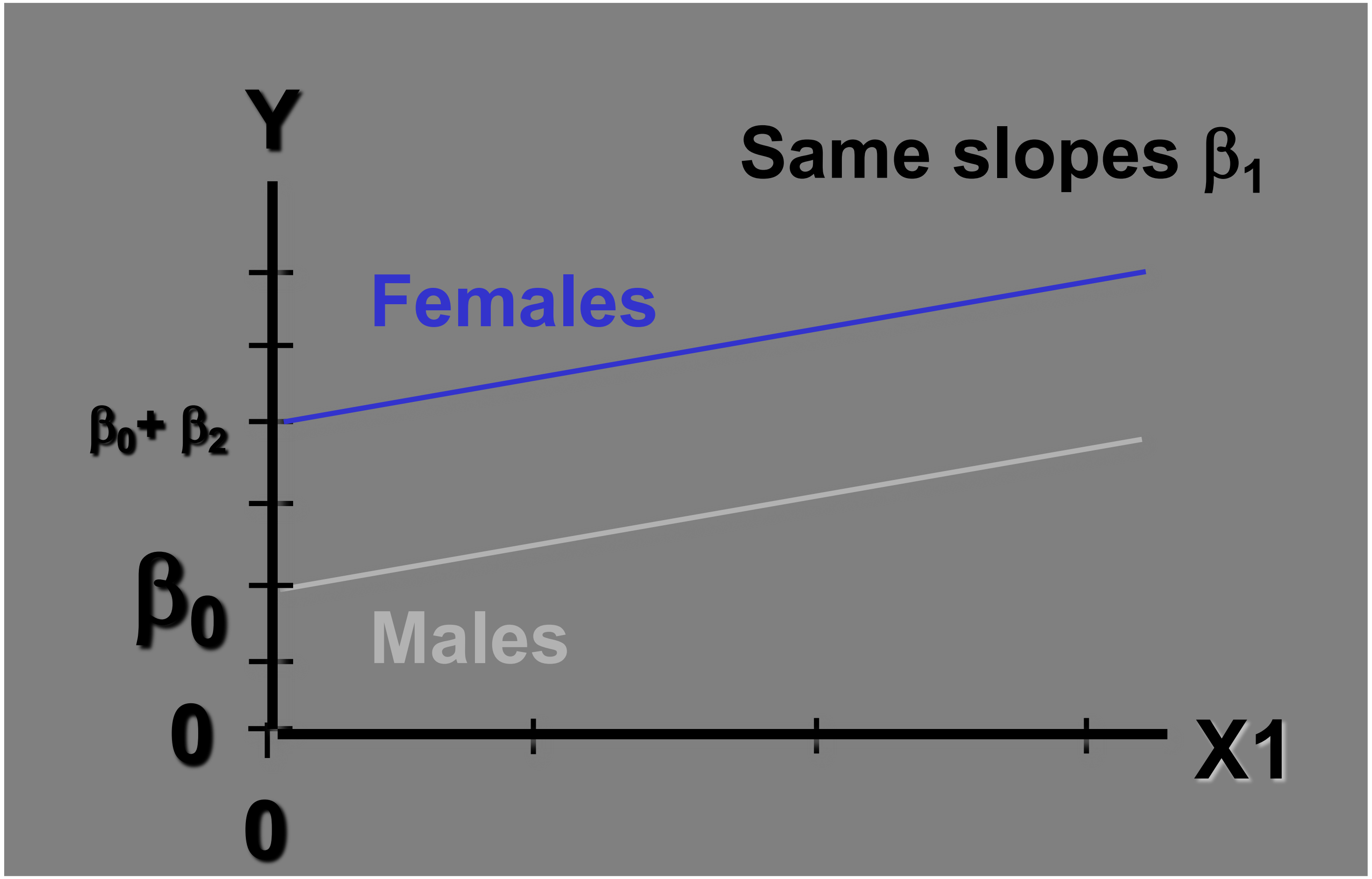
FEMALE is 0 in case of male, so males are modeled as:

$\widehat{\beta_0} + \widehat{\beta_1} INCOME.$

- What if the case *for* a female?
- DFEMALE=1 and so females are modeled using a different regression line:

$$\widehat{(\beta_0 + \widehat{\beta_2})} + \beta_1 INCOME$$

  - Thus, the coefficient of β_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 (β) indicates that women are consistently higher compared to men (on dep. var.)

- A negative coefficient indicated women are lower

- Example: If FEMALE coeff = 1.2:

  "Women are on average 1.2 points higher than men with respect to level of satisfaction".

- 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 FEMALE is included, the contrast is with males
  - If BLACK and OTHER 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?

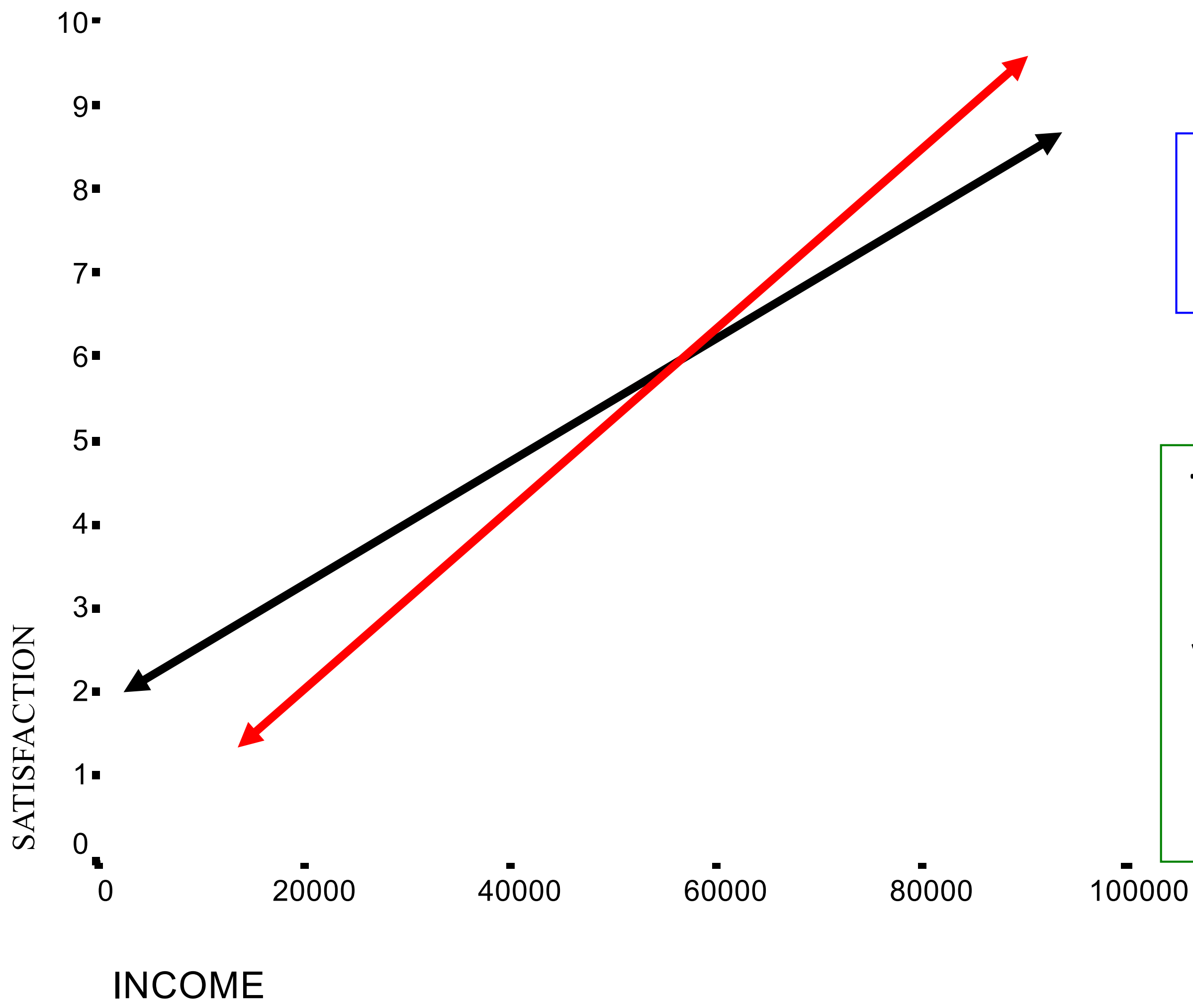
- Example: Income and Satisfaction with life – gender
  - Perhaps for men an extra euro increases their satisfaction a lot
  - Whereas for women each euro has a smaller effect on satisfaction (compared to men)

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

More in general, an interaction occurs when an independent variable has a different effect on the outcome depending on the values of another independent variable.

Interactions

here women, have a less steep income-satisfaction relationship compared to men



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: (FEMALE, INCOME) to create a new variable
    - – Called: FEMALE*INCOME
    - – Add that variable to the multiple regression model.

Interactions

Example, Y is satisfaction

$$Y = \beta_0 + \beta_1 INCOME + \beta_2 FEMALE + \beta_3 INC * FEM + \varepsilon$$

if the case *of* male:

FEMALE is 0, so $\widehat{\beta_3}$ (FEM*INC)=0 and males are modeled using the regression equation:

$\widehat{\beta_0} + \widehat{\beta_1} INC.$

$$Y = \beta_0 + \beta_1 INCOME + \beta_2 FEMALE + \beta_3 INC * FEM + \varepsilon$$

Females are then modeled using a different regression line:

$$(\widehat{\beta_0} + \widehat{\beta_2}) + (\widehat{\beta_1} + \widehat{\beta_3})INC$$

Now the regression lines have different intercepts, β0+β2 versus β0, as well as different slopes, β1+β3 versus β

Interactions

- Interpreting interaction terms:
- A positive b for FEMALE*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: FEMALE*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 “FEMALE * INCOME” should also include FEMALE 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:

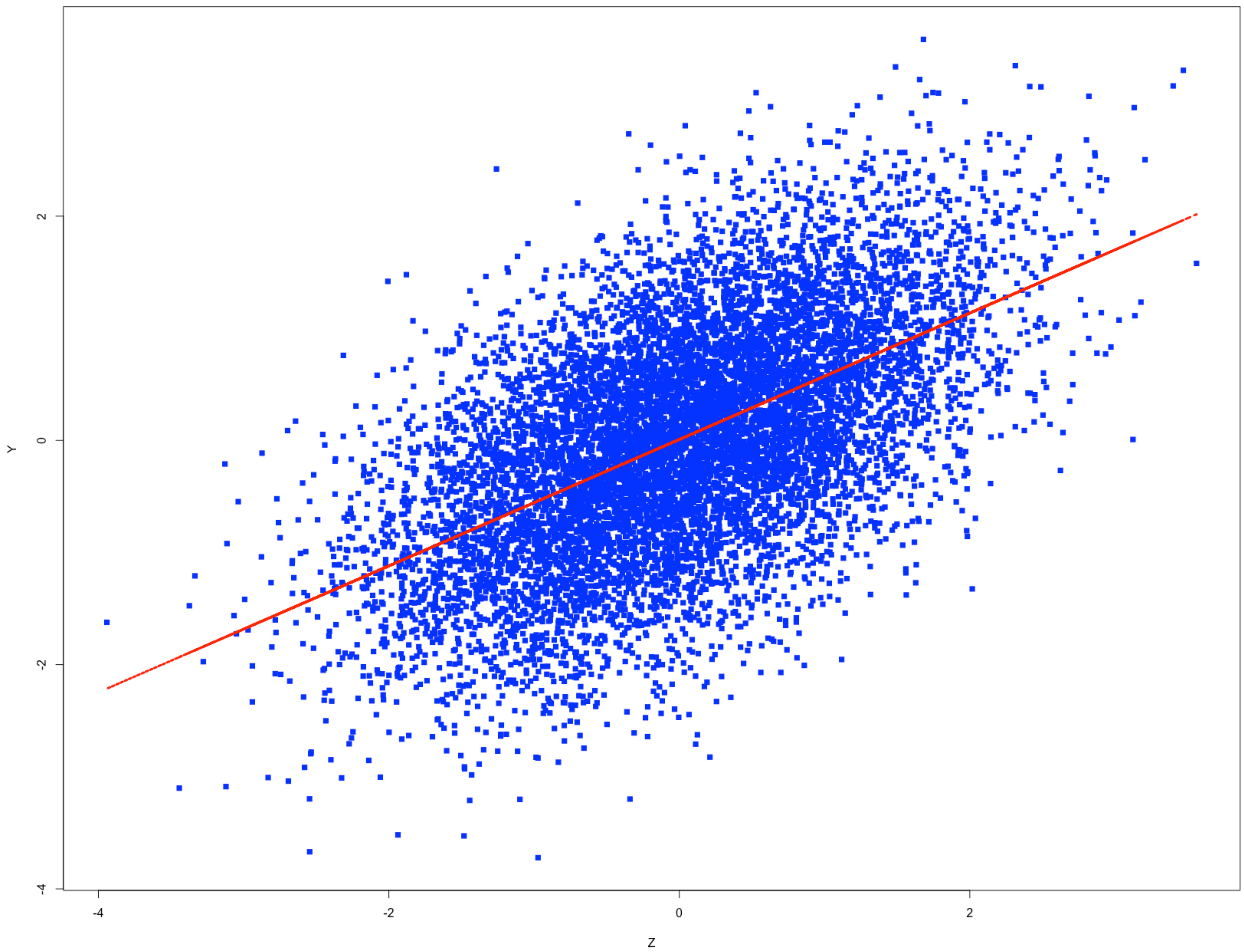
Predictor not significant and has the expected sign -> Keep it
Predictor not significant and does not have the expected sign -> Drop it
Predictor is significant and has the expected sign -> Keep it
Predictor is significant but does not have the expected sign -> Review, you may need more variables, it 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

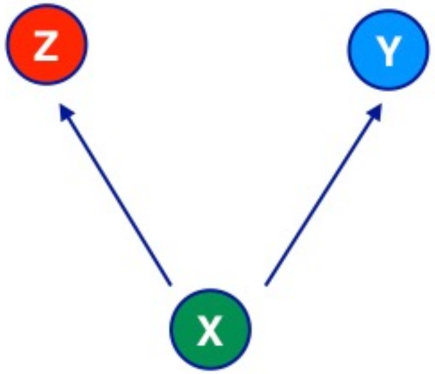

Y
Z
-4
-2
0
2

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!