Quantitative Methods in Social Research

Burak Sonmez

UCL

Feb 19, 2025 (UPF)

Day 2 – Outline

1. Understanding the Basics of Trends and Associations

2. Correlation

3. The Linear Regression Model

Day 2 – Outline

1. Understanding the Basics of Trends and Associations

2. Correlation

3. The Linear Regression Model

Association

- In everyday language, dependence, association and correlation are used interchangeably.

Association

- In everyday language, dependence, association and correlation are used interchangeably.
- However, association is synonymous with *dependence* and is different from correlation.

Association

- In everyday language, dependence, association and correlation are used interchangeably.
- However, association is synonymous with *dependence* and is different from correlation.
- Association is a very general relationship: one variable provides information about another.In other words, two variables are associated if there is a pattern in the scatter-plot, which is too strong to be likely to arise simply by chance.

Association

- In everyday language, dependence, association and correlation are used interchangeably.
- However, association is synonymous with *dependence* and is different from correlation.
- Association is a very general relationship: one variable provides information about another.In other words, two variables are associated if there is a pattern in the scatter-plot, which is too strong to be likely to arise simply by chance.
- Correlation measures a specific form of association: two variables are correlated when they display an increasing or decreasing trend. This is relevant to relationships with a linear trend.

Association

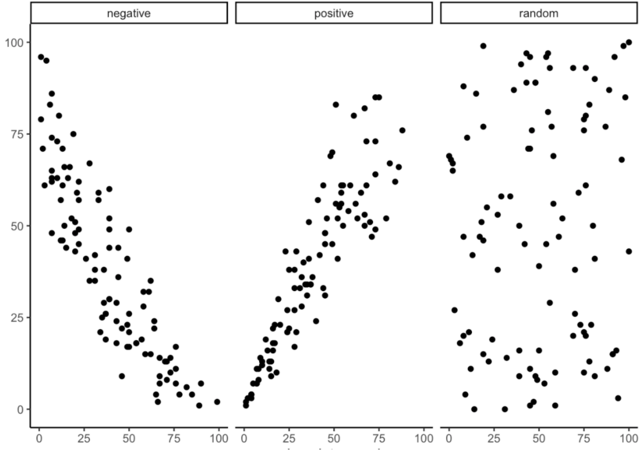
- In everyday language, dependence, association and correlation are used interchangeably.
- However, association is synonymous with *dependence* and is different from correlation.
- Association is a very general relationship: one variable provides information about another.In other words, two variables are associated if there is a pattern in the scatter-plot, which is too strong to be likely to arise simply by chance.
- Correlation measures a specific form of association: two variables are correlated when they display an increasing or decreasing trend. This is relevant to relationships with a linear trend.
- We can find patterns that look like one variable (X) is correlated with another (Y), even when that one variable does not **cause** the other variable.

Day 2 – Outline

1. Understanding the Basics of Trends and Associations
2. Correlation
3. The Linear Regression Model

Positive, Negative, and No-Correlation

Imagine the scatter plot would look like if the relationship between variables X and Y was negative and positive or if there was no relationship.



The Idea of Co-variance

- What co-variance refers to, the idea that the pattern of varying numbers in one measure is shared by the pattern of varying numbers in another measure.

The Idea of Co-variance

- What co-variance refers to, the idea that the pattern of varying numbers in one measure is shared by the pattern of varying numbers in another measure.
- Statistically speaking, co-variance is the multiplication of the deviations in X from the mean of X, and the deviation in Y from the mean of Y.

$$cov(X, Y) = \frac{\sum_i^n (x_i - \bar{X})(y_i - \bar{Y})}{N} \tag{1}$$

The Idea of Co-variance

- What co-variance refers to, the idea that the pattern of varying numbers in one measure is shared by the pattern of varying numbers in another measure.
- Statistically speaking, co-variance is the multiplication of the deviations in X from the mean of X, and the deviation in Y from the mean of Y.

$$cov(X,Y) = \frac{\sum_i^n (x_i - \bar{X})(y_i - \bar{Y})}{N} \tag{1}$$

- When we look at a co-variation statistic, we can see what direction it points, positive or negative. Yet, we don't know how big or small it is compared to the maximum or minimum possible value, so we don't know the relative size, which means we cannot say how strong the correlation is!

Pearson's Correlation r

- Therefore, for quantitative and ordinal data, we often use a certain measure of correlation, which is called Pearson's correlation (r). It measures linear trends!

$$r = \frac{cov(X,Y)}{\sigma_X \sigma_Y} = \frac{cov(X,Y)}{SD_X SD_Y} \tag{2}$$

Pearson's Correlation r

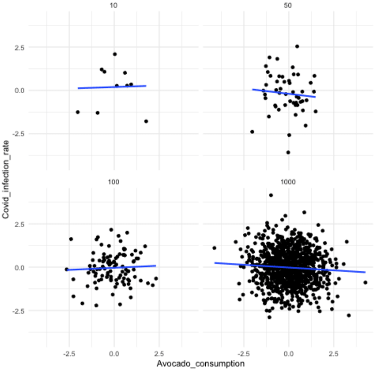
- Therefore, for quantitative and ordinal data, we often use a certain measure of correlation, which is called Pearson's correlation (r). It measures linear trends!

$$r = \frac{cov(X,Y)}{\sigma_X \sigma_Y} = \frac{cov(X,Y)}{SD_X SD_Y} \tag{2}$$

- Pearson's r is the co-variance of X and Y, divided by the product of the standard deviation of X and the standard deviation of Y. Why are we dividing the co-variance by the product of the standard deviations? Because this operation has the effect of normalising the co-variance into the range -1 to 1.

Spurious correlation

- This simulation illustrates some conundrums in interpreting correlations. They can be produced by random chance. This means that you can find a positive or negative correlation between two measures, even when they have absolutely nothing to do with one another.



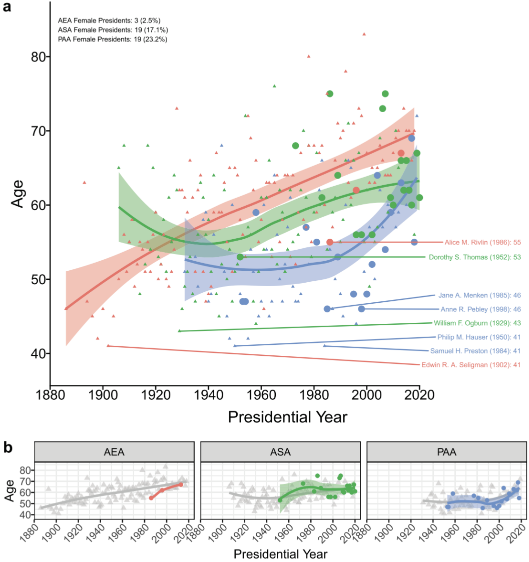
No correlation doesn’t mean no causality

- Just because there is no observable relationship does not mean there is no causal one!

No correlation doesn’t mean no causality.

CAUSAL INFERENCE: THE MIXTAPE, SCOTT CUNNINGHAM

Example: The Trends in Presidential Ages (Hauer, 2019)



What are the Trends?

- Since the 1940s, economic presidents are approximately five years older than sociology presidents while exhibiting a nearly linear increase in average age over time.

What are the Trends?

- Since the 1940s, economic presidents are approximately five years older than sociology presidents while exhibiting a nearly linear increase in average age over time.
- Sociology presidential ages demonstrate a nonlinear trend in the twentieth century, with presidents getting younger until the 1940s before aging at a similar pace as economists.

What are the Trends?

- Since the 1940s, economic presidents are approximately five years older than sociology presidents while exhibiting a nearly linear increase in average age over time.
- Sociology presidential ages demonstrate a nonlinear trend in the twentieth century, with presidents getting younger until the 1940s before aging at a similar pace as economists.
- Demography presidents have tended to be considerably younger than both sociology and economic presidents, though PAA presidents have aged considerably since the year 2000, approaching parity with the ASA only recently.

What are the Trends?

- Since the 1940s, economic presidents are approximately five years older than sociology presidents while exhibiting a nearly linear increase in average age over time.
- Sociology presidential ages demonstrate a nonlinear trend in the twentieth century, with presidents getting younger until the 1940s before aging at a similar pace as economists.
- Demography presidents have tended to be considerably younger than both sociology and economic presidents, though PAA presidents have aged considerably since the year 2000, approaching parity with the ASA only recently.
- The first female presidents for ASA and PAA appear almost simultaneously in the early 1950s, but female association presidents for ASA and PAA were still relatively rare until the mid-1980s.

Let's Discuss

- What type of data is used in this research?
- What is the analytical strategy in this research?
- What are the main methodological strengths and limitations of the study?

Let’s Discuss

- What type of data is used in this research?
- What is the analytical strategy in this research?
- What are the main methodological strengths and limitations of the study?
- Methodologically speaking, how would you approach these questions?
- Do young academics lack the credentials of previous generations; is the age composition of association membership also aging; does the increasing number of female presidents reflect the slow erosion of patriarchy in sociology and demography; why has this shift not occurred in economics?

Day 2 – Outline

1 Understanding the Basics of Trends and Associations

2 Correlation

3 The Linear Regression Model

What Do We Mean by Model?

- A model is a simplified abstraction of reality

What Do We Mean by Model?

- A model is a simplified abstraction of reality
- All models are wrong, but some are useful

What Do We Mean by Model?

- A model is a simplified abstraction of reality
- All models are wrong, but some are useful
- We will be using statistical models which will always be ‘wrong’, but some will be useful
- Simply put, a linear regression model is an approximation of the relationship between our independent variable X and our response variable Y

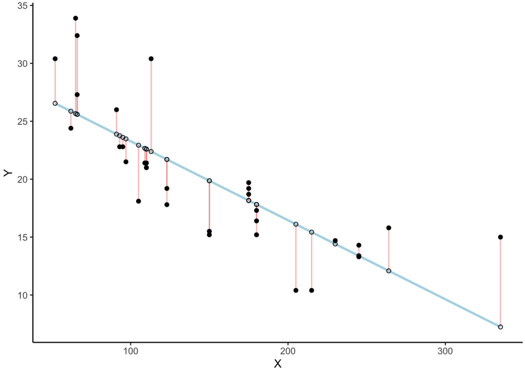
What is linear regression?

- Linear regression is a method that summarizes in a linear equation how the values of a numerical outcome variable vary over subpopulations defined by the values of predictor variables

What is linear regression?

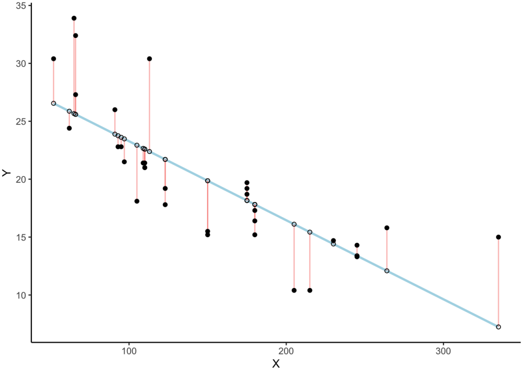
- Linear regression is a method that summarizes in a linear equation how the values of a numerical outcome variable vary over subpopulations defined by the values of predictor variables
- By observing a vast number of individuals with varied combinations of characteristics, we seek to isolate the impact of one characteristic one is interested in on the outcome by looking at the outcome of groups that are similar on all counts except for the characteristic one is interested in

Linear Relationships



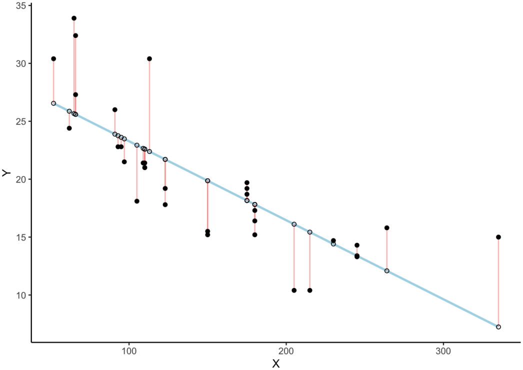
- The line can be represented by this function $Y = \alpha + \beta X$

Linear Relationships



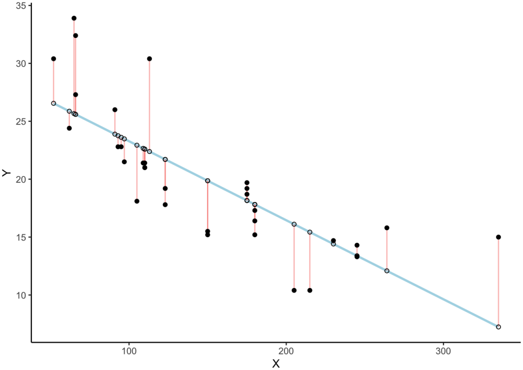
- The line can be represented by this function $Y = \alpha + \beta X$
- It could also be shown as $Y = \text{y-intercept} + \textit{slope} * X$

Linear Relationships



- The line can be represented by this function $Y = \alpha + \beta X$
- It could also be shown as $Y = \text{y-intercept} + slope * X$
- The slope (β) is the slant of the line, and the y-intercept is where the line crosses the y-axis when X is zero

Linear Relationships



- The line can be represented by this function $Y = \alpha + \beta X$
- It could also be shown as $Y = \text{y-intercept} + slope * X$
- The slope (β) is the slant of the line, and the y-intercept is where the line crosses the y-axis when X is zero
- β is the amount that Y increases when X increases by one unit

Linear Relationships

- When we have two variables, and plot them together, we now have a two-dimensional space. Hence, for two dimensions we could use a line, to summarize the central tendency of the relationship between the two variables.

Linear Relationships

- When we have two variables, and plot them together, we now have a two-dimensional space. Hence, for two dimensions we could use a line, to summarize the central tendency of the relationship between the two variables.
- Would we able to draw the best line to describe the general pattern of all the data points?

Linear Relationships

- When we have two variables, and plot them together, we now have a two-dimensional space. Hence, for two dimensions we could use a line, to summarize the central tendency of the relationship between the two variables.
- Would we able to draw the best line to describe the general pattern of all the data points?
- What's is even the best line? It is the one that has the least error.

Linear Relationships

- When we have two variables, and plot them together, we now have a two-dimensional space. Hence, for two dimensions we could use a line, to summarize the central tendency of the relationship between the two variables.
- Would we able to draw the best line to describe the general pattern of all the data points?
- What's is even the best line? It is the one that has the least error.
- Our goal is to estimate the line that 'best' fits our data

The Linear Regression Model

- We can express linearly related two variables with the bivariate linear regression model:

$$Y_i = \beta_0 + \beta_1 X_i + u_i \tag{3}$$

The Linear Regression Model

- We can express linearly related two variables with the bivariate linear regression model:

$$Y_i = \beta_0 + \beta_1 X_i + u_i \tag{3}$$

where:

- Observations i = 1, . . . , n
- Y is the dependent variable.
- X is the independent variable.
- β_0 is the intercept or constant.
- β_1 is the slope.
- u_i is the error term or residuals.
- β_0 and β_1 are known as the coefficients of the regression line

The Linear Regression Model

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

- β_0 gives the average value of Y when X is equal to 0

The Linear Regression Model

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

- β_0 gives the average value of Y when X is equal to 0
- β_0 gives the average change in Y that results from a one-unit change in X

The Linear Regression Model

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

- β_0 gives the average value of Y when X is equal to 0
- β_0 gives the average change in Y that results from a one-unit change in X
- These form the population regression line: the relationship that holds, on average, between X and Y in the population

The Linear Regression Model

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

- β_0 gives the average value of Y when X is equal to 0
- β_0 gives the average change in Y that results from a one-unit change in X
- These form the population regression line: the relationship that holds, on average, between X and Y in the population
- The final term is the random part of the equation u_i represents all the other factors aside from X that determine the value of Y

The Linear Regression Model

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

- β_0 gives the average value of Y when X is equal to 0
- β_0 gives the average change in Y that results from a one-unit change in X
- These form the population regression line: the relationship that holds, on average, between X and Y in the population
- The final term is the random part of the equation u_i represents all the other factors aside from X that determine the value of Y
- We want to know the population value of β_1 and β_0. We must estimate the values using a sample from the population

Ordinary Least Squares

- In order to estimate population value of β_1 and β_0, the most widely used approach is the ordinary least squares (OLS) method.

Ordinary Least Squares

- In order to estimate population value of β_1 and β_0, the most widely used approach is the ordinary least squares (OLS) method.
- The OLS estimator chooses the regression coefficients so that the estimated line is "as close as possible" to the data.

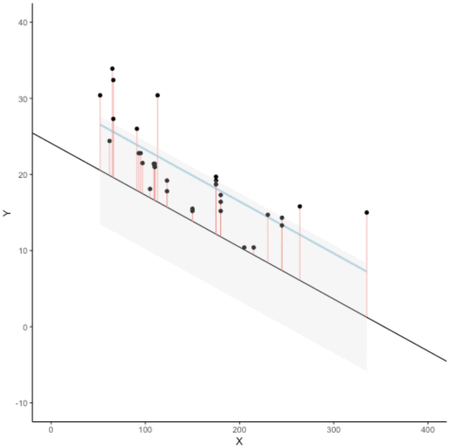
Ordinary Least Squares

- In order to estimate population value of β_1 and β_0, the most widely used approach is the ordinary least squares (OLS) method.
- The OLS estimator chooses the regression coefficients so that the estimated line is "as close as possible" to the data.
- It minimises the sum of the squared differences between the actual values of each observation (Y_i) and the predicted value of each value based on the estimated line ($\hat{Y}_i$).

The best fit line – regression

- For this line, the sum of the squared distances between Y_i and $\hat{Y_i}$:

$$\sum_{i=1}^{n}[Y_i - (\hat{\beta}_0 + \beta_1\hat{}X_i]^2 \tag{4}$$



Residuals

- These lines drop down from each dot, and they land on the line. Each of these little lines is called a residual. They show you how far off the line is for different dots. It's measure of error, it shows us just how wrong the line is. This means the line does not actually represent all of the dots. The best fit line is the least wrong of all the wrong lines.

Residuals

- These lines drop down from each dot, and they land on the line. Each of these little lines is called a residual. They show you how far off the line is for different dots. It's measure of error, it shows us just how wrong the line is. This means the line does not actually represent all of the dots. The best fit line is the least wrong of all the wrong lines.
- If we wanted to know how wrong this line was, we could simply gather up all the red lines, measure how long they are, and then add all the wrongness together. This would give us the total amount of wrongness.

Residuals

- These lines drop down from each dot, and they land on the line. Each of these little lines is called a residual. They show you how far off the line is for different dots. It's measure of error, it shows us just how wrong the line is. This means the line does not actually represent all of the dots. The best fit line is the least wrong of all the wrong lines.
- If we wanted to know how wrong this line was, we could simply gather up all the red lines, measure how long they are, and then add all the wrongness together. This would give us the total amount of wrongness.
- What we will actually be doing with the red lines, is computing the sum of the squared deviations from the line. That sum is the total amount of error. Now, this blue line here minimizes the sum of the squared deviations. Any other line would produce a larger total error.

Regression analysis

- Simple (bivariate) regression analysis involves one response variable (Y): continuous or discrete (linear or non-linear model, respectively) and one explanatory variable (X) (continuous and discrete).

Regression analysis

- Simple (bivariate) regression analysis involves one response variable (Y): continuous or discrete (linear or non-linear model, respectively) and one explanatory variable (X) (continuous and discrete).
- Using more than one explanatory variables (continuous and discrete) forms multiple regression model.

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 X_i ... + \beta_n X_i + u_i$$

OLS: Coefficient Interpretation

- What is the interpretation of $\hat{\beta}_1 = -0.45$?

OLS: Coefficient Interpretation

- What is the interpretation of $\hat{\beta}_1 = -0.45$?
- A one-unit increase in X is associated with a $\hat{\beta}_1$ change in Y, on average

OLS: Coefficient Interpretation

- What is the interpretation of $\hat{\beta}_1 = -0.45$?
- A one-unit increase in X is associated with a $\hat{\beta}_1$ change in Y, on average
- A one year increase in the age of job candidates is associated with a decrease of 0.45 in hiring favourability, on average, holding constant other variables constant

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 X_i ... + \beta_n X_i + u_i$$

Key assumptions in the simple linear model

- **Validity**: Most importantly, the data you are analyzing should map to the research question you are trying to answer (reflections on variable(s) and sample selection).

Key assumptions in the simple linear model

- **Validity**: Most importantly, the data you are analyzing should map to the research question you are trying to answer (reflections on variable(s) and sample selection).
- **Additivity and linearity**: The most important mathematical assumption of the regression model is that its deterministic component is a linear function of the separate predictors.

Key assumptions in the simple linear model

- **Validity**: Most importantly, the data you are analyzing should map to the research question you are trying to answer (reflections on variable(s) and sample selection).
- **Additivity and linearity**: The most important mathematical assumption of the regression model is that its deterministic component is a linear function of the separate predictors.
- **Independence of errors**: The simple regression model assumes that the errors from the prediction line are independent.

Key assumptions in the simple linear model

- **Validity**: Most importantly, the data you are analyzing should map to the research question you are trying to answer (reflections on variable(s) and sample selection).
- **Additivity and linearity**: The most important mathematical assumption of the regression model is that its deterministic component is a linear function of the separate predictors.
- **Independence of errors**: The simple regression model assumes that the errors from the prediction line are independent.
- **Equal variance of errors**: If the variance of the regression errors are unequal, estimation is more efficiently performed using weighted least squares, where each point is weighted inversely proportional to its variance. However, this issue is ”minor”. Unequal variance does not affect the most important aspect of a regression model, which is the form of the predictor $X\beta$

Key assumptions in the simple linear model

- **Validity**: Most importantly, the data you are analyzing should map to the research question you are trying to answer (reflections on variable(s) and sample selection).
- **Additivity and linearity**: The most important mathematical assumption of the regression model is that its deterministic component is a linear function of the separate predictors.
- **Independence of errors**: The simple regression model assumes that the errors from the prediction line are independent.
- **Equal variance of errors**: If the variance of the regression errors are unequal, estimation is more efficiently performed using weighted least squares, where each point is weighted inversely proportional to its variance. However, this issue is ”minor”. Unequal variance does not affect the most important aspect of a regression model, which is the form of the predictor $X\beta$

Key assumptions in the simple linear model

- **Normality of errors**: The regression assumption that is generally least important is that the errors are normally distributed. In fact, for the purpose of estimating the regression line (as compared to predicting individual data points), the assumption of normality is barely important at all. Even if there's no clear-cut consensus, we do not recommend diagnostics of the normality of regression residuals.

Model Fit

- How does our model perform?
- How tightly are the observations clustered around the line?
- What proportion of the variation in the dependent variable can be explained by the explanatory variable?

R^2 – coefficient of determination

- The coefficient of determination R^2 shows how much of the variation of the dependent variable (y) can be explained by our model.

R^2 – coefficient of determination

- The coefficient of determination R^2 shows how much of the variation of the dependent variable (y) can be explained by our model.
- The coefficient of determination R^2 can be derived from a simple variance decomposition. $R^2 = \frac{SSE}{SST}$

R^2 – coefficient of determination

- The coefficient of determination R^2 shows how much of the variation of the dependent variable (y) can be explained by our model.
- The coefficient of determination R^2 can be derived from a simple variance decomposition. $R^2 = \frac{SSE}{SST}$
- The Sum of Squares Explained (SSE) is a measure of the variability in the outcome variable that is explained by the explanatory variables $SSE = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2$

R^2 – coefficient of determination

- The coefficient of determination R^2 shows how much of the variation of the dependent variable (y) can be explained by our model.
- The coefficient of determination R^2 can be derived from a simple variance decomposition. $R^2 = \frac{SSE}{SST}$
- The Sum of Squares Explained (SSE) is a measure of the variability in the outcome variable that is explained by the explanatory variables $SSE = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2$
- The Sum of Squares Residual (SSR) is a measure of the variability in the outcome variable that is not explained by your regression $SSR = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} e_i^2$

R^2 – coefficient of determination

- The coefficient of determination R^2 shows how much of the variation of the dependent variable (y) can be explained by our model.
- The coefficient of determination R^2 can be derived from a simple variance decomposition. $R^2 = \frac{SSE}{SST}$
- The Sum of Squares Explained (SSE) is a measure of the variability in the outcome variable that is explained by the explanatory variables $SSE = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2$
- The Sum of Squares Residual (SSR) is a measure of the variability in the outcome variable that is not explained by your regression $SSR = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} e_i^2$
- You can show mathematically that SSE + SSR is equal to the following expression, which is referred to as the total sum of squares (TSS) $TSS = \sum_{i=1}^{n}(\bar{y}_i - y_i)^2$

Hypothesis Test: Multiple Linear Regression

- More than one explanatory variables, continuous or discrete. Coefficients describe partial associations of explanatory variable with response variable, controlling for the other explanatory variable(s) in the model.
- H_0: in the population, there is no partial association between X_j and Y, controlling for the other explanatory variable(s) in the model.
- For a continuous variable, null hypothesis is that the slope or beta is 0.
- For a dummy variable, null hypothesis is no difference between the two categories.