Lecture 8: Regression

Naijia Liu

Feb. 15, 2024

Notations

- Y_i: Outcome Variable / Dependent Variable
- X_i: Independent Variables
- β: Coefficient for IVs
- β_0: Coefficient for Intercept
- ϵ_i: Error term

Linear Regression: A Model for the Mean

Assume a model for an observed simple random sample Y_i:

$$\underbrace{Y_i}_{\text{Outcome}} = \underbrace{\beta_0}_{\text{Mean or intercept}} + \underbrace{\epsilon_i}_{\text{Residual, error term}}$$

Linear Regression: A Model for the Mean

Assume a model for an observed simple random sample Y_i:

$$\underbrace{Y_i}_{\text{Outcome}} = \underbrace{\beta_0}_{\text{Mean or intercept}} + \underbrace{\epsilon_i}_{\text{Residual, error term}}$$

How to choose $\widehat{\beta}_0$?

$$\widehat{\beta}_0 = \underset{\widetilde{\beta}_0}{\operatorname{argmin}} \mathcal{L}\left(\widetilde{\beta}_0\right)$$

We want $\widehat{\beta}_0$ to be the one to minimize some type of error, in predicting Y_i.

And $\mathcal{L}\left(\widetilde{\beta}_0\right)$ is a **loss** function.

Commonly Encountered Loss Functions

- Criterion of Least Squares (OLS): **We want to minimize the sum of squared error between true data and our predictions.**

$$\widehat{\beta}_0 = \underset{\widetilde{\beta}_0}{\operatorname{argmin}} \sum_{i=1}^{N} \left(Y_i - \widetilde{\beta}_0 \right)^2$$

Commonly Encountered Loss Functions

- Criterion of Least Squares (OLS): **We want to minimize the sum of squared error between true data and our predictions.**

$$\widehat{\beta}_0 = \underset{\widetilde{\beta}_0}{\text{argmin}} \sum_{i=1}^{N} \left(Y_i - \widetilde{\beta}_0 \right)^2$$

- Criteron of Least Absolute Deviation: **We want to minimize the sum of absolute error between true data and our predictions.**

$$\widehat{\beta}_0 = \underset{\widetilde{\beta}_0}{\text{argmin}} \sum_{i=1}^{N} \left| Y_i - \widetilde{\beta}_0 \right|$$

Commonly Encountered Loss Functions

- Criterion of Least Squares (OLS): **We want to minimize the sum of squared error between true data and our predictions.**

$$\widehat{\beta}_0 = \underset{\widetilde{\beta}_0}{\operatorname{argmin}} \sum_{i=1}^{N} \left(Y_i - \widetilde{\beta}_0 \right)^2$$

- Criteron of Least Absolute Deviation: **We want to minimize the sum of absolute error between true data and our predictions.**

$$\widehat{\beta}_0 = \underset{\widetilde{\beta}_0}{\operatorname{argmin}} \sum_{i=1}^{N} \left| Y_i - \widetilde{\beta}_0 \right|$$

- Penalized Least Squares: **We want to minimize the sum of squared error between true data and our predictions, plus something else (later).**

$$\widehat{\beta}_0 = \underset{\widetilde{\beta}_0}{\operatorname{argmin}} \sum_{i=1}^{N} \left(Y_i - \widetilde{\beta}_0 \right)^2 + \lambda \widetilde{\beta}_0^2$$

Review on Taking Derivatives

For a function:

$$f(x) = (x + a)^2$$

The first derivative of it is:

$$f'(x) = \frac{d}{dx} f(x) = 2(x + a)$$

Review on Taking Derivatives

For a function:

$$f(x) = (x + a)^2$$

The first derivative of it is:

$$f'(x) = \frac{d}{dx} f(x) = 2(x + a)$$

The second derivative of it is:

$$f''(x) = 2$$

Review on Taking Derivatives

For a function:

$$f(x) = (x + a)^2$$

The first derivative of it is:

$$f'(x) = \frac{d}{dx} f(x) = 2(x + a)$$

The second derivative of it is:

$$f''(x) = 2$$

If we want to find the point that minimizes the function, we want to set first derivative to 0.
In this case we have:

$$x = -a$$

OLS Solution

$$\widehat{\beta}_0 = \underset{\widetilde{\beta}_0}{\operatorname{argmin}} \sum_{i=1}^{N} \left(Y_i - \widetilde{\beta}_0 \right)^2$$

OLS Solution

$$\widehat{\beta}_0 = \underset{\widetilde{\beta}_0}{\operatorname{argmin}} \sum_{i=1}^{N} \left(Y_i - \widetilde{\beta}_0 \right)^2$$

$$\Rightarrow \left. \frac{\partial \mathcal{L}(\widetilde{\beta}_0)}{\partial \widetilde{\beta}_0} \right|_{\widetilde{\beta}_0 = \widehat{\beta}_0} = 0$$

$$\frac{\partial}{\partial \widetilde{\beta}_0} \sum_{i=1}^{N} \left(Y_i - \widetilde{\beta}_0 \right)^2 = 0$$

$$\sum_{i=1}^{N} \frac{\partial}{\partial \widetilde{\beta}_0} \left(Y_i - \widetilde{\beta}_0 \right)^2 = 0$$

$$\sum_{i=1}^{N} -2 \cdot \left(Y_i - \widehat{\beta}_0 \right) = 0$$

OLS Solution

$$\widehat{\beta}_0 = \underset{\widetilde{\beta}_0}{\operatorname{argmin}} \sum_{i=1}^{N} \left(Y_i - \widetilde{\beta}_0 \right)^2$$

$$\Rightarrow \left. \frac{\partial \mathcal{L}(\widetilde{\beta}_0)}{\partial \widetilde{\beta}_0} \right|_{\widetilde{\beta}_0 = \widehat{\beta}_0} = 0$$

$$\frac{\partial}{\partial \widetilde{\beta}_0} \sum_{i=1}^{N} \left(Y_i - \widetilde{\beta}_0 \right)^2 = 0$$

$$\sum_{i=1}^{N} \frac{\partial}{\partial \widetilde{\beta}_0} \left(Y_i - \widetilde{\beta}_0 \right)^2 = 0$$

$$\sum_{i=1}^{N} -2 \cdot \left(Y_i - \widehat{\beta}_0 \right) = 0$$

$$\sum_{i=1}^{N} \left(Y_i - \widehat{\beta}_0 \right) = 0$$

$$\sum_{i=1}^{N} Y_i = \sum_{i=1}^{N} \widehat{\beta}_0$$

$$\sum_{i=1}^{N} Y_i = N\widehat{\beta}_0$$

$$\frac{1}{N} \sum_{i=1}^{N} Y_i = \overline{Y}_i = \widehat{\beta}_0$$

Least Absolute Deviation Solution (optional)

$$\widehat{\beta}_0 = \underset{\widetilde{\beta}_0}{\text{argmin}} \sum_{i=1}^{N} |Y_i - \widetilde{\beta}_0|$$

$$\Rightarrow \frac{\partial \mathcal{L}(\widetilde{\beta}_0)}{\partial \widetilde{\beta}_0}\bigg|_{\widetilde{\beta}_0 = \widehat{\beta}_0} = 0$$

$$\frac{\partial}{\partial \widetilde{\beta}_0} \left\{ \sum_{i=1}^{N} |Y_i - \widetilde{\beta}_0| \right\} = 0$$

$$\sum_{i=1}^{N} \frac{\partial}{\partial \widetilde{\beta}_0} |Y_i - \widetilde{\beta}_0| = 0$$

$$\sum_{i=1}^{N} sgn(Y_i - \widehat{\beta}_0) = 0$$

$$\Rightarrow \widetilde{Y} = \widehat{\beta}_0$$

- If we rank all observations from small to large:

  $Y_{(1)} = \min(Y_i);\ Y_{(N)} = \max(Y_i)$

Least Absolute Deviation Solution (optional)

$$\widehat{\beta}_0 = \underset{\widetilde{\beta}_0}{\text{argmin}} \sum_{i=1}^{N} | Y_i - \widetilde{\beta}_0|$$

$$\Rightarrow \frac{\partial \mathcal{L}(\widetilde{\beta}_0)}{\partial \widetilde{\beta}_0}\bigg|_{\widetilde{\beta}_0 = \widehat{\beta}_0} = 0$$

$$\frac{\partial}{\partial \widetilde{\beta}_0} \left\{ \sum_{i=1}^{N} | Y_i - \widetilde{\beta}_0| \right\} = 0$$

$$\sum_{i=1}^{N} \frac{\partial}{\partial \widetilde{\beta}_0} | Y_i - \widetilde{\beta}_0| = 0$$

$$\sum_{i=1}^{N} sgn(Y_i - \widehat{\beta}_0) = 0$$

$$\Rightarrow \widetilde{Y} = \widehat{\beta}_0$$

- If we rank all observations from small to large:

  $Y_{(1)} = \min(Y_i);\ Y_{(N)} = \max(Y_i)$

- Median value of Y_i: $Y_{(N+1)/2}$.

Least Absolute Deviation Solution (optional)

$$\widehat{\beta}_0 = \underset{\widetilde{\beta}_0}{\text{argmin}} \sum_{i=1}^{N} |Y_i - \widetilde{\beta}_0|$$

$$\Rightarrow \frac{\partial \mathcal{L}(\widetilde{\beta}_0)}{\partial \widetilde{\beta}_0}\bigg|_{\widetilde{\beta}_0 = \widehat{\beta}_0} = 0$$

$$\frac{\partial}{\partial \widetilde{\beta}_0} \left\{ \sum_{i=1}^{N} |Y_i - \widetilde{\beta}_0| \right\} = 0$$

$$\sum_{i=1}^{N} \frac{\partial}{\partial \widetilde{\beta}_0} |Y_i - \widetilde{\beta}_0| = 0$$

$$\sum_{i=1}^{N} sgn(Y_i - \widehat{\beta}_0) = 0$$

$$\Rightarrow \widetilde{Y} = \widehat{\beta}_0$$

- If we rank all observations from small to large:

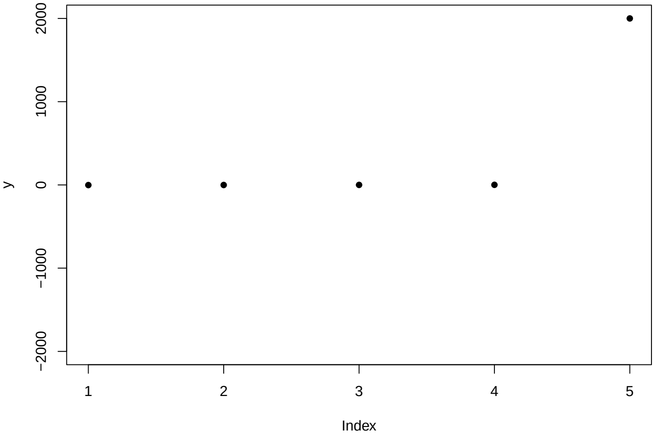
  $Y_{(1)} = \min(Y_i);\ Y_{(N)} = \max(Y_i)$
- Median value of Y_i: $Y_{(N+1)/2}$.
- Median more robust to extreme values than mean.

A toy example

- Five observations:

$$Y_1 = -2,\ Y_2 = -1,\ Y_3 = 0,\ Y_4 = 1,\ Y_5 = 2000$$

- Mean as 399.6, median as 0

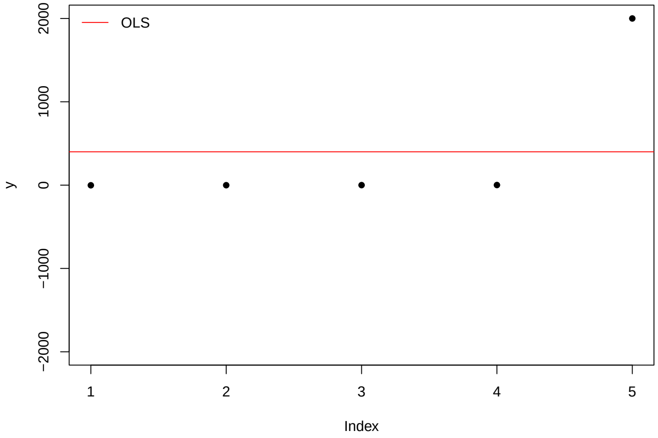


A toy example

- Five observations:

$$Y_1 = -2,\ Y_2 = -1,\ Y_3 = 0,\ Y_4 = 1,\ Y_5 = 2000$$

- Mean as 399.6, median as 0

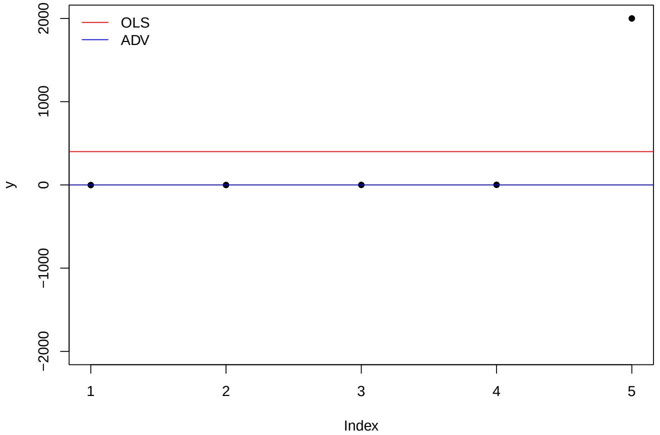


A toy example

- Five observations:

$$Y_1 = -2,\ Y_2 = -1,\ Y_3 = 0,\ Y_4 = 1,\ Y_5 = 2000$$

- Mean as 399.6, median as 0



Linear Regression Model

- A model for a **linear** relationship between two variables

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

- X: Independent (explanatory) variable
- Y: Dependent (outcome, response) variable
- ϵ: error (disturbance) term

Linear Regression Model

- A model for a **linear** relationship between two variables

$$Y_i \;=\; \beta_0 + \beta_1 X_i + \epsilon_i$$

- X: Independent (explanatory) variable
- Y: Dependent (outcome, response) variable
- ϵ: error (disturbance) term

- Given a value of X, the model predicts the average of Y
- Abuse of regression: extrapolation, causal misinterpretation

  Correlation is not causation!

Regression

Regression analysis answers:

1. What is the **best** line that describes an outcome variable (aka dependent variable) in terms of an independent variable?

Regression

Regression analysis answers:

1. What is the **best** line that describes an outcome variable (aka dependent variable) in terms of an independent variable?
2. Given a value of the independent variable, what is my best guess for the dependent variable?

Regression

Regression analysis answers:

1. What is the **best** line that describes an outcome variable (aka dependent variable) in terms of an independent variable?
2. Given a value of the independent variable, what is my best guess for the dependent variable?
3. How close is the line to the data?

   Loss function of choice

Estimating GDP

Given GDP growth rate in 2007, how can we estimate GDP growth rate in 2008?

- Assume: GDP growth in 2008 is GDP growth in 2007 times a constant plus an intercept

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

Estimating GDP

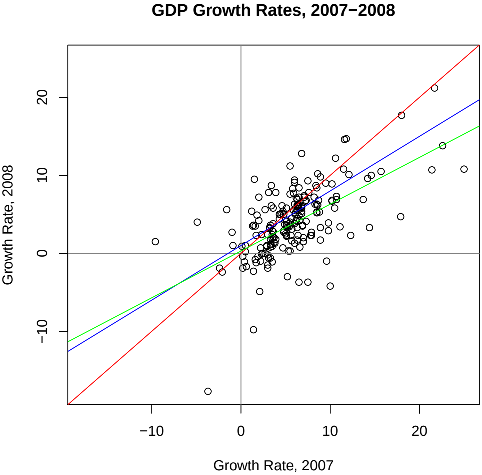
Given GDP growth rate in 2007, how can we estimate GDP growth rate in 2008?

- Assume: GDP growth in 2008 is GDP growth in 2007 times a constant plus an intercept

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

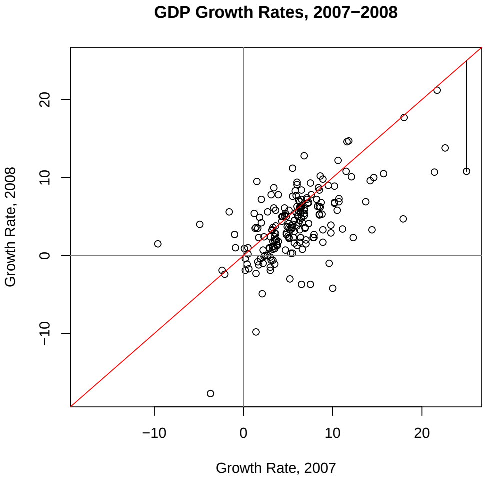
- Do we expect the coefficient estimate of GDP 2007 on GDP 2008 to be positive or negative?

Which Line to Choose



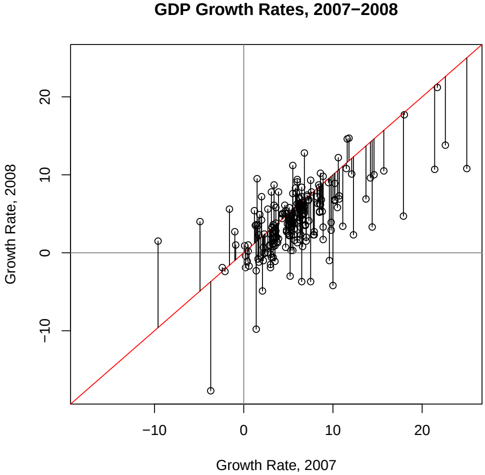
How far is a point from the line?

The distance from one point to the line, called the residual



How far are all of the points from the line?

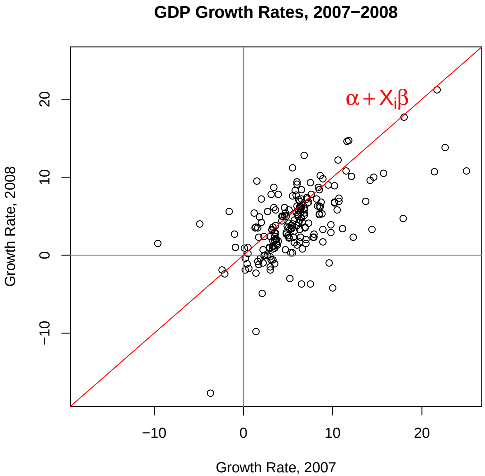
The total distances from the data to the line (residuals)



Determining the line of best fit

Determining the line of best fit (aka the line of least squares)

- Y_i: 2008 GDP growth rate for country i
- X_i: 2007 GDP growth rate for country i



How far are all of the points from the line?

To allow for some difference between Y_i and $\beta_0 + X_i\beta_1$, we say

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

This is our **assumed model**

How far are all of the points from the line?

To allow for some difference between Y_i and $\beta_0 + X_i\beta_1$, we say

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

This is our **assumed model**

After we see our data, we are going to estimate a model,

$$\widehat{Y}_i = \widehat{\beta}_0 + X_i\widehat{\beta}_1$$

This is our **fitted model** or <u>estimated model</u>

How far are all of the points from the line?

To allow for some difference between Y_i and $\beta_0 + X_i\beta_1$, we say

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

This is our **assumed model**

After we see our data, we are going to estimate a model,

$$\widehat{Y}_i = \widehat{\beta}_0 + X_i\widehat{\beta}_1$$

This is our **fitted model** or estimated model

The fitted model varies from sample to sample (like in a survey). The assumed model does not necessarily.

Assumptions

- Linearity among variables and error terms.

  age and age square, income

Assumptions

- Linearity among variables and error terms.

  age and age square, income

- Error terms have a mean of zero

Assumptions

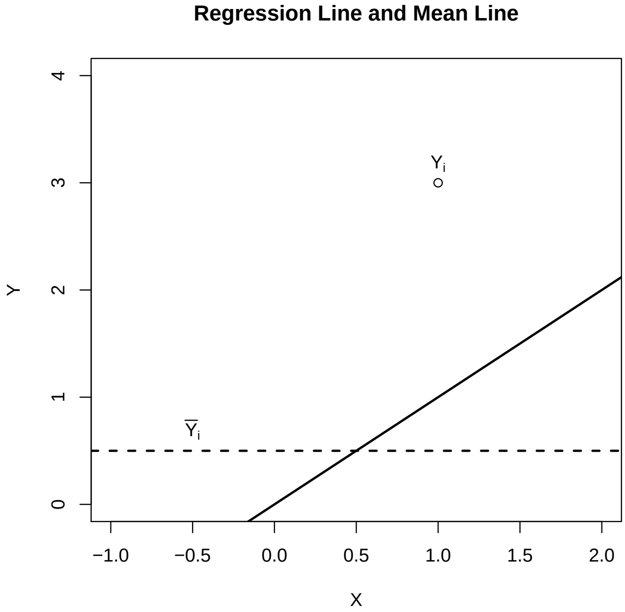
- Linearity among variables and error terms.

  age and age square, income

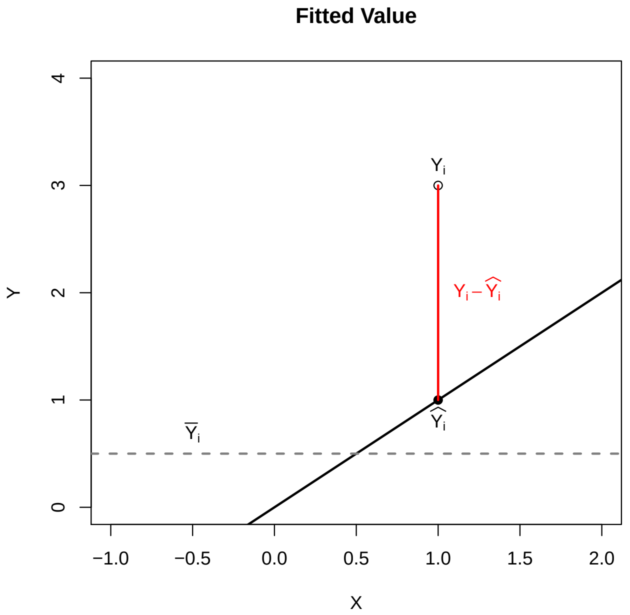
- Error terms have a mean of zero
- Error terms are uncorrelated with each other.

  One observation of the error term should not predict the next observation

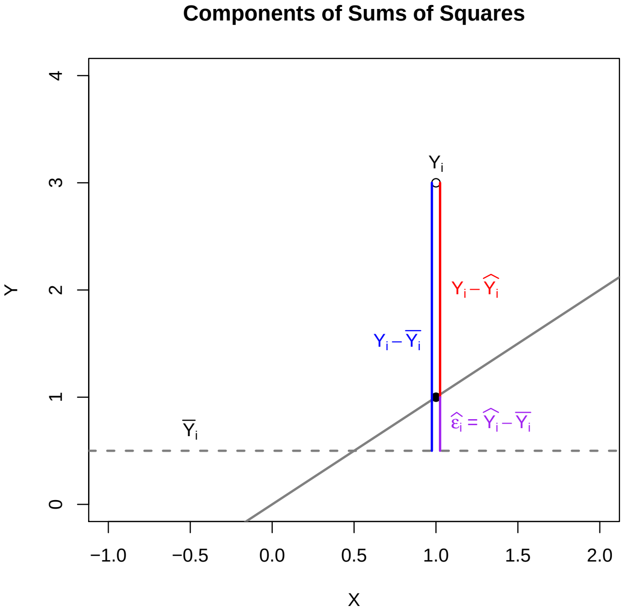
Components of Sums of Squares



Components of Sums of Squares



Components of Sums of Squares



Sums of Squares

- Total Sum of Squares–how much variance in Y_i is there to explain?

$$TSS: \sum_{i=1}^{N} (Y_i - \overline{Y}_i)^2$$

- Estimated Sum of Squares–how much of this variance do we explain?

$$ESS: \sum_{i=1}^{N} (\widehat{Y}_i - \overline{Y}_i)^2$$

- Residual Sum of Squares–how much variance is unexplained?

$$RSS: \sum_{i=1}^{N} (Y_i - \widehat{Y}_i)^2$$

Geometric Projection

When $\widehat{Y}_i$ are the fitted value from a linear regression, the total variance to explain equals the explained variance plus the unexplained variance.

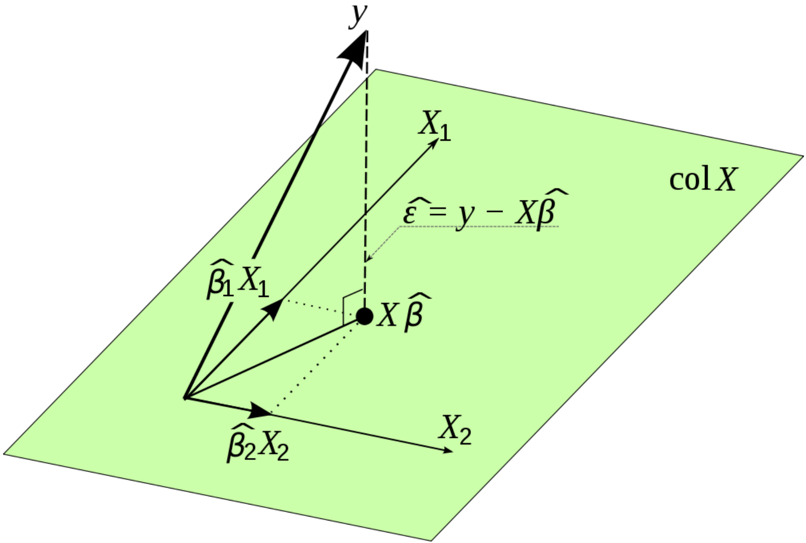
$$\underbrace{\sum_{i=1}^{N}(Y_i - \overline{Y}_i)^2}_{TSS} = \underbrace{\sum_{i=1}^{N}(\widehat{Y}_i - \overline{Y}_i)^2}_{ESS} + \underbrace{\sum_{i=1}^{N}(Y_i - \widehat{Y}_i)^2}_{RSS}$$

- This is a simple ANOVA (ANalysis Of VAriance) decomposition

Derivation: Use put-and-take

$$
\begin{aligned}
\sum_{i=1}^{N}(Y_i - \overline{Y}_i)^2 &= \sum_{i=1}^{N}(Y_i - \widehat{Y}_i + \widehat{Y}_i - \overline{Y}_i)^2 \\
&= \sum_{i=1}^{N}(Y_i - \widehat{Y}_i)^2 + \sum_{i=1}^{N}(\widehat{Y}_i - \overline{Y}_i)^2 + 2\sum_{i=1}^{N}(Y_i - \widehat{Y}_i)(\widehat{Y}_i - \overline{Y}_i) \\
&= \sum_{i=1}^{N}(Y_i - \widehat{Y}_i)^2 + \sum_{i=1}^{N}(\widehat{Y}_i - \overline{Y}_i)^2 + 2\sum_{i=1}^{N}\widehat{\epsilon}_i(\widehat{Y}_i - \overline{Y}_i) \\
&= \sum_{i=1}^{N}(Y_i - \widehat{Y}_i)^2 + \sum_{i=1}^{N}(\widehat{Y}_i - \overline{Y}_i)^2 + 2\underbrace{\sum_{i=1}^{N}\widehat{\epsilon}_i\widehat{Y}_i}_{=0} - 2\sum_{i=1}^{N}\widehat{\epsilon}_i\overline{Y}_i \\
&= \sum_{i=1}^{N}(Y_i - \widehat{Y}_i)^2 + \sum_{i=1}^{N}(\widehat{Y}_i - \overline{Y}_i)^2 - 2\overline{Y}_i\underbrace{\sum_{i=1}^{N}\widehat{\epsilon}_i}_{=0}
\end{aligned}
$$

Sums of Squares Identity



R^2: The Coefficient of Determination

Motivation:

$$TSS = ESS + RSS \Rightarrow \quad 1 = \frac{ESS}{TSS} + \frac{RSS}{TSS}$$

R^2: What proportion of the total variation in Y_i are we explaining with $\widehat{Y}_i$?

$$R^2 = \frac{\sum_{i=1}^{N}(\widehat{Y}_i - \overline{Y}_i)^2}{\sum_{i=1}^{N}(Y_i - \overline{Y}_i)^2} = \frac{ESS}{TSS} = 1 - \frac{RSS}{TSS}$$

Variance of Simple Least Squares Coefficient

What we will need:

1. $\mathrm{Var}(\beta_0) = \mathrm{Var}(X_i\beta_1) = 0$. Why?
2. $\mathrm{Var}(Y_i) = \mathrm{Var}(\beta_0 + X_i\beta_1 + \epsilon_i) = \mathrm{Var}(\epsilon_i) = \sigma^2$.
3. $\mathrm{Cov}(\epsilon_i, \epsilon_{i'}) = 0$
4. For constant a, $\mathrm{var}(aY_i) = a^2\mathrm{Var}(Y_i) = a^2\sigma^2$

$$
\begin{aligned}
\mathrm{Var}(\widehat{\beta}_1) &= \mathrm{Var}\left(\frac{\sum_{i=1}^{N}(Y_i - \overline{Y}_i)(X_i - \overline{X}_i)}{\sum_{i=1}^{N}(X_i - \overline{X}_i)^2}\right) \\
&= \left(\frac{1}{\sum_{i=1}^{N}(X_i - \overline{X}_i)^2}\right)^2 \mathrm{Var}\left(\sum_{i=1}^{N}(Y_i - \overline{Y}_i)(X_i - \overline{X}_i)\right) \\
&= \left(\frac{1}{\sum_{i=1}^{N}(X_i - \overline{X}_i)^2}\right)^2 \sum_{i=1}^{N}\mathrm{Var}\left((Y_i - \overline{Y}_i)(X_i - \overline{X}_i)\right) \\
&= \left(\frac{1}{\sum_{i=1}^{N}(X_i - \overline{X}_i)^2}\right)^2 \sum_{i=1}^{N}(X_i - \overline{X}_i)^2\mathrm{Var}\left(Y_i - \overline{Y}_i\right)
\end{aligned}
$$

$$
\begin{aligned}
&= \left(\frac{1}{\sum_{i=1}^{N}(X_i - \overline{X}_i)^2}\right)^2 \sum_{i=1}^{N}(X_i - \overline{X}_i)^2\sigma^2 \\
&= \frac{\sigma^2}{\sum_{i=1}^{N}(X_i - \overline{X}_i)^2}
\end{aligned}
$$

Inference on Slope

Feasible estimator:

$$\widehat{\sigma}^2 = \frac{\sum_{i=1}^N (Y_i - \widehat{Y}_i)^2}{N-2} \Rightarrow$$

$$\widehat{\mathrm{Var}}(\widehat{\beta}) = \frac{\widehat{\sigma}^2}{\sum_{i=1}^N (X_i - \overline{X}_i)^2}$$

Inference

CI: $\left[\widehat{\beta} - c_{1-\alpha/2}\sqrt{\widehat{\mathrm{Var}}(\widehat{\beta})}, \widehat{\beta} - c_{\alpha/2}\sqrt{\widehat{\mathrm{Var}}(\widehat{\beta})}\right]$

$$\frac{\widehat{\beta}}{\sqrt{\widehat{\mathrm{Var}}(\widehat{\beta})}} \sim t_{N-2}$$