Lecture 6

Normal Linear Regression

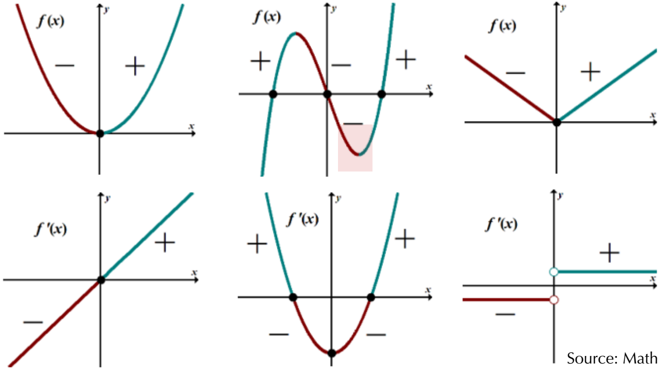
Haoyu Yue / yohaoyu@washington.edu
Ph.D. Student, Interdisciplinary Urban Design and Planning
University of Washington

RE 519 Real Estate Data Analytics and Visualization
Course Website: www.yuehaoyu.com/data-analytics-visualization/
Autumn 2025

Statistics Review

Derivative and Extreme Points

Derivative $f(x)'$ quantifies the sensitivity to change of a function's output with respect to its input.



Source: Math Hints

$f(x)$ will get a min/max when $f(x)' = 0$ and second derivative $f(x)'' \neq 0$. **But we cannot guarantee a global min/max.**

Function f(x)	**Derivative** f′(x)
c	0
x^n	$n\,x^{n-1}$
ax	a
e^x	e^x
$\ln(x)$	$1/x$

Rules
$(f + g)' = f' + g'$
$(fg)' = f'g + fg'$
$\left(\frac{f}{g}\right)' = \frac{(f'g - fg')}{g^2}$
$\left(f(g(x))\right)' = f'(g(x)) \cdot g'(x)$

Statistics Review

Distribution and Probability Density/Mass Function (PDF/PMF)

For discrete variables

A **probability distribution** is a function that gives the probabilities of the occurrence of possible events. Refer to the Seeing Theory.

Example: Normal Distribution. If $X \sim \mathrm{N}\left(\mu, \sigma^2\right)$

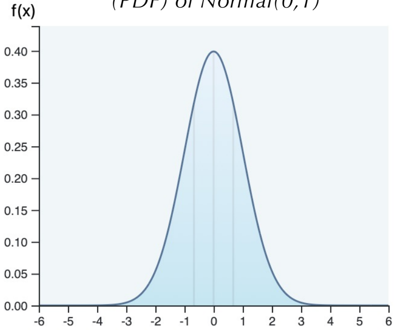
$$p(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{-(x-\mu)^2}{2\sigma^2}}$$

$$\begin{aligned} \text{Mean:} \quad & E[X] = \mu \\ \text{Variance:} \quad & \mathrm{Var}[X] = \sigma^2 \end{aligned}$$

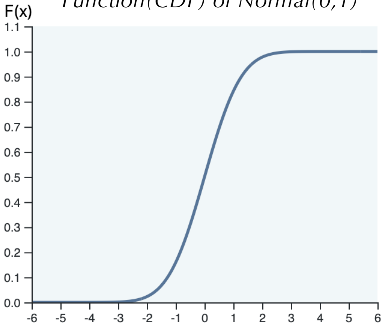
Expectation and Variance

Refer to Seeing Theory: https://seeing-theory.brown.edu/basic-probability/index.html

Probability Density Function (PDF) of Normal(0,1)



Cumulative Distribution Function(CDF) of Normal(0,1)

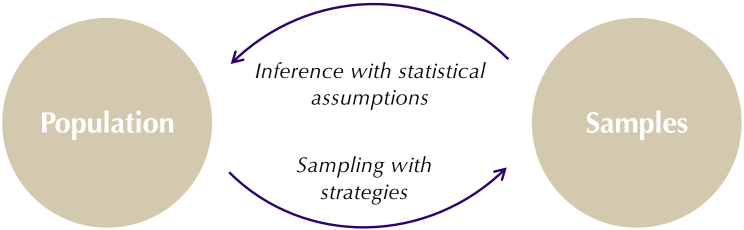


Statistics Review

Inference Statistics – Frequentist View

- We will never see the population
  - Q: If we want to study the crime rate and the property sale price, I have all the transaction data from 2024. Should I treat it as a population?
- There are some **consistent population parameters** (e.g., mean μ)
- We will only use samples to estimate those population parameters
  - e.g., sample mean $\overline{\mathrm{Y}}$ as an unbiased estimate of population mean μ
- However, some features of the samples are related to the estimations
  - Sample size n: larger n, better estimation with lower uncertainty
  - Sample variance S^2: lower S^2, better estimation with lower uncertainty

Note: There are two major beliefs in statistics -- - **Frequentist** and **Bayesian**. In Bayesian statistics (seeing theory), parameters are random variables; they will be updated based on prior knowledge and new evidence.



Contexts

Predicting Housing Index by Humans

As humans, we always want to predict the future. Based on some known data, I can make a guess on the housing index…

Region	Population	Other Information	My Guess Housing Index	Real Housing Index (Unknown)
New York, NY	19,940,274	…	22000	22000
Los Angeles, CA	12,927,614	…	19000	22200
Chicago, IL	9,408,576	…	17000	15600
Dallas, TX	8,344,032	…	10000	12900
Houston, TX	7,796,182	…	11000	12400

Known Data Prediction Real Value

Clearly, I cannot make an accurate guess using my brain. But, how to **define an accurate guess**?

In this section, please try to understand the flow and concepts. Do not worry too much about proofs or equations.

Contexts

Mean Squared Error (MSE)

My Guess Housing Index	Real Housing Index (Unknown)
22000	22000
19000	22200
17000	15600
10000	12900
11000	12400

Prediction Real Value

But, how to define an accurate guess?
An intuitive way is to get the average difference between **My Guess** and **the Real Housing Index**.

$$MAE = \frac{1}{n}\sum_{i=1}^{n} |\text{My Guess for City } i - \text{Real Housing Index for City } i|$$

That is a good way to measure the difference. It is called Mean Absolute Error (MAE), but the problem is that we don't like to work with the absolute value.

Instead, we like **square**. We use the average squares of the difference between My Guess and the Real Housing Index, which is called Mean Squared Error (MSE).

$$MSE = \frac{1}{n}\sum_{i=1}^{n} (\text{My Guess for City } i - \text{Real Housing Index for City } i)^2$$

Notes: MSE is a way to define the error (difference), but there are more ways. In regression, MSE is commonly used.

Contexts

Best MSE Predictors

Region	Population	Other Information	My Guess Housing Index	Real Housing Index (Unknown)
New York, NY	19,940,274	...	22000	22000
Los Angeles, CA	12,927,614	...	19000	22200
Chicago, IL	9,408,576	...	17000	15600
Dallas, TX	8,344,032	...	10000	12900
Houston, TX	7,796,182	...	11000	12400

Known Data *Prediction* *Real Value*

We want to use **known data (*X*)** to **minimize MSE to** make the difference as small as possible.
After some calculations (omitted), we can get:

$$\text{Best MSE Predictor} = E[\text{Housing Index}|\text{Known Data}] = E[Y|X]$$

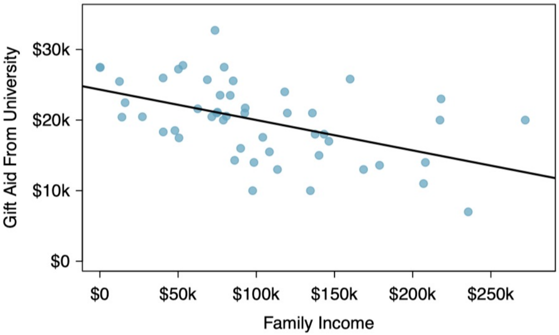
Given known data, the expectation of the housing index is the best predictor in terms of MSE!

What is Linear Regression

Purpose

$$\text{Best MSE Predictor} = E[\text{Housing Index}|\text{Known Data}] = E[Y|X]$$

But it is often **too ambitious**, as we have limited data, many variables, etc.
We need a simple way to find the relationship between X and Y.



Source: David Diez, et al., OpenIntro Statistics.

X	Y
1	2
1	4
1	6
0	4
0	4
0	8

$$E[Y|X=1] = 4$$

Best MSE Predictor when $X = 1$ in new predictions.

Linear Regression is a simple approach to finding the relationship when Y is continuous.

We also have other types of regression, such as Logistic (1/0), Poisson (Count), and Multinomial (categories).

What is Linear Regression

Linear Regression and Its Common Terms

Dependent variables
Response variables
Outcomes

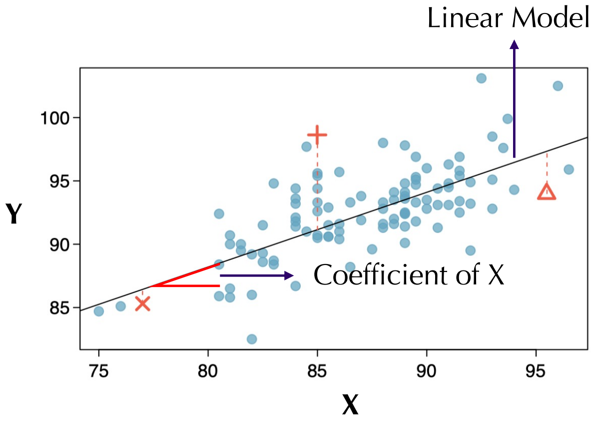
Independent variables
Explanatory variables
Features

$$Y_i = \beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k + \varepsilon_i$$

Intercept

Slope
Coefficient

Random
Error term



Source: David Diez, et al., OpenIntro Statistics.

People have different names for those terms, which is somewhat confusing.

What is Linear Regression

Linear Regression for Prediction

Independent variables
Explanatory variables
Features

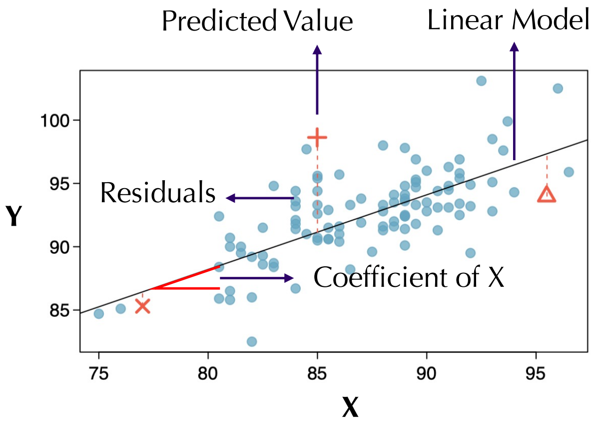
Fitted value
Predicted value

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \cdots + \hat{\beta}_k X_k$$

Estimated Intercept **Estimated Coefficient**

Residuals ← $e_i = Y_i - \hat{Y}_i$

Ture value



Source: David Diez, et al., OpenIntro Statistics.

What is Linear Regression

Error and Residual

$$Y_i = \beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k + \boxed{\varepsilon_i}$$

Error is part of the model we assumed to account for random effects that cannot be captured by the model. We cannot really know the errors.

$$\boxed{e_i} = Y_i - \hat{Y}_i \quad \text{where} \quad \hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \cdots + \hat{\beta}_k X_k$$

Residual is the difference between our predicted values and true values, which can be calculated. Ideally, $e_i \approx \varepsilon_i$

Looks like we get a good estimate for the relationship. **But what's the cost?**

Normal Linear Regression

Assumptions

Here, we are talking about the most classical regression – **Normal Linear Regression.**
We **make some assumptions** in order to use this linear regression. In a rigorous statistical analysis, we need to test those assumptions. See lab 7.

No perfect multicollinearity among X
X should not be a linear combination of each other
Example: total score (A+B), part A score, part B score ➔ all as X

Error terms are independent
Time series/spatial data is not independent: (spatial) autocorrelation (Wikipedia)

$$\boxed{Y_i} = \beta_0 + \beta_1 \boxed{X_1} + \cdots + \beta_k \boxed{X_k} + \boxed{\varepsilon_i} \longrightarrow \varepsilon_i \sim N(0, \sigma^2)$$

Normality and homoscedasticity for error terms
Errors should be normally distributed with the same variance

Linearity: Y changes linearly with X
We can do a transformation to X, such as log(X) and X^2, but not to beta

Normal Linear Regression

Parameters Point Estimation

To simplify the calculation, we use a model with just one explanatory variable.

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

$$MSE(\beta_0, \beta_1) = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \beta_0 - \beta_1 X_i)^2$$

Take the partial derivative of both β_0 and β_1, then set them to 0.
We can the make estimates:

$$\hat{\beta}_1 = \frac{\sum(X_i - \bar{X})(Y_i - \bar{Y})}{\sum(X_i - \bar{X})^2} \qquad \hat{\beta}_0 = \bar{Y} - \beta_1 \bar{X}$$

Note: $\bar{X}$ is the mean of X

X	1	3	3	5	5	6	8	9
Y	2	3	5	4	6	5	7	8

Can you calculate the estimates of β_0 and β_1 by hand?

Normal Linear Regression

Student-t tests for Individual Coefficients

We can always estimate β_1, but **whether the result is convincing?**

$$\hat{\beta}_1 = \frac{\sum(X_i - \bar{X})(Y_i - \bar{Y})}{\sum(X_i - \bar{X})^2}$$

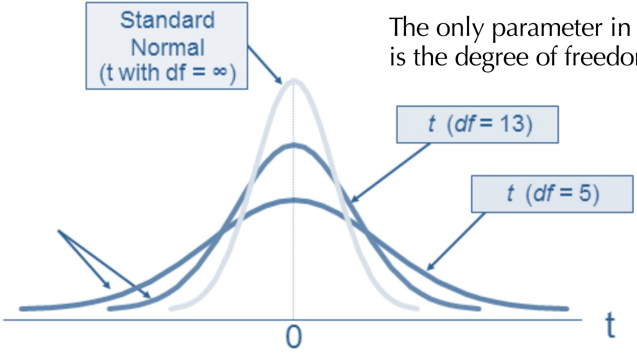
Luckily, can prove that $\hat{\beta}$ follows a certain distribution.

True Parameter

$$\frac{(\hat{\beta}_1 - \beta_1)\sqrt{\sum(x_i - \bar{X})^2}}{\hat{\sigma}} \sim t_{n-k-1} = t_{n-2}$$

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{n} e_i^2}{n-k-1}$$

K is the number of explanatory variables. You don't have to understand the t-test, but you need to know that $\hat{\beta}$ follows a certain distribution.



Source: financetrain.com

The only parameter in the t-distribution is the degree of freedom (df)

Normal Linear Regression

Hypothesis Testing and Errors

Suppose we know: $\dfrac{(\hat{\beta}-\beta)\sqrt{\sum(x_i-\bar{X})^2}}{\hat{\sigma}} \sim t_{n-2}$

We would like to test whether $\beta = 0$. Intuitively, the test means **whether X has a significant relationship with Y**.

We frame this question into a hypothesis testing framework:

$$H_0 : \beta = 0 \quad H_1 : \beta \neq 0$$

- H_0: null hypothesis – “bad/normal”, we want to reject it
- H_1: alternative hypothesis

Thinking about the process, we may have 4 possible outcomes about the truth and our results:

Real-world truth, only one is true

Our test result

	H_0 is True	H_1 is True
Reject H_0	Type I Error (α)	Correct
Accept H_0	Correct	Type II Error
Sum	1	1

Normal Linear Regression

t-values and p-values

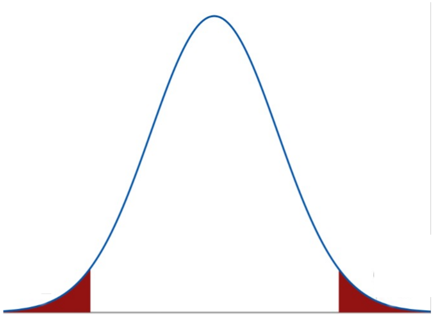
Suppose we know: $$\frac{(\hat{\beta}-\beta)\sqrt{\sum(x_i-\bar{X})^2}}{\hat{\sigma}} \sim t_{n-2}$$

And we are testing: $H_0 : \beta = 0 \quad H_1 : \beta \neq 0$

Under the null, the **t-value** we observed: $$\frac{\hat{\beta}\sqrt{\sum(x_i-\bar{X})^2}}{\hat{\sigma}}$$

If H_0 is true, it should follow t_{n-2}

Draw PDF of this distribution



Source: Department of Statistics, Penn State University.

p-value is the probability that the t-value is more extreme (the area of **red color**)

- If **p-value < 0.05**, we can say it is statistically significant. But picking 0.05 is just a convention, not a law.

"The difference between significant and not significant is not itself significant."

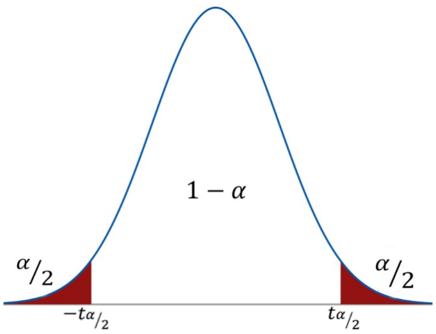
-- *Andrew Gelman and Hal Stern*

Normal Linear Regression

Confidence Interval of True Parameter $\boldsymbol{\beta}$

Suppose we know: $\dfrac{(\hat{\beta}-\beta)\sqrt{\sum(x_i-\bar{X})^2}}{\hat{\sigma}} \sim t_{n-2}$

We have the estimated $\hat{\boldsymbol{\beta}}$, and we can get a possible range of $\boldsymbol{\beta}$ under a certain significance level.



Source: Department of Statistics, Penn State University.

$\dfrac{(\hat{\beta}-\beta)\sqrt{\sum(x_i-\bar{X})^2}}{\hat{\sigma}}$ should be within $(-t_{\alpha/2}, t_{\alpha/2})$ under a 1-α significance level

After the calculation, β should be within $\left(\hat{\beta}-\dfrac{t_{\alpha/2}\hat{\sigma}}{\sqrt{\sum(x_i-\bar{X})^2}}, \hat{\beta}+\dfrac{t_{\alpha/2}\hat{\sigma}}{\sqrt{\sum(x_i-\bar{X})^2}}\right)$, also called **Confidence Interval**. (Seeing Theory)

R will give us the results of all things here, so we don't have to calculate by ourselves.

CI tells us the uncertainty around our estimates.

Normal Linear Regression

How much does the model explain the data?

First, let's think about why we care about the difference in data.

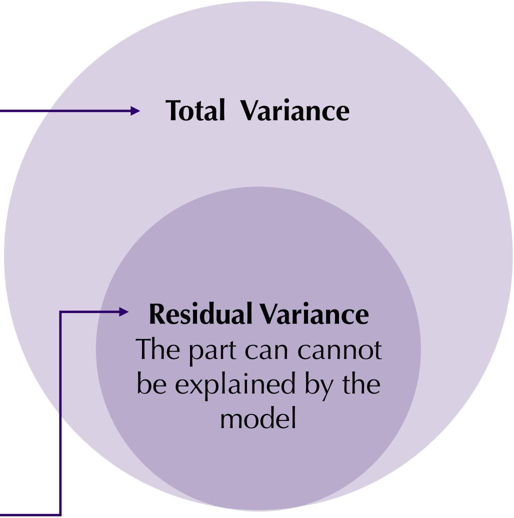
- People's ability to pay the mortgage
  - Income, occupation, disease, families, pandemic
- Sale price
  - Locations, area, view, luck, timing
- Trying to understand the reasons for the difference.

Which statistics did we use to represent the difference in data?

- Variance

We can define a value to show how much variance can be explained by the model - R^2

$$R^2 = 1 - \frac{\mathrm{SSR}}{\mathrm{SST}} \quad \text{where} \quad SST = \sum (y_i - \bar{y})^2; \quad SSR = \sum (y_i - \hat{y})^2$$



Normal Linear Regression

Results from R and Interpretations

Y – response variable **X – explanatory variable**

```
Call:
lm(formula = sale_price ~ sqft + year_built, data = sales)

Residuals:
     Min       1Q   Median       3Q      Max
-1877471  -322074   -92840   173424 13832047

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  7.982e+06  3.268e+05   24.42   <2e-16 ***
sqft         5.082e+02  5.975e+00   85.06   <2e-16 ***
year_built  -4.009e+03  1.672e+02  -23.98   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 653700 on 15651 degrees of freedom
Multiple R-squared:  0.3162,    Adjusted R-squared:  0.3161
F-statistic:  3619 on 2 and 15651 DF,  p-value: < 2.2e-16
```

Regression details in the lab page.

Distribution of Residuals

$$e_i = Y_i - \hat{Y}_i$$

Estimated coefficient $\hat{\beta}_1$

Interpretation: expected change in price for 1 sqrt increase in area, holding all year variables constant.

Many times, in social science, we care about the sign of β instead of the absolute value.

R-squared $R^2 = 1 - \mathrm{SSR/SST}$

Interpretation: the model can explain 31.62% of data variance.

Standard Error of beta

Can be used to calculate confidence intervals

t-values and p-values

Interpretation: t-value is 85.06, and the probability (<2e-16) of seeing such an extreme t-value if the true coefficient were 0.

Adjusted R-squared

Penalizes complexity (the number of X) as well. **More helpful than R^2 in multivariate regression.**

Thank you!

Haoyu Yue / yohaoyu@washington.edu
Ph.D. Student, Interdisciplinary Urban Design and Planning
University of Washington

RE 519 Real Estate Data Analysis and Visualization
Course Website: www.yuehaoyu.com/data-analytics-visualization/
Autumn 2025

The course was developed based on previous instructors: Christian Phillips, Siman Ning, Feiyang Sun
Cover page credits: Visax

Optional, for Reference

Type I Errors and the Controls

Suppose we know: $$\frac{(\hat{\beta}-\beta)\sqrt{\sum(x_i-\bar{X})^2}}{\hat{\sigma}} \sim t_{n-2}$$

And we are testing: $H_0: \beta = 0 \quad H_1: \beta \neq 0$

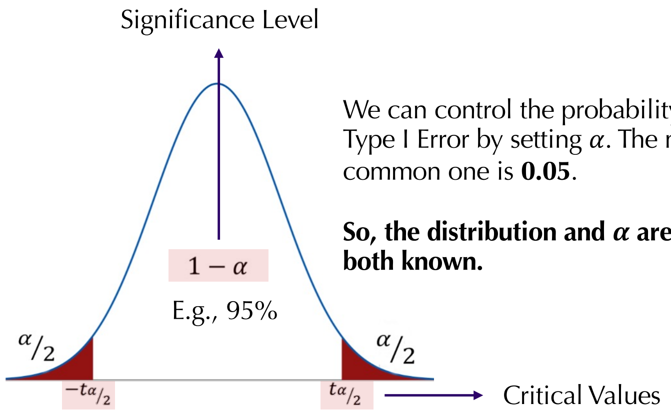
Thinking about the process, we may have 2 types of errors, and we want to minimize them for sure.

	H_0 is True	H_1 is True
Reject H_0	Type I Error (α)	Correct
Accept H_0	Correct	Type II Error
Sum	1	1

Type I Error can be rewritten as:

$$\alpha = P(\text{reject } H_0 \mid H_0 \text{ true})$$

When H_0 is true ($\beta = 0$), we know everything about the $\hat{\beta}$.



Source: Department of Statistics, Penn State University.

We can control the probability of Type I Error by setting α. The most common one is **0.05**.

So, the distribution and α are both known.

Optional, for Reference

Type II Errors and the Controls

Suppose we know: $$\frac{(\hat{\beta}-\beta)\sqrt{\sum(x_i-\bar{X})^2}}{\hat{\sigma}} \sim t_{n-2}$$

And we are testing: $H_0 : \beta = 0 \quad H_1 : \beta \neq 0$

Thinking about the process, we may have 2 types of errors, and we want to minimize them for sure.

	H_0 is True	H_1 is True
Reject H_0	Type I Error (α)	Correct
Accept H_0	Correct	Type II Error
Sum	1	1

Type II Error can be rewritten as:

$$P(\text{fail to reject } H_0 \mid H_1 \text{ true})$$

When H_1 is true, we don't know the true β is.

So, we cannot directly control Type II Error!
There are some ways to minimize it:

- Increase sample size
- Reduce noise (better measure, controlled experiment, etc.)
- Better models
- ……