Introduction to Statistical Inference

Marta Bofill Roig

March 18, 2025

UNIVERSITAT POLITÈCNICA
DE CATALUNYA
BARCELONATECH

Outline

Warming up

Generic hypothesis test

- Consider two hypotheses:
  - $H_0 : \mu = \mu_0$
  - $H_1 : \mu \neq \mu_0$
- μ_0 is a (known) value of interest.
- Assume H_0 is true: $\mu = \mu_0$.
- We want to determine the probability of seeing $\overline{Y}$ (our sample mean).

Generic hypothesis test (cont'd)

Using the t-distribution

- If Y is normally distributed, under H_0:

$$\frac{\overline{Y} - \mu_0}{s_{\overline{Y}}} \sim t_{n-1}$$

  where $s_{\overline{Y}}$ is the standard error of the mean.

- This statistic follows a t-distribution with $n - 1$ degrees of freedom.

Generic hypothesis test (cont'd)

- Define a small probability α (e.g., 0.05).
- If the probability of seeing our $\overline{Y}$ is less than α, we reject H_0.
- This means the evidence is strong enough to refute H_0.

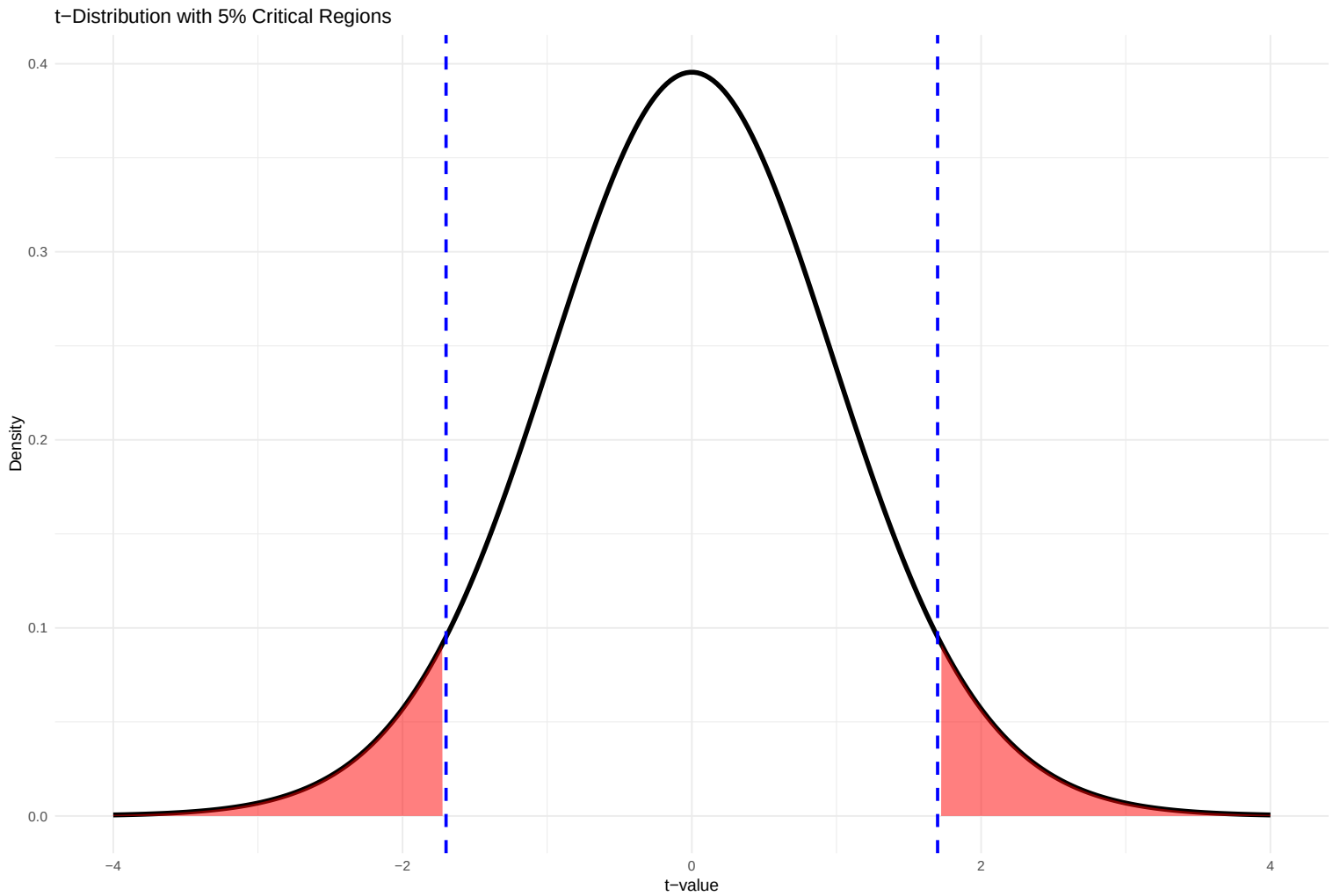
Critical values and the t-distribution

- The statistic follows a t_{n-1} **distribution**.
- There exists a **critical value** $t_{n-1,\alpha/2}$ such that:

$$P\left(\left|\frac{\overline{Y}-\mu_0}{s_{\overline{Y}}}\right| > t_{n-1,\alpha/2}\right) = \alpha$$

- $t_{n-1,\alpha/2}$ defines the **rejection region**.

Critical values and the t-distribution



- Each shaded region has area $\alpha/2$.
- Total shaded area is α.

Wrapping up – let's put all together

- We use the t-**distribution** to assess the **likelihood** of our **sample mean**.
- We define a **critical value** based on α.
- If the test statistic falls in the **rejection region**, we reject H_0.

Decision tools for hypothesis testing (cont'd)

- Instead of comparing the test statistic to a critical value, we can use probabilities.
- Find the probability of observing a test statistic as extreme as the one we calculated.
- This probability is called the p**-value**.

Decision tools for hypothesis testing (cont'd)

Calculating the p-value:

- If t_{n-1} is a t-distributed random variable with $n-1$ degrees of freedom, find:

$$P(|t_{n-1}| > \text{value of test statistic we saw})$$

- Compare this probability (p-value) to α.
- If p-value $< \alpha$, reject H_0.

Linear Regression Models

Regression Models

Another problem that arises in the biological and physical sciences, economics, industrial applications, and biomedical settings is that of

investigating the **relationship** between two (or more) variables.

Linear Regression Models (cont'd)

- We will focus on the **simplest** case:
  - **Two variables**
  - Relationship is assumed to be a **straight line**
- We will explore the relationship between two variables using a linear model.

Linear Regression Models (cont'd)

- "Linear" refer to how a component of an equation describing a relationship enters that relationship.

Linear vs. Nonlinear Examples

- **Linear Example:** $Y_{ij} = \mu + \tau_i + \epsilon_{ij}$
  - Parameters $(\mu, \tau_i, \epsilon_{ij})$ enter directly.
- **Nonlinear Example:** $Y_{ij} = \exp(\mu + \tau_i) + \epsilon_{ij}$
  - Parameters (μ, τ_i) enter through an exponential function.

Linear Regression Models (cont'd)

- "Linear regression" indicates that parameters enter in a straightforward rather than complicated way.
- "Simple linear regression" refers to the particular case where the relationship is a straight line.

Linear Regression Models (cont'd)

Which is now the situation?

- We observe pairs of X and Y values from a sample of experimental units.
- We want to understand the relationship between X and Y.
- The nature of the relationship depends on the data source.

Linear Regression Models (cont'd)

In experimental studies/randomised studies...

- Observations are planned as a result of an experiment.
- Examples:
  - X = Drug dose, Y = Blood pressure response
  - X = Toxic substance concentration, Y = Mutant offspring count

Linear Regression Models (cont'd)

- In experimental data, we **control** the values of X.
- We observe the resulting values of Y.

Linear Regression Models (cont'd)

In observational studies...

- We observe both X and Y values.
- Neither X nor Y is under our control.
- Examples:
  - X = Weight, Y = Height of a human
  - X = Average plant height, Y = Yield

Linear Regression Models (cont'd)

- In experimental data, X is controlled by the investigator.
- In observational data, no such distinction exists.
- Y is always the observed outcome, while X may or may not be manipulated.

Linear Regression Models (cont'd)

- Variables can be functionally related: $Y = g(X)$.
- g can be derived from theory or provide an empirical description.
- We seek to identify systematic relationships between X and Y.

Linear Regression Models (cont'd)

- Formally describe and assess the relationship between X and Y.
- Based on a sample of observations.

Linear Regression Models (cont'd)

In experimental studies:

- We model the relationship between X and Y using:

$$Y = g(X) + \epsilon$$

- $g(X)$ describes the relationship.
- ϵ represents the additive error.

Linear Regression Models (cont'd)

- Y = Response or Dependent Variable
- X = Independent Variable, Covariate

Linear Regression Models (cont'd)

- We characterize how the response (Y) changes with the value of X.
- We control X independently.
- The distinction between X and Y is clear in experimental data.

Summary

- Experimental data is modeled with an additive error term.
- Y is the response, X is the controlled independent variable.

Simple regression model

Simple regression model – Straight Line Model

- For experimental data with fixed X, assume a linear relationship between Y and X.
- The model is:

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

- Where:
  - β_0 and β_1 are constants.
  - ϵ is the error term.

Simple regression model – Intercept and Slope

- **Intercept (β_0):**
  - The value of Y when $X = 0$.
- **Slope (β_1):**
  - The rate of change in Y per unit change in X.
  - $\beta_1 = \frac{\text{change in Y}}{\text{change in X}}$.

Summary

- Simple linear regression models a linear relationship between Y and X.
- β_0 is the intercept, and β_1 is the slope.

Simple regression model (cont'd)

Unknown β_0 and β_1

- We need to estimate β_0 and β_1 to understand the relationship.
- We collect data: (X_i, Y_i) for $i = 1, ..., n$.
- Goal: Fit the model $Y = \beta_0 + \beta_1 X + \epsilon$ to the data.

Simple regression model (cont'd)

Potential applications of the fitted model:

- **Quantify** the relationship between Y and X.
- **Predict** a new response Y_0 for a given X_0.
- **Calibrate**: Estimate X_0 for a new observed Y_0.

Simple regression model (cont'd)

- Model: $Y = \beta_0 + \beta_1 X + \epsilon$
- **Regression relationship**: Straight line $\beta_0 + \beta_1 X$
- β_0 and β_1 are **regression coefficients/parameters**.

Simple regression model (cont'd)

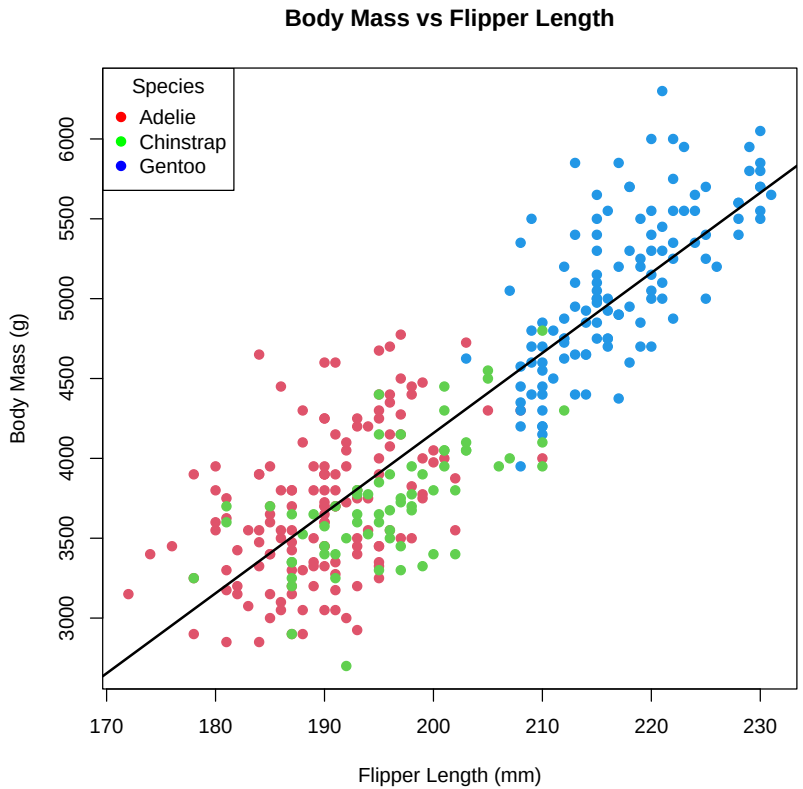
- $Y_i = \beta_0 + \beta_1 X_i + \epsilon_i = \mu_i + \epsilon_i$
- $\mu_i = \beta_0 + \beta_1 X_i$
- μ_i is the mean of observations at X_i.

Simple regression model (cont'd)

- The "errors" ϵ_i represent inherent variation.
- They cause Y_i to deviate from its mean $\mu_i = \beta_0 + \beta_1 X_i$.
- This is considered **experimental error**.

Simple regression model (cont'd)

- $\mu_i = \beta_0 + \beta_1 X_i$ represents the mean of Y_i at X_i.
- Means lie on a **straight line**.
- Observed Y_i values **vary** around μ_i and are scattered.



Simple regression model (cont'd)

- **Fit** the line to the data.
- **Estimate** β_0 and β_1.
- This characterizes the mean at any X value.

Summary

- Simple linear regression allows mean estimation with single observations.
- Error terms represent inherent variation.
- We aim to find the line that best fits the data, estimating β_0 and β_1.

Fitting the Simple Linear Regression Model

- We aim to describe the practical implementation of fitting a simple linear regression model.
- Assume X values are fixed.
- Model: $Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$ for $i = 1, ..., n$.
- We estimate β_0 (intercept) and β_1 (slope).

Assumptions for Inference

To make inferences about β_0 and β_1, make predictions, etc., we make the following assumptions:

- Assumption 1: Independence
- Assumption 2: Constant Variance
- Assumption 3: Normality

Assumption 1: Independence

- The observations $Y_1, ..., Y_n$ are **independent**.
- They are not related in any way.

Assumption 2: Constant Variance

- The observations $Y_1, ..., Y_n$ have the **same variance** σ^2.
- Each Y_i has mean $\mu_i = \beta_0 + \beta_1 X_i$.
- The variation in possible Y_i values is the same for all X_i.

Assumption 3: Normality

- The observations Y_i are **normally distributed**.
- Mean $\mu_i = \beta_0 + \beta_1 X_i$.
- Variance σ^2 (same as in Assumption 2).
- For each X_i, possible Y_i values are well-represented by a normal distribution.

Method: Least Squares

- We fit the model using the method of **least squares**.
- (Further details on the least squares method would be included in subsequent frames.)

Estimating β_0 and β_1

- We use the **method of least squares** to estimate β_0 and β_1.
- Intuitively appealing and mathematically appropriate under normality.

Estimating β_0 and β_1 (cont'd)

- Deviation: $Y_i - (\beta_0 + \beta_1 X_i) = \epsilon_i$
- Measures vertical distance of Y_i from the line.
- Overall deviation: $\sum_{i=1}^{n} \{Y_i - (\beta_0 + \beta_1 X_i)\}^2$
- Similar to sample variance, ignoring signs but accounting for magnitude.

Estimating β_0 and β_1 (cont'd)

- Choose $\hat{\beta}_0$ and $\hat{\beta}_1$ to **minimize** the overall variation.
- Minimize: $\sum_{i=1}^{n}(Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i)^2$
- Attributing as much variation as possible to the straight line relationship.

Estimating β_0 and β_1 (cont'd)

- Least squares line minimizes the sum of squared vertical discrepancies.
- Least squares minimizes the sum of squared deviations.
- This finds the "best fit" line by attributing variation to the linear relationship.

Estimating β_0 and β_1 (cont'd)

- Calculus is used to find the estimators $\hat{\beta}_0$ and $\hat{\beta}_1$ that minimize the sum of squared deviations.

Estimating β_0 and β_1 (cont'd)

- $S_{XY} = \sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y}) = \sum_{i=1}^{n} X_i Y_i - \frac{(\sum_{i=1}^{n} X_i)(\sum_{i=1}^{n} Y_i)}{n}$
- $S_{XX} = \sum_{i=1}^{n}(X_i - \overline{X})^2 = \sum_{i=1}^{n} X_i^2 - \frac{(\sum_{i=1}^{n} X_i)^2}{n}$
- $S_{YY} = \sum_{i=1}^{n}(Y_i - \overline{Y})^2 = \sum_{i=1}^{n} Y_i^2 - \frac{(\sum_{i=1}^{n} Y_i)^2}{n}$
- $\overline{X}$ and $\overline{Y}$ are sample means of X_i and Y_i, respectively.

Estimator Formulas

- $\hat{\beta}_1 = \frac{S_{XY}}{S_{XX}}$
- $\hat{\beta}_0 = \overline{Y} - \hat{\beta}_1 \overline{X}$

Estimating β_0 and β_1 (cont'd)

- The fitted straight line is: $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i$
- $\hat{Y}_i$ are the **predicted values** or "best guesses" for the means at each X_i.
- Actual observed values Y_i may not fall on the line.

Example: Oxygen Consumption in Birds

- Oxygen consumption rates (Y) measured at different temperatures (X).
- Temperatures (X) were set by the investigator, justifying fixed X assumption.

Example: Oxygen Consumption in Birds (cont'd)

X (degrees Celsius)	Y (ml/g/hr)
-18	5.2
-15	4.7
-10	4.5
-5	3.6
0	3.4
5	3.1
10	2.7
19	1.8

Example: Oxygen Consumption in Birds (cont'd)

- Analyse the relationship between temperature (X) and oxygen consumption (Y) using simple linear regression.
- Estimate the regression coefficients.
- Interpret the results.

Example: Oxygen Consumption in Birds (cont'd)

- We have $n = 8$ data points.
- $\sum_{i=1}^{n} Y_i = 29$, $\overline{Y} = 3.625$, $\sum_{i=1}^{n} Y_i^2 = 114.04$
- $\sum_{i=1}^{n} X_i = -14$, $\overline{X} = -1.75$, $\sum_{i=1}^{n} X_i^2 = 1160$
- $\sum_{i=1}^{n} X_i Y_i = -150.4$

Example: Oxygen Consumption in Birds (cont'd)

- $S_{XY} = -150.4 - \frac{(29)(-14)}{8} = -99.65$
- $S_{XX} = 1160 - \frac{(-14)^2}{8} = 1135.5$
- $S_{YY} = 114.04 - \frac{29^2}{8} = 8.915$

Example: Oxygen Consumption in Birds (cont'd)

- $\hat{\beta}_1 = \frac{S_{XY}}{S_{XX}} = \frac{-99.65}{1135.5} = -0.0878$
- $\hat{\beta}_0 = \overline{Y} - \hat{\beta}_1 \overline{X} = 3.625 - (-0.0878)(-1.75) = 3.4714$

Example: Oxygen Consumption in Birds (cont'd)

- We calculated the summary statistics, S values, and regression coefficients.
- The fitted line is $\hat{Y}_i = 3.4714 - 0.0878 X_i$.

Standard Deviations for β_1 and β_0

- We need to assess the precision of our estimates for β_0, β_1, and the regression line.

Standard Deviations for β_1 and β_0 (cont'd)

- Under our assumptions, the standard deviations of $\hat{\beta}_1$ and $\hat{\beta}_0$ are:

$$SD(\hat{\beta}_1) = \frac{\sigma}{\sqrt{S_{XX}}}$$

$$SD(\hat{\beta}_0) = \frac{\sigma\sqrt{\sum_{i=1}^{n} X_i^2}}{\sqrt{nS_{XX}}}$$

- σ is unknown, so we estimate it with s.

Standard Deviations for β_1 and β_0 (cont'd)

- Estimated standard deviations (standard errors) are:

$$EST\ SD(\hat{\beta}_1) = \frac{s}{\sqrt{S_{XX}}}$$

$$EST\ SD(\hat{\beta}_0) = \frac{s\sqrt{\sum_{i=1}^{n} X_i^2}}{\sqrt{nS_{XX}}}$$

Confidence Intervals for β_1 and β_0

- Under our assumptions:

$$\frac{\hat{\beta}_1 - \beta_1}{EST\ SD(\hat{\beta}_1)} \sim t_{n-2} \quad \text{and} \quad \frac{\hat{\beta}_0 - \beta_0}{EST\ SD(\hat{\beta}_0)} \sim t_{n-2}$$

- Similar to single mean and difference of means, using t-distribution due to estimating σ with s (n-2 degrees of freedom).

Confidence Intervals for β_1 and β_0 (cont'd)

- We provide confidence intervals for the true values of β_1 and β_0.
- Similar to confidence intervals for means or differences of means.
- Uses the t-distribution with $n - 2$ degrees of freedom.

Confidence Intervals for β_1 and β_0 (cont'd)

- Interval for β_1:

$$\{\hat{\beta}_1 - t_{n-2,\alpha/2} EST\ SD(\hat{\beta}_1), \hat{\beta}_1 + t_{n-2,\alpha/2} EST\ SD(\hat{\beta}_1)\}$$

- Interval for β_0:

$$\{\hat{\beta}_0 - t_{n-2,\alpha/2} EST\ SD(\hat{\beta}_0), \hat{\beta}_0 + t_{n-2,\alpha/2} EST\ SD(\hat{\beta}_0)\}$$

Confidence Intervals for β_1 and β_0 (cont'd)

- Imagine conducting many experiments with the same X values.
- For each experiment, calculate confidence intervals for β_1 and β_0.
- In $100(1-\alpha)\%$ of these experiments, the true values of β_1 and β_0 will fall within the intervals.
- The intervals depend on the data and the experimental procedure.

Example: Oxygen Consumption Data

- $n = 8$
- $t_{6,0.025} = 2.447$
- $\hat{\beta}_0 = 3.4714$
- $\hat{\beta}_1 = -0.0878$
- $s^2 = 0.028$

Example: Oxygen Consumption in Birds (cont'd)

- $EST\ SD(\hat{\beta}_0) = \frac{\sqrt{0.028}\sqrt{1160}}{\sqrt{8(1135.5)}} = 0.06012$
- $EST\ SD(\hat{\beta}_1) = \frac{\sqrt{0.028}}{\sqrt{1135.5}} = 0.0050$

Example: Oxygen Consumption in Birds (cont'd)

- For β_0: $3.4714 \pm (2.447)(0.0601) = (3.3216, 3.6185)$
- For β_1: $-0.0878 \pm (2.447)(0.0050) = (-0.0999, -0.0755)$

Regression functions in R

Regression in R

Regression in R is as simple as `lm(y ~ x)`, in which "`lm`" stands for "linear model".

```
> tips = read.table("RestaurantTips.txt",h=T)
> lm(Tip ~ Bill, data=tips)
Call:
lm(formula = Tip ~ Bill, data = tips)
Coefficients:
(Intercept)  Bill
-0.2923      0.1822
```

It is better to save the model as an object,

```
lmtips = lm(Tip ~ Bill, data=tips)
```

and then we can get a more detailed output by viewing the `summary()` of the model object. The output is shown in the next slide.

Regression in R

```
> summary(lmtips)

Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.292267   0.166160  -1.759  0.0806 .
Bill         0.182215   0.006451  28.247 <2e-16 ***
---
Residual standard error: 0.9795 on 155 degrees of freedom
Multiple R-squared: 0.8373, Adjusted R-squared: 0.8363
F-statistic: 797.9 on 1 and 155 DF, p-value: < 2.2e-16
```

- The column *Estimate* gives the LS estimate for the intercept $b_0 = -0.292267$ and the slope $b_1 = 0.182215$
- The column *Std. Error* gives $SE(b_0)$ and $SE(b_1)$:

$$SE(b_0) = 0.166160, \quad SE(b_1) = 0.006451$$

Example: Confidence Interval for β_1

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-0.292267	0.166160	-1.759	0.0806
Bill	0.182215	0.006451	28.247	<2e-16

As $df = n - 2 = 157 - 2 = 155$, t^* for a 95% CI is 1.975 (between 1.97 and 1.98).

	0.1	0.05	0.025	0.01	0.005
one tail	0.1	0.05	0.025	0.01	0.005
two tails	0.2	0.10	0.050	0.02	0.010
df 150	1.29	1.66	1.98	2.35	2.61
200	1.29	1.65	1.97	2.35	2.60

Hence the 95% CI for β_1 is:

$$b_1 \pm t^* SE(b_1) = 0.182215 \pm 1.975 \times 0.006451$$
$$= 0.182215 \pm 0.01274 \approx (0.169, 0.195).$$

Interpretation: With 95% confidence, for each additional dollar in the bill, the customers gave 16.9 cents to 19.5 cents more tips on average.

Example: Confidence Interval for β_1

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-0.292267	0.166160	-1.759	0.0806
Bill	0.182215	0.006451	28.247	<2e-16

- Note t-values $b_i/SE(b_i)$ are simply the ratio of the numbers in the "Estimate" column and the numbers in the "Std. Error" column, e.g.,

$$-1.759 = \frac{-0.292267}{0.166160}, \quad 28.247 = \frac{0.182215}{0.006451}$$

- Testing $H_0 : \beta_1 = 0$ is equivalent to testing whether the amount of tips is linearly related to the amount of the bill. The small P-value $< 2 \times 10^{-16}$ asserts that the relation is significant.

Example: Test for the Slope β_1

A general rule for waiters is to tip 15 to 20% of the pre-tax bill. That is, $\beta_0 = 0$ and β_1 is between 0.15 to 0.20.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-0.292267	0.166160	-1.759	0.0806
Bill	0.182215	0.006451	28.247	<2e-16

- R tests $\beta_0 = 0$ for us: t-statistic = -1.759, 2-sided p-value = 0.0806
- To test $H_0 : \beta_1 = 0.2$ v.s. $H_A : \beta_1 < 0.2$. The t-statistic is

$$t = \frac{b_1 - 0.2}{SE(b_1)} = \frac{0.182215 - 0.2}{0.006451} = -2.757$$

with df = 155, the one-sided p-value is < 0.005.

	0.1	0.05	0.025	0.01	0.005
one tail	0.1	0.05	0.025	0.01	0.005
two tails	0.2	0.10	0.050	0.02	0.010
df 150	1.29	1.66	1.98	2.35	2.61
200	1.29	1.65	1.97	2.35	2.60

Conclusion: Customers of this restaurant gave less than 20% the bill as tips on average.

How to Read R Outputs for Regression?

```
Residual standard error:  0.9795 on 155 degrees of freedom

Multiple R-squared:  0.8373, Adjusted R-squared:  0.8363

F-statistic:  797.9 on 1 and 155 DF, p-value:  < 2.2e-16
```

- `Residual standard error:  0.9795 on 155 degrees of freedom`
  - This gives the estimate s_e of σ, which is 0.9795.
  - $df = n - 2 = 157 - 2 = 155$
- `Multiple R-squared:  0.8373` gives $r^2 = 0.8373$, Bill size explained 83.73% of the variation in tipping amount.
  - The correlation between bill size and tips is $r = \sqrt{r^2} = \sqrt{0.8373} = 0.915$.
- `Adjusted R-squared`: Ignore this.
- `F-statistic:  797.9 on 1 and 155 DF, p-value:  < 2.2e-16` Skip.

Overview of Regression functions in R

R provides several functions for different types of regression models:

- **Linear Regression:** `lm()`
- **Multiple Linear Regression:** `lm()`
- **Logistic Regression:** `glm()` with `family = binomial`
- **Poisson Regression:** `glm()` with `family = poisson`
- **Ridge/Lasso Regression:** `glmnet()`

References and other materials

1. D.R. Cox, E. J. Snell. Applied Statistics - Principles and Examples. Chapman & Hall, 1981.

2. Spiegelhalter, David. The art of statistics: Learning from data. Penguin UK, 2019.

3. James, G., Witten, D., Hastie, T., Tibshirani, R. An introduction to statistical learning. Springer, 2013.

4. Teaching courses from `https://www4.stat.ncsu.edu/~davidian/`

5. Teaching courses from `https://www.drizopoulos.com/`

That's all folks!

UNIVERSITAT POLITÈCNICA
DE CATALUNYA
BARCELONATECH