Linear Regression: Review, Extension and Diagnosis

Cheng Peng

STA200: Statistics II

Contents

1 Introduction

Simple Linear Regression (SLR) is a statistical method used to model the relationship between a single independent (predictor) variable (X) and a dependent (response) variable (Y). It assumes a linear relationship between the two variables and is widely used for prediction and inference.

Purpose of Simple Linear Regression

- To quantify the relationship between X and Y. *For example, price vs. demand in economics, drug dosage vs. effect in medicine, stress vs. material strain in engineering, etc..*
- To predict the value of Y for a given value of X. *For example, we may want to use house footage to predict the house price.*
- To test hypotheses about the relationship between X and Y. *For example, we may want to test whether hours of study influence the course grade.*

2 Structure of the Simple Linear Regression Model

The simple linear regression (SLR) model is represented as

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

Where Y is a dependent variable (response), X is an independent variable (predictor) **which is assumed to be non-random**. β_0 is the intercept (value of Y when $X = 0$), β_1 is the slope (the change in Y per unit change in X). ϵ = " "residual** random error term (assumed $\epsilon \to N(0, \sigma^2)$, see the assumptions of the linear regression model in the subsequent section).

For example, the practical regression between price and demand is expressed in the following.

$$\text{price} = \beta_0 + \beta_1 \times \text{demand} + \epsilon.$$

With a given dataset, we can estimate β_0 and β_1 using the least squares method, denoted respectively by b_0 and b_1. The estimated regression line (fitted model) is:

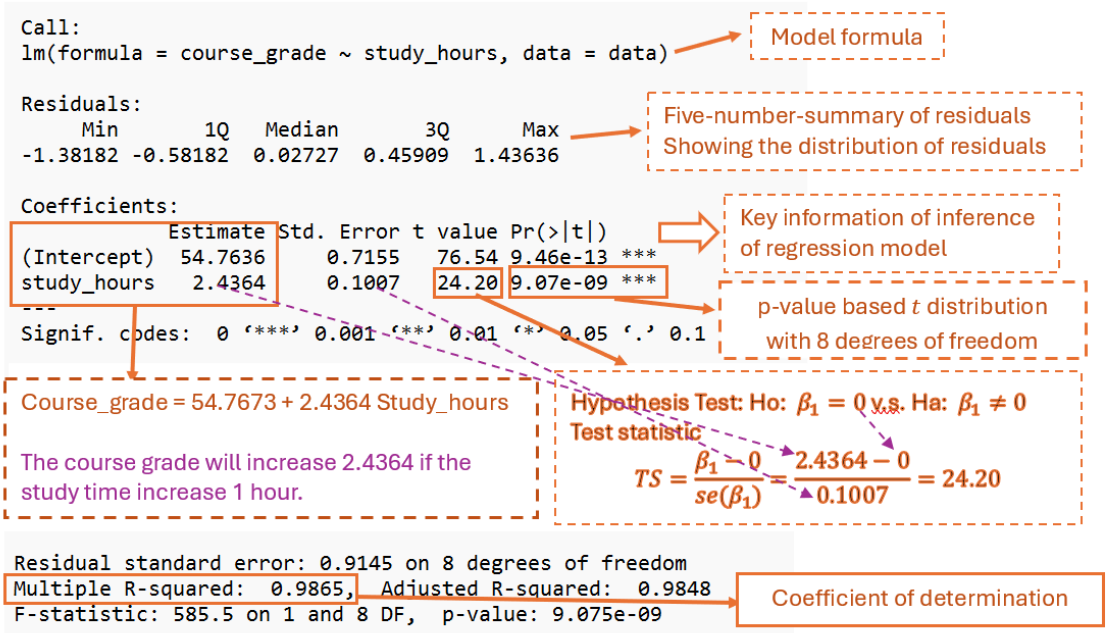
$$\hat{Y} = b_0 + b_1 X.$$

The **estimated** residuals are defined to be the estimated value $\hat{Y}$ and the observed Y from the data set. In R, we can extract estimated $\hat{Y}$ and residuals using R commands `fitted(model.name)` and ‘resid(model.name“

Example 1: Simple linear regression to assess the relationship between *study hours* and *course grade*. We us simulated data in the following R code

```
##
## Call:
## lm(formula = course_grade ~ study_hours, data = example.data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.38182 -0.58182  0.02727  0.45909  1.43636
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  54.7636     0.7155   76.54 9.46e-13 ***
## study_hours   2.4364     0.1007   24.20 9.07e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.9145 on 8 degrees of freedom
## Multiple R-squared:  0.9865, Adjusted R-squared:  0.9848
## F-statistic: 585.5 on 1 and 8 DF,  p-value: 9.075e-09
```

The annotation of the output is depicted in the following.



The estimated Y, denoted by $\hat{Y}$ (commonly called fitted Y) and estimated residual $e = Y - \hat{Y}$ can be extracted from the above model in the following R code.

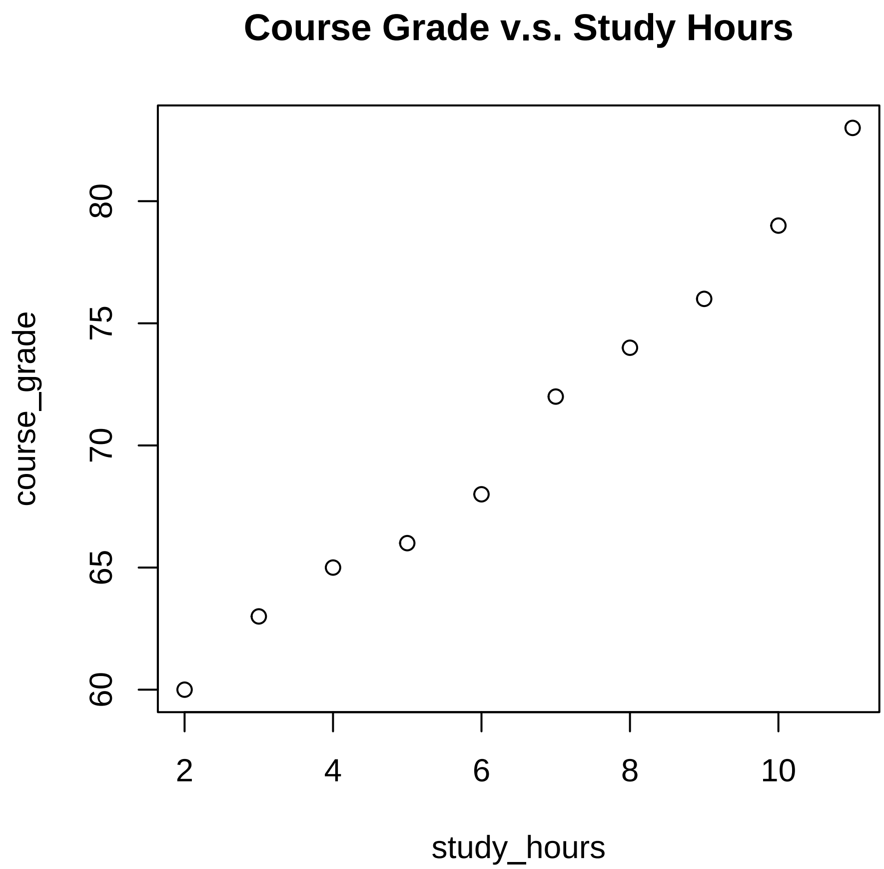
```
##     Y    Y.hat       e.hat
## 1  60 59.63636  0.3636364
## 2  63 62.07273  0.9272727
## 3  65 64.50909  0.4909091
## 4  66 66.94545 -0.9454545
## 5  68 69.38182 -1.3818182
## 6  72 71.81818  0.1818182
## 7  74 74.25455 -0.2545455
## 8  76 76.69091 -0.6909091
## 9  79 79.12727 -0.1272727
## 10 83 81.56364  1.4363636
```

You see from the above output that `Y - Y.hat = e.hat`.

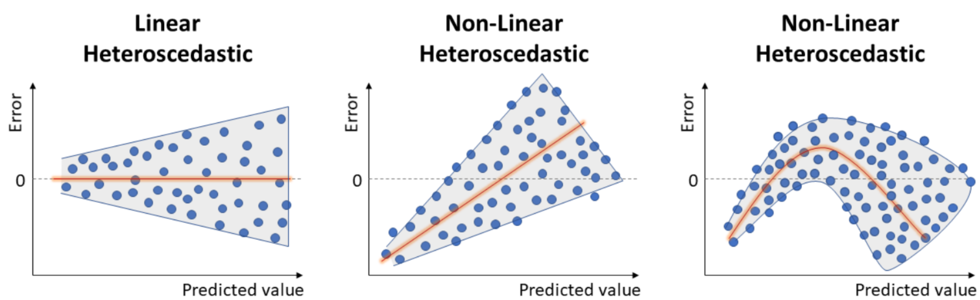
3 Assumptions of Simple Linear Regression

For valid inference and predictions, SLR relies on the following assumptions:

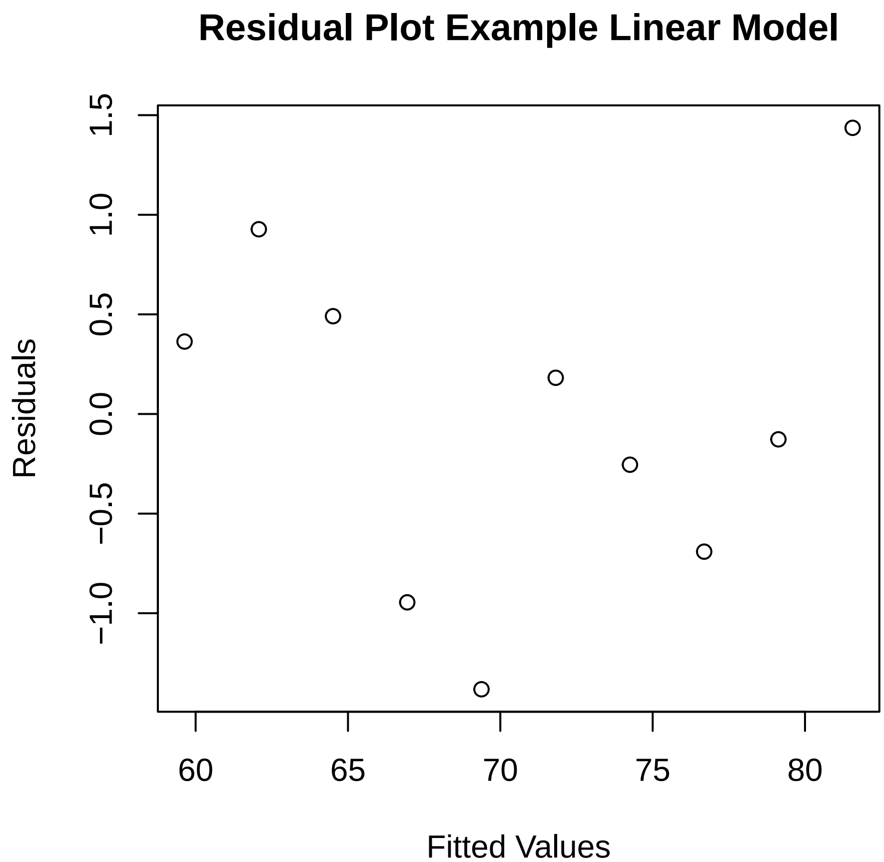
- **Linearity**: The relationship between X and Y is linear. This can be visually checked by a simple scatter plot for SLR (i.e., with only one predictor variable). We can check the linearity of the above example using the following R code.



- **Independence**: Errors (ϵ) are independent (no autocorrelation). This is hard to check in this level of statistics course.
- Homoscedasticity: Constant variance of errors across X. Since X is assumed to be non-random, the variance of Y is the same as the residual ϵ. In practice, we check whether the variance of the estimated residual errors $e = Y - \hat{Y}$.



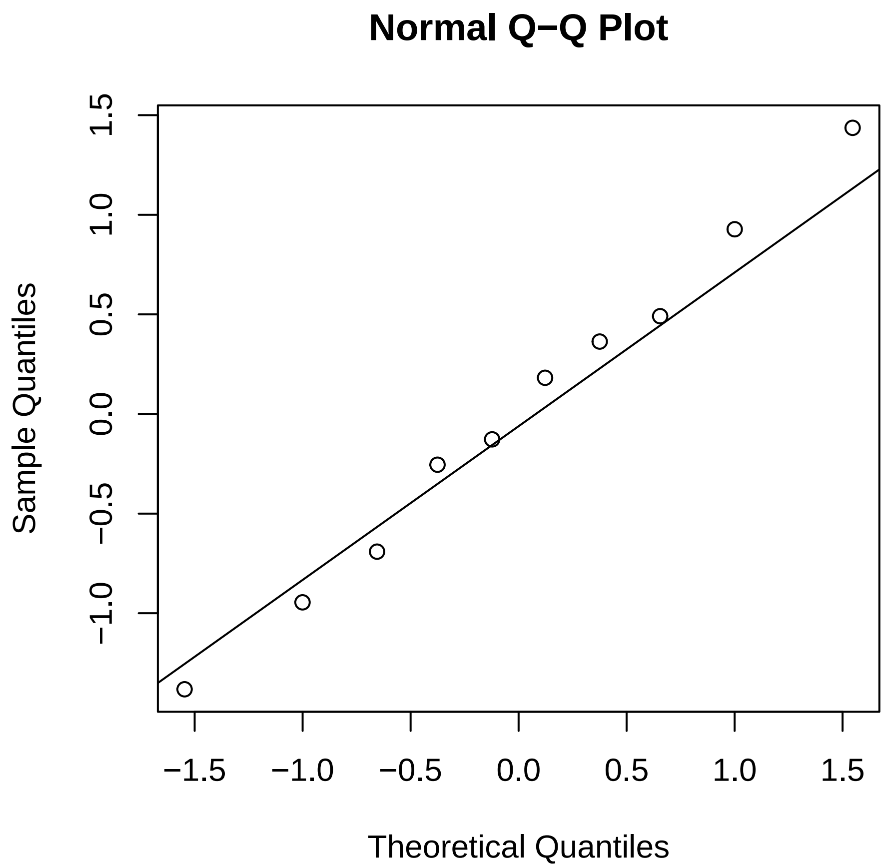
In the above grade-study example, we can plot the residual in the following.



It turns out that the residuals **does not** have a constant variance across the fitted values.

- **Normality**: Errors are normally distributed. There are testing hypothesis procedures for normality. In practice, we only use visual checks for normality. **QQ-plot** (quantile-quantile plot) is a commonly used visual tool. It compares the distribution of the estimated residuals with a set of true normally distributed values. The following **QQ-plot** was generated from R code.

The R functions `qqnorm()` and `qqline()` in package **{stats}** can draw the qq-plot and the reference line. The following code shows how to draw the QQ-plot and reference line based on the above example.

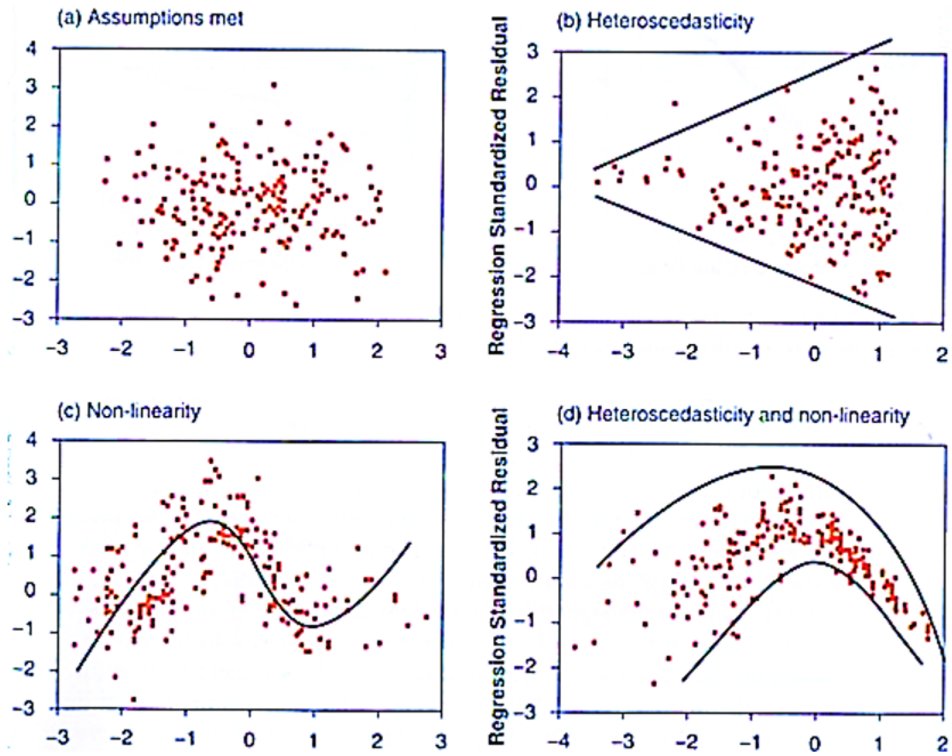


4 Model Diagnostics

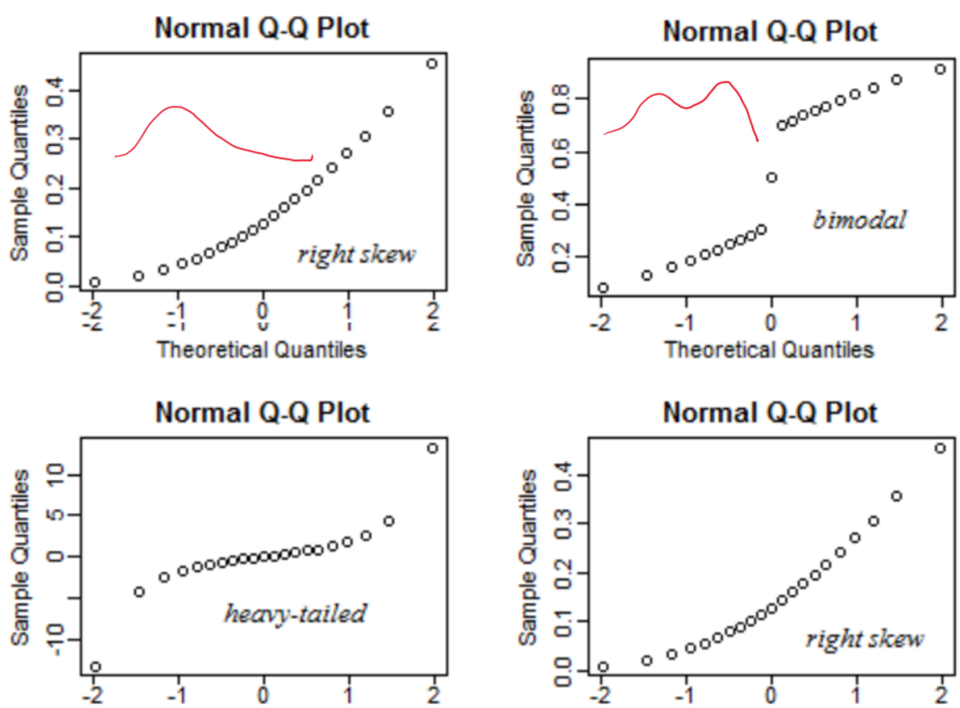
In practice, it is quite often to have different methods under different assumptions to solve the same problem. The goals of an analyst are to validate each individual model based on the related assumption and then choose the best model to implement (report). The process of validating a model is called **model diagnosis**.

As discussed in the previous section, most of the assumptions are related to the residual errors. This section will focus on **residual analysis** by analyzing the patterns in various residual plots.

- **Residual Plots** - Residuals vs Fitted: Checking for non-linearity & heteroscedasticity.



- **Q-Q Plot**: Check normality of residuals.



- **Influence Measures - Cook's Distance**: Detects influential observations that impact the regression model (line) significantly. To read the plot of Cook's distance, you need to have a clear understanding of three definitions:

1. **Outlier** - an observation that lies an **abnormal distance** from other values in the dataset.

2. Leverage - a measure of how far an **independent variable (X)** deviates from its mean.

3. Influential Point - an observation that significantly alters the regression coefficients if removed.

The following YouTube video explains these concepts clearly.

The following table explains the three concepts and diagnostic measures.

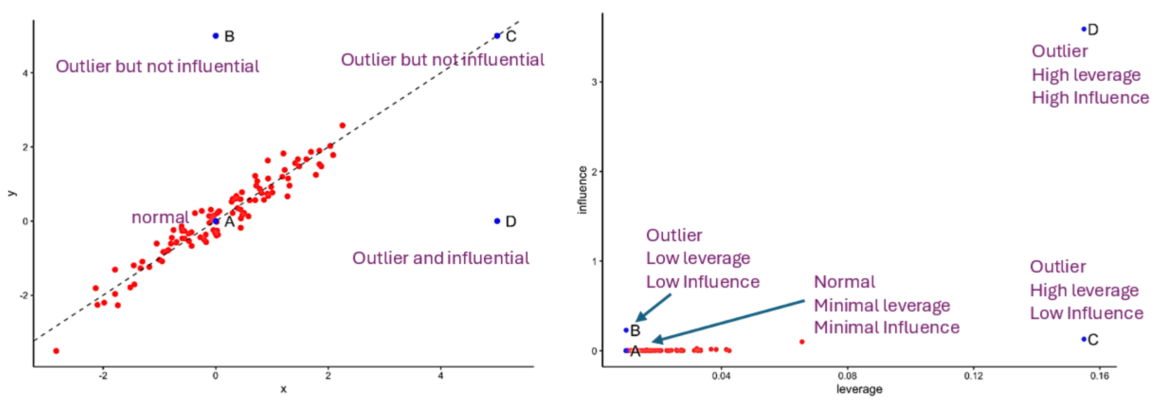
Concept	Detects...	Depends on...	Diagnostic Measure
Outlier	Atypical Y value	Residual magnitude	Standardized residuals
Leverage	Extreme X value	X position only	leverage measure
Influence	Impact on model	Both X and Y	Cook's Distance

- **Relationship Between Outliers, Leverage, and Influential Points**

The following table explains the relationship.

Scenario	Leverage	Residual	Influence
Normal point	Low	Small	No
Vertical outlier	Low	Large	Minimal
Good leverage point	High	Small	No
Bad leverage point	High	Large	Yes

- **Graphical Illustration of These Concepts**



5 Inference About Regression Coefficients

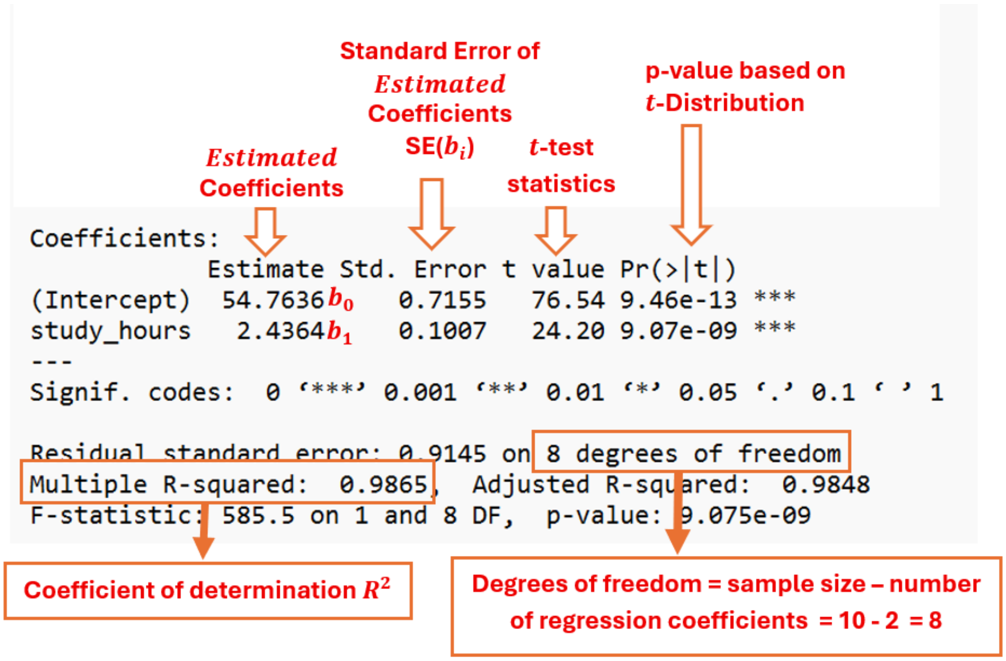
We perform hypothesis tests to determine if the relationship is statistically significant. Recall that the equation of the simple linear regression model is

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

X is a numeric or binary (has two possible distinct values, such as `STEM-major` and `non-STEM-major`, `yes` and `no`, `disease` and `disease-free`, etc.).

If the slope $\beta_1 = 0$, the above regression model is reduced to $Y = \beta_0 + \epsilon$, In this case, any changes in X will **not** influence Y. We call X and Y uncorrelated.

In practice, we only need to test whether $\beta_1 = 0$ to assess the linear relationship between Y and X. As shown earlier, the output in R provides hypothesis testing on both $beta_0 = 0$ and $\beta_1 = 0$ respectively. The following screenshot from R linear regression provides all the information for inference of regression coefficients.



NOTE: The degrees of freedom of the t-test in linear regression is defined to be: df = n - # coefficients!

5.1 t-test for Slope (β_1)

$$H_0: \quad \beta_1 = 0 \quad \text{v.s.} \quad \beta_1 \neq 0$$

In the **Course-grade and Study-time** example, the sample size = 10. The test Statistic for testing the above hypothesis is defined as

$$TS = \frac{b_1 - 0}{\text{SE}(b_1)} \to t_{10-2}$$

We use the information in the output to evaluate the above test statistic and obtain

$$TS = \frac{2.4364 - 0}{0.1007} = 24.02.$$

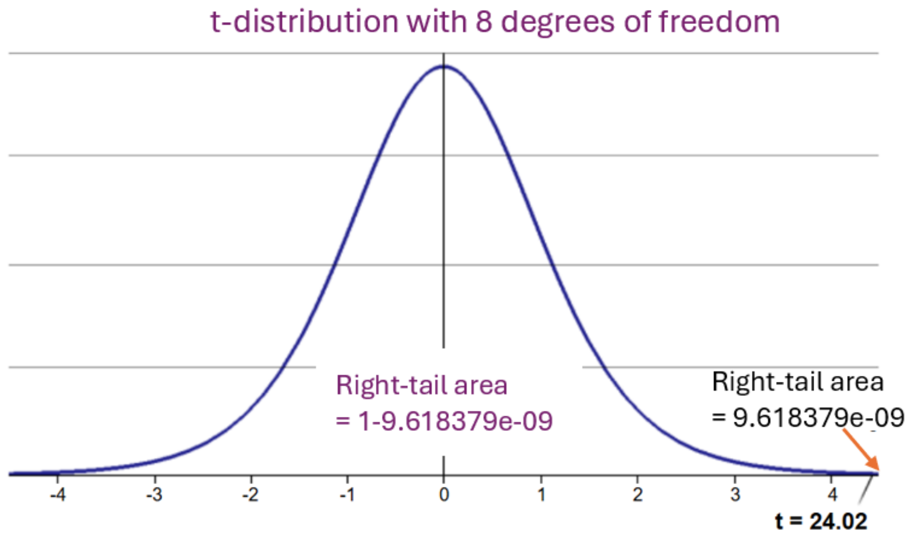
The p-value of the two-tailed test based on the distribution with 8 degrees of freedom is p-value $= P(TS > 24.02) = 9.618379e - 09 \approx 0$. We can use the following R code to calculate the p-value.

```
## [1] 9.618379e-09
```

Using significance level $\alpha = 0.05$, we fail to **reject** the null hypothesis $H_0 : \beta = 0$. This implies that study-hours influence the course grade **significantly** since p-value $\approx 0 <$ the significance level 0.05.

Remarks:

(1). `pt(24.02, 8)` gives the **left-tail area**. The **right-tail area** is `(1-pt(24.02, 8))`.



(2) We have learned from the previous introductory statistics that the p-value of a two-tailed test is equal to two times the smaller tail area.

5.2 Confidence Interval for Slope

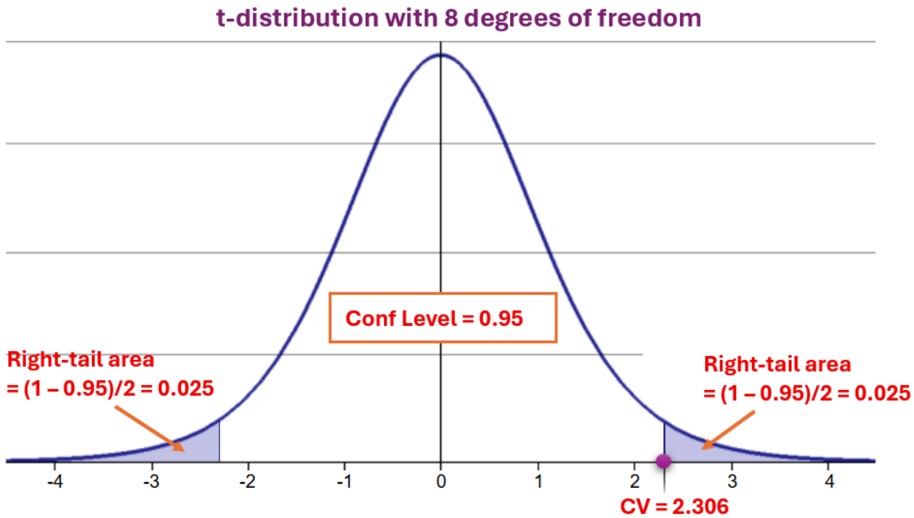
Using the information in the output, we can also construct the 95% confidence interval of the slope. Recall that the confidence of the population mean has the following general form.

$$\bar{x} \pm CV \times \frac{s}{\sqrt{n}}$$

The confidence interval of the slope (β_1) has a similar form given below.

$$b_1 \pm CV \times \text{SE}(b_1),$$

Where b_1 and SE(b_1) are given in the R output (see the above screenshot). CV (critical value) is based on the t-distribution with `n - # coefficients` (in this study-time and grade example, 10 -2 = 8). Let's use a confidence level of 95%, the critical value can be found using the R command `qt(0.975,8)`, which gives 2.306. The confidence level and critical value are labeled on the following t-density curve.



Based on the above explanation, the 95% confidence interval of β_1 is given by

$$b_1 \pm CV \times \text{SE}(b_1) = 2.4364 \pm 2.306 \times 0.1007 = (2.204186, 2.668614).$$

6 Summary

1. Structure

Models the relationship between one predictor (X) and one response (Y).

Equation: $Y = \beta_0 + \beta_1 X + \epsilon$. Two regresson coefficients are intercept (β_0) and slope (β_1). ϵ is the random error.

2. Assumptions

- *Linearity*: The Relationship between X and Y is linear.
- *Independence*: Residuals are uncorrelated (no autocorrelation).
- *Homoscedasticity*: Constant variance of residuals across X.
- *Normality*: Residuals are normally distributed (important for inference).

3. Diagnostics

Residual plots: Check for patterns (non-linearity, heteroscedasticity).

Q-Q plot: Assess normality of residuals.

Leverage/Cook's distance: Identify influential points.

R-squared: Proportion of variance explained by X.

4. Interpretation

- *Slope (β_1)*: Change in Y per unit increase in X.
- *Intercept (β_0*: Expected Y when X = 0 (may not always be meaningful).
- *R-squared:* between 0 and 1; higher = better fit (but **doesn't imply causation**).

5. Inference

- *Hypothesis test on β_1*: $H_0 : \beta_1 = 0$ v.s $H_a : \beta_1 \neq 0$. Use a t-test for significance.

- *Confidence interval of* β_1: Range for β_1. $b_1 \pm t_{\text{df},\ 1-\alpha/2}$
- *p-value*: Probability of observing the slope under $H_0 : \beta_1 = 0$.

6. Key R Functions

- Fit model: `lm(Y ~ X, data)`
- Summarize: `summary(model).` need to understand estimated regression coefficients, R-squared, p-values, etc.
- Diagnostics:
  - plot(model) (residual plots).
  - qqnorm(resid(model)) (normality check).
- Predictions: predict(model, newdata)