Hypothesis Test for Linear Regession

TrungTin Nguyen

STATIFY team, Inria centre at the University Grenoble Alpes

Statistical analysis and document mining
Complementary course, MSIAM

Outline

1. Simple linear regression
   - Estimation of the parameters by least squares
   - Motivation: advertising data
   - Assessing the accuracy of the coefficient estimates
2. Hypothesis tests on the coefficients
   - Review of hypothesis testing and p-values
   - The t-test versus Wald test
   - Applying for simple linear regression
   - Assessing the overall accuracy of the model

Simple linear regression

- We assume a model
$$Y = \beta_0 + \beta_1 X + \epsilon,$$
where ϵ is the error term, and two unknown constants (also known as coefficients or parameters)
  - β_0: intercept,
  - β_1: slope.
- The **hat** symbol denotes an estimated value. Given some estimates $\widehat{\beta}_0$ and $\widehat{\beta}_1$ for β_0 and β_1, respectively, we define a prediction of Y based on the basis of $X = x$ as follows
$$\widehat{y} = \widehat{\beta}_0 + \widehat{\beta}_1 x.$$

Simple linear regression

- We assume a model
$$Y = \beta_0 + \beta_1 X + \epsilon,$$
where ϵ is the error term, and two unknown constants (also known as coefficients or parameters)
    - β_0: intercept,
    - β_1: slope.
- The **hat** symbol denotes an estimated value. Given some estimates $\widehat{\beta}_0$ and $\widehat{\beta}_1$ for β_0 and β_1, respectively, we define a prediction of Y based on the basis of $X = x$ as follows
$$\widehat{y} = \widehat{\beta}_0 + \widehat{\beta}_1 x.$$
- Given n independent observations $\left(x_{[N]}, y_{[N]}\right) \equiv \{(x_n, y_n)\}_{n\in[N]}, [N] \equiv \{1, \ldots, N\}$, our goal is to obtain coefficient estimates $\widehat{\beta}_0$ and $\widehat{\beta}_1$ such that $y_n \approx \widehat{\beta}_0 + \widehat{\beta}_1 x_n, n \in [N]$.

Outline

Estimation of the parameters by least squares

- The least squares approach chooses $\widehat{\beta}_0$ and $\widehat{\beta}_1$ to minimize the residual sum of squares (RSS)

$$\mathrm{RSS} = \sum_{n=1}^{N} \left(y_n - \widehat{\beta}_0 - \widehat{\beta}_1 x_n \right)^2 .$$

Estimation of the parameters by least squares

- The least squares approach chooses $\widehat{\beta}_0$ and $\widehat{\beta}_1$ to minimize the residual sum of squares (RSS)

$$\mathrm{RSS} = \sum_{n=1}^{N} \left(y_n - \widehat{\beta}_0 - \widehat{\beta}_1 x_n \right)^2 .$$

- By using calculus, $\left(\frac{\partial \, \mathrm{RSS}}{\partial \widehat{\beta}_1}, \frac{\partial \, \mathrm{RSS}}{\partial \widehat{\beta}_0} \right) = (0, 0)$, the minimizing values can be shown to be (see for example chapter 3 from [Hastie et al., 2009, James et al., 2021])

$$\widehat{\beta}_1 = \frac{\sum_{n=1}^{N} (x_n - \bar{x})(y_n - \bar{y})}{\sum_{n=1}^{N} (x_n - \bar{x})^2}, \quad \widehat{\beta}_0 = \bar{y} - \widehat{\beta}_1 \bar{x},$$

where $\bar{x} = \frac{1}{N} \sum_{n=1}^{N} x_n$ and $\bar{y} = \frac{1}{N} \sum_{n=1}^{N} y_n$ are the sample means.

Estimation of the parameters by least squares

- The least squares approach chooses $\widehat{\beta}_0$ and $\widehat{\beta}_1$ to minimize the residual sum of squares (RSS)

$$\mathrm{RSS} = \sum_{n=1}^{N} \left(y_n - \widehat{\beta}_0 - \widehat{\beta}_1 x_n \right)^2 .$$

- By using calculus, $\left(\frac{\partial\,\mathrm{RSS}}{\partial \widehat{\beta}_1}, \frac{\partial\,\mathrm{RSS}}{\partial \widehat{\beta}_0} \right) = (0, 0)$, the minimizing values can be shown to be (see for example chapter 3 from [Hastie et al., 2009, James et al., 2021])

$$\widehat{\beta}_1 = \frac{\sum_{n=1}^{N} (x_n - \bar{x})(y_n - \bar{y})}{\sum_{n=1}^{N} (x_n - \bar{x})^2}, \quad \widehat{\beta}_0 = \bar{y} - \widehat{\beta}_1 \bar{x},$$

  where $\bar{x} = \frac{1}{N} \sum_{n=1}^{N} x_n$ and $\bar{y} = \frac{1}{N} \sum_{n=1}^{N} y_n$ are the sample means.
- This is a minimum (and not a maximum or saddle point): RSS is a quadratic function and has positive coefficients of the squared term of $\widehat{\beta}_0$ and $\widehat{\beta}_1$.

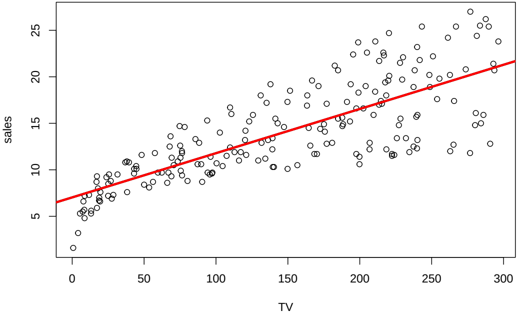
Outline

Motivation: advertising data

- **Description:** set consists of the sales of that product in 200 different markets, along with advertising budgets for the product in each of those markets for three different media: TV, radio, and newspaper [James et al., 2021, Chapter 2].

Motivation: advertising data

- **Description:** set consists of the sales of that product in 200 different markets, along with advertising budgets for the product in each of those markets for three different media: TV, radio, and newspaper [James et al., 2021, Chapter 2].
- **Goal:** develop an accurate model that can be used to predict sales on the basis of the three media budgets $\leftarrow$ Linear regression in $\boldsymbol{R}$.



Goal: understand how linear regression works in R

```
Call:
lm(formula = sales ~ TV)

Residuals:
    Min      1Q  Median      3Q     Max 
-8.3860 -1.9545 -0.1913  2.0671  7.2124 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 7.032594   0.457843   15.36   <2e-16 ***
TV          0.047537   0.002691   17.67   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.259 on 198 degrees of freedom
Multiple R-squared:  0.6119,    Adjusted R-squared:  0.6099 
F-statistic: 312.1 on 1 and 198 DF,  p-value: < 2.2e-16
```

Outline

1. Simple linear regression
   - Estimation of the parameters by least squares
   - Motivation: advertising data
   - Assessing the accuracy of the coefficient estimates

2. Hypothesis tests on the coefficients
   - Review of hypothesis testing and p-values
   - The t-test versus Wald test
   - Applying for simple linear regression
   - Assessing the overall accuracy of the model

Assessing the accuracy of the coefficient estimates

1. Note that the estimated parameters $\widehat{\beta}_0$ and $\widehat{\beta}_1$ are unbiased (TD1), we wonder how close $\widehat{\beta}_0$ and $\widehat{\beta}_1$ are to the true values β_0 and β_1 $\rightarrow$ computing the standard error, $\mathrm{SE}(\widehat{\beta}_i) = \mathrm{var}\left(\widehat{\beta}_i\right)^{1/2}, i = 0, 1$.

Assessing the accuracy of the coefficient estimates

1. Note that the estimated parameters $\widehat{\beta}_0$ and $\widehat{\beta}_1$ are unbiased (TD1), we wonder how close $\widehat{\beta}_0$ and $\widehat{\beta}_1$ are to the true values β_0 and $\beta_1 \to$ computing the standard error, $\mathrm{SE}(\widehat{\beta}_i) = \mathrm{var}\left(\widehat{\beta}_i\right)^{1/2}, i = 0, 1$.
2. When $\epsilon_n \overset{i.i.d.}{\sim} \mathcal{N}(0, \sigma^2), n \in [N]$, we can show that (TD1)

$$\mathrm{var}\left(\widehat{\beta}_0\right) = \frac{\sigma^2}{N}\left(1 + \frac{\bar{x}^2}{s_X^2}\right), \quad \mathrm{var}\left(\widehat{\beta}_1\right) = \frac{\sigma^2}{N}\frac{1}{s_X^2}.$$

where $\sigma^2 = \mathrm{var}\left(\epsilon\right)$, $s_X^2 = \frac{1}{N}\sum_{n=1}^{N}(x_n - \bar{x})^2$.

Assessing the accuracy of the coefficient estimates

1. Note that the estimated parameters $\widehat{\beta}_0$ and $\widehat{\beta}_1$ are unbiased (TD1), we wonder how close $\widehat{\beta}_0$ and $\widehat{\beta}_1$ are to the true values β_0 and β_1 $\rightarrow$ computing the standard error, $\mathrm{SE}(\widehat{\beta}_i) = \mathrm{var}\left(\widehat{\beta}_i\right)^{1/2}, i = 0, 1$.
2. When $\epsilon_n \overset{i.i.d.}{\sim} \mathcal{N}(0, \sigma^2), n \in [N]$, we can show that (TD1)
$$\mathrm{var}\left(\widehat{\beta}_0\right) = \frac{\sigma^2}{N}\left(1 + \frac{\bar{x}^2}{s_X^2}\right), \quad \mathrm{var}\left(\widehat{\beta}_1\right) = \frac{\sigma^2}{N}\frac{1}{s_X^2}.$$
where $\sigma^2 = \mathrm{var}\left(\epsilon\right)$, $s_X^2 = \frac{1}{N}\sum_{n=1}^{N}(x_n - \bar{x})^2$.
3. These standard errors can be used to compute confidence intervals. A 95% confidence interval is defined as a range of values such that with 95% probability, the range will contain the true unknown value of the parameter. Using normal-based confidence interval [Wasserman, 2004, Theorem 6.16], for $i = 0, 1$, it has the form
$$[\widehat{\beta}_i - 2 \times \mathrm{SE}(\widehat{\beta}_i), \widehat{\beta}_i + 2 \times \mathrm{SE}(\widehat{\beta}_i)] \text{ since } \widehat{\beta}_i \sim \mathcal{N}\left(\beta_i, \mathrm{SE}(\widehat{\beta}_i)\right).$$

Recall the normal-based confidence interval

Recall the normal-based confidence interval [Wasserman, 2004]:

6.16 Theorem (Normal-based Confidence Interval). *Suppose that* $\widehat{\theta}_n \approx N(\theta, \widehat{\mathsf{se}}^2)$. *Let* Φ *be the* CDF *of a standard Normal and let* $z_{\alpha/2} = \Phi^{-1}(1-(\alpha/2))$, *that is,* $\mathbb{P}(Z > z_{\alpha/2}) = \alpha/2$ *and* $\mathbb{P}(-z_{\alpha/2} < Z < z_{\alpha/2}) = 1-\alpha$ *where* $Z \sim N(0,1)$. *Let*

$$C_n = (\widehat{\theta}_n - z_{\alpha/2}\,\widehat{\mathsf{se}},\ \widehat{\theta}_n + z_{\alpha/2}\,\widehat{\mathsf{se}}). \tag{6.10}$$

Then

$$\mathbb{P}_\theta(\theta \in C_n) \to 1-\alpha. \tag{6.11}$$

PROOF. Let $Z_n = (\widehat{\theta}_n - \theta)/\widehat{\mathsf{se}}$. By assumption $Z_n \rightsquigarrow Z$ where $Z \sim N(0,1)$. Hence,

$$\begin{aligned}
\mathbb{P}_\theta(\theta \in C_n) &= \mathbb{P}_\theta\left(\widehat{\theta}_n - z_{\alpha/2}\,\widehat{\mathsf{se}} < \theta < \widehat{\theta}_n + z_{\alpha/2}\,\widehat{\mathsf{se}}\right) \\
&= \mathbb{P}_\theta\left(-z_{\alpha/2} < \frac{\widehat{\theta}_n - \theta}{\widehat{\mathsf{se}}} < z_{\alpha/2}\right) \\
&\to \mathbb{P}\left(-z_{\alpha/2} < Z < z_{\alpha/2}\right) \\
&= 1-\alpha. \quad \blacksquare
\end{aligned}$$

For 95 percent confidence intervals, $\alpha = 0.05$ and $z_{\alpha/2} = 1.96 \approx 2$ leading to the approximate 95 percent confidence interval $\widehat{\theta}_n \pm 2\,\widehat{\mathsf{se}}$.

Outline

1. Simple linear regression
   - Estimation of the parameters by least squares
   - Motivation: advertising data
   - Assessing the accuracy of the coefficient estimates

2. Hypothesis tests on the coefficients
   - Review of hypothesis testing and p-values
   - The t-test versus Wald test
   - Applying for simple linear regression
   - Assessing the overall accuracy of the model

Hypothesis tests on the coefficients

Standard errors can also be used to perform hypothesis tests on the coefficients. The most common hypothesis test involves testing the null hypothesis of

- $\mathcal{H}_0$: There is no relationship between X and Y versus the alternative hypothesis
- $\mathcal{H}_1$: There is some relationship between X and Y.

Outline

Review of hypothesis testing and p-values

- We partition the parameter space Θ into two disjoint sets Θ_0 and Θ_1.

Review of hypothesis testing and p-values

- We partition the parameter space Θ into two disjoint sets Θ_0 and Θ_1.
- We wish to test null hypothesis $\mathcal{H}_0 : \theta \in \Theta_0$ versus alternative hypothesis $\mathcal{H}_1 : \theta \notin \Theta_1$.

Review of hypothesis testing and p-values

- We partition the parameter space Θ into two disjoint sets Θ_0 and Θ_1.
- We wish to test null hypothesis $\mathcal{H}_0 : \theta \in \Theta_0$ versus alternative hypothesis $\mathcal{H}_1 : \theta \notin \Theta_1$.
- Let X be a random variable and let $\mathcal{X}$ be the range of X. Given rejection region R, then
  - $X \in R \implies$ reject $\mathcal{H}_0$,
  - $X \notin R \implies$ retain (do not reject) $\mathcal{H}_0$.

Review of hypothesis testing and p-values

- We partition the parameter space Θ into two disjoint sets Θ_0 and Θ_1.
- We wish to test null hypothesis $\mathcal{H}_0 : \theta \in \Theta_0$ versus alternative hypothesis $\mathcal{H}_1 : \theta \notin \Theta_1$.
- Let X be a random variable and let $\mathcal{X}$ be the range of X. Given rejection region R, then
  - $X \in R \Longrightarrow$ reject $\mathcal{H}_0$,
  - $X \notin R \Longrightarrow$ retain (do not reject) $\mathcal{H}_0$.
- Usually, the rejection region R is of the form $R = \{x : T(x) \geq c\}$, where T is a test statistic and c is a critical value.
  $\Longrightarrow$ Hypothesis testing $\longleftrightarrow$ find appropriate T and c.

The size α Wald test

- The power function of a test with rejection region R is defined by $\beta(\theta) = \mathbb{P}_{\theta}(X \in R)$.

The size α Wald test

- The power function of a test with rejection region R is defined by $\beta(\theta) = \mathbb{P}_\theta(X \in R)$.
- The size of a test is defined to be $\alpha = \sup_{\theta \in \Theta_0} \beta(\theta)$.

The size α Wald test

- The power function of a test with rejection region R is defined by $\beta(\theta) = \mathbb{P}_\theta(X \in R)$.
- The size of a test is defined to be $\alpha = \sup_{\theta \in \Theta_0} \beta(\theta)$.
- A test is said to have level α if its size is less than or equal to α.

The size α Wald test

- The power function of a test with rejection region R is defined by $\beta(\theta) = \mathbb{P}_\theta(X \in R)$.
- The size of a test is defined to be $\alpha = \sup_{\theta \in \Theta_0} \beta(\theta)$.
- A test is said to have level α if its size is less than or equal to α.
- The Wald test: $\mathcal{H}_0 : \theta = \theta_0$ versus $\mathcal{H}_1 : \theta \neq \theta_0$.
  - Assume that $\hat{\theta}$ is asymtotically Normal: $\frac{\hat{\theta}-\theta_0}{\mathsf{SE}(\hat{\theta})} \rightsquigarrow \mathcal{N}(0,1)$, where $\hat{\theta}$ and $\mathsf{SE}(\hat{\theta})$ are estimate of θ and estimated standard error of $\hat{\theta}$, respectively.

The size α Wald test

- The power function of a test with rejection region R is defined by $\beta(\theta) = \mathbb{P}_\theta(X \in R)$.
- The size of a test is defined to be $\alpha = \sup_{\theta \in \Theta_0} \beta(\theta)$.
- A test is said to have level α if its size is less than or equal to α.
- The Wald test: $\mathcal{H}_0 : \theta = \theta_0$ versus $\mathcal{H}_1 : \theta \neq \theta_0$.
  - Assume that $\hat{\theta}$ is asymtotically Normal: $\frac{\hat{\theta}-\theta_0}{\mathsf{SE}(\hat{\theta})} \rightsquigarrow \mathcal{N}(0,1)$, where $\hat{\theta}$ and $\mathsf{SE}(\hat{\theta})$ are estimate of θ and estimated standard error of $\hat{\theta}$, respectively.
  - The size α Wald test is: reject $\mathcal{H}_0$ when $|W| > z_{\alpha/2}$ where $W = \frac{\hat{\theta}-\theta_0}{\mathsf{SE}(\hat{\theta})}$ and $z_{\alpha/2}$ satisfies $\mathbb{P}(Z \geq z_{\alpha/2}) = \alpha/2$, where $Z \sim \mathcal{N}(0,1)$.

The size α Wald test

- The power function of a test with rejection region R is defined by $\beta(\theta) = \mathbb{P}_\theta(X \in R)$.
- The size of a test is defined to be $\alpha = \sup_{\theta \in \Theta_0} \beta(\theta)$.
- A test is said to have level α if its size is less than or equal to α.
- The Wald test: $\mathcal{H}_0 : \theta = \theta_0$ versus $\mathcal{H}_1 : \theta \neq \theta_0$.
    - Assume that $\hat{\theta}$ is asymtotically Normal: $\frac{\hat{\theta} - \theta_0}{\mathrm{SE}(\hat{\theta})} \rightsquigarrow \mathcal{N}(0, 1)$, where $\hat{\theta}$ and $\mathrm{SE}(\hat{\theta})$ are estimate of θ and estimated standard error of $\hat{\theta}$, respectively.
    - The size α Wald test is: reject $\mathcal{H}_0$ when $|W| > z_{\alpha/2}$ where $W = \frac{\hat{\theta} - \theta_0}{\mathrm{SE}(\hat{\theta})}$ and $z_{\alpha/2}$ satisfies $\mathbb{P}(Z \geq z_{\alpha/2}) = \alpha/2$, where $Z \sim \mathcal{N}(0, 1)$.
    - We can show that, asymptotically, the Wald test has size α.

The size α Wald test

- The power function of a test with rejection region R is defined by $\beta(\theta) = \mathbb{P}_\theta(X \in R)$.
- The size of a test is defined to be $\alpha = \sup_{\theta \in \Theta_0} \beta(\theta)$.
- A test is said to have level α if its size is less than or equal to α.
- The Wald test: $\mathcal{H}_0 : \theta = \theta_0$ versus $\mathcal{H}_1 : \theta \neq \theta_0$.
  - Assume that $\hat{\theta}$ is asymtotically Normal: $\frac{\hat{\theta} - \theta_0}{\mathsf{SE}(\hat{\theta})} \rightsquigarrow \mathcal{N}(0, 1)$, where $\hat{\theta}$ and $\mathsf{SE}(\hat{\theta})$ are estimate of θ and estimated standard error of $\hat{\theta}$, respectively.
  - The size α Wald test is: reject $\mathcal{H}_0$ when $|W| > z_{\alpha/2}$ where $W = \frac{\hat{\theta} - \theta_0}{\mathsf{SE}(\hat{\theta})}$ and $z_{\alpha/2}$ satisfies $\mathbb{P}(Z \geq z_{\alpha/2}) = \alpha/2$, where $Z \sim \mathcal{N}(0, 1)$.
  - We can show that, asymptotically, the Wald test has size α. Indeed, by using asymptotically Normal,

$$\mathbb{P}_{\theta_0}\left(|W| > z_{\alpha/2}\right) = \mathbb{P}_{\theta_0}\left(\frac{|\hat{\theta} - \theta_0|}{\mathsf{SE}(\hat{\theta})} > z_{\alpha/2}\right) \to \mathbb{P}(|Z| \geq z_{\alpha/2}) = \alpha.$$

Theorem (Scientific significance versus statistical significance)

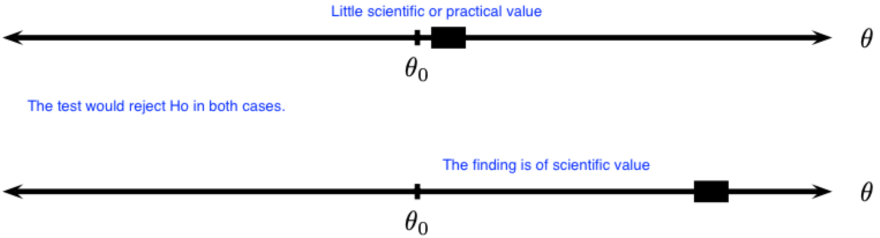
The size α Wald test rejects $\mathcal{H}_0 : \theta = \theta_0$ (say statistically significant) versus $\mathcal{H}_1 : \theta \neq \theta_0$ if and only if $\theta_0 \notin C$ where $C = \left(\hat{\theta} - \mathsf{SE}(\hat{\theta}) z_{\alpha/2}, \hat{\theta} + \mathsf{SE}(\hat{\theta}) z_{\alpha/2}\right)$ is $1 - \alpha$ asymptotic confidence interval.

Therefore, testing the hypothesis $\Longleftrightarrow$ checking whether the null value is in the confidence interval.

Theorem (Scientific significance versus statistical significance)

The size α Wald test rejects $\mathcal{H}_0 : \theta = \theta_0$ (say statistically significant) versus $\mathcal{H}_1 : \theta \neq \theta_0$ if and only if $\theta_0 \notin C$ where $C = \left(\hat{\theta} - \mathrm{SE}(\hat{\theta}) z_{\alpha/2}, \hat{\theta} + \mathrm{SE}(\hat{\theta}) z_{\alpha/2}\right)$ is $1 - \alpha$ asymptotic confidence interval.
Therefore, testing the hypothesis $\Longleftrightarrow$ checking whether the null value is in the confidence interval.

Statistical significance $\nrightarrow$ scientific importance.
Confidence intervals are often more informative than tests.



Definition (p-values)

Suppose that for every $\alpha \in (0,1)$, we have a size α test with rejection region $\mathbb{R}_\alpha$. Then. p-value $= \inf \{\alpha : T(x) \in R_\alpha\}$. That is, the p-value is the smallest level at which we can reject $\mathcal{H}_0$.

Definition (p-values)

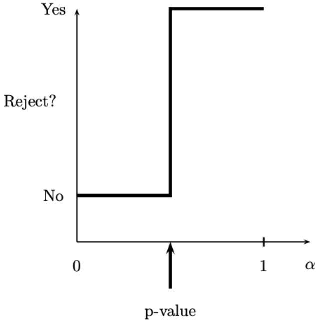
Suppose that for every $\alpha \in (0, 1)$, we have a size α test with rejection region $\mathbb{R}_\alpha$. Then. p-value $= \inf \{\alpha : T(x) \in R_\alpha\}$. That is, the p-value is the smallest level at which we can reject $\mathcal{H}_0$.

Informally, the smaller the p-value, the stronger the evidence against $\mathcal{H}_0$.

Definition (p-values)

Suppose that for every $\alpha \in (0,1)$, we have a size α test with rejection region $\mathbb{R}_\alpha$. Then. p-value $= \inf\left\{\alpha : T(x) \in R_\alpha\right\}$. That is, the p-value is the smallest level at which we can reject $\mathcal{H}_0$.

Informally, the smaller the p-value, the stronger the evidence against $\mathcal{H}_0$.
BUT, large p-value is not strong evidence in favor of $\mathcal{H}_0$: (i) $\mathcal{H}_0$ is true or (ii) $\mathcal{H}_0$ is false but the test has low power.
DO NOT CONFUSE: p-value $\neq \mathbb{P}(\mathcal{H}_0|\text{Data})$.



Theorem (Compute the p-values)

Suppose that the size α test is of the form reject $\mathcal{H}_0$ if and only if $T(X_{[N]}) \geq c_\alpha$. Then, given the observed value $x_{[N]}$ of random sample $X_{[N]}$,

$$p\text{-}value = \sup_{\theta \in \Theta_0} \mathbb{P}_{\theta_0}\left(T(X_{[N]}) \geq T(x_{[N]})\right).$$

Theorem (Compute the p-values)

Suppose that the size α test is of the form reject $\mathcal{H}_0$ if and only if $T(X_{[N]}) \geq c_\alpha$. Then, given the observed value $x_{[N]}$ of random sample $X_{[N]}$,

$$p\text{-}value = \sup_{\theta \in \Theta_0} \mathbb{P}_{\theta_0} \left(T(X_{[N]}) \geq T(x_{[N]}) \right).$$

Let $w = \hat{\theta} - \theta_0 / \operatorname{SE}(\hat{\theta})$ denote the observed value of the Wald statistic W,

$$p\text{-}value = \mathbb{P}_{\theta_0}(|W| \geq |w|) \approx \mathbb{P}(|Z| \geq |w|) = 2\Phi(-|w|), Z \sim \mathcal{N}(0,1).$$

Theorem (Compute the p-values)

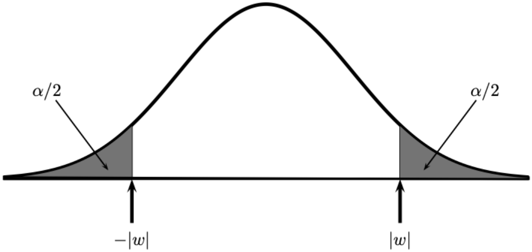
Suppose that the size α test is of the form reject $\mathcal{H}_0$ if and only if $T(X_{[N]}) \geq c_\alpha$. Then, given the observed value $x_{[N]}$ of random sample $X_{[N]}$,

$$\textit{p-value} = \sup_{\theta \in \Theta_0} \mathbb{P}_{\theta_0}\left(T(X_{[N]}) \geq T(x_{[N]})\right).$$

Let $w = \hat{\theta} - \theta_0 / \operatorname{SE}(\hat{\theta})$ denote the observed value of the Wald statistic W,

$$\textit{p-value} = \mathbb{P}_{\theta_0}(|W| \geq |w|) \approx \mathbb{P}(|Z| \geq |w|) = 2\Phi(-|w|), Z \sim \mathcal{N}(0,1).$$

Informally, p-value = the probability (under H0) of observing a value of the test statistic the same as or more extreme than what was actually observed.



Outline

1. Simple linear regression
   - Estimation of the parameters by least squares
   - Motivation: advertising data
   - Assessing the accuracy of the coefficient estimates
2. Hypothesis tests on the coefficients
   - Review of hypothesis testing and p-values
   - The t-test versus Wald test
   - Applying for simple linear regression
   - Assessing the overall accuracy of the model

Definition (The t-test versus Wald test)

Let $X_n \sim \mathcal{N}(\mu, \sigma^2), n \in [N]$ where μ, σ^2 are both unknown. Suppose we want to test $\mathcal{H}_0 : \mu = \mu_0$ versus $\mathcal{H}_1 : \mu \neq \mu_0$. We choose

$$T = \frac{\sqrt{N}(\bar{X}_N - \mu_0)}{S_n} \sim t_{N-1} \text{ under } \mathcal{H}_0,$$

where $S_n^2 = \frac{1}{N-1} \sum_{n=1}^{N} (X_n - \bar{X}_N)^2$ is the sample variance and t_{N-1} is Student's t-distributiont with $N - 1$ degrees of freedom.

Definition (The t-test versus Wald test)

Let $X_n \sim \mathcal{N}(\mu, \sigma^2), n \in [N]$ where μ, σ^2 are both unknown. Suppose we want to test $\mathcal{H}_0 : \mu = \mu_0$ versus $\mathcal{H}_1 : \mu \neq \mu_0$. We choose

$$T = \frac{\sqrt{N}(\bar{X}_N - \mu_0)}{S_n} \sim t_{N-1} \text{ under } \mathcal{H}_0,$$

where $S_n^2 = \frac{1}{N-1}\sum_{n=1}^{N}(X_n - \bar{X}_N)^2$ is the sample variance and t_{N-1} is Student's t-distributiont with $N-1$ degrees of freedom.
We reject $\mathcal{H}_0$ if $|T| > t_{N-1,\alpha/2}$ then we get a size α test.

Definition (The t-test versus Wald test)

Let $X_n \sim \mathcal{N}(\mu, \sigma^2), n \in [N]$ where μ, σ^2 are both unknown. Suppose we want to test $\mathcal{H}_0 : \mu = \mu_0$ versus $\mathcal{H}_1 : \mu \neq \mu_0$. We choose

$$T = \frac{\sqrt{N}(\bar{X}_N - \mu_0)}{S_n} \sim t_{N-1} \text{ under } \mathcal{H}_0,$$

where $S_n^2 = \frac{1}{N-1} \sum_{n=1}^{N} (X_n - \bar{X}_N)^2$ is the sample variance and t_{N-1} is Student's t-distributiont with $N-1$ degrees of freedom.
We reject $\mathcal{H}_0$ if $|T| > t_{N-1,\alpha/2}$ then we get a size α test.

When N is moderately large, $T \approx \mathcal{N}(0,1)$ under $\mathcal{H}_0$: the t-test is essentially identical to the Wald test.

Outline

1. Simple linear regression
   - Estimation of the parameters by least squares
   - Motivation: advertising data
   - Assessing the accuracy of the coefficient estimates
2. Hypothesis tests on the coefficients
   - Review of hypothesis testing and p-values
   - The t-test versus Wald test
   - Applying for simple linear regression
   - Assessing the overall accuracy of the model

Hypothesis tests on the coefficients

- Standard errors can also be used to perform hypothesis tests on the coefficients. The most common hypothesis test involves testing the null hypothesis of
  - $\mathcal{H}_0$: There is no relationship between X and Y versus the alternative hypothesis
  - $\mathcal{H}_1$: There is some relationship between X and Y.

Hypothesis tests on the coefficients

- Standard errors can also be used to perform hypothesis tests on the coefficients. The most common hypothesis test involves testing the null hypothesis of
  - $\mathcal{H}_0$: There is no relationship between X and Y versus the alternative hypothesis
  - $\mathcal{H}_1$: There is some relationship between X and Y.
- Mathematically, this corresponds to testing

$$\mathcal{H}_0 : \beta_1 = 0 \text{ versus } \mathcal{H}_1 : \beta_1 \neq 0.$$

Hypothesis tests on the coefficients

- Standard errors can also be used to perform hypothesis tests on the coefficients. The most common hypothesis test involves testing the null hypothesis of
  - $\mathcal{H}_0$: There is no relationship between X and Y versus the alternative hypothesis
  - $\mathcal{H}_1$: There is some relationship between X and Y.
- Mathematically, this corresponds to testing

$$\mathcal{H}_0 : \beta_1 = 0 \text{ versus } \mathcal{H}_1 : \beta_1 \neq 0.$$

- To test the null hypothesis, we compute a t-statistics, given by

$$t = \frac{\widehat{\beta}_1 - 0}{\mathrm{SE}(\widehat{\beta}_1)} \sim t_{N-2} \text{ assuming } \beta_1 = 0.$$

- Using statistical software, it is easy to compute the probability of observing any value equal to $|t|$ or larger, p-value $= \mathbb{P}(|T| \geq |t|)$.

Outline

1. Simple linear regression
   - Estimation of the parameters by least squares
   - Motivation: advertising data
   - Assessing the accuracy of the coefficient estimates
2. Hypothesis tests on the coefficients
   - Review of hypothesis testing and p-values
   - The t-test versus Wald test
   - Applying for simple linear regression
   - Assessing the overall accuracy of the model

Assessing the overall accuracy of the model

- Given the Residual sum-of-squares $\mathrm{RSS} = \sum_{n=1}^{N}(y_n - \hat{y}_n)^2$, we compute the Residual Standard Error

$$\mathrm{RSE} = \sqrt{\frac{1}{N-2}\mathrm{RSS}} = \sqrt{\frac{1}{N-2}\sum_{n=1}^{N}(y_n - \hat{y}_n)^2}.$$

Assessing the overall accuracy of the model

- Given the Residual sum-of-squares $\text{RSS} = \sum_{n=1}^{N}(y_n - \hat{y}_n)^2$, we compute the Residual Standard Error

$$\text{RSE} = \sqrt{\frac{1}{N-2}\text{RSS}} = \sqrt{\frac{1}{N-2}\sum_{n=1}^{N}(y_n - \hat{y}_n)^2}.$$

- R-squared or fraction of variance explained is

$$R^2 = \frac{\text{TSS} - \text{RSS}}{\text{TSS}} = 1 - \frac{\text{RSS}}{\text{TSS}}.$$

where the total sum of square is $\text{TSS} = \sum_{n=1}^{N}(y_n - \bar{y})^2$.

Goal: understand how linear regression works in R

```
Call:
lm(formula = sales ~ TV)

Residuals:
    Min      1Q  Median      3Q     Max
-8.3860 -1.9545 -0.1913  2.0671  7.2124

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 7.032594   0.457843   15.36   <2e-16 ***
TV          0.047537   0.002691   17.67   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.259 on 198 degrees of freedom
Multiple R-squared:  0.6119,    Adjusted R-squared:  0.6099
F-statistic: 312.1 on 1 and 198 DF,  p-value: < 2.2e-16
```

References I

Hastie, T., Tibshirani, R., Friedman, J. H., & Friedman, J. H. (2009).
The elements of statistical learning: data mining, inference, and prediction, volume 2.
Springer.

(Cited on pages 6, 7, and 8.)

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2021).
An Introduction to Statistical Learning: with Applications in R.
Springer Texts in Statistics. Springer US.

(Cited on pages 6, 7, 8, 10, and 11.)

Wasserman, L. (2004).
All of statistics: a concise course in statistical inference, volume 26.
Springer.

(Cited on pages 14, 15, 16, 17, and 18.)