STA3010 Regression Analysis

Feng YIN

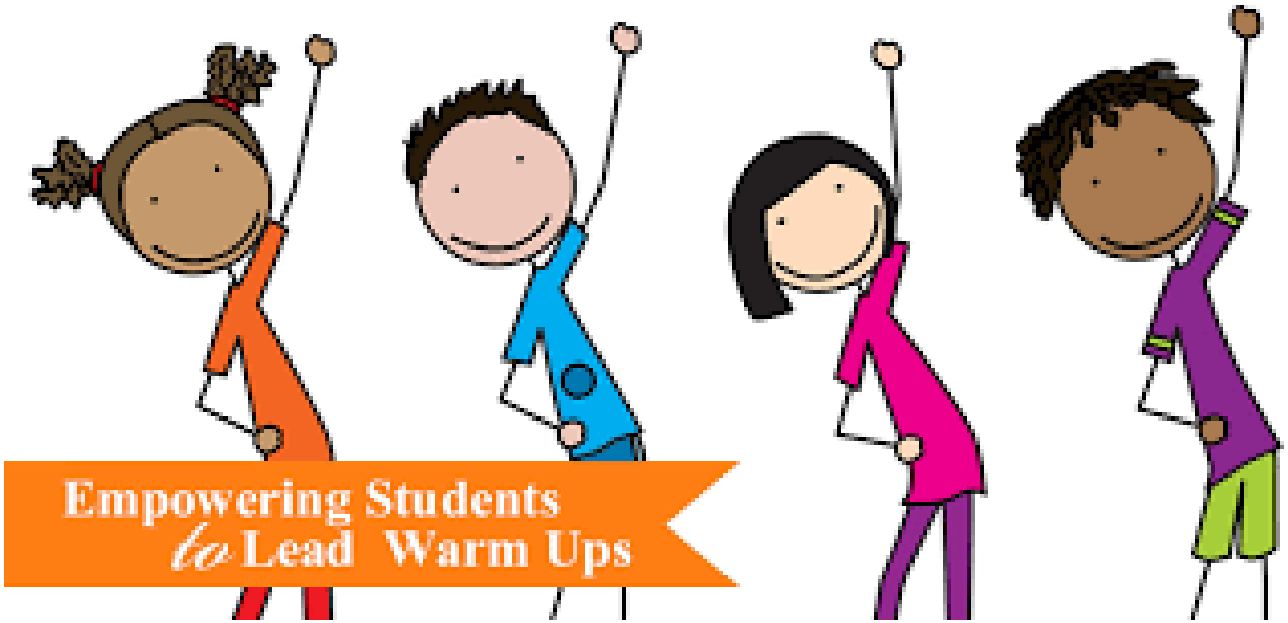
The Chinese University of Hong Kong (Shenzhen)

yinfeng@cuhk.edu.cn

January 13, 2020

Overview

Warm Up for Multiple Linear Regression



Linear Regression Model

The simple linear regression model is given by

$$y = f(x; \beta_0, \beta_1) + \varepsilon = \beta_0 + \beta_1 x + \varepsilon \tag{1}$$

where

- intercept β_0 and slope β_1 are unknown model parameters;
- ε is random error term with zero mean and unknown variance σ^2; (Note that, no specific error distribution is assumed herein)
- x is the input/regressor and y is the output/response;
- input x is assumed to be deterministic and precisely known throughout this lecture.

Linear Regression Model

The simple linear regression model is given by

$$y = \beta_0 + \beta_1 x + \varepsilon \tag{2}$$

Ask ourselves:

1. Which are known?
2. Which are unknown?
3. Which are deterministic?
4. Which are random/stochastic?

Linear Regression Model

We need to collect a data set $\mathcal{S} \triangleq \{(x_1, y_1), (x_2, y_2), ..., (x_n, y_n)\}$ with n data points, for which

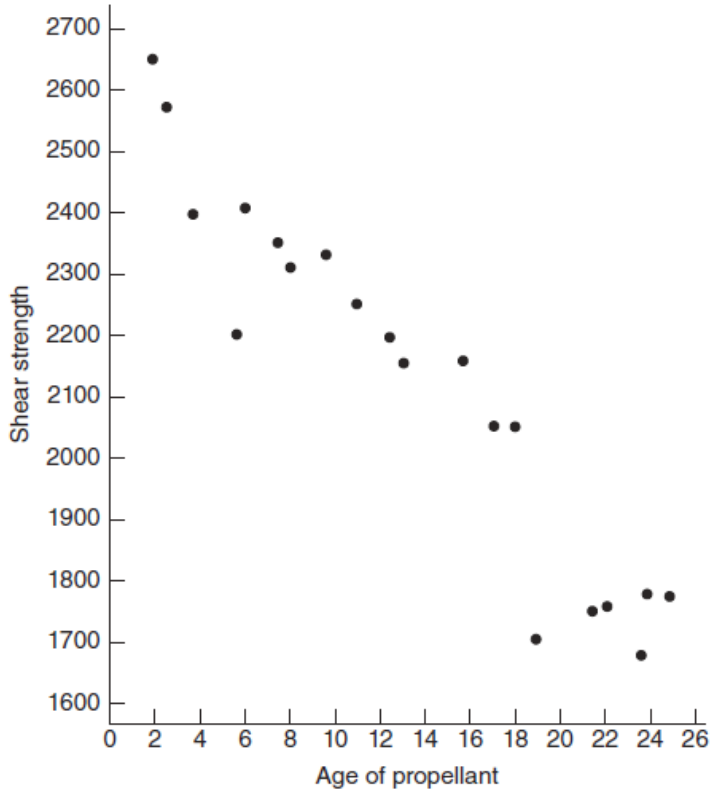
$$\begin{aligned} y_1 &= \beta_0 + \beta_1 x_1 + \varepsilon_1 \\ y_2 &= \beta_0 + \beta_1 x_2 + \varepsilon_2 \\ \vdots &= \qquad \vdots \\ y_n &= \beta_0 + \beta_1 x_n + \varepsilon_n \end{aligned} \tag{3}$$

Special Note: The model is called linear regression model solely because $f(x; \beta_0, \beta_1)$ is linear in terms of the parameters β_0 and β_1.

Linear Regression Model: An Example

Input: $x =$ age of rocket propellant, Output: $y =$ shear strength

In this example, $\mathcal{S} = \{(x_1, y_1), (x_2, y_2), ..., (x_n, y_n)\}$, $n = 20$ samples



From textbook: shear strength versus age of the rocket propellant

Linear Regression Model

Questions:

1. Is linear regression model a good model?
2. Why do we learn linear regression?

Parameter Estimation: General Procedure

After having selected the model, next we aim at finding a method to fit the model parameters. A general method comprises the following two steps:

1. Select a measure of fitness.
2. Optimize (minimize or maximize) the pre-selected measure of fitness with respect to the model parameters for a data set.

Two classic parameter estimation methods:

1. The least-squares (LS) estimation adopts the squared error loss as the measure of fitness and then minimize the total squared error over all data points (w.r.t. the parameters).
2. The maximum likelihood (ML) estimation adopts the likelihood as the measure of fitness and then maximize the likelihood of all observed data points (w.r.t. the parameters).

Parameter Estimation: Estimator Vs. Estimate

We differentiate parameter estimate and parameter estimator as follows:

1. The parameter estimate has a specific value, which is the result derived from a parameter estimation method acting on a specific data set.
2. The parameter estimator is a random variable, which is the result derived from a parameter estimation method where we don't specify any data set.

A parameter estimate can be treated as one realization of the corresponding parameter estimator.

LS Parameter Estimation: β_0 and β_1

Derive (in non-matrix manner) the estimators of β_0 and β_1 via least-squares (LS) criterion:

$$[\hat{\beta}_0, \hat{\beta}_1] = \arg \min_{\beta_0, \beta_1} S(\beta_0, \beta_1) \triangleq \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2 \tag{4}$$

where the LS estimators (in non-matrix form) are given by

$$\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}} \tag{5}$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} \tag{6}$$

The definitions of $\bar{y}$, $\bar{x}$, S_{xx} and S_{xy} as well as detailed derivations of the above results are given both on the white board and in the manuscript.

LS Parameter Estimation: σ^2

- Compute the residuals, $e_i \triangleq y_i - \hat{y}_i = y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i)$, $i = 1, 2, ..., n$.
- Define residual/error sum of squares to be

$$SS_{Res} = \sum_{i=1}^{n} e_i^2 \tag{7}$$

- Construct an estimator of σ^2 by

$$\hat{\sigma}^2 = \frac{SS_{Res}}{n-2} = MS_{Res} \tag{8}$$

- Prove that MS_{Res} is an unbiased estimator of σ^2. (Will be given in lecture 3)

Properties of the LS Estimators

Prove (given in the manuscript):

1. The sum of the residuals in the assumed model, cf. Eq.(1), is always zero.
2. The sum of the observed values y_i equals the sum of the fitted values.
3. The LS regression line always passes through the centroid of the data.
4. The sum of the residuals weighted by the corresponding value of the regressor variable always equal to zero.
5. The sum of the residuals weighted by the corresponding fitted value always equals zero, that is $\sum_{i=1}^{n} \hat{y}_i e_i = 0$.

Proofs of (1)-(5) are fairly simple, just exploit Eq.(2.5) and Eq.(2.6) given in the textbook. Details are given in the manuscript.

Properties of the LS Estimators: β_0 and β_1

Additional assumptions: The underlying function $f(x; \beta_0, \beta_1)$ is indeed a linear function with $\beta_0 = \beta_{00}$ and $\beta_1 = \beta_{10}$. Moreover, the noise terms are uncorrelated.

Prove (on the white board and in the manuscript):

$$E(\hat{\beta}_1) = \beta_1 = \beta_{10}, \qquad var(\hat{\beta}_1) = \frac{\sigma^2}{S_{xx}} \tag{9}$$

$$E(\hat{\beta}_0) = \beta_0 = \beta_{00}, \qquad var(\hat{\beta}_0) = \sigma^2 \left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}} \right) \tag{10}$$

For the special case where the error terms $\varepsilon \sim \mathcal{N}(0, \sigma^2)$, we have

$$\hat{\beta}_1 \sim \mathcal{N}\left(\beta_{10}, \frac{\sigma^2}{S_{xx}} \right) \tag{11}$$

$$\hat{\beta}_0 \sim \mathcal{N}\left(\beta_{00}, \sigma^2 \left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}} \right) \right) \tag{12}$$

ML Parameter Estimation

Major difference between the LS parameter estimation and maximum likelihood (ML) parameter estimation:

LS Parameter Estimation

The LS estimation does not assume any specific distribution of the random error ε.

ML Parameter Estimation

The ML estimation does assume a known statistical distribution of the random error ε.

ML Parameter Estimation

The idea of the ML parameter estimation is to find the estimator $\hat{\boldsymbol{\theta}}$ that maximizes the likelihood function, defined to be $p(\mathbf{y};\boldsymbol{\theta})$ herein, in other words, given any specific data set, the estimate makes the observed data (i.e., the output) most probable (in terms of probability density)!

In math language, we obtain the ML parameter estimator via

$$\hat{\boldsymbol{\theta}} = \max_{\boldsymbol{\theta}} p(\mathbf{y};\boldsymbol{\theta}) \tag{13}$$

For most of the error distributions, there is NO closed-form expression of $\hat{\boldsymbol{\theta}}$. But Gaussian error distribution is an exception!

ML Parameter Estimation: An Example

For the simple linear regression model with i.i.d. Gaussian errors, the likelihood function is

$$p(\mathbf{y}; \boldsymbol{\beta}, \sigma^2) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} \exp\left[\frac{-(y_i - \beta_0 - \beta_1 x_i)^2}{2\sigma^2}\right] \tag{14}$$

Very often, we take the logarithm of the likelihood function, short for log-likelihood, namely

$$\ln p(\mathbf{y}; \boldsymbol{\beta}, \sigma^2) = -\frac{n}{2} \ln 2\pi - \frac{n}{2} \ln \sigma^2 - \frac{1}{2\sigma^2} \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2 \tag{15}$$

Why can we take the logarithm?

ML Parameter Estimation: An Example

We obtain the ML parameter estimator via:

$$\hat{\boldsymbol{\theta}} = \max_{\boldsymbol{\theta}} \ln p(\mathbf{y}; \boldsymbol{\theta}). \tag{16}$$

We take the derivative of $\ln p(\mathbf{y}; \boldsymbol{\theta})$ with respect to $\boldsymbol{\theta}$ and set it equal to zeros, yielding:

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} \tag{17}$$

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} y_i (x_i - \bar{x})}{\sum_{i=1}^{n} (x_i - \bar{x})^2} = \frac{S_{xy}}{S_{xx}} \tag{18}$$

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2}{n} \tag{19}$$

Is this result as same as the LS result?

Prediction with Fitted Parameters

When we are given a novel input x_*, the output is predicted by $y_* = \hat{\beta}_0 + \hat{\beta}_1 x_*$.

Given a training data set, when the parameter estimation is done, the prediction model is solely determined by the trained model parameters and the original data set will be discarded totally. Such model is also called parametric model.

As a summary, the linear regression model we learned in this lecture is a deterministic linear parametric regression model.

Hypothesis Testing on the Model Parameters

Additional Assumption

The error terms $\varepsilon_1, \varepsilon_2, ..., \varepsilon_n$ are Gaussian/normally and independently distributed with zero mean and variance σ^2.

We wish to test the hypothesis that the slope equals a constant, say the true slope β_{10}. The hypotheses are:

$$H_0 : \beta_1 = \beta_{10}, \quad H_1 : \beta_1 \neq \beta_{10} \tag{20}$$

We use the following test statistic:

$$t_0 = \frac{\hat{\beta}_1 - \beta_{10}}{\sqrt{MS_{Res}/S_{xx}}} = \frac{\hat{\beta}_1 - \beta_{10}}{se(\hat{\beta}_1)} \sim t_{n-2} \tag{21}$$

for which we need to prove (will be given in lecture 2):

1. $\frac{(n-2)MS_{Res}}{\sigma^2} = \frac{SS_{Res}}{\sigma^2} \sim \chi^2_{n-2}$
2. MS_{Res} and $\hat{\beta}_1$ are independent

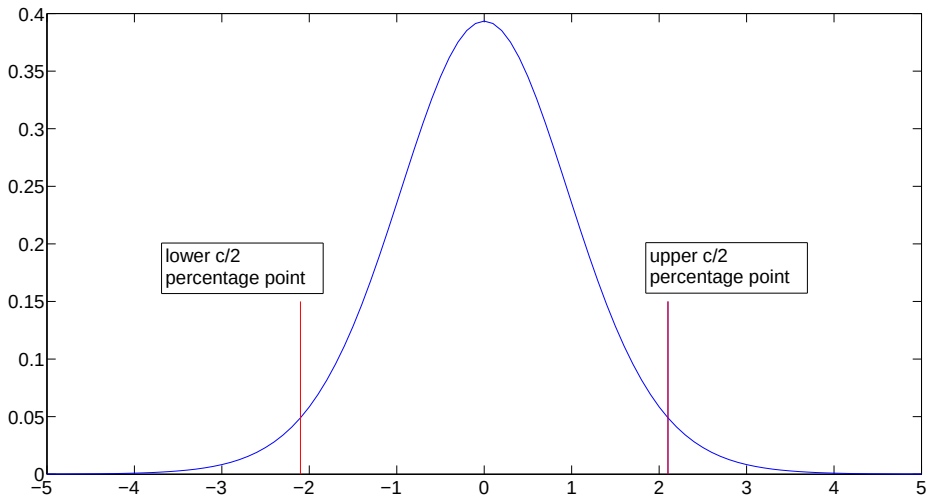
Hypothesis Testing on Model Parameters

Test procedure:

1. Compute a realization of t_0, given the observed data $\mathcal{S}$.
2. Compare the value of t_0 with the upper $c/2$ percentage point of the t_{n-2} distribution $t_{c/2,n-2}$.
3. Reject the null hypothesis $H_0 : \beta_1 = \beta_{10}$, if $|t_0| > t_{c/2,n-2}$.

Hypothesis Testing on Model Parameters

Example: For the Rocket propellant example, there are $n = 20$ samples and we got $\hat{\beta}_1 = -37.15$ and the standard error of the slope estimator $se(\hat{\beta}_1) = 2.89$. Assuming $\beta_{10} = -39$ is the true slope, yields $t_0 = 0.64$.



Central t-distribution with 18 degrees of freedom and the upper $c/2 = 2.5$ percentage point $t_{0.025,18} = 2.101$

Hypothesis Testing on Model Parameters

Additional Assumption

The error terms $\varepsilon_1, \varepsilon_2, ..., \varepsilon_n$ are Gaussian/normally and independently distributed with zero mean and variance σ^2.

Similarly, we can test the hypothesis that the intercept equals a constant, say β_{00}.
The hypotheses are:

$$H_0 : \beta_0 = \beta_{00}, \quad H_1 : \beta_0 \neq \beta_{00} \tag{22}$$

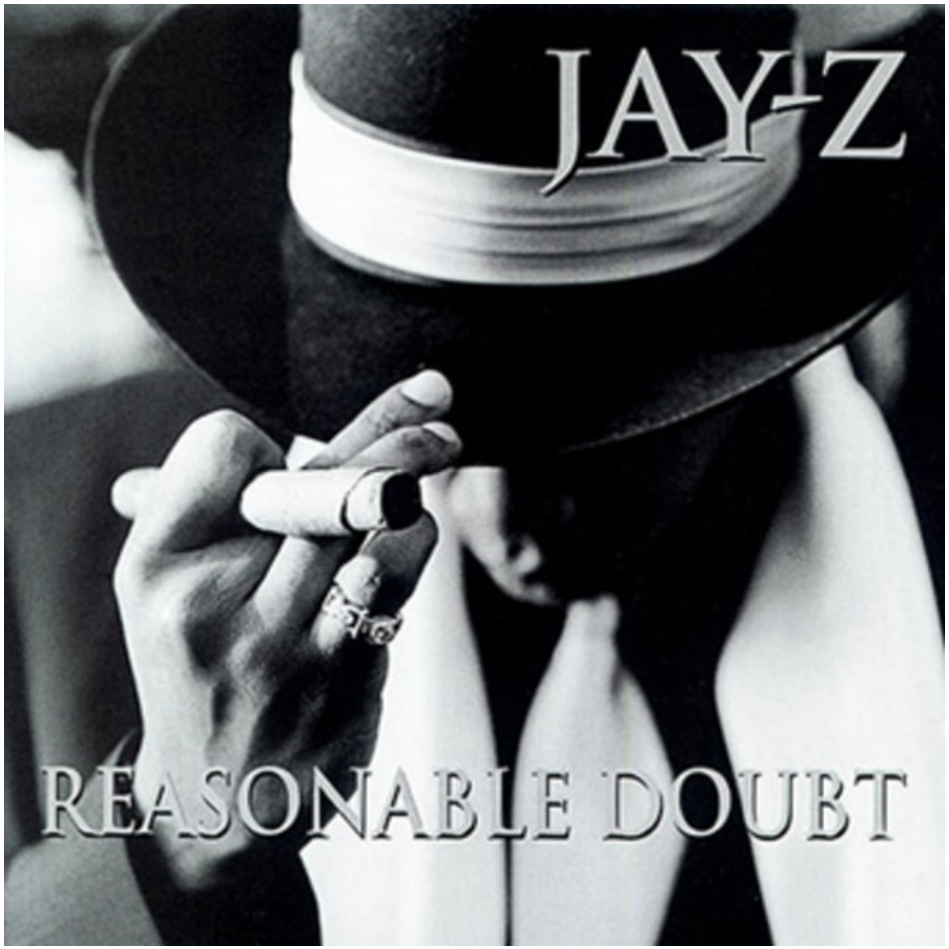
We use the following test statistic:

$$t_0 = \frac{\hat{\beta}_0 - \beta_{00}}{\sqrt{MS_{Res}(1/n + \bar{x}^2/S_{xx})}} = \frac{\hat{\beta}_0 - \beta_{00}}{se(\hat{\beta}_0)} \sim t_{n-2} \tag{23}$$

Similarly, we reject the null hypothesis $H_0 : \beta_0 = \beta_{00}$, if $|t_0| > t_{c/2,n-2}$.

Hypothesis Testing on Model Parameters

What is the problem with the above hypothesis testing?



Hypothesis Testing on Model Parameters

A special (but more reasonable) case of the above hypothesis test related to the significance of regression is:

$$H_0 : \beta_1 = 0, \quad H_1 : \beta_1 \neq 0 \tag{24}$$

Key conclusions:

- Accepting H_0 implies:
  - a there is NO linear relationship between x and y
  - b there may exist nonlinear relationship between x and y
- Rejecting H_0 implies:
  - a there is linear relationship between x and y
  - b there may exist nonlinear relationship between x and y (e.g. when the true function $f(x) = \beta_1 x + \beta_2 x^2$)

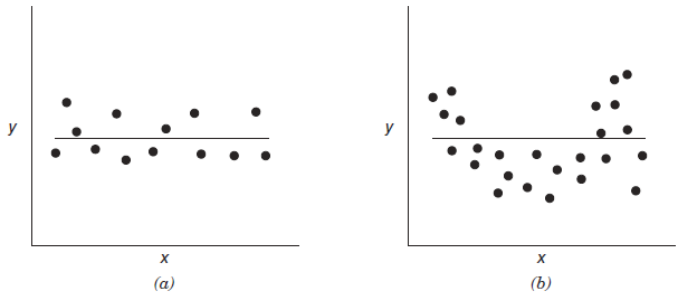


Figure 2.2 Situations where the hypothesis H_0: $\beta_1 = 0$ is not rejected.

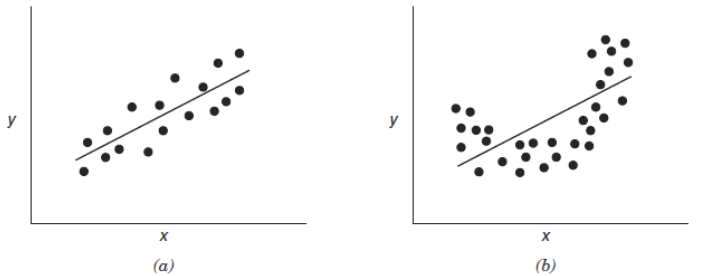
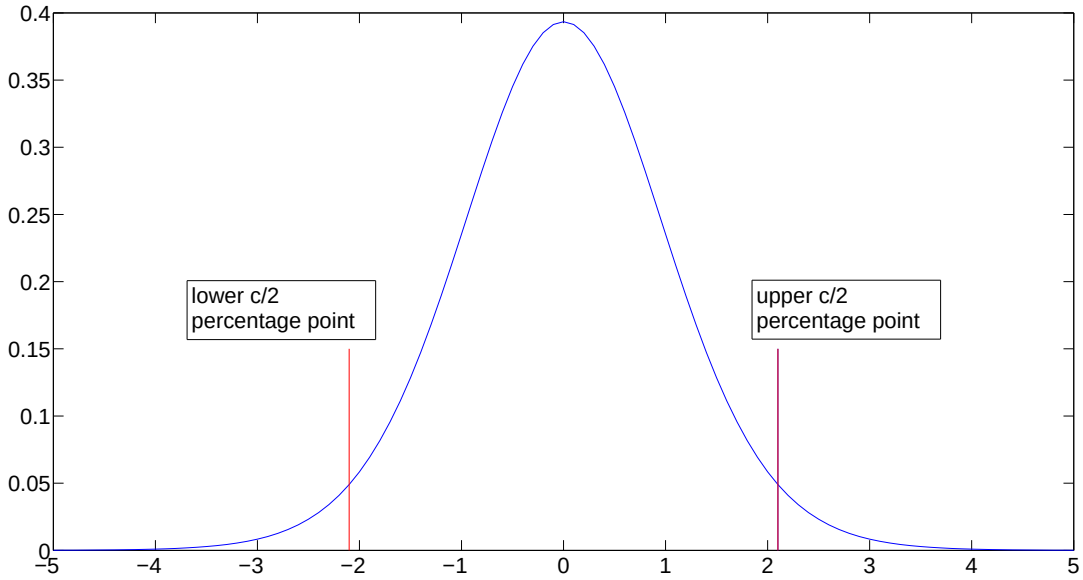


Figure 2.3 Situations where the hypothesis H_0: $\beta_1 = 0$ is rejected.

Hypothesis Testing on Model Parameters

Example: For the rocket propellant example, there are $n = 20$ samples and we got $\hat{\beta}_1 = -37.15$ and the standard error of the slope estimator $se(\hat{\beta}_1) = 2.89$. Testing if $\beta_{10} = 0$, yields $t_0 = -12.85$.



Central t-distribution with 18 degrees of freedom and the upper $c/2 = 2.5$ percentage point $t_{0.025,18} = 2.101$

Summary

To summarize with some keywords:

1. linear, deterministic, parametric model
2. measure of fitness: LS vs. ML
3. parameter estimator vs. estimate
4. residual
5. Gaussian error
6. hypothesis test: t-test
7. test for significance

Appendix

TABLE A.3 Percentage Points of the t Distribution

	α							
v	.40	.25	.10	.05	.025	.01	.005	.0025
1	.325	1.000	3.078	6.314	12.706	31.821	63.657	127.32
2	.289	.816	1.886	2.920	4.303	6.965	9.925	14.089
3	.277	.765	1.638	2.353	3.182	4.541	5.841	7.453
4	.271	.741	1.533	2.132	2.776	3.747	4.604	5.598
5	.267	.727	1.476	2.015	2.571	3.365	4.032	4.773
6	.265	.718	1.440	1.943	2.447	3.143	3.707	4.317
7	.263	.711	1.415	1.895	2.365	2.998	3.499	4.029
8	.262	.706	1.397	1.860	2.306	2.896	3.355	2.833
9	.261	.703	1.383	1.833	2.262	2.821	3.250	3.690
10	.260	.700	1.372	1.812	2.228	2.764	3.169	3.581
11	.260	.697	1.363	1.796	2.201	2.718	3.106	3.497
12	.259	.695	1.356	1.782	2.179	2.681	3.055	3.428
13	.259	.694	1.350	1.771	2.160	2.650	3.012	3.372
14	.258	.692	1.345	1.761	2.145	2.624	2.977	3.326
15	.258	.691	1.341	1.753	2.131	2.602	2.947	3.286
16	.258	.690	1.337	1.746	2.120	2.583	2.921	3.252
17	.257	.689	1.333	1.740	2.110	2.567	2.898	3.222
18	.257	.688	1.330	1.734	2.101	2.552	2.878	3.197
19	.257	.688	1.328	1.729	2.093	2.539	2.861	3.174
20	.257	.687	1.325	1.725	2.086	2.528	2.845	3.153
21	.257	.686	1.323	1.721	2.080	2.518	2.831	3.135
22	.256	.686	1.321	1.717	2.074	2.508	2.819	3.119
23	.256	.685	1.319	1.714	2.069	2.500	2.807	3.104
24	.256	.685	1.318	1.711	2.064	2.492	2.797	3.091
25	.256	.684	1.316	1.708	2.060	2.485	2.787	8.078

Percentage points of Central t-distribution with different numbers of degree of freedom.