THE UNIVERSITY OF QUEENSLAND
AUSTRALIA

Analysis of Engineering and Scientific Data

Semester 1 – 2019

Sabrina Streipert s.streipert@uq.edu.au

Chapter 9: Hypothesis Testing – continued

So far:

- Formulate $H_0 : \mu = \mu_0$ and $H_1 : \mu > \mu_0$
- Fix α-error (often 0.05)
- Find c for the rejection Region $\mathcal{R}$

$$\mathcal{R} = \{\bar{x} > \mu_0 + c\}$$

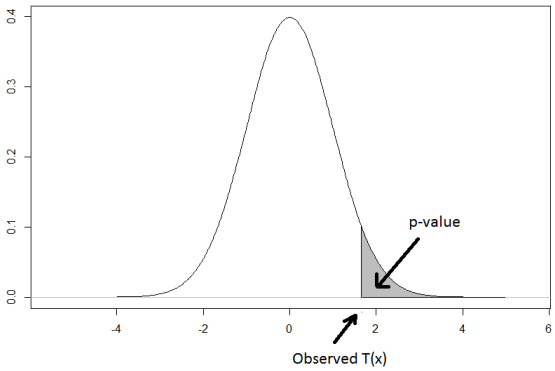
- Check if sample mean is in rejection region or not.

Alternatively: p-value Approach

The **p-value** is smallest level of significance α that would lead to rejection of the null hypothesis H_0 with the given data.

p-value Method

1. Collect data.
2. Give the null and alternative hypotheses (H_0 and H_1).
3. Choose an appropriate test statistic.
4. Determine the distribution of the test statistic under H_0.
5. Evaluate the outcome of the test statistic.
6. Calculate the p-value.
7. Accept or reject H_0 based on the p-value.



p-value low $\Rightarrow$ reject H_0

p-value	evidence
< 0.01	very strong evidence against H_0
$0.01 - 0.05$	moderate evidence against H_0
$0.05 - 0.10$	suggestive evidence against H_0
> 0.1	little or no evidence against H_0

Example revisited

- Let $X_1, \ldots, X_n \sim \mathsf{N}(\mu, \sigma^2)$, ($\sigma$ is known).
- We would like to test $H_0 : \mu = \mu_0$, $H_1 : \mu > \mu_0$

Solution:

$T = \bar{X}$, p-value:

$$P(T > \bar{x}) = P(Z > \frac{\bar{x} - \mu_0}{\frac{\sigma}{\sqrt{n}}}) = 1 - \Phi\left(\frac{\bar{x} - \mu_0}{\frac{\sigma}{\sqrt{n}}}\right)$$

Alien example - revisited

Recall $\overline{X} = 0.14$. Define $T = \bar{X}$, $\quad H_0 : \mu = 0.1$, $\quad H_1 : \mu > 0.1$

Solution:

$$P(T > 0.14) = P(Z > \frac{0.14 - 0.1}{\frac{0.25}{\sqrt{400}}}) = 1 - \Phi(3.2) = 0.00069 < 0.01$$

$\Longrightarrow$ reject H_0

Would the alien's opinion change if they would have picked a different sample?

Solution:

Possibly. Take for example $n = 40$, then

$$P(T > 0.14) = P(Z > \frac{0.14 - 0.1}{\frac{0.25}{\sqrt{40}}}) = 1 - \Phi(1.0119) = 0.1563$$

so there is nearly no evidence to reject H_0.

Types of tests

- **Right one-sided test:** H_0 is rejected $T \geq t$

- **Left one-sided test:** H_0 is rejected if $T \leq t$

- **Two-sided test:** H_0 is rejected if $T \geq t$ or $T \leq -t$

Example:

A manufacturer produces crankshafts for an automobile engine. The wear of the crankshaft after 100.000 miles in 0.0001 inches is of interest due to warranty claims. A random sample of 15 shafts is tested with a sample mean of 2.78 and known standard deviation of 0.9. Test the claim that the expected wear is not equal to 0.0003 inches.

Solution:

$n = 15, \qquad \sigma = 0.9, \qquad \bar{x} = 2.78$

$H_0 : \mu = 3, \qquad H_1 : \mu \neq 3$

Pick $\alpha = 0.05$.

$\mathcal{R} = \{T < \mu - c\} \cup \{T > \mu + c\}$

$$0.05 = \alpha = P_{H_0}(T < \mu - c) + P_{H_0}(T > \mu + c) = P\left(\frac{T-\mu_0}{\sigma/\sqrt{n}} < \frac{-c}{\sigma/\sqrt{n}}\right) + P\left(\frac{T-\mu_0}{\sigma/\sqrt{n}} > \frac{c}{\sigma/\sqrt{n}}\right)$$
$$= \underbrace{\Phi\left(\frac{-c}{\sigma/\sqrt{n}}\right)}_{=1-\Phi\left(\frac{c}{\sigma/\sqrt{n}}\right)} + 1 - \Phi\left(\frac{c}{\sigma/\sqrt{n}}\right) = 2\left[1 - \Phi\left(\frac{c}{\sigma/\sqrt{n}}\right)\right]$$

$$\longrightarrow \qquad 1 - \tfrac{0.05}{2} = \Phi\left(\tfrac{c}{\sigma/\sqrt{n}}\right) \quad \longrightarrow \qquad 0.975 = \Phi\left(\tfrac{c}{\sigma/\sqrt{n}}\right)$$
$$\longrightarrow \qquad 1.96 = \tfrac{c}{\sigma/\sqrt{n}} \qquad c = 1.96\tfrac{\sigma}{\sqrt{n}} = 1.96\tfrac{0.9}{\sqrt{15}} = 0.4555$$

Reject H_0 if $T > 3 + 0.4555 = 3.4555$ or if $T < 3 - 0.4555 = 2.5445$. Since $2.5445 < \bar{x} = 2.78 < 3.4555$, we do NOT reject H_0.

Exam Cheat Sheet: Testing Hypotheses on the Mean, Variance Known

➢ **Testing Hypotheses on the Mean, Variance Known (Z-Tests)**

Model: $X_i \overset{i.i.d.}{\sim} N(\mu, \sigma^2)$ with μ unknown but σ^2 known.

Null hypothesis: $H_0 : \mu = \mu_0$.

Test statistic: $z = \dfrac{\overline{x} - \mu_0}{\sigma/\sqrt{n}}, \qquad Z = \dfrac{\overline{X} - \mu_0}{\sigma/\sqrt{n}}.$

Alternative Hypotheses	***P*-value**	**Rejection Criterion for Fixed-Level Tests**
$H_1 : \mu \neq \mu_0$	$P = 2\big[1 - \Phi\big(\lvert z\rvert\big)\big]$	$z > z_{1-\alpha/2}$ or $z < z_{\alpha/2}$
$H_1 : \mu > \mu_0$	$P = 1 - \Phi\big(z\big)$	$z > z_{1-\alpha}$
$H_1 : \mu < \mu_0$	$P = \Phi\big(z\big)$	$z < z_{\alpha}$

Example:

Given $H_0 : \mu = 5, \qquad H_1 : \mu < 5$, σ known. Calculate the p-value for the test statistic $z_0 = 0.4$.

Solution:

$$p\text{-value} = P_{H_0}(T < \bar{x}) = P\left(\tfrac{T-5}{\frac{\sigma}{\sqrt{n}}} < \tfrac{\bar{x}-5}{\frac{\sigma}{\sqrt{n}}}\right) = \Phi\left(\underbrace{\frac{\bar{x}-5}{\frac{\sigma}{\sqrt{n}}}}_{z_0}\right) = \Phi(0.4) = 0.6554$$

Since $p > 0.1$, we would not reject H_0.

Recall

We distinguish between the two cases:

- Unrealistic (but simpler): The population variance, σ^2, is known.
- More realistic: The variance is not known and estimated by the sample variance, s^2.

Private school - revisited

Private school claims that its students have a higher IQ. Recall $IQ_{all} \sim N(100, 16)$ and $IQ_A \sim N(105, 25)$.

- Should we try to place our child in this school?
- Is the observed result *significant* **(can be trusted?)**, or due to a *chance*?

Variance is known to be 16.

Medical treatment - revisited

14 subjects were randomly assigned to control or treatment group of the experimental medical treatment. The survival times (in days) were

	Data	Mean
Treatment group	91, 140, 16, 32, 101, 138, 24	77.428
Control group	3, 115, 8, 45, 102, 12, 18	43.285

The variance is not known and estimated by the sample variance, s^2.

Known variance — the Z-test

- A Z-test is any statistical test for which the distribution of the test statistic (the mean) under the null hypothesis can be approximated by a normal distribution (with known variance).
- Thanks to the central limit theorem, many test statistics are approximately **normally distributed** for large enough samples.

PROBLEM: Z test uses CLT (required: σ is known and n is "large")

- If σ^2 is unknown $\rightarrow$ estimate using sample variance s^2

  Recall: Sample variance $= s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}$

  $\longrightarrow$ the test-statistic reads as

$$T = \frac{\overline{X} - \mu_0}{s/\sqrt{n}},$$

- If n small or σ unknown $\rightarrow$ T no longer follows a Normal distribution!

- **Good news:** T follows The Student-t Distribution.

The t-distribution

- The pdf of Student-t Distribution with parameter k (degrees of freedom) is:

$$f(x) = \frac{\Gamma((k+1)/2)}{\sqrt{\pi k}\Gamma(k/2)} \frac{1}{[(x^2/k)+1]^{(k+1)/2}}, \quad -\infty < x < \infty,$$

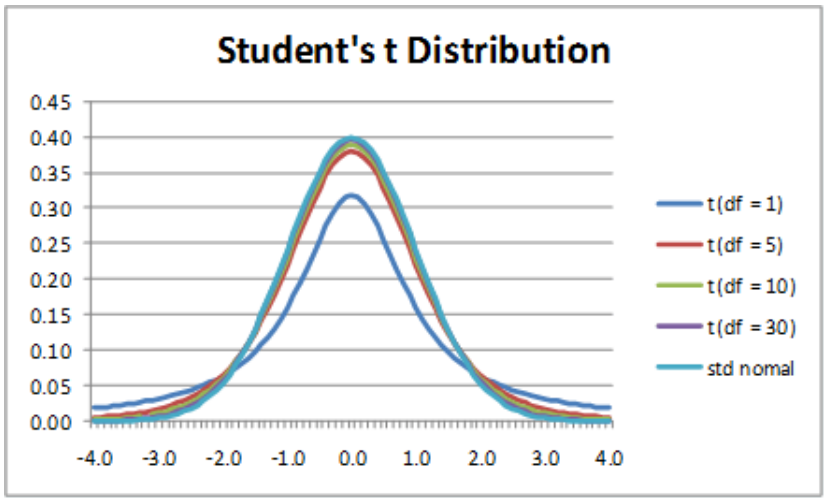
  where $\Gamma(\cdot)$ is the Gamma-function:

$$\Gamma(k) = \int_0^\infty x^{k-1} e^{-x} \mathrm{d}x.$$

- It is a symmetric distribution about 0 and as $k \rightarrow \infty$, it approaches a standard Normal distribution.

t-test

The t statistic (introduced in 1908 by William Sealy Gosset, a chemist working for the Guinness brewery in Ireland) with $n-1$ degrees of freedom is



$$T_{n-1} = \frac{\overline{X} - \mu_0}{s/\sqrt{n}},$$

where s is the estimated standard deviation:

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2.$$

- For large n, t-test is indistinguishable from the z-test.

Confidence and prediction intervals

- If $\overline{x}$ and s are the mean and standard deviation of a random sample from a normal distribution with unknown variance σ^2, a $100(1-\alpha)$ confidence interval on μ is given by:

$$\overline{x} - t_{1-\alpha/2,n-1}\frac{s}{\sqrt{n}} \leq \mu \leq \overline{x} + t_{1-\alpha/2,n-1}\frac{s}{\sqrt{n}},$$

where $t_{1-\alpha/2,n-1}$ is the $1-\alpha/2$ quantile of the t distribution with n - 1 degrees of freedom.

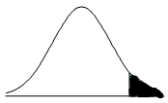
- A related concept is a $100(1-\alpha)$ prediction interval (PI) on a single future observation from a normal distribution is given by

$$\overline{x} - t_{1-\alpha/2,n-1}s\sqrt{1+\frac{1}{n}} \leq X_{n+1} \leq \overline{x} + t_{1-\alpha/2,n-1}s\sqrt{1+\frac{1}{n}}$$

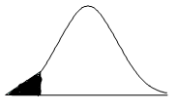
  This is the range where we expect the $n+1$ observation to be, after observing n observations and computing $\overline{x}$ and s.

Types of tests

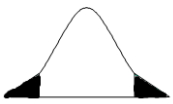
- **Right one-sided test:** where H_0 is rejected for the p-value defined by $P_{H_0}(T \geq t)$.



- **Left one-sided test:** where H_0 is rejected for the p-value defined by $P_{H_0}(T \leq t)$.



- **Two-sided test:** where H_0 is rejected for the p-value defined by $P_{H_o}(T \geq t) + P_{H_o}(T \leq -t) = 2P_{H_o}(T \geq t)$.



Model:	$X_i \overset{i.i.d.}{\sim} N(\mu, \sigma^2)$	with both μ and σ^2 unknown.
Null hypothesis:	$H_0 : \mu = \mu_0$.	
Test statistic:	$t = \dfrac{\bar{x} - \mu_0}{s/\sqrt{n}}$, $\quad T = \dfrac{\overline{X} - \mu_0}{S/\sqrt{n}}$.	

Alternative Hypotheses	***P*-value**	**Rejection Criterion for Fixed-Level Tests**
$H_1 : \mu \neq \mu_0$	$P = 2\big[1 - F_{n-1}\big(\lvert t\rvert\big)\big]$	$t > t_{1-\alpha/2,n-1}$ or $t < t_{\alpha/2,n-1}$
$H_1 : \mu > \mu_0$	$P = 1 - F_{n-1}(t)$	$t > t_{1-\alpha,n-1}$
$H_1 : \mu < \mu_0$	$P = F_{n-1}(t)$	$t < t_{\alpha,n-1}$

Note, since the pdf of the Student t-distribution is symmetric, $t_{\frac{\alpha}{2}} = -t_{1-\frac{\alpha}{2}}$.

- In the p-value calculation, $F_{n-1}(\cdot)$ denotes the CDF of the t-distribution with $n-1$ degrees of freedom.
- As opposed to $\Phi(\cdot)$, the CDF of t is not tabulated in standard tables. So to calculate p-values, we use software.

Example 1:

Five samples of a material were tested in a structure, and the average interior temperature in Celcius reported:

23.01 22.22 22.04 22.62 22.59

Test the claim that the interior temperature is 22.5.

Solution:

$\bar{x} = 22.496, \qquad s = 0.3783$

$H_0 : \mu = 22.5, \qquad H_1 : \mu \neq 22.5$

Pick $\alpha = 0.05$

Step 1: Get $t_{1-\frac{\alpha}{2},n-1}$, which is the value such that the cdf of the t-distribution (with $n-1=4$ degrees of freedom) is equal to $1-\frac{\alpha}{2} = 0.975$. Looking it up in the table yields $t_{1-\frac{\alpha}{2},n-1} = 2.776$.

Hence, we reject H_0 if $\frac{\bar{x}-\mu_0}{\frac{s}{\sqrt{n}}} < -2.776$ or if $\frac{\bar{x}-\mu_0}{\frac{s}{\sqrt{n}}} > 2.776$.

Step 2: Calculate

$$\frac{\bar{x}-\mu_0}{\frac{s}{\sqrt{n}}} = \frac{22.496-22.5}{\frac{0.3783}{\sqrt{5}}} = -0.0236$$

Step 3: Decision:

Since $-2.776 < \frac{\bar{x}-\mu_0}{\frac{s}{\sqrt{n}}} = \frac{22.496-22.5}{\frac{0.3783}{\sqrt{5}}} = -0.0236 < 2.776$,

we do not reject H_0.

t-Distribution Quantiles

ν	$Q(.9)$	$Q(.95)$	$Q(.975)$	$Q(.99)$	$Q(.995)$	$Q(.999)$	$Q(.9995)$
1	3.078	6.314	12.706	31.821	63.657	318.317	636.607
2	1.886	2.920	4.303	6.965	9.925	22.327	31.598
3	1.638	2.353	3.182	4.541	5.841	10.215	12.924
4	1.533	2.132	2.776	3.747	4.604	7.173	8.610
5	1.476	2.015	2.571	3.365	4.032	5.893	6.869
6	1.440	1.943	2.447	3.143	3.707	5.208	5.959
7	1.415	1.895	2.365	2.998	3.499	4.785	5.408
8	1.397	1.860	2.306	2.896	3.355	4.501	5.041
9	1.383	1.833	2.262	2.821	3.250	4.297	4.781
10	1.372	1.812	2.228	2.764	3.169	4.144	4.587
11	1.363	1.796	2.201	2.718	3.106	4.025	4.437
12	1.356	1.782	2.179	2.681	3.055	3.930	4.318
13	1.350	1.771	2.160	2.650	3.012	3.852	4.221
14	1.345	1.761	2.145	2.624	2.977	3.787	4.140
15	1.341	1.753	2.131	2.602	2.947	3.733	4.073
16	1.337	1.746	2.120	2.583	2.921	3.686	4.015
17	1.333	1.740	2.110	2.567	2.898	3.646	3.965
18	1.330	1.734	2.101	2.552	2.878	3.610	3.922
19	1.328	1.729	2.093	2.539	2.861	3.579	3.883
20	1.325	1.725	2.086	2.528	2.845	3.552	3.849
21	1.323	1.721	2.080	2.518	2.831	3.527	3.819
22	1.321	1.717	2.074	2.508	2.819	3.505	3.792
23	1.319	1.714	2.069	2.500	2.807	3.485	3.768
24	1.318	1.711	2.064	2.492	2.797	3.467	3.745
25	1.316	1.708	2.060	2.485	2.787	3.450	3.725
26	1.315	1.706	2.056	2.479	2.779	3.435	3.707
27	1.314	1.703	2.052	2.473	2.771	3.421	3.690
28	1.313	1.701	2.048	2.467	2.763	3.408	3.674
29	1.311	1.699	2.045	2.462	2.756	3.396	3.659
30	1.310	1.697	2.042	2.457	2.750	3.385	3.646
40	1.303	1.684	2.021	2.423	2.704	3.307	3.551
60	1.296	1.671	2.000	2.390	2.660	3.232	3.460
120	1.289	1.658	1.980	2.358	2.617	3.160	3.373
∞	1.282	1.645	1.960	2.326	2.576	3.090	3.291

This table was generated using the "INVCDF" command in Minitab.

Example 2:

A golf club producer claims that the mean coefficient of restitution exceeds 0.82. Suppose you take a sample with the following restitution coefficient measurements:

0.8411	0.8191	0.8182	0.8125	0.8750
0.8580	0.8532	0.8483	0.8276	0.7983
0.8042	0.8730	0.8282	0.8359	0.8660

Solution:

$\bar{x} = 0.83725, s = 0.02456, n = 15$

$H_0 : \mu = 0.82, \qquad H_1 : \mu > 0.82$

Pick $\alpha = 0.05$.

Step 1: Get $t_{1-\alpha,n-1}$:

$t_{1-\alpha,n-1}$ is the value such that the cdf of the t-distribution (with $n-1 = 14$ degrees of freedom) is equal to $1 - \alpha = 0.95$.

Looking it up in the table yields $t_{1-\alpha,n-1} = 1.761$.

Hence, we reject H_0 if $\frac{\bar{x}-\mu_0}{\frac{s}{\sqrt{n}}} > 1.761$.

Step 2: Calculate

$$\frac{\bar{x} - \mu_0}{\frac{s}{\sqrt{n}}} = \frac{0.83725 - 0.82}{\frac{0.02456}{\sqrt{15}}} = 2.72$$

Step 3: Decision:

Since $\frac{\bar{x}-\mu_0}{\frac{s}{\sqrt{n}}} = 2.72 > 1.761$, we would reject H_0.

Chapter 11: Simple Linear Regression

- Aim: Study or analysis of the relationship between two or more variables

  e.g. Pressure of a gas in a container versus its temperature

> We examine a dependent variable and one or more independent variables (= predictors).
> $\Longrightarrow$ Regression Analysis

- Key importance (conditional expectation):

$$\mathbb{E}[Y \mid x] = \mu_{Y|x}$$

Suppose for now, the variable Y depends linearly on only one predictor, i.e.:

$$\mathbb{E}[Y \mid x] = \mu_{Y|x} = \beta_0 + \beta_1 x$$

$\Longrightarrow$

$$Y = \beta_0 + \beta_1 x + \epsilon,$$

where:

- x is a (non-random) predictor, and
- ϵ is a R.V.(=noise) with $\mathbb{E}[\epsilon] = 0$, $\mathrm{Var}(\epsilon) = \sigma^2$.

Assumptions:

- Normality of residuals,
- Constant variance, and,
- Independence of observations

Method:

- Collect data:

$$(x_1, y_1), \ldots, (x_n, y_n).$$

- Assume linear relation:

$$y \approx \beta_0 + \beta_1 x \qquad \leftrightarrow \qquad y = \beta_0 + \beta_1 x + \epsilon$$

- Since we do not have all possible tuples, we can only estimate β_0 and β_1 by $\hat{\beta}_0$ and $\hat{\beta}_1$, respectively, i.e.,

$$y_i = \hat{\beta}_0 + \hat{\beta}_1 x_i + e_i, \quad i = 1, \ldots, n.$$

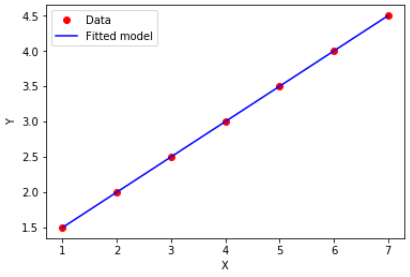
- $e_i =$ **residual**.

Use $\hat{\beta}_0, \hat{\beta}_1$ for predictions.

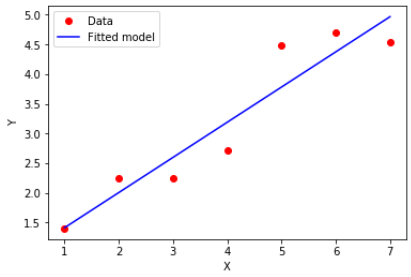
$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x.$$

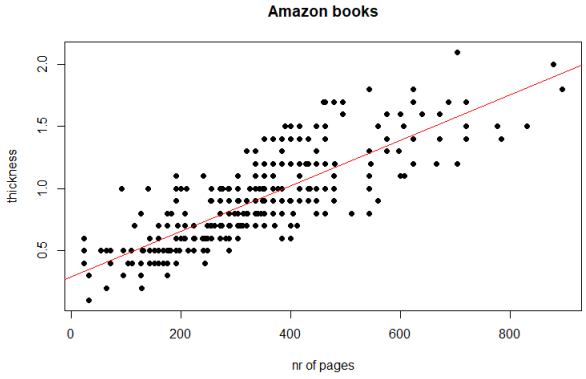
Note that we can also compute predicted observations for our data $(x_i, y_i)_{\{1 \leq i \leq n\}}$.

Ideally, we would like to find $\hat{\beta}_0$ and $\hat{\beta}_1$, such that $y_i = \hat{y}_i$, that is, $e_i = 0$.



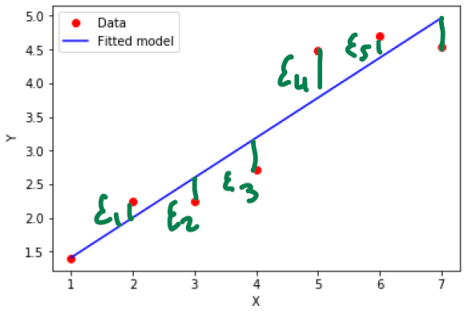
Much more likely:






```
D <- read.delim("amazon-books.txt")
plot(D$NumPages, D$Thick)
abline(lm(D$Thick~D$NumPages), col='red')
```

Total mean squared error



Total Mean Squared Error:

$$L = SS_E = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2$$

The least squares estimators

- To find the best estimators $\hat{\beta}_0$ and $\hat{\beta}_1$, we would like to minimize

$$L = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2.$$

- Specifically, solve $\hat{\beta}_0, \hat{\beta}_1 = \text{argmin}_{\beta_0,\beta_1} \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2$.

- The solution, called the *least squares estimators* must satisfy:

  Since we want to minimize L, we take the (partial) derivative and set them equal to zero.

$$0 = \frac{\partial}{\partial \beta_0} L = \frac{\partial}{\partial \beta_0} \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2 = \sum_{i=1}^{n} \frac{\partial}{\partial \beta_0} (y_i - \beta_0 - \beta_1 x_i)^2 = \sum_{i=1}^{n} 2 (y_i - \beta_0 - \beta_1 x_i) (-1)$$

$$0 = \frac{\partial}{\partial \beta_1} L = \sum_{i=1}^{n} \frac{\partial}{\partial \beta_1} (y_i - \beta_0 - \beta_1 x_i)^2 = \sum_{i=1}^{n} 2 (y_i - \beta_0 - \beta_1 x_i) (-x_i)$$