50.034 - Introduction to Probability and Statistics

Week 13 – Lecture 23

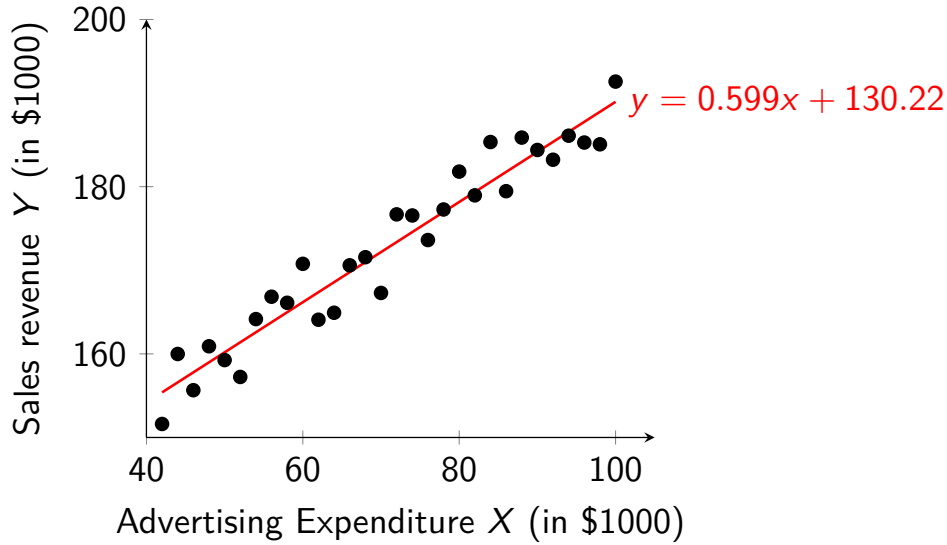
January–May Term, 2019

Lecture 23 Assigned Readings: 11.3, 11.5

Outline of Lecture

- Simple linear regression
- Least squares estimators
- Joint distribution of least squares estimators
- Hypothesis testing on regression coefficients
- Multiple linear regression

Recall: Intuition for Simple Linear Regression



Suppose we treat advertising expenditure (in \$1000) as a R.V. X, and we treat sales revenue (in \$1000) as a R.V. Y.

- We have direct control over how much we want to invest in advertising (X), but we have no direct control over Y.
- From the plot, it seems that Y is roughly a linear function of X, say $Y \approx \beta_0 + \beta_1 X$, but there seems to be some random deviation from the linear function $\beta_0 + \beta_1 X$.
- Thus, we can model $Y = \beta_0 + \beta_1 X + E$, where E is some R.V. representing noise.

Model set-up for Simple Linear Regression

For our model $Y = \beta_0 + \beta_1 X + E$ to be precise, we first need to state our model set-up.

- There are 30 points in our plot, so there are 30 R.V.'s $X_1, \ldots, X_{30}$, which are called **predictor variables**, and 30 R.V.'s $Y_1, \ldots, Y_{30}$, which are called **response variables**.
  - A predictor variable is a R.V. for which we have direct control over its observed value.
  - Each X_i is a specific amount of advertising expenditure (in \$1000) that we can choose to invest, so we could decide to try different values

    $$X_1 = 42, X_2, = 44, X_3 = 46, \ldots X_{30} = 100,$$

    and then observe the values of the corresponding response variables $Y_1, \ldots, Y_{30}$.
  - A response variable is a R.V. that depends on the observed value of a predictor variable.
  - The observed value of Y_i (sales revenue) is a "response" to the given value $X_i = x_i$ (advertising expenditure).

A closer look at the model set-up

Let X, Y, E be R.V.'s related by $Y = \beta_0 + \beta_1 X + E$, where β_0, β_1 are unknown parameters.

- Let $\{X_1, \ldots, X_n\}$ and $\{Y_1, \ldots, Y_n\}$ be two random samples, where each X_i has the same distribution as X, and each Y_i has the same distribution as Y.
- **Interpretation:** If $(x_1, y_1), \ldots, (x_n, y_n)$ are the observed values of $(X_1, Y_1), \ldots, (X_n, Y_n)$, then we can think of these n pairs $(x_1, y_1), \ldots, (x_n, y_n)$ as pairs being sampled from the joint distribution of (X, Y).

Assumption: E is a normal R.V. with mean 0 and variance σ^2.

- **Consequence:** If we are given $X = x$, then $Y = \beta_0 + \beta_1 x + E$ is a normal R.V. with mean $\beta_0 + \beta_1 x$ and variance σ^2.
- Similarly, given $X_i = x_i$, then $Y_i = \beta_0 + \beta_1 x_i + E$ is a normal R.V. with mean $\beta_0 + \beta_1 x_i$ and variance σ^2.

More Assumptions on Model Set-up

Assumption: Predictor variables are known.

- The observed values $x_1, \ldots, x_n$ of the predictor variables $X_1, \ldots, X_n$ are known beforehand.
  - In our example where X is the advertising expenditure and Y is the sales revenue, we could never know what the sales revenue is beforehand, but we could decide beforehand how much money to invest into the advertising expenditure.
  - So in this example (and other similar examples), we decide in advance what the observed values $x_1, \ldots, x_n$ should be, and our aim is to observe the corresponding values for $Y_1, \ldots, Y_n$.
- Thus, we are interested in the condition distribution of $(Y_1, \ldots, Y_n)$ given $X_1 = x_1, \ldots, X_n = x_n$.
  - Intuitively, we can treat the observed values $x_1, \ldots, x_n$ as constants given to us.
  - In particular, we are implicitly assuming that $x_1, \ldots, x_n$ are error-free (i.e. no measurement errors).

Assumption: $Y_1, \ldots, Y_n$ are conditionally independent given the observed values $x_1, \ldots, x_n$.

- $Y_1, \ldots, Y_n$ are the response variables (in "response" to $x_1, \ldots, x_n$)

Simple Linear Regression

The model $Y = \beta_0 + \beta_1 X + E$, together with the assumptions given in the previous slides, form the **statistical model** that is used in "simple linear regression".

- Simple linear regression is a type of **statistical inference** to determine the values of β_0 and β_1.
- **Note:** $\mathbf{E}[Y|X = x] = \beta_0 + \beta_1 x$, which is linear in terms of x.
  - This explains why simple linear regression is called "linear".
  - Simple linear regression is called "simple" because the response variable depends only on **one** predictor variable.
- When $\beta_0 + \beta_1 x$ is treated as a function of the observed value x, this function is called the regression function of Y on X, or more simply, the regression of Y on X.
  - β_0 and β_1 are called regression coefficients.
  - **Intuition:** Assuming that Y is roughly linear w.r.t. X, i.e. $Y = \beta_0 + \beta_1 X +$ "noise", the idea of simple linear regression is that we can determine estimates for β_0 and β_1, based on some sample points $(x_1, y_1), \ldots, (x_n, y_n)$.

Important Note: Simple linear regression makes sense only when our assumptions are satisfied.

Least square estimators

In simple linear regression ($Y = \beta_0 + \beta_1 X + E$), we are treating β_0 and β_1 as unknown parameters.

- We can use estimation techniques to estimate β_0 and β_1.

Theorem: Suppose $Y = \beta_0 + \beta_1 X + E$ is a simple linear regression model (i.e. the assumptions on the previous slides are satisfied, including the assumption that E has variance σ^2). If $X_1 = x_1, \ldots, X_n = x_n$ are the given observed predictor variables, then the **maximum likelihood estimators** of β_0 and β_1 are:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(Y_i - \overline{Y}_n)(x_i - \overline{x}_n)}{\sum_{i=1}^{n}(x_i - \overline{x}_n)^2};$$

$$\hat{\beta}_0 = \overline{Y}_n - \beta_1 \overline{x}_n;$$

where $\overline{Y}_n$ is the sample mean of $\{Y_1, \ldots, Y_n\}$, and $\overline{x}_n = \frac{x_1 + \cdots + x_n}{n}$.

- $\hat{\beta}_0$ and $\hat{\beta}_1$ are statistics of $\{Y_1, \ldots, Y_n\}$.
- $\hat{\beta}_0$ and $\hat{\beta}_1$ are called least square estimators, because they have the same expressions as the coefficients computed in the **least squares method**.

Conditional variance of response variables

Recall: The conditional distribution of Y given $X = x$ is the normal distribution with mean $\beta_0 + \beta_1 x$ and variance σ^2.

- We can also use estimation techniques to estimate σ^2.

Theorem: If $X_1 = x_1, \ldots, X_n = x_n$ are the given observed predictor variables, then the **maximum likelihood estimator** of σ^2 (given $x_1, \ldots, x_n$) is

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

Note: $\hat{\sigma}^2$ is a statistic of $\{Y_1, \ldots, Y_n\}$.

- Given $Y_1 = y_1, \ldots, Y_n = y_n$, the observed value of $\hat{\sigma}^2$ (M.L.E.) looks very similar to $\frac{1}{n}$ times the observed sum of the square deviates that we saw in the least squares method.
  - In the least squares method, we want to minimize $\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)^2$, treated as a function of β_0 and β_1.
  - In contrast, the the maximum likelihood **estimate** of $\hat{\sigma}^2$ is $\frac{1}{n}$ times $\sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2$. To compute this, we first have to compute the maximum likelihood **estimates** of $\hat{\beta}_0$ and $\hat{\beta}_1$.

Joint distribution of least square estimators

$\hat{\beta}_0$ and $\hat{\beta}_1$ are estimators, so in particular, they are R.V.'s.

- It makes sense to consider the joint distribution of $\hat{\beta}_0$ and $\hat{\beta}_1$.

Recall: By assumption, the observed values $x_1, \ldots, x_n$ are known beforehand. Let $\overline{x}_n = \frac{x_1 + \cdots + x_n}{n}$, and define the real number

$$s_{\mathbf{x}} = \sqrt{\sum_{i=1}^{n} (x_i - \overline{x}_n)^2}.$$

Theorem: The joint distribution of $\hat{\beta}_0$ and $\hat{\beta}_1$ is a **bivariate normal distribution**.

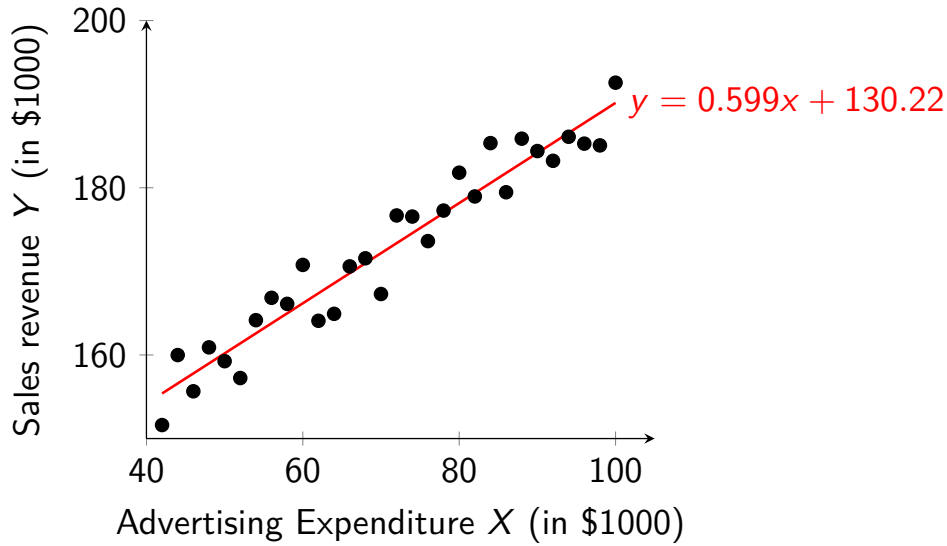
- $\hat{\beta}_0$ has mean β_0 and variance $\sigma^2\left(\frac{1}{n} + \frac{\overline{x}_n^2}{s_{\mathbf{x}}^2}\right)$.
- $\hat{\beta}_1$ has mean β_1 and variance $\frac{\sigma^2}{s_{\mathbf{x}}^2}$.
- The covariance of $\hat{\beta}_0$ and $\hat{\beta}_1$ is $\operatorname{cov}(\hat{\beta}_0, \hat{\beta}_1) = -\frac{\overline{x}_n \sigma^2}{s_{\mathbf{x}}^2}$.
  - Thus, the correlation is $\frac{\operatorname{cov}(\hat{\beta}_0, \hat{\beta}_1)}{\sqrt{\operatorname{var}(\hat{\beta}_0)\operatorname{var}(\hat{\beta}_1)}} = \frac{-\overline{x}_n}{\sqrt{\frac{1}{n} + \frac{\overline{x}_n^2}{s_{\mathbf{x}}^2}}}$.

Corollary: $\hat{\beta}_0$, $\hat{\beta}_1$ are **unbiased** estimators of β_0, β_1 respectively.

- "$\hat{\beta}_i$ is unbiased" means "$\mathbf{E}[\hat{\beta}_i] = \beta_i$".

Using simple linear regression for prediction

After computing estimates for the regression coefficients β_0, β_1, we can use the simple linear regression model to make **predictions**.



In this model, X represents advertising expenditure (in \$1000), and Y represents sales revenue (in \$1000).

- After computation, we have $\mathbf{E}[Y|X = x] = 0.599x + 130.22$.

Question: What if we want to predict the expected sales revenue if we spend \$200k on advertising expenditure?

- \$200k on advertising expenditure corresponds to $x = 200$, so our model predicts $\mathbf{E}[Y|X = 200] = 0.599(200) + 130.22 = 250.02$, i.e. the sales revenue is approximately \$250k.

A closer look at prediction in simple linear regression

Starting with the statistical model $Y = \beta_0 + \beta_1 X + E$, we have computed estimates $\hat{\beta}_0$, $\hat{\beta}_1$ for the unknown parameters β_0, β_1, based on the observed values $(x_1, y_1), \ldots, (x_n, y_n)$.

- We obtained a "best-fit line" model $Y = \hat{\beta}_0 + \hat{\beta}_1 X$, where $\hat{\beta}_0 = \hat{\beta}_0(y_1, \ldots, y_n)$ and $\hat{\beta}_1 = \hat{\beta}_1(y_1, \ldots, y_n)$ are real numbers.
  - This model is called a simple linear regression model.
- Simple linear regression model as a **prediction model**.
  - Given **any** "new" value $X = x$, our simple linear regression model gives the prediction

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

Intuition: Once we have obtained sufficiently many observed values $(x_1, y_1), \ldots, (x_n, y_n)$ to get a "good" simple linear regression model, we could then use this model to predict the expected value of Y for any given value $X = x$.

- **Question:** How do we determine how "good" our simple linear regression model is?

Mean squared error of prediction

Given any prediction model that gives a prediction $\hat{y}$ for some random variable Y, the real number $\mathbf{E}[(Y - \hat{y})^2]$ is called the mean squared error (M.S.E.) of the prediction $\hat{y}$.

- i.e. a measurement of the "error" of $\hat{y}$ from the "true" value.

Theorem: In simple linear regression, the prediction $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$ (given $X = x$) has the **mean squared error**

$$\mathbf{E}[(Y - \hat{y})^2] = \sigma^2 \left[1 + \frac{1}{n} + \frac{(x - \overline{x}_n)}{s_{\mathbf{x}}^2}\right],$$

where $\overline{x}_n = \frac{x_1 + \cdots + x_n}{n}$ is the average of the given observed values $x_1, \ldots, x_n$ of the predictor variables, and $s_{\mathbf{x}} = \sqrt{\sum_{i=1}^{n} (x_i - \overline{x}_n)^2}$.

Interpretation:

- The M.S.E. becomes smaller and closer to σ^2 with a larger n (i.e. more sample points $(x_1, y_1), \ldots, (x_n, y_n)$).
- The M.S.E. is larger as the input value x becomes further away from $\overline{x}_n$. Hence, it is "harder" to predict the value of Y given a value of X that is further away from the values $x_1, \ldots, x_n$ used to compute the model.

Distribution of M.L.E. of σ^2

Recall: In a simple linear regression model $Y = \hat{\beta}_0 + \hat{\beta}_1 X$, we know $(\hat{\beta}_0, \hat{\beta}_1)$ has a bivariate normal distribution, and Y conditioned on $X = x$ is normal with mean $\beta_0 + \beta_1 x$ and variance σ^2.

- **Recall:** $\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2$.

Important Theorem: If $n \geq 3$, then $\hat{\sigma}^2$ is independent of $(\hat{\beta}_0, \hat{\beta}_1)$, and $\frac{n\hat{\sigma}^2}{\sigma^2}$ has the χ^2 distribution with $n-2$ degrees of freedom.

Interpretation:

- To estimate the value of the unknown parameter σ^2, we can use the estimator $\hat{\sigma}^2$ to compute an estimate $\hat{\sigma}^2(y_1, \ldots, y_n)$.
- Even though $\hat{\sigma}^2$ is computed in terms of the estimators $\hat{\beta}_0$ and $\hat{\beta}_1$, this theorem says that any estimate of $\hat{\sigma}^2$ is actually independent of the corresponding estimates of $\hat{\beta}_0$ and $\hat{\beta}_1$, i.e. it does not matter what the estimates of $\hat{\beta}_0$ and $\hat{\beta}_1$ are.

Hypothesis testing on regression coefficients

Recall: β_0 and β_1 are called regression coefficients.

- The M.L.E.'s of β_0 and β_1 are $\hat{\beta}_0$ and $\hat{\beta}_1$.
- The joint distribution of $(\hat{\beta}_0, \hat{\beta}_1)$ is bivariate normal.
  - So any (non-zero) linear combination of $\hat{\beta}_0$ and $\hat{\beta}_1$ is normal.

Theorem: Given any (non-zero) linear combination $c_0\beta_0 + c_1\beta_1$, its M.L.E. is $c_0\hat{\beta}_0 + c_1\hat{\beta}_1$, which has a normal distribution with mean $c_0\beta_0 + c_1\beta_1$ and variance $\sigma^2\left(\frac{c_0^2}{n} + \frac{(c_0\overline{x}_n - c_1)^2}{s_x^2}\right)$.

Note: Let c_0, c_1 be constants, not both zero. Since $c_0\beta_0 + c_1\beta_1$ can be estimated by its M.L.E. $c_0\hat{\beta}_0 + c_1\hat{\beta}_1$, which is normal, we could perform hypothesis testing on $c_0\beta_0 + c_1\beta_1$.

- Given a fixed $c \in \mathbb{R}$, we could consider a hypothesis test with the hypotheses:
  - $H_0 : c_0\beta_0 + c_1\beta_1 = c$ and $H_1 : c_0\beta_0 + c_1\beta_1 \neq c$;

  or the hypotheses
  - $H_0 : c_0\beta_0 + c_1\beta_1 \leq c$ and $H_1 : c_0\beta_0 + c_1\beta_1 > c$;

  or the hypotheses
  - $H_0 : c_0\beta_0 + c_1\beta_1 \geq c$ and $H_1 : c_0\beta_0 + c_1\beta_1 < c$.

Regression coefficients and t-distributions

Note: $Y_1, \ldots, Y_n$ are observable R.V.'s; $x_1, \ldots, x_n$ are given values.

- Let $\overline{x}_n = \frac{x_1 + \cdots + x_n}{n}$, and let $\overline{Y}_n = \frac{Y_1 + \cdots + Y_n}{n}$.
- Let $\hat{\beta}_0$ and $\hat{\beta}_1$ be the M.L.E.'s of β_0 and β_1 respectively.
- Let $s_{\mathbf{x}} = \sqrt{\sum_{i=1}^n (x_i - \overline{x}_n)^2}$; $\sigma' = \sqrt{\frac{1}{n-2} \sum_{i=1}^n (Y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2}$.
  - **Recall:** $\hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^n (Y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2$.
  - Hence, $(\sigma')^2 = \frac{n}{n-2} \hat{\sigma}^2$.

Important Theorem: Let c_0, c_1 be constants, not both zero, and suppose that $\mathbf{E}[c_0 \beta_0 + c_1 \beta_1] = c$. If $n \geq 3$, then the statistic

$$T = \left[\frac{c_0^2}{n} + \frac{(c_0 \overline{x}_n - c_1)^2}{s_{\mathbf{x}}^2} \right]^{-0.5} \left(\frac{c_0 \hat{\beta}_0 + c_1 \hat{\beta}_1 - c}{\sigma'} \right)$$

has the t-distribution with $(n-2)$ degrees of freedom.

- **Note:** T is a statistic of $\{Y_1, \ldots, Y_n\}$.
  - In particular, the values $x_1, \ldots, x_n, c_0, c_1, c$ are given.

t-tests for linear combinations of regression coefficients

Same as before: $Y_1, \ldots, Y_n$ are observable R.V.'s; $x_1, \ldots, x_n$ are given values. Let $\overline{x}_n = \frac{x_1 + \cdots + x_n}{n}$.

- Let $\hat{\beta}_0$ and $\hat{\beta}_1$ be the M.L.E.'s of β_0 and β_1 respectively.
- Let $s_{\mathbf{x}} = \sqrt{\sum_{i=1}^n (x_i - \overline{x}_n)^2}$; $\sigma' = \sqrt{\frac{1}{n-2} \sum_{i=1}^n (Y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2}$.

Important consequences of theorem on the previous slide: Assume $n \geq 3$. Let c_0, c_1, c be constants such that c_0, c_1 are not both zero, and such that $\mathbf{E}[c_0\beta_0 + c_1\beta_1] = c$. Define the statistic

$$T = \left[\frac{c_0^2}{n} + \frac{(c_0\overline{x}_n - c_1)^2}{s_{\mathbf{x}}^2}\right]^{-0.5} \left(\frac{c_0\hat{\beta}_0 + c_1\hat{\beta}_1 - c}{\sigma'}\right).$$

- If $\mathcal{H}$ is a hypothesis test with $H_0 : c_0\beta_0 + c_1\beta_1 = c$, test statistic T, and rejection region $[k, \infty)$, then $\mathcal{H}$ is a t-test.
- If $\mathcal{H}$ is a hypothesis test with $H_0 : c_0\beta_0 + c_1\beta_1 \leq c$, test statistic T, and rejection region $[k, \infty)$, then $\mathcal{H}$ is a t-test.
- If $\mathcal{H}$ is a hypothesis test with $H_0 : c_0\beta_0 + c_1\beta_1 \geq c$, test statistic $|T|$, and rejection region $(-\infty, k]$, then $\mathcal{H}$ is a t-test.

Special Case: t-test for regression coefficient β_0

Assume $n \geq 3$. For each $c \in \mathbb{R}$, define the statistic

$$T_c = \left[\frac{1}{n} + \frac{\overline{x}_n^2}{s_{\mathbf{x}}^2}\right]^{-0.5} \left(\frac{\hat{\beta}_0 - c}{\sigma'}\right) = \frac{\hat{\beta}_0 - c}{\sigma' \sqrt{\frac{1}{n} + \frac{\overline{x}_n^2}{s_{\mathbf{x}}^2}}},$$

where:

- $\hat{\beta}_0$ and $\hat{\beta}_1$ are the M.L.E.'s of β_0 and β_1 respectively;
- $s_{\mathbf{x}} = \sqrt{\sum_{i=1}^n (x_i - \overline{x}_n)^2}$; $\sigma' = \sqrt{\frac{1}{n-2} \sum_{i=1}^n (Y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2}$.

[Here, we substitute $c_0 = 1$, $c_1 = 0$ into the expression for T on slide 16.]

Fact: The hypothesis test $\mathcal{H}$ with null hypothesis $H_0 : \beta_0 = c$ and test statistic T_c given above is a t-test, since if H_0 is true, then T_c has the t-distribution with $(n-2)$ degrees of freedom.

- Similar statements hold if $\mathcal{H}$ has null hypothesis $H_0 : \beta_0 \leq c$ or $H_0 : \beta_0 \geq c$.
- **Note:** The special subcase $H_0 : \beta_0 = 0$ is a hypothesis test that checks if the regression line $Y = \beta_0 + \beta_1 X$ passes through the origin.

Special Case: t-test for β_0 (continued)

Assume $n \geq 3$. For each $c \in \mathbb{R}$, define the statistic

$$T_c = \left[\frac{1}{n} + \frac{\overline{x}_n^2}{s_{\mathbf{x}}^2}\right]^{-0.5} \left(\frac{\hat{\beta}_0 - c}{\sigma'}\right) = \frac{\hat{\beta}_0 - c}{\sigma' \sqrt{\frac{1}{n} + \frac{\overline{x}_n^2}{s_{\mathbf{x}}^2}}}.$$

Theorem: Suppose that k_0 is the $100(1 - \frac{\alpha_0}{2})$-percentile of the t-distribution with $n-2$ degrees of freedom. If $\mathcal{H}$ is the t-test with null hypothesis $H_0 : \beta_0 = c$, test statistic T_c, and rejection region $[k, \infty)$, then $\mathcal{H}$ has significance level α_0 if and only if $k \geq k_0$.

Theorem: Suppose that k_0 is the $100(1 - \alpha_0)$-percentile of the t-distribution with $n-2$ degrees of freedom.

- If $\mathcal{H}$ is the t-test with null hypothesis $H_0 : \beta_0 \leq c$, test statistic T_c, and rejection region $[k, \infty)$, then $\mathcal{H}$ has significance level α_0 if and only if $k \geq k_0$.
- If $\mathcal{H}$ is the t-test with null hypothesis $H_0 : \beta_0 \geq c$, test statistic T_c, and rejection region $(-\infty, k]$, then $\mathcal{H}$ has significance level α_0 if and only if $k \leq k_0$.

Special Case: t-test for regression coefficient β_1

Assume $n \geq 3$. For each $c \in \mathbb{R}$, define the statistic

$$T_c = \left[\frac{1}{s_{\mathbf{x}}^2}\right]^{-0.5}\left(\frac{\hat{\beta}_1 - c}{\sigma'}\right) = \frac{s_{\mathbf{x}}(\hat{\beta}_1 - c)}{\sigma'},$$

where:

- $\hat{\beta}_0$ and $\hat{\beta}_1$ are the M.L.E.'s of β_0 and β_1 respectively;
- $s_{\mathbf{x}} = \sqrt{\sum_{i=1}^{n}(x_i - \overline{x}_n)^2}$; $\sigma' = \sqrt{\frac{1}{n-2}\sum_{i=1}^{n}(Y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2}$.

[Here, we substitute $c_0 = 0$, $c_1 = 1$ into the expression for T on slide 16.]

Fact: The hypothesis test $\mathcal{H}$ with null hypothesis $H_0 : \beta_1 = c$ and test statistic T_c given above is a t-test, since if H_0 is true, then T_c has the t-distribution with $(n-2)$ degrees of freedom.

- Similar statements hold if $\mathcal{H}$ has null hypothesis $H_0 : \beta_1 \leq c$ or $H_0 : \beta_1 \geq c$.
- **Note:** The special subcase $H_0 : \beta_1 = 0$ is a hypothesis test that checks if Y is related to X (under the assumption that $Y = \beta_0 + \beta_1 + E$ for some parameters β_0, β_1).

Special Case: t-test for β_1 (continued)

Assume $n \geq 3$. For each $c \in \mathbb{R}$, define the statistic

$$T_c = \left[\frac{1}{s_x^2}\right]^{-0.5}\left(\frac{\hat{\beta}_1 - c}{\sigma'}\right) = \frac{s_x(\hat{\beta}_1 - c)}{\sigma'}.$$

Theorem: Suppose that k_0 is the $100(1 - \frac{\alpha_0}{2})$-percentile of the t-distribution with $n-2$ degrees of freedom. If $\mathcal{H}$ is the t-test with null hypothesis $H_0 : \beta_1 = c$, test statistic T_c, and rejection region $[k, \infty)$, then $\mathcal{H}$ has significance level α_0 if and only if $k \geq k_0$.

Theorem: Suppose that k_0 is the $100(1 - \alpha_0)$-percentile of the t-distribution with $n-2$ degrees of freedom.

- If $\mathcal{H}$ is the t-test with null hypothesis $H_0 : \beta_1 \leq c$, test statistic T_c, and rejection region $[k, \infty)$, then $\mathcal{H}$ has significance level α_0 if and only if $k \geq k_0$.
- If $\mathcal{H}$ is the t-test with null hypothesis $H_0 : \beta_1 \geq c$, test statistic T_c, and rejection region $(-\infty, k]$, then $\mathcal{H}$ has significance level α_0 if and only if $k \leq k_0$.

Special Case: Testing if given point lies on regression line

Assume $n \geq 3$. Given a point $(x_0, y_0) \in \mathbb{R}^2$, define the statistic

$$T = \left[\frac{1}{n} + \frac{(\overline{x}_n - x_0)^2}{s_{\mathbf{x}}^2}\right]^{-0.5} \left(\frac{\hat{\beta}_0 + x_0\hat{\beta}_1 - y_0}{\sigma'}\right),$$

where:

- $\hat{\beta}_0$ and $\hat{\beta}_1$ are the M.L.E.'s of β_0 and β_1 respectively;
- $s_{\mathbf{x}} = \sqrt{\sum_{i=1}^n (x_i - \overline{x}_n)^2}$; $\sigma' = \sqrt{\frac{1}{n-2}\sum_{i=1}^n (Y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2}$.

[Here, we substitute $c_0 = 1$, $c_1 = x_0$ into the expression for T on slide 16.]

[We also substitute $c = y_0$.]

Fact: The hypothesis test $\mathcal{H}$ with $H_0 : \beta_0 + \beta_1 x_0 = y_0$ and test statistic T given above is a t-test, since if H_0 is true, then T has the t-distribution with $(n-2)$ degrees of freedom.

- **Interpretation:** $H_0 : \beta_0 + \beta_1 x_0 = y_0$ is the hypothesis that the point (x_0, y_0) lies on the regression line $Y = \beta_0 + \beta_1 X$. Hence $\mathcal{H}$ is a t-test to decide whether to reject or not reject the hypothesis that (x_0, y_0) lies on the regression line.

Multiple Linear Regression

So far, we have dealt with simple linear regression.

- The statistical model used is "simple" because the response variable depends only on **one** predictor variable.

More generally, we could consider statistical models where a response variable depends on **multiple** predictor variables.

- The corresponding statistical inference procedure is called multiple linear regression, or more simply, linear regression.

Model: $Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k + E$.

- $\beta_0, \beta_1, \ldots, \beta_k$ are called regression coefficients.
- **Assumption:** E is normal with mean 0 and variance σ^2
- **Sample points:** Let $\mathbf{z}_1, \ldots, \mathbf{z}_n \in \mathbb{R}^k$ be vectors sampled from the joint distribution of $(X_1, \ldots, X_k)$.
  - For each $\mathbf{z}_i = (z_{i,1}, \ldots, z_{i,k}) \in \mathbb{R}^k$, let $y_i \in \mathbb{R}$ be the observed value of Y (given $X_1 = z_{i,1}, \ldots, X_k = z_{i,k}$).
- **Assumption:** $Y_1, \ldots, Y_n$ are conditionally independent given $\mathbf{z}_1, \ldots, \mathbf{z}_n$, which are values already known beforehand.

Multiple Linear Regression (continued)

Note: $\mathbf{E}[Y|X_1 = x_1, \ldots, X_k = x_k] = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k$, which is a linear k-variate function in terms of $x_1, \ldots, x_k$.

- This explains why linear regression is called "linear".

When $\beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k$ is treated as a function of $x_1, \ldots, x_k$, then this function is called the regression function of Y on $X_1, \ldots, X_k$, or more simply, the regression of Y on $X_1, \ldots, X_k$.

- **Intuition:** Assuming that Y is $\approx$ linear w.r.t. $X_1, \ldots, X_k$, i.e. $Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k +$ "noise", the idea of linear regression is that we can determine estimates for $\beta_0, \ldots, \beta_k$, based on some sample pairs $(\mathbf{z}_1, y_1), \ldots, (\mathbf{z}_n, y_n)$.

Important Note: Linear regression makes sense only when our assumptions are satisfied.

- Similar to the simple case, the **maximum likelihood estimators** $\hat{\beta}_0, \ldots, \hat{\beta}_k$ for $\beta_0, \ldots, \beta_k$ respectively have the same expressions as the coefficients computed in the **least square method** (for the least squares hyperplane).

Summary

- Simple linear regression
- Least squares estimators
- Joint distribution of least squares estimators
- Hypothesis testing on regression coefficients
- Multiple linear regression

Note: In this week's cohort class, we shall work through **several examples for linear regression**.

Reminders:

The **Final Exam** will be held on 3rd May (Friday), 9–11am, at the **Indoor Sports Hall 2** (61.106).

- Tested on all materials covered in this course
  - Lectures 1–24 and Cohort classes weeks 1–13.
- 1 piece of A4-sized double-sided **handwritten** cheat sheet is allowed for the final exam.
  - A formula sheet similar to that given in the mid-term exam will also be provided. Details of this formula sheet will be announced soon.
- Tomorrow's lecture (Lecture 24) will be a review lecture.