476 Statistics, Spring 2022.
Lecturer: Sergey Nadtochiy.

Lecture 11. LS in 1 dim: inferences about Y, testing hypotheses on β, correlation and R^2, nonlinear extensions. (Sections 11.6–11.9)

1 Inferences about Y

Q 1. *How to make inferences about Y?*

Given $X = x$, we have

$$Y = \beta_0 + \beta_1 x + \varepsilon.$$

A natural estimator of Y, then, is

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

But it is not really clear what it means to estimate a random variable...

Def 1. *A predicted value of Y, given $X = x$, is $\mathbb{E}(Y|X = x) = \beta_0 + \beta_1 x$.*

A natural estimator for $\theta := \mathbb{E}(Y|X = x)$ is

$$\hat{\beta}_0 + \hat{\beta}_1 x$$

To find confidence intervals for θ (and to test hypotheses on θ), we need to recall the distributional properties of linear functions of $\hat{\beta}$ (see Lecture 10):

- If If $\{X_i = x_i\}$ are deterministic, then

$$V(a\hat{\beta}_0 + b\hat{\beta}_1) = \sigma^2 \sigma_{ab}^2, \quad \sigma_{ab}^2 = \frac{\frac{a^2}{n}\sum_{i=1}^n x_i^2 + b^2 - 2ab\bar{x}}{S_{xx}}.$$

- If$\{\varepsilon_i\}$ are normal, then

$$\frac{a\hat{\beta}_0 + b\hat{\beta}_1 - (a\beta_0 + b\beta_1)}{\sigma\sigma_{ab}} \sim N(0,1),$$

$$\frac{a\hat{\beta}_0 + b\hat{\beta}_1 - (a\beta_0 + b\beta_1)}{\tilde{S}\sigma_{ab}} \sim T(n-2).$$

Ex 1. *Assume a 1-dimensional linear regression model and consider the following observations:*

$$\{x_i\} : -2, -1, 0, 1, 2$$

$$\{y_i\} : 0, 0, 1, 1, 3$$

Q 2. *Assuming that $\{X_i = x_i\}$ are deterministic, estimate the predicted value of Y_4 and put a 90%-confidence interval on it.*

$$\theta = \mathbb{E}(Y|X=1) = \beta_0 + \beta_1.$$

Recall

$$\hat{\beta}_0 = 1, \quad \hat{\beta}_1 = 0.7, \quad \tilde{s} = \sqrt{0.367} \approx 0.606, \quad S_{xx} = 10.$$

Then, the estimated predicted value of Y, given $x = 1$, is

$$\hat{\beta}_0 + \hat{\beta}_1 = 1 + 0.7 = 1.7.$$

To construct a confidence interval, we use the pivot

$$G = \frac{\hat{\beta}_0 + \hat{\beta}_1 - (\beta_0 + \beta_1)}{\tilde{S}\sigma_{11}} \sim T(3),$$

$$\sigma_{11}^2 = \frac{\frac{1}{5}\sum_{i=1}^5 x_i^2 + 1 - 2\bar{x}}{S_{xx}} = \frac{\frac{1}{5}10 + 1}{10} = 0.3.$$

The resulting confidence interval is

$$[\hat{\beta}_0 + \hat{\beta}_1 \,\pm\, t_{0.05}\,\tilde{s}\,\sigma_{11}] \approx [1.7 \pm 2.353 \cdot 0.606\sqrt{0.3}] \approx [1.7 \pm 0.781].$$

2 Testing hypotheses on β

The most common test used when fitting a linear regression model, which is interpreted as a check on the **existence of predictive power** of X for Y, is

$$H_0 = \{\beta_1 = 0\}, \quad H_a = \{\beta_1 \neq 0\}.$$

More generally, one may consider

$$H_0 = \{\beta_j = \beta_j^0\}, \quad H_a = \{\beta_j \neq \beta_j^0\}.$$

The results of Lecture 10 (also stated in the previous section) imply that, under H_0:

$$U = \frac{\hat{\beta}_j - \beta_j^0}{\sigma\sqrt{c_j}} \sim N(0,1), \quad j = 0, 1, \quad c_0 := \frac{\sum_{i=1}^n X_i^2}{nS_{xx}}, \quad c_1 := \frac{1}{S_{xx}}.$$

We can use the above as a test statistic, with $RR = (-\infty, -z_{\alpha/2}] \cup [z_{\alpha/2}, \infty)$, if σ is known.

When σ is not known, we use the fact that, under H_0,

$$U = \frac{\hat{\beta}_j - \beta_j^0}{\tilde{S}\sqrt{c_j}} \sim T(n-2), \quad j = 0, 1,$$

as a test statistic, with $RR = (-\infty, -t_\alpha] \cup [t_\alpha, \infty)$.

Ex 2. *Assume a 1-dimensional linear regression model and consider the following observations:*

$$\{x_i\} : -2, -1, 0, 1, 2$$

$$\{y_i\} : 0, 0, 1, 1, 3$$

Q 3. *Does this data present sufficient evidence to argue that $\beta_1 \neq 0$, at level 0.05? Compute the p-value.*

First, we compute:

$$\hat{\beta}_1 = 0.7, \quad c_1 = \frac{1}{S_{xx}} = 1/10 = 0.1.$$

Then, we choose the test statistic

$$U = \frac{\hat{\beta}_1}{\tilde{S}\sqrt{c_1}} \sim T(3),$$

whose value is

$$u = \frac{\hat{\beta}_1}{\tilde{s}\sqrt{0.1}} \approx \frac{0.7}{\sqrt{0.367}\sqrt{0.1}} \approx 3.65.$$

The rejection region is

$$RR = (-\infty, -t_{\alpha/2}] \cup [t_{\alpha/2}, \infty) \approx (-\infty, -3.182] \cup [3.182, \infty).$$

Thus, we reject the null hypothesis.

The p-value is

$$p = 2\min(F_U(3.65), 1 - F_U(3.65)) \in (0.02, 0.05).$$

3 Correlation and R^2

By testing the hypothesis $H_0 = \{\beta_1 = 0\}$ vs. $H_a = \{\beta_1 \neq 0\}$ we can answer the question: **does X have any predictive power for Y**? The main theme of this section is the following question.

Q 4. *How to estimate* **how much of predictive power X has for Y***?*

The amount of information about Y released by X can be measured by the **fraction of variance of Y explained by X**.

Assuming $\{\varepsilon_i\}$ and $\{X_i\}$ are i.i.d. and independent of each other, and denoting $\sigma_y^2 := V(Y)$, $\sigma_x^2 = V(X)$, we notice:

$$\sigma_y^2 = \beta_1^2\sigma_x^2 + \sigma^2.$$

Thus,

$$\frac{\beta_1^2\sigma_x^2}{\sigma_y^2}$$

is a natural measure of predictive power of X on Y.

By a routine computation, one can verify that

$$\beta_1 = \frac{\sigma_y}{\sigma_x}\rho,$$

where $\rho := \text{cor}(X, Y)$ is the correlation between X and Y. Hence, the squared correlation

$$\rho^2 = \frac{\beta_1^2\sigma_x^2}{\sigma_y^2}$$

is the **measure of predictive power** of X on Y.

Thm 1. *If $\{X_i\}$ and $\{\varepsilon_i\}$ are i.i.d. normal, then, the MLE for ρ is*

$$R := \frac{S_{xy}}{\sqrt{S_{xx}}\sqrt{S_{yy}}} = \frac{\sum(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum(X_i - \bar{X})^2 \sum(Y_i - \bar{Y})^2}}.$$

In particular, R is consistent.

Exercise 1. *Prove the above theorem.*

Thm 2. *If $\{\varepsilon_i\}$ are i.i.d., then R is an unbiased estimator of ρ. If, in addition, $\{X_i\}$ are i.i.d., then, R is a consistent estimator of ρ.*

Exercise 2. *Prove the above theorem.*

Since ρ^2 is the measure of fit quality, it is natural to **use "R-squared", R^2, as an estimator for the predictive power of a linear regression model**.

A routine computation shows:

$$R^2 = \hat{\beta}_1^2 \frac{S_{xx}}{S_{yy}}, \quad R^2 = 1 - \frac{(n-2)\tilde{S}^2}{S_{yy}} = 1 - \frac{\sum_{i=1}^n (Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i)^2}{\sum_{i=1}^n (Y_i - \bar{Y})^2}.$$

In view of the last expression above (which shows that R^2 is the fraction of variation of $\{Y_i\}$ explained by $\{X_i\}$), R^2 is often also interpreted as a measure of "fit quality" of a linear regression model.

Rem 1. *Note that R^2 does not completely replace the p-value of a test for $H_0 = \{\beta_1 = 0\}$ vs. $H_a = \{\beta_1 \neq 0\}$. The latter provides an answer to a less ambitious question, but it contains information about the accuracy of the answer. Since R^2 is just a point estimator, its value does not contain information about its precision.*

It turns out that we can use R to compute the test statistic for $H_0 = \{\rho = 0\} = \{\beta_1 = 0\}$. Recall the result from Lecture 10: whenever $\{\varepsilon_i\}$ are i.i.d. normal and independent of $\{X_i\}$, under H_0, we have

$$U = \frac{\hat{\beta}_1}{\tilde{S}/\sqrt{S_{xx}}} \sim T(n-2).$$

Using

$$R = \hat{\beta}_1 \sqrt{S_{xx}/S_{yy}},$$

we can deduce

$$\frac{R\sqrt{n-2}}{\sqrt{1-R^2}} = U \sim T(n-2),$$

under H_0, whenever $\{\varepsilon_i\}$ **are normal**. This allows us to test hyp. $H_0 = \{\rho = 0\} = \{\beta_1 = 0\}$ via R.

Ex 3. *Test the existence of non-zero correlation between the students' test scores and final grades (at level 0.05).*

$$\{x_j\} : 39, 43, 21, 64, 57, 47, 28, 75, 34, 52,$$

$$\{y_j\} : 65, 78, 52, 82, 92, 89, 73, 98, 56, 75.$$

$$s_{xx} = 2474, \quad s_{yy} = 2056, \quad s_{xy} = 1894$$

$$r = \frac{s_{xy}}{\sqrt{s_{xx}s_{yy}}} = \frac{1894}{\sqrt{2474 \cdot 2056}} \approx 0.8398, \quad r^2 \approx 0.7053,$$

$$U = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} = \frac{0.8398\sqrt{8}}{\sqrt{1-0.7053}} \approx 4.375,$$

$$t_{0.025} \approx 2.306$$

We reject the null hyp. of zero correlation.

$$p \approx 0.01.$$

The following theorem is useful for testing more general hypothesis $H_0 = \{\rho = \rho_0\}$ and for constructing asymptotic conf. int-ls.

Thm 3. *Assume that $\{\varepsilon_i\}$ and $\{X_i\}$ are i.i.d. and independent of each other. Then, under additional technical assumptions, we have:*

$$V = \frac{\frac{1}{2}\log\frac{1+R}{1-R} - \frac{1}{2}\log\frac{1+\rho}{1-\rho}}{1/\sqrt{n-3}} \to N(0,1),$$

as $n \to \infty$.

4 Nonlinear extensions

In some cases, it does not make sense to fit a linear function to explain Y via X. This could be due to the nature of data (e.g. if Y must be positive) or could be deduced from visual representation. Then, we may be able to guess the type of nonlinear function and linearize the problem.

For example,

$$Y \approx \alpha_0 X^{\beta_1}$$

can be equivalently rewritten as

$$\log Y \approx \log\alpha_0 + \beta_1 \log X.$$

Ex 4. *(Table 11.5) We approximate the weight W (in lb) of a crocodile as a function of its length L (in ft). Since both are positive (and weight is roughly proportional to a cube of length), it makes sense (although must be checked against visualized data) to fit:*

$$\log W = \log\alpha_0 (=: \beta_0) + \beta_1 \log L + \varepsilon.$$

Sample of size $n = 15$ gives:

$$\{x_j = \log l_j\} : 3.87, 3.61, 4.33, 3.43, \ldots, 3.78,$$

$$\{y_j = \log w_j\} : 4.87, 3.93, 6.46, 3.33, \ldots, 4.25$$

Q 5. *Compute the LS estimator of (β_0, β_1).*

$$s_{xx} = 0.8548, \quad s_{yy} = 10.26, \quad s_{xy} = 2.933,$$

$$\hat\beta_1 = \frac{s_{xy}}{s_{xx}} \approx 3.4312, \quad \hat\beta_0 = \bar y - \hat\beta_1 \bar x \approx -8.476,$$

$$\hat\alpha_0 = e^{\hat\beta_0} = e^{-8.476} \approx 0.0002$$

Q 6. *Estimate the predicted value of W given $\log L = 4$ (assume normal residuals).*

The quantity we need to estimate (i.e., the predicted value of Y) is

$$\theta = \mathbb{E}(W \mid \log L = 4) = \mathbb{E}\exp(\beta_0 + 4\beta_1 + \varepsilon) = e^{\beta_0 + 4\beta_1}\mathbb{E}e^{\varepsilon}.$$

One may be tempted to estimate θ using the estimator $\hat\beta_0 + 4\hat\beta_1 =: \widehat{\log W}$ of

$$\mathbb{E}(\log W \mid \log L = 4) = \beta_0 + 4\beta_1,$$

and considering

$$\exp(\widehat{\log W}) = \exp(\hat\beta_0 + 4\hat\beta_1) \approx \exp(-8.476 + 4 \cdot 3.4312) \approx \exp(5.2488) \approx 190.3377$$

as an estimator of θ. This estimator is typically biased, but the main problem is that it is **not consistent***: if $\{X_i\}$ are i.i.d., as $n \to \infty$,*

$$\exp(\widehat{\log W}) = \exp(\hat\beta_0 + 4\hat\beta_1) \to e^{\beta_0 + 4\beta_1} \neq \theta := \mathbb{E}(W \mid \log L = 4) = e^{\beta_0 + 4\beta_1}\mathbb{E}e^{\varepsilon},$$

because $\mathbb{E}e^{\varepsilon} = e^{\sigma^2/2} \neq 1$.

A **better estimator** *of θ is*

$$\widehat{W} := \exp(\hat\beta_0 + 4\hat\beta_1)\mathbb{E}e^{\varepsilon} = \exp(\hat\beta_0 + 4\hat\beta_1)e^{\sigma^2/2},$$

if σ^2 is known. If not, σ^2 needs to be replaced by its estimator: $\tilde S^2$. The above may also be biased, but it is **consistent** *if $\{X_i\}$ are i.i.d.*

Q 7. *Construct a 90%-confdence interval for $\tilde\theta := \exp(\mathbb{E}(\log W \mid \log L = 4))$, assuming normal errors.*

Recall that $\hat\beta_0 + 4\hat\beta_1$ is a good estimator of

$$\log\tilde\theta = \mathbb{E}(\log W \mid \log L = 4) = \beta_0 + 4\beta_1.$$

Thus, we use $\hat\beta_0 + 4\hat\beta_1$ to construct a pivot:

$$\frac{\hat\beta_0 + 4\hat\beta_1 - \log\tilde\theta}{\tilde S\sqrt{1/n + (4 - \bar X)^2/S_{xx}}} \sim T(n-2),$$

$$\tilde s = 0.123, \quad \bar x = 3.758.$$

Using the above pivot, we obtain the confidence interval for $\log\tilde\theta$:

$$\hat\beta_0 + 4\hat\beta_1 \pm t_{0.05}\,\tilde s\sqrt{1/n + (4 - \bar x)^2/s_{xx}} \approx (-8.476 + 4 \cdot 3.4312 \pm 1.771 \cdot 0.123\sqrt{1/15 + (4 - 3.758)^2/0.8548})$$

$$\approx (-8.476 + 4 \cdot 3.4312 \pm 1.771 \cdot 0.123 \cdot 0.3681) \approx (5.2488 \pm 0.08) \approx (5.1688, 5.3288).$$

To obtain a confidence interval for $\tilde\theta = \exp(\mathbb{E}(\log W \mid \log L = 4))$, we compute the exponential of the above interval.