MAT 2377 (Summer 2009)
Simple Linear Regression (Inference)
Sections 11.4.1, 11.5, 11.6, 11.8

§11-4: Hypothesis tests in simple linear regression

§11-4.1: t-test concerning β_0 or β_1.

The **Simple Linear Regression Model** with **normal random errors** is

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

where β_0 and β_1 are unknown constants, x is a value taken by the predictor X and ϵ is **random error**.

What is new? We will assume that ϵ is a **normal** random variable with mean 0 and variance σ^2 ($\epsilon \sim N(0, \sigma^2)$). That is, we assume:

$$E(\epsilon) = 0 \quad \text{and} \quad V(\epsilon) = \sigma^2$$

Consequences (see 5-5):

- Y is follows $N(\beta_0 + \beta_1 x, \sigma^2)$ distribution.
- The estimators $\widehat{\beta}_0$ and $\widehat{\beta}_1$ are linear combinations of $Y_1, \ldots, Y_n$, that is linear combinations of independent normals (recall last lecture formulae on page 9), thus they are both normal random variables. That is,

$$\widehat{\beta}_1 \sim N(\beta_1, \sigma^2_{\widehat{\beta}_1}) \quad \text{and} \quad \widehat{\beta}_0 \sim N(\beta_0, \sigma^2_{\widehat{\beta}_0}).$$

  Recall that $\sigma_{\widehat{\beta_1}} = \sqrt{\frac{\sigma^2}{S_{xx}}}$ and $E(\widehat{\beta}_1) = \beta_1$, AND

  $\sigma_{\widehat{\beta_0}} = \sqrt{\sigma^2 \left[\frac{1}{n} + \frac{\overline{x}^2}{S_{xx}}\right]}$ and $E(\widehat{\beta}_0) = \beta_0$.

- The standardized estimators are standard normals, i.e.

$$\frac{\widehat{\beta}_0 - \beta_0}{\sigma_{\widehat{\beta}_0}} \sim N(0,1) \quad \text{and} \quad \frac{\widehat{\beta}_1 - \beta_1}{\sigma_{\widehat{\beta}_1}} \sim N(0,1).$$

- As we replace the standard errors (see definition on page 10—last lecture: the standard deviation of the estimator) by the estimated standard errors we get t random variables with $\nu = n-2$ degrees of freedom, that is

$$\frac{\widehat{\beta}_0 - \beta_0}{\widehat{\sigma}_{\widehat{\beta}_0}} \sim t(n-2) \quad \text{and} \quad \frac{\widehat{\beta}_1 - \beta_1}{\widehat{\sigma}_{\widehat{\beta}_1}} \sim t(n-2),$$

where

$$\widehat{\sigma}_{\widehat{\beta}_1} = \sqrt{\frac{\hat{\sigma}^2}{S_{xx}}} \quad , \quad \widehat{\sigma}_{\widehat{\beta}_0} = \sqrt{\hat{\sigma}^2 \left[\frac{1}{n} + \frac{\overline{x}^2}{S_{xx}}\right]} \quad \text{and} \quad \hat{\sigma}^2 = \frac{SS_E}{n-2}$$

- Recall that $\hat{\sigma}^2 = \frac{SS_E}{n-2} = \frac{S_{yy} - \widehat{\beta}_1 S_{xy}}{n-2}$.

t-tests concerning β_0: Suppose that we have a hypothesis test with the following null hypothesis $H_0 : \beta_0 = \beta_{0,0}$ where $\beta_{0,0}$ is some real number. We will use the following test statistic

$$T_0 = \frac{\widehat{\beta}_0 - \beta_{0,0}}{\widehat{\sigma}_{\widehat{\beta}_0}} = \frac{\widehat{\beta}_0 - \beta_{0,0}}{\sqrt{\widehat{\sigma}^2 \left[\frac{1}{n} + \frac{\overline{x}^2}{S_{xx}}\right]}}.$$

where T_0 follows a t distribution with $\nu = n-2$ degrees of freedom when H_0 is true.

t-tests concerning β_1: Suppose that we have a hypothesis test with the following null hypothesis $H_0 : \beta_1 = \beta_{1,0}$ where $\beta_{1,0}$ is some real number. We will use the following test statistic

$$T_0 = \frac{\widehat{\beta}_1 - \beta_{1,0}}{\widehat{\sigma}_{\widehat{\beta}_1}} = \frac{\widehat{\beta}_1 - \beta_{1,0}}{\sqrt{\widehat{\sigma}^2 / S_{xx}}}.$$

where T_0 follows a t distribution with $\nu = n-2$ degrees of freedom when H_0 is true.

Test For the Significance of the Regression: The following test allows us to test the significance of the predictor X in predicting the response Y.

$$H_0 : \beta_1 = 0 \quad \text{against} \quad H_1 : \beta_1 \neq 0.$$

Interpretation:

- Failure to reject H_0 means that there is no linear relationship between Y and X.
- Rejecting H_0 means the linear relationship between Y and X is significant.

We are going to recall here the critical regions for the t tests: Let t_0 be the observed value of our test statistic T_0. Then

- Suppose that $H_0 : \beta_1 = \beta_{1,0}$ and $H_1 : \beta_1 \neq \beta_{1,0}$. Then we would reject $H_0 : \beta_1 = \beta_{1,0}$ IF $|t_0| > t_{\alpha/2,n-2}$ (just recall page 12 from Lecture June 16, 2009) Of course:

  [Right-Sided Alternative] $t_0 > t_{\alpha,n-2}$, where $H_1 : \beta_1 > \beta_{1,0}$,

  [Left-Sided Alternative] $t_0 < -t_{\alpha,n-2}$, where $H_1 : \beta_1 < \beta_{1,0}$;

- Suppose that $H_0 : \beta_0 = \beta_{0,0}$ and $H_1 : \beta_0 \neq \beta_{0,0}$. Then we would reject $H_0 : \beta_0 = \beta_{0,0}$ IF $|t_0| > t_{\alpha/2,n-2}$. Again:

  [Right-Sided Alternative] $t_0 > t_{\alpha,n-2}$, where $H_1 : \beta_0 > \beta_{0,0}$

  [Left-Sided Alternative] $t_0 < -t_{\alpha,n-2}$, where $H_1 : \beta_0 < \beta_{0,0}$

Example 1: Consider the data from Examples 1, 2 and 3 from Lecture June 23, 2009. Recall that the point estimates for β_0 and β_1 are respectively $\widehat{\beta_0} = 0.05$ and $\widehat{\beta_1} = 0.0039$. Furthermore, the estimated standard errors are $\widehat{\sigma}_{\widehat{\beta_0}} = 0.081453$ and $\widehat{\sigma}_{\widehat{\beta_1}} = 0.000370$.

(a) Test for the significance of the regression with $\alpha = 5\%$.

(b) Do the data support the claim that $\beta_0 > 0.1$ at a level of significance of 5%?

Sol: a) We have $\alpha = 0.05$, $H_0 : \beta_1 = 0$, $H_1 : \beta_1 \neq 0$, so $\beta_{1,0} = 0$ in this 2-sided hypothesis. Compute:

- $t_0 = \frac{\widehat{\beta_1} - \beta_{1,0}}{\widehat{\sigma}_{\widehat{\beta_1}}} = \frac{0.0039 - 0}{0.000370} = 10.54054054,$
- $t_{0.05/2,11-2} = t_{0.025,9} = 2.262$

Since $|t_0| = 10.5405 > 2.262 = t_{0.05/2,11-2}$ we reject $H_0 : \beta_1 = 0$, i.e., the linear relationship between Y and X is significant. (See again the picture in the previous lecture, see the dots, and see that indeed the slope of the best fitted line is NOT zero!)

b) We have $\alpha = 0.05$, $H_0 : \beta_0 = 0.1$, $H_1 : \beta_0 > 0.1$, so $\beta_{0,0} = 0.1$ in this right-sided hypothesis! Compute:

- $t_0 = \frac{\widehat{\beta_0} - \beta_{0,0}}{\sqrt{\widehat{\sigma}^2 \left[\frac{1}{n} + \frac{\overline{x}^2}{S_{xx}}\right]}} = \frac{0.05 - 0.1}{0.081453} = -0.613850,$
- $t_{\alpha,n-2} = t_{0.05,9} = 1.833.$

Since $t_0 = -0.613850$ is NOT greater than $t_{0.05,9} = 1.833$, we fail to reject $H_0 : \beta_0 = 0.1$. Anyway we got the estimation $\widehat{\beta_1} = 0.05$ which is close to 0.1.

§11.5: Interval Estimation

In other words, we want to get the confidence intervals! We known that

$$\frac{\widehat{\beta}_0 - \beta_0}{\widehat{\sigma}_{\widehat{\beta}_0}} \sim t(n-2) \quad \text{and} \quad \frac{\widehat{\beta}_1 - \beta_1}{\widehat{\sigma}_{\widehat{\beta}_1}} \sim t(n-2),$$

where the estimated standard errors are given by:

$$\widehat{\sigma}_{\widehat{\beta_1}} = \sqrt{\frac{\hat{\sigma}^2}{S_{xx}}} \quad \text{and} \quad \widehat{\sigma}_{\widehat{\beta_0}} = \sqrt{\hat{\sigma}^2 \left[\frac{1}{n} + \frac{\overline{x}^2}{S_{xx}}\right]}.$$

Hence a $(1-\alpha) \times 100\%$ **confidence interval for** β_0 is

$$\widehat{\beta}_0 \pm t_{\alpha/2,n-2}\, \widehat{\sigma}_{\widehat{\beta_0}} = \widehat{\beta}_0 \pm t_{\alpha/2,n-2} \sqrt{\hat{\sigma}^2 \left[\frac{1}{n} + \frac{\overline{x}^2}{S_{xx}}\right]},$$

and a $(1-\alpha) \times 100\%$ **confidence interval for** β_1 is

$$\widehat{\beta}_1 \pm t_{\alpha/2,n-2}\, \widehat{\sigma}_{\widehat{\beta_1}} = \widehat{\beta}_1 \pm t_{\alpha/2,n-2} \sqrt{\frac{\hat{\sigma}^2}{S_{xx}}}.$$

Example 2: Consider the data from Examples 1, 2 and 3 from Lecture June 23, 2009. Recall that the point estimates for β_0 and β_1 are respectively $\widehat{\beta}_0 = 0.05$ and $\widehat{\beta}_1 = 0.0039$. Furthermore, the estimated standard errors are $\widehat{\sigma}_{\widehat{\beta}_0} = 0.081453$ and $\widehat{\sigma}_{\widehat{\beta}_1} = 0.000370$.

(a) Construct a 95% confidence interval for β_0.

(b) Give a 95% confidence interval for β_1.

Sol: a) $\alpha = 0.05$ and the CI is

$$\left[\widehat{\beta}_0 - t_{\alpha/2,n-2}\sqrt{\hat{\sigma}^2\left[\frac{1}{n}+\frac{\overline{x}^2}{S_{xx}}\right]}, \widehat{\beta}_0 + t_{\alpha/2,n-2}\sqrt{\hat{\sigma}^2\left[\frac{1}{n}+\frac{\overline{x}^2}{S_{xx}}\right]}\right] =$$

$$\left[0.05 - t_{0.05/2,9} \times 0.081453, 0.05 + t_{0.05/2,9} \times 0.081453\right] =$$

$$\left[0.05 - 2.262 \times 0.081453, 0.05 + 2.262 \times 0.081453\right] = \left[-0.134246, 0.234246\right].$$

b) $\alpha = 0.05$ and the CI is:

$$\left[\widehat{\beta}_1 - t_{\alpha/2,n-2}\sqrt{\frac{\hat{\sigma}^2}{S_{xx}}}, \widehat{\beta}_1 + t_{\alpha/2,n-2}\sqrt{\frac{\hat{\sigma}^2}{S_{xx}}}\right] =$$

$$\left[0.0039 - 2.262 \times 0.000370, 0.0039 + 2.262 \times 0.000370\right] = [0.003063, 0.039837].$$

Compare this intervals to our $\widehat{\beta}_0$ and $\widehat{\beta}_1$.

Estimating the mean response

Given a specified value of the predictor X, say x_0, we would like to estimate the mean response, that is

$$\mu_{Y|x_0} = \beta_0 + \beta_1\, x_0.$$

We can use the value on the estimated regression line as a point estimate, i.e.

$$\widehat{\mu}_{Y|x_0} = \widehat{\beta}_0 + \widehat{\beta}_1\, x_0.$$

Properties of the estimated mean response:

- Its expectation is

$$E[\widehat{\mu}_{Y|x_0}] = \beta_0 + \beta_1\, x_0 = \mu_{Y|x_0}. \quad Why?$$

  Hence it is unbiased for estimating $\mu_{Y|x_0}$.

- Its variance is

$$V[\widehat{\mu}_{Y|x_0}] = \sigma^2 \left[\frac{1}{n} + \frac{(x_0 - \overline{x})^2}{S_{xx}}\right]$$

  — see page 9 of last lecture!

- Standardization with the estimated standard error:

$$\frac{\widehat{\mu}_{Y|x_0} - \mu_{Y|x_0}}{\sqrt{\widehat{\sigma}^2 \left[\frac{1}{n} + \frac{(x_0-\overline{x})^2}{S_{xx}}\right]}} \sim t(n-2).$$

Interval Estimation:

A $100\,(1-\alpha)\%$ confidence interval for the mean response at a value $x = x_0$, say $\mu_{Y|x}$, is

$$\widehat{\mu}_{Y|x_0} \pm t_{\alpha/2,n-2} \sqrt{\widehat{\sigma}^2 \left[\frac{1}{n} + \frac{(x_0 - \overline{x})^2}{S_{xx}}\right]}.$$

§Section 11.6: Prediction of new observations

Goal: To predict a new or future response Y_0 corresponding to a specified level x_0 of the predictor.

Prediction: We can use the following

$$\widehat{Y}_0 = \widehat{\beta}_0 + \widehat{\beta}_1 \, x_0$$

as a point estimator of the new or future value of the response Y_0.

Error in Prediction: We will define the error in prediction as

$$\mathbf{e} = Y_0 - \widehat{Y}_0.$$

The expectation of the error in prediction is

$$E[\mathbf{e}] = E[Y_0] - E[\widehat{Y}_0] = (\beta_0 + \beta_1 \, x_0) - (\beta_0 + \beta_1 \, x_0) = 0$$

and the variance of the error in prediction is

$$V[\mathbf{e}] = V[Y_0] + V[\widehat{Y}_0] = \sigma^2 + \sigma^2 \left[\frac{1}{n} + \frac{(x_0 - \overline{x})^2}{S_{xx}} \right] = \sigma^2 \left[1 + \frac{1}{n} + \frac{(x_0 - \overline{x})^2}{S_{xx}} \right],$$

since we are assuming that new or future value Y_0 is independent of the current observations $Y_1, \ldots, Y_n$.

If we use $\widehat{\sigma}^2$ to estimate σ^2, it can be shown that then we have:

$$\frac{Y_0 - \widehat{Y}_0}{\sqrt{\widehat{\sigma}^2 \left[1 + \frac{1}{n} + \frac{(x_0 - \overline{x})^2}{S_{xx}} \right]}} \sim t(n-2).$$

A $100\,(1-\alpha)\%$ **PREDICTION interval for new or future value response** Y_0 at the value x_0 is given by

$$\widehat{y}_0 \pm t_{\alpha/2, n-2} \sqrt{\widehat{\sigma}^2 \left[1 + \frac{1}{n} + \frac{(x_0 - \overline{x})^2}{S_{xx}} \right]}$$

where $\widehat{y}_0 = \widehat{\beta}_0 + \widehat{\beta}_1 x_0$ (recall here the estimated regression line $\widehat{y} = \widehat{\beta}_0 + \widehat{\beta}_1 x$).

Example 3: Consider the data from Examples 1, 2 and 3 of the previous lecture. Recall that the estimated regression line is

$$\widehat{y} = 0.05 + 0.0039\,x,$$

the point estimate for σ^2 is $\widehat{\sigma}^2 = 0.0231$. Furthermore, $n = 11$, $S_{xx} = 168363.64$ and $\overline{x} = 2000/11 = 181.818$.

(a) Give a 95% confidence interval for the mean evaporation coefficient at a velocity of $x_0 = 140$.

(b) Give a 95% prediction interval for a new or future evaporation coefficient at a velocity of $x_0 = 140$.

Sol: a) We have $\alpha = 0.05$, $x_0 = 140$, and the CI is:

$$\left[\widehat{\mu}_{Y|x_0} - t_{\alpha/2,n-2}\sqrt{\widehat{\sigma}^2\left[\frac{1}{n} + \frac{(x_0-\overline{x})^2}{S_{xx}}\right]}, \widehat{\mu}_{Y|x_0} + t_{\alpha/2,n-2}\sqrt{\widehat{\sigma}^2\left[\frac{1}{n} + \frac{(x_0-\overline{x})^2}{S_{xx}}\right]}\right] =$$

$$\Big[\widehat{\beta}_0 + \widehat{\beta}_1 x_0 - t_{0.025,9}\sqrt{0.0231\{\frac{1}{11} + \frac{(140-181.818)^2}{168363.64}\}},$$
$$\widehat{\beta}_0 + \widehat{\beta}_1 x_0 + t_{0.025,9}\sqrt{0.0231(\frac{1}{11} + \frac{(140-181.818)^2}{168363.64})}\Big]$$

$= [0.596 - 0.00529308, 0.596 + 0.00529308] = [0.59070692, 0.60129308]$ — a small interval around our estimate 0.596.

b) We have $\alpha = 0.05$, $x_0 = 140$, and the CI is:

$$\left[\widehat{y}_0 - t_{\alpha/2,n-2}\sqrt{\widehat{\sigma}^2\left[1 + \frac{1}{n} + \frac{(x_0-\overline{x})^2}{S_{xx}}\right]}, \widehat{y}_0 + t_{\alpha/2,n-2}\sqrt{\widehat{\sigma}^2\left[1 + \frac{1}{n} + \frac{(x_0-\overline{x})^2}{S_{xx}}\right]}\right]$$

$$= \left[\widehat{\beta}_0 + \widehat{\beta}_1 x_0 - t_{0.025,9}\sqrt{0.0231\left[1 + \frac{1}{11} + \frac{(x_0-\overline{x})^2}{S_{xx}}\right]},\right.$$
$$\left.\widehat{\beta}_0 + \widehat{\beta}_1 x_0 + t_{0.025,9}\sqrt{0.0231\left[1 + \frac{1}{11} + \frac{(x_0-\overline{x})^2}{S_{xx}}\right]}\right]$$

$= [0.596 - 0.360787, 0.596 + 0.360787] = [0.235213, 0.956787]$ — of course the second interval is larger: it is about a prediction!

§**Section 11-8: Correlation Analysis**

Scenario: We will assume that both X and Y are random variables. We would like to measure the linear association between the two random variables.

We could use the correlation coefficient

$$\rho = \frac{\sigma_{XY}}{\sigma_X \, \sigma_Y}$$

to measure the linear association between X and Y. In practice, the joint distribution of X and Y is unknown so we must estimate ρ.

Consider the following random sample $(X_1, Y_1), \ldots, (X_n, Y_n)$, we define the **sample correlation coefficient** as

$$R = \frac{\sum_{i=1}^n (X_i - \overline{X})\,(Y_i - \overline{Y})}{\sqrt{\sum_{i=1}^n (X_i - \overline{X})^2 \cdot \sum_{i=1}^n (Y_i - \overline{Y})^2}} =$$

$$= \frac{\sum_{i=1}^n (X_i - \overline{X})\,Y_i}{\sqrt{\sum_{i=1}^n (X_i - \overline{X})^2 \cdot \sum_{i=1}^n (Y_i - \overline{Y})^2}} = \frac{S_{XY}}{\sqrt{S_{XX}\,S_{YY}}}.$$

Remark: Recall that the slope of the estimated regression line is

$$\widehat{\beta}_1 = \frac{S_{XY}}{S_{XX}}.$$

Thus,

$$R = \frac{S_{XY}}{\sqrt{S_{XX}\,S_{YY}}} = \frac{S_{XY}}{S_{XX}}\sqrt{\frac{S_{XX}}{S_{YY}}} = \widehat{\beta}_1\sqrt{\frac{S_{XX}}{S_{YY}}}.$$

Hence R and $\widehat{\beta}_1$ are closely related.

Testing $\rho = 0$: Suppose that we would like to test

$$H_0 : \rho = 0 \quad \text{against} \quad H_1 : \rho \neq 0.$$

We will use the following test statistic

$$T_0 = \frac{R\sqrt{n-2}}{\sqrt{1-R^2}} = \frac{\widehat{\beta}_1 - 0}{\widehat{\sigma}_{\widehat{\beta}_1}} \sim t(n-2).$$

Critical regions (rule):

Since we are dealing with a t distribution with $n-2$ degrees of freedom, we would **reject the null hypothesis** $(H_0 : \rho = 0)$ if the observed value t_0 of the test statistic T_0 satisfies $|t_0| > t_{\alpha/2,n-2}$.

Example 4: Consider the data from Examples 1, 2 and 3 from previous lecture. Recall that

$$S_{xx} = 168363.64, \quad S_{xy} = 657.22, \quad S_{yy} = 2.7713.$$

Indeed, $S_{yy} = (\sum_{i=1}^{11} y_i^2) - n\overline{y}^2 = 9.1097 - \frac{8.35^2}{11} = 9.1097 - 6.3384 = 2.7713$.
(a) Compute the sample correlation coefficient between X and Y.

(b) Test $H_0 : \rho_{XY} = 0$ against $H_1 : \rho_{XY} \neq 0$ at $\alpha = 5\%$.

Sol: a) $R = \frac{S_{xy}}{\sqrt{S_{xx}S_{yy}}} = \frac{657.22}{\sqrt{168363.64 \times 2.7713}} = \frac{657.22}{683.071} \cong 0.962155 \in [-1, 1]$.

b) We compute $t_0 = \frac{R\sqrt{n-2}}{\sqrt{1-R^2}} = \frac{0.962155 \times \sqrt{9}}{\sqrt{1-0.925742}} = 10.592415$ and $t_{0.025,9} = 2.262$. Since $|t_0| > t_{0.025,9}$ we reject the null hypothesis $H_0 : \rho_{XY} = 0$. (Recall the picture in the previous lecture!)

DO 11-65/page 425 From the statement we got: $R = 0.75$, $n = 20$, $\alpha = 0.05$, $H_0 : \rho = 0$ and $H_1 : \rho > 0$ (right-sided hypothesis). We compute:

- $t_0 = \frac{0.75\sqrt{20-2}}{\sqrt{1-0.75^2}} = \frac{0.75\sqrt{18}}{\sqrt{0.4375}} = \frac{3.181980515}{0.661437827} = 4.81$,
- Since it is right-sided we have $t_{\alpha,n-2} = t_{0.05,18} = 1.734$ by table V.
- Compare: since $t_0 = 4.81 > 1.734 = t_{\alpha,n-2}$ (recall the critical regions!) we reject H_0 and accept H_1.