3Y03 - 3J04 PROBABILITY & STATISTICS FOR (CIVIL) ENGINEERING

C01 - Lecture 31

Last time LINEAR REGRESSION

Model linear relationship between X and Y using

$$Y = \beta_0 + \beta_1 X + \varepsilon$$

← random error with $E(\varepsilon) = 0$ $[V(\varepsilon) = \sigma^2]$

least squares estimates

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

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

Estimate σ^2 with

$$\hat{\sigma}^2 = \frac{1}{n-2} SS_E = \frac{1}{n-2}(SS_T - SS_R)$$

where $SS_T = \sum_{i=1}^n y_i^2 - n\bar{y}^2$ $= \sum_{i=1}^n (y_i - \bar{y})^2$ (see below), $SS_R = \hat{\beta}_1 S_{xy}$

11.3 Properties of Least Squares Estimators $(\hat{\beta}_0, \hat{\beta}_1)$.

For fixed x, $Y = \beta_0 + \beta_1 x + \varepsilon$ is a r.v. ← $E(\varepsilon) = 0$, $V(\varepsilon) = \sigma^2$

So $E(Y) = \beta_0 + \beta_1 x$

$V(Y) = \sigma^2$.

Recall $\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}} = \frac{\sum x_i \cancel{y_i}^{\,Y_i} - n\bar{x}\cancel{\bar{y}}\,\bar{Y}}{\sum x_i^2 - n\bar{x}^2}$

(Now, here, $\hat{\beta}_1$ is a function of r.v.s — an estimator for β_1, while in the last lecture it stood for the estimate for β_1 from the data)

$\sum = \sum_{i=1}^n$ everywhere

$$E(\hat{\beta}_1) = \frac{\sum x_i E(Y_i) - n\bar{x} E(\bar{Y})}{\sum x_i^2 - n\bar{x}^2}$$

$$= \frac{\sum x_i(\beta_0 + \beta_1 x_i) - n\bar{x}(\beta_0 + \beta_1 \bar{x})}{\sum x_i^2 - n\bar{x}^2}$$

$\bar{Y} = \frac{1}{n}\sum_{i=1}^n Y_i$

$E(\bar{Y}) = \frac{1}{n}\sum_{i=1}^n E(Y_i) = \frac{1}{n}\sum_{i=1}^n (\beta_0 + \beta_1 x_i)$ $= \frac{1}{n}\cdot n\beta_0 + \beta_1 \frac{1}{n}\sum_{i=1}^n x_i = \beta_0 + \beta_1 \bar{x}$.

$$= \frac{\beta_0 \left(\cancel{\sum x_i - n\bar{x}} \right) + \beta_1 \left(\cancel{\sum x_i^2 - n\bar{x}^2} \right)}{\cancel{\sum x_i^2 - n\bar{x}^2}}$$

$$= \beta_1$$

$\hat{\beta}_1$ is an unbiased estimator for β_1.

And $V(\hat{\beta}_1) = \underset{\text{Check!}}{\ldots}\ \sigma^2 / S_{xx}$ (similarly! Use 5.4!)

& $E(\hat{\beta}_0) = E(\bar{Y} - \hat{\beta}_1 \bar{x})$

(Just to remind us that today we are talking estimators — functions of r.v.s Y_i, not of data points y_i)

$$= E(\bar{Y}) - \bar{x} E(\hat{\beta}_1)$$

$$= \beta_0 + \cancel{\beta_1 \bar{x}} - \cancel{\bar{x} \beta_1} = \beta_0$$

so $\hat{\beta}_0$ is an unbiased estimator for β_0.

And $V(\hat{\beta}_0) = V(\bar{Y} - \hat{\beta}_1 \bar{x})$ (See 5.4!)

$$= V(\bar{Y}) + \bar{x}^2 V(\hat{\beta}_1)$$

$$= \frac{\sigma^2}{n} + \frac{\bar{x}^2 \sigma^2}{S_{xx}} = \sigma^2 \left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}} \right).$$

Both variances depend on σ^2, so we can estimate with

$$\hat{\sigma}^2 = \frac{1}{n-2} SS_E.$$ (See last lecture.)

The estimated standard errors of $\hat{\beta}_1$ and $\hat{\beta}_0$ are their estimated standard deviations i.e.

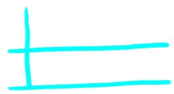
$$se(\hat{\beta}_1) = \sqrt{\frac{\hat{\sigma}^2}{S_{xx}}} \quad \text{and} \quad se(\hat{\beta}_0) = \sqrt{\hat{\sigma}^2 \left(\frac{1}{n} - \frac{\bar{x}^2}{S_{xx}} \right)}.$$

11.4 Hypothesis Tests in Simple Linear Regression

We want to test $H_0 : \beta_1 = (\beta_1)_0$ ← some #

H_1: e.g. $\beta_1 \neq (\beta_1)_0$

Slope = 0



e.g. $(\beta_1)_0 = 0$, which would mean NO linear relationship between X & Y.

We can't get anywhere without assuming something, so we assume $\varepsilon \sim N(0, \sigma^2)$ & so

i.e. this section will only apply in situations where this is a valid assumption

We found these above.

$$Y_i \sim N(\beta_0 + \beta_1 x_i, \sigma^2)$$ (linear comb. of Normal r.v.s)

We found these above.

So then $\hat{\beta}_1 \sim N\left(\beta_1, \dfrac{\sigma^2}{S_{xx}}\right)$ (" is Normal)

Under H_0 ($: \beta_1 = (\beta_1)_0$), $\dfrac{\hat{\beta}_1 - (\beta_1)_0}{\sigma/\sqrt{S_{xx}}} \sim N(0,1)$ (Standardizing.)

Since we don't know σ^2 we're in similar territory to "Tests on Mean of Normal pop. Variance Unknown" setting.

We replace σ^2 with $\hat{\sigma}^2$ & get as test statistic:

$$T_0 = \frac{\hat{\beta}_1 - (\beta_1)_0}{\hat{\sigma}/\sqrt{S_{xx}}} \sim t\text{-distribution with } n-2 \text{ degrees of freedom}$$

As usual, with 2-sided H_1, reject H_0 if

$$|t_0| > t_{\frac{\alpha}{2}, n-2} \quad \text{or} \quad -|t_0| < -t_{\frac{\alpha}{2}, n-2}.$$

<u>Example</u> In the example from last time, we found

$$\hat{\beta}_1 = -0.49, \; S_{xx} = 32.8, \; \hat{\sigma}^2 = 0.72.$$

Test the claim that there is a linear relationship between X and Y, at significance level $\alpha = 0.05$.

($n=5$).

<u>Solution</u> $H_0 : \beta_1 = 0$ (no lin. relation.)

$H_1 : \beta_1 \neq 0$

↳ We must put the "equality statement" as H_0.

$$t_0 = \frac{\hat{\beta}_1 - 0}{\hat{\sigma} / \sqrt{S_{xx}}} = \frac{-0.49}{\sqrt{0.72} / \sqrt{32.8}} = -3.31.$$

Compare against $-t_{\frac{0.05}{2}, 5-2} = -t_{0.025, 3} = -3.182.$

$-3.31 < -3.182$ so reject H_0 for H_1. (Yes, there is a linear relationship.)

<u>Tests for β_0</u> Exactly same idea. With assumption $\varepsilon \sim N(0, \sigma^2)$ we have

$$\hat{\beta}_0 \sim N\left(\beta_0, \sigma^2\left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right)\right).$$

t-test : $H_0 : \beta_0 = (\beta_0)_0$ Use $T_0 = \dfrac{\hat{\beta}_0 - (\beta_0)_0}{\sqrt{\hat{\sigma}^2\left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right)}}$

H_1 e.g. $\beta_0 \neq (\beta_0)_0$

$\sim$ t-distribution

also with $n-2$ degrees of freedom.

Analysis of Variance Approach (ANOVA)

Recall: $SS_E = SS_T - SS_R$

$SS_E = \sum_{i=1}^{n} e_i^2$

Sum of Squares Error

$= \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$

$SS_T = \sum y_i^2 - n\bar{y}^2$

Total sum of squares

$= \sum (y_i - \bar{y})^2$

$SS_R = \hat{\beta}_1 S_{xy}$

$= \sum (\hat{y}_i - \bar{y})^2$

Regression sum of squares; residual amount only explained by regression.

So $SS_T = SS_E + SS_R$

SS_E: $n-2$ degrees of freedom

SS_R: 1 degree of freedom

We have $E(SS_E / n-2) = \sigma^2$ $\quad E(SS_R / 1) = \ldots = \sigma^2 + \hat{\beta}_1 S_{xx}$ Check!

$E(\hat{\beta}_1 S_{xy}) = \ldots$

And so

$$F_0 = \frac{SS_R/1}{SS_E/(n-2)}$$ ~ F-distribution with 1 d.o.f. in numerator & n-2 d.o.f. in denominator.

T. B. C. ...

(Next time we'll explain something more about this — for now, the point is that we've come up with a test statistic — F_0 — whose underlying distribution is known & understood.)