STAT-1302; Lecture 18; March 19, '24

Test 2 Info. on Nexus

In Ch. 13, § 13.1

Last lecture:

Data:

(x): Age (in years)

(y): Price (in hundreds of dollars)

See Summary Stats. in Lecture 17.

a) least squares regression line:

$$\widehat{Price} = 318.83 - 33.49 * Age$$

b) Give a brief interpretation of "a" and "b" calculated in part (a).

b: -33.49 is the estimated decrease ($b < 0$)

in the mean price when age(x) is increased one unit (1 year).

That is, the average price of the vehicle goes down by \$3349 when the age of the vehicle goes up by one year.

a: Statistically, 318.83 is the mean price when Age(x) is Zero. For this problem, 318.83 is not meaningful.

(C) Predict the price of a 7-year old car.

$$\widehat{Price} = 318.83 - 33.49 \times Age$$

$$\widehat{Price} = 318.83 - 33.49 \times 7 = 84.4$$

(i.e. $84.4 \times 100 = \$\,8440$).

↑ units of x

Remark: It is safe to use the estimated

regression line to predict the price of a 7-year old car because $x = 7$ falls within the range of x-values observed in this data set (See Lecture 17).

d) Predict the price of an 18-year old car. Comment on the finding.

$$\widehat{\text{Price}} = 318.83 - 33.49 \times 18 = -283.99$$

i) negative price !

ii) Even if the predicted price for an 18-year old car were positive, $x = 18$ is outside the range of x-values for which the data is collected; it is not safe to use the regression line for $x = 18$.

Regression is for interpolation, not extrapolation.

§ 13.2 Standard Deviation of Errors & Coefficient of Determination:

Recall the idea of a residual:

$$e = Y - \hat{Y} \qquad \text{where,}$$

Y = Observed response,

$\hat{Y}$ = Predicted response from the estimated least squares regression line.

Recall: The Simple Linear Regression Model:

$$Y_i = A + BX_i + \varepsilon_i$$

$$\varepsilon_i \overset{ind}{\sim} N(0, \sigma^2), \text{ for } i = 1, \ldots, n.$$

e_i is an estimate of ε_i.

Long-hand formula for estimating σ:

$$S_e = \hat{\sigma} = \sqrt{\frac{SSE}{n-2}}$$

$$= \sqrt{\frac{\sum (y_i - \hat{y}_i)^2}{n-2}} = \sqrt{\frac{\sum e_i^2}{n-2}}$$

Short-cut formula (Given):

$$S_e = \sqrt{\frac{SS_{yy} - b\,S_{xy}}{n-2}}$$

where

$$SS_{xy} = \sum xy - \frac{(\sum x)(\sum y)}{n}$$

$$SS_{yy} = \sum y^2 - \frac{(\sum y)^2}{n}$$

Ex. Consider the food expenditure (y) and monthly income (x) example in a previous lecture.

Summary Statistics:

$\sum x = 212$, $\sum y = 64$, $\sum xy = 2150$,

$\sum x^2 = 7{,}222$, $\sum y^2 = \ldots\ldots$,

$$SS_{yy} = \sum y^2 - \frac{(\sum y)^2}{n} = 60.8571$$

Q'n: What is an estimate of the standard deviation of the residuals?

(Model: Expenditure = A + B*Income + ε

where $\varepsilon \overset{ind}{\sim} N(0, \sigma^2)$.

This q'n is asking you to estimate σ.

$$S_e = \sqrt{\frac{SS_{yy} - bS_{xx}}{n-2}}$$

$$= \sqrt{\frac{60.8571 - 0.2642(211.7143)}{7-2}} = 0.9922$$

$\therefore \quad \hat{\sigma}_e = 0.9922.$

[Extra: STAT-3103

$$SST = SSR + SSE$$

$$SST = \sum_{i=1}^{n} (y_i - \bar{y})^2 \quad \text{Total Sums of Squares}$$

$$SSR = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2 \quad \text{Regression Sums of Squares}$$

$$SSE = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \quad \text{Error Sums of Squares.}$$

For the least Squares line to be useful, we want SSR to be "large" relative to SSE.]

A statistic to assess the usefulness of the regression line is the Coefficient of determination (r^2).

It measures the proportion of the variation in Y (the response variable) that is explained by our X (the explanatory variable).

$$r^2 = \frac{SSR}{SST}$$

Short-cut formula:

$$r^2 = b \frac{SS_{xy}}{SS_{yy}}$$

(Not given!)

But, to find r, Square the value of

$$r = \frac{SS_{xy}}{\sqrt{SS_{xx}SS_{yy}}}$$ ← Correlation between x and y

Given on Final

Property of r^2: $0 \le r^2 \le 1$

Ex. Consider the monthly food expenditure (y) and monthly income (x) regression example.

Q'n: What is the percentage of the variation in monthly food expenditure that is explained by monthly income (x)?

(ie What is r^2 (expressed as a percentage)?)

$$r = \frac{SS_{xy}}{\sqrt{SS_{xx} \cdot SS_{yy}}} = \frac{211.7143}{\sqrt{801.4286(60.8571)}} = 0.9587 \approx 0.96$$

$\therefore \; r^2 = 0.96^2 = 0.92$

Interpretation: Approx. 92% of the variation in monthly food expenditure (y) is explained by monthly income (x).

Compare with:

$$r^2 = \frac{b\,SSxy}{SSyy} = 0.2642 \times \frac{211.7143}{60.8571} = 0.92$$

§13.3 Inference about B

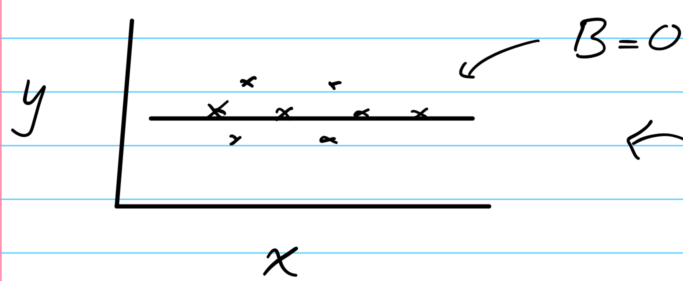
Idea: want to test whether the regression line is "useful". We do this by testing.

$H_0: B = 0$ "regression line not useful"

against

$H_1: B > 0$ or $H_1: B < 0$ or

$H_1: B \neq 0$.



← regression not useful

⇓

use $\bar{y}$ to predict y rather than x.

Test Statistic:

$$t = \frac{b - B_0}{S_e / \sqrt{SS_{xx}}}$$

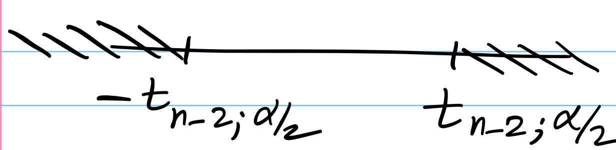
(see S_e on formula sheet)

Under H_0 (ie. assuming H_0 is true),

$$t = \frac{b - B_0}{S_e / \sqrt{SS_{xx}}} \sim t_{n-2}$$

$H_0 : B = 0$

<u>Test</u>	<u>Rejection Region:</u>
$H_0 : B = 0$ vs. $H_1 : B > 0$	$t > t_{n-2;\alpha}$
$H_0 : B = 0$ vs. $H_1 : B < 0$	$t < -t_{n-2;\alpha}$
$H_0 : B = 0$ vs. $H_1 : B \neq 0$	$\lvert t \rvert > t_{n-2;\alpha/2}$ $\Leftrightarrow$ $t < -t_{n-2;\alpha/2}$ or $t > t_{n-2;\alpha/2}$



Omit p-value for testing $H_0: B=0$.