3703-3J04 Probability & Statistics for (Civil) Engineering

C01 - Lecture 33

Last time Prediction Intervals in Linear Regression

We use $\hat{y}_0 = \hat{\beta}_0 + \hat{\beta}_1 x_0$ as a predictor of Y_0 the next observation of Y at $X = x_0$.

Error in prediction $e_{\hat{p}} = Y_0 - \hat{Y}_0$ has mean 0, variance $\sigma^2\left(1 + \frac{1}{n} + \frac{(\bar{x} - x_0)^2}{S_{xx}}\right)$.

A $100(1-\alpha)\%$ Prediction Interval for prediction $\hat{Y}_0$ is

$$\hat{y}_0 \pm t_{\frac{\alpha}{2}, n-2} \sqrt{\hat{\sigma}^2\left(1 + \frac{1}{n} + \frac{(\bar{x} - x_0)^2}{S_{xx}}\right)}.$$

Example For our running example [with $\hat{\beta}_0 = 4.05$, $\hat{\beta}_1 = -0.49$, $\bar{x} = 4.2$, $S_{xx} = 32.8$, $n = 5$, $\hat{\sigma}^2 = 0.72$] a 95% Prediction Interval for a single future observation Y_0 at $X = 3$ is given by ...

Solution $2.58 \pm 3.182 \sqrt{0.72\left(1 + \frac{1}{5} + \frac{(4.2-3)^2}{32.8}\right)}$

$= 2.58 \pm 3.182(0.94)$

↖ contrast 0.42 from C.I. example above

i.e. $(-0.43, 5.59)$ ← C.I. was by contrast $(1.25, 3.91)$

Bigger interval for 1 prediction as opposed to mean.

11.7 Adequacy of the Regression Model

What have we assumed so far in our regression model?

For each i $\quad Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$ where

(i.e. there is a linear relationship between X & Y!)

$E(\varepsilon_i) = 0$

$V(\varepsilon_i) = \sigma^2$ for all i.

& $Cov(\varepsilon_i, \varepsilon_j) = 0$ for all $i \neq j$.

To do tests/find C.I.s we also assumed $\varepsilon_i \sim N(0, \sigma^2)$.

Our goal: conduct analyses to check up on these assumptions

We can check our assumption $\varepsilon_i \sim N(0, \sigma^2)$ for all i with a probability plot.

Recall: Find residuals $e_i = y_i - \hat{y}_i$ ← $\hat{\beta}_0 + \hat{\beta}_1 x_i$

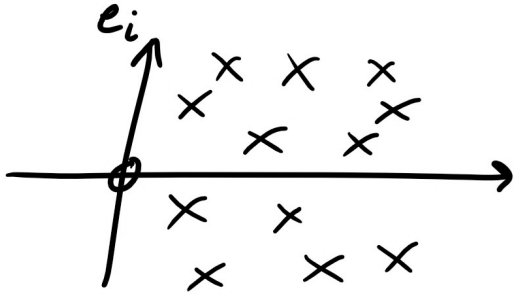
Reorganize the e_is as $e_{(1)} < e_{(2)} < \ldots < e_{(n)}$

Plot $e_{(i)}$ against z_i where $\Phi(z_i) = \dfrac{i - 0.5}{n}$.

If ε_i Normally distributed, get a straight line.

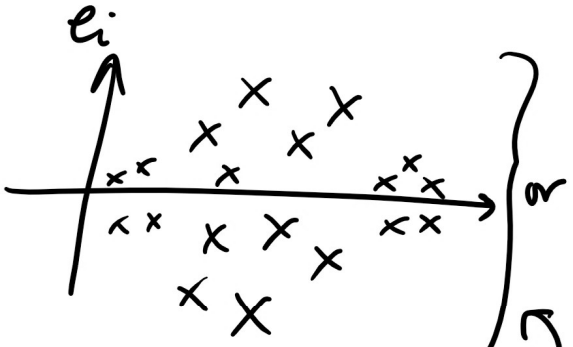
To check $\text{Cov}(\varepsilon_i, \varepsilon_j) = 0$ for $i \neq j$, can plot e_is against x_i or y_i or time

Should look like

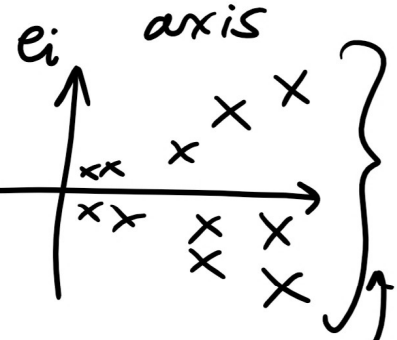


} evenly spread either side of the horizontal axis

& not look like

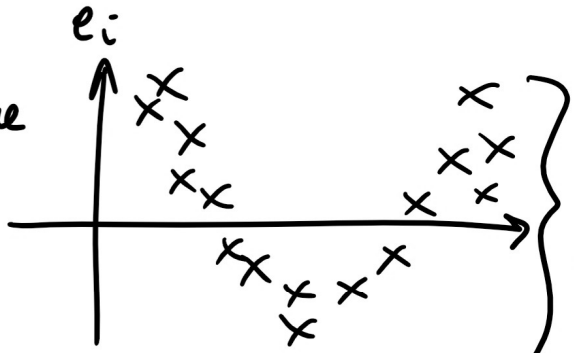


or



} spread is not even, may depend on x or y or on time

& not like



} suggests the relationship between X and Y NOT linear.

To check linearity we can go back to the ANOVA :

The <u>coefficient of determination</u> is

$$R^2 = \frac{SS_R}{SS_T}$$

} Proportion of variation in Y explained by regression

Recall

$SS_T = SS_E + SS_R$

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

↳ "Total variation in y"

↓ Residuals

↳ Variation in Y "explained by regression on x"

The bigger the proportion, the more that variation in Y is explained by a linear relationship to X.

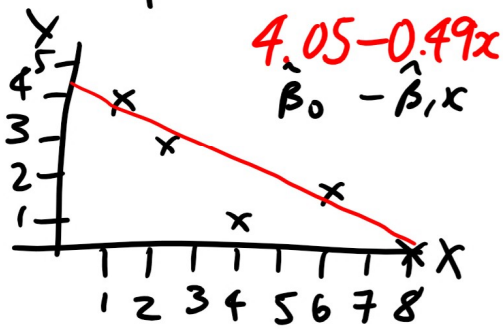
i.e. the bigger R^2 is (notice: must be in $[0,1]$), the closer to being linear the relationship between X and Y is.

Example Our running example had

x_i	1	2	4	6	8
y_i	4	3	1	2	0

($S_{xy} = -16$, $\sum y_i^2 = 30$, $\bar{y} = 2$)

Find R^2.



Solution $R^2 = SS_R / SS_T$

$SS_R = \hat{\beta}_1 S_{xy} = (-0.49)(-16) = 7.84$

$SS_T = \sum y_i^2 - n\bar{y}^2 \quad [= (n-1)S_y^2]$ (S_Y^2 = sample variance of Y)

$= 30 - 5.2^2$

$= 10$

So $R^2 = \frac{7.84}{10} = 0.784.$

i.e. the model accounts for 78.4% of the variability of Y in the data.

(0% – model accounts for no variability i.e. no linear relationship

100% – model accounts for all variability i.e. all data points lie exactly on line).

Notice Value of R^2 does not tell us anything about the steepness of the line.

11.8 Correlation

So far trying to understand how Y varies for fixed $X = x$. (Perhaps we can even control x-value.)

Now we'll look at how X & Y vary together.

We assume X, Y jointly distributed (with jointpdf $f(x)$)

Recall correlation coefficient: $\rho = \dfrac{\sigma_{XY}}{\sigma_X \sigma_Y}$

$\sigma_{XY} = \text{Cov}(X,Y)$

$= E((X-\mu_X)(Y-\mu_Y))$

$\sigma_X = \sqrt{V(X)}$ $\sigma_Y = \sqrt{V(Y)}$

Under certain assumptions $E(Y|x) = \beta_0 + \beta_1 x$

↳ has to do with conditional pdfs, which we didn't cover – see textbook.

where $\beta_1 = \dfrac{\sigma_Y}{\sigma_X}\rho$

i.e. we get a simple linear regression model and $\beta_1 = 0$

exactly when $\rho = 0$.

We have an estimator of ρ:

$$R = \frac{\sum_{i=1}^{n} (Y_i - \bar{Y})(X_i - \bar{X})}{\sqrt{\underbrace{\sum_{i=1}^{n} (X_i - \bar{X})^2}_{S_{xx}} \underbrace{\sum_{i=1}^{n} (Y_i - \bar{Y})^2}_{SS_T}}} = \frac{S_{xy}}{\sqrt{S_{xx} SS_T}}$$

Watch out for textbook typo.

Note $\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}} = \frac{S_{xy}}{\sqrt{S_{xx} SS_T}} \cdot \frac{\sqrt{SS_T}}{\sqrt{S_{xx}}} = R\sqrt{\frac{SS_T}{S_{xx}}}$ ← spread of y-values; ↖ spread of x-values

$\hat{\beta}_1$ measures predicted change in mean of Y given change in X } So can actually use x-values to predict y-values.

R measures linear association between X & Y but o/w # has no meaning (it's always in $[-1,1]$ no matter what values X and Y take).