Confidence and Prediction Intervals for Response (CI & PI)

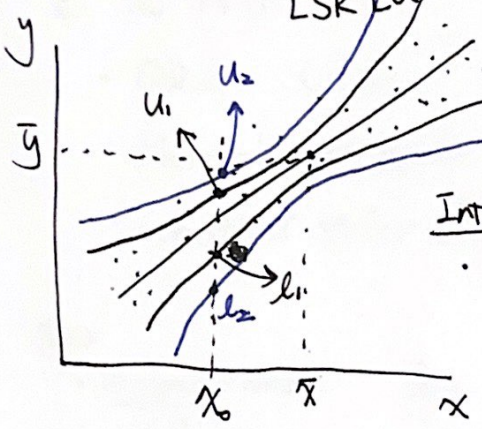
Confidence intervals give a reasonable range of values are used for the mean response at a specific value of x. (for response)

CI for avg y at a particular value of x.

Prediction intervals give a range of values for NEW individual response (i.e. if we were to predict the response y at a new observation at that value of x, what is a good interval of values in which the true y should fall?). There is a formula for the prediction interval, but typically, we just let a computer do it and interpret the output, instead of calculating this by hand.

PI for an individual y at a particular value of x.

($1-\alpha$ confidence level)



$(\bar{x}, \bar{y})$ is on the line $\hat{y} = a + bx$.

CI for response

PI for response

Interpretation:

- (If $x = x_0$) we are $1-\alpha$ confident that the avg response for all subjects whose x-value is x_0 is somewhere between l_1 and u_1.
- We are $1-\alpha$ confident that the response for a new individual whose x-value is x_0 is somewhere btwn l_2 and u_2.

Remarks:

1. PI always wider than CI.
2. Both PI and CI are the narrowest at $x = \bar{x}$.

Example: Interpret the CI and PI for Ex1=80 for the example that predicts Final Score in the class from Exam 1.

- 95% CI for response (85.370, 86.457)

We are 95% confident that the average final score in class for all students who scored 80 on Exam 1 is btwn 85.37 and 86.457.

- 95% PI for response (74.568, 97.259)

We are 95% conf that the true final score in class for one student who would score 80 on Exam 1 would be btwn 74.568 and 97.259.

Residual Analysis – Checking Assumptions

We can look at residual plots to determine if the assumptions are met. Recall that

$$e = \text{resid} = \text{obs } y - \text{pred } y = y_i - \hat{y}_i$$

The assumptions we need to check are:

(random)

ε = theoretical prediction errors

$e_i = y_i - \hat{y}_i$ = residual for each obs in data set.

$\varepsilon \overset{iid}{\sim} N(0, \sigma^2)$

- SRS (iid)
- Normal distribution of response (N)
- Constant stder of pts around line (σ^2)

4 in 1

Plots:

- Histogram of residual (N)

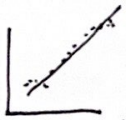
Not bell shape → departure from normality

We want bell shape → Normality

- Normal Probability plot (N)

NPP

We want diagnal/straight line



NPP is a graphical technique for assessing whether or not a data set is approximately normally distributed.

Departure from a straight line → Departure from normality.

- Residual vs. Fit (randomness σ^2)
- Residuals vs. Order

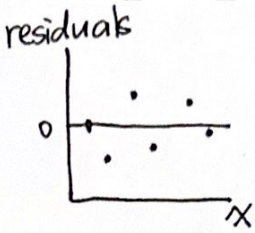
= Want totally random plots with no pattern.

Example: Interpret the 4in1 Residual Plots for the example that predicts Final Score in the class from Exam 1.

Residual Plots magnify any problems in Model or Assumptions:

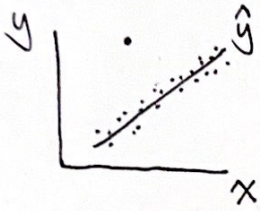
Original data

Good data for reg

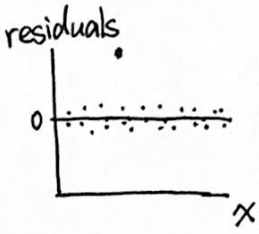


Residual plot (versus x)

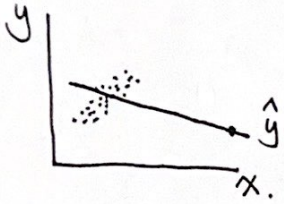
Random
Good



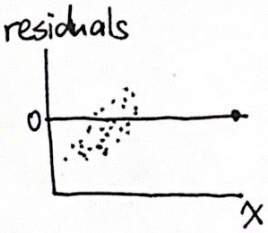
outlier



magnifies outlier
perhaps violation of normality ~~or~~ and/or constant variance.

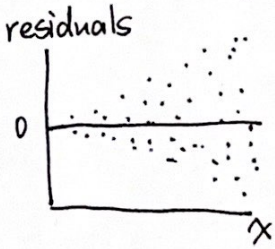


influential outlier (high leverage point)
awful in reg

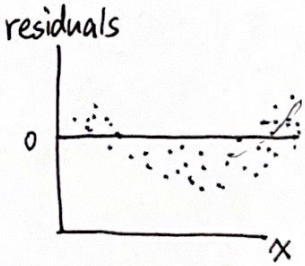


maybe wrong model
we might need to adjust/change the model and/or get more data.

funnel shape



constant variance is violated.



pattern noticed
pos → neg → pos
curvy
NOT SLR but QUAD Reg

Remark:

- Residual plots are more important in Multiple Linear Regression, but easier to understand in SLR.