In HW2, Q2 (i) asked you to provide:

* **95% confidence interval for mean salary (option CLM in SAS) of the student:**

A common objective in regression analysis is to estimate the mean for one or more probability distributions of Y. Let Xh: level of X (gpa) for which we wish to estimate the mean response (denoted by E{Yh}). It's easy to get the point estimator   of E{Yh}:

= b0+ b1 Xh

The sampling distribution of refers to the different values of that would be obtained if repeated samples were selected, each holding the levels of the predictor variable X constant, and calculating for each sample.

Konda has demonstated in class:

E{} = E{b0+ b1 Xh } = E{b0} + Xh E{b1} = β0 + β1 Xh

{} = MSE

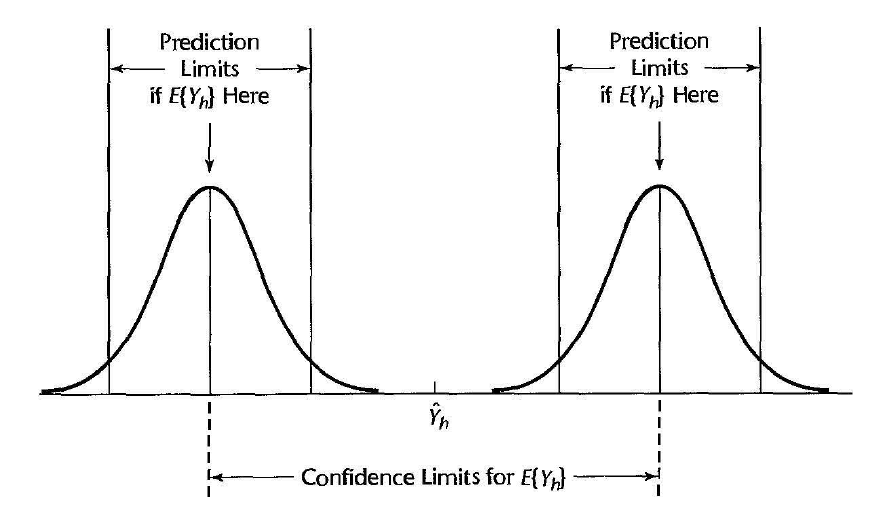
Thus, confidence interval ± t(1-α/2; n-2) {}

* **95% prediction interval for the true(individual) salary (CLI in SAS).**

Now we consider the prediction of a new observation Y corresponding to a given level X (gpa). Prediction interval for Yh(new) must take account of 2 elements:

1- Variation in possible location of the distribution of Y (in other words, the mean of the distribution of Y is estimated by , and the variance of the distribution of Y is estimated by MSE).

2- Variation within the probability distribution of Y (i.e. there are two probability distributions of Y, corresponding to the upper and lower limits of a confidence interval for E{Yh}. Since we do not know the mean E{Yh} and only estimate it by a CI, we cannot be certain of the location of the distribution of Y.



(figure from Kutner, Nachtsheim, Neter and Li, Applied Linear Statistical Model 5th)

E{} = E{b0+ b1 Xh } = E{b0} + Xh E{b1} = β0 + β1 Xh

{pred} = MSE + {} = MSE

Thus, prediction interval ± t(1-α/2; n-2) {pred}

What is the difference between these two types of confidence intervals?

- 95% prediction interval obtained is always wider than the 95% confidence interval. The reason is that we encounter both the variability in from sample to sample as well as the variation within the probability distribution of Y. We can see the addtional element added up in the s{pred} compared to s{}.

- PI and CI are different conceptually. A CI represents an inference on a parameter and is an interval that is intended to cover the value of parameter. A PI is a statement about the value to be taken by a random variable, the new observation Yh(new).

SAS code is quite simple:

PROC IMPORT OUT=WORK.SAL

            DATAFILE="C:/Users/minhh/OneDrive - University of Illinois at Chicago/BSTT401\_TASpring2021/HW2/salary.xlsx"

DBMS=XLSX REPLACE;

RUN;

DATA gpa30sth;

INPUT student gpa salary;

DATALINES;

31 3.25 .

32 2.01 .

;

RUN;

DATA COMBINED;

SET gpa30sth SAL;

RUN;

PROC PRINT DATA=COMBINED(OBS=5);

title "Combined data";

RUN;

PROC REG DATA=COMBINED;

id student;

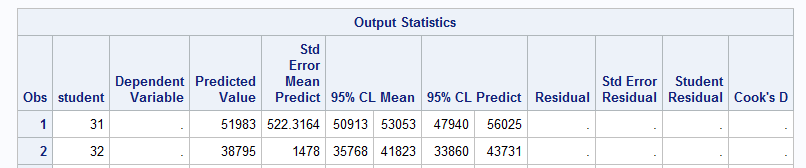
model salary=gpa/ r clm cli;

\*OUTPUT OUT = predictions

predicted = yhat LCLM = lower UCLM = upper STDP = se\_mean ;

title "Regression results for the combined data";

RUN;



Unfortunately, SAS output printed out the whole dataset, we may need to use output statement to extract then print (proc print) the necessary observations. Here's a small sample so it's still convenient to see.

References:

- Konda's lecture notes in class

- Kutner, Nachtsheim, Neter and Li, Applied Linear Statistical Model 5th

- <https://documentation.sas.com/?cdcId=pgmsascdc&cdcVersion=9.4_3.4&docsetId=statug&docsetTarget=statug_reg_syntax08.htm&locale=en#statug.reg.modelclb>

Let me know if you have any questions.

Hai (TA).