BUS 232 Residuals

Terminology

Residuals

- the distances from each point to the least-squares regression line.
- the sum of the residuals is always 0

Standardized Residuals

- z-score residuals
- if over or under 2 than outside of 95% and outlier

Lurking Variable

- A variable not included in the study design that does have an effect of the variables studied
- lurking variables can falsely suggest a relationship
- eg) strong positive association between number of firefighters at a fire site and the amount of damage a fire does.
  Here the lurking variable is the size of the fire
- can include a lurking variable. For example using number of beers **/ weight**.

outlier

- only an outlier if far away from the residual line.
- usually 2 standard deviations from the residual line.

Two-way / Block design

- an experiment of two *categorical* factors that are studied with several levels of each factor.

Cautions in Regression:

1. **association does not imply causation.**
   Changes in x does not necessarily mean changes in y.
   A regression model should have x causes y.
   To detect a causation we need to do an experiment.

2. **Beware correlations based on averaged data.**

3. **Beware predicting outside range of data.**
   Usually having zero of something doesn't make sense. Therefore we can't use the *y* intercept for much sometimes.

Residual plot

The residual plot shows how the points line up with the regression line.
residual plots can be useful for finding outliers and determining if the linear regression line is appropriate.

- if there is a curved pattern on the residual plot then it is not linear.
- if the variability is growing across the plot this is a warning size because predictions will probably be bad.

Influential observation

(removing it would have a large effect of the regression)
We have to do an experiment to see if there is a correlation/influence.

Can have data that influences the regression line majorly.

Relations in categorical data

two way tabels

Present the data of two categorical variables that are most likely correlated.

organize data about two categorical variables obtained from a two-way/block design.

Example data:
3 age groups as columns
each group has 4 factors (eg 4 years of high school, finished high school ...)

Here education is the row variable and age group is the column variable
Each combination of values for these two variables is called a cell

Then under the columns and beside the columns you can put the total (marginal distributions).

marginal distributions

Written to the right or under like in a margin.
The total or percentages.

The marginal distributions summarize each categorical variable independently.

conditional distributions

Just look at one column or one row in the two-way table.
For example, what is the percentage of those who completed highschool in a given age group.

relationships between categorical variables

This (https://youtu.be/ULkLzSJZOFE?t=1044) explains how conditional distributions are related.

if the conditional distributions are different than the two categorical variables are associative. (Because one of the variables actually has an effect on another one.)