DUMMY REGRESSION

The focus this week is on **dummy regression.** It can be used to examine the relationship between multiple *independent* variables and a *dependent* variable to predict outcomes or answer business questions.

Variables in Hypotheses

One Variable

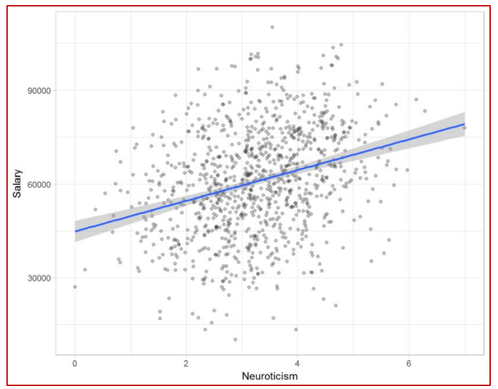
Numeric: Use one-sample t-test for mean comparison.
Categorical: Use chi-square goodness-of-fit test for proportions

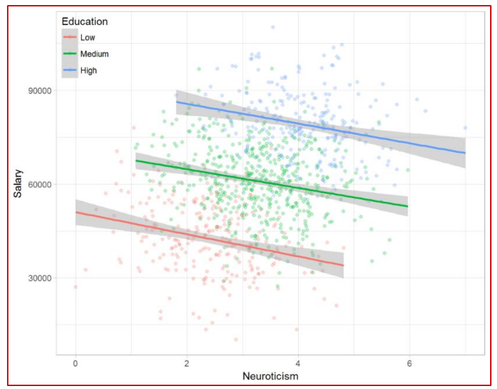
Two Variable

Two Numeric: Use regession to test linear relationship slope
One Numeric, One Categorical: Use independent samples t-test (between groups) or paired t-test (related groups/time)

Simpson's Paradox

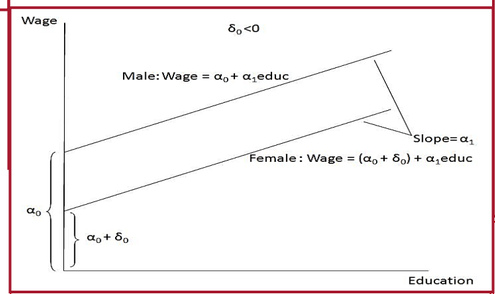
An association between two variables in a population emerges, disappears or reverses when the population is divided into subpopulations.





Dummy Variables

- Represent categories as binary variables (0/1)
- Created using functions (e.g. IF in Excel)
- Important in regression models to handle categorical data (e.g. day of the week).



Left example source:
https://paulvanderlaken.com/2017/09/27/simpsons-paradox-two-hr-examples-with-r-code/

Regression Models

Multiple Regression Equation

$$\hat{y} = b_0 + b_1 x_1 + \cdots + b_p x_p$$

Comparing Models

Occam's Razor:
Prefer simpler models when possible.

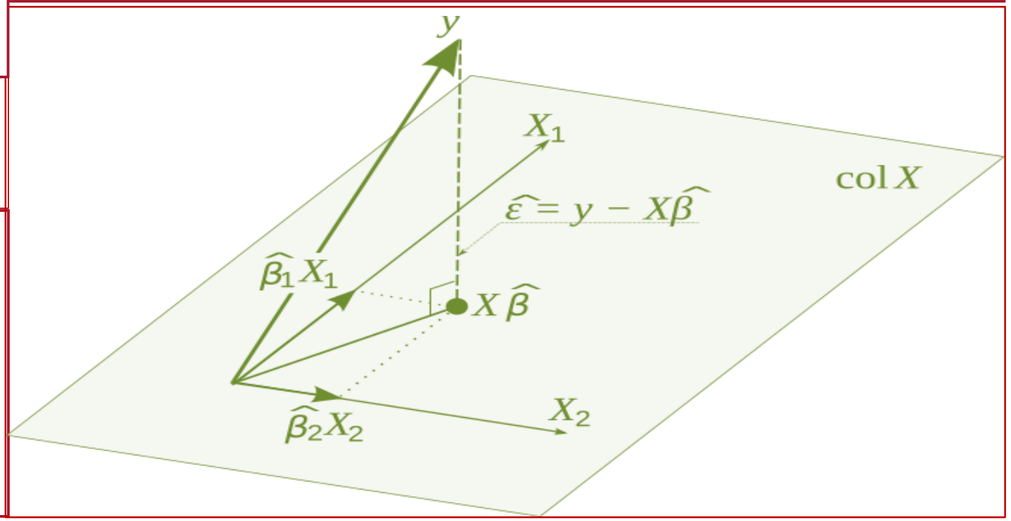
Adjusted R^2:
Used for model comparisons when there are differing numbers of independent variables.

Interpreting Coefficients

Coefficient Interpretation: Shows the expected change in the dependent variable (y) for a one-unit increase in the independent variable (x), assuming other variables remain constant.

Positive Coefficient: Predicts an increase in y with an increase in x

Negative Coefficient: Predicts a decrease in y with an increase in x



R^2

Definition: R^2 represents the proportion of variability in the outcome variable that is explained by the regression model.
Higher R^2: The higher the R^2 value, more of the total variation is explained by the model.
Compare models based on adjusted R^2 for a fair evaluation across different numbers of variables.
P-value (<0.05): Tests significance of slopes to identify important variables.