Towards Regression

Review LS 7 Comparing Means Hypothesis Tests

- Perform hypothesis testing and estimation to generalize
- Important to know and sample from target popl'n
  - Often the available popl'n ≠target popl'n
- Null hypotheses are crisp assertion, e.g.
  - Null hypothesis; H(0): $\theta=a$ (a is a real constant) vs
  - Select apriori alternative one of H(a): $\theta \neq a$ or H(a): $\theta<a$ or H(a): $\theta>a$
  - Use P_value determined from computed value of Z* or t* and compared against distribution of transformed sampling distribution under H(0)
- Other concerns
  - Data types: metric or ordinal
  - Test types: unpaired (two groups), paired (diff)
  - Variance: pooled vs not pooled
  - Assumption: normality, iid, skewness, metric➔parametric test, otherwise non-parametric

Preview of LS 8 Regression

- Start the last of the course's major topics
- Set up the approach and machinery for creating a formal structure to model a functional relationship between elements or variables
- Starting out focusing on simple linear regression (SLR) ➔ one predictor, one response variable and move on to >1 predictor and one response variable multiple linear regression
- Use an estimation method called Ordinary Least Squares (OLS) which is built upon assumptions of MOM estimators
- Recall for the BP the following result $E[(Y-g(X))^2|X]$ called mean square error was the minimized when $g(x) = E[Y|X=x]$
- In the world of statistics we want $BLP = \sum(Y-E[Y|X])^2 = \sum(\varepsilon^2)$ which under the assumptions $\sum(\varepsilon)=0$, $cov(\varepsilon,X) = 0$ and Y (and ε) ~ IID ➔ $E[Y|X]) = \beta(0) + \beta(1)*X(i)$
- The model correct on average; error term is unpredictable by the X's, referred to as exogeneity, otherwise OLS attributes (incorrectly) some of the variance the error explains to the predictors.

Moment Conditions

- Moment Condition Assumptions $\sum(\varepsilon)=0$, $cov(\varepsilon,X) = 0$
- Let $Y(i) = \beta(0) + \beta(1)*X(i) + \varepsilon(i)$

  $Ybar = 1/n * \sum(\beta(0) + \beta(1)*X(i) + \cancel{\varepsilon(i)}\ )$

  $= \beta(0) + \beta(1)*\sum(X(i))/n \ + \sum(\varepsilon(i))$

- $Cov(X,Y) = Cov(X, \beta(0) + \beta(1)*X + \varepsilon\ )$

  $= Cov(X, \cancel{\beta(0)}) + Cov(X, \beta(1)*X) + Cov(\cancel{X, \varepsilon})$

  $Cov(X,Y) = \ \beta(1) * Var(X)$

OLS Normal Equations

- $y(i) = \beta(0) + \beta(1)*x(i) + \varepsilon(i)$ or $\varepsilon(i) = y(i) - \beta(0) + \beta(1)*x(i)$
- Let $Q = \sum \varepsilon(i)\text{^}2$; want to $\beta(0)$, $\beta(1)$ which minimize Q or min $\sum(y(i) - \beta(0) - \beta(1)*x(i))\text{^}2$
- For $\beta(0)$: $\partial Q/\partial\beta(0) = 2*\sum(y(i) - \beta(0) - \beta(1)*x(i))(-1)$
- $= -2n*(\text{ybar}-\beta(0) - \beta(1)*\text{xbar})$
- Min: $0=\text{ybar}-\beta\text{hat}(0) - \beta\text{hat}(1)*\text{xbar} \Rightarrow \beta\text{hat}(0) = \text{ybar} - \beta\text{hat}(1)*\text{xbar}$ (note hat)
- For $\beta(1)$: $\partial Q/\partial\beta(1) = -2*\sum(y(i) - \beta(0) - \beta(1)*x(i))(x(i))$
- Min: $0 = -2\sum x(i)(y(i) - \beta\text{hat}(0) - \beta\text{hat}(1)*x(i))$
- $= \sum x(i)y(i) - \beta\text{hat}(0)\sum x(i) - \beta\text{hat}(1)*\sum x(i)\text{^}2 \Rightarrow$
- $\boxed{\sum x(i)y(i) - \beta\text{hat}(1)*\sum x(i)\text{^}2 - (\text{ybar} - \beta\text{hat}(1)*\text{xbar})*\sum x(i)}$
- Then substituting in $\beta\text{hat}(0)$ & rearranging (left as exercise for you)
- $\Rightarrow \beta\text{hat}(1) = (\sum(x(i)-\text{xbar})(y(i) - \text{ybar}))/\sum(x(i)-\text{xbar})\text{^}2$