Midterm_review

==

Note One

==

- Experimental vs non-experimental data
    - Random assignment of “x”
- Types of data
    - Cross-sections
        - Observe many units in one period
    - Time-series
        - Observe on unit over many periods
    - Panels
        - Observe many units over many periods

==

Note Two

==

- P(B | A)
    - P(B && A) / P(A)

- Random variables
    - X is a random variable if for every real number *a* there exists a probability $P(X \leq a)$, that is, this is the probability that the random variable X will take on a value that is less or equal than the number *a*

- Density Functions
    - f(a) is a formula or a table giving the probability that X takes on each possible value a

- **Expected value**
    - $E(Y) = y_1p_1 + y_2p_2 + \ldots + y_kp_k$
    - Value that expect on average from the random variable
- **Variance and Standard Deviation**
    - $var(Y) = E(r - \mu_Y)^2 = \sum(r - \mu_Y)^2 * p$
    - How far is the value from the expected value

- Joint Distributions
    - Multiple random variables
    - f(x, y) = P(X = x, Y = y)
    - Marginal distributions
        - Probabilities that X and Y take on different values (**margin column or row in the table**)

- Conditional distributions
    - Probabilities that one variable take on each value **given that the other has taken a given value** → P(X = a | Y = b) = P(X = a, Y = b) / P(Y = b)
- **Independence**
    - X and Y are independent if
        - Conditional distributions are equal to the marginals for all possible values of Y
        - P(X = a | Y = b) = P(X = a, Y = b) / P(Y = b)
        - = P(X = a) * P(Y = b) / P(Y = b)
        - = P(X = a)
- **Covariance**
    - cov(X, Y) = $E[(X - u_X)(Y - u_Y)]$
    - cov(X, Y) > 0 → positive relation
    - If independent, cov = 0
    - cov(X, X) = $E[(X - u_X)^2]$ = var(X)
- **Correlation coefficient**
    - corr(X, Y) = $\frac{cov(X, Y)}{\sqrt{var(X)var(Y)}} = \frac{\sigma_{XY}}{\sigma_X \sigma_Y} = \rho_{XY}$
    - -1 ≤ corr(X, Y) ≤ 1
    - = 1 means perfect positive linear association
    - = -1 means perfect negative linear association
    - = 0 means no linear association

- Normal Distribution
    - u +/- 1.96 standard deviation = 95% of the data
    - Z score = (Y - u) / standard deviation

- Estimator and Estimates
    - Estimator
        - A function
    - Estimates
        - A value
- Estimation of the Mean
    - Y_bar = ∑(Y) / n
    - Y_bar is an unbiased estimator of u_Y.
    - var(Y_bar) = $(\text{standard deviation})^2$ / n

- Sample size matters
    - Convergence in probability, consistency, and the law of large numbers

- **Central Limit Theorem**
    - As n → infinity, the distribution of (Y-bar - u_Y) / (standard deviation) becomes the standard normal distribution

- Estimation of the Mean

Example: Suppose Y takes on 0 or 1 (a ***Bernoulli*** random variable) with the probability distribution

$\Pr[Y = 0] = .22, \Pr(Y = 1) = .78$

$E(Y) = p\times 1 + (1 - p)\times 0 = p = .78$
$\sigma^2_Y = E[Y - E(Y)]^2 = p(1 - p) = .78\times(1 - .78) = 0.1716$

The sampling distribution of $\bar{Y}$ depends on n (the sample size):

– Consider $n = 2$: The sampling distribution of $\bar{Y}$ is
  - $\Pr(\bar{Y} = 0) = (.22)^2 = .0484$
  - $\Pr(\bar{Y} = 1/2) = 2\times.22\times.78 = .3432$
  - $\Pr(\bar{Y} = 1) = (.78)^2 = .6084$

-

- Summary
    - The exact sampling distribution of Y-bar has mean u_Y and variance <u>$(\text{delta})^2/n$</u>
    - When n → infinity

        $\boxed{\bar{Y} \xrightarrow{p} \mu_Y}$ (Law of large numbers)

        $\boxed{\frac{\bar{Y} - E(\bar{Y})}{\sqrt{\text{var}(\bar{Y})}} \text{ is approximately } N(0,1)}$ (CLT)

        -
        - **Note: Y_bar does not equal to μ_Y**

==

Note Three

==

- **Confidence Intervals**
    - A 95% confidence interval for μ_Y is an interval that contains the true value of μ_Y in 95% of repeated samples.
    - For r.v Y which has only 1 or 0.
        - E(Y) = p
        - var(Y) = p(1 - p)

- **Hypothesis Testing**
    - $H_0 : E(Y) = u_Y$
    - $H_1 : E(Y) \neq u_Y$
    - p-value
        - Calculation

- $$\boxed{p\text{-value} = \Pr_{H_0}[|\bar{Y} - \mu_{Y,0}| > |\bar{Y}^{act} - \mu_{Y,0}|]}$$

$$\begin{aligned} p\text{-value} &= \Pr_{H_0}[|\bar{Y} - \mu_{Y,0}| > |\bar{Y}^{act} - \mu_{Y,0}|], \\ &= \Pr_{H_0}[|\frac{\bar{Y} - \mu_{Y,0}}{\sigma_Y / \sqrt{n}}| > |\frac{\bar{Y}^{act} - \mu_{Y,0}}{\sigma_Y / \sqrt{n}}|] \\ &= \Pr_{H_0}[|\frac{\bar{Y} - \mu_{Y,0}}{\sigma_{\bar{Y}}}| > |\frac{\bar{Y}^{act} - \mu_{Y,0}}{\sigma_{\bar{Y}}}|] \\ &= \text{probability under left+right } N(0,1) \text{ tails} \end{aligned}$$

- where $\sigma_{\bar{Y}}$ = std. dev. of the distribution of $\bar{Y} = \sigma_Y/\sqrt{n}$.
- Link between p-value and the significance level
    - If the significance level is 5%
        - Reject the null hypothesis if $|t^{act}| \geq 1.96$
        - Reject if $p \leq 0.05$
- **t-table and degrees of freedom**
    - Degrees of freedom = n - 1

<u>Digression: the Student *t* distribution</u>

If Y_i, $i = 1,\ldots, n$ is i.i.d. $N(\mu_Y,\sigma_Y^2)$, then the *t*-statistic has the Student *t*-distribution with $n - 1$ degrees of freedom.

The critical values of the Student *t*-distribution is tabulated in the back of all statistics books. Remember the recipe?

1. Compute the *t*-statistic
2. Compute the degrees of freedom, which is $n - 1$
3. Look up the 5% critical value
4. If the *t*-statistic exceeds (in absolute value) this critical value, reject the null hypothesis.

- 11
- T table is typically used when the sample size is small.

==

Note Four

==

- Regression Analysis
    - How to draw a line through a set of points
        - Sample - set of points
        - Want to know å and ß following some estimation technique
        - **<u>OLS</u>**
            - **<u>Ordinary Least Squares</u>**

- Linear regression with one regressor

- Estimate the **causal effect** on Y of a unit change in X
- Hypothesis testing
    - How to test if the slope is zero
- Confidence intervals
    - How to construct a confidence interval for the slope
- Ex:
    - Effect of STR on Test Scores
    - Linear regression
        - Population regression line
            - Test Score = $ß_0$ + $ß_1$STR
            - $ß_1$ is the slope of population regression line
            - In general, the relation will not hold exactly
                - Omitted variables & errors in measurement
    - **Population Linear Regression Model**

$$Y_i = \beta_0 + \beta_1 X_i + u_i,\ i = 1,\dots, n$$

        - X is the ***independent variable*** or ***regressor***
        - Y is the ***dependent variable***
        - β_0 = ***intercept***
        - β_1 = ***slope***
        - u_i = the regression ***error***
        -
        - The ordinary least squares estimator
            - How to estimate $ß_0$ and $ß_1$ from data?
            - **Ordinary least Squares - OLS**

$$\min_{b_0, b_1} \sum_{i=1}^{n} [Y_i - (b_0 + b_1 X_i)]^2$$

                -
                - Minimizes the sum of squared difference between the actual values of Y_i and the prediction based on the estimated line
                - **Minimization of OLS (Discussion)**

- **THE OLS ESTIMATOR, PREDICTED VALUES, AND RESIDUALS**

  The OLS estimators of the slope β_1 and the intercept β_0 are

  $$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})}{\sum_{i=1}^{n}(X_i - \overline{X})^2} = \frac{s_{XY}}{s_X^2} \tag{4.7}$$

  $$\hat{\beta}_0 = \overline{Y} - \hat{\beta}_1\overline{X}. \tag{4.8}$$

  The OLS predicted values $\hat{Y}_i$ and residuals $\hat{u}_i$ are

  $$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i,\ i = 1, \ldots, n \tag{4.9}$$

  $$\hat{u}_i = Y_i - \hat{Y}_i,\ i = 1, \ldots, n. \tag{4.10}$$

  The estimated intercept ($\hat{\beta}_0$), slope ($\hat{\beta}_1$), and residual ($\hat{u}_i$) are computed from a sample of n observations of X_i and Y_i, $i = 1, \ldots, n$. These are estimates of the unknown true population intercept (β_0), slope (β_1), and error term (u_i).

- The OLS estimator → **STATA output**
- Measure of Fit
  - Regression R^2
    - The fraction of the variance of Y that is explained by X. [0, 1]
    - $$R^2 = \frac{ESS}{TSS} = \frac{\sum_{i=1}^{n}(\hat{Y}_i - \overline{Y})^2}{\sum_{i=1}^{n}(Y_i - \overline{Y})^2}$$ → estimated VS true
    - 0 → predict the mean for Y_i (gain nothing)
    - 1 → ESS = TSS → perfect fit
  - Standard error of the regression (SER)
    - The magnitude of a typical regression residual in the units of Y
    - $$SER = \sqrt{\frac{1}{n-2}\sum_{i=1}^{n}\hat{u}_i^2}$$
    - The smaller the residuals, the better the fit.
    - Measure the average size of the OLS residual
  - Ex: **R^2 = 0.05, SER = 18.6**
    - Only a small fraction of the variation in test scores (5%)
- The least squares assumptions
  - Properties of the OLS estimator
    - Unbiased

        - Small variance
    - **Assumptions (LSA)**
        - The conditional distribution of u given X has mean 0 → $ß_1$ is unbiased
        - (X_i, Y_i) are **<u>i.i.d</u> (Independent Identical Distribution)**
            - Non-i.i.d when data are recorded over time
        - Large outliers in X/ Y are rare
    - All about random sampling
    - Sampling distribution of the OLS estimator
        - Difference samples give different $ß_1$
        - We want to
            - Quantify the sampling uncertainty associated with $ß_1$
            - Use $ß_1$ to test hypothesis
            - Construct a confidence interval

- Probability Framework for Linear Regression
- The sampling distribution of $ß_1$

    - What is $E(\hat{\beta}_1)$? (where is it centered?)
        - If $E(\hat{\beta}_1) = \beta_1$, then OLS is unbiased – a good thing!
    - What is $\text{var}(\hat{\beta}_1)$? (measure of sampling uncertainty)
    - What is the distribution of $\hat{\beta}_1$ in small samples?
        - It can be very complicated in general
    - What is the distribution of $\hat{\beta}_1$ in large samples?
        - It turns out to be relatively simple – in large samples, $\hat{\beta}_1$ is normally distributed.

    - 38
    - The mean
        - Unbiased estimator of $ß_1$
    - The variance
        - $$\text{var}(\hat{\beta}_1) = \frac{1}{n} \times \frac{\text{var}[(X_i - \mu_x)u_i]}{\sigma_X^4}$$
        - **The larger the variance of X, the smaller the variance of $ß_1$**
        - Inversely proportional to n
            - The larger the sample, the smaller the $ß_1$
    - Large sample distribution
    - **Summary**

If the three Least Squares Assumptions hold, then

- In **large samples**
  - $\hat{\beta}_1 \xrightarrow{p} \beta_1$ (that is, $\hat{\beta}_1$ is consistent, LLN)
  - $\dfrac{\hat{\beta}_1 - \beta_1}{\sqrt{\dfrac{\sigma_v^2}{n\sigma_X^4}}} \sim N(0,1)$ (CLT)

- - *This parallels the sampling distribution of* $\bar{Y}$

==

Note Four_One

==

- $E(ß_1) = ß_1$ → OLS is unbiased
- The mean and variance of the sampling distribution of $ß_1$

Some preliminary algebra:

$$Y_i = \beta_0 + \beta_1 X_i + u_i$$
$$\bar{Y} = \beta_0 + \beta_1 \bar{X} + \bar{u}$$
so $$Y_i - \bar{Y} = \beta_1(X_i - \bar{X}) + (u_i - \bar{u})$$

Remember from our previous derivation:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{n}(X_i - \bar{X})^2}$$

- Plugging the equation above we have

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})[\beta_1(X_i - \bar{X}) + (u_i - \bar{u})]}{\sum_{i=1}^{n}(X_i - \bar{X})^2}$$

- The result is the **OLS estimators of β_0 and β_1**, which we will
- denote $\hat{\beta}_0$ and $\hat{\beta}_1$
- **Question notes4_1 page 5**

- Final equation is

- $$\boxed{\hat{\beta}_1 - \beta_1 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})u_i}{\sum_{i=1}^{n}(X_i - \bar{X})^2}}$$

- **<u>LIE?</u>**
- $E(ß_1\text{-bar}) - ß_1 = 0$
- Variance of $ß_1$
    - $$\text{var}(\hat{\beta}_1) = \frac{1}{n} \times \frac{\text{var}[(X_i - \mu_x)u_i]}{\sigma_X^4}$$

===

Note Five

===

$$Y_i = \beta_0 + \beta_1 X_i + u_i,\ i = 1,\ldots, n$$

$\beta_1 = \Delta Y/\Delta X$ (***causal effect***)

The Least Squares Assumptions:

1. $E(u|X = x) = 0$.
2. (X_i, Y_i), $i = 1,\ldots,n$, are i.i.d.
3. Large outliers are rare

The Sampling Distribution of $\hat{\beta}_1$:

Under the LSA's, for n large, $\hat{\beta}_1$ is approximately distributed,

- $$\boxed{\hat{\beta}_1 \sim N\left(\beta_1, \frac{\sigma_v^2}{n\sigma_X^4}\right), \text{ where } v_i = (X_i - \mu_X)u_i}$$
- $SE(ß_1\text{-bar})$

- **Hypothesis testing**
    - General approach
        - t = (estimator - hypothesized value) / (standard error of the estimator)

    - • ***For testing the mean of Y***: $t = \dfrac{\bar{Y} - \mu_{Y,0}}{s_Y/\sqrt{n}}$

      • ***For testing β_1***, $\boxed{t = \dfrac{\hat{\beta}_1 - \beta_{1,0}}{SE(\hat{\beta}_1)}}$

    - SE($ß_1$) is a square root of an estimator of the variance of the sampling distribution of $ß_1$ (given by STATA)
- **Reject at 5% significance level if |t| > 1.96**
- **Reject if the p-value < 5%**