Unit 1: Description Statistics

Sample Variance

Denoted by s^2 , is given by

$$s^2 = \frac{\sum(x_i - x)^2}{n-1}$$

Standard Deviation

$$s = \sqrt{x^2}$$

Properties:

1. If $y_1 = x_1 + c$, $y_2 = x_2 + c$ … $y_n = x_n + c$, then $s_y^2 = s_x^2$
2. If $y_1 = cx_1$, $y_2 = cx_2$ … $y_n = cx_n$, then $s_y = |c|\ s_x$

Sample covariance

$$cov(x,y) = \frac{\sum_i (x_i - x)(y_i - y)}{n-1}$$

Sample correlation

$$r = \frac{cov(x,y)}{s_x s_y}$$

Measures the linear relationship between to variables between -1 and 1.

Properties:

1. Adding c to var x or y doesn't change $r_{x,y}$
2. Multiplying a non-zero c to var x or y doesn't change $r_{x,y}$

Unit 3: Probability Theory

Sample space S = set of outcomes (> 1)

Event space F a set of events

Axioms

1. P(A) >= 0 for any event A in F
2. P(S) = 1
3. For any countably infinite set of disj events {A_1, A_2, …}

$$P(A_1 \cup A_2 \cup \dots) = \sum_{i=1}^{\infty} P(A_i)$$

Probability space is the triple (S, F, P)

Conditional Probability

$$P(X|Y) = \frac{P(X \cap Y)}{P(Y)}$$

Multiplication Rule

$$P(X \cap Y) = P(X|Y)P(Y)$$
$$P(X \cap Y \cap Z) = P(X|Y)P(Y)P(Z|X \cap Y)$$

Independence

$$P(X \cap Y) = P(X)P(Y)$$

If X and Y are independent, or P(X|Y) = P(X)

Unit 5: Joint Distributions

Unit 3: Probability Theory (con't)

Baye's Rule

$$P(Y|X)P(X) = P(X|Y)P(Y)$$
$$P(X|Y) = \frac{P(Y|X)P(X)}{P(Y)}$$

Law of Total Probability

$$P(Y) = P(Y|X)P(X) + P(Y|!X)P(!X)$$

Unit 4: Random Variables

Discrete Random Variable

A D.R. V. X is described by a prob func. f

$$f(x) = P(X = x)$$

Probability space (S, F, P)

$$Ox = \{x \in R : f(x) > 0\}$$

Cumulative Probability Function

$$F(x) = P(X \le x) = \sum_{y \in O_x : y \ge x} f(y)$$

Expectation

X be a D.R.V. with outcomes O={x_1, x_2, …, x_k} and prob func. F

$$E(X) = \sum_{j=1}^{k} x_j f(x_j)$$

Expectation for Uniform Distribution

$$E(X) - \frac{1}{k}\sum_j x_j$$

Expectations of Functions of Random Variables

$$E(g(x) = \sum_x g(x)f(x)$$

Expectation as a Linear Function

Let g(x) = ax + b

$$E\big(g(x)\big) = g\big(E(X)\big)$$

Continue Random Variable

A C.R.V> X is described by a prob func. f

$$P(a \le X \le b) = \int_{x=a}^{b} f(x)dx$$

Cumulative Probability Function

$$F(X) = \int_{-\infty}^{x} f(x)dx$$

$$f(x) = F'(x)\ when\ F'(x)\ exists$$

Unit 4: Random Variables (con't)

Expectation

$$E(X) = \int_{-\infty}^{\infty} xf(x)dx$$

If h: R→R, H(X) is a R.V.

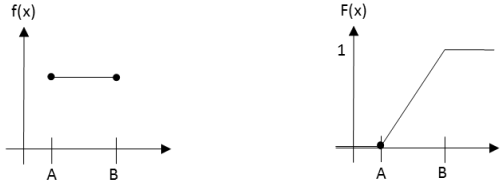
$$E\big(h(x)\big) = \int_{-\infty}^{\infty} h(x)f(x)dx$$

If h(x) = ax + b

$$E\big(h(x)\big) = aE(x) + b$$

Uniform Random Variable

X has a uniform distribution on [A,B] if it has pdf



$$f = \begin{cases} \frac{1}{B-A}, A \le x \le B \\ 0, otherwise \end{cases}$$

$$F(X) = \frac{x - A}{B - A}$$

$$E(X) = \frac{A + B}{2}$$

Variance Page

$$var(X) = E(X^2) = [E(X)]^2$$

$$\sigma_x = \sqrt{var(X)}$$

$$var(X + c) = var(X)$$

$$var(cX) = c^2 var(X)$$

$$f_Y(y) = \int_{x=-\infty}^{\infty} f_{X,Y}(x,y)dx = \int_{x=-\infty}^{\infty} f_{Y|X}(y|x)f_X(x)dx$$

This is just the law of total probability

Independence

$$f_{X|Y}(x|\,y) = \frac{f_{X,Y}(x,y)}{f_Y(y)}$$

$$Pr(X = x|Y = y) = \frac{\Pr\,(X = x\ and\ Y = y)}{Pr(Y = y)}$$

X and Y are independent if

$$f_{X,Y}(x,y) = f_x(x)\,f_y(y)$$

Expectations of Sums

$$E(X+Y) = E(X) + E(Y)$$

Expectations of Products

If X and Y are independent

$$E(XY) = E(X)\,E(Y)$$

Covariance

R. V. X, Y, μ_x=E(X), μ_y=E(Y),

$$cov(X,Y) = E((X-\mu_x)(Y-\mu_y)$$

1. $cov(X,X) = var(X)$
2. $cov(X,Y) = E(XY) - E(X)E(Y)$
3. $Let\ a,b \in R$
   $cov(aX,By) = ab\ cov(X,Y)$
4. $Let\ R.V.\ Z, \mu z = E(Z)$
   $cov(X,Y+Z) = cov(X,Y) + cov(X,Z)$
5. $\rho_{X,Y} = \frac{cov(X,Y)}{\sigma_x\sigma_y}$, $\rho_{X,Y}\ is\ the\ correlation$

Variance of Sums

$$var(X+Y) = var(X) + var(Y) + 2cov(X,Y)$$

If X and Y are independent,

$$var(X+Y) = var(X) + var(Y)$$

Variance of Differences

If X and Y are independent,

$$var(X-Y) = var(X) + var(Y)$$

The variance of differences of RVs is the **sum** of the two RV's variances.

Law of Iterated Expectations

$$E(Y) = E_x(E(Y|X))$$

Unit 8: Hypothesis Testing (con't)

Unit 6: Sampling and the Central Limit Theorem

The Central Limit Theorem

Let $X_1, X_2, \ldots X_n$ be a random sample from a distribution with mean μ and variance σ^2. Then if n is sufficiently large (>30), $\overline{X}$ has approximately a normal distribution with $\mu_{\overline{X}} = \mu$ and $\sigma_{\overline{X}} = \frac{\sigma^2}{n}$.

Sampling Distribution of the Mean

Let $X_1, X_2, \ldots X_n$ be iid with mean μ and variance σ^2.

$$E(\overline{X}) = \mu$$

$$\sigma_{\overline{X}} = \frac{\sigma}{\sqrt{n}}$$

Unit 7: Estimators

- Is calculated from our sample
- Approximates the population parameter
- Changes from sample to sample (random variable)

Bias

If you took a huge # of samples, the average of your estimator from each one would equal to your parameter.

$$E(\hat{\theta}) = \theta$$

The **bias** of your estimator is defined as

$$Bias(\hat{\theta}) = E(\hat{\theta}) - \theta$$

Consistency

A **consistent** estimator is one for which bias approaches 0 for large samples.

$$Prob(|\hat{\theta} - \theta| > \epsilon) \to 0\ as\ n \to \infty$$

Efficiency

Estimator $\hat{\theta}_1$ is relatively **efficient** to estimator $\hat{\theta}_2$ if $var(\hat{\theta}_1) \leq var(\hat{\theta}_2)$ for all possible values of θ with at least one θ such that $var(\theta_1) < var(\theta_2)$.

Efficiency means we are more confident of being closer to the true population parameter.

Method of Moments

- The idea is we write down equations that involve the moments of the population model
- A population model includes the expectation of X and high powers of X: E(X), $E(X^2)$, $E(X^3)$, etc.
- Equations involving these moments are written down (ie moment conditions)
- The population moments are then replaced with the sample counterparts

Unit 7: Estimators (con't)

Maximum Likelihood

Let $X_1, X_2, \ldots X_n$ with pdf $f(x_1, x_2, \ldots, x_n; \theta)$

Define: Likelihood, $L(\theta) = f(x_1, x_2, \ldots, x_n; \theta)$

Idea: $\hat{\theta}_{ML}$ = argmax $L(\theta)$

Steps to estimate θ:

1. Define $L(\theta) = f(x_1, x_2, \ldots, x_n; \theta)$
2. Take the ln of $L(\theta)$, makes calculations easier
3. Take the derivative w.r.t. θ
4. Set derivative to 0 and solve for θ

Confidence Interval

95% C.I. is

$$(\overline{X} - 1.96 * \frac{\sigma}{\sqrt{n}}, \overline{X} + 1.96 * \frac{\sigma}{\sqrt{n}})$$

based on z-distribution.

- Centered at sample mean
- Extends
- The interval is random, since each endpoint is a R.V.
- μ is fixed. The C.I. varies from sample to sample
- 95% C.I. if we repeat the same calculation for different samples, there is a 95% change C.I. contains μ

Unit 8: Hypothesis Testing

- We cannot prove H_A
- All we can do is reject H_0

The Z-test as a Hypothesis Test

$$z = \frac{\overline{X} - \mu_0}{\frac{\sigma}{\sqrt{n}}}$$

z-test assumes we know the population standard deviation, which is not a realistic situation. When we don't know the population standard deviation, we use the t-test.

t-Test

$$T = \frac{\overline{X} - \mu_0}{\frac{s}{\sqrt{n}}} \sim T_{n-1}$$

The Meaning of p-Value

The p-val of a test is the chance of getting a big test statistic – assuming the null hypothesis to be right. P-pval is **not** the chance of the null hypothesis being right.

Type I and Type II Errors

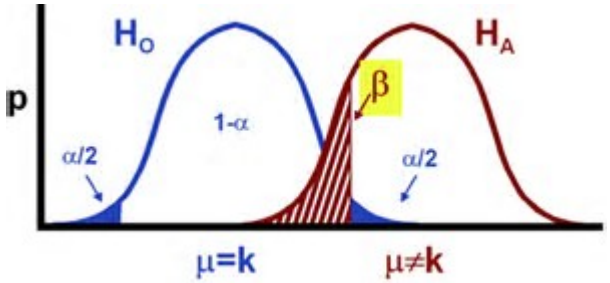
Type 1 Error: Falsely rejecting a null hypothesis, denoted by α.

- The α value is the statistical significance level (e.g. 0.05)

Type 2 Error: Failing to reject the null hypothesis when it is false (false negative).

- Denoted by β. Anything we do to reduce α will increase β

Visualizing α and β



Statistical Power

The probability of a Type II error (β) is the probability to miss a potentially important finding by retaining the null when it is actually false.

1- β is know as **statistical power.**

Unit 9: Comparing Two Means - Independent & Dependent Tests

Independent Sample t-Test

$$t = \frac{M_1 - M_2}{S_{DM}}$$

$$S_{DM} = sqrt\left(\frac{(N_1-1)^2 + (N_2-1)^2}{N_1 + N_2 - 2} * \left(\frac{1}{N_1} + \frac{1}{N_2}\right)\right)$$

Assumptions Underlying the Independent Sample Test

1. Assumption of Normality: Variables are normally distributed within each group.
2. Assumption of Homogeneity: The variation of scores in the two groups are roughly equal (Levene's test)

Doing an Independent Sample t-Test

1. State your hypothesis (one-sided or two sided)
2. Check the assumption of normality – If variables are not normal, either transform variable or use non-parametric test
3. Check assumption of equal variances (Levene's test). If does not hold true, use Welch's t-test)

Effect Size (Cohen's d)

- Measure of practical significance, not statistical significance
- Estimates a population parameter, and is not effected by sample size

$$d = \frac{M_1 - M_2}{S_{pooled}}$$

$$s_{pooled} = sqrt\left(\frac{(N_1-1)S_1^2 + (N_2-1)S_2^2}{N_1 + N_2 - 2}\right)$$

- > 0.8 → Large, around 0.5 → Medium, < 0.2 → Small

Unit 9: Comparing Two Means - Independent and Dependent Tests (con't)

Effect Size Correlation (r)

$$r = \frac{t}{sqrt(t^2 + df)}$$

- r = 0.5 → large, r = 0.3 → medium, r = 0.1 → small

Dependent t-Test

- Dependent t-test compares two means (of the same variable) taken at two different points in time for the same group of cases.

Attempts to reduce the underlying variation so that the difference in means can be measured with precision

Non-Parametric Tests

- For small sample and have major deviation from normality → use a non-parametric test
- Works on principle of ranking data
  - List scores from highest to lowest
  - Just consider the ranks instead of looking at the metric value of the variable
  - Use the order of variables to construct statistics that we can use to test hypothesis

Type of Design	Parametric Tests	Non-Parametric Tests
Two Independent Samples	Independent sample t-test	Wilcoxon Rank-Sum Test (Mann-Whitney test)
Two Dependent Samples	Dependent sample t-test	Wilcoxon Signed-Rank Test

Wilcoxon Rank-Sum Test

- Default is two-sided test
- Null-hypothesis: No difference in ranks
- There are always two W values, one per group. Lowest W is typically used

Wilcoxon Signed-Rank Test for Related Conditions

- Nonparametric equivalent of the dependent test
- Allows us to examine the diff in scores among the same cases
- Calculation of ranks similar to rank-sum test, excel the focus is on the diff between the first score and the second score
- Diff can be positive or negative (0s excluded)
- Scores ranked by absolute value
- After ranking, the positive and negative ranks are summed separately
- Between the positive and negative sum of ranks, smaller value is used as the test statistic

Unit 10: Bivariate Ordinary Least Squares Estimation

- Regression assumes a one-way link between X and Y
- Assumption has to be justified by our theory
- Regression **does not** assume a causal relationship
- Correlation measures whether two variables have a linear relationship (as well as strength of that relationship
- Regression is a more precision description of linearity

Simple Regression Population Model

$$y = \beta_0 + \beta_1 x + u$$ ← Error term

Dependent variable | Intercept | Slope parameter | Independent variable

Constraining the Error

- Assume the error has mean 0, E(u) = 0
- Zero Conditional mean: E(u|x) = 0
  - Even at a specific value of x, we expect errors to average 0
  - Explanatory variable must not contain info about the mean of any unobserved factors

OLS as Error Minimization

The regression residuals are our estimated errors

$$\hat{u} = y_i - \hat{y}_i = y_i - \widehat{\beta_0} - \widehat{\beta_1} x$$

We minimize the sum of the squared residuals

$$\min \sum_{i=1}^{n} \widehat{u^2} \rightarrow \widehat{\beta_0}, \widehat{\beta_1}$$

Solving the minimization problem, we arrive at the OLS estimates:

$$\beta_1 = \frac{cov(x_i, y_i)}{var(x_i)}$$

$$\beta_0 = \bar{y} - \widehat{\beta_1}\,\bar{x}$$

Algebraic Properties of the OLS Estimators

The (estimated errors sum to zero $\sum_{i=0}^{n} \widehat{u_i} = 0$

Correlation between residuals and regressors is zero $\sum_{i=0}^{n} x_i \widehat{u_i} = 0$

The sample averages of y and x lie on a regression lie $\bar{y} = \widehat{\beta_0} + \widehat{\beta_1} x$

OLS Coefficients as Random Variables

- The coefficeints are random variables
- They will differe from sample to sample
- So the line we fit varies and will be different form the true population regression line

Unit 10: Bivariate Ordinary Least Squares Estimation (con't)

Goodness of Fit: Measure of Variation

1. Total Sum of Squares
   Represents total variation in the dependent variable
$$SST = \sum_{i=1}^{n}(y_i - \bar{y})^2$$
2. Explained Sum of Square
   Represents variation explained by regression
$$SST = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2$$
3. Residual Sum of Squares
   Represents variation not explained by regression
$$SSR = \sum_{i=1}^{n}\hat{u}_i^{\,2}$$

SST = SSE + SSR

Total variation = explained part + unexplained part

R-Squared

- Measures the fraction of total variation explained by the regression
- Requires all OLS assumptions for correct interpretation
- High R-squared tells us that a lot of variation in our y variable is explained by our model
- R-squared can be considered a measure of predictive accuracy

$$R^2 = \frac{SSE}{SST} = 1 - \frac{SSR}{SST}$$

Standard Assumptions for Bivariate Linear Regression

Assumption SLR.1 (Linear in Parameters)
In the population, the relationship between x and y is linear.
$$y = \beta_0 + \beta_1 x + u$$

Assumption SLR.2 (Random Sampling)
The data is a random sample drawn from the population.
$$\{(x_i, y_i): i = 1..n\}$$

Assumption SLR.3 (Sample Variation in Explanatory Variable)
Not all values of the explanatory variables are equal.
$$\sum_{i=1}^{n}(x_i - \bar{x})^2 > 0$$

Assumption SLR.4 (Zero Conditional Mean)
The value of the explanatory variable must contain no information about the mean of the unobserved factors.
$$E(u_i|x_i) = 0$$

Unit 10: Bivariate Ordinary Least Squares Estimation (con't)

Unbiasedness of OLS (Theorem 2.1)
$$SLR.1 - SLR.4 \rightarrow E(\widehat{\beta_0}) = \beta_0, E(\widehat{\beta_1}) = \beta_1$$

Unit 11: Multiple OLS Regression

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + u$$
$$E(u) = 0, cov(x_i, u) = 0, \forall i$$

β_j represents the expected change in y from a unit change in x_j, holding u and all other x terms constant.

Matrix Notation
$$Y = X\beta + u$$
$$\beta = (X^T X)^{-1} X^T Y$$

Measure of Fit: Multiple R^2

- Multiple R^2: the square of the correlation between the observed Y values and the Y values predicted by the multiple regression
- Multiple R^2 **always** goes up as more variables are incorporated into the model.

Measure of Fit: Akaike Information Criterion (AIC)

- Aka "parsimony-adjusted measure of fit"
- Penalizes model as # of variables increases
- Only useful when applied to models with the same data and same dependent variable
- Larger AIC → worse fit

BLUE – Best Linear Unbiased Estimator

Best well-known benchmark for OLS performance

- Best: relative efficiency
- Linear: OLS estimates are a linear function of the y's
- Unbiased: each $\widehat{\beta_j}$ is an unbiased estimator of the true parameter β_j. $E(\widehat{\beta_j}) = \beta_j$

Theorem 3.1 (Unbiasedness of OLS)

Under MLR.1-4, OLS estimates are unbiased. β_j. $E(\widehat{\beta_j}) = \beta_j$

MLR Assumptions

Assumption MLR.1 (Linear in Parameters)
In the population, the relationship between x and y is linear.
$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + u$$

Assumption MLR.2 (Random Sampling)
The data is a random sample drawn from the population.
$$\{(x_{i1}, x_{i2}, \ldots, x_{ik}, y_i): i = 1..n\}$$

Unit 11: Multiple OLS Regression (con't)

Assumption MLR.3 (No Perfect Collinearity)
None of the independent variables are constant, and no exact relationships among the independent variables. Rules out only perfect collinearity between explanatory variables, imperfect correlation is allowed but can greatly increase errors.

Assumption MLR.4 (Zero Conditional Mean)
The value of the explanatory variables must contain no information about the mean of the unobserved factors.
$$E(u_i|x_{i1}, x_{i2}, \ldots, x_{ik}) = 0$$

Troubleshooting the Bias Assumptions

Random Sampling
Common ways where assumption fails:
1. Clustering
2. Autocorrelation or serial correlation - Occurs when error for one data point is correlated for error for next data point

Multicollinearity
Drop redundant variables

Zero Conditional Mean
Use residuals vs fitted values plot.
- If conditional mean of error is not constant, change functional form
- Adding new variables may fix issue
- Sometimes may not be able to meet zero-conditional mean, but able to meet a weaker assumption: exogeneity

Endogeneity

- Endogeneous – When explanatory variables are correlated with error term
- Endogeneous **not** a statement about causality
- Endogeneity is a violation of zero-conditional mean, and presence implies that OLS coefficients are biased and inconsistent.

Exogeneity

- Assumption MLR.4' (Exogeneity): $cov(x_j, u) = 0, \forall j$
- A weaker assumption than zero-conditional mean, but easier to meet and more realistic.

Theorem:

Under MLR.1-3 and MLR.4', the OLS estimators are consistent.
$$plim_{n\to\infty}(\widehat{\beta_j}) = \beta_j$$

Unit 11: Multiple OLS Regression (con't)

Leverage vs Influence

- Leverage – Amount of potential each data point has to change the regression
- Influence – Amount by which each data point actually changes the regression

Leverage

- h_{ii} = measure of how sensitive the regression is to a given data point i
- h_{ii} = ratio of how much the regression line moves compared to how much you move the data point
- imagine moving a data point up and down the y-axis and observing how much the regression line follows it.

Influence

- Cook's distance measures the actual effect of deleting given observation
- Each point with a large cook's distance should be examined and potentially eliminated from analysis
- Cook's dist > 1 may be problematic
- Some cases can exert huge influence and still produce small residuals

---

Unit 12: Multiple OLS Regression Inference

Homoscedasticity

Assumption MLR.5 (Homoscedasticity)**:** Variance of the error term is constant. E.g. error term cannot vary more for some value of x's than others.

$$var(u_i|x_{i1}, x_{i2}, \dots, x_{ik}) = \sigma^2$$

Sample Variance of OLS Estimators

$$var(\widehat{\beta_j}) = \frac{\sigma^2}{SST_j(1 - R_j^2)}, j = 1, \dots, k$$

σ^2 = variance of error term

SST_j = total sample variation in explanatory variable x_j

R_j^2 - R_2 from a regression of x_j on all other independent variables

The Gauss-Markov Theorem

Under Theorem 3.4 (Gauss-Markov Theorem), the OLS estimators are the best linear unbiased estimators (BLUE) of the regression coefficients.

For any other linear unbiased estimators with coefficients, $\widetilde{\beta_j}$

$$var(\widehat{\beta_j}) \leq var(\widetilde{\beta_j}), j = 1, \dots, k$$

Unit 12: Multiple OLS Regression Inference (con't)

Troubleshooting Homoscedasticity

- Residuals vs fitted plot – If errors were homoscedastic, the band would be of uniform thickness
- Scale-location plot – homoscedastic errors should appear as a horizontal band of points
- Breush-Pagan test can be used to check for heteroscedasticity. Null hypothesis is that there is homoscedasticity. Sample size matters greatly.

Responding to Heteroscedasticity

- Simplest solution is to switch to heteroscedasticity-robust tools (e.g. White standard errors)
- Try modelling the log of variable.

Normal Error Term Assumption

To infer the sampling distribution of our OLS coefficients, add assumption about shape of the error distribution.

Assumption MLR.6 (Normality of Error Terms)**:**
Assume errors are drawn from normal distribution with mean zero.

$$u_i \sim N(0, \sigma^2) indepndently\ of\ x_{i1}, x_{i2}, \dots, x_{ik}$$

Classical Linear Model Assumptions

MLR.1 – MLR.5: Gauss-Markov assumptions
MLR.1-MLR.6**:** Classical Linear Model (CLM) assumptions

Theorem 4.1 (Normal sampling distribution)

Under MLR.1-MLR.6, the OLS coefficients are normally distributed. Each $\widehat{\beta_j}$ is normally distributed around the true parameter.

$$\widehat{\beta_j} \sim N(\widehat{\beta_j}, var(\widehat{\beta_j}))$$

Troubleshooting the Normality Assumption

- Example regression diagnostics after you fit a linear regression
- Hist of residuals
- QQ-plot of residuals
- Shapiro-Wilk test on your residuals

Responding to Normality Violations

- If you have a large dataset, rely on asymptotic properties of OLS. OLS estimates are normally distributed for large sample sizes (about N >= 30)
- If n is small and residuals do not look normal, look for an alternate specification that meets normality (log y)
- Bootstrapping – resample from data in order to estimate sampling dist of coefficients.

Unit 12: Multiple OLS Regression Inference (con't)

Large-Sample Properties

- As long as we have a large sample and use heteroscedasticity-robust standard errors, we generally focus on MLR.1-3 and MLR4'.
- Under these, OLS estimators are consistent
- CLT tells us that our coefficients have an asymptotically normal distribution
- For large samples, there are two key assumptions to focus on
  - Random sampling: are observations correlated in some way?
  - Exogeneity: Is any x correlated with the error? And is there some unmeasured factor that ends up in the error that is related to an x?

Theorem 5.2 (Asymptotic normality of OLS):

$$\frac{\widehat{\beta_j} - \beta_j}{SE(\widehat{\beta_j})} \sim N(0,1) AS\ n \to \infty, also\ plim_{n\to\infty}\ \widehat{\sigma^2} = \sigma^2$$

Unit 13: Linear Model Specification

Testing if Parameters are Different

H_0: $\beta_1 - \beta_2 = 0$

$$t = \frac{\widehat{\beta_1} - \widehat{\beta_2}}{se(\widehat{\beta_1} - \widehat{\beta_2})}$$

$$se(\widehat{\beta_1} - \widehat{\beta_2}) = sqrt\left(\widehat{var}(\widehat{\beta_1} - \widehat{\beta_2})\right) = sqrt(var(\widehat{\beta_1}) + var(\widehat{\beta_2}) - 2cov(\widehat{\beta_1}, \widehat{\beta_2}))$$

Alternative – Change the variables
Define $\theta_1 = \beta_1 - \beta_2 \rightarrow H_0: \theta_1 = 0$
This method can be generalized to test any linear combination of parameters.

Joint Significance

Exclusion restriction – testing whether variables could be excluded from model.

$$H_0: \beta_3 = 0, \beta_4 = 0, \beta_5 = 0$$

Testing Coefficients Jointly

Remove from regression, see how much worse model fit is. Now model = restricted model. Model fit: SSR.

Unit 13: Linear Model Specification (con't)

Forming Test Statistic

$$F = \frac{(SSR_r - SSR_{ur})/q}{SSR_{ur}/(n-k-1)} \sim F_{q,n-k-1}$$

- Measures relative change in SSR, without constant scaling factors
- Under null hypothesis, and assuming CLM assumptions (MLR.1-6), test statistic follows an F-distribution

Model Significance

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + u$$

- Application of join significance: testing regression model as a whole.
- Omnibus test: can we exclude every x variable at the same time?
- Does the model have any predictive power on the whole?
- Automatically reported in R
- Most of the time, null hypothesis automatically rejected.

Variable Transformations

Base 10 Log

Helpful when thinking y in terms of powers of 10

Natural Log

Gives elegant interpretation for slope coefficients.
Say $\beta_1 = 0.15$, expect +1 in x_1 to result in in 15% increase in y

Log-log Form

Measuring % increase in y, per 1% increase in x.
In economics, coefficient log-log model is called elasticity.

Logarithm Rules of Thumb

- Look for variables that are naturally always positive
- Look for variables that have meaningful zero point but no obvious max
- Look for variables where % change is meaningful
- Log mitigates influence of outliers in positive direction
- Logs can help secure normality and homoscedasticity for OLS

Unit 13: Linear Model Specification (con't)

Quadratic

- Y on X^2
- Powers less than 1: corrects negative skew

Higher Order Polynomials

Include all lower-order terms (if x^2, then x too)
Interpret all coefficients simultaneously to understand effect of variable.

Indicator Variables

- Categorical variables can be nominal or ordinal
- To put categorical variables into regression model, typically use indicator variables 1 (true), 0 (false)
- Must omit one category (base category) to avoid collinearity.

Ordinal Variables

- Generally wrong to place ordinal variable directly into population model, would impose linear structure on variable.
- Use indicator variables for each category, allowing the effect of each one to vary independently.

Interaction Term

- Term in regression where two variables are multiplied together.
- Simple way to make one variable's effect depend on another variable, can make population model more realistic.
- Must be careful interpreting coefficients, understanding dependent on whether continuous or indicator variables.

Interaction of Indicator and Metric Variable

- Interaction terms sometime include three variables
- Need to include all subsets

Interaction of Indicator and Metric Variable

- Interaction term allows both intercepts and slopes to vary independently
- Interaction is the difference between the slops
- OLS will fit best line to both groups independently

Unit 14 : Causal Inference

Causal Models

1. Point 1 – Causal Model
   Causality is a property of our model
   - Assumption is made whether model is causal
   - Our coefficient has a causal interpretation as long as the error term doesn't change as we manipulate x.
2. Point 2 - Ceteris Paribus
   Causality is bound up in our ceteris paribus assumption "all other things equal"
   - To make a causal model, we imagine taking every factor except our x and putting them into the error term.
   - All things we can't compute or measure go to the error term
   - With this assumption, it's plausible that $\frac{\partial u}{\partial x} = 0$
   - We're just manipulating x, all other natural factors that affect y are constant
3. Point 3 - Causality and Exogeneity
   Causality is **not** the same as exogeneity
   - Causality is about whether manipulations to x influences the error term
   - Exogeneity is about whether OLS can correctly estimate our coefficients

Omitted Variable Bias in Simple Regression

True model: $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + u, E(u) = 0$
$cov(x_1, u) = cov(x_2, u) = 0$
Estimated model: $y = \alpha_0 + \alpha_1 x_1 + w$
Let $x_2 = \delta_0 + \delta_{1x_1} + v, E(v) = 0, cov(x_1, v) = 0$

$y = \beta_0 + \beta_1 x_1 + \beta_2(\delta_0 + \delta_1 x_1 + v) + u$
$= (\beta_0 + \beta_2\delta_0) + (\beta_1 + \beta_2\delta_1)x_1 + (u + \beta_2 v)$
$E(u + \beta_2 v) = 0$
$cov(x_1, u + \beta_2 v) = cov(x_1, u) + cov(x_1, \beta_2 v) = 0$

$\alpha_1 = \beta_1 + \beta_2\,\delta_1, \beta_2\,\delta_1$ is the omitted variable bias.

Identification Strategies

An identification strategy is our plan for consistently measuring a causal effect in the face of endogeneity.

1. A True Experiment
2. Difference in Difference – A tool that allows us to remove the effects of time-constant confounders
3. Instrumental variables – Allows us to study situations in which we can't randomly assign treatment, but we can identify a source of variation that we believe is exogeneous and affects the treatment
4. Regression Discontinuity – Idea: exploit a decision rule that determines what individuals receive a treatment based on an attribute

Useful R Commands (Units 1 – 8)

Distribution Functions

dnorm – density function
pnorm – cumulative probability function
qnorm – quantile function (e.g. inverse cumulative prob function)
rnorm – random draws from distribution

dnorm(0) → 0.3989 e.g. 40% chance you get 0 from a random draw of a normal distribution
dnorm(0, mean=1, sd=1) → 0.24197

pnorm(1.96) → 0.975
pnorm(0) – pnorm(-1) → 0.6827 E.g. area between 1 and -1 sd from the mean

qnorm(0.5) → 0 E.g. inverse of the cumulative prob function
qnorm(0.975) → 1.96

rnorm(1) → returns a numeric
rnorm(100) → returns a vector of length 100

Hypothesis Testing

```
t.test(x, y = NULL,
        alternative = c("two.sided", "less", "greater"),
        mu = 0,
        paired = FALSE,
        var.equal = FALSE,
        conf.level = 0.95)

z.test = function(x, mu, var) {
       zeta = (mean(x) - mu) / (sqrt(var / length(a)))
       return(zeta)
}

power.t.test(n=n,
      delta=1.5, sd=s, sig.level=0.05,
       type="one.sample",alternative="two.sided",strict = TRUE)
```

Replication

```
replicate(1000, sample(mean(sample(eruptions, 3, replace=T))))
```

Maximum Likelihood

```
f <- function(p) {
       -p^2 + p + 2
}
optimize(f, interval = c(0,100), maximum = TRUE)
```

Other Commands

cumsum(x) – cumulative sum
seq(1, 1000, 1) – returns a sequence vector from 1 to 1000, by 1
sample (x, n, replace=T) – takes n samples from vector x with replace.
qqnorm(x)

Useful R Commands (Units 9 – 13)

```
library(effsize) # used for Cohen's d
t.test(math4 ~ ben, data=schools) # or
t.test(high_bs_scores, low_bs_scores)
by(Schools$avgsalary, Schools$ben, mean)
cohen.d(math4 ~ ben, data=schools) # measures how many pop sd separates the mean

wilcox.test(math4 ~ ben, data=schools)
t.test(Schools$math4, Schools$story4, paired=T) #paired t-test

plot(model, which=1) # residuals vs fitted values plot
plot(model, which=5) # influence
AIC(model) # Akaike information criterion
BIC(model) # Bayesian information criterion
Stargazer(model1, model2) # regression tables.

library(car)
library(lmtest)
library(sandwich)
shapiro.test(model1$residuals) # check for normality
bptest(model1) # Breush-Pagan test, check for heteroscedasticity

To address heteroscedasticity, use robust standard errors:
Coeftest(model1, vcov=vcovHC) # use instead of summary
vcovHC(model1) # variance-covariance matrix

Regression tables:
se.model1 = sqrt(diag(vcovHC(model1)))
se.model2 = sqrt(diag(vcovHC(model2)))

stargazer(model1, model2, type="text", omit.stat="f",
          se=list(se.model1, se.model2),
          star.cutoffs=c(0.05, 0.01, 0.001))

waldtest(model1, model_rest, vcov=vcovHC) # Generalizes the usual F-test of overall significance, but allows for a heteroskedasticity robust covariance matrix)

linearHypothesis(model1, c("bavg=0", "hrunsyr=0",  "rbisyr=0"), vcov=vcovHC) # from car package
linearHypothesis(model1, c("hits=bb"), vcov=vcovHC)
```

Useful Links (Units 1 – 8)

Distributions

- Uniform distribution (discrete)
- Binomial distribution (discrete)
- Bernoulli distribution (discrete)
- Poisson distribution (discrete)
- Uniform distribution (continuous)
- Normal distribution (continuous)
- Exponential distribution (continuous)
- Gamma distribution (continuous)

Solving Integrations

- Wolfram Alpha (q=integrate x^2 from 0 to 100)

Solving Derivatives

- Wolfram Alpha (q=derive x^2)

Probability

- Probability Calculator

R

Distribution functions

OLS Assumptions

Assumption	Assumption Description	How to Test Assumption	Responding to Assumption Violation
MLR.1 Linear in Parameters	In the population, the relationship between x and y is linear. $y = \beta_0 + \beta_1 x + u$	MLR.1 is always met if the model is specified such that the dependent variable is a linear function of the explanatory variables.	N/A
MLR.2 Random Sampling	The data is a random sample drawn from the population. $\{(x_i, y_i): i = 1..n\}$	Common ways where assumption fails: 1. Clustering 2. Autocorrelation or serial correlation - Occurs when error for one data point is correlated for error for next data point	Re-sample
MLR.3 No Perfect Collinearity	None of the independent variables are constant, and no exact relationships among the independent variables.	• Correlation plot to see if any two independent variables are perfectly correlated, and check their Variance Inflation Factors (VIF) (e.g. if they are less than 10).	• Drop redundant variables • Rules out only perfect collinearity between explanatory variables, imperfect correlation is allowed but can greatly increase errors.
MLR.4 Zero Conditional Mean	The value of the explanatory variables must contain no information about the mean of the unobserved factors. $E(u_i \mid x_{i1}, x_{i2}, \dots, x_{ik}) = 0$	• Use residuals vs fitted values plot • Check if the covariances of the three independent variables with the residuals are very close to 0, indicating that they are exogenous.	. • If conditional mean of error is not constant, change functional form • Adding new variables may fix issue • Sometimes may not be able to meet zero-conditional mean, but able to meet a weaker assumption: exogeneity
MLR.4' Exogeneity	The independent variables are uncorrelated with the error term $cov(x_j, u) = 0, \forall j$	• Check if this is true: $cov(x_j, u) = 0, \forall j$	• Understand if there is an independent variable that is correlated to an unmeasured variable (that is currently being captured in the error term)
MLR.5 Heteroscedasticity	Variance of the error term is constant. E.g. error term cannot vary more for some value of x's than others. $var(u_i \mid x_{i1}, x_{i2}, \dots, x_{ik}) = \sigma^2$	• Residuals vs fitted plot – If errors were homoscedastic, the band would be of uniform thickness • Scale-location plot – homoscedastic errors should appear as a horizontal band of points • Breush-Pagan test can be used to check for heteroscedasticity. Null hypothesis is that there is homoscedasticity. Sample size matters greatly.	• Simplest solution is to switch to heteroscedasticity-robust tools (e.g. White standard errors) • Try modelling the log of variable.
MLR.6 Normality of Error Terms	Assume errors are drawn from normal distribution with mean zero. $u_i \sim N(0, \sigma^2) indepndently\ of\ x_{i1}, x_{i2}, \dots, x_{ik}$	• Example regression diagnostics after you fit a linear regression • Histogram of residuals (do they look normally distributed?) • QQ-plot of residuals (linear, diagonal line in the plot?) • Shapiro-Wilk test on your residuals	• If you have a large dataset, rely on asymptotic properties of OLS. OLS estimates are normally distributed for large sample sizes (about N >= 30) • If n is small and residuals do not look normal, look for an alternate specification that meets normality (log y) • Bootstrapping – resample from data in order to estimate sampling dist of coefficients.

OLS Theorems

Theorem	Description	Assumptions Needed to Hold True
Theorem 2.1 Unbiasedness of OLS	$SLR.1 - SLR.4 \rightarrow E(\widehat{\beta_0}) = \beta_0, E(\widehat{\beta_1}) = \beta_1$	SLR.1 Linear in Parameters SLR.2 Random Sampling SLR.3 Sample Variation in Explanatory Variable SLR.4 Zero Conditional Mean
Theorem 3.1 Unbiasedness of OLS	Under MLR.1-4, OLS estimates are unbiased. β_j. $E(\widehat{\beta_j}) = \beta_j$	MLR.1 Linear in Parameters MLR.2 Random Sampling MLR.3 No Perfect Collinearity MLR.4 Zero Conditional Mean
Theorem 3.4 The Gauss-Markov Theorem	The OLS estimators are the best linear unbiased estimators (BLUE) of the regression coefficients. For any other linear unbiased estimators with coefficients, $\widetilde{\beta_j}$ $var(\widehat{\beta_j}) \leq var(\widetilde{\beta_j}), j = 1, \dots, k$	MLR.1 Linear in Parameters MLR.2 Random Sampling MLR.3 No Perfect Collinearity MLR.4 Zero Conditional Mean MLR.5 Heteroscedasticity
Theorem	Under MLR.1-3 and MLR.4', the OLS estimators are consistent. $plim_{n\to\infty}(\widehat{\beta_j}) = \beta_j$	MLR.1 Linear in Parameters MLR.2 Random Sampling MLR.3 No Perfect Collinearity MLR.4' Exogeneity
Theorem 4.1 Normal Sampling Distribution	Under MLR.1-MLR.6, the OLS coefficients are normally distributed. Each $\widehat{\beta_j}$ is normally distributed around the true parameter. $\widehat{\beta_j} \sim N(\widehat{\beta_j}, var(\widehat{\beta_j}))$	MLR.1 Linear in Parameters MLR.2 Random Sampling MLR.3 No Perfect Collinearity MLR.4 Zero Conditional Mean MLR.5 Heteroscedasticity MLR.6 Normality of Error Terms
Theorem 5.2 Asymptotic normality of OLS	$\frac{\widehat{\beta_j} - \beta_j}{SE(\widehat{\beta_j})} \sim N(0,1)\ AS\ n \to \infty, also\ plim_{n\to\infty}\ \widehat{\sigma^2} = \sigma^2$	MLR.1 Linear in Parameters MLR.2 Random Sampling MLR.3 No Perfect Collinearity MLR.4 Zero Conditional Mean MLR.5 Heteroscedasticity

Considerations in Hypothesis Testing

Consideration	What is Needed	Example in R
One-sided vs Two Sided	Context or Domain Knowledge – Do you have reason to believe that the difference in mean will be greater? (or less than?)	t.test(S_q1$libcpre_self_scaled, S_q1$libcpo_self_scaled, **alternative="two.sided"**, paired=T) # Default is one-sided
Independent vs Dependent Test	Data structure - Are the two means from the same variable taken at two different points in time for the same group of cases?	E.g. t.test(Schools$math4, Schools$story4, **paired=T**) Paired=T → dependent test Paired=F → independent test
Parametric vs Non-Parametric Test	Assumption of normality Variables are normally distributed within each group. Assumption of homogeneity The variation of scores in the two groups are roughly equal (Levene's test) **Type of Design** / **Parametric Tests** / **Non-Parametric Tests** Two Independent Samples / Independent sample t-test / Wilcoxon Rank-Sum Test (Mann-Whitney test) Two Dependent Samples / Dependent sample t-test / Wilcoxon Signed-Rank Test	Parametric t.test(S_q1$libcpre_self_scaled, S_q1$libcpo_self_scaled, alternative="two.sided", paired=T) Non-Parametric (Wilcoxon Rank-Sum) wilcox.test(score_diff ~ dr, data=S_q4, alternative="two.sided", **paired=F**, conf.int=T) Non-Parametric (Wilcoxon Signed-Rank) wilcox.test(score_diff ~ dr, data=S_q4, alternative="two.sided", **paired=T**, conf.int=T)