Simple Regression

Theorem 2.1 (Unbiasedness of OLS)

$SLR.1 \rightarrow SLR.4 \Rightarrow E(\hat{\beta}_0) = \beta_0, E(\hat{\beta}_1) = \beta_1$

Bivariate Linear Regression Assumptions

SLR.1 (Linear in Params)
- In population, rel'shp b/w x and y is linear
- not very restrictive since error not restrained yet

SLR.2 (Random Sampling)
- data is a random sample from the population

SLR.3 (Sample variation in explanatory variable)
- not all values of explanatory var are equal
- not something to worry about

SLR.4 (Zero conditional mean)

$$E(u_i|x_i) = 0$$

- value of explanatory var must contain no info about the mean of the unobserved factors

OLS as Error Minimization

$$\hat{\beta}_1 = \frac{cov(x_i, y_i)}{var(x_i)}$$

$$\beta_0 = \overline{y} - \hat{\beta}_1 \overline{x}$$

Algebraic Properties of OLS Estimators

Est. Errors Sum to 0	$\sum_{i=1}^{n} \hat{u}_i = 0$
Correlation b/w residuals and regressors is 0	$\sum_{i=1}^{n} (x_i, \hat{u}_i) = 0$
Sample avgs of y and x lie on a regression line	$\overline{y} = \hat{\beta}_0 + \hat{\beta}_1 \overline{x}$

Goodness of Fit: Measures of Variation

Total Sum of Squares **(SST)** $\sum_{i=1}^{n} (y_i - \overline{y})^2$

Meaning: Total Variation in dependent variable

Explained Sum of Squares **(SSE)** $\sum_{i=1}^{n} (\hat{y}_i - \overline{y})^2$

Meaning: Variation explained by regression

Total Sum of Squares **(SST)** $\sum_{i=1}^{n} \hat{u}_i^2$

Meaning: Variation not explained by regression

$$SST = SSE + SSR$$

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

Meaning: variation of y explained by regression
Requirements: All OLS assumptions

Theorems

Theorem 2.4 (Gauss-Markov Theorem)

$$MLR.1 - MLR.5 \Rightarrow OLS\ ests.\ are\ BLUE$$

- BLUE - Best Linear Unbiased Estimators or regression coefficients

Theorem 4.1 (Normal Sampling Distributions)

$$MLR.1 - MLR.6 \Rightarrow OLS\ coeffs. \sim N$$
$$\hat{\beta}_j \sim N(\hat{\beta}_j, Var(\hat{\beta}_j))$$

Theorem 3.1 (Unbiasedness of OLS)

$$MLR.1 - MLR.4 \Rightarrow E(\hat{\beta}_j) = \hat{\beta}_j$$

Gauss-Markov Assumptions = MLR.1 - 5
Classical Linear Model Assumptions = MLR.1 - 6

Multiple Regression

Partialling Out (Multiple Regression)

Step 1: Regress x_1 on all the other x's

$$x_1 = \delta_0 + \delta_2 x_2 + ... + \delta_k x_k + r_1$$

Step 2: Regress y on the residuals of x_1 from step 1

$$y = \lambda_0 + \lambda_1 r_1 + v$$

- beta1 is the same as the coeff on r1 in this new regression

Regression Anatomy Formula $\Longrightarrow \beta_1 = \frac{cov(r_1, y)}{var(r_1)}$

Measures of Fit

Multiple R^2
- Single value that can be used for multiple explanatory vars
- Higher signifies better fit

Con: Will always increase when predictor variables are added, even if they are junk

Adjusted R^2
- increases only if the new var improves the model morethan would be expected by change

Pro: Only increases when the model improves

AIC (Akaike Information Criterion)
- AKA - Parsimony-adjusted measure of fit
- Way to look at several models (w/same data and same dependent vars) to find the most parsimonious model that has good fit

Pro: Penalizes the model when variables are added

Leverage and Influence

Leverage
- amt of potential ea data point has to change the regression
- Range is 0-1; low # = low leverage

Influence
- amt that ea data point actually changes the regression
- large residual needed for high influence
- Measured by Cook's Distance,
  > 1 = too much influence

Multiple Regression, cont.

Theorem 3.1 (Unbiasedness of OLS)

$MLR.1 \rightarrow MLR.4 \Rightarrow E(\hat{\beta}_j) = \beta_j$

Bivariate Linear Regression Assumptions

MLR.1 (Linear in Params)
- In population, rel'shp b/w x and y is linear
- not very restrictive since error not restrained yet

MLR.2 (Random Sampling)
- data is a random sample from the population
- data must be iid

MLR.3 (No perfect collinearity)
- no exact rel'shps b/w indep vars and none are constant
- Only perfect collinearity is not allowed

MLR.4 (Zero conditional mean)

$$E(u_i | x_{i1}, x_{i2}, \ldots, x_{ik}) = 0$$

- strongest assumptions so far
- enforces linearity

MLR.4' (Exogeneity)

$$Cov(x_j, u) = 0, \; for \; all \; j$$

- more critical assumption for real world data sets
- estimators are biased, but consistent
- bias goes to zero for large sample sizes

$MLR.1 - 3 \;\&\; MLR.4' \Rightarrow plim_{n\rightarrow\infty}(\hat{\beta}_j) = \beta_j$

$MLR.1 - 3 \;\&\; MLR.4' \Rightarrow consistency \; achieved$

Types of Residuals

Unstandardized	- measured in same units as outcome variable - Do not indicate which residual is too large, only applicable for single model
Standardized	- normal residual divided by their standard error - Used to compare residuals across different models

- Used to ID outliers
- If 1% or more of cases have residuals > 2.5, model has too much error
- If 5% or more cases have residuals >2, model has too much error

Studentized	- difference from standardized: before computing standard error, we remove one data point to make numerator and denom. indep.

- follows a student's t distribution; lets us apply precise tests to ID significant outliers

Causality

Counterfactual	- What if x were some other value? - Would y change in the way our population model predicts?
Manipulation	- We can imagine making different choices and imagine what the results would be

$$\frac{\delta y}{\delta x} = \beta_1 \;\; as \; long \; as \;\; \frac{\delta u}{\delta x} = 0$$

- beta1 is the rate of change of y w/respect to x but only if the rate of change of u w/respect to x is 0

Causality vs Exogeneity	- Causality is about whether manipulations to x do not influence the error term - Exogeneity is about whether OLS can correctly estimate beta1

Outcome variables only go on the left!!

MLR.5 - Homoskedasticity

MLR.5 (Homoskedasticity)
- Variance of the error term is constant

$Var(u_i | x_1, x_2, \ldots, x_k) = \sigma^2$

- error term cannot vary more for some values of x than others
- strong assumption that is unrealistic for many real datasets
- homoskedasticity = even thickness of residuals band on scatter plot, indicates equivalent variance for all values of x

Sampling Variance of OLS Estimators

$$Var(\hat{\beta}_j) = \frac{\sigma^2}{SST_j(1-R_j^2)}, \; j = 1, \ldots, k$$

- **sigma squared** - variance of error term
    - more error varies, more noise exists to throw estimates off, variance increases
- **SST** - total sample variation in x
    - more variation in x, more precise the estimate
- **(1 - R2)** - fraction of variation in x not explained by other vars; only unique variation in x left in denom.
- If multicollinearity, unique variation small and precision is lost

MLR.6 - Normality of Error Terms

MLR.6 (Normality of Error Terms)
- distribution of error terms is normal, $u_i \sim N(0, \sigma^2)$
- rather strong assumption

Theorem 2.4 (Gauss-Markov Theorem)

$$MLR.1 - MLR.5 \Rightarrow \; OLS \; ests. \; are \; BLUE$$

- **BLUE** - Best Linear Unbiased Estimators or regression coefficients
- There are sometimes biased estimators that are still consistent and can outperform OLS

Unbiasedness vs Efficiency

- **<u>Consistency</u>:** est. of param is consistent if the est converges to the true value of the param in the plim as the sample size increases; accuracy improves as n incr
    - only tells us that we're right in the expectation
    - but how close are the coefficients to true values?

- **<u>Efficiency</u>:** refers to variance of estimator; we want est. that varies less across diff samples
    - Since our model coefficients are rv's, we need to know how much they vary b/w draws

Regression Steps

1.) Inspect data
- look for NAs
- inspect data types
- check sample distros w/histogram
- correlationmatrix or correlation plot

2.) Create model
- remember to use heteroskedastistic robust standard errors

3.) Check assumptions
- Inspect diagnostic plots
- Run necessary statistical tests (R2, VIF, Condition Number, Breusch-Pagan test, Shapiro-Wilk test, etc.
- Adapt as necessary to correct violations

4.) Inspect for influential cases
- Use residuals vs leverage plot to detect highly influential cases

5.) Present results of model
- Regression Table (Stargazer)

Multiple Regression, cont.

Diagnostic Plots

Residuals vs Fitted Plot

Used for testing:
- homoskedasticity: looking for uniform thickness of band
- zero-conditional mean: if plot shows curvature, zero cond. mean is violated

Notes:
- red smoothing line that approx. mean of residuals

Scale-Location Plot

Used for testing:
- homoskedasticity: if fitted line is horizontal, this indicates homoskedasticity; if fitted line is not horizontal, this indicates heteroskedasticity

Notes:
- same as residuals vs. fitted plot + two transformations
    1.) calculate absolute value of data points
    2.) calculate the square root to reduce skew and move pts away from x-axis

Q-Q Plot

Used for testing:
- Normality of residuals: more deviation from the diagonal line, less normality in residual distro.

Residuals vs Leverage Plot

Used for testing:
- Influence: Cook's Dist. > 1 is too much influence

Notes:
- If a value has Cook's Dist > 1, don't automatically remove; must assess whether the data point is an error or if the value is meanginful

Troubleshooting Assumption Violations

Linearity of Parameters

- Since the error terms are not constrained yet, this assumption is a freebie and doesn't require any testing or diagnostic summaries

Random Sampling

- Confirming this assumption requires knowledge of the data and how it was collected; there are no diagnostic summaries to explicitly confirm or deny this assumption

Homoskedasticity

- Use heteroskedasticity-robust standard errors
    - AKA: Huber-White, Eicker-White, Eicker-Huber-White, White std errors
- **Diagnostic Summaries**:
    - Residuals vs Fitted: looking for uniform thickness of band = homoskedasticity
    - Scale-Location Plot: if fitted line is horizontal, this indicates homoskedasticity; if fitted line is not horizontal, this indicates heteroskedasticity
- **Tests:**
    - Breusch-Pagan test: Null Hyp = there is homosked.
        - signific. result = evidence supporting heterosked.
        - sample size is crucial: for large datasets, nearly any heterosked. will appear as signific.; vice versa for small datasets
        - use in conjunction w/diagnostic plots
        - bptest(model)
- **Mitigation Options:**
    - Switch to heterskedastic robust tools:
        - White standard errors **(**to be safe, use these all the time!!**)**
        - coeftest(model, vcov = vcovHC)
        - vcovHC(model)

Troubleshooting Assumption Violations - cont.

Normality of Residuals

- **Diagnostic Summaries:**
    - Residuals vs. Fitted Plot: if plot shows curvature, normality violated
    - Scale-Location Plot: If plot shows curvature, normality violated
    - Q-Q Plot: more deviation from the diagonal line, less normality in residual distro.
    - Histogram of residual values
- **Tests:**
    - Shapiro-Wilk Test: Null Hyp = errors are normal
        - don't directly tell how large deviations from normality are; large datasets will show signific. for even tiny deviations; small datasets will rarely be signific. regardless of devation from normality
        - use in conjunction w/diagnostic plot
        - shapiro.test(model)
- **Mitigation Options:**
    - CLT applies for large datasets; inspect Q-Q plot if 30 < n < 100
    - transform y variable for small datasets
    - bootstrap: simulate repeated samples from population by resampling from our one existing sample; generally not used w/OLS

Multicollinearity

- **Diagnostic Summaries:**
    - VIF (Variance Inflation Factor): VIF = 1/(1-R2)
        - score of 10 or higher provides evidence of serious multicollinearity
    - Condition Number (k): sqr root of the ratio of the largest eigenvalue to the smallest; k of 30 or larger indicates serious multicollinearity
    - R2: (**THIS IS R2 FROM THE VIF EQUATION**)as it increases toward 1, magnitude of potential problems assoc w/multicollinearity increases correspondingly
    - Correlation Matrix: strong correlation between explanatory variables is an indicator of multicollinearity
- **Tests:** None
- **Mitigation Options:**
    For perfect multicollinearity:
    - If an explanatory variable is a perfect linear combination of other explan vars, it's superflous and can be removed; redundant vars can be dropped

    For strong multicollinearity:
    - impact: betas will have a lot of noise and high standard error
    - Decide how much you care about the beta being estimated; if you care a lot, consider dropping other less important vars

Zero-Conditional Mean

- **Diagnostic Summaries:**
    - Residuals vs. Fitted Plot: If plot shows curvature, zero cond. mean is violated
- **Tests:** None
- **Mitigation Options:**
    Small Sample Sizes (< 30): Add more flexibility to your model by adjusting the specification; consider transformations; consier omitted vars; requires domain expertise and knowledge of the data
    For Large Sample Sizes (>30): Violations of zero-conditional mean are not a major concern; we are only concerned about establishing consistency through exogeneity

Exogeneity

- when present, this feature indicates the presence of consistency; this is really just saying that there are no omitted variables missing from our model
- **Diagnostic Summaries:** None
- **Tests:** None
- **Mitigation Options:** None covered in this class