Math 455 Class 14 February 20

Please see me if you are having problems using Gradescope.

Homeworks 4 & 5 have been reopened.

---

A type of question we haven't looked at much: multiple choice.

- these questions partly test your reading comprehension.
- you should take time and read carefully.
- enormous numbers of M.C. questions are out there.
- you can find these with web search or an LLM.
- check carefully!

Remember: 2 points of view on our data.

Say we have $n$ data points $(x_i, y_i) \in \mathbb{R}^2$.

This is also 2 vectors $\vec{Y} = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}$ $\vec{X} = \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix}$

in $\mathbb{R}^n$. The problem of minimizing.

$$\sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2 = Q(\beta_0, \beta_1)$$

is the same as the problem of finding $\beta_0, \beta_1$

to minimize

$$\| \vec{Y} - (\beta_0 \vec{1} + \beta_1 \vec{X}) \|^2 \qquad \vec{1} = \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix}$$

↑

"response vector"

Minimizing the distance from $\vec{Y}$ to the

plane generated by $\vec{1}$ and $\vec{X} = \{\beta_0 \vec{1} + \beta_1 \vec{X}\}$.

is dropping a perpendicular.

So the residual vector $\vec{e}$ is

$$\vec{e} = \vec{Y} - (\hat{\beta}_0 \vec{1} + \hat{\beta}_1 \vec{X})$$

and $\vec{e}$ is perp. to plane generated by $\vec{1}$ and $\vec{X}$.

"predictor matrix" $= \begin{pmatrix} 1 & X_1 \\ \vdots & \vdots \\ 1 & X_n \end{pmatrix}$

also "model matrix"
or "design matrix"

Column space of the predictor matrix is the plane generated by $\vec{1}$ and $\vec{X}$.

Hill Criteria

Strength

Consistency

Specificity

Temporality

Gradient

Plausibility

Experiment.

Goal for midterm:

be able to answer qualitative questions about Hill criteria.

---

Read Section 5.7 in Faraway.

---

Note that the Hill criteria are heuristics, not theorems.

Strength: $\hat{\beta}$ is "large in practical terms".

Consistency: Does it hold in all data sets?

After adjusting for covariates?

Specificity: the causal factor is associated with a specific response.

Temporality: if X causes Y, X must happen before Y.

Gradient: increasing $X$ corresponds to a larger change in $Y$.

Plausibility: Some reasonable theory suggesting the causal response.

Experiment: the "gold standard" is randomized controlled trials. But sometimes natural experiments exist.

Sample MC questions follow.

# Linear Regression Concepts: Multiple Choice Question Bank

Conceptual Fundamentals, Gauss-Markov, and Hill Criteria

## Part 1: Fundamental Objects of Regression Analysis

1. In the classical linear regression model $y = X\beta + \epsilon$, which of the following statements best describes the relationship between the parameter $\beta$ and the OLS estimator $\hat{\beta}$?

   (A) $\beta$ is a random variable that depends on the sample, while $\hat{\beta}$ is a fixed, unknown constant.

   (B) $\beta$ is a fixed, unknown constant representing the population, while $\hat{\beta}$ is a random variable whose value depends on the specific sample drawn.

   (C) Both $\beta$ and $\hat{\beta}$ are fixed constants that define the true relationship between variables.

   (D) Both $\beta$ and $\hat{\beta}$ change every time you take a new sample from the same population.

2. While often used interchangeably in casual conversation, "errors" and "residuals" have distinct roles in regression. Which of the following is true?

   (A) Errors ($\epsilon$) are observable once the data is collected, while residuals ($e$) must be estimated.

   (B) Residuals are the vertical distances from the data points to the population regression line.

   (C) Errors represent the deviation of the data from the true population model, whereas residuals represent the deviation of the data from the estimated sample line.

   (D) The sum of the errors is always zero by construction in an OLS model.

3. The Ordinary Least Squares (OLS) method chooses the estimators $\hat{\beta}$ by minimizing a specific quantity. Conceptually, what is OLS "trying" to do?

   (A) Minimize the sum of the absolute distances between the observed $y$ and the predicted $\hat{y}$.

   (B) Maximize the correlation between the independent variables $X$ and the error term $\epsilon$.

   (C) Minimize the sum of the squared vertical distances between the observed data points and the fitted regression line.

   (D) Ensure that the variance of the estimators is exactly zero.

4. In the context of linear algebra and the geometry of OLS, which of the following best describes the residual vector $e$?

   (A) It is the projection of the response vector $y$ onto the column space of the predictor matrix $X$.

   (B) It is a vector that is always parallel to the predicted values vector $\hat{y}$.

   (C) It represents the portion of the response vector $y$ that cannot be expressed as a linear combination of the columns of $X$.

(D) It is the vector that connects the origin to the estimated parameter $\hat{\beta}$.

5. If we say that an OLS estimator $\hat{\beta}$ is "unbiased," what does this mean conceptually?

   (A) The estimator will always be equal to the true population parameter $\beta$ for any given sample.

   (B) If we were to take many samples from the population and calculate $\hat{\beta}$ for each, the average of those estimates would equal the true population parameter $\beta$.

   (C) The estimator has the smallest possible variance among all linear estimators.

   (D) The model accounts for all possible variables that could influence the outcome $y$.

## Part 2: The Gauss-Markov Theorem

6. The assumption $E[\epsilon|X] = 0$ is arguably the most critical for causal inference. Conceptually, what does this assumption imply about the relationship between your independent variables ($X$) and the error term ($\epsilon$)?

   (A) The independent variables must be perfectly correlated with the error term.

   (B) There are no unobserved factors grouped in the error term that are also correlated with $X$.

   (C) The error term must follow a standard Normal distribution with a mean of zero.

   (D) The independent variables $X$ are constants and cannot be random variables.

7. The Gauss-Markov theorem requires the error terms to be "homoscedastic." In plain English, what does this mean for your data?

   (A) The variance of the residuals increases as the predicted value $\hat{y}$ increases.

   (B) The errors are all equal to the same constant value.

   (C) The "spread" or uncertainty of the error term remains constant regardless of the value of the independent variables $X$.

   (D) The independent variables $X$ all have the same variance.

8. One of the "no-go" zones for OLS is "Perfect Collinearity." Why is it a problem if one independent variable is a perfect linear combination of the others?

   (A) It causes the OLS estimator to become biased toward zero.

   (B) It makes the error terms correlate with each other (autocorrelation).

   (C) It makes it mathematically impossible to "isolate" the individual effect of one variable, as the model cannot distinguish between them.

   (D) It forces the $R^2$ value to be exactly 1.0, which is unrealistic.

9. Under the Gauss-Markov assumptions, OLS is the "Best Linear Unbiased Estimator" (BLUE). In this specific context, what does the word "Best" actually mean?

   (A) It is the estimator that is the easiest to calculate by hand.

   (B) It is the estimator that has the smallest variance among all other linear unbiased estimators.

(C) It is the estimator that provides the most accurate individual predictions for every data point.

(D) It is the only estimator that produces a high $R^2$.

10. The Gauss-Markov theorem does not require the error terms $\epsilon$ to be Normally distributed to prove that OLS is BLUE. Why, then, do we often assume $\epsilon \sim N(0, \sigma^2)$ in introductory courses?

(A) Without normality, the OLS estimators $\hat{\beta}$ will always be biased.

(B) The normality assumption is required to perform hypothesis tests (like t-tests and F-tests) and construct confidence intervals in small samples.

(C) Normality is required to ensure that the sum of the residuals is zero.

(D) If the errors aren't normal, the OLS line will not pass through the mean of the data.

## Part 3: The Hill Criteria for Causality

11. Among all of Austin Bradford Hill's criteria, "Temporality" is often cited as the only one that is strictly necessary for causality. In a regression context, what does this imply?

(A) The independent variable ($X$) and the dependent variable ($Y$) must be measured at the exact same moment in time.

(B) The cause ($X$) must precede the effect ($Y$) in time.

(C) The regression model must include a "time" variable as a control.

(D) The relationship between $X$ and $Y$ must be linear over time.

12. Hill argued that "Strong associations are more likely to be causal than weak associations." In a linear regression output, which statistic most directly relates to the "Strength" of the association?

(A) The p-value of the coefficient.

(B) The $R^2$ of the model.

(C) The magnitude (size) of the estimated coefficient $\hat{\beta}$ relative to its standard error.

(D) The number of observations ($n$) in the dataset.

13. The "Gradient" criterion suggests that if a relationship is causal, we should see a "dose-response" effect. How is this typically represented in a simple linear regression model?

(A) By the fact that the intercept $\beta_0$ is non-zero.

(B) By a significant slope coefficient $\beta_1$, indicating that as $X$ increases, $Y$ changes in a predictable, monotonic direction.

(C) By the requirement that the residuals follow a Normal distribution.

(D) By ensuring that the $X$ variable is a binary (dummy) variable.

14. The "Consistency" criterion is satisfied when the same association is found by different researchers in different locations using different samples. Why does this strengthen a causal claim?

(A) It proves that the "Zero Conditional Mean" assumption is perfectly met in all datasets.

(B) It reduces the likelihood that the observed association is due to a fluke or a specific quirk of one particular dataset.

(C) It ensures that the $R^2$ will be identical across all models.

(D) It eliminates the need for a control group.

15. If a regression shows a statistically significant relationship that contradicts all established theory in the field (violating "Plausibility"), what is the most cautious interpretation?

(A) The theory is definitely wrong, and the regression has discovered a new truth.

(B) The relationship is likely "spurious," potentially caused by an omitted variable or a coincidence.

(C) The OLS estimator is biased, but the relationship is still causal.

(D) The $R^2$ is too high, indicating a calculation error.

# Answer Key

| Question | Answer | Core Concept |
|---|---|---|
| 1 | B | Parameters (fixed) vs Estimators (random/sample-dependent) |
| 2 | C | Errors (theoretical/unobservable) vs Residuals (sample-based) |
| 3 | C | OLS Objective Function (Minimizing SSR) |
| 4 | C | Residuals as the orthogonal component to the column space |
| 5 | B | Unbiasedness as a property of the sampling distribution |
| 6 | B | Exogeneity / No Endogeneity |
| 7 | C | Homoscedasticity (constant error variance) |
| 8 | C | Identification / Multicollinearity issues |
| 9 | B | Efficiency (Minimum variance among linear unbiased estimators) |
| 10 | B | Normality's role in Small-Sample Inference |
| 11 | B | Temporality as a prerequisite for causality |
| 12 | C | Effect size and signal-to-noise ratio |
| 13 | B | Dose-Response represented by the slope |
| 14 | B | Replication and robustness against sampling error |
| 15 | B | Spurious correlation vs Theoretical Plausibility |