Exam 1 Practice

MATH 2025

Instructions

- Exam length: 75 minutes. This exam must be completed in class and entirely by hand. Show your work where appropriate.
- You may use a non-programmable calculator and the provided formula sheet. No other aids (notes, textbooks, computers, phones) are permitted.
- Read each question carefully. If a question appears ambiguous, state your interpretation clearly and proceed.
- Unless explicitly prompted, you are not required to carry out lengthy calculations to decimal answers; exact values, simplified algebraic expressions, or clearly set-up calculations earn full credit.
- Do **not** write any code from scratch. When code is presented, interpret or analyze it as directed.
- Clearly label each answer. Partial credit may be awarded when reasoning is shown.
- You are allowed one sheet of handwritten notes (front and back).

The exam is divided into four sections. Suggested time allocations are provided to help you manage the 75-minute limit.

Section A. Multiple Choice (Suggested time: 20 minutes)

For each question, circle the single best answer. Each question is worth 4 points.

1. Which step of the data analysis cycle focuses on transforming raw data into a tidy, analysable form?

    - (A) Pose a question
    - (B) Obtain data
    - (C) Wrangle/clean data

2. When creating scatterplots using `ggformula`, which function is primarily used to build the scatterplot?

   - (A) `gf_plot()`
   - (B) `gf_point()`
   - (C) `gf_scatter_plot()`

3. Suppose we fit a simple linear regression model to predict monthly apartment rent (in hundreds of dollars) from size (in hundreds of square feet). Which of the following statements about the slope coefficient is **true**?

   - (A) It represents the average change in size for a one-dollar increase in rent.
   - (B) It represents the predicted change in rent for a one-foot increase in size.
   - (C) It is guaranteed to be positive whenever there is a positive correlation.

4. When conducting a randomization (simulation-based) hypothesis test for a regression slope, which option best describes how the null distribution is generated?

   - (A) Resampling residuals with replacement and refitting the model many times under the assumption the null hypothesis is true.
   - (B) Shuffling the response values relative to the predictor and refitting the model many times to represent no relationship.
   - (C) Using a theoretical t distribution with $n - 2$ degrees of freedom.

5. Which statement correctly distinguishes a prediction interval from a confidence interval for the mean response at the same predictor value?

   - (A) A prediction interval is narrower because it only concerns a single future case.
   - (B) A prediction interval accounts for both the uncertainty in the regression line and the variability of individual responses.
   - (C) Only a confidence interval requires the equal variance (constant spread) condition.

Section B. Short Answer (Suggested time: 20 minutes)

Show all reasoning. Each question is worth 8 points.

1. **Conditions for Simple Linear Regression**: List and briefly describe the four core conditions for simple linear regression. Provide a concise explanation of how you would check each condition using plots or summaries.

2. **Interpreting Model Output**: Consider the following R output from a model relating park accessibility points (`pct_near_park_points`) to spending per resident (`spend_per_resident_data`):

```r
lm(pct_near_park_points ~ spend_per_resident_data, data = parks) |>
   tidy()
```

```
# A tibble: 2 x 5
  term                   estimate std.error statistic  p.value
  <chr>                     <dbl>     <dbl>     <dbl>    <dbl>
1 (Intercept)               18.5      1.49       12.4 4.73e-32
2 spend_per_resident_data    0.133    0.0122     10.9 1.04e-25
```

What is the predicted `pct_near_park_points` score for a city that spends $150 per resident?

3. Interpret the slope and intercept in the context of the problem. Comment on whether the intercept has a meaningful interpretation.

4. **Residual Patterns**: A residual vs. fitted plot for a commuting time model shows a clear funnel shape, with residual spread increasing as fitted values increase. Which model condition is being violated? Explain how this impacts the reliability of slope inference and suggest one possible remedy.

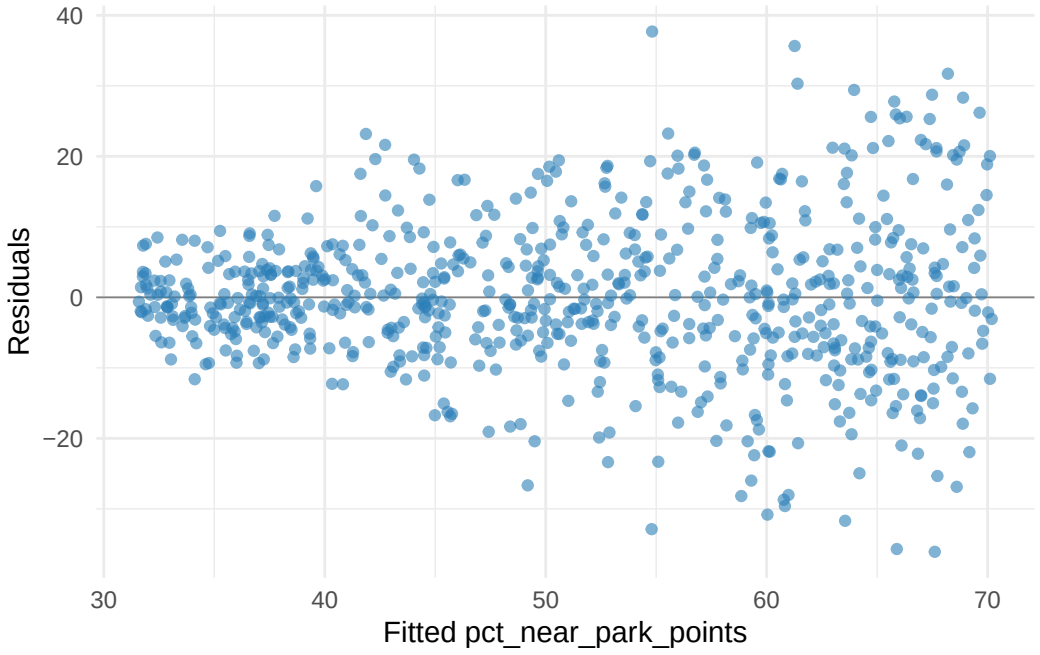

Figure 1: Residuals vs Fitted (synthetic, parks-like) showing a clear funnel shape.

5. **Confidence Interval Interpretation**: A simulation-based procedure for the slope produced a 95% confidence interval of $(1.2, 2.8)$ minutes per mile when predicting commute

time from travel distance. State what this interval means in context, including the interpretation of the confidence level.

Section C. True/False with Justification (Suggested time: 15 minutes)

For each statement, circle True or False and justify your answer in 1–2 sentences. Each question is worth 5 points.

1. **True / False**: If the population correlation between study hours and exam score is near zero, then fitting a simple linear regression to a sample will always yield a slope that is exactly zero.
2. **True / False**: When creating a scatterplot in `ggplot`, swapping the roles of the explanatory and response variables on the axes changes the underlying correlation between them.
3. **True / False**: In a randomization test for the slope, a p-value of 0.03 means there is only a 3% chance the alternative hypothesis is true.
4. **True / False**: A 95% confidence interval for the slope will likely get more accurate as you increase the sample size.
5. **True / False**: Prediction intervals widen as you move away from the mean of the predictor values even when the constant spread condition is satisfied.

Section D. Evaluate a Response (Suggested time: 20 minutes)

Each item below presents a student's written response. Evaluate the response by (i) stating whether the student's conclusion is *correct*, *partially correct*, or *incorrect*, and (ii) explaining your reasoning. Each question is worth 10 points.

1. **False Positives**: Suppose there is no relationship between study hours and exam score in the population of interest. However, 100 different researchers all, independently, collect samples and conduct their own hypothesis tests using a significance level of $\alpha = 0.05$. We should expect that all 100 researchers fail to reject the null hypothesis.

2. **Hypothesis Test Conclusion**: Student B used a randomization test for the slope in a model relating greenhouse gas emissions to average household energy bills and obtained a p-value of 0.18. They conclude, "Because the p-value is larger than 0.05, there is strong evidence of a positive relationship between the variables." Assess the correctness of this conclusion.

3. **Confidence Interval Interpretation**: Student C reports, "A 95% confidence interval for the mean commute time of (12, 28) minutes means there is a 95% probability that the true average commute time lies between 12 and 28 minutes." Evaluate this interpretation.

Point Summary

Section	Points
A	20
B	40
C	25
D	30
Total	**115**

Solutions

Section A. Multiple Choice

1. **(C)** The wrangle/clean step transforms raw data into a tidy structure that supports analysis.
2. **(B)** `gf_point()` adds points for a scatterplot within the plotting system.
3. **(B)** The slope estimates the average change in rent for a one-unit increase in size (as defined in the model's units).
4. **(B)** Shuffling the response relative to the predictor simulates a world with no association between them.
5. **(B)** Prediction intervals include uncertainty from both the fitted mean and individual variability, making them wider.

Section B. Short Answer

1. **Regression conditions**: (i) *Linearity* — the relationship between predictor and response is approximately linear; check with a scatterplot or residuals vs. fitted plot. (ii) *Independence* — observations do not influence each other; check study design or order of data collection. (iii) *Constant spread* — residual variance is roughly equal across fitted values; assess with a residuals vs. fitted plot for uniform scatter. (iv) *Normal residuals* — residual distribution is roughly symmetric without extreme tails; assess with a histogram or normal Q-Q plot.
2. **Predicted value**: Use the fitted line from the output. If the fitted coefficients are $\hat{\beta}_0$ (intercept) and $\hat{\beta}_1$ (slope on spending per resident in dollars), then the prediction at $x = 150$ is $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 \times 150$. If spending were expressed in hundreds of dollars, then use $x = 1.5$ accordingly.
3. **Interpret slope/intercept**: The slope represents the expected change in `pct_near_park_points` for a \$1 increase in `spend_per_resident_data` (in the units used in the model). The intercept is the predicted `pct_near_park_points` when spending is \$0; this may be outside the data range and not substantively meaningful.
4. **Condition violation**: The funnel shape signals a violation of the constant spread (homoscedasticity) condition. Inference on the slope may be unreliable because standard errors, and thus our p-values and confidence intervals, are misestimated.
5. **Confidence interval meaning**: We are 95% confident that the true increase in commute time per mile of travel lies between 1.2 and 2.8 minutes. Over many similar studies using the same procedure, about 95% of the computed intervals would capture the true slope.

Section C. True/False with Justification

1. **False.** A near-zero correlation suggests a weak linear association but the sample slope need not be exactly zero; sampling variability can produce small nonzero slopes.
2. **False.** Switching axes changes which variable is treated as predictor or response but does not alter the underlying correlation value.
3. **False.** A p-value of 0.03 measures the probability of obtaining a result as or more extreme under the null, not the probability the alternative is true.
4. **True.** t-based intervals rely on the sampling distribution of the slope being approximately normal, which becomes more plausible with large samples or nearly normal residuals.
5. **True.** Prediction intervals account for individual variability and tend to widen for predictor values farther from the sample mean.

Section D. Evaluate a Response

1. **Incorrect.** With $\alpha = 0.05$ and a true null in all studies, about 5% of tests will (by chance) falsely reject the null in the long run. Across 100 independent studies, we expect around 5 false positives—not that all 100 will fail to reject.
2. **Incorrect.** A p-value of 0.18 indicates the observed slope is plausible under the null; it does not provide strong evidence for a positive relationship. The student should conclude there is insufficient evidence to claim a positive association.
3. **Partially correct.** The interval targets the population mean, and the correct frequentist interpretation is that the method yields intervals that capture the true mean in about 95% of repeated samples. It is acceptable to say "we are 95% confident the true mean lies between 12 and 28 minutes," but it is not a literal 95% probability statement about a fixed parameter. The statement should avoid implying probability on the parameter.