DATA 505 (Statistics Using R): Exam 2 Part 1 Practice

Name: _____

2025-11-25

Instructions:

This is a practice exam designed to help you prepare for the actual Exam 2. The format, question types, and topics covered are similar to what you will encounter on the real exam. Use this to identify areas where you need additional study.

The actual exam will have two parts:

- *Part 1 (approximately 45-50 minutes):*
  - *Closed computer and closed notes*
  - *Handwritten answers*
- *Part 2 (approximately 45-50 minutes):*
  - *Open computer and open resources (except "alive" resources like other people)*
  - *Electronic submission on Moodle: you will fill in a Quarto script with R code and written answers*

Once you turn in Part 1, you may open your computer and download the blank Quarto script for Part 2. You may not return to Part 1 after turning it in and beginning the open-computer portion.

PRACTICE EXAM - NOT FOR CREDIT

Part 1A (closed computer): Multiple choice

(3 points)

Which of the following correctly fits a multiple regression model predicting `y` from `x1` and `x2` in R? *(Circle the correct answer)*

A. `lm(y ~ x1, x2, data = df)`

B. `lm(y ~ x1 + x2, data = df)`

C. `lm(y = x1 + x2, data = df)`

D. `regression(y, predictors = c(x1, x2), df)`

(3 points)

In the regression equation $Y = \beta_0 + \beta_1 X + \varepsilon$, what does β_0 represent?

A. The change in Y for a one-unit increase in X

B. The expected value of Y when X equals zero

C. The correlation between X and Y

D. The standard error of the regression

(3 points)

You fit a regression model with R-squared = 0.81. What does this tell you?

A. The model correctly predicts 81% of observations

B. There is an 81% chance the relationship is causal

C. 81% of the variance in Y is explained by the predictor(s)

D. The slope coefficient is 0.81

(3 points)

Which R function extracts the estimated regression coefficients from a fitted model object called `model`?

A. `coefficients(model)`

B. `coef(model)`

C. `model$coefficients`

D. All of the above

(3 points)

Why is a prediction interval wider than a confidence interval for the same predictor value?

A. Prediction intervals use a different formula

B. Prediction intervals account for uncertainty in individual observations, not just the mean

C. Confidence intervals are always wrong

D. Prediction intervals are calculated at a higher confidence level

(3 points)

In a regression model with a categorical predictor that has 4 levels, how many dummy variables will R create?

A. 4

B. 3

C. 5

D. 2

(3 points)

What is the primary reason to write functions in R?

A. Functions are required for all R scripts to run

B. To avoid code repetition and make code more maintainable

C. Functions automatically optimize code performance

D. To satisfy R syntax requirements

(3 points)

Consider a function defined as:

```
my_function <- function(x, y = 10) {
  x + y
}
```

What will `my_function(5)` return?

A. 5

B. 10

C. 15

D. An error

(3 points)

Two regression models are fit to different datasets, both predicting Y from X. Model A has R-squared = 0.64 and Model B has R-squared = 0.86. What can we conclude?

A. Model A has a larger slope coefficient than Model B

B. Model B explains more variance in Y than Model A

C. Model A is always preferred over Model B

D. The two models use different predictors

(3 points)

What is the main advantage of organizing R code into separate files for different functions (e.g., `01_load.R`, `02_clean.R`)?

A. It reduces memory usage

B. It improves code modularity, making it easier to test and reuse components

C. It makes the code run faster

D. It's required by R syntax

Part 1B (closed computer): Short answer

Write your answers in complete sentences where appropriate.

Consider the following regression output predicting test scores from hours studied:

```
Call:
lm(formula = score ~ hours, data = test_data)

Residuals:
     Min       1Q   Median       3Q      Max
-19.2671  -5.0501  -0.3149   5.5238  16.7507

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  61.6282     1.6841   36.59   <2e-16 ***
hours         3.7783     0.2875   13.14   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 7.426 on 78 degrees of freedom
Multiple R-squared:  0.6889,    Adjusted R-squared:  0.6849
F-statistic: 172.7 on 1 and 78 DF,  p-value: < 2.2e-16
```

(4 points)

Write the fitted regression equation for this model (in the form: $\widehat{\text{score}} = \hat{\beta}_0 + \hat{\beta}_1 \times \text{hours}$).

(4 points)

Interpret the slope coefficient in context. What does it tell you about the relationship between hours studied and test score?

(4 points)

Based on the p-value for the slope, is there statistically significant evidence of a linear relationship between hours studied and test score at the $\alpha = 0.05$ level? Justify your answer.

(4 points)

Using this model, what would you predict for the test score of a student who studied for 6 hours? Show your calculation.

(4 points)

Write R code that would add a regression line to a scatterplot of score vs. hours using ggplot2. Your code doesn't need to include all cosmetic details.

(3 points)

Explain why it would be problematic to use this model to predict the test score for a student who studied 50 hours.