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

2025-11-25

PRACTICE EXAM SOLUTIONS

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)` ← CORRECT

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

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

Explanation: The formula syntax requires `~` to separate the response from predictors, and `+` to add multiple predictors. Option A is missing the `+`. Option C uses `=` instead of `~`. Option D is not valid R syntax.

(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 ← CORRECT

C. The correlation between X and Y

D. The standard error of the regression

Explanation: The intercept β_0 is the expected value of Y when $X = 0$. The slope β_1 represents the change in Y for a one-unit increase in X.

(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) ← CORRECT

D. The slope coefficient is 0.81

Explanation: R-squared measures the proportion of variance in the response variable that is explained by the predictor(s). It ranges from 0 to 1, with higher values indicating better fit.

(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 ← CORRECT

Explanation: All three methods work to extract coefficients from a fitted model. `coef()` is the most commonly used shorthand, `coefficients()` is the full function name, and `$coefficients` accesses the component directly from the model object.

(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 ← CORRECT

C. Confidence intervals are always wrong

D. Prediction intervals are calculated at a higher confidence level

Explanation: Confidence intervals estimate the mean response for a given predictor value, while prediction intervals estimate where a single new observation will fall. Prediction intervals must account for both the uncertainty in estimating the mean AND the variability of individual observations around that mean, making them wider.

(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 ← CORRECT

C. 5

D. 2

Explanation: R creates $k - 1$ dummy variables for a categorical predictor with k levels, using one level as the reference category. With 4 levels, R creates 3 dummy variables.

(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 ← CORRECT

C. Functions automatically optimize code performance

D. To satisfy R syntax requirements

Explanation: The main benefit of writing functions is to follow the DRY (Don't Repeat Yourself) principle. Functions improve code readability, reduce errors, and make code easier to test and reuse.

(3 points)

Consider a function defined as:

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

What will `my_function(5)` return?

A. 5

B. 10

C. 15 ← CORRECT

D. An error

Explanation: The function has a default value of `y = 10`. When called with only one argument `my_function(5)`, it uses `x = 5` and the default `y = 10`, returning `5 + 10 = 15`.

(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 ← CORRECT

C. Model A is always preferred over Model B

D. The two models use different predictors

Explanation: The R-squared value tells us what proportion of variance is explained. Model B has a higher R-squared ($0.86 > 0.64$), meaning it explains more variance in Y. We cannot determine slope coefficients or model preference from R-squared alone, and the question states both models predict Y from X.

(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 ← CORRECT

C. It makes the code run faster

D. It's required by R syntax

Explanation: Organizing code into separate files improves modularity, making it easier to test individual components, reuse functions across projects, and collaborate with others. It doesn't directly affect performance or memory usage, and it's not required by R syntax.

Part 1B (closed computer): Short answer

Solutions with explanations.

```
summary(test_model)
```

```
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: $\widehat{\text{score}} = \hat{\beta}_0 + \hat{\beta}_1 \times \text{hours}$).

SOLUTION:

From the output, the coefficients are: - Intercept: 61.628 - Slope: 3.778

Therefore, the fitted regression equation is:

$$\widehat{\text{score}} = 61.63 + 3.78 \times \text{hours}$$

(4 points)

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

SOLUTION:

The slope coefficient is approximately 3.78. This means that for each additional hour studied, the test score is expected to increase by about 3.78 points, on average. This indicates a positive relationship between study time and test performance.

(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.

SOLUTION:

The p-value for the slope coefficient is 1.8e-21, which is much less than 0.05.

Since the p-value < 0.05, we reject the null hypothesis that the slope is zero. There is statistically significant evidence of a linear relationship between hours studied and test score at the $\alpha = 0.05$ significance level.

(4 points)

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

SOLUTION:

Using the regression equation:

$$\widehat{\text{score}} = 61.63 + 3.78 \times 6$$

$$\widehat{\text{score}} = 61.63 + 22.67 = 84.3$$

We would predict a test score of approximately 84.3 points for a student who studied 6 hours.

Alternative using R:

```
predict(test_model, newdata = data.frame(hours = 6))
```

```
       1
84.29802
```

(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.

SOLUTION:

```r
ggplot(test_data, aes(x = hours, y = score)) +
  geom_point() +
  geom_smooth(method = "lm", se = FALSE)
```

Alternative:

```r
ggplot(test_data, aes(x = hours, y = score)) +
  geom_point() +
  geom_abline(intercept = coef(test_model)[1],
              slope = coef(test_model)[2])
```

(3 points)

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

SOLUTION:

This would be **extrapolation** - predicting outside the range of observed data. The observed hours studied range from approximately 0 to 9.9 hours.

Predicting at 50 hours is far beyond this range. We have no evidence that the linear relationship holds at such extreme values. The relationship might level off, become non-linear, or behave differently for very high study times. Making predictions outside the observed range is unreliable and can lead to nonsensical results.