Ten-Question Regression Capstone

Notes for SHS | Topic 78 | Fictional educational statistics | November 2026

Task 1: Dataset inspection

Load main CSV; validate unique codes, numeric fields and 0–100 outcome.

Verified reference: 240 fictional unique rows, no missing values generated.

Task 2: Descriptive summaries

Find sample means and sample SD of hours and scores.

Verified reference: Hours mean 3.6265 SD 1.6925; score mean 51.0179 SD 12.1735.

Task 3: Pearson association

Calculate Pearson correlation hours vs score.

Verified reference: r=0.6784; description only.

Task 4: Fit model on training data

Split 80/20 with random_state 78, fit simple LinearRegression.

Verified reference: 192 train, 48 test. Slope 4.9728; intercept 32.9489.

Task 5: Make a fitted prediction

Predict at five hours using the training-fit model.

Verified reference: Predicted score 57.8128; in-range model prediction.

Task 6: Interpret residuals

Take observed minus predicted for rows in test CSV.

Verified reference: Positive residual lies above fitted line; negative residual lies below.

Task 7: Evaluate held-out data

Calculate MAE, RMSE and R2 on the 48 test records.

Verified reference: Test MAE 7.9784, RMSE 9.6863, R2 0.3516.

Task 8: Compare training-only baseline

Predict the training mean for every test case.

Verified reference: Baseline test MAE 9.7208, RMSE 12.0522, R2 -0.0038.

Task 9: Review limitations

Inspect test residuals, outliers and artificial construction.

Verified reference: The single-feature linear model is descriptive and predictive within the synthetic setting, not causal.

Task 10: Write findings report

Use methods, coefficients, held-out metrics and caveats.

Verified reference: Include random seed 78078, split seed 78, 240-row CSV and caution about generalization.