The University of Texas at Austin
MOCK EXAM 1
Predictive Analytics

February 25, 2026

Instructions: Provide your <u>complete solution</u> to the following problems. Final answers only, without appropriate justification, will receive zero points even if correct.

DEFINITIONS

Problem 1.1. (10 points) Provide the definition of *bias*.

Solution. See the solutions to the first homework assignment.

Problem 1.2. (10 points) Provide the definition of the *mean-squared error* in the context of parameter estimation.

Solution. See the solutions to the first homework assignment.

CONCEPTUAL QUESTIONS

Problem 1.3. (10 points) Describe the difference between classification and regression.

Solution. Solutions will vary. The salient point of any response which is to earn credit must be that classification has a categorical response variable whereas regression has a numerical response variable.

FREE RESPONSE PROBLEMS

Please, explain carefully all your statements and assumptions. Numerical results or single-word answers without an explanation (even if they're correct) will be worth 0 points.

Problem 1.4. (10 points) Consider a simple linear regression fitted on 20 observations. In our usual notation, you are given the following:

- $\sum (y_i - \hat{y}_i)^2 = 10$
- $\sum (\hat{y}_i - \bar{y})^2 = 112$

Find the coefficient of determination.

Solution. You should know that

$$TSS = \sum (y_i - \bar{y})^2 = \sum (y_i - \hat{y}_i)^2 + \sum (\hat{y}_i - \bar{y})^2 = RSS + \sum (\hat{y}_i - \bar{y})^2.$$

The first given fact tells us that the RSS equals 10. The second given fact tells us that $TSS - RSS$ equals 112. So,

$$R^2 = 1 - \frac{RSS}{TSS} = \frac{TSS - RSS}{TSS} = \frac{112}{122}.$$

For your convenience, I am providing the proof of the above equality. You did not need to prove this fact in the exam. Proving

$$\sum (y_i - \bar{y})^2 = \sum (y_i - \hat{y}_i)^2 + \sum (\hat{y}_i - \bar{y})^2$$

is equivalent to proving

$$\sum (y_i - \hat{y}_i)(\hat{y}_i - \bar{y}) = 0.$$

As we did in class, let us denote $\varepsilon_i = y_i - \hat{y}_i$ for all $i = 1, ..., n$. Recall that the sum of residuals is equal to zero, i.e., $\sum \varepsilon_i = 0$. Also, by the independence condition, $\sum \varepsilon_i x_i = 0$. Then,

$$\sum (y_i - \hat{y}_i)(\hat{y}_i - \bar{y}) = \sum \varepsilon_i (\beta_0 + \beta_1 x_i - \beta_0 - \beta_1 \bar{x})$$
$$= \sum \varepsilon_i (\beta_1 x_i - \beta_1 \bar{x})$$
$$= \beta_1 \sum \varepsilon_i x_i - \beta_1 \bar{x} \sum \varepsilon_i = 0.$$

Problem 1.5. (15 points) *Source: An old SOA exam.*
Consider a multiple linear regression where predictor X_1 stands for the amount of precipitation in a month (in inches), predictor X_2 stands for the traffic volume in a month, and predictor X_3 stands for the indicator of whether a holiday weekend occurred during a particular month.

The response variable Y corresponds to the number of fatal car accidents.

You fit the data for the past year, and get the following coefficients

Intercept	−2.358
Precipitation	0.245
Volume	1.129
Holiday	2.334

- (5 points) What is your prediction for a month with no holiday weekends, with the precipitation equal to 1.5 inches, and the traffic volume equal to 4?

- (5 points) How do you interpret the coefficient for *Precipitation*?

- (5 points) How do you interpret the coefficient for *Holiday*?

Solution.

$$-2.358 + 0.245(1.5) + 1.129(4) = 2.5255$$

The coefficient for *Precipitation* tells us that for every additional inch of precipitation there are 0.245 more accidents per our model on average (with all else kept fixed).

The coefficient for *Holiday* tells us that months with holiday weekends have 2.334 more fatal car accidents than comparable months without holiday weekends.

Problem 1.6. (10 points) *Source: MAS-I, Spring 2018.*
You are given the following information about an insurance policy:

- The probability of a policy renewal $p(X)$ follows a logistic model with an intercept and one explanatory variable.

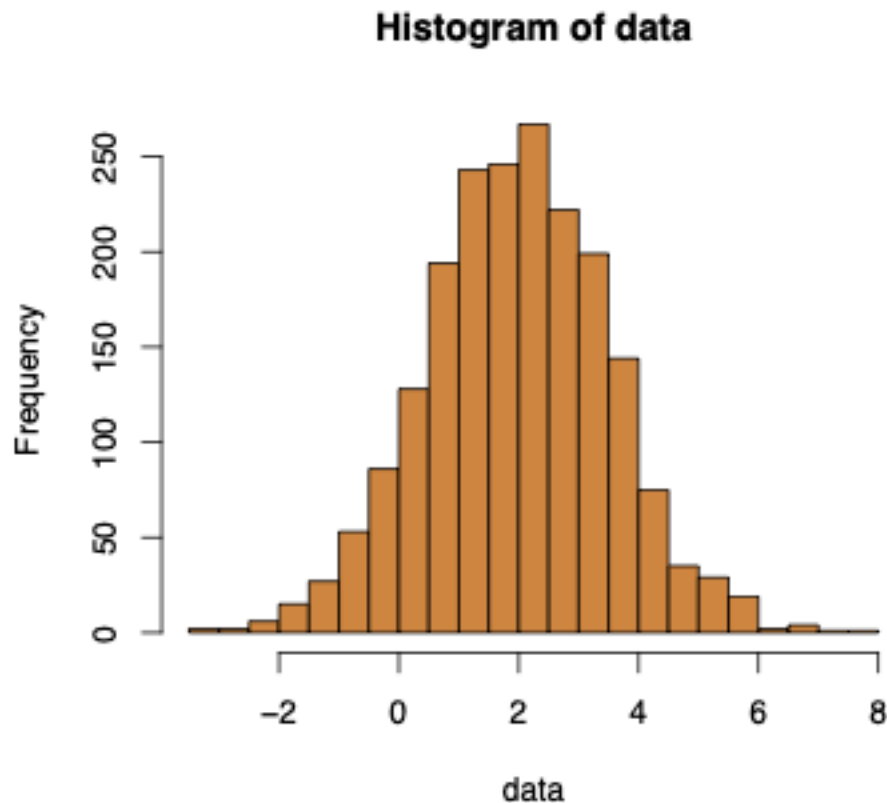- $\hat{\beta}_0 = 5$

- $\hat{\beta}_1 = -0.65$

Calculate the fitted odds of renewal at $x = 5$.

Solution. In the logistic model, the odds are $e^{\beta_0 + \beta_1 x}$. So, the coefficients given above yield

$$e^{5 - 0.65(5)} = 5.754603$$

MULTIPLE CHOICE QUESTIONS

Problem 1.7. (5 points) You have a sample of size 252 from a distribution that you know from past experience looks like this:

Histogram of data

Your task is to estimate its mean. Of the following, what procedure(s) would be acceptable in this case? Choose f all that apply.

- A 95% bootstrap confidence interval using quantiles.

- Using the 't.test' command in f R.

- A $2SE$ bootstrap confidence interval.

- The standard z −procedure 95%-confidence interval.

- None of the above.

Solution. (a, b, c, d)