AML 5201
Advanced Applications

of Probability & Statistics

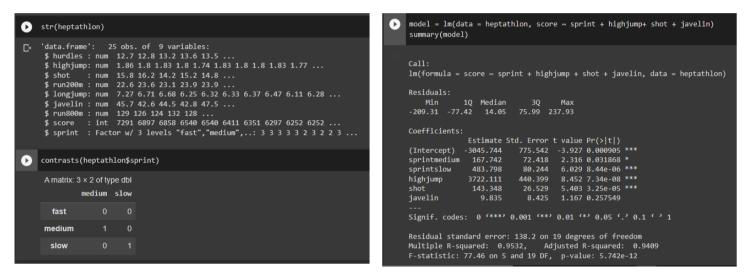
Even Sem. 2023



Sessional-1

[10 points] [TLO 2.2, CO 1] Say true or false with a brief explanation (nor more than 1-2 lines):

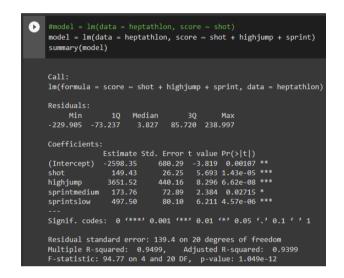
- (a) It is possible to build a linear regression model for the following scenario: response variable is customer credit card default status (yes/no) and predictor variables are customer age, annual income, level of education, marital status, and credit card limit.
- (b) In a linear regression model, a categorical predictor with 3 levels will result in 2 dummy (new) variables.
- (c) Suppose we build a linear regression model for predicting house price based on square feet area and number of bedrooms. An R^2 value of 60% means that the remaining 40% of the variance in the house price is due to noise.
- (d) To understand the relationship between a continuous and a categorical variable, we use a scatter plot.
- [10 points] [TLO 2.1, CO 1] Consider the heptathlon dataset with the following details for predicting *score* as a function of selected predictors:



- (a) How many continuous predictors have we used to build the model?
- (b) How many categorical predictors have we used to build the model?
- (c) For each categorical predictor that is used to build the model, state the number of levels, the reference level, and the names of the dummy variables introduced.
- (d) How accurate is the model?
- (e) Which predictor is most likely to <u>not</u> to have a linear relationship with *score*?
- (f) Which predictor has the smallest standard error in a <u>relative</u> sense?
- 3. [10 points] [TLO 2.1, CO 2] Consider the heptathlon dataset with the following details for predicting *score* as a function of the predictor *sprint*:

The linear regression model is $\hat{y}^{(i)} = \hat{\beta}_0 + \hat{\beta}_1 x_1^{(i)} + \hat{\beta}_2 x_2^{(i)}$. Write the equations for predicting *score* for *slow*, *medium*, and *fast* athletes. What is the average of *score* for *slow*, *medium*, and *fast* athletes?

4. [10 points] [TLO 2.1, CO 2] Consider the heptathlon dataset with the following details for predicting *score* as a function of the predictors *shot*, *highjump* and *sprint*:



- (a) Compared to the model in the previous question which had only the predictor *sprint*, has the model accuracy improved?
- (b) Considering a p-value of 5% as threshold, are there any insignificant features?
- (c) For a 1 metre increase in shot put throw and with the same *highjump* and *sprint* performance, we can say with 95% confidence that the athlete's score will increase/decrease by an amount in the interval [?, ?].
- 5. [10 points] [TLO 2.1, CO 2] Suppose we want to predict starting salary after graduation (in thousands of dollars) using the following predictors:
 - (1) GPA (2) IQ (3) Gender (female and male) (4) Interaction between GPA and IQ (5) Interaction between GPA and Gender.

The results of fitting a linear regression model are:

$$\hat{\beta}_0 = 50, \hat{\beta}_1 = 20, \hat{\beta}_2 = 0.07, \hat{\beta}_3 = 35, \hat{\beta}_4 = 0.01, \hat{\beta}_5 = -10.$$

- (a) Write the predicted salary in terms of the coefficient estimates and predictor variables for the *i*th individual.
- (b) Write the regression estimate in terms of the coefficient estimates and predictor variables when the *i*th individual is (a) male (b) female.
- (c) Predict the salary of a female with IQ of 110 and a GPA of 4.0.
- (d) Is the small value of the coefficient for the GPA/IQ interaction term indicating that there is little evidence of an interaction effect between those two predictors? Give a one-line explanation.