

Sessional-1

- [10 points] [TLO 2.2, CO 1] Say true or false with a brief explanation (nor more than 1-2 lines):
 - 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.
 - In a linear regression model, a categorical predictor with 3 levels will result in 2 dummy (new) variables.
 - 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.
 - 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:

```
str(heptathlon)

'data.frame': 25 obs. of 9 variables:
 $ hurdles : num 12.7 12.8 13.2 13.6 13.5 ...
 $ highjump: num 1.86 1.8 1.83 1.8 1.74 1.83 1.8 1.8 1.83 1.77 ...
 $ shot : num 15.8 16.2 14.2 15.2 14.8 ...
 $ run200m : num 22.6 23.6 23.1 23.9 23.9 ...
 $ longjump: num 7.27 6.71 6.68 6.25 6.32 6.33 6.37 6.47 6.11 6.28 ...
 $ javelin : num 45.7 42.6 44.5 42.8 47.5 ...
 $ run800m : num 129 126 124 132 128 ...
 $ score : int 7291 6897 6858 6540 6540 6411 6351 6297 6252 6252 ...
 $ sprint : Factor w/ 3 levels "fast","medium",...: 3 3 3 3 3 2 3 2 2 3 ...

contrasts(heptathlon$sprint)

A matrix: 3 x 2 of type dbl
      medium slow
fast      0      0
medium    1      0
slow      0      1
```

```
model = lm(data = heptathlon, score ~ sprint + highjump + shot + javelin)
summary(model)

Call:
lm(formula = score ~ sprint + highjump + shot + javelin, data = heptathlon)

Residuals:
    Min       1Q   Median       3Q      Max
-209.31  -77.42   14.05   75.99  237.93

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -3045.744    775.542   -3.927 0.000905 ***
sprintmedium    167.742     72.418    2.316 0.031868 *
sprintslow     483.798     80.244    6.029 0.000000 ***
highjump      3722.111    440.399    8.452 7.34e-08 ***
shot          143.348     26.529    5.403 3.25e-05 ***
javelin         9.835       8.425    1.167 0.257549
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 138.2 on 19 degrees of freedom
Multiple R-squared:  0.9532,    Adjusted R-squared:  0.9409
F-statistic: 77.46 on 5 and 19 DF,  p-value: 5.742e-12
```

- How many continuous predictors have we used to build the model?
 - How many categorical predictors have we used to build the model?
 - 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.
 - How accurate is the model?
 - Which predictor is most likely to not to have a linear relationship with *score*?
 - Which predictor has the smallest standard error in a relative sense?
- [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*:

```
model = lm(data = heptathlon, score ~ sprint)
summary(model)

Call:
lm(formula = score ~ sprint, data = heptathlon)

Residuals:
    Min       1Q   Median       3Q      Max
-1090.5  -278.6    89.5   284.5   678.5

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  5656.5      151.0    37.454 < 2e-16 ***
sprintmedium   356.1       207.6    1.715 0.100328
sprintslow     956.0       213.6    4.476 0.000189 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 427.2 on 22 degrees of freedom
Multiple R-squared:  0.4824,    Adjusted R-squared:  0.4353
F-statistic: 10.25 on 2 and 22 DF,  p-value: 0.0007144
```

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?

- [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*:

```
#model = lm(data = heptathlon, score ~ shot)
model = lm(data = heptathlon, score ~ shot + highjump + sprint)
summary(model)

Call:
lm(formula = score ~ shot + highjump + sprint, data = heptathlon)

Residuals:
    Min       1Q   Median       3Q      Max
-229.905  -73.237    3.827   85.720  238.997

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -2598.35     680.29   -3.819 0.00107 **
shot         149.43       26.25    5.693 1.43e-05 ***
highjump     3651.52     440.16    8.296 6.62e-08 ***
sprintmedium  173.76       72.89    2.384 0.02715 *
sprintslow    497.50       80.10    6.211 4.57e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 139.4 on 20 degrees of freedom
Multiple R-squared:  0.9499,    Adjusted R-squared:  0.9399
F-statistic: 94.77 on 4 and 20 DF,  p-value: 1.049e-12
```

- Compared to the model in the previous question which had only the predictor *sprint*, has the model accuracy improved?
 - Considering a p-value of 5% as threshold, are there any insignificant features?
 - 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 [?, ?].
- [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:

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

- Write the predicted salary in terms of the coefficient estimates and predictor variables for the i th individual.
- Write the regression estimate in terms of the coefficient estimates and predictor variables when the i th individual is (a) male (b) female.
- Predict the salary of a female with IQ of 110 and a GPA of 4.0.
- 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.