Chapter 3 Concept Test

Due Apr 8 at 12pm **Points** 1 **Questions** 6 **Time Limit** None

Instructions

Read the chapter, review the pre-class material, then take this concept test

Attempt History

	Attempt	Time	Score
LATEST	Attempt 1	9 minutes	1 out of 1

Score for this quiz: **1** out of 1
Submitted Apr 8 at 9:52am
This attempt took 9 minutes.

Question 1

0 / 0 pts

You are fitting a simple linear regression model to predict the prices of the items for sale in a store (Y_i) based on the weight of the item (X_i). You have some training data with many items for sale (this data includes the weight of each item and its price). Suppose that you saw a horizontal line at $y = 1$ when you fit the model to the data. What would this tell you about the relationship between the input variable and the output?

○ For all input values, X_i, $Y_i = 1 \times X_i + \varepsilon$

○ The linear model only shows a relationship between input and output when $y = 1$

○ For all input values, X_i, $Y_i = 1 \times X_i$ (the model is a perfect fit without errors)

Correct!

● The linear model does not reveal a relationship between the input variable and the output variable

Correct - since the model always predicts y=1, the model's prediction is unaffected by the weight of the item.

Question 2

0 / 0 pts

You have a dataset with *n* observations and *p* features. Assuming that each feature will be used at most once in your model (first-order terms only, no interaction terms), how many possible individual multiple linear regression models could be formed using the *p* features?

○ p^2

○ $p+1$

○ $\frac{p(p-1)}{2}$

Correct!

● 2^p

Correct. Since you must decide whether or not to use each feature in a model, the total number of possible models is the number of possible subsets of the original set of features. In other words, the answer is the cardinality of the powerset: 2^p

Question 3

0 / 0 pts

Suppose you train a multiple linear regression to predict *y* on a dataset with 4 input variables. During preprocessing you normalized the data so that all inputs were gaussian normalized with mean 0 and variance 1. You then trained the model. Your trained linear model coefficients are: β_1 = 0.5, β_2 =0.4 β_3 =0.001 and β_4 = -0.85. What can you infer about the importance of inputs x_1, x_2, x_3, and x_4? (select all that apply)

Correct!

☑ X_4 has a stronger relationship with y than X_3

Yes, due to its magnitude, the fourth feature has a stronger relationship to y than the third feature.

□ A model using X_4 alone will produce y values which are 85% accurate

□ The hypothesis that X_3 has no relationship with y can be rejected since $0.001 < 0.05$

'ou Answered

☑ A model using only X_1 and X_2 will outperform a model using only X_4

From the coefficients alone, it is not possible to know whether a different model would perform because the coefficients do not directly indicate the performance of the model and the coefficient's values are applicable only indicative of feature importance in the current model. In other words, the coefficients in the four-feature model do not provide information about the values of the respective coefficients in a model with fewer features.

Since the data was pre-processed and all the features used the same scale, the coefficients can be directly compared. The larger the magnitude of a coefficient, the more the effect of that input on the model. Thus, given that we build the model with all four features, x_4 has a large affect on the model, x_1 has the second largest, x_2 has the third largest, and x_3 has a negligible affect on the model. The coefficients do not directly indicate model performance or probability that the coefficient's value is due to chance.

Question 4

0 / 0 pts

You have a dataset with n>30 observations and p features and each observation has a real value output y. Suppose you form a hypothesis H_0 that none of the p predictors (features) of a dataset has a linear relationship with the output variable y. The alternative hypothesis H_a is that at least one of the predictors (features) of the dataset has a linear relationship with the output variable y. Which of the following techniques is appropriate for evaluating that hypothesis if you would like to have inappropriately rejected the null no more often than 5% of the time?

○ For each predictor compute the variance inflation factor (VIF) on each coefficient. If all VIFs are below 5 then reject the null hypothesis H_0.

○ Compute the correlation between each pair of features in the dataset. If at least two are correlated with 0.95 or higher, then we can reject the null hypothesis H_0.

Correct!

● Compute the t-statistic for each predictor. Compute the probability of that value of t occurring. If the probability < 0.05 for any predictor then reject the null hypothesis H_0.

Some software packages provide p values for each coefficient. These values can be used to determine whether the null hypothesis can be rejected.

○ Compute the Residual Sum of Squares (RSS) for a simple linear regression using each individual member of the possible features which could be used for regression. If there exists at least one with RSS ≤ 0.05 then reject the null hypothesis H_0.

Question 5

0 / 0 pts

What was the **most interesting** concept in the material you reviewed for this pre-class assignment? Be specific - wherever possible, include page numbers, filenames, concept names to help your instructor understand what you are referring to:

Your Answer:

The section on qualitative predictions was the most interesting.

Question 6

1 / 1 pts

What was the **most confusing** aspect of the material you reviewed? Be specific - wherever possible, include page numbers, filenames, concept names to help your instructor understand what you are referring to:

Your Answer:

Interpreting the results of the regression, i.e. interpreting what the coefficients that come out of the model are was the most confusing part.

Quiz Score: **1** out of 1