GSCM 410/510 – Summer 2021

Solutions Week 3

Directions

It is OK to work together on the Homework. Of course, make sure that everyone working together on an assignment is learning the material. Each person still turns in their own homework even if exactly the same as that of other group members. Learning this material well enough to do the homework problems is how you succeed in this course.

Short-Answer Problems

These concepts can appear on the short-answer part of the tests. As part of this homework, answer the following questions, usually just several sentences that include the definition.

1. *Compare the two primary data visualization Python packages: seaborn and matplotlib.*

matplotlib is the Python standard for visualizations. *seaborn* is a newer more modern alternative that is somewhat easier to use and tends to produce more elegant visualizations.

2. *What is the distinction between a stacked and unstacked bar chart? What do stacked and unstacked bar charts display?*

Stacked and unstacked bar charts display the relation between two categorical variables. The values of one categorical variable are listed on one axis with the associated bars, and the associated counts or proportions listed on the other axis. For a stacked bar chart, each bar is divided into the corresponding values of the second categorical variable. For an unstacked bar chart, there is a separate bar for each level of the second categorical variable.

3. *How is the concept of data aggregation related to a pivot table?*

Data aggregation summarizes the values of a numerical variable with descriptive statistics across different groups defined by levels of one or more categorical variables. The pivot table displays these summary statistics for the different groups.

4. *Describe two different data visualizations for different groups of data.*

Bar chart: Visualize the numerical value associated with each category of a numerical variable, such as the count of each category in a data set.

Histogram: Visualize the count or proportion of data values in each bin of similar values for a numerical variable.

Scatterplot: Visualize the relationship between two numerical variables by plotting each point defined with coordinates of the two values for the two variables.

5. How does a heat map differ from a standard correlation matrix?

The heat map uses colors to indicate the value of the correlation between each pair of a set of two numerical variables. The correlation matrix lists the numerical correlations directly. A modified heat map can show both the colors and the correlations.

6. What is supervised machine learning?

Based on a prediction equation, or in more complex cases, a network of inter-related prediction equations, the supervised machine learning forecasts unknown values of a variable of interest based on the values of known variables related to the unknown variable of interest.

7. What are the two goals of supervised machine learning?

Two important goals are accomplished with regression analysis:

- Understand the relationship between a predictor variable (feature) and the response variable
- Forecast the unknown future values of a response variable

Adding relevant predictor variables with new information contributes to both goals. Additional relevant predictor variables contribute to our understanding of the relation between a predictor and response variables with the values of all other predictors held constant, and increase the forecasting accuracy of the model, and via the imposition of statistical control.

8. What is a linear model? What are its parameters?

A weighted sum of variables plus a constant term.

9. What is the shape of the visualization of a linear model?

A straight surface. In two dimensions, that is a line. In three dimensions, a cube. And beyond.

10. What is y^? What are the two primary situations in which it is applied?

Given a regression model, $\hat{y}$ is the value calculated from the values of the predictor variables.

If applied to the data from which the model was estimated, $\hat{y}$ is the of y, that is, fit by the model for the associated values of the predictor variables, called the fitted value.

If predicting a future event for which the value of y was not available or used to estimate the model, then $\hat{y}$ is called the predicted value or the forecasted value.

11. Graph of X with y vs the graph of X with y^.

The graph of X with $\hat{y}$ is a single line (for a linear function to predict y). The graph of X and y is a scatter plot.

Note: For multiple regression, the topic of this homework, there are at last three variables in the analysis, y, X1 and X2. This plot would be in three dimensions, and more dimensions for more variables. However, the concept of the basic error term, the residual $e = y - \hat{y}$, still applies. Just that the plot is multi-dimensional.

12. Meaning of the slope coefficient in y^ = b0 + b1X1.

In this regression model with a single predictor (feature), b_1 is the slope coefficient, estimated from the data. The slope coefficient determines, on average, how much y changes with a increase of 1 unit in X. When applied to a regression model, this change in y is the average (or expected) change. (If there is more than one predictor variable in the model, then the values of all the other predictor variables are held constant.)

With multiple regression, each slope coefficient is interpreted with the values of the remaining predictor variables (features) held constant.

13. Meaning and interpretation of the hypothesis test of the slope coefficient.

Each predictor variable in the model is associated with a slope-coefficient. Each test evaluates the hypothesis of no relationship between predictor and target variables. The null hypothesis is that the population slope coefficient that relates x to y is 0, beta=0. The meaning of the slope coefficient is that as x increases by one unit, y, on average, changes by the value of beta.

14. Meaning and interpretation of the confidence interval of the slope coefficient.

Each predictor variable in the model is associated with a slope-coefficient. The slope-coefficient specifies the average change in y for a unit increase in the change in the predictor variable. For each estimated slope coefficient, the sample slope coefficient, b, there is a corresponding population value beta. If zero is in the confidence interval, then the analysis is unable to demonstrate a relationship between the predictor variable and the response variable, with all other predictor variables held constant.

15. Meaning of the residual variable e.

Residual variable

e is the difference between the actual value of y and the estimated value of y, $\hat{y}$. The residual or error represents the influences on the value y not explained or accounted for by the model.

16. Criterion of ordinary least squares regression to obtain the estimated model.

The least squares criterion is the choice of the regression model that minimizes the sum of squared residuals across all the rows of data in the analysis. That is, this estimation process yields values of each

b_j such that, as a set, yield the linear function that results in the smallest possible sum of squared residuals $y - \hat{y}$.

17. How is the least-squares regression model obtained with a gradient descent solution?

An initial, even arbitrary solution for the model parameters is given. Then, to minimize the squared errors across all the rows of data, the parameter values are changed. Then again. Then again, each time getting closer to the smallest possible sum of squared errors. The process stops when changing the parameter estimates results in virtually no change in the sum of squared errors.

18. Model Fit: The standard deviation of the residuals to interpret model fit.

If the residuals are normally distributed as the result of a random process, as they usually are, then +2 and -2 standard deviations on either side of zero contains about 95% of the forecasting errors. How is the standard deviation of the residuals used to interpret model fit?

19. Why is R-squared called a relative index of fit?

R-sq literally compares the residuals from two models: the specified model, and the null model where the X's are unrelated to y so that the forecast is just the mean of y, which plots as a straiht line.

20. What is model validation and what is the problem training data to validate a model?

Every data set sampled from a population differs from any other data set sampled from the same population. Every sample reflects the underlying population values, but every corresponding sample value, such as the mean, does not equal the corresponding population value. So fitting sample data from which the model trained fits random sampling error as well as true, stable population characteristics. A model can fit training data perfectly, but have no useful ability to forecast on new (testing) data.