Stat 462/862 Final Project

Aaron Springford

Due: November 26 2014 in class

This project has three parts. The number of marks for each part is indicated at the beginning of each part. For each part you will provide a maximum two-page description of the approach you took to your solution. Your description should be typed in a minimum 12pt font and use minimum 1 inch page margins. You will be graded according to how close you come to the correct answer, relative to your classmates (objective score), as well as your rationale and description of your approach (subjective score).

1 Parameter estimation

This question is worth 15 points – 10 for objective score and 5 for subjective score.

Refer to the dataset contained in the file *dataset1.csv*

For this part, you are to provide an estimate of the linear relationship between two covariates of interest (denoted $X.1$ and $X.2$, in columns 2 and 3 of the provided dataset) and the response variable (denoted y, in column 1 of the provided dataset). Additional (possibly confounding) covariates $X.3, ..., X.9$ are provided in columns 4-10. You can assume that the true relationships between each of the two covariates of interest $X.1$ and $X.2$ and the response variable y is linear (because they *are* linear). Provide a point estimate $\hat{\beta}$ as well as 100 bootstrap 50% confidence intervals for each of the two linear coefficients. Also detail your approach – how did you arrive at your estimates?

Your objective score will be calculated as follows:

1. Calculate squared error loss: $L = (\hat{\beta}_1 - \beta_1)^2 + (\hat{\beta}_2 - \beta_2)^2$. Calculate your rank in the class based on L – for example, if there are 5 students who had a smaller L, then your rank would be 6. If you and a friend submit identical answers, and there are 5 students who had a smaller L, then you and your friend will both be rank 7.

2. Determine the proportion of intervals p_1 and p_2 that contain the true parameter value. Calculate the total absolute difference $D = |p_1 - 0.5| + |p_2 - 0.5|$, which will be zero for exact coverage. Calculate your rank in the class based on 1. above.

3. Calculate the sum of your ranks in 1. and 2. (call this R). Your objective score will be
$$S = 10 - 5\frac{(R-2)}{2(n-1)}$$
where n is the number of students in the class who submitted a response.

Your subjective score will be based on my reading of your written document, as follows:

Subjective Score	Description
5	*Excellent* - Precise and detailed description of your approach, including rationale. Exceeds expectations.
4	*Good* - Excellent description of approach, including rationale. Some minor problems. Met expectations.
3	*Acceptable* - Description of approach has some issues. Rationale not perfectly sound or unclear.
2	*Poor* - Description of approach and/or rationale lacking coherence.
1	*Very poor* - Severe issues with understanding.
0	*Missing*

2 Response curve approximation

This question is worth 15 points – 10 for objective score and 5 for subjective score.

Refer to the dataset contained in the file *dataset2.csv*

For this part, you are to approximate the correct model structure (i.e. response curve f) for the data given. In the file are (x, y) data. Fit a function $\hat{f}$ of your choosing through the (x, y) pairs over the domain of x. Be sure to detail your approach clearly.

Your objective score will be calculated as follows:

1. Calculate the average squared difference L between f (the true function) and $\hat{f}$ at 100 points evenly spaced over the domain of x. Calculate your rank in the class based on L – for example, if there are 5 students who had a smaller L, then your rank would be 6. If you and a friend submit identical answers, and there are 5 students who had a smaller L, then you and your friend will both be rank 7.

2. Call R your rank from 1. Your objective score will be
$$S = 10 - 5\frac{(R-1)}{(n-1)}$$

where n is the number of students in the class who submitted a response.

Your subjective score will be based on my reading of your written document, as follows:

Subjective Score	Description
5	*Excellent* - Precise and detailed description of your approach, including rationale. Exceeds expectations.
4	*Good* - Excellent description of approach, including rationale. Some minor problems. Met expectations.
3	*Acceptable* - Description of approach has some issues. Rationale not perfectly sound or unclear.
2	*Poor* - Description of approach and/or rationale lacking coherence.
1	*Very poor* - Severe issues with understanding.
0	*Missing*

3 Prediction

This question is worth 20 points – 10 for objective score and 10 for subjective score.

Refer to the training dataset contained in the file *dataset3training.csv* and the test dataset contained in the file *dataset3test.csv*

For this part, your task is to predict the responses of the subjects in the test dataset (whose covariate information is provided in the file *dataset3test.csv*) using the information provided in the training dataset (*dataset3training.csv*). You will notice that the data contain some missing values (coded as NA's), as well as some categorical variables among the covariates. The response variable of interest is labeled y in the training data, and the covariates are labeled $x.1, ..., x.30$. Your job is to use the training data to come up with a method of prediction (using the covariates) for the test data. (You will notice that I have witheld the y values in the test data so you won't be tempted.) Once you have your method of prediction, generate predictions $\hat{y}$ for each of the subjects in the test dataset. Be sure to detail your approach clearly.

Your objective score will be calculated as follows:

1. Calculate the total squared difference L between y and $\hat{y}$ for each of the subjects in the test dataset. Calculate your rank in the class based on L – for example, if there are 5 students who had a smaller L, then your rank would be 6. If you and a friend submit identical answers, and there are 5 students who had a smaller L, then you and your friend will both be rank 7.

2. Call R your rank from 1. Your objective score will be

$$S = 10 - 5\frac{(R-1)}{(n-1)}$$

where n is the number of students in the class who submitted a response.

Your subjective score will be based on my reading of your written document, as follows:

Subjective Score	Description
10	*Excellent* - Precise and detailed description of your approach, including rationale. Exceeds expectations.
8	*Good* - Excellent description of approach, including rationale. Some minor problems. Met expectations.
6	*Acceptable* - Description of approach has some issues. Rationale not perfectly sound or unclear.
4	*Poor* - Description of approach and/or rationale lacking coherence.
2	*Very poor* - Severe issues with understanding.
0	*Missing*