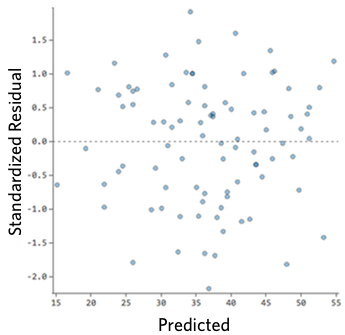
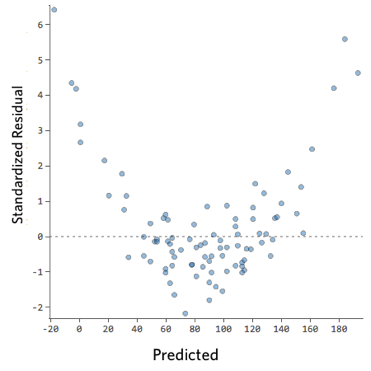
1530: Data Mining

Midterm Fall 2020

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Instructor: | Konstantinos Pelechrinis |  | Name: | Angela Tseng |

1. Read carefully each multiple-choice question and circle the correct answer. Every question is worth 3 points.
2. You are building a linear regression model for predicting the total monthly revenue of a restaurant chain. You have tried 3 different models and the following are the plots of the (standardized) residuals versus the predicted values. For which one **can** linear regression a suitable model?
   1. 
   2. 
   3. 
   4. None of the above
3. You have built three models for predicting the probability of a biking accident for trips through a specific road and you have calculated their reliability curve. Which model would you choose to use?
   1. 
   2. 
   3. 
   4. None of the above
4. You are trying to build a model to predict the probability of the stock of a specific company will increase the upcoming day. You are evaluating the *training* accuracy and you obtain an almost 98% accuracy. Which of the following is the action that you would **NOT** take:
   1. Use this model to make my future investment decisions
   2. Build a model with less complexity
   3. Perform cross validation
   4. Try using regularization
5. Consider all the digital cameras on Amazon. Each camera has on average 450 reviews and a rating of 4.6. Consider now digital camera X, with an average rating of 3.8. The Bayesian average will be very close to 3.8 if the number of reviews for this camera is:
   1. 1
   2. 15206
   3. Any value less than 10
   4. 13
6. A Poisson regression for a dependent variable Y that follows a Poisson distribution models:
   1. The expected value λ of the distribution through a linear combination of the independent variables **x**, i.e., λ = **b**T**x**.
   2. The expected value of the distribution through a normal distribution
   3. +The logarithm of the expected value of the distribution through a linear combination of the independent variables, i.e., log(λ) = **b**T**x**.
   4. None of the above
7. Each of the following questions is worth 2 points. Read carefully and answer choose whether the sentence is true or false.

1. Bayesian linear regression requires the dependent variable to follow a normal distribution. T

2. A model with increased complexity will have low variance. F

3. Among two probability models M1 and M2, with Brier score 0.21 and 0.19 respectively, model M2 is the better calibrated one. F

4. The validation set is used for model and/or feature selection. T

5. For count data with variance 5 times higher than their mean, a Poisson regression is still a good model. T

1. The following question is worth 5 points and it is for **extra credit**.

(i). An NBA finals series, is a “best of 7” series, that is, there is a maximum of 7 games played between the two teams and the team that wins 4 of them is the series winner (and the NBA champion). This means that the total number of games in the series can be anywhere between 4 to 7. Let X be the random variable representing the number of games in an NBA finals series. Assuming the teams matched at the NBA finals are equally matched, i.e., the win probability for each team, for each game, is 0.5, what is the probability distribution and the expected value of X?

With four games, there are 8 scenarios (2\*1/2^4), five games there are 16 (8\*1/2^5), for six games there are 10 scenarios, etc. Only some of those results in a team actually winning. For 7, c(6,3) is 20 because you only look at the first 6 games, times 2 since there are two ways the last game could go, then put it over 2^7. The results are…

4 games: 1/8

5 games: 1/4

6 games: 5/16

7 games: 5/16

The length of the NBA final series for the past 57 NBA finals is given in the following table:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Length | 4 games | 5 games | 6 games | 7 games |
| **# series** | 7 | 13 | 22 | 15 |

(ii). Do you think teams are usually equally matched? Describe the steps that you would take to answer this question with statistical arguments.

For your reference, the number of possible combination of r objects from a set of n objects is given by: , where .

Logically, it makes sense that the teams are not always equally matched. I tried using a chi-squared test for this part (I seem to like using it for a lot of problems like in HW 1) but the idea is to test the null hypothesis that the teams are not equally matched. We result in accepting that they are not equally matched. However, something to consider is that the calculation has to be done over 57, but in this case is done over 64.

Table

Description automatically generated