HW4: Team [12]

Yunxuan Li, Wenxin Liao, Yan Zhao

Due October 13, 2017

This problem set has several dependent parts, so plan accordingly. Here is a suggested outline to finish the assignment on time:

- Start problems 1-2, 4, 7 and 10 prior to Monday individually and with your group adding working code and minimal documentation for now. It is important to get a head start on the model building.
- Try problem 3 and 6 prior to lab Wednesday so that you will be prepared to ask questions about simulation and coding with the goal of having a minimal working version for 3 and 6 by the end of lab. This will help with the later questions where you apply it to the other models. (work as much on those as well)
- don't forget midterm and take time to enjoy fall break
- Try problem 12, 14 and 15 before Lab on the 11th; use lab time to refine code, ask questions about interpretation, theory etc.
- finish write up and turn in on Sakai on the 13th. Please let us know if there are problems with missing packages for wercker as you go so that you have a passing badge. Remove any instructions like this and above to clean up the presentation.

Preliminaries

Load the college application data from Lab1 and create the variable `Elite` by binning the `Top10perc` variable. We are going to divide universities into two groups based on whether or not the proportion of students coming from the top 10% of their high school classes exceeds 50 %. We will also save the College names as a new variable and remove `Accept` and `Enroll` as temporally they occur after applying, and do not make sense as predictors in future data.

```
data(College)
College = College %>%
  mutate(college = rownames(College)) %>%
  mutate(Elite = factor(Top10perc > 50)) %>%
  mutate(Elite =
           recode(Elite, 'TRUE' = "Yes", 'FALSE'="No")) %>%
  select(c(-Accept, -Enroll))
```

We are going to create a training and test set by randomly splitting the data. First set a random seed by

```
# do not change this; for a break google `8675309`
set.seed(8675309)
n = nrow(College)
n.train = floor(.75*n)
train = sample(1:n, size=n.train, replace=FALSE)
College.train = College[train,]
College.test = College[-train,]
```

1. Create scatter plots of predictors versus `Apps` using the training data only. If you use pairs or preferably `ggpairs` make sure that `Apps` is on the y-axis in plots versus the other predictors. (Make sure that the plots are legible, which may require multiple plots.)

Comment on any features in the plots, such as potential outliers, non-linearity, needs for transformations etc.

2. Build a linear regression model to predict `Apps` from the other predictors using the training data. Present model summaries and diagnostic plots. Based on diagnostic plots using residuals, comment on the adequacy of your model.

3. Generate 1000 replicate data sets using the coefficients from the model you fit above. Using RMSE as a statistic,
$$\sqrt{\sum_i (y^{\text{rep}} - \hat{y}_i^{\text{rep}})^2/n}$$
, how does the RMSE from the model based on the training data compare to RMSE's based on the replicated data. What does this suggest about model adequacy? Provide a histogram of the RMSE's with a line showing the location of the observed RMSE and compute a p-value. Hint: write a function to calculate RMSE.

4. Build a second model, considering transformations of the response and predictors, possible interactions, etc with the goal of trying to achieve a model where assumptions for linear regression are satisfied, providing justification for your choices. Comment on how well the assumptions are met and and issues that diagnostic plots may reveal.

5. Repeat the predictive checks described in problem 3, but using your model from problem 4. If you transform the response, you will need to back transform data to the original units in order to compute the RMSE in the original units. Does this suggest that the model is adequate? Do the two graphs provide information about which model is better?

6. Use your two fitted models to predict the number of applications for the testing data, `College.test`. Plot the predicted residuals $y_i - \hat{y}_i$ versus the predictions. Are there any cases where the model does a poor job of predicting? Compute the RMSE using the test data where now RMSE $= \sqrt{\sum_{i=1}^{n.test}(y_i - \hat{y}_i)^2/n.test}$ where the sum is over the test data. Which model is better for the out of sample prediction?

7. As the number of applications is a count variable, a Poisson regression model is a natural alternative for modelling this data. Build a Poisson model using main effects and possible interactions/transformations. Comment on the model adequacy based on diagnostic plots and other summaries. Is there evidence that there is lack of fit?

8. Generate 1000 replicate data sets using the coefficients from the Poisson model you fit above. Using RMSE as a statistic, $\sqrt{\sum_i (y^{\text{rep}} - \hat{y}_i^{\text{rep}})^2/n}$, how does the RMSE from the model based on the training data compare to RMSE's based on the replicated data. What does this suggest about model adequacy? Provide a histogram of the RMSE's with a line showing the location of the observed RMSE and compute a p-value.

9. Using the test data set, calculate the RMSE for the test data using the predictions from the Poisson model. How does this compare to the RMSE based on the observed data? Is this model better than the linear regression models in terms of out of sample prediction?

10. Build a model using the negative binomial model (consider transformations and interactions if needed) and examine diagnostic plots. Are there any suggestions of problems with this model?

11. Carry out the predictive checks for the negative model model using simulated replicates with RMSE and add RMSE from the test data and observed data to your plot. What do these suggest about 1) model adequacy and 2) model comparison? Which model out of all that you have fit do you recommend?

12. While RMSE is a popular summary for model goodness of fit, coverage of confidence intervals is an alternative. For each case in the test set, find a 95% prediction interval. Now evaluate if the response is in the test data are inside or outside of the intervals. If we have the correct coverage, we would expect that at least 95% of the intervals would contain the test cases. Write a function to calculate coverage (the input should be the fitted model object and the test data-frame) and then evaluate coverage for

each of the models that you fit (the two normal, the Poisson and the negative binomial). Include plots of the confidence intervals versus case number ordered by the prediction, with the left out data added as points. Comment on the plots, highlighting any unusual colleges where the model predicts poorly.

13. Provide a table with the 1) RMSE's on the observed data, 2) RMSE's on the test data, 3) coverage, 4) the predictive check p-value with one row for each of the models and comment the results. Which model do you think is best and why? Consider the job of an administrator who wants to ensure that there are enough staff to handle reviewing applications. Explain why coverage might be useful.

14. For your "best" model provide a nicely formatted table (use `kable()` or `xtable()`) of relative risks and 95% confidence intervals. Pick 5 of the most important variables and provide a paragraph that provides an interpretation of the parameters (and intervals) that can be provided to a university admissions officer about which variables increase admissions.

Some Theory

15. Gamma mixtures of Poissons: From class we said that

$$\begin{aligned}
Y \mid \lambda &\sim P(\lambda) && (1)\\
p(y \mid \lambda) &= \frac{\lambda^y e^{-\lambda}}{y!} && (2)\\
& && (3)\\
\lambda \mid \mu, \theta &\sim G(\theta, \theta/\mu) && (4)\\
p(\lambda \mid \mu, \theta) &= \frac{(\theta/\mu)^\theta}{\Gamma(\theta)} \lambda^{\theta - 1} e^{-\lambda\theta/\mu} && (5)\\
& && (6)\\
p(Y \mid \mu, \theta) &= \int p(Y \mid \lambda) p(\lambda \mid \theta, \theta/\mu) d\lambda && (7)\\
&= \frac{\Gamma(y+\theta)}{y!\Gamma(\theta)} \left(\frac{\theta}{\theta+\mu}\right)^\theta \left(\frac{\mu}{\theta+\mu}\right)^y && (8)\\
Y \mid \mu, \theta &\sim NB(\mu, \theta) && (9)
\end{aligned}$$

Derive the density of $Y \mid \mu, \theta$ in (8) showing your work using LaTeX expressions. (Note this may not display if the output format is html, so please use pdf.) Using iterated expectations with the Gamma-Poisson mixture, find the mean and variance of Y, showing your work.