Discrete Response Model
Lecture 5

Models for Count Response, Discrete Response Model Evaluation, and Model Selection

datascience@berkeley

Model Evaluation

Model Assumptions

- Any statistical model comes with a set of statistical assumptions.
- When estimating a GLM, we make the following assumptions:

1. The data are generated independently by an identical distribution we specify, be that binomial, Poisson, and so on.
2. The mean of the distribution is linked to the explanatory variables by the “link” function we specify.
3. The link relates to the explanatory variables in a linear fashion.

On Residuals

- "Raw" residuals, the difference between observed and predicted values, are not useful to evaluate the underling assumptions of GLM and count response models because the residuals of these models depend on their means.
- As such, we introduce the concepts of Pearson residuals and standardized Pearson residuals
- The idea is to account for the variance related to the response variables, be that categorical or count response.

$$\boxed{y_i - \hat{y}_i}$$

$$e_i = \frac{y_i - \hat{y}_i}{\sqrt{\hat{Var}(Y_i)}}$$

$$r_i = \frac{y_i - \hat{y}_i}{\sqrt{\hat{Var}\left(Y_i - \hat{Y}_i\right)}} = r_m = \frac{y_i - \hat{y}_i}{\sqrt{\hat{Var}(Y_i -)(1 - h_i)}},$$

where $i = 1, \ldots, n$ and h_i is the i^{th} diagonal element of the hat matrix.

On Residuals

- Although there are other kinds of residuals, such as *deviance residuals*, we will use *standardized residuals*.
  - It is easy to compute.
  - It is easy to interpret: It can be viewed approximately as "observations from a standard normal distribution."
- Of course, we have to use this "approximation" with caution.
- Residuals from binomial models when there are only a small number of exploratory variables combinations are poorly approximated by a normal distribution.
  - The normal approximation is also not appropriate when is near 0 or 1.
  - A similar situation happens with Poisson model when the estimated means are very small.

Computing Residuals in R

- The generic function **residuals()** has methods functions that work on model-estimated objects produced by **glm()**.
- These functions use the **type** argument to select the type of residuals computed.
  - "pearson" or "response" are used to selected Pearson and Raw residuals.
- **Residuals.glm()** methods also has a "deviance" argument value to produce deviance residuals.
- Standardized Pearson and deviance residuals are avaiable for the glm-class objects using **rstandard()**.

Residual Diagnostic Assessment

- The information content contained in a set of residuals can be easily revealed in graphs.
- Many of these graphs are similar to those used in the context of classical linear regression models.
- However, interpretation in GLM can be different.
- Residuals in GLM can be used to diagnose problems in the conditional mean model, possible outliers, and model assumptions, such as the choice of the probability family, that are inappropriate.
- As mentioned above, residual plots of count response model is subject to the caveats already noted, that is, in situation where there are a limited number of possible responses.

Residual Diagnostic Assessment

1. Residuals against each of the explanatory variables

- A plot of standardized residuals against each explanatory variable can show whether the form for the explanatory variable is appropriate.
- The plot should show:
  - Same variance throughout the range of the explanatory variable
  - No serious fluctuations in the mean value
- Add a smoother, such as *loess* smoother, to the plot to add the visualization.
- Watch out for a clear pattern of curvature.
- In binomial models, the loess curve can be weighted by so that larger numbers of trails contribute relatively more to the curve placement than those with relatively few.
- Note: Loess curves are highly variable when data are sparse or near the extreme values.
  - ➔ Do not overreact to the changes in the curve at the edge of the plot.

Residual Diagnostic Assessment

2. Residuals against the fitted values (or in the binomial model)

- It is useful to examine if the link function is appropriate.
- The plot should have:
  - Constant variance throughout the range of the response
  - No clear curvature
- A plot of the residuals against the linear predictor,:
  - Shows patterns of change in the mean residuals more clearly
  - Help diagnose how the link function g(.) should be changed to better fit the data

Residual Diagnostic Assessment

3. Any of these plots should be used to check for extreme residuals

- As noted before, only about 5% of standardized residuals should be beyond +/-2 and typically none beyond +/-3.
- Presence of a large amount of extreme residuals indicates overdisperson, meaning that there is more variability to the counts than what the model assumes.
- It may also indicate that there are missing explanatory variables from the model.

Slight different in interpretation for binomial models:

- Extreme residuals occur more often in regions of high or low estimated probability of success.
- To account for this situation, calculate the binomial probability that such an extreme value of could occur trails with probability

Berkeley SCHOOL OF INFORMATION