StatR 502 Homework 4

Gregor Thomas

Due Thursday, Feb. 4, 2016, 6:30 pm

Submission guidelines: please submit a knitted PDF or Word file.

1: Offsets and Complaints

In the `faraway` package, the `esdcomp` data set is about the **count** of complaints received about doctors working in an emergency room.

We could consider a logistic model, with `visits` as the number of "attempts", and `complaints` the number of "successes" (for a strange definition of "success"). But for rare events where we don't observe anywhere near the full range of possible probabilities (0 to 1), it can better to model the events as counts.

(a) Verify my assertion that the ratio of complaints per visits is low. What are the mean and maximum observed probabilities of a complaint?

(b) We want to fit a Poisson GLM modeling complaints, but we need like to explicitly account for `visits` as having a direct one-to-one effect on the rate of complaints. This can be done with an `offset()` term in the model formula. Fit a Poisson GLM using all the predictors *except* `visits`, then add to the model formula `+ offset(log(visits))`. This tells GLM to add `log(visits)` to the linear predictor, but it's coefficient will be fixed at 1 instead of estimated. (Why are we using `log(visits)`? Because the log link for the Poisson distribution means the linear predictor predicts `log(complaints)`.) The offset term not show up as a fitted coefficient in the model summary, but it will be in the model call. This use of Poisson GLM is called a **rate model**.

(*Note:* the most common use of `offset()` is exactly this, but if for any other reason you want to "fix" one or more regression coefficient, while still estimating other coefficients, `offset()` is the way to do it.)

(c) Run the function `drop1()` on your model. It will show you the AIC of the resulting model if you were to drop each of the model terms individually. Run the function `add1()` on your model. Set the `scope` argument to `~ (residency + gender + revenue + hours)^2`, which will tell `add1` to consider adding two-way interactions. Similar to `drop1`, it will show you the AIC if you were to add each of the possible terms. Add and/or drop a term or two to/from your model based on the `add1` and `drop1` advice. Show the summary of your model after a couple iterations. (*Note:* Do these things "behind the scenes". The only output you should include in your homework write-up is the model summary after an a couple iterations and perhaps a sentence saying what terms you added or dropped.)

(d) Try a quasipoisson model. What is its dispersion parameter? (Switching from Poisson to Quasi-Poisson will scale the standard errors by the square root of the dispersion parameter, so a dispersion parameter of 1 indicates no difference. The Galapagos example in class had a dispersion parameter of about 32.)

(e) Make a coefficient plot (inluding standard errors) of both the Poisson and Quasi Poisson models, as in error bars, as in slides 29-31 of Lecture 3 or in G&H Figure 4.6 (page 74) and Figure 9.9 (page 182). You can compare them in separate plots, in separate facets of the same plot, or in one plot using color to distinguish the two models. (If you choose to use one plot, you might want to use `position = position_dodge()` to avoid overplotting.) You can use `broom::tidy()` on your models to get nice data frames with the coefficient estimates and standard errors. In a sentence, what is the difference in estimated coefficients between the Poisson and Quasipoisson models?

2: New data exploration

Use the abalone dataset (posted on the website, with accompanying descriptive file). The response variable is the number of rings observed on the shell, which is directly related to the age of the abalone. We would like to predict the number of rings from other, more easily measured characteristics of the abalone.

(a) Make a few exploratory plots of the data.

(b) Fit a linear model (OLS) with rings as the response against all the other predictors (`rings ~ .`). Calculate the Cook's Distance and the leverages (you can use `hatvalues()` and `cooks.distance()` or `broom::augment`). Identify any concerning outliers and briefly discuss two possible coping strategies.

Book Problem (singular!)

Do G&H #6 and the additional part (d) below (pp. 133). Rather than using the data from the book website, use the `congress.txt` data on the course website. I added better column names, calculated vote proportions, and removed rows where the seat was uncontested. The `incumbent` column is coded as `-1` for republican incumbent, `0` for open race, and `1` for democratic incumbent. For part (b), you can use any robust regression method. The book suggests t-regression via `hett::tlm`, but feel free to explore others. Quantile regression (e.g., `quantreg::rq`) and `robustbase::lmrob` which uses an "MM estimator" are popular.

(d) Do you like the way `incumbent` is coded? You *could* change it to a factor, what additional comparisons would that allow you to make?