Fox Module 23 Generalized linear models probabilities HW

(The attached PDF file has better formatting.)

Homework assignment: Renewal Rates and Years Insured

Policy renewal rates increase the longer the policyholder has been insured.

- New policyholders often switch to another insurer at the end of the policy term.
- Policyholders who have stayed for ten years are very likely to stay another year.

The type of non-renewal affects the statistical modeling. At older ages, some policyholders do not renew because they die or they no longer have any exposure. Auto policyholders may give up driving and Homeowners policyholders may move to a retirement home.

- For pricing, it may not matter why the policyholder fails to renew. The insurer loses the value of future business whether the policyholder is alive or dead, driving or not driving, or still living in the home.
- For statistical modeling, the type of non-renewal affects the relation. The renewal rate for auto insurance rises with each renewal. As the number of years insured increases beyond the expected driving life of the policyholder, the renewal ratio may decrease. If the renewal rate increases and then decreases, a logit GLM may not work well.

The logit GLM in this exercise uses the percentage of renewals with the existing insurer vs all insurers. The data are from a telephone questionnaire with policyholders who did not renew. One question was whether the policyholder has coverage with another insurer or does not have insurance. The renewal rate is the number of policyholders who renew divided by the number who still have a policy with any insurer.

We model renewal rates as a function of years insured using a logit link function.

We examine *six month* policies that come up for renewal in 20X1. Each record shows

- The years already insured at the renewal data, ranging from 0.5 to 30.
  - 0.5 years means the policy is at its first renewal.
  - 30 years means the policy is at its 60^{th} renewal.
- Whether the insured renews the policy: True = renews and False = does not renew

From the individual records, we form an aggregate data base with 60 records. Each record has three fields.

- The policy age, ranging from 0.5 to 30 (1 half year to 60 half years).
- The exposures at that policy age, from 10,000 at 0.5 years to *** at 30 years.
- The percentage of policies that renew.

We relate the renewal rate to years insured. We examine regressions of the renewal rate to years insured and of the log odds of the renewal rate to years insured.

The exposures indicate the quality of the empirical data at each renewal date.

- At 0.5 years insured, the data are highly credible (many exposures).
- At 30 years, the data are less credible (few exposures).

GLMs use *weighted* regressions. The weights depend on the exposures and the conditional distribution function.

This homework assignment does not require you to fit the GLM. Fitting the GLM is hard by pencil and paper, but it requires only a single function in R: glm(renewals ~ years, …).

The homework assignment asks which points are more influential. The answer depends on the exposures at each point and the variance of the distribution at each point.

The data are in an Excel spread-sheet and data files. Use the format that is convenient. You can complete the homework using Excel and its *REGRESSION* add-in. If you want to fit the GLM, use the data files and R.

A. Graph the renewal rate as a function of years insured. Is the curve convex or concave? [You can answer this intuitively, since the renewal rate is bounded by 100%.]
B. Form a regression line linking renewal rates to years insured. What are the least squares estimates for α and β? [Ignore the exposures for this part.]
C. Does the regression line over-estimate or under-estimate the renewal rate for (i) 1 to 2 years-insured, (ii) 29 to 30 years-insured, (iii) 14.5 to 15.5 years-insured? [You can answer by comparing the observed vs fitted values or by comparing the curve with the regression line.]
D. The first point looks like an outlier. It is an influential point, so it skews the regression line. If we exclude this point from the regression, are α and β higher or lower? Which regression line has the higher R^2? Which regression line has the lower estimated σ^2? You can answer all these questions intuitively. If you want, check your work with Excel.
E. A simple regression gives the same weight to each point. Based on the exposures, which renewal rates have more random fluctuation: high or low years insured? Should we give more weight to high or low years insured in fitting the regression line? You don't have to do a weighted regression for the homework assignment.
F. The GLM uses link functions, distributions, and exposures. Form logits of the renewal rates. (The logits are the log odds.) Graph these logits as a function of years insured. What is the shape of this curve: convex, concave, or straight? Are there any outliers?
G. Regress the logit of the renewal rate on the years insured. Solve for α and β.

Form the graphs with Excel, R, or other software. *You don't have to submit the graphs with your homework assignment.* The graphs help you visualize the data and spot outliers. The Excel spreadsheet attached to this homework assignment has the graphs. Your graphs should look similar (or the same).

You should always graph data for a statistical study, such as the student project. Excel, R, and most statistical packages have excellent graphing tools.

Jacob: In this homework assignment, the logit transformation creates a linear curve. Is this generally true in real applications? What if the logit does not form a linear curve?

Rachel: The logit transformation of probabilities often creates linear curves, but not always. For some skewed distributions, we might us *complementary log log* transformations. If the distribution is symmetric but the thickness of the tails doesn't fit the logit distribution, we might try a probit distribution. We use whatever transformation creates a linear relation. The logit is a simple transformation, and it works well in many scenarios.

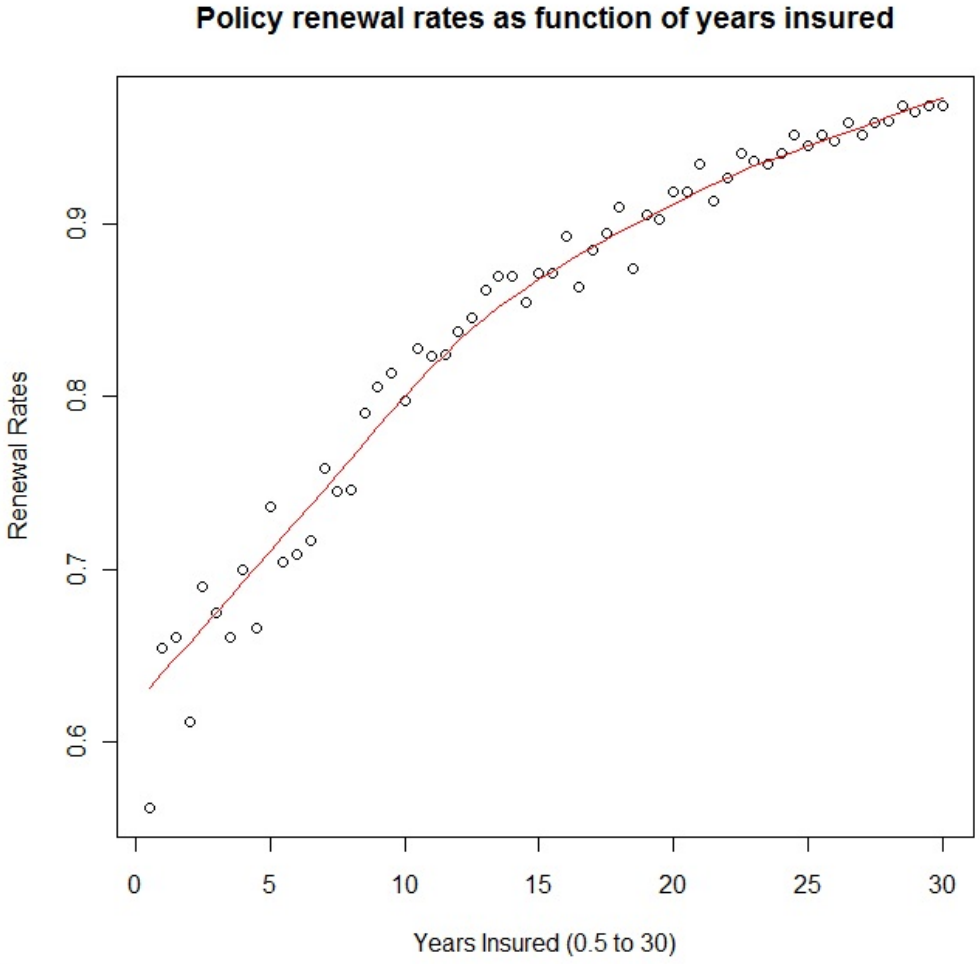
Jacob: How does the binomial distribution affect the GLM? We didn't use the distribution in this homework assignment.

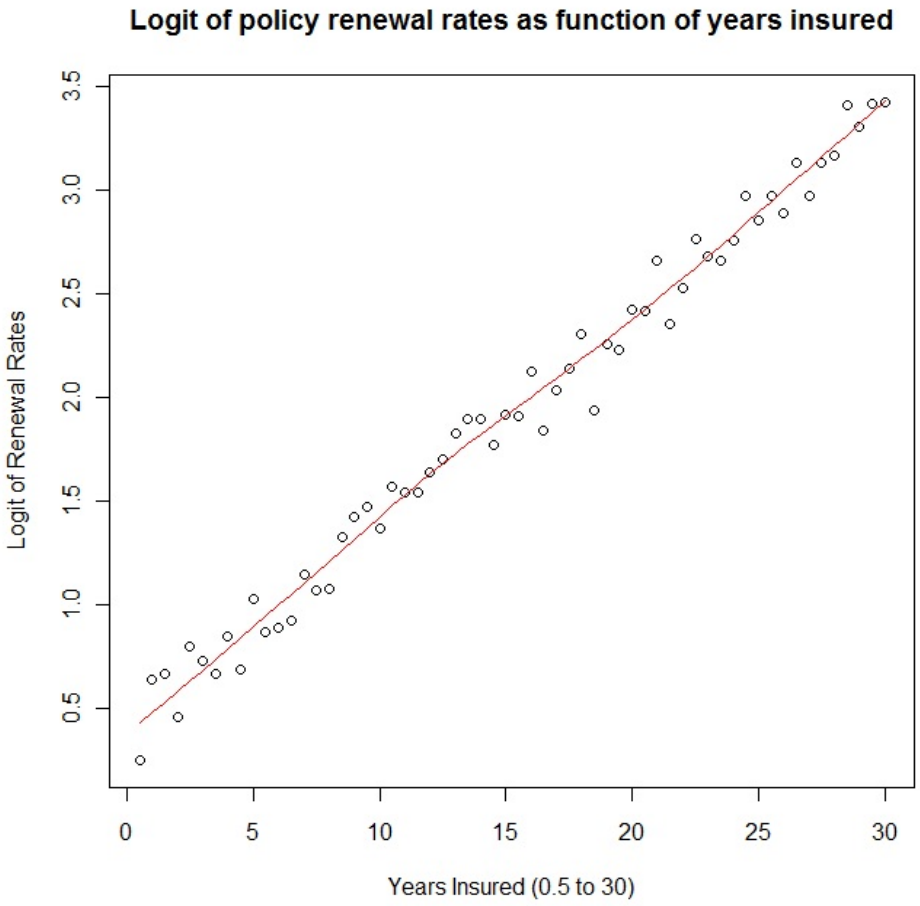
Rachel: The distribution is important, but it is harder to grasp. The homework assignment for the previous module deals with Poisson and Gamma distributions, and the same logic applies to binomial distributions. The variance of a binomial distribution is highest at the center and lowest in the tails.

Jacob: The GLM predicts the logits of the renewal rate. How do we get the renewal rate?

Rachel: Given the logit, the renewal rates is $1/(1 + e^{-\text{logit}}) = e^{\text{logit}} /(1 + e^{\text{logit}})$.

{The graphs below are what you should find.}


Policy renewal rates as function of years insured
Renewal Rates
0.6
0.7
0.8
0.9
0
5
10
15
20
25
30
Years Insured (0.5 to 30)


Logit of policy renewal rates as function of years insured
Logit of Renewal Rates
0.5
1.0
1.5
2.0
2.5
3.0
3.5
0
5
10
15
20
25
30
Years Insured (0.5 to 30)