Regression analysis Poisson GLM practice problems

(The attached PDF file has better formatting.)

** Exercise 22.1: Poisson GLM

An actuary uses a Poisson GLM with a log-link function to relate claim frequency to sex (male vs female) and location (urban vs rural), using dummy regressors of

- sex: 0 = female and 1 = male
- location: 0 = rural and 1 = urban

The estimates for the coefficients are

- β_1 = sex (male) = 0.2
- β_2 = location (urban) = 0.3

If female-rural drivers have expected claim frequencies of 5%, what are the expected claim frequencies for

A. male-rural drivers
B. female-urban drivers
C. male-urban drivers

Part A: The expected claim frequency for any driver is $\exp(\alpha + \beta_1 \times \text{sex} + \beta_2 \times \text{location})$

The expected claim frequency for female-rural drivers is $\exp(\alpha + 0.2 \times 0 + 0.3 \times 0) = \exp(\alpha)$.

The expected claim frequency for male-rural drivers is $\exp(\alpha + 0.2 \times 1 + 0.3 \times 0) = \exp(\alpha + 0.2) = \exp(\alpha) \times e^{0.2} = \exp(\alpha) \times 1.22140 = 5\% \times 1.22140 = 6.11\%$.

Part B: The expected claim frequency for male-rural drivers is $\exp(\alpha + 0.2 \times 0 + 0.3 \times 1) = \exp(\alpha + 0.3) = \exp(\alpha) \times e^{0.3} = \exp(\alpha) \times 1.34986 = 5\% \times 1.34986 = 6.75\%$.

Part C: The expected claim frequency for male-urban drivers is $\exp(\alpha + 0.2 \times 1 + 0.3 \times 1) = \exp(\alpha + 0.2 + 0.3) = \exp(\alpha) \times e^{0.2} \times e^{0.3} = \exp(\alpha) \times 1.22140 \times 1.34986 = 5\% \times 1.34986 = 8.24\%$.

Jacob: What is the value of α?

Rachel: $e^{\alpha} = 5\% \Rightarrow \alpha = ln(5\%) = -2.99573$.

** Exercise 22.2: Poisson GLM

An actuary uses a Poisson GLM with a log-link function to relate claim frequency to sex (male vs female) and location (urban vs rural), using dummy regressors of sex (male vs female) and location (urban vs rural).

The maximized likelihoods using neither sex nor location, sex only, location only, or both sex and location are

- H_0: neither sex nor location: 0.0001
- H_1: sex only: 0.0003
- H_2: location only: 0.0019
- H_3: both sex and location: 0.0021
- H_4: saturated model: 0.0028

The actuary uses a 5% significance level to test hypotheses, and $\chi^2_{0.05,1}$ (χ^2 test with 1 degree of freedom at a 5% significance level) = 3.84.

A. What are the residual deviances for models H_0, H_1, H_2, H_3, and H_4?
B. Is sex plus location a significant improvement on sex alone?
C. Is sex plus location a significant improvement on location alone?
D. What is the (pseudo-) R^2 for the models using sex only, location alone, and sex plus location?
E. Would the answers to Parts B and C change if the saturated model has a greater maximized likelihood?

Part A: The residual deviance is $2 \times (ln(L_s) - ln(L_m))$, where

- L_m = the maximized likelihood under the model in question.
- L_s = the maximized likelihood under a saturated model.

The residual deviances for the four models here are

- H_0: neither sex nor location: $2 \times (ln(0.0028) - ln(0.0001)) = 6.664$
- H_1: sex only: $2 \times (ln(0.0028) - ln(0.0003)) = 4.467$
- H_2: location only: $2 \times (ln(0.0028) - ln(0.0019)) = 0.776$
- H_3: both sex and location: $2 \times (ln(0.0028) - ln(0.0021)) = 0.575$
- H_3: saturated model: $2 \times (ln(0.0028) - ln(0.0028)) = 0.000$

The residual deviance for the saturated model is always zero.

Part B: The difference in the residual deviances for two models (G_0^2), one of which is nested within the other, has a χ^2 distribution with degrees of freedom equal to the number of extra parameters in the larger model. For sex only vs sex plus location, G_0^2 = the difference in the residual deviances = 4.467 – 0.575 = 3.892, which is more than $\chi^2_{0.05,1}$ (χ^2 test with 1 degree of freedom at a 5% significance level), which is 3.84. We reject the smaller model (sex alone) in favor of the larger model (sex plus location).

Jacob: What does this significance test mean?

Rachel: If the χ^2 value is 3.84, the probability of a difference in residual deviances of 3.84 by chance alone if the smaller model (sex alone) is true is 5%. Since the difference in residual deviances is more than 3.84, the probability of the smaller model (sex alone) being true is less than 5%.

Part C: For location only vs sex plus location, the difference in the residual deviances is 0.776 – 0.575 = 0.201, less than $\chi^2_{0.05,1}$ (χ^2 test with 1 degree of freedom at a 5% significance level), which is 3.84. We do not reject the smaller model (location alone) in favor of the larger model (sex plus location).

Jacob: Can we use this significance test to compare H_1 (sex only) with H_2 (location only)?

Rachel: We use this significance test to compare nested models, not models with unrelated explanatory variables. H_1 (sex only) is not nested in H_2 (location only) and H_2 is not nested in H_1, so we can't compare them. In this example, if H_1 had a residual deviance of 0.776, just like H_2, both explanatory variables may be significant; comparing the residual deviances of these two models doesn't tell us anything.

Jacob: Why do GLMs use the χ^2 distribution whereas classical regression analysis uses the *F* distribution?

Rachel: Both GLMs and classical regression analysis use both the χ^2 distribution and the *F* distribution. Use the χ^2 distribution if the variance is known; use the *F* distribution if the variance must be estimated from the data. For the binomial and Poisson conditional distributions of the response variable, the variance is known if the mean is known:

- Binomial distribution: $\sigma^2 = \mu \times (1 - \mu)$
- Poisson distribution: $\sigma^2 = \mu$

For the normal distribution assumed by classical regression analysis, the variance is unrelated to the mean.

If one uses a pseudo-Poisson or pseudo-binomial distribution for the GLM, and the dispersion parameter must be estimated from the data, one uses an *F* distribution, not a χ^2 distribution.

Part D: We define D_0 as the residual deviance for the model including only the regression constant α (termed the null deviance) and D_1 the residual deviance for the model in question. The (pseudo-) $R^2 = 1 - D_1 / D_0$ represents the proportion of the null deviance accounted for by the model.

- For the model using sex only, the R^2 is $1 - 4.467 / 6.664 = 32.97\%$.
- For the model using location only, the R^2 is $1 - 0.776 / 6.664 = 88.36\%$.
- For the model using sex plus location, the R^2 is $1 - 0.575 / 6.664 = 91.37\%$.

Jacob: Why is this called the pseudo-R^2 instead of the R^2?

Rachel: The R^2 for classical regression is the square of the correlation between the response variable and the explanatory variable. The pseudo- R^2 for a GLM is not the square of the correlation. The R^2 for classical regression analysis is the percentage of the RSS accounted for by the regression equation, and the pseudo-R^2 for the GLM is the percentage of the residual deviance account for by the GLM.

Part E: The maximized likelihood for the saturated model does not affect the G_0^2 used to compare nested models. If the maximized log-likelihoods for two models W and Z are L_w and L_z, and the maximized log-likelihood for the saturated model is L_s,

$$G_{w,z}^2 = 2 \times (ln(L_s) - ln(L_w)) - 2 \times (ln(L_s) - ln(L_z)) = 2 \times (ln(L_z) - ln(L_w))$$