STA2201 Assignment 1

Matthew Scicluna

2016-02-02

Short Answer

A Simulation Study

On the Coverage Probabilities

We simulated 100 data sets by simulating random draws from a poisson distribution and fitting the correct model to the data. We used the point estimate of the coefficient of x along with its standard error to compute a 2 standard error confidence interval and computed the coverage probability of this interval. We found that it was 0.96.

The coverage probability is quite high, however computing it requires knowing beforehand the parameter value, so this approach is pretty pointless.

Approximation by a Normal Distrubution

We now check whether the coefficient for x can be approximated by a Normal distribution centred on 0.2.

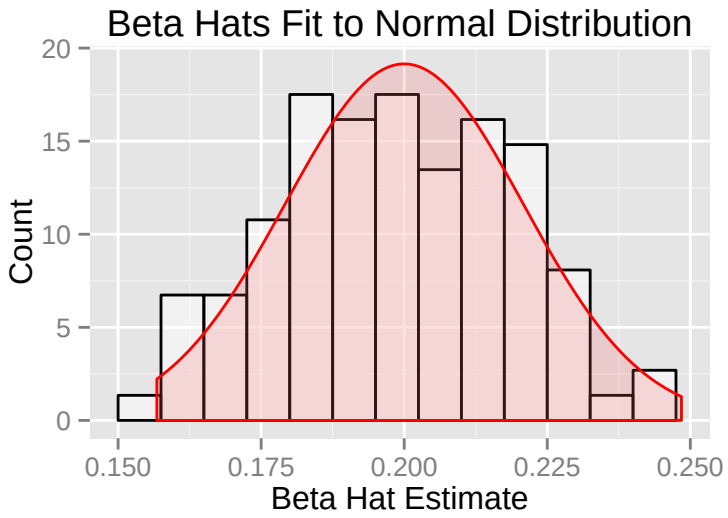

Figure 1: Checking if the distribution of the estimates can be fit to a normal distribution

Note that for the variance of this distribution we used 0.02 – the standard error of the coefficient in place of its (unknown) variance. This appears to be a reasonable fit. We further did a Shapiro-Wilks test of Normality for which we rejected the null hypothesis that the data did not come from a Normally distribution population with p>0.1.

Distribution of the Likelihood Ratio Statistics

We now calculate 100 likelihood ratio statistics for testing if the x coefficient is 0.2. We can see that this appears roughly chi-squared with one degree of freedom, which is what we would expect if the null hypothesisthat we are testing were true. In this case our null hypothesis is in fact true – that the true value of x coefficient is 0.2.

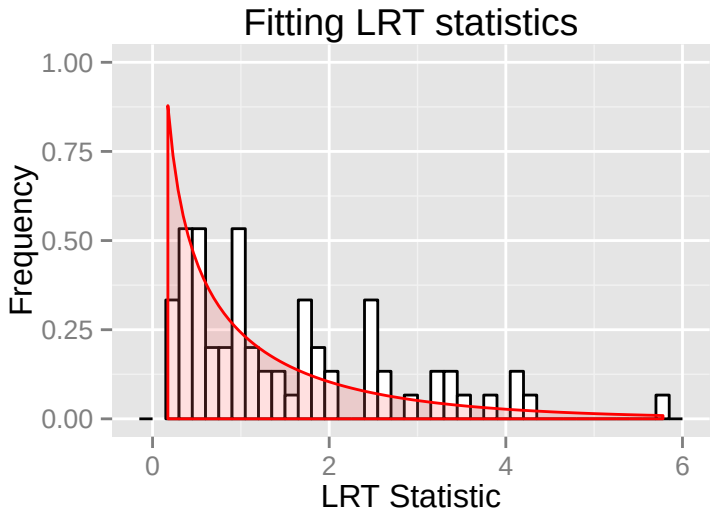

Figure 2: Fitting LRT statistics to a Chi Squared Distribution with 1 df

Distribution Functions

We now derive the parameters for each distribution which will result in random variables with mean 2 and variance 3.

Parameter Derivations

Zero-inflated Poisson

The zero-inflated Poisson distribution has the following mass function:

$$P(y_j = 0) = \pi + (1 - \pi)e^{-\lambda}$$

$$P(y_j = h_i) = (1 - \pi)\frac{\lambda^{h_i} e^{-\lambda}}{h_i!}, \qquad h_i \geq 1$$

The mean is $(1 - \pi)\lambda$, and the variance is $\lambda(1 - \pi)(1 + \lambda\pi)$. From the above see that

$$\lambda(1 - \pi)(1 + \lambda\pi) = 3 \Rightarrow 2(1 + \lambda\pi) = 3 \Rightarrow \lambda = \frac{1}{2\pi}$$

And finally we substitute this into the equation for the mean to get $\pi = \frac{1}{5}$ and $\lambda = \frac{5}{2}$

Gamma

The Gamma distribution has the following density:

$$f(x;\alpha,\beta) = \begin{cases} \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x} & x \geq 0, \\ 0 & x < 0, \end{cases}$$

It has the familiar formulas $\alpha\beta$ and $\alpha\beta^2$ for mean and variance respectively. Note that α is the shape parameter and β is the rate (inverse of the scale) parameter. Clearly

$$\alpha\beta = 2 \Rightarrow \alpha = \frac{2}{\beta} \Rightarrow 2\beta = 3$$

And after some simple algebraic manipulations we have that $\alpha = \frac{4}{3}$ and $\beta = \frac{3}{2}$

Weibull

The Weibull distribution has the following density:

$$f(x;\lambda,k) = \begin{cases} \frac{k}{\lambda}\left(\frac{x}{\lambda}\right)^{k-1} e^{-(x/\lambda)^k} & x \geq 0, \\ 0 & x < 0, \end{cases}$$

The mean and vairance for a Weibull distribution is $\lambda\Gamma(1+1/k)$ and $\lambda^2\left(\Gamma\left(1+\frac{2}{k}\right) - \left(\Gamma\left(1+\frac{1}{k}\right)\right)^2\right)$ respectively.

Rather than analytically solve these equations for the parameters we can use a numeric optimizer. We did this in R using the `nleqslv` function in the nleqslv package. We got $\lambda = 2.11$ $k = 1.16$.

Log-Normal

The Log Normal distribution has the following density:

$$f(x;\mu,\sigma) = \begin{cases} \frac{1}{x\sigma\sqrt{2\pi}}\, e^{-\frac{(\ln x-\mu)^2}{2\sigma^2}} & x \geq 0, \\ 0 & x < 0, \end{cases}$$

The mean and variance of the Log-Normal distribution are $e^{\mu+\sigma^2/2}$ and $(e^{\sigma^2}-1)e^{2\mu+\sigma^2}$ respectively. Equating these to 2 and 3 we solve for μ and σ in the following way.

$$e^{\mu+\sigma^2/2} = 2 \Rightarrow \ln 4 = 2\mu + \sigma^2 \Rightarrow e^{\sigma^2} - 1)4 = 3 \Rightarrow e^{\sigma^2} - 1) = \frac{3}{4} \Rightarrow \sigma^2 = \ln\frac{7}{4}$$

and

$$\ln 4 = 2\mu + \ln\frac{7}{4} \Rightarrow \mu = \frac{\ln\frac{16}{7}}{2}$$

Negative Binomial

The negative Binomial distribution has the following mass function:

$$P(X = k) = \binom{k + r - 1}{k} \cdot (1 - p)^r p^k,\ k = 0, 1, 2, ...$$

The mean and variance of a Negative Binomial random variable is $\frac{pr}{1-p}$ and $\frac{pr}{(1-p)^2}$ respectively. And setting the

$$\frac{pr}{1-p} = 2 \Rightarrow \frac{2}{1-p} = 3 \Rightarrow p = \frac{1}{3}$$

And we substitute this into the equation for the mean to get

$$2 = \frac{\frac{1}{3}r}{1 - \frac{1}{3}} \Rightarrow \frac{r}{2} = 2 \Rightarrow r = 4$$

A plot of all the distributions together

A plot of all the distributions together can be found in figure 3.

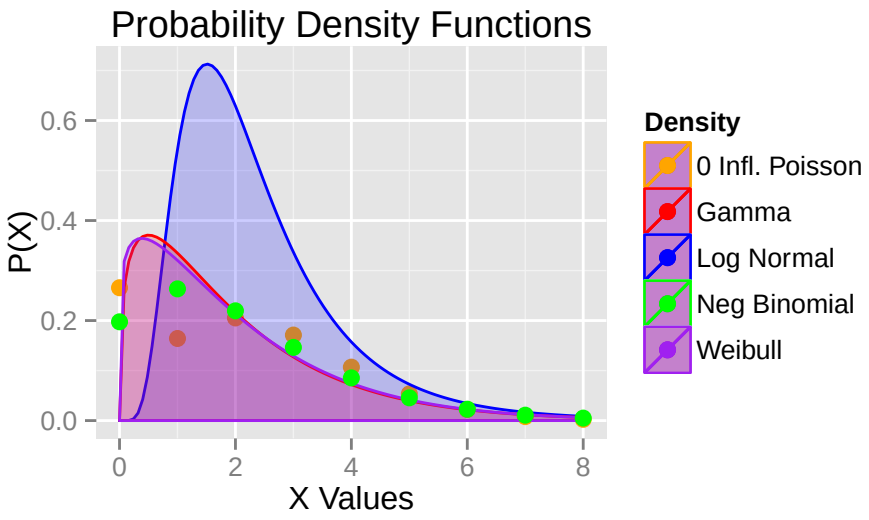

Figure 3: A Plot of each of the five distributions mentioned in this report

99% Upper Quantiles for each distribution

The 99% upper quantiles for each distribution are as follows:

- Gamma = 8
- Weibull = 7.87
- Negative Binomial = 7
- Log Normal = 5.56
- Zero Inflated Poisson = 8

Simulations drawn from each distribution

	Gamma	Weibull	Neg Binomial	Log Normal	0 Infl. Poisson
Sample Mean	1.83	2.57	2.20	1.92	2.00
Sample Variance	3.74	4.07	3.22	1.42	2.63

Table 1: Sample means of 20 random draws from each distribution

We can see that the means and variances are roughly where we would expect them to be.

Are Fertile Women Dangerous to Men?

Maybe, according to a dataset from the Faraday package. Fruitflies were forced to cohabilitate with either one or many women, either fertile or pregnant (unwilling to mate). The lifetime (in days) of 125 fruitflies were measured controlling for thorax length (which is known to be correlated with lifespan). The mean lifespan for each group is listed in the following table:

	Longevity (Days)
Isolated	64
With 1 Pregnant Fly	65
With 1 Virgin Fly	57
With 8 Pregnant Flies	65
With 8 Virgin Flies	39

Table 2: Marginal means of each fruit fly group

After we fit a Gamma Regression model to properly control for the effect of thorax size, we found that flies cohabitating with one virgin fly lived 11% shorter than flies in isolaton, and that flies cohabilitating with 8 virgin flies had their lifetimes reduced by a third! This can be inferred from the exponentiated parameter estimates given in the following table:

	Exp. Estimate	Std. Error	t value	P-Value
Intercept	6.62	0.19	9.73	0.00
Thorax Length	14.73	0.23	11.80	0.00
With 1 Pregnant Fly	1.06	0.05	1.04	0.30
With 1 Virgin Fly	0.89	0.05	-2.18	0.03
With 8 Pregnant Flies	1.08	0.05	1.52	0.13
With 8 Virgin Flies	0.66	0.05	-7.69	0.00

Table 3: Estimated parameters from the Gamma GLM model of the fruitflies

Finally, if you are curious about how good a fit the model was to the data,we present the empirical distribution along with the model fit in figure 4.

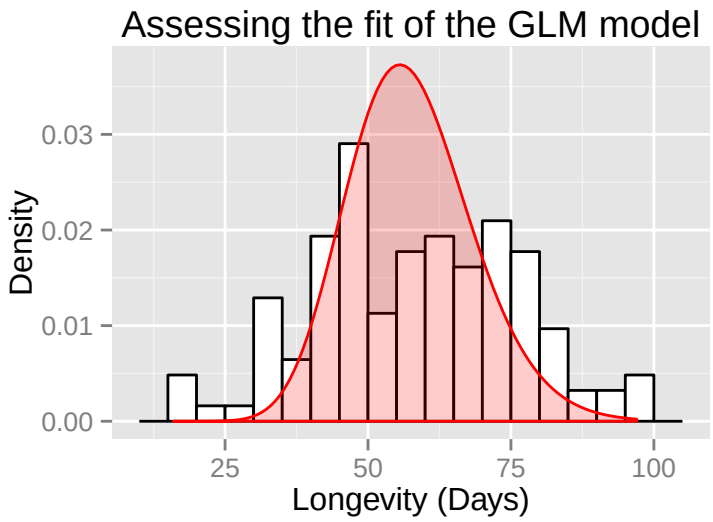

Figure 4: Assessing if the Gamma GLM (shown in red) was a good fit to the fly data

The American National Youth Tobacco Survey and the Two Statistical Cultures

In 2001 Leo Breiman wrote a paper about what he saw as the two prevalent cultures growing in the statistics discipline. Amazingly, this commentary has held true over a decade later. We use our recent analysis of the 2014 American National Youth Tobacco Survey to demonstrate the truth of Breimans commentary.

Our analysis was twofold in its aims– mainly to compare certain demographic effects on smoking habits and secondly to quantify these effects. Specifically, our research hypothesis was to investigate the effect of race on the habit of regularly chewing tobacco and the effect of sex of youth trying hookah.

For the primary question we used two logistic regression models; one predicting the odds of chewing tobacco regularly and the other predicting the odds of trying a Hookah at least once. Model interpretability (simplicity) was crucial here, since the models were built to answer specific research questions. Breiman rallies against simplicity in favor of complex models with greater predictive ability. His opinion on this matter is no doubt based on his experience as a consultant. These jobs did not seem to require him to answer research questions concerning understanding undlerlying mechanisms of the response variable, only having to predict it. The goals in our study are quite different, as we wish to know *what* causes youth smoking habits.

In Breimans paper he argues against fitting data to a model, as the assumptions of that model are taken as true and any violations of model assumptions can have serious repurcussions. We note that the logistic regression model we used has far less assumptions then, say, an OLS would have. It does not assume linearity between dependant and independant variables (only linearity between log odds and covariates). There is no normality or homoscedasticity assumptions for the residuals. Despite this, Breiman would have a point to argue that there is no good way to know whether the linearity assumption holds (he rejects goodness-of-fit tests).

If we had used an algorithmic model like a neural net, random forest or a SVM instead we could avoid this linearity assumption, but at the sacrifice of model interpretability. The algorithmic model may be able to be more powerful a predictor, but would not help identify the specific causes of smoking use among youth. This would defeat the purpose of what we were trying to do, since predicting smoking rates in youth was not the goal: analyzing its causality was. Despite this, breiman would argue that We could have teased some causailty from a smartly prunned random forest or related algorithmic method to determine relationships between the variables. He himself demonstrates that it can be done in part 11 of his paper. We argue that any algorithmic model simple enough to be interpretable should not be any more powerful than a much simpler parametric model (like logistic regression), which has a much more natural and widely accepted interpretation (the exponentiated coefficients represent changes in odds).

Breiman mentions that utilizing cross validation and incorperating averages of different models with perturbed training sets can capture more aspects of the data. We again acknowlege that this may have improved our model, provided we were looking to maximize the models predicibility. But we were not, and so his suggestions for improving our models seem needlessly complicated and not necessary to answer the research question we explored.

Section 2: Report

Summary

We analyzed the results of the 2014 American National Youth Tobacco Survey to look for indicators that correlated with inceases in the odds of chewing tobacco regularly or trying a hookah. We found that white people were the most likely to chew tobacco followed by hispanics and black people, who chewed tobacco regularly at half the rate and 1/5th of the rate respectively. Perhaps not surprisingly, older males living in rural areas had the largest odds of chewing tobacco regularly.

Unlike chewing tobacco habits, When it came to the odds of people using a Hookah at least once, the difference between men and women were not statistically different. Like before, older people were more likely to try using a Hookah, and black people were less likely. Unlike before, hispanics and urban dwellers were actually significantly more likely to try Hookah, the reverse of what we saw with trends in chewing tobacco.

Introduction

We analyzed the 2014 American National Youth Tobacco Survey using an R version of the dataset available at pbrown.ca. The original dataset was released by the Center for Disease Control. The data was collected from a survey administered to 258 Schools across the United States. We wanted to explore the relationship between demographics and smoking habits. Mainly we explored whether the odds of Regular use of chewing tobacco, snuff or dip increased with race. Note that the survey defined chewing tobacco regularly to be at least once in the last month. We also explored whether the probability of having used a hookah or waterpipe at least once was affected by gender. For both of these analysis we controlled for age and whether the respondant was from a rural area or an urban one. Additionally, we quantified how the use of chewing tobacco changes with age, sex, and ethnic group.

Methods

For our analysis we used a model that included the three aformentioned races, Asians, Natives and Pacific islanders. Our primary research question did not include these races, but we found our model did not change significantly upon the removal of these races from the data set, so we kept them in for greater generalizability. The model with the restricted dataset in included in the code for this document for the readers interest.

Being that we seeked to model probabilities, it was a natural choice to use the logistic regression model. We considered the following model for each of the aforementioned analysis:

$$\ln Odds = \beta_0 + \beta_1 x_{Age} + \beta_2 I_{Female} + \beta_3 I_{Black} + \beta_4 I_{Hisp} + \beta_5 I_{Asian} + \beta_6 I_{Native} + \beta_7 I_{Pacif} + \beta_8 I_{Rural}$$

Where *Odds* was either the Odds of regularly chewing Tobacco or the odds of ever using a Hookah,depending on context. Specifically, we tested whether race was a significant predictor of chewing tobacco: $H_0 \colon \beta_3 = \beta_4 = \beta_5 = \beta_6 = \beta_7 = 0$ and we tested whether gender was a significant predictor of using a hookah: $H_0 \colon \beta_2 = 0$. Finally, we compare the values of the β_i coefficients from each model to see what has the largest effect in predicting chewing tobacco and hookah use respectively.

Results

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Age	1	8.60	8.60	257.84	0
Sex	1	11.99	11.99	359.53	0
Race	5	7.02	1.40	42.12	0
Rural	1	4.68	4.68	140.37	0
Residuals	20422	680.91	0.03	NA	NA

Table 4: ANOVA summary table for modelling odds of regular use of chewing tobacco

From the ANOVA table from our first model, we can see that race was a significant predictor of chewing tobacco use, even after controlling for age, sex and whether participant was from a rural location. We also found that the rates of chewing tobacco were significantly different between ages, the sexes and between youth living in rural areas versus urban ones.

	Exp. Estimate	Std. Error	z value	P-Value
Intercept	0.00	0.32	-22.87	0.00
Age	1.34	0.02	14.54	0.00
Female	0.18	0.10	-16.44	0.00
Black	0.22	0.17	-8.98	0.00
Hispanic	0.53	0.10	-6.33	0.00
Asian	0.24	0.32	-4.40	0.00
Native	1.05	0.28	0.17	0.87
Pacific	2.69	0.36	2.76	0.01
Rural	2.56	0.09	11.01	0.00

Table 5: Modeling odds of regular use of Chewing tobacco

After looking at the exponentiated coefficients of our model (which represents the odds ratio increase/decrease between groups) we see that black people and hispanics are about 20 percent and half as likely to chew tobacco as whites respectively. This is after controlling for all the aforementioned covariates. Additionally, we can see that women where only 20 percent as likely to chew tobacco as men, and that the chances someone regularly chews tobacco increases 34 percent for each year of life. Not surprisingly, we see that people from rural areas are over 2.5 times more likely to chew tobacco, when compared to their urban dwelling counterparts.

	Exp. Estimate	Std. Error	z value	P-Value
Intercept	0.00	0.18	-42.71	0.00
Age	1.51	0.01	35.75	0.00
Female	1.04	0.04	0.92	0.36
Black	0.53	0.07	-8.95	0.00
Hispanic	1.42	0.05	7.28	0.00
Asian	0.53	0.12	-5.35	0.00
Native	1.20	0.19	0.96	0.34
Pacific	2.61	0.27	3.55	0.00
Rural	0.68	0.04	-8.81	0.00

Table 6: Modeling odds of ever using a hookah

We found similar trends in age and trying a Hookah as with age and chewing tobacco habits. Older people were about likely 30 percent more likely to try a Hookah for each year of life. Black people were half as likely and hispanics about 40 percent more likely than whites to trying Hookah. Also, the habit of trying Hookah among rural dwellers was the reverse the the trend of rural dwellers chewing tobacco, with 64% as many rural respondants trying Hookah as Urbanites. This may be because cities tend to be more multicultural than rural regions, and the Hookah is an import of the middle east.

Finally, we see that the odds of using a hookah are about 4 percent higher for women then men, but this difference is not statistically significant ($p = 0.32$) and so we cannot conclude that, controlling for age, race and geographic location, women and men are no more likely to use a hookah.

Appendix

This file was made using the R markdown package. All code used in this paper can be accessed from within the code blocks of the markdown document.