Assignment 3.   STAC51
Ankit 1003697065

1) $f(y, k, u) = \binom{y+k-1}{y}\left(1-\frac{k}{u+k}\right)^{y}\left(\frac{k}{u+k}\right)^{k}$

$y = 0, 1, 2 —$

⇒ Natural Parameter :-

The natural exponential families are a subset of the exponential families.
A Natural Exponential Family is an exponential family in which the natural parameter η and the natural statistic $T(x)$ are both the identity. A distribution in an exponential family with parameter θ can be written with PDF :-

~~$f_x(x|\theta) = h(x) \exp$~~

$$f(x|\theta) = h(x)\exp\left(\eta(\theta)T(x) - A(\theta)\right)$$

- $h(x)$ and $A(\theta)$ are known parameters or known functions.

- Given the Negative Binomial where k is the number of successes.
- $y+k$ is the total no. of trials for k successes.

- $p = \frac{k}{u+k}$

log
$$\Rightarrow \ln(f) = \ln\binom{y+k-1}{y} + y\ln\left(1-\frac{k}{u+k}\right) + k\ln\left(\frac{k}{u+k}\right)$$

$$\Rightarrow f = \binom{y+k-1}{y} \exp^{y\ln\left(1-\frac{k}{u+k}\right) + \ln\left(\frac{k}{u+k}\right)}$$

Negative binomial distribution is comparing with standard form,

$= \eta(\theta) = \ln\left(1 - \frac{k}{\mu+k}\right) = \ln\left(\frac{\mu}{\mu+k}\right)$

$T(y) = y$

$\rightarrow A(\theta) = k \ln\left(\frac{k}{\mu+k}\right) = k\ln(1-e^{(\eta)\theta})$

$\Rightarrow \therefore$ The neutral parameter is

$\eta(\theta) = \ln\left(\frac{\mu}{\mu+k}\right)$

where, $\theta = \frac{\mu}{\mu+k}$

2)

b) Using poisson regression we have:-

$\ln(\mu_A) = \alpha + \beta(0) = \alpha$

$\ln(\mu_B) = \alpha + \beta(1) = \alpha + \beta$

$\therefore$ we have $\ln(\mu_B) - \ln(\mu_A) = \beta$

$\beta = \ln\left(\frac{\mu_B}{\mu_A}\right)$

$\mu_B = \mu_A \times e^{\beta} = \ln \alpha\, e^{\beta}$

- Mean no. of ~~treatments~~ imperfections for treatment B is e^{β} times the mean no. of treatments for A.

2) c) $\frac{\mu_B}{\mu_A} = e^{\beta}$

→ Test $H_0: \mu_A = \mu_B$ against $\mu_B \neq \mu_A$

$H_0: e^{\beta} = 1$ or $\beta = 0$

$H_a: e^{\beta} \neq 1$ or $\beta \neq 0$

From the R output $\beta = 0.5878$ &

$\sigma_{\beta} = 0.1767$

$$Z = \frac{\beta}{\sigma_\beta} = \frac{0.5878}{0.1767} = 3.3322$$

$$P(Z > 3.3322) = 0.00\,043\,08116$$

Assuming a 95% CI with ~~α~~ 0.05 or 5%, level of signifigance we observe that the pvalue of the test statistic is very low. Hence, we reject the Null hypothesis that $\beta = 0$, which simultaneously means that there is significant evidence to think that $\mu_B \neq \mu_A$.

d) using part (c) we construct a 95% Wald CI for β.

$$(0.5878 - Z_{2.5\%}\,\sigma_\beta,\ 0.5878 + Z_{97.5\%}\,\sigma_\beta)$$

$$= (0.5878 - 1.96 \times 0.1764,\ 0.5878 + 1.96 \times 0.1764)$$

$$CI = (0.242056,\ 0.933544)$$

since $\frac{\mu_B}{\mu_A} = e^{\beta}$, 95% CI ⇒

$$CI = (e^{0.242056},\ e^{0.933544})$$

$$= (1.273866,\ 2.5435071)$$

5)

a) Wald test

$$W = \left(\frac{\tilde{\beta} - \beta_0}{\sigma_\beta} \right)^2 \sim \chi^2(1)$$

Hence, $\sqrt{W} = \frac{\tilde{\beta}_0 - \beta_0}{\sigma_\beta} \sim N(0,1)$

The p-value of the test $\beta = 0$ against alternate hypothesis. $\beta \neq 0$ is shown against each variable.

to test significance of Race, we look at the significance of variable "Race whites"

$\therefore$ parameter coeff = 0.79129 with pvalue 0.005.

thus shows that it is significant in predicting the dependent variable.

For GLM2 → the test compares the likelihood of the data under two models. If p-value is small, it means that the model with more ~~pred~~ predictors ~~variables~~ glm1 (District + Race) is significant compared to the model with the fewer predictors (only District) → glm2.

The simpler model (GLm1). Null hypothesis is that the coeff. of the predictors variable Race, in model 2 is not significant & model 1 is good enough as model 2.

→ from the Analysis of Variance (ANOVA) of both glm 1 & 2.

test statistic :- 7.39 with 1 df.
p-value of the model is 0.006553 < 5%.

∴ the coeff of Race is significant not zero.
So we conclude that the model with District & Race is significant compared to the simpler model with only one district.
~~the conclusion~~

b)

~~the odds~~

$\ln(\text{odds}(\text{White}) \mid \text{District}) = \alpha + \beta_W + \beta_D$

$\ln(\text{odds}(\text{Black}) \mid \text{District}) = \alpha + \beta_D$

$$\ln\left(\frac{\text{odds}(\text{White}) \mid \text{District}}{\text{odds}(\text{Black}) \mid \text{District}}\right) = \alpha + \beta_W + \beta_D - (\alpha + \beta_D)$$

$$= \beta_W$$

$$\ln(OR) = \beta_W = 0.79129$$

Where $OR = \dfrac{\text{Odds}(\text{White}) \mid \text{District}}{\text{Odds}(\text{Black}) \mid \text{District}}$ is OR of white

$$OR = e^{\beta_{white}} = e^{0.79129} = 2.206$$

Wald CI

CI for β_w = $(0.79129 - 1.96(0.28532), 0.79129 + 1.96(0.28532))$

$CI = (0.2320628, 1.350517)$

CI for OR b/w Race & Merit Pay

$= (e^{0.2320628}, e^{1.350517})$

$= (1.261199, 3.85942)$

→ If the study was repeated and OR was calculated each time, you would expect the true value to lie within the ranges on 95% of occasions. The odds of success for white is at least 1.26 times than for blacks with 95% confidence.

→ CI for OR b/w Race & Merit Pay

(1.251974, 3.844984)

) From the ANOVA result, 2.07 ~~df~~ test statistic and 4 df, p-value of 0.7227. High p-value > 0.05 which means that coeff. of interaction variables b/w that of Race & District are not significantly diff. from zero.

∴ There is no significant interaction & the association b/w Race & District is uniform.