Workers’ Compensation

Molly Hischke & Kyle Hancock

3/13/2020

# Introduction (4 pts - Kyle)

* Background
* Identify research question(s)
* State response and predictor variables
* Identify each variable as continuous or categorical (with levels)
* Identify observational/experimental units and number of observations
* Further details of design

# Summary statistics and/or graphics (4 pts - Kyle)

library(readr)  
library(tidyr)  
library(dplyr)  
library(ggplot2)  
library(car)  
library(emmeans)  
library(knitr)  
  
## Reading in raw data  
wc <- read\_csv("../data/craftbeer\_wc.csv")  
  
## Adding column for number of years a brewery had a policy with insurance comp  
policy\_years <- wc %>%   
 group\_by(Policy\_Dim\_Key, year) %>%   
 summarize(total\_claims = n()) %>%   
 ungroup() %>%   
 group\_by(Policy\_Dim\_Key) %>%   
 summarize(years\_with\_policy = n()) %>%   
 ungroup()  
  
## CLEANED DATA FRAME  
wc2 <- wc %>%   
 group\_by(Policy\_Dim\_Key, year) %>%   
 summarize(avg\_payroll = mean(adj\_payroll),  
 total\_claims = sum(Claim\_Count)) %>%   
 mutate(claim\_filed = ifelse(total\_claims == 0, 0, 1)) %>%   
 ungroup() %>%   
 full\_join(policy\_years) %>%   
 mutate(Policy\_Dim\_Key = as.factor(Policy\_Dim\_Key),  
 year = as.factor(year)) %>%   
 rename(brewery = Policy\_Dim\_Key)  
  
  
data\_table <- wc2 %>%   
 rename(Brewery = brewery,  
 Year = year,  
 'Average Payroll' = avg\_payroll,  
 'Total Claims' = total\_claims,  
 'Years with Policy' = years\_with\_policy)  
  
kable(sample\_n(data\_table, 10, replace = TRUE))

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Brewery | Year | Average Payroll | Total Claims | claim\_filed | Years with Policy |
| 3561828 | 2017 | 287114.45 | 1 | 1 | 6 |
| 3613425 | 2018 | 97220.06 | 0 | 0 | 3 |
| 3611858 | 2018 | 527086.91 | 0 | 0 | 5 |
| 3466166 | 2018 | 43343.96 | 0 | 0 | 3 |
| 3561828 | 2016 | 235325.99 | 1 | 1 | 6 |
| 3599745 | 2017 | 384997.65 | 0 | 0 | 5 |
| 3591062 | 2018 | 92284.21 | 0 | 0 | 5 |
| 3537607 | 2014 | 16305.34 | 0 | 0 | 6 |
| 3606524 | 2018 | 99870.94 | 0 | 0 | 5 |
| 3506826 | 2016 | 190532.93 | 0 | 0 | 3 |

# Analysis (6 pts - Molly)

## Binary Response with Repeated Measures (Model1)

Model1 analysis was done using R and the geepack (Hojsgaard, Halekoh, & Yan 2006; Yan & Fine, 2004; Yan, 2002). Generalized Estimating Equations (GEE) was used to fit a model with the response being if a brewery filed a claim (1 = yes, 0 = no). Fixed effects include payroll (proxy for brewery size). Breweries were used as clusters and we assumed a exchangeable correlation structure.

library(geepack)

## Warning: package 'geepack' was built under R version 3.6.3

Model1 <- geeglm(claim\_filed ~ avg\_payroll, id = brewery,  
 family = binomial(link = "logit"),   
 corstr = "exchangeable", data = wc2)

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

summary(Model1)

##   
## Call:  
## geeglm(formula = claim\_filed ~ avg\_payroll, family = binomial(link = "logit"),   
## data = wc2, id = brewery, corstr = "exchangeable")  
##   
## Coefficients:  
## Estimate Std.err Wald Pr(>|W|)   
## (Intercept) -2.370e+00 4.102e-01 33.372 7.61e-09 \*\*\*  
## avg\_payroll 3.210e-06 1.555e-06 4.263 0.039 \*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Correlation structure = exchangeable   
## Estimated Scale Parameters:  
##   
## Estimate Std.err  
## (Intercept) 249715 2.718e+10  
## Link = identity   
##   
## Estimated Correlation Parameters:  
## Estimate Std.err  
## alpha 2.09e-07 0.01438  
## Number of clusters: 130 Maximum cluster size: 6

citation("geepack")

##   
## To cite geepack in publications use:  
##   
## Højsgaard, S., Halekoh, U. & Yan J. (2006) The R Package geepack  
## for Generalized Estimating Equations Journal of Statistical  
## Software, 15, 2, pp1--11  
##   
## Yan, J. & Fine, J.P. (2004) Estimating Equations for Association  
## Structures Statistics in Medicine, 23, pp859--880.  
##   
## Yan, J (2002) geepack: Yet Another Package for Generalized  
## Estimating Equations R-News, 2/3, pp12-14.  
##   
## To see these entries in BibTeX format, use 'print(<citation>,  
## bibtex=TRUE)', 'toBibtex(.)', or set  
## 'options(citation.bibtex.max=999)'.

## Count Response with Repeated Measures

Model1 analysis was done using R and the geepack (Hojsgaard, Halekoh, & Yan 2006; Yan & Fine, 2004; Yan, 2002). Generalized Estimating Equations (GEE) was used to fit a model with the response being the number of times a brewery filed a claim within a year. Fixed effects include payroll (proxy for brewery size). The number of years the brewery held a policy with the insurance company was included as a covariate. Breweries were used as clusters and we assumed a exchangeable correlation structure.

library(geepack)  
Model2 <- geeglm(total\_claims ~ avg\_payroll\*year + years\_with\_policy, id = brewery,  
 family = poisson(link = "log"),   
 corstr = "exchangeable", data = wc2)  
summary(Model2)

##   
## Call:  
## geeglm(formula = total\_claims ~ avg\_payroll \* year + years\_with\_policy,   
## family = poisson(link = "log"), data = wc2, id = brewery,   
## corstr = "exchangeable")  
##   
## Coefficients:  
## Estimate Std.err Wald Pr(>|W|)   
## (Intercept) -6.41e+00 1.35e+00 22.67 1.9e-06 \*\*\*  
## avg\_payroll 3.63e-08 4.29e-07 0.01 0.9326   
## year2014 3.00e-01 1.19e-01 6.38 0.0115 \*   
## year2015 3.93e-01 1.85e-01 4.52 0.0335 \*   
## year2016 4.77e-01 1.48e-01 10.30 0.0013 \*\*   
## year2017 4.38e-01 1.41e-01 9.67 0.0019 \*\*   
## year2018 2.51e-02 2.08e-01 0.01 0.9038   
## years\_with\_policy 1.10e+00 2.58e-01 18.28 1.9e-05 \*\*\*  
## avg\_payroll:year2014 3.56e-07 3.58e-07 0.99 0.3194   
## avg\_payroll:year2015 1.67e-07 5.07e-07 0.11 0.7419   
## avg\_payroll:year2016 1.17e-07 4.36e-07 0.07 0.7889   
## avg\_payroll:year2017 1.73e-07 4.32e-07 0.16 0.6886   
## avg\_payroll:year2018 1.55e-07 4.35e-07 0.13 0.7206   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Correlation structure = exchangeable   
## Estimated Scale Parameters:  
##   
## Estimate Std.err  
## (Intercept) 4.07 2.86  
## Link = identity   
##   
## Estimated Correlation Parameters:  
## Estimate Std.err  
## alpha 0.804 0.619  
## Number of clusters: 130 Maximum cluster size: 6

# Results and Conclusions (4 pts - Molly)

* ANOVA style likelihood ratio tests and/or table of estimated coefficients
* Other results as appropriate (ex: pairwise comparisons for categorical predictor)
* Interpretation and discussion (even if nothing is “significant”)
* Respond to your research questions

# R code appendix

Add the code chunk from the HW assignments.