QBS 121: Assignment on GLM and Modeling Counts

February 2, 2022

Do the following questions: 1.1, 1.2, 2.1, 3.1, 3.2

1 Data Analyses

1.1 Modelling Student Absences

Analyze the dataset *quine* which comes with the R library *MASS*. The dependent variable is number of student absences.

1. Put together a table of univariable (1 covariate at at time) results on how each of the covariates relate to student absences.
2. Put together a table of multivariable results, i.e., run a multivariable model using all of the variables (or a subset if you choose).
3. Do this in two ways, (i) using Poisson regression in conjunction with sandwich variance to determine standard errors (or by selecting family=quasipoisson in the glm function) and (ii) negative binomial regression. Comment on the difference or similarity between the two sets of results.

1.2 Cancer Counts in Danish Cities

Access the data *eba1977* in the R library ISwR. This is a small dataset on cancer counts by city and age group in Denmark.

1. Which variable makes sense to use as an offset.
2. a. Use Poisson regression to model the association with age group. b. Test the significance of age using anova(o.glm, test=”Chisq”) c. Test the association of age as an ordinal variable (hint: create ageOrdinal = as.numeric(age)).
3. a. Use Poisson regression to model the association with city. b. Test the significance of city.
4. Run the multivariable model with city and age.
5. For interest, instead of using an offset include log(population) as a covariate. Is the coefficient significantly different from 1.0 ?

1.3 Modelling Incidence of Hypoglycemia in Children with Type 1 Diabetes

The event of interest (dependent variable) is hypoglycemia, an episode in which an individual needs assistance because their glucose level is too low usually because they had taken too much insulin. The dependent variable comes in two piecies of information, (i) the number of hypoglycemic events and (ii) the follow-up time.

1. Estimate the incidence of hypoglycemia in this cohort as the total number of hypoglycemia divided by total follow-up time.
2. Run a Poisson model for the incidence of hypoglycemia (account for variable follow-up time using an offset or adding log of follow-up time as a covariate) without covariates.
3. Run univariable analyses for each of the covariates. Account for variable follow-up time using an offset or adding log of follow-up time as a covariate. Account for over (or under

dispersion) by employing a sandwich correction or specifying family=quassipoisson within glm.

4. Repeat this using negative binomial (gamma-Poisson) regression.
5. Develop a multivariable model using either LASSO, forwards or backwards stepwise regression.
6. Using the results of the multivariable model comment on the effect of insurance status on hypoglycemic incidence.

2 Simulate and Analyze

2.1 Large Counts: Linear Regression vs Poisson

If the dependent variable is a count that takes large values (e.g. counts that are zero with very low frequency) it may be preferable to use linear regression.

1. Choose a sample size, e.g. n=500
2. Generate a couple continuous variables, Z1=rnorm(n) and Z2=rnorm(n)
3. Generate a large count Y=rpois(n, lambda=100*1.5**Z1/1.2**Z2)
4. Plot this count vs Z1, and then versus Z2
5. Use multivariable Poisson regression to model Y vs Z1 and Z2.
6. Use multivariable linear regression to model Y vz Z1 and Z2
7. Use multivariable linear regression to model log(Y) vz Z1 and Z2
8. Assess similarities and differences of the estimates, standard errors and Z-values from these three models

2.2 Binary Endpoint: Logistic vs Poisson

Sometime Poisson regression is used when the dependent variable is binary in order to get a coefficient which can be interpreted as a log risk ratio, as opposed to logistic regression which yields a log odds ratio.

1. Choose a sample size, say $n = 1000$.
2. Generate a treatment indicator X to be 0/1 with 50-50 probability, X=runif(n) ≤ 0.5
3. Let Z be a covariate that is standard normal, e.g. Z=rnorm(n).
4. Generate a binary endpoint that depends on X and Z according to the log-linear model as follows, Y = runif(n) < exp(-1.0 + 0.5*X - 0.1*Z)
5. Use the following three approaches for modelling the dependence of Y on X and Z: (i) logistic regression, (ii) Poisson regression without robust standard errors, (iii) Poisson regression corrected for over or under dispersion.
6. Discuss any substantive differences in the findings.

3 Simulation

3.1 AUROC as Measure of Difference of Two Distributions

The AUROC of a score that predicts an event equals the probability that a subject with the event will have a higher score than a person without the event. If the distribution of the scores in subjects with the event is normal with mean m1 and s1 and the distribution of scores in subjects without the event is normal with mean m0 and s0, then the following R line of code estimates the concordancy.

```
mean( rnorm(n=n<−10ˆ6, mean=m0, sd=s0) < rnorm(n=n, mean=m1, sd=s1) )
```

a. Create a table of Concordancy vs the following choices, m0=0,sd=1, m1 = 0.0, 0.25,0.5,0.75,1,1.5,2,3 and s1=1.

b. Suppose a score for the risk of an event is such that its distribution in those who will have the event is normal with mean m1 and standard deviation of s1, and its distribution in those who will not have the event is mean 0 and standard deviation 1. Simulate the score of 1000 events (cases) and 1000 controls and plot the corresponding ROC curve for the following 4 scenarios (m1=0.5, s1=1), (m1=0.5, s1=2), (m1=2.0, s1=1), (m1=2.0, s1=2).

To do this the following code could be helpful:

```
  Score1 <- rnorm(n=n, mean=m1,sd=s1)
  Score0 <- rnorm(n=n, mean=0,sd=1)
  Event <- rep(c(1,0), each=n)
  Score <- c(Score1, Score0)
  o <- roc(Event, Score)
  plot(o$specificities, o$sensitivities, type=”line”,
```

3.2 ADR vs APC

The most recommended modality for colorectal cancer screening in the USA is colonoscopy. During a colonoscopy a clinician uses a camera at the end of a tube (colonoscope) to examine the colon. The colonoscope is also equipped with features to remove pre-cancerous lesions (polyps, adenomas). Colonoscopists vary in their ability to detect polyps. One measure of detection ability is the Adenoma Detection Rate (ADR). It is defined as the proportion of colonoscopies in which at least one adenoma is detected; like the proportion of games in which an athlete gets at least one point. An alternative metric is the APC (adenomas per colonoscopies); like the average number of points per game. Explain what the following simulation is doing and interpret the results.

```
R <- 1000
cor.ADR.true <- cor.APC.true <- R
n.endoscopists <- 200 # number of endoscopists in the cohort
for (r in 1:R) {
  # number of patients each endoscopists scopes in a year
  n.pt.endoscopist <- ceiling(rgamma(n=n.endoscopists, shape=10, scale=30))
  N <- sum(n.pt.endoscopist)
  ID.Endo <- rep(1:n.endoscopists, times=n.pt.endoscopist)
  true.endo.rate <- runif(n.endoscopists, min=0.35,max=0.99) # given a unifor
  long.true.endo.rate <- rep(true.endo.rate, times=n.pt.endoscopist)
  n.polyps <- rpois(n=N, lambda=0.6) # lambda is the average actual adenomas
  n.polyps.detected <- rbinom(n=N, size=n.polyps, prob=long.true.endo.rate)
  at.least.one <- n.polyps.detected>0
  ADR <- tapply(at.least.one, ID.Endo, mean)
  APC <- tapply(n.polyps.detected, ID.Endo, mean)
  cor.ADR.true[r] <- cor(ADR, true.endo.rate)
  cor.APC.true[r] <- cor(APC, true.endo.rate)
}
pairs(cbind(true.endo.rate, ADR, APC))
summary(cbind(cor.ADR.true, cor.APC.true))
```

3.3 Ordinal Probit Regression

Comment on what the following code is doing and illustrating. How you might interpret the coefficients in terms of a continuous latent variable?

```
n <- 5000
Age <- 80*runif(n)
Sex <- ifelse(runif(n) < 0.5, "M", "F")
DM <- ifelse(runif(n) < 0.1, "Diabetes", "no Diabetes")
CKD <- ifelse(runif(n) < 0.05, "KidneyDisease", "No CKD")
coef <- c(Age=0.05, Male = 0.20, DM = 0.40, CKD = 0.60)
Linear <- cbind(Age, Sex=="M", DM=="Diabetes", CKD == "KidneyDisease") %*% co
Latent.Continuous <- Linear + rnorm(n)
cut.offs <- quantile(Latent.Continuous, c(0.60,0.85))
Discharge <- factor(cut(Latent.Continuous, c(-Inf, cut.offs, Inf), labels=c("

table(Discharge)

head(data.frame(Discharge, Age, Sex, DM, CKD))

library(MASS)
summary(o <- polr(Discharge ~ Age + Sex + DM + CKD, method="probit"))
o$coef - coef
o$coef / coef
```