4.1 **Logistic regression:** we consider data on wage of professors in the United States over nine month periods. The `profsalary` dataset contains information about the participants.

- `sex`: binary, either man (0) or woman (1);
- `rank`: categorical, one of assistant (1), associate (2) or full professor (3);
- `degree`: highest degree, either masters (0) or doctorate (1);
- `yd`: number of years since last degree;
- `yr`: number of years in academic rank;
- `salary`: salary in USD over nine months;

(a) Fit a logistic regression to model the probability that a professor has a salary superior to 105K USD as a function of `degree`, `sex`, `yr` and `yd`. Write the equation for the mean and interpret the estimated coefficients of the model.

(b) If you add the covariate `rank`, what happens? Do you identify any problem with the model? If so, try to find an explanation.

4.2 An education researcher is interested in the association between the number of awards students receive at a high school as a function of their math scores and the type of school they attend. The `awards` data contains

- `awards`: response variable indicating the number of prize received throughout the year
- `math`: score of students on their math final exam
- `prog`: student program in which the student, one of `general` (1), `academic` (2) and `vocational` (3).

Fit a Poisson model and a negative binomial model with `math` and `prog` as covariates and interpret the parameters. Compare the results between the two and say which model is more appropriate, if any.

4.3 **Rate data** The `ceb` data contains information about the number of children ever born (CEB) from the Fiji Fertility Survey. The variable measured for each group of women are

- `nwom`: number of women in the group.
- `nceb`: response, number of children ever born.
- `dur`: time (in years) since wedding, either 0–4 (1), 5–9 (2), 10–14 (3), 15–19 (4), 20–24 (5) and greater than 25 (6).
- `res`: categorical variable for residence, one of Suva (1), urban (2) or rural (3).
- `educ`: ordinal variable giving the educational achievements, one of none (1), lower primary (2), upper primary (3), high school or higher (4).
- `var`: estimated within-group variance in number of children ever born per group.

(a) Plot (a) the number of children ever born (`nceb`) as a function of the number of women in the group (`nwom`) and (b) the mean number of children ever born per group against the variance of the number of children ever born. Comment on the two plots.

(b) Should an offset term be included? Explain.

- If no offset is used, which function (if any) of `nwom` should be included in the mean model?
- If one considers using an offset, how does it compare relative to the model with log(`nwom`)?

(c) Fit a Poisson regression model with an offset including the three categorical covariates `dur`, `res` and `educ` as main effects. Which of the three predictors is the most significant?

(d) Interpret the coefficients of the fitted model.

(e) Assess whether there is need for an interaction between `educ` and `dur` by performing a likelihood ratio test.

(f) It is possible to assess goodness-of-fit using diagnostic plots for the so-called deviance residuals from a Poisson generalized linear model.[1] Using software, produce diagnostic plots of (a) fitted values against deviance-based residuals, (b) quantile-quantile plot of deviance-based residuals, (c) leverage and (d) Cook distance against observation number. Comment on the adequacy of the model with all three categorical covariates and an offset. *To produce the plots in SAS , use the options*

[1] In general, however, these diagnostics are harder to interpret because the observations are discrete whereas the fitted mean is continuous.

Compiled 02/10/2020 at 12:34

```
proc genmod data=... plots=(resdev(xbeta) leverage cooksd)
```

The quantile-quantile plot can be produced with the procedure `univariate` *using the standardized deviance residuals* (`stdresdev`). *In R , use the function* `boot::glm.diag.plots` *to produce graphical diagnostics.*

4.4 **Understanding the drivers of BIXI rentals**: BIXI is a Montreal-based bicycle rental company. We examine the data for 500 days during the period 2017–2019 at the Edouard Montpetit bike docking station in front of HEC. Our interest is in explaining variability in daily bike rental (measured through the number of users) at that station based on time of the week and meteorological factors. The data consist of

- `nusers`: number of daily users at the station.
- `temp`: temperature (in degree Celcius)
- `relhumid`: percentage of relative humidity, taking values between 0 and 100.
- `weekday`: categorical variable for week day, between Sunday (1) and Saturday (7).
- `weekend`: binary variable taking value zero if rental is on a weekend (Saturday or Sunday) and one otherwise.

We consider four competing models for the data

- Model 8.4.1 is a Poisson regression model with `weekend` as covariate.
- Model 8.4.2 is a Poisson regression model with `weekend`, `relhumid` and `temp` as covariates.
- Model 8.4.3 is a negative binomial model with `weekend`, `relhumid` and `temp` as covariates.
- Model 8.4.4 is a negative binomial model with `weekday` (categorical), `relhumid` and `temp` as covariates.

(a) Is Model 8.4.1 an adequate simplification of Model 8.4.2? Assess this hypothesis formally.

(b) Interpret the coefficients for the intercept and for `relhumid` in Model 8.4.2.

(c) Compare the model fit of the negative binomial model (Model 8.4.3) with that of the Poisson (Model 8.4.2) using all of the following methods: (a) the deviance statistic (b) a likelihood ratio test and (c) information criteria.

(d) Suppose that, rather than including `weekend`, we instead consider `weekday` as covariate in the models. Explain how the model would differ if we included `weekday` as an integer-valued variable as opposed to declaring it categorical? Explain which of the two makes more sense in the present context.

(e) The equation for the mean of Model 8.4.4 is

$$\begin{aligned}\mathsf{E}(\texttt{nusers}) = \exp\big(&\beta_0 + \beta_1 \texttt{temp} + \beta_2 \texttt{relhumid} + \beta_3 \mathbf{1}_{\texttt{weekday=2}} + \beta_4 \mathbf{1}_{\texttt{weekday=3}} \\ &+ \beta_5 \mathbf{1}_{\texttt{weekday=4}} + \beta_6 \mathbf{1}_{\texttt{weekday=5}} + \beta_7 \mathbf{1}_{\texttt{weekday=6}} + \beta_8 \mathbf{1}_{\texttt{weekday=7}}\big).\end{aligned}$$

Write the null hypothesis for the test comparing Model 8.4.3 to Model 8.4.4 in terms of the model parameters $\boldsymbol{\beta}$, thereby showing that Model 8.4.3 is nested within Model 8.4.4. Does the number of user significantly vary between weekdays and between weekend days?

4.5 **Two-way contingency tables:** Counts data are often stored in two-way contingency tables, with two factors taking respectively J and K levels. The same format is used to store the numbers of successes/failures. The **saturated** mean model for cell j,k (interaction plus two main effects) is

$$\text{logit}(p_{jk}) = \alpha + \beta_j + \gamma_k + \nu_{jk}, \qquad j = 1,\ldots,J-1; k = 1,\ldots,K-1. \tag{M_s}$$

which has $JK = 1 + (J-1) + (K-1) + (J-1)(K-1)$ parameters. We can consider simpler models:

- M_0: the null model $\text{logit}(p_{jk}) = \alpha$ with 1 parameter
- M_1: the main effect model $\text{logit}(p_{jk}) = \alpha + \beta_j (j = 1,\ldots,J-1)$ with J parameters
- M_2: the main effect model $\text{logit}(p_{jk}) = \alpha + \gamma_k (k = 1,\ldots,K-1)$ with K parameters
- M_3: the model with both additive main effects $\text{logit}(p_{jk}) = \alpha + \beta_j + \gamma_k (j = 1,\ldots,J-1; k = 1,\ldots,K-1)$ with $J+K-1$ parameters.

The deviance measures the **discrepancy** in fit between the saturated and the fitted models. Under regularity con-

ditions and assuming the number of observations in each of the JK levels goes to ∞,

$$D(\widehat{\boldsymbol{\beta}}_{M_i}) = 2\{\ell(\widehat{\boldsymbol{\beta}}_{M_s}) - \ell(\widehat{\boldsymbol{\beta}}_{M_i})\} \dot{\sim} \chi^2_{JK-p_i}$$

under the null hypothesis that the simpler model M_i with p_i parameters is adequate. We proceed by backward selection and compare the difference in deviance between two nested models $M_i \subset M_j$; the difference $D(\widehat{\boldsymbol{\beta}}_{M_i}) - D(\widehat{\boldsymbol{\beta}}_{M_j})$ follows $\chi^2_{p_j-p_i}$ asymptotically if M_i is an adequate simplification of M_j (the comparison in deviance is another way to express the likelihood ratio statistic for nested models).

Once the selection is complete, we end up with model M_i, say. If model M_i is adequate, then $D(\widehat{\boldsymbol{\beta}}_{M_i}) \dot{\sim} \chi^2_{JK-p_i}$ and its expectation should be roughly $JK - p_i$.

Fit these models using a binomial likelihood with logit link to the `cancer` data set, which contains categorical two regressors (`age` and `malignant`) taking respectively 3 and 2 levels. Perform an analysis of deviance and select the best model using backward elimination, starting from the saturated model.

4.6 **Equivalence of Poisson and binomial models:** Suppose $Y_j \sim \mathcal{B}(m_j, p_j)$ and $m_j p_j \to \mu_j$ as $m_j \to \infty$, we can approximate the distribution of Y_j by that of a Poisson $\mathcal{P}(\mu_j)$. Therefore, we may consider a generalized linear model with

$$\log(\mu_j) = \log(m_j) + \log(p_j).$$

In this model, m_j is a fixed constant, so the coefficient for the predictor $\log(m_j)$ is set to one. Such a term is called offset.

(a) Fit the models $M_0, \ldots, M_3$ using a binomial likelihood with logistic link function to the `smoking` data set, which contains counts of lung cancers as a function of `age` category and `smoking` habits. Write down the deviance and the number of degrees of freedom for the model and perform an analysis of deviance using backward elimination. Repeat the analysis using instead a Poison model with log link and an offset term.

(b) Compare your results with those obtained using the logistic model in terms of fitted probability of death.

Compiled 02/10/2020 at 12:34