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.

Solution

The equation is

$$\frac{\pi_i}{1-\pi_i} = \exp\left(\beta_0 + \beta_1 \texttt{degree}_i + \beta_2 \texttt{sex}_i + \beta_3 \texttt{yr}_i + \beta_4 \texttt{yd}_i\right)$$

- expit(β_0) is the probability that a new assistant professor with a master degree and no experience will earn more than 105K USD. The estimated probability is 0.000286.
- $\widehat{\beta}_1 = 18.58$; the odds for a professor with a PhD degree are 18.58 times higher, everything else being constant.
- $\widehat{\beta}_2 = 0.30$; the odds for women are 0.3 times those of men, ceteris paribus.
- The two last coefficients cannot be interpreted separately, unless the person changes academic rank (in which case `yr` decreases from x to 1. In general, the odds for a given person staying at the same rank increase by $\exp(\widehat{\beta}_3 + \widehat{\beta}_4) = \exp(1.276 + 1.171) = 11.55$.

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

Solution

The information from `rank` and the number of years since diploma and in academic rank are partly redundant (collinear). The estimated intercept is $\widehat{\beta}_0 = -25.3$ in R /-11.1 in SAS , the coefficient for associated is $\widehat{\beta}_{\texttt{associate}} = 0.46$ in R /-5.72 in SAS and that of full rank is $\widehat{\beta}_{\texttt{full}} = 26.1$ in R /11.92 in SAS ; this means that the model predicts that everyone who is assistant or associate professor has a (essentially) zero probability of having a salary exceeding 105K USD. This is an example of quasi-complete separation of variables problem.

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.

Solution

The likelihood ratio test compares the negative binomial model to the Poisson model (corresponding to the test of $\mathscr{H}_0 : k = 0$ in the negative binomial variance formula) and the p-value is 0.096; this means we fail to reject the null $k = 0$; the estimated coefficient is $\widehat{k} = 0.1635 = 1/6.114$. The ratio of deviance to degrees of freedom for the Poisson model is 0.97, suggesting the simpler model is also adequate.

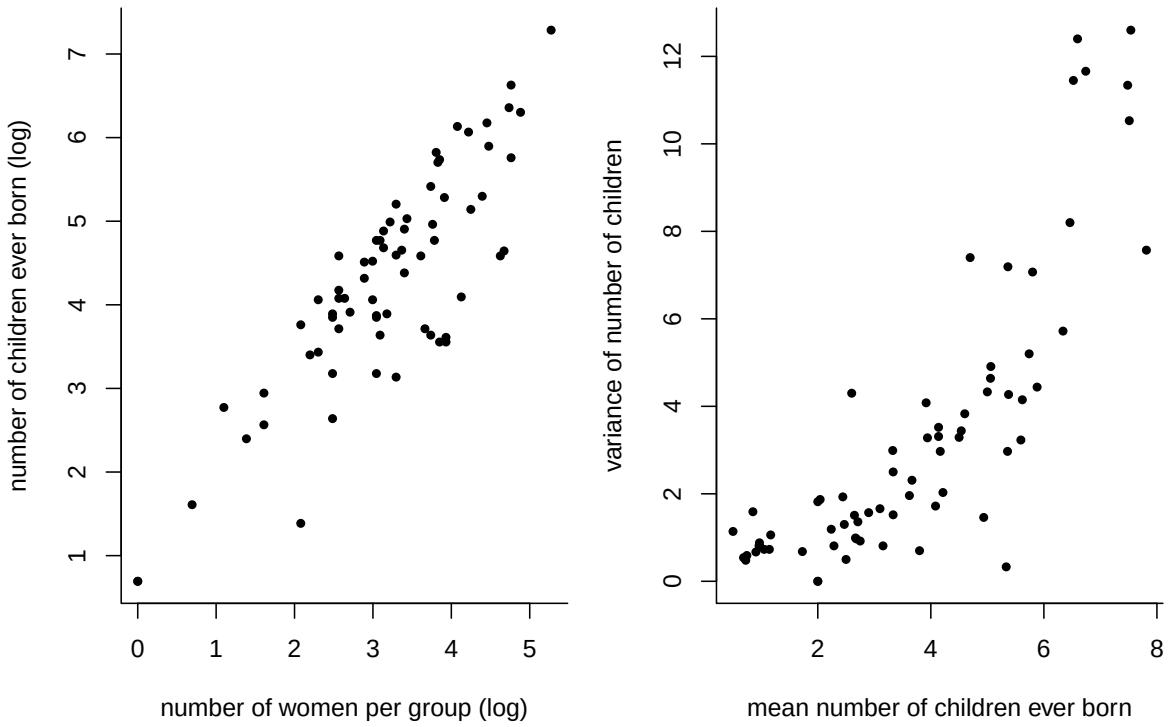

Figure 1: Number of children ever born as function of the number of women per group on log-log scale (left) and mean versus variance of the number of children ever born per group (right).

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.

Solution

There appears to be a clear linear relationship between the two variables on the log scale. The mean-variance relationship appears nonlinear, with somewhat higher variability for the largest counts (but could depend on covariates).

(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`)?

Solution

Counts are clearly not comparable, so an offset term is warranted. The proper term to include in the mean model is log(`nwom`). We can fit the model with the three categorical covariates and check if the coefficient associated to log(`nwom`) could be unity; the 95% profile likelihood confidence interval is $[0.97, 1.06]$, suggesting no evidence against inclusion of the number of women as offset term.

(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?

Solution

Duration of marriage is the most significant global predictor by far; it has the highest likelihood ratio statistic (meaning its p-value is smallest after accounting for the five degrees of freedom).

(d) Interpret the coefficients of the fitted model.

Solution

The parameter estimates and their interpretation depends on the baseline; the following corresponds to choosing the lowest level of each category as baseline (0–4 years since wedding, living in Suva island, no education).

- The estimated mean number of chidren ever born per woman in the baseline category is $\exp(\widehat{\beta}_0) = 0.89$.
- The estimated mean increase is 2.7 for 5–9 years (respectively 3.93 for 10–14, 5.02 for 15–19, 5.96 for 20–24, 7.2 for 25 and above) relative to the baseline of less than five years of marriage, *ceteris paribus*.
- The estimated increase relative to the baseline is 1.02, 0.90 and 0.73 for higher education levels; the higher the educational achievement, the lower the average number of children ever born.
- People in urban areas have 1.12 times more children and those in rural area 1.16 times more than in Suva, for the same number of years since wedding and same educational achievements.

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

Solution

Adding an interaction adds 15 parameters to the model. The likelihood ratio test statistic is 15.86, and the p-value for the null hypothesis that all interaction parameters are zero is 0.3912. We conclude that no evidence that the model with the interaction fits significantly better.

(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*

```
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.*

Solution

The diagnostics plots in Figure 2 look okay; one residual value is has a high value and leverage, but outside of that the model fit is excellent.

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

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

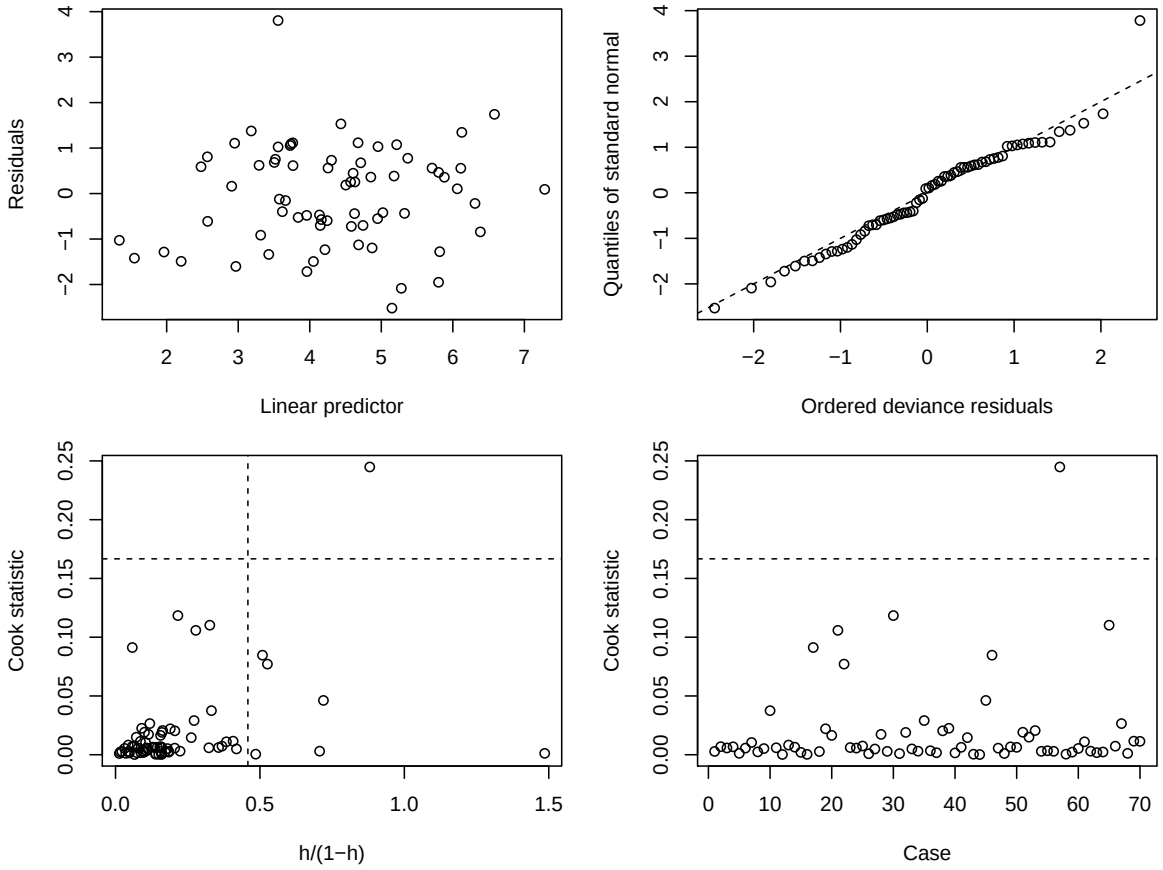

Figure 2: Diagnostic plots for the `ceb` data: deviance residuals versus linear predictor (top left), quantile-quantile plot of deviance residuals (top right), Cook distance against weighted leverage (bottom left), and Cook's distance statistics (bottom right)

- `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 4.4.1 is a Poisson regression model with `weekend` as covariate.
- Model 4.4.2 is a Poisson regression model with `weekend`, `relhumid` and `temp` as covariates.
- Model 4.4.3 is a negative binomial model with `weekend`, `relhumid` and `temp` as covariates.
- Model 4.4.4 is a negative binomial model with `weekday` (categorical), `relhumid` and `temp` as covariates.

(a) Is Model 4.4.1 an adequate simplification of Model 4.4.2? Assess this hypothesis formally.

Solution

The models are nested, so we can use a likelihood ratio test to compare them with $\mathscr{H}_0 : \beta_{\texttt{temp}} = \beta_{\texttt{relhumid}} = 0$. The likelihood ratio statistic is $2 \times (2577.3604 - 2190.8777) = 772.97$, to be compared with a χ^2_2 null distribution. The linear effects of the additional covariates `relhumid` and `temp` are statistically significant.

(b) Interpret the coefficients for the intercept and for `relhumid` in Model 4.4.2.

Solution

- When the relative humidity is zero and the temperature is $0°C$, the average number of users on weekdays is $\exp(\widehat{\beta}_0) = 13,07$.
- For every percentage increase in the relative humidity, *ceteris paribus*, the estimated mean number of users is multiplied by a factor $\exp(\widehat{\beta}_{\texttt{relhumid}}) = \exp(-0.0066) = 0.9934217$, corresponding to a decrease of 0.657%.

(c) Compare the model fit of the negative binomial model (Model 4.4.3) with that of the Poisson (Model 4.4.2) using all of the following methods: (a) the deviance statistic (b) a likelihood ratio test and (c) information criteria.

Solution

- Deviance statistic: the ratio of the deviance, 1954, relative to the degrees of freedom for the Poisson regression, 496, is 3.94, while that of the negative binomial regression model is $522.3013/496 = 1.0530$. Both models are compared to the saturated models using a likelihood ratio test and the negative binomial is adequate (ratio should be approximately one).
- Likelihood ratio test between **Model 4.4.2** and **Model 4.4.3** (non-regular). The "Full Log Likelihood" gives ℓ, which is -1808.0756 for the negative binomial and -2190.8777 for the Poisson. The LRT statistic is twice this difference, 765.6042, to be compared to a $\frac{1}{2}\chi^2_1$. The Poisson model is clearly not an adequate simplification due to overdispersion.
- Information criteria: the value of AIC for the Poisson model is 4389.75, versus 3626.27 for the negative binomial, suggesting the latter is preferable. Same for BIC ($4406.6137 > 3647.22$)

(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. Which of the two makes more sense in the present context?

Solution

Only the categorical variable makes sense. Integer-valued implies that there is a linear effect, but the number of days is arbitrary. What would make more sens is a cyclical effect, but this cannot be accomodated with a linear trend.

(e) The equation for the mean of Model 4.4.4 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 4.4.3 to Model 4.4.4 in terms of the model parameters $\boldsymbol{\beta}$, thereby showing that Model 4.4.3 is nested within Model 4.4.4. Does the number of user significantly vary between weekdays and between weekend days?

Solution

This is yet another likelihood ratio test to compare Model 4.4.3 and Model 4.4.4. The null hypothesis (in terms of the model parameters) is $\mathscr{H}_0 : \beta_3 = \beta_4 = \beta_5 = \beta_6 = \beta_7, \beta_8 = 0$, hence the models are nested. The statistic is $2 \times (1808.0756 - 1801.7758) = 12.6$, to be compared with a χ^2_5. The p-value is 0.027, hence we reject $\mathscr{H}_0$ at level 5% and conclude that there is evidence that the effect differs across within week-days and week-ends.

4.5 **Soccer matches**: Let Y_{ij} (resp. Z_{ij}) denote the score of the home (resp. visitor) team for a soccer match opposing teams i and j. Maher (1982) suggested modelling the scores as

$$Y_{ij1} \sim \mathsf{Po}\{\exp(\delta + \alpha_i + \beta_j)\}, \quad Y_{ij2} \sim \mathsf{Po}\{\exp(\alpha_j + \beta_i)\}, \qquad i \neq j; i, j \in \{1, \ldots, 20\}, \tag{E4.5.1}$$

where α_i represent the offensive strength of the team, β_j the defensive strength of team j and δ is the common home advantage. The scores in a match and between matches are assumed to be independent of one another. The data set `soccer` contains the results of football (soccer) matches for the 2015 season of the English Premier Ligue (EPL) and contains the following variables

- `score`: number of goals of `team` during a match
- `team`: categorical variable giving the name of the team which scored the goals
- `opponent`: categorical variable giving the name of the adversary
- `home`: binary variable, 1 if `team` is playing at home, 0 otherwise.

(a) A common home advantage δ makes sense provided that the scores at home and away are independent, i.e., there is no interaction between the two. To validate this hypothesis, we consider aggregates over multiple seasons of the scores, cross-classified in terms of number of points for the team at home and the team away, for each match (Table 1). The file `socceragg` contains the two-way contingency table in long-format. Using the latter, test the assumption of independence.

	away						
home	0	1	2	3	4	5	6
0	32	33	9	14	3	0	1
1	37	41	28	11	3	1	0
2	27	25	29	7	2	1	0
3	18	15	10	5	2	0	0
4	9	6	3	0	0	1	0
5	0	4	0	0	0	0	0
6	0	1	2	0	0	0	0

Table 1: Frequency of scores for EPL soccer matches in 2015

Solution

Testing "independence" amounts to testing for the statistical significance of the interaction term, which corresponds to the saturated model. The deviance statistic for the Poisson model is 43.8 for 36 degrees of freedom. Under the null hypothesis of independence, the deviance follows a χ^2_{36} and the p-value is 0.174; we fail to reject the null, suggesting here that a common home advantage is adequate.

(b) Fit the model characterized by Equation (E4.5.1) and answer the following questions:

i. Using the fitted model, give the expected number of goals for a match between Manchester United (at home) against Liverpool.

ii. Report and interpret the estimated home advantage $\widehat{\delta}$.

iii. Test whether the home advantage δ is significantly different from zero.

iv. The asymptotic null distribution of the deviance statistic D is χ^2_{n-p-1}, but the latter is only valid when the number of observations in each group is large. In our analysis, there are only 38 matches in a given year at home/visiting for each team. We can instead approximate the null distribution of D using a simulation: specifically, we repeat the following steps $B = 10\,000$ times

   A. generate new Poisson data from the fitted model
   B. fit the Poisson regression specified by Equation (E4.5.1) on the simulated data
   C. calculate the deviance statistic.

   Table 2 gives quantiles of the simulated null distribution of the deviance based on these $B = 10\,000$ simulations. Comment on the adequacy of the fit based on the deviance statistic and Table 2 and contrast with the conclusions obtained by using the asymptotic null distribution of the deviance.

level	2.5%	5%	10%	25%	50%	75%	90%	95%	97.5%
quantile	760.53	770.33	782.30	803.41	826.25	849.92	871.23	885.74	897.44

Table 2: Quantiles of the simulated deviance statistics based on the model of Maher (1982)

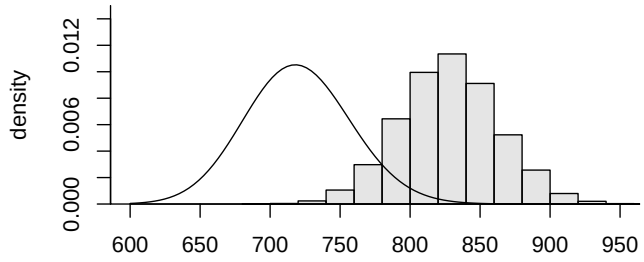

Figure 3: Asymptotic (continuous curve) and simulated (grey histogram) null distribution for the deviance statistic of Model E4.5.1.

Solution

i. We plug-in the coefficients estimates in eq. (E4.5.1) to get the prediction, of 1.385 goals for Manchester United and 1.01 for Liverpool.

ii. The estimated home advantage is $\widehat{\delta} = 0.21$ with 95% profiled-based confidence interval of $[0.0088, 0.33]$. It corresponds to a 23.5% increase in the average number of goals scored by the home team during a match.

iii. The likelihood ratio statistic for comparing the model in which home advantage is zero is 11.39. Compared to a χ^2_1, this result is unlikely and we strongly reject the null hypothesis that $\delta = 0$ (p-value of 0.00074).

iv. The deviance (829.78) is seemingly large relative to the degrees of freedom (720) is 1.15, with an asymptotic p-value of 0.0027; see Figure 3. However, the later is unreliable. Based on the simulated values, the deviance is close to the median, so there is no evidence against the null hypothesis that the model is as adequate as a saturated model.

(c) Maher also suggested more complex models, including one in which the offensive and defensive strength of each team changed depending on whether they were at home or visiting another team, i.e.

$$Y_{ij1} \sim \text{Po}\{\exp(\alpha_i + \beta_j + \delta)\}, \quad Y_{ij2} \sim \text{Po}\{\exp(\gamma_j + \omega_i)\}, \qquad i \neq j;\ i, j \in \{1, \ldots, 20\} \tag{E4.5.2}$$

Does Model E4.5.2 fit significantly better than Model E4.5.1?

Solution

To fit Model E4.5.2, we include the interaction `home·opponent` and `home·team` in the model — this is a contrast parametrization. Both models are nested. The first model has 40 parameters, the second has 78 parameters and so the likelihood ratio statistic, worth 47.86, follows approximately a χ^2_{38} distribution. The p-value is 0.13, suggesting the more complex model is not a significative improvement over eq. (E4.5.1)

(d) Would a similar Poisson be adequate to model basketball scores? Justify your answer

Solution

Using data from the last ten years for the NBA matches, the average number of points per match is 103, with a standard deviation of 12.89. This suggests overdispersion: because of the mean-variance relationship, the Poisson distribution might be too inflexible for modelling the scores.

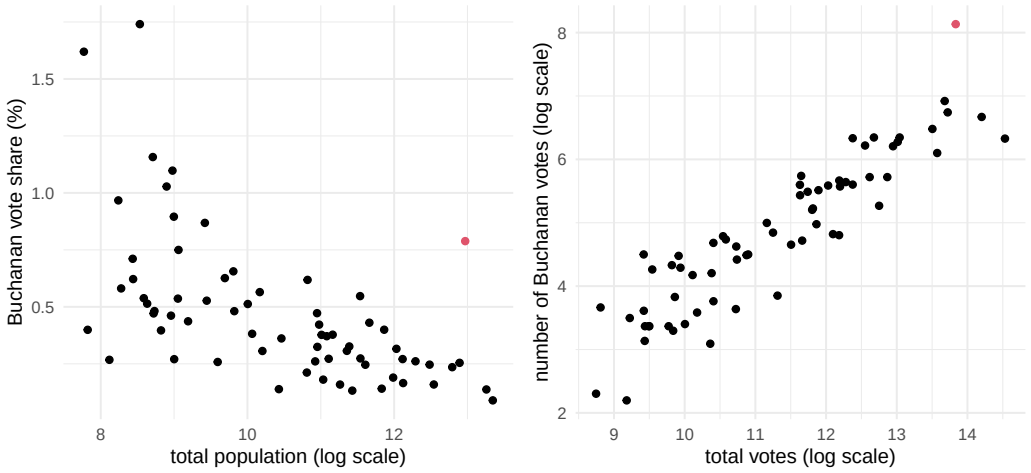

Figure 4: Buchanan's share of vote as a function of log of population in county (left) and the number of votes for Buchanan as a function of total ballots cast. The point in red, an outlier, corresponds to Palm Beach County.

4.6 **Bush vs Gore**: the 2000 US presidential election opposed Georges W. Bush (GOP) and Albert A. Gore (Democrat), as well as marginal third party candidates. The tipping state was Florida, worth 25 electors, which Bush won by a narrow margin of 537 votes. There have been many claims that the design of so-called butterfly ballots used in poor neighborhoods of Palm Beach county led to confusion among voters and that this deprived Gore of some thousands of votes that were instead assigned to a paleoconservative third-party candidate, Patrick Buchanan (Reform). Smith (2002) analysed the election results in Palm Beach country, in which a unusually high number of ballots (3407) were cast for Buchanan.

We are interested in building a model to predict the expected number of votes for Buchanan in Palm Beach county, based only on the information from other county votes. The `buchanan` data contains the following variables:

- `county`: name of county
- `popn`: population of the county in 1997.
- `white`: percentage of "white" (*sic*) in 1996 (per US Census definitions; people having origins in any of the original peoples of Europe, the Middle East, or North Africa).
- `black`: percentage of Black and African Americans in 1996 (origins in sub-saharian Africa).
- `hisp`: percentage of Hispanics in 1996.
- `geq65`: percentage of the population aged 65 and above based on 1996 and 1997 population estimates.
- `highsc`: percentage of the population with a high school degree (1990 Census data).
- `coll`: percentage of the population that are college graduates (1990 Census data).
- `income`: mean personal income in 1994.
- `buch`: total ballots cast for Pat Buchanan (Reform).
- `bush`: total ballots cast for Georges W. Bush (GOP).
- `gore`: total ballots cast for Al Gore (Democrat).
- `totmb`: total number of votes cast for the presidential election in each county, minus Buchanan votes.

(a) Calculate the total proportion of votes for Buchanan in Florida.

(b) Plot the percentage of votes obtained by Buchanan, `buch/(buch+totmb)`, against ln(`popn`) and comment.

Solution

The vote share for Buchanan is higher is small (rural) counties. There is a clear outlier in the left panel of Figure 4 at approximately $[13, 0.75\%]$ that corresponds to Palm Beach county. There seems to be more hetero-

geneity in less populated counties.

Exclude the results of Palm Beach county for the rest of the question.

(c) We consider first a Poisson model for the percentage of votes for Buchanan, `buch/totmb`, as a function of `white`, ln(`hisp`), `geq65`, `highsc`, ln(`coll`), `income`.

i. Explain why an offset is necessary in this case.

ii. Why is `totmb` a better choice of denominator than `popn` for the rate? Explain.

iii. Is the Poisson model appropriate? Justify your answer.

iv. Explain why, if there is evidence of overdispersion, this means the binomial model is also inadequate. *Hint: what is the variance of the binomial distribution and how does it relate to the Poisson distribution?*

Solution

i. The population in counties differ drastically, from 6.3K to 2 million voters.

ii. We want the percentage of votes, so `totmb` — the percentage of Buchanan's votes is less than 1.5%, so omitting it won't affect the results much. `popn` would not be adequate because the percentage of inhabitants who cast a ballot differs a lot across counties, ranging from 25% to 58%. Part of this has to do with citizenship status (immigrants may not be entitled to vote), number of felons or age (only adults can vote).

iii. The number of trials is large, so we expect the χ^2 approximation to the deviance to be adequate. The deviance statistic for the sample excluding county 50 is $D = 596.25$ with $\nu = 58$ residual degrees of freedom, a ratio of almost 10! There is clear evidence of overdispersion; testing whether a negative binomial model fits better yields a likelihood ratio statistic of 562.71, suggesting overwhelming evidence against the null of equal mean-variance.

iv. The variance for the binomial distribution is $N_i p_i(1-p_i)$, compared to the Poisson distribution which has $N_i p_i$. Since $p_i \approx 0.003$, this second term is negligible and there is also overdispersion for the fraction that cannot be handled directly with a binomial model.

(d) Use a negative binomial model with the same covariates to predict the expected number of Buchanan votes in Palm Beach county. Comment hence on the discrepancy between this forecast and the number of votes received in the election.

Solution

The predicted number of voters (rounded to the nearest unit) is 504 voters for the negative binomial model, compared to 438 for the Poisson model: the change in the likelihood affects the parameter estimates.

Standard software only return confidence intervals for the mean, but not for predictions. The prediction interval combines two sources of error: the uncertainty associated to the estimated coefficients and the uncertainty arising from the distribution. Since maximum likelihood estimators are asymptotically normal, we can obtain uncertainty by using this approximation. We can approximate the latter using a Monte Carlo simulation, where for each iteration

- we sample a draw $\boldsymbol{\beta}_b \sim \mathsf{No}_{p+1}\{\widehat{\boldsymbol{\beta}}, j^{-1}(\widehat{\boldsymbol{\beta}})\}$ and similarly for $k_b \sim \mathsf{No}\{\widehat{k}, \mathsf{se}(\widehat{k})\}$ [uncertainty of the estimated coefficients].
- we simulate one new observation $Y_b \sim \mathsf{NB}(\mu_b = \exp(\boldsymbol{\beta}_b \cdot \mathbf{x}_{\mathrm{PB}} + o), k = k_b)$, where o is the offset, $\mathbf{x}_{\mathrm{PB}}$ is the row of the model matrix corresponding to Palm Beach (including the intercept) [uncertainty arising from the distribution of the response]

Based on the B values for the prediction, we simply compute the sample quantiles of μ_b (confidence interval for the mean) and the sample quantiles of Y_b to get an approximate prediction interval With $B = 10000$ replications and looking at the 0.025 and 0.975 quantiles of the simulated draws, this yields a 95% prediction interval of [260, 861], far from the observed counts.

Compiled 16/12/2020 at 10:23