Exercise 4 - solutions

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.

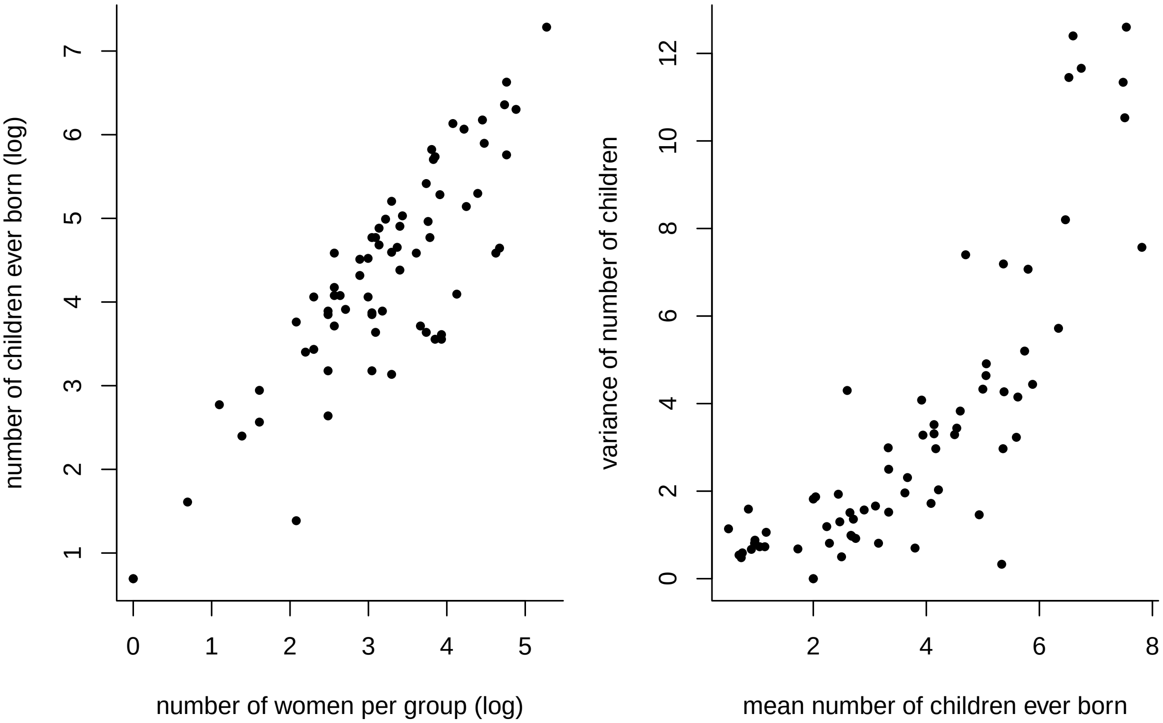


Figure 1: Number of children ever born as function of the number of women per group on log-log scale (left) and mean versus variance of the number of children ever born per group (right).

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.

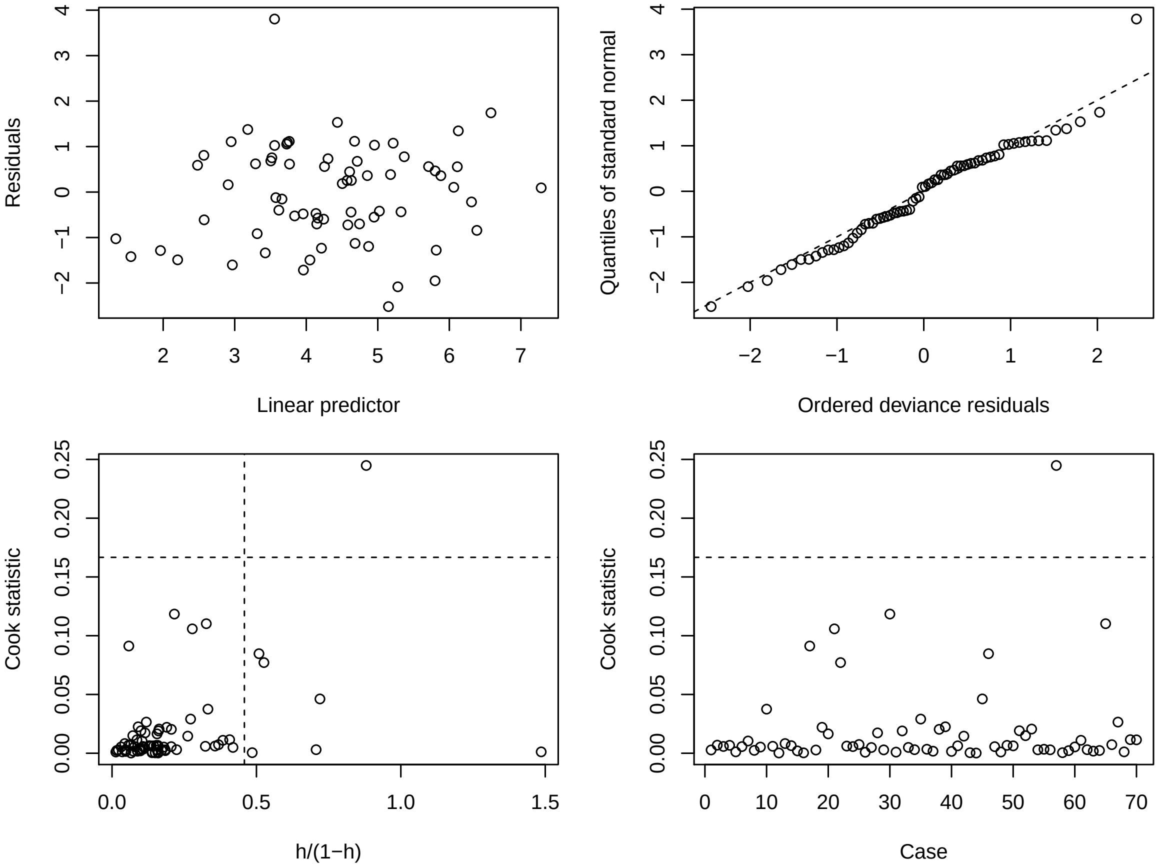


Figure 2: Diagnostic plots for the `ceb` data: deviance residuals versus linear predictor (top left), quantile-quantile plot of deviance residuals (top right), Cook distance against weighted leverage (bottom left), and Cook's distance statistics (bottom right)

- `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 $0^\circ C$, 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 **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

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

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

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

The deviance measures the **discrepancy** in fit between the saturated and the fitted models. Under regularity conditions and assuming the number of observations in each of the JK levels goes to ∞,

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

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.

Solution

There are $JK = 6$ parameters in the saturated model.

model	deviance	p	variables
M_0	12.66	1	`none`
M_1	6.64	3	`age`
M_2	5.96	2	`malignant`
M_3	0.49	4	`age`, `malignant`

Table 1: Analysis of deviance for the `cancer` data set

We first compare the additive model M_3 against the saturated model with $JK = 6$ coefficients. The deviance is 0.49 and under the null hypothesis $\mathscr{H}_0 : \nu_{jk} = 0, j = 1, \ldots, J-1, k = 1, \ldots, K-1$, and $D(\widehat{\boldsymbol{\beta}}_{M_3}) = 0.49 \sim \chi^2_2$ asymptotically. The p-value for this hypothesis is 0.78, so we fail to reject the null that the simpler model M_3 is an adequate simplification of the saturated model.

The next step is to compare M_3 against model M_2 or M_1. Consider first M_3 versus M_2, corresponding to the null hypothesis $\mathscr{H}_0 : \beta_j = 0 (j = 1, \ldots, J-1)$ and $D(\widehat{\boldsymbol{\beta}}_{M_2}) - D(\widehat{\boldsymbol{\beta}}_{M_3}) \dot{\sim} \chi^2_{p_3 - p_2}$; we compare the value of the test statistic, 5.46, against the quantiles of a χ^2_2. The 95% of the χ^2_2 distribution is 5.99, so we fail to reject the null at the 5% level. The p-value associated to $\mathscr{H}_0$ is 0.065.

We could have performed a test for M_3 against M_1; the test statistic is $6.64 - 0.49 = 6.15$, to compare against the 95% of the χ^2_1 distribution, which is 3.84. The p-value is 0.013, so the coefficient corresponding to `malignant` is significative.

Lastly, we could compare M_2 against M_0. The null hypothesis is $\mathscr{H}_0 : \gamma_k = 0 (k = 1, \ldots, K-1)$. The p-value is 0.009.

If model M_2 was appropriate, its deviance should be approximately χ^2_4 but the true distribution depends on the unknown parameters. The p-value of observing a deviance for M_2 as extreme as 5.96 is approximately 0.25, so reasonable.

This is due to the fact that the design is strongly unbalanced (for older age categories with lower counts). A better way to assess whether the null χ^2_4 distribution is approximately correct is to simulate replicates from the fitted model (conditioning on the number of counts per age and malignant status) to obtain an empirical distribution for the deviance and compare the latter to the asymptotic χ^2_4 distribution; we find that the latter is appropriate.

4.6 **Equivalence of Poisson and binomial models:** Suppose $Y_j \sim \mathsf{Bin}(m_j, p_j)$ 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 $\mathsf{Po}(\mu_j)$. 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.

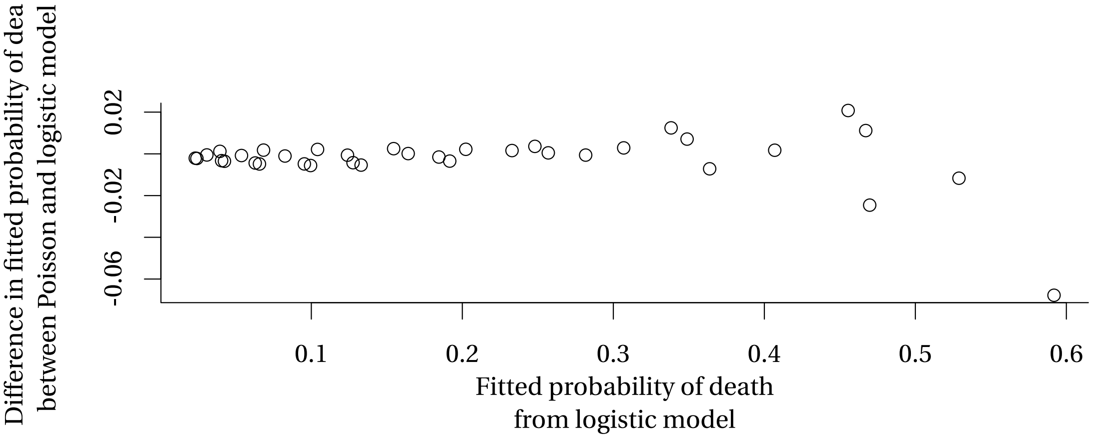
Solution

(a) We perform backward elimination; going from the saturated model with $m = 36$ parameters to the model M_3 corresponds to $\mathscr{H}_0 : \nu_{jk} = 0, j = 1, \ldots, J-1; k = 1, \ldots, K-1$. The p-value is 0.61 for the Poisson model with offset and 0.55 for the binomial model, so we fail to reject the null that M_3 is an adequate simplification. The additive model cannot be further simplified and it appears adequate, since $D(\widehat{\boldsymbol{\beta}}_{M_3}) \approx m - p = 24$.

model	deviance (binom.)	deviance (Poisson)	p
M_0	4055.98	4917.03	1
M_1	3910.70	4740.34	4
M_2	191.72	247.94	9
M_3	21.49	22.44	12

Table 2: Analysis of deviance for the `smoking` data set

(b) We can see that, for some categories, the proportion of death from lung cancer is high and so the Poisson approximation is poor.


Difference in fitted probability of dea
between Poisson and logistic model
Fitted probability of death
from logistic model