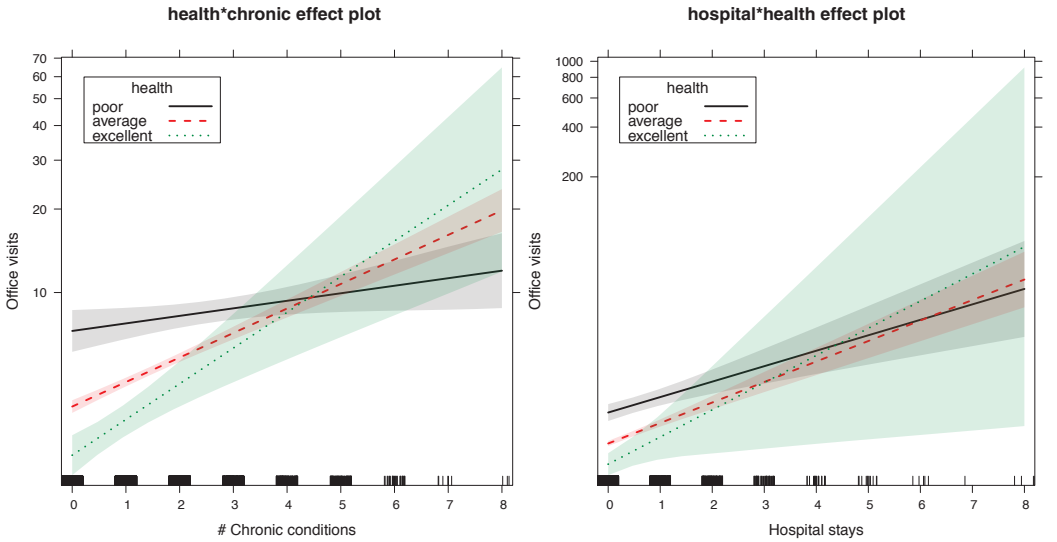


Figure 11.29: Effect plots for the interactions of chronic conditions and hospital stays with perceived health status in the model nmes.nbin2.

stays and number of chronic conditions (left panel of Figure 11.30) has a clearly interpretable pattern: for those with few chronic conditions, there is a strong positive relationship between hospital stays and office visits. As the number of chronic conditions increases, the relation with hospital stays decreases in slope.

```
> plot(eff_nbin2, "hospital:chronic", multiline = TRUE, ci.style = "bands",
+      ylab = "Office visits", xlab = "Hospital stays",
+      key.args = list(x = 0.05, y = .70, corner = c(0, 0), columns = 1))
>
> plot(eff_nbin2, "health:school", multiline = TRUE,  ci.style = "bands",
+      ylab = "Office visits", xlab = "Years of education",
+      key.args = list(x = 0.65, y = .1, corner = c(0, 0), columns = 1))
```

Finally, the interaction of `health:school` is shown in the right panel of Figure 11.30. It can be readily seen that for those of poor health, office visits are uniformly high, and have no relation to years of education. Among those of average or excellent health, office visits increase with years of education in roughly similar ways. △

11.5.2.3 More model wrinkles: Nonlinear terms

Effect plots such as those above are much easier to interpret than tables of fitted coefficients. However, we emphasize that these only reflect the *fitted model*. It might be that the effects of both `hospital` and `chronic` are nonlinear (on the scale of `log(visits)`). In assessing this question, we increase the complexity of model and try to balance parsimony against goodness-of-fit, but also assure that the model retains a sensible interpretation.

EXAMPLE 11.16: Demand for medical care

The simplest approach is to use `poly(hosp,2)` and/or `poly(numchron,2)` to add possible quadratic (or higher power) relations to the model `nmes.nbin2` containing interactions studied above. A slightly more complex model could use `poly(hosp, numchron, degree=2)` for a response-surface model in these variables. A significantly improved fit of such a model is evidence for nonlinearity of the effects of these predictors. This is easily done using `update()`:

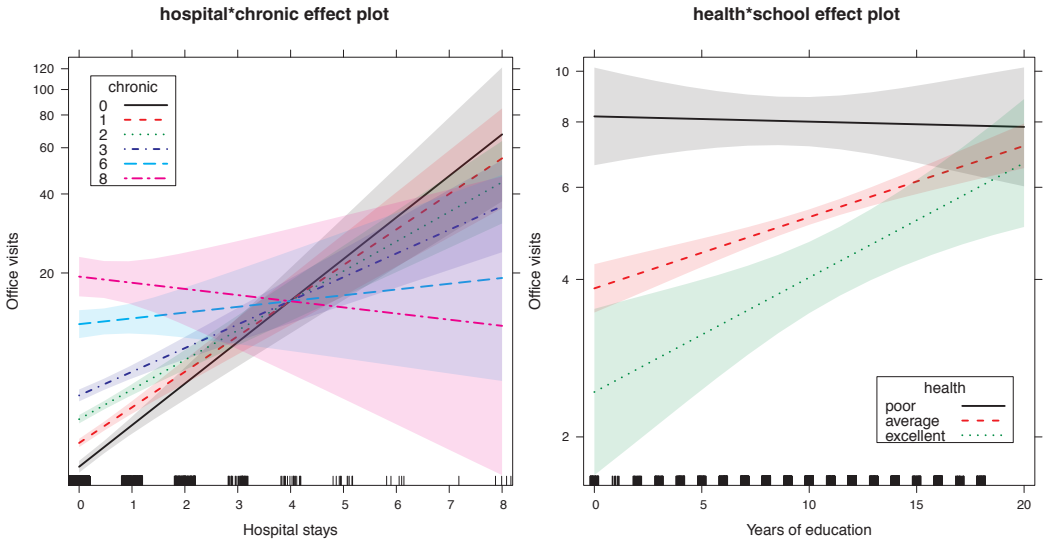


Figure 11.30: Effect plots for the interactions of chronic conditions and hospital stays and for health status with years of education in the model nmes.nbin2.

```
> nmes.nbin3 <- update(nmes.nbin2, . ~ . + I(chronic^2) + I(hospital^2))
```

This model is equivalent to the long-form version below:

```
> nmes.nbin3 <- glm.nb(visits ~ poly(hospital, 2) + poly(chronic, 2) +
+                      insurance + school + gender +
+                      (health + chronic + hospital)^2 + health : school,
+                      data = nmes)
```

Comparing these models using `anova()`, we see that there is a substantial improvement in the model fit by including these nonlinear terms. The quadratic model also fits best by AIC and BIC.

```
> ret <- anova(nmes.nbin, nmes.nbin2, nmes.nbin3)
> ret$Model <- c("nmes.nbin", "nmes.nbin2", "nmes.nbin3")
> ret

Likelihood ratio tests of Negative Binomial Models

Response: visits
       Model  theta Resid. df    2 x log-lik.   Test    df
1  nmes.nbin 1.2066      4398          -24341
2 nmes.nbin2 1.2354      4391          -24267 1 vs 2     7
3 nmes.nbin3 1.2446      4389          -24245 2 vs 3     2
  LR stat.    Pr(Chi)
1
2   74.307 1.9829e-13
3   22.278 1.4537e-05

> LRstats(nmes.nbin, nmes.nbin2, nmes.nbin3)

Likelihood summary table:
             AIC   BIC LR Chisq   Df Pr(>Chisq)
nmes.nbin  24359 24417     5045 4398    2.2e-11 ***
nmes.nbin2 24299 24401     5047 4391    1.1e-11 ***
nmes.nbin3 24281 24396     5049 4389    8.5e-12 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

However, effect plots for this model quickly reveal a *substantive* limitation of this approach using polynomial terms. Figure 11.31 shows one such plot for the interaction of health and number of chronic conditions that you should compare with Figure 11.28.

```
> eff_nbin3 <- allEffects(nmes.nbin3,
+   xlevels = list(hospital = c(0 : 3, 6, 8),
+                  chronic = c (0 : 3, 6, 8),
+                  school = seq(0, 20, 5)))
> plot(eff_nbin3, "health : chronic", layout = c(3, 1))
```

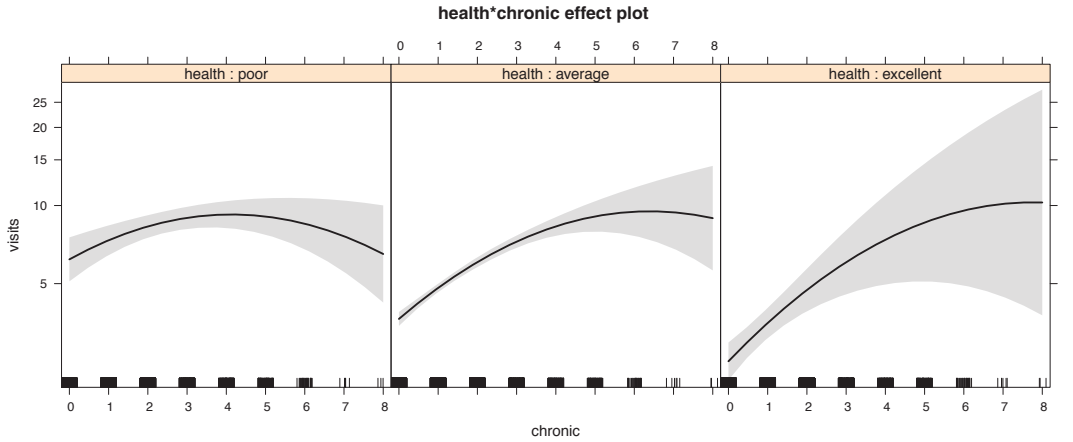


Figure 11.31: Effect plot for the interaction of health and number of chronic conditions in the quadratic model nmes.nbin3.

The quadratic fits for each level of health in Figure 11.31 imply that office visits increase with chronic conditions up to a point and then decrease—with a quadratic, what goes up must come down, the same way it went up! This makes no sense here, particularly for those with poor health status. As well, the confidence bands in this figure are uncomfortably wide, particularly at higher levels of chronic conditions, compared to those in Figure 11.28. The quadratic model is thus preferable statistically and descriptively, but serves less well for explanatory, substantive, and predictive goals.

An alternative approach to handle nonlinearity is to use regression splines (as in Example 7.9) or a ***generalized additive model*** (Hastie and Tibshirani, 1990) for these terms. The latter specifies the linear predictor as a sum of smooth functions,

$$g(\mathcal{E}(y)) = \beta_0 + f_1(x_1) + f_2(x_2) + \cdots + f_m(x_m)\,.$$

where each $f_j(x_j)$ may be a function with a specified parametric form (for example, a polynomial) or may be specified non-parametrically, simply as "smooth functions," to be estimated by non-parametric means.

In R, a very general implementation of the generalized additive model (GAM) is provided by `gam()` in the mgcv (Wood, 2015) package and described in detail by Wood (2006). Particular features of the package are facilities for automatic smoothness selection (Wood, 2004), and the provision of a variety of smooths of more than one variable. This example just scratches the surface of GAM methodology.

In the context of the NB model we are considering here, the analog of model `nmes.nbin3` fitted using `gam()` is `nmes.gamnb` shown below. The negative-binomial distribution can be specified using `family=nb()` when the parameter θ is also estimated from the data (as with

`glm.nb()`), or `family=negbin(theta)` when θ is taken as fixed, for example using the value `theta=1.24` available from models `nmes.nbin2`, and `nmes.nbin3`.

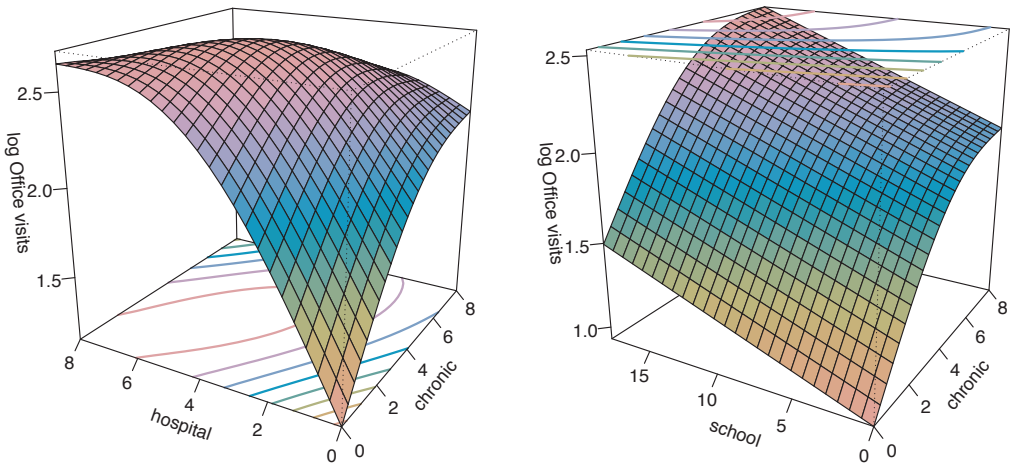
```
> library(mgcv)
> nmes.gamnb <- gam(visits ~ s(hospital, k = 3) + s(chronic, k = 3) +
+                             insurance + school + gender +
+                             (health + chronic + hospital)^2 +
+                             health : school,
+                   data = nmes, family = nb())
```

The key feature here is the specification of the smooth terms for `s(hospital, k=3)` and `s(chronic, k=3)`, where `k=3` specifies the dimension of the basis used to represent the smooth term. There are many other possibilities with `gam()`, but these are beyond the scope of this example.

We could again visualize the predicted values from this model using effect plots. However, a different approach is to visualize the *fitted surface* in 3D, using a range of values for two of the predictors, and controlling for the others.

The rsm package provides extensions of the standard `contour()`, `image()` and `persp()` functions for this purpose. The package provides S3 methods (e.g., `persp.lm()`) for "lm" objects, or classes (such as "negbin" and "glm") that inherit methods from `lm`. The calculation of fitted values in these plots use the applicable `predict()` method for the model object. As in effect plots, the remaining predictors are controlled at their average values (or other values specified in the `at` argument).

Two such plots are shown in Figure 11.32. The left panel shows the interaction of hospital stays and chronic conditions, included in the model with smoothed terms for their main effects. The right panel shows the joint effects of years of education and chronic conditions on office visits, but there is no interaction of these variables in the GAM model `nmes.gamnb`. These plots use `rainbow()` colors (in HCL space) to depict the predicted values of office visits. Contours of these values are projected into the bottom or top plane with corresponding color coding.[17]



Slice at school = 10, health = average

Slice at hospital = 0.3, health = average

Figure 11.32: Fitted response surfaces for the relationships among chronic conditions, number of hospital stays, and years of education to office visits in the generalized additive model, nmes.gamnb.

[17] The vignette `vignette("rsm-plots", package="rsm")` illustrates some of these options.

```
> library(rsm)
> library(colorspace)
> persp(nmes.gamnb, hospital ~ chronic, zlab = "log Office visits",
+   col = rainbow_hcl(30), contour = list(col = "colors", lwd = 2),
+   at = list(school = 10, health = "average"), theta = -60)
>
> persp(nmes.gamnb, school ~ chronic, zlab = "log Office visits",
+   col = rainbow_hcl(30),
+   contour = list(col = "colors", lwd = 2, z = "top"),
+   at = list(hospital = 0.3, health = "average"), theta = -60)
```

A simple, credible interpretation of the plot in the left panel is that office visits rise steeply initially with both hospital stays and number of chronic conditions, and then level off. For those with no chronic conditions, the effect of hospital stays rises to a higher level compared with the effect of chronic conditions among those who have had no hospital stays. However, as we have seen before, the data is quite thin at the upper end of these predictors, and this plot does not show model uncertainty.

The right panel of Figure 11.32 illustrates the form of model predictions for a term where one variable (`chronic`) is treated as possibly nonlinear using a smooth `s()` effect, the other is treated as linear (`school`), and no interaction between these is included in the model. At each fixed value of `chronic`, increasing education results in greater office visits. At each fixed value of `school`, the number of chronic conditions shows a steep increase in office visits initially, leveling off toward higher levels, but these all have the same predicted shape.

△

11.6 Diagnostic plots for model checking

> Models, of course, are never true, but fortunately it is only necessary that they be useful.
>
> G. E. P. Box, *Some Problems of Statistics of Everyday Life*, 1979, p. 2

Most of the model diagnostic methods for classical linear models extend in a relatively direct way to GLMs. These include (a) plots of residuals of various types, (b) diagnostic measures and plots of leverage and influence, as well as some (c) more specialized plots (component-plus-residual plots, added-variable plots) designed to show the specific contribution of a given predictor among others in a linear model. These methods were described in Section 7.5 in the context of logistic regression, and most of that discussion is applicable here in wider GLM class.

One additional complication here is that in any GLM we are specifying: (a) the distribution of the random component, which for count data models may also involve a dispersion parameter or other additional parameters; (b) the form of the linear predictor, $\eta = \boldsymbol{x}^\mathsf{T}\beta = \beta_0 + \beta_1 x_1 + \cdots$, where all important regressors have been included, and on the right scale; (c) the correct link function, $g(\mu) = \eta$ transforming the conditional mean of the response y to the predictor variables where they have linear relationships.

Thus, there are a lot of things that could go wrong, but the famous quote from George Box should remind us that all models are approximate, and the goal for model diagnosis should be an adequate model, useful for description, estimation, or prediction as the case may be. What is most important is that our models should not be misleadingly wrong, that is, they should not affect substantive conclusions or interpretation.

11.6.1 Diagnostic measures and residuals for GLMs

Estimation of GLMs by maximum likelihood uses an iterative weighted least squares (IWLS) algorithm, and many of the diagnostic measures for these models are close counterparts of their forms

for classical linear models. Roughly speaking, these follow from replacing $\boldsymbol{y}$ and $\widehat{\boldsymbol{y}}$ in least squares diagnostics by a "working response" and $\widehat{\boldsymbol{\eta}}$, replacing the residual variance $\widehat{\sigma}^2$ by $\widehat{\phi}$, and using a weighted form of the Hat matrix.

11.6.1.1 Leverage

Hat values, h_i, measuring ***leverage*** or the potential of an observation to affect the fitted model, are defined as the diagonal elements of the hat matrix $\boldsymbol{H}$, using the weight matrix $\boldsymbol{W}$ from the final IWLS iteration. This has the same form as in a weighted least squares regression using a fixed $\boldsymbol{W}$ matrix:

$$\boldsymbol{H} = \boldsymbol{W}^{1/2}\boldsymbol{X}(\boldsymbol{X}^{\mathsf{T}}\boldsymbol{W}\boldsymbol{X})^{-1}\boldsymbol{X}^{\mathsf{T}}\boldsymbol{W}^{1/2} \, .$$

In contrast to OLS, the weights depend on the $\boldsymbol{y}$ values as well as the $\boldsymbol{X}$ values, so high leverage observations do not necessarily reflect only unusualness in the space of the predictors.

11.6.1.2 Residuals

Several types of residuals can be defined starting from the goodness-of-fit measures discussed in Section 11.1.3. The ***raw residual*** or ***response residual*** is simply the difference $y_i - \widehat{\mu}_i$ between the observed response y_i and the estimated mean, $\widehat{\mu} = g^{-1}(\widehat{\eta}_i) = g^{-1}(\boldsymbol{x}_i^{\mathsf{T}}\widehat{\beta})$.

From this, the ***Pearson residual*** is defined as

$$r_i^P = \frac{y_i - \widehat{\mu}_i}{\sqrt{\widehat{\mathcal{V}}(y_i)}} \tag{11.11}$$

and the ***deviance residual*** is defined as the signed square root of the contribution of observation i to the deviance in Eqn. (11.4).

$$r_i^D = \text{sign}(y_i - \widehat{\mu}_i)\sqrt{d_i} \, . \tag{11.12}$$

The Pearson and deviance residuals do not account for dispersion or for differential leverage (which makes their variance smaller), so ***standardized residuals*** (sometimes called *scaled* residuals) can be calculated as

$$\widetilde{r}_i^P = \frac{r_i^P}{\sqrt{\widehat{\phi}(1-h_i)}} \, . \tag{11.13}$$

$$\widetilde{r}_i^D = \frac{r_i^D}{\sqrt{\widehat{\phi}(1-h_i)}} \, . \tag{11.14}$$

These have approximate standard normal $\mathcal{N}(0,1)$ distributions, and will generally have quite similar values (except for small values in $\widehat{\mu}$). Consequently, convenient thresholds like $|\widetilde{r}_i| > 2$ or $|\widetilde{r}_i| > 4$ are useful for identifying unusually large residuals.

Finally, the ***studentized residual*** (or *deletion* residual) gives the standardized residual that would result from omitting each observation in turn and calculating the change in the deviance. Calculating these exactly would require refitting the model n times, but an approximation is

$$\widetilde{r}_i^S = \text{sign}(y_i - \widehat{\mu}_i)\sqrt{(\widetilde{r}_i^D)^2 + (\widetilde{r}_i^P)^2 h_i/(1-h_i)} \, . \tag{11.15}$$

From the theory of classical linear models, these provide formal outlier tests for individual observations (Fox, 2008, Section 11.3) as a *mean-shift* outlier model that dedicates an additional parameter to fit observation i exactly. To correct for multiple testing and a focus on the largest absolute residuals, it is common to apply a Bonferroni adjustment to the p-values of these tests, multiplying them by n.

For a class "glm" object, the function `residuals(object, type)` returns the unstandardized residuals for `type="pearson"` or `type="deviance"`.[18] The standardized versions are obtained using `rstandard()`, again with a `type` argument for the Pearson or deviance flavor. `rstudent()` calculates the studentized deletion residuals.

11.6.1.3 Influence

As discussed in Section 7.5 in the context of logistic regression, influence measures attempt to evaluate the effect that an observation exerts on the parameters, fitted values, or goodness-of-fit statistics by comparing a statistic calculated for all the data with the value obtained omitting each observation in turn. Again, approximations are used to estimate these effects without laboriously refitting the model n times.

Overall measures of influence include

- Cook's distance (Eqn. (7.10)), a squared measure of the difference $\widehat{\boldsymbol{\beta}} - \widehat{\boldsymbol{\beta}}_{(-i)}$ in all p coefficients in the model. The approximation used in `cooks.distance()` is

$$C_i = \frac{\widetilde{r}_i h_i}{\widehat{\phi}\, p\, (1 - h_i)} \, .$$

  This follows Williams (1987), but scales the result by the estimated dispersion $\widehat{\phi}$ as an approximate $F_{p,n-p}$ statistic rather than χ^2_p.
- DFFITS, the standardized signed measure of the difference of the fitted value $\widehat{\mu}_i$ using all the data and the value $\widehat{\mu}_{(-i)}$ omitting observation i.

EXAMPLE 11.17: Publications of PhD candidates

For models that inherit methods from the "glm" class (including NB models fit using `glm.nb()`), the simplest initial diagnostic plots are provided by the `plot()` method. Figure 11.33 shows the default *regression quartet* of plots for the negative-binomial model `phd.nbin` examined in earlier examples. By default, the `id.n=3` most noteworthy observations are labeled with their row names from the original data set.

```
> plot(phd.nbin)
```

The plot of residuals against predicted values in the upper left panel of Figure 11.33 should show no overall systematic trend for a well-fitting model. The smoothed loess curve in red suggests that this is not the case.

Several functions in the car package make these plots more flexibly and with greater control of the details. Figure 11.34 shows the plot of residuals against predicted values two ways. The right panel explains the peculiar pattern of diagonal band of points. These correspond to the different discrete values of the response variable, number of articles published.

```
> library(car)
> residualPlot(phd.nbin, type = "rstandard", col.smooth = "red", id.n = 3)
> residualPlot(phd.nbin, type = "rstandard",
+                groups = PhdPubs$articles, key = FALSE, linear = FALSE,
+                smoother = NULL)
```

Other useful plots show the residuals against each predictor. For a good-fitting model, the average residual should not vary systematically with the predictor. As shown in Figure 11.35, `residualPlot()` draws a lowess smooth, and also computes a curvature test for each of the plots by adding a quadratic term and testing the quadratic to be zero.

[18] Other types include raw response residuals (`type="response"`), working residuals (`type="working"`), and partial residuals (`type="partial"`).

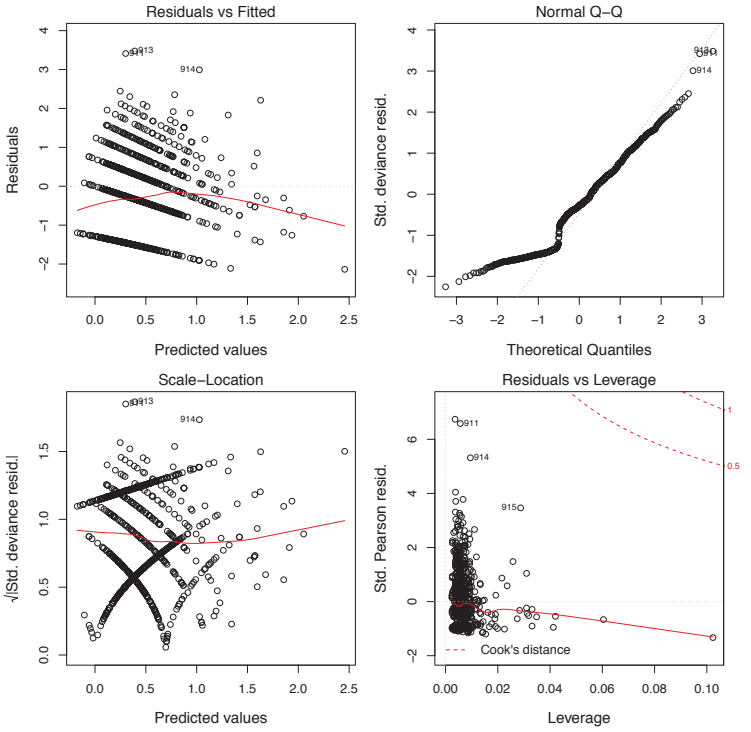


Figure 11.33: Default diagnostic plots for the negative-binomial model fit to the PhdPubs data.

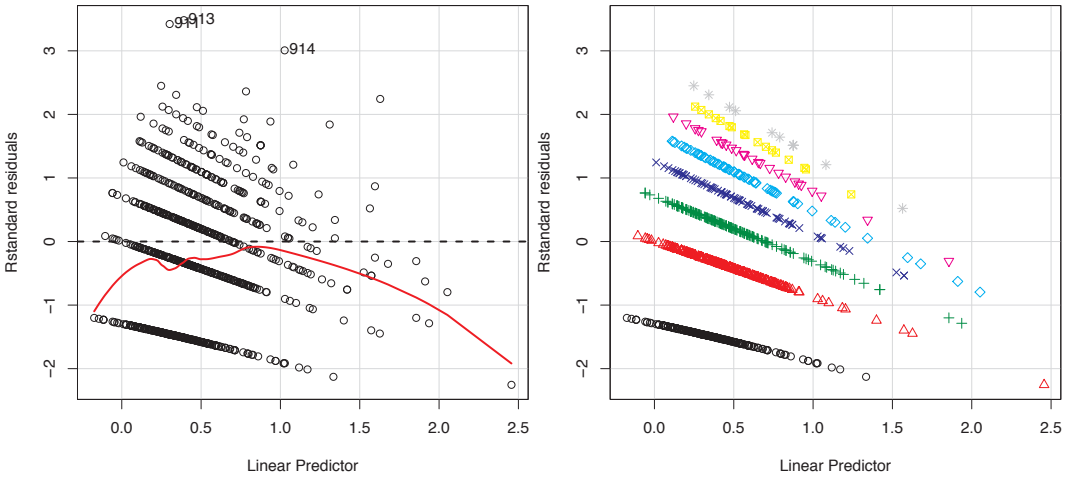


Figure 11.34: Plots of residuals against the linear predictor using residualPlot(). The right panel shows that the diagonal bands correspond to different values of the discrete response.

```
> residualPlot(phd.nbin, "mentor", type = "rstudent",
+                quadratic = TRUE, col.smooth = "red", col.quad = "blue",
+                id.n = 3)
> residualPlot(phd.nbin, "phdprestige", type = "rstudent",
+                quadratic = TRUE, col.smooth = "red", col.quad = "blue",
+                id.n = 3)
```

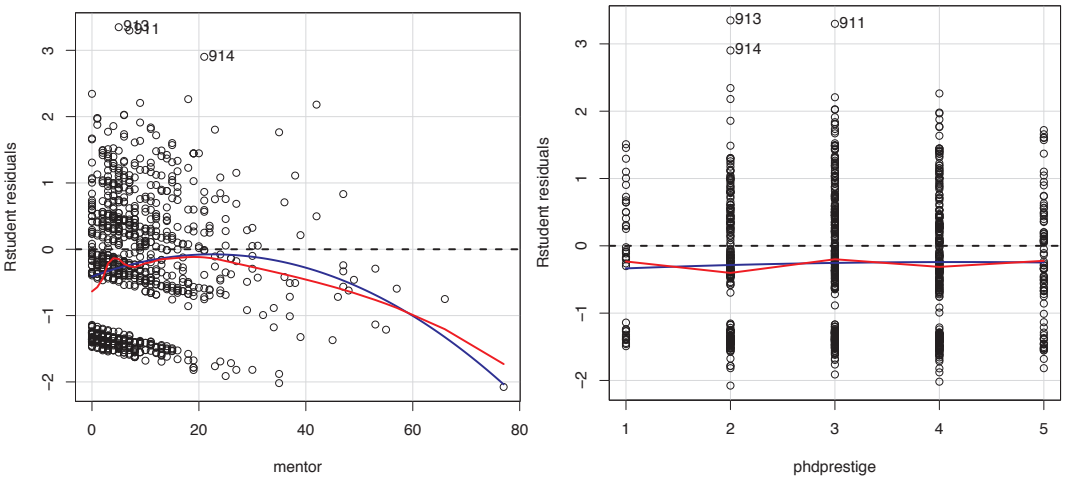


Figure 11.35: Plots of residuals against two predictors in the phd.nbin model. Such plots should show no evidence of a systematic trend for a good-fitting model.

In the plot at the left for number of articles by the student's mentor, the curvature is quite pronounced: at high values of `mentor`, nearly all of the residuals are negative, these students publishing fewer articles than would be expected. This would indicate a problem in the scale for `mentor` if there were more observations at the high end; but only about 1.5% points occur for `mentor>45`, so this can be discounted.

Figure 11.36 gives a better version of the influence plot shown in the lower right panel of Figure 11.33. This plots studentized (deletion) residuals against leverage, showing the value of Cook's distance by the area of the bubble symbol.

```
> influencePlot(phd.nbin)

     StudRes        Hat    CookD
328 -2.0762 0.1023449 0.18325
913  3.3488 0.0036473 0.16652
915  2.1810 0.0287496 0.24345
```

Several observations are considered noteworthy, because of one or more of large absolute residual, large leverage, or large Cook's distance. `influencePlot()` uses different default rules for point labeling than does the `plot()` method, but provides many options to control the details. Observation 328 stands out as having the largest leverage and a large negative residual; case 913 has the largest absolute residual, but is less influential than case 915.[19]

The `outlierTest()` function in car gives a formal test of significance of the largest absolute studentized residuals, with a Bonferroni-adjusted p-value accounting for choosing the largest values

[19]The higher case numbers appear in these plots and diagnostics because the data set *PhdPubs* had been sorted by the response, `articles`.

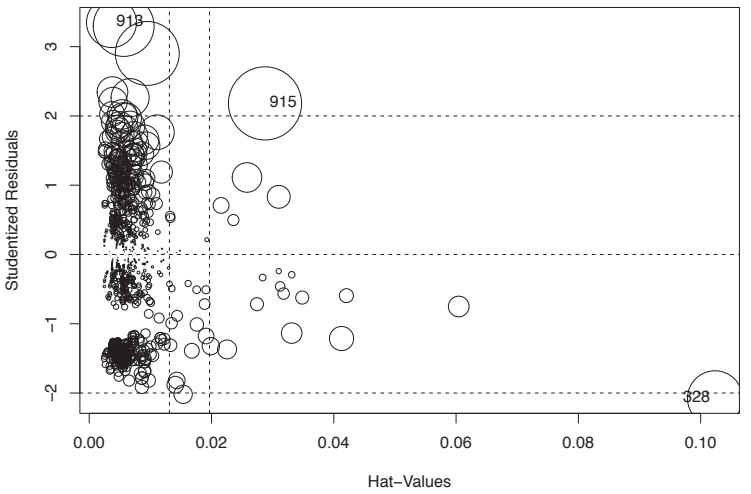


Figure 11.36: Influence plot showing leverage, studentized residuals, and Cook's distances for the negative-binomial model fit to the PhdPubs data. Conventional cutoffs for studentized residuals are shown by dashed horizontal lines at ± 2; vertical lines show 2 and 3 times the average hat-value.

among n such tests. Individually, case 913 is extreme, but it is not at all extreme among $n = 915$ such tests, each using $\alpha = .05$.

```
> outlierTest(phd.nbin)


No Studentized residuals with Bonferonni p < 0.05
Largest |rstudent|:
    rstudent unadjusted p-value Bonferonni p
913   3.3488         0.00084491      0.77309
```

This example started with the negative-binomial model, the best-fitting from the previous examples. It highlighted a few features of the data not seen previously and worth considering, but doesn't seriously challenge the substantive interpretation of the model. This is what we hope for from model diagnostic plots.

△

11.6.2 Quantile–quantile and half-normal plots

As we noted above, in theory the standardized and studentized Pearson and deviance residuals have approximate standard normal $\mathcal{N}(0, 1)$ distributions (in large samples) when the fitted model is correct. This suggests a plot of the sorted residuals, $r_{(i)}$, against the corresponding expected values, $z_{(i)}$, an equal-sized sample of size n would have in a normal distribution.[20]

If the distribution of the residuals is approximately normal, the points $(r_{(i)}, z_{(i)})$ should lie along a line with unit slope through the origin; systematic or individual departure from this line signals a potential violation of assumptions. The expected values are typically calculated as $z_{(i)} =$

[20]The subscripted notation $r_{(i)}$ (and $z_{(i)}$) denotes an *order statistic*, i.e., the i^{th} largest value in a set arranged in increasing order.

$\Phi^{-1}\{(i-\frac{3}{8})/(n+\frac{1}{4})\}$, where $\Phi^{-1}(\bullet)$ is the inverse normal, or normal quantile function, `qnorm()` in R.

Such plots, called ***normal quantile plots*** or ***normal QQ plots***, are commonly used for GLMs with a quantitative response variable. The upper right panel of Figure 11.33 illustrates the form of such plots produced by `plot()` for a "glm" object.

One difficulty with the default plots is that it is hard to tell to what extent the points deviate from the unit line because there is no visual reference for the line or envelope to indicate expected variability about that line. This problem is easily remedied using `qqPlot()` from car.

Figure 11.37 shows the result for the model `phd.nbin`. The envelope lines used here are at the quartiles of the expected normal distribution. They suggest a terrible fit, but, surprisingly, the largest three residuals are within the envelope.

```
> qqPlot(rstudent(phd.nbin), id.n = 3,
+        xlab = "Normal quantiles", ylab = "Studentized residuals")

913 911 914
915 914 913
```

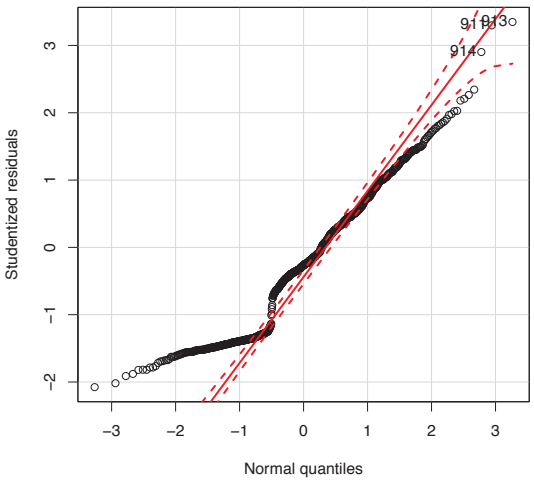


Figure 11.37: Normal QQ plot of the studentized residuals from the NB model for the PhdPubs data. The normal-theory reference line and confidence envelope are misleading here.

For GLMs with discrete responses, such plots are often disappointing, even with a reasonably good-fitting model, because: (a) possible outliers can appear at both the lower and upper ends of the distribution of residuals; (b) the theoretical normal distribution used to derive the envelope may not be well approximated in a given model.

Atkinson (1981, 1987) suggested a more robust and useful version of these QQ plots: half normal plots, with simulated confidence envelopes. The essential ideas are:

- Model departures and outliers are often easier to see for discrete data when the *absolute values* of residuals are plotted, because large positive and negative values are sorted together. This gives the ***half-normal plot***, in which the absolute values of residuals, arranged in increasing order, $|r|_{(i)}$, are plotted against $|z|_{(i)} = \Phi^{-1}\{(n+i-\frac{1}{8})/(2n+\frac{1}{2})\}$. All outliers will then appear in the upper right of such a plot, as points separated from the trend of the remaining cells.

- The normal-theory reference line, $|r|_{(i)} = |z|_{(i)}$ and the normal-theory confidence envelope can be replaced by simulating residuals from the assumed distribution, that need not be normal. The reference line is taken as the mean of S simulations and the envelope with $1 - \alpha$ coverage is taken as the $(\alpha/2, 1 - \alpha/2)$ quantiles of their values.
- Specifically, for a GLM, S sets of random observations $\boldsymbol{y}_j, j = 1, 2, \ldots S$ are generated from the fitted model, each with mean $\widehat{\boldsymbol{\mu}}$, the fitted values under the model and with the *same* distribution. In R, this is readily accomplished using the generic `simulate()` function; the random variation around $\widehat{\mu}$ uses `rnorm()`, `rpois()`, `rnegbin()`, etc., as appropriate for the family of the model.
- The same model is then fit to each simulated $\boldsymbol{y}_j$, giving a new set of residuals for each simulation. Sorting their absolute values then gives the simulation distribution used as reference for the observed residuals.

At the time of writing there is no fully general implementation of these plots in R, but the technique is not too difficult and is sufficiently useful to illustrate here.

EXAMPLE 11.18: Publications of PhD candidates

First, calculate the sorted absolute values of the residuals $|r|_{(i)}$ and their expected normal values, $|z|_{(i)}$. The basic plot will be `plot(expected, observed)`.

```
> observed <- sort(abs(rstudent(phd.nbin)))
> n <- length(observed)
> expected <- qnorm((1:n + n - 1/8) / (2*n + 1/2))
```

Then, use `simulate()` to generate $S = 100$ simulated response vectors around the fitted values in the model. Here this uses the negative-binomial random number generator (`rnegbin()`) with the same dispersion value ($\widehat{\theta} = 2.267$) estimated in the model. The result, called `sims` here, is a data frame of $n = 915$ rows and $S = 100$ columns, named `sim_1, sim_2, ...`.

```
> S <- 100
> sims <- simulate(phd.nbin, nsim = S)
> simdat <- cbind(PhdPubs, sims)
```

The next step is computationally intensive, because we have to fit the NB model $S = 100$ times, and a little bit tricky, because we need to use the same model formula as the original, but with the simulated $\boldsymbol{y}$. We first define a function `resids` to do this for a given `y`, and then use a loop to calculate them all. To save computing time, the coefficients from the `phd.nbin` model are used as starting values.

```
> # calculate residuals for one simulated data set
> resids <- function(y)
+   rstudent(glm.nb(y ~ female + married + kid5 + phdprestige + mentor,
+                   data=simdat, start=coef(phd.nbin)))
> # do them all ...
> simres <- matrix(0, nrow(simdat), S)
> for(i in 1:S) {
+         simres[,i] <- sort(abs(resids(dat[,paste("sim", i, sep="_")])))
+ }
```

We can then use `apply()` to compute the summary measures defining the center and limits for the simulated confidence interval.

```
> envelope <- 0.95
> mean <- apply(simres, 1, mean)
> lower <- apply(simres, 1, quantile, prob = (1 - envelope) / 2)
> upper <- apply(simres, 1, quantile, prob = (1 + envelope) / 2)
```

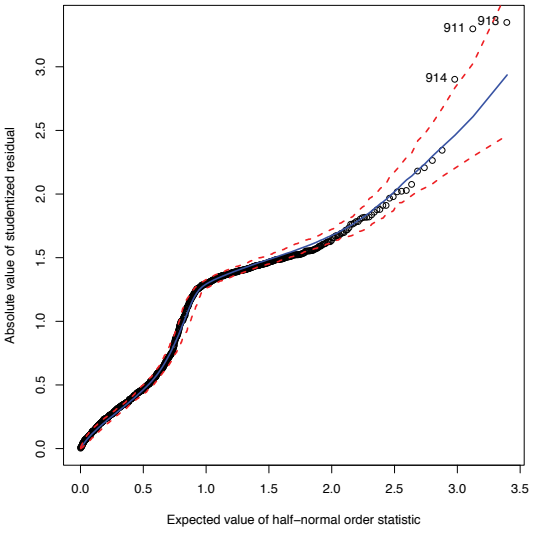


Figure 11.38: Half-normal QQ plot of studentized residuals for the NB model fit to the PhdPubs data. The reference line and confidence envelope reflect the mean and (2.5%, 97.5%) quantiles of the simulation distribution under the negative-binomial model for the same data.

Finally, plot the observed against expected absolute residuals as points, and add the lines for the confidence envelope, producing Figure 11.38.

```
> plot(expected, observed,
+      xlab = "Expected value of half-normal order statistic",
+      ylab = "Absolute value of studentized residual")
> lines(expected, mean, lty = 1, lwd = 2, col = "blue")
> lines(expected, lower, lty = 2, lwd = 2, col = "red")
> lines(expected, upper, lty = 2, lwd = 2, col = "red")
> identify(expected, observed, labels = names(observed), n = 3)
```

The shape of the QQ plot in Figure 11.37 shows a peculiar bend at low values and the half-normal version in Figure 11.38 has a peculiar hump in the middle. What could be the cause?

Figure 11.39 shows two additional plots of the studentized residuals that give a clear answer. The density plot at the left shows a strongly bimodal distribution of the residuals. An additional plot at the right of residuals against the log(response) confirms the guess that the lower mode corresponds to those students who published no articles—excess zeros again!

```
> # examine distribution of residuals
> res <- rstudent(phd.nbin)
> plot(density(res), lwd = 2, col = "blue",
+      main = "Density of studentized residuals")
> rug(res)
>
> # why the bimodality?
> plot(jitter(log(PhdPubs$articles + 1), factor = 1.5), res,
+      xlab = "log (articles + 1)", ylab = "Studentized residual")
```

Now we have something to worry about that *could* affect substantive interpretation or conclusions from this analysis using the NB model, but not accounting for excess zeros. If we believe, following Long (1997), that there is a separate latent class of students who don't publish, it would be sensible to fit a zero-inflated NB model, perhaps with a different subset of predictors for the

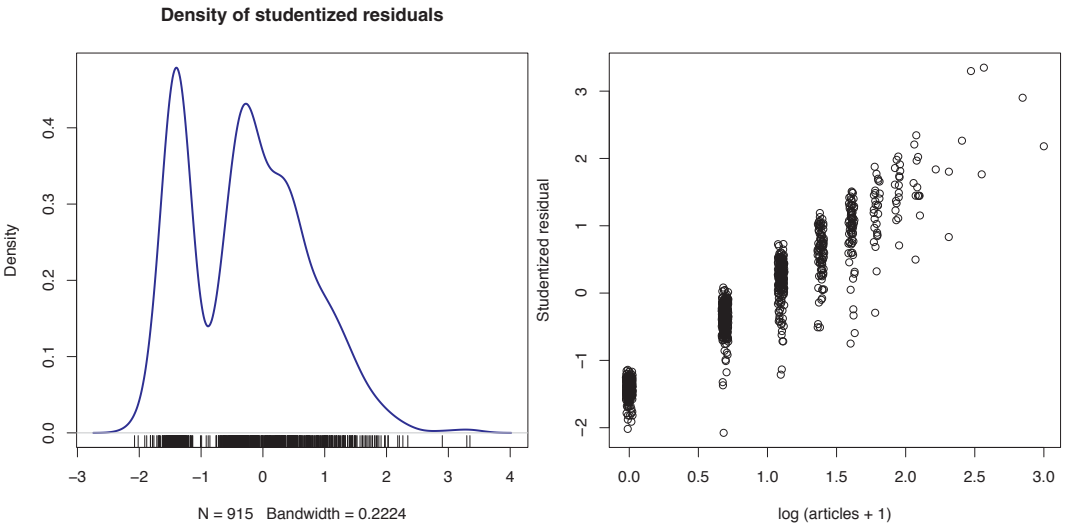


Figure 11.39: Further plots of studentized residuals. Left: density plot; right: residuals against log(articles+1).

zero component. The alternative theory of a "hurdle" to a first publication suggests fitting a hurdle model. We leave these as exercises for the reader.

△

11.7 Multivariate response GLM models*

> Far better an approximate answer to the right question, which is often vague, than an exact answer to the wrong question, which can always be made precise.
>
> John W. Tukey (1962), *The future of data analysis*

As noted in Section 10.4, in many studies, there may be several response variables along with one or more explanatory variables, and it is useful to try to model some properties of their joint distribution as well as their separate dependence on the predictors. In the current chapter, the case study (Section 11.5.2) of demand for medical care by the elderly provides a relevant example. There are actually four indicators of medical care, a 2×2 set of (office vs. hospital) place and (physician vs. non-physician) practitioner. That case study analyzed only the office visits by physicians.

This section describes a few steps in this direction. To provide some context, we begin with a capsule overview of classical multivariate response models.

In the case of classical linear models with Gaussian error distributions, the model for a univariate response, $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$, with $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma})$, extends quite readily to the ***multivariate linear model*** (MLM) for q response variables, $\boldsymbol{Y} = \{\boldsymbol{y}_1, \boldsymbol{y}_2, \ldots, \boldsymbol{y}_q\}$. The MLM has the form

$$\underset{(n\times q)}{\boldsymbol{Y}} = \underset{(n\times p)}{\boldsymbol{X}}\,\underset{(p\times q)}{\boldsymbol{B}} + \underset{(n\times q)}{\boldsymbol{E}} \tag{11.16}$$

where $\boldsymbol{Y}$ is a matrix of n observations on q response variables; $\boldsymbol{X}$ is a model matrix with columns for p regressors, typically including an initial column of 1s for the regression constant; $\boldsymbol{B}$ is a matrix of regression coefficients, one column for each response variable; and $\boldsymbol{E}$ is a matrix of errors.

It is important to note that:

- The maximum likelihood estimator of $\boldsymbol{B}$ in the MLM is equivalent to the result of fitting q separate univariate models for the individual responses and joining the coefficients columwise, giving
$$\widehat{\boldsymbol{B}} = \{\widehat{\boldsymbol{\beta}}_1, \widehat{\boldsymbol{\beta}}_2, \ldots, \widehat{\boldsymbol{\beta}}_q\} = (\boldsymbol{X}^\mathsf{T}\boldsymbol{X})^{-1}\boldsymbol{X}^\mathsf{T}\boldsymbol{Y} \; .$$

- Procedures for statistical inference (hypothesis tests, confidence intervals), however, take account of the correlations among the responses. Multivariate tests can therefore be more powerful than separate univariate tests under some conditions.
- A unique feature of the MLM stems from the assumption of multivariate normality of the errors, so that each row ϵ_i^T of $\boldsymbol{E}$ is assumed to be distributed independently, $\epsilon_i^\mathsf{T} \sim \mathcal{N}_q(\boldsymbol{0}, \boldsymbol{\Sigma})$, where $\boldsymbol{\Sigma}_{q\times q}$ is the error covariance matrix, constant across observations, like σ^2 in univariate models. Then, the conditional distributions of $\boldsymbol{y}_j \mid \boldsymbol{X}$ are all univariate normal, all bivariate distributions, $\boldsymbol{y}_j, \boldsymbol{y}_k \mid \boldsymbol{X}$ are bivariate normal, and any linear combination of the conditional ys is univariate normal.
- Consequently, all relationships among the $\boldsymbol{y}$s can be summarized by correlations and relationships between the $\boldsymbol{y}$s and $\boldsymbol{x}$s by linear regressions. These can be visualized using ***data ellipse***s (Friendly et al., 2013) and hypothesis tests in the MLM can be visualized by ellipses using ***hypothesis-error plot***s (Friendly, 2007, Fox et al., 2009).

This generality of the MLM is lost, however, when we move to multivariate response models in the non-Gaussian case. For binomial responses, Section 10.4 described several approaches toward a multivariate logistic regression model that attempt to separate the marginal dependence of each $\boldsymbol{y}$ on the $\boldsymbol{x}$s from the relationship of the association among the $\boldsymbol{y}$s on the $\boldsymbol{x}$s. The bivariate logistic model for $(\boldsymbol{y}_1, \boldsymbol{y}_2)$, for example, was parameterized (see Eqn. (10.15)) in terms of submodels for a logit for each response, $\eta_1 = \boldsymbol{x}^\mathsf{T}\boldsymbol{\beta}_1$, $\eta_2 = \boldsymbol{x}^\mathsf{T}\boldsymbol{\beta}_2$ and a submodel for the log odds ratio, $\theta_{12} = \boldsymbol{x}^\mathsf{T}\boldsymbol{\beta}_{12}$.

The situation becomes more difficult for multivariate count data responses, because parametric approaches to their joint distribution (e.g., a multivariate Poisson distribution) given a set of explanatory variables are computationally and analytically intractable. Cameron and Trivedi (2013, Chapter 8) provide a detailed description of the problems and some solutions for the bivariate case, including bivariate Poisson, negative-binomial and hurdle models.

Consequently, only a few special cases have been worked out theoretically, and mostly for the bivariate case. For example, King (1989) described a seemingly unrelated bivariate Poisson model for two correlated count variables. This models the separate linear predictors for $\boldsymbol{y}_1$ and $\boldsymbol{y}_2$ as

$$\begin{aligned} g(\boldsymbol{\mu}_1) &= \boldsymbol{x}_1^\mathsf{T}\boldsymbol{\beta}_1 \\ g(\boldsymbol{\mu}_2) &= \boldsymbol{x}_2^\mathsf{T}\boldsymbol{\beta}_2 \; , \end{aligned}$$

with the covariance between $\boldsymbol{y}_1$ and $\boldsymbol{y}_2$ represented as ξ. As in the MLM, the coefficients have the same point estimates as in equation-by-equation Poisson models. However, there is a gain in efficiency (reduced standard errors) resulting from a bivariate full-information maximum likelihood solution, and efficiency increases with the covariance ξ between the two count variables.

As a result, for lack of a fully general model for multivariate count data, one simple approach is to employ a method for simultaneous estimation of the equation-by-equation coefficients, accepting some loss of efficiency. This allows for hypothesis tests that may not be the most powerful, but provide approximate answers to more interesting questions. We can supplement this with separate analysis of the dependencies among the responses, and how these vary with the explanatory variables.

In R, the VGAM package is the most general available package for analysis of multivariate response GLMs. For multivariate count data, it provides for both Poisson and negative-binomial models. For NB models, the dispersion parameters $\theta_j = \alpha_j^{-1}$ can be allowed to vary with the predictors via a GLM of the form $\log\theta_j = \boldsymbol{x}^\mathsf{T}\boldsymbol{\gamma}_j$ or can be constrained to be "intercept-only,"

$\log \theta_j = \gamma_{0j}$, giving separate global dispersion estimates for each response. In the latter case, the resulting coefficients are the same as fitting a separate model for each response using `glm.nb()`.

EXAMPLE 11.19: Demand for medical care

In the examples in Section 11.5.2 we considered a variety of models for the number of office visits to physicians (`visits`) as the primary outcome variable in the study of demand for medical care by the elderly. We noted that other indicators of demand included office visits to non-physicians and hospital visits to both physicians and non-physicians. A more complete analysis of this data would consider all four response indicators together.

A special feature of this example is that the four response variables constitute a 2×2 set of the combinations of *place of visit* (office vs. hospital) and (physician vs. non-physician) *practitioner*. These are all counts, and could be transformed to two binary responses according to place and practitioner. Instead, we treat them individually here.

We start by selecting the variables to consider from the *NMES1988* data, giving a new working data set `nmes2`.

```
> data("NMES1988", package = "AER")
> nmes2 <- NMES1988[, c(1 : 4, 6 : 8, 13, 15, 18)]
> names(nmes2)[1 : 4]     # responses

[1] "visits"   "nvisits"  "ovisits"  "novisits"

> names(nmes2)[-(1 : 4)]  # predictors

[1] "hospital"  "health"    "chronic"   "gender"    "school"
[6] "insurance"
```

11.7.1 Analyzing correlations: HE plots

For purely descriptive purposes, a useful starting point is often an analysis of the $\log(\boldsymbol{y})$ on the predictor variables using the classical MLM, a rough analog of a multivariate Poisson regression with a log link. Inferential statistics will be biased, but we can use the result to visualize the pairwise linear relations that exist among all responses and all predictors compactly using hypothesis-error (HE) plots (Friendly, 2007).

Zero counts cause problems because the log of zero is undefined, so we add 1 to each y_{ij} in the call to `lm()`. The result is an object of class "mlm".

```
> clog <- function(x) log(x + 1)
> nmes.mlm <- lm(clog(cbind(visits, nvisits, ovisits, novisits)) ~ .,
+                data = nmes2)
```

An HE plot provides a visualization of the covariances of effects for the linear hypothesis (H) for each term in an MLM in relation to error covariances (E) using data ellipsoids in the space of dimension q, the number of response variables. The size of each H ellipsoid in relation to the E ellipsoid indicates the strength of the linear relations between the responses and the individual predictors.[21] The orientation of each H ellipsoid shows the direction of the correlations for that term with the response variables. For 1 degree of freedom terms (a covariate or factor with two levels), the corresponding H ellipsoid collapses to a line.

The heplots (Fox and Friendly, 2014) package contains functions for 2D plots (`heplot()`) of pairs of y variables, 3D plots (`heplot3d()`), and all pairwise plots (`pairs()`). We illustrate this here using `pairs()` for the MLM model, giving the plot shown in Figure 11.40.

[21] When the errors, $\boldsymbol{\mathcal{E}}$ in Eqn. (11.16), are approximately multivariate normal, the H ellipsoid provides a visual test of significance: the H ellipsoid projects outside the E ellipsoid *if and only if* Roy's test is significant at a chosen α level.

```
> library(heplots)
> vlabels <- c("Physician\noffice visits",
+              "Non-physician\n office visits",
+              "Physician\nhospital visits",
+              "Non-physician\nhospital visits")
> pairs(nmes.mlm, factor.means = "health",
+       fill = TRUE, var.labels = vlabels)
```

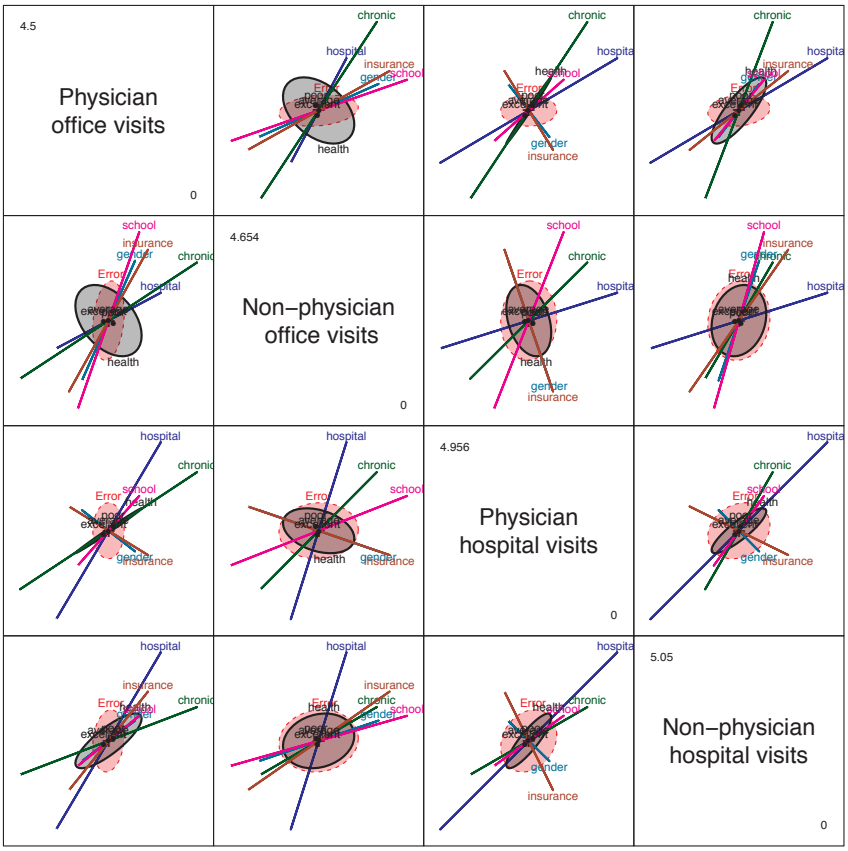


Figure 11.40: Pairwise HE plots for all responses in the nmes2 data.

The top row in Figure 11.40 shows the relationship of physician office visits to the other types of medical services. It can be seen that chronic conditions and hospital stays are positively correlated with both responses, as they also are in all other pairwise plots. Having private health insurance is positively related to some of these outcomes, and negatively to others. Except for difficulties with overlapping labels and the obvious violation of statistical assumptions of the MLM here, such plots give reasonably useful overviews on the relationships among the y and x variables.

11.7.2 Analyzing associations: Odds ratios and fourfold plots

In the analysis below, we first attempt to understand the association among these response variables and how these associations relate to the explanatory variables. It is natural to think of this in terms of the (log) odds ratio of a visit to a physician vs. a non-physician, given that the place is in an office as opposed to a hospital. Following this, we consider some multivariate negative binomial models relating these counts to the explanatory variables.

In order to treat the four response variables as a single response (`visit`), distinguished by `type`, it is necessary to reshape the data from a wide format to a long format with four rows for each input observation.

```
> vars <- colnames(nmes2)[1 : 4]
> nmes.long <- reshape(nmes2,
+   varying = vars,
+   v.names = "visit",
+   timevar = "type",
+   times = vars,
+   direction = "long",
+   new.row.names = 1 : (4 * nrow(nmes2)))
```

Then, the `type` variable can be used to create two new variables, `practitioner` and `place`, corresponding to the distinctions among visits. While we are at it, we create factors for two of the predictors.

```
> nmes.long <- nmes.long[order(nmes.long$id),]
> nmes.long <- transform(nmes.long,
+   practitioner = ifelse(type %in% c("visits", "ovisits"),
+                         "physician", "nonphysician"),
+   place = ifelse(type %in% c("visits", "nvisits"), "office", "hospital"),
+   hospf = cutfac(hospital, c(0 : 2, 8)),
+   chronicf = cutfac(chronic))
```

Then, we can use `xtabs()` to create a frequency table of `practitioner` and `place` classified by any one or more of these factors. For example, the total number of visits of the four types is given by

```
> xtabs(visit ~ practitioner + place, data = nmes.long)

              place
practitioner   hospital office
  nonphysician     2362   7129
  physician        3308  25442
```

From this, we can calculate the odds ratio and visualize the association with a fourfold or mosaic plot. More generally, by including more factors in the call to `xtabs()`, we can calculate and visualize how the *conditional* association varies with these factors. For example, Figure 11.41 shows fourfold plots conditioned by health status. It can be seen that there is a strong positive association, except for those with excellent health: people are more likely to see a physician in an office visit, and a non-physician in a hopsiptal visit. The corresponding log odds ratios are shown numerically using `loddsratio()`.

```
> library(vcdExtra)
> fourfold(xtabs(visit ~ practitioner + place + health, data = nmes.long),
+          mfrow=c(1,3))
> loddsratio(xtabs(visit ~ practitioner + place + health,
+                  data = nmes.long))

log odds ratios for practitioner and place by health

     poor   average excellent
 1.140166  0.972777  0.032266
```

Going further, we can condition by more factors. Figure 11.42 shows the fourfold plots conditioned by the number of chronic conditions (in the rows) and the combinations of gender and private insurance (columns).

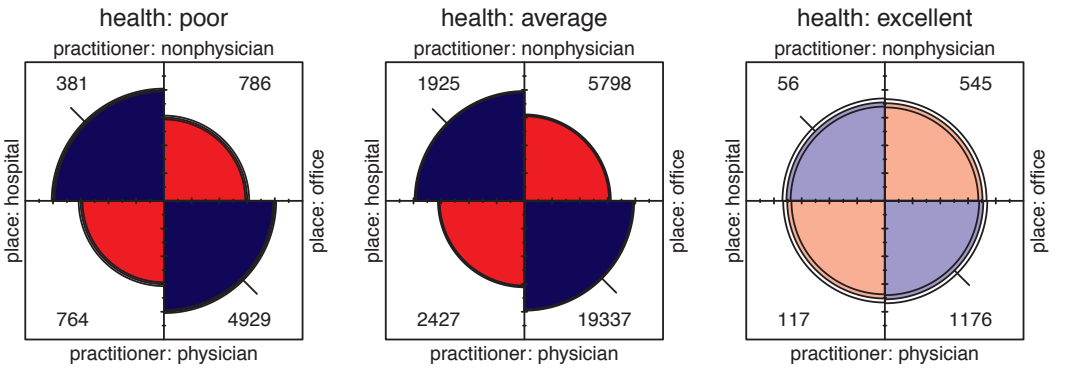


Figure 11.41: Fourfold displays for the association between practitioner and place in the nmes.long data, conditioned on health status.

```
> tab <- xtabs(visit ~ practitioner + place + gender +
+                       insurance + chronicf,
+              data = nmes.long)
> fourfold(tab, mfcol=c(4,4), varnames=FALSE)
```

The systematic patterns seen here are worth exploring further by graphing the log odds ratios directly. The call `as.data.frame(loddsratio(tab))` converts the result of `loddsratio(tab)` to a data frame with factors for these variables, and variables `LOR` and `ASE` containing the estimated log odds ratio ($\widehat{\theta}$) and its asymptotic standard error ($\text{ASE}(\widehat{\theta})$). Figure 11.43 shows the plot of these values as line graphs with associated ± 1 error bars produced using ggplot2.[22]

```
> lodds.df <- as.data.frame(loddsratio(tab))
> library(ggplot2)
> ggplot(lodds.df, aes(x = chronicf, y = LOR,
+                      ymin = LOR - 1.96 * ASE, ymax = LOR + 1.96 * ASE,
+                      group = insurance, color = insurance)) +
+   geom_line(size = 1.2) + geom_point(size = 3) +
+   geom_linerange(size = 1.2) +
+   geom_errorbar(width = 0.2) +
+   geom_hline(yintercept = 0, linetype = "longdash") +
+   geom_hline(yintercept = mean(lodds.df$LOR), linetype = "dotdash") +
+   facet_grid(. ~ gender, labeller = label_both) +
+   labs(x = "Number of chronic conditions",
+        y = "log odds ratio (physician|place)") +
+   theme_bw() + theme(legend.position = c(0.1, 0.9))
```

It can be seen that for those with private insurance, the log odds ratios are uniformly positive, but males and females exhibit a somewhat different pattern over number of chronic conditions. Among those with no private insurance, the log odds ratios generally increase over number of chronic conditions, except for females with 3 or more such conditions.

Beyond this descriptive analysis, you can test hypotheses about the effects of the predictors on the log odds ratios using a simple ANOVA model. Under the null hypothesis, $H_0 : \theta_{ijk\ldots} = 0$, the $\widehat{\theta}$ are each distributed normally, $\mathcal{N}(0, \text{ASE}(\widehat{\theta}))$, so a weighted ANOVA can be used to test for differences according to the predictors. This analysis gives the results below.

[22] A similar plot can be obtained using `cotabplot(~practitioner + place + chronicf + insurance | gender, tab, panel = cotab_loddsratio)`.

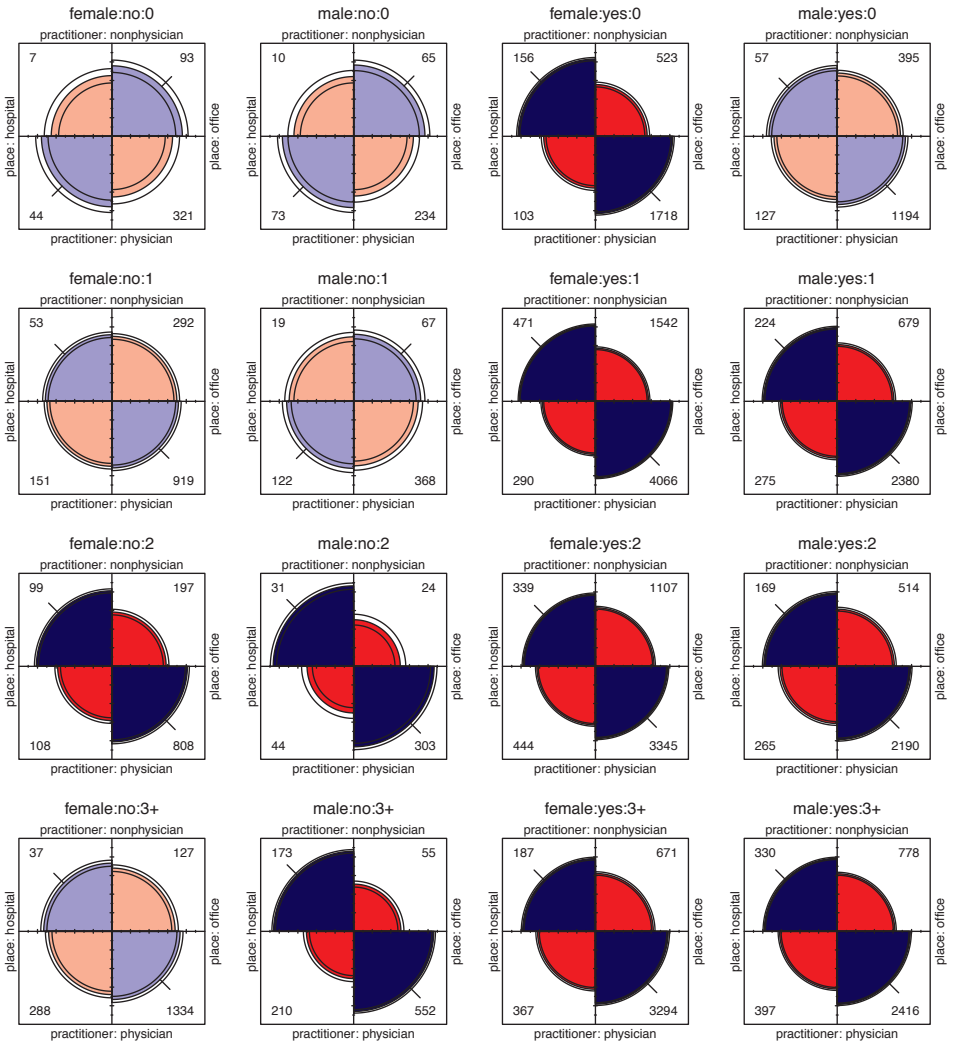


Figure 11.42: Fourfold displays for the association between practitioner and place in the nmes.long data, conditioned on gender, insurance, and number of chronic conditions. Rows are levels of chronic; columns are the combinations of gender and insurance.

```
> lodds.mod <- lm(LOR ~ (gender + insurance + chronicf)^2,
+                 weights = 1 / ASE^2, data = lodds.df)
> anova(lodds.mod)

Analysis of Variance Table

Response: LOR
                   Df Sum Sq Mean Sq F value Pr(>F)
gender              1    0.8     0.8    0.17  0.707
insurance           1    5.3     5.3    1.17  0.358
chronicf            3    4.6     1.5    0.34  0.802
gender:insurance    1   32.5    32.5    7.20  0.075 .
gender:chronicf     3   54.1    18.0    3.99  0.143
insurance:chronicf  3  114.1    38.0    8.43  0.057 .
Residuals           3   13.5     4.5
---
```

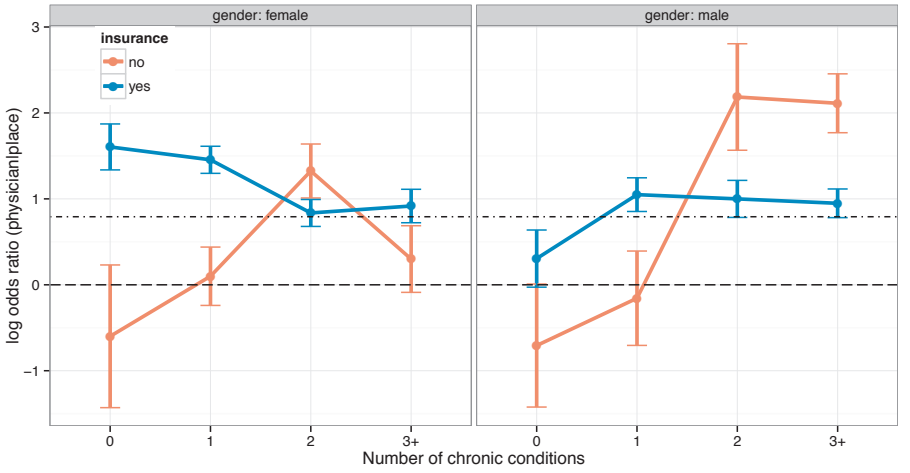


Figure 11.43: Plot of log odds ratios with 1 standard error bars for the association between practitioner and place, conditioned on gender, insurance, and number of chronic conditions. The horizontal lines show the null model (longdash) and the mean (dot–dash) of the log odds ratios.

```
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

As might be expected from the graph in Figure 11.43, having private insurance is a primary determinant of the decision to seek an office visit with a physician, but this effect interacts slightly according to number of chronic conditions and gender.

△

11.7.2.1 Fitting and testing multivariate count data models

With a multivariate response, `vglm()` in the VGAM package estimates the separate coefficients for each response jointly. A special feature of this formulation is that constraints can be imposed to force the coefficients for a given term in a model to to be the same for all responses. A likelihood-ratio test against the unconstrained model can then be used to test for differences in the effects of predictors across the response variables.

This is achieved by formulating the linear predictor as a sum of terms,

$$\eta(\boldsymbol{x}) = \sum_{k=1}^{p} \boldsymbol{H}_k \boldsymbol{\beta}_k \boldsymbol{x}_k \; ,$$

where $\boldsymbol{H}_1, \ldots, \boldsymbol{H}_p$ are *known* full-rank constraint matrices. With no constraints, the $\boldsymbol{H}_k$ are identity matrices $\boldsymbol{I}_q$ for all terms. With `vglm()`, the constraint matrices for a given model are returned using `constraints()`, and can be set for a new, restricted model using the `constraints` argument. To constrain the coefficients for a term k to be equal for all responses, use $\boldsymbol{H}_k = \mathbf{1}_q$, a unit vector.

More general Wald tests of hypotheses can be carried out without refitting using `linearHypothesis()` in the car package. These include (a) joint tests that a subset of predictors for a given response have null effects; (b) across-response tests of equality of coefficients for one or more model terms.

EXAMPLE 11.20: Demand for medical care

In the examples in Section 11.5.2, we described a series of increasingly complex models for

physician office visits, including interactions and nonlinear terms. The multivariate case is computationally more intensive, and estimation can break down in complex models. We can illustrate the main ideas here using the multivariate analog of the simple main effects model discussed in Example 11.14.

Using `vglm()`, the response variables are specified as the matrix form $\boldsymbol{Y}$ using `cbind()` on the left-hand side of the model formula. The right-hand side, `~ .` here specifies all other variables as predictors. `family = negbinomial` uses the NB model for each $\boldsymbol{y}_j$, with an intercept-only model for the dispersion parameters by default.

```
> nmes2.nbin <- vglm(cbind(visits, nvisits, ovisits, novisits) ~ .,
+                    data = nmes2, family = negbinomial)
```

The estimated parameters from this model are returned by the `coef()` method as pairs of columns labeled `log(mu)`, `logsize` for each response. For example, the parameters for the `visits` response are in the first two columns, and are the same as those estimated for the model `nmes.nbin` using `glm.nb()`.

```
> # coefficients for visits
> coef(nmes2.nbin, matrix = TRUE)[,c(1, 2)]

                loge(mu1) loge(size1)
(Intercept)      0.929257     0.18781
hospital         0.217772     0.00000
healthpoor       0.305013     0.00000
healthexcellent -0.341807     0.00000
chronic          0.174916     0.00000
gendermale      -0.126488     0.00000
school           0.026815     0.00000
insuranceyes     0.224402     0.00000

> # theta for visits
> exp(coef(nmes2.nbin, matrix = TRUE)[1, 2])

[1] 1.2066
```

The `log(mu)` coefficients for all four response variables are shown below.

```
> coef(nmes2.nbin, matrix = TRUE)[,c(1, 3, 5, 7)]

                loge(mu1) loge(mu2) loge(mu3) loge(mu4)
(Intercept)      0.929257 -0.747798  -1.11284 -1.341793
hospital         0.217772  0.144645   0.41506  0.483883
healthpoor       0.305013 -0.179822   0.16491  0.033509
healthexcellent -0.341807 -0.038121  -0.42449 -1.006527
chronic          0.174916  0.093430   0.27664  0.243570
gendermale      -0.126488 -0.255508   0.33456  0.052044
school           0.026815  0.068269   0.03559 -0.027475
insuranceyes     0.224402  0.492793  -0.53105  0.484062
```

We notice that the coefficients for `hospital` and `chronic` have values with the same signs for all four responses. If it is desired to test the hypothesis that their coefficients are all the same for each of these predictors, first extract the $\boldsymbol{H}$ matrices for the unconstrained model using `constraints()`.

```
> clist <- constraints(nmes2.nbin, type = "term")
> clist$hospital[c(1, 3, 5, 7),]

     [,1] [,2] [,3] [,4]
[1,]    1    0    0    0
```

```
[2,]    0    1    0    0
[3,]    0    0    1    0
[4,]    0    0    0    1
```

Then, reset the constraints for these terms to be unit vectors, forcing them to be all equal.

```
> clist2 <- clist
> clist2$hospital <- cbind(rowSums(clist$hospital))
> clist2$chronic  <- cbind(rowSums(clist$chronic))
> clist2$hospital[c(1, 3, 5, 7), 1, drop = FALSE]

     [,1]
[1,]    1
[2,]    1
[3,]    1
[4,]    1
```

Now, fit the same model as before, but using the constraints in `clist2`.

```
> nmes2.nbin2 <- vglm(cbind(visits, nvisits, ovisits, novisits) ~ .,
+                     data = nmes2, constraints = clist2,
+                     family = negbinomial(zero = NULL))
```

The coefficients for the constrained model are shown below. As you can see, the coefficients for `hospital` and `chronic` have the same estimates for all four responses.

```
> coef(nmes2.nbin2, matrix = TRUE)[,c(1, 3, 5, 7)]

                loge(mu1) loge(mu2) loge(mu3) loge(mu4)
(Intercept)      0.918002 -0.835090 -0.864251 -1.175650
hospital         0.244655  0.244655  0.244655  0.244655
healthpoor       0.293334 -0.315479  0.366404  0.172403
healthexcellent -0.334959  0.047830 -0.538294 -1.044784
chronic          0.178563  0.178563  0.178563  0.178563
gendermale      -0.127956 -0.272264  0.330587  0.071228
school           0.026829  0.064401  0.031764 -0.028986
insuranceyes     0.222531  0.477475 -0.529976  0.507129
```

A likelihood-ratio test prefers the reduced model with equal coefficients for these two predictors. The degrees of freedom for this test (6) is the number of constrained parameters in the smaller model.

```
> lrtest(nmes2.nbin, nmes2.nbin2)

Likelihood ratio test

Model 1: cbind(visits, nvisits, ovisits, novisits) ~ .
Model 2: cbind(visits, nvisits, ovisits, novisits) ~ .
    #Df LogLik Df Chisq Pr(>Chisq)
1 35212 -25394
2 35218 -25413  6  39.2    6.4e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Alternatively, these tests can be performed as tests of linear hypotheses (see Section 11.1.2) on the coefficients $\boldsymbol{B}$ from the original model without refitting. Using `linearHypothesis()`, a hypothesis matrix $\boldsymbol{L}$ specifying equality of the coefficients for a given predictor can be easily generated using a character vector of the coefficient names.

```
> lh <- paste("hospital:", 1 : 3, " = ", "hospital:", 2 : 4, sep="")
> lh

[1] "hospital:1 = hospital:2" "hospital:2 = hospital:3"
[3] "hospital:3 = hospital:4"
```

Using `lh` as the `linear.hypothesis` argument then gives the following result for the coefficients of `hospital`, rejecting the hypothesis that they are all equal across response variables.

```
> car::linearHypothesis(nmes2.nbin, lh)

Linear hypothesis test

Hypothesis:
hospital:1 - hospital:2 = 0
hospital:2 - hospital:3 = 0
hospital:3 - hospital:4 = 0

Model 1: restricted model
Model 2: cbind(visits, nvisits, ovisits, novisits) ~ .

  Res.Df Df Chisq Pr(>Chisq)
1  35215
2  35212  3  26.4    7.8e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

△

To pursue this analysis further, you could investigate whether any interactions of these effects were interesting and important as in Example 11.14, but now for the multivariate response variables.

To interpret a given model visually, you could use effect plots for the terms predicting each of the responses, as in Example 11.15. The effects package cannot handle models fit with VGAM directly, but you can use `glm()` or `glm.nb()` to fit the equivalent submodels for each response separately, and then use the `plot(Effect())` methods to display the effects for interesting terms. Figure 11.44 shows one such plot, for the effects of health status on each of the four response variables.

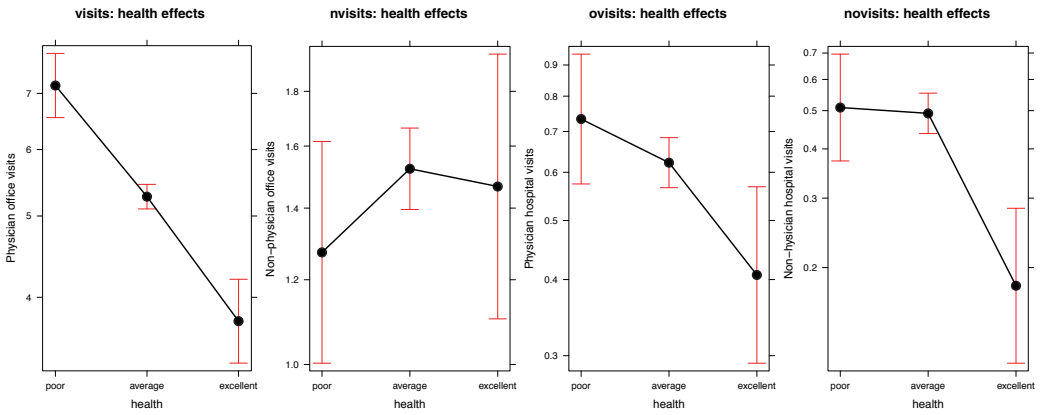


Figure 11.44: Effect plots for the effects of health status on the four response variables in the nmes2 data.

11.8 Chapter summary

- The generalized linear model extends the familiar classical linear models for regression and ANOVA to encompass models for discrete responses and continuous responses for which the assumption of normality of errors is untenable.
- It does this by retaining the idea of a *linear predictor*—a linear function of the regressors, $\eta_i = \boldsymbol{x}_i^\mathsf{T}\boldsymbol{\beta}$, but then allowing:
  - a *link function*, $g(\bullet)$, connecting the linear predictor η_i to the mean, $\mu_i = \mathcal{E}(y_i)$, of the response variable, so that $g(\mu_i) = \eta_i$. The link function formalizes the more traditional approach of analyzing an ad-hoc transformation of y, such as $\log(y)$, $\sqrt{y}$, y^2, or Box-Cox (Box and Cox, 1964) transformations y^λ to determine an empirical optimal power transformation.
  - a *random component*, specifying the conditional distribution of $y_i \,|\, \boldsymbol{x}_i$ as any member of the exponential family, including the normal, binomial, Poisson, gamma, and other distributions.
- For the analysis of discrete response variables, and count data in particular, a key feature of the GLM is recognition of a *variance function* for the conditional variance of y_i, not forced to be constant, but rather allowed to depend on the mean μ_i and possibly a dispersion parameter, ϕ.
- From this background, we focus on GLMs for discrete count data response variables that extend considerably the loglinear models for contingency tables treated in Chapter 9. The Poisson distribution with a log link function is an equivalent starting point; however, count data GLMs often exhibit overdispersion in relation to the Poisson assumption that the conditional variance is the same as the mean, $\mathcal{V}(y_i \,|\, \eta_i) = \mu_i$.
  - One simple approach to this problem is the quasi-Poisson model, which estimates the dispersion parameter ϕ from the data, and uses this to correct standard errors and inferential tests.
  - Another is the wider class of negative-binomial models that allow a more flexible mean-variance function such as $\mathcal{V}(y_i \,|\, \eta_i) = \mu_i + \alpha\mu_i^2$.
- In practical application, many sets of empirical count data also exhibit a greater prevalence of zero counts than can be fit well using (quasi-) Poisson or negative-binomial models. Two simple extensions beyond the GLM class are
  - zero-inflated models, which posit a latent class of observations that always yield $y_i = 0$ counts, among the rest that have a Poisson or negative-binomial distribution including some zeros;
  - hurdle (or zero-altered) models, with one submodel for the zero counts and a separate submodel for the positive counts.
- Data analysis and visualization of count data therefore requires flexible tools and graphical methods. Some useful exploratory methods include jittered scatterplots and boxplots of $\log(y)$ against predictors enhanced by smoothed curves and trend lines, spine plots, and conditional density plots. Rootograms are quite helpful in visualizing the goodness-of-fit of count data models.
- Effect plots provide a convenient visual display of the high-order terms in a possibly complex GLM. They show the fitted values of the linear predictor $\widehat{\eta}^\star = \boldsymbol{X}^\star\widehat{\beta}$, using a score matrix $\boldsymbol{X}^\star$ that varies the predictors in a given term over their range while holding all other predictors constant. It is important to recognize, however, that like any model summary these show only the fitted effects under a given model, not the data.

- Model diagnostic measures (leverage, residuals, Cook's distance, etc.) and plots of these provide important ancillary information about the adequacy of a given model as a summary of relationships in the data. These help to detect problems of violations of assumptions, unusual or influential observations, or patterns that suggest that an important feature has not been accounted for.

- For multivariate response count data, there is no fully general theory as there is for the MLM with multivariate normality assumed for the errors. Nevertheless, there is a lot one can do to analyze such data combining the ideas of estimation for the separate responses with analysis of dependencies among the responses, conditioned by the explanatory variables.

11.9 Lab exercises

Exercise 11.1 Poole (1989) studied the mating behavior of elephants over 8 years in Amboseli National Park, Kenya. A focal aspect of the study concerned the mating success of males in relation to age, since larger males tend to be more successful in mating. Her data were used by Ramsey and Schafer (2002, Chapter 22) as a case study, and are contained in the Sleuth2 (Ramsey et al., 2012) package (Ramsey et al., 2012) as *case2201*.

For convenience, rename this to `elephants`, and study the relation between `Age` (at the beginning of the study) and number of successful `Matings` for the 41 adult male elephants observed over the course of this study, ranging in age from 27–52.

```
> data("case2201", package="Sleuth2")
> elephants <- case2201
> str(elephants)

'data.frame': 41 obs. of  2 variables:
 $ Age    : num  27 28 28 28 28 29 29 29 29 29 ...
 $ Matings: num  0 1 1 1 3 0 0 0 2 2 ...
```

(a) Create some exploratory plots of `Matings` against `Age` in the styles illustrated in this chapter. To do this successfully, you will have to account for the fact that `Matings` has a range of only 0–9, and use some smoothing methods to show the trend.
(b) Repeat (a) above, but now plotting `log(Matings+1)` against `Age` to approximate a Poisson regression with a log link and avoid problems with the zero counts.
(c) Fit a linear Poisson regression model for `Matings` against `Age`. Interpret the fitted model *verbally* from a graph of predicted number of matings and/or from the model coefficients. (*Hint*: Using `Age-27` will make the intercept directly interpretable.)
(d) Check for nonlinearity in the relationship by using the term `poly(Age,2)` in a new model. What do you conclude?
(e) Assess whether there is any evidence of overdispersion in these data by fitting analogous quasi-Poisson and negative-binomial models.

Exercise 11.2 The data set *quine* in MASS gives data on absenteeism from schools in rural New South Wales, Australia. 146 children were classified by ethnic background (`Eth`), age (`Age`, a factor), `Sex`, and Learner status (`Lrn`), and the number of days absent (`Days`) from school in a particular school year was recorded.

(a) Fit the all main-effects model in the Poisson family and examine the tests of these effects using `summary()` and car::`Anova()`. Are there any terms that should be dropped according to these tests?

(b) Re-fit this model as a quasi-Poisson model. Is there evidence of overdispersion? Test for overdispersion formally, using `dispersiontest()` from AER.
(c) Carry out the same significance tests and explain why the results differ from those for the Poisson model.

Exercise 11.3 The data set *AirCrash* in vcdExtra was analyzed in Exercise 5.2 and Exercise 6.3 in relation to the `Phase` of the flight and `Cause` of the crash. Additional variables include the number of `Fatalities` and `Year`. How does `Fatalities` depend on the other variables?

(a) Use the methods of this chapter to make some exploratory plots relating fatalities to each of the predictors.
(b) Fit a main effects poisson regression model for `Fatalities`, and make effects plots to visualize the model. Which phases and causes result in the largest number of fatalities?
(c) A linear effect of `Year` might not be appropriate for these data. Try using a natural spline term, `ns(Year, df)` to achieve a better, more adequate model.
(d) Use a model-building tool like `add1()` or MASS::`stepAIC()` to investigate whether there are important two-way interactions among the factors and your chosen effect for `Year`.
(e) Visualize and interpret your final model and write a brief summary to answer the question posed.

Exercise 11.4 Male double-crested cormorants use advertising behavior to attract females for breeding. The *Cormorants* data set in vcdExtra gives some results from a study by Meagan Mc Rae (2015) on counts of advertising males observed two or three times a week at six stations in a tree-nesting colony for an entire breeding season. The number of advertising birds was counted and these observations were classified by characteristics of the trees and nests. The goal was to determine how this behavior varies temporally over the season and spatially over observation stations, as well as with characteristics of nesting sites. The response variable is `count` and other predictors are shown below. See `help(Cormorants, package="vcdExtra")` for further details.

```
> data("Cormorants", package = "vcdExtra")
> car::some(Cormorants)

      category week station    nest height  density tree_health count
9          Pre    1      C2      no    mid      few        dead     8
48         Pre    1      B2 partial    mid      few     healthy     2
66         Pre    2      C2 partial   high      few     healthy     2
141        Pre    3      B1    full   high      few     healthy     1
143        Pre    3      B2      no    mid      few        dead     1
214 Incubation    5      C3    full   high      few        dead     1
217 Incubation    5      C4      no   high      few        dead    10
219 Incubation    5      C4 partial   high      few        dead     2
319       <NA>   10      C1      no   high     high        dead     1
342       <NA>   13      C2    full    mid moderate        dead     1
```

(a) Using the methods illustrated in this chapter, make some exploratory plots of the number of advertising birds against week in the breeding season, perhaps stratified by another predictor, like tree `height`, nest condition, or observation `station`. To see anything reasonable, you should plot `count` on a log (or square root) scale, jitter the points, and add smoothed curves. The variable `category` breaks the weeks into portions of the breeding season, so adding vertical lines separating those will be helpful for interpretation.
(b) Fit a main-effects Poisson GLM to these data and test the terms using `Anova()` from the car package.
(c) Interpret this model using an effects plot.
(d) Investigate whether the effect of `week` should be treated as linear in the model. You could

try using a polynomial term like `poly(week, degree)` or perhaps better, using a natural spline term like `ns(week, df)` from the splines package.

(e) Test this model for overdispersion, using either a `quasipoisson` family or `dispersiontest()` in AER.

Exercise 11.5 For the *CodParasites* data, recode the `area` variable as an ordered factor as suggested in footnote 13. Test the hypotheses that prevalence and intensity of cod parasites is linearly related to area.

Exercise 11.6 In Example 11.10, we ignored other potential predictors in the *CodParasites* data: `depth`, `weight`, `length`, `sex`, `stage`, and `age`. Use some of the graphical methods shown in this case study to assess whether any of these are related to prevalence and intensity.

Exercise 11.7 The analysis of the *PhdPubs* data in the examples in this chapter were purposely left incomplete, going only as far as the negative binomial model.

(a) Fit the zero-inflated and hurdle models to this data set, considering whether the count component should be Poisson or negative-binomial, and whether the zero model should use all predictors or only a subset. Describe your conclusions from this analysis in a few sentences.

(b) Using the methods illustrated in this chapter, create some graphs summarizing the predicted counts and probabilities of zero counts for one of these models.

(c) For your chosen model, use some of the diagnostic plots of residuals and other measures shown in Section 11.6 to determine if your model solves any of the problems noted in Example 11.17 and Example 11.18, and whether there are any problems that remain.

Exercise 11.8 In Example 11.19 we used a simple analysis of $\log(\boldsymbol{y} + 1)$ for the multivariate responses in the *NMES1988* data using a classical MLM (Eqn. (11.16)) as a rough approximation of a multivariate Poisson model. The HE plot in Figure 11.40 was given as a visual summary, but did not show the data. Examine why the MLM is not appropriate statistically for these data, as follows:

(a) Calculate residuals for the model `nmes.mlm` using

```
> resids <- residuals(nmes.mlm, type="deviance")
```

(b) Make univariate density plots of these residuals to show their univariate distributions. These should be approximately normal under the MLM. What do you conclude?

(c) Make some bivariate plots of these residuals. Under the MLM, each should be bivariate normal with elliptical contours and linear regressions. Add 2D density contours (`kde2d()`, or `geom_density2d()` in ggplot2) and some smoothed curve. What do you conclude?

This page intentionally left blank

References

Aberdein, J. and Spiegelhalter, D. (2013). Have London's roads become more dangerous for cyclists? *Significance*, 10(6), 46–48.

Adler, D. and Murdoch, D. (2014). *rgl: 3D visualization device system (OpenGL)*. R package version 0.95.1201.

Agresti, A. (1984). *Analysis of Ordinal Categorical Data*. New York: Wiley.

Agresti, A. (1990). *Categorical Data Analysis*. New York: Wiley-Interscience.

Agresti, A. (1996). *An Introduction to Categorical Data Analysis*. New York: Wiley Interscience.

Agresti, A. (2002). *Categorical Data Analysis*. Wiley Series in Probability and Statistics. New York: Wiley-Interscience [John Wiley & Sons], 2nd edn.

Agresti, A. (2007). *An Introduction to Categorical Data Analysis*. New York: Wiley, 2nd edn.

Agresti, A. (2013). *Categorical Data Analysis*. Wiley Series in Probability and Statistics. New York: Wiley-Interscience [John Wiley & Sons], 3rd edn.

Agresti, A. and Winner, L. (1997). Evaluating agreement and disagreement among movie reviewers. *Chance*, 10(2), 10–14.

Aitchison, J. (1986). *The Statistical Analysis of Compositional Data*. London: Chapman and Hall.

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In B. N. Petrov and F. Czaki, eds., *Proceedings of the 2nd International Symposium on Information*. Budapest: Akademiai Kiado.

Andersen, E. B. (1991). *Statistical Analysis of Categorical Data*. Berlin: Springer-Verlag, 2nd edn.

Anderson, E. (1935). The irises of the Gaspé peninsula. *Bulletin of the American Iris Society*, 35, 2–5.

Andrews, D. F. and Herzberg, A. M. (1985). *Data: A Collection of Problems from Many Fields for the Student and Research Worker*. New York, NY: Springer-Verlag.

Antonio, A. L. M. and Crespi, C. M. (2010). Predictors of interobserver agreement in breast imaging using the breast imaging reporting and data system. *Breast Cancer Research and Treatment*, 120(3), 539–546.

Arbuthnot, J. (1710). An argument for divine providence, taken from the constant regularity observ'd in the births of both sexes. *Philosophical Transactions*, 27, 186–190. Published in 1711.

Ashford, J. R. and Sowden, R. D. (1970). Multivariate probit analysis. *Biometrics*, 26, 535–546.

Atkinson, A. C. (1981). Two graphical displays for outlying and influential observations in regression. *Biometrika*, 68, 13–20.

Atkinson, A. C. (1987). *Plots, Transformations and Regression: An Introduction to Graphical Methods of Diagnostic Regression Analysis*. New York: Oxford University Press.

Bangdiwala, S. I. (1985). A graphical test for observer agreement. In *Proceeding of the International Statistics Institute*, vol. 1, (pp. 307–308). Amsterdam: ISI.

Bangdiwala, S. I. (1987). Using SAS software graphical procedures for the observer agreement chart. *Proceedings of the SAS User's Group International Conference*, 12, 1083–1088.

Bartlett, M. S. (1935). Contingency table interactions. *Journal of the Royal Statistical Society, Supplement*, 2, 248–252.

Bates, D., Maechler, M., Bolker, B., and Walker, S. (2014). *lme4: Linear mixed-effects models using Eigen and S4*. R package version 1.1-7.

Becker, M. P. and Clogg, C. C. (1989). Analysis of sets of two-way contingency tables using association models. *Journal of the American Statistical Association*, 84(405), 142–151.

Bentler, P. M. (1980). Multivariate analysis with latent variables: Causal modeling. *Annual Review of Psychology*, 31(1), 419–456.

Benzécri, J.-P. (1977). Sur l'analyse des tableaus binaries associés a une correspondense multiple. *Cahiers de l'Analyse des Données*, 2, 55–71.

Bertin, J. (1981). *Graphics and Graphic Information-processing*. New York: de Gruyter. (trans. W. Berg and P. Scott).

Bertin, J. (1983). *Semiology of Graphics*. Madison, WI: University of Wisconsin Press. (trans. W. Berg).

Bickel, P. J., Hammel, J. W., and O'Connell, J. W. (1975). Sex bias in graduate admissions: Data from Berkeley. *Science*, 187, 398–403.

Birch, M. W. (1963a). An algorithm for the logarithmic series distributions. *Biometrics*, 19, 651–652.

Birch, M. W. (1963b). Maximum likelihood in three-way contingency tables. *Journal of the Royal Statistical Society, Series B*, 25, 220–233.

Bishop, Y. M. M., Fienberg, S. E., and Holland, P. W. (1975). *Discrete Multivariate Analysis: Theory and Practice*. Cambridge, MA: MIT Press.

Bliss, C. I. (1934). The method of probits. *Science*, 79(2037), 38–39.

Böhning, D. (1983). Maximum likelihood estimation of the logarithmic series distribution. *Statistische Hefte (Statistical Papers)*, 24(1), 121–140.

Bollen, K. A. (2002). Latent variables in psychology and the social sciences. *Annual Review of Psychology*, 53(1), 605–634.

von Bortkiewicz, L. (1898). *Das Gesetz der Kleinen Zahlen*. Leipzig: Teubner.

Bouchet-Valat, M. (2015). *logmult: Log-Multiplicative Models, Including Association Models*. R package version 0.6.1.

Box, G. E. P. (1979). Some problems of statistics and everyday life. *Journal of the American Statistical Association*, 74(365), 1–4.

Box, G. E. P. and Cox, D. R. (1964). An analysis of transformations (with discussion). *Journal of the Royal Statistical Society, Series B*, 26, 211–252.

Box, G. E. P. and Draper, N. R. (1987). *Empirical Model Building and Response Surfaces*. New York, NY: John Wiley & Sons.

Bradu, D. and Gabriel, K. R. (1978). The biplot as a diagnostic tool for models of two-way tables. *Technometrics*, 20, 47–68.

Brinton, W. C. (1939). *Graphic Presentation*. New York, NY: Brinton Associates.

Brockmann, H. J. (1996). Satellite male groups in horseshoe crabs, *Limulus polyphemus*. *Ethology*, 102(1), 1–21.

Brown, P. J., Stone, J., and Ord-Smith, C. (1983). Toxaemic signs during pregnancy. *Journal of the Royal Statistical Society, Series C (Applied Statistics)*, 32, 69–72.

Brunswick, A. F. (1971). Adolescent health, sex, and fertility. *American Journal of Public Health*, 61(4), 711–729.

Burt, C. (1950). The factorial analysis of qualitative data. *British Journal of Statistical Psychology*, 3, 166–185.

Cameron, A. C. and Trivedi, P. K. (1990). Regression-based tests for overdispersion in the poisson model. *Journal of Econometrics*, 46, 347–364.

Cameron, A. C. and Trivedi, P. K. (1998). *Regression analysis of count data*. Econometric society monographs. Cambridge (U.K.), New York: Cambridge University Press.

Cameron, A. C. and Trivedi, P. K. (2013). *Regression analysis of count data*. Econometric society monographs. Cambridge (U.K.), New York: Cambridge University Press, 2nd edn.

Carlyle, T. (1840). *Chartism*. London: J. Fraser.

Caussinus, H. (1966). Contribution à l'analyse statistique des tableaux de corrélation. *Annales de la Faculté des Sciences de l'Université de Toulouse*, 39 (année 1965), 77–183.

Chambers, J. M., Cleveland, W. S., Kleiner, B., and Tukey, P. A. (1983). *Graphical Methods for Data Analysis*. Belmont, CA: Wadsworth.

Chang, W. and Wickham, H. (2015). *ggvis: Interactive Grammar of Graphics*. R package version 0.4.1.

Christensen, R. (1997). *Log-Linear Models and Logistic Regression*. New York, NY: Springer, 2nd edn.

Cicchetti, D. V. and Allison, T. (1971). A new procedure for assessing reliability of scoring EEG sleep recordings. *American Journal of EEG Technology*, 11, 101–109.

Cleveland, W. S. (1993a). A model for studying display methods of statistical graphics. *Journal of Computational and Graphical Statistics*, 2, 323–343.

Cleveland, W. S. (1993b). *Visualizing Data*. Summit, NJ: Hobart Press.

Cleveland, W. S., McGill, M. E., and McGill, R. (1988). The shape parameter of a two-variable graph. *Journal of the American Statistical Association*, 83, 289–300.

Cleveland, W. S. and McGill, R. (1984). Graphical perception: Theory, experimentation and application to the development of graphical methods. *Journal of the American Statistical Association*, 79, 531–554.

Cleveland, W. S. and McGill, R. (1985). Graphical perception and graphical methods for analyzing scientific data. *Science*, 229, 828–833.

Cohen, A. (1980). On the graphical display of the significant components in a two-way contingency table. *Communications in Statistics—Theory and Methods*, A9, 1025–1041.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20, 37–46.

Cohen, J. (1968). Weighted kappa: Nominal scale agreement with provision for scaled diasgreement or partial credit. *Psychological Bulletin*, 70, 213–220.

Cook, R. D. (1993). Exploring partial residual plots. *Technometrics*, 35(4), 351–362.

Cook, R. D. and Weisberg, S. (1999). *Applied Regression Including Computing and Grapics*. New York: Wiley.

Cragg, J. G. (1971). Some statistical models for limited dependent variables with application to the demand for durable goods. *Econometrica*, 39, 829–844.

Croissant, Y. (2013). *mlogit: multinomial logit model*. R package version 0.2-4.

de la Cruz Rot, M. (2005). Improving the presentation of results of logistic regression with r. *Bulletin of the Ecological Society of America*, 86, 41–48.

Dahl, D. B. (2014). *xtable: Export tables to LaTeX or HTML*. R package version 1.7-4.

Dalal, S., Fowlkes, E. B., and Hoadley, B. (1989). Risk analysis of the space shuttle: Pre-*Challenger* prediction of failure. *Journal of the American Statistical Association*, 84(408), 945–957.

Dalgaard, P. (2008). *Introductory Statistics with R*. Springer, 2nd edn.

Deb, P. and Trivedi, P. K. (1997). Demand for medical care by the elderly: A finite mixture approach. *Journal of Applied Econometrics*, 12, 313–336.

Dragulescu, A. A. (2014). *xlsx: Read, write, format Excel 2007 and Excel 97/2000/XP/2003 files*. R package version 0.5.7.

Edwards, A. W. F. (1958). An analysis of geissler's data on the human sex ratio. *Annals of Human Genetics*, 23(1), 6–15.

Emerson, J. W. (1998). Mosaic displays in S-PLUS: A general implementation and a case study. *Statistical Graphics and Computing Newsletter*, 9(1), 17–23.

Emerson, J. W. and Green, W. A. (2014). *gpairs: The Generalized Pairs Plot*. R package version 1.2.

Emerson, J. W., Green, W. A., Schloerke, B., Crowley, J., Cook, D., Hofmann, H., and Wickham, H. (2013). The generalized pairs plot. *Journal of Computational and Graphical Statistics*, 22(1), 79–91.

Evers, M. and Namboordiri, N. K. (1977). A Monte Carlo assessment of the stability of log-linear estimates in small samples. In *Proceedings of the Social Statistics Section*. Alexandria, VA: American Statistical Association.

Feynman, R. P. (1988). *What Do You Care What Other People Think? Further Adventures of a Curious Character*. New York: W. W. Norton.

Fienberg, S. E. (1975). Perspective Canada as a social report. *Social Indicators Research*, 2, 153–174.

Fienberg, S. E. (1980). *The Analysis of Cross-Classified Categorical Data*. Cambridge, MA: MIT Press, 2nd edn.

Fienberg, S. E. and Rinaldo, A. (2007). Three centuries of categorical data analysis: Log-linear models and maximum likelihood estimation. *Journal of Statistical Planning and Inference*, 137(11), 3430–3445.

Finney, D. J. (1947). *Probit analysis*. Cambridge, England: Cambridge University Press.

Firth, D. (2003). Overcoming the reference category problem in the presentation of statistical models. *Sociological Methodology*, 33, 1–18.

Firth, D. and Menezes, R. X. d. (2004). Quasi-variances. *Biometrika*, 91, 65–80.

Fisher, R. A. (1925). *Statistical Methods for Research Workers*. London: Oliver & Boyd.

Fisher, R. A. (1936a). Has Mendel's work been rediscovered? *Annals of Science*, 1, 115–137.

Fisher, R. A. (1936b). The use of multiple measurements in taxonomic problems. *Annals of Eugenics*, 8, 379–388.

Fisher, R. A. (1940). The precision of discriminant functions. *Annals of Eugenics*, 10, 422–429.

Fisher, R. A., Corbet, A. S., and Williams, C. B. (1943). The relation between the number of species and the number of individuals. *Journal of Animal Ecology*, 12, 42.

Fleiss, J. L. (1973). *Statistical Methods for Rates and Proportions*. New York: John Wiley and Sons.

Fleiss, J. L. and Cohen, J. (1972). The equivalence of weighted kappa and the intraclass correlation coefficient as measures of reliability. *Educational and Psychological Measurement*, 33, 613–619.

Fleiss, J. L., Cohen, J., and Everitt, B. S. (1969). Large sample standard errors of kappa and weighted kappa. *Psychological Bulletin*, 72, 332–327.

Fowlkes, E. B. (1987). Some diagnostics for binary logistic regression via smoothing. *Biometrika*, 74(3), 503–5152.

Fox, J. (1987). Effect displays for generalized linear models. In C. C. Clogg, ed., *Sociological Methodology, 1987*, (pp. 347–361). San Francisco: Jossey-Bass.

Fox, J. (2003). Effect displays in R for generalized linear models. *Journal of Statistical Software*, 8(15), 1–27.

Fox, J. (2008). *Applied Regression Analysis and Generalized Linear Models*. Thousand Oaks, CA: SAGE Publications, 2nd edn.

Fox, J. (2015). Appendices to *Applied Regression Analysis, Generalized Linear Models, and Related Methods*, third edition. Online document. Available at `http://socserv.socsci.mcmaster.ca/jfox/Books/Applied-Regression-3E/Appendices.pdf`.

Fox, J. and Andersen, R. (2006). Effect displays for multinomial and proportional-odds logit models. *Sociological Methodology*, 36, 225–255.

Fox, J. and Friendly, M. (2014). *heplots: Visualizing Hypothesis Tests in Multivariate Linear Models*. R package version 1.0-12.

Fox, J., Friendly, M., and Monette, G. (2009). Visualizing hypothesis tests in multivariate linear models: The *heplots* package for R. *Computational Statistics*, 24(2), 233–246. (Published online: May 15, 2008).

Fox, J. and Weisberg, S. (2011a). *An R Companion to Applied Regression*. Thousand Oaks CA: SAGE Publications, 2nd edn.

Fox, J. and Weisberg, S. (2011b). *An R Companion to Applied Regression*. Thousand Oaks CA: SAGE Publications, 2nd edn.

Fox, J. and Weisberg, S. (2015a). *car: Companion to Applied Regression*. R package version 2.0-25/r421.

Fox, J. and Weisberg, S. (2015b). Visualizing fit and lack of fit in complex regression models: Effect plots with partial residuals. submitted, *Journal of Computational and Graphical Statistics*.

Fox, J., Weisberg, S., Friendly, M., and Hong, J. (2015). *effects: Effect Displays for Linear, Generalized Linear, and Other Models*. R package version 3.0-4/r200.

Friendly, M. (1991). *SAS System for Statistical Graphics*. Cary, NC: SAS Institute, 1st edn.

Friendly, M. (1992). Mosaic displays for loglinear models. In *ASA, Proceedings of the Statistical Graphics Section*, (pp. 61–68). Alexandria, VA.

Friendly, M. (1994a). A fourfold display for 2 by 2 by K tables. Tech. Rep. 217, York University, Psychology Dept.

Friendly, M. (1994b). Mosaic displays for multi-way contingency tables. *Journal of the American Statistical Association*, 89, 190–200.

Friendly, M. (1994c). SAS/IML graphics for fourfold displays. *Observations*, 3(4), 47–56.

Friendly, M. (1995). Conceptual and visual models for categorical data. *The American Statistician*, 49, 153–160.

Friendly, M. (1997). Conceptual models for visualizing contingency table data. In M. Greenacre and J. Blasius, eds., *Visualization of Categorical Data*, chap. 2, (pp. 17–35). San Diego, CA: Academic Press.

Friendly, M. (1999a). Extending mosaic displays: Marginal, conditional, and partial views of categorical data. *Journal of Computational and Graphical Statistics*, 8(3), 373–395.

Friendly, M. (1999b). Extending mosaic displays: Marginal, conditional, and partial views of categorical data. *Journal of Computational and Graphical Statistics*, 8(3), 373–395.

Friendly, M. (2000). *Visualizing Categorical Data*. Cary, NC: SAS Institute.

Friendly, M. (2002). Corrgrams: Exploratory displays for correlation matrices. *The American Statistician*, 56(4), 316–324.

Friendly, M. (2003). Visions of the past, present and future of statistical graphics: An ideo-graphic view. American Psychological Association. Toronto, ON, URL: `http://datavis.ca/papers/apa-2x2.pdf`.

Friendly, M. (2007). HE plots for multivariate general linear models. *Journal of Computational and Graphical Statistics*, 16(2), 421–444.

Friendly, M. (2013). Comment on the generalized pairs plot. *Journal of Computational and Graphical Statistics*, 22(1), 290–291.

Friendly, M. (2014a). *HistData: Data sets from the history of statistics and data visualization*. R package version 0.7-5.

Friendly, M. (2014b). *Lahman: Sean Lahman's Baseball Database*. R package version 3.0-1.

Friendly, M. (2015). *vcdExtra: vcd Extensions and Additions*. R package version 0.6-7.

Friendly, M. and Denis, D. (2005). The early origins and development of the scatterplot. *Journal of the History of the Behavioral Sciences*, 41(2), 103–130.

Friendly, M. and Kwan, E. (2003). Effect ordering for data displays. *Computational Statistics and Data Analysis*, 43(4), 509–539.

Friendly, M. and Kwan, E. (2011). Comment (graph people versus table people). *Journal of Computational and Graphical Statistics*, 20(1), 18–27.

Friendly, M., Monette, G., and Fox, J. (2013). Elliptical insights: Understanding statistical methods through elliptical geometry. *Statistical Science*, 28(1), 1–39.

Gabriel, K. R. (1971). The biplot graphic display of matrices with application to principal components analysis. *Biometrics*, 58(3), 453–467.

Gabriel, K. R. (1980). Biplot. In N. L. Johnson and S. Kotz, eds., *Encyclopedia of Statistical Sciences*, vol. 1, (pp. 263–271). New York: John Wiley and Sons.

Gabriel, K. R. (1981). Biplot display of multivariate matrices for inspection of data and diagnosis. In V. Barnett, ed., *Interpreting Multivariate Data*, chap. 8, (pp. 147–173). London: John Wiley and Sons.

Gabriel, K. R., Galindo, M. P., and Vincente-Villardón, J. L. (1997). Use of biplots to diagnose independence models in three-way contingency tables. In M. Greenacre and J. Blasius, eds., *Visualization of Categorical Data*, chap. 27, (pp. 391–404). San Diego, CA: Academic Press.

Gabriel, K. R. and Odoroff, C. L. (1990). Biplots in biomedical research. *Statistics in Medicine*, 9, 469–485.

Galton, F. (1886). Regression towards mediocrity in hereditary stature. *Journal of the Anthropological Institute*, 15, 246–263.

Gart, J. J. and Zweiful, J. R. (1967). On the bias of various estimators of the logit and its variance with applications to quantal bioassay. *Biometrika*, 54, 181–187.

Geissler, A. (1889). Beitrage zur frage des geschlechts verhaltnisses der geborenen. *Z. K. Sachsischen Statistischen Bureaus*, 35(1), n.p.

Gesmann, M. and de Castillo, D. (2015). *googleVis: R Interface to Google Charts*. R package version 0.5.8.

Gifi, A. (1981). *Nonlinear Multivariate Analysis*. The Netherlands: Department of Data Theory, University of Leiden.

Glass, D. V. (1954). *Social Mobility in Britain*. Glencoe, IL: The Free Press.

Goodman, L. A. (1970). The multivariate analysis of qualitative data: Interactions among multiple classifications. *Journal of the American Statistical Association*, 65, 226–256.

Goodman, L. A. (1971). The analysis of multidimensional contingency tables: Stepwise procedures and direct estimates for building models for multiple classifications. *Technometrics*, 13, 33–61.

Goodman, L. A. (1972). Some multiplicative models for the analysis of cross classified data. In *Proceedings of the sixth Berkeley Symposium on Mathematical Statistics and Probability*, (pp. 649–696). Berkeley, CA: University of California.

Goodman, L. A. (1973). The analysis of multidimensional contingency tables when some variables are posterior to others: A modified path analysis approach. *Biometrika*, 60, 179–192.

Goodman, L. A. (1978). *Analyzing Qualitative Categorical Data: Log-Linear Models and Latent-Strucutre Analysis*. Cambridge, MA: Abt Books.

Goodman, L. A. (1979). Simple models for the analysis of association in cross-classifications having ordered categories. *Journal of the American Statistical Association*, 74, 537–552.

Goodman, L. A. (1981). Association models and canonical correlation in the analysis of cross-classifications having ordered categories. *Journal of the American Statistical Association*, 76(374), 320–334.

Goodman, L. A. (1983). The analysis of dependence in cross-classifications having ordered categories, using log-linear models for frequencies and log-linear models for odds. *Biometrics*, 39, 149–160.

Goodman, L. A. (1985). The analysis of cross-classified data having ordered and/or unordered categories: Association models, correlation models, and asymmetry models for contingency tables with or without missing entries. *Annals of Statistics*, 13(1), 10–69.

Goodman, L. A. (1986). Some useful extensions of the usual correspondence analysis approach and the usual log-linear models approach in the analysis of contingency tables. *International Statistical Review*, 54(3), 243–309. With a discussion and reply by the author.

Gower, J., Lubbe, S., and Roux, N. (2011). *Understanding Biplots*. Wiley.

Gower, J. C. and Hand, D. J. (1996). *Biplots*. London: Chapman & Hall.

Grayson, D. K. (1990). Donner party deaths: A demographic assessment. *Journal of Anthropological Research*, 46(3), 223–242.

Greenacre, M. (1984). *Theory and Applications of Correspondence Analysis*. London: Academic Press.

Greenacre, M. (1989). The Carroll-Green-Schaffer scaling in correspondence analysis: A theoretical and empirical appraisal. *Journal of Marketing Research*, 26, 358–365.

Greenacre, M. (1990). Some limitations of multiple correspondence analysis. *Computational Statistics Quarterly*, 3, 249–256.

Greenacre, M. (1994). Multiple and joint correspondence analysis. In M. J. Greenacre and B. Jörg, eds., *Correspondence Analysis in the Social Sciences*. London: Academic Press.

Greenacre, M. (1997). Diagnostics for joint displays in correspondence analysis. In J. Blasius and M. Greenacre, eds., *Visualization of Categorical Data*, (pp. 221–238). Academic Press.

Greenacre, M. (2007). *Correspondence analysis in practice*. Boca Raton: Chapman & Hall/CRC.

Greenacre, M. (2013). Contribution biplots. *Journal of Computational and Graphical Statistics*, 22(1), 107–122.

Greenacre, M. and Hastie, T. J. (1987). The geometric interpretation of correspondence analysis. *Journal of the American Statistical Association*, 82, 437–447.

Greenacre, M. and Nenadic, O. (2014). *ca: Simple, Multiple and Joint Correspondence Analysis*. R package version 0.58.

Greenwood, M. and Yule, G. U. (1920). An inquiry into the nature of frequency distributions of multiple happenings, with particular reference to the occurrence of multiple attacks of disease or repeated accidents. *Journal of the Royal Statistical Society, Series A*, 83, 255–279.

Haberman, S. J. (1972). Statistical algorithms: Algorithm AS 51: Log-linear fit for contingency tables. *Applied Statistics*, 21(2), 218–225.

Haberman, S. J. (1973). The analysis of residuals in cross-classified tables. *Biometrics*, 29, 205–220.

Haberman, S. J. (1974). *The Analysis of Frequency Data*. Chicago: University of Chicago Press.

Haberman, S. J. (1979). *The Analysis of Qualitative Data: New Developments*, vol. II. New York: Academic Press.

Haldane, J. B. S. (1955). The estimation and significance of the logarithm of a ratio of frequencies. *Annals of Human Genetics*, 20, 309–311.

Hamilton, N. (2014). *ggtern: An extension to ggplot2, for the creation of ternary diagrams.* R package version 1.0.3.2.

Harrell, Jr, F. E. (2001). *Regression Modeling Strategies: With Applications to Linear Models, Logistic Regression, and Survival Analysis*. Graduate Texts in Mathematics. New York: Springer.

Harrell, Jr., F. E. (2015). *rms: Regression Modeling Strategies*. R package version 4.3-0.

Hartigan, J. A. and Kleiner, B. (1981). Mosaics for contingency tables. In W. F. Eddy, ed., *Computer Science and Statistics: Proceedings of the 13th Symposium on the Interface*, (pp. 268–273). New York, NY: Springer-Verlag.

Hartigan, J. A. and Kleiner, B. (1984). A mosaic of television ratings. *The American Statistician*, 38, 32–35.

Hastie, T. J. and Tibshirani, R. J. (1990). *Generalized Additive Models*. London: Chapman & Hall.

Hauser, R. M. (1979). Some exploratory methods for modeling mobility tables and other cross-classified data. In K. F. Schuessler, ed., *Sociological Methodology 1980*, (pp. 413–458). San Francisco: Jossey-Bass.

Hedeker, D. (2005). Generalized linear mixed models. In *Encyclopedia of Statistics in Behavioral Science*. John Wiley & Sons, Ltd.

van der Heijden, P. G. M., de Falguerolles, A., and de Leeuw, J. (1989). A combined approach to contingency table analysis using correspondence analysis and log-linear analysis. *Applied Statistics*, 38(2), 249–292.

van der Heijden, P. G. M. and de Leeuw, J. (1985). Correspondence analysis used complementary to loglinear analysis. *Psychometrika*, 50, 429–447.

Hemmingsen, W., Jansen, P. A., and Mackenzie, K. (2005). Crabs, leeches and trypanosomes: an unholy trinity? *Marine Pollution Bulletin*, 50(3), 336–339.

Heuer, J. (1979). *Selbstmord Bei Kinder Und Jugendlichen*. Stuttgard: Ernst Klett Verlag. [Suicide by children and youth.].

Hilbe, J. (2011). *Negative Binomial Regression*. Cambridge University Press, 2nd edn.

Hilbe, J. M. (2014). *Modeling Count Data*. New York, NY: Cambridge University Press.

Hoaglin, D. C. (1980). A poissonness plot. *The American Statistician*, 34, 146–149.

Hoaglin, D. C. and Tukey, J. W. (1985). Checking the shape of discrete distributions. In D. C. Hoaglin, F. Mosteller, and J. W. Tukey, eds., *Exploring Data Tables, Trends and Shapes*, chap. 9. New York: John Wiley and Sons.

Hocking, T. D. (2013). *directlabels: Direct labels for multicolor plots in lattice or ggplot2*. R package version 2013.6.15.

Hofmann, H. (2000). Exploring categorical data: Interactive mosaic plots. *Metrika*, 51(1), 11–26.

Hofmann, H. (2001). Generalized odds ratios for visual modeling. *Journal of Computational and Graphical Statistics*, 10(4), 628–640.

Hofmann, H. and Theus, M. (2005). Interactive graphics for visualizing conditional distributions. Unpublished manuscript.

Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, 6(2), 65–70.

Hosmer, Jr, D. W., Lemeshow, S., and Sturdivant, R. X. (2013). *Applied Logistic Regression*. New York: John Wiley and Sons, 3rd edn.

Hothorn, T., Zeileis, A., Farebrother, R. W., and Cummins, C. (2014). *lmtest: Testing Linear Regression Models*. R package version 0.9-33.

Hout, M., Duncan, O. D., and Sobel, M. E. (1987). Association and heterogeneity: Structural models of similarities and differences. *Sociological Methodology*, 17, 145–184.

Hummel, J. (1996). Linked bar charts: Analyzing categorical data graphically. *Computational Statistics*, 11, 23–33.

Hurley, C. B. (2004). Clustering visualizations of multidimensional data. *Journal of Computational and Graphical Statistics*, 13, 788–806.

Husson, F., Josse, J., Le, S., and Mazet, J. (2015). *FactoMineR: Multivariate Exploratory Data Analysis and Data Mining*. R package version 1.29.

Ihaka, R., Murrell, P., Hornik, K., Fisher, J. C., and Zeileis, A. (2015). *colorspace: Color Space Manipulation*. R package version 1.2-6.

Immer, F. R., Hayes, H., and Powers, L. R. (1934). Statistical determination of barley varietial adaptation. *Journal of the American Society of Agronomy*, 26, 403–419.

Jackman, S., Tahk, A., Zeileis, A., Maimone, C., and Fearon, J. (2015). *pscl: Political Science Computational Laboratory, Stanford University*. R package version 1.4.9.

Jansen, J. (1990). On the statistical analysis of ordinal data when extravariation is present. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 39(1), 75–84.

Jinkinson, R. A. and Slater, M. (1981). Critical discussion of a graphical method for identifying discrete distributions. *The Statistician*, 30, 239–248.

Johnson, K. (1996). *Unfortunate Emigrants: Narratives of the Donner Party*. Logan, UT: Utah State University Press.

Johnson, N. L., Kotz, S., and Kemp, A. W. (1992). *Univariate Discrete Distributions*. New York, NY: John Wiley and Sons, 2nd edn.

Kemp, A. W. and Kemp, C. D. (1991). Weldon's dice data revisited. *The American Statistician*, 45, 216–222.

Kendall, M. G. and Stuart, A. (1961). *The Advanced Theory of Statistics*, vol. 2. London: Griffin.

Kendall, M. G. and Stuart, A. (1963). *The Advanced Theory of Statistics*, vol. 1. London: Griffin.

King, G. (1989). A seemingly unrelated Poisson regression model. *Sociological Methods and Research*, 17(3), 235–255.

Kleiber, C. and Zeileis, A. (2008). *Applied Econometrics with R*. New York: Springer-Verlag. ISBN 978-0-387-77316-2.

Kleiber, C. and Zeileis, A. (2014). Visualizing count data regressions using rootograms. Working papers, Faculty of Economics and Statistics, University of Innsbruck.

Kleiber, C. and Zeileis, A. (2015). *AER: Applied Econometrics with R*. R package version 1.2-3.

Koch, G. and Edwards, S. (1988). Clinical efficiency trials with categorical data. In K. E. Peace, ed., *Biopharmaceutical Statistics for Drug Development*, (pp. 403–451). New York: Marcel Dekker.

Kosambi, D. D. (1949). Characteristic properties of series distributions. *Proceedings of the National Institute of Science of India*, 15, 109–113.

Kosslyn, S. M. (1985). Graphics and human information processing: A review of five books. *Journal of the American Statistical Association*, 80, 499–512.

Kosslyn, S. M. (1989). Understanding charts and graphs. *Applied Cognitive Psychology*, 3, 185–225.

Kundel, H. L. and Polansky, M. (2003). Measurement of observer agreement. *Radiology*, 228(2), 303–308.

Labby, Z. (2009). Weldon's dice, automated. *Chance*, 22(4), 6–13.

Lambert, D. (1992). Zero-inflated poisson regression, with an application to defects in manufacturing. *Technometrics*, 34, 1–14.

Landis, J. R. and Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, 33, 159–174.

Landis, R. J., Heyman, E. R., and Koch, G. G. (1978). Average partial association in three-way contingency tables: A review and discussion of alternative tests,. *International Statistical Review*, 46, 237–254.

Landwehr, J. M., Pregibon, D., and Shoemaker, A. C. (1984). Graphical methods for assessing logistic regression models. *Journal of the American Statistical Association*, 79, 61–71.

Lang, J. B. and Agresti, A. (1994). Simultaneously modeling joint and marginal distributions of multivariate categorical responses. *Journal of the American Statistical Association*, 89(426), 625–632.

Larsen, W. A. and McCleary, S. J. (1972). The use of partial residual plots in regression analysis. *Technometrics*, 14, 781–790.

Lavine, M. (1991). Problems in extrapolation illustrated with space shuttle O-ring data. *Journal of the American Statistical Association*, 86, 912–922.

Lawrance, A. J. (1995). Deletion influence and masking in regression. *Journal of the Royal Statistical Society. Series B (Methodological)*, 57(1), 181–189.

Lazarsfeld, P. F. (1950). The logical and mathematical foundation of latent structure analysis. In S. A. Stouffer, L. Guttmann, E. A. Suchman, P. F. Lazarsfeld, S. A. Star, and J. A. Clausen, eds., *Studies in Social Psychology in World War II, vol. IV, Measurement and Prediction*, (pp. 362–412). Princeton, NJ: Princeton University Press.

Lazarsfeld, P. F. (1954). A conceptual introduction to latent structure analysis. In P. F. Lazarsfeld, ed., *Mathematical Thinking in the Social Sciences*, (pp. 349–387). Glencoe, IL: Free Press.

Lebart, L., Morineau, A., and Warwick, K. M. (1984). *Multivariate Descriptive Statistical Analysis: Correspondence Analysis and Related Techniques for Large Matrices*. New York: John Wiley and Sons.

Lee, A. J. (1997). Modelling scores in the Premier League: Is Manchester United really the best? *Chance*, 10(1), 15–19.

Leifeld, P. (2013). texreg: Conversion of statistical model output in R to LaTeX and HTML tables. *Journal of Statistical Software*, 55(8), 1–24.

Leifeld, P. (2014). *texreg: Conversion of R regression output to LaTeX or HTML tables.* R package version 1.34.

Lemeshow, S., Avrunin, D., and Pastides, J. S. (1988). Predicting the outcome of intensive care unit patients. *Journal of the American Statistical Association*, 83, 348–356.

Lenth, R. V. (2014). *rsm: Response-surface analysis*. R package version 2.07.

Lenth, R. V. and Hervé, M. (2015). *lsmeans: Least-Squares Means*. R package version 2.16.

Lewandowsky, S. and Spence, I. (1989). The perception of statistical graphs. *Sociological Methods & Research*, 18, 200–242.

Lindsey, J. K. (1995). *Modelling Frequency and Count Data*. Oxford, UK: Oxford University Press.

Lindsey, J. K. and Altham, P. M. E. (1998). Analysis of the human sex ratio by using overdispersion models. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 47(1), 149–157.

Lindsey, J. K. and Mersch, G. (1992). Fitting and comparing probability distributions with log linear models. *Computational Statistics and Data Analysis*, 13, 373–384.

Linzer, D. and Lewis., J. (2014). *poLCA: Polytomous variable Latent Class Analysis*. R package version 1.4.1.

Long, J. S. (1990). The origins of sex differences in science. *Social Forces*, 68(4), 1297–1316.

Long, J. S. (1997). *Regression Models for Categorical and Limited Dependent Variables*. Thousand Oaks, CA: SAGE Publications.

Lumley, T. and Zeileis, A. (2015). *sandwich: Robust Covariance Matrix Estimators*. R package version 2.3-3.

Maindonald, J. and Braun, J. (2007). *Data Analysis and Graphics Using R*. Cambridge: Cambridge University Press, 2nd edn.

Mc Rae, M. (2015). *Spatial, Habitat and Frequency Changes in Double-crested Cormorant Advertising Display in a Tree-nesting Colony*. Masters project, environmental studies, York University.

McCullagh, P. (1980). Regression models for ordinal data. *Journal of the Royal Statistical Society*, 42, 109–142.

McCullagh, P. and Nelder, J. A. (1989). *Generalized Linear Models*. London: Chapman and Hall.

McCulloch, C. E. and Neuhaus, J. M. (2005). Generalized linear mixed models. In *Encyclopedia of Biostatistics*. John Wiley & Sons, Ltd.

Mendenhall, W. and Sincich, T. (2003). *A Second Course in Statistics: Regression Analysis*. Prentice Hall / Pearson Education.

Merkle, E. C. and You, D. (2014). *nonnest2: Tests of Non-nested Models*. R package version 0.2.

Mersey, L. (1912). Report on the loss of the "Titanic" (S. S.). Parliamentary command paper 6352.

Meyer, D., Zeileis, A., and Hornik, K. (2006). The strucplot framework: Visualizing multi-way contingency tables with vcd. *Journal of Statistical Software*, 17(3), 1–48.

Meyer, D., Zeileis, A., and Hornik, K. (2015). *vcd: Visualizing Categorical Data*. R package version 1.3-3.

Milan, L. and Whittaker, J. (1995). Application of the parametric bootstrap to models that incorporate a singular value decomposition. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 44(1), 31–49.

Mirai Solutions GmbH (2015). *XLConnect: Excel Connector for R*. R package version 0.2-11.

Mosteller, F. and Wallace, D. L. (1963). Inference in an authorship problem. *Journal of the American Statistical Association*, 58(302), 275–309.

Mosteller, F. and Wallace, D. L. (1984). *Applied Bayesian and Classical Inference: The Case of the Federalist Papers*. New York, NY: Springer-Verlag.

Mullahy, J. (1986). Specification and testing of some modified count data models. *Journal of Econometrics*, 33, 341–365.

Murrell, P. (2011). *R Graphics*. Boca Raton, FL: Chapman & Hall/CRC.

Nelder, J. A. and Wedderburn, R. W. M. (1972). Generalized linear models. *Journal of the Royal Statistical Society, Series A*, 135, 370–384.

Neter, J., Wasserman, W., and Kutner, M. H. (1990). *Applied Linear Statistical Models : Regression, Analysis of Variance, and Experimental Designs*. Homewood, IL: R. D. Irwin, Inc., 3rd edn.

Noack, A. (1950). A class of random variables with discrete distributions. *Annals of Mathematical Statistics*, 21, 127–132.

Ord, J. K. (1967). Graphical methods for a class of discrete distributions. *Journal of the Royal Statistical Society, Series A*, 130, 232–238.

Pareto, V. (1971). *Manuale di economia politica ("Manual of political economy").* New York: A.M. Kelley. Translated by Ann S. Schwier. Edited by Ann S. Schwier and Alfred N. Page.

Pearson, K. (1900). On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen by random sampling. *Philosophical Magazine*, 50(5th Series), 157–175.

Peterson, B. and Harrell, Jr, F. E. (1990). Partial proportional odds models for ordinal response variables. *Applied Statistics*, 39, 205–217.

Pilhoefer, A. (2014). *extracat: Categorical Data Analysis and Visualization*. R package version 1.7-1.

Pinheiro, J., Bates, D., and R-core (2015). *nlme: Linear and Nonlinear Mixed Effects Models*. R package version 3.1-120.

Poole, J. H. (1989). Mate guarding, reproductive success and female choice in African elephants. *Animal Behavior*, 37, 842–849.

Powers, D. A. and Xie, Y. (2008). *Statistical Methods for Categorical Data Analysis*. Bingley, UK: Emerald, 2nd edn.

Pregibon, D. (1981). Logistic regression diagnostics. *Annals of Statistics*, 9, 705–724.

R Core Team (2015). *foreign: Read Data Stored by Minitab, S, SAS, SPSS, Stata, Systat, Weka, dBase, ...* R package version 0.8-63.

Raftery, A. E. (1986). Choosing models for cross-classifications. *American Sociological Review*, 51, 146–146.

Ramsey, F. L. and Schafer, D. W. (2002). *The Statistical Sleuth: A Course in Methods of Data Analysis*. Belmont, CA: Duxbury, 2nd edn.

Ramsey, F. L., Schafer, D. W., Sifneos, J., and Turlach, B. A. (2012). *Sleuth2: Data sets from Ramsey and Schafer's* Statistical Sleuth *(2nd ed).* R package version 1.0-7.

Revelle, W. (2015). *psych: Procedures for Psychological, Psychometric, and Personality Research*. R package version 1.5.1.

Riedwyl, H. and Schüpbach, M. (1983). Siebdiagramme: Graphische darstellung von kontingenztafeln. Tech. Rep. 12, Institute for Mathematical Statistics, University of Bern, Bern, Switzerland.

Riedwyl, H. and Schüpbach, M. (1994). Parquet diagram to plot contingency tables. In F. Faulbaum, ed., *Softstat '93: Advances In Statistical Software*, (pp. 293–299). New York: Gustav Fischer.

Ripley, B. (2015a). *MASS: Support Functions and Datasets for Venables and Ripley's MASS.* R package version 7.3-40.

Ripley, B. (2015b). *nnet: Feed-Forward Neural Networks and Multinomial Log-Linear Models*. R package version 7.3-9.

Roberto, C., Giordano, S., Cazzaro, M., and Lang, J. (2014). *hmmm: hierarchical multinomial marginal models*. R package version 1.0-3.

le Roux, N. and Lubbe, S. (2013). *UBbipl: Understanding Biplots: Data Sets And Functions*. R package version 3.0.4.

RStudio, Inc. (2011). *manipulate: Interactive Plots for RStudio*. R package version 0.98.977.

RStudio, Inc. (2015). *shiny: Web Application Framework for R*. R package version 0.11.1.

Sarkar, D. (2015). *lattice: Lattice Graphics*. R package version 0.20-31.

Schloerke, B., Crowley, J., Cook, D., Hofmann, H., Wickham, H., Briatte, F., Marbach, M., and Thoen, E. (2014). *GGally: Extension to ggplot2*. R package version 0.5.0.

Schwartz, G. (1978). Estimating the dimensions of a model. *Annals of Statistics*, 6, 461–464.

Searle, S. R., Speed, F. M., and Milliken, G. A. (1980). Population marginal means in the linear model: An alternative to least squares means. *The American Statistician*, 34(4), 216–221.

Shneiderman, B. (1992). Tree visualization with treemaps: A 2-D space-filling approach. *ACM Transactions on Graphics*, 11(1), 92–99.

Shrout, P. E. and Fleiss, J. L. (1979). Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin*, 86, 420–428.

Simpson, E. H. (1951). The interpretation of interaction in contingency tables. *Journal of the Royal Statistical Society, Series B*, 30, 238–241.

Skellam, J. G. (1948). A probability distribution derived from the binomial distribution by regarding the probability of success as variable between the sets of trials. *Journal of the Royal Statistical Society. Series B (Methodological)*, 10(2), 257–261.

Snee, R. D. (1974). Graphical display of two-way contingency tables. *The American Statistician*, 28, 9–12.

Snow, G. (2013). *TeachingDemos: Demonstrations for teaching and learning*. R package version 2.9.

Spence, I. (1990). Visual psychophysics of simple graphical elements. *Journal of Experimental Psychology: Human Perception and Performance*, 16, 683–692.

Spence, I. and Lewandowsky, S. (1990). Graphical perception. In J. Fox and J. S. Long, eds., *Modern Methods of Data Analysis*, chap. 1, (pp. 13–57). SAGE Publications.

Srole, L., Langner, T. S., Michael, S. T., Kirkpatrick, P., Opler, M. K., and Rennie, T. A. C. (1978). *Mental Health in the Metropolis: The Midtown Manhattan Study*. New York: NYU Press.

Stokes, M. E., Davis, C. S., and Koch, G. G. (2000). *Categorical Data Analysis Using the SAS System*. Cary, NC: SAS Institute, 2nd edn.

Stubben, C., Milligan, B., and Nantel, P. (2012). *popbio: Construction and analysis of matrix population models*. R package version 2.4.

Temple Lang, D., Swayne, D., Wickham, H., and Lawrence, M. (2014). *rggobi: Interface between R and GGobi*. R package version 2.1.20.

Theus, M. and Lauer, S. R. W. (1999). Visualizing loglinear models. *Journal of Computational and Graphical Statistics*, 8(3), 396–412.

Thornes, B. and Collard, J. (1979). *Who Divorces?* London: Routledge & Kegan.

Thurstone, L. L. (1927). A law of comparitive judgment. *Psychological Review*, 34, 278–286.

Tufte, E. (2006). *Beautiful Evidence*. Cheshire, CT: Graphics Press.

Tufte, E. R. (1983). *The Visual Display of Quantitative Information*. Cheshire, CT: Graphics Press.

Tufte, E. R. (1990). *Envisioning Information*. Cheshire, CT: Graphics Press.

Tufte, E. R. (1997). *Visual Explanations: Images and Quantities, Evidence and Narrative*. Cheshire, CT: Graphics Press.

Tukey, J. W. (1962). The future of data analysis. *Annals of Mathematical Statistics*, 33, 1–67 and 81.

Tukey, J. W. (1977). *Exploratory Data Analysis*. Reading, MA: Addison Wesley.

Tukey, J. W. (1990). Data-based graphics: Visual display in the decades to come. *Statistical Science*, 5(3), 327–339.

Tukey, J. W. (1993). Graphic comparisons of several linked aspects: Alternative and suggested principles. *Journal of Computational and Graphical Statistics*, 2(1), 1–33.

Turner, H. and Firth, D. (2014). *gnm: Generalized Nonlinear Models*. R package version 1.0-7.

Upton, G. J. G. (1976). The diagrammatic representation of three-party contests. *Political Studies*, 24, 448–454.

Upton, G. J. G. (1994). Picturing the 1992 British general election. *Journal of the Royal Statistical Society, Series A*, 157(Part 2), 231–252.

Urbanek, S. and Wichtrey, T. (2013). *iplots: iPlots - interactive graphics for R*. R package version 1.1-7.

Vaidyanathan, R. (2013). *rCharts: Interactive Charts using Javascript Visualization Libraries*. R package version 0.4.5.

Von Eye, A. and Mun, E. (2006). *Analyzing Rater Agreement: Manifest Variable Methods*. New York: Psychology Press, Taylor & Francis.

Vuong, Q. H. (1989). Likelihood ratio tests for model selection and non-nested hypotheses. *Econometrica*, 57(2), pp. 307–333.

Wainer, H. (1996). Using trilinear plots for NAEP state data. *Journal of Educational Measurement*, 33(1), 41–55.

Wand, M. (2015). *KernSmooth: Functions for Kernel Smoothing Supporting Wand & Jones (1995)*. R package version 2.23-14.

Wang, P. C. (1985). Adding a variable in generalized linear models. *Technometrics*, 27, 273–276.

Warnes, G. R., Bolker, B., Gorjanc, G., Grothendieck, G., Korosec, A., Lumley, T., MacQueen, D., Magnusson, A., Rogers, J., and others (2014). *gdata: Various R programming tools for data manipulation*. R package version 2.13.3.

Wei, T. (2013). *corrplot: Visualization of a correlation matrix*. R package version 0.73.

Wickham, H. (2009). *ggplot2: Elegant Graphics for Data Analysis*. Springer New York.

Wickham, H. (2014a). *plyr: Tools for splitting, applying and combining data*. R package version 1.8.1.

Wickham, H. (2014b). *reshape2: Flexibly Reshape Data: A Reboot of the Reshape Package.* R package version 1.4.1.

Wickham, H. and Chang, W. (2015). *ggplot2: An Implementation of the Grammar of Graphics*. R package version 1.0.1.

Wilkinson, L. (2005). *The Grammar of Graphics*. New York: Springer, 2nd edn.

Williams, D. A. (1987). Generalized linear model diagnostics using the deviance and single case deletions. *Applied Statistics*, 36, 181–191.

Wimmer, G. and Altmann, G. (1999). *Thesaurus of univariate discrete probability distributions*. Essen: Stamm.

Wong, R. S.-K. (2001). Multidimensional association models: A multilinear approach. *Sociological Methods and Research*, 30(2), 197–240.

Wong, R. S.-K. (2010). *Association Models*. Quantitative Applications in the Social Sciences. Los Angeles: SAGE Publications.

Wood, S. (2015). *mgcv: Mixed GAM Computation Vehicle with GCV/AIC/REML Smoothness Estimation*. R package version 1.8-6.

Wood, S. N. (2004). Stable and efficient multiple smoothing parameter estimation for generalized additive models. *Journal of the American Statistical Association*, 99(467), 673–686.

Wood, S. N. (2006). *Generalized Additive Models: An Introduction with R*. Chapman and Hall/CRC Press.

Woolf, B. (1995). On estimating the relation between blood group and disease. *Annals of Human Genetics*, 19, 251–253.

Wright, K. (2013). Revisiting Immer's barley data. *The American Statistician*, 67(3), 129–133.

Wright, K. (2015). *agridat: Agricultural Datasets*. R package version 1.11.

Xie, Y. (2014). *animation: A gallery of animations in statistics and utilities to create animations*. R package version 2.3.

Xie, Y. (2015). *knitr: A General-Purpose Package for Dynamic Report Generation in R*. R package version 1.9.

Yee, T. W. (2015). *VGAM: Vector Generalized Linear and Additive Models*. R package version 0.9-7.

Yip, K. C. and Yau, K. K. (2005). On modeling claim frequency data in general insurance with extra zeros. *Insurance: Mathematics and Economics*, 36(2), 153–163.

Zeileis, A. (2004). Econometric computing with hc and hac covariance matrix estimators. *Journal of Statistical Software*, 11(10), 1–17.

Zeileis, A. (2006). Object-oriented computation of sandwich estimators. *Journal of Statistical Software*, 16(9), 1–16.

Zeileis, A., Hornik, K., and Murrell, P. (2009). Escaping RGBland: Selecting colors for statistical graphics. *Computational Statistics & Data Analysis*, 53, 3259–3270.

Zeileis, A. and Kleiber, C. (2014). *countreg: Count Data Regression*. R package version 0.1-2/r88.

Zeileis, A., Kleiber, C., and Jackman, S. (2008). Regression models for count data in R. *Journal of Statistical Software*, 27(8).

Zeileis, A., Meyer, D., and Hornik, K. (2007). Residual-based shadings for visualizing (conditional) independence. *Journal of Computational and Graphical Statistics*, 16(3), 507–525.

Zelterman, D. (1999). *Models for Discrete Data*. New York: Oxford University Press.

Colophon

This book was produced using R version 3.2.1 (2015-06-18) and knitr (1.11). Writing, editing and compositing was done using RStudio. Hence, we can be assured that the code examples produced the output in the text.

The principal R package versions used in examples and illustrations are listed below. At the time of writing, most of these were current on CRAN repositories (e.g., `http://cran.us.r-project.org/`) but some development versions are indicated in the "source" column. "R-Forge" refers to the development platform (`https://r-forge.r-project.org`) used by many package authors to prepare and test new versions. By the time you read this, most of these should be current on CRAN.

package	version	date	source
AER	1.2-4	2015-06-06	CRAN
ca	0.58	2014-12-31	CRAN
car	2.1-0	2015-09-03	CRAN
colorspace	1.2-6	2015-03-11	CRAN
corrplot	0.73	2013-10-15	CRAN
countreg	0.1-3	2015-04-18	R-Forge
directlabels	2013.6.15	2013-07-23	CRAN
effects	3.0-5	2015-09-10	R-Forge
ggparallel	0.1.2	2015-08-21	CRAN
ggplot2	1.0.1	2015-03-17	CRAN
ggtern	1.0.6.1	2015-10-12	CRAN
gmodels	2.16.2	2015-07-22	CRAN
gnm	1.0-8	2015-04-22	CRAN
gpairs	1.2	2014-03-09	CRAN
heplots	1.0-16	2015-07-13	CRAN
Lahman	4.0-1	2015-09-15	CRAN
lattice	0.20-33	2015-07-14	CRAN
lmtest	0.9-34	2015-06-06	CRAN
logmult	0.6.2	2015-04-22	CRAN
MASS	7.3-45	2015-11-10	CRAN
mgcv	1.8-9	2015-10-30	CRAN
nnet	7.3-11	2015-08-30	CRAN
plyr	1.8.3	2015-06-12	CRAN
pscl	1.4.9	2015-03-29	CRAN
RColorBrewer	1.1-2	2014-12-07	CRAN
reshape2	1.4.1	2014-12-06	CRAN
rms	4.4-0	2015-09-28	CRAN
rsm	2.7-4	2015-10-07	CRAN
sandwich	2.3-4	2015-09-24	CRAN
vcd	1.4-2	2015-10-18	R-Forge
vcdExtra	0.6-11	2015-09-17	CRAN
VGAM	1.0-0	2015-10-29	CRAN
xtable	1.8-0	2015-11-02	CRAN

To prepare your R installation for running the examples in this book, you can use the following commands to install these packages.

```
> packages <- c("AER", "ca", "car", "colorspace", "corrplot", "countreg",
+     "directlabels", "effects", "ggparallel", "ggplot2", "ggtern",
+     "gmodels", "gnm", "gpairs", "heplots", "Lahman", "lattice", "lmtest",
+     "logmult", "MASS", "mgcv", "nnet", "plyr", "pscl", "RColorBrewer",
+     "reshape2", "rms", "rsm", "sandwich", "splines", "vcd", "vcdExtra",
+     "VGAM", "xtable")
> install.packages(packages)
> # if countreg is not yet on CRAN:
> install.packages("countreg", repos = "http://R-Forge.R-project.org")
```

Discrete Data Analysis with R: Visualization and Modeling Techniques for Categorical and Count Data presents an applied treatment of modern methods for the analysis of categorical data, both discrete response data and frequency data. It explains how to use graphical methods for exploring data, spotting unusual features, visualizing fitted models, and presenting results. Along with describing the necessary statistical theory, the authors illustrate the practical application of the techniques to a large number of substantive problems, including how to organize data, conduct an analysis, produce informative graphs, and evaluate what the graphs reveal about the data.

The first part of the book contains introductory material on graphical methods for discrete data, basic R skills, and methods for fitting and visualizing one-way discrete distributions. The second part focuses on simple, traditional nonparametric tests and exploratory methods for visualizing patterns of association in two-way and larger frequency tables. The final part of the text discusses model-based methods for the analysis of discrete data.

Features

- Provides an accessible introduction to the major methods of categorical data analysis for data exploration, statistical testing, and statistical models
- Emphasizes computing, visualizing, understanding, and communicating the results of the analyses
- Helps you translate theory into practical application by developing your skills and providing software tools for carrying out the methods
- Includes many examples using real data, which are often treated from several perspectives
- Contains lab exercises that encourage you to work through applications of the methods
- Offers the data sets and R code on a supplementary website

CRC Press
Taylor & Francis Group
an **informa** business
www.crcpress.com

6000 Broken Sound Parkway, NW
Suite 300, Boca Raton, FL 33487
711 Third Avenue
New York, NY 10017
2 Park Square, Milton Park
Abingdon, Oxon OX14 4RN, UK

K25797
ISBN: 978-1-4987-2583-5
90000
9 781498 725835
www.crcpress.com