

Chapter 11

Poisson Regression

If the response variable Y is a count, then the Poisson regression model is often useful. For example, counts often occur in wildlife studies where a region is divided into subregions and Y_i is the number of a specified type of animal found in the subregion. The following definition makes simulation of Poisson regression data simple. See Section 1.3.

11.1 Poisson Regression

Definition 11.1. The **Poisson regression model** states that Y_1, \dots, Y_n are independent random variables with

$$Y_i \sim \text{Poisson}(\mu(\mathbf{x}_i)).$$

The **loglinear Poisson regression (LLR) model** is the special case where

$$\mu(\mathbf{x}_i) = \exp(\alpha + \boldsymbol{\beta}^T \mathbf{x}_i). \quad (11.1)$$

Model (11.1) can be written compactly as $Y|SP \sim \text{Poisson}(\exp(SP))$. Notice that the conditional mean and variance functions are equal: $E(Y|SP) = V(Y|SP) = \exp(SP)$. For the LLR model, the Y are independent and

$$Y \approx \text{Poisson}(\exp(ESP)),$$

or $Y|SP \approx Y|ESP \approx \text{Poisson}(\hat{\mu}(ESP))$. For example, $Y|(SP = 0) \sim \text{Poisson}(1)$, and $Y|(ESP = 0) \approx \text{Poisson}(1)$.

In the response plot for loglinear regression, the shape of the estimated mean function $\hat{\mu}(ESP) = \exp(ESP)$ depends strongly on the range of the ESP. The variety of shapes occurs because the plotting software attempts to fill the vertical axis. Hence the range of the ESP is narrow, then the exponential function will be rather flat. If the range of the ESP is wide, then the exponential curve will look flat in the left of the plot but will increase sharply in the right of the plot.

Definition 11.2. The estimated sufficient summary plot (ESSP) or *response plot*, is a plot of the $ESP = \hat{\alpha} + \hat{\boldsymbol{\beta}}^T \mathbf{x}_i$ versus Y_i with the estimated mean function

$$\hat{\mu}(ESP) = \exp(ESP)$$

added as a visual aid. A scatterplot smoother such as lowess is also added as a visual aid.

This plot is very useful as a goodness of fit diagnostic. The lowess curve is a nonparametric estimator of the mean function and is represented as a jagged curve to distinguish it from the estimated LLR mean function (the exponential curve) in Figure 1.9. If the number of predictors $k < n/10$, if there is no overdispersion, and if the lowess curve follows the exponential curve closely (except possibly for the largest values of the ESP), then the LLR mean function may be a useful approximation for $E(Y|\mathbf{x})$. **A useful lack of fit plot** is a plot of the ESP versus the *deviance residuals* that are often available from the software.

The deviance test described in Section 11.2 is used to test whether $\boldsymbol{\beta} = \mathbf{0}$, and is the analog of the ANOVA F test for multiple linear regression. If the LLR model is a good approximation to the data but $\boldsymbol{\beta} = \mathbf{0}$, then the predictors \mathbf{x} are not needed in the model and $\hat{\mu}(\mathbf{x}_i) \equiv \hat{\mu} = \bar{Y}$ (the sample mean) should be used instead of the LLR estimator

$$\hat{\mu}(\mathbf{x}_i) = \exp(\hat{\alpha} + \hat{\boldsymbol{\beta}}^T \mathbf{x}_i).$$

If the exponential curve clearly fits the lowess curve better than the line $Y = \bar{Y}$, then H_o should be rejected, but if the line $Y = \bar{Y}$ fits the lowess curve about as well as the exponential curve (which should only happen if the exponential curve is approximately linear with a small slope), then Y may be independent of the predictors. Figure 1.10 shows the ESSP when only X_4 and X_5 are used as predictors for the artificial data, and Y is independent of

these two predictors by construction. It is possible to find data sets that look like Figure 1.10 where the p-value for the deviance test is very small. Then the LLR relationship is statistically significant, but the investigator needs to decide whether the relationship is practically significant.

Warning: For many count data sets where the LLR mean function is correct, the LLR model is not appropriate but the LLR MLE is still a consistent estimator of β . The problem is that for many data sets where $E(Y|\mathbf{x}) = \mu(\mathbf{x}) = \exp(SP)$, it turns out that $V(Y|\mathbf{x}) > \exp(SP)$. This phenomenon is called **overdispersion**. Adding parametric and nonparametric estimators of the standard deviation function to the response plot can be useful. See Cook and Weisberg (1999a, p. 401-403). Alternatively, if the response plot looks good and $G^2/(n - k - 1) \approx 1$, then the LLR model is likely useful. Here the deviance G^2 is described in Section 11.2.

A useful alternative to the LLR model is a negative binomial regression (NBR) model. If Y has a (generalized) negative binomial distribution, $Y \sim NB(\mu, \kappa)$, then the probability mass function of Y is

$$P(Y = y) = \frac{\Gamma(y + \kappa)}{\Gamma(\kappa)\Gamma(y + 1)} \left(\frac{\kappa}{\mu + \kappa} \right)^\kappa \left(1 - \frac{\kappa}{\mu + \kappa} \right)^y$$

for $y = 0, 1, 2, \dots$ where $\mu > 0$ and $\kappa > 0$. Then $E(Y) = \mu$ and $V(Y) = \mu + \mu^2/\kappa$. (This distribution is a generalization of the negative binomial (κ, ρ) distribution with $\rho = \kappa/(\mu + \kappa)$ and $\kappa > 0$ is an unknown real parameter rather than a known integer.)

Definition 11.3. The **negative binomial regression (NBR) model** states that Y_1, \dots, Y_n are independent random variables where

$$Y_i \sim NB(\mu(\mathbf{x}_i), \kappa)$$

with $\mu(\mathbf{x}_i) = \exp(\alpha + \beta^T \mathbf{x}_i)$. Hence $Y|SP \sim NB(\exp(SP), \kappa)$, $E(Y|SP) = \exp(SP)$ and

$$V(Y|SP) = \exp(SP) \left(1 + \frac{\exp(SP)}{\kappa} \right).$$

The NBR model has the same mean function as the LLR model but allows for overdispersion. As $\kappa \rightarrow \infty$, the NBR model converges to the LLR model. Since the Poisson regression model is simpler than the NBR model, graphical

diagnostics for the goodness of fit of the LLR model would be useful. The following plot was suggested by Winkelmann (2000, p. 110).

Definition 11.4. To check for overdispersion, use the **OD plot** of the estimated model variance $\hat{V}(Y|SP)$ versus the squared residuals $\hat{V} = [Y - \hat{E}(Y|SP)]^2$. For the LLR model, $\hat{V}(Y|SP) = \exp(ESP) = \hat{E}(Y|SP)$ and $\hat{V} = [Y - \exp(ESP)]^2$.

Numerical summaries are also available. The deviance G^2 is a statistic used to assess the goodness of fit of the Poisson regression model much as R^2 is used for multiple linear regression. For Poisson regression, G^2 is approximately chi-square with $n - p - 1$ degrees of freedom. Since a χ_d^2 random variable has mean d and standard deviation $\sqrt{2d}$, the 98th percentile of the χ_d^2 distribution is approximately $d + 3\sqrt{d} \approx d + 2.121\sqrt{2d}$. If $G^2 > (n - p - 1) + 3\sqrt{n - p - 1}$, then a more complicated count model than (11.1) may be needed. A good discussion of such count models is in Simonoff (2003).

For model (11.1), Winkelmann (2000, p. 110) suggested that the plotted points in the OD plot should scatter about the identity line through the origin with unit slope and that the OLS line should be approximately equal to the identity line if the LLR model is appropriate. But in simulations, it was found that the following two observations make the OD plot much easier to use for Poisson regression.

First, recall that a normal approximation is good for both the Poisson and negative binomial distributions if the count Y is not too small. Notice that if $Y = E(Y|SP) + 2\sqrt{V(Y|SP)}$, then $[Y - E(Y|SP)]^2 = 4V(Y|SP)$. Hence if both the estimated mean and estimated variance functions are good approximations, the plotted points in the OD plot for Poisson regression will scatter about a wedge formed by the $\hat{V} = 0$ line and the line through the origin with slope 4: $\hat{V} = 4\hat{V}(Y|SP)$. If the normal approximation is good, only about 5% of the plotted points should be above this line.

Second, the evidence of overdispersion increases from slight to high as the scale of the vertical axis increases from 5 to 10 times that of the horizontal axis. (The scale of the vertical axis tends to depend on the few cases with the largest $\hat{V}(Y|SP)$, and $P[(Y - \hat{E}(Y|SP))^2 > 10\hat{V}(Y|SP)]$ can be approximated with a normal approximation or Chebyshev's inequality.) There is considerable evidence of overdispersion if the scale of the vertical axis is more than 10 times that of the horizontal, or if the percentage of points above the slope 4 line through the origin is much larger than 5%. Hence the identity

line and slope 4 line are added to the OD plot as visual aids, and one should check whether the scale of the vertical axis is more than 10 times that of the horizontal.

Combining the response plot with the OD plot is a powerful method for assessing the adequacy of the Poisson regression model. It is easier to use the OD plot to check the variance function than the response plot since judging the variance function with the straight lines of the OD plot is simpler than judging two curves. Also outliers are often easier to spot with the OD plot.

For LLR Poisson regression, judging the mean function from the ESSP may be rather difficult for large counts for two reasons. First, the mean function is curved. Secondly, for real and simulated Poisson regression data, it was observed that scatterplot smoothers such as lowess tend to underestimate the mean function for large ESP.

The basic idea of the following two plots for Poisson regression is to transform the data towards a linear model, then make the response plot and residual plot for the transformed data. The plots are based on weighted least squares (WLS) regression. For the equivalent least squares (OLS) regression without intercept of W on \mathbf{u} , the ESSP is the (weighted fit) response plot of \hat{W} versus W . The mean function is the identity line and the vertical deviations from the identity line are the WLS residuals $W - \hat{W}$. Since $P(Y_i = 0) > 0$, the estimators given in the following definition are useful. Let $Z_i = Y_i$ if $Y_i > 0$, and let $Z_i = 0.5$ if $Y_i = 0$.

Definition 11.5. The **minimum chi-square estimator** of the parameters $(\alpha, \boldsymbol{\beta})$ in a loglinear regression model are $(\hat{\alpha}_M, \hat{\boldsymbol{\beta}}_M)$, and are found from the weighted least squares regression of $\log(Z_i)$ on \mathbf{x}_i with weights $w_i = Z_i$. Equivalently, use the ordinary least squares (OLS) regression (without intercept) of $\sqrt{Z_i} \log(Z_i)$ on $\sqrt{Z_i}(1, \mathbf{x}_i^T)^T$.

The minimum chi-square estimator tends to be consistent if n is fixed and all n counts Y_i increase to ∞ , while the loglinear regression maximum likelihood estimator $(\hat{\alpha}, \hat{\boldsymbol{\beta}})$ tends to be consistent if the sample size $n \rightarrow \infty$. See Agresti (2002, p. 611-612). However, the two estimators are often close for many data sets. Use the OLS regression (without intercept) of $\sqrt{Z_i} \log(Z_i)$ on $\sqrt{Z_i}(1, \mathbf{x}_i^T)^T$. Then the plot of the “fitted values” $\sqrt{Z_i}(\hat{\alpha}_M + \hat{\boldsymbol{\beta}}_M^T \mathbf{x}_i)$ versus the “response” $\sqrt{Z_i} \log(Z_i)$ should have points that scatter about the identity line. These results and the equivalence of the minimum chi-square estimator to an OLS estimator suggest the following diagnostic plots.

Definition 11.6. For a loglinear Poisson regression model, a **weighted fit response plot** is a plot of $\sqrt{Z_i}ESP = \sqrt{Z_i}(\hat{\alpha} + \hat{\beta}^T \mathbf{x}_i)$ versus $\sqrt{Z_i} \log(Z_i)$. The **weighted residual plot** is a plot of $\sqrt{Z_i}(\hat{\alpha} + \hat{\beta}^T \mathbf{x}_i)$ versus the “WLS” residuals $r_{Wi} = \sqrt{Z_i} \log(Z_i) - \sqrt{Z_i}(\hat{\alpha} + \hat{\beta}^T \mathbf{x}_i)$.

If the loglinear regression model is appropriate and the LLR estimator is good, then the plotted points in the weighted fit response plot should follow the identity line. When the counts Y_i are small, the “WLS” residuals can not be expected to be approximately normal. Often the larger counts are fit better than the smaller counts and hence the residual plots have a “left opening megaphone” shape. This fact makes residual plots for Poisson regression rather hard to use, but cases with large “WLS” residuals may not be fit very well by the model. Both the weighted fit response and residual plots perform better for simulated LLR data with many large counts than for data where all of the counts are less than 10.

Example 11.1. For the Ceriodaphnia data of Myers, Montgomery and Vining (2002, p. 136-139), the response variable Y is the number of Ceriodaphnia organisms counted in a container. The sample size was $n = 70$ and seven concentrations of jet fuel (x_1) and an indicator for two strains of organism (x_2) were used as predictors. The jet fuel was believed to impair reproduction so high concentrations should have smaller counts. Figure 11.1 shows the 4 plots for this data. In the response plot of Figure 11.1a, the lowess curve is represented as a jagged curve to distinguish it from the estimated LLR mean function (the exponential curve). The horizontal line corresponds to the sample mean \bar{Y} .

The OD plot in Figure 11.1b suggests that there is little evidence of overdispersion since the vertical scale is less than ten times that of the horizontal scale and all but one of the plotted points are close to the wedge formed by the horizontal axis and slope 4 line. The plotted points scatter about the identity line in Figure 11.1c and there are no unusual points in Figure 11.1d. The four plots suggest that the LLR Poisson regression model is a useful approximation to the data. Hence $Y|ESP \approx \text{Poisson}(\exp(ESP))$. For example, when $ESP = 1.61$, $Y \approx \text{Poisson}(5)$ and when $ESP = 4.5$, $Y \approx \text{Poisson}(90)$. Notice that the Poisson mean can be roughly estimated by finding the height of the exponential curve in Figure 11.1a.

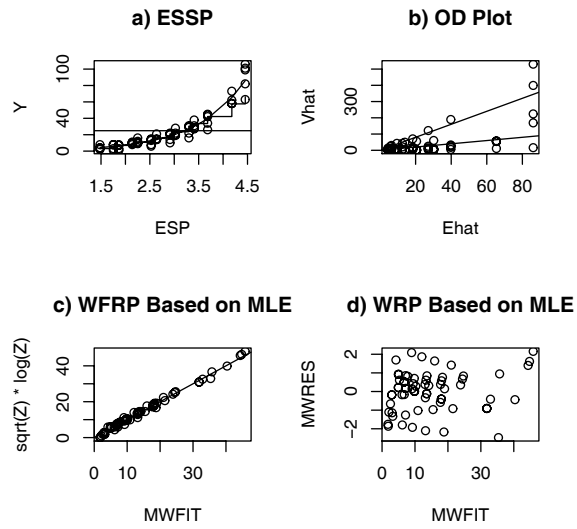


Figure 11.1: Plots for Ceriodaphnia Data

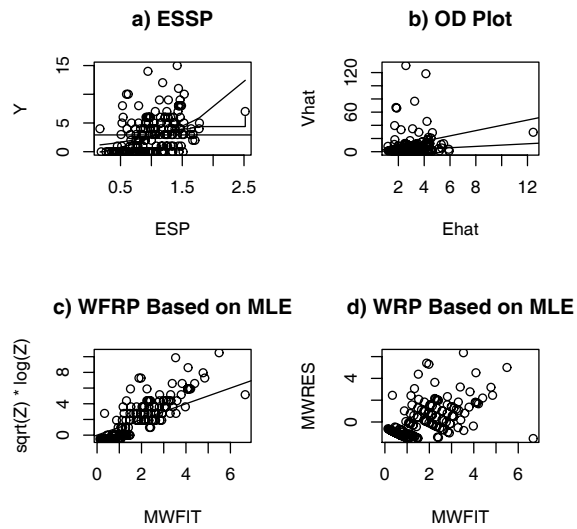


Figure 11.2: Plots for Crab Data

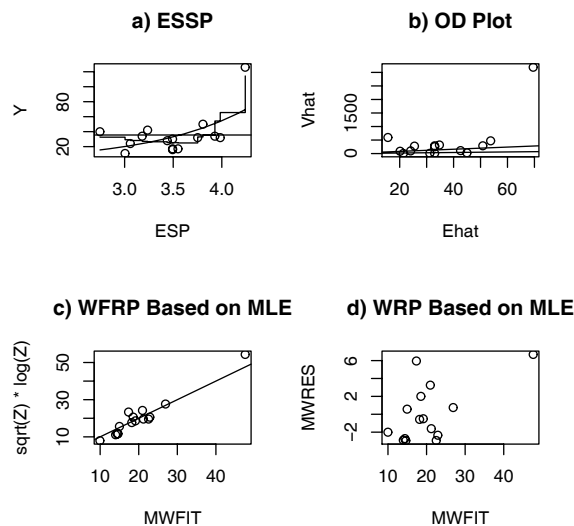


Figure 11.3: Plots for Popcorn Data

Example 11.2. Agresti (2002, p. 126-131) uses Poisson regression for data where the response Y is the number of satellites (male crabs) near a female crab. The sample size $n = 173$ and the predictor variables were the *color* (2: light medium, 3: medium, 4: dark medium, 5: dark), *spine condition* (1: both good, 2: one worn or broken, 3 both worn or broken), carapace *width* in cm and *weight* of the female crab in grams.

The model used to produce Figure 11.2 used the ordinal variables color and spine condition as coded. An alternative model would use spine condition as a factor. Figure 11.2a suggests that there is one case with an unusually large value of the ESP. Notice that the lowess curve does not track the exponential curve very well. Figure 11.2b suggests that overdispersion is present since the vertical scale is about 10 times that of the horizontal scale and too many of the plotted points are large and higher than the slope 4 line. The lack of fit may be clearer in Figure 11.2c since the plotted points fail to cover the identity line. Although the exponential mean function fits the lowess curve better than the line $Y = \bar{Y}$, alternative models suggested by Agresti (2002) may fit the data better.

Example 11.3. For the popcorn data of Myers, Montgomery and Vining (2002, p. 154), the response variable Y is the number of inedible popcorn

kernels. The sample size was $n = 15$ and the predictor variables were *temperature* (coded as 5, 6 or 7), amount of *oil* (coded as 2, 3 or 4) and popping *time* (75, 90 or 105). One batch of popcorn had more than twice as many inedible kernels as any other batch and is an outlier that is easily detected in all four plots in Figure 11.3. Ignoring the outlier in Figure 11.3a suggests that the line $Y = \bar{Y}$ will fit the data and lowess curve better than the exponential curve. Hence Y seems to be independent of the predictors. Notice that the outlier sticks out in Figure 11.3b and that the vertical scale is well over 10 times that of the horizontal scale. If the outlier was not detected, then the Poisson regression model would suggest that temperature and time are important predictors, and overdispersion diagnostics such as the deviance would be greatly inflated.

11.2 Inference

This section gives a brief discussion of inference for the loglinear Poisson regression (LLR) model. Inference for this model is very similar to inference for the multiple linear regression, survival regression and logistic regression models. For all of these models, Y is independent of the $k \times 1$ vector of predictors $\mathbf{x} = (x_1, \dots, x_k)^T$ given the sufficient predictor $\alpha + \boldsymbol{\beta}^T \mathbf{x}$:

$$Y \perp \mathbf{x} | (\alpha + \boldsymbol{\beta}^T \mathbf{x}).$$

To perform inference for LLR, computer output is needed. The computer output looks nearly identical to that needed for logistic regression.

Point estimators for the mean function are important. Given values of $\mathbf{x} = (x_1, \dots, x_k)^T$, a major goal of loglinear regression is to estimate the mean $E(Y|\mathbf{x}) = \mu(\mathbf{x})$ with the estimator

$$\hat{\mu}(\mathbf{x}) = \exp(\hat{\alpha} + \hat{\boldsymbol{\beta}}^T \mathbf{x}). \quad (11.2)$$

Investigators also sometimes test whether a predictor X_j is needed in the model given that the other $k - 1$ nontrivial predictors are in the model with a **4 step Wald test of hypotheses**:

- i) State the hypotheses $H_0: \beta_j = 0$ $H_a: \beta_j \neq 0$.
- ii) Find the test statistic $z_{o,j} = \hat{\beta}_j / se(\hat{\beta}_j)$ or obtain it from output.
- iii) The p-value = $2P(Z < -|z_{o,j}|) = 2P(Z > |z_{o,j}|)$. Find the p-value from output or use the standard normal table.

iv) State whether you reject H_0 or fail to reject H_0 and give a nontechnical sentence restating your conclusion in terms of the story problem.

If H_0 is rejected, then conclude that X_j is needed in the LLR model for Y given that the other $k - 1$ predictors are in the model. If you fail to reject H_0 , then conclude that X_j is not needed in the LLR model for Y given that the other $k - 1$ predictors are in the model. Note that X_j could be a very useful LLR predictor, but may not be needed if other predictors are added to the model.

The Wald confidence interval (CI) for β_j can also be obtained from the output: the large sample $100(1 - \delta)\%$ CI for β_j is $\hat{\beta}_j \pm z_{1-\delta/2} se(\hat{\beta}_j)$.

The Wald test and CI tend to give good results if the sample size n is large. Here $1 - \delta$ refers to the coverage of the CI. Recall that a 90% CI uses $z_{1-\delta/2} = 1.645$, a 95% CI uses $z_{1-\delta/2} = 1.96$, and a 99% CI uses $z_{1-\delta/2} = 2.576$.

For a LLR, often 3 models are of interest: the **full model** that uses all k of the predictors $\mathbf{x}^T = (\mathbf{x}_R^T, \mathbf{x}_O^T)$, the **reduced model** that uses the r predictors \mathbf{x}_R , and the **saturated model** that uses n parameters $\theta_1, \dots, \theta_n$ where n is the sample size. For the full model the $k + 1$ parameters $\alpha, \beta_1, \dots, \beta_k$ are estimated while the reduced model has $r + 1$ parameters. Let $l_{SAT}(\theta_1, \dots, \theta_n)$ be the likelihood function for the saturated model and let $l_{FULL}(\alpha, \boldsymbol{\beta})$ be the likelihood function for the full model. Let

$$L_{SAT} = \log l_{SAT}(\hat{\theta}_1, \dots, \hat{\theta}_n)$$

be the log likelihood function for the saturated model evaluated at the maximum likelihood estimator (MLE) $(\hat{\theta}_1, \dots, \hat{\theta}_n)$ and let

$$L_{FULL} = \log l_{FULL}(\hat{\alpha}, \hat{\boldsymbol{\beta}})$$

be the log likelihood function for the full model evaluated at the MLE $(\hat{\alpha}, \hat{\boldsymbol{\beta}})$. Then the **deviance**

$$D = G^2 = -2(L_{FULL} - L_{SAT}).$$

The degrees of freedom for the deviance $= df_{FULL} = n - k - 1$ where n is the number of parameters for the saturated model and $k + 1$ is the number of parameters for the full model.

The saturated model for loglinear regression states that Y_1, \dots, Y_n are independent $\text{Poisson}(\mu_i)$ random variables where $\hat{\mu}_i = Y_i$. The saturated model is usually not very good for Poisson data, but the saturated model may be good if n is fixed and all of the counts Y_i are large.

If $X \sim \chi_d^2$ then $E(X) = d$ and $\text{VAR}(X) = 2d$. An observed value of $x > d + 3\sqrt{d}$ is unusually large and an observed value of $x < d - 3\sqrt{d}$ is unusually small.

When the saturated model is good, a rule of thumb is that the loglinear regression model is ok if $G^2 \leq n - k - 1$ (or if $G^2 \leq n - k - 1 + 3\sqrt{n - k - 1}$). The χ_{n-k+1}^2 approximation for G^2 may not be good even for large sample sizes n . For LLR, the response and OD plots and $G^2 \leq n - k - 1 + 3\sqrt{n - k - 1}$ should be checked.

The *Arc* output below, shown in symbols and for a real data set, is used for the deviance test described after the output. Assume that the estimated sufficient summary plot has been made and that the loglinear regression model fits the data well in that the lowess estimated mean function follows the estimated model mean function closely. The deviance test is used to test whether $\beta = \mathbf{0}$. If this is the case, then the predictors are not needed in the LLR model. If $H_o : \beta = \mathbf{0}$ is not rejected, then for loglinear regression the estimator $\hat{\mu} = \bar{Y}$ should be used.

Response = Y
 Terms = (X_1, \dots, X_k)
 Sequential Analysis of Deviance

Predictor	df	Total Deviance	df	Change Deviance
Ones	$n - 1 = df_o$	G_o^2		
X_1	$n - 2$		1	
X_2	$n - 3$		1	
\vdots	\vdots	\vdots	\vdots	
X_k	$n - k - 1 = df_{FULL}$	G_{FULL}^2	1	

The 4 step **deviance test** follows.

- i) $H_o : \beta = \mathbf{0}$ $H_A : \beta \neq \mathbf{0}$
- ii) test statistic $G^2(o|F) = G_o^2 - G_{FULL}^2$
- iii) The p-value = $P(\chi^2 > G^2(o|F))$ where $\chi^2 \sim \chi_k^2$ has a chi-square

distribution with k degrees of freedom. Note that $k = k + 1 - 1 = df_o - df_{FULL} = n - 1 - (n - k - 1)$.

iv) Reject H_o if the p-value $< \delta$ and conclude that there is a LLR relationship between Y and the predictors X_1, \dots, X_k . If p-value $\geq \delta$, then fail to reject H_o and conclude that there is not a LLR relationship between Y and the predictors X_1, \dots, X_k .

The output shown on the following page, both in symbols and for a real data set, can be used to perform the change in deviance test. If the reduced model leaves out a single variable X_i , then the change in deviance test becomes $H_o : \beta_i = 0$ versus $H_A : \beta_i \neq 0$. This likelihood ratio test is a competitor of the Wald test. The likelihood ratio test is usually better than the Wald test if the sample size n is not large, but the Wald test is currently easier for software to produce. For large n the test statistics from the two tests tend to be very similar (asymptotically equivalent tests).

Response = Y Terms = (X_1, \dots, X_k) (Full Model)

Label	Estimate	Std. Error	Est/SE	p-value
Constant	$\hat{\alpha}$	$se(\hat{\alpha})$	$z_{o,0}$	for Ho: $\alpha = 0$
x_1	$\hat{\beta}_1$	$se(\hat{\beta}_1)$	$z_{o,1} = \hat{\beta}_1/se(\hat{\beta}_1)$	for Ho: $\beta_1 = 0$
\vdots	\vdots	\vdots	\vdots	\vdots
x_k	$\hat{\beta}_k$	$se(\hat{\beta}_k)$	$z_{o,k} = \hat{\beta}_k/se(\hat{\beta}_k)$	for Ho: $\beta_k = 0$

Degrees of freedom: $n - k - 1 = df_{FULL}$

Deviance: $D = G_{FULL}^2$

Response = Y Terms = (X_1, \dots, X_r) (Reduced Model)

Label	Estimate	Std. Error	Est/SE	p-value
Constant	$\hat{\alpha}$	$se(\hat{\alpha})$	$z_{o,0}$	for Ho: $\alpha = 0$
x_1	$\hat{\beta}_1$	$se(\hat{\beta}_1)$	$z_{o,1} = \hat{\beta}_1/se(\hat{\beta}_1)$	for Ho: $\beta_1 = 0$
\vdots	\vdots	\vdots	\vdots	\vdots
x_r	$\hat{\beta}_r$	$se(\hat{\beta}_r)$	$z_{o,r} = \hat{\beta}_r/se(\hat{\beta}_r)$	for Ho: $\beta_r = 0$

Degrees of freedom: $n - r - 1 = df_{RED}$

Deviance: $D = G_{RED}^2$

If the reduced model is good, then the **EE plot** of $ESP(R) = \hat{\alpha}_R + \hat{\beta}_R^T \mathbf{x}_{Ri}$ versus $ESP = \hat{\alpha} + \hat{\beta}^T \mathbf{x}_i$ should be highly correlated with the identity line

with unit slope and zero intercept.

After obtaining an acceptable full model where

$$SP = \alpha + \beta_1 x_1 + \cdots + \beta_k x_k = \alpha + \boldsymbol{\beta}^T \mathbf{x} = \alpha + \boldsymbol{\beta}_R^T \mathbf{x}_R + \boldsymbol{\beta}_O^T \mathbf{x}_O$$

try to obtain a **reduced model**

$$SP = \alpha_R + \beta_{R1} x_{R1} + \cdots + \beta_{Rr} x_{Rr} = \alpha_R + \boldsymbol{\beta}_R^T \mathbf{x}_R$$

where the reduced model uses r of the predictors used by the full model and \mathbf{x}_O denotes the vector of $k - r$ predictors that are in the full model but not the reduced model. For loglinear regression the reduced model is $Y_i | \mathbf{x}_{Ri} \sim$ independent Poisson($\exp(\boldsymbol{\beta}_R^T \mathbf{x}_{Ri})$) for $i = 1, \dots, n$.

Assume that the full model looks good (so the response and OD plots look good). Then we want to test H_o : the reduced model is good (can be used instead of the full model) versus H_A : use the full model (the full model is significantly better than the reduced model). Fit the full model and the reduced model to get the deviances G_{FULL}^2 and G_{RED}^2 .

The 4 step **change in deviance test** follows.

- i) H_o : the reduced model is good H_A : use the full model
- ii) test statistic $G^2(R|F) = G_{RED}^2 - G_{FULL}^2$
- iii) The p-value = $P(\chi^2 > G^2(R|F))$ where $\chi^2 \sim \chi_{k-r}^2$ has a chi-square distribution with k degrees of freedom. Note that k is the number of non-trivial predictors in the full model while r is the number of nontrivial predictors in the reduced model. Also notice that $k - r = (k + 1) - (r + 1) = df_{RED} - df_{FULL} = n - r - 1 - (n - k - 1)$.
- iv) Reject H_o if the p-value $< \delta$ and conclude that the full model should be used. If p-value $\geq \delta$, then fail to reject H_o and conclude that the reduced model is good.

Interpretation of coefficients: if $x_1, \dots, x_{i-1}, x_{i+1}, \dots, x_k$ can be held fixed, then increasing x_i by 1 unit increases the sufficient predictor SP by β_i units. In loglinear Poisson regression, increasing a predictor x_i by 1 unit (while holding all other predictors fixed) multiplies the estimated mean function by a factor of $\exp(\hat{\beta}_i)$.

Label	Estimate	Std. Error	Est/SE	p-value
Constant	-0.406023	0.877382	-0.463	0.6435
bombload	0.165425	0.0675296	2.450	0.0143
exper	-0.0135223	0.00827920	-1.633	0.1024
type	0.568773	0.504297	1.128	0.2594

Example 11.4. Use the above output to perform inference on the number of locations where aircraft was damaged. The output is from a loglinear regression. The variable *exper* = total months of aircrew experience while type of aircraft was coded as 0 or 1. There were $n = 30$ cases. Data is from Montgomery, Peck and Vining (2001).

a) Predict $\hat{\mu}(\mathbf{x})$ if *bombload* = $x_1 = 7.0$, *exper* = $x_2 = 80.2$ and *type* = $x_3 = 1.0$.

b) Perform the 4 step Wald test for $H_0 : \beta_2 = 0$.

c) Find a 95% confidence interval for β_3 .

Solution: a) $ESP = \hat{\alpha} + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \hat{\beta}_3 x_3 = -0.406023 + 0.165426(7) - 0.0135223(80.2) + 0.568773(1) = 0.2362$. So $\hat{\mu}(\mathbf{x}) = \exp(ESP) = \exp(0.2360) = 1.2665$.

b) i) $H_0 : \beta_2 = 0$ $H_A : \beta_2 \neq 0$

ii) $t_{02} = -1.633$.

iii) $pval = 0.1024$

iv) Fail to reject H_0 , *exper* is not needed in the LLR model for number of locations given that *bombload* and *type* are in the model.

c) $\hat{\beta}_3 \pm 1.96SE(\hat{\beta}_3) = 0.568773 \pm 1.96(0.504297) = 0.568773 \pm 0.9884 = (-0.4196, 1.5572)$.

11.3 Variable Selection

This section gives some rules of thumb for variable selection for loglinear Poisson regression. Before performing variable selection, a useful full model needs to be found. The process of finding a useful full model is an iterative process. Given a predictor x , sometimes x is not used by itself in the full model.

The full model will often contain factors and interaction. If w is a nominal variable with J levels, make w into a factor by using $J - 1$ (indicator or) dummy variables $x_{1,w}, \dots, x_{J-1,w}$ in the full model. For example, let $x_{i,w} = 1$ if

w is at its i th level, and let $x_{i,w} = 0$, otherwise. An interaction is a product of two or more predictor variables. Interactions are difficult to interpret. Often interactions are included in the full model, and then the reduced model without any interactions is tested. The investigator is often hoping that the interactions are not needed.

A **scatterplot** of x versus Y is used to visualize the conditional distribution of $Y|x$. A **scatterplot matrix** is an array of scatterplots and is used to examine the marginal relationships of the predictors and response. Place Y on the top or bottom of the scatterplot matrix. Variables with outliers, missing values or strong nonlinearities may be so bad that they should not be included in the full model. Suppose that all values of the variable x are positive. The **log rule** says add $\log(x)$ to the full model if $\max(x_i)/\min(x_i) > 10$.

To make a full model, use the above discussion and then make the response and OD plots to check that the full model is good. The number of predictors in the full model should be much smaller than the number of data cases n . For loglinear regression, a rough rule of thumb is that the full model should use no more than $n/5$ predictors and the final submodel should use no more than $n/10$ predictors.

Variable selection, also called subset or model selection, is the search for a subset of predictor variables that can be deleted without important loss of information. A *model for variable selection* for LLR can be described by

$$SP = \alpha + \boldsymbol{\beta}^T \mathbf{x} = \alpha + \boldsymbol{\beta}_S^T \mathbf{x}_S + \boldsymbol{\beta}_E^T \mathbf{x}_E = \alpha + \boldsymbol{\beta}_S^T \mathbf{x}_S \quad (11.3)$$

where $\mathbf{x} = (\mathbf{x}_S^T, \mathbf{x}_E^T)^T$ is a $k \times 1$ vector of nontrivial predictors, \mathbf{x}_S is a $r_S \times 1$ vector and \mathbf{x}_E is a $(k - r_S) \times 1$ vector. Given that \mathbf{x}_S is in the model, $\boldsymbol{\beta}_E = \mathbf{0}$ and E denotes the subset of terms that can be eliminated given that the subset S is in the model.

Since S is unknown, candidate subsets will be examined. Let \mathbf{x}_I be the vector of r terms from a candidate subset indexed by I , and let \mathbf{x}_O be the vector of the remaining terms (out of the candidate submodel). Then

$$SP = \alpha + \boldsymbol{\beta}_I^T \mathbf{x}_I + \boldsymbol{\beta}_O^T \mathbf{x}_O. \quad (11.4)$$

Definition 11.7. The model with $SP = \alpha + \boldsymbol{\beta}^T \mathbf{x}$ that uses all of the predictors is called the *full model*. A model with $SP = \alpha + \boldsymbol{\beta}_I^T \mathbf{x}_I$ that only uses the constant and a subset \mathbf{x}_I of the nontrivial predictors is called a *submodel*. The full model is always a submodel.

Suppose that S is a subset of I and that model (11.3) holds. Then

$$SP = \alpha + \boldsymbol{\beta}_S^T \mathbf{x}_S = \alpha + \boldsymbol{\beta}_S^T \mathbf{x}_S + \boldsymbol{\beta}_{(I/S)}^T \mathbf{x}_{I/S} + \mathbf{0}^T \mathbf{x}_O = \alpha + \boldsymbol{\beta}_I^T \mathbf{x}_I \quad (11.5)$$

where $\mathbf{x}_{I/S}$ denotes the predictors in I that are not in S . Since this is true regardless of the values of the predictors, $\boldsymbol{\beta}_O = \mathbf{0}$ if the set of predictors S is a subset of I . Let $(\hat{\alpha}, \hat{\boldsymbol{\beta}})$ and $(\hat{\alpha}_I, \hat{\boldsymbol{\beta}}_I)$ be the estimates of $(\alpha, \boldsymbol{\beta})$ obtained from fitting the full model and the submodel, respectively. Denote the ESP from the *full model* by $ESP = \hat{\alpha} + \hat{\boldsymbol{\beta}}^T \mathbf{x}_i$ and denote the ESP from the *submodel* by $ESP(I) = \hat{\alpha}_I + \hat{\boldsymbol{\beta}}_I^T \mathbf{x}_{Ii}$.

Definition 11.8. An **EE plot** is a plot of $ESP(I)$ versus ESP .

Variable selection is closely related to the change in deviance test for a reduced model. You are seeking a subset I of the variables to keep in the model. The $AIC(I)$ statistic is used as an aid in backward elimination and forward selection. The full model and the model I_{min} found with the smallest AIC are always of interest. Burnham and Anderson (2004) suggest that if $\Delta(I) = AIC(I) - AIC(I_{min})$, then models with $\Delta(I) \leq 2$ are good, models with $4 \leq \Delta(I) \leq 7$ are borderline, and models with $\Delta(I) > 10$ should not be used as the final submodel. Create a full model. The full model has a deviance at least as small as that of any submodel. The final submodel should have an EE plot that clusters tightly about the identity line. As a rough rule of thumb, a good submodel I has $\text{corr}(ESP(I), ESP) \geq 0.95$. Look at the submodel I_I with the smallest number of predictors such that $\Delta(I_I) \leq 2$, and also examine submodels I with fewer predictors than I_I with $\Delta(I) \leq 7$. Model I_I is a good initial submodel to examine.

Backward elimination starts with the full model with k nontrivial variables, and the predictor that optimizes some criterion is deleted. Then there are $k - 1$ variables left, and the predictor that optimizes some criterion is deleted. This process continues for models with $k - 2, k - 3, \dots, 3$ and 2 predictors.

Forward selection starts with the model with 0 variables, and the predictor that optimizes some criterion is added. Then there is 1 variable in

the model, and the predictor that optimizes some criterion is added. This process continues for models with 2, 3, ..., $k - 1$ and k predictors. Both forward selection and backward elimination result in a sequence of k models $\{x_1^*\}, \{x_1^*, x_2^*\}, \dots, \{x_1^*, x_2^*, \dots, x_{k-1}^*\}, \{x_1^*, x_2^*, \dots, x_k^*\} = \text{full model}$. The two sequences found by forward selection and backward elimination need not be the same.

All subsets variable selection can be performed with the following procedure. Compute the LLR ESP and the OLS ESP found by the OLS regression of Y on \mathbf{x} . Check that $|\text{corr}(\text{LLR ESP}, \text{OLS ESP})| \geq 0.95$. This high correlation will exist for many data sets. Then perform multiple linear regression and the corresponding all subsets OLS variable selection with the $C_p(I)$ criterion. If the sample size n is large and $C_p(I) \leq 2(r + 1)$ where the subset I has $r + 1$ variables including a constant, then $\text{corr}(\text{OLS ESP}, \text{OLS ESP}(I))$ will be high by the proof of Proposition 3.2, and hence $\text{corr}(\text{LLR ESP}, \text{LLR ESP}(I))$ will be high. In other words, if the OLS ESP and LLR ESP are highly correlated, then performing multiple linear regression and the corresponding MLR variable selection (eg forward selection, backward elimination or all subsets selection) based on the $C_p(I)$ criterion may provide many interesting submodels.

Know how to find good models from output. The following rules of thumb (roughly in order of decreasing importance) may be useful. It is often not possible to have all 10 rules of thumb to hold simultaneously. Let submodel I have $r_I + 1$ predictors, including a constant. Do not use more predictors than submodel I_I , which has no more predictors than the minimum AIC model. It is possible that $I_I = I_{min} = I_{full}$. Then the submodel I is good if

- i) the response plot for the submodel looks like the response plot for the full model.
- ii) Want $\text{corr}(\text{ESP}, \text{ESP}(I)) \geq 0.95$.
- iii) The plotted points in the EE plot cluster tightly about the identity line.
- iv) Want the p-value ≥ 0.01 for the change in deviance test that uses I as the reduced model.
- v) Want $r_I + 1 \leq n/10$.
- vi) Want the deviance $G^2(I)$ close to $G^2(full)$ (see iv): $G^2(I) \geq G^2(full)$ since adding predictors to I does not increase the deviance).
- vii) Want $AIC(I) \leq AIC(I_{min}) + 7$ where I_{min} is the minimum AIC model found by the variable selection procedure.

- viii) Want hardly any predictors with p-values > 0.05 .
- ix) Want few predictors with p-values between 0.01 and 0.05.
- x) Want $G^2(I) \leq n - r_I - 1 + 3\sqrt{n - r_I - 1}$.

Heuristically, backward elimination tries to delete the variable that will increase the deviance the least. An increase in deviance greater than 4 (if the predictor has 1 degree of freedom) may be troubling in that a good predictor may have been deleted. In practice, the backward elimination program may delete the variable such that the submodel I with j predictors has a) the smallest $AIC(I)$, b) the smallest deviance $G^2(I)$ or c) the biggest p-value (preferably from a change in deviance test but possibly from a Wald test) in the test $H_0 \beta_i = 0$ versus $H_A \beta_i \neq 0$ where the model with $j + 1$ terms from the previous step (using the j predictors in I and the variable x_{j+1}^*) is treated as the full model.

Heuristically, forward selection tries to add the variable that will decrease the deviance the most. A decrease in deviance less than 4 (if the predictor has 1 degree of freedom) may be troubling in that a bad predictor may have been added. In practice, the forward selection program may add the variable such that the submodel I with j nontrivial predictors has a) the smallest $AIC(I)$, b) the smallest deviance $G^2(I)$ or c) the smallest p-value (preferably from a change in deviance test but possibly from a Wald test) in the test $H_0 \beta_i = 0$ versus $H_A \beta_i \neq 0$ where the current model with j terms plus the predictor x_i is treated as the full model (for all variables x_i not yet in the model).

Suppose that the full model is good and is stored in M1. Let M2, M3, M4 and M5 be candidate submodels found after forward selection, backward elimination, etc. Make a scatterplot matrix of the ESPs for M2, M3, M4, M5 and M1. Good candidates should have estimated sufficient predictors that are highly correlated with the full model estimated sufficient predictor (the correlation should be at least 0.9 and preferably greater than 0.95).

The final submodel should have few predictors, few variables with large Wald p-values (0.01 to 0.05 is borderline), good response and OD plots, and an EE plot that clusters tightly about the identity line. If a factor has $I - 1$ dummy variables, either keep all $I - 1$ dummy variables or delete all $I - 1$ dummy variables, do not delete some of the dummy variables.

	P1	P2	P3	P4
df	144	147	148	149
# of predictors	6	3	2	1
# with $0.01 \leq$ Wald p-value ≤ 0.05	1	0	0	0
# with Wald p-value > 0.05	3	0	1	0
G^2	127.506	131.644	147.151	149.861
AIC	141.506	139.604	153.151	153.861
corr(P1:ETA'U,Pi:ETA'U)	1.0	0.954	0.810	0.792
p-value for change in deviance test	1.0	0.247	0.0006	0.0

Example 11.5. The above table gives summary statistics for 4 models considered as final submodels after performing variable selection. Poisson loglinear regression was used. The response plot for the full model P1 was good. Model P2 was the minimum AIC model found.

Which model is the best candidate for the final submodel? Explain briefly why each of the other 3 submodels should not be used.

Solution: P2 is best. P1 has too many predictors with large pvalues and more predictors than the minimum AIC model. P3 and P4 have corr and pvalue too low and AIC too high.

11.4 Complements

Cameron and Trivedi (1998), Long (1997) and Winkelmann (2008) cover Poisson regression. Also see Hilbe (2007) and texts on categorical data analysis and generalized linear models.

The response plot is essential for understanding the loglinear Poisson regression model and for checking goodness and lack of fit if the estimated sufficient predictor $\hat{\alpha} + \hat{\beta}^T \mathbf{x}$ takes on many values. The response plot and OD plot are examined in Olive (2007b). Goodness of fit is also discussed by Cheng and Wu (1994), Kauermann and Tutz (2001), Pierce and Schafer (1986), Spinelli, Lockart and Stephens (2002), Su and Wei (1991).

For Poisson regression, the OD plot can be used to complement tests and diagnostics for overdispersion such as those given in Breslow (1990), Cameron and Trevedi (1998), Dean (1992), Ganio and Schafer (1992), Lambert and Roeder (1995), and Winkelmann (2008).

The same 4 plots for LLR Poisson regression can be used for NBR, but the OD plot should use $\hat{V}(Y|SP) = \exp(ESP)(1 + \exp(ESP)/\hat{\kappa})$ on the

horizontal axis. As overdispersion increases, larger sample sizes are needed for the OD plot. The weighted fit response plot will be linear but the weights $w_i = Z_i$ will be suboptimal. For Example 11.2, the WFRP will again look like Figure 11.2c, suggesting that the NBR model is not appropriate.

Olive and Hawkins (2005) give the simple all subsets variable selection procedure that can be applied to Poisson regression using readily available OLS software. The procedures of Lawless and Singhai (1978) and Nordberg (1982) are much more complicated. Variable selection using the AIC criterion is discussed in Burnham and Anderson (2004), Cook and Weisberg (1999) and Hastie (1987).

Results from Cameron and Trivedi (1998, p. 89) suggest that if a loglinear Poisson regression model is fit using OLS software for MLR, then a rough approximation is $\hat{\beta}_{LLR} \approx \hat{\beta}_{OLS}/\sqrt{\bar{Y}}$. So a rough approximation is LLR ESP \approx (OLS ESP)/ $\sqrt{\bar{Y}}$.

To motivate the weighted fit response plot and weighted residual plot, assume that all n of the counts Y_i are large. Then

$$\log(\mu(\mathbf{x}_i)) = \log(\mu(\mathbf{x}_i)) + \log(Y_i) - \log(Y_i) = \alpha + \beta^T \mathbf{x}_i,$$

or

$$\log(Y_i) = \alpha + \beta^T \mathbf{x}_i + e_i$$

where

$$e_i = \log\left(\frac{Y_i}{\mu(\mathbf{x}_i)}\right).$$

The error e_i does not have zero mean or constant variance, but if $\mu(\mathbf{x}_i)$ is large

$$\frac{Y_i - \mu(\mathbf{x}_i)}{\sqrt{\mu(\mathbf{x}_i)}} \approx N(0, 1)$$

by the central limit theorem. Recall that $\log(1+x) \approx x$ for $|x| < 0.1$. Then, heuristically,

$$e_i = \log\left(\frac{\mu(\mathbf{x}_i) + Y_i - \mu(\mathbf{x}_i)}{\mu(\mathbf{x}_i)}\right) \approx \frac{Y_i - \mu(\mathbf{x}_i)}{\mu(\mathbf{x}_i)} \approx \frac{1}{\sqrt{\mu(\mathbf{x}_i)}} \frac{Y_i - \mu(\mathbf{x}_i)}{\sqrt{\mu(\mathbf{x}_i)}} \approx N\left(0, \frac{1}{\mu(\mathbf{x}_i)}\right).$$

This suggests that for large $\mu(\mathbf{x}_i)$, the errors e_i are approximately 0 mean with variance $1/\mu(\mathbf{x}_i)$. If the $\mu(\mathbf{x}_i)$ were known, and all of the Y_i were large,

then a weighted least squares of $\log(Y_i)$ on \mathbf{x}_i with weights $w_i = \mu(\mathbf{x}_i)$ should produce good estimates of $(\alpha, \boldsymbol{\beta})$. Since the $\mu(\mathbf{x}_i)$ are unknown, the estimated weights $w_i = Y_i$ could be used.

11.5 Problems

The following three problems use the possums data from Cook and Weisberg (1999a).

Output for Problem 11.1

Data set = Possums, Response = possums

Terms = (Habitat Stags)

Coefficient Estimates

Label	Estimate	Std. Error	Est/SE	p-value
Constant	-0.652653	0.195148	-3.344	0.0008
Habitat	0.114756	0.0303273	3.784	0.0002
Stags	0.0327213	0.00935883	3.496	0.0005

Number of cases:	151
Degrees of freedom:	148
Pearson X2:	110.187
Deviance:	138.685

11.1*. Use the above output to perform inference on the number of possums in a given tract of land. The output is from a loglinear regression.

- Predict $\hat{\mu}(\mathbf{x})$ if $habitat = x_1 = 5.8$ and $stags = x_2 = 8.2$.
- Perform the 4 step Wald test for $H_0 : \beta_1 = 0$.
- Find a 95% confidence interval for β_2 .

Output for Problem 11.2

Response	= possums		Terms	= (Habitat Stags)	
Predictor	df	Total Deviance		df	Change Deviance
Ones	150	187.490			
Habitat	149	149.861		1	37.6289
Stags	148	138.685		1	11.1759

11.2*. Perform the 4 step deviance test for the same model as in Problem 11.1 using the output above.

Output for Problem 11.3

Terms	= (Acacia Bark Habitat Shrubs Stags Stumps)			
Label	Estimate	Std. Error	Est/SE	p-value
Constant	-1.04276	0.247944	-4.206	0.0000
Acacia	0.0165563	0.0102718	1.612	0.1070
Bark	0.0361153	0.0140043	2.579	0.0099
Habitat	0.0761735	0.0374931	2.032	0.0422
Shrubs	0.0145090	0.0205302	0.707	0.4797
Stags	0.0325441	0.0102957	3.161	0.0016
Stumps	-0.390753	0.286565	-1.364	0.1727
Number of cases:		151		
Degrees of freedom:		144		
Deviance:		127.506		

11.3*. Let the reduced model be as in Problem 11.1 and use the output for the full model be shown above. Perform a 4 step change in deviance test.

Arc Problems

The following two problems use data sets from Cook and Weisberg (1999a).

11.4*. a) Activate *possums.lsp* in *Arc* with the menu commands “File > Load > Data > Arcg > possums.lsp.” Scroll up the screen to read the data description.

From *Graph&Fit* select *Fit Poisson response*. Select *y* as the response and select *Acacia*, *bark*, *habitat*, *shrubs*, *stags* and *stumps* as the predictors. Include the output in *Word*. This is your full model

b) (Response plot): From *Graph&Fit* select *Plot of*. Select *P1:Eta'U* for the H box and *y* for the V box. From the OLS popup menu select *Poisson* and move the slider bar to 1. Move the *lowess* slider bar until the *lowess* curve tracks the exponential curve well. Include the response plot in *Word*.

c) From *Graph&Fit* select *Fit Poisson response*. Select *y* as the response and select *bark*, *habitat*, *stags* and *stumps* as the predictors. Include the output in *Word*.

d) (Response plot): From *Graph&Fit* select *Plot of*. Select *P2:Eta'U* for the H box and *y* for the V box. From the OLS popup menu select *Poisson* and move the slider bar to 1. Move the *lowess* slider bar until the lowess curve tracks the exponential curve well. Include the response plot in *Word*.

e) Deviance test. From the *P2* menu, select *Examine submodels* and click on OK. Include the output in *Word* and perform the 4 step deviance test.

f) Perform the 4 step change of deviance test.

g) EE plot. From *Graph&Fit* select *Plot of*. Select *P2:Eta'U* for the H box and *P1:Eta'U* for the V box. Move the OLS slider bar to 1. Click on the *Options* popup menu and type “y=x”. Include the plot in *Word*. Is the plot linear?

11.5*. In this problem you will find a good submodel for the *possums* data.

a) Activate *possums.lsp* in *Arc* with the menu commands “File > Load > Data > Arcg> possums.lsp.” Scroll up the screen to read the data description.

b) From *Graph&Fit* select *Fit Poisson response*. Select *y* as the response and select *Acacia, bark, habitat, shrubs, stags* and *stumps* as the predictors.

In Problem 11.4, you showed that this was a good full model.

c) Using what you have learned in class find a good submodel and include the relevant output in *Word*.

(Hints: Create a full model. The full model has a deviance at least as small as that of any submodel. Consider forward selection and backward elimination. For each method, find the submodel I_{min} with the smallest AIC. Let $\Delta(I) = AIC(I) - AIC(I_{min})$, and find submodel I_I with the smallest number of predictors such that $\Delta(I_I) \leq 2$, and also examine submodels I with fewer predictors than I_I that have $\Delta(I) \leq 7$. The final submodel should have an EE plot that clusters tightly about the identity line. As a rough rule of thumb, a good submodel I has $\text{corr}(ESP(I), ESP) \geq 0.95$. Submodel I_I is your initial candidate model. Fit this candidate model and look at the Wald test p-values. Try to eliminate predictors with large p-values but make sure that the deviance does not increase too much. You may have several

models, say P2, P3, P4 and P5 to look at. Make a scatterplot matrix of the $\text{Pi:ETA}'U$ from these models and from the full model P1. Make the EE and response plots for each model. The correlation in the EE plot should be at least 0.9 and preferably greater than 0.95. As a very rough guide for Poisson regression, the number of predictors in the full model should be less than $n/5$ and the number of predictors in the final submodel should be less than $n/10$. WARNING: do not delete part of a factor. Either keep all $J - 1$ factor dummy variables or delete all $J - 1$ factor dummy variables. WARNING: if an important factor is in the full model but not the reduced model, then the plotted points in the EE plot may follow more than 1 line.)

d) Make a response plot for your final submodel, say P2. From *Graph&Fit* select *Plot of*. Select $P2:\text{Eta}'U$ for the H box and y for the V box. From the OLS popup menu select *Poisson* and move the slider bar to 1. Move the *lowess* slider bar until the lowess curve tracks the exponential curve well. Include the response plot in *Word*.

e) Suppose that P1 contains your full model and P2 contains your final submodel. Make an EE plot for your final submodel: from *Graph&Fit* select *Plot of*. Select $P1:\text{Eta}'U$ for the V box and $P2:\text{Eta}'U$, for the H box. After the plot appears, click on the *options* popup menu. A window will appear. Type $y = x$ and click on OK. This action adds the identity line to the plot. Also move the OLS slider bar to 1. Include the plot in *Word*.

f) Using c), d), e) and any additional output that you desire (eg AIC(full), AIC(min) and AIC(final submodel), explain why your final submodel is good.

Warning: The following problems use data from the book's webpage. Save the data files on a disk. Get in Arc and use the menu commands "File > Load" and a window with a *Look in box* will appear. Click on the black triangle and then on *3 1/2 Floppy(A:)*. Then click twice on the data set name.

11.6*. a) This problem uses a data set from Myers, Montgomery and Vining (2002). Activate *popcorn.lsp* in Arc with the menu commands "File > Load > Floppy(A:) > popcorn.lsp." Scroll up the screen to read the data description. From *Graph&Fit* select *Fit Poisson response*. Use *oil*, *temp* and *time* as the predictors and y as the response. From *Graph&Fit* select *Plot of*. Select $P1:\text{Eta}'U$ for the H box and y for the V box. From the OLS

popup menu select *Poisson* and move the slider bar to 1. Move the *lowess* slider bar until the lowess curve tracks the exponential curve. Include the EY plot in *Word*.

b) From the *P1* menu select *Examine submodels*, click on *OK* and include the output in *Word*.

c) Test whether $\beta_1 = \beta_2 = \beta_3 = 0$.

d) From the *popcorn* menu, select *Transform* and select *y*. Put 1/2 in the *p* box and click on *OK*. From the *popcorn* menu, select *Add a variate* and type $yt = \sqrt{y} * \log(y)$ in the resulting window. Repeat three times adding the variates $oilt = \sqrt{y} * oil$, $tempt = \sqrt{y} * temp$ and $timet = \sqrt{y} * time$. From *Graph&Fit* select *Fit linear LS* and choose $y^{1/2}$, *oilt*, *tempt* and *timet* as the predictors, *yt* as the response and click on the *Fit intercept* box to remove the check. Then click on *OK*. From *Graph&Fit* select *Plot of*. Select *L2:Fit-Values* for the H box and *yt* for the V box. A plot should appear. Click on the *Options* menu and type $y = x$ to add the identity line. Include the weighted fit response plot in *Word*.

e) From *Graph&Fit* select *Plot of*. Select *L2:Fit-Values* for the H box and *L2:Residuals* for the V box. Include the weighted residual plot in *Word*.

f) For the plot in e), highlight the case in the upper right corner of the plot by using the mouse to move the arrow just above and to the left the case. Then hold the rightmost mouse button down and move the mouse to the right and down. From the *Case deletions* menu select *Delete selection from data set*, then from *Graph&Fit* select *Fit Poisson response*. Use *oil*, *temp* and *time* as the predictors and *y* as the response. From *Graph&Fit* select *Plot of*. Select *P3:Eta'U* for the H box and *y* for the V box. From the OLS popup menu select *Poisson* and move the slider bar to 1. Move the *lowess* slider bar until the lowess curve tracks the exponential curve. Include the response plot in *Word*.

g) From the *P3* menu select *Examine submodels*, click on *OK* and include the output in *Word*.

h) Test whether $\beta_1 = \beta_2 = \beta_3 = 0$.

i) From *Graph&Fit* select *Fit linear LS*. Make sure that $y^{1/2}$, *oilt*, *tempt* and *timet* are the predictors, *yt* is the response, and that the *Fit intercept*

box does not have a check. Then click on *OK* From *Graph&Fit* select *Plot of*. Select *L4:Fit-Values* for the H box and *yt* for the V box. A plot should appear. Click on the *Options* menu and type $y = x$ to add the identity line. Include the weighted fit response plot in *Word*.

j) From *Graph&Fit* select *Plot of*. Select *L4:Fit-Values* for the H box and *L4:Residuals* for the V box. Include the weighted residual plot in *Word*.

k) Is the deleted point influential? Explain briefly.

l) From *Graph&Fit* select *Plot of*. Select *P3:Eta'U* for the H box and *P3:Dev-Residuals* for the V box. Include the deviance residual plot in *Word*.

m) Is the weighted residual plot from part j) a better lack of fit plot than the deviance residual plot from part l)? Explain briefly.

R/Splus problems

Download functions with the command `source("A:/regpack.txt")`. See **Preface or Section 17.1**. Typing the name of the `regpack` function, eg `llressp`, will display the code for the function. Use the `args` command, eg `args(llressp)`, to display the needed arguments for the function.

11.7. a) Obtain the function `llrdata` from `regpack.txt`. Enter the commands

```
out <- llrdata()
x <- out$x
y <- out$y
```

b) Obtain the function `llressp` from `regpack.txt`. Enter the commands `llressp(x,y)` and include the resulting plot in *Word*.

c) Obtain the function `llrwtfrp` from `regpack.txt`. Enter the commands `llrwtfrp(x,y)` and include the resulting plot in *Word*.