Plots for Binomial and Poisson Regression

David J. Olive*

Southern Illinois University

August 15, 2009

Abstract

The binomial and Poisson regression models state that the conditional distribution of a count Y given the sufficient predictor (SP) follows a binomial(m, F(SP)) or Poisson(exp(SP)) distribution where the sufficient predictor is a linear combination of predictor variables and F is a distribution function. Two new plots for Poisson regression as well as modifications to two plots from the literature are used to visualize the regression model, to check for lack of fit and overdispersion, and to detect outliers.

KEY WORDS: Diagnostics, Generalized Linear Models, Goodness of Fit, Outliers, Overdispersion.

*David J. Olive is Associate Professor, Department of Mathematics, Southern Illinois University, Mailcode 4408, Carbondale, IL 62901-4408, USA. E-mail address: dolive@math.siu.edu. This research was supported by NSF grant DMS-0600933.

1 INTRODUCTION

Regression models are used to study the conditional distribution $Y|\boldsymbol{x}$ of the response variable Y given the $p \times 1$ vector of nontrivial predictors $\boldsymbol{x}$. The Poisson regression model states that $Y_1, ..., Y_n$ are independent random variables with

$$Y_i \sim \text{Poisson}(\mu(\boldsymbol{x}_\text{i})).$$

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

$$\mu(\boldsymbol{x}_i) = \exp(\alpha + \boldsymbol{\beta}^T \boldsymbol{x}_i). \tag{1.1}$$

The LLR model, a special case of a generalized linear model, is often used to analyze categorical data when the response variable Y is a count. Let the linear predictor = sufficient predictor $SP = \alpha + \boldsymbol{\beta}^T \boldsymbol{x}$. Then (1.1) can be written compactly as $Y|SP \sim$ Poisson(exp(SP)). For example, $Y|SP = 0 \sim$ Poisson(1). Also note that the conditional mean and variance functions are equal: $E(Y|SP) = V(Y|SP) = \exp(SP)$.

The *binomial regression model* states that $Y_1, ..., Y_n$ are independent random variables with

$$Y_i \sim \text{binomial}(\text{m}_\text{i}, \rho(\alpha + \boldsymbol{\beta}^\text{T} \boldsymbol{x}_\text{i})),$$

or

$$Y_i|SP_i \sim \text{binomial}(\text{m}_\text{i}, \rho(\text{SP}_\text{i})). \tag{1.2}$$

Note that the conditional mean function $E(Y_i|SP_i) = m_i \rho(SP_i)$ and the conditional variance function $V(Y_i|SP_i) = m_i \rho(SP_i)(1 - \rho(SP_i))$. The *binary regression model* is the special case where $m_i \equiv 1$ for $i = 1, ..., n$.

Typically $\rho(SP) = F(SP)$ where F is the distribution function (DF) of a location scale family. The *logistic regression (LR) model* is the special case of binomial regression where

$$P(\text{success}|\boldsymbol{x}_\text{i}) = \rho(\boldsymbol{x}_\text{i}) = \frac{\exp(\alpha + \boldsymbol{\beta}^\text{T}\boldsymbol{x}_\text{i})}{1 + \exp(\alpha + \boldsymbol{\beta}^\text{T}\boldsymbol{x}_\text{i})}. \tag{1.3}$$

Equivalently,

$$\rho(SP) = \frac{\exp(SP)}{1 + \exp(SP)}.$$

Note that $\rho(x)$ is the DF of a logistic(0,1) distribution. The probit regression model uses $\rho(SP) = \Phi(SP)$ where $\Phi(x)$ is the DF of a normal(0,1) distribution. The choice $\rho(SP) = \exp[-\exp(-SP)]$ corresponds to the DF of the largest extreme value distribution, and the choice $\rho(SP) = 1 - \exp[-\exp(SP)]$ corresponds to the DF of the smallest extreme value distribution. If the successes Y_i can be modelled by one extreme value distribution, then the failures $m_i - Y_i$ can be modelled by the other.

Often the LR mean function is a good approximation to the data and the LR MLE is a consistent estimator of $\boldsymbol{\beta}$, but the LR model is not appropriate. The problem is that for many data sets where $E(Y_i|\boldsymbol{x}_i) = m_i\rho(SP_i)$, it turns out that $V(Y_i|\boldsymbol{x}_i) > m_i\rho(SP_i)(1 - \rho(SP_i))$. Similarly, for many data sets where $E(Y|\boldsymbol{x}) = \mu(\boldsymbol{x}) = \exp(SP)$, it turns out that $V(Y|\boldsymbol{x}) > \exp(SP)$, and model (1.1) is not appropriate. See Cameron and Trivedi (1998, p. 64). This phenomenon is called *overdispersion*.

Section 2 reviews the estimated sufficient summary plot and adds visual aids to make a graphical diagnostic for overdispersion easy to use. Two new plots for Poisson regression are also discussed. Section 3 gives examples that show how to assess the adequacy of the binomial and Poisson regression models with the plots. Section 4 gives conclusions.

2 Four Plots

For regression models where Y is independent of $\boldsymbol{x}$ given $SP = \alpha + \boldsymbol{\beta}^T\boldsymbol{x}$, the conditional distribution of $Y|\boldsymbol{x}$ can be visualized with an *estimated sufficient summary plot* (ESSP) of the estimated sufficient predictor $ESP = \hat{\alpha} + \hat{\boldsymbol{\beta}}^T\boldsymbol{x}_i$ versus Y_i. Since regression is the study of $Y|\boldsymbol{x}$, the plot is crucial for analyzing regression models. See Cook (1998, p. 10). A closely related plot of $c + \boldsymbol{a}^T\boldsymbol{x}_i$ versus Y_i for some constant c (often zero) and some vector $\boldsymbol{a}$ is called a model checking plot by Cook and Weisberg (1999, p. 397) and a marginal response plot by Cook and Weisberg (1997). Adding the parametric mean function and a scatterplot smoother to the plot is the graphical analog of goodness of fit tests such as those of Hosmer and Lemeshow (1980) and Pardoe (2001). Other goodness of fit tests and diagnostics include those given in Cheng and Wu (1994), Collett (1999), Landwehr, Pregibon and Shoemaker (1984), Pardoe and Cook (2002), Pierce and Schafer (1986), Pregibon (1981), Simonoff (1998), Spinelli, Lockart and Stephens (2002) and Su and Wei (1991).

To check for overdispersion in parametric models, we suggest using the *OD plot* of the estimated model variance $\hat{V}(Y|SP)$ versus the squared residuals $\hat{V} = [Y - \hat{E}(Y|SP)]^2$. This plot has been used by Winkelmann (2000, p. 110) for the LLR model where $\hat{V}(Y|SP) = \hat{E}(Y|SP) = \exp(ESP)$. For binomial and Poisson regression, the OD plot can be used to complement tests and diagnostics for overdispersion such as those given in Breslow (1990), Cameron and Trivedi (1998), Collett (1999, ch. 6), Dean (1992), Ganio and Schafer (1992), Lambert and Roeder (1995) and Winkelmann (2000).

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 (and binomial regression if the counts are neither too small nor too large), G^2 is approximately chi-square with $n - p - 1$ degrees of freedom. Since a χ^2_d random variable has mean d and standard deviation $\sqrt{2d}$, the 98th percentile of the χ^2_d 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 (1.1) or (1.3) may be needed. A good discussion of such count models is in Simonoff (2003).

Next the ESSP is tailored to the Poisson regression model (1.1). The estimated mean function

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

is added to the ESSP as a visual aid. The scatterplot smoother lowess is a nonparametric estimator of the mean function, 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 $E(Y|\boldsymbol{x})$.

Let $Z_i = Y_i/m_i$. Then the conditional distribution $Z_i|\boldsymbol{x}_i$ of the binomial regression model (1.3) can be visualized with a plot of the ESP versus Z_i with the estimated mean function of the Z_i

$$\hat{\rho}(ESP) = \frac{\exp(ESP)}{1 + \exp(ESP)}$$

added as a visual aid. Cook and Weisberg (1999, p. 515) add a lowess curve to the plot. Alternatively, divide the ESP into J slices with approximately the same number of cases in each slice. Then compute $\hat{\rho}_s = \sum_s Y_i / \sum_s m_i$ where the sum is over the cases in slice s. Then plot the resulting step function. For binary data the step function is

simply the sample proportion in each slice. Both the lowess curve and step function are simple nonparametric estimators of the mean function $\rho(SP)$. If the lowess curve or step function tracks the logistic curve (the estimated mean) closely, then the LR mean function is a reasonable approximation to the data. The plot is called an *ESS plot* because of the "ESS" shape of the mean function. The plot of the step function and logistic curve is a graphical approximation of the goodness of fit tests described in Hosmer and Lemeshow (1980).

Although the ESSP is used to visualize $Y|\boldsymbol{x}$, examining the mean function is simpler than examining the variance function. Cook and Weisberg (1999, pp. 401-403) suggest adding parametric and nonparametric estimators of the standard deviation function to the ESSP.

For model (1.1), 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 binomial and Poisson regression.

First, recall that a normal approximation is good for the Poisson distribution 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: $\hat{V} = 4\hat{V}(Y|SP)$. Only about 5% of the plotted points should be above this line. Similar remarks apply to binomial regression if the counts are neither too big nor

too small.

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. It is easier to use the OD plot to check the variance function than the ESSP 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.

Suppose the ESSP and OD plot suggest that the model is reasonable. If a scatterplot smoother fits the horizontal line $W = \hat{\theta}$ (where $\hat{\theta}$ is the MLE of $W = Y$ or $W = Z$ without any predictors) about as well as the estimated mean function in the ESSP, then the predictors are not much more useful than using $\hat{\theta}$ for prediction (analogous to R^2 being low). If the scatterplot smoother fits the estimated parametric mean function far better than any horizontal line, then the model may explain a large proportion of the variability of the response (analogous to R^2 being high). This possibly new graphical diagnostic is a competitor of those suggested by Agresti and Caffo (2002), Liao and McGee (2003) and Menard (2000).

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 $\boldsymbol{u}$, the ESSP is the (weighted forward) 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 $W - \hat{W}$.

The *weighted forward response plot* is a plot of $\sqrt{Z_i}ESP = \sqrt{Z_i}(\hat{\alpha} + \hat{\boldsymbol{\beta}}^T \boldsymbol{x}_i)$ versus $\sqrt{Z_i}\log(Z_i)$ where $Z_i = Y_i$ if $Y_i > 0$, and $Z_i = 0.5$ if $Y_i = 0$. The *weighted residual plot* is a plot of $\sqrt{Z_i}(\hat{\alpha}+\hat{\boldsymbol{\beta}}^T \boldsymbol{x}_i)$ versus the "WLS" residuals $r_{Wi} = \sqrt{Z_i}\log(Z_i) - \sqrt{Z_i}(\hat{\alpha}+\hat{\boldsymbol{\beta}}^T \boldsymbol{x}_i)$. The WLS residuals are often highly correlated with the deviance residuals. 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 forward 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.

To motivate the above two plots, recall that the minimum chi–square estimator $(\hat{\alpha}_M, \hat{\boldsymbol{\beta}}_M)$ for Poisson regression is found from the WLS regression of $\log(Z_i)$ on $\boldsymbol{x}_i$ with weights $w_i = Z_i$. Equivalently, use the OLS regression (without intercept) of $\sqrt{Z_i}\log(Z_i)$

on $\sqrt{Z_i}(1, \boldsymbol{x}_i^T)^T$. Then the plot of the "fitted values" $\sqrt{Z_i}(\hat{\alpha}_M + \hat{\boldsymbol{\beta}}_M^T \boldsymbol{x}_i)$ versus the "response" $\sqrt{Z_i}\log(Z_i)$ should have points that scatter about the identity line. The minimum chi–square estimator tends to be consistent if n is fixed and all n counts Y_i increase to ∞ while the LLR MLE tends to be consistent if the sample size $n \to \infty$. See Agresti (2002, pp. 611-612). Since the two estimators are often close for many data sets, the plotted points in the weighted forward response plot should scatter about the identity line if $\hat{E}(Y|SP) = \exp(ESP)$ is a good approximation to the mean function $E(Y|SP)$.

3 Examples

The first three examples are for Poisson regression where the OD plot of $\exp(ESP)$ versus $(Y - \exp(ESP))^2$ is a plot of fitted values versus squared residuals. Notice that $\hat{Y} = \exp(ESP) = \hat{E}(Y|SP)$.

Example 1. Myers, Montgomery and Vining (2002, Example 4.5) give data where 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 1 shows the 4 plots for this data. In the ESSP of Figure 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 $\overline{Y}$. Scatter about this line is analogous to R^2 being low for linear regression. Since the exponential function gives a good fit to the data while the horizontal line does not, the Poisson regression is useful

for explaining the variation of Y (analogous to R^2 being high). Notice that the lowess curve underestimates the mean function for large ESP.

The OD plot in Figure 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 1c and there are no unusual points in Figure 1d. The four plots suggest that the LLR Poisson regression model is a useful approximation to the data. Hence $Y|ESP \approx$ Poisson(exp(ESP)). For example, when ESP = 1.61, $Y \approx$ Poisson(5) and when ESP = 4.5, $Y \approx$ Poisson(90). Notice that the Poisson mean can be roughly estimated by finding the height of the exponential curve in Figure 1a.

Example 2. Agresti (2002, pp. 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 2 used the ordinal variables color and spine condition as coded. An alternative model would use spine condition as a factor. Figure 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 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 2c since the plotted points fail to cover the identity line. Although the exponential mean function fits the lowess curve better than the line $Y = \overline{Y}$, alternative models suggested by Agresti (2002) may fit the data better.

Example 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 3. Ignoring the outlier in Figure 3a suggests that the line $Y = \overline{Y}$ 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 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. See Figure 3b.

The next two examples are for binomial regression. The mean function may be useful if the step function tracks the logistic curve. The OD plot is a plot of $\hat{V}_{mod} = \hat{V}(Y_i|SP) = m_i\rho(ESP_i)(1-\rho(ESP_i))$ versus $\hat{V} = (Y_i - m_i\rho(ESP_i))^2$. The wedge formed by the horizontal line and slope 4 line tends to be useful if the Z_i scatter about the logistic curve (so that the counts are neither too large nor too small) in the ESS plot. For binary data, the OD plot is not needed but the ESS plot is still very useful.

Example 4. Abraham and Ledolter (2006, pp. 360-364) describe death penalty sentencing in Georgia. The predictors are aggravation *level* from 1 to 6 (treated as a

continuous variable) and *race* of victim coded as 1 for white and 0 for black. There were 362 jury decisions and 12 level–race combinations. The response variable was the number of death sentences in each combination. The ESS plot in Figure 4a shows that the Y_i/m_i are close to the estimated LR mean function (the logistic curve), and the step function based on 5 slices tracks the logistic curve well. The horizontal line is $\hat{\rho} = \sum_{i=1}^{n} Y_i / \sum_{i=1}^{n} m_i$. Scatter of the step function about this line is analogous to R^2 being low. Since the step function based on 5 slices tracks the logistic curve well, but does not track the horizontal line, the binomial regression is useful for explaining the variation of Y (analogous to R^2 being high). Notice that this interpretation is also useful for binary data where the Z_i will not scatter about the logistic curve.

The OD plot is shown in Figure 4b with the identity, slope 4 and OLS lines added as visual aids. The vertical scale is less than the horizontal scale and there is no evidence of overdispersion. The logistic regression model suggests that $Y_i \approx \text{binomial}(m_i, \rho(ESP))$ where the logistic curve $\rho(ESP)$ can be estimated by its height in Figure 4a. Thus $Y_i \approx \text{binomial}(m_i, 0.018)$ when ESP $= -4$, and $Y_i \approx \text{binomial}(m_i, 0.5)$ when ESP $= 0$ while $Y_i \approx \text{binomial}(m_i, 0.982)$ when ESP $= 4$.

Example 5. Collett (1999, pp. 216-219) describes a data set where the response variable is the number of rotifers that remain in suspension in a tube. A rotifer is a microscopic invertebrate. The two predictors were the *density* of a stock solution of Ficolli and the *species* of rotifer coded as 1 for polyarthra major and 0 for keratella cochlearis. The sample size $n = 40$, and Figure 5a shows the ESS plot. Both the observed proportions and the step function track the logistic curve well, suggesting that the LR mean function is a good approximation to the data. The OD plot suggests that there is

overdispersion since the vertical scale is about 30 times the horizontal scale. Notice that the OLS line has slope much larger than 4 and two outliers seem to be present.

Example 6. The ICU data is available from STATLIB (http://lib.stat.cmu.edu/DASL/Datafiles/ICU.html). The survival of 200 patients following admission to an intensive care unit was studied with logistic regression. The response variable was STA (0 = Lived, 1 = Died). Predictors were AGE, SEX (0 = Male, 1 = Female), RACE (1 = White, 2 = Black, 3 = Other), SER= Service at ICU admission (0 = Medical, 1 = Surgical), CAN= Is cancer part of the present problem? (0 = No, 1 = Yes), CRN= History of chronic renal failure (0 = No, 1 = Yes), INF= Infection probable at ICU admission (0 = No, 1 = Yes), CPR= CPR prior to ICU admission (0 = No, 1 = Yes), SYS= Systolic blood pressure at ICU admission (in mm Hg), HRA= Heart rate at ICU admission (beats/min), PRE= Previous admission to an ICU within 6 months (0 = No, 1 = Yes), TYP= Type of admission (0 = Elective, 1 = Emergency), FRA= Long bone, multiple, neck, single area, or hip fracture (0 = No, 1 = Yes), PO2= PO2 from initial blood gases (0 = >60, 1 = 60), PH= PH from initial blood gases (0 = 7.25, 1 <7.25), PCO= PCO2 from initial blood gases (0 = 45, 1 = >45), Bic= Bicarbonate from initial blood gases (0 = 18, 1 = <18), CRE= Creatinine from initial blood gases (0 = 2.0, 1 = >2.0), and LOC= Level of consciousness at admission (0 = no coma or stupor, 1= deep stupor, 2 = coma).

Factors LOC and RACE had two indicator variables. The response plot in Figure 6 shows that the logistic regression model using the 19 predictors is useful for predicting survival. After variable selection, the submodel using AGE, CAN, SYS, TYP and LOC was chosen. The EE plot of ESP(sub) versus ESP(full) is shown in Figure 7. Olive and

Hawkins (2005) show that the plotted points in the EE plot should cluster tightly about the identity line if the full model and the submodel are good. This clustering did not occur in Figure 7. The lowest cluster of points and the case furthest to the right near the identity line correspond to black patients. The main cluster and upper right cluster correspond to patients who are not black. Figure 8 shows the EE plot when RACE is added to the submodel. Then all of the points cluster about the identity line. Although variable selection did not suggest that RACE is important, the two EE plots suggest that RACE is important. Also the RACE variable could be replaced by an indicator for black. This example illustrates how the plots can be used to quickly improve the model obtained by following logistic regression with variable selection.

4 Conclusions

The ESSP can be used to visualize $Y|\boldsymbol{x}$ for models such as generalized linear models where Y is independent of $\boldsymbol{x}$ given the sufficient predictor $\alpha + \boldsymbol{\beta}^T\boldsymbol{x}$. Adding the estimated mean function and a scatterplot smoother as visual aids is again useful. This plot is one of the simplest ways to improve the analysis of important regression models such as multiple linear regression, logistic regression and Poisson regression. The plot is also useful for teaching regression to students and for explaining the model to consulting clients. An assumption is that the ESP takes on many values. More research is needed to determine when these plots are useful for contingency tables.

Similarly, the OD plot can be made for regression models where $E(Y|SP)$ and $V(Y|SP)$ can be estimated, but the simple visual aids may need to be changed. No-

tice that the OD plot is a check both for overdispersion and for the variance function.

The estimated sufficient summary plot, where the parametric estimated mean function and a scatterplot smoother are added as visual aids, is not new. The OD plot has been used for the LLR Poisson regression model although the visual aids added to the plot are new. The combination of the ESSP with the OD plot is a powerful method for assessing the adequacy of Poisson and binomial regression models, and these plots should be made before performing inference. Influential cases and outliers will often appear in the plots, and information from case diagnostics such as analogs for Cook's distances and leverage can be incorporated into the plots by highlighting cases corresponding to diagnostics larger than some cutoff value. The Poisson and binomial regression models are simpler than most alternative count models, so plots for goodness of fit of these models are useful.

A useful alternative to the LLR model is the 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 $y = 0, 1, 2, ...$ where $\mu > 0$ and $\kappa > 0$. Then $E(Y) = \mu$ and $\mathrm{V}(Y) = \mu + \mu^2/\kappa$. (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.)

The NBR model states that $Y_1, ..., Y_n$ are independent random variables where $Y_i \sim$ NB$(\mu(\boldsymbol{x}_\mathrm{i}), \kappa)$ with $\mu(\boldsymbol{x}_i) = \exp(\alpha + \boldsymbol{\beta}^T \boldsymbol{x}_i)$. Hence $Y|SP \sim$ NB(exp(SP), κ), $E(Y|SP) =$ exp(SP) and

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

This model has the same mean function as the LLR model but allows for overdispersion. As $\kappa \to \infty$, the NBR model converges to the LLR model.

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 forward response plot will be linear but the weights $w_i = Z_i$ will be suboptimal. For Example 2, the WFRP will again look like Figure 2c, suggesting that the NBR model is not appropriate.

A useful alternative to the binomial regression model is a beta–binomial regression (BBR) model. Following Simonoff (2003, pp. 93-94) and Agresti (2002, pp. 554-555), let $\delta = \rho/\theta$ and $\nu = (1-\rho)/\theta$, so $\rho = \delta/(\delta+\nu)$ and $\theta = 1/(\delta+\nu)$. Let

$$B(\delta, \nu) = \frac{\Gamma(\delta)\Gamma(\nu)}{\Gamma(\delta+\nu)}.$$

If Y has a beta–binomial distribution, $Y \sim \text{BB}(\text{m}, \rho, \theta)$, then the probability mass function of Y is

$$P(Y = y) = \binom{m}{y} \frac{B(\delta + y, \nu + m - y)}{B(\delta, \nu)}$$

for $y = 0, 1, 2, ..., m$ where $0 < \rho < 1$ and $\theta > 0$. Hence $\delta > 0$ and $\nu > 0$. Then $E(Y) = m\delta/(\delta+\nu) = m\rho$ and $\text{V}(Y) = m\rho(1-\rho)[1 + (m-1)\theta/(1+\theta)]$. If $Y|\pi \sim$ binomial(m, π) and $\pi \sim$ beta(δ, ν), then $Y \sim \text{BB}(\text{m}, \rho, \theta)$.

The BBR model states that $Y_1, ..., Y_n$ are independent random variables where $Y_i|SP_i \sim \text{BB}(\text{m}_\text{i}, \rho(\text{SP}_\text{i}), \theta)$. Hence $E(Y_i|SP_i) = m_i\rho(SP_i)$ and

$$V(Y_i|SP_i) = m_i\rho(SP_i)(1 - \rho(SP_i))[1 + (m_i - 1)\theta/(1+\theta)].$$

The BBR model has the same mean function as the LR model, but allows for overdispersion. As $\theta \to 0$, it can be shown that $V(\pi) \to 0$ and the BBR model converges to the binomial LR model.

The ESS plot can again be used to visualize the BBR model, but the OD plot should use $\hat{V}(Y|SP) = m_i\rho(ESP)(1-\rho(ESP))[1+(m_i-1)\hat{\theta}/(1+\hat{\theta})]$ on the horizontal axis. As overdispersion increases, larger sample sizes are needed for the OD plot.

If the binomial LR OD plot is used but the data follows a beta–binomial regression model, then $\hat{V}_{mod} = \hat{V}(Y_i|ESP) \approx m_i\rho(ESP)(1-\rho(ESP))$ while $\hat{V} = [Y_i - m_i\rho(ESP)]^2 \approx (Y_i - E(Y_i))^2$. Hence $E(\hat{V}) \approx V(Y_i) \approx m_i\rho(ESP)(1-\rho(ESP))[1+(m_i-1)\theta/(1+\theta)]$, so the plotted points with $m_i = m$ should scatter about a line with slope $\approx$

$$1 + (m-1)\frac{\theta}{1+\theta} = \frac{1+m\theta}{1+\theta}.$$

The website (www.math.siu.edu/olive/ol-bookp.htm) has links to `robdata.txt` that contains the five data sets and `rpack.txt` that contains *R software*. The function `llrplot` makes the four plots for Poisson regression, and the function `llrsim` simulates these four plots for LLR and NBR data. The function `lrplot` makes the ESS and OD plots for binomial data while `lrplot2` makes the ESS plot for binary data.

5 References

Abraham, B., Ledolter, J., 2006. Introduction to Regression Modeling. Thomson Brooks/Cole, Belmont, CA.

Agresti, A., 2002. Categorical Data Analysis. 2nd ed. Wiley, Hoboken, NJ.

Agresti, A., Caffo, B., 2002. Measures of relative model fit. Comput. Statist. Data Anal. 39, 127-136.

Breslow, N., 1990. Tests of hypotheses in overdispersed Poisson regression and other quasi-likelihood models. J. Amer. Statist. Assoc. 85, 565-571.

Cameron, A.C., Trivedi, P.K., 1998. Regression Analysis of Count Data. Cambridge University Press, Cambridge, UK.

Cheng, K.F., Wu, J.W., 1994. Testing goodness of fit for a parametric family of link functions. J. Amer. Statist. Assoc. 89, 657-664.

Collett, D., 1999. Modelling Binary Data. Chapman & Hall/CRC, Boca Raton, Florida.

Cook, R.D., 1998. Regression Graphics: Ideas for Studying Regression Through Graphics. Wiley, New York.

Cook, R.D., Weisberg, S., 1997. Graphics for assessing the adequacy of regression models. J. Amer. Statist. Assoc. 92, 490-499.

Cook, R.D., Weisberg, S., 1999. Applied Regression Including Computing and Graphics. Wiley, New York.

Dean, C.B., 1992. Testing for overdispersion in Poisson and binomial regression models. J. Amer. Statist. Assoc. 87, 441-457.

Ganio, L.M., Schafer, D.W., 1992. Diagnostics for overdispersion. J. Amer. Statist. Assoc. 87, 795-804.

Hosmer, D.W., Lemeshow, S., 1980. A goodness of fit test for the multiple logistic regression model. Commun. Statist. A10, 1043-1069.

Lambert, D., Roeder, K., 1995. Overdispersion diagnostics for generalized linear models. J. Amer. Statist. Assoc. 90, 1225-1236.

Landwehr, J.M., Pregibon, D., Shoemaker, A.C., 1984. Graphical models for assessing logistic regression models. J. Amer. Statist. Assoc. 79, 61-83.

Liao, J.G., McGee, D., 2003. Adjusted coefficients of determination for logistic regression. Amer. Statist. 57, 3, 161-165.

Menard, S., 2000. Coefficients of determination for multiple logistic regression analysis. Amer. Statist. 54, 17-24.

Myers, R.H., Montgomery, D.C., Vining, G.G., 2002. Generalized Linear Models with Applications in Engineering and the Sciences. Wiley, New York.

Olive, D.J., Hawkins, D.M. 2005. Variable Selection for 1D Regression Models. Technom. 47, 43-50.

Pardoe, I., 2001. A Bayesian sampling approach to regression model checking. J. Computat. Graphic. Statist. 10, 617-627.

Pardoe, I., Cook, R.D., 2002. A graphical method for assessing the fit of a logistic regression model. Amer. Statist. 56, 263-272.

Pierce, D.A., Schafer, D.W., 1986. Residuals in generalized linear models. J. Amer. Statist. Assoc. 81, 977-986.

Pregibon, D., 1981. Logistic regression diagnostics. Ann. Statist. 9, 705-724.

Simonoff, J.S., 1998. Logistic regression, categorical predictors, and goodness-of-fit: it depends on who you ask. Amer. Statist. 52, 10-14.

Simonoff, J.S., 2003. Analyzing Categorical Data, Springer-Verlag, New York.

Spinelli, J.J., Lockart, R.A., Stephens, M.A., 2002. Tests for the response distribution in a Poisson regression model. J. Statist. Plann. Infer. 108, 137-154.

Su, J.Q., Wei, L.J., 1991. A lack–of–fit test for the mean function in a generalized linear

model. J. Amer. Statist. Assoc. 86, 420-426.

Winkelmann, R., 2000. Econometric Analysis of Count Data. 3rd ed. Springer-Verlag, New York.

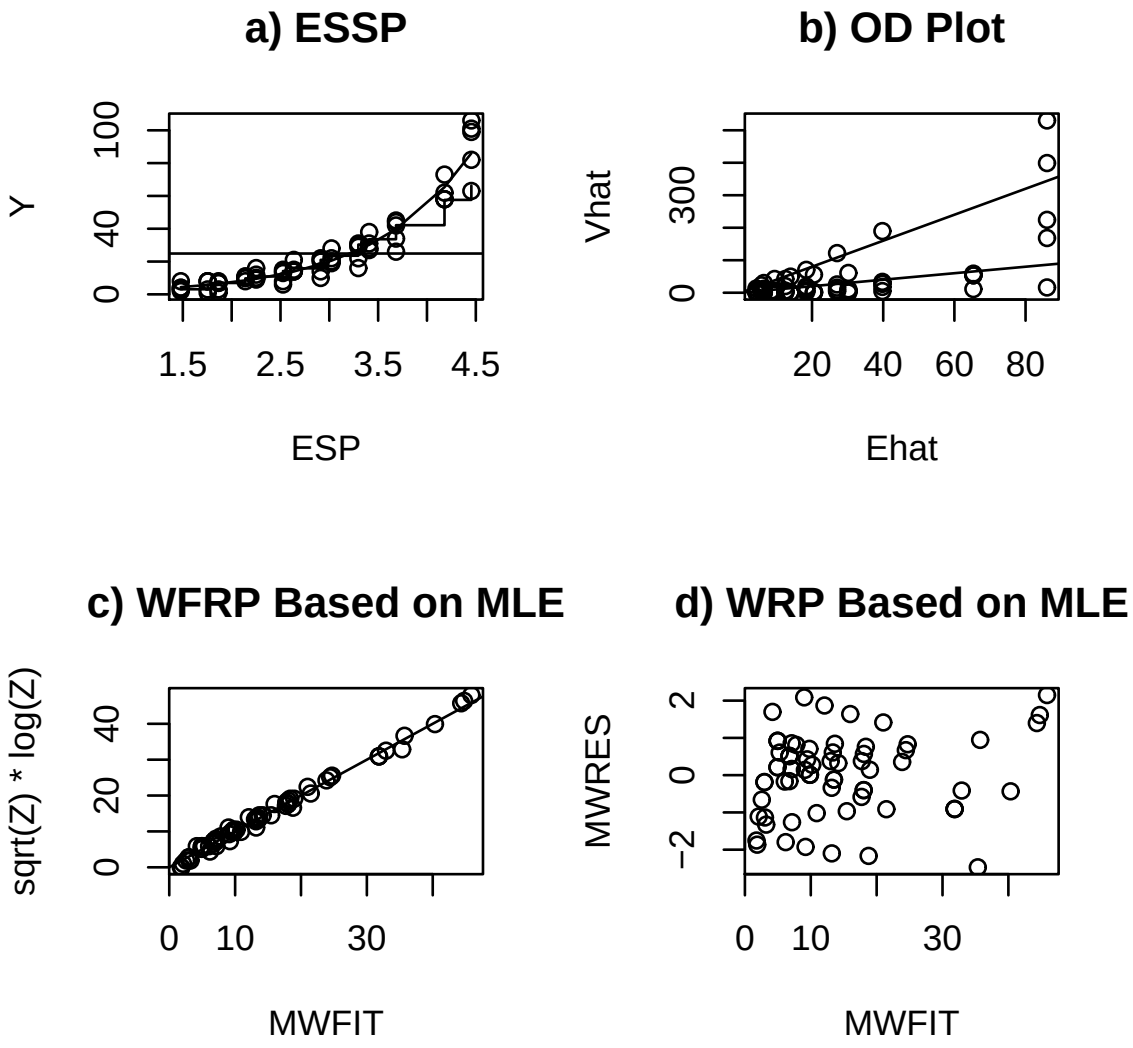

Figure 1: Plots for Ceriodaphnia Data

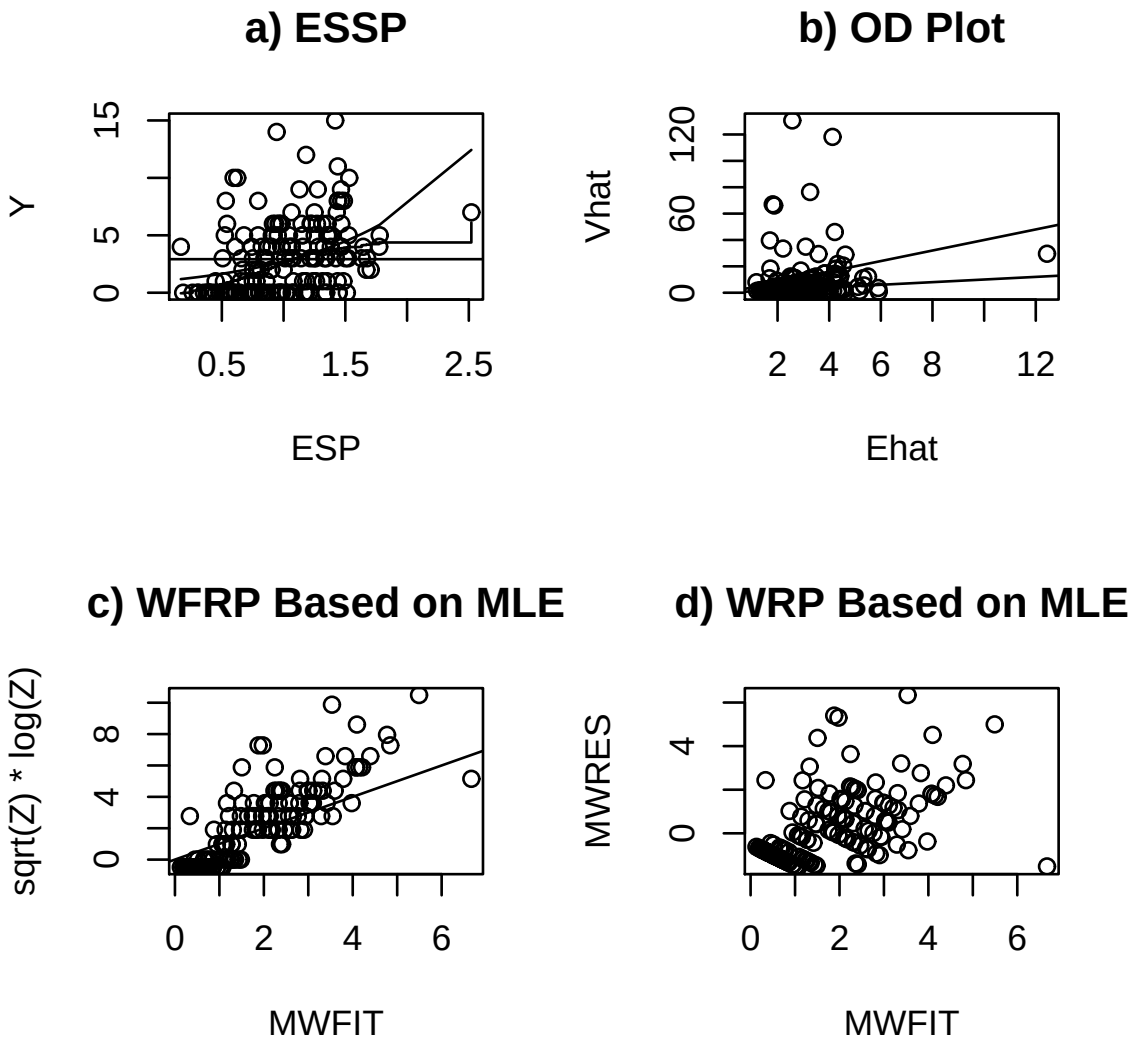

Figure 2: Plots for Crab Data

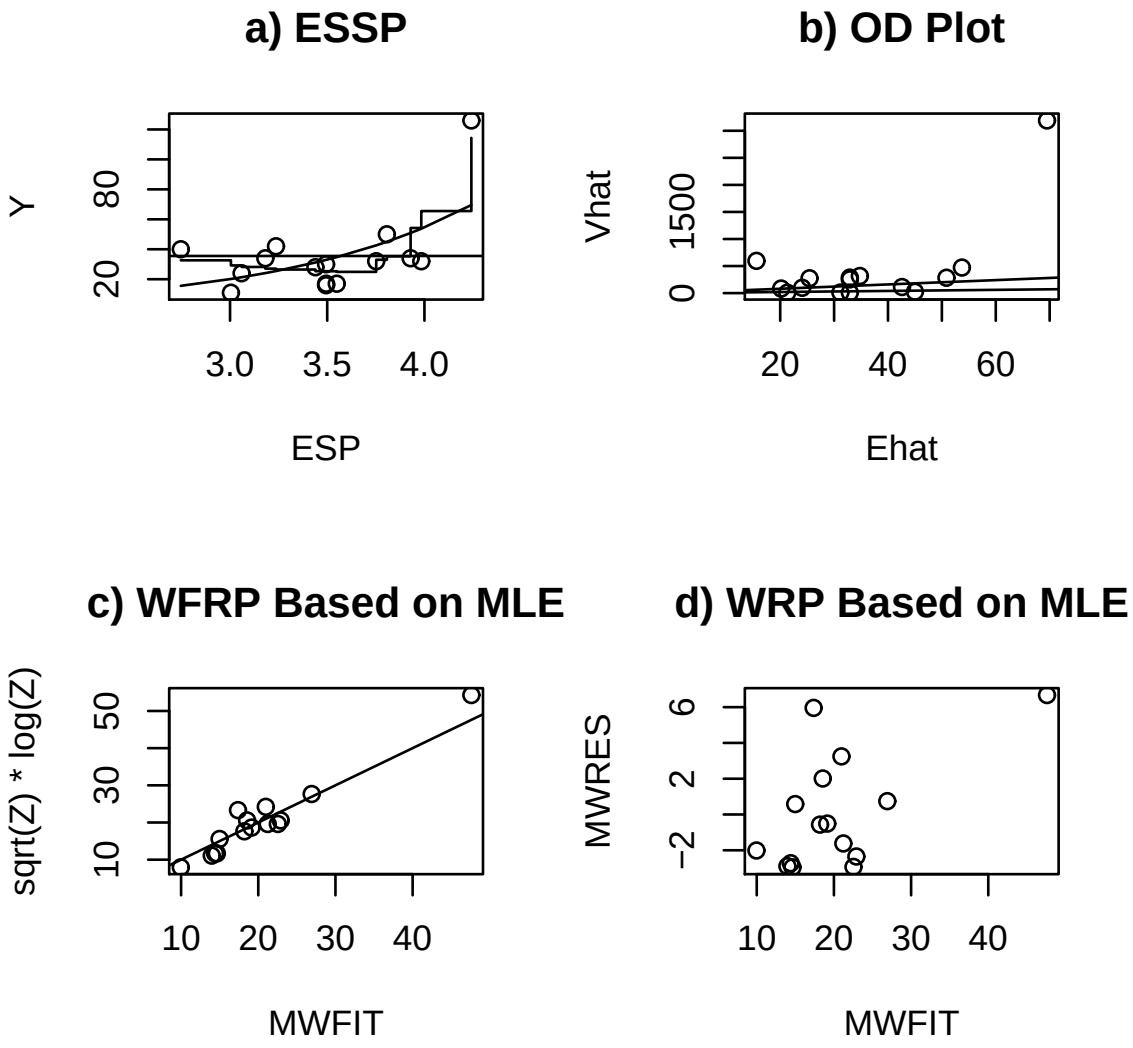

Figure 3: Plots for Popcorn Data

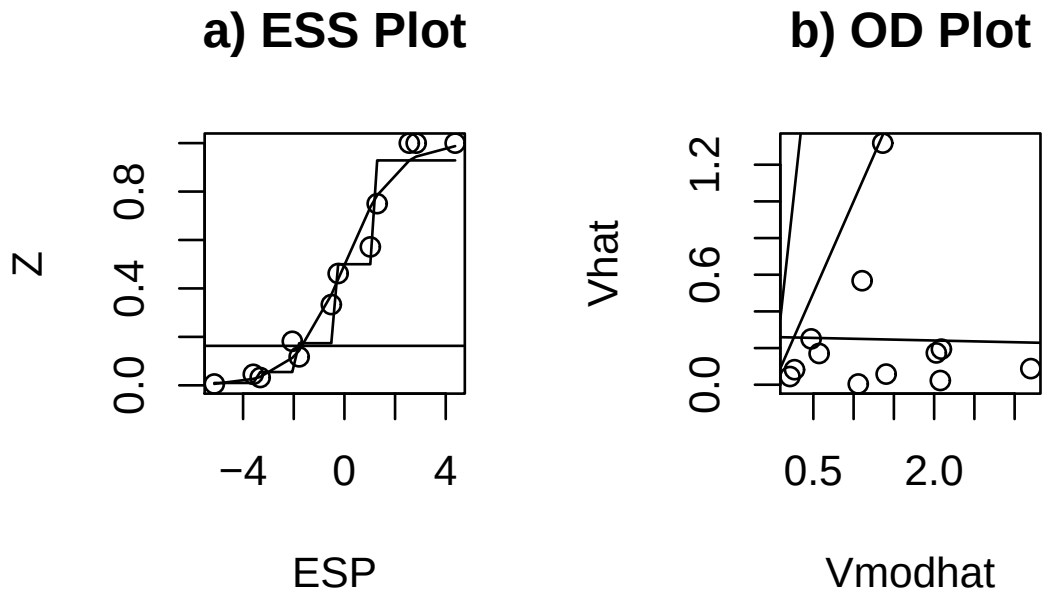

Figure 4: Visualizing the Death Penalty Data

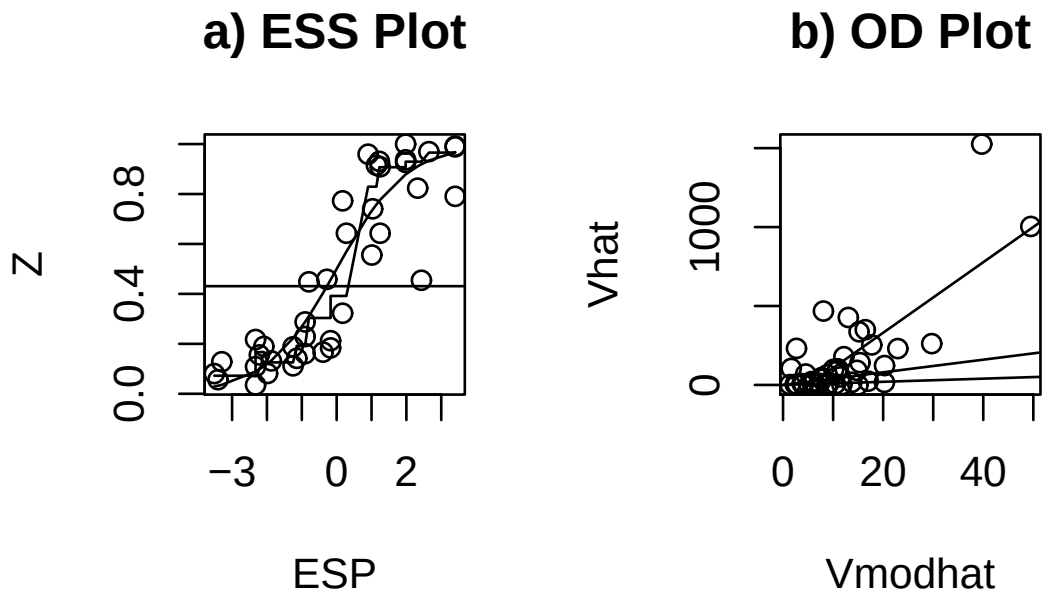

Figure 5: Plots for Rotifer Data

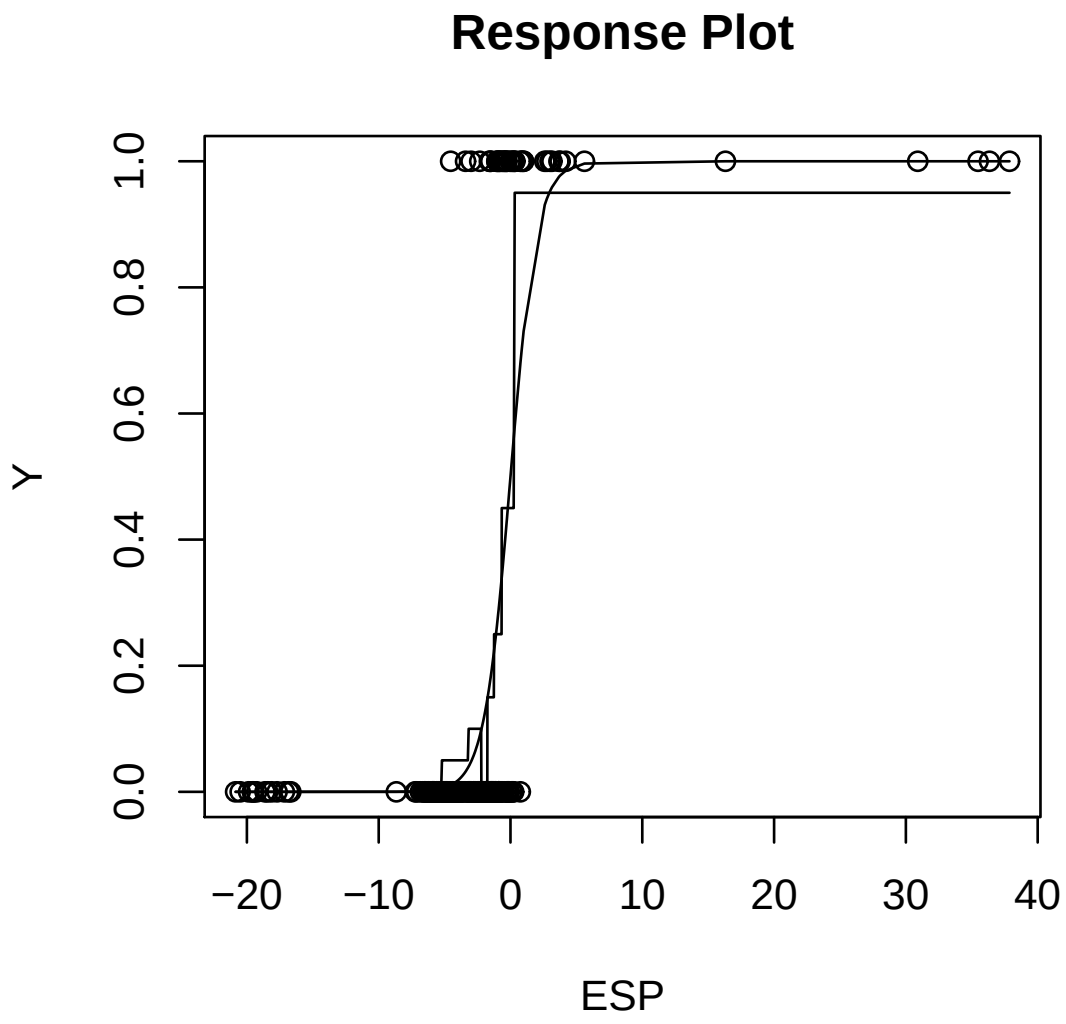

Figure 6: Visualizing the ICU Data

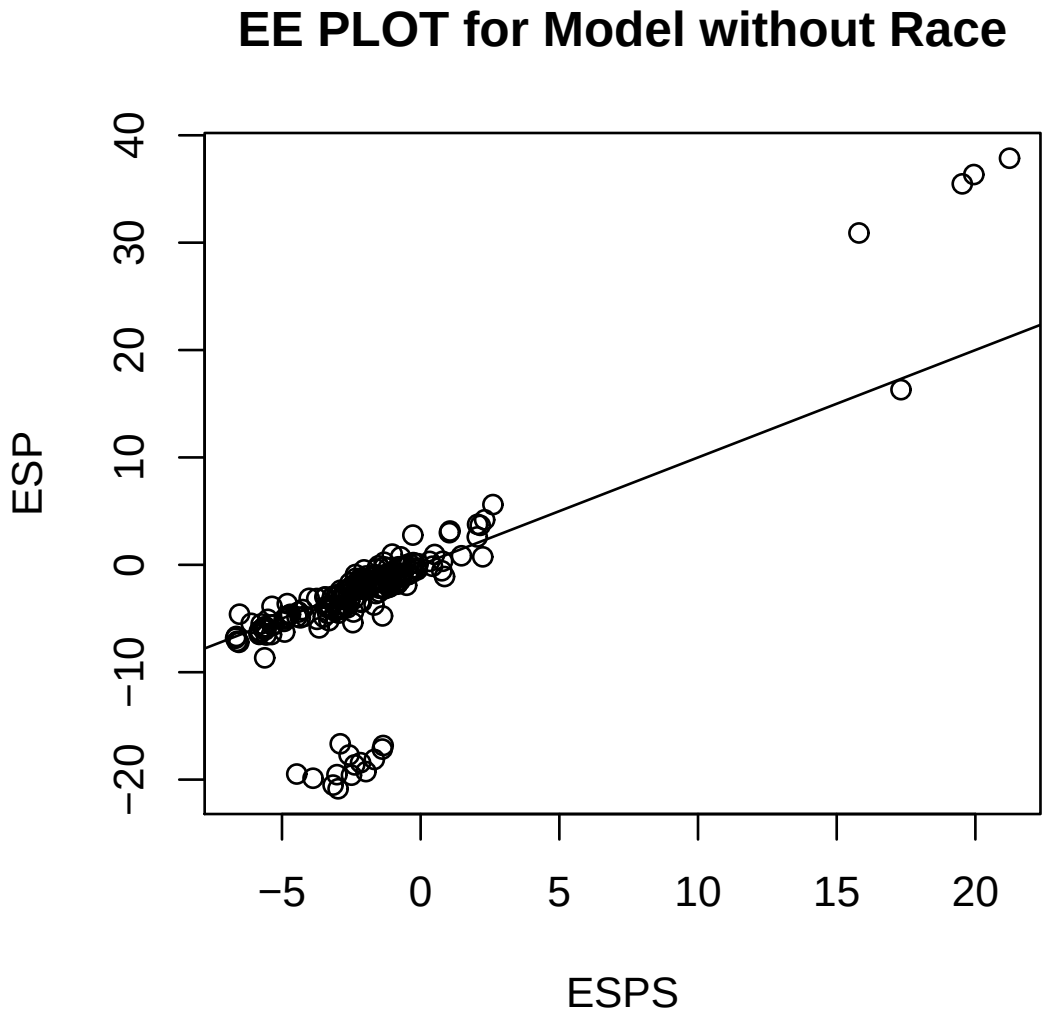

Figure 7: EE Plot Suggests Race is an Important Predictor

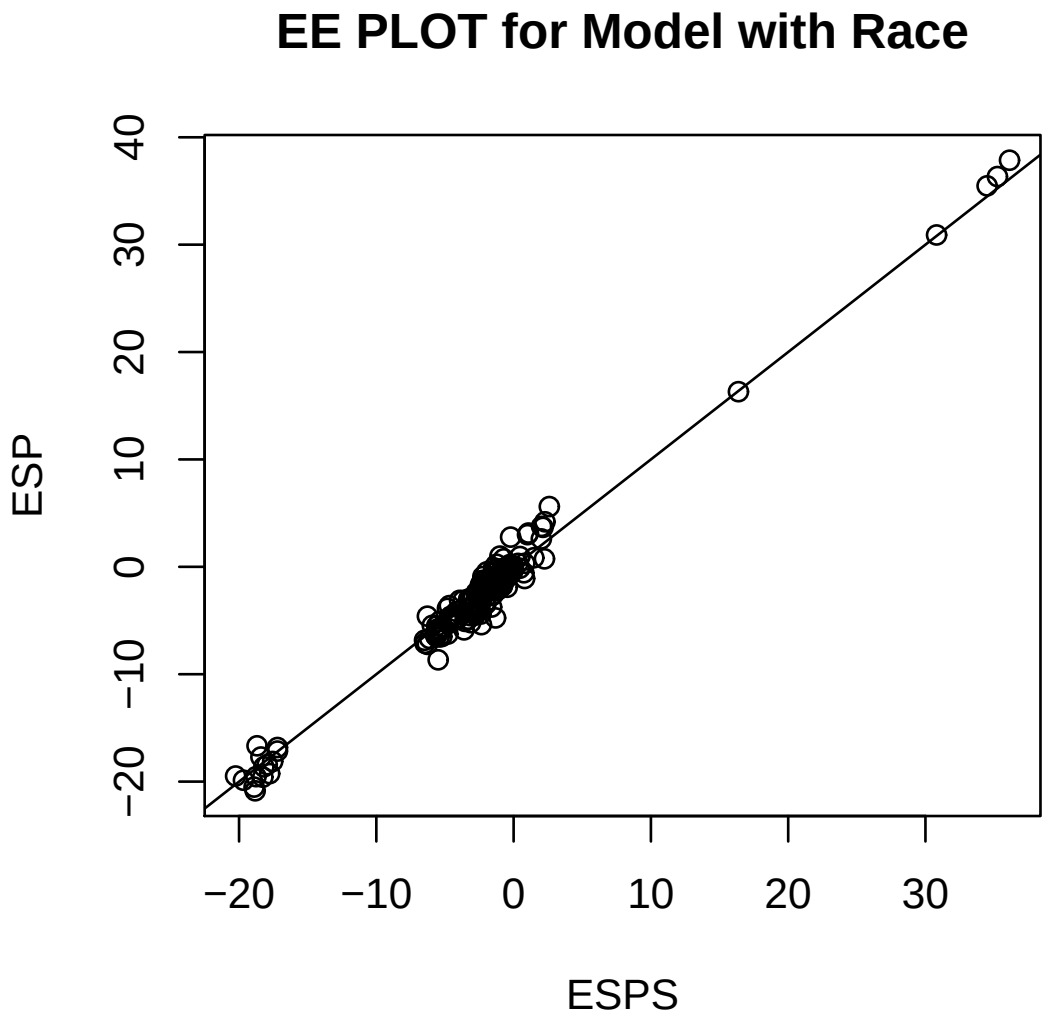

Figure 8: EE Plot Suggests Race is an Important Predictor