Response Plots and Related Plots for Regression

David Olive
Southern Illinois University

In a *1D regression model*, the response variable of interest Y is independent of the vector of predictors $\boldsymbol{x}$ given a single linear combination $\boldsymbol{x}^T\boldsymbol{\beta}$ of the predictors, written

$$Y \perp\!\!\!\perp \boldsymbol{x} | \boldsymbol{x}^T\boldsymbol{\beta}.$$

See Cook and Weisberg (1999, pp. 414-415).

The **main point of this paper** is that the estimated sufficient predictor

$$ESP = \boldsymbol{x}^T\hat{\boldsymbol{\beta}}$$

should be used to form plots to examine the regression model.

i) For the linear model $Y = \boldsymbol{x}^T\boldsymbol{\beta} + e$, the *residual plot* of $ESP = \boldsymbol{x}^T\hat{\boldsymbol{\beta}} = \hat{Y}$ vs. the residual $r = Y - \hat{Y}$ is widely used. See Draper and Smith (1966), Anscombe (1961) and Anscombe and Tukey (1963).

ii) An important plot for 1D regression is the *response plot* of $ESP = \boldsymbol{x}^T\hat{\boldsymbol{\beta}}$ vs. the response Y. The response plot is useful for visualizing the regression model in the background of the data, for description and for outlier detection.

iii) The *EE plot* is a plot of one estimated sufficient predictor vs. another. For example, the EE plot can be used to compare the semiparametric Cox proportional hazards regression model with the parametric Weibull proportional hazards model. The EE plot can also be used with variable selection or hypothesis testing to compare a submodel or reduced model with the full model.

iv) To check for goodness of fit of the regression model, order ESP from the smallest to largest values and divide ESP into several groups of roughly equal size called slices. For generalized linear models, make the response plot and compare the slice means with the estimated model conditional mean function. For survival regression models, compute the estimated model survival function for an $\boldsymbol{x}$ selected from each slice with the Kaplan Meier estimator computed from all of the censored survival times in the slice.

From the dimension reduction and regression graphics literature, if

$$Y \perp\!\!\!\perp \boldsymbol{x} | \boldsymbol{x}^T\boldsymbol{\beta},$$

then

$$Y \perp\!\!\!\perp \boldsymbol{x} | (a + c\ \boldsymbol{x}^T\boldsymbol{\beta})$$

for any constants a and $c \neq 0$. The quantity $a + c\boldsymbol{\beta}^T\boldsymbol{x}$ is called a *sufficient predictor* (SP), and an *estimated sufficient predictor* (ESP) is $\tilde{\alpha} + \boldsymbol{x}^T\tilde{\boldsymbol{\beta}}^T$ where $\tilde{\boldsymbol{\beta}}$ is an estimator of $d\boldsymbol{\beta}$ for some nonzero constant d. If $Y \perp\!\!\!\perp \boldsymbol{x} | (\alpha + \boldsymbol{x}^T\boldsymbol{\beta})$, then often the $ESP = \hat{\alpha} + \boldsymbol{x}^T\hat{\boldsymbol{\beta}}$ will be used. An *estimated sufficient summary plot* of ESP vs. Y is a response plot, and is used to study the conditional distribution of $Y|\boldsymbol{x}$. See Cook (1998, p. 10) and Cook and Weisberg (1999, p. 417).

If there is one predictor x, then the widely used *scatterplot* of x vs. Y is a response plot, where the estimated model conditional mean function is often added to the plot.

See, for example, the cover of Agresti (2002) for logistic regression and the cover of Cook and Weisberg (1999) for simple linear regression.

For more than one nontrivial predictor, the response plot has long been used to visualize the coefficient of determination R^2 in multiple linear regression. See Chambers, Cleveland, Kleiner and Tukey (1983, p. 280). Brillinger (1983) recognized that the response plot can be used to visualize the conditional mean function $E(Y|SP) = m(SP)$ of an additive error single index model

$$Y = m(SP) + e.$$

Also see Chang and Olive (2010). Response plots are called *marginal model plots* and *model checking plots* by Cook and Weisberg (1997, 1999, p. 396). Also see Sheather (2009, pp. 193-195).

Next, linear models such as multiple linear regression and experimental design models; response transformation models; generalized least squares; generalized linear models such as binary logistic regression, binomial logistic regression and Poisson regression; survival regression models such as the Cox proportional hazards regression, Weibull proportional hazards regression and Weibull accelerated failure time models give 1D regression examples. See Cook and Olive (2001), Olive (2004, 2009ab) and Olive and Hawkins (2005, 2010). *R software* is described in Olive (2008, 2010). Also see Sheather (2009). The Cook and Weisberg (1999) *Arc* software can be used for multiple linear, binary and Poisson regression.

Linear models: $Y = \boldsymbol{x}^T\boldsymbol{\beta} + e$. Assume that the conditional distribution $Y|\boldsymbol{x}^T\boldsymbol{\beta}$ has unknown pdf $f(y - \boldsymbol{x}^T\boldsymbol{\beta})$, a location family with location parameter $= SP = \boldsymbol{x}^T\boldsymbol{\beta}$. Then $Y|SP$ has pdf $f(y - SP)$, the error distribution has pdf $f(y)$, and the linear model is $Y = SP + e$ with conditional mean function $E(Y|SP) = SP$. The constant variance assumption is that the conditional variance function $V(Y|SP) \equiv \sigma^2$. If the error distribution is normal, $e \sim N(0, \sigma^2)$, then $Y|SP \sim N(SP, \sigma^2)$. The estimated sufficient predictor $ESP = \boldsymbol{x}^T\hat{\boldsymbol{\beta}} = \hat{Y}$ is the fitted value, and the residual $r = Y - \hat{Y}$ where $\hat{\boldsymbol{\beta}}$ is an estimator of $\boldsymbol{\beta}$. Note that $Y = ESP + r$, and the estimated conditional mean function is $\hat{E}(Y|\boldsymbol{x}) = \hat{E}(Y|SP) = ESP = \hat{Y} = \boldsymbol{x}^T\hat{\boldsymbol{\beta}}$. The estimated mean function $\hat{E}(Y|SP) = ESP$ is the identity line with unit slope and zero intercept.

Let the *iid error model* be the linear model where the zero mean constant variance errors are iid from a unimodal distribution that is not highly skewed. If the fitted values take on many values, then the plotted points should scatter about the identity line in a (roughly) evenly populated band while the plotted points in the residual plot should scatter about the $r = 0$ line in a (roughly) evenly plotted band. Deviations from the evenly populated band suggest that something is wrong with the iid error model. The response and residual plots are best for $n > 5p$.

Example 1. Tremearne (1911) presents a data set of about 17 measurements on 112 people of Hausa nationality. We used *height* as the response variable Y. Along with a constant $x_{i,1} \equiv 1$, the five additional predictor variables used were *height when sitting*, *height when kneeling*, *head length, nasal breadth*, and *span*. The OLS response and residual plots in Figure 1 show that multiple linear regression should be a useful model for the data since the plotted points in the response plot follow the identity line while the plotted points in the residual plot follow the $r = 0$ line with no other pattern.

To use the response plot to visualize the conditional distribution of $Y|\boldsymbol{x}^T\boldsymbol{\beta}$, use the fact that the fitted values $\hat{Y} = \boldsymbol{x}^T\hat{\boldsymbol{\beta}}$. For example, suppose the height given fit = 1700 is of interest. Mentally examine the plot about a narrow vertical strip about fit = 1700, perhaps from 1675 to 1725. The cases in the narrow strip have a mean close to 1700 since they fall close to the identity line. Similarly, when the fit = w for w between 1500 and 1850, the cases have heights near w, on average.

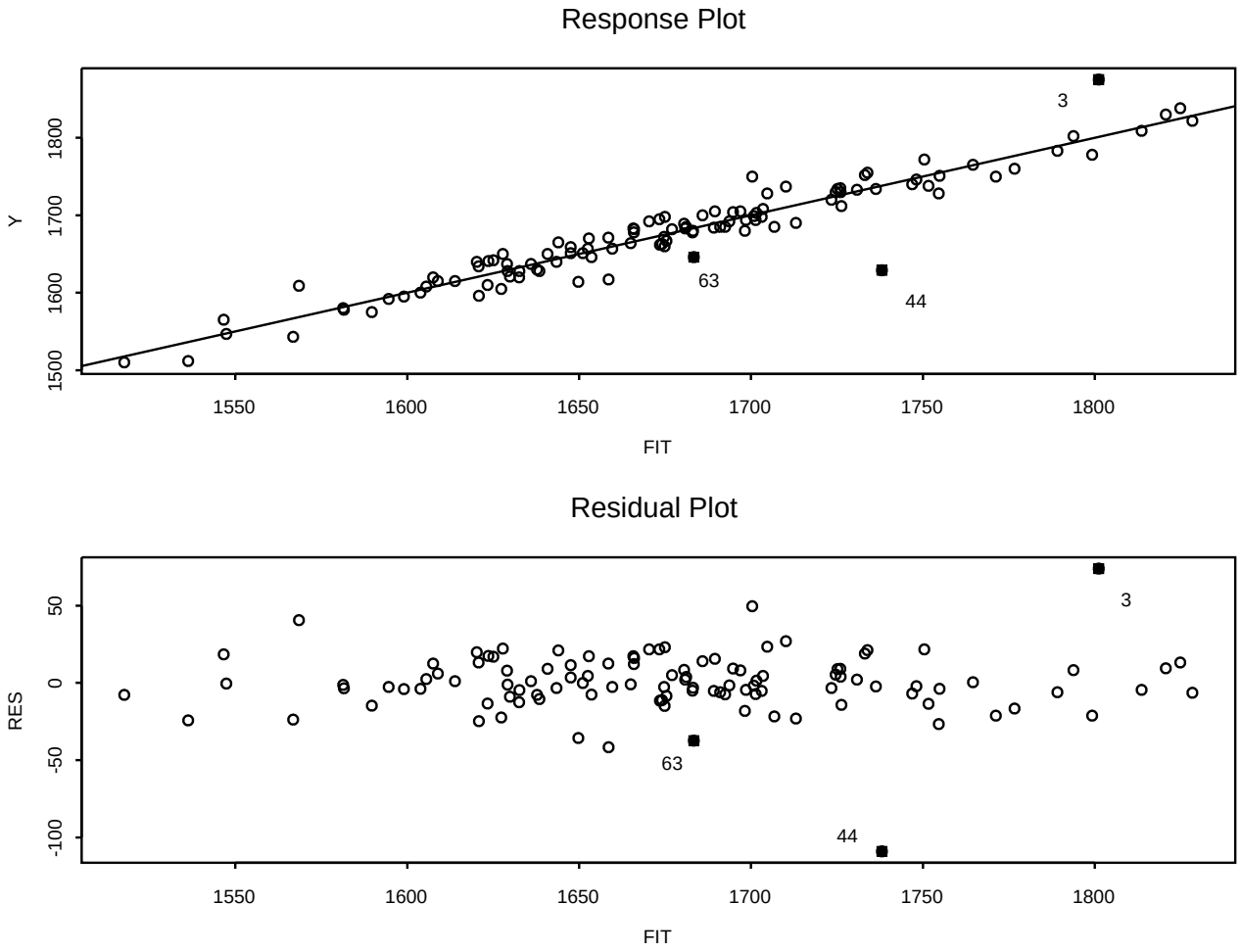

Figure 1: Residual and Response Plots for the Tremearne Data

Many experimental design models satisfy the iid error model, but often the fitted values do not take on many values. Consider the one way anova model with k treatments and $m \geq 5$ replications per treatment. The plotted points still scatter about the identity or $r = 0$ line, but there are k dot plots corresponding to the k treatments. The dot plots should have similar spread and shape if the one way anova model assumptions are reasonable.

Example 2. SAS Institute (1985, p. 126) uses clover data to illustrate the one way anova model. The response variable is the nitrogen content of red clover plants

inoculated with six strains of bacteria, and each strain has five replicates. Figure 2 shows the response and residual plots. The one way anova F test is approximately correct if $\max(R_1, ..., R_k) \leq 2\min(R_1, ..., R_k)$ where R_i is the range of the ith dot plot. Linearity seems reasonable, but the approximately constant variance assumption may not hold.

The response transformation model $Y = t(Z) = \boldsymbol{x}^T\boldsymbol{\beta} + e$ can be checked by making the response and residual plots. If $t_\lambda(Z) = Z^\lambda$ for $\lambda \neq 0$ and $t_0(Z) = \log(Z)$ for $\lambda \in \{-1, -1/3, -1/2, 0, 1/3, 1/2, 1\}$, make the 7 plots and choose the transformation with the best response and residual plots.

Generalized least squares (GLS): The GLS model $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{e}$ is a linear model (written in matrix form) with $E(\boldsymbol{e}) = \boldsymbol{0}$, but $\mathrm{Cov}(\boldsymbol{e}) = \sigma^2\boldsymbol{V}$ where $\boldsymbol{V}$ is a known $n \times n$ positive definite matrix. The *weighted least squares* (WLS) model with weights $w_1, ..., w_n$ is the special case of the GLS model where $\boldsymbol{V}$ is diagonal: $\boldsymbol{V} = \mathrm{diag}(\mathrm{v}_1, ..., \mathrm{v}_\mathrm{n})$ and $w_i = 1/v_i$. The *GLS estimator*

$$\hat{\boldsymbol{\beta}}_{GLS} = (\boldsymbol{X}^T\boldsymbol{V}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{V}^{-1}\boldsymbol{Y}.$$

The fitted values are $\hat{\boldsymbol{Y}}_{GLS} = \boldsymbol{X}\hat{\boldsymbol{\beta}}_{GLS}$.

Following Freedman (2005, p. 54), the *feasible generalized least squares* (FGLS) model is the same as the GLS estimator except that $\boldsymbol{V} = \boldsymbol{V}(\boldsymbol{\theta})$ is a function of an unknown $q \times 1$ vector of parameters $\boldsymbol{\theta}$. Let the estimator of $\boldsymbol{V}$ be $\hat{\boldsymbol{V}} = \boldsymbol{V}(\hat{\boldsymbol{\theta}})$. Then the FGLS estimator

$$\hat{\boldsymbol{\beta}}_{FGLS} = (\boldsymbol{X}^T\hat{\boldsymbol{V}}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}^T\hat{\boldsymbol{V}}^{-1}\boldsymbol{Y}.$$

The fitted values are $\hat{\boldsymbol{Y}}_{FGLS} = \boldsymbol{X}\hat{\boldsymbol{\beta}}_{FGLS}$. The *feasible weighted least squares* (FWLS) estimator is the special case of the FGLS estimator where $\boldsymbol{V} = \boldsymbol{V}(\boldsymbol{\theta})$ is diagonal. Hence the estimated weights $\hat{w}_i = 1/\hat{v}_i = 1/v_i(\hat{\boldsymbol{\theta}})$.

The GLS estimator can be transformed to a linear model $\boldsymbol{Z} = \boldsymbol{U}\boldsymbol{\beta} + \boldsymbol{\epsilon}$ where $E(\boldsymbol{\epsilon}) = \boldsymbol{0}$ and $\mathrm{Cov}(\boldsymbol{\epsilon}) = \sigma^2\boldsymbol{I}_n$. From the spectral theorem, there is a symmetric, nonsingular $n \times n$ matrix $\boldsymbol{R}$ such that $\boldsymbol{V} = \boldsymbol{R}\boldsymbol{R}$. Let $\boldsymbol{Z} = \boldsymbol{R}^{-1}\boldsymbol{Y}$, $\boldsymbol{U} = \boldsymbol{R}^{-1}\boldsymbol{X}$ and $\boldsymbol{\epsilon} = \boldsymbol{R}^{-1}\boldsymbol{e}$. This method has better numerical properties than the transformation based on the Cholesky decomposition.

The response and residual plots can be made for the transformed model to check the linearity and constant variance assumptions as in the previous section, assuming that the distribution of ϵ is not highly skewed. If the plots are good, then the GLS model may be a reasonable approximation for the data. Similar plots can be made for FGLS since the FGLS estimator can also be found from the OLS regression (without an intercept) of $\boldsymbol{Z}$ on $\boldsymbol{U}$ where $\boldsymbol{V}(\hat{\boldsymbol{\theta}}) = \boldsymbol{R}\boldsymbol{R}$. But now $\boldsymbol{U}$ is a random matrix instead of a constant matrix.

The plots based on the transformed model give both a check on linearity and on whether the model using $\boldsymbol{V}$ (or $\hat{\boldsymbol{V}}$) gives a good approximation of the data, provided that $n > 5(p + q)$ where $q = 0$ for GLS.

If the plots for the transformed model show high leverage points or outliers while the response plot based on OLS is linear without outliers, then the GLS model may be poor. Then it may be better to use the consistent but inefficient OLS estimator along with the sandwich estimator.

Sheather (2009, ch. 9, ch. 10) makes the residual plots based on the Cholesky decomposition and shows that many linear models with serially correlated errors (e.g.

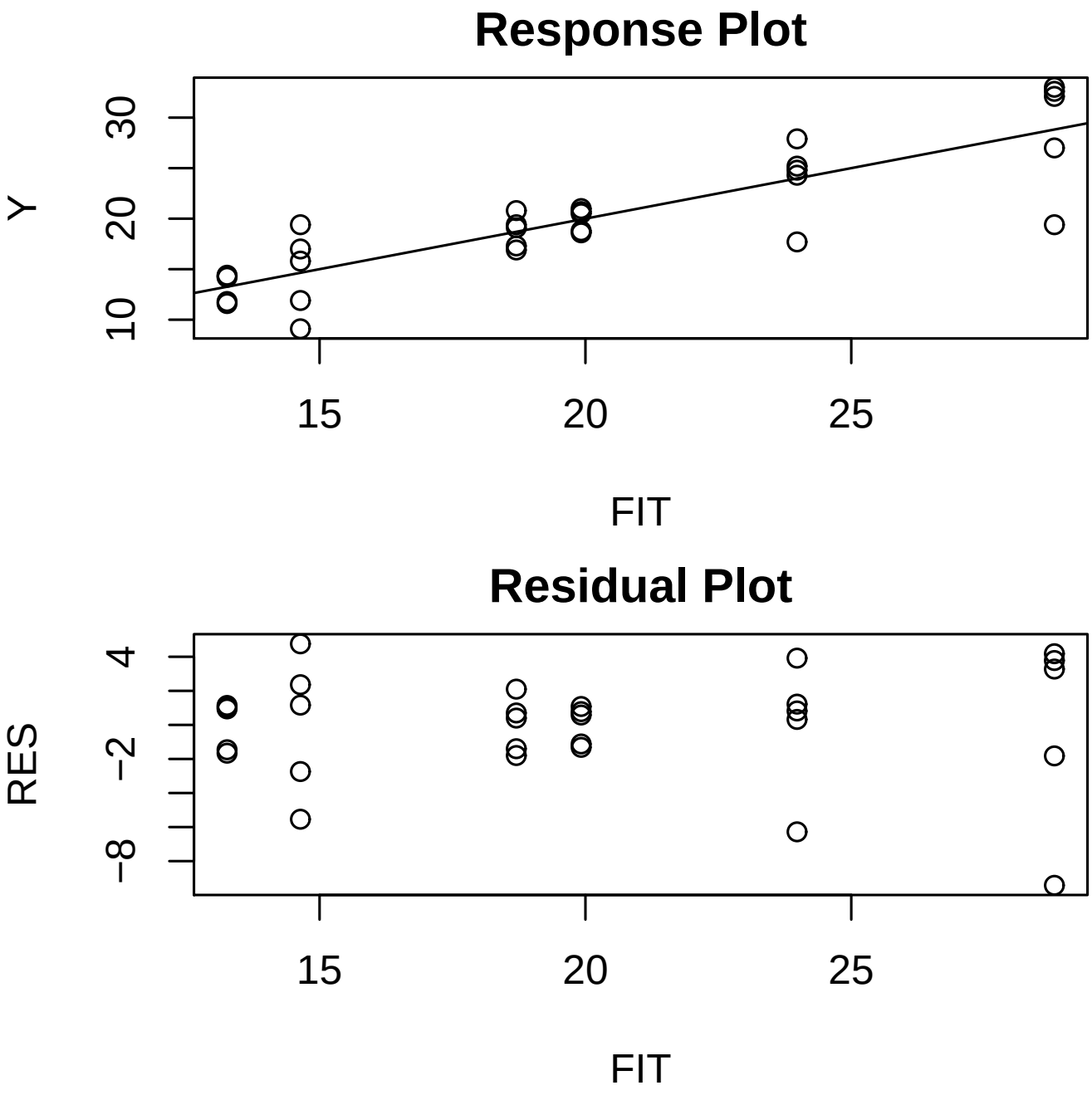

Figure 2: SAS One Way Anova Data

AR(1) errors) and many linear mixed models can be fit with FGLS. Houseman, Ryan and Coull (2004) also use the Cholesky decomposition. Montgomery, Peck and Vining (2006, pp. 182-183) make residual plots based on the spectral theorem.

Generalized Linear Models (GLM): 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 $\hat{V} = [Y - \hat{E}(Y|SP)]^2$. For binomial and Poisson regression (Winkelmann 2000, p. 110), the OD plot can be used to complement tests and diagnostics for overdispersion.

For Poisson regression, 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. 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%. Similar remarks apply to binomial regression if the counts are neither too big nor too small.

For these two regression models, the deviance G^2 is approximately chi-square with $df = n - p - 1$. The 98th percentile of the χ^2_d distribution is approximately $d + 3\sqrt{d}$. If $G^2 > (n - p - 1) + 3\sqrt{n - p - 1}$, then more complicated models may be needed.

The *binary regression model* has $Y|\boldsymbol{x} \sim \text{binomial}(1, \rho(\alpha + \boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{x}))$, or

$$Y|SP \sim \text{binomial}(1, \rho(\text{SP})).$$

Note that the conditional mean function $E(Y|SP) = \rho(SP)$ and the conditional variance function $V(Y|SP) = \rho(SP)(1 - \rho(SP))$.

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

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

or

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

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 *logistic regression (LR) model* is the special case of binomial regression where

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

Equivalently,

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

Note that $\rho(x)$ is the CDF of a logistic(0,1) distribution.

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

$$\hat{E}(Z|SP) = \frac{\exp(ESP)}{1 + \exp(ESP)}$$

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. 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).

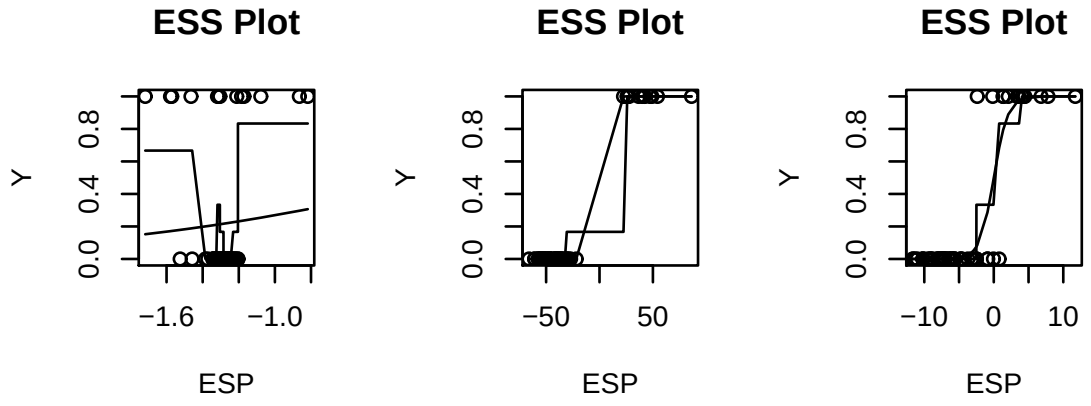

Figure 3: Plots for Museum Data

Example 3. Schaaffhausen (1878) gives data on skulls at a museum. The 1st 47 skulls are humans while the remaining 13 are apes. The response variable *ape* is 1 for an ape skull. The left plot in Figure 3 uses the predictor *face length*. The model fits very poorly since the probability of a 1 decreases then increases. The middle plot uses the predictor

head height and perfectly classifies the data since the ape skulls can be separated from the human skulls with a vertical line at ESP = 0. The right plot uses predictors *lower jaw length, face length,* and *upper jaw length*. None of the predictors is good individually, but together provide a good LR model since the observed proportions (the step function) track the model proportions (logistic curve) closely. For binary regression, overdispersion is not a problem and residuals behave poorly.

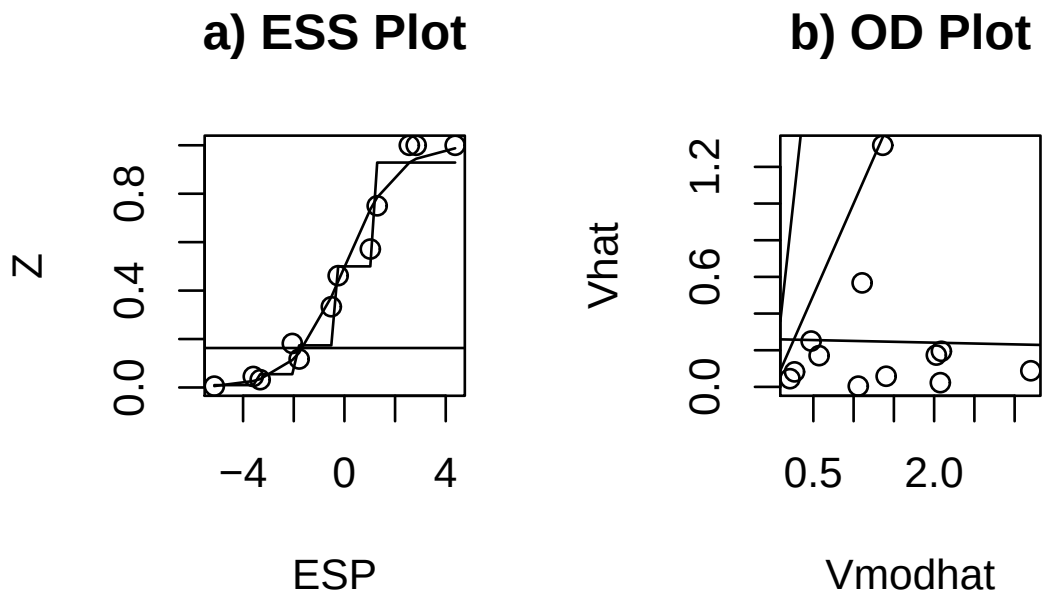

Figure 4: Visualizing the Death Penalty Data

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). 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 Poisson regression model is

$$Y|SP \sim \text{Poisson}(\mu(\text{SP}))$$

where $\mu(SP) = \exp(SP)$ is both the conditional mean and conditional variance function. For Poisson regression, the data can be transformed towards a linear model, then make the response plot and residual plot for the transformed data. 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.

Example 5. 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 5 shows the 4 plots for this data. In the ESSP of Figure 5a, the lowess curve is represented as a jagged curve to distinguish it from the estimated 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 5c and there are no unusual points in Figure 5d. The four plots suggest that the Poisson regression model is a useful approximation to the data. Hence $Y|ESP \approx \text{Poisson}(\exp(\text{ESP}))$. 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 5a.

Example 6. The ICU data studies the survival of 200 patients following admission

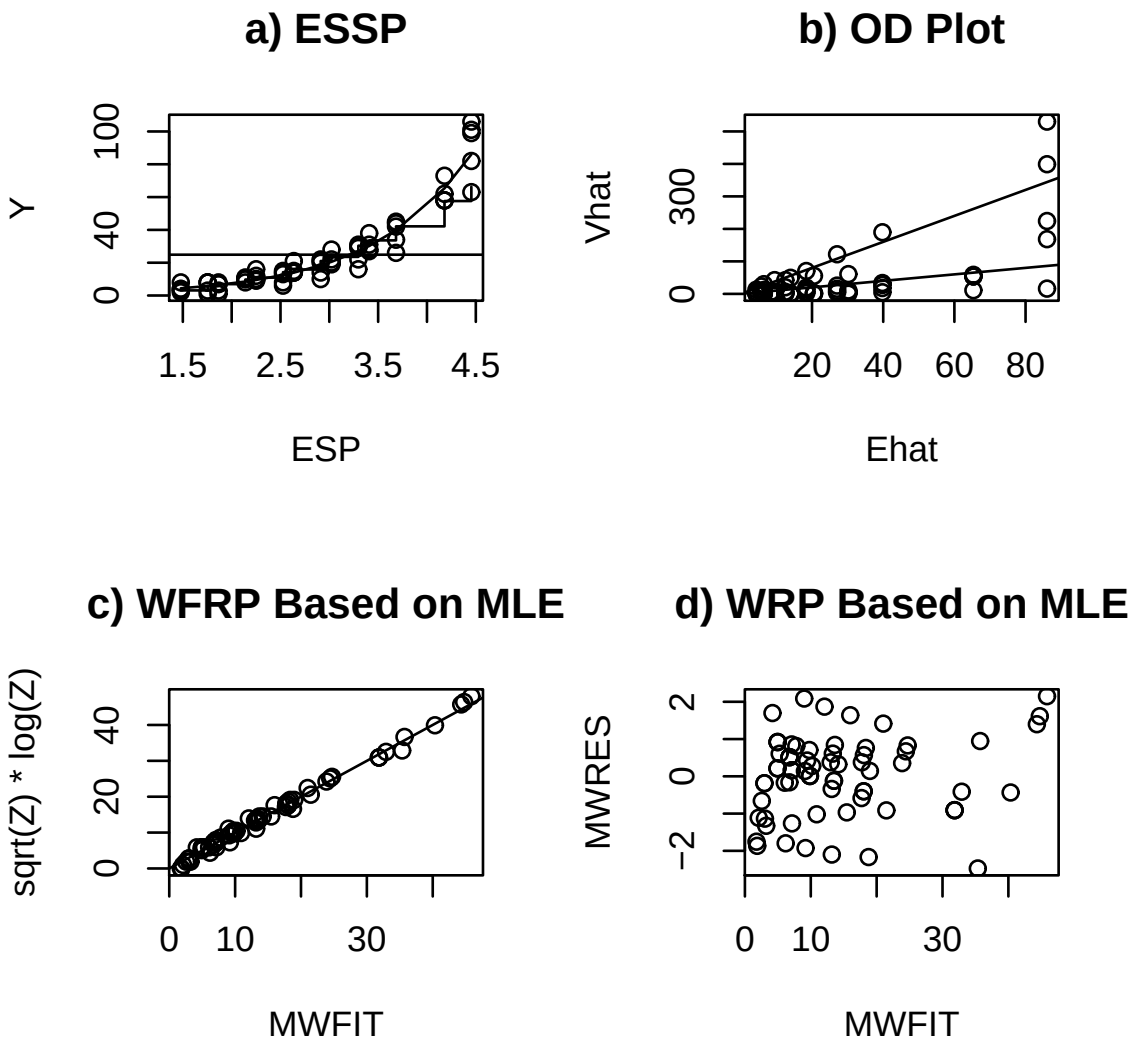

Figure 5: Plots for Ceriodaphnia Data

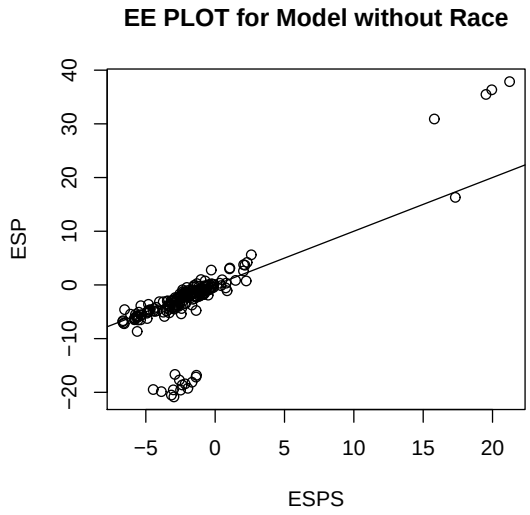

Figure 6: EE Plot Suggests Race is an Important Predictor

to an intensive care unit. The response variable was STA (0 = Lived, 1 = Died). The 19 predictors were primarily indicator variables describing the health of the patient at time of admission, but two factors had 3 levels including RACE (1 = White, 2 = Black, 3 = Other). The response plot showed that the full model using the 19 predictors was useful for predicting survival. Variable selection suggested a submodel using five predictors. The EE plot of the submodel ESP vs. full model ESP is shown in Figure 6. 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 6. The lowest cluster of points and the case on the right nearest to the identity line correspond to black patients. The main cluster and upper right cluster correspond to patients who are not black. When RACE is added to the submodel, all of the points cluster about the identity line. Although variable selection did not suggest that RACE is important, the above results suggest that RACE is important. Also the RACE variable could be replaced by an indicator for black.

Survival Regression: ESP = estimated risk score. The conditional distribution $Y|\boldsymbol{x}$ is completely determined by pdf $f_{\boldsymbol{x}}(t)$, the survival function

$$S_{\boldsymbol{x}}(t) \equiv S_{Y|SP}(t) = P(Y > t|SP = \boldsymbol{\beta}^T\boldsymbol{x}),$$

or the hazard function $h_{\boldsymbol{x}}(t) = f_{\boldsymbol{x}}(t)/S_{\boldsymbol{x}}(t)$ for $t > 0$. High hazard implies low survival times while low hazard implies long survival times.

Survival data is usually right censored so Y is not observed. Instead, the survival time $T_i = \min(Y_i, Z_i)$ where $Y_i \perp\!\!\!\perp Z_i$ and Z_i is the censoring time. Also $\delta_i = 0$ if $T_i = Z_i$ is censored and $\delta_i = 1$ if $T_i = Y_i$ is uncensored. Hence the data is $(T_i, \delta_i, \boldsymbol{x}_i)$ for $i = 1, ..., n$.

The *Cox proportional hazards* (PH) regression model (Cox 1972) is a semiparametric model with $SP = \boldsymbol{\beta}_C^T\boldsymbol{x}$ and

$$h_{\boldsymbol{x}}(t) \equiv h_{Y|SP}(t) = \exp(\boldsymbol{\beta}_C^T\boldsymbol{x})h_0(t) = \exp(SP)h_0(t)$$

where the baseline hazard function $h_0(t)$ is left unspecified. The survival function is

$$S_{\boldsymbol{x}}(t) \equiv S_{Y|SP}(t) = [S_0(t)]^{\exp(\boldsymbol{\beta}_C^T \boldsymbol{x})} = [S_0(t)]^{\exp(SP)}.$$

First $\boldsymbol{\beta}_C$ is estimated by the maximum partial likelihood estimator $\hat{\boldsymbol{\beta}}_C$, then estimators $\hat{h}_0(t)$ and $\hat{S}_0(t)$ can be found (see Breslow 1974), and

$$\hat{S}_{\boldsymbol{x}}(t) = [\hat{S}_0(t)]^{\exp(\hat{\boldsymbol{\beta}}_C^T \boldsymbol{x})} = [\hat{S}_0(t)]^{\exp(ESP)}. \tag{1}$$

For *parametric proportional hazards* regression models, the baseline function is parametric and the parameters are estimated via maximum likelihood. Then $SP = \boldsymbol{\beta}_P^T \boldsymbol{x}$,

$$h_{\boldsymbol{x}}(t) = \exp(\boldsymbol{\beta}_P^T \boldsymbol{x}) h_{0,P}(t),$$

the survival function is

$$S_{\boldsymbol{x}}(t) \equiv S_{Y|SP}(t) = [S_{0,P}(t)]^{\exp(\boldsymbol{\beta}_P^T \boldsymbol{x})} = [S_{0,P}(t)]^{\exp(SP)}, \tag{2}$$

and

$$\hat{S}_{\boldsymbol{x}}(t) = [\hat{S}_{0,P}(t)]^{\exp(\hat{\boldsymbol{\beta}}_P^T \boldsymbol{x})} = [\hat{S}_{0,P}(t)]^{\exp(ESP)}. \tag{3}$$

For a parametric *accelerated failure time* (AFT) model,

$$\log(Y) = \alpha + \boldsymbol{\beta}_A^T \boldsymbol{x} + \sigma e \tag{4}$$

where the e are iid from a location scale family. The parameters are again estimated by maximum likelihood and the survival function is

$$S_{\boldsymbol{x}}(t) \equiv S_{Y|\boldsymbol{x}}(t) = S_0\left(\frac{t}{\exp(\boldsymbol{\beta}_A^T \boldsymbol{x})}\right), \tag{5}$$

and

$$\hat{S}_{\boldsymbol{x}}(t) = \hat{S}_0\left(\frac{t}{\exp(\hat{\boldsymbol{\beta}}_A^T \boldsymbol{x})}\right) \tag{6}$$

where $\hat{S}_0(t)$ depends on $\hat{\alpha}$ and $\hat{\sigma}$. If $SP = -\boldsymbol{\beta}_A^T \boldsymbol{x}$, then $h(t|SP) = h_0(t\ e^{SP}) e^{SP}$.

Chen and Jewell (2001) suggest the accelerated hazards model $h(t|SP) = h_0(t\ e^{SP})$ and the 2D regression model $h(t|\boldsymbol{x}) = h_0(t\ e^{\boldsymbol{\beta}_1^T \boldsymbol{x}}) e^{\boldsymbol{\beta}_2^T \boldsymbol{x}}$.

For Weibull regression, the AFT has $e_i \sim SEV(0,1)$. Thus
$\log(Y)|\boldsymbol{x} \sim SEV(\alpha + \boldsymbol{\beta}_A^T \boldsymbol{x}, \sigma)$, and as a proportional hazards model,
$Y|\boldsymbol{x} \sim W(\gamma = 1/\sigma, \lambda_{\boldsymbol{x}})$ where

$$\lambda_{\boldsymbol{x}} = \exp\left[-\left(\frac{\alpha}{\sigma} + \frac{\boldsymbol{\beta}_A^T \boldsymbol{x}}{\sigma}\right)\right] = \lambda_0 \exp(\boldsymbol{\beta}_P^T \boldsymbol{x})$$

with $\lambda_0 = \exp(-\alpha/\sigma)$ and $\boldsymbol{\beta}_P = -\boldsymbol{\beta}_A/\sigma$. Thus for $t > 0$, $P(Y > t|\boldsymbol{x}) = S_{\boldsymbol{x}}(t)$

$$= \exp(-\lambda_{\boldsymbol{x}} t^\gamma) = \exp(-\lambda_0 \exp(\boldsymbol{\beta}_P^T \boldsymbol{x}) t^\gamma) = [\exp(-\lambda_0 t^\gamma)]^{\exp(\boldsymbol{\beta}_P^T \boldsymbol{x})} = [S_{0,P}(t)]^{\exp(\boldsymbol{\beta}_P^T \boldsymbol{x})}.$$

Exponential regression is the special case where $\sigma = 1$.

Grambsch and Therneau (1994) give a useful graphical check for the proportional hazards model. For each variable, a plot based on scaled Schoenfeld residuals is made with the loess curve added. If the loess curve is approximately horizontal for each of the p plots, then the proportional hazards assumption is reasonable. Alternatively, fit a line to each plot and test that each of the p slopes is equal to 0. The *R/Splus* function `cox.zph` makes both the plots and tests.

The *slice survival plot* divides the ESP into J groups of roughly the same size. For each group j, $\hat{S}_j(t)$ is computed using an $\boldsymbol{x}$ corresponding to the middle ESP of the group. (The "middle ESP" is the kth order statistic of the ESP in group j, where $k = 1 + floor[(n_j - 1)/2]$ and n_j is the number of cases in group j.) Let $\hat{S}_{KMj}(t)$ be the Kaplan Meier estimator computed from the survival times (T_i, δ_i) in the jth group. For each group, $\hat{S}_j(t)$ is plotted and $\hat{S}_{KMj}(t_i)$ as circles at the uncensored event times t_i. The survival regression model is reasonable if the circles "track the curve well" in each of the J plots.

For the Cox model, if pointwise confidence interval (CI) bands are added to the plot, then $\hat{S}_{KMj}$ "tracks $\hat{S}_j$ well" if most of the plotted circles do not fall very far outside the pointwise CI bands since these pointwise bands are not as wide as simultaneous bands. Collett (2003, pp. 241-243) places several observed Kaplan Meier curves with fitted curves on the same plot.

The slice survival plot tailored to the Cox model is closely related to May and Hosmer (1998) test, and the plot has been suggested by several authors with $\boldsymbol{x}$ divided into J groups instead of the ESP. For example, see Miller (1981, p. 168). Hosmer and Lemeshow (1999, pp. 141–145) suggests making plots based on the quartiles of the ith predictor x_i, and note that a problem with Cox survival curves is that they may use inappropriate extrapolation. Using the ESP results in narrow slices with many cases, and adding Kaplan Meier curves shows if there is extrapolation.

A *censored response plot* is a plot of the ESP versus T with plotting symbol 0 for censored cases and + for uncensored cases. Slices in this plot correspond to the slices used in the slice survival plot. Suppose the ESP is a good estimator of the SP. Consider a narrow vertical slice taken in the censored response plot about $ESP = w$. The points in the slice are a censored sample with $S_{Y|SP}(t) \approx S_{Y|w}(t)$. For proportional hazards models, $h_{Y|SP}(t) \approx \exp(ESP)h_0(t)$, and the hazard increases while the survival decreases as the ESP increases. Gentleman and Crowley (1991) make this plot for 1 predictor.

Let $\log(T_i) = \hat{\alpha} + \hat{\boldsymbol{\beta}}_A^T \boldsymbol{x}_i + r_i$. For accelerated failure time models, a *log censored response (LCR) plot* is a plot of $\hat{\alpha} + \hat{\boldsymbol{\beta}}_A^T \boldsymbol{x}_i$ versus $\log(T_i)$ with plotting symbol 0 for censored cases and + for uncensored cases. The identity line with unit slope and zero intercept is added to the plot, and the vertical deviations from the identity line $= r_i$.

For parametric proportional hazards models, an *EE plot* is a plot of the parametric ESP $\hat{\boldsymbol{\beta}}_P^T \boldsymbol{x}$ versus the Cox semiparametric ESP $\hat{\boldsymbol{\beta}}_C^T \boldsymbol{x}$. If the parametric proportional hazards model is good, then the plotted points should track the identity line with unit slope and zero intercept. As $n \to \infty$, the correlation of the plotted points goes to 1 in probability for any finite interval, e.g., from the 1st percentile to the 99th percentile of $\hat{\boldsymbol{\beta}}_C^T \boldsymbol{x}$. Lack of fit is suggested if the plotted points do not cluster tightly about the

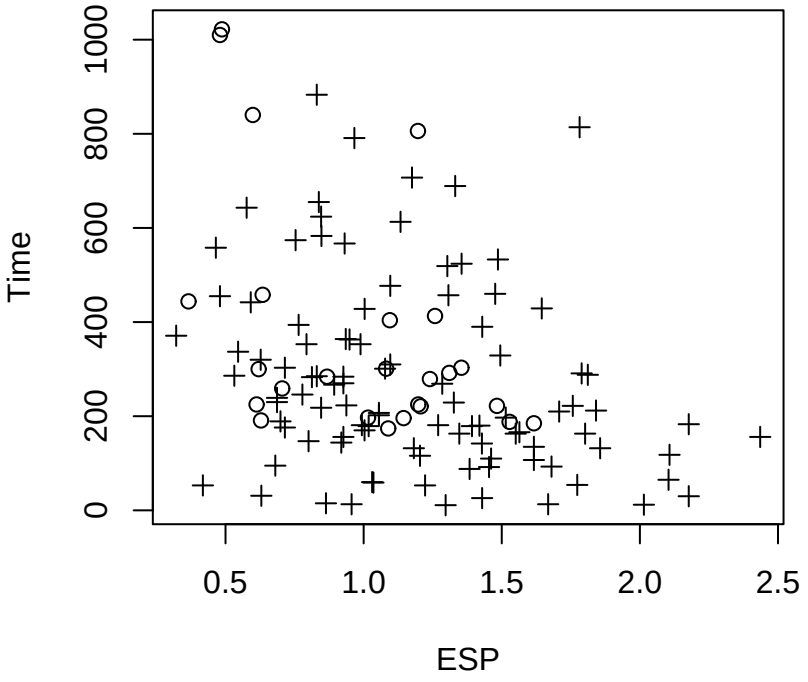

Figure 7: Censored Response Plot for R Lung Cancer Data

identity line. For the Exponential regression model, $\sigma = 1$ and $\boldsymbol{\beta}_C = -\boldsymbol{\beta}_A$, and the Exponential EE plot is a plot of

$$ESPE = -\hat{\boldsymbol{\beta}}_A^T \boldsymbol{x} \text{ versus } \text{ESPC} = \hat{\boldsymbol{\beta}}_{\text{C}}^{\text{T}} \boldsymbol{x}.$$

For the Weibull regression model, $\boldsymbol{\beta}_C = -\boldsymbol{\beta}_A/\sigma$, and the Weibull EE plot is a plot of

$$ESPW = \frac{-1}{\hat{\sigma}} \hat{\boldsymbol{\beta}}_A^T \boldsymbol{x} \text{ versus } \text{ESPC} = \hat{\boldsymbol{\beta}}_{\text{C}}^{\text{T}} \boldsymbol{x}.$$

Example 7. *R* contains a data set *lung* where the response variable Y is the time until death for patients with lung cancer. See MathSoft (1999, p. 268). Consider the data set for males with predictors *ph.ecog* = Ecog performance score 0-4, *ph.karno* = a competitor to ph.ecog, *pat.karno* = patient's assessment of their karno score and *wt.loss* = weight loss in last 6 months. Figure 7 shows the censored response plot. Notice that the survival times decrease rapidly as the ESP increases and that there is one time that is unusually large for ESP ≈ 1.8. Figure 8 shows the slice survival plots. The ESP was divided into 4 groups and correspond to the upper left, upper right, lower right and lower left corners of the plot where $\hat{S}(400) \approx (0.70, 0.60, 0.55, 0.30)$. The circles corresponding to the Kaplan Meier estimator are "close" to the Cox survival curves in that the circles do not fall very far outside the pointwise CI bands.

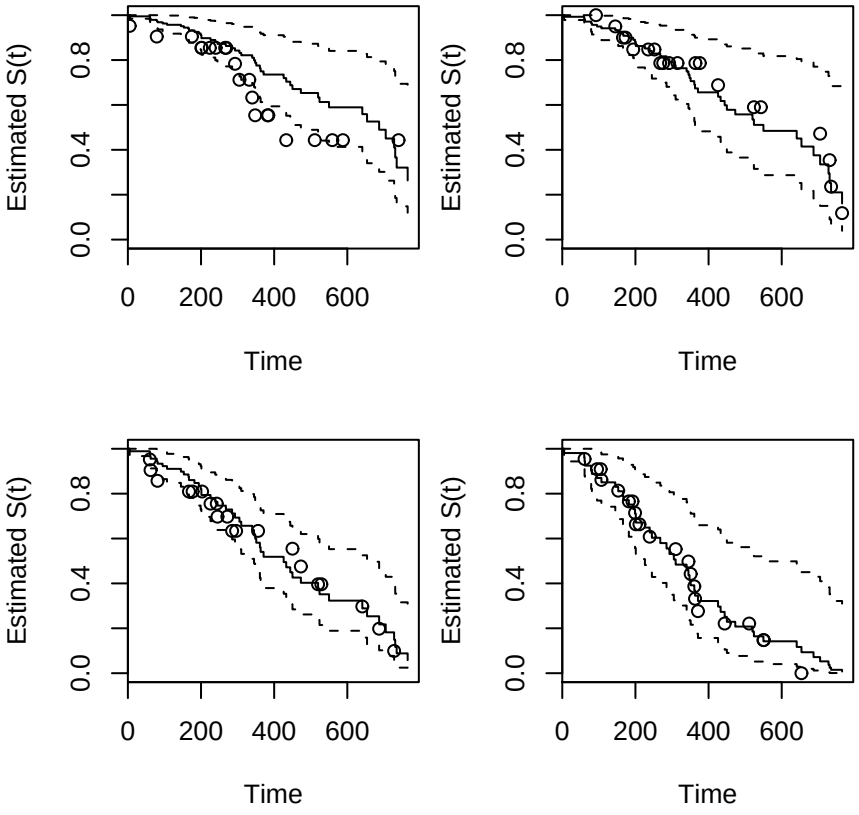

Figure 8: Slice Survival Plots for R Lung Cancer Data

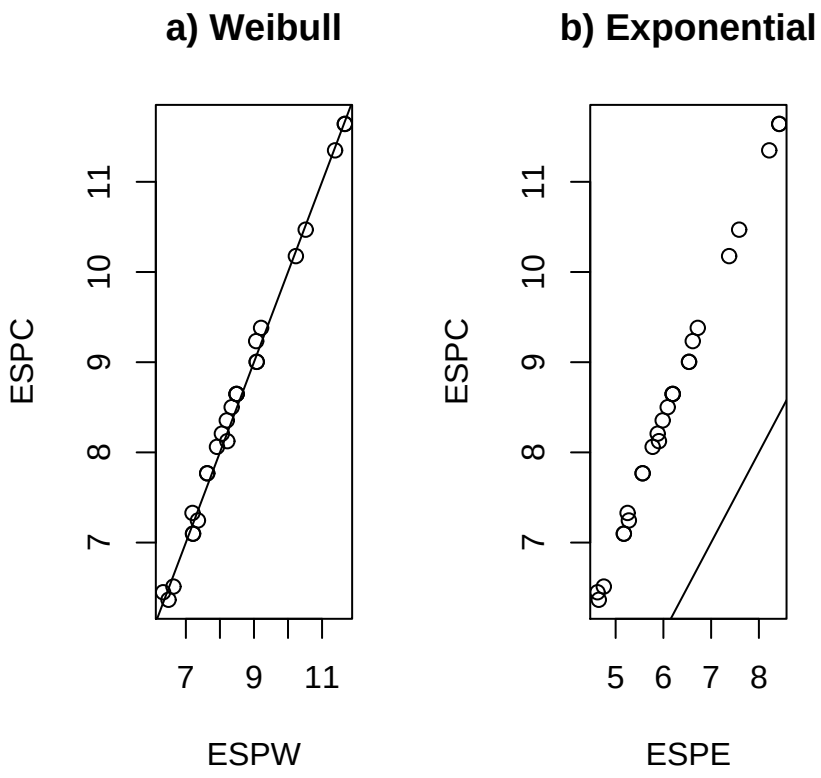

Figure 9: EE Plots for Ovarian Cancer Data

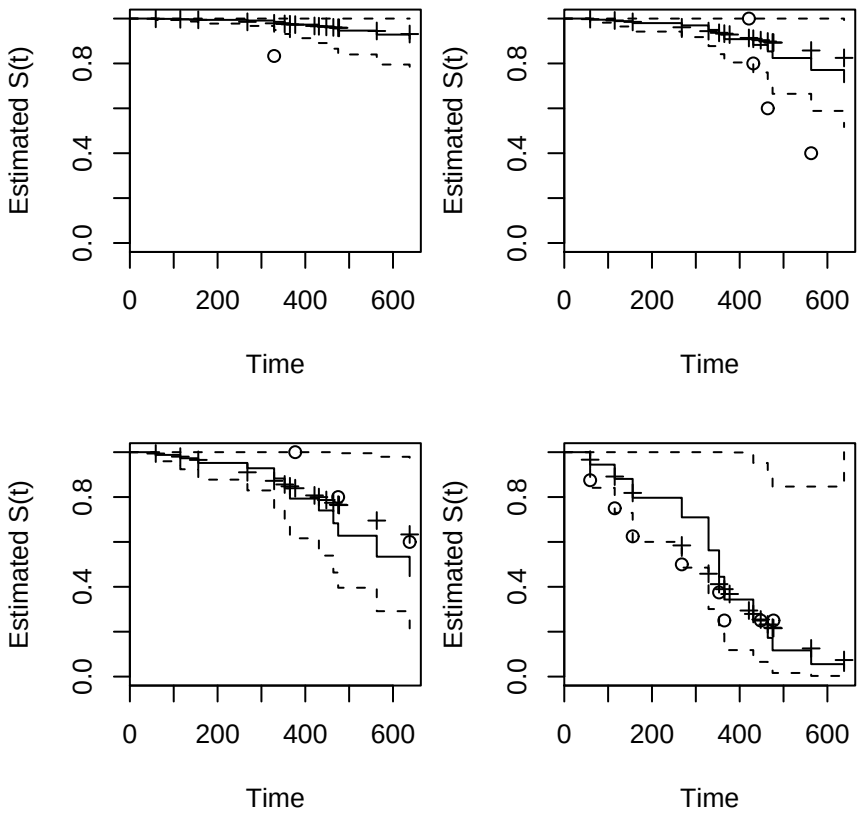

Figure 10: Slice Survival Plots for Ovarian Cancer Data

Example 8. The ovarian cancer data is from Collett (2003, pp. 187-190). The response variable is the survival time of $n = 26$ patients in days with predictors *age* in years and *treat* (1 for cyclophosphamide alone and 2 for cyclophosphamide combined with adriamycin). Figure 9 shows the Weibull and Exponential regression EE plots. Notice that the estimated risk scores from the Cox regression and Weibull regression are nearly the same with correlation = 0.997. The points from the Exponential regression do not cluster about the identity line. Hence Exponential regression should not be used. Figure 10 gives the slice survival plot for the Cox model with the Weibull survival function $\hat{S}_{\boldsymbol{x}}(t) = \exp[-\exp(-\hat{\gamma}\hat{\boldsymbol{\beta}}_A^T\boldsymbol{x})\ \exp(-\hat{\gamma}\hat{\alpha})\ t^{\hat{\gamma}}\ ]$ represented by crosses where $\hat{\gamma} = 1/\hat{\sigma}$. Notice that the Weibull and Cox estimated survival functions are close and thus similar. Again the circles corresponding to the Kaplan Meier estimator are "close" to the Cox survival curves in that the circles do not fall very far outside the pointwise CI bands.

Example 9. *R* contains a data set *nwtco* where the response variable Y is the time until relapse with $n = 4028$. The model used predictors *histol* = tumor histology from central lab, *instit* = tumor histology from local institution, *age* in months, and *stage* of disease from 1 to 4 (treated as an continuous variable). Figure 11 shows the Grambsch and Therneau (1994) plots which look fairly flat, but with such a large sample, all slopes are significantly different from zero, and the global test has p-value $\approx 5.66 \times 10^{-11}$. The slice survival plot in Figure 12 shows that the Cox survival estimators and Kaplan

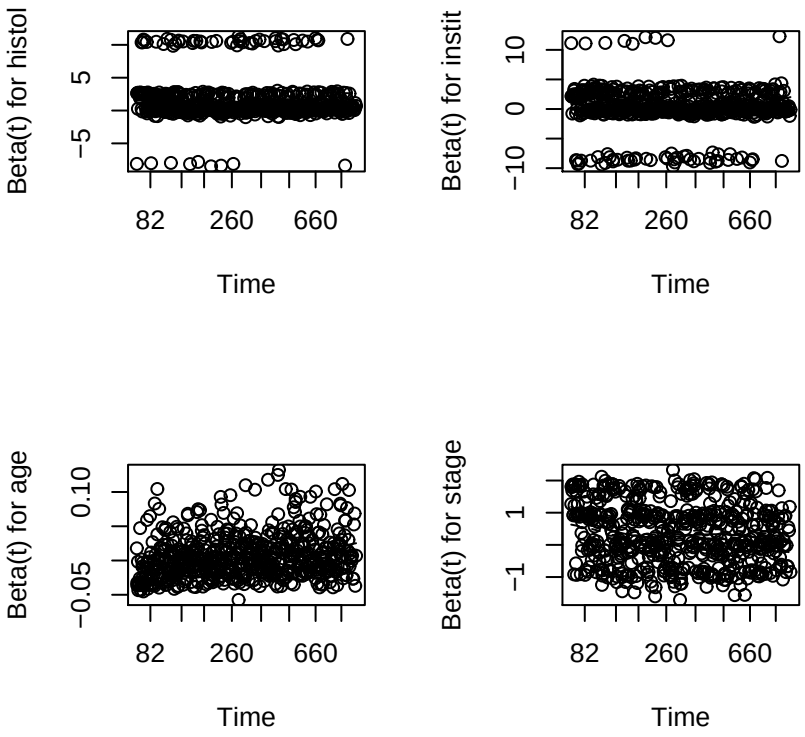

Figure 11: Grambsch and Therneau Plots for NWTCO Data

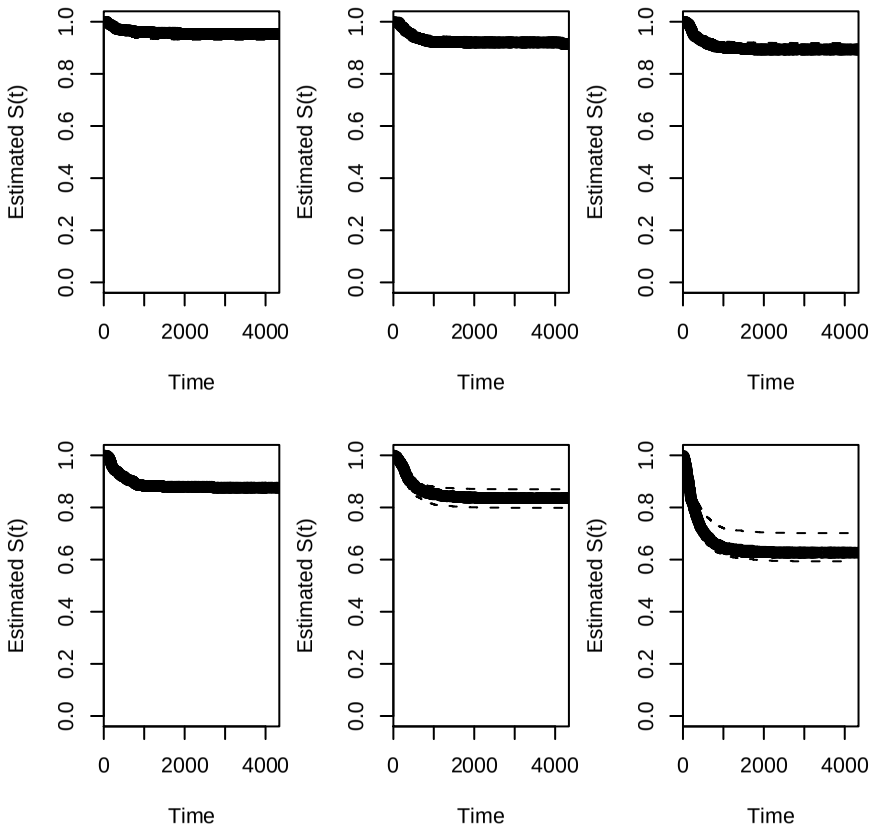

Figure 12: Slice Survival Plot for NWTCO Data: Horizontal Axis is the Estimated Survival Function S(t)

Meier estimators are nearly identical in the six slices, suggesting that the Cox model is a reasonable approximation to the data.

References

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

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

Anscombe, F.J. (1961), "Examination of Residuals," in *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability*, ed. J. Neyman, University of California Press, Berkeley, CA, 1-31.

Anscombe, F.J., and Tukey, J.W. (1963), "The Examination and Analysis of Residuals," *Technometrics,* 5, 141-160.

Breslow, N.E. (1974), "Covariance Analysis of Censored Survival Data," *Biometrics*, 30, 89-100.

Brillinger, D.R. (1983), "A Generalized Linear Model with "Gaussian" Regressor Variables," in *A Festschrift for Erich L. Lehmann,* eds. Bickel, P.J., Doksum, K.A., and Hodges, J.L., Wadsworth, Pacific Grove, CA, 97-114.

Chambers, J.M., Cleveland, W.S., Kleiner, B., and Tukey, P. (1983), *Graphical Methods for Data Analysis,* Duxbury Press, Boston, MA.

Chang, J., and Olive, D.J. (2010), "OLS for 1D Regression Models," *Communications in Statistics: Theory and Methods*, to appear.

Chen, Y.Q., and Jewell, N.P. (2001), "On a General Class of Semiparametric Hazards Regression Models," *Biometrika*, 88, 687-702.

Collett, D. (2003), *Modelling Survival Data in Medical Research*, 2nd ed., Chapman & Hall/CRC, Boca Raton, FL.

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

Cook, R.D., and Olive, D.J. (2001), "A Note on Visualizing Response Transformations in Regression," *Technometrics,* 43, 443-449.

Cook, R.D., and Weisberg, S. (1997), "Graphs for Assessing the Adequacy of Regression Models," *Journal of the American Statistical Association,* 92, 490-499.

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

Cox, D.R. (1972), "Regression Models and Life-Tables," *Journal of the Royal Statistical Society, B,* 34, 187-220.

Draper, N.R., and Smith, H. (1966), *Applied Regression Analysis,* Wiley, New York, NY.

Freedman, D. A. (2005), *Statistical Models Theory and Practice,* New York, NY: Cambridge University Press.

Gentleman, R., and Crowley, J. (1991), "Graphical Methods for Censored Data," *Journal of the American Statistical Association,* 86, 678-683.

Grambsch, P.M., and Therneau, T.M. (1994), "Proportional Hazards Tests and Diagnostics Based on Weighted Residuals," *Biometrika*, 81, 515-526.

Hosmer, D.W., and Lemeshow, S. (1980), "A Goodness of Fit Test for the Multiple Logistic Regression Model," *Communications in Statistics,* A10, 1043-1069.

Hosmer, D.W., and Lemeshow, S. (1999), *Applied Survival Analysis: Regression Modeling of Time to Event Data*, Wiley, NY.

Houseman, E. A., Ryan, L. M., and Coull, B. A. (2004), "Cholesky Residuals for Assessing Normal Errors in a Linear Model with Correlated Errors," *Journal of the American Statistical Association*, 99, 383-394.

MathSoft (1999), *S-Plus 2000 Guide to Statistics,* Volume 2, Data Analysis Products Division, MathSoft, Seattle, WA.

May, S., and Hosmer, D.W. (1998), "A Simple Method for Calculating a Goodness-of-Fit Test for the Proportional Hazards Model," *Lifetime Data Analysis*, 4, 109-120.

Miller, R. (1981), *Survival Analysis*, Wiley, New York, NY.

Montgomery, D.C., Peck, E.A., and Vining, G. (2006), *Introduction to Linear Regression Analysis,* 4th ed., Wiley, Hoboken, NJ.

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

Olive, D.J. (2004), "Visualizing 1D Regression," in *Theory and Applications of Recent Robust Methods*, eds. Hubert, M., Pison, G., Struyf, A., and Van Aelst S., Series: Statistics for Industry and Technology, Birkhäuser, Basel, Switzerland, 221-233.

Olive, D.J. (2008), *Applied Robust Statistics*, Preprint, see (www.math.siu.edu/olive/).

Olive, D.J. (2009a), *Plots for Survival Regression*, Preprint, see (www.math.siu.edu/olive/).

Olive, D.J. (2009b), "Plots for Binomial and Poisson Regression," Unpublished Manuscript available from (www.math.siu.edu/olive/ppgfit.pdf).

Olive, D.J. (2010), *Multiple Linear and 1D Regression Models*, Unpublished Online Text available from (www.math.siu.edu/olive/regbk.htm).

Olive, D.J., and Hawkins, D.M. (2005), "Variable Selection for 1D Regression Models," *Technometrics*, 47, 43-50.

Olive, D.J., and Hawkins, D.M. (2010), "Response Plots for Linear Models," Preprint, see (http://www.math.siu.edu/olive/preprints.htm).

Schaaffhausen, H. (1878), "Die Anthropologische Sammlung Des Anatomischen Der Universitat Bonn," *Archiv fur Anthropologie,* 10, 1-65, Appendix.

Sheather, S.J. (2009), *A Modern Approach to Regression with R,* Springer, New York, NY.

SAS Institute, (1985), *SAS User's Guide: Statistics*, Version 5, Cary, NC: SAS Institute.

Tremearne, A.J.N. (1911), "Notes on Some Nigerian Tribal Marks," *Journal of the Royal Anthropological Institute of Great Britain and Ireland,* 41, 162-178.

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