6. INFLUENTIAL CASES IN GENERALIZED LINEAR MODELS

The generalized linear model (GLM) extends from the general linear model to accommodate dependent variables that are not normally distributed, including those that are not continuous. This chapter starts with a brief description of the GLM. It then provides a brief discussion of diagnostic methods for detecting unusual cases in the GLM. It ends with an introduction to robust generalized linear models, providing empirical examples for logistic and Poisson regression models.

The Generalized Linear Model

I provide only a basic description of the GLM, emphasizing information that is necessary to understand robust generalized linear models. For a more extensive and detailed description of GLMs, see McCullagh and Nelder's (1989) classic book on the topic (see also Dobson 1990; Fahrmeir and Tutz 2001; and Lindsey 1997 for good general treatments of the GLM). For other discussions of the GLM geared toward social scientists, there are three books in the present series (Gill 2001; Dunteman and Ho 2005; Liao 1994).

Recall that the linear model is written as

$$y_i = \sum_{j=1}^{k} x_{ij}\beta_j + \varepsilon_i, \tag{6.1}$$

where y is assumed to be linearly related to the xs, and the errors are assumed to be uncorrelated, have constant variance, and be normally distributed. In other words, the linear model represents the conditional mean of y given the xs as

$$\mu_i = \sum_{j=1}^{k} x_{ij}\beta_j. \tag{6.2}$$

The generalized linear model loosens these assumptions to predict the conditional mean of a dependent variable with any exponential distribution, taking the following general form

$$f(y_i;\theta_i;\varphi) = \exp\left[\frac{y\theta - b(\theta)}{a(\varphi)} + c(y,\varphi)\right], \tag{6.3}$$

where θ is the *canonical parameter* that represents the estimate of location, and φ is the *dispersion parameter* that represents the scale. In other words, the GLM allows the distribution of y to take the shape of many different exponential families:

$$y_i \left| xs \sim \begin{cases} \text{Gaussian} \\ \text{Binomial} \\ \text{Poisson} \\ \text{gamma} \\ \text{etc.} \end{cases} \right.$$

The exponential family is defined by the a, b, and c functions in Equation 6.3.

The assumption of linearity remains for the GLM but it is with respect to a *linear predictor* η rather than to y itself

$$\eta_i = \sum_{j=1}^{k} x_{ij}\beta_j. \qquad [6.4]$$

In other words, the canonical parameter θ in Equation 6.3 depends on the *linear predictor*. More specifically, the conditional mean μ_i of the dependent variable is linked to this linear predictor through a transformation, called the *link function* $g(.)$:

$$g(\mu_i) = \eta_i \qquad [6.5]$$

The link function must be monotonic and differentiable, and take any value (positive or negative) that ensures the linear dependence of η on the explanatory variables. An OLS regression is fitted when the identity link and the Gaussian family are specified. Any other link function results in a nonlinear relationship between the expectation of the dependent variable y_i and the independent variables x_{ij}. Table 6.1 displays some important families included in the GLM framework and some associated link functions.

Maximum likelihood estimates for GLMs are found by regarding Equation 6.3 as a function of the parameters $\boldsymbol{\beta}$. Typically, this means maximizing the log-likelihood function with respect to $\boldsymbol{\beta}$:

$$\begin{aligned} l(\beta) &= \log L(\boldsymbol{\beta}) = \log \prod_{i=1}^{n} f(y_i; \mu_i) \\ &= \log \prod_{i=1}^{n} f(y_i; \mathbf{x}_i, \boldsymbol{\beta}) = \sum_{i=1}^{n} \log f(y_i; \mathbf{x}_i, \boldsymbol{\beta}) \end{aligned} \qquad [6.6]$$

Maximum likelihood estimates can be obtained using the Newton-Raphson method or iteratively reweighted least squares (see Nelder and Wedderburn 1972; McCullagh and Nelder 1989). For IRLS estimation of GLMs, the

TABLE 6.1
Important Exponential Families and Their Link Functions

Distribution	*Range of μ*	*Link Function, (g)*	
Normal	$(-\infty, +\infty)$	Identity link	$g(\mu) = \mu$
Binomial	$(0, 1)$	Logit link	$g(\mu) = \log[\mu/(1-\mu)]$
	$(0, 1)$	Probit link	$g(\mu) = \Phi^{-1}(\mu)$
Poisson	$(0, \infty)$	Log link	$g(\mu) = \log(\mu)$
Gamma	$(0, \infty)$	Reciprocal link	$g(\mu) = \mu^{-1}$
	$(0, \infty)$	Log link	$g(\mu) = \log(\mu)$

dependent variable is not y itself, but the *adjusted dependent variable z*, which is a linearized form of the link function applied to y. We start by defining the linear predictor for the first iteration

$$\underset{(n\times 1)}{\hat{\eta}^{(0)}} = \underset{(n\times p)}{X^T} \underset{(p\times 1)}{\beta^{(0)}} \qquad [6.7]$$

with initial fitted values of $\hat{\mu}^{(0)}$ resulting from $g^{-1}(\hat{\eta}^{(0)})$. We then define z as

$$z^{(0)} = \hat{\eta}^{(0)} + \left(\left.\frac{\partial \eta}{\partial \mu}\right|_{\hat{\mu}^{(0)}} \right) \left(y - \hat{\mu}^{(0)} \right). \qquad [6.8]$$

The *quadratic weight matrix* to be used in the IRLS is defined by

$$W_{(0)}^{-1} = \left(\left.\frac{\partial \eta}{\partial \mu}\right|_{\hat{\mu}^{(0)}} \right)^2 V(\mu)|_{\hat{\mu}^{(0)}}, \qquad [6.9]$$

where $V(\mu)$ is the variance function defined at $\hat{\mu}^{(0)}$. Both z and $W_{(0)}$ depend on the current fitted value, and thus an iterative process is needed to find a solution. We first regress $z^{(0)}$ on the xs with weight $W_{(0)}$ to find new estimates of the regression coefficients $\hat{\beta}^{(1)}$, and from these a new estimate of the linear predictor. Using the new estimates of z and W, the estimation process is continually repeated until convergence, resulting in normal equations of the general form

$$\hat{\boldsymbol{\beta}} = \left(\mathbf{X}^T \mathbf{W} \mathbf{X} \right)^{-1} \mathbf{X}^T \mathbf{W} \mathbf{z}, \qquad [6.10]$$

where **z** represents the adjusted dependent variable transformed by the link function and **W** is the final weight matrix. GLMs are further extended by *quasi-likelihood estimation*, which, along with the usual specification of the link function, allows specification of the dispersion parameter φ instead of the entire distribution of y (see Wedderburn 1974 for more details).

The deviance of the model parallels the residual sum of squares for least squares regression in that it compares the model under investigation with the saturated model $\boldsymbol{\beta}_S$ for the data. A saturated model with n coefficients for the n observations matches the data exactly, meaning that it achieves the highest possible likelihood. The likelihood of this saturated model provides a baseline to which the likelihood of a less than saturated model can be compared. The deviance measures the discrepancy in fit between the two models. More specifically, it is twice the difference between the log likelihood of the saturated model and the log likelihood achieved by the model under investigation

$$\begin{aligned} D(\boldsymbol{\beta}, \mathbf{y}) &= 2[\log L(\boldsymbol{\beta}_S)] - 2[\log L(\boldsymbol{\beta})] \\ &= -2[\log L(\boldsymbol{\beta})]. \end{aligned} \qquad [6.11]$$

The deviance plays an important role in assessing the fit of the model and in statistical tests for parameters in the model, and also provides one method for calculating residuals that can be used for detecting outliers.

Detecting Unusual Cases in Generalized Linear Models

As for OLS regression, unusual cases can distort estimates for GLMs. For some models, such as the binary logit and probit models, the impact of unusual cases is usually less severe because the dependent variable has only two possible values, but it is still possible for such observations to affect the regression estimates. For other models, like Poisson regression, highly unusual values of the dependent variable are more likely. It is important, then, to explore for outliers in GLMs. Many diagnostic tools for OLS regression have been adapted for the GLM, and those for assessing unusual observations are quite effective.

Residuals From the GLM

Residuals from GLMs can be defined in several ways. Some of these include the response residuals, which are simply the difference between the observed value of y and its fitted value, $y_i - \hat{\mu}_i$; the *deviance residuals*, which are derived from the case-wise components of the deviance of the model; and the *working residuals*, which are the residuals from the final iteration of weighted least squares. There are also approximations of the studentized residuals. This book is most concerned with *Pearson residuals* because they play a central role in many robust GLM models. Pearson residuals are simply the response residuals scaled by the standard deviation of the expected value:

$$e_{\text{Pearson}_i} = \frac{y_i - \hat{\mu}_i}{\sqrt{V(\hat{\mu})}}. \qquad [6.12]$$

For more details of the relative merits of the various types of residuals, see Gill (2001). Each has its uses, and none of them is best for all purposes.

Hat Values and Leverage

As with OLS regression, leverage in the GLM is assessed by the *hat values* h_i, which are taken from the final IWLS fit. Unlike in linear regression, however, the hat values for GLMs depend on the values of y *and* the values of x. Following from Pregibon (1981), the hat matrix is defined by

$$\mathbf{H} = \mathbf{W}^{1/2}\mathbf{X}\left(\mathbf{X}^T\mathbf{W}\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{W}^{\mathbf{1/2}}, \qquad [6.13]$$

where $\mathbf{W}$ is the weight matrix from the final iteration of the IWLS fit. This differs from the general form of $\mathbf{H}$ (Equation 3.7) by replacing $\mathbf{X}$ with $\mathbf{W}^{1/2}\mathbf{X}$. Doing so allows for a change in the variance of $\mathbf{y}$, and thus the hat values depend on both $\mathbf{y}$ and $\mathbf{X}$ (see McCullagh and Nelder 1989:405).

Assessing Influence

Following the linear model, DFBETAs and Cook's distances are helpful for detecting influence in GLMs. DFBETAs are calculated by finding the difference in an estimate before and after a particular observation is removed, $D_{ij} = \hat{\beta}_j - \hat{\beta}_{j(-i)}$, for $i = 1, \ldots, n$ and $j = 0, 1, \ldots, k$. An approximation of Cook's D measure of influence is also available:

$$D_i = \frac{e^2_{\text{Pearson}_i}}{\hat{\varphi}(k+1)} \times \frac{h_i}{1 - h_i}, \qquad [6.14]$$

where $\hat{\varphi}$ is the estimated dispersion of the model and k is the number of parameters being estimated excluding the constant (see Fox 2002).

Robust Generalized Linear Models

Methods for robust estimation of GLMs have developed much more slowly than robust methods for linear regression. Although there were several early attempts to make logistic regression more robust (e.g., Pregibon 1981; Copas 1988; Carroll and Pederson 1993; Bianco and Yohai 1996), the extension to other GLMs was seldom considered. Still today there are very few statistical programs that have routines for robust GLMs, and those that do are usually limited to the logit and Poisson model.

M-Estimation for GLMs

As with the linear model, the most widely used robust methods for the GLM are based in some way on *M*-estimation. Like early *M*-estimators for linear regression, many early attempts at *M*-estimation for GLMs suffered from an unbounded influence function (see Stefanski, Carroll, and Ruppert 1986; Kunsch, Stefanski, and Carroll 1989). Often, the resulting estimators were also undesirable because they were Fisher inconsistent.[1] In recent years, however, consistent bounded influence methods based on quasi-likelihood estimation have developed. One of these methods is due to Cantoni and Ronchetti (2001).[2]

Cantoni and Ronchetti's estimator evolved from the quasi-likelihood generalized estimating equations of Preisser and Qaqish (1999):

$$\sum_{i=1}^{n} \frac{\partial}{\partial \boldsymbol{\beta}} Q(y_i;\mu_i) = \sum_{i=1}^{n} \frac{(y_i-\mu_i)}{V(\mu_i)}\mu_i' = 0, \qquad [6.15]$$

where $\mu_i' = \frac{\partial}{\partial \boldsymbol{\beta}}\mu_i$ and $Q(y_i;\mu_i)$ is the quasi-likelihood function. The solution is an *M*-estimator defined by the score function

$$\Psi(y_i;\mu_i) = \frac{(y_i-\mu_i)}{v(\mu_i)}\mu_i'. \qquad [6.16]$$

Unfortunately, this estimator is limited for robust regression because its influence is proportional to Ψ, and thus unbounded.

Cantoni and Ronchetti follow the logic of Mallow's *GM*-estimates for regression (see Chapter 4) to improve Equation 6.16. Recall that the general *M*-estimator is the solution to

$$\sum_{i=1}^{n} \Psi(y;\theta) = 0, \qquad [6.17]$$

or, in the specific case of the generalized linear model,

$$\sum_{i=1}^{n} \Psi(y;\mu) = 0, \qquad [6.18]$$

where Ψ gives weight to the observations. As for *MM*-estimation for linear regression, if is odd and bounded, meaning $\rho(\cdot)$ is symmetric around 0, the breakdown point of the estimator is $BDP = 0.5$. Cantoni and Ronchetti accomplish this by solving

$$\Psi(y;\mu) = v(y;\mu)w(\mathbf{x})\mu' - a(\boldsymbol{\beta}), \qquad [6.19]$$

where

$$a(\boldsymbol{\beta}) = \frac{1}{2}\sum_{i=1}^{n} E[v(y_i;\mu_i)]w(\mathbf{x}_i)\mu_i' \qquad [6.20]$$

and the v_i and w_i are weight functions that consider the residuals and the hat values of the observations respectively. An adaptation of the Huber function ensures that the weights are robust to unusual y values

$$v_i(y_i;\mu_i) = \Psi(e_i)\frac{1}{V^{1/2}(\mu_i)}. \qquad [6.21]$$

Following the original Mallows *GM*-estimator for linear models, a possible choice for $w_i(\mathbf{x}_i)$ is $w_i(x_i) = \sqrt{1-h_i}$. As we have already seen, however, this produces a low breakdown point, so the inverse of the robust distances is employed instead (recall the discussions in Chapter 4 on robust distances). The end result is an estimator that is efficient, has bounded influence, and is asymptotically normal. More important, it has been shown that inference from this model is much more reliable than from ordinary GLMs when the data are contaminated (see Cantoni and Ronchetti 2001).

Example 6.1: Logistic Regression Predicting Vote for the Labour Party in Britain, 2001

This example uses data from the 1997–2001 British Election Panel Study (Heath, Jowell, and Curtice 2002). We concentrate only on respondents who participated in the final wave in 2001. After missing data are removed, the analytical sample size is 1,421. The goal is to assess the impact of the leader of the Labour Party, Tony Blair, on vote for his party during the 2001 British election. The dependent variable is vote for the Labour Party (coded 1) versus otherwise (coded 0). Evaluations of Blair were tapped with a five-point Likert item asking respondents how well they thought Blair was doing as prime minister (high values indicated a good job). The analysis also controls for age; gender; education (degree, some postsecondary, a-level, o-level, and none); social class (managers/professionals, routine nonmanual, self-employed, manual working class); retrospective sociotropic economic perceptions (a five-point scale, with high values indicating that the respondent felt the economy improved in the past year); and retrospective egocentric economic perceptions (a five-point scale, with high values indicating that the respondent felt his or her own personal economic position improved in the past year).[3] Both a regular logistic regression and a robust regression are fitted to the data.

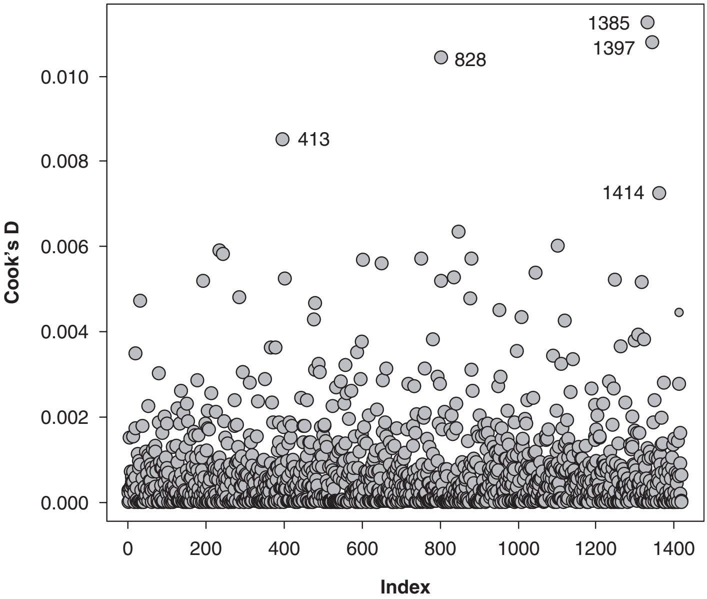

Figure 6.1 Index Plot of Cook's Ds for Logistic Regression Predicting Labour Vote in Britain, 2001

We start by assessing influence in the regular logit model. As we see from the index plot of Cook's Ds in Figure 6.1, a handful of observations have relatively high influence on the regression surface. Further diagnostics, including close examination of the DFBETA$_i$ for each coefficient in the model, failed to uncover any obvious problems, however. In other words, although some cases have unusually high influence overall, they do not appear to substantially influence any of the coefficients, at least not individually. Still, given the high overall influence of these cases, we explore whether or not a robust logistic regression tells a different story than the regular logistic regression.

Table 6.2 shows the results of the two regressions. Despite the presence of some observations with relatively high influence, the regular logistic regression has performed quite well. In fact, substantive conclusions are similar regardless for the two models—we would conclude that appraisals of Blair had a profound effect on whether or not someone voted for the Labour Party. Although the coefficient for the impact of appraisals of Tony Blair is slightly larger for the robust logistic regression (1.205 versus 1.127), the difference between the two coefficients is not statistically significant. The regular

TABLE 6.2
Logistic Regression Models Predicting Labour Vote in Britain, 2001

	Maximum Likelihood Logit Model		*Robust Logit Model*	
	$\hat{\beta}$	SE $\hat{\beta}$	$\hat{\beta}$	SE $\hat{\beta}$
Intercept	–5.15	0.468	–5.42	0.525
Age	–0.003	0.004	–0.003	0.004
Male	0.117	0.141	0.129	0.144
Education				
Degree	–0.372	0.269	–0.350	0.276
A-level	–0.391	0.255	–0.321	0.261
O-level	–0.190	0.181	0.163	0.187
Some post-sec	–0.462	0.235	–0.374	0.241
None	0	—	0	—
Social Class				
Professionals/managers	–0.055	0.184	–0.040	0.189
Routine nonmanual	–0.271	0.193	–0.213	0.197
Self-employed	–0.548	0.259	–0.551	0.266
Manual working class	0	—	0	—
Economic Perceptions				
Retrospective sociotropic	0.496	0.087	0.476	0.090
Retrospective egocentric	0.268	0.077	0.266	0.079
Opinions of Tony Blair	1.127	0.101	1.205	0.122
n	1,421		1,421	

logistic regression should be preferred for these data, then, because of its simplicity relative to the robust regression. This example is typical in that it is difficult for unusual observations to exert strong influence on the regression surface in logistic regression because the dependent variable can take on only two values. As we shall see below, however, unusual cases are more likely to exert high influence in Poisson regression.

Example 6.2: Robust Poisson Regression Predicting Voluntary Association Membership in Quebec

This example uses data from the Canadian Equality, Security, and Community Survey of 2000. Although the data set contains information on respondents from across Canada, only Quebec respondents are included in the analysis ($n = 949$). The dependent variable is the number of voluntary associations to which respondents belonged. The independent variables are gender (with women as the reference category), Canadian born (the

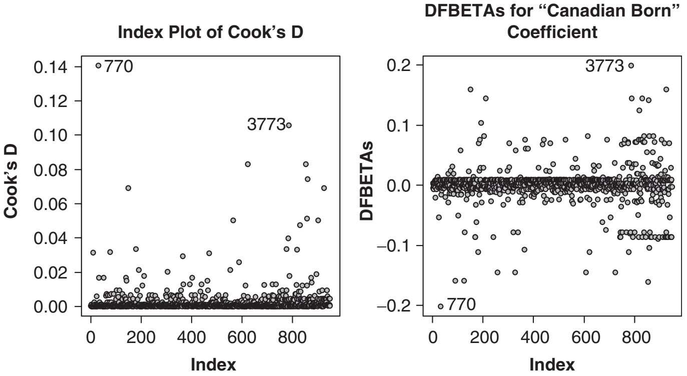

Figure 6.2 Diagnostic Plots for a Poisson Regression Model Predicting Voluntary Association Involvement in Quebec

reference category is ‘‘not born in Canada’’), and language spoken in the home (divided into English, French, and other, with French coded as the reference category). Given that the dependent variable is a count variable (and follows a Poisson distribution), Poisson regression models are employed. Both a regular generalized linear model using maximum likelihood and a robust GLM using quasi-likelihood are fitted. Before discussing the results, we turn to diagnostic plots for the OLS regression.

Although extensive diagnostics were carried out, only those that uncovered potentially problematic observations are reported. In this respect, Figure 6.2 displays index plots for Cook's distances and the DFBETA$_i$ for the ‘‘Canadian born’’ coefficient. The Cook's distances indicate that there are perhaps 10 observations with fairly large influence on the regression, two of which may be particularly problematic (observations 770 and 3773). Analysis of the DFBETA$_i$ indicates that the influence of these two cases is largely with respect to the effect of Canadian born, although as the plot indicates, their influences are in opposite directions.

Table 6.3 displays the results from the regular Poisson regression and the robust Poisson regression. We see clearly that the coefficient for Canadian born for the regular GLM was affected by unusual observations that did not fit with the bulk of the data. The coefficient for the robust regression model is nearly 10 times as large as the regular GLM coefficient. The difference in effect makes for very different substantive interpretations. We would conclude from the regular GLM that, holding the other predictors constant, there is no difference between those born in Canada and those born elsewhere in terms of participation in voluntary associations ($e^{0.027} = 1.03$;

TABLE 6.3
Poisson Regression Models Predicting Voluntary Association Membership

	Maximum Likelihood GLM			*Robust GLM*		
	$\hat{\beta}$	SE $\hat{\beta}$	$e^{\hat{\beta}}$	$\hat{\beta}$	SE $(\hat{\beta})$	$e^{hat\beta}$
Intercept	0.586	0.077	1.79	0.120	0.095	1.13
Men	0.079	0.045	1.08	0.084	0.053	1.09
Canadian born	0.027	0.072	1.03	0.258	0.088	1.29
Language						
English	0.357	0.061	1.43	0.537	0.068	1.71
Other	–0.014	0.094	0.98	0.079	0.112	1.08
French	0	0	1.00	0	0	1.00
n	949			949		

$p = .71$). On the other hand, the robust regression suggests that, on average, those born in Canada belong to 30 percent more associations at fixed values of the other predictors ($e^{0.258} = 1.29$; $p = .0035$).

The examples in this chapter are informative for two reasons. First, the Poisson regression example clearly showed that estimates from GLMs can be drastically altered by unusual observations. Conclusions based on the regular GLM were quite different from those based on the robust GLM. In this case, it makes the most sense to report the robust GLM. Second, the logistic regression example showed that even with a handful of observations with relatively high influence, the substantive conclusions from a robust GLM will not necessarily differ from those based on the regular GLM. Because the dependent variable can take on only two values—and hence it is usually impossible for the residuals to get extremely large—this is often the case for logistic regression. In these situations, the regular GLM is preferred because of its simplicity relative to the robust GLM. Nevertheless, it is worth exploring the robust GLM if only as a diagnostic tool.

Notes

1. An *M*-estimator is considered conditionally Fisher-consistent if

$$E_\beta[\Psi(y,x,\beta)|x] = \iint \Psi(y,x,\beta)P_\beta(dy|x) = 0 \text{ for all } \beta \text{ and } x.$$

Maximum likelihood estimators for linear and generalized linear models are conditionally Fisher-consistent if the distribution of x does not depend on β.

[2] This is the method employed by the glmrob function in the robustbase library for the statistical package **R**.

[3] For more detailed information on the coding of the variables see Andersen and Evans (2003).

7. CONCLUSIONS

This book emphasizes the importance of detecting and properly handling unusual observations in regression analysis. The empirical examples demonstrate that if such cases go undetected, they could seriously distort the regression estimates. Evidence was also provided to indicate that vertical outliers, and more generally, heavy-tailed distributions can decrease the precision of regression estimates. These problems pertain to models fitted using both OLS and the more general GLM, and thus highlight the importance of diagnostics. Several traditional methods for detecting vertical outliers, leverage points, and influence were explored. Combined, these methods were effective in identifying problematic observations in the empirical examples.

Having identified problematic observations, the researcher can consider several options to accommodate them. The simple "fix" is to remove the offending observations from the analysis. This is a sensible strategy if there are good reasons for doing so, for example, if an observation is miscoded or known to be unique for some particular reason. Sometimes, however, the unusual observations reflect something systematic for which the model was unable to account. This is an important issue that implies that unusual observations are not always synonymous with "bad" data. In fact, the outliers could be the most intriguing part of the data. If there are many unusual observations, we should try to accommodate them by adding terms to the model—either new variables or interactions between existing variables—that account for the discrepancy. If no sound justification for the removal of the unusual observations or changes to the model specification can be determined, robust regression techniques are a suitable option.

On one hand, the strategy of robust regression is not much different from removing the observations. Both strategies have the goal of finding the best fitting model for the bulk of the data. In this respect, some might argue that both strategies result in a form of data truncation bias. In other words, by removing or down-weighting observations when we don't know if they are truly "contaminated" data, we are biasing the regression estimates. I disagree with this argument. We use statistical models to describe patterns in the data. The goal should be to tell the best possible "story" from the data. It doesn't make sense, then, to talk about a relationship between y and x,