Statistica Sinica 1(1991), 93-125

REGRESSION PERCENTILES USING ASYMMETRIC SQUARED ERROR LOSS

B. Efron

Stanford University

Abstract: We consider the problem of estimating *regression percentiles*, for example the 75th conditional percentile of the response variable y given the covariate vector x. Asymmetric Least Squares (ALS) is a variant of ordinary least squares, in which the squared error loss function is given different weight depending on whether the residual is positive or negative. ALS estimates of regression percentiles are easy to compute. They are reasonably efficient under normality conditions. There is an interesting connection between ALS estimates and absolute residual regression for detecting heteroscedasticity. Three examples are given to demonstrate the utility of estimated regression percentiles for understanding regression data, particularly when the covariate x is multi-dimensional.

Key words and phrases: Conditional percentiles, heteroscedasticity, absolute residual regression, regression quantiles.

1. Introduction

The data for a typical regression analysis is a cloud of points in Euclidean space

$$(x_i, y_i), \qquad i = 1, 2, \dots, n, \tag{1.1}$$

where the x_i are $1 \times p$ covariate vectors, and the y_i are scalar responses. The primary output of the usual analysis is an estimated regression function $\hat{\mu}(x)$, which describes the middle of the point cloud, in the y direction, as a function of the covariate x. But what if the statistician is interested in the higher or lower regions of the point cloud, as well as its middle? This paper describes a useful method, borrowed from the econometrics literature, for estimating *regression percentiles*, for example the 25th or 75th conditional percentile of y as a function of x.

Figure 1 displays two simple regressions, "simple" meaning that the covariate vector x_i is a function of a one-dimensional regressand z_i, so that the data can be displayed by a scatterplot of the (z_i, y_i) pairs. The left panel concerns the four divisions of major league baseball in the United States. The regressand z_i is the lead at the half-way point of the baseball season of the first-place team over its nearest runnerup (for each of the four divisions in each of the 18 years 1970–1987, for a total of $n = 72$ cases). The response y_i is that same team's lead at the end of the season, taken negative if that team did not finish in first place. For example the arrow shows a team that was two games ahead of its division at the half-way point, but finished the season 22 games behind the division champion. The line marked "OLS" is the ordinary least squares regression line for the model $y_i = b_0 + b_1 z_i + \text{error}_i$.

The right panel of Figure 1 concerns a clinical trial in which subjects received cholostyramine, a drug believed to lower blood cholesterol levels. The regressand z_i is the compliance of the ith patient, defined as the percentage of prescribed packets of cholostyramine actually taken; response y_i is the observed reduction in

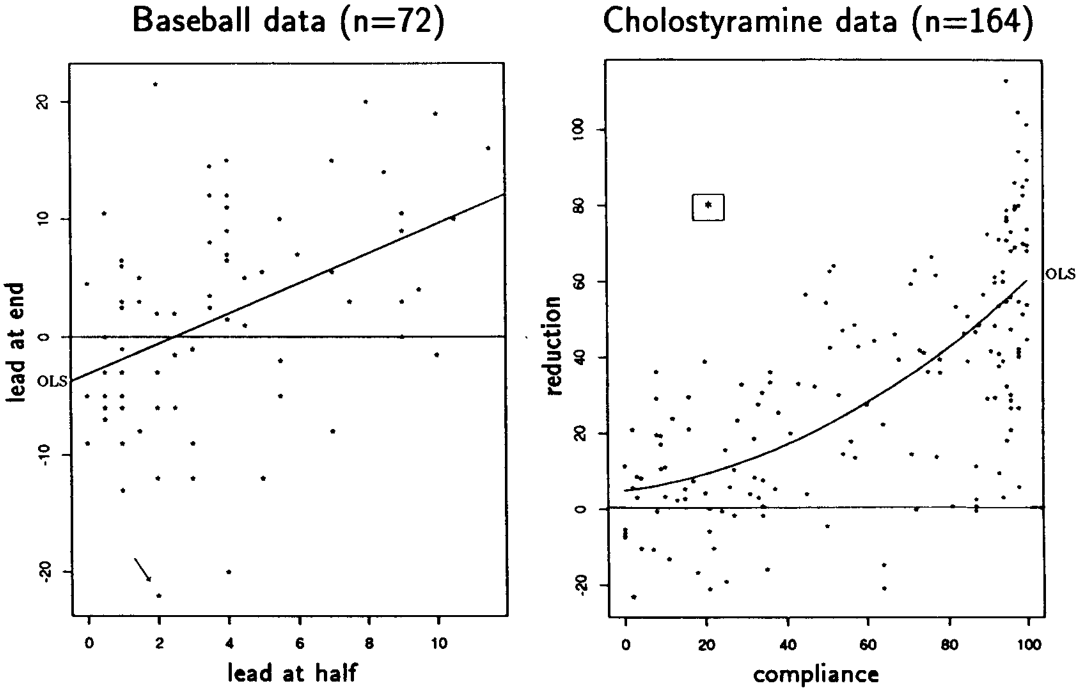

Figure 1. Two simple regression problems. Left panel: regressand z_i is the lead at mid-season of first-place baseball team; response y_i is that same team's lead at the end of the season, taken negative if the team did not finish in first palce; $n = 72$ data points, coming from 18 years of data for each of the four major league divisions. Right panel: regressand z_i is the observed compliance of patients in a clinical trial of cholostyramine, a drug intended to reduce blood cholesterol levels; response y_i is the observed reduction in blood cholesterol; $n = 164$ data points, after removing one outlying patient (boxed). (From the Lipid Research Clinics Primary Prevention Trial, with thanks to Drs. D. Feldman and J. Farquhar for generous help with the data retrieval.)

blood total cholesterol level. The OLS curve is the estimated quadratic regression $y_i = b_0 + b_1 z_i + b_2 z_i^2 + \text{error}_i$. (Only the treatment arm of the Stanford portion of the study appears in Figure 1, $n = 164$ patients after removal of the boxed outlier. See Lipid Research Clinics Program (1984) for a discussion of the entire project and its results. This data appears courtesy of Drs. D. Feldman and J. Farquhar of the Standford Medical School and the Lipid Research Clinics Program.)

Figure 2 shows regression percentiles for the two examples, as estimated by the asymmetric least squares method that is the subject of this paper. The regression percentiles convey more information than the OLS line by itself. For instance the 25th regression percentile for the baseball data is seen to cross the horizontal axis at $z = 6.5$, so if your favorite team is $6\frac{1}{2}$ games ahead at the half-way point of the season it has about 75% chance of winning or tying for the division championship. For 60% compliance in the cholostyramine trial (about average), the central 80% of the response, 10th through 90th percentile, is estimated to be a decrease of between -6 to 54 units in total cholesterol count.

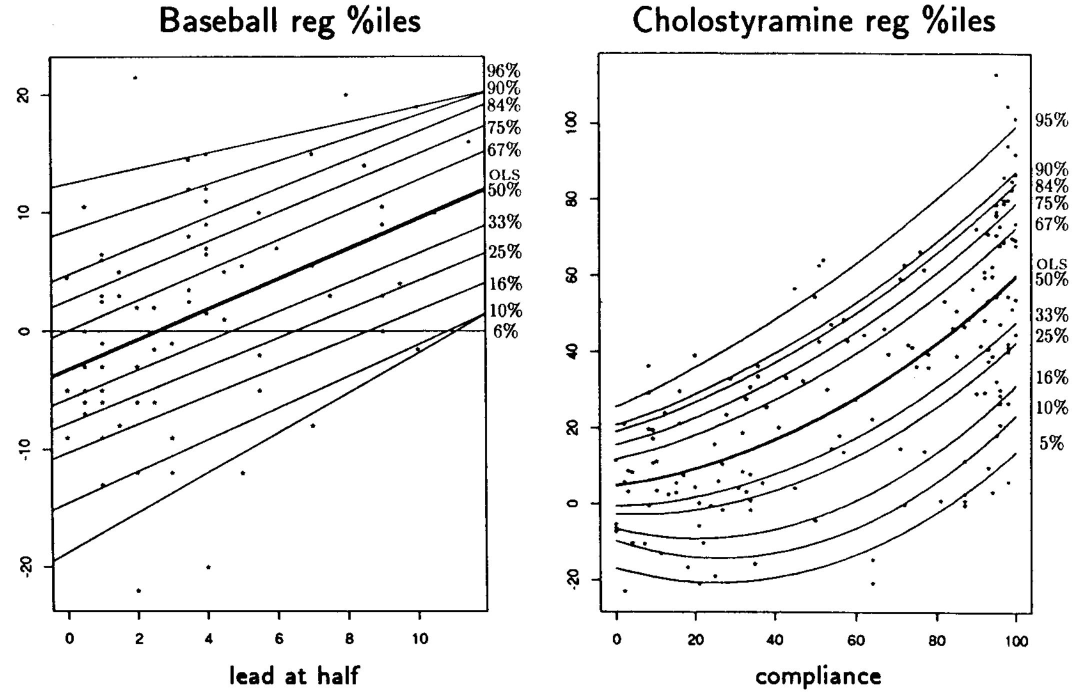

Figure 2. Regression percentiles for the baseball and the cholostyramine examples. Regression precentiles for the examples of Figure 1 are estimated by the method of asymmetric least squares described in Section 2. Marginal numbers indicate the observed percentage of data points below the regression percentile curves. The regression percentiles are nearly parallel to the OLS line in the baseball example, but diverge toward the high end of the compliance scale for the cholostyramine data.

The usual method of assigning regression percentiles goes as follows: assume that the error term in the regression model is homogeneous and normal, error$_i \sim$ N$(0,\sigma^2)$ for $i = 1,2,\dots,n$; estimate σ^2 by $\hat{\sigma}^2$ in the usual way; then estimate the 100αth regression percentile by

$$\hat{\mu}(x) + \hat{\sigma} z^{(\alpha)}, \tag{1.2}$$

where $\hat{\mu}(x)$ is the OLS estimated regression and $z^{(\alpha)}$ is the 100αth percentile of a standard normal distribution, $z^{(\alpha)} = \Phi^{-1}(\alpha)$. The curves (1.2) are parallel to $\hat{\mu}(x)$, displaced by amount $\hat{\sigma} z^{(\alpha)}$.

Method (1.2) agrees reasonably well with the regression percentiles in the baseball example, but not for the cholostyramine data where the regression percentiles diverge noticeably toward the high end of the compliance scale. (The curves are slightly more than twice as far apart at $z = 100$ as at $z = 0$.) Regression percentiles offer an easy way to account for heteroscedasticity in a linear model. A surprising theoretical connection between asymmetric least squares and Glejser's (1969) test for heteroscedasticity using absolute residuals is discussed in Section 3.

Another approach to estimating regression percentiles is to group data along the z-axis, estimate the 100αth percentile value for each group, and connect the values between groups by the use of some smoothing device. This approach will usually be practical only for simple regression, like our two examples, where the data can be displayed as a two-dimensional scattergram.

Asymmetric least squares, like ordinary least squares, is just as easy to apply in general regression situations, where the covariate vectors x_i are *not* functions of a one-dimensional regressand z_i. Section 4 describes an example of this type, which illustrates an interesting point: how a covariate which is important for the OLS curve, that is for describing the middle of point cloud, can have little predictive value at the higher or lower percentiles of the data set.

The basic idea of this paper was invented by Koenker and Bassett (1978, 1982a, 1982b), who used asymmetric absolute loss, rather than asymmetric squared loss, to define what they call *regression quantiles.* Breckling and Chambers (1988) consider asymmetric M-estimators. Newey and Powell (1987), following earlier work by Aigner, Amemiya, and Poirier (1976), use asymmetric squared error as in this paper, calling the resulting curves *expectiles.* (The term "regression percentiles" is intended to apply to all the various forms of this same basic idea, though in what follows it refers mainly to curves calculated by asymmetric least squares.) These authors' results are further discussed in subsequent sections. Their papers are primarily, though not exclusively concerned with using regression percentiles to test and to robustify the usual OLS model, as in

Ruppert and Carroll (1980), going back to Bickel (1973). Here we will be more interested in regression percentiles for their own sake, as useful descriptors of a regression data set. In this spirit, we give a simple and efficient algorithm for calculating *all* the asymmetric least squares regression percentiles for a particular data set.

2. Asymmetric Least Squares Regression

This section concerns the definition and calculation of regression percentiles, such as those in Figure 2, by the method of asymmetric least squares. As Figure 2 indicates, it is useful to compute *all* of the regression percentiles, or at least a wide range of them. Fortunately this turns out to be computationally quite simple to do. Except for computational methods, the discussion here follows that of Newey and Powell (1987), who go into considerably more detail on the formal properties of asymmetric least squares percentiles.

We begin with the data set (x_i, y_i), $i = 1, 2, \ldots, n$, as in (1.1), thought of as a point cloud in $(p+1)$-dimensional Euclidean space $\mathcal{R}^{p+1}$, the x_i being $1 \times p$ covariate vectors and the y_i being scalar responses. All of our calculations relate to the usual linear model

$$y_i = x_i\beta + \epsilon_i, \qquad i = 1, 2, \ldots, n, \tag{2.1}$$

where β is an unknown $p \times 1$ parameter vector, and the ϵ_i are error terms. For convenience we can write (2.1) in matrix form

$$\boldsymbol{y} = \boldsymbol{X}\beta + \epsilon, \tag{2.2}$$

$\boldsymbol{y} = (y_1, \ldots, y_n)'$, $\epsilon = (\epsilon_1, \ldots, \epsilon_n)'$, and $\boldsymbol{X}$ being the $n \times p$ matrix having x_i as ith row, with $\boldsymbol{X}$ assumed to be of full rank p.

A trial value b for the unknown parameter vector β produces a *residual vector* $\boldsymbol{r}(b)$,

$$r_i(b) = y_i - x_i b, \qquad i = 1, 2, \ldots, n. \tag{2.3}$$

A good choice of b is one that makes the residuals small. To quantify this notion in a manner appropriate to regression percentiles define the asymmetric squared error loss function

$$Q_w\{r\} = \begin{cases} r^2 & r \leq 0, \\ wr^2 & r > 0, \end{cases} \tag{2.4}$$

w is a positive constant. A trial value b for β results in total asymmetric squared error loss

$$S_w(b) = \sum_{i=1}^{n} Q_w\{r_i(b)\}. \tag{2.5}$$

By definition the best choice of b for a given value of w is $\hat{\beta}_w$, the minimizer of $S_w(b)$ over b,

$$\hat{\beta}_w = b : \sum_{i=1}^{n} Q_w\{r_i(b)\} = \text{min!}. \tag{2.6}$$

We call $\hat{\beta}_w$ the asymmetric least squares (ALS) estimate of β. Notice that for $w = 1$, $Q_w\{r\} = r^2$ so $\hat{\beta}_1$ is the OLS estimate of β,

$$\hat{\beta}_1 = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}. \tag{2.7}$$

We will present an algorithm for computing $\hat{\beta}_w$ for all values of w ranging from 0 to ∞. The p-dimensional hyperplane $\hat{\mathcal{L}}_w$ defined by $\hat{\beta}_w$ in the $(p+1)$-dimensional (x, y) space,

$$\hat{\mathcal{L}}_w \equiv \{y = x\hat{\beta}_w, x \in \mathcal{R}^p\}, \tag{2.8}$$

called the "w-regression plane", moves smoothly from the bottom to the top of the data point cloud as w increases from 0 to ∞. (See Remark L in Section 6.)

Let $p(w)$ indicate the proportion of data points (x_i, y_i) lying below $\hat{\mathcal{L}}_w$, that is the proportion of points having $r_i(\hat{\beta}_w) \leq 0$. For $p(w) = \alpha$, some given value between 0 and 1, define

$$\hat{\beta}^{(\alpha)} = \hat{\beta}_w. \tag{2.9}$$

In other words, for a given α, $\hat{\beta}^{(\alpha)}$ is the vector $\hat{\beta}_w$ for that value of w having proportion α of the data points lying below $\hat{\mathcal{L}}_w$.

The plane corresponding to $\hat{\beta}^{(\alpha)}$,

$$\hat{\mathcal{L}}^{(\alpha)} = \{y = x\hat{\beta}^{(\alpha)}, x \in \mathcal{R}^p\} \qquad (p(w) = \alpha), \tag{2.10}$$

is by definition the 100αth *regression percentile*. In the simple regression situation where x is a function of a scalar z, the curve $y = x(z)\hat{\beta}^{(\alpha)}$ in the (z, y) space will also be called the 100αth regression percentile, as on the right side of Figure 2.

Notice that there are two distinct parts to the definition of the regression percentiles: (i) the method of asymmetric least squares, (2.6), determines the family of w-regression planes $\hat{\mathcal{L}}_w$, $0 < w < \infty$; (ii) the various planes $\hat{\mathcal{L}}_w$ are labelled $\hat{\mathcal{L}}^{(\alpha)}$ according to the proportion of data points (x_i, y_i) lying below them. In other words, we are using regression methods to estimate conditional percentiles of y given x, for which we may have few or no direct observations $y_i|x$, but are calibrating these estimates by the overall empirical percentiles of the $\hat{\mathcal{L}}_w$, which are based on all n observations (x_i, y_i).

How can we calculate the vector $\hat{\beta}_w$ minimizing $S_w(b) = \sum_{i=1}^{n} Q_w\{r_i(b)\}$? It is easy to show that $S_w(b)$ is strictly convex and continuously differentiable as a function of b, and goes to ∞ as b goes to infinity in any direction. This implies that the minimizer $\hat{\beta}_w$ exists uniquely, and equals the solution of

$$\dot{S}_w(b) \equiv \nabla_b S_w(b) \equiv \begin{pmatrix} \vdots \\ \frac{\partial S_w(b)}{\partial b_j} \\ \vdots \end{pmatrix} = 0. \tag{2.11}$$

Define the step function

$$W(r) = \begin{cases} 1 & \text{if } r \le 0, \\ w & \text{if } r > 0. \end{cases} \tag{2.12}$$

Then $Q_w\{r\} = W(r)r^2$ and $dQ_w\{r\}/dr = 2W(r)r$, a continuous function of r. We conclude that the gradient vector $\dot{S}_w(b)$ is a continuous function of b,

$$\begin{aligned} \dot{S}_w(b) &= \sum_{i=1}^{n} \nabla_b Q_w\{y_i - x_i b\} = -2\sum_{i=1}^{n} x_i' W(r_i(b)) r_i(b) \\ &= -2\boldsymbol{X}'\boldsymbol{W}(b)\boldsymbol{r}(b), \end{aligned} \tag{2.13}$$

where $\boldsymbol{W}(b) = \text{diag}[W(r_i(b))]$, the $n \times n$ diagonal matrix having $W(r_i(b))$ as its ith diagonal element. Combining (2.11) and (2.13), we see that $\hat{\beta}_w$ is the solution of

$$\boldsymbol{X}'\boldsymbol{W}(b)\boldsymbol{r}(b) = 0. \tag{2.14}$$

Iterative methods are needed to actually solve (2.14). Define

$$\beta(w, b) = [\boldsymbol{X}'\boldsymbol{W}(b)\boldsymbol{X}]^{-1}\boldsymbol{X}'\boldsymbol{W}(b)\boldsymbol{y}. \tag{2.15}$$

Notice that $\beta(w, b)$ is the Gauss-Markov solution for choosing β to minimize the modified sum of squares $\sum W_i \cdot (y_i - x_i\beta)^2$, when the weights $W_i \equiv W(r_i(b))$ are considered fixed. Then

$$\beta(w, b) - b = [\boldsymbol{X}'\boldsymbol{W}(b)\boldsymbol{X}']^{-1}\boldsymbol{X}'\boldsymbol{W}(b)\boldsymbol{r}(b), \tag{2.16}$$

and we see from (2.14) that $\hat{\beta}_w$ is the stationary value of $\beta(w, b)$,

$$\hat{\beta}_w = \beta(w, \hat{\beta}_w). \tag{2.17}$$

This last result can be thought of as follows: a trial value of b determines a plane in the (x, y) space; say $\mathcal{L}(b) \equiv \{y = xb, x \in \mathcal{R}^p\}$; this plane determines weights W_i on the data points (x_i, y_i), weight w if (x_i, y_i) is above $\mathcal{L}(b)$ and weight 1 if (x_i, y_i) is below $\mathcal{L}(b)$; these weights produce a weighted Gauss-Markov solution vector $\beta(w, b)$ according to (2.15), and thus a new plane $\mathcal{L}(\beta(w, b)) = \{y = x\beta(w, b), x \in \mathcal{R}^p\}$. The solution vector $\hat{\beta}_w$ is that value of b for which the plane $\mathcal{L}(b)$ that produces the weights coincides with the plane $\mathcal{L}(\beta(w, b))'$ that these weights produce.

The second derivative matrix is

$$\ddot{S}_w(b) \equiv \left(\frac{\partial^2 S_w(b)}{\partial b_j \partial b_h} \right)_{j,h=1,2,\dots,p} = 2\boldsymbol{X}'\boldsymbol{W}(b)\boldsymbol{X}. \tag{2.18}$$

This formula assumes that none of the $r_i(b) = 0$, a necessary assumption to avoid evaluating the discontinuous second derivative of $Q_w\{r\}$ at $r = 0$.

Starting from a trial value b, the usual Newton-Raphson updating formula suggests b_{NEW} as the solution to $\dot{S}_w(b) = 0$, where

$$\begin{aligned} b_{\mathrm{NEW}} - b = -\ddot{S}_w(b)^{-1}\dot{S}_w(b) &= [\boldsymbol{X}'\boldsymbol{W}(b)\boldsymbol{X}]^{-1}[\boldsymbol{X}'\boldsymbol{W}(b)\boldsymbol{r}(b)] \\ &= \beta(w, b) - b, \end{aligned} \tag{2.19}$$

or equivalently $b_{\mathrm{NEW}} = \beta(w, b)$. In this problem the obvious iterative method for solving (2.17), $b_1 = \beta(w, b_0)$, $b_2 = \beta(w, b_1)$, $b_3 = \beta(w, b_2)$, etc., is the same as a Newton-Raphson search for the minimizer of $S_w(b)$.

There is another way to search for solutions $\hat{\beta}_w$ to (2.17): by letting w vary as well as b. The following useful formula is verified in Section 6:

$$\frac{d\hat{\beta}_w}{dw} = \frac{1}{1+w}[\boldsymbol{X}'\boldsymbol{W}(\hat{\beta}_w)\boldsymbol{X}]^{-1}\boldsymbol{X}'|\boldsymbol{r}(\hat{\beta}_w)|, \tag{2.20}$$

where $|\boldsymbol{r}|$ indicates the vector of absolute values $(|r_1|, |r_2|, \dots, |r_n|)'$. Having found $\hat{\beta}_w$ for some value of w, we can approximate the solution at a nearby value $w + \Delta w$ by

$$\hat{\beta}_{w+\Delta w} \doteq \hat{\beta}_w + \frac{d\hat{\beta}_w}{dw}\Delta w. \tag{2.21}$$

The regression percentiles for the baseball and cholostyramine examples were found by an algorithm that alternated steps (2.21) and (2.19). Starting from a solution $\hat{\beta}_w$, two new values were found,

$$\hat{\beta}^{(0)}_{w+\Delta w} = \hat{\beta}_w + \frac{d\hat{\beta}_w}{dw}\Delta w \quad \text{and} \quad \hat{\beta}^{(1)}_{w+\Delta w} = \beta(w + \Delta w, \hat{\beta}^{(0)}_{w+\Delta w}). \tag{2.22}$$

Keeping Δw small, $\hat{\beta}^{(1)}_{w+\Delta w}$ was then a quite satisfactory approximation to $\hat{\beta}_{w+\Delta w}$, the solution for w now equal to $w + \Delta w$. It is convenient to start the calculations at $w = 1$, for which $\hat{\beta}_w = [\boldsymbol{X}'\boldsymbol{X}]^{-1}\boldsymbol{X}'\boldsymbol{y}$. The two-step algorithm (2.22) was then executed for 39 successively larger values $w_{i+1} = w_i \cdot (1+\Delta)$ and also for 39 successively smaller values $w_{j+1} = w_j/(1+\Delta)$. The choice $\Delta = .15$ gave excellent accuracy over the range of w values necessary to construct Figure 2. The Fortran program ran in 6.8 seconds, on a SUN 3/50 workstation. A more cautious version of the program, which iterated the second step in algorithm (2.22) until convergence was recorded, took about twice as long.

Remark A. If $\hat{\beta}^{(0)}_{w+\Delta w}$ agrees with $\hat{\beta}_{w+\Delta w}$ in the sense that for $i = 1, 2, \ldots, n$ the residuals $r_i(\hat{\beta}^{(0)}_{w+\Delta w})$ and $r_i(\hat{\beta}_{w+\Delta w})$ have the same sign, then $\hat{\beta}^{(1)}_{w+\Delta w}$ *exactly* equals the true solution $\hat{\beta}_{w+\Delta w}$. This occurred most of the time in the calculations for the baseball and cholostyramine examples. The underlying reason for this nice behavior, and the generally tractable character of asymmetric least squares calculations, is that $S_w(b)$ in (2.5) is " piecewise quadratic" in b.

Table 1 shows the first ten steps in the computation of the $\hat{\beta}_w$ vectors for the cholostyramine data. Altogether (2.22) was executed 79 times, for w ranging from $(1+.15)^{-39}$ to $(1+.15)^{39}$. (Notice that (2.22) included one step each of (2.21) and (2.19), neither of which by itself was sufficient.) Linear interpolation in $\alpha = p(w)$ was used to compute regression percentiles for the "nice" values of α, e.g. $\alpha = .67$, displayed in Figure 2. The covariate vectors were expressed as $x_i = (1, (z_i - \bar{z}), (z_i - \bar{z})^2)$, where $\bar{z}$ was the average compliance 60.116, in order to better condition the calculations.

Table 1. Regression percentile calculations: cholostyramine data

w	------------	$\hat{\beta}_w$	------------	$\alpha = p(w)$
	constant	linear	quadratic	
1.000	27.789	0.634	0.00415	0.49
1.150	29.069	0.640	0.00410	0.52
1.322	30.342	0.646	0.00405	0.52
1.521	31.599	0.652	0.00400	0.54
1.749	32.848	0.657	0.00394	0.55
2.011	34.083	0.662	0.00388	0.57
2.313	35.330	0.666	0.00381	0.62
2.660	36.566	0.670	0.00374	0.63
3.059	37.775	0.675	0.00369	0.65
3.518	38.949	0.679	0.00364	0.66
4.046	40.083	0.684	0.00361	0.68

Note: First ten steps of the regression percentile calculations for the cholostyramine data are given. Each successive line was produced from its predecessor via the two computations in (2.22). The columns *constant, linear, quadratic* refer to the coefficients in the linear model $y_i = \beta(0) + \beta(1) \cdot (z_i - \bar{z}) + \beta(2) \cdot (z_i - \bar{z})^2 +$ error$_i$, $\bar{z} = 60.116$. The last column gives $\alpha = p(w)$, the proportion of data points lying below the regression percentile.

Remark B. In order to get a rough idea of the relationship between w and α, consider the following special case: there are no covariates (all $x_i = 1$); $n \to \infty$; and the histogram of $y_1, y_2, \dots, y_n$ goes to a standard N(0,1) density. We denote $\hat{\beta}_w$ by β_w in this case. Let $w^{(\alpha)}$ indicate the value of w such that β_w equals $z^{(\alpha)} = \Phi^{-1}(\alpha)$, the 100$\alpha$th standard normal percentile point. Then formula (2.7) of Newey and Powell (1987) gives

$$w^{(\alpha)} = 1 + z^{(\alpha)}/\{\phi(z^{(\alpha)}) - (1-\alpha)z^{(\alpha)}\}, \tag{2.23}$$

$\phi(z) = e^{-\frac{1}{2}z^2}/\sqrt{2\pi}$, which can also be derived easily from (2.13).

Formula (2.23) yields values

$$\begin{array}{rcccccc} \alpha: & .5 & .67 & .75 & .84 & .90 & .95 \\ w^{(\alpha)}: & 1 & 3.02 & 5.52 & 12.81 & 28.11 & 80.04 \end{array} \tag{2.24}$$

(with $w^{(1-\alpha)} = 1/w^{(\alpha)}$). These values can be compared with those from the cholostyramine calculations, e.g., $\hat{w}^{(.67)} = 3.60$, $\hat{w}^{(.75)} = 5.00$, $\hat{w}^{(.84)} = 15.84$, $\hat{w}^{(.90)} = 40.13$. See also the w values in Table 2, Section 4.

Consider changing a standard normal density $\phi(z)$ to

$$\phi^{(.75)}(z) = \begin{cases} \phi(z)/c & z \le z^{(.75)} = .674 \\ 5.52\phi(z)/c & z > .674, \end{cases} \tag{2.25}$$

where the constant c is chosen to make $\int_{-\infty}^{\infty} \phi^{(.75)}(z)dz = 1$. Then $\phi^{(.75)}$ has expectation

$$\int_{-\infty}^{\infty} z\phi^{(.75)}(z)dz = .675, \tag{2.26}$$

this being relationship (2.17). Interestingly enough, it can be shown that $\phi^{(\alpha)}$ always has variance 1, i.e.,

$$\phi^{(\alpha)} \sim (z^{(\alpha)}, 1), \tag{2.27}$$

for all values of α, so $\phi^{(\alpha)}$ behaves somewhat like a location family. This result is special to the normal case.

Remark C. The asymmetric squared error loss function (2.4) leads to an important invariance property, as mentioned in Theorem 1 of Newey and Powell (1987): if the response vector is changed from $(y_1, \dots, y_n)$ to $(cy_1, cy_2, \dots, cy_n)$ then the solution vector (2.6) changes from $\hat{\beta}_w$ to $c\hat{\beta}_w$. (We actually have scale and location invariance: $\boldsymbol{y} \to c\boldsymbol{y} + d\boldsymbol{1}$ implies $\hat{\beta}_w \to c\hat{\beta}_w + (d, 0, 0, \dots, 0)'$, assuming that the covariate vectors x_i have first coordinate equal to one.)

Without the scale invariance property, our estimates of the regression percentiles would depend on the scale we chose to work with. In order to control the properties of the estimates we would need to augment the percentile estimation procedure with a supplementary estimate of scale. This is a familiar difficulty in the theory of robust estimation, discussed in Chapters 6 and 7 of Huber (1981). However it would be considerably more vexing here since we are particularly interested in situations where the scale varies with x_i, as in (3.12) and Section 5.

Scale invariance is not an exclusive property of asymmetric squared error loss. Any power loss function

$$Q_w\{r\} = \begin{cases} |r|^p & r \leq 0 \\ w|r|^p & r > 0, \end{cases} \quad (p > 0) \tag{2.28}$$

results in scale invariance for the corresponding estimates $\hat{\beta}_w$, the crucial property of (2.28) being $Q_w\{cr\} = |c|^p Q_w\{r\}$. Koenker and Bassett's regression quantiles (1982) are based on the choice $p = 1$. Some interesting asymmetric loss functions which have appeared in the econometric literature are ruled out by the invariance requirement: in particular Varian's LINEX loss, see Zellner (1986). Breckling and Chambers' (1988) asymmetric M-estimates include some non-invariant choices.

The power loss function with $p = 1.5$ is appealing as a compromise between the robustness of $p = 1$ and the high normal theory efficiency of $p = 2$ (see Section 5). Most of the calculations in this paper go through for all values of $p > 1$, though the case $p = 2$ has definite computational advantages (see Remark G of Section 3). The clerical workers example of Section 4 was investigated using $p = 1.5$ regression percentiles, but the results were similar to the $p = 2$ percentiles, and will not be reported here.

Remark D. One might worry that asymmetric least squares regression percentiles are estimating something other than the true regression percentiles $y^{(\alpha)}|x$, the 100αth percentile of y given x. However in cases where $y^{(\alpha)}|x$ is linear in x, as in model (3.12), discussed in Section 5, it is easy to show that the regression percentiles calculated by asymmetric least squares are consistent for $y^{(\alpha)}|x$. Newey and Powell (1987), in their Lemma and subsequent remarks, present a much more general consistency result for the convergence of regression percentiles to their population counterparts.

3. Absolute Residual Regression and the Tilt Statistic

The formula for the derivative of $\hat{\beta}_w$ with respect to w, (2.20), is particularly simple at $w = 1$,

$$\frac{d\hat{\beta}_w}{dw}\Big|_{w=1} = \frac{1}{2}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{R}, \tag{3.1}$$

$\boldsymbol{R}$ being the vector of absolute values of the ordinary residuals $r_i(\hat{\beta}_1)$,

$$R_i = |r_i(\hat{\beta}_1)|. \tag{3.2}$$

Statistic (3.1) is interesting in its own right on three counts: it is very easy to compute, and gives a quick approximation to $\hat{\beta}_w$ for values of w near one, $\hat{\beta}_w \doteq \hat{\beta}_1 + \frac{d\hat{\beta}_w}{dw}|_{w=1}(w-1)$; it helps answer questions like whether or not the decrease in curvature of the regression percentiles for the cholostyramine data, observed as we move upwards on the right side of Figure 2, is statistically significant; it is equivalent to Glejser's (1969) test for heteroscedasticity in the ordinary linear model.

In this section we assume that the covariate vectors are of the form $x_i = (1, x_{(1)i})$, where $\sum_{i=1}^n x_{(1)i}/n = 0$ (but see Remark C). Then we can write

$$\boldsymbol{X} = (\boldsymbol{1}, \boldsymbol{X}_{(1)}) \qquad (\boldsymbol{1}'\boldsymbol{X}_{(1)} = 0), \tag{3.3}$$

$\boldsymbol{X}_{(1)}$ being the $n \times (p-1)$ matrix with ith row $x_{(1),i}$. It is more convenient to deal with a multiple of statistic (3.1), namely

$$T \equiv c(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{R} \qquad \left(c = \sqrt{\pi/2} \cong 1.25\right). \tag{3.4}$$

The motivation for the constant $c = \sqrt{\pi/2}$ goes as follows: let

$$z_w = \frac{\hat{\beta}_w(0) - \hat{\beta}_1(0)}{\bar{\sigma}} \qquad \left(\bar{\sigma} \equiv \left(\sum_{i=1}^n R_i^2/n\right)^{\frac{1}{2}}\right), \tag{3.5}$$

where $\hat{\beta}_w(0)$ is the first coordinate of $\hat{\beta}_w$, and $\hat{\beta}_1(0)$ is the same quantity evaluated at $w = 1$. Because of (3.3), z_w is the difference of *intercepts* between the w-regression plane $\mathcal{L}_w$ and the OLS regression plane $\mathcal{L}_1$, measured in units of the empirical standard error of the ordinary residuals, $\bar{\sigma}$.

The first coordinate of equation (3.1) is

$$\frac{d\hat{\beta}_w(0)}{dw}\Big|_{w=1} = \frac{1}{2}\bar{R} \qquad \left(\bar{R} = \sum_{i=1}^n R_i/n\right), \tag{3.6}$$

because of (3.3), so

$$\frac{dz_w}{dw}\Big|_{w=1} = \frac{\bar{R}}{2\bar{\sigma}}. \tag{3.7}$$

Therefore

$$\frac{d\hat{\beta}_w}{dz_w}\Big|_{w=1} = \hat{c}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{R} \equiv \hat{T} \qquad (\hat{c} = \bar{\sigma}/\bar{R}). \tag{3.8}$$

The statistic $d\hat{\beta}_w/dz_w|_{w=1}$ is a little easier to interpret than (3.1). It measures how quickly the w-regression planes $\mathcal{L}_w$ "tilt" relative to $\mathcal{L}_1$, the derivative being taken with respect to units of the standardized intercept z_w. We could use $\hat{T}$ directly, but in our examples $\hat{c}$ nearly equaled its normal theory value $\sigma/E|z| = \sqrt{\pi/2}$, so $\hat{T} \doteq T$. We will use the term "Tilt Statistic" for both (3.4) and (3.8).

The tilt statistic is easy to calculate: ordinary linear regression provides $\hat{\beta}_1$ and the ordinary residual vector $\boldsymbol{r}(\hat{\beta}_1) = \boldsymbol{y} - \boldsymbol{X}\hat{\beta}_1$; this gives $\boldsymbol{R} = |\boldsymbol{r}(\hat{\beta}_1)|$, and also $\hat{c}$; then an ordinary linear regression of $\boldsymbol{R}$ on $\boldsymbol{X}$ gives $(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{R}$ as regression coefficients, providing T and $\hat{T}$.

Example: The cholostyramine data. Writing the quadratic regression of Table 1 in centered form $y_i = \beta(0) + \beta(1)(z_i - \overline{z}) + \beta(2)[(z_i - \overline{z})^2 - 1207.9]+\text{error}_i$, 1207.9 being the average of the $(z_i - \overline{z})^2$, ordinary linear regression gave

$$\hat{\beta}_1 = (32.80,\ .634,\ .00415). \tag{3.9}$$

Regression of $\boldsymbol{R}$ as the response variable in the same quadratic model then gave

$$T = (22.29,\ .112,\ -.00093) \tag{3.10}$$

(almost the same as $\hat{T}$ since $\hat{c} = (\overline{R^2})^{\frac{1}{2}}/\overline{R} = 1.254$).

The interpretation of T as (approximately) equal to $d\hat{\beta}_w/dz_w|_{w=1}$ allows us to approximate the regression percentiles. For example

$$\begin{aligned}\hat{\beta}^{(.75)} &\doteq \hat{\beta}_1 + T \cdot z^{(.75)} = \hat{\beta}_1 + T \cdot .674\\ &= (48.28,\ .709,\ .00353).\end{aligned} \tag{3.11}$$

The actual value of $\hat{\beta}^{(.75)}$ for the cholostyramine data was (49.53, .704, .00348). The reader is warned that not all cases work as well as this one. Whenever possible, this actual calculation of the regression percentiles is preferred.

Calculation (3.11) tacitly assumes that if $z_w = z^{(.75)} = .674$, then about 75% of the data points lie below $\mathcal{L}_w$. This is an obvious, though crude, interpretation of (3.5). We avoid using this interpretation if we calibrate the approximate regression percentiles obtained from the tilt statistic, by directly counting the proportion of data points lying below them, as we did with the actual regression percentiles.

The first coordinate of the tilt statistic relates to the intercept, and provides no tilting information, so we can concentrate on the last $(p-1)$ coordinates, say $T_{(1)}$. $T_{(1)}$ estimates the tilting parameter vector $\tau_{(1)}$, as discussed below in Section 5. The regression of $\boldsymbol{R}$ on $\boldsymbol{X}$, which gives T, also gives standard errors and t-values in the usual way. In the cholostyramine example $T_{(1)} = (.112, -.00093)$, with estimated standard errors (.040, .00153) and t-values $(2.80, -.61)$, on 161 degrees of freedom. The t-value corresponding to the linear term in the regression is quite significant, but not the t-value corresponding to the quadratic term. We have good reason to believe that the increasing slope of the regression percentiles with increasing α, observed on the right side of Figure 2, is genuine, but *not* the decreasing curvature.

Example: The baseball data. Ordinary least squares gave $\hat{\beta}_1 = (-3.15, 1.274)$ as the estimated coefficients in the model (final lead) $= b_0 + b_1 \cdot$ (lead at half) $+$ error. The tilt statistic for the slope b_1 was $T_{(1)} = .045 \pm .250$, giving a t-value of only 0.18 on 70 degrees of freedom. The regression percentiles for the baseball data in Figure 2 look parallel to the OLS line. This impression is verified by the tilt statistic being not significantly different from zero.

Remark E. Formula (3.4) for T gives the same value for the last $(p-1)$ coordinates $T_{(1)}$ whether or not $\mathbf{1}'\boldsymbol{X}_{(1)} = 0$, that is whether or not we have centered the regression problem. It is a convenient fact, which we used for the baseball data, that no centering is necessary to carry out the estimation and testing theory for $T_{(1)}$ or $\hat{T}_{(1)}$.

Remark F. An extension of the normal linear model which allows for heterogeneous variance is

$$y_i|x_i \sim \mathrm{N}(x_i\gamma, (x_i\tau)^2), \qquad i = 1, 2, \ldots, n, \tag{3.12}$$

where γ is the unknown regression parameter vector and τ is an unknown parameter vector such that $x_i\tau > 0$ for all i. The true absolute residuals $|r_i(\gamma)| \equiv R_i(\gamma)$ have expectation vector

$$E\{\boldsymbol{R}(\gamma)\} = \frac{1}{c}\boldsymbol{X}\tau, \tag{3.13}$$

so

$$c(\boldsymbol{X}'\boldsymbol{X}^{-1})\boldsymbol{X}'\boldsymbol{R}(\gamma) \tag{3.14}$$

is an unbiased estimate of τ.

Comparing (3.14) with (3.4), we see that T is an obvious estimator for τ, so $T_{(1)}$ is a candidate test statistic for heteroscedasticity. (The first coordinate of τ has no effect on heteroscedasticity.) In fact, Glejser's (1969) absolute residual regression test for heterogeneous variances is based on a generalized version of (3.4). Newey and Powell (1987) compare a difference analog of $d\hat{\beta}_1/dw|_{w=1}$, namely $\hat{\beta}_w - \hat{\beta}_{1/w}$, with Glejser's test, and note almost identical behavior in a Monte Carlo study. This is now explained by equation (3.4).

Remark G. For asymmetric power loss $Q_w\{r\} = W(r)|r|^p$, (2.28), formula (3.1) becomes

$$\frac{d\hat{\beta}_w}{dw}\Big|_{w=1} = \frac{1}{2(p-1)}[X'\mathrm{diag}(R^{p-2})X]^{-1}X'R^{p-1} \tag{3.15}$$

(for $p > 1$), where $\mathrm{diag}(R^{p-2})$ is the $n \times n$ diagonal matrix with ith diagonal element $R_i^{p-2} = |r_i(\hat{\beta}_1)|^{p-2}$. Notice that for $p < 2$, small values of R_i can cause numerical difficulties.

4. Influence Calculations

Regression percentiles based on asymmetric least squares minimization can be sensitive to outlying data points, especially for extreme values of α. This can be seen in the baseball example, left side of Figure 2, where the two extreme regression percentiles tilt sharply away from the inner ones. This section gives simple formulas for the influence of an individual data point as a given regression percentile. The section includes a multiple regression example, illustrating the influence calculations and how they might be used to robustify the regresssion percentiles.

We will derive, in Section 6, two measures of influence for the ith data point (x_i, y_i) on the w-regression plane $\mathcal{L}_w = \{y = x\hat{\beta}_w, x \in \mathcal{R}^p\}$:

$$\begin{aligned} D^2_{w,i} &= [W(r_i(\hat{\beta}_w)) \cdot r_i(\hat{\beta}_w)]^2 x_i M x_i' \\ M &\equiv [X'W(\hat{\beta}_w)X]^{-1}[X'X][X'W(\hat{\beta}_w)X]^{-1} \end{aligned}$$

and (4.1)

$$\begin{aligned} \tilde{D}^2_{w,i} &= [W(r_i(\hat{\beta}_w)) \cdot r_i(\hat{\beta}_w)]^2 x_i \tilde{M} x_i' \\ \tilde{M} &= [X'W(\hat{\beta}_w)X]^{-1}[\tilde{X}'\tilde{X}][X'W(\hat{\beta}_w)X]^{-1}, \end{aligned}$$

where $\tilde{X} = X - 1\overline{x}$, $\overline{x} = \sum x_i/n$. The notation here follows (2.12)–(2.14).

Both $D^2_{w,i}$ and $\tilde{D}^2_{w,i}$ are closely related to Cook's distance, as described in Section 4.2.5.1 of Chatterjee and Hadi (1988). These "distances" are overall measures of influence of the ith data point on the entire w-regression plane : $\tilde{D}^2_{w,i}$ relates to the tilting of $\mathcal{L}_w$ relative to the OLS plane, discounting the influence of (x_i, y_i) on the intercept of $\mathcal{L}_w$. See Remark M in Section 6.

The left panel of Figure 3 displays a summary of the data for our third example, a multiple regression in which the two regressands x_1 = age and x_2 = years clerical experience are used to predict y = log salary, for $n = 122$ clerical workers at a large pharmaceutical firm. The OLS estimate in the model

$$y_i = \beta(0) + \beta(1)x_{1i} + \beta(2)x_{2i} + \mathrm{error}_i \qquad i = 1, 2, \ldots, 122 \tag{4.2}$$

is

$$\hat{\beta} = (9.69,\ -.0181,\ .0319).$$

The horizontal axis in Figure 3 is the predicted log salary using model (4.2), $\hat{\beta}(0) + \hat{\beta}(1)x_{1i} + \hat{\beta}(2)x_{2i}$.

The estimated slopes $\hat{\beta}(1) = -.0181$ and $\hat{\beta}(2) = .0319$ are both significantly non-zero, with t-values 4.09 and 6.15 respectively, on 119 degrees of freedom. The age slope is negative, so for a given level of experience, the predicted log salary *decreases* with increasing age.

The right panel of Figure 3 displays the maximum percentage influence of each point. The plotted number is

$$\max_{w} 100 \cdot \tilde{D}^2_{w,i} \Big/ \sum_{j=1}^{122} \tilde{D}^2_{w,j}, \tag{4.3}$$

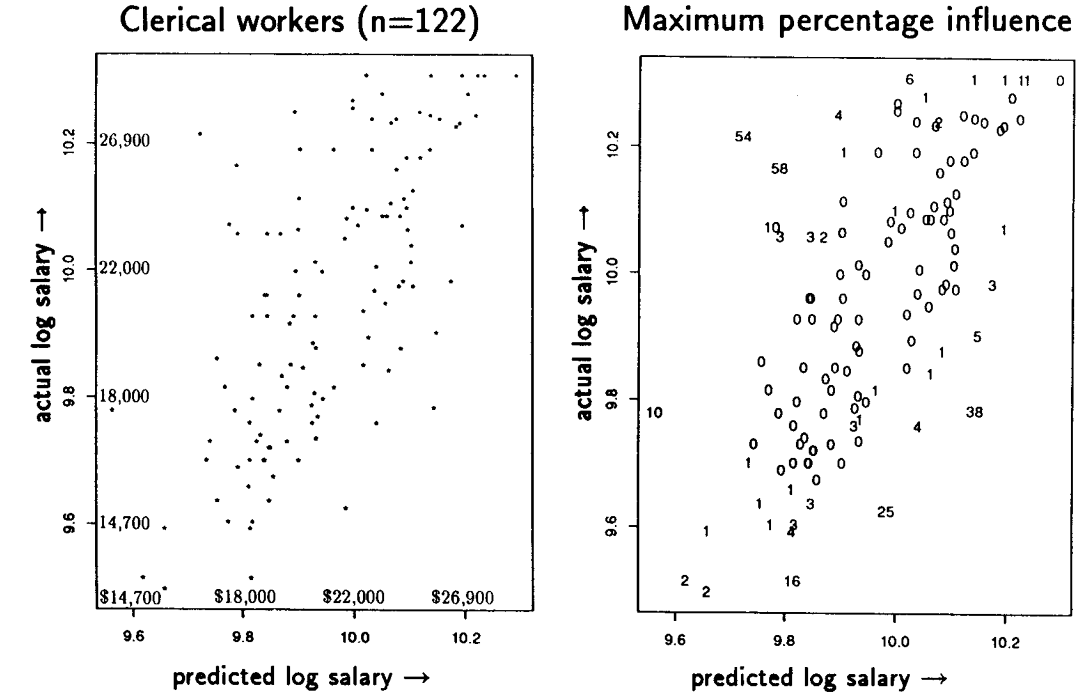

Figure 3. The clerical workers example. Left panel shows data for 122 clerical workers; vertical axis is log salary, horizontal axis is predicted log salary using model (4.2). Right panel: each of the 122 points is labelled with the maximum value of its percentage influence, (4.3), the maximum over $\alpha = .10, .16, .25, .33, .50, .67, .75, .84, .90$. Four of the points are seen to have percentage influence $\geq 20\%$ for at least one of the nine α values.

the maximum being taken over the nine values of w corresponding to the regression percentiles for α =.10, .16, .25, .33, .50, .67, .75, .84, .90. Four of the points, two above and two below the central cloud, are seen to have maximum percentage influence $\geq$ 20%.

Regression percentiles were calculated for the clerical workers data, using algorithm (2.22). Table 2 show $\hat{\beta}^{(\alpha)}$ for α =.05, .10, .16, .25, .33, .50, .67, .75, .84, .90, .95. Two analyses were run, one for all n = 122 points and one for the 118 points remaining after deletion of the four outlying points having maximum influence $\geq$20%.

Both analyses give the same picture, but in a smoother way after the outliers were removed: the estimated regression slope for experience, $\hat{\beta}^{(\alpha)}(2)$, stays reasonably constant as α changes. However the slope for age, $\hat{\beta}^{(\alpha)}(1)$ becomes almost negligible at the larger values of α. This says that age is not an important predictive variable for the better-paid workers, though it is important for the middle and poorly-paid workers.

Table 2. Regression percentile analysis for the clerical workers

		— $\hat{\beta}^{(\alpha)}$, all 122 points —			— $\hat{\beta}^{(\alpha)}$, all 118 points —			
α	w	($\hat{\beta}^{(\alpha)}(0)$,	$\hat{\beta}^{(\alpha)}(1)$,	$\hat{\beta}^{(\alpha)}(2)$)	($\hat{\beta}^{(\alpha)}(0)$,	$\hat{\beta}^{(\alpha)}(1)$,	$\hat{\beta}^{(\alpha)}(2)$)	w
.05	.0075	(9.20	−.0248	.0191)	(9.74	−.0258	.0295)	.0064
.10	.03	(9.77	−.0221	.0234)	(9.78	−.0212	.0308)	.022
.16	.06	(9.80	−.0212	.0256)	(9.81	−.0208	.0323)	.056
.25	.13	(9.84	−.0211	.0281)	(9.84	−.0210	.0332)	.11
.33	.20	(9.86	−.0208	.0293)	(9.87	−.0217	.0335)	.21
.50	1.00	(9.96	−.0181	.0319)	(9.96	−.0198	.0352)	1.00
.67	3.48	(10.04	−.0122	.0324)	(10.02	−.0152	.0360)	3.22
.75	5.55	(10.07	−.0098	.0323)	(10.06	−.0127	.0365)	5.95
.84	26.1	(10.15	−.0006	.0297)	(10.10	−.0107	.0365)	12.60
.90	49.4	(10.18	.0022	.0287)	(10.17	−.0065	.0355)	52.07
.95	354	(10.23	.0028	.0200)	(10.21	.0010	.0392)	233
		intercept	age	experience	intercept	age	experience	

Note: Left: $\hat{\beta}^{(\alpha)}$ based on all 122 workers. Right: $\hat{\beta}^{(\alpha)}$ based on 118 workers after removing the four workers with the largest influences, as shown in the right panel of Figure 3. The estimated regression slope for experience is reasonably constant, but the slope for age approaches zero as α gets large.

Remark H. The geometry of the clerical workers point cloud is moderately complicated, generally sloping upwards in experience and downwards in age, but with the age slope flattening out toward the top of the cloud. It is difficult to spot this structure by eye, even using a good 3-dimensional scatterplot rotation and viewing program. Also, this geometry is not consistent with a simple model of heteroscedasticity, like (3.12), where a tilt below the OLS plane must be balanced out by a reverse tilt above (Theorem 2 of Newey and Powell (1987)). The methods of regression percentiles are particularly useful in multiple regression situations, where simple scattergram pictures like those in Figure 1 are generally not available, and visual inspection is difficult.

The tilt statistic $T_{(1)}$, the last $p-1$ coordinates of expression (3.4), based on all 122 workers, for (age, experience), was

$$T_{(1)} = (.0093 \pm .0025,\ \ .0025 \pm .0029). \tag{4.4}$$

(The standard errors are obtained from the usual formulas applied to regression (3.4), as in Section 3.) We see that the age tilt is significantly non-zero, while the experience tilt is not. This conclusion agrees with the actual tilting, as shown in Table 2. Using only the data for the 118 workers gave

$$T_{(1)} = (.0066 \pm .0023,\ \ .0010 \pm .0027), \tag{4.5}$$

leading to the same qualitative interpretation.

5. Efficiency Calculations

Asymmetric least squares turns out to be a reasonably efficient way of estimating the true regression percentiles, in a normal-theory model like (3.12), where we compute asymptotic efficiencies relative to the maximum likelihood estimates. This section presents the main results, with details and proofs deferred until Section 6.

We first consider a simple case where there are no covariates. The data consists of n observations from a scale-location family

$$y_i = \mu + \sigma Z_i, \qquad i = 1, \dots, n, \tag{5.1}$$

$Z_1, Z_2, \dots, Z_n$ being independent and identically distributed (i.i.d.) variates drawn from a known probability density function $f^0(z)$ on the real line. The Z_i are assumed to be *standardized*, that is $E\{Z_i\} = 0$, $\text{var}\{Z_i\} = 1$.

Let β_w^0 indicate the "true w-mean for Z", i.e., the minimizer of $EQ_w\{Z-b\}$ over the choice of b, as in (2.4)–(2.6). (Notice that $\beta_1^0 = EZ = 0$.) The w-mean for $y = \mu + \sigma Z$, the minimizer over b of $EQ_w\{y-b\}$, is easily seen to be

$$\beta_w = \mu + \sigma \beta_w^0. \tag{5.2}$$

In other words, the w-mean is scale and location invariant. Our first efficiency result compares the asymptotic variance of two consistent estimates for β_w : $\hat{\beta}_w$, the sample asymmetric mean, versus $\tilde{\beta}_w$, the maximum likelihood estimate (MLE).

The Fisher information matrix for estimating (μ, σ) in (5.1) is

$$\mathcal{I}(\mu, \sigma) = \frac{1}{\sigma^2} \begin{pmatrix} i_{11} & i_{12} \\ i_{12} & i_{22} \end{pmatrix} \tag{5.3}$$

where the i_{hj} are computed in terms of $h(z) = d \log f^0(z)/dz$: $i_{11} = Eh(Z)^2$, $i_{12} = Eh(Z)\{h(Z)Z - 1\}$, and $i_{22} = E\{h(Z)Z - 1\}^2$. The asymptotic variance $\text{AVAR}(\tilde{\beta}_w) \equiv \lim_{n \to \infty} n \cdot \text{var}(\tilde{\beta}_w)$ of the MLE $\tilde{\beta}_w = \tilde{\mu} + \tilde{\sigma}\beta^0_w$ [$\tilde{\mu}$ and $\tilde{\sigma}$ being the MLEs of μ and σ] is then

$$\frac{\text{AVAR}(\tilde{\beta}_w)}{\sigma^2} = \frac{i_{22} - 2i_{12}\beta^0_w + i_{11}(\beta^0_w)^2}{i_{11}i_{22} - i^2_{12}}. \tag{5.4}$$

The sample w-mean $\hat{\beta}_w$ based on $y_1, y_2, \ldots, y_n$ is by definition the minimizer over b of

$$\sum_{i=1}^{n} Q_w\{y_i - b\}, \tag{5.5}$$

as in (2.6). A standard argument based on the theory of M-estimation, Corollary 2.5 of Huber (1981), shows that $\hat{\beta}_w$ has asymptotic variance

$$\frac{\text{AVAR}(\hat{\beta}_w)}{\sigma^2} = \frac{E[W(Z - \beta^0_w) \cdot (Z - \beta^0_w)]^2}{[1 + (w-1)\text{Prob}\{Z > \beta^0_w\}]^2}. \tag{5.6}$$

The asymptotic relative efficiency (ARE) of $\hat{\beta}_w$, also called the asymptotic efficiency, compared to the fully efficient estimate $\tilde{\beta}_w$ is

$$\text{ARE}_{\text{SL}}(\hat{\beta}_w) = \frac{\text{AVAR}(\tilde{\beta}_w)}{\text{AVAR}(\hat{\beta}_w)}. \tag{5.7}$$

The subscript SL indicates the scale-location family situation (5.1).

We are particularly interested in the case where the Z_i in (5.1) are standard normal variates, $Z_i \sim \text{N}(0, 1)$. Suppose we set w equal to $w^{(\alpha)}$, (2.23), so that $\beta^0_w = z^{(\alpha)} = \Phi^{-1}(\alpha)$. Formulas (5.4), (5.6), (5.7) become

$$\begin{aligned} \frac{\text{AVAR}(\tilde{\beta}_{w^{(\alpha)}})}{\sigma^2} &= [1 + z^{(\alpha)^2}/2] \\ \frac{\text{AVAR}(\hat{\beta}_{w^{(\alpha)}})}{\sigma^2} &= \frac{[1+z^{(\alpha)^2}][1+(w^{(\alpha)^2}-1)\{(1-\alpha)-\phi(z^{(\alpha)})\cdot z^{(\alpha)}/(1+z^{(\alpha)^2})\}]}{[1 + (w^{(\alpha)} - 1)(1-\alpha)]^2} \end{aligned} \tag{5.8}$$

and

$$\mathrm{ARE}_{\mathrm{SL}}(\hat{\beta}_{w^{(\alpha)}}) = \frac{[1+z^{(\alpha)^2}/2][1+(w^{(\alpha)}-1)(1-\alpha)]^2}{[1+z^{(\alpha)^2}][1+(w^{(\alpha)^2}-1)\{(1-\alpha)-\phi(z^{(\alpha)})z^{(\alpha)}/(1+z^{(\alpha)^2})\}]}. \quad (5.9)$$

Table 3 presents $\mathrm{ARE}_{\mathrm{SL}}(\hat{\beta}_{w^{(\alpha)}})$ for several values of α. The efficiency is seen to be quite high, especially for $.25 \leq \alpha \leq .75$. For comparison, the asymptotic relative efficiency of $\overline{\beta}^{(\alpha)}$, the 100$\alpha$th sample percentile of $y_1, y_2, \ldots, y_n$, is also presented based on the standard formula

$$\frac{\mathrm{AVAR}(\overline{\beta}^{(\alpha)})}{\sigma^2} = \frac{\alpha(1-\alpha)}{\phi(z^{(\alpha)})^2}, \quad (5.10)$$

Kendall and Stuart (1958), Section 10.15.

In this simple situation, $\overline{\beta}^{(\alpha)}$ corresponds to Koenker and Bassett's regression quantile estimate. As might be expected, estimates $\overline{\beta}^{(\alpha)}$ based on asymmetric least absolute deviations are less efficient than asymmetric least squares estimates, in a normal scale-location model. This comparison ignores one fact discussed later, Remark K: for a given value of α we need to know $w^{(\alpha)}$ in order to calculate $\hat{\beta}_{w^{(\alpha)}}$, which in this case means knowing that the Z_i are normally distributed. No such knowledge is required for $\overline{\beta}^{(\alpha)}$.

Table 3. Asymptotic relative efficiency of $\hat{\beta}_{w^{(\alpha)}}$ and $\overline{\beta}^{(\alpha)}$

α :	.50	.67 or .33	.75 or .25	.84 or .16	.90 or .10	.95 or .05
$\mathrm{ARE}_{\mathrm{SL}}(\hat{\beta}_{w^{(\alpha)}})$:	1	.98	.95	.87	.74	.55
$\mathrm{ARE}_{\mathrm{SL}}(\overline{\beta}^{(\alpha)})$:	.64	.65	.66	.66	.62	.53

Note: Table shows asymptotic relative efficiency of the asymmetric least squares estimator $\hat{\beta}_{w^{(\alpha)}}$ compared to the MLE, in the scale-location model $y_i = \mu + \sigma Z_i$, $Z_i \sim \mathrm{N}(0,1)$ independently for $i = 1, 2, \ldots, n$. Also shown is the asymptotic relative efficiency of $\overline{\beta}^{(\alpha)}$, the sample 100$\alpha$th percentile. Efficiency of the asymmetric least squares estimator is quite high in this model, especially for $.25 \leq \alpha \leq .75$. The sample percentiles, which are Koenker and Bassett's regression quantiles here, are less efficient.

Returning to regression problems, we consider a generalization of the heteroscedastic normal linear model (3.12),

$$y_i = x_i\gamma + (x_i\tau)Z_i, \quad (5.11)$$

where, as in (5.1), $Z_1, Z_2, \ldots, Z_n$ is an i.i.d. sample of standardized variates with density $f^0(z)$; γ and τ are unknown $p \times 1$ parameter vectors controlling the mean and standard deviation of y as a function of x; and τ is assumed to satisfy $x_i\tau > 0$ for all i. The parameter γ is the usual regression vector. Note: The efficiency results below are particularly simple for the usual homoscedastic linear model where $x_i\tau$ equals a constant.

The true w-mean of y_i given x_i, i.e., the minimizer of $E\{Q_w\{y_i - \mu\}|x_i\}$ over the choice of μ, is

$$x_i\gamma + (x_i\tau)\beta_w^0 = x_i(\gamma + \tau\beta_w^0), \tag{5.12}$$

where as before β_w^0 is the true w-mean of Z. The vector

$$\beta_w \equiv \gamma + \tau\beta_w^0 \tag{5.13}$$

is defined to be the *true w-regression vector.* Our main efficiency result compares the asymptotic variance of two estimates of β_w: the asymmetric least squares estimate $\hat{\beta}_w$ versus the MLE $\tilde{\beta}_w = \tilde{\gamma} + \tilde{\tau}\beta_w^0$.

The efficiency results depend on two $p \times p$ matrices,

$$\boldsymbol{M}_0 = [\boldsymbol{X}'\text{diag}(x_i\tau)^{-2}\boldsymbol{X}]^{-1}$$

and

$$\boldsymbol{M}_1 = [\boldsymbol{X}'\boldsymbol{X}]^{-1}[\boldsymbol{X}'\text{diag}(x_i\tau)^2\boldsymbol{X}][\boldsymbol{X}'\boldsymbol{X}]^{-1}, \tag{5.14}$$

where the diagonal matrices have ith diagonal elements $1/(x_i\tau)^2$ and $(x_i\tau)^2$ respectively. Assumptions 1–4 of Newey and Powell (1987), which we will follow here, imply that $n\boldsymbol{M}_0$ and $n\boldsymbol{M}_1$ approach limiting matrices in probability as $n \to \infty$, say

$$n\boldsymbol{M}_0 \to \boldsymbol{m}_0 \quad \text{and} \quad n\boldsymbol{M}_1 \to \boldsymbol{m}_1. \tag{5.15}$$

(In the homoscedastic case $x_i\tau \equiv \sigma$, both $\boldsymbol{m}_0$ and $\boldsymbol{m}_1$ equal $\lim \sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1}/n$.)

We then have the following results:

$$\text{AVAR}(\tilde{\beta}_w) \equiv \lim_{n\to\infty} n \cdot \text{cov}(\tilde{\beta}_w) = a_0(w)\boldsymbol{m}_0, \tag{5.16}$$

where $a_0(w)$ is the right side of (5.4);

$$\text{AVAR}(\hat{\beta}_w) = a_1(w)\boldsymbol{m}_1, \tag{5.17}$$

where $a_1(w)$ is the right side of (5.6); and

$$\begin{aligned}\text{ARE}(\hat{\beta}_w) &\equiv \text{AVAR}(\hat{\beta}_w)^{-\frac{1}{2}}\text{AVAR}(\tilde{\beta}_w)\text{AVAR}(\hat{\beta}_w)^{-\frac{1}{2}} \\ &= \text{ARE}_{\text{SL}}(\hat{\beta}_w)\boldsymbol{m}_1^{-\frac{1}{2}}\boldsymbol{m}_0\boldsymbol{m}_1^{-\frac{1}{2}},\end{aligned} \tag{5.18}$$

where $\mathrm{ARE}_{\mathrm{SL}}(\hat{\beta}_w)$ is (5.7), the asymptotic efficiency of $\hat{\beta}_w$ for estimating β_w in the scale-location family (5.1). (In the homoscedastic case, $x_i\tau \equiv \sigma^2$, we have $\mathrm{ARE}(\hat{\beta}_w) = \mathrm{ARE}_{\mathrm{SL}}(\hat{\beta}_w)$.)

The case where the Z_i in (5.11) are N(0,1) variates (3.12), is a heteroscedastic version of the usual linear model which allows the standard deviation of y_i to depend linearly on the covariate x_i. In this case the OLS estimate $\hat{\gamma}$ for γ has asymptotic efficiency

$$\mathrm{ARE}(\hat{\gamma}) = m_1^{-\frac{1}{2}} m_0 m_1^{-\frac{1}{2}}, \tag{5.19}$$

relative to the MLE $\tilde{\gamma}$. (5.19) is just the familiar expression for the efficiency of ordinary least squares compared to weighted least squares, using the optimum weights $(x_i\tau)^{-2}$ appropriate to (5.11).

For the normal case, (5.18) can be written in the following evocative form:

$$\mathrm{ARE}(\hat{\beta}_w) = \mathrm{ARE}_{\mathrm{SL}}(\hat{\beta}_w) \cdot \mathrm{ARE}(\hat{\gamma}), \tag{5.20}$$

where $\mathrm{ARE}_{\mathrm{SL}}(\hat{\beta}_{w(\alpha)})$ is given by (5.9), with the numerical values shown in Table 3. In other words, the asymptotic relative efficiency of $\hat{\beta}_w$ is composed of two factors: the ARE of asymmetric least squares estimation in the scale-location problem, times the ARE of ordinary least squares for estimating γ. It is not surprising that this last factor appears, since the asymmetric least squares estimate $\hat{\beta}_w$ *is* the OLS estimator $\hat{\gamma}$ for $w = 1$.

Remark I. It is not difficult to obtain a "Gauss-Markov" version of the asymmetric least squares estimation procedure which eliminates the factor $\mathrm{ARE}(\hat{\gamma})$ from its asymptotic relative efficiency.

In a certain sense (5.20), and Remark I, say that the asymptotic efficiency of the asymmetric least squares estimator $\hat{\beta}_w$ in the normal regression model (5.11) is the same as its efficiency in the normal scale-location model (5.1), as given in Table 3. For example, the 25th and 75th regression percentiles, which were of particular interest in the clerical workers salary study described in Section 4, are estimated with 95% asymptotic efficiency, ignoring the factor $\mathrm{ARE}(\hat{\gamma})$. A result similar to (5.20) holds for regression quantiles, at least in the homoscedastic case discussed in Theorem 3.2 of Koenker and Bassett (1982), so that the 25th and 75th percentiles would be estimated with 66% efficiency by that method.

We now give a different way of writing (5.18), that applies *whether or not the Z_i in* (5.11) *are normal*:

$$\mathrm{ARE}(\hat{\gamma})^{-\frac{1}{2}} \cdot \mathrm{ARE}(\hat{\beta}_w) \cdot \mathrm{ARE}(\hat{\gamma})^{-\frac{1}{2}} = \mathrm{ARE}_{\mathrm{SL}}(\hat{\beta}_w)/\mathrm{ARE}_{\mathrm{SL}}(\hat{\gamma}), \tag{5.21}$$

where $\mathrm{ARE}(\hat{\gamma})$ is the asymptotic relative efficiency of the least squares estimate $\hat{\gamma}$ for γ. This result, like (5.20), says that the asymptotic efficiency of $\hat{\beta}_w$ in the heteroscedastic regression model (5.11) is determined by its efficiency in the corresponding scale-location model (5.1). Result (5.20) is simpler than (5.21) because

$$\text{ARE}_{\text{SL}}(\hat{\gamma}) = [i_{11} - i_{12}^2/i_{22}]^{-1} \tag{5.22}$$

equals 1 when the Z_i are N(0,1).

We now consider the asymptotic behavior of the tilt statistic T, (3.4), under the assumptions of model (3.12), which is model (5.11) with the $Z_i \sim$ N(0,1). The absolute residuals $R_i = |r_i(\hat{\gamma})|$ are asymptotically independent with mean and variance

$$R_i \sim [x_i\tau/c, (x_i\tau)^2(1 - 1/c^2)]. \tag{5.23}$$

(Formula (5.23) gives exactly the mean and variance for $|r_i(\gamma)| = (x_i\tau)Z_i$, $Z_i \sim$ N(0,1).) Then $T = c(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{R}$ has asymptotic mean vector and covariance matrix

$$T \to (\tau, (c^2 - 1)\boldsymbol{m}_1/n), \tag{5.24}$$

where $c = \sqrt{\pi/2} \cong 1.253$ as in (3.4), and $\boldsymbol{m}_1$ is the limit of n times matrix $\boldsymbol{M}_1$ in (5.14). Moreover a symmetry argument, Section 6, shows that T is uncorrelated with $\hat{\gamma}$,

$$\text{cov}(\hat{\gamma}, T) = \mathbf{0}. \tag{5.25}$$

Taken together, (5.23)–(5.25) provide a convenient way to test for tilting effects of the regression percentiles relative to the OLS regression plane. Suppose that x_i equals $(1, x_{(1)i})$ as in Section 3, and write $\tau = (\tau_0, \tau_{(1)})'$. The hypothesis of no tilting effects is equivalent to

$$H_0 : \tau_{(1)} = 0. \tag{5.26}$$

If H_0 is true then true w-regression planes $\mathcal{L}_w = \{y = x\beta_w\}$ are parallel to each other, $\mathcal{L}_w = \{y = x\gamma + \tau_0\beta_w^0\}$, according to (5.12), (5.13).

Under H_0, (5.24) becomes

$$T \to \left(\begin{pmatrix}\tau_0\\ 0\end{pmatrix}, \tau_0^2(c^2 - 1)\boldsymbol{m}_{00}/n\right) \qquad \left(\boldsymbol{m}_{00} = \lim(\boldsymbol{X}'\boldsymbol{X}/n)^{-1}\right). \tag{5.27}$$

We can use $T_{(1)}$, the last $p - 1$ coordinates of T, to test H_0. This was done in Sections 3 and 4 by approximating the null-hypothesis covariance matrix of T by

$$c^2\hat{\sigma}^2(\boldsymbol{R})(\boldsymbol{X}'\boldsymbol{X})^{-1} \qquad \left(\hat{\sigma}^2(\boldsymbol{R}) \equiv \|(I - \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}')\boldsymbol{R}\|^2/(n - p)\right). \tag{5.28}$$

In other words we treated $T = c(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{R}$ as if $\boldsymbol{R}$ was a response vector in an ordinary linear model. This is justified by the null hypothesis version of (5.23),

$$R_i \to (\tau_0/c, \tau_0^2(1 - 1/c^2)), \tag{5.29}$$

in which the R_i are homoscedastic.

Individual tests for the coordinates of $\tau_{(1)}$ being zero were used in Sections 3 and 4, based on the approximate t statistics $T_i/\{c^2\hat{\sigma}^2(R)(\boldsymbol{X}'\boldsymbol{X})_{ii}^{-1}\}^{\frac{1}{2}}$. An omnibus F-test is also possible of course, but won't be discussed here. Newey and Powell (1987) present a more elaborate testing theory.

Remark J. The estimated w-regression vectors $\hat{\beta}_w$ are highly correlated with one another, especially for nearby values of w, which complicates hypothesis testing. If all the coordinates of $\hat{\gamma} = \hat{\beta}_1$ are significantly non-zero, how should we assess the significance of the coordinates of say $\hat{\beta}_5$? The tilt statistic offers an uncorrelated local decomposition of the $\hat{\beta}_w$ values,

$$\hat{\beta}_w \doteq \hat{\beta}_1 + \frac{w-1}{2c}T = \hat{\gamma} + \frac{w-1}{2c}T \tag{5.30}$$

for w near 1, (3.1), with $\hat{\gamma}$ and T uncorrelated, (5.25). This simplifies the testing problem. We can decide in the usual way the significance of coordinates of $\hat{\gamma}$, and then, nearly independently, test for differences between $\gamma = \beta_1$ and β_w.

Remark K. The efficiency calculations (5.17), (5.18) for $\hat{\beta}_w$ require correction if they are to be applied to the regression percentiles $\hat{\mathcal{L}}^{(\alpha)}$ rather than the w-regression planes $\hat{\mathcal{L}}_w$. We need to take account of the fact that $\hat{\beta}^{(\alpha)}$ in (2.9) is chosen to have proportion α of the n data points lying below $\hat{\mathcal{L}}^{(\alpha)}$. Roughly speaking, this makes the intercept of $\hat{\mathcal{L}}^{(\alpha)}$ have the asymptotics of Koenker and Bassett's regression quantile intercept, while the $p-1$ slopes of $\hat{\mathcal{L}}^{(\alpha)}$ behave as indicated by (5.17).

We now present a formula for AVAR($\hat{\beta}^{(\alpha)}$), in the case where the Z_i in (5.11) are N(0,1), (Model 3.12):

$$\text{AVAR}(\hat{\beta}^{(\alpha)}) = (I_p - \tau x^0, -\tau\phi(z^{(\alpha)}))\begin{pmatrix} a_1^{(\alpha)}m_1 & -c^{(\alpha)}\tau' \\ -\tau c^{(\alpha)} & \alpha(1-\alpha) \end{pmatrix}\begin{pmatrix} I_p - x^{0\prime}\tau' \\ -\tau'\phi(z^{(\alpha)}) \end{pmatrix}. \tag{5.31}$$

The new definitions are as follows: $a_1^{(\alpha)}$ is the right side of the lower expression in (5.8); $c^{(\alpha)} = \{\phi(z^{(\alpha)}) + \alpha z^{(\alpha)}\}/\{1 + (w^{(\alpha)} - 1)(1-\alpha)\}$; I_p is the $p \times p$ identity matrix; and $x^0 \equiv E\{x/x\tau\}$, the expectation being taken over the random selection of x, as in Newey and Powell's (1987) Assumption 1.

Result (5.31) should be compared with (5.17), with $w = w^{(\alpha)}$, for the normal situation $Z_i \sim$ N(0,1). In the special case $x = (1, x_{(1)})$, $\tau = (\tau_0, 0)$, and $E\{x_{(1)}\} = 0$, (5.31) gives the same result as (5.17) with $w = w^{(\alpha)}$.

6. Proofs and Details

We now complete some of the more important arguments that were left open in previous sections.

Section 2. In order to verify the useful formula (2.20), define

$$\delta(w,b) \equiv \beta(w,b) - b = [\boldsymbol{X}'\boldsymbol{W}(b)\boldsymbol{X}]^{-1}\boldsymbol{X}'\boldsymbol{W}(b)\boldsymbol{r}(b), \tag{6.1}$$

as in (2.16). The solution vector $\hat{\beta}_w$ satisfies $\delta(w, \hat{\beta}_w) = 0$, according to (2.17). If none of the residuals $r_i(b) = 0$, then $\nabla_b \boldsymbol{W}(b) = \boldsymbol{0}$ and so

$$\nabla_b \delta(w,b) = -\boldsymbol{I}_p, \tag{6.2}$$

where $\boldsymbol{I}_p$ is the $p \times p$ identity matrix. Formula (2.22) can fail if any of the $r_i(b) = 0$, but the more special result

$$\nabla_b \delta(w,b)|_{b=\hat{\beta}_w} = -\boldsymbol{I}_p \tag{6.3}$$

is always valid. (6.3) requires a careful but straightforward accounting of boundary cases $r_i(\hat{\beta}_w) = 0$, which is omitted here.

Write $\delta(w,b) = A_w^{-1} B_w$, where $A_w \equiv \boldsymbol{X}'\boldsymbol{W}(b)\boldsymbol{X}$ and $B_w \equiv \boldsymbol{X}'\boldsymbol{W}(b)\boldsymbol{r}(b)$ are now thought of as functions of w with b held fixed. Then since

$$\frac{\partial W(r_i(b))}{\partial w} = I_+(r_i(b)) \qquad \left(I_+(r) = \begin{cases} 0 & \text{if } r \le 0 \\ 1 & \text{if } r > 0 \end{cases}\right), \tag{6.4}$$

we calculate

$$\begin{aligned} \frac{\partial \delta(w,b)}{\partial w} &= A_w^{-1}\left[\frac{\partial B_w}{\partial w} - \frac{\partial A_w}{\partial w} A_w^{-1} B_w\right] \\ &= [\boldsymbol{X}'\boldsymbol{W}(b)\boldsymbol{X}]^{-1}\left[\boldsymbol{X}'\boldsymbol{I}_+(b)\boldsymbol{r}(b) - \frac{\partial A_w}{\partial w}\delta(w,b)\right], \end{aligned} \tag{6.5}$$

$\boldsymbol{I}_+(b)$ being the $n \times n$ diagonal matrix with ith diagonal element $I_+(r_i(b))$.

If $b = \hat{\beta}_w$ then $\delta(w, \hat{\beta}_w) = 0$, and (6.5) reduces to

$$\frac{\partial \delta(w,b)}{\partial w}\Big|_{b=\hat{\beta}_w} = [\boldsymbol{X}'\boldsymbol{W}(\hat{\beta}_w)\boldsymbol{X}]^{-1}\boldsymbol{X}'\boldsymbol{I}_+(\hat{\beta}_w)\boldsymbol{r}(\hat{\beta}_w). \tag{6.6}$$

Notice that (2.14) can be written as $\boldsymbol{X}'[w\boldsymbol{I}_+(\hat\beta_w) + (1 - \boldsymbol{I}_+(\hat\beta_w))]\boldsymbol{r}(\hat\beta_w) = 0$, which implies a further equality: $\boldsymbol{X}'|\boldsymbol{r}(\hat\beta_w)| = \boldsymbol{X}'[\boldsymbol{I}_+(\hat\beta_w) - (1 - \boldsymbol{I}_+(\hat\beta_w))]\boldsymbol{r}(\hat\beta_w)$ equals $(1+w)\boldsymbol{X}'I_+(\hat\beta_w)\boldsymbol{r}(b)$. Substitution into (6.6) gives

$$\frac{\partial\delta(w,b)}{\partial w}\Big|_{b=\hat\beta_w} = \frac{1}{1+w}[\boldsymbol{X}'\boldsymbol{W}(\hat\beta_w)\boldsymbol{X}]^{-1}\boldsymbol{X}'|\boldsymbol{r}(\hat\beta_w)|. \tag{6.7}$$

The local linear expansion

$$\delta(w+\Delta w, \hat\beta_w + \Delta\beta) \doteq \frac{\partial\delta(w,b)}{\partial w}\Big|_{b=\hat\beta_w}\Delta w + \nabla_b\delta(w,b)|_{b=\hat\beta_w}\Delta\beta \tag{6.8}$$

implies that along the curve $\delta(w,\hat\beta_w) = 0$ in (w,b) space, we have

$$\frac{d\hat\beta_w}{dw} = -\left[\nabla_b\delta(w,b)|_{b=\hat\beta_w}\right]\left[\frac{\partial\delta(w,b)}{\partial w}\Big|_{b=\hat\beta_w}\right]. \tag{6.9}$$

Result (2.20) follows from (6.3) and (6.7).

Some care is needed in interpreting (2.20) for values of w such that one or more of the residuals $r_i(\hat\beta_w)$ equals zero. Although both (6.3) and (6.7) are valid at such w values, the local expansion can fail. At such points formula (2.20) gives different left and right hand derivatives for $d\hat\beta_w/dw$, depending on how the two possible values 1 or w for $W(r_i(b))$ are assigned. This ambiguity made little numerical difference in the baseball and cholostyramine examples, where (2.20) performed its role in algorithm (2.22) excellently.

Remark L. The vector of fitted values corresponding to $\hat\beta_w$, say $\hat\mu_w = \boldsymbol{X}\hat\beta_w$, has derivative

$$\frac{d\hat\mu_w}{dw} = \boldsymbol{X}\frac{d\hat\beta_w}{dw} = \frac{1}{1+w}\boldsymbol{X}[\boldsymbol{X}'\boldsymbol{W}(\hat\beta_w)\boldsymbol{X}]^{-1}\boldsymbol{X}'|\boldsymbol{r}(\hat\beta_w)|. \tag{6.10}$$

For any vector $\boldsymbol{x}_v = \boldsymbol{X}v$, in the column space $\mathcal{L}_{\text{col}}(\boldsymbol{X})$ of $\boldsymbol{X}$, the weighted inner product $\langle \boldsymbol{x}_v, d\hat\mu_w/dw\rangle_w \equiv \boldsymbol{x}_v'\boldsymbol{W}(\hat\beta_w)d\hat\mu_w/dw$ is given by

$$\langle \boldsymbol{x}_v, d\hat\mu_w/dw\rangle_w = \frac{1}{1+w}\boldsymbol{x}_v'|\boldsymbol{r}(\hat\beta_w)|. \tag{6.11}$$

If $\boldsymbol{x}_v$ has all non-negative components, then (6.11) says that $\langle \boldsymbol{x}_v, d\hat\mu_w/dw\rangle_w \geq 0$ for all w. That is, $d\hat\mu_w/dw$ has non-negative inner product with any vector $\boldsymbol{x}_v$ in the intersection of $\mathcal{L}_{\text{col}}(\boldsymbol{X})$ and the positive orthant of $\mathcal{R}^n$. In this sense, $\hat\mu_w$ moves upwards as w increases. If the vector $\boldsymbol{1} = (1,1,\ldots,1)'$ lies in $\mathcal{L}_{\text{col}}(\boldsymbol{X})$, which is usually the case, then (6.11) gives

$$\langle \boldsymbol{1}, d\hat\mu_w/dw\rangle_w / \sum W_i(\hat\beta_w) = \frac{\sum_{i=1}^n W_i(\hat\beta_w)\cdot d\hat\mu_{wi}/dw}{\sum_{i=1}^n W_i(\hat\beta_w)} \geq 0, \tag{6.12}$$

so that the (weighted) average derivative of the predicted values is increasing with w.

It is easy to see that as $w \to \infty$, $\mathcal{L}_w$ approaches a bounding hyperplane to the convex hull of data points, with all of the (x_i, y_i) lying below $\mathcal{L}_w$. Likewise as $w \to 0$, $\mathcal{L}_w$ approaches a lower bounding hyperplane to the convex hull.

Section 4. We now derive the two influence measures $D^2_{w,i}$ and $\tilde{D}^2_{w,i}$ proposed in (4.1). The influence of the ith data point (x_i, y_i) on a statistic of interest like $\hat{\beta}_w$ is, by definition, the differential change in the statistic corresponding to a small measure in the mass, or weight, attached to (x_i, y_i). For a vector of masses $\boldsymbol{m} = (m_1, m_2, \ldots, m_n)'$, all $m_i \geq 0$, define

$$\beta_{\boldsymbol{m}}(w,b) = [\boldsymbol{X}'\{m\boldsymbol{W}(b)\}\boldsymbol{X}]^{-1}X'\{m\boldsymbol{W}(b)\}\boldsymbol{y} \tag{6.13}$$

where $\{m\boldsymbol{W}(b)\}$ is the $n \times n$ diagonal matrix having ith diagonal element $m_i\boldsymbol{W}(r_i(b))$.

If $\boldsymbol{m} = \boldsymbol{1} = (1, 1, \ldots, 1)'$ then $\beta_{\boldsymbol{m}}(w,b) = \beta(w,b)$, as defined in (2.15). For other values of $\boldsymbol{m}$, $\beta_{\boldsymbol{m}}(w,b)$ is the value of $\beta(w,b)$ when the ith data point (x_i, y_i) is treated as if it occurred m_i times in the sample. Corresponding to (2.17) we define $\hat{\beta}_{\boldsymbol{m},w}$ to be the stationary value

$$\hat{\beta}_{\boldsymbol{m},w} = \beta_{\boldsymbol{m}}(w, \hat{\beta}_{\boldsymbol{m},w}), \tag{6.14}$$

so $\hat{\beta}_{\boldsymbol{m},w}$ is the value of b that minimizes $\sum_{i=1}^n m_i Q_w\{r_i(b)\}$, as in (2.6).

A calculation much like (2.27) gives

$$\frac{\partial\beta_{\boldsymbol{m}}(w,b)}{\partial m_i} = [\boldsymbol{X}'\{mW(b)\}\boldsymbol{X}]^{-1}x_i'W_i(b)r_i(\beta_{\boldsymbol{m}}(w,b)), \tag{6.15}$$

$W_i(b) \equiv W(r_i(b))$. In order to smoothly increase the mass attached to the ith data point, define the family of mass vectors

$$\boldsymbol{m}(\epsilon_i) = n(\epsilon_i\boldsymbol{e}_i + (1-\epsilon_i)\boldsymbol{1}/n), \tag{6.16}$$

$\boldsymbol{e}_i$ being the ith unit vector $(0, 0, \ldots, 1, \ldots, 0)'$ with 1 in the ith place. Then

$$\frac{d\boldsymbol{m}(\epsilon_i)}{d\epsilon_i} = n(\boldsymbol{e}_i - \boldsymbol{1}/n). \tag{6.17}$$

Combining (6.15) and (6.17) gives

$$\begin{aligned}&\frac{\partial\beta_{\boldsymbol{m}}(w,b)}{\partial\epsilon_i}\\ &= n[\boldsymbol{X}'\{mW(b)\}\boldsymbol{X}]^{-1}[x_i'W_i(b)r_i(\beta_{\boldsymbol{m}}(w,b)) - \boldsymbol{X}'\boldsymbol{W}(b)\boldsymbol{r}(\beta_{\boldsymbol{m}}(w,b))].\end{aligned} \tag{6.18}$$

In particular

$$\frac{\partial \beta_{\boldsymbol{m}}(w,b)}{\partial \epsilon_i}\Big|_{\epsilon_i=0,b=\hat{\beta}_w} = n[\boldsymbol{X}'\boldsymbol{W}(\hat{\beta}_w)\boldsymbol{X}]^{-1}x_i'W_i(\hat{\beta}_w)r_i(\hat{\beta}_w), \tag{6.19}$$

where we have used $\beta_1(w,b) = \beta(w,b)$, $\beta(w,\hat{\beta}_w) = \hat{\beta}_w$, and $\boldsymbol{X}'\boldsymbol{W}(\hat{\beta}_w)\boldsymbol{r}(\hat{\beta}_w) = 0$, (2.14). It is also true, as in (6.3), that

$$\nabla_b \beta_{\boldsymbol{m}}(w,b)|_{b=\hat{\beta}_{\boldsymbol{m},w}} = 0. \tag{6.20}$$

Together (6.19) and (6.20) show that

$$\frac{\partial \hat{\beta}_{\boldsymbol{m},w}}{\partial \epsilon_i}\Big|_{\epsilon_i=0} = n[\boldsymbol{X}'\boldsymbol{W}(\hat{\beta}_w)\boldsymbol{X}]^{-1}x_i'W_i(\hat{\beta}_w)r_i(\hat{\beta}_w), \tag{6.21}$$

assuming that no $r_i(\hat{\beta}_w) = 0$. Expression (6.21) is the *empirical influence function* of the ith data point on $\hat{\beta}_w$, as defined for instance at (6.16) of Efron (1982).

Now let $\hat{\mu}_{\boldsymbol{m},w}$ indicate the entire vector of predictions $x_i\hat{\beta}_{\boldsymbol{m},w}$ corresponding to $\hat{\beta}_{\boldsymbol{m},w}$,

$$\hat{\mu}_{\boldsymbol{m},w} = \boldsymbol{X}\hat{\beta}_{\boldsymbol{m},w}. \tag{6.22}$$

The empirical influence function of the ith data point on $\hat{\mu}_w \equiv \hat{\mu}_{1,w} = \boldsymbol{X}\hat{\beta}_w$ is the vector

$$\boldsymbol{U}_i \equiv \frac{\partial \hat{\mu}_{\boldsymbol{m},w}}{\partial \epsilon_i}\Big|_{\epsilon_i=0} = n\boldsymbol{X}[\boldsymbol{X}'\boldsymbol{W}(\hat{\beta}_w)\boldsymbol{X}]^{-1}x_i'W_i(\hat{\beta}_w)r_i(\hat{\beta}(w)). \tag{6.23}$$

Comparing (6.23) with (4.1), we see that $D^2_{w,i}$ is a scalar summary statistic for the vector influence $\boldsymbol{U}_i$,

$$D^2_{w,i} = \frac{1}{n^2}||\boldsymbol{U}_i||^2. \tag{6.24}$$

Changing the masses on the data points from $\boldsymbol{1}$ to $\boldsymbol{m}(\epsilon_i) = n(\epsilon_i\boldsymbol{e}_i + (1-\epsilon_i)\boldsymbol{1}/n)$ makes an overall change in the vector of predicted values of magnitude

$$||\hat{\mu}_{\boldsymbol{m}(\epsilon_i),w} - \hat{\mu}_w|| \doteq ||\boldsymbol{U}_i||\epsilon_i = nD_{w,i}\epsilon_i. \tag{6.25}$$

Comparison with formula (4.44a) of Chatterjee and Hadi (1988) shows that $D^2_{w,i}$ follows the same basic definition as *Cook's distance*.

Another motivation for $D^2_{w,i}$ comes from the delta method (or *infinitesimal jackknife* or influence function) estimate of variance (see Efron (1982), formula (6.18)).

The estimated variance of the jth prediction $x_j\hat{\beta}_w$ by any of these equivalent methods is

$$\widehat{\text{var}}_j = \frac{1}{n^2}\sum_{i=1}^{n} U_{ij}^2 \qquad (U_i = (\ldots, U_{ij}, \ldots)'), \tag{6.26}$$

so

$$\sum_{j=1}^{n} D_{w,j}^2 = \sum_{j=1}^{n} \widehat{\text{var}}_j. \tag{6.27}$$

In this sense $100 \cdot D_{w,i}^2 / \sum_{j=1}^{n} D_{w,j}^2$, Figure 3, is the percentage of the total variance of the estimated w-regression plane attributable to the ith data point.

The component of U_i orthogonal to $\mathbf{1}$ is

$$\tilde{U}_i = U_i - \bar{U}_i\mathbf{1} \qquad \left(\bar{U}_i \equiv \sum_{j=1}^{n} U_{ij}/n\right). \tag{6.28}$$

From (6.23),

$$\tilde{U}_i = n\tilde{X}[X'W(\hat{\beta}_w)X]^{-1}x_i'W_i(\hat{\beta}_w)r_i(\hat{\beta}_w), \tag{6.29}$$

and so

$$\tilde{D}_{w,i}^2 = \frac{1}{n^2}\|\tilde{U}_i\|^2 \tag{6.30}$$

by definition (4.1). The version of (6.28) relevant to $\tilde{D}_i$ is

$$\|P^{(1)}(\hat{\mu}_{m(\epsilon_i),w} - \hat{\mu}_w)\| \doteq \|\tilde{U}_i\|\epsilon_i = n\tilde{D}_{w,i}\epsilon_i, \tag{6.31}$$

where $P^{(1)} = I - \mathbf{1}\mathbf{1}'/n$ is the projection matrix orthogonal to $\mathbf{1}$.

Remark M. We see that $\tilde{D}_{w,i}$ is a Cook's distance type of influence function, of (x_i, y_i) on $\hat{\mu}_w$, where we exclude from the influence measure *translations* of the regression plane in the y direction, $\hat{\mu}_w \to \hat{\mu}_w + c\mathbf{1}$. To put it another way, $\tilde{D}_{w,i}$ measures the influence of (x_i, y_i) on the tilt of $\hat{\mu}_w$. This is appealing in our context for two reasons: the tilting of the regression percentiles relative to the OLS plane is perhaps the most interesting output of our analysis; secondly, $\tilde{D}_{w,i}^2$ more accurately portrays the influence of (x_i, y_i) on the regression percentiles $x\hat{\beta}^{(\alpha)}$, as opposed to the influence on the w-regression planes $x\hat{\beta}_w$.

Section 5. The asymptotic efficiency results of Section 5, (5.20), (5.21), are easily derived as special cases of Newey and Powell's (1987) Theorem 3. Newey and Powell's results apply far beyond model (5.11). The reason here for concentrating on (5.11), and especially its normal-theory version (3.12), was to derive specific efficiency comparisons such as those in (5.20)–Table 3.

Here is a heuristic argument supporting result (5.31). We are dealing with the heteroscedastic normal model (3.12), so with $w^{(\alpha)}$ given by formula (2.23), the true $w^{(\alpha)}$–mean of Z is $\beta^0_{w(\alpha)} = z^{(\alpha)} = \Phi^{-1}(\alpha)$ as in Remark B, and the true $w^{(\alpha)}$–regression vector is

$$\beta^{(\alpha)} = \beta_{w(\alpha)} = \gamma + \tau z^{(\alpha)}, \tag{6.32}$$

as in (5.13). The notation $\beta^{(\alpha)} = \beta_{w(\alpha)}$ is permissible here because the plane $\mathcal{L}_{w(\alpha)} = \{y = x\beta_{w(\alpha)}\}$ is the true 100αth regression percentile $\mathcal{L}^{(\alpha)}$ in model (3.12). However we must distinguish between $\hat\beta_{w(\alpha)}$, the sample w-regression vector with $w = w^{(\alpha)}$, and $\hat\beta^{(\alpha)}$, the vector giving the sample 100αth regression percentile $\hat{\mathcal{L}}^{(\alpha)}$, (2.9). Usually $\hat{\mathcal{L}}_{w(\alpha)}$, (2.8), will not have exactly proportion α of the n data points (x_i, y_i) lying beneath it. An important part of variance formula (5.3) comes from the difference between $\hat\beta_{w(\alpha)}$ and $\hat\beta^{(\alpha)}$.

A "one-step" approximation for $\hat\beta_{w(\alpha)}$ is

$$\begin{aligned}\hat\beta_{w(\alpha)} - \beta^{(\alpha)} &\doteq \beta(w^{(\alpha)}, \beta^{(\alpha)}) - \beta^{(\alpha)} \\ &= \left[\frac{\boldsymbol{X}'\boldsymbol{W}(\beta^{(\alpha)})\boldsymbol{X}}{n}\right]^{-1}\left[\frac{1}{n}\boldsymbol{X}'\boldsymbol{W}(\beta^{(\alpha)})\boldsymbol{r}(\beta^{(\alpha)})\right].\end{aligned} \tag{6.33}$$

Formula (2.19), with $b = \beta^{(\alpha)}$, shows that (6.33) is the usual one-step approximation for finding the minimizer of $S_{w(\alpha)}(b)$, and will err by only $o_p(n^{-\frac{1}{2}})$ under reasonable regularity conditions. Assuming that the (x_i, y_i) pairs are i.i.d. observations, as in Newey and Powell's Assumption 1, then

$$[\boldsymbol{X}'\boldsymbol{W}(\beta^{(\alpha)})\boldsymbol{X}/n]^{-1} \rightarrow E\{X'X\}^{-1}/D^{(\alpha)} \quad \left(D^{(\alpha)} \equiv 1 + (w^{(\alpha)} - 1)(1-\alpha)\right) \tag{6.34}$$

in probability as $n \to \infty$. We can then use (6.33) to approximate $\hat\beta_{w(\alpha)} - \beta^{(\alpha)}$ in the influence function form

$$\hat\beta_{w(\alpha)} - \beta^{(\alpha)} \doteq \frac{1}{n}\sum_{i=1}^{n} U_i^\beta \qquad \left(U_i^\beta \equiv \frac{E\{X'X\}^{-1}}{D^{(\alpha)}} x_i' W(r_i(\beta^{(\alpha)})) r_i(\beta^{(\alpha)})\right). \tag{6.35}$$

Next let

$$\hat\alpha_0 = \sum_{i=1}^{n} I_-(r_i(\beta^{(\alpha)}))/n \qquad \left(I_-(r) = \begin{cases} 1 & \text{if } r \le 0 \\ 0 & \text{if } r > 0 \end{cases}\right), \tag{6.36}$$

so $\hat\alpha_0$ is the proportion of data points lying below the true 100αth regression percentile $\mathcal{L}_{w(\alpha)}$. Then

$$\hat\alpha_0 - \alpha = \frac{1}{n}\sum_{i=1}^{n} U_i^\alpha \qquad \left(U_i^\alpha \equiv I_-(r_i(\beta^{(\alpha)})) - \alpha\right). \tag{6.37}$$

The $(p+1) \times (p+1)$ covariance matrix of (U_i^β, U_i^α) is

$$\operatorname{cov}\begin{pmatrix} U_i^\beta \\ U_i^\alpha \end{pmatrix} = \begin{pmatrix} a_1^{(\alpha)} m_1 & -c^{(\alpha)}\tau' \\ -\tau c^{(\alpha)} & \alpha(1-\alpha) \end{pmatrix}, \tag{6.38}$$

as in (5.31). Formula (6.38) depends on these facts: $r_i(\beta^{(\alpha)}) = (x_i\tau)(Z_i - z^{(\alpha)})$ (from (5.11) and (6.32)); $U_i^\beta = E\{x'x\}^{-1}x_i'x_i\tau W(Z_i - z^{(\alpha)})(Z_i - z^{(\alpha)})/D^{(\alpha)}$; $Z_i \sim$ N(0,1), independent of x_i; $EW(Z_i - z^{(\alpha)})(Z_i - z^{(\alpha)}) = 0$ (from the population analogue of (2.14)); $EW(Z_i - z^{(\alpha)})(Z_i - z^{(\alpha)}) \cdot I_-(Z_i - z^{(\alpha)}) = E(Z_i - z^{(\alpha)})I_-(Z_i - z^{(\alpha)}) = -[\phi(z^{(\alpha)}) + \alpha]$; and $EU_i^\beta U_i^{\beta'} = \{Ex'x\}^{-1}E\{x'(x\tau)^2x\}\{Ex'x\}^{-1}E\{W(Z - z^{(\alpha)})(Z - z^{(\alpha)})\}^2/D^{(\alpha)^2} = a_1^{(\alpha)}m_1$.

Let $P\{b\}$ equal the probability mass, under model (3.12), lying below the plane $\{y = xb\}$. If b_1 and b are approaching $\beta^{(\alpha)}$ at rate $O_p(n^{-\frac{1}{2}})$, we compute

$$P\{b_1\} - P\{b\} \doteq \int_{\mathcal{R}^p} x(b_1 - b)\phi(z^{(\alpha)})dx = x^0(b_1 - b)\phi(z^{(\alpha)}), \tag{6.39}$$

with $x^0 = E\{x/x\tau\}$ as in (5.31), the relative error in (6.39) being $O_p(n^{-\frac{1}{2}})$. Choosing $b_1 = \hat{\beta}_{w(\alpha)}$ and $b = \beta^{(\alpha)}$ gives

$$P\{\hat{\beta}_{w(\alpha)}\} - \alpha \doteq x^0(\hat{\beta}_{w(\alpha)} - \beta^{(\alpha)})\phi(z^{(\alpha)}). \tag{6.40}$$

Then

$$\hat{\alpha} - \alpha \doteq (\hat{\alpha}_0 - \alpha) + x^0(\hat{\beta}_{w(\alpha)} - \beta^{(\alpha)})\phi(z^{(\alpha)}), \tag{6.41}$$

where $\hat{\alpha}$ equals the proportion of data points lying below $\hat{\mathcal{L}}_{w(\alpha)} = \{y = x\hat{\beta}_{w(\alpha)}\}$.

We need two more relationships to complete the verification of (5.31),

$$\frac{d\beta_{w(\alpha)}}{d\alpha} \doteq \frac{\tau}{\phi(z^{(\alpha)})} \tag{6.42}$$

and

$$dP\{\hat{\beta}_{w(\alpha)}\}/d\alpha \doteq 1. \tag{6.43}$$

Result (6.42) follows by taking the limiting value as $n \to \infty$ of (2.20), or by a continuity argument using (6.32). Then (6.43) is derived from (6.39) and (6.42), and the fact that $x^0\tau = E\{x\tau/x\tau\} = 1$.

By definition, the 100αth sample regression percentile is $\hat{\beta}_{w(\alpha_1)}$, where α_1 is chosen so that the proportion of the data points lying below $\hat{\mathcal{L}}_{w(\alpha_1)} = \{y = x\hat{\beta}_{w(\alpha_1)}\}$ equals α. From (6.43) and (6.41) we see that

$$\alpha_1 - \alpha \doteq -(\hat{\alpha} - \alpha) \doteq -(\hat{\alpha}_0 - \alpha) - x^0(\hat{\beta}_{w(\alpha)} - \beta^{(\alpha)})\phi(z^{(\alpha)}). \tag{6.44}$$

Moreover (6.42) gives

$$\hat{\beta}_{w(\alpha_1)} - \hat{\beta}_{w(\alpha)} \doteq \frac{\tau}{\phi(z^{(\alpha)})}(\alpha_1 - \alpha). \tag{6.45}$$

Writing $\hat{\beta}_{w(\alpha_1)} - \beta^{(\alpha)} = (\hat{\beta}_{w(\alpha_1)} - \hat{\beta}_{w(\alpha)}) + (\hat{\beta}_{w(\alpha)} - \beta^{(\alpha)})$, we have

$$\hat{\beta}_{w(\alpha_1)} - \beta^{(\alpha)} \doteq (I_p - \tau x^0, -\tau/\phi(z^{(\alpha)}))\begin{pmatrix} \hat{\beta}_{w(\alpha)} - \beta^{(\alpha)} \\ \hat{\alpha} - \alpha \end{pmatrix}. \tag{6.46}$$

This verifies (5.31), by combining (6.35)–(6.38) with (6.46).

Remark N. Suppose that we use maximum likelihood to estimate $\beta^{(\alpha)}$ in model (3.12), but that we empirically calibrate the $\tilde{\beta}^{(\alpha)}$; that is we estimate $\beta^{(\alpha)}$ by $\tilde{\beta}^{(\alpha_1)}$, where α_1 is chosen so that proportion α of the data points lie below $\{y = x\tilde{\beta}^{(\alpha_1)}\}$. Then calculations like those above give

$$\text{AVAR}(\tilde{\beta}^{(\alpha_1)}) = (I_p - \tau x^0, -\tau/\phi(z^{(\alpha)}))\begin{pmatrix} a_0^{(\alpha)} m_0 & -q^{(\alpha)} \\ -q^{(\alpha)\prime} & \alpha(1-\alpha) \end{pmatrix}\begin{pmatrix} I_p - x^{0\prime}\tau' \\ -\tau'/\phi(z^{(\alpha)}) \end{pmatrix}, \tag{6.47}$$

where $a_0^{(\alpha)} = 1 + z^{(\alpha)^2}/2$ as in (5.8), so $a_0^{(\alpha)} m_0 = \text{AVAR}(\tilde{\beta}^{(\alpha)})$, and $q^{(\alpha)} = m_0 x^{0\prime}[\phi(z^{(\alpha)})(1 + z^{(\alpha)^2}/2)]$. The asymptotic relative efficiency of $\hat{\beta}^{(\alpha)}$ might better be defined with respect to $\tilde{\beta}^{(\alpha_1)}$ than $\tilde{\beta}^{(\alpha)}$, since both $\hat{\beta}^{(\alpha)}$ and $\tilde{\beta}^{(\alpha_1)}$ are empirically calibrated. This form of the ARE can be computed from (5.31) and (6.47).

To verify (5.25), let $\boldsymbol{r}^0 = \boldsymbol{y} - \boldsymbol{X}\gamma = \boldsymbol{r}(\gamma)$, the true residual vector, so $r_i^0 = (x_i\tau)Z_i$. Then $\hat{\gamma} - \gamma = ((\boldsymbol{X}'\boldsymbol{X}^{-1})\boldsymbol{X}'\boldsymbol{r}^0)$ and $T = c(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'|\overset{\perp}{\boldsymbol{P}}\boldsymbol{r}^0|$, where $\overset{\perp}{\boldsymbol{P}} \equiv (\boldsymbol{I}_n - \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}')$. If $Z_i \sim$ N(0,1), (or any other distribution symmetric about 0), then $\boldsymbol{Z} = (Z_1, \dots, Z_n)$ and $-\boldsymbol{Z}$ are equally likely. But $\boldsymbol{Z} \to -\boldsymbol{Z}$ implies $\boldsymbol{r}^0 \to -\boldsymbol{r}^0$ and so $(\hat{\gamma} - \gamma) \to -(\hat{\gamma} - \gamma)$ while $T \to T$. This shows that $\text{cov}(\hat{\gamma}, T) = 0$.

References

Aigner, D., Amemiya, T. and Poirier, D. (1976). On the estimation of production frontiers: maximum likelihood estimation of the parameters of a discontinuous density function. *Internat. Econom. Rev.* **17**, 372–396.

Bassett, G and Koenker, R. (1982b). An empirical quantile function for linear models with i.i.d. errors. *J. Amer. Statist. Assoc.* **77**, 407–415.

Bickel, P. J. (1973). On some analogues to linear combinations of order statistics in the linear model. *Ann. Statist.* **1**, 597–616.

Breckling, J. and Chambers, R. (1988). *M*-quantiles. *Biometrika* **75**, 761–771.

Chatterjee, S. and Hadi, A. S. (1988). *Sensitivity Analysis in Linear Regression.* John Wiley, New York.

Efron, B. (1982). The jackknife, the bootstrap and other resampling plans. *SIAM CBMS Monograph* **38**.

Glejser, H. (1969). A new test for heteroskedasticity. *J. Amer. Statist. Assoc.* **64**, 316–323.

Huber, P. J. (1981). *Robust Statistics.* John Wiley, New York.

Kendall, M. and Stuart, A. (1958). *The Advanced Theory of Statistics,*Vol.1. Charles Griffin, London.

Koenker, R. and Bassett, G. (1978). Regression quantiles. *Econometrica* **46**, 33–50.

Koenker, R. and Bassett, G. (1982a). Robust tests for heteroscedasticity based on regression quantiles. *Econometrica* **50**, 43–61.

Lipid Research Clinics Program (1984). The Lipid Research Clinics Coronary Primary Prevention Trial Results (Parts I and II). *J. Amer. Med. Assoc.* **251**, 351–374.

Newey, W. K. and Powell, J. L. (1987). Asymmetric least squares estimation and testing. *Econometrica* **55**, 819–847.

Ruppert, D. and Carroll, R. J. (1980). Trimmed least squares estimation in the linear model. *J. Amer. Statist. Assoc.* **75**, 828–838.

Zellner, A. (1986). Bayesian estimation and prediction using asymmetric loss functions. *J. Amer. Statist. Assoc.* **81**, 446–451.

Department of Statistics, Stanford University, Stanford, CA 94305, U.S.A.

(Received August 1989; accepted June 1990)