Bringing Order to Outlier Diagnostics in Regression Models

D. R. Jensen and D. E. Ramirez
Virginia Polytechnic Institute and State University
and University of Virginia
der@virginia.edu
http://www.math.virginia.edu/˜der/home.html

0.1 INTRODUCTION

Consider the model

$$\{Y_i = \beta_0 + \beta_1 X_{i1} + \cdots + \beta_k X_{ik} + \varepsilon_i \, (1 \leq i \leq n)\}$$

relating a response Y_i to fixed regressors $\{X_{i1}, X_{i2}, \ldots, X_{ik}\}$ through $p = k + 1$ unknown parameters $\boldsymbol{\beta} = [\beta_0, \beta_1, \ldots, \beta_k]'$.

THE FULL MODEL $\mathbf{Y} = \mathbf{X}_0\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ is partitioned based on the proposed outliers $\{Y_i : i \in I\}$ of cardinality r.

$$\begin{bmatrix} \mathbf{Y}_1 \\ \mathbf{Y}_2 \end{bmatrix} = \begin{bmatrix} \mathbf{X} \\ \mathbf{Z} \end{bmatrix} \boldsymbol{\beta} + \begin{bmatrix} \boldsymbol{\varepsilon}_1 \\ \boldsymbol{\varepsilon}_2 \end{bmatrix} \quad \begin{matrix} ((n-r) \times 1) \\ (r \times 1) \end{matrix}$$

with least-squares estimator $\hat{\boldsymbol{\beta}}$ $(p \times 1)$. All matrices are assume to be of full rank.

THE REDUCED MODEL $\mathbf{Y}_1 = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}_1$ with the proposed outliers deleted

$$\mathbf{Y}_1 = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}_1$$

with least-squares estimator $\hat{\boldsymbol{\beta}}_I$.

0.2 STANDARD OUTLIER DIAGNOSTICS

0.2.1 GROUP I DIAGNOSTICS (t-like statistics)

- STUDENTIZED DELETED RESIDUAL TEST (Gold Standard)

$$\frac{y_i - \hat{y}_{(i)i}}{s_{(i)}\sqrt{1 + \mathbf{x}_i(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_i'}} = \frac{y_i - \hat{y}_{(i)i}}{s_{(i)}\sqrt{1 + h_{ii}/(1-h_{ii})}} \sim t(n-p-1) \tag{1}$$

- THE EXTERNALLY STUDENTIZED RESIDUALS ($R-$STUDENT statistic)

Easy to compute

$$t_i = \frac{y_i - \hat{y}_i}{s_{(i)}\sqrt{1 - \mathbf{x}_i(\mathbf{X}_0'\mathbf{X}_0)^{-1}\mathbf{x}_i'}} = \frac{y_i - \hat{y}_i}{s_{(i)}\sqrt{1-h_{ii}}} \sim t(n-p-1) \tag{2}$$

- DFFITS

$$\begin{aligned} DFFITS_i &= \frac{\hat{y}_i - \hat{y}_{(i)i}}{s_{(i)}\sqrt{h_{ii}}} \\ DFFITS_i &= t_i\sqrt{\frac{h_{ii}}{1-h_{ii}}} \\ \sqrt{\frac{h_{ii}}{1-h_{ii}}} &= \begin{cases} 0.5774 \text{ with } h_{ii} = 0.25 \\ 1.0000 \text{ with } h_{ii} = 0.50 \\ 4.3589 \text{ with } h_{ii} = 0.95 \end{cases} \end{aligned}$$

- DFBETAS

$$DFBETAS_{(i)j} = \frac{\widehat{\beta}_j - \widehat{\beta}_{(i)j}}{s_{(i)}\sqrt{(\mathbf{X}_0'\mathbf{X}_0)^{-1}_{jj}}} \tag{3}$$

$$DFBETAS_{(i)j} = t_i \left[\frac{[(\mathbf{X}_0'\mathbf{X}_0)^{-1}\,\mathbf{X}']_{ji}}{\sqrt{(\mathbf{X}_0'\mathbf{X}_0)^{-1}_{jj}\,(1-h_{ii})}} \right]$$

USEFUL RELATIONSHIPS

$$\begin{aligned} \widehat{\boldsymbol{\beta}}-\widehat{\boldsymbol{\beta}}_{(i)} &= \frac{(\mathbf{X}_0'\mathbf{X}_0)^{-1}\mathbf{x}_i(y_i-\widehat{y}_i)}{1-h_{ii}} \\ \frac{(y_i-\widehat{y}_i)^2}{1-h_{ii}} &= (n-p)s^2-(n-p-1)s^2_{(i)} \end{aligned}$$

STANDARD RECOMMENDATION (Myers (1990)): "As we indicated earlier, these two measures (DFFITS and DFBETAS) are t-like. Surely any analyst who is familiar with the concept of a standard error knows that if $DFBETAS_{(i)j}$ exceeds 2.0 in magnitude, the influence of the data point in unquestioned."

Note: $\mathbf{Y}-\widehat{\mathbf{Y}}$ or $\widehat{\boldsymbol{\beta}}-\widehat{\boldsymbol{\beta}}_{(i)}$ are OUTLIER measures while h_{ii} and $1/(1-h_{ii})$ are measures of INFLUENCE.

CONCLUSION: All Group I diagnostics yield the same p-values: Jensen and Ramirez (1996) # 1 and 2, Jensen (1998, 2000) # 2, 3 and 3 and LaMotte (1999) # 1, 2, 3 and 3

0.2.2 GROUP II DIAGNOSTICS (heuristics based on s_I^2/s^2)

The ratio s_I^2/s^2 is the Neyman-Pearson likelihood ratio test statistics.

- R-FISHER parallels the R-Student statistic (Jensen (1999, 2001) (Gold Standard)

$$\begin{aligned} F_I &= \frac{\mathbf{e}_I'(\mathbf{I}_r - \mathbf{Z}(\mathbf{X}_0'\mathbf{X}_0)^{-1}\mathbf{Z}')^{-1}\mathbf{e}_I}{rs_I^2} = \frac{\mathbf{e}_I'(\mathbf{I}_r - \mathbf{H}_{II})^{-1}\mathbf{e}_I}{rs_I^2} \quad (4) \\ &= \frac{((n-p)s^2 - (n-p-r)s_I^2)/r}{s_I^2} \sim F_{r,n-p-r} \end{aligned}$$

where

$$\mathbf{X}(\mathbf{X}'\mathbf{X})\mathbf{X}' = \mathbf{H} = \begin{bmatrix} \mathbf{H}_{00} & \mathbf{H}_{0I} \\ \mathbf{H}_{I0} & \mathbf{H}_{II} \end{bmatrix}$$

- MEAN SHIFT OUTLIER MODEL is based on

$$\begin{bmatrix} \mathbf{Y}_1 \\ \mathbf{Y}_2 \end{bmatrix} = \begin{bmatrix} \mathbf{X} \\ \mathbf{Z} \end{bmatrix} \boldsymbol{\beta} + \begin{bmatrix} \mathbf{0} \\ \mathbf{I}_r \end{bmatrix} \boldsymbol{\delta} + \begin{bmatrix} \boldsymbol{\varepsilon}_1 \\ \boldsymbol{\varepsilon}_2 \end{bmatrix}$$

Outliers can be checked jointly with model building.
Outliers are tested based on the null hypothesis $H_0 : \boldsymbol{\delta} = \mathbf{0}$ with

$$\begin{aligned} F_I &= \frac{(RSS(H_0) - RSS(H_a))/r}{RSS(H_a)/(n-p-r)} \quad (5) \\ &= \frac{((n-p)s^2 - (n-p-r)s_I^2)/r}{s_I^2} \sim F_{r,n-p-r} \end{aligned}$$

With

$$F_{(i)} = t_i^2$$

- OUT_I $(r \geq 1)$ Barnett and Lewis (1984)

$$\begin{aligned} OUT_I &= 1 - \frac{s_I^2}{s^2} && (6) \\ -\frac{1}{n-p-r} &\leq OUT_I = \frac{r(F_I - 1)}{rF_I + (n-p-r)} \leq 1 \end{aligned}$$

- COVRATIO $(r \geq 1)$

$$\begin{aligned} CR_{(i)} &= \frac{|s_{(i)}^2(\mathbf{X}'\mathbf{X})^{-1}|}{|s^2(\mathbf{X}_0'\mathbf{X}')^{-1}|} = \frac{1}{1-h_{ii}}\left(\frac{s_{(i)}^2}{s^2}\right)^p && (7) \\ 0 &\leq CR_i = \frac{\left(\frac{n-p}{t_i^2+(n-p-1)}\right)^p}{1-h_{ii}} \leq \frac{\left(\frac{n-p}{n-p-1}\right)^p}{1-h_{ii}} \\ \frac{1}{1-h_{ii}} &= \begin{cases} 1.3333 \text{ with } h_{ii} = 0.25 \\ 2.0000 \text{ with } h_{ii} = 0.50 \\ 20.0000 \text{ with } h_{ii} = 0.95 \end{cases} && (8) \end{aligned}$$

- AP_i of Andrews and Pregibon (1978) $(r \geq 1)$

$$\begin{aligned} AP_i &= 1 - \frac{n-p-1}{n-p}\frac{s_{(i)}^2|(\mathbf{X}'\mathbf{X})^{-1}|}{s^2|(\mathbf{X}_0'\mathbf{X})^{-1}|} && (9) \\ h_{ii} &\leq AP_i = \frac{t_i^2 + (n-p-1)h_{ii}}{t_i^2 + (n-p-1)} \leq 1 \end{aligned}$$

- $FVARATIO$ of Belsley, Kuh and Welsch (1980) ($r \geq 1$)

$$FV_i = \left[\frac{1}{1-h_{ii}}\right] \frac{s^2_{(i)}}{s^2} \tag{10}$$

$$0 \leq FV_i = \frac{(n-p)/(t_i^2+n-p-1)}{1-h_{ii}} \leq \frac{(n-p)/(n-p-1)}{1-h_{ii}}$$

$$\frac{1}{1-h_{ii}} = \begin{cases} 1.3333 \text{ with } h_{ii} = 0.25 \\ 2.0000 \text{ with } h_{ii} = 0.50 \\ 20.0000 \text{ with } h_{ii} = 0.95 \end{cases}$$

USEFUL RELATIONSHIPS

$$\widehat{\boldsymbol{\beta}} - \widehat{\boldsymbol{\beta}}_I = (\mathbf{X}_0'\mathbf{X}_0)^{-1}\mathbf{Z}'(I-H_{II})^{-1}(Y_I-\widehat{Y}_I)$$

$$(Y_I-\widehat{Y}_I)'(I-H_{II})^{-1}(Y_I-\widehat{Y}_I) = (n-p)s^2-(n-p-1)s_I^2$$

CONCLUSION: All Group II statistics yield the same p-values: Jensen (1999, 2001) # 4, 6, 7, 9, 10, Jensen and Ramirez (1998) # 5 and LaMotte (1999) # 7

0.2.3 GROUP III DIAGNOSTICS (Distance measures)

- COOK'S D_I STATISTICS (1977) take the form

$$D_I(\hat{\boldsymbol{\beta}}, \mathbf{M}, c\hat{\sigma}^2) = \frac{(\hat{\boldsymbol{\beta}}_I - \hat{\boldsymbol{\beta}})^{'} \mathbf{M} (\hat{\boldsymbol{\beta}}_I - \hat{\boldsymbol{\beta}})}{c\hat{\sigma}^2}$$

where $\mathbf{M}(p \times p)$ is non-negative definite and $\hat{\sigma}^2$ is some estimator for the variance σ^2 and c is a user-defined constant.

Standard Choices

- C_i of Cook (1977): $\hat{\sigma}^2 = s^2, c = p$, and $\mathbf{M} = \mathbf{X}_0^{'}\mathbf{X}_0 = \sigma^2 cov(\hat{\boldsymbol{\beta}})^{-1}$

$$\begin{aligned} C_i &= \frac{(\hat{\boldsymbol{\beta}}_{(i)} - \hat{\boldsymbol{\beta}})\mathbf{X}_0^{'}\mathbf{X}_0(\hat{\boldsymbol{\beta}}_{(i)} - \hat{\boldsymbol{\beta}})}{ps^2} = \frac{|\widehat{Y}_{(i)} - \widehat{Y}|^2}{ps^2} \qquad (11) \\ C_i &= \left[\frac{(n-p)h_{ii}}{p(1-h_{ii})}\right] \frac{t_i^2}{t_i^2 + (n-p-1)} \\ \frac{h_{ii}}{1-h_{ii}} &= \begin{cases} 0.3333 \text{ with } h_{ii} = 0.25 \\ 1.0000 \text{ with } h_{ii} = 0.50 \\ 19.0000 \text{ with } h_{ii} = 0.95 \end{cases} \end{aligned}$$

- WK_i of Welsch and Kuh (1977): $\hat{\sigma}^2 = s_I^2, c = p$ and $\mathbf{M} = \mathbf{X}_0^{'}\mathbf{X}_0 = \sigma^2 cov(\hat{\boldsymbol{\beta}})^{-1}$

$$\begin{aligned} WK_i &= \frac{(\hat{\boldsymbol{\beta}}_{(i)} - \hat{\boldsymbol{\beta}})\mathbf{X}_0^{'}\mathbf{X}_0(\hat{\boldsymbol{\beta}}_{(i)} - \hat{\boldsymbol{\beta}})}{ps_{(i)}^2} = \frac{|\widehat{Y}_{(i)} - \widehat{Y}|^2}{ps_{(i)}^2} \qquad (12) \\ WK_i &= \left[\frac{h_{ii}}{p(1-h_{ii})}\right] t_i^2 \\ \frac{h_{ii}}{1-h_{ii}} &= \begin{cases} 0.3333 \text{ with } h_{ii} = 0.25 \\ 1.0000 \text{ with } h_{ii} = 0.50 \\ 19.0000 \text{ with } h_{ii} = 0.95 \end{cases} \end{aligned}$$

Our recommendation

$$\begin{aligned} D_I(\hat{\beta}, \mathbf{X}_0^{'}\mathbf{X}_0, rs_I^2) &= \frac{(\hat{\boldsymbol{\beta}}_I - \hat{\boldsymbol{\beta}})^{'}\mathbf{X}_0^{\prime}\mathbf{X}_0(\hat{\boldsymbol{\beta}}_I - \hat{\boldsymbol{\beta}})}{rs_I^2} && (13) \\ &\sim F_r(t; \gamma_1^2, \cdots, \gamma_r^2; n-p-r) && (14) \end{aligned}$$

$\{\gamma_1^2 \geq \cdots \geq \gamma_r^2 > 0\}$ are the ordered eigenvalues of $\mathbf{Z}(\mathbf{X}^{'}\mathbf{X})^{-1}\mathbf{Z}^{\prime}$
STANDARD RECOMMENDATION: "Cook's D ... is an F-like statistic with degrees of freedom p and $n-p$."

The critical value has been suggested by Cook to be $F_{p,n-p}(0.50)$.

- W_i of Welsch (1982): $\hat{\sigma}^2 = s_I^2, c = p$, and $\mathbf{M} = \mathbf{X}^{'}\mathbf{X} = \sigma^2 cov(\hat{\boldsymbol{\beta}}_I)^{-1}$

$$\begin{aligned} W_i &= \frac{(\hat{\boldsymbol{\beta}}_{(i)} - \hat{\boldsymbol{\beta}})\mathbf{X}^{'}\mathbf{X}(\hat{\boldsymbol{\beta}}_{(i)} - \hat{\boldsymbol{\beta}})}{ps_{(i)}^2} && (15) \\ W_i &= \left[\frac{h_{ii}}{p}\right] t_i^2 \\ h_{ii} &= \begin{cases} 0.25 \text{ with } h_{ii} = 0.25 \\ 0.50 \text{ with } h_{ii} = 0.50 \\ 0.95 \text{ with } h_{ii} = 0.95 \end{cases} \end{aligned}$$

Our recommendation $\hat{\sigma}^2 = s_I^2, c = r$ and $\mathbf{M} = \mathbf{X}^{'}\mathbf{X} = \sigma^2 cov(\hat{\boldsymbol{\beta}}_I)^{-1}$

$$\begin{aligned} D_I(\hat{\beta}, \mathbf{X}^{'}\mathbf{X}, rs_I^2) &= \frac{(\hat{\boldsymbol{\beta}}_I - \hat{\boldsymbol{\beta}})^{'}\mathbf{X}^{\prime}\mathbf{X}(\hat{\boldsymbol{\beta}}_I - \hat{\boldsymbol{\beta}})}{rs_I^2} \\ &\sim F_r(t; \lambda_1, \cdots, \lambda_r; n-p-r) \end{aligned}$$

$\{\lambda_1 \geq \cdots \geq \lambda_r > 0\}$ are the ordered eigenvalues of $\mathbf{Z}(\mathbf{X}_0^{'}\mathbf{X}_0)^{-1}\mathbf{Z}^{\prime}$, the canonical subset leverages.

- D_I of Jensen and Ramirez (1998) (Gold Standard) Normalized Cook Statistic
  $\hat{\sigma}^2 = s_I^2, c = r$, and $\mathbf{M} = ((\mathbf{X}^{'}\mathbf{X})^{-1} - (\mathbf{X}_0^{\prime}\mathbf{X}_0)^{-1})^{+} = \sigma^2 cov(\hat{\boldsymbol{\beta}}_I - \hat{\boldsymbol{\beta}})^{+}$

$$\begin{aligned} D_I &= \frac{(\hat{\boldsymbol{\beta}}_I - \hat{\boldsymbol{\beta}})((\mathbf{X}^{\prime}\mathbf{X})^{-1} - (\mathbf{X}_0^{\prime}\mathbf{X}_0)^{-1})^{+}(\hat{\boldsymbol{\beta}}_I - \hat{\boldsymbol{\beta}})}{r s_I^2} \\ D_I &= F_I \sim F_{r,n-p-r} \\ D_i &= t_i^2 \end{aligned}$$

CONCLUSION:: All Group III statistics yield the same p-values when $r = 1$: Jensen (1999, 2001) # 11, 12, and 15, Jensen and Ramirez (1996) # 11 and LaMotte (1999) # 11.

For $r > 1$, the p-values are different. Their distribution is now known.

Notation:

$\{\gamma_1^2 \geq \cdots \geq \gamma_r^2\}$ are the ordered eigenvalues of $\mathbf{Z}(\mathbf{X}^{'}\mathbf{X})^{-1}\mathbf{Z}^{\prime}$

$\{\lambda_1 \geq \cdots \geq \lambda_r\}$ are the ordered eigenvalues of $\mathbf{Z}(\mathbf{X}_0^{'}\mathbf{X}_0)^{-1}\mathbf{Z}^{\prime}$

$$\begin{aligned} 0 &< \lambda_i = \gamma_i^2/(1+\gamma_i^2) = h_{ii} < 1, 1 \leq i \leq r \\ \gamma_i^2 &= \begin{cases} 0.3333 \text{ with } h_{ii} = 0.25 \\ 1.0000 \text{ with } h_{ii} = 0.50 \\ 19.0000 \text{ with } h_{ii} = 0.95 \end{cases} \end{aligned}$$

SINGLE OUTLIER THEOREM (Jensen and Ramirez 1996): When $r = 1$, the following tests are all equivalent.

$$\begin{aligned}
\frac{\mathbf{y}_i - \hat{\mathbf{y}}_{(i)}}{s_{(i)}\sqrt{1 + \mathbf{x}_i(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_i'}} &\sim t(n-p-1) \\
\frac{\mathbf{y}_i - \hat{\mathbf{y}}_i}{s_{(i)}\sqrt{1 - \mathbf{x}_i(\mathbf{X}_0'\mathbf{X}_0)^{-1}\mathbf{x}_i'}} &\sim t(n-p-1) \\
F_i &\sim F(1, n-p-1) \\
D_i(\hat{\boldsymbol{\beta}}, \mathbf{X}_0^{'}\mathbf{X}_0, s_{(i)}^2) &\sim \gamma_1^2 F(1, n-p-1) \\
D_i(\hat{\boldsymbol{\beta}}, \mathbf{X}^{'}\mathbf{X}, s_{(i)}^2) &\sim \lambda_1 F(1, n-p-1) \\
D_i(\hat{\boldsymbol{\beta}}, \sigma^2 cov(\hat{\boldsymbol{\beta}}_i - \hat{\boldsymbol{\beta}})^+, s_{(i)}^2) &\sim F(1, n-p-1)
\end{aligned}$$

GENERALIZED F DISTRIBUTION (Jensen and Ramirez, 1998) Suppose that elements of $\mathbf{U} = [U_1, \cdots, U_r]'$ are independent $\{N_1(\omega_i, 1); 1 \leq i \leq r\}$ random variables; let $\{\alpha_1, \cdots, \alpha_r\}$ be non-increasing positive weights; and identify

$$T = \alpha_1 U_1^2 + \cdots \alpha_r U_r^2$$

If $L(V) = \chi^2(\nu)$ independently of U, then the cdf of

$$W = \frac{T/r}{V/\nu}$$

is denoted by

$$F_r(t; \alpha_1, \cdots, \alpha_r; \boldsymbol{\omega}; \nu)$$

THEOREM (Ramirez and Jensen (1991)) For $\boldsymbol{\omega} = \mathbf{0}$, the pdf of the central generalized F distribution $(T/r)/(V/\nu) = W$ has the representation in terms of central F distributions as

$$\sum_{i=0}^{\infty} \frac{c_i}{\delta} \frac{r}{r+2i} f_F(\frac{r}{r+2i} \frac{w}{\delta}; r+2i, \nu) \tag{16}$$

with f_F as the pdf of $F(\nu_1, \nu_2)$, and whose coefficients $\{c_i\}$ are defined recursively with $0 < \delta < \alpha_r$.

Ramirez (2000) gave the Fortran code for computing p-values for $F_r(t; \alpha_1, \cdots, \alpha_r; \nu)$.

Dunkl and Ramirez (2001) have developed a fast algorithm for computing the p-values for W without having to integrate the density function in terms of Lauricella $F_D^{(r)}$ functions.

STOCHASTIC BOUNDS (Jensen and Ramirez (1991)) for F_r are

$$F_r(t; \alpha_1; \nu) \leq F_r(t; \alpha_1, \cdots, \alpha_r; \nu) \leq F_r(t; \alpha^*; \nu)$$

with α^* the geometric mean of $\{\alpha_1, \cdots, \alpha_r\}$.

THEOREM (Jensen and Ramirez (200x)) For $\boldsymbol{\omega} \neq \mathbf{0}$, the pdf of the non-central generalized F distribution $(T/r)/(V/\nu) = W$ has the representation in terms of central F distributions as

$$\sum_{i=0}^{\infty} \frac{c_i}{\delta} \frac{r}{r+2i} f_F\left(\frac{r}{r+2i} \frac{w}{\delta}; r+2i, \nu\right)$$

with f_F as the pdf of $F(\nu_1, \nu_2)$, and whose coefficients $\{c_i\}$ are defined recursively with $0 < \delta < \alpha_r$ and where $\{c_i\}$ are functions of $\boldsymbol{\omega}$.

NORMAL THEORY RESULTS (Jensen and Ramirez (1998)): If $L(\mathbf{Y}) = N_N(\mathbf{X}_0\boldsymbol{\beta}, \sigma^2 I_n)$, then with $\nu = n - p - r$

$$D_I(\hat{\boldsymbol{\beta}}, \mathbf{X}_0^{'}\mathbf{X}_0, rs_I^2) \sim F_r(t; \gamma_1^2, \cdots, \gamma_r^2; n-p-r)$$

where $\{\gamma_1^2 \geq \cdots \geq \gamma_r^2\}$ are the ordered eigenvalues of $\mathbf{Z}(\mathbf{X}^{'}\mathbf{X})^{-1}\mathbf{Z}'$;

$$D_I(\hat{\boldsymbol{\beta}}, \mathbf{X}^{'}\mathbf{X}, rs_I^2) \sim F_r(t; \lambda_1, \cdots, \lambda_r; n-p-r)$$

where $\{\lambda_1 \geq \cdots \geq \lambda_r\}$ are the ordered eigenvalues of $\mathbf{Z}(\mathbf{X}_0^{'}\mathbf{X}_0)^{-1}\mathbf{Z}'$ and $\lambda_i = \gamma_i^2/(1+\gamma_i^2) = h_{ii}, 1 \leq i \leq r$; and

$$D_I(\hat{\boldsymbol{\beta}}, \sigma^2 cov(\hat{\boldsymbol{\beta}}_I - \hat{\boldsymbol{\beta}})^+, rs_I^2) \sim F_r(t; 1, \cdots, 1; n-k)$$

JOINT OUTLIER RESULTS: When $r > 1$, outliers can be tested with

$$\begin{aligned}
F_I &\sim F_r(t; 1, \cdots, 1; n-p-r) \\
D_I(\hat{\boldsymbol{\beta}}, \sigma^2 cov(\hat{\boldsymbol{\beta}}_I - \hat{\boldsymbol{\beta}})^+, rs_I^2) &= D_I \sim F_r(t; 1, \cdots, 1; n-p-r) \\
D_I(\hat{\boldsymbol{\beta}}, \mathbf{X}_0^{'}\mathbf{X}_0, rs_I^2) &\sim F_r(t; \gamma_1^2, \cdots, \gamma_r^2; n-p-r) \quad (17) \\
D_I(\hat{\boldsymbol{\beta}}, \mathbf{X}^{'}\mathbf{X}, rs_I^2) &\sim F_r(t; \lambda_1, \cdots, \lambda_r; n-p-r) \quad (18)
\end{aligned}$$

The mean shift statistic F_I and the normalized Cook statistic D_I yield the same p-values. For the Cook-like statistics, we prefer (18) over (17) for computationally reasons, since the number of terms required in the series expansion for the cdf of $(T/r)/(V/\nu)$ is smaller. This number increases with the condition numbers, which satisfy $\gamma_1^2/\gamma_r^2 \geq \lambda_1/\lambda_r$.

0.3 EXAMPLE

We use the Drill Data Set from Cook and Weisberg (1982, p. 149) with $n = 31$ and $k = 10$. With $r = 1$, we find the influential rows (with $p < .025$) to be rows 9, 28, and row 31. The p-values are (necessarily) the same using $D_i(\hat{\boldsymbol{\beta}}, \mathbf{X}_0^{'}\mathbf{X}_0, s_i^2)$, $D_i = D_i(\hat{\boldsymbol{\beta}}, \mathbf{X}^{'}\mathbf{X}, s_{(i)}^2)$, $D_i(\hat{\boldsymbol{\beta}}, \sigma^2 cov(\hat{\boldsymbol{\beta}}_I - \hat{\boldsymbol{\beta}})^+, s_{(i)}^2)$, or t_i (RSTUDENT).

Row	$s_{(i)}^2$	D_i	λ_i	p-value	D_i/λ_i
9	.0182	7.485	.663	.0031	11.28
28	.0205	**2.953**	.376	.0110	**7.85**
31	.0215	**3.582**	.550	.0190	**6.51**
5	.0228	2.186	.439	.0366	5.02
26	.0237	1.763	.436	.0580	4.04
30	.0259	0.828	.404	.1677	2.05

With $r = 2$, we find 39 influential pairs (with $p < .01$ and using the average of the p-bounds) using $D_I(\hat{\boldsymbol{\beta}}, \mathbf{X}_0^{'}\mathbf{X}_0, 2s_I^2)$, 35 pairs using $D_I(\hat{\boldsymbol{\beta}}, \mathbf{X}^{'}\mathbf{X}, 2s_I^2)$, and 16 pairs (directly) using $D_I(\hat{\boldsymbol{\beta}}, \sigma^2 cov(\hat{\boldsymbol{\beta}}_I - \hat{\boldsymbol{\beta}})^+, 2s_I^2)$. All contain rows from {9,28,31} except for the pair (5,26). For this pair using $D_I(\hat{\boldsymbol{\beta}}, \mathbf{X}'\mathbf{X}, 2s_I^2)$, we have

Rows	s_I^2	D_I	λ_1	λ_2	LB	p	UB
5,26	.0119	14.75	.654	.217	.000005	.000115	.000161

The p-values for this pair, using $D_I(\hat{\beta}, X_0'X_0, 2s_I^2)$, $D_I(\hat{\beta}, X'X,$

$2s_I^2)$, and $D_I(\hat{\beta}, \sigma^2 cov(\hat{\beta}_I - \hat{\beta})^+, 2s_I^2)$:

Test	p-value	N
$D_I(\hat{\beta}, X_0^{'}X_0, 2s_I^2)$	.000129	50
$D_I(\hat{\beta}, X^{'}X, 2s_I^2)$	.000115	20
$D_I(\hat{\beta}, \sigma^2 cov(\hat{\beta}_I - \hat{\beta})^+, 2s_I^2)$	.000161	1

0.3.1 Other Heuristics

We now consider joint outliers for the Intercountry Life-Cycle Savings Data from Belsley *et al.* (1980, p. 41). We will only consider pairs of observations with $I = \{i_1, i_2\}$. The extension to larger subsets is immediate. Belsley *et al.* (1980) used a number of different empirical procedures to identify 15 pairs of observations that may be influential. These procedures include MDFFIT (MD), COVRATIO (CR), RESRATIO (RR), MEWDFFIT (ME), Wilks' Λ statistic (WL), and the Andrews and Pregibon statistic (AP). Table I shows those pairs of observations are of note based on these criteria. The last column records the p-values from $D_I(\hat{\beta}, X_0^{'}X_0, 2s_I^2)$. These empirical procedures do not agree with each other, nor do they detect the joints outliers which we are able to determine using D_I as an omnibus test of model fit under the mean shift and variance shift outlier model. Outliers are often influential, however not all influential subsets are outliers (see Barnett and Lewis (1994, p. 317)).

Table I: Multiple-Row Influence Using Empirical Procedures

Rows	MD	CR	RR	ME	WL	AP	p-value
34,46		*	*				.001
33,46		*	*				.001
23,46	*		*	*		*	.007
19,23	*						.030
47,49				*	*	*	.081
7,46		*	*				.143
24,49				*			.146
33,49				*			.161
37,49							.242
23,49					*		.329
6,49							.339
46,49						*	.381
44,49					*	*	.420
21,49					*	*	.438
6,44					*		.699

MDFFIT (MD) (numerator of $D_I(\hat{\boldsymbol{\beta}}, \mathbf{X}'\mathbf{X}, 2s_I^2)$)

$$(\hat{\boldsymbol{\beta}}_I - \hat{\boldsymbol{\beta}})' \mathbf{X}'\mathbf{X}(\hat{\boldsymbol{\beta}}_I - \hat{\boldsymbol{\beta}})$$

COVRATIO (CR)

$$CR_{(i)} = \frac{|s^2_{(i)}(\mathbf{X}'\mathbf{X})^{-1}|}{|s^2(\mathbf{X}_0'\mathbf{X}')^{-1}|}$$

RESRATIO (RR) (same as the mean-shift statistic)

$$F_I = \frac{((n-p)s^2 - (n-p-r)s_I^2)/r}{s_I^2}$$

MEWDFFIT (ME)

$$ME = \sum_{i,j \in I} h_{ij} \frac{e_i e_j}{(1-h_i)(1-h_j)}$$

Wilks' Λ statistic (WL)

$$WL = 1 - \frac{n}{r(n-r)} \mathbf{1}'\mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{1}$$

Andrews and Pregibon (AP)

$$AP_I = 1 - \frac{n-p-r}{n-p} \frac{s_I^2 |(\mathbf{X}'\mathbf{X})^{-1}|}{s^2 |(\mathbf{X}_0'\mathbf{X})^{-1}|}$$

0.4 R-Fisher distribution

Assumptions A.

A1. $E(\boldsymbol{\varepsilon}) = 0 \in \Re^{n-r}$ and $E(\boldsymbol{\varepsilon}_I) = \boldsymbol{\delta} \in \Re^r$;

A2. $V(\boldsymbol{\varepsilon}_0) = Diag(\sigma^2, \sigma_1^2)$; and

A3. $\mathcal{L}(\boldsymbol{\varepsilon}, \boldsymbol{\varepsilon}_I - \boldsymbol{\delta}) = N_n(\mathbf{0}, Diag(\sigma^2, \sigma_1^2))$.

This model allows for shifts in both location and scale at design points in $\mathbf{Z}$.

$$\left[\begin{array}{c} \mathbf{Y}_1 \\ \mathbf{Y}_2 \end{array}\right] = \left[\begin{array}{c} \mathbf{X} \\ \mathbf{Z} \end{array}\right] \boldsymbol{\beta} + \left[\begin{array}{c} \mathbf{0} \\ \mathbf{I}_r \end{array}\right] \boldsymbol{\delta} + \left[\begin{array}{l} \boldsymbol{\varepsilon} \\ \boldsymbol{\varepsilon}_I - \delta \end{array}\right]$$

With $\delta = 0$ and $\kappa = \sigma_1^2/\sigma^2 = 1$ (Jensen (1999, 2001))

$$\begin{aligned} F_I &= \frac{\mathbf{e}_I'(\mathbf{I}_r - \mathbf{Z}(\mathbf{X}_0'\mathbf{X}_0)^{-1}\mathbf{Z}')^{-1}\mathbf{e}_I}{rs_I^2} = \frac{\mathbf{e}_I'(\mathbf{I}_r - \mathbf{H}_{II})^{-1}\mathbf{e}_I}{rs_I^2} \\ &= \frac{((n-p)s^2 - (n-p-r)s_I^2)/r}{s_I^2} \sim F_{r,n-p-r} \end{aligned}$$

For the general model with $\delta \neq 0$ and $\kappa = \sigma_1^2/\sigma^2 > 1$, we have

THEOREM (Jensen (1999) for $\kappa = 1$ and Jensen and Ramirez (200x) for $\kappa \neq 1$) Consider the R-Fisher diagnostic F_I under Assumptions A; let $\{\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_r > 0\}$ comprise the canonical subset leverages (the eigenvalues of $\mathbf{H}_{II}$); set $\kappa = \sigma_1^2/\sigma^2 \geq 1$; let $\boldsymbol{\theta} = \mathbf{Q}'\boldsymbol{\delta}$, such that $\mathbf{Q}'\mathbf{H}_{II}\mathbf{Q} = Diag(\lambda_1, \lambda_2, \cdots, \lambda_r)$; and identify $\nu = n - p - r$.

i) The *cdf* of F_I is given by $F_s(w; \alpha'; \omega'; \nu)$, with weights $\{\alpha_i = \kappa - (\kappa - 1)\lambda_i;\ 1 \leq i \leq r\}$ satisfying $\{\alpha_r \geq \ldots \geq \alpha_1 \geq 1\}$, and with location parameters $\{\omega_i = \theta_i/\sigma[\kappa + \lambda_i/(1 - \lambda_i)]^{1/2}; 1 \leq i \leq r\}$.

ii) Bounds for the *cdf*s in terms of Fisher's distribution are given by

$$F(w/\alpha_r; r, \nu, \lambda) \ \leq \ F_s(w; \boldsymbol{\alpha}'; \boldsymbol{\omega}'; \nu) \leq \ F(w/\alpha^*; r, \nu, \lambda) \tag{19}$$

where $\lambda = \sum_{i=1}^{r} \theta_i^2/\sigma^2[\kappa + \lambda_i/(1 - \lambda_i)]$ and $\alpha^* = (\alpha_1 \cdots \alpha_r)^{1/r}$ is the geometric mean.

0.4.1 Case Study

Data (Myers (1990)) regarding the administration of Bachelor Officers Quarters *(BOQ)* were reported for sites at $n = 25$ naval installations. Monthly man-hours (Y) were related linearly to average daily occupancy (X_1), monthly number of check-ins (X_2), weekly service desk operation in hours (X_3), size of common use area (X_4), number of building wings (X_5), operational berthing capacity (X_6), and number of rooms (X_7). The data are reported in Myers (1990), p. 218 ff, together with detailed analyses using single-case deletion diagnostics. Subset deletion diagnostics are not reported there.

We focus on sites $\{15, 20, 21, 23, 24\}$, having individual leverages $\{0.5576, 0.3663, 0.0704, 0.9885, 0.8762\}$, their individual R-Student values exceeding the widely used ± 2 rule. Pairs of sites selected are $S_1 = \{20, 21\}$, $S_2 = \{15, 20\}$, and $S_3 = \{23, 24\}$, reflecting smaller, intermediate, and larger individual leverages.

Table II of Low, Medium, and High Leverages

		S_1			S_2			S_3	
		κ			κ			κ	
$\frac{\mathbf{1}}{\sigma}$	1	2	3	1	2	3	1	2	3
1	.1573	.2703	.3592	.1240	.2045	.2759	.0551	.0634	.0716
1	.1573	.2715(6)	.3606(6)	.1240	.2061(6)	.2779(6)	.0551	.0640(4)	.0734(5)
1	.1573	.2987	.3999	.1240	.2237	.3072	.0551	.0723	.0907
2	.5088	.5501	.5836	.3776	.4293	.4707	.0707	.0789	.0869
2	.5088	.5518(10)	.5848(9)	.3776	.4327(9)	.4743(9)	.0707	.0805(5)	.0904(6)
2	.5088	.5828	.6231	.3776	.4551	.5058	.0707	.0893	.1085
3	.8631	.8262	.8099	.7199	.7063	.7041	.0981	.1057	.1133
3	.8631	.8273(15)	.8105(13)	.7199	.7100(13)	.7080(12)	.0981	.1089(6)	.1195(7)
3	.8631	.8472	.8364	.7199	.7290	.7339	.0981	.1185	.1389
4	.9852	.9620	.9424	.9319	.9007	.8801	.1388	.1451	.1514
4	.9852	.9623(21)	.9426(18)	.9319	.9028(18)	.8826(16)	.1388	.1504(7)	.1613(8)
4	.9852	.9688	.9534	.9319	.9123	.8971	.1388	.1609	.1823

0.4.2 Conclusions

The single-case deletion diagnostics of Groups I, II, and III are all equivalent. Moreover, the Group II subset deletion diagnostics and the Group III diagnostic D_I are all equivalent to the R-Fisher diagnostic F_I. These subset deletion diagnostics have been studied under outliers arising from shifts in both location and scale. In these circumstances the distribution of F_I has been shown to be a noncentral generalized F distribution. We have given series expansions for these distributions, as well as global error bounds for their partial sums. Bounds for the *cdf*s of noncentral generalized F distributions also have been given in terms of the noncentral Fisher distributions.

The noncentral generalized F distributions have been used to compute the power for the diagnostic F_I, and thus for all equivalent diagnostics, under shifts in location and scale at selected subsets of the BOQ data. These case studies demonstrate that subsets with large canonical leverages tend to associate with tests having low power, so that shifts in both location and scale may be masked at points of high leverages.

References

[1] Andrews, D. F. and Pregibon, D. (1978), Finding outliers that matter, *J. Royal Statist. Soc. B* 40, 85-93.

[2] Barnett, V. and Lewis, T. (1994). *Outliers in Statistical Data.* John Wiley and Sons, New York.

[3] Belsley, D. A., Kuh, E. and Welsch, R. E. (1980). *Regression, Diagnostics: Identifying Influential Data and Sources of Collinearity.* John Wiley and Sons, New York.

[4] Cook, R. D. and Weisberg, S. (1982). *Residuals and Influence in Regression.* Chapman and Hall.

[5] Dunkl, C. F. and Ramirez, D. E. (2001), Computation of the generalized F distribution, *Australian and New Zealand Journal of Statistics*, **43**, 21-31.

[6] Jensen, D. R. and Ramirez, D. E. (1996), Computing the cdf of Cook's D_I statistic, In *Proceedings of the 12th Symposium in Computational Statistics*, Barcelona, Spain, 1996, Prat, A. and Ripoll.

[7] LaMotte, L. R. (1999), Collapsibility hypotheses and diagnostic bounds in regression analysis, *Metrika* **50**, 109-119.

[8] Jensen, D. R. (1998), The use of standardized diagnostics in regression, *Research Report* R079, Department of Statistics, Virginia Polytechnic Institute and State University, October 1998.

[9] Jensen, D. R. (1999), Properties of selected subset diagnostics in regression, *Research Report* R080, Department of Statistics, Virginia Polytechnic Institute and State University, March 1999.

[10] Jensen, D. R. (2000), The use of Standardized diagnostics in regression, *Metrika,* **52***, 213-223.*

[11] Jensen, D. R. (2001), Properties of selected subset diagnostics in regression, *Statist. Probab. Letters,* **51***, 377-388.*

[12] Jensen, D. R. and Ramirez, D. E. (1991), Misspecified Tˆ2 tests, I. Location and scale, *Comm. Stat. Theory Meth. A* **20**, No. 1, 249–259

[13] Jensen, D. R. and Ramirez, D. E. (1998), Some exact properties of Cook's D_I, *Handbook of Statistics*, Vol. 16, Balakrishnan, N. and Rao, C. R., eds., pp. 387-402 Elsevier Science Publishers B. V.

[14] Jensen, D. R. and Ramirez, D. E. (200x), Detecting mean-shift outliers via distances, Journal of Statistical Computation and Simulation.

[15] Jensen, D. R. and Ramirez, D. E. (200x), Detecting shifts in location and scale in regression, South African Statistical Journal.

[16] Myers, R. H. (1990). *Classical and Modern Regression with Applications*, Second ed. PWS-Kent Publishing Company, Boston, MA.

[17] Ramirez, D E. (2000), The generalized F distribution, *Journal of Statistical Software* **5(1)**, 1-13.

[18] Ramirez, D. E. and Jensen, D. R. (1991). Misspecified T^2 tests. II. Series expansions. *Commun. Statist. Simula.* **20:** 97-108.