

Dimension reduction in regression and an example robustness study of inverse response plot estimation

Luke Prendergast
Department of Mathematics and Statistics
La Trobe University
Melbourne, Australia

6 October, 2010

Overview

1. Dimension reduction in regression
2. Inverse response plot (IRP) estimation
3. The influence function
4. Influence functions for IRP's
5. A robust proposal for IRP estimation

Some notation and the MLR model

Let $\mathbf{x} = [x_1, \dots, x_p]^\top \in \mathbb{R}^p$ be a p -dimensional regressor vector.

Let $\boldsymbol{\beta} = [\beta_1, \dots, \beta_p]^\top$ be a p -dimensional vector of regressor coefficients.

The the Multiple Linear Regression (MLR) model is of the form

$$Y = \beta_0 + \boldsymbol{\beta}^\top \mathbf{x} + \epsilon$$

where

- ▶ ϵ is a random error term
- ▶ β_0 is the intercept coefficient
- ▶ Y is the random response

The Single-Index Model (SIM)

For the MLR model note that Y depends on \mathbf{x} only through

$$\boldsymbol{\beta}^\top \mathbf{x} = \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_p x_p$$

but in a very simple linear sense.

Now, consider

$$Y = f(\boldsymbol{\beta}^\top \mathbf{x}, \epsilon)$$

where

- ▶ f is the unknown 'link function'

Note that Y still depends on \mathbf{x} only through $\boldsymbol{\beta}^\top \mathbf{x}$ though not necessarily in a strictly linear sense.

Some more models for dimension reduction

A generalization of the SIM is (Li, 1991)

$$Y = f(\beta_1^\top \mathbf{x}, \dots, \beta_K^\top \mathbf{x}, \epsilon) \quad (1)$$

Under this models we can replace the p -dimensional \mathbf{x} with the K -dimensional $\beta_1^\top \mathbf{x}, \dots, \beta_K^\top \mathbf{x}$ without loss of information (i.e. dimension reduction when $K < p$).

The purpose of dimension reduction (DR) methods

Let $\mathcal{S} = \text{span}(\beta_1, \dots, \beta_K)$.

Since, for e.g., f is unknown, the β_j 's cannot be uniquely determined.

However, given $\gamma_1, \dots, \gamma_K$ such that

$$\text{span}(\gamma_1, \dots, \gamma_K) = \mathcal{S}$$

we can similarly use

$$\gamma_1^\top \mathbf{x}, \dots, \gamma_K^\top \mathbf{x}$$

in place of \mathbf{x} and achieve dimension reduction.

Importantly: DR methods seek any basis for \mathcal{S} .

Some modern DR methods

Method	Pros	Cons
OLS	✓ Mild conditions on \mathbf{x}	✗ Only with $K = 1$
SIR ¹	✓ Mild conditions on \mathbf{x} ✓ Can deal with $K \geq 1$	✗ For some models may only find part of \mathcal{S}
SAVE ²	✓ Fewer model limitations than SIR	✗ more restrictions on \mathbf{x}

¹ Sliced Inverse Regression (Li, 1991)

² Sliced Average Variance Estimates (Cook & Weisberg, 1991)

Sufficient Summary Plots

A plot of Y versus $\beta_1^\top \mathbf{x}, \dots, \beta_K^\top \mathbf{x}$ is called a Sufficient Summary Plot (SSP, see for e.g. Cook, 1998).

An SSP can be used to detect structure linking Y and \mathbf{x} .

Let $(y_1, \mathbf{x}_1), \dots, (y_n, \mathbf{x}_n)$ denote n sample realizations of (Y, \mathbf{x}) .

Let $\hat{\beta}_1, \dots, \hat{\beta}_K$ be an estimated basis for \mathcal{S} .

A plot of the y_i 's versus the $\hat{\beta}_1^\top \mathbf{x}_i, \dots, \hat{\beta}_K^\top \mathbf{x}_i$ is an Estimated SSP (ESSP).

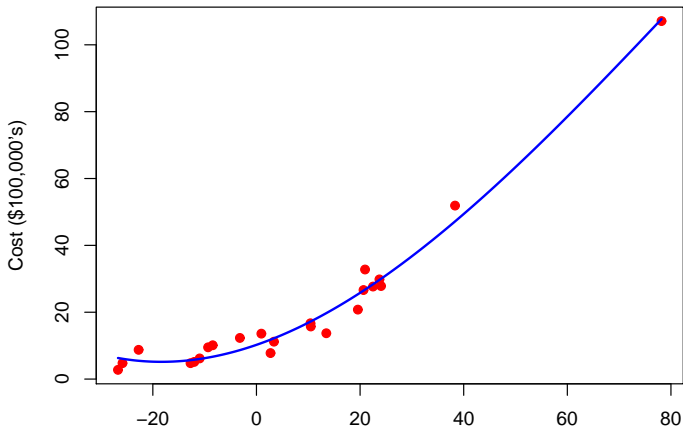
The aircraft cost data ($n = 23$): OLS (+ improvements)

Consider data for $n = 23$ aircraft where

- ▶ Regressors: Aspect Ratio (AR), Lift-to-Drag Ratio (LDR), Weight (W), Thrust (T)
- ▶ Response is cost in \$100,000's.

Method of estimation: OLS with view improved via additional residual minimization based on a polynomial fit (Prendergast & Healey, 2009).

Aircraft data ESSP



$$-7.818 \times AR + 3.393 \times LDR + 0.003 \times W - 0.003 \times T$$

Power transformation models

Consider the model

$$Y = (\beta_0 + \beta_1^\top \mathbf{x} + \varepsilon)^{1/\lambda} \quad (\text{or } Y = \exp(\beta_0 + \beta_1^\top \mathbf{x} + \varepsilon))$$

where the aim is to determine λ for the purpose of linearization.

That is, transform the response so that

$$t(Y) = Y^\lambda = \beta_0 + \beta_1^\top \mathbf{x} + \varepsilon. \quad (\text{or } t(Y) = \ln(Y))$$

Inverse response plots

To help seek λ , Cook & Weisberg (1994) suggest plotting the

$$\hat{\beta}_0 + \hat{\beta}_1^T \mathbf{x}_i$$

versus the y_i 's.

Usually, the y_i 's are on the vertical axis, but here they are on the horizontal axis \Rightarrow Inverse Response Plot (IRP)

We could just use the $\hat{\beta}_1^T \mathbf{x}_i$'s, but the question arises

How do we estimate β_1 ?

Cook & Weisberg note that under mild conditions OLS can be used to estimate β_1 (Duan & Li, 1989).

Scaled power transformations

For $Y > 0$, a popular choice for estimating λ in practice is the scaled power transformation family where Y is transformed using

$$\Psi(Y, \lambda) = \begin{cases} \frac{Y^\lambda - 1}{\lambda}, & \lambda \neq 0 \\ \ln(Y) & \lambda = 0 \end{cases}$$

To do this carry out the non-linear least squares regression where $\hat{\lambda}$ is chosen such that the sum of squared residuals for the regression of the

$$\hat{\beta}_1^\top \mathbf{x}_i \text{'s on the } \Psi(y_i, \lambda) \text{'s}$$

is minimized.

Cook & Weisberg (1994) considered IRP estimation for the wool data.

$n = 27$ samples measured on the 4 variables:

- ▶ Y : cycles to failure of worsted yarn.
- ▶ X_1 : length of specimen.
- ▶ X_2 : amplitude of loading cycle.
- ▶ X_3 : load.

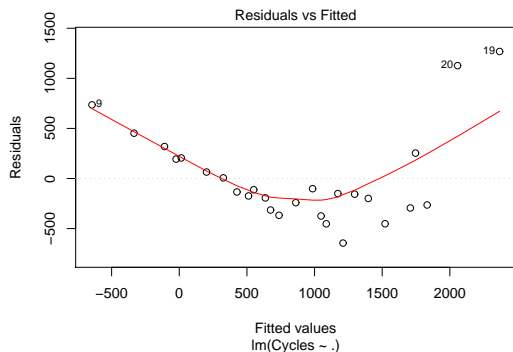
Can a model be found that can be useful for predicting Y given values for X_1, X_2, X_3 ?

Let's start by considering the MLR model

$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \varepsilon$ and let $\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2, \hat{\beta}_3$ denote the OLS estimates.

MLR model diagnostics

$$R_{adj}^2 = 0.694$$



Clear pattern in the residuals versus fits suggests this model is definitely not appropriate.

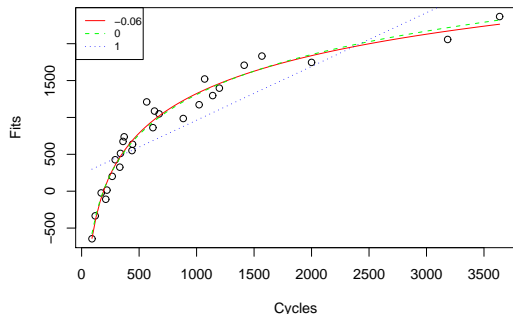
IRP approach

1. Assume a model of the form

$$Y^\lambda(\text{or } \ln(Y)) = b_0 + b_1X_1 + b_2X_2 + b_3X_3 + \varepsilon$$

2. Plot the $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1x_{i1} + \hat{\beta}_2x_{i2} + \hat{\beta}_3x'_{i3}$ s versus the y_i 's (IRP).
3. Estimate λ from the IRP.
4. Transform the y_i 's using this estimate, then use OLS to estimate b_0, \dots, b_3 .

IRP for the wool data

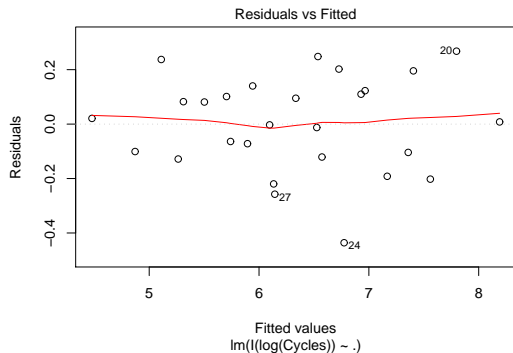


$\hat{\lambda} = -0.06 \approx 0$ so let's try transforming the y_i 's using the natural log. That is, use OLS to estimate the model assumed to be

$$\ln(Y) = b_0 + b_1X_1 + b_2X_2 + b_3X_3 + \varepsilon.$$

Log transformed model diagnostics

$$R_{adj}^2 = 0.961$$



No obvious pattern in the residuals versus fits and high R_{adj}^2 suggest this is an excellent model.

Statistical functionals

Consider a population distribution function F .

Suppose that we have a sample of n observations observed from this population which are denoted x_1, \dots, x_n and let F_n denote the empirical distribution for this data.

Also suppose θ is a population parameter of interest which is to be estimated and denote this estimate as $\hat{\theta}$.

Let t denote the statistical functional for the estimator of θ . Then

- ▶ $t(F) = \theta$
- ▶ $t(F_n) = \hat{\theta}$

Example: Functionals for the mean and variance estimators

Let μ and σ^2 denote the population mean and variance.

Let T and V denote the functionals for the usual sample mean estimator and the MLE variance estimator.

$$T(F) = \int xf(x)dx = \int xdF = \mu \text{ and } T(F_n) = \frac{1}{n} \sum_{i=1}^n x_i = \bar{x}$$

$$V(F) = \sigma^2 \text{ and } V(F_n) = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2 = \hat{\sigma}^2$$

The Contamination Distribution

Let

$$F_\epsilon = (1 - \epsilon)F + \epsilon\Delta_{x_0}$$

where

- ▶ Δ_{x_0} is the probability distribution that puts all mass at the point x_0 .
- ▶ x_0 is the contamination point.
- ▶ $0 \leq \epsilon \leq 1$ is the proportion of contamination.

Question: what is the effect of contamination on the statistical estimator with functional t ?

The influence function

Let us consider $t(F_\epsilon)$ with respect to the Power Series

$$t(F_\epsilon) = t(F) + \epsilon \text{IF}(t, F; x_0) + O(\epsilon^2)$$

$\text{IF}(t, F; x_0)$ is called the influence function and is defined as

$$\text{IF}(t, F; x_0) = \lim_{\epsilon \downarrow 0} \frac{t(F_\epsilon) - t(F)}{\epsilon} = \left. \frac{\partial}{\partial \epsilon} t(F_\epsilon) \right|_{\epsilon=0}.$$

We say that $\text{IF}(t, F; x_0)$ exists in a 'closed form' when it may be written in terms of x_0 and parameters of F only.

If $\text{IF}(t, F; x_0)$ is large then the contamination has been highly influential since $t(F_\epsilon)$ is very different from $t(F)$.

Influence functions for the mean and variance estimators

For the mean

$$\text{IF}(T, F; x_0) = x_0 - \mu$$

- ▶ zero influence when $x_0 = \mu$.
- ▶ influence of x_0 increases as x_0 is moved further from μ .

For the variance

$$\text{IF}(V, F; x_0) = (x_0 - \mu)^2 - \sigma^2$$

- ▶ zero influence when $(x_0 - \mu)^2 = \sigma^2$.
- ▶ influence of x_0 increases as $(x_0 - \mu)^2$ is moved further from σ^2 .

Sample versions

Recall $\hat{\theta}$ is the estimate to θ based on the n observed sample observations.

Now let $\hat{\theta}_{(i)}$ denote the estimate of θ when the i th observation is ignored.

The sample influence function is

$$\text{SIF}_i = (n - 1) \left[\hat{\theta} - \hat{\theta}_{(i)} \right]$$

The empirical influence function denoted EIF_i is the $\text{IF}(t, F; x_0)$ but where

- ▶ x_i replaces x_0 .
- ▶ Unknown parameters are replaced with their estimates based on the n observations.

Sample versions (ctd)

Importantly: $\text{EIF}(t, F_{n_i}; x_i) \approx \text{SIF}(t, F_{n_i}; x_i)$

For the mean

$$\text{EIF}(T, F_{n_i}; x_i) = x_i - \bar{x} = (n-1)(\bar{x} - \bar{x}_{(i)}) = \text{SIF}_i.$$

For the variance

$$\text{EIF}(V, F_{n_i}; x_i) = (x_i - \bar{x})^2 - \hat{\sigma}^2 \approx (n-1)(\hat{\sigma}^2 - \hat{\sigma}_{(i)}^2) = \text{SIF}_i.$$

That is, we may use the influence to

- ▶ assess the sensitivity of an estimator
- ▶ construct influence diagnostics that:
 - ▶ are efficient to calculate
 - ▶ have interpretive strengths

Influence function for the estimator of λ

Let $IF(I, F; y_0, \mathbf{x}_0)$ denote the IF for the estimator of λ where y_0 and \mathbf{x}_0 are the contamination response and predictor.

Prendergast and Sheather (2010) show that

$$IF(I, F; y_0, \mathbf{x}_0) \propto c(y_0, \mathbf{x}_0) + \mathbf{v}^\top IF(\beta_R, F; y_0, \mathbf{x}_0)$$

where

- ▶ $IF(\beta_R, F; y_0, \mathbf{x}_0)$ is the influence function for the initial estimator of β_1 .
- ▶ $\mathbf{v} \in \mathbb{R}^p$ free of (y_0, \mathbf{x}_0) .
- ▶ $c : \mathbb{R}^{p+1} \mapsto \mathbb{R}$ is for the influence associated with the second step (i.e. estimation of λ following the estimation of β_1).

Influence function for the estimator of λ (ctd)

But, when the conditions hold for the initial estimator of β_1 to be applicable:

$$\text{IF}(l, F; y_0, \mathbf{x}_0) \propto c(y_0, \mathbf{x}_0) + \mathbf{v}^\top \text{IF}(\beta_F, F; y_0, \mathbf{x}_0)$$

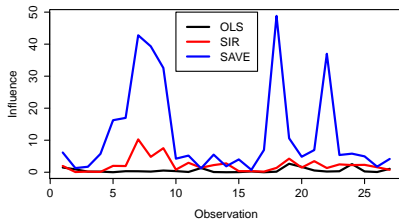


$$\text{IF}(l, F; y_0, \mathbf{x}_0) \propto c(y_0, \mathbf{x}_0)$$

The final influence is independent of influence in the initial step!

Sample influence for the wool data

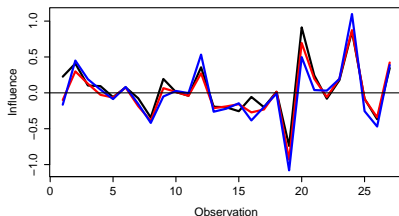
A: Influence on initial estimation step



Estimates of λ after OLS, SIR and SAVE used: -0.061, -0.061, -0.062

High influence in initial estimation (measure from Prendergast, 2008) does not necessarily mean high influence in estimation of λ .

B: Influence on estimation of lambda



Influence in initial estimation can be very different for OLS, SIR and SAVE yet overall influence on estimation of λ is similar.

Consequences

Poor IRP's can still result in reasonable estimates of λ .

Question: So does this mean we don't have to worry about poor estimation in the first step?

Answer: Not necessarily! A good initial estimate and therefore good IRP can help us determine whether we have a good estimate to λ .

Robust IRP estimation

- ▶ Use a robust estimator of β_1 in an effort to obtain an improved IRP.
- ▶ Use robust non-linear least squares to estimate λ .

We will use robust M -estimators using the Huber weight function.

An example

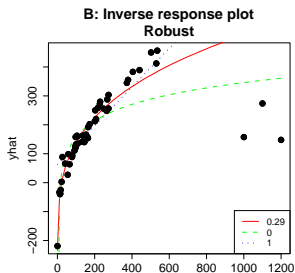
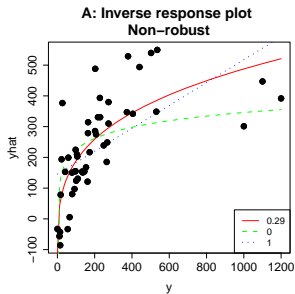
Consider the following model:

$$Y = (\beta_0 + \beta_1^\top \mathbf{x} + 0.1\varepsilon)^3$$

where

- ▶ $\mathbf{x} = [x_1, \dots, x_p]^\top \sim N_p([2.5, \dots, 2.5]^\top, \text{diag}[0.8^2, \dots, 0.8^2])$
- ▶ $\varepsilon \sim N(0, 1)$
- ▶ $\beta_1 = [2, -1, 0, \dots, 0]^\top$

We will generate $n = 50$ observations according to this model and replace 3 of the generated y_i 's with large values.



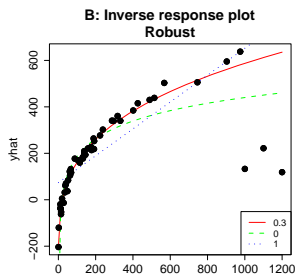
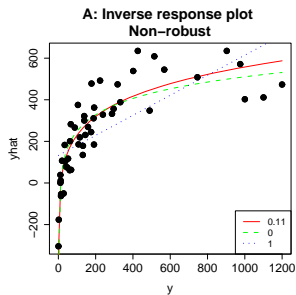
Simulation 1:

Target is $\lambda = 0.33$.

Both estimators of λ estimate $\hat{\lambda} = 0.29$.

Easier to have faith in the robust estimator because the corresponding IRP provides a sharper view.

Suspicious observations easily identified on the robust IRP.



Simulation 2:

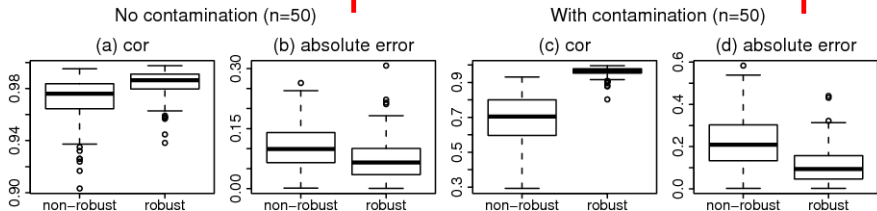
Non-robust estimate is poor; $\hat{\lambda} = 0.11$.

No obvious suspicious observations in non-robust IRP.

Robust estimate is very good; $\hat{\lambda} = 0.3$.

Suspicious observations easily identified on the robust IRP.

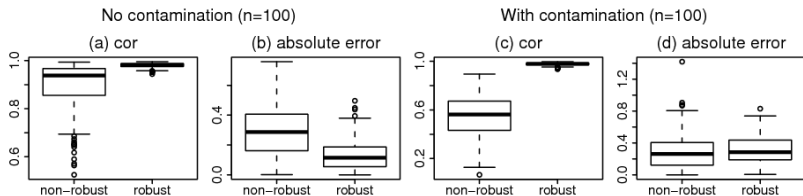
Close to 0 \Leftrightarrow good estimate of λ



Close to 1 \Leftrightarrow good IRP

Even without contamination, the robust approach typically provides a better IRP and $\hat{\lambda}$

This time $Y = (\beta_0 + \beta_1^T \mathbf{x} + 0.5\epsilon)^{-1}$



Again, even without contamination, the robust approach typically provides a better IRP and $\hat{\lambda}$

With contamination: Much better IRP's via robust approach but similar estimates to λ .

Conclusions

Dimension reduction methods can be applied to a very general class of models.

Inverse response plots are a popular tool for detecting suitable response transformation for linearization following a dimension reduction.

The influence function helps to explain some unexpected behavior; sometimes poor initial estimates can still result in reasonable final estimates.

A simple robust approach can be beneficial even for 'well behaved' data.

References

- Li, K.C. (1991), Sliced inverse regression for dimension reduction, *J. Amer. Statist. Assoc.*, **86**, 316–327.
- Cook, R. D. and Weisberg, S. (1991), Discussion of “Sliced inverse regression for dimension reduction”, *J. Amer. Statist. Assoc.*, **86**, 328–332.
- Cook, R. D. and Weisberg, S. (1994), Transforming a response variable for linearity, *Biometrika*, **81**, 731–737.
- Prendergast, L. A. (2008), Trimming influential observations for improved single-index model estimated sufficient summary plots, *Comput. Stat. Data. An.*, **52**, 5319–5327.
- Prendergast, L. A. and Healey, A. (2009), Improving estimated sufficient summary plots in dimension reduction using minimization criteria based on initial estimates. Being revised for resubmission.
- Prendergast, L. A. and Sheather, S. J. (2010), On sensitivity of inverse response plot estimation and the benefits of a robust estimation approach. Submitted.