

Outliers & Robust Bayesian Regression

Readings: Hoff Chapter 9, West JRSSB 1984, Fúquene, Pérez
& Pericchi 2015

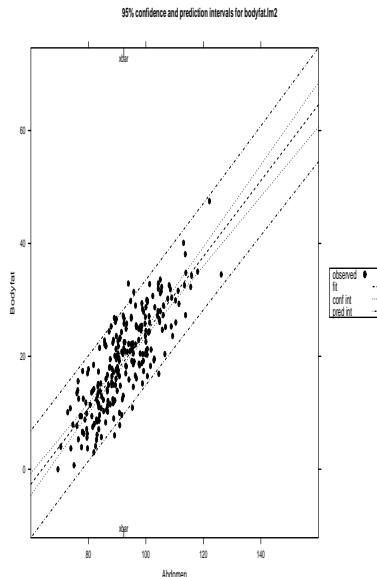
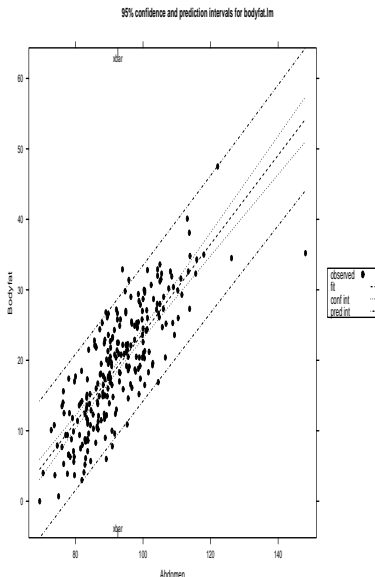
STA 702 Duke University

Duke University

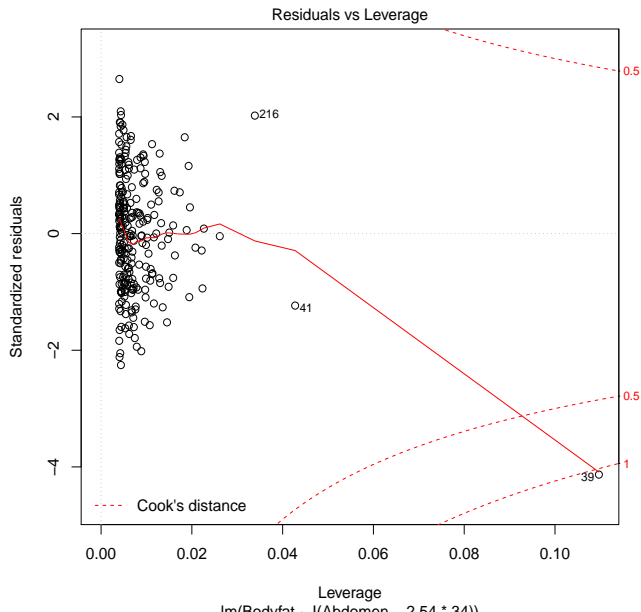
November 7, 2022

Body Fat Data: Intervals w/ All Data

Response % Body Fat and Predictor Waist Circumference



Cook's Distance



Outliers in Regression

- ▶ Hoeting, Madigan and Raftery (in various permutations) consider the problem of simultaneous variable selection and outlier identification.
- ▶ This is implemented in the package BMA in the function `MC3.REG`. This has the advantage that more than 2 points may be considered as outliers at the same time.
- ▶ The function uses a Markov chain to identify both important variables and potential outliers, but is coded in Fortran so should run reasonably quickly.
- ▶ Can also use BAS or other variable selection programs

Options for Handling Influential Cases

- ▶ Are there scientific grounds for eliminating the case?
- ▶ Test if the case has a different mean than population
- ▶ Report results with and without the case
- ▶ Model Averaging to Account for Model Uncertainty?
- ▶ Full model $Y = X\beta + I_n\delta + \epsilon$
- ▶ 2^n submodels $\gamma_i = 0 \Leftrightarrow \delta_i = 0$
- ▶ If $\gamma_i = 1$ then case i has a different mean “mean shift” outliers.

Mean Shift = Variance Inflation

► Model $Y = X\beta + I_n\delta + \epsilon$

► Prior

$$\delta_i \mid \gamma_i \sim N(0, V\sigma^2\gamma_i)$$

$$\gamma_i \sim \text{Ber}(\pi)$$

Then ϵ_i given σ^2 is independent of δ_i and

$$\epsilon_i^* \equiv \epsilon_i + \delta_i \mid \sigma^2 \begin{cases} N(0, \sigma^2) & \text{wp } (1 - \pi) \\ N(0, \sigma^2(1 + V)) & \text{wp } \pi \end{cases}$$

Model $Y = X\beta + \epsilon^*$ “variance inflation”

$V + 1 = K = 7$ in the paper by Hoeting et al. package BMA

Simultaneous Outlier and Variable Selection

```
MC3.REG(all.y = bodyfat$Bodyfat, all.x = as.matrix(bodyfat$Abdom  
num.its = 10000, outliers = TRUE)
```

Model parameters: $PI=0.02$ $K=7$ $\nu=2.58$ $\lambda=0.28$ $\phi=2.85$

15 models were selected

Best 5 models (cumulative posterior probability = 0.9939):

	prob	model 1	model 2	model 3	model 4	model 5
variables						
all.x	1	x	x	x	x	x
outliers						
39	0.94932	x	x	.	x	.
204	0.04117	.	.	.	x	.
207	0.10427	.	x	.	.	x
post prob		0.815	0.095	0.044	0.035	0.004

Change Error Assumptions

$$Y_i \stackrel{\text{ind}}{\sim} t(\nu, \alpha + \beta x_i, 1/\phi)$$

$$L(\alpha, \beta, \phi) \propto \prod_{i=1}^n \phi^{1/2} \left(1 + \frac{\phi(y_i - \alpha - \beta x_i)^2}{\nu} \right)^{-\frac{(\nu+1)}{2}}$$

Use Prior $p(\alpha, \beta, \phi) \propto 1/\phi$

Posterior distribution

$$p(\alpha, \beta, \phi \mid Y) \propto \phi^{n/2-1} \prod_{i=1}^n \left(1 + \frac{\phi(y_i - \alpha - \beta x_i)^2}{\nu} \right)^{-\frac{(\nu+1)}{2}}$$

Bounded Influence - West 1984 (and references within)

Treat σ^2 as given, then *influence* of individual observations on the posterior distribution of β in the model where $E[Y_i] = x_i^T \beta$ is investigated through the score function:

$$\frac{d}{d\beta} \log p(\beta | Y) = \frac{d}{d\beta} \log p(\beta) + \sum_{i=1}^n x_i g(y_i - x_i^T \beta)$$

where

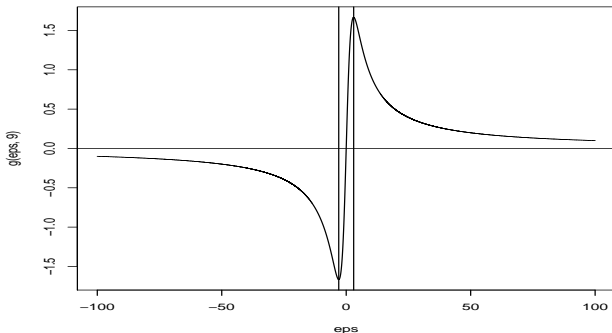
$$g(\epsilon) = -\frac{d}{d\epsilon} \log p(\epsilon)$$

is the influence function of the error distribution (unimodal, continuous, differentiable, symmetric)

An outlying observation y_j is accommodated if the posterior distribution for $p(\beta | Y_{(j)})$ converges to $p(\beta | Y)$ for all β as $|Y_j| \rightarrow \infty$. Requires error models with influence functions that go to zero such as the Student t (O'Hagan, 1979)

Choice of df

- ▶ Score function for t with α degrees of freedom has turning points at $\pm\sqrt{\alpha}$



- ▶ $g'(\epsilon)$ is negative when $\epsilon^2 > \alpha$ (standardized errors)
- ▶ Contribution of observation to information matrix is negative and the observation is doubtful
- ▶ Suggest taking $\alpha = 8$ or $\alpha = 9$ to reject errors larger than $\sqrt{8}$ or 3 sd.

Scale-Mixtures of Normal Representation

$$Z_i \stackrel{\text{iid}}{\sim} t(\nu, 0, \sigma^2) \Leftrightarrow$$

$$Z_i \mid \lambda_i \stackrel{\text{iid}}{\sim} N(0, \sigma^2 / \lambda_i)$$

$$\lambda_i \stackrel{\text{iid}}{\sim} G(\nu/2, \nu/2)$$

Integrate out “latent” λ 's to obtain marginal distribution.

Latent Variable Model

$$\begin{aligned} Y_i \mid \alpha, \beta, \phi, \lambda &\stackrel{\text{ind}}{\sim} N\left(\alpha + \beta x_i, \frac{1}{\phi \lambda_i}\right) \\ \lambda_i &\stackrel{\text{iid}}{\sim} G(\nu/2, \nu/2) \\ p(\alpha, \beta, \phi) &\propto 1/\phi \end{aligned}$$

Joint Posterior Distribution:

$$\begin{aligned} p((\alpha, \beta, \phi, \lambda_1, \dots, \lambda_n \mid Y) &\propto \phi^{n/2} \exp \left\{ -\frac{\phi}{2} \sum \lambda_i (y_i - \alpha - \beta x_i)^2 \right\} \times \\ &\phi^{-1} \\ &\prod_{i=1}^n \lambda_i^{\nu/2-1} \exp(-\lambda_i \nu/2) \end{aligned}$$

Model Specification via R2jags

```
rr.model = function() {  
  for (i in 1:n) {  
    mu[i] <- alpha0 + alpha1*(X[i] - Xbar)  
    lambda[i] ~ dgamma(9/2, 9/2)  
    prec[i] <- phi*lambda[i]  
    Y[i] ~ dnorm(mu[i], prec[i])  
  }  
  phi ~ dgamma(1.0E-6, 1.0E-6)  
  alpha0 ~ dnorm(0, 1.0E-6)  
  alpha1 ~ dnorm(0, 1.0E-6)  
}
```

Specifying which Parameters to Save

The parameters to be monitored and returned to R are specified with the variable `parameters`

```
parameters = c("beta0", "beta1", "sigma",  
               "mu34", "y34", "lambda[39]")
```

- ▶ All of the above (except `lambda`) are calculated from the other parameters. (See R-code for definitions of these parameters.)
- ▶ `mu34` and `y34` are the mean functions and predictions for a man with a 34 in waist.
- ▶ `lambda[39]` saves only the 39th case of λ
- ▶ To save a whole vector (for example all lambdas, just give the vector name)

Output

	mean	sd	2.5%	50%	97.5%
beta0	-41.70	2.75	-46.91	-41.67	-36.40
beta1	0.66	0.03	0.60	0.66	0.71
sigma	4.48	0.23	4.05	4.46	4.96
mu34	15.10	0.35	14.43	15.10	15.82
y34	14.94	5.15	4.37	15.21	24.65
lambda[39]	0.33	0.16	0.11	0.30	0.72
95% HPD interval for expected bodyfat (14.5, 15.8)					
95% HPD interval for bodyfat (5.1, 25.3)					

Comparison

- ▶ 95% Probability Interval for β is (0.60, 0.71) with t_9 errors
- ▶ 95% Confidence Interval for β is (0.58, 0.69) (all data normal model)
- ▶ 95% Confidence Interval for β is (0.61, 0.73) (normal model without case 39)

Results intermediate without having to remove any observations
Case 39 down weighted by λ_{39}

Full Conditional for λ_j

$$\begin{aligned} p(\lambda_j \mid \text{rest}, Y) &\propto p(\alpha, \beta, \phi, \lambda_1, \dots, \lambda_n \mid Y) \\ &\propto \phi^{n/2-1} \prod_{i=1}^n \exp \left\{ -\frac{\phi}{2} \lambda_i (y_i - \alpha - \beta x_i)^2 \right\} \times \\ &\quad \prod_{i=1}^n \lambda_i^{\frac{\nu+1}{2}-1} \exp(-\lambda_i \frac{\nu}{2}) \end{aligned}$$

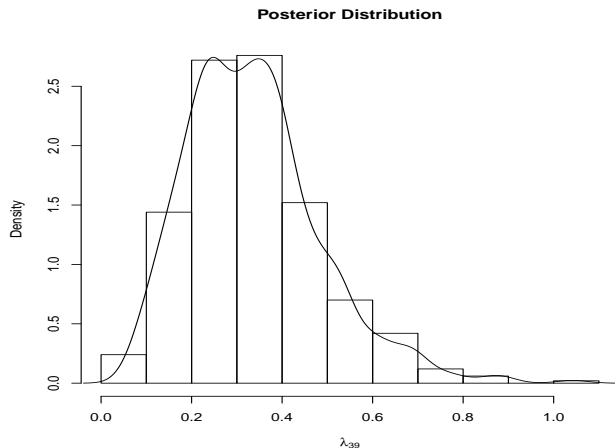
Ignore all terms except those that involve λ_j

$$\lambda_j \mid \text{rest}, Y \sim G \left(\frac{\nu+1}{2}, \frac{\phi(y_j - \alpha - \beta x_j)^2 + \nu}{2} \right)$$

Weights

Under prior $E[\lambda_i] = 1$

Under posterior, large residuals are down-weighted (approximately those bigger than $\sqrt{\nu}$)



Prior Distributions on Parameter

As a general recommendation, the prior distribution should have “heavier” tails than the likelihood

- ▶ with t_9 errors use a t_α with $\alpha < 9$
- ▶ also represent via scale mixture of normals
- ▶ Horseshoe, Double Pareto, Cauchy all have heavier tails

Summary

- ▶ Classical diagnostics useful for EDA (checking data, potential outliers/influential points) or posterior predictive checks
- ▶ BMA/BVS and Bayesian robust regression avoid interactive decision making about outliers
- ▶ Robust Regression (Bayes) can still identify outliers through distribution on weights
- ▶ continuous versus mixture distribution on scale parameters
- ▶ Other mixtures (sub populations?) on scales and β ?