Supervised Models

- Classical Linear Models
  - Linear Regression
  - ANOVA/ANCOVA
- Decision Trees
- Neural Networks
- Deep Learning
- LASSO

	Rain	Yield
[1,]	9.6	24.5
[2,]	12.9	33.7
[3,]	9.9	27.9
[4,]	8.7	27.5
[5,]	6.8	21.7
[6,]	12.5	31.9
[7,]	13.0	36.8
[8,]	10.1	29.9
[9,]	10.1	30.2
[10,]	10.1	32.0
[11,]	10.8	34.0
[12,]	7.8	19.4
[13,]	16.2	36.0
[14,]	14.1	30.2

….

Assume data collected in a single season on n independent plots

Estimate Yield when Rain=10.1

Approach 1:

Mean(29.9,30.2,32.0) = 30.7

Approach 2:

Postulate: mean(Y/X=x) = f(x)

Simple Case: Y scalar, X scalar

General case:

Assume data collected on n independent plots, several predictors X. The dimension of X (number of covariates) may be large compared to n.

Functional data analysis (fda):

- y and/or x assumed smooth functions of time
- Functional linear models:
  - Can handle y and/or x functional or scalar
- Kernels and penalty functionals often used to achieve optimal fit/smoothing, high dimensions, etc.

Ref: Ramsey & Silverman, 2002, "*Applied Functional Data Analysis*"

General case:

Given (X_1,Y_1), ..., (X_n,Y_n), find a function $f \in \mathcal{H}$, that approximates the relationship between X and Y

One approach: Regularization class

Find f which minimizes:

$$\sum_{i=1}^{n} \mathcal{C}(y_i, f(x_i)) + J_\lambda (f)$$

where $\mathcal{C}(y, f)$: a measure of goodness of fit

J_λ (f): Penalty functional

Several special cases available. One early example:

Cubic smoothing spline

(Craven & Wahba, 1979):

- $y \in$ R
- $X \in [0,1]$
- $\mathcal{H}$: Functions with square integrable second derivatives
- $\mathcal{C}(y, f) := (y - f(x))^2$
- $J_\lambda (f) = \lambda \int_0^1 (f''(x))^2 \, dx$

Specific topics

- Linear models
- Robust regression
- Ridge, Lasso, etc.

Simple Linear Regression

Consider the Corn Rain (X) and Corn Yield (Y) data discussed earlier. Denote the individual paired observations by: $(X_i, Y_i), i = 1, \cdots, n$.

Given $X = x$, let the expected value of Y be

$$\mu_{Y/x} = \beta_o + \beta_1 x$$

for any (X_i, Y_i), one has the simple linear regression model:

$$Y_i = \beta_o + \beta_1 X_i + \epsilon_i$$

where ϵ_i are assumed to be uncorrelated random variables, with mean 0 and unknown variance σ^2.

Under the above formulation, the main objectives of regression analysis are:

- To find reasonable estimates of the unknown parameters
- To make inference about the unknown parameters, and
- To assess model adequacy.

Estimation: OLS

Let

$$Q(\beta_o, \beta_1) = \sum_{j=1}^{n} (Y_j - \beta_o - \beta_1 X_j)^2.$$

Then the least squares estimators $\hat{\beta}_0$ and $\hat{\beta}_1$ are such that

$$Q(\hat{\beta}_o, \hat{\beta}_1) = min$$

$$\hat{\beta}_1 = \frac{\Sigma_{j=1}^{n}(X_j - \bar{X})(Y_j - \bar{Y})}{\Sigma_{j=1}^{n}(X_j - \bar{X})^2}$$

and

$$\hat{\beta}_o = \bar{Y} - \hat{\beta}_1 \bar{X}$$

Properties of OLS Estimators

Theorem 1. Under the model assumptions, the OLS estimators are the best, linear unbiased estimators (BLUE).

A reasonable estimator of the error variance, σ^2 is given by

$$\hat{\sigma}^2 = \frac{\Sigma_{j=1}^{n}(Y_j - \hat{Y}_j)^2}{n-2}$$

where $\hat{Y}_j = \hat{\beta}_o + \hat{\beta}_1 X_j$ is the fitted value.

Goodness-of-fit measures

It can be shown that:

$$\sum_{j=1}^{n}(Y_j - \bar{Y})^2 = \sum_{j=1}^{n}(\hat{Y}_j - \bar{Y})^2 + \sum_{j=1}^{n}(Y_j - \hat{Y}_j)^2.$$

Then a measure of goodness of fit is:

$$R^2 = 1 - \frac{\Sigma_{j=1}^{n}(Y_j - \hat{Y}_j)^2}{\Sigma_{j=1}^{n}(Y_j - \bar{Y}_j)^2}$$

-

Values of R^2 close to 1 indicate good fit, while values near 0 suggest lack of fit.

Inference

$$H_o : \beta_1 = c$$

against the alternative

$$H_1 : \beta_1 > c$$

A reasonable test statistic is given by

$$T = \frac{\hat{\beta}_1 - c}{SE(\hat{\beta}_1)}$$

where

$$SE(\hat{\beta}_1) = \frac{\hat{\sigma}}{\sqrt{\Sigma_{j=1}^{n}(X_j - \bar{X})^2}}$$

Under the above assumptions, T has a $t_{(n-2)}$ distribution when H_o holds.

Inference

A $100(1-\alpha)\%$ confidence interval for β_1 is given by

$$\hat{\beta}_1 \pm t_{\alpha/2,n-2} SE(\hat{\beta}_1)$$

Similar results may be obtained for the intercept.

```
corn.reg <- lm(corn.yield ~ corn.rain)
summary(corn.reg)
```

```
Coefficients:
               Value Std. Error t value Pr(>|t|)
(Intercept) 23.5521  3.2365     7.2771  0.0000
  corn.rain  0.7755  0.2939     2.6391  0.0122

Residual standard error: 4.049 on 36 degrees of
Multiple R-Squared: 0.1621
```

Departures from Assumptions

1. Functional form

Diagnosis:

- Look at the scatter plot of the data
- Compute R^2
- Plots of residuals versus X

Corrective measures

- Simple transformations, e.g., log
- Non-linear model
- Other predictors

Departures from Assumptions

2. Non-constancy of the Error Variance

Impact

- OLS estimators not optimal
- Associated inferential results unreliable

Diagnosis

- Plot residuals vs. X, and see whether error variance changes with X
- Variance homogeneity test to the residual variances based on data divided into two groups by values of X.

Corrective measures

- Transformation
- Build variance structure into model: WLS

Departures from Assumptions

3. Non-normality

Impact

- p-values and confidence intervals may not be reliable.

Diagnosis

- Simple graphical displays, e.g., histograms, qqnorm of the residuals
- Goodness-of-t tests, e.g., the Kolmogorov-Smirnov or Shapiro-Wilk test, on the residuals.

Corrective measures

- Transformation
- Robust regression methods

Departures from Assumptions

4. Correlated Errors

Impact

- p-values and confidence intervals may not be reliable.

Diagnosis

- Plot data over time/Look at design of study
- Durbin-Watson test for 1^{st} order AR

Corrective measures

- Transformation: Cochrane-Orcutt Procedure
- Use models that incorporate the correlation structure
  - Generalized Estimating Equations (GEE)

To construct the Durbin-Watson test, let

$$\epsilon_j = \rho\epsilon_{j-1} + u_j$$

where u_j is $N(0, \sigma^2)$, and $\rho = cor(\epsilon_j, \epsilon_{j-1})$. We wish to test

$$H_o : \rho = 0$$

against

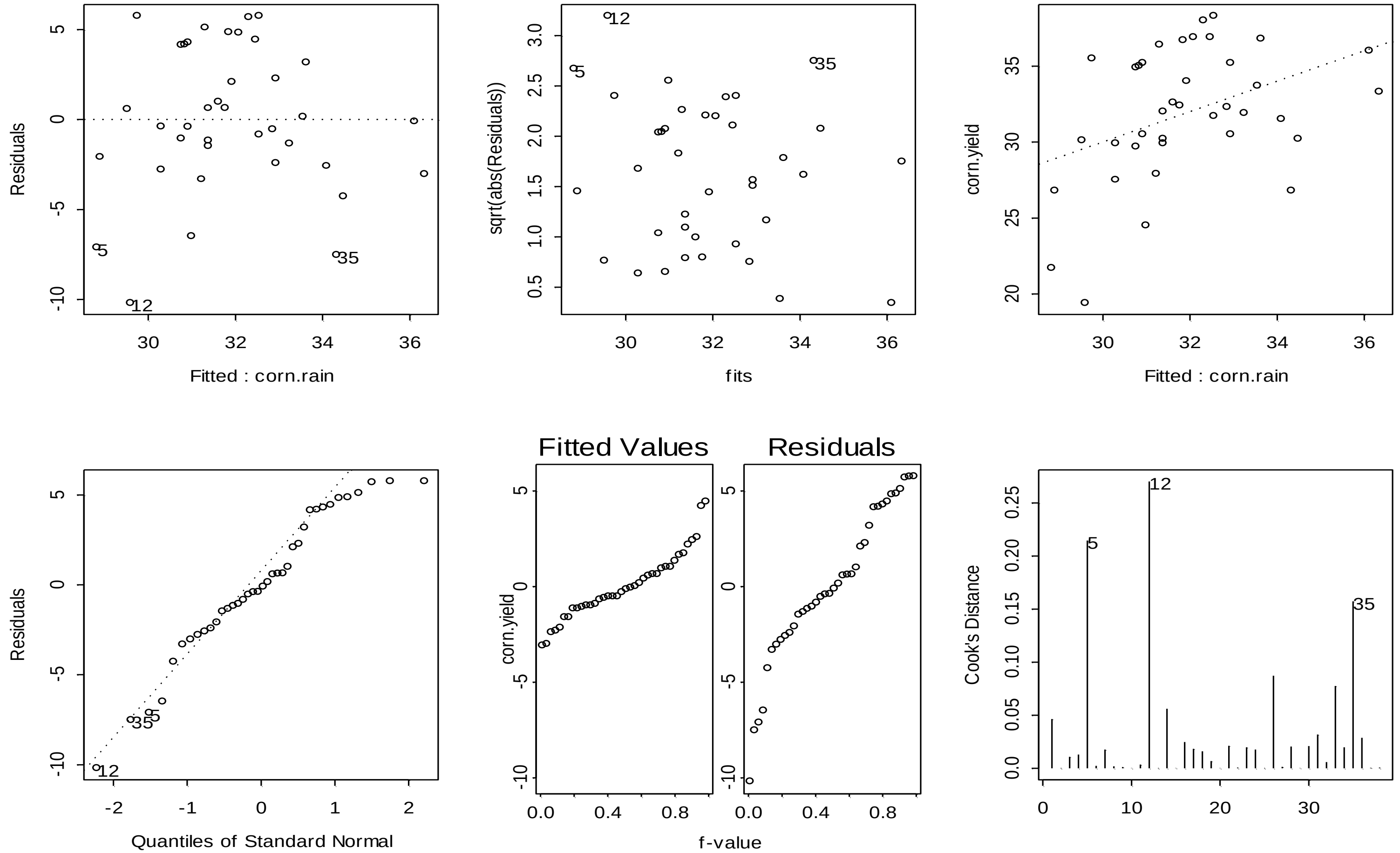
$$H_1 : \rho > 0$$

Put

$$D = \frac{\Sigma_{j=2}^{n}(\hat{\epsilon}_j - \hat{\epsilon}_{j-1})^2}{\Sigma_j \, \hat{\epsilon}_{j=1}^2}$$

The Durbin-Watson test rejects H_o when D is too small. Tables of critical values are available

```
> corn.reg <- lm(corn.yield~corn.rain)
> plot(corn.reg)
```



Estimation and Prediction

Let x_o be a value of the explanatory variable.

Then an estimator of the mean of Y corresponding to $X = x_o$ is

$$\hat{Y}_o = \hat{\beta}_o + \hat{\beta}_1 x_o$$

and has estimated SE given by

$$\hat{\sigma}\sqrt{\left(1/n + \frac{(x_o - \bar{X})^2}{\Sigma(X_i - \bar{X})^2}\right)}$$

Estimation and Prediction

Next suppose we wish to predict a future value of Y corresponding to $X = x_{new}$. This is given by

$$\hat{Y}_{new} = \hat{\beta}_o + \hat{\beta}_{\mathbf{1}} x_{new}$$

and has estimated SE given by

$$\hat{\sigma}\sqrt{\left(1 + 1/n + \frac{(x_o - \bar{X})^2}{\Sigma(X_i - \bar{X})^2}\right)}$$

Regression Through the Origin

An appropraite model is

$$Y_i = \beta_1 X_i + \epsilon_i$$

$$\hat{\beta}_1 = \frac{\Sigma_i X_i Y_i}{\Sigma_i X_i^2}$$

$$\hat{\sigma}^2 = \frac{\Sigma_i \epsilon_i^2}{n-1}$$

.

It follows that

$$SE(\hat{\beta}) = \frac{\hat{\sigma}}{\sqrt{\Sigma_i X_i^2}}$$

Inverse Prediction (Calibration)

Given a new observation Y_{new}, we wish to estimate the corresponding X_{new}. Under normality, the MLE is given by

$$\hat{X}_{new} = (Y_{new} - \hat{\beta}_o)/\hat{\beta}_{\mathbf{1}}$$

The variance is estimated by

$$\hat{\sigma}^2_{\hat{X}_{new}} = \frac{\hat{\sigma}^2}{\hat{\beta}^2_{\mathbf{1}}}[1 + 1/n + \frac{(\hat{X}_{new} - \bar{X})^2}{\Sigma_i(X_i - \bar{X})^2}$$

Multiple Regression

Given Y, a dependent variable, and $X_1, \cdots, X_p$, p explanatory variables, the mean of Y given $X_j = x_j, j = 1, \cdots, p$, may be expressed, under the linear model assumption, as

$$\mu_{Y|x_1,\cdots,x_p} = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p$$

Let $(Y_i, X_{i,1}, \cdots, X_{i,p})$, $i = 1, \cdots, n$, be a random sample. It is often convenient to use the corresponding matrix formulations for the model:

$$\mathbf{Y} = \mathbf{X}\beta + \epsilon,$$

where $\mathbf{Y}$ is $n \times 1$, β is $p+1 \times 1$, $\mathbf{X}$ is $n \times p+1$, and ϵ is $n \times 1$. As in the simple linear model case, the error terms are assumed to be uncorrelated, with mean 0 and constant variance σ^2.

Then under the model assumptions, the LS estimators are obtained as solutions to the normal equations:

$$\mathbf{X}'\mathbf{X}\beta = \mathbf{X}'\mathbf{Y}$$

When $\mathbf{X}$ is full rank, the BLUE is given by

$$\hat{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$$

The Gauss-Markov Theorem.

Under the model assumptions, OLSE is UMVUE in the class of linear unbiased estimators.

Assume **X** (nxp+1) is of full rank. Then

$$Var(\hat{\beta}) = (\mathbf{X}'\mathbf{X})^{-\mathbf{1}}\sigma^{\mathbf{2}}$$

The distribution of $\hat{\beta}$ is (p+1) variate normal

As in the simple linear model case, a reasonable estimator of σ^2 is

$$\hat{\sigma}^2 = \frac{\Sigma_{j=1}^{n}(Y_j - \hat{Y}_j)^2}{n-p-1}$$

Inference

Consider the testing problem of $H_o : \beta_k = \beta_{k,o}$

$$T = \frac{(\hat{\beta}_k - \beta_{k,o})}{SE(\hat{\beta}_k)}$$

Null distribution?

Similarly, a $100(1-\alpha)\%$ confidence interval for β_k is given by

$$\hat{\beta}_k \pm T_{n-p-1,\alpha/2} SE(\hat{\beta}_k)$$

Outliers and Influential Points

$$\mathbf{H} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$$

$$\mathbf{h} = \mathbf{diag}(\mathbf{H})$$

The hat matrix plays an important role in model diagnostics.

h_{ii} is sometimes referred to as the leverage of the ith case. Recall that $0 \leq h_{ii} \leq 1$, and $\Sigma_i h_{ii} = p + 1$. When h_{ii} is large, i.e., $h_{ii} > 2(p+1)/n$, the ith case is said to have a high leverage in determining $\hat{Y}_i$.

$$var(\hat{\epsilon}_j) = (1 - h_{jj})\sigma^2$$

Outliers and Influential Points

- Semi-studentized residuals

  Let
  $$T_j^* = \frac{\hat{\epsilon}_j}{\hat{\sigma}}$$
  Large values of T_j^* may indicate outliers in the Y values.

- Studentized Residuals

  Define
  $$T_j = \frac{\hat{\epsilon}_j}{\hat{\sigma}\sqrt{1 - h_{jj}}}$$

Deleted Residuals

Let $\hat{Y}_{(i)}$ be the fitted value obtained after deleting the ith record. It can be shown that the deleted residual

$$\hat{\epsilon}_{(i)} = \frac{\hat{\epsilon}_i}{\sqrt{1-h_{ii}}}$$

with

$$SE(\hat{\epsilon}_{(i)}) = \frac{\hat{\sigma}_{(i)}}{\sqrt{1-h_{ii}}}$$

$$(n-p-1)\hat{\sigma}^2 = (n-p-1)\hat{\sigma}^2_{(i)} - \frac{\hat{\epsilon}_i^2}{(1-h_{ii})}$$

Studentized deleted residuals

$$T_{(i)} = \frac{\hat{\epsilon}_{(i)}}{SE(\hat{\epsilon}_{(i)})} = \frac{\hat{\epsilon}_i}{\hat{\sigma}_{(i)}\sqrt{1-h_{ii}}}$$

The following Bonferroni confidence set may be used to identify Y-outliers:

$$|\, T_{(i)} \,| > t_{\alpha/2n, n'-p-1}$$

where $n' = n - 1$.

DFFITS

The influence of the ith observation on the fitted value $\hat{Y}_i$ may be assessed based on

$$DFFITS_i = \frac{\hat{Y}_i - \hat{Y}_{(i)}}{\hat{\sigma}_{(i)}\sqrt{h_{ii}}}$$

$$= T_{(i)} \left(\frac{h_{ii}}{1 - h_{ii}} \right)^{1/2}$$

The ith case is considered influential if $|DFFITS_i| > 1$ for small to medium n, and exceeds $2\sqrt{\frac{p+1}{n}}$ for large n.

Cook's Distance

An aggregate measure of influence

Cook's Distance, defined as

$$D_i = \frac{(\hat{\mathbf{Y}} - \hat{\mathbf{Y}}_{(\mathbf{i})})'(\hat{\mathbf{Y}} - \hat{\mathbf{Y}}_{(\mathbf{i})})}{(p+1)\hat{\sigma}^2}$$

which is distributed as $F_{p+\mathbf{1},n-p-\mathbf{1}}$. The ith case is said to be an influential point over all the n fitted values, if $D_i > F_{\alpha,p+\mathbf{1},n-p-\mathbf{1}}$.

DFBETAS

$$DFBETAS_{k(i)} = \frac{\hat{\beta}_k - \hat{\beta}_{k(i)}}{\hat{\sigma}_{(i)}\sqrt{c_{kk}}}$$

where c_{kk} is the kth diagonal element of $\mathbf{(X'X)^{-1}}$

Large values of $|DFBETAS_{k(i)}|$ indicate the influence of the i^{th} case on the k^{th} regression coefficient estimate.

Typically > 1 or 2/sqrt(n), depending on n

```
> fit1 <-lm(Y~x1+x2)
> summary(fit1)
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.94425    4.11174  -0.230 0.839714
x1           3.22012    0.33886   9.503 0.010893 *
x2           6.05792    0.08339  72.643 0.000189 ***
```

> lmi <- lm.influence(fit1)

> names(lmi)

```
[1] "coefficients" "sigma"        "hat"
```

> names(lms)

```
 [1] "call"         "terms"        "residuals"    "coefficients" "sigma"
 [6] "df"           "r.squared"    "fstatistic"   "cov.unscaled" "correlation"
```

>lms <- summary(fit1)

```
e <- resid(fit1)
s <-lms$sigma
si <-lmi$sigma
xxi <-diag(lms$cov.unscaled)
h <-lmi$hat
```

```
coef(lmi)
 (Intercept)          x1           x2
1 -15.0108430  0.20726437  0.468425836
2  -3.8149505  0.28318345  0.058954961
3  11.7849186 -1.70357945 -0.184475162
4   0.4542068 -0.02554101 -0.007153742
5   2.3582137 -0.08404465 -0.044319891
```

```
bi <- coef(fit1)-t(coef(lmi))
dfbetas  <- bi/t(si%o%xxi^0.5)
 stand.resid <- e/(s*(1-h)^0.5)
student.resid <-  e/(si*(1-h)^0.5)
DFFITS <- h^0.5*e/(si*(1-h))
```

```
> x <- cbind(x1,x2)
> Y2 <- Y
> Y
[1] 350.3001 203.9001 202.2531 202.7426 177.9737
```

```
> fit1 <-lm(Y~x1+x2)
> coef(fit1)
(Intercept)          x1          x2
 -0.9442516   3.2201178   6.0579185
```

```
> Y2 <- Y

> Y2
[1]   350.3001   203.9001   202.2531   202.7426 17797.3689
> fit2 <-lm(Y2~x1+x2)
coef(fit2)

(Intercept)          x1          x2
 19118.9442   -678.1958   -353.2782
```

Robust Regression

The goals of robust regression are:

- To perform as well as the OLS when the latter works.
- To perform better than the OLS when the latter fails.
- Not complex to compute or understand.

$$\epsilon_i(\beta) = Y_i - \beta_0 - \beta_1 X_{i1} - \cdots - X_{ip}$$

Least Absolute Deviation (L1) Regression

- Find estimators which minimize:

$$\sum_{i}^{n} \mid \epsilon_i(\beta) \mid$$

Remarks:

- Minimization is not as straightforward as the LS case, and may require linear programming techniques.
- Sum of the residuals may not be 0.
- Estimators may be susceptible to high leverage points

:

Least Median of Squares Regression

The least median of squares regression finds estimators which minimize

$$median\{\epsilon_i^2(\beta), i = 1, \cdots, n\}$$

- The procedure has a high (50%) breakdown point.
- Computation cumbersome

:

Least Trimmed Squares of Regression

Denote the ith ordered residual squared by $\epsilon^2_{(i)}(\beta)$. Then the least trimmed squares robust regression minimizes the trimmed sum:

$$\sum_{i=1}^{q} \epsilon^2_{(i)}(\beta)$$

where q is a suitably chosen trimming quantity.

Remarks

- Relatively high breakdown point, but < 50%.
- Calculation is complex, and uses random algorithms to get approximate solutions.

M-Estimates of Robust Regression

Given an objective function $\rho()$, M-estimates of robust regression estimates are obtained minimizing

$$\sum_{i=1}^{n} \rho\left(\frac{\epsilon_i(\beta)}{\sigma}\right)$$

Remarks

- When $\rho(x) = x^2$, we get the OLS, whereas $\rho(x) = |x|$, gives L1 regression estimates.
- The procedure protects against Y outliers, but may be sensitive to leverage points in $\mathbf{X}$.

M-Estimates of Robust Regression

- Compared to trimmed regression, easier to compute. Computation involves iterated weighted least squares, with weights given by

$$w_i = \frac{\rho'\left(\frac{\epsilon_i(\beta)}{\sigma}\right)}{\frac{\epsilon_i(\beta)}{\sigma}}$$

  – Huber: Quadratic in the center, but linear in the tails.

$$\rho(u) = \frac{u^2}{2}, \mid u \mid \leq k$$

$$= k \mid u \mid - \frac{k^2}{2}, \mid u \mid > k$$

```
> x <- cbind(x1,x2)
> Y2 <- Y
> Y
[1] 350.3001 203.9001 202.2531 202.7426 177.9737

> Y2[5] <- Y[5] *100
> Y2
[1]   350.3001   203.9001   202.2531   202.7426 17797.3689
```

```
> fit1 <-lm(Y~x1+x2)
> fit2 <-lm(Y2~x1+x2)
> coef(fit1)
(Intercept)          x1          x2
 -0.9442516   3.2201178   6.0579185

> coef(fit2)
(Intercept)          x1          x2
 19118.9442   -678.1958   -353.2782
```

```
> fit1.lms <-lmsreg(x,Y)
> fit2.lms <-lmsreg(x,Y2)
> coef(fit1.lms)
(Intercept)          x1          x2
 -12.153527    4.888467    6.227675
> coef(fit2.lms)
(Intercept)          x1          x2
  -5.794811    3.426881    6.144069
```

```
> fit1.rreg <- rlm(x,Y)
> fit2.rreg <- rlm(x,Y2)
> coef(fit1.rreg)
      x1       x2
3.106378 6.047713
> coef(fit2.rreg)
      x1       x2
3.306668 6.042112
```

R functions

library(MASS)

help(lqs)

lqs(x,y,method=“lts”,”lqs”,”lms”,”S”)

lmsreg()

ltsreg()

huber(); rlm()

Problem Set

Reading Assignment:

Chapter 7,8, and 10: The Statistical Sleuth: A Course in Methods of Data Analysis. Ramsey & Schafer

Consider the *Pima.te dataset, in R library MASS, on Diabetes in Pima Indian Women*.

a) Fit a multiple linear regression model of predict 'glu', plasma glucose concentration in an oral glucose tolerance test, using the following set of predictors:

- 'npreg' number of pregnancies
- 'bp' diastolic blood pressure (mm Hg)
- 'skin' triceps skin fold thickness (mm)
- 'bmi' body mass index (weight in kg/(height in m)^2)
- 'age' age in years

b) State and assess the validity of the underlying assumptions:

- Linearity/functional form, including the need for any interaction terms
- Normality
- Homoscedasticity
- Uncorrelated error, and
- Check for outliers and influential points.

c) Propose remedial measures in case of violations of any of the underlying assumptions

c) Repeat (a) using Least Median of Squares Regression and compare the results with those obtained in (a).