HW -10: Today (4-23)

Quiz - 8 :Today (4-23)

Class - 13

Final Exam:
Covers Chapters 6-11,
No electronic devices except a calculator,
One double-sided hand written sheet is allowed.

Need to Know:
How to read a computer output for SLR model, ANOVA table, etc.
How to find table values for Standard normal, t , and F-distributions.

Extra Office Hours:
T : 11am - 12 pm, 1-2 pm,
W: 3 pm-4.30 pm,
R : 2 pm - 5 pm,
F : 3 pm - 5 pm.
M : 10.30 am -12 pm, 4 - 5.20 pm

MA 542 SPRING 2018

Applied Regression Analysis

Chapter 11

Building the Regression Model III: Remedial Measures

Remedial Measures

When the fitted regression model is not appropriate or one or several cases are very influential, remedial measures may be used to fix those.

Remedial measures so far: Transformations

- to linearize the regression relation. (X)
- to make the error distribution (nearly) normal. (Y)
- to make the error variance (nearly) constant. (Y)

Remedial measures in this chapter:

- deal with unequal error variance.
- deal with high degree multicollinearity.
- deal with influential observations.

Nonparametric Methods: 1. lowess, 2. Regression trees

General Approach: Bootstrapping. .

Unequal Error Variances

- **So far in Ch 3 and 6 :** Transformation of Y: reducing or eliminating unequal variances.
- **Limitation:** sometimes they may create an inappropriate regression relationship.
- **Alternative Method :** Weighted Least Squares : procedure based on a generalization of multiple regression model.

Consider the generalized multiple regression model

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots \beta_n X_{in} + \epsilon_i, \quad i = 1, 2, \cdots, n.$$

where $\epsilon_i \overset{ind}{\sim} N(0, \sigma_i^2)$ $\quad \sigma^2 I = \begin{bmatrix} \sigma^2 & & \\ & \sigma^2 & \\ & & \ddots & \\ & & & \sigma^2 \end{bmatrix}$

$$\Rightarrow \sigma^2_{n\times n}\{\epsilon\} = \begin{bmatrix} \sigma_1^2 & 0 & \cdots & 0 \\ 0 & \sigma_2^2 & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & \sigma_n^2 \end{bmatrix}$$

Weighted Least Squares

- **Least Squares Estimators :** $b = (X'X)^{-1}X'Y \Rightarrow$ unbiased, consistent but dont have minimum variance.
- To obtain unbiased estimators with minimum variance, we have to use different weights for different Y observations.
- We define the reciprocal of the variance σ_i^2 as the weight w_i (i.e. $w_i = 1/\sigma_i^2$).

Here we consider three cases:

- Error Variances Know $(w_i = 1/\sigma_i^2)$(unrealistic)
- Error Variances Know up to Proportionality Constant $(w_i = k(1/\sigma_i^2))$
- Error Variances Unknown: estimation of variance function or standard deviation function $(w_i = \frac{1}{(\hat{s}_i)^2}, \quad w_i = \frac{1}{\hat{v}_i})$

Weighted Least Squares (σ_i^2 known))

Likelihood function

$$
\begin{aligned}
L(\beta) = & \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma_i^2}} exp \left[-\frac{1}{2\sigma_i^2} (Y_i - \beta_0 - \beta_1 - \cdots - \beta_{p-1} X_{i,p-1})^2 \right] \\
= & \prod_{i=1}^{n} \sqrt{\frac{w_i}{2\pi}} exp \left[-\frac{1}{2} \underbrace{\sum_{i=1}^{n} w_i (Y_i - \beta_0 - \beta_1 - \cdots - \beta_{p-1} X_{i,p-1})^2}_{Q_w} \right]
\end{aligned}
$$

Maximum likelihood estimator:

$$b = arg \ max \ L(\beta)$$

$$\Rightarrow b = arg \ min \ Q_w$$

where $Q_w = \sum_{i=1}^{n} w_i (Y_i - \beta_0 - \beta_1 X_{i,1} - \cdots - \beta_{p-1} X_{i,p-1})^2$.

- Here b is called weighted least squares estimator and it same as the maximum likelihood estimator.
- Ordinary least square (OLS) is a special case of this and for OLS $w_i = 1$ for all i.

Weighted Least Squares (σ_i^2 known)

- $w_i = \frac{1}{\sigma_i^2}$: reflects the amount of information contained in the observation Y_i. ($Var(Y_i) = \sigma_i^2 \uparrow \Rightarrow w_i \downarrow$)

$X'X b = X'y$

- The normal equations: $(X'WX)b_w = X'WY$
  where

$$W_{n\times n} = \begin{bmatrix} w_1 & 0 & \cdots & 0 \\ 0 & w_2 & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & w_n \end{bmatrix}$$

- The weighted least squares estimator (MLE) for β:

$$\left.\begin{aligned} b_w &= (X'WX)^{-1}(X'WY) \\ \Rightarrow \sigma^2_{p\times p}\{b_w\} &= (X'WX)^{-1} \end{aligned}\right\}$$

- b_w : unbiased, consistent, have minimum variance among unbiased linear estimators.

WLS (σ_i^2 known up to Proportionality Constant)

Here we assume that the relative magnitudes of the variances are known.

$$w_i = k\frac{1}{\sigma_i^2},$$ where **k** is proportionality constant.

- The weighted least squares estimator (MLE) for β:

MSE

$$b_w = (X'WX)^{-1}(X'WY) \quad (unaffected)$$

$$\Rightarrow \sigma^2_{p\times p}\{b_w\} = k(X'WX)^{-1}$$

$\sigma^2\{b\} = \sigma^2(X'X)^{-1}$

- When k is unknown : $\sigma^2_{p\times p}\{b_w\}$ can be estimated by

$$s^2_{p\times p}\{b_w\} = MSE_w(X'WX)^{-1}$$

where

$$MSE_w = \frac{\sum w_i(Y_i - \hat{Y}_i)^2}{n-p} = \frac{\sum w_i e_i^2}{n-p}$$

Res for OLS

- MSEw : an estimator of the proportionality constant k.

WLS (σ_i^2 unknown)

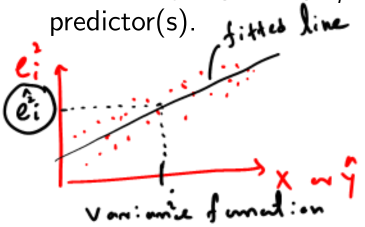
Here we estimate σ_i^2 by e_i^2 where e_i : i^{th} least square residual.
Further σ_i is estimated by $|e_i|$.

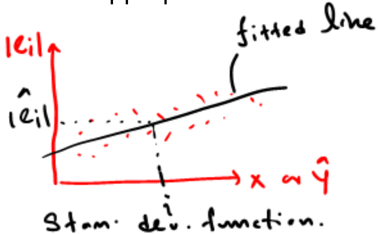
$\hat{\sigma}_i^2 = |e_i|^2$

$$\hat{\sigma}_i^2 = Var(\epsilon_i) = E(\epsilon_i^2) - [\overset{0}{\overset{\|}{E(\epsilon_i)}}]^2 = E(\epsilon_i^2)$$

There for we can estimate the variance (σ_i^2) or the standard deviation (σ_i) as a function of relevent predictor variables following the process:

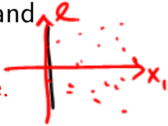
1. Fit the regression model by OLS and analyze e_i
2. Estimate the variance function or the standard deviation function by regressing e_i^2 or $|e_i|$ on the appropriate predictor(s).





WLS (σ_i^2 unknown)

The following illustrates the use of some possible variance and standard deviation functions:



1. A residual plot against X_1 exhibits a megaphone shape. Regress the absolute residuals against X_1.
2. A residual plot against $\hat{Y}$ exhibits a megaphone shape. Regress the absolute residuals against $\hat{Y}$.
3. A plot of the squared residuals against X_4 exhibits an upward tendency. Regress the squared residuals against X_4.

4. A plot of the residuals against X_2 suggests that the variance increases rapidly with increases in X_2 up to a point and then increases more slowly. Regress the absolute residuals against X_2 and X_2^2.

WLS (σ_i^2 unknown)

After the variance function or the standard deviation function is estimated, the fitted values from this function are used to obtain the estimated weights:

$$w_i = \frac{1}{\hat{s}_i^2} \qquad \text{OR} \qquad w_i = \frac{1}{\hat{v}_i}$$

where s_i is fitted value from standard deviation function and $\hat{v}_i$ is fitted value from variance function.

- The weighted least squares estimator (MLE) for β:

$$b_w = (X'WX)^{-1}(X'WY)$$
$$\Rightarrow \sigma^2_{p\times p}\{b_w\} = (X'WX)^{-1}$$

where $W = diag\{w_1, w_2, \cdots, w_n\}$.

Note: If the estimated coefficients differ substantially from OLS coefficients : do severel iterations (iteratively reweighted lease squares) .

Multicollinearity Remedial Measures (Ridge Regression)

LSM

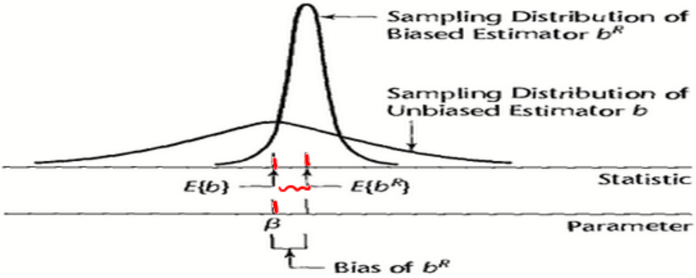
- Ridge Regression is a modified version of least square method that overcomes multicollinearity problems.

$\lim_{n\to\infty} \hat{\theta}_n = \theta$

$MSE(\hat{\theta}) = Var(\hat{\theta}) + (Bias(\hat{\theta}))^2$

- Ridge Regression estimators (b^R) are biased but more precise than OLS estimators (b). i.e. $MSE(b^R) << MSE(b)$ where $MSE(b^R)$: mean square error of b^R and

$$MSE(b^R) = E\{b^R - \beta\}^2 = \sigma^2\{b^R\} + (bias\{b^R\})^2$$



Ridge Estimators

Consider the transformed (stanrardized) variables;

$$Y_i^* = \frac{1}{\sqrt{n-1}}\left(\frac{Y_i - \bar{Y}}{s_Y}\right), \qquad X_{ik}^* = \frac{1}{\sqrt{n-1}}\left(\frac{X_{ik} - \bar{X}_k}{s_k}\right),$$

where $k = 1, 2, \cdots, p-1$.

Normal equations for OLS: $\boxed{r_{XX} b = r_{YX}}$

where r_{XX} - correlation matrix of X^* and

r_{YX} - vector of correlations between Y^* and each X^*.

Ridge estimators are given by

$$(r_{XX} + c\mathbb{I})b^R = r_{YX}$$

$$c\mathbb{I} = \begin{bmatrix} c & & & \\ & c & & \\ & & c & \\ & & & c \end{bmatrix}$$

where $b^R = [b_1^R, b_2^R, \cdots, b_{p-1}^R]^T$ and $\mathbb{I}$ is $(p-1) \times (p-1)$ identity matrix.

$$\Rightarrow b^R = (r_{XX} + c\mathbb{I})^{-1} r_{YX}.$$

Note : • c: amount of bias in the estimator.

• When $c = 0$, $b^R = b$ (OLSE) .

Choice of Biasing Constant c

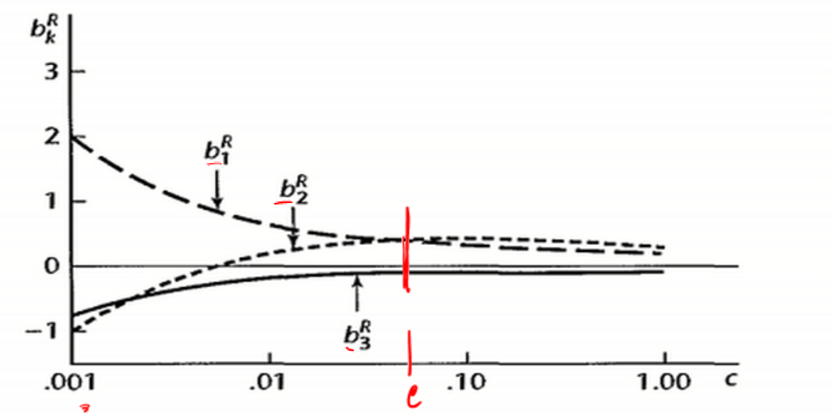
- There exist a value of c such that $MSE(b^R) < MSE(b)$.
- Bias $(b^R) \uparrow$ as $c \uparrow$ while $\sigma^2\{b^R\} \downarrow$.
- Difficulty : the optimum value of c varies from one application to another and is unknown.

Commenly used methods:

- Ridge trace : The ridge trace is a simultaneous plot of the values of b^R for different values of c, usually between 0 and 1.
- $(VIF)_k$: variance inflation factors

Choice of Biasing Constant c

Choose the value of c at which the ridge trace start to become stable and $(VIF)_k$ becomes sufficiently small.
PICTURE



Remedial Measures for Influential Case

5' 8' → $\underbrace{8' \; 5''}$ 5' 4'' → 4' 5''

- outlier
  - → due to an error
    - → can correct it → correct it (use)
    - → can't corret it → dis card
  - → not Sure
    - → influential → change the modul or discard
    - → not influential → keep it.

Alternative Methods: Robust Regression Methods.

Robust regression procedures incoporate the influence of outlying cases and provide a better fit for majority of the cases.

Robust Regression Methods

LAR: Least absolute residuals or least absolute deviations (LAD) regression,

$\sum(Y_i-(\beta_0+\beta_1X_{i1}+\cdots\beta_{p-1}X_{i\,p-1})^*)^2$

Regression coefficients are estimated by minimizing

$$L_1=\sum_{i=1}^{n}|Y_i-(\beta_0+\beta_1X_{i1}+\beta_2X_{i2}+\cdots+\beta_{p-1}X_{p-1})|.$$

LMS Regression: Least median of squares regression

Regression coefficients are estimated by minimizing

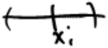
$$median\{[Y_i-(\beta_0+\beta_1X_{i1}+\beta_2X_{i2}+\cdots+\beta_{p-1}X_{p-1})]^2\}.$$

IRLS Regression: Iteratively reweighted least squares regression

This method with an appropriate weight function can also be used to reduc the influence of outlying cases.

Nonparametric Regression

- Non parametric regression estimates the response surface considering X levels seperately.



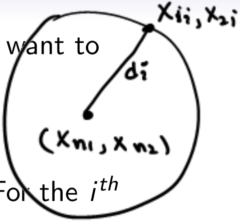
- Fitted value for each X level consider only a part of the data set that are nearest to the point (neighborhood).
- Unlike parametric regression, no analytic expression for the response surface is provided by nonparametric regression.
- Nonparametric regression fits are useful for exploring the nature of the response function.

Here we discuss two nonparametric Regression Methods:

1. Lowess Method
   (locally weighted regression scatter plot smoothing method)
2. Regression Trees

Lowess Method

Consider two predictor variable case and suppose we want to obtain the fitted value for $(X_{h1},\ X_{h2})$.



- **Distance Measure :**

Usually, a Euclidean distance measure is employed. For the i^{th} case, the distance is

100 1000

$$d_i = [(X_{i1} - X_{h1})^2 + (X_{i2} - X_{h2})^2]^{1/2}$$

- When the predictor variables are measured on different scales, each should be scaled by dividing it by its standard deviation.
- The median absolute deviation estimator can be used in place of the standard deviation if outliers are present.

$$MAD = \frac{|Y_i - \bar{Y}|}{n-1} \qquad \sqrt{S^2} = \sqrt{\frac{\Sigma (Y_i - \bar{Y})^2}{n-1}}$$

Lowess Method (Weight Function)

- The neighborhood about the point (X_{h1}, X_{h2}) is defined in terms of the proportion q of cases that are nearest to the point.
  50%

- Let d_q denote the Euclidean distance of the furthest case in the neighborhood. Then the weight function is

d_1
d_2
$50\% \longrightarrow d_q$
d_n

$$\boxed{w_i = \begin{cases} [1 - (d_i/d_q)^3]^3 & d_i < d_q \\ 0 & d_i \geq d_q \end{cases}}$$

- The choice of the proportion q:
  - larger q: smoother fit (but may be a bias fit).
  - smaller q: Unbiased fit (but fitted surface may not be smooth).

A choice of q between .4 and .6 may often be appropriate.

Lowess Method (Local Fitting)

$\hat{Y}$

- Weighted least squares (first order or second order based on the cases in the neighborhood) used to fit the model.

- The fitted value $\hat{Y}_h$ at (X_{h1}, X_{h2}) is the estimate of the mean response at this X level.

- Repeat the procedure for all the X levels.

we obtain information about the response surface without making any assumptions about the nature of the response function.

.

Regression Trees

Regression trees are a very powerful, but conceptually simple, nonparametric methods.

- **Steps:**

  1. X space is partitioned into sub regions.

  - For predictor: the range of X is partitioned into intervals.
  - For two or more predictors : X space is partitioned into rectangular regions.

  2. The fitted value for each region is estimated by mean of the responses in the region.

$(x'x)^{-1}$

This method is better for extremely large data sets.

.

Remedial Measures for Evaluating Precision in Nonstandard Situations

- In many nonstandard situations, (ex: nonconstant error variances estimated by iteratively reweighted least squares), standard methods for evaluating the precision may not be available
  OR
- these methods may only be approximately applicable when the sample size is large
- Bootstrapping: provide estimates of the precision of sample estimats for these complex cases

$\theta \longrightarrow \hat{\theta}$

Bootstrapping (General)

Let $\hat{\theta}$ be an estimated model parameter of θ from some nonstandard method. The following are the steps for bootstrapping.

$Y_1, Y_2, \ldots\ Y_n \quad , \quad \boxed{Y_1^*, Y_2^* \cdots\ Y_n^*}$

1. Selecting the bootstrap sample : select a random sample of size n with replacement from observed sample. (This may contain some duplicates and may omit some other data)

2. Estimate the model parameter using the bootstrap sample and using the same method ($\hat{\theta}^*$).

3. Repeat large number of times. $\hat{\theta}_1^*, \ldots\ldots\ \hat{\theta}_{500}^*$

4. Calculate the standard deveation $s^*\{\hat{\theta}^*\}$ of $\hat{\theta}^*$s.

Then $s^*\{\hat{\theta}^*\}$ can be used as an estimate of the variability of the sampling distribution of $\hat{\theta}$.

Bootstrapping in Regression

β

There are two basic ways to obtain a bootstrap sample in regression. Let the parameter of interest is β_1 and $\hat{\beta}_1 = b_1$.

1. When the regression function being fitted is a good model for the data, the error terms have constant variance, and fixed X sampling is appropriate.

$\{e_1, e_2, \cdots e_n\}$

- Obtaine the residuals e_is from the original fit and take a bootstrap sample of size n from residuals $\{e_1^*, e_2^*, \cdots, e_n^*\}$.
- Calculate $Y_i^* = \hat{Y}_i + e_i^*$
- Regress Y^* on Xs and obtained the estimate for β_1 (b_1^*)
- Repeat the process.

$\underbrace{b_{11}^*, b_{12}^*, b_{13}^* \cdots b_{1500}^*}$

Bootstrapping in Regression

2. When there is some doubt about the adequacy of the regression function being fitted, the error variance are not constant, and/or random X sampling is appropriate.

( )

- Take a bootstrap sample of size n from the original sample.
  $\{(X_1^*, Y_1^*),\ (X_2^*, Y_2^*),\ \cdots,\ (X_n^*, Y_n^*)\}$
- Regress Y^* on X^*s and obtained the estimate for β_1 (b_1^*)
- Repeat the process.

Y_1, X_1
Y_2, X_2
$\vdots$
Y_n, X_n

Once a set of b_1^*s is obtained, calculate the standard deviation $s^*\{b_1^*\}$ and that can be used as an estimate of the variability of the sampling distribution of b_1.

Bootstrap Confidence Intervals

- From the bootstrap distribution of b_1^* , find the $\alpha/2$ and $1\alpha/2$ quantiles $b_1^*(\alpha/2)$ and $b_1^*(1-\alpha/2)$

$b^*_{11}, b^*_{12} \cdots\cdots b^*_{1500}$

- $1-\alpha$ bootstrap confidence interval :

$$(\ b_1^*(\alpha/2),\quad b_1^*(1-\alpha/2)\ )$$

- Large number of bootstrap samples (about 500) are required for these confidence intervals.