

PubH 7405: REGRESSION ANALYSIS



MLR: MORE DIAGNOSTICS & SOME REMEDIES

The data are in the form :

$$\{(y_i; x_{1i}, x_{2i}, \dots, x_{ki})\}_{i=1, \dots, n}$$

Multiple Regression Model :

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_k x_k + \varepsilon$$

$$\varepsilon \in N(0, \sigma^2)$$

**Do data fit the Normal Error Regression Model?
If not , could we do something to make them fit?**

REMEDIAL MEASURES FOR NONLINEARITY

- If a “linear model” is found not appropriate for the added value of certain predictor, there are two choices:
 - (1) Add in a power terms (**quadratic**), or
 - (2) Use some **transformation** on the data to create a fit for the transformed data

Each has advantages & disadvantages:
first approach (quadratic model) may
yield better insights but may lead to more
technical difficulties; transformations
(log, reciprocal, etc...) are more simple
but may obscure the fundamental real
relationship between Y and that predictor

It's **more time-consuming to detect non-linearity**, but it's **more simple to fix it**: A log transformation of an X or addition of its quadratic term would normally solve the problem on non-linearity.

It's **more simple to detect a non-constant variance**; added-value plot is not needed. However, it would be **more difficult to fix** because, in order to change the variance of Y but we to make a transformation on Y .

Transformations of Y maybe helpful in reducing or eliminating unequal variances of the error terms.

However, **a transformations of Y also changes the regression relation/function**. In many circumstances an appropriate **linear regression relationship has been found** but the variances of the error terms are unequal; a transformation would **make that linear relationship non-linear** which is a more severe violation.

An alternative to data transformations: – which are more difficult to find - using method “**weighted least squares**” instead of regular least squares.

With the **Weighted Least Squares (WLS)**, estimators for regression coefficients are obtained by minimizing the quantity Q_w where “w” is a “weight” (associated with the error term); setting the partial derivatives equal to zero to obtain the “normal equations”:

$$Q_w = \sum w(Y - \beta_0 - \sum_{i=1}^k \beta_i X_i)^2$$

The **optimal choice for the weight** is the **inverse of variance**; when the variance is constant, ordinary and weighted least squares estimators are identical. For example, in the marginal SLR, when standard deviation is proportional to X_5 (or variance is kX_5^2), we minimize:

$$Q = \sum \frac{1}{X_5^2} (Y - \beta_0 - \sum_{i=1}^5 \beta_i X_i)^2$$

When **error variances are known**, estimators for regression coefficients are obtained by minimizing the quantity Q_w where “w” is a “weight” (associated with the error term, optimal choice is inverse of the known variance); setting the partial derivatives equal to zero to obtain the “normal equations”. The weighted least squares estimators of the regression coefficients are unbiased, consistent, and have minimum variance among unbiased linear estimators

$$Q_w = \sum w_i (Y_i - \beta_0 - \beta_1 X_{i1} - \cdots - \beta_k X_{ik})^2$$

Let the matrix \mathbf{W} be a diagonal matrix containing the weights w_i 's, we have

$$(\mathbf{X}'\mathbf{W}\mathbf{X})\mathbf{b}_w = \mathbf{X}'\mathbf{W}\mathbf{Y}$$

$$\mathbf{b}_w = (\mathbf{X}'\mathbf{W}\mathbf{X})^{-1} \mathbf{X}'\mathbf{W}\mathbf{Y}$$

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

With Ordinary Least squares, $\mathbf{W} = \mathbf{I}$.

If the error variances were known, the use of WLS would be straight forward: **easy and simple**. The resulting estimators exhibit less variability than the ordinary least squares estimators. **Unfortunately, it is an realistic assumption** that we know the error variances. We are forced to use estimates of the **variances** and perform Weighted Least squares estimation using these estimated variances.

The process to estimate error variances is rather **tedious** and can be summarized as follows:

- (1) Fit the model by un-weighted (or ordinary) least squares and obtained residuals,
- (2) Regress the **squared residual** (or absolute value of residual) **against the predictor variables** to obtain a “variance function” (or standard deviation function)
- (3) Use the **fitted values from the estimated variance** (or standard deviation) **to obtain the weight** (for example, inverse of estimated variance)
- (4) **Perform WLS using estimated weights.**

A SIMPLE SAS PROGRAM

If error **variances are known**, the WEIGHT statement (not “option”) allow users to specify the variable to use as the weight in the weighted least squares procedure:

```
PROC REG;
```

```
    WEIGHT W;
```

```
    MODEL Y = X1 X2 X3 X4;
```

```
RUN;
```

If error variances are unknown, it would take a few step to estimate them **before you can use** the WEIGHT statement.

A SAMPLE OF “SAS” FOR WLS

```
proc REG data = SURV;  
  model SurTime = LTest ETest PIndex Clotting;  
  output out=Temp1 R=SurTLSR;
```

```
run;  
data TEMP2;  
set Temp1;  
sqr=SurTLSR*SurTLSR;
```

Here we use “Variance Function”; could choose Standard Deviation Function

```
run;  
proc reg data=TEMP2;  
  model sqr=LTest ETest PIndex Clotting;  
  output out = temp3 P=Esqr;
```

```
run;  
data Temp4;  
set temp3;  
w=1/Esqr;
```

```
run;  
proc reg data=temp4;  
weight w;  
model SurTime = LTest ETest PIndex Clotting;  
run;
```

The condition of the error variance **not being constant** over all cases is called **heteroscedasticity** in contrast to the condition of equal error variances, called **homoscedasticity**. Heteroscedasticity is inherent when the response in regression analysis follows a distribution in which the **variance is functionally related to the mean** (so it is related to at least one predictor variable).

The **remedy for heteroscedasticity** is complicated and, **most of the times, may not worth the efforts.**

CONSEQUENCES OF MULTICOLLINEARITY

- Adding or deleting a ‘predictor variable’ (not a case) changes the regression coefficients
- The “extra sum of square” associated with a predictor variable varies depending on which other variables are already included in the model.
- The estimated standard deviations of many regression coefficients are large.
- The estimated regression coefficients may not be significant even their presence improve prediction.

INFORMAL DIAGNOSTICS

- Indications of the presence of multicollinearity are given by the following informal diagnostics:
- **Large coefficients of correlation** between pairs of predictor variables in matrix \mathbf{r}_{XX} .
- **Non-significant results in individual tests** – even for some important predictor variables identified in advance; some estimated regression coefficients may even have wrong algebraic sign.
- Large changes in estimated regression coefficients when a predictor variable is added or deleted.

For the purpose of measuring and formally detecting the impact of multicollinearity, **it is easier to work with the standardized regression model** which is obtained by transforming all variables (Y and all X's) by means of the **correlation transformation**. The estimated coefficients are now denoted by b^* 's; and there is no intercept.

Correlation transformation: $x^* = \frac{1}{\sqrt{n-1}} \left(\frac{x - \bar{x}}{s_x} \right)$

With transformed variables Y^* 's and all X^* 's, the result is called the Standardized Regression Model:

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + \varepsilon$$

&

$$Y^* = \beta_1^* x_1^* + \beta_2^* x_2^* + \cdots + \beta_k^* x_k^* + \varepsilon^*$$

Results:

$$\mathbf{b}^* = (\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}'\mathbf{Y}$$

$$= \mathbf{r}_{\mathbf{XX}}^{-1} \mathbf{r}_{\mathbf{YX}}$$

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

$$= \sigma^{*2} \mathbf{r}_{\mathbf{XX}}^{-1}$$

$$\begin{aligned}\sigma^2(\mathbf{b}^*) &= \sigma^{*2}(\mathbf{X}'\mathbf{X})^{-1} \\ &= \sigma^{*2}\mathbf{r}_{\mathbf{XX}}^{-1}\end{aligned}$$

It is obvious that the variances of estimated regression coefficients **depend on the correlation matrix $\mathbf{r}_{\mathbf{XX}}$ between predictor variables**. Let the j th diagonal element of the matrix $\mathbf{r}_{\mathbf{XX}}^{-1}$ be denoted by **$(\mathbf{VIF})_j$** (“variance inflation factor”), we have – where σ^{*2} is the error term variance for the (standardized regression) transformed model:

$$\sigma^2(b_j^*) = \sigma^{*2}(VIF)_j$$

VARIANCE INFLATION FACTORS

Let R_j^2 be the coefficient of multiple determination when predictor variable X_j is regressed on the other predictor variables, then we have more simple formulas for the variance inflation factor and the variance of the estimated regression coefficient b_j^* . These formulas indicate that:

- (1) If $R_j^2=0$ (that is X_j is not linearly related at all to the other predictor variables), $(VIF)_j=1$,
- (2) If $R_j^2 \neq 0$, then $(VIF)_j > 1$ indicating an “inflated variance” for b_j^* , and
- (3) If $R_j^2=1$, both $(VIF)_j$ and the variance of b_j^* are unbounded (go to infinity).

$$(VIF)_j = (1 - R_j^2)^{-1}$$

$$\sigma^2(b_j^*) = \frac{\sigma^{*2}}{1 - R_j^2}$$

FORMAL DIAGNOSTICS

Mean VIF values considerably larger than 1.0 are indicative of serious multicollinearity problems because, in addition to having inflated variances, large VIF values also results in larger differences between the estimated and true regression coefficients (“bias” problem):

$$E \left\{ \sum_{j=1}^k (b_j^* - \beta_j^*)^2 \right\} = \sigma^{*2} \sum_{j=1}^k (VIF)_j$$

COMMON REMEDIAL MEASURES

- (1) The presence of multicollinearity often **does not affect the usefulness of the model** in making predictions provided that the values of the predictor variables for inferences are intended follow the same multicollinearity pattern as seen in the data. One simple remedial measure is to **restrict inferences to target subpopulations having the same pattern of multicollinearity.**
- (2) **In polynomial regression models, use centered values for predictors**

COMMON REMEDIAL MEASURES

- (3) **One or more predictor variables may be dropped** from model in order to lessen the degree of multicollinearity. This practice an important problem: no information is obtained about the dropped predictor variables.
- (4) To **add more cases** that would **break the pattern of multicollinearity**; this option is not often available.

COMMON REMEDIAL MEASURES

- (5) In some economic studies, it is possible to estimate the regression coefficients for different predictor variables from different sets of data; however, in other fields, this option is not often available.
- (6) To form a **composite index or indices to represent the highly correlated predictor variables**. Methods such as “principal components” can help, but implementation is not easy
- (7) Finally, “**Ridge Regression**” has proven to be useful to remedy multicollinearity problems.

About “Option #3”, **dropping one or a few predictor variables from the Regression Model**: Still unsettling but remain **most popular with practitioners** (after all if two predictor variables are highly correlated, information in one are almost all contained in the other; **why would we need them both?**)

When **data are representative** of the target population, the presence of multicollinearity – **even severe multicollinearity** – **does not affect the prediction**. The problem is how to reduce the standard errors of the estimated regression coefficients. The method of “ridge regression” modifies the method of least squares to allow biased but more precise (smaller variances) estimators. When an estimator has only a small bias but much smaller variance, it may be preferred because it corresponds to a larger probability of being close to the true parameter value.

The “value” of an estimator b^R is judged by the “**mean squared error**”, a measure of the **combined effect of bias and sampling variation** (variance); if an estimator is unbiased, its mean squared error is equal to its variance.

$$\begin{aligned} E\{b^R - \beta\}^2 &= \{E(b^R)\}^2 + \sigma^2(b^R) \\ &= (\textit{Bias})^2 + \textit{Variance} \end{aligned}$$

Recall that, for the **ordinary unweighted least squares**, the “normal equations are:

$$(\mathbf{X}'\mathbf{X})\mathbf{b} = \mathbf{X}'\mathbf{Y}$$

Or, when all variables are transformed by the “correlation transformation”, the transformed or standardized regression model (the one with stars) is given below with its least squares normal equations:

$$\mathbf{r}_{\mathbf{X}\mathbf{X}}\mathbf{b} = \mathbf{r}_{\mathbf{Y}\mathbf{X}}$$

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + \varepsilon$$

&

$$Y^* = \beta_1^* x_1^* + \beta_2^* x_2^* + \cdots + \beta_k^* x_k^* + \varepsilon^*$$

RIDGE ESTIMATORS

The ridge standardized regression estimators are obtained by introducing into the least squares normal equations a “**biasing constant c**” as follows. The constant “c” reflects the amount of bias in the estimators (that we accept in exchange for smaller variances). When $c=0$, they reduce to the ordinary least squares estimators; when $c>0$, the **ridge estimators are biased but tend to be more stable.**

$$(\mathbf{r}_{XX} + c\mathbf{I})\mathbf{b}^R = \mathbf{r}_{YX}$$

$$\mathbf{b}^R = (\mathbf{r}_{XX} + c\mathbf{I})^{-1}\mathbf{r}_{YX}$$

BIASING CONSTANT “c”

When the biasing constant “c” gets larger, the **biased component** of the total mean squared error of the ridge regression estimator increases while the variance component becomes smaller. There always **exists a value c_{op} for which the ridge regression estimator has a smaller total mean squared error** the ordinary least squares estimator but the optimal value varies from one application to another and is unknown.

When the biasing constant “c” is increased slowly from zero, the **ridge estimators may fluctuate widely but gradually these fluctuations cease**. In the meantime, the values of **VIF first fall rapidly then gradually slow down as c is increased further**. The smallest value of c where the ridge estimators first become stable and the VIF values become sufficiently small is often chosen. But it’s a judgmental call; there is no magic rule to tell how stable is stable and how small is small. The simultaneous graph of ridge estimators against c is called the “ridge trace”. **You can use SAS to form this graph.**

In the presence of severe multicollinearity, the ordinary least squares regression estimates may be highly ‘unstable’ in the sense that their values might be changed substantially for small changes in the data. Ridge regression estimates, on the other hand, are more “stable”; they are little affected by small changes in the data on which the fitted regression model is based.

Keep in mind that the process is extremely time consuming and, therefore, less popular with practitioners.

INFLUENTIAL CASES

- **Hat matrix** and studentized deleted residuals are valuable tools for identifying outlying cases
- We can measure the influence of the outlying cases by means of **DDFFITS**, Cook's distance, and **DFBETAS** measures.
- Reason for our concern: method of least squares is susceptible to these cases possibly leading to seriously distorted results/fitted models.
- **Question:** How to handle highly influential cases, i.e. to reduce their influence of LS results?

First source to look for: **recording error; fix them if confirmed and true values available**, if not discard them. The problem is, more often, it is not possible to confirm that an outlying/influential case is clearly erroneous.

Next stop: check for the adequacy of the model; may be a term (such as quadratic/power) or an **important predictor variable is omitted**.

Discarding of outlying influential cases that are not clearly erroneous and that cannot be accounted for by model improvements should be done only rarely & with caution.

ROBUST REGRESSION

Robust regression procedures reduce the influence of outlying cases, as **compared to** ordinary least squares estimation. They do not require that the assumption of a normal distribution for the error terms and they could be used with or without the identification of outlying cases – even to fit data that are “noisy” with numerous outlying cases. There are **many robust regression procedures**.

EXPLORATION OF SHAPE of Regression Function

Instead of checking the appropriateness of a linear regression function by means graphs such scatter plots or marginal scatter plots, it is also helpful to explore the nature or shape of the regression relationship by fitting a smoothed curve without any prior constraints on the regression function. These smoothed curves are called “non-parametric regression curves”; the most well-known one is the “**Lowess Method**”.

Many smoothing methods have been developed for obtaining smoothed curves for “time series data”. **The “method of moving averages” uses the mean of the Y observations for (two or three) adjacent time periods to obtain smoothed values.** For example, use the mean of $\{Y_1, Y_2, Y_3\}$ as new first Y value, then mean of $\{Y_2, Y_3, Y_4\}$, then mean of $\{Y_3, Y_4, Y_5\}$, etc... The “method of running medians” is similar to the method of moving averages except that the median is used as new observation instead of the average in order to further reduce the influence of outlying cases.

Another smoothing method, called “Band Regression”, divide the data into a number of groups or bands consisting of adjacent cases according to the X levels. For each band, the **median of X** and the **median of Y** values are calculated and used in fitting, say, a straight line by least squares.

AUTOCORRELATION

The basic multiple regression models have assume that the random **error terms are independent** normal random variables or, at least, uncorrelated random variables. In some fields – for example in economics, regression applications may involve “time series”; the assumption of uncorrelated or independent error terms may not be appropriate. In time series data, error terms are often (positively) correlated over time – **auto-correlated** or serially correlated.

$$Y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_k x_{ki} + \varepsilon_i$$

PROBLEMS OF AUTOCORRELATION

- Least squares estimates of regression coefficients are still unbiased but no longer have minimum variance
- MSE may seriously under-estimate variance of error terms
- Standard errors of estimated regression coefficients may seriously under-estimate the true standard deviations of the estimated regression coefficients; confident intervals of regression coefficients and of response means, therefore, may not have the correct coverage.
- t and F tests may no longer applicable, have wrong size.

FIRST-ORDER
AUTOREGRESSIVE ERROR
MODEL

$$Y_t = \beta_0 + \beta_1 x_{1t} + \beta_2 x_{2t} + \cdots + \beta_k x_{kt} + \varepsilon_t$$

$$\varepsilon_t = \rho \varepsilon_{t-1} + u_t$$

where:

$$|\rho| < 1$$

u_i 's are independent $N(0, \sigma^2)$

$$Y_t = \beta_0 + \beta_1 x_{1t} + \beta_2 x_{2t} + \cdots + \beta_k x_{kt} + \varepsilon_t$$

$$\varepsilon_t = \rho \varepsilon_{t-1} + u_t$$

where:

$$|\rho| < 1$$

u_i 's are independent $N(0, \sigma^2)$

Note that **each error term consists of a fraction of the previous error term plus a new disturbance term u_i ; the parameter ρ – often positive - is called the “**autocorrelation parameter**”.**

First, we can easily expand from the definition of the first-order autoregressive error model to show that each error term is a linear combination of current and preceding disturbance terms. This is used to prove that the **mean is zero and the variance is constant.**

$$\begin{aligned}\varepsilon_t &= \rho\varepsilon_{t-1} + \mathbf{u}_t \\ &= \rho(\rho\varepsilon_{t-2} + u_{t-1}) + u_t \\ &= \rho^2\varepsilon_{t-2} + \rho u_{t-1} + u_t \\ &= \rho^2(\rho\varepsilon_{t-3} + u_{t-2}) + \rho u_{t-1} + u_t \\ &= \rho^3\varepsilon_{t-3} + \rho^2 u_{t-2} + \rho u_{t-1} + u_t \\ &= \dots\end{aligned}$$

$$\begin{aligned}\mathbf{E}(\varepsilon_t) &= \sum_{s=0}^{\infty} \rho^s \mathbf{E}(\mathbf{u}_{t-s}) \\ &= \mathbf{0}\end{aligned}$$

$$\begin{aligned}\sigma^2(\varepsilon_t) &= \sigma^2 \sum_{s=0}^{\infty} \rho^{2s} \\ &= \frac{\sigma^2}{1 - \rho^2}\end{aligned}$$

The error terms of the first-order autoregressive model still have mean zero and constant variance but a **positive covariance between consecutive terms**:

$$Y_t = \beta_0 + \beta_1 x_{1t} + \beta_2 x_{2t} + \cdots + \beta_k x_{kt} + \varepsilon_t$$

$$\varepsilon_t = \rho \varepsilon_{t-1} + u_t$$

where:

$$|\rho| < 1$$

u_i 's are independent $N(0, \sigma^2)$

$$E(\varepsilon_t) = 0$$

$$\sigma^2(\varepsilon_t) = \frac{\sigma^2}{1 - \rho^2}$$

$$\sigma\{\varepsilon_t, \varepsilon_{t-1}\} = \rho \left(\frac{\sigma^2}{1 - \rho^2} \right)$$

The autocorrelation parameter is also the coefficient of correlation between two consecutive error terms:

$$\frac{\sigma(\varepsilon_t, \varepsilon_{t-1})}{\sigma(\varepsilon_t)\sigma(\varepsilon_{t-1})} = \frac{\rho \left(\frac{\sigma^2}{1-\rho^2} \right)}{\sqrt{\frac{\sigma^2}{1-\rho^2}} \sqrt{\frac{\sigma^2}{1-\rho^2}}} = \rho$$

The coefficient of correlation below, of two error terms that are **s periods apart**, shows that the error terms are also positively correlated but the further apart they are the less the correlation between them:

$$\frac{\sigma(\varepsilon_t, \varepsilon_{t-s})}{\sigma(\varepsilon_t)\sigma(\varepsilon_{t-s})} = \frac{\rho^s \left(\frac{\sigma^2}{1-\rho^2} \right)}{\sqrt{\frac{\sigma^2}{1-\rho^2}} \sqrt{\frac{\sigma^2}{1-\rho^2}}} = \rho^s$$

DURBIN-WATSON TEST

- The **Durbin-Watson** test for autocorrelation assumes the first-order autoregressive error model (with values of predictor variables fixed)
- The test consists of determining whether or not the autocorrelation parameter (which is also the coefficient of correlation between consecutive error terms) is zero (if so, errors terms are equal to distance terms which are i.i.d. normal):

$$H_0 : \rho = 0$$

$$H_A : \rho > 0$$

The Durbin-Watson test statistic **D** is obtained by first using ordinary least squares method to fit the regression function, calculating residuals and then **D**. Small values of **D** support the conclusion of autocorrelation because, under the first-order autoregressive error model, adjacent error terms tend to be of the same magnitude (because they are positively correlated) – leading to small differences and a small total of those difference:

$$D = \frac{\sum_{t=2}^n (e_t - e_{t-1})^2}{\sum_{t=1}^n e_t^2}$$

Exact critical values for the Durbin-Watson test statistics D are difficult to obtain, but Durbin and Watson have provided lower and upper bounds d_L and d_U such that an observed value of the test statistic D outside these bounds leads to a definitive decision. Values of the bound depend on the alpha level, number of predictor variables, and sample size n (**Table B7, page 675**).

$\left\{ \begin{array}{l} \text{If } D > d_U : \text{Data support } \mathbf{H}_0 \\ \text{If } D < d_L : \text{Data support } \mathbf{H}_A \\ \text{If } d_L < D < d_U : \text{Test is } \mathbf{inconclusive} \end{array} \right.$

SAS implementation is very simple: Use option DW

PROC REG;

MODEL Y = X1 X2 X3/DW;

An estimate of the autocorrelation parameter is provided using the following formula:

$$r = \frac{\sum_{t=2}^n e_t e_{t-1}}{\sum_{t=1}^n e_t^2}$$

When the presence of autocorrelation is confirmed, the problem could be remedied by adding in another predictor variable or variables: **one of the major cause is the omission from the model of one or more key predictors:**

Readings & Exercises

- Readings: A scan through the text's Chapter 11 and sections 12.1-12.3 would help.
- Exercises: 11.6, 11.7, 12.5, 12.6, and 12.7
- Due as Homework: **11.6(a-d) and 12.5**