PubH 7405: REGRESSION ANALYSIS

DIAGNOSTICS IN MULTIPLE REGRESSION

The data are in the form :

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

Multiple Regression Model :

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

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

The error terms are identically and independently distributed as normal with a constant variance.

A simple strategy for the building of a regression model consists of some, most, or all of the following five steps or phases:

(1) Data collection and preparation,

(2) Preliminary model investigation,

(3) Reduction of the predictor variables,

(4) Model refinement and selection, and

(5) Model validation

Let say we finished step #3; we tentatively have a good model: it’s time for some fine tuning!

D**iagnostics** to identify possible violations of model’s assumptions are often focused on these “major issues”:

(1) Non-linearity,

(2) Outlying & influential cases,

(3) Multi-collinearity,

(4) **Non-constant variance** &

(5) Non-independent errors.

DETECTION OF NON-LINEARITY

A limitation of the usual residual plots (say, residuals against values of a predictor variable): **they may not show the nature of the "additional contribution"** of a predictor variable to those by other variables already in the model.

For example, we consider a multiple regression model with 2 independent variables X_1 and X_2; is the relationship between Y and X_1 linear?

"**Added-variable plots**" (also called "partial regression plots" or "adjusted variable plots") are **refined residual plots** that provide graphic information about the marginal importance of a predictor variable given the other variables already in the model.

ADDED-VARIABLE PLOTS

In an added-variable plot, both the response variable **Y** and the predictor variable under investigation (say, $\mathbf{X_1}$) are **both regressed against the other predictor variables** already in the regression model and the **residuals are obtained for each**. These two sets of residuals reflect the part of each (Y and X_1) that is not linearly associated with the other predictor variables.

The plot of one set of residuals against the other set would **show the marginal contribution of the candidate predictor** in reducing the residual variability as well as the information about the nature of its marginal contribution.

A SIMPLE & SPECIFIC EXAMPLE

Consider the case in which we already have a regression model of Y on predictor variable X_2 and is now considering **if we should add X_1** into the model (if we do, we would have a multiple regression model of Y on (X_1,X_2)). In order to decide, we investigate 2 simple linear regression models: (a) The **regression of Y on X_2** and (b) The **regression of X_1 on X_2** and **obtain 2 sets of residuals as follows**:

Regression#1 :

Y on X_2 (first or old model)

$$\hat{Y}_i = b_0 + b_2 X_{i2}$$

$$e_i(Y \mid X_2) = Y_i - \hat{Y}_i$$

These residuals represent the **part of Y not explained by X_2**

Regression#2 :

X_1 on X_2 (second or new model)

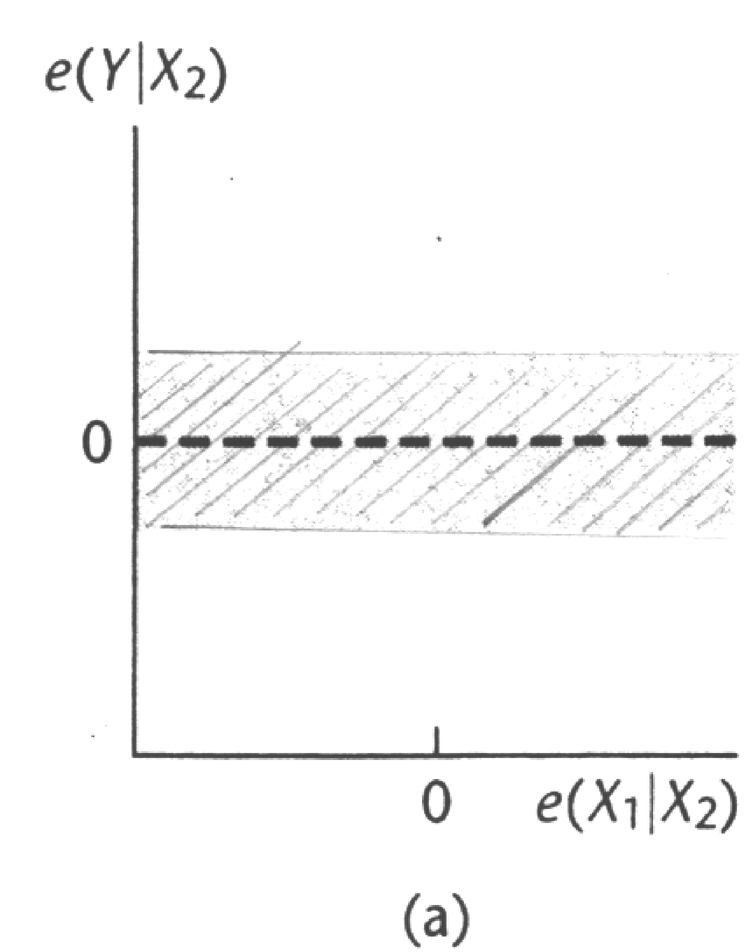
$$\hat{X}_{1i} = b_0^* + b_2^* X_{i2}$$

$$e_i(X_1 \mid X_2) = X_{1i} - \hat{X}_{1i}$$

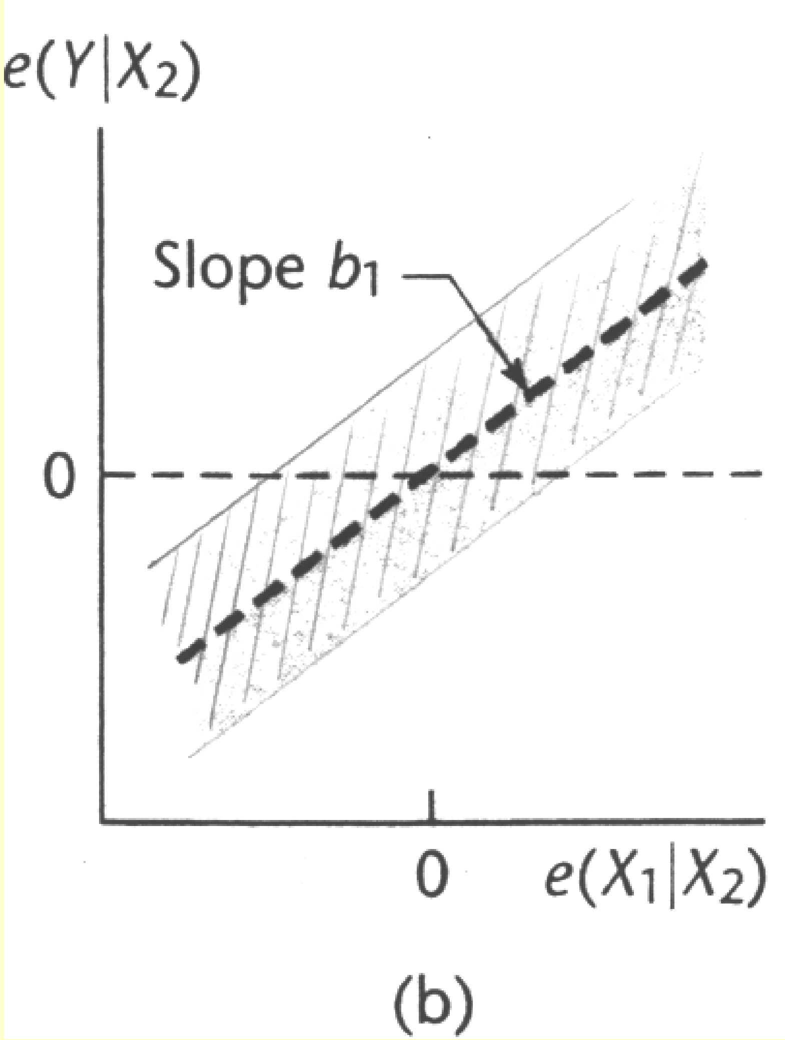
These residuals represent the

part of $\mathbf{X}_1$ not contained in $\mathbf{X}_2$

We now do a “regression of $e_i(Y|X_2)$ – as new dependent variable on $e_i(X_1|X_2)$ – as independent variable: That is **to see if “the part of X_1 not contained in X_2” can further explained “the part of Y not explained by X_2”**; (if it can, X_1 should be added to the model for Y which already has X_2 in it). There are three possibilities:



The “horizontal band” shows that X_1 contains **no** “additional information” useful for the prediction of Y beyond that contained in and provided for by X_2



This "linear band with a non-zero slope" indicates that "the part of X_1 not contained in X_2" is **linearly related to** "the part of Y not explained by X_2" alone. That is, **X_1 should be added** to form a 2-variable model. Note that if we do a "**regression through the origin**", **the slope b_1 should be the same as the coefficient of X_1 if it is added to the regression model already containing X_2.**

AN EXAMPLE

Suppose :

$$(1)\ \hat{Y}(X_2) = 50.70 + 15.54X_2$$

$$(2)\ \hat{X}_1(X_2) = 40.78 + 1.72X_2$$

and :

$$(3)\ e(Y \mid X_2) = 6.29e(X_1 \mid X_2)$$

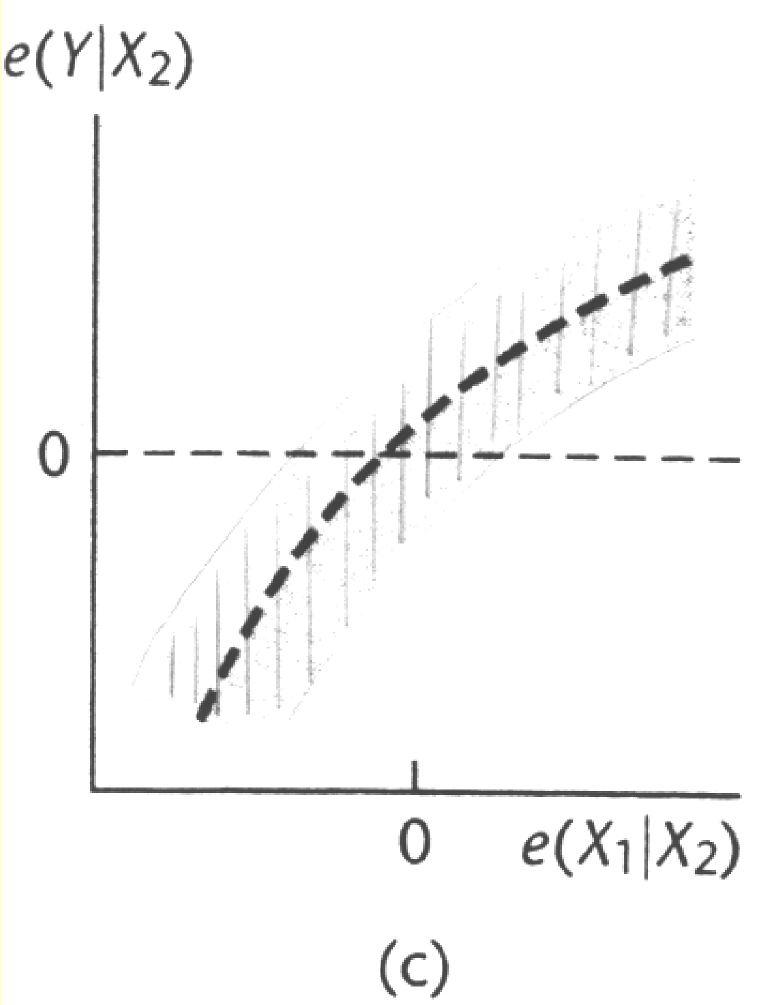
Then :

$$[Y - (50.70 + 15.54X_2)] = 6.29[X_1 - (40.78 + 1.72X_2)],$$

we have the same result as in multiple regression fitting :

$$\mathrm{Y} = [50.70 - (6.29)(40.78)] + 6.29X_1 + [15.54 - (6.29)(1.72)]X_2$$

$$\mathbf{Y} = \mathbf{-205.81} + \mathbf{6.29X_1} + \mathbf{4.72X_2}$$



The "curvilinear band" indicates that, like the case (b) previously, X_1 should be added to the model already containing X_2. However, it further suggests that the **inclusion of X_1 is justified but some power terms – or some type of data transformation** – are needed.

The fact that an added-variable plot may suggest "the nature of the **functional form**" in which a predictor variable should be added to the regression model is more important **than that the variable possible inclusion** (which can be resolved without graphical help). **The added-variable plots play the role that we use scatter diagrams for in simple linear regression**; **they would tell if data transformation or if certain polynomial model is desirable.**

ADDED-VARIABLE PLOT

```
PROC REG data = Origset;
  Model Y X1 = X2/noprint;
  Output Out = SaveMe R = YResid X1Resid;
Run;
```

2 Simple Linear Regression here

```
PROC PLOT data = SaveMe;
  Plot YResid*X1Resid;
Run;
```

Another Example:

Suppose we have a dependent variable Y and 5 predictor variables X1-X5; and assume that we **focus on X5**, our primary predictor variable:
(1) we regress **Y on X1-X4** and obtain residuals (say, **YR**), and
(2) we regress **X5 on X1-X4** and obtain residuals (say, **X5R**), then
(3) we regress **YR on X5R** (**SLR, no intercept**) – **linear assumption about X5 can be studied** from this last SLR & its (**added-value) plot**.

“MARGINAL” SL REGRESSION

```
PROC REG data = Origset;
Model Y X5 = X1 X2 X3 X4/noprint;
Output Out = SaveMe R = YR X5R;
Run;

PROC PLOT data = SaveMe;
  Plot YR*X5R;
Run;

PROC REG data = SaveMe;
Model YR = X5R/ noint;
Run;
```

2 Multiple Linear Regression here

New:

Marginal Simple Linear Regression

REMEDIAL MEASURES

- 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. **One is hard “to do” and one is hard “to explain”**

OUTLYING & INFLUENTIAL CASES

SEMI-STUDENTIZED RESIDUALS

$$e_i^* = \frac{e_i - \bar{e}}{\sqrt{MSE}} = \frac{e_i}{\sqrt{MSE}}$$

If √MSE were an estimate of the standard deviation of the residual e, we would call e* a studentized (or standardized) residual. However, the standard deviation of the residual is complicated and varies for different residuals, and √MSE is only an approximation. Therefore, e* is call a "**semi-studentized residual**". But "exact" standard deviations can be found!

THE HAT MATRIX

$$\overset{\wedge}{\mathbf{Y}} = \mathbf{Xb}$$

$$= \mathbf{X[(X'\,X)^{-1}X'\,Y]}$$

$$= \mathbf{[X(X'\,X)^{-1}X'\,]Y}$$

$$= \mathbf{HY}$$

$$\mathbf{H} = \mathbf{X(X'\,X)^{-1}X'}\ \text{is called the "Hat Matrix"}$$

The Hat matrix can be obtained from data

RESIDUALS

Model :

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

Fitted Value :

$$\overset{\wedge}{\mathbf{Y}} = \mathbf{X}\mathbf{b}$$

Residuals :

$$\begin{aligned} \mathbf{e} &= \mathbf{Y} - \overset{\wedge}{\mathbf{Y}} \\ &= \mathbf{Y} - \mathbf{X}\mathbf{b} \\ &= \mathbf{Y} - \mathbf{H}\mathbf{Y} \\ &= (\mathbf{I} - \mathbf{H})\mathbf{Y} \end{aligned}$$

Like the Hat Matrix **H**, (**I-H**) is symmetric & idempotent

VARIANCE OF RESIDUALS

$$
\begin{aligned}
\mathbf{e} &= (\mathbf{I}-\mathbf{H})\mathbf{Y} \\
\boldsymbol{\sigma}^2(\mathbf{e}) &= (\mathbf{I}-\mathbf{H})\boldsymbol{\sigma}^2(\mathbf{Y})(\mathbf{I}-\mathbf{H})' \\
&= (\mathbf{I}-\mathbf{H})(\sigma^2\mathbf{I})(\mathbf{I}-\mathbf{H})' \\
&= \sigma^2(\mathbf{I}-\mathbf{H})(\mathbf{I}-\mathbf{H})' \\
&= \sigma^2(\mathbf{I}-\mathbf{H})(\mathbf{I}-\mathbf{H}) \\
&= \sigma^2(\mathbf{I}-\mathbf{H}) \\
&\overset{\wedge}{=} MSE(\mathbf{I}-\mathbf{H})
\end{aligned}
$$

$\text{For the ith observatio n :}$

$$\boldsymbol{\sigma}^2(\mathbf{e}_i) = (\mathbf{1} - \mathbf{h}_{ii})\boldsymbol{\sigma}^2$$

$$s^2(e_i) = (1 - h_{ii})MSE$$

$$\sigma(e_i, e_j) = -h_{ij}\sigma^2; i \neq j$$

$$s(e_i, e_j) = -h_{ij}MSE; i \neq j$$

$\mathbf{h}_{ii}$ is the ith element on the main diagonal &

h_{ij} is on ith row and jth column of the hat matrix

STUDENTIZED RESIDUALS

$$r_i = \frac{e_i}{s(e_i)}$$

$$= \frac{e_i}{\sqrt{(1-h_{ii})MSE}}$$

Studentized residuals is a "refine" version of the semi-studentized residuals; in "r" we use the exact standard deviation of the residual "e" – not an approximation.

Semi-studentized residuals and Studentized Residuals are tools for detecting Outliers.

DELETED RESIDUALS

The second refinement to make residuals more effective for detecting outlying or extreme observations is to measure the ith residual when the fitted regression is **based on all cases except the ith case** (similar to the concept of jackknifing). The reason is **to avoid the influence of the ith case – especially if it is an outlying observation -** on the fitted value. If the ith case is an outlying observation, **its exclusion will make the “deleted residual” larger** and, therefore, more likely to “confirm” its outlying status:

$$d_i = Y_i - \hat{Y}_{i(i)}$$

STUDENTIZED DELETED RESIDUALS

Combining the two refinements**, the studentized residual and the deleted residual,** we get the "**Studentized Deleted Residual**". A studentized residual is also called an **internal studentized residual** and a studentized deleted residual is an **external studentized residual**; **MSE$_{(i)}$ is the mean square error when the ith case is omitted in fitting the regression model.**

$$t_i = \frac{d_i}{s(d_i)}$$

$$= \frac{e_i}{\sqrt{(1-h_{ii})MSE_{(i)}}}$$

$$t_i = \frac{e_i}{\sqrt{(1-h_{ii})MSE_{(i)}}}$$

The studentized deleted residuals can be calculated without having to fit new regression functions each time a different case is omitted; that is, we need to fit the model only once – with all n cases to obtain MSE – not the same model n times, here “p” is the number of parameters, p = k+1:

$$(n-p)MSE = (n-p-1)MSE_{(i)} + \frac{e_i^2}{1-h_{ii}}$$

$$(n-p)MSE = (n-p-1)MSE_{(i)} + \frac{e_i^2}{1-h_{ii}}$$

leading to:

$$\mathbf{t_i = e_i\left[\frac{n-p-1}{SSE(1-h_{ii})-e_i^2}\right]}$$

which is distributed as "t" with $(n-p-1)$ degrees of freedom under H_0 of no outliers.

This can be used as a statistical test for outliers – with allowance for multiple decisions

The diagonal elements h_{ii} of the hat matrix , called “leverages”, have the same type of influence on the studentized deleted residuals and, therefore, serve as good indicators in identifying outlying X observations; **the larger this value the more likely that the case is an outliers**. A leverage value is usually considered to be **large** if it is more than twice as large as the mean leverage value (which is 2p/n).

$$\hat{\mathbf{Y}} = \mathbf{HY}$$

$$\boldsymbol{\sigma}^2(\hat{\mathbf{Y}}) = \mathbf{H}\boldsymbol{\sigma}^2(\mathbf{Y})\mathbf{H'}$$

$$= \mathbf{H}(\sigma^2\mathbf{I})\mathbf{H'}$$

$$= \sigma^2\mathbf{HH}\ (\mathbf{H}\text{ is symmetric, }\ \mathbf{H} = \mathbf{H'}\,)$$

$$= \sigma^2\mathbf{H}\ (\mathbf{H}\text{ is idempotent , }\mathbf{HH} = \mathbf{H})$$

$$\sigma^2(\hat{Y}_i) = \sigma^2 h_{ii};\text{ so large }h_{\mathrm{ii}}\text{ is "not good"}$$

IDENTIFICATION OF INFLUENTIAL CASES

What we have done is to identify cases that are **outlying** with respect to their Y values (say, using their studentized deleted residuals) and/or X values (using their leverages). The next step is to ascertain **whether or not these outlying cases are influential**; case is influential if its exclusion causes major changes in the fitted regression function. (If a case is determined to be outlying and influential, the next step would be investigating – using other sources – **to see if it should be taken out**).

INFLUENCE ON A FITTED VALUE

Recall the idea of “deleted residuals” in which we measure the ith residual from the fitted value where regression fitting is based on **all cases except the ith case** so as to avoid the influence of the ith case itself. We can use the very same idea to measure the “influence of a case on its own fitted value; that is to measure the difference (“DF”) **between the fitted values** when all data are used to the fitted value where regression fitting is based on all cases except the ith case; MSE(i) is the mean square error when the ith case is omitted

$$(\textbf{DFFITS})_i = \frac{\hat{Y}_i - \hat{Y}_{i(i)}}{\sqrt{MSE_{(i)} h_{ii}}}$$

INFLUENCE ON ALL FITTED VALUES

Taking the same idea of deleting a case to investigate its influence but, in contrast to DDFITS which consider the influence of a case on its own fitted value, **“Cook’s Distance”** shows the effect of ith case on “**all**” fitted values. The denominator serves only as a standardized measure so as to reference Cook’s Distance to the **F(p,n-p) percentiles**.

$$D_i = \frac{\sum_{j=1}^{n}(\hat{Y}_j - \hat{Y}_{j(i)})^2}{pMSE}$$

INFLUENCE ON
REGRESSION COEFFICIENTS

Another measure of influence, DFBETAS, is defined similar to DDFITS but DFBETAS focuses on the effect of (deleting) ith case **on the values of all regression coefficients**; in this formula, **c_{jj} is the jth diagonal element of matrix $(X'X)^{-1}$.** As a guideline for identifying influential cases , a case is considered "influential" if the absolute value of a DFBETAS **exceeds 1** for medium data sets and **2/√n** for larger sets:

$$(DFBETAS)_{j(i)} = \frac{b_j - b_{j(i)}}{\sqrt{MSE_{(i)} c_{jj}}}$$

THE ISSUE OF MULTICOLINEARITY

When **predictor variables are highly correlated among themselves we have** "**multicollinearity**": the estimated regression coefficients tend to have **large standard errors**.

Outlying and influential effects are case-based whereas multicollinearity effects are variable-based.

CONSEQUENCES OF MULTICOLLINEARITY

- **Adding or deleting a “predictor variable” changes some regression coefficients substantially.**
- **Estimated standard deviations of some regression coefficients are very 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** –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:

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

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

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

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

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

It is obvious that the variances of estimated regression coefficients **depend on the correlation matrix $\mathbf{r}_{\mathbf{xx}}$ between predictor variables**. The **jth diagonal element of the matrix $\mathbf{r}_{\mathbf{xx}}^{-1}$ be denoted by (VIF)$_j$ and is called the “variance inflation factor”)**:

VARIANCE INFLATION FACTORS

Let $\mathbf{R_j^2}$ be the coefficient of multiple determination when predictor **variable $\mathbf{X_j}$ 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 $R_j^2\neq0$, then $(VIF)_j>1$ indicating an “inflated variance” for b^*_j

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

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

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 presents 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

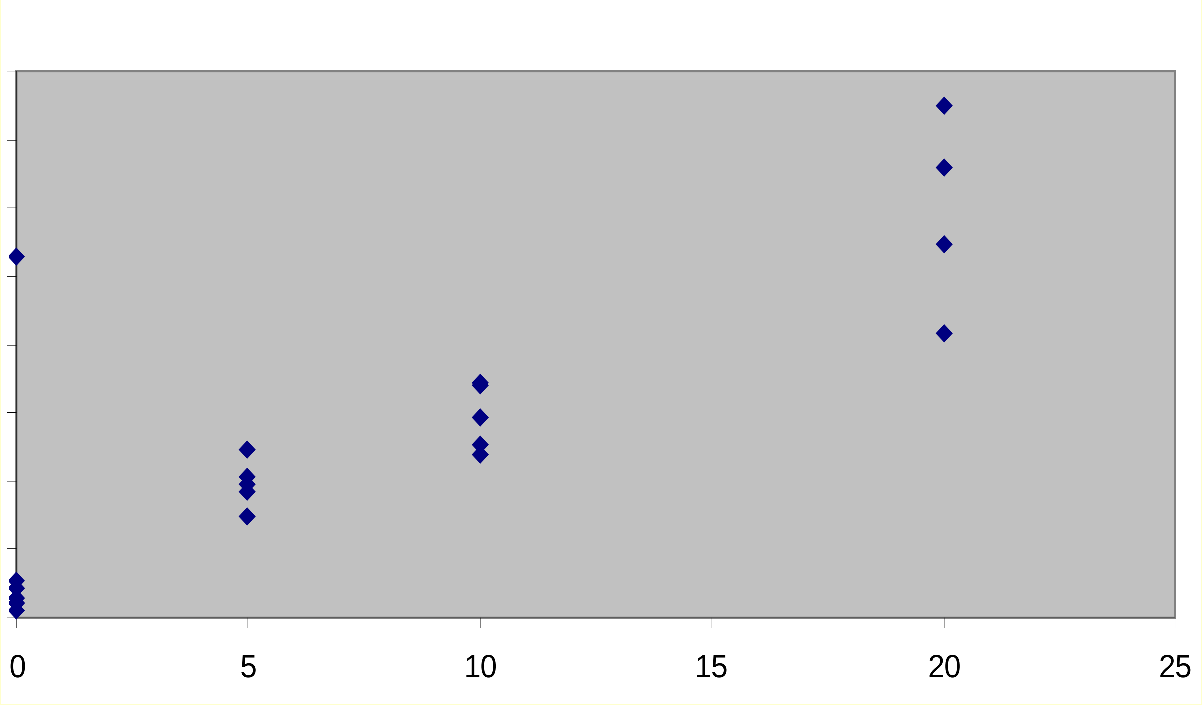
(5) In some economic studies, it is possible to estimate the regression coefficients for different predictor variables from different sets of data; however, in other fields, this option is not often available.

(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; **Ridge regression** is one of the advanced methods which modifies the method of least squares to **allow biased** but **more precise estimators** of the regression coefficients.

NON-CONSTANT VARIANCE

- **Graph of Residuals against Predictor** variable or **against the fitted values** is helpful to see **if the variance of error terms are constant**; if model fits, it shows a band centered around zero with **a constant width.**
- **Lack of fit result in a graph showing the residuals departing from zeros in a systematic fashion – a "megaphone" shape.**
- **No new fancy method/graph are needed here!**



Example: Residual Plot for Non-constant Variance

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 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**. For example, when standard deviation is proportional to X_5 (or variance is kX_5^2), we minimize:

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

Let consider, in more depth, the generalized multiple regression model:

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

$$\varepsilon_i \in N(0, \sigma_i^2)$$

And **first look at the case where error variances are known** and then relax this unrealistic assumption.

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 **W** be a diagonal matrix containing the weights w_{i}'s, we have

$$(\mathbf{X'WX})\mathbf{b}_{\mathrm{w}} = \mathbf{X'WY}$$

$$\mathbf{b}_{w} = (\mathbf{X'WX})^{-1}\mathbf{X'WY}$$

$$\boldsymbol{\sigma}^{2}(\mathbf{b}_{w}) = \sigma^{2}(\mathbf{X'WX})^{-1}$$

With Ordinary Least squares, **W** = **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 <u>**process**</u> **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 against the predictor variables** to obtain a “variance function” (variance as a function of all predictors)

(3) Use the **fitted values from the estimated variance to obtain the weight** (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;
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;
```

To form “Variance Function”

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, **sometimes, it may not worth the efforts**. **Transformations of Y could get you into troubles and Weighted Least Squares is less often used because it is "hard to implement"**

THE ISSUE OF CORRELATED ERRORS

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 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 independen t $N(0, \sigma^2)$

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

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

$$where:$$

$$|\rho| < 1$$

$$u_i\text{'s are independen t } N(0, \sigma^2)$$

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** – the variance is larger.

$$\begin{aligned}
\boldsymbol{\varepsilon}_t &= \boldsymbol{\rho}\boldsymbol{\varepsilon}_{t-1} + \mathbf{u}_t \\
&= \rho(\rho\varepsilon_{t-2} + u_{t-1}) + u_t \\
&= \rho^2\varepsilon_{t-2} + \rho u_{t-1} + u_t \\
&= \rho^2(\rho\varepsilon_{t-3} + u_{t-2}) + \rho u_{t-1} + u_t \\
&= \rho^3\varepsilon_{t-3} + \rho^2 u_{t-2} + \rho u_{t-1} + u_t \\
&= \cdots \\
&= \sum_{s=0}^{\infty} \rho^s u_{t-s}
\end{aligned}$$

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

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

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

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

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

$$where:$$

$$|\rho| < 1$$

$$u_i\text{'s are independen t } N(0,\sigma^2)$$

$$\longrightarrow$$

$$E(\varepsilon_t) = 0$$

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

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

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

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

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:

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

The text presents a small simulated data set, pages 482-483, the regression line by least squares method varies from case to case **depending on the value of the first error term.** The **estimated slope ranges from a negative value to a positive value**. It could be a serious problem unless we know how to deal with autocorrelation. And one could find example of time series data not only in economics and business but also in biomedical longitudinal data as well.

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**).

$$\begin{cases} \text{If } D > d_U: \text{ Data support } \mathbf{H_0} \\ \text{If } D < d_L: \text{ Data support } \mathbf{H_A} \\ \text{If } d_L < D < d_U: \text{Test is } \mathbf{inconclusi\ ve} \end{cases}$$

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**:

BASIC SAS OPTIONS

- CORR: between all variables in model statement
- P: (or PRED) predicted values
- R: (or RESID) residuals
- COVB: (“B” for estimated regression coefficient)
- CORRB: Variance-covariance matrix (among **B**)
- CLB: confidence interval for “B”
- CLM: confidence interval for Mean Response
- CLI: confidence interval for individual predicted.

SAS OPTIONS FOR SELECTION

- RSQUARE
- ADJRSQ
- MAXR
- AIC
- FORWARD
- BACKWARD
- STEPWISE

SAS OPTIONS FOR DIAGNOSTICS

- CORR: (individual $|r|>.7$, say)
- STUDENT: studentized residuals
- PRESS: deleted residuals (square & add to obtain the PRESS statistic)
- RSTUDENT: studentized deleted residuals ($t(1-\alpha/2;n-p-1)$
- H: Leverages ($2p/n$)
- COOKD: Cook’s distance (50% percentile of $F(p,n-p)$
- DFFITS: (1 or $2\sqrt{p/n}$)
- DFBETAS: (1 or $1/\sqrt{n}$)

Readings & Exercises

- Readings: A thorough reading of the text's sections 10.1-10.5 (pp.384-410) and skim over the example in section 10.6 is recommended.
- Exercises: The following exercises are good for practice, all from chapter 10: 10.7, 10.11, 10.17, and 10.21.

Due As Homework

#19.1 Refer to dataset “Cigarettes”, let Y = log(NNAL) and consider a model with three independent variables, X_1 = CPD, X_2 = Age, and X_3 = Gender:

a) Fit the multiple regression model with 3 predictor variables. Does it appear that all predictors should be retained?

b) Prepare an added-variable plot for X_1 and X_2. Do your plots suggest that the corresponding regression relationships in (a) are appropriate? Explain your answers.

c) Obtain the variance inflation factors. Are there indication that serious multicollinearity problems exist here? Why or why not?

d) Obtain the residuals and prepare a normal probability plot; Also plot the residuals against X_1 and X_2, and draw conclusions.

#19.2 Answer the 4 questions of Exercise 19.1 using dataset “Infants” with Y = Birth Weight, X_1 = Gestational Weeks, X_2 = Mother’s Age, and X_3 = Toxemia (toxemia = 0/1 is a pregnancy condition resulting from metabolic disorder).