Multicollinearity

Dr. Kiah Wah Ong

Model Assumptions

Recall the major assumptions we have made in linear regression models

$$y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ik} + \epsilon_i, \quad i = 1, 2, \cdots, n$$

are

- The relationship between the response and regressors is linear.
- The error terms ϵ_i have mean zero.
- The error terms ϵ_i have constant variance σ^2 (homoscedasticity)
- The error terms ϵ_i are normally distributed.
- The error terms ϵ_i and ϵ_j are uncorrelated for $i \neq j$.
- The regressors $x_1, \cdots, x_k$ are nonrandom.
- The regressors $x_1, \cdots, x_k$ are measured without error.
- The regressors are linearly independent.

Multicollinearity

Multiple linear regression model given a set of n-data is given by

$$y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ik} + \epsilon_i, \quad i = 1, \cdots, n$$

and the above equation can be written as

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

where

$$\mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}, \quad \mathbf{X} = \begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1k} \\ 1 & x_{21} & x_{22} & \cdots & x_{2k} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{nk} \end{bmatrix}, \quad \boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{bmatrix}, \quad \boldsymbol{\epsilon} = \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix}$$

Multicollinearity

Using the least squares method, the estimator of $\boldsymbol{\beta} = (\beta_0, \beta_1, \cdots \beta_k)^T$ is given by

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$$

provided that $\mathbf{X}^T\mathbf{X}$ is invertiable.

This will requires that the columns of $\mathbf{X}$ be linearly independent.

When the predictor variables are dependent (correlated) with each other, then we will have the problem called multicollinearity.

Multicollinearity

Example

Suppose we have a regression model with the fitted value given by

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \hat{\beta}_3 x_3$$

but $x_3 = 4x_1 + 3x_2$.

Then x_3 is redundant and carry no new information. Worst, it will make the estimated slopes in the regression model arbitrary as

$$\begin{aligned}
\hat{y} &= \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \hat{\beta}_3 x_3 \\
&= \hat{\beta}_0 + (\hat{\beta}_1 + 4\hat{\beta}_3) x_1 + (\hat{\beta}_2 + 3\hat{\beta}_3) x_2 \\
&= \hat{\beta}_0 + \left(\hat{\beta}_2 - \frac{3}{4}\hat{\beta}_1\right) x_2 + \left(\hat{\beta}_3 + \frac{\hat{\beta}_1}{4}\right) x_3 \\
&= \cdots
\end{aligned}$$

Multicollinearity

Examples of correlated predictor variables are:

- A person's height and weight in relation to the body mass index (BMI)
- Level of education and age of a person in relation to their annual salary
- Arm length, leg length, height, weight of a person in relation to their physical strength.

Effect of Multicollinearity

Suppose we have only two predictor variables x_1 and x_2. Let us consider a regression model

$$y = \beta_1 x_1 + \beta_2 x_2 + \epsilon$$

where x_1, x_2 and y are scaled to unit length, that is $\overline{x}_1 = \overline{x}_2 = 0$ with

$$x_{11}^2 + x_{21}^2 = 1 \;\text{ and }\; x_{12}^2 + x_{22}^2 = 1.$$

Hence with

$$\mathbf{X} = \begin{bmatrix} x_{11} & x_{12} \\ x_{21} & x_{22} \end{bmatrix} \;\text{ we see that }\; \mathbf{X}^T\mathbf{X} = \begin{bmatrix} 1 & r_{12} \\ r_{12} & 1 \end{bmatrix}$$

where r_{12} is the correlation between x_1 and x_2

$$r_{12} = \frac{\sum (x_{i1} - \overline{x}_1)(x_{i2} - \overline{x}_2)}{\sqrt{\sum (x_{i1} - \overline{x}_1)^2 \sum (x_{i2} - \overline{x}_2)^2}}$$

Effect of Multicollinearity

From our previous discussion, we have

$$\mathrm{Var}(\hat{\boldsymbol{\beta}}) = \sigma^2(\mathbf{X}^T\mathbf{X})^{-1} = \frac{\sigma^2}{1 - r_{12}^2}\begin{bmatrix} 1 & -r_{12} \\ -r_{12} & 1 \end{bmatrix}$$

This gives

$$\mathrm{Var}(\hat{\beta}_1) = \frac{\sigma^2}{1 - r_{12}^2} = \mathrm{Var}(\hat{\beta}_2)$$

and

$$\mathrm{Cov}(\hat{\beta}_1, \hat{\beta}_2) = -\frac{r_{12}\sigma^2}{1 - r_{12}^2}.$$

From here we see that

$$\mathrm{Var}(\hat{\beta}_1) = \mathrm{Var}(\hat{\beta}_2) \to \infty$$

as $r_{12} \to 1$ (or -1).

Effect of Multicollinearity

The equation below

$$\mathrm{Var}(\hat{\beta}_1) = \mathrm{Var}(\hat{\beta}_2) \to \infty$$

as $r_{12} \to 1$ (or -1)

tell us that when there is a strong multicollinearity, i.e. when $|r_{12}| \approx 1$, we will have a large variances and covariances for the least squares estimators.

Hence a slightly different sample can lead to a vastly different estimates of the model parameters.

When there are more than two predictor variables, multicollinearity produces similar effect.

Effect of Multicollinearity

Another effect of having Multicollinearity in a model is that the parameters tend to be overestimates in magnitude.

For this, we look at $||\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}||^2 = \sum_j (\hat{\beta}_j - \beta_j)^2$. Taking the expectation leads to

$$\begin{aligned}
E\left(||\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}||^2\right) &= \sum_j E\left[(\hat{\beta}_j - \beta_j)^2\right] \\
&= \sum_j \mathrm{Var}(\hat{\beta}_j) \\
&= \mathrm{Trace}\left(\mathrm{Var}(\hat{\boldsymbol{\beta}})\right) \\
&= \sigma^2 \mathrm{Trace}\left((\mathbf{X}^T\mathbf{X})^{-1}\right) \\
&= \frac{2\sigma^2}{1 - r_{12}^2}
\end{aligned}$$

Effect of Multicollinearity

The equation

$$E\left(||\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}||^2\right) = \frac{2\sigma^2}{1 - r_{12}^2}$$

shows that when there is a strong multicollinearity, $\hat{\boldsymbol{\beta}}$ is far from $\boldsymbol{\beta}$ on average.

Also, we can show that

$$E\left(||\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}||^2\right) = E\left(||\hat{\boldsymbol{\beta}}||^2\right) - ||\boldsymbol{\beta}||^2$$

Therefore

$$E\left(||\hat{\boldsymbol{\beta}}||^2\right) = ||\boldsymbol{\beta}||^2 + \frac{2\sigma^2}{1 - r_{12}^2}$$

This implies that when there is a strong multicollinearity, the length of the vector $\hat{\boldsymbol{\beta}}$ is on average, much larger than $\boldsymbol{\beta}$.

Multicollinearity Diagnostics

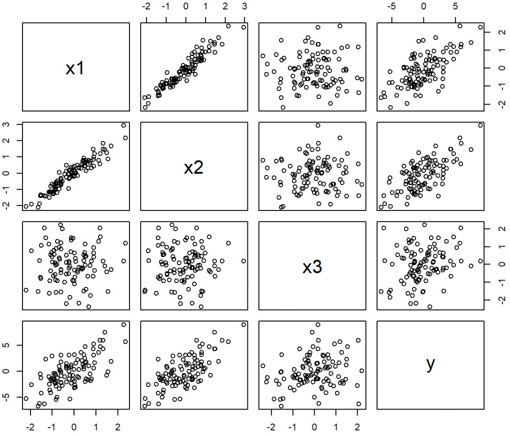
A very simple measure of multicollinearity is the inspection of the scatter plot and correlation matrix.

```
Multicol1<-read.csv("Multicol1.CSV", header=TRUE, sep=",")
x1<-Multicol1$x1
x2<-Multicol1$x2
x3<-Multicol1$x3
y<-Multicol1$y

plot(Multicol1)

#Calculate the correlation matrix
myCorr=cor(Multicol1)
myCorr
```

```
           x1         x2         x3         y
x1 1.00000000 0.95257460 0.01582233 0.6732043
x2 0.95257460 1.00000000 0.02461845 0.6911211
x3 0.01582233 0.02461845 1.00000000 0.2165981
y  0.67320428 0.69112108 0.21659812 1.0000000
```

Multicollinearity Diagnostics

Just like what we have seen in the covariance matrix, the scatter plot also indicates the collinearity of x_1 and x_2.



Variance Inflation Factors

The scatter plot and covariance matrix are bivariate method in the sense that they detect the relationship between two variables.

We now develop a method to detect whether there is a relationship between one predictor with a linear combination of the rest of the predictors.

For that, we regress each single predictor x_j, $j = 1, \cdots, k$ on the remaining ones, i.e.

$$x_j \sim x_1 + \cdots + x_{j-1} + x_{j+1} + \cdots + x_k$$

and compute the corresponding coefficients of determination R_j^2.

Variance Inflation Factors

From

$$x_j \sim x_1 + \cdots + x_{j-1} + x_{j+1} + \cdots + x_k$$

the value of R_j^2 tells us how well is x_j describable by the other variables.

Hence a large value of R_j^2 indicates strong linear dependence of x_j on the other predictors.

This then implies that there is multicollinearity of the predictors in the model.

Variance Inflation Factors

Instead of comparing each R_j^2, we define the variance inflation factors (VIF) of the predictors as

$$\mathrm{VIF}_j = \frac{1}{1 - R_j^2}$$

Notice that with VIF_j so defined,

- when x_j is nearly a combination of the other predictors:

$$R_j^2 \approx 1 \quad \text{hence} \quad \mathrm{VIF}_j \ \text{ is large}$$

- when x_j is orthogonal to all the other predictors:

$$R_j^2 = 0 \quad \text{hence} \quad \mathrm{VIF}_j = 1$$

Variance Inflation Factors

In general,

$$\mathrm{VIF}_j = \frac{1}{1 - R_j^2}$$

the larger these factors are, the more worry you should be about having multicollinearity in the model. As a rule of thumb:

If $\mathrm{VIF}_j > 4$ we should investigate, while

If $\mathrm{VIF}_j > 10$ we should act and remediate.

Variance Inflation Factors

Running the VIF on R

```
model1=lm(y~x1+x2+x3)
#variance inflation factor
install.packages('car')
library(car)
vif(model1)
```

```
> vif(model1)
       x1        x2        x3
10.805740 10.809586  1.001236
```

We see that x_1 and x_2 have a high VIF value, the result is in accordance to previous diagnostic analysis using scatter plot and the covariance matrix.

Dealing with Multicollinearity

When dealing with multicollinearity, these are the strategies that one can try:

- Collecting additional data to break up the multicollinearity in the existing data.
- Model respecification.
  - (i) If x_1, x_2 and x_3 are nearly linearly dependent, it may be possible to find some function such as $x = (x_1 + 2x_2)/x_3$ that preserves the original predictors but reduces the ill-conditioning.
  - (ii) Variable elimination: If x_1, x_2 and x_3 are nearly linearly dependent, eliminating one of the predictor (say x_1) may help in reducing the effect of multicollinearity.

Ridge Regression

As we have observed in the previous slide. When a data contains multicollinearity, the magnitude of $\boldsymbol{\beta}$ will be inflated (on average).

This implies that the confidence intervals for the slope parameters will tend to be wide and estimation of the slopes will be unstable.

Ridge Regression

The main idea behind ridge regression is to add a small bias to help shrinks the estimated coefficients towards zero to fix the magnitude inflation.

That is, we find the minimum of

$$S = \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{k} \beta_j x_{ji} \right)^2 + \lambda \sum_{j=1}^{k} (\beta_j)^2$$

The added term

$$\lambda \sum_{j=1}^{k} (\beta_j)^2$$

is the penalty that shrinks our coefficients.

Ridge Regression

In the minimization of S below

$$S = \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{k} \beta_j x_{ji} \right)^2 + \lambda \sum_{j=1}^{k} (\beta_j)^2$$

the parameter λ needs to fit the linear model and shrinking the coefficients.

The selection of λ will be done by using the smallest value of Generalized Cross Validation (GCV) value.

Ridge Regression

Remark:

We need to perform unit normal scaling on the columns of $\mathbf{X}$ before performing ridge regression.

This is because the penalty term would be unfair to other predictors if they are not on the same scale.

Ridge Regression in R

Looking at the same data set $\mathrm{Multicol1.csv}$, we create a unit normal scaling on the column of $\mathbf{X}$ as follows:

```
Multicol1<-read.csv("Multicol1.CSV", header=TRUE, sep=",")
x1<-Multicol1$x1
x2<-Multicol1$x2
x3<-Multicol1$x3
y<-Multicol1$y

#Standardize each variable(subtract mean, divided by sd)
Multicol1S=data.frame(scale(Multicol1))

xs1<-Multicol1S$x1
xs2<-Multicol1S$x2
xs3<-Multicol1S$x3
y<-Multicol1$y
```

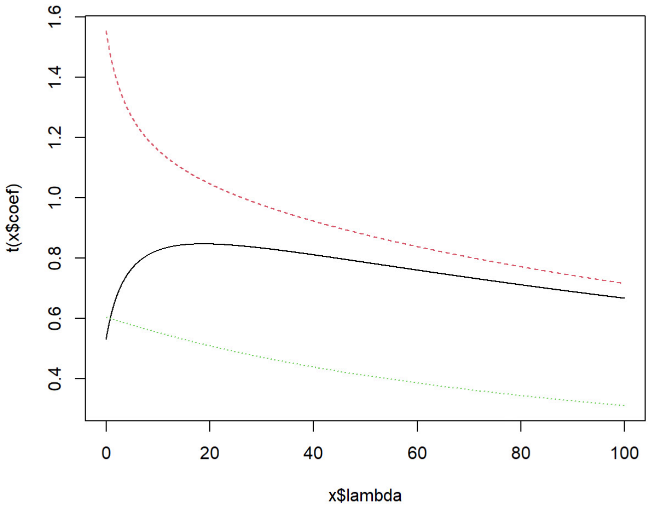
Ridge Regression in R

The ridge regression in R is performed as follows:

```
#Ridge Regression
install.packages('MASS')
library(MASS)

reg_seq=seq(0,100,0.001)
fit=lm.ridge(y~xs1+xs2+xs3, lambda = reg_seq)
plot(fit)
```

Ridge Regression in R

From the values we tried for various λ, we see how the coefficients shrink as λ grows larger:



Ridge Regression in R

Using $\mathrm{select(fit)}$ we obtain the following outputs:

```
> select(fit)
modified HKB estimator is 1.471051
modified L-W estimator is 0.9599733
smallest value of GCV  at 9.806
```

We will use the smallest value of Generalized Cross Validation (GCV) to be the value of our λ. This gives $\lambda = 9.806$ Running the ridge regression with this λ gives

```
model1=lm.ridge(y~xs1+xs2+xs3, lambda = 9.806)
model1$coef
```

```
> model1$coef
      xs1       xs2       xs3
0.8253835 1.1616990 0.5544383
```