Statistical Linear Models

Collinearity and its effects

In multiple regression analysis, the nature and significance of the relations between the predictor variables and the response variable are often of particular interest.

Some questions frequently asked are:

1. What is the relative importance of the effects of the different predictor variables?
2. What is the magnitude of the effect of a given predictor variable on the response variable?
3. Can any predictor variable be dropped from the model because it has little or no effect on the response variable?
4. Should any predictor variables not yet included in the model be considered for possible inclusion?

Relatively simple answers can be given to these uestions If the predictor variables included in the model are

i. Uncorrelated among themselves

ii. Uncorrelated with any other predictor variables that are related to the response but are omitted from the model.

Unfortunately, in many non-experimental situations in business, economics & social & biological sciences, the predictor variables tend to be correlated among themselves and other variables related to the response but not included in the model.

Example: $Y_i = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \varepsilon_i$

Y_i: Family food expenditures; X_1: family income; X_2: family savings; X_3: age of head of household

Are predictor variables correlated? Are predictor variables correlated with the response? Are they correlated with other variables not included in the model that affect food expenditures?
(e.g. family size)

> Collinearity occurs when predictor variables are correlated among themselves.

Example: Uncorrelated Predictor Variables

Suppose we have data for a small-scale experiment on the effect of crew size (size) and level of bonus pay (pay) on crew productivity (prod).

The predictor variables are uncorrelated.

```
> summary(lm(prod ~ size + pay))
Call:
lm(formula = prod ~ size + pay)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   0.3750     4.7405   0.079 0.940016
size          5.3750     0.6638   8.097 0.000466 ***
pay           9.2500     1.3276   6.968 0.000937 ***

Residual standard error: 1.877 on 5 degrees of freedom
Multiple R-squared: 0.958,     Adjusted R-squared: 0.9412
F-statistic: 57.06 on 2 and 5 DF,  p-value: 0.000361

> summary(lm(prod ~ size))
Call:
lm(formula = prod ~ size)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   23.500     10.111   2.324   0.0591 .
size           5.375      1.983   2.711   0.0351 *
Residual standard error: 5.609 on 6 degrees of freedom
Multiple R-squared: 0.5505,     Adjusted R-squared: 0.4755
F-statistic: 7.347 on 1 and 6 DF,  p-value: 0.03508

> summary(lm(prod ~ pay))
Call:
lm(formula = prod ~ pay)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   27.250     11.608   2.348   0.0572 .
pay            9.250      4.553   2.032   0.0885 .

Residual standard error: 6.439 on 6 degrees of freedom
Multiple R-squared: 0.4076,     Adjusted R-squared: 0.3088
F-statistic: 4.128 on 1 and 6 DF,  p-value: 0.08846
```

Note: The regression coefficient for size (5.3750) whether or not pay is included in the model.

(The same holds true for size (9.25).)

This is a result of the predictor variables being uncorrelated.

Adding or removing uncorrelated predictors to the regression model does not change the regression coefficient.

Nature of Problem

Suppose we have the following data:

Data

Case	X_{1i}	X_{2i}	Y_i
1	2	6	23
2	8	9	83

Ricky fits a multiple regression function.

$$\hat{Y} = -87 + X_1 + 18 X_2$$

Case	X_{1i}	X_{2i}	Y_i
1	2	6	23
2	8	9	83
3	6	8	63
4	10	10	103

$$\hat{Y} = -87 + X_1 + 18X_2$$

He's proud because his response function fits the data perfectly!

$$23 = -87 + 2 + 18(6)$$

Kinjal also fits a multiple regression function.

$$\hat{Y} = -7 + 9X_1 + 2X_2$$

Her response function likewise fits the data perfectly!

$$23 = -7 + 9(2) + 2(6)$$

In fact, we can show that infinitely many response functions will fit the data perfectly. The reason is that the predictor variables, X_1 + X_2, are perfectly related.

$$X_2 = 5 + 0.5X_1$$

Two key implications of this example are:

1. The perfect relation between X_1 + X_2 did not inhibit our ability to obtain a good fit to the data.
2. Since many different response functions provide the same good fit, we cannot interpret any one set of regression coefficients as reflecting the effects of the different predictor variables.

Ricky: $\hat{\beta}_1 = 1$ $\quad \hat{\beta}_2 = 18$

Kinjal: $\hat{\beta}_1 = 9$ $\quad \hat{\beta}_2 = 2$

In practice, we seldom find predictor variables that are perfectly related or data that do not contain some random error component.

Note. If collinearity is perfect, the mathematics that underlie regression analysis will fail because $X'X$ will not be invertible.

However, when predictor variables are highly correlated ($\sim > 0.8$) collinearity can occur.

Consequences of Collinearity

1. Coefficient estimates are unstable.
   - coefficient standard errors are large due to imprecision of

β's estimation.

- Confidence intervals are broad.
- Sensitive to changes in model specification. (e.g. dropping a variable or excluding some observations).

2. It becomes difficult or impossible to separate the effects of changes in the individual variables. (although we can still get good predictions)

Detection of Collinearity

1. High R^2 or adjusted R^2 and insignificant coefficients (p-values > 0.10)

This is the most evident sign! You have a model with good explanatory power but very few or no significant estimated coefficients. This situation should scream "Collinearity" to you.

2. High correlation coefficients between predictor variables.
3. Variance Inflation Factors (VIFs)

To calculate VIF's we regress each X_i on the other predictors

$$X_{1i} = \alpha_0 + \alpha_1 X_{2i} + \alpha_2 X_{3i} + \dots + \delta_i$$

Then take the R^2 from this regression to calculate

$$\boxed{VIF = \frac{1}{1-R^2}}$$

A VIF of greater than maybe 5 or 6 suggests that the variable contains some sort of collinearity.

Advantage of VIF's over correlation coefficients

VIF's allow the collinearity to be between more than two explanatory variables while correlation coefficients allow only for bivariate (two variable relationships)

Remedies for Collinearity

1. Collect new data in such a manner that the problem is avoided.

<u>Ex:</u> Y = athletic performance

X_1 = heat $\Longrightarrow$ X_1 = heat (on hot dry days)

X_2 = humidity $\quad$ X_2 = humidity (on cold, damp days)

2. Drop one or more of the collinear variables.

However, we run the risk of having biased coefficients & standard errors that are too small.

3. Respecify the model. (not ideal)

Perhaps several regressors can be combined or one can be chosen to represent the others.

<u>Note:</u> Respecification is possible only where the original model was poorly thought out or where the researcher is willing to abandon some of the goals of the research.

Errors in Predictors

The regression model $Y = X\beta + \varepsilon$ allows for Y being measured with error by having the error term. But what if the X is measured with error?

Consider observing (x_i^0, y_i^0) for $i = 1, \dots, n$ which are related to the true values by

$$y_i^0 = y_i^A + \varepsilon_i \qquad \varepsilon_i \text{ \& } \delta_i \text{ independent}$$

$$x_i^0 = x_i^A + \delta_i$$

The true relationship is:

$$y_i^A = \beta_0 + \beta_1 x_i^A \qquad \left(\text{but we only observe } x_i^0 \text{ \& } y_i^0\right)$$

Putting it together, we get:

$$y_i^0 = \beta_0 + \beta_1 x_i^0 + (\varepsilon_i - \beta_1 \delta_i)$$

Suppose we use least squares to estimate β_0 & β_1.

Assume $E[\varepsilon_i] = E[\delta_i] = 0$ and

$$var[\varepsilon_i] = \sigma_\varepsilon^2, \quad var[\delta_i] = \sigma_\delta^2.$$

Let

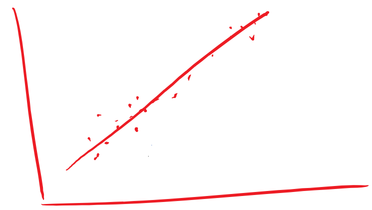
$var[X] = \sigma_x^2$. Then we can show that

$$E[\hat{\beta}_1] = \beta_1 \frac{1}{1 + \frac{\sigma_\delta^2}{\sigma_x^2}}$$

$\hat{\beta}_1$ will be biased towards 0, regardless of the sample size.

```
> x <- 10*runif(50)
> y <- x+rnorm(50)
> gx <- lm(y~x)
> summary(gx)
```

$y = x +$ noise



```
Call:
lm(formula = y ~ x)

Residuals:
     Min       1Q   Median       3Q      Max
-1.83966 -0.83433  0.00421  1.00956  1.69865

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.23419    0.25597  -0.915    0.365
x            1.07443    0.04734  22.697   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.031 on 48 degrees of
freedom
Multiple R-squared: 0.9148,  Adjusted R-squared:
0.913
F-statistic: 515.2 on 1 and 48 DF,  p-value: < 2.2e-16
```

$Y = 1.074\ x$ ✓

$R^2 = .915$

```
> z <- x + rnorm(50)
> gz <- lm(y~z)
> summary(gz)
```

Before $y = x +$ noise

Now $z = x +$ noise

correlated

$y = \beta_0 + \beta_1 z$

```
Call:
lm(formula = y ~ z)

Residuals:
    Min      1Q  Median      3Q     Max
-2.0384 -0.9196 -0.1296  0.8948  3.7659

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.44366    0.31181   1.423    0.161
z            0.96508    0.05792  16.663   <2e-16 ***
```

$Y = .96\ z$

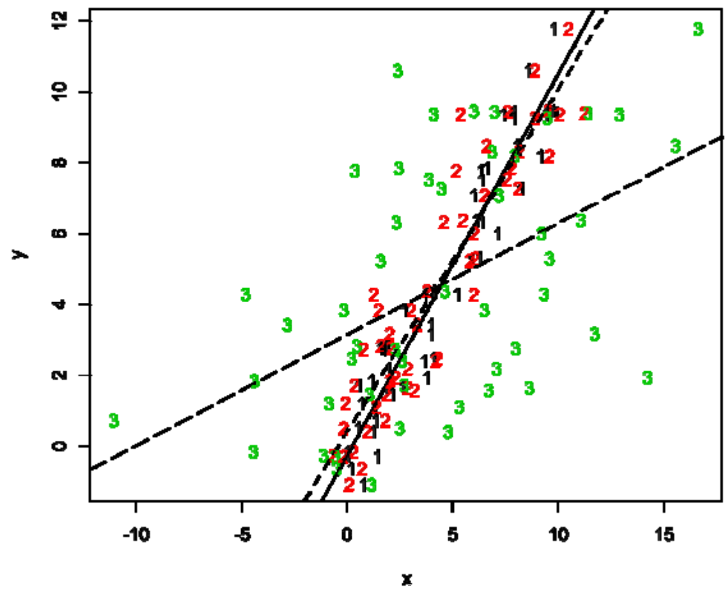
```
z          0.96508    0.05792  16.663   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.355 on 48 degrees of
freedom
Multiple R-squared: 0.8526,     Adjusted R-squared:
0.8495
F-statistic: 277.7 on 1 and 48 DF,  p-value: < 2.2e-16


> z2 <- x+5*rnorm(50)
> gz2 <- lm(y ~z2)
> summary(gz2)

Call:
lm(formula = y ~ z2)

Residuals:
    Min      1Q  Median      3Q     Max
-5.6468 -2.6922 -0.1855  2.6678  6.7303

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  3.15538    0.55136   5.723 6.63e-07 ***
z2           0.31364    0.07768   4.038 0.000194 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.05 on 48 degrees of
freedom
Multiple R-squared: 0.2535,     Adjusted R-squared:
0.238
F-statistic:  16.3 on 1 and 48 DF,  p-value: 0.0001936
```

unstable!

```
> matplot(cbind(x,z,z2),y,xlab="x",ylab="y")
> abline(gx, lty=1)
> abline(gz, lty=2)
> abline(gz2, lty=5)
```

$z_2 = x + 5(\text{noise})$

$y = \beta_0 + \beta_1 z_2$



Do this 1000 times!

```
> bc <- numeric(1000)
> for(i in 1:1000){
+ y <- x+rnorm(50)
+ z<- x+5*rnorm(50)
+ g <- lm(y~z)
+ bc[i] <- g$coef[2]
+ }
>
> summary(bc)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 0.0926  0.2449  0.2820  0.2818  0.3223  0.4829
```

$y = x + \text{noise}$

$z = x + 5 * \text{noise}$

$y = \beta_0 + \beta_1 z$

range of β_1 coefficients.