ST430: Introduction to Regression Analysis, Chapter 7

Luo Xiao

October 28, 2015
Regression pitfalls

Pitfall
Noun:

- A hidden or unsuspected danger or difficulty.
- A covered pit used as a trap.

Multiple regression is a widely used and powerful tool.

It is also one of the most abused statistical techniques.
Observational versus experimental data

Recall:
In some investigations, the independent variables $X_1, X_2, \ldots, X_k$ can be \textit{controlled}; that is, held at desired values.

The resulting data are called \textit{experimental}.

In other cases, the independent variables cannot be controlled, and their values are simply observed.

The resulting data are called \textit{observational}.
Observational example

“Cocaine Use During Pregnancy Linked To Development Problems”

- Two groups of new mothers, 218 used cocaine during pregnancy, 197 did not.
- IQ tests of infants at age 2 showed lower scores for children of users.

“Correlation does not imply causation.”
The study does not show that cocaine use causes development problems.

It does show association, which might be used in prediction.

For instance, it could help identify children at high risk of having development problems.
Experimental example
Animal-assisted therapy.

76 heart patients randomly assigned to three therapies:
- T: visit from a volunteer and a trained dog;
- V: visit from a volunteer only;
- C: no visit.

Response $y$ is decrease in anxiety.
Result:

\[ \bar{y}_T = 10.5, \quad \bar{y}_V = 3.9, \quad \bar{y}_C = 1.4. \]

Model:

\[ E(Y) = \beta_0 + \beta_1 X_1 + \beta_2 X_2, \]

where \( X_1 \) is the indicator variable for group T and \( X_2 \) is the indicator variable for group V.

The model-utility \( F \)-test shows significant differences among groups.

Because of random assignment, the differences can be assumed to be caused by the treatments.
Parameter estimability

Recall

The *normal* equations

\[ X'X\hat{\beta} = X'y \]

that define least squares parameter estimates always have a solution.

But if \( X'X \) is singular, they have *many* solutions. So we need at least \( n \geq k + 1 \) observations.

An individual parameter that is not uniquely estimated is called *nonestimable.*
Multicollinearity

Two independent variables are *orthogonal* if their sample correlation coefficient is zero.

If *all* pairs of independent variables are orthogonal, $\mathbf{X}'\mathbf{X}$ is diagonal, and the normal equations are trivial to solve.

In a controlled experiment, the variables are often orthogonal by design.

If some pairs are far from orthogonal, the equations may be *nearly* singular.
If $\mathbf{X}'\mathbf{X}$ is nearly singular, its inverse $(\mathbf{X}'\mathbf{X})^{-1}$ exists but will have large entries.

So the least squares estimates

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

are very sensitive to small changes in $\mathbf{y}$.

The covariance matrix of least squares estimates

$$\text{Cov}(\hat{\beta}) = \sigma^2 (\mathbf{X}'\mathbf{X})^{-1}.$$ 

That makes their standard errors large.
Example

Carbon monoxide from cigarettes (output in "output1.txt")

```r
setwd("~/Dropbox/teaching/2015Fall/R_datasets/Exercises&Examples")
load("FTCCIGAR.Rdata")
cigar = FTCCIGAR

pairs(cigar)
cor(cigar)
summary(lm(CO ~ TAR, cigar))
summary(lm(CO ~ TAR + NICOTINE + WEIGHT, cigar))
```
The standard error of $\hat{\beta}_{\text{TAR}}$ increases nearly five-fold when NICOTINE is added to the model.

When the response and all predictors are standardized, then it can be shown that

$$s_{\beta_i}^2 = s^2 \left( \frac{1}{1 - R_i^2} \right),$$

where $s^2$ is the estimate of $\sigma^2$ and $R_i^2$ is the R-squared for the model that regresses $X_i$ on all other independent variables $X_1, X_2, \ldots, X_{i-1}, X_{i+1}, \ldots, X_k$.

**Variance inflation factor:**

$$\text{VIF}_i = 1/(1 - R_i^2)$$
VIF\textsubscript{i} is related to the increase in the standard error of $\hat{\beta}_i$ when the other variables are included.

\[ VIF_i = 1 \text{ if } X_i \text{ is orthogonal to the other independent variables.} \]

How to detect multicollinearity?

- High correlation between pairs of independent predictors;
- High VIF ($>10$) for some parameter;
- When the overall model utility $F$-test is significant but none of the parameters are significant.

One solution to multicollinearity:

Remove one or more highly correlated variables.
Extrapolation

A regression model is an approximation to the complexities of the real world.

It may fit the sample data well.

If it fits well, it will usually give a reliable prediction for a new context that is similar to those in the sample data.

With several variables, deciding when the new context is too different for reliable prediction may be difficult, especially in the presence of multicollinearity.
Transformation

In many problems, one or more of the variables (dependent and independent) may be measured and recorded in a form that is not the best from a modeling perspective.

Linear transformations are usually pointless, as a linear model is essentially unchanged by it.

Among nonlinear transformations, logarithms are most widely useful, followed by powers of the variables.
The primary goal of transformation is to find a good approximation to the way $E(Y)$ depends on $X$.

Another goal is to make the variance of the random error

$$
\epsilon = Y - E(Y)
$$

reasonably constant: *variance stabilizing transform*.

Finally, if a transformation makes $\epsilon$ approximately normally distributed, that is worth achieving.
Example 7.8

Impact of price of coffee on demand:

Example 7.8 models \( Y ('DEMAND') \) against \( X \), where \( X = 1/\text{PRICE} \).

```r
setwd("~/Dropbox/teaching/2015Fall/R_datasets/Exercises&Examples")
load("COFFEE.Rdata")

par(mfrow=c(1,2))
plot(DEMAND~PRICE,COFFEE,pch=20)
plot(DEMAND~X,COFFEE,pch=20) #X is 1/PRICE
```
R code:

```r
setwd("~/Dropbox/teaching/2015Fall/R_datasets/Exercises&Examples")
load("COFFEE.Rdata")

par(mfrow=c(1,2))
plot(DEMAND~PRICE,COFFEE)
plot(DEMAND~X,COFFEE)  #X is 1/PRICE

fit = lm(DEMAND~PRICE,COFFEE)
summary(fit)

fit2 = lm(DEMAND~X,COFFEE)
summary(fit2)

extractAIC(fit)
extractAIC(fit2)
```
The model with 1/’PRICE’ as predictor has higher $R_a^2$ and smaller AIC.