STAT430.Multiple.Reg

Benjamin Kedem

June 2020

1 Multiple Regression

We have observations $y_1, ..., y_n$ such that each y_i depends on its covariates $x_{1i}, .., x_{ki}$ by a **linear model**:

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

where, as in simple linear regression, the y_i are random variables, the x's are design non-random variables, and the ϵ_i are random errors such that:
$E(\epsilon_i) = 0$
$Var(\epsilon_i) = \sigma^2$
$Cov(\epsilon_i, \epsilon_j) = 0, \ i \neq j$

So we have:

$$\begin{aligned} y_1 &= \beta_0 + \beta_1 x_{11} + \cdots \beta_k x_{k1} + \epsilon_1 \\ y_2 &= \beta_0 + \beta_1 x_{12} + \cdots \beta_k x_{k2} + \epsilon_2 \\ &\cdots\cdots\cdots\cdots\cdots\cdots \\ &\cdots\cdots\cdots\cdots\cdots\cdots \\ y_n &= \beta_0 + \beta_1 x_1 + \cdots \beta_k x_{kn} + \epsilon_i \end{aligned}$$

It is convenient to use matrix notation:

$$\begin{pmatrix} y_1 \\ y_2 \\ . \\ . \\ . \\ y_n \end{pmatrix} = \begin{pmatrix} 1 & x_{11} & x_{21} & . & . & . & x_{k1} \\ 1 & x_{12} & x_{22} & . & . & . & x_{k2} \\ & . & . & . & . & . & . \\ & . & . & . & . & . & . \\ & . & . & . & . & . & . \\ 1 & x_{1n} & x_{2n} & . & . & . & x_{kn} \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \\ . \\ . \\ . \\ \beta_k \end{pmatrix} + \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ . \\ . \\ . \\ \epsilon_n \end{pmatrix}$$

Or

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

To estimate $\boldsymbol{\beta}$ we use the **least squares** method by minimizing $\boldsymbol{\epsilon}'\boldsymbol{\epsilon}$ w.r.t. $\boldsymbol{\beta}$:

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

where we assume that X has full rank for the inverse to exist.

We can show:
$E(\hat{\boldsymbol{\beta}}) = \beta$
$Var(\hat{\boldsymbol{\beta}}) = \sigma^2(X'X)^{-1}$
Gauss-Markov: $\mathbf{c}'\hat{\boldsymbol{\beta}}$ is the Best Linear Unbiased Estimate (BLUE) of $\mathbf{c}'\boldsymbol{\beta}$.

Again, we have the same basic decomposition of the total (corrected) sum of squares:

$$\sum(y_i - \bar{y})^2 = \sum(y_i - \hat{y}_i)^2 + \sum(\hat{y}_i - \bar{y})^2$$

Or with $p = k + 1$, $k = p - 1$ (number of slopes),

$$SST(df = n - 1) = SSE(df = n - p) + SSR(p - 1)$$

and to test $H_0 : \beta_1 = \cdots \beta_k = 0$ we use the test statistics,

$$\frac{SSR/k}{SSE/(n-p)} \sim F_{k,n-p}$$

Example: Antelope

```
The data (X1, X2, X3, X4) are for each year.
X1 = spring fawn count/100
X2 = size of adult antelope population/100
X3 = annual precipitation (inches)
X4 = winter severity index (1=mild,
5=severe)

DATA ANTELOPE;\\
INPUT   X1 X2 X3 X4;\\
DATALINES;\\
2.9 9.2 13.2 2
2.4 8.7 11.5 3
2.0 7.2 10.8 4
2.3 8.5 12.3 2
3.2 9.6 12.6 3
1.9 6.8 10.6 5
3.4 9.7 14.1 1
2.1 7.9 11.2 3
;

PROC REG DATA=ANTELOPE;
/*PRESICTED, RESIDUALS*/
MODEL X1=X2 X3 X4/P R;
RUN;
```

```
The REG Procedure
Model: MODEL1
Dependent Variable: x1
Number of Observations Read 8
Number of Observations Used 8

Analysis of Variance
Source          DF SS          MS    F Value    Pr > F
Model           3   2.21651 0.73884   50.52    0.0012
Error           4   0.05849 0.01462
Corrected Total 7    2.27500

Root MSE       0.12093           R-Square 0.9743
Dependent Mean 2.52500           Adj R-Sq 0.9550
Coeff Var      4.78921


Parameter Estimates

Variable    DF Parameter Estimate        Standard Error t Value   Pr > |t|
Intercept  1    -5.92201                    1.25562      -4.72    0.0092
x2         1     0.33822                    0.09947       3.40    0.0273
x3         1     0.40150                    0.10990       3.65    0.0217
x4         1     0.26295                    0.08514       3.09    0.0366


The REG Procedure
Model: MODEL1
Dependent Variable: x1
Output Statistics
```

Obs	y	$\hat{y}$	SE $\hat{y}$	Resid	SE Resid	Student Resid	Cook's D
1	2.9	3.0153	0.0645	-0.1153	0.102	-1.128	0.126
2	2.4	2.4266	0.0847	-0.0266	0.0863	-0.308	0.023
3	2.0	1.9012	0.0684	0.0988	0.0997	0.991	0.116
4	2.3	2.4172	0.0728	-0.1172	0.0965	-1.214	0.210
5	3.2	3.1727	0.1054	0.0273	0.0593	0.461	0.167
6	1.9	1.9485	0.1058	-0.0485	0.0585	-0.830	0.564
7	3.4	3.2828	0.0955	0.1172	0.0742	1.580	1.034
8	2.1	2.0356	0.0758	0.0644	0.0943	0.683	0.075

Application of Multiple Regression: Fitting a Sinusoid

We wish to fit a sinusoid to data x_t.

$$x_t = \mu + \alpha\cos(\omega t) + \beta\sin(\omega t) + \epsilon_t, \quad t = 1, ...N$$

where ϵ_t are iid $N(0, \sigma^2)$, and N is even.

The problem is to estimate ω. For that, we'll fix ω and first estimate μ, α, β by least squares. This will give us a clue as to how to estimate ω.

For $\omega, \lambda \in \Omega = \{\frac{2\pi k}{N}, k = 1, ..., \frac{N}{2} - 1\}$ we have the following orthogonality relationships.

$$\sum_{t=1}^{N}\cos(\omega t) = \sum_{t=1}^{N}\sin(\omega t) = 0$$

$$\sum_{t=1}^{N}\cos(\omega t)\sin(\lambda t) = 0, \ \forall\lambda, \omega \in \Omega$$

$$\begin{aligned}
\sum_{t=1}^{N}\cos(\omega t)\cos(\lambda t) &= 0, \ \lambda \neq \omega \\
&= N/2, \ \lambda = \omega \\
\sum_{t=1}^{N}\sin(\omega t)\sin(\lambda t) &= 0, \ \lambda \neq \omega \\
&= N/2, \ \lambda = \omega
\end{aligned}$$

Now, in matrix notation we have,

$$\begin{pmatrix} x_1 \\ x_2 \\ . \\ . \\ . \\ x_N \end{pmatrix} = \begin{pmatrix} 1 & \cos(\omega) & \sin(\omega) \\ 1 & \cos(2\omega) & \sin(2\omega) \\ & . & . \\ & . & . \\ & . & . \\ 1 & \cos(N\omega) & \sin(N\omega) \end{pmatrix} \begin{pmatrix} \mu \\ \alpha \\ \beta \end{pmatrix} + \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ . \\ . \\ . \\ \epsilon_N \end{pmatrix}$$

Or

$$\boldsymbol{x} = \boldsymbol{A\theta} + \boldsymbol{\epsilon}$$

Therefore

$$\hat{\boldsymbol{\theta}} = (A'A)^{-1}A'x$$

Applying the orthogonality relationships we get:

$$\hat{\theta} = \begin{pmatrix} \hat{\mu} \\ \hat{\alpha} \\ \hat{\beta} \end{pmatrix} = \begin{pmatrix} \bar{x} \\ \frac{2}{N}\sum_{t=1}^{N} x_t \cos(\omega t) \\ \frac{2}{N}\sum_{t=1}^{N} x_t \sin(\omega t) \end{pmatrix}$$

Therefore,

$$R^2 = \frac{\sum_{t=1}^{N}(\hat{x}_t - \bar{x})^2}{\sum_{t=1}^{N}(x_t - \bar{x})^2} = \frac{\frac{N}{2}(\hat{\alpha}^2 + \hat{\beta}^2)}{\sum_{t=1}^{N}(x_t - \bar{x})^2}$$

But $\hat{\alpha}, \hat{\beta}$ are functions of ω! Therefore

$$R^2 = R^2(\omega)$$

and *we choose* ω *which maximizes* $R^2(\omega)$.

We can show that

$$R^2(\omega) \propto \frac{2}{N}\left|\sum_{t=1}^{N} x_t \exp(i\omega t)\right|^2$$

The resulting estimate $\hat{\omega}$ is very precise.

An Unbiased Estimate for σ^2

Using non-bold notation:

$$e = x - \hat{x} = A\theta + \epsilon - A\hat{\theta} = A\theta + \epsilon - A[(A'A)^{-1}A'(A\theta + \epsilon)] = [I - A(A'A)^{-1}A']\epsilon$$

Or with **idempotent** $M = I - A(A'A)^{-1}A'$,

$$e = M\epsilon$$

Hence,

$$E(e'e) = E[tr(\epsilon' M \epsilon)] = E[tr(M\epsilon\epsilon')] = tr(\sigma^2 M)$$

Or

$$E(e'e) = \sigma^2[tr(I) - tr[(A'A)^{-1}A'A]] = tr[I_{(N\times N)}] - tr[I_{(3\times 3)}] = \sigma^2(N-3)$$

Therefore,

$$S^2 = \frac{e'e}{N-3}$$

is unbiased for σ^2. In general, in the full rank case with p β's (including intercept):

$$S^2 = \frac{e'e}{n-p}$$

is unbiased for σ^2.

Model Selection Methods

When fitting a regression model, it is a good idea to fit several models and select the "best" model based on some criterion. SAS offsrs several criteria as follows.

1. Forward selection. It is a step-wise selection method by which a variable which enters never leaves when other variables are entertained.

2. Stepwise selection. It is a step-wise selection method by which a variable which enters could leave the model in subsequent steps.

3. A Information Criterion (AIC) invented by Hirotugo Akaike (1927-2009). We choose a model which minimizes with respect to p the quantity:

$$AIC(p) = -2 \log L(\hat{\boldsymbol{\beta}}) + 2p$$

where $\boldsymbol{\beta}$ is p-dimensional. Thus, p is the number of estimated parameters. Note that as p increases, $-2 \log L(\hat{\boldsymbol{\beta}})$ decreases while the "penalty" term $2p$ increases.

4. Bayesian Information Criterion (BIC) invented by Gideon Schwartz (1933-2007). As in the AIC, we choose a model which minimizes with respect to p the quantity:

$$BIC(p) = -2 \log L(\hat{\boldsymbol{\beta}}) + p \log(N)$$

where N is the number of data points. In general, the AIC and BIC results are close. That is, the optimal models are similar.

5. Mallows' C_p invented by Colin Mallows (1930-). It is a predecessor of the AIC. Again we choose a model which minimizes with respect to p the quantity:

$$C_p = \frac{SSE_p}{S^2} - N + 2p$$

where SSE_p is the residual SS from a reduced model with p parameters, N is the number of data points, and $S^2 = \hat{\sigma}^2$ from the full model with all the covariates.

Cook's distance D

Cook's distance D_i measures the influence of an observation y_i on the regression estimates. That is, D_i tells us if y_i stands out. $|D_i| > 2$ is considered large. If so, y_i deserves some special scrutiny.

Get the jth predicted value $\hat{y}_j$ from $y_1, ..., y_n$. Similarly, get the predicted value $\hat{y}_{j(i)}$ after deleting y_i. Then,

$$D_i = \frac{\sum_{i=j}^{n} (\hat{y}_j - \hat{y}_{j(i)})^2}{pS^2}$$

where

$$S^2 = \hat{\sigma}^2 = \frac{1}{n-p}\sum_{j=1}^{n}(y_j - \hat{y}_j)^2$$