Lecture 5: Multiple Linear Regression

Nancy R. Zhang

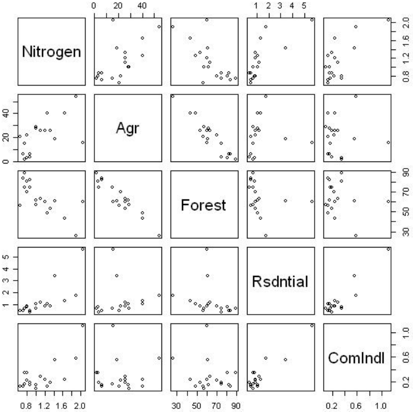
Statistics 203, Stanford University

January 19, 2010

Agenda

- Today: multiple linear regression.
- This week: comparing nested models in multiple linear regression.
- Finish diagnostics slides next lecture.

How does land use affect river pollution?



$$\text{Nitrogen} = \beta_0 + \beta_1 \text{Agr} + \beta_2 \text{Forest} + \beta_3 \text{Rsdntial} + \beta_4 \text{ComIndl} + \text{error}$$

Multiple Linear Regression

Design matrix:

$$X = \begin{pmatrix} 1 & X_{11} & X_{21} & \cdots & X_{p1} \\ 1 & X_{12} & X_{22} & & X_{p2} \\ \vdots & \ldots & \ddots & \vdots & \\ 1 & X_{1n} & X_{2n} & & X_{pn} \end{pmatrix} \quad y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}$$

Squared error loss function:

$$L(\beta) = \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j X_{ij} \right)^2 .$$

In matrix notation:

$$L(\beta) = (y - X\beta)'(y - X\beta).$$

Linear Subspaces and Projections

With p predictors we have $p+1$ vectors in $\Re^n$:

$$X_i = \begin{pmatrix} X_{i1} \\ X_{i2} \\ \vdots \\ X_{in} \end{pmatrix}, \quad i = 0, \ldots, p.$$

From now on we will always let X_0 be the vector of ones.

We denote by $\mathcal{L}(X_0, \ldots, X_p)$ the linear space spanned by the vectors $X_0, \ldots, X_p$:

$$\mathcal{L}(X_0, \ldots, X_p) = \left\{ \sum_{i=0}^{p} a_i X_i : (a_0, \ldots, a_p) \in \Re^{p+1} \right\}.$$

This is a linear subspace of $\Re^n$. We use the shorthand $\mathcal{L}(X)$.

- The *dimension* of $\mathcal{L}(X_0, \ldots, X_p)$ is equivalent to the rank of the matrix

$$X = \begin{pmatrix} X_{01} & X_{11} & \cdots & X_{p,1} \\ X_{02} & X_{12} & & X_{p,2} \\ \vdots & \ldots & \ddots & \vdots \\ X_{0n} & X_{1n} & & X_{p,n} \end{pmatrix}$$

- The rank of a matrix is equal to the number of linearly independent columns.
- The linear map that projects any vector $v \in \Re^n$ onto $\mathcal{L}(X)$ can be obtained by

$$P_X = X(X'X)^{-1}X'$$

.

Projection Matrices

Thus, for any $n \times p+1$ matrix X, we can construct a projection matrix $P_X = X(X'X)^{-1}X'$ that projects vectors onto the column space of X. Projection matrices enjoy some special properties:

1. $P_X^2 = P_X$.
2. $rank(P_X) = rank(X)$.
3. For any $v \in \mathcal{L}(X)$, $P_X v = v$.

For any linear space $\mathcal{L}_X$, its *null* space is the set

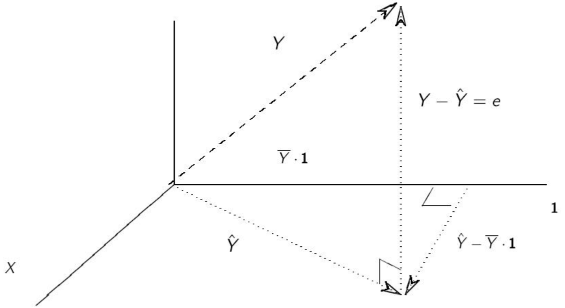
$$\mathcal{L}(X^{\perp}) = \{v \in \Re^n : Xv = 0\}$$

.

The projection matrix onto $\mathcal{L}(X^{\perp})$ is $I - P_X$.

Linear Regression by Least Squares = Projection

Simple linear regression with intercept:



$$\text{Project} \quad \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} \quad \text{onto} \quad \begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix}, \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix}.$$

Multiple Linear Regression

The solution

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

can also be obtained directly from the concept of a projection:

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

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

.

Calculating variances

Multivariate Gaussians

Let $Z \sim N(\boldsymbol{\mu}, \Sigma)$, and $\boldsymbol{a} \in \Re^n$, B an $n \times n$ matrix, then

$$a + BZ \sim N(a + B\mu, B\Sigma B').$$

$$\hat{\beta} = (X'X)^{-1}X'y, \quad y \sim N(X\beta, \sigma^2 I)$$

$$\begin{aligned} \Rightarrow E(\hat{\beta}) &= \beta \\ Var(\hat{\beta}) &= \sigma^2(X'X)^{-1} \end{aligned}$$

Calculating variance

Multivariate Gaussians

Let Z $N(\boldsymbol{\mu}, \Sigma)$, and $\boldsymbol{a} \in \Re^n$, B an $n \times n$ matrix, then

$$a + BZ \sim N(a + B\mu, B\Sigma B').$$

$$\hat{y} = P_X y$$

$$\begin{aligned} \Rightarrow E(\hat{y}) &= X\beta \\ Var(\hat{y}) &= \sigma^2 P_X \end{aligned}$$

The diagonal of P_X are the leverage values from last lecture.

t-tests for $\hat{\beta}_i$

$$\hat{\beta} \sim N(\beta, \sigma^2 (X'X)^{-1}).$$

As before, estimate σ^2 using

$$\hat{\sigma}^2 = SSE/(n-p-1)$$

Then, we can construct t-test by:

$$t_{\hat{\beta}_i} = \frac{\hat{\beta}_i}{s.e.(\hat{\beta}_i)}.$$

As before, reject the hypothesis $H_{i,0}: \ \hat{\beta}_i = 0$ at level α if

$$t_{\hat{\beta}_i} > t(n-p-1, \alpha/2).$$

Interpreting the $\hat{\beta}_i$'s

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p$$

The $\hat{\beta}_i$ obtained from a multiple regression is sometimes called partial regression coefficients because they correspond to a simple regression of Y on X_i, after taking out the effects of $X_j : \; j \neq i$.

1. Regress X_i on $\{X_j : j \neq i\}$, get residuals $r^i = X_i - \hat{X}_i$.
2. Regress Y_i on $\{X_j : j \neq i\}$, get residuals $r^Y = Y - \hat{Y}_{\sim i}$.
3. Do simple linear regression of r^Y on r^i, the slope will give you $\hat{\beta}_i$.

This gives us:

$$Var(\hat{\beta}_i) = \sigma^2 / \|r^i\|^2.$$

High correlation among the X's can “mask" out each other's effects.

Goodness of fit

Sums of squares

$$SSE = \sum_{i=1}^{n}(Y_i - \widehat{Y}_i)^2 = \sum_{i=1}^{n}(Y_i - \widehat{\beta}_0 - \widehat{\beta}_1 X_i)^2$$

$$SSR = \sum_{i=1}^{n}(\overline{Y} - \widehat{Y}_i)^2 = \sum_{i=1}^{n}(\overline{Y} - \widehat{\beta}_0 - \widehat{\beta}_1 X_i)^2$$

$$SST = \sum_{i=1}^{n}(Y_i - \overline{Y})^2 = SSE + SSR$$

$$R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST}.$$

$R = \sqrt{R^2}$ is called the multiple correlation coefficient.
R^2 is large: a lot of the variability in $\boldsymbol{Y}$ is explained by $\boldsymbol{X}$.

Large R^2 may not indicate a good model.

Hypothetical scenario:

n observations, n linearly independent covariates.

What would you get for R^2?

As you add predictors to the model, R^2 will always increase, no matter what those predictors are!

"Goodness of Fit" Measures

R provides the following measures:

$$R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST}.$$

$$R_a^2 = 1 - \frac{SSE/(n-p-1)}{SST/(n-1)} = 1 - \frac{n-1}{n-p-1}(1-R^2)$$

1. R^2 is easy to interpret. It is the proportion of the "variation" in the data explained by the model.
2. R^2 does not adjust for the model size, while R_a^2 does. When comparing models of different sizes, use R_a^2.
3. However, for hypothesis testing the F statistic should be used.

F-tests for R^2

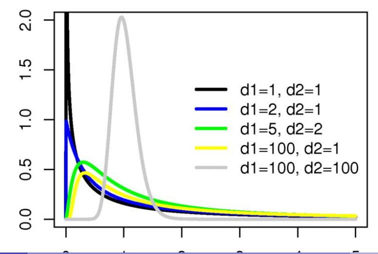
Assume model has intercept (design matrix has p columns).

$$F = \frac{SSR/(p+1)}{SSE/(n-p-1)}$$

F-distribution

If $W \sim \chi^2_q$ is independent of $Z \sim \chi^2_r$, then

$$\frac{W/q}{Z/r} \sim F_{q,r}.$$



F-Table

Source	Sum of Squares	d.f.	Mean Square	F
Regression	SSR	$p+1$	$MSR = \frac{SSR}{p+1}$	$F = \frac{MSR}{MSE}$
Residuals	SSE	$n-p-1$	$MSE = \frac{SSE}{n-p-1}$	

Reject at level α if $F > F(p+1, n-p-1, \alpha)$.

This tests the hypothesis $H_0 : \beta_1 = \beta_2 = \cdots = \beta_p = 0$.

Nested models

Test the hypothesis that a *subset* of β_i's are zero:

$$H_0 : \beta_1 = \beta_2 = \cdots = \beta_r = 0.$$

That is, we have the model

$$RM : \quad Y = \beta_{r+1} X_{r+1} + \cdots + \beta_p X_p + \text{error}$$

nested within

$$FM : \quad Y = \beta_1 X_1 + \cdots + \beta_p X_p + \text{error}$$

.

Does $X_1, \ldots, X_r$ have a significant marginal effect, after adjusting for the other predictors?

Nested models

$$RM: \quad Y = \beta_{r+1}X_{r+1} + \cdots + \beta_p X_p + \text{error}$$

$$FM: \quad Y = \beta_1 X_1 + \cdots + \beta_p X_p + \text{error}$$

-

$$\Delta df = df(FM) - df(RM)$$

$$F = \frac{[SSE(RM) - SSE(FM)]/[\Delta df]}{SSE(FM)/[n - df(FM)]}.$$

$$F \sim F_{\Delta df, n-df(FM)}.$$

Testing Constraints

In some situations you want to test that your model parameters satisfy some constraint. Say you have model:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \text{error},$$

and want to test:

$$H_0 : \beta_1 = \beta_2.$$

This is equivalent to the model:

$$Y = \beta_0 + \beta_1 (X_1 + X_2) + \text{error}.$$

.

Fit these two models, apply F test with $\Delta df = 1$.

Testing Constraints – Another example

Another example:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \text{error},$$

and want to test:

$$H_0 : \beta_1 + \beta_2 = 1.$$

This is equivalent to the model:

$$Y = \beta_0 + \beta_1 X_1 + (1 - \beta_1) X_2 + \text{error}.$$

.

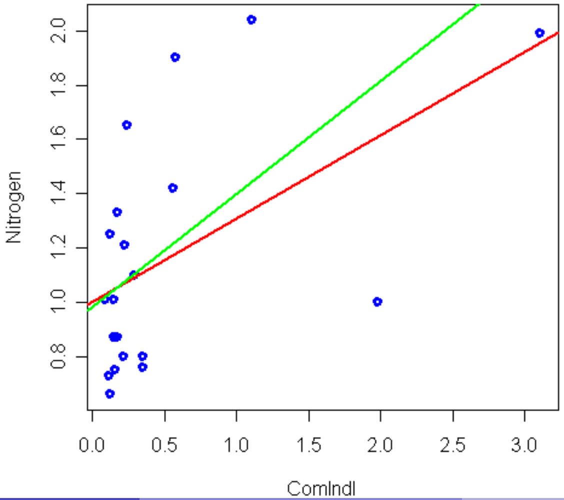
which can be simplified to:

$$Y - X_2 = \beta_0 + \beta_1 (X_1 - X_2) + \text{error}.$$

Fit these two models, apply F test with $\Delta df = 1$.
Usually Δdf equals the number of constraints.

Influence of Outliers

How much influence does the data point have on the model fit?



Different measures of Influence

1. How much influence does observation *i* have on its own fit?

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

   DFFITS exceeding $2\sqrt{p/n}$ is considered large.

2. How much influence does observation *i* have on the fitted β's?

$$(DFBETAS)_i = \frac{\hat{\beta}_1 - \hat{\beta}_{1(i)}}{\sqrt{MSE_{(i)} c_{11}^{-1}}},$$

   where $c_{11} = \sum_i (x_i - \bar{x})^2$. *DFBETA* exceeding $2/\sqrt{n}$ is considered large.

These conventional rules work for "reasonably sized" data sets.

Different measures of Influence: Cook's Distance

1. Cook's distance is defined as:

$$D_i = \frac{\sum_{j=1}^{n} (\hat{y}_j - \hat{y}_{j(i)})^2}{(p+1)MSE}$$

2. Considers the influence of Y_i on all of the fitted values, not just the i-th case.
3. It can be shown that D_i is equivalent to

$$\frac{\tilde{r}_i^2}{p+1} \frac{h_{ii}}{1-h_{ii}}.$$

4. Compare D_i to the $F_{p+1,n-p-1}$ distribution.