Regression Analysis

A regression model is a linear model

$$Y = X\beta + \epsilon$$

where $X_{n\times p}$ is usually of full rank p. The columns of X consist of all categorical variables, then the model is NOT a regression model in the usual sense. Thus, a regression model needs al least one continuous variable. The variables that make up the columns of X are typically called covariates, predictors, regressors, explanatory variables or independent variables. We will assume throughout our discussion of regression models that X has full rank p so that X^TX is nonsingular.

When X has full rank p, β is estimable, and every linear of β is estimable. To see that β is estimable, note that we can write

$$\beta = P^TX\beta$$

where

$$P^T = (X^TX)^{-1}X^T$$

Simple Linear Regression

The simple linear regression model including the intercept can be writtes as

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i, \quad i = 1, \ldots, n,$$

where

$$X = \begin{pmatrix} 1 & X_1 \\ 1 & X_2 \\ \vdots & \vdots \\ 1 & X_n \end{pmatrix} : n \times 2$$

We see here that rank$(X) = 2$ if the X_i's are not all the same. We know from previous results that the BLUE of β is

$$\hat{\beta} = \begin{pmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \end{pmatrix} = (X^TX)^{-1}X^TY.$$

We have

$$X^TX = \begin{pmatrix} n & n\bar{X} \\ n\bar{X} & \sum_{i=1}^n X_i^2 \end{pmatrix}$$

where $\bar{X} = \frac{1}{n}\sum_{i=1}^n X_i$. It is easily see that

$$(X^TX)^{-1} = \frac{1}{\sum_{i=1}^n (X_i - \bar{X})^2} \begin{pmatrix} \sum_{i=1}^n X_i^2 & -n\bar{X} \\ -n\bar{X} & n \end{pmatrix}.$$

Note also that

$$X^TY = \begin{pmatrix} n\bar{Y} \\ \sum_{i=1}^n X_iY_i \end{pmatrix}.$$

Thus

$$\hat{\beta} = \begin{pmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \end{pmatrix} = \begin{pmatrix} \bar{Y} - \bar{X}\left(\frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^n (X_i - \bar{X})^2}\right) \\ \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^n (X_i - \bar{X})^2} \end{pmatrix} = \begin{pmatrix} \bar{Y} - \hat{\beta}_1 \bar{X} \\ \hat{\beta}_1 \end{pmatrix}.$$

where

$$\hat{\beta}_1 = \frac{\widehat{\mathrm{Cov}(X, Y)}}{\widehat{\mathrm{Var}}(X)} \quad \text{and} \quad \hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}.$$

Exercise: Explain some relation(s) between $\hat{\beta}_1$ and $r = \widehat{\mathrm{Cor}}(X, Y)$ as much as you can.

By doing the matrix multiplication, it is easily shown that the orthogonal projection operator $\mathcal{C}(X)$ has ij-the element given by

$$m_{ij} = \frac{1}{n} + \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^n (X_i - \bar{X})^2}.$$

Also,

$$\mathrm{Cov}(\hat{\beta}) = \sigma^2 (X^T X)^{-1}.$$

Confidence intervals and tests of hypotheses are carried out by assuming $\epsilon \sim N_n(0, \sigma^2 I)$.

Multiple Linear Regression

In matrix form, we can write the multiple regression model as

$$Y = X\beta + \epsilon.$$

If X is of full rank, the BLUE of β is

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

and

$$\mathrm{Cov}(\hat{\beta}) = \sigma^2 (X^T X)^{-1}.$$

In multiple linear regression, we can decompose the total sums of squares as

$$\begin{aligned} Y^T Y &= Y^T M Y + Y^T (I - M) Y \\ &= \mathrm{SSR}(X) + \mathrm{SSE}. \end{aligned}$$

The column space that makes up the regression sums of squares is $\mathcal{C}(X) = \mathcal{C}(M)$. Thus

$$R^n = \mathcal{C}(X) + \mathcal{C}(X)^\perp = \mathcal{C}(M) + \mathcal{C}(I - M)$$

Since $r(M) = p$, $\mathcal{C}(M)$ can be decomposed into a sum of p orthogonal subspaces, each of deimension 1. Thus

$$\mathcal{C}(M) = \mathcal{C}(M_1) + \mathcal{C}(M_2) + \ldots + \mathcal{C}(M_p)$$

where $r(M_i) = 1, M_i M_j = 0, i \neq j$ and M_i is an orthogonal projection operator. Let X_i denote the i-th column of $X, i = 1, \ldots, p$ so that $X = (X_1, \ldots, X_p)$. Define

$$Z_i = (X_1, \ldots, X_i), i = 1, \ldots, p$$

and define $Z_0 = 0$. Let

$$P_{Z_i} = Z_i(Z_i^T Z_i)^{-1} Z_i$$

denote the orthogonal projection operator onto $\mathcal{C}(Z_i), i = 0, 1, \ldots, p$. Now let

$$M_i = P_{Z_i} - P_{Z_{i-1}}, i = 1, \ldots, p.$$

It is easily seen that each M_i is is an orthogonal projection operator, $r(M_i) = 1, M_i M_j = 0, i \neq j$. Thus

$$\begin{aligned} M &= M_1 + \ldots + M_p \\ &= P_{Z_1} + (P_{Z_2} - P_{Z_1}) + (P_{Z_3} - P_{Z_2}) + \ldots + (P_{Z_p} - P_{Z_{p-1}}) = P_{Z_p} \end{aligned}$$

and

$$Y^T M Y = Y^T M_1 Y + \ldots + Y^T M_p Y$$

Now let's denote

$$\text{SSR}(X_i | X_1, \ldots, X_{i-1}) = Y^T M_i Y, i = 1, \ldots, p$$

Thus $\text{SSR}(X_i | X_1, \ldots, X_{i-1})$ can be interpreted as the sum of squares due to adding X_i given that $X_1, \ldots, X_{i-1}$ are already in the model. Thus we can rite

$$\text{SSR}(X) = \text{SSR}(X_1) + \text{SSR}(X_2 | X_1) + \ldots + \text{SSR}(X_p | X_1, \ldots, X_{p-1}).$$

This orthogonal decomposition of $\mathcal{C}(X)$ depends on the order of the variables being fit into the model. Another ordering of the variables gives a different orthogonl breakdown. Thus the orthogonal decomposition of $\mathcal{C}(X)$ is not unique. This is not surprising since we already encountered this when we discussed unbalanced ANOVA. The orthogonal decomposition of $\mathcal{C}(X)$, however, is unque if $\mathcal{C}(X_i) \perp \mathcal{C}(X_j), i \neq j$.

We can construct the ANOVA table for the multiple regression model as

Source	df	SS	MS
Regression	p	$Y^T M Y$	$\frac{Y^T M Y}{p}$
Error	$n-p$	$Y^T(I-M)Y$	$\frac{Y^T(I-M)Y}{n-p}$
Total	n	$Y^T Y$	

Breaking $Y^T M Y$ into a sum of p independent quadratic forms yields

Source	df	SS	MS
X_1	1	$Y^T M_1 Y$	$Y^T M_1 Y$
$X_2 \mid X_1$	1	$Y^T M_2 Y$	$Y^T M_2 Y$
$\vdots$	$\vdots$	$\vdots$	$\vdots$
$X_p \mid X_1, \ldots, X_{p-1}$	1	$Y^T M_p Y$	$Y^T M_p Y$
Error	$n-p$	$Y^T(I-M)Y$	$\frac{Y^T(I-M)Y}{n-p}$
Total	n	$Y^T Y$	

If an intercept is included in the model, then we can let $J_n = X_1$, and $M_1 = \frac{1}{n} J_n J_n^T$.

To carry out tests of hypotheses and construct confidence regions, we assume $\epsilon \sim N_n(0, \sigma^2 I)$, and carry out the procedures as we did for the general linear model.

Best Linear Prediction

One of the main goals and uses of regression models is for prediction. We often wnat to construct models so that we can use the model to predict future observations. One can argue that the main goal in any statistical analysis is to predict and ,and inference about observations quantities, and that parameters should not be the main focus in inference. That is, parameters should be viewed as a tool for making predictors and developing models, but they are not an end in themselves. There is a lot of philosophical discussion on estimation versus prediction in many text book and articles.

We can view the regression problem as a prediction issue. That is, regression can be considered as the problem of predicting Y on the basis of $X_1, \ldots, X_p$. Let $X = (X_1, \ldots, X_p)^T$ be a $p \times 1$ vector, and Y is a scalar. Further, we assume that Y and X are random. A reasonable criterion for choosing a predictor for Y is to pick a predictor $f(X)$ that minimizes the mean square error(MSE), thst is,

$$\min_{f(X)} \mathbf{E}(Y - f(X))^2$$

Here, the expectation is taken with respect to the joint distribution (X, Y).

Theorem 6.3.1 Let $m(x) \equiv \mathbf{E}(y|x)$. Then for any other predictor $f(x)$,

$$\mathbf{E}[y - m(x)]^2 \leq \mathbf{E}[y - f(x)]^2$$

In other words, $m(x) = \mathbf{E}(y|x)$ is the best predictor of y.

For example, let

$$\begin{pmatrix} Y \\ X \end{pmatrix} \sim N_{p+1}(\mu, \Sigma) = N_{p+1}\left(\begin{pmatrix} \mu_Y \\ \mu_X \end{pmatrix}, \begin{pmatrix} \sigma_Y^2 & \Sigma_{YX} \\ \Sigma_{XY} & \Sigma_{XX} \end{pmatrix}\right).$$

Then

$$\mathbf{E}(Y|X) = (\mu_Y - \Sigma_{YX}\Sigma_{XX}^{-1}\mu_X) + \Sigma_{YX}\Sigma_{XX}^{-1}X$$

Now suppose the joint distribution of (Y, X) is not known,

$$\begin{pmatrix} Y \\ X \end{pmatrix} \sim \left(\begin{pmatrix} \mu_Y \\ \mu_X \end{pmatrix}, \begin{pmatrix} \sigma_Y^2 & \Sigma_{YX} \\ \Sigma_{XY} & \Sigma_{XX} \end{pmatrix}\right).$$

but means, variances, and covariances of (Y, X) (or their estimates) are all known. In this case, we can find the best *linear* predictor oy Y. We seek a linear predictor of the form

$$\alpha + X^T\beta$$

that minimizes

$$\mathbf{E}(Y - \alpha - X^T\beta)^2$$

for all scalars α and $p \times 1$ vector β.

Proposition: Let β_* be a solution to

$$\Sigma_{XX}\beta = \Sigma_{XY}.$$

Then

$$\mu_Y + (X - \mu_X)^T\beta_*$$

is the best liner predictor.

Proof: exercise

If (Y, X) have a multivariate normal distribution, the best predictor is linear, so that it is also the best linear predictor. Thus, if (Y, X) are multivariate normal, then

Best predictor=Best Linear predictor=conditional Expectation of $Y|X$.

Now we apply these results to the multiple linear regression model. Let $X_{n\times p} = (X_1, \ldots, X_p)$ be the matrix of covariates based on n observations, where X_i is an $n \times 1$ vector of i-th predictor. Also, let Y be the $n \times 1$ vector of responses. Then

$$S_{XX} = \frac{X^T \left(I - \frac{1}{n} J_n^n\right) X}{n-1}$$

and

$$S_{XY} = \frac{X^T \left(I - \frac{1}{n} J_n^n\right) Y}{n-1}.$$

Thus S_{XX} is the sample covariance matrix of the covariates, and S_{XY} is the vector of sample covariance between (X, Y).

Next,

$$\hat{\mu}_X = \bar{X} = \frac{1}{n} X = \left(\frac{1}{n} \sum_{i=1}^{n} X_{ij}\right) : p \times 1,\ j = 1, \ldots, p$$

and

$$\hat{\mu}_Y = \bar{Y} = \frac{1}{n} X = \frac{1}{n} \sum_{i=1}^{n} Y_i.$$

For multiple linear regression, the best linear predictor of Y_i is

$$\hat{Y}_i = \bar{Y} + (X_i - \bar{X})^T \hat{\beta}_*$$

where $\hat{\beta}_*$ is a solution to

$$X^T \left(I - \frac{1}{n} J_n^n\right) X \hat{\beta}_* = X^T \left(I - \frac{1}{n} J_n^n\right) Y.$$

If $X^T \left(I - \frac{1}{n} J_n^n\right) X$ is nonsingular, then

$$\hat{\beta}_* = \left[X^T \left(I - \frac{1}{n} J_n^n\right) X\right]^{-1} X^T \left(I - \frac{1}{n} J_n^n\right) Y.$$