Stat 242/542 Supplementary Notes on Regression 4-19-02

Andrew Barron, Yale University

1 Introduction

In these notes I provide some perspectives on regression. Much of what is said here is in the Chapter 14 of the text by Rice. However, I provide some different points of emphasis, as well as some supplementary information, especially with regard to model selection by predicted squared error criteria. I advocate the orthogonalization method for understanding regression here, as it avoids lots of fancy matrix manipulations, and it is closely tied to the geometry. Nevertheless, as Rice shows, if you like working with computations of inverses and the products of many matrices and their traces, you can get to the same conclusions.

These notes are written with three purposes. To the students to whom I have been lecturing in Stat 242/542 to provide explicit details for some of the concepts mentioned more loosely in the lectures, especially in matters that are not covered much in Rice, in response to requests from students. To Robb Muirhead and the Pfizer students (Robb knows more than I about most of these matters – he has written a book on multivariate analysis), so that they in the Pfizer section of the class can see some of the perspective I have tried to communicate. To myself, to organize my thinking on some of what I know about regression.

Some thoughts on what is expected to be remembered and understood are given in a final section.

2 Simple Projection

Let X consist of n values which have been set for an explanatory variable (arranged in a column vector if you like), and let Y denote the corresponding observations of a response variable. We think of X and Y as points in n dimensional Euclidean space. In regression on to a single variable (taking first, for simplicity the case of no intercept), we consider multiples $X\beta$ (which are on the line spanned by the vector X) and seek that choice of $\hat{\beta}$ such that the point $\hat{Y} = X\hat{\beta}$ is closest to Y, in the sense of minimizing the squared norm of the error (sum of squared errors) $||Y - X\beta||^2 = \sum_{i=1}^n (Y_i - x_i\beta)^2$. Then geometrical considerations (or a little calculus) lead to the suggestion to set $\hat{\beta}$ in such a way that the error $Y - X\hat{\beta}$ is orthogonal to X, that is, the sum of products $\sum_{i=1}^n (Y_i - x_i\hat{\beta})x_i$ is equal to zero. The $\hat{\beta}$ that satisfies this so-called normal equation is $\hat{\beta} = \sum_{i=1}^n x_i Y_i / \sum_{i=1}^n x_i^2$, the sum of products of x_i and Y_i divided by the sum of squares of x_i. [The sum of products $\sum_{i=1}^n x_i Y_i$ is called the inner product between the vectors X and Y]. The best way to confirm that this suggested solution does indeed produce the minimum sum of squared errors (as well as to visualize the geometry involved) is to note that our choice makes the cross term $2\sum_{i=1}^n (Y_i - x_i\hat{\beta})(x_i\beta - x_i\hat{\beta})$ in an expansion of the sum of squared errors vanish, such that we have the following Pythagorean identity for the squared lengths (associated with the right triangle that the error vector $Y - X\hat{\beta}$ makes with $X\beta - X\hat{\beta}$), for all β,

$$||Y - X\beta||^2 = ||Y - X\hat{\beta}||^2 + ||X\beta - X\hat{\beta}||^2.$$

Hence any $X\beta$ would produce a larger squared norm of error than $X\hat{\beta}$. The chosen $\hat{Y} = X\hat{\beta}$ is thus the unique projection of Y onto the line spanned by X.

Suppose it happens that the observations $Y_i = x_i\beta + e_i$ are random with mean $x_i\beta$ and variance σ^2, with errors uncorrelated for $i = 1, 2, \ldots, n$ (a standard linear statistical model with one explanatory variable). One finds then that the mean (expected value) of $\hat{\beta}$ is equal to β (the estimator is unbiased), and moreover, noting that the coefficient $\hat{\beta} = \sum_{i=1}^n a_i Y_i$ is a linear combination of the uncorrelated random variables Y_i, where $a_i = x_i / \sum_{i=1}^n x_i^2$, we have that the variance $\sigma_{\hat{\beta}}^2 = VAR(\hat{\beta})$ is the sum of the variances $\sigma_{\hat{\beta}}^2 = \sum_{i=1}^n a_i^2 \sigma^2$, which simplifies to $\sigma_{\hat{\beta}}^2 = \sigma^2 / \sum_{i=1}^n x_i^2$.

3 Projection onto Orthogonal Variables

Let $\tilde{X}_1, \tilde{X}_2, \ldots, \tilde{X}_k$ be k explanatory variables, n observations for each (which we may think of as the column vectors of an n by k design matrix $\tilde{X}$), which are used to fit a vector Y of observations of a response variable, by taking linear combinations of these explanatory variables, of the form $\tilde{X}_1\tilde{\beta}_1 + \ldots + \tilde{X}_k\tilde{\beta}_k$. Here we take the simplifying assumption that these expanatory variables are orthogonal to each other, that is, the inner products between any two of them is zero. [The tilde on the $\tilde{X}\tilde{\beta}$ is used to make a distinction with its counterpart $X\beta$ expressed in terms of variables that might not be orthogonal, as explained further below]. By inspection of suitable vanishing cross product terms, one finds that the task of minimizing the sum of squared errors in the fit by linear combination (of orthogonal variables) reduces to one-dimensional projections onto the k variables separately. Thus the least squares fit uses $\hat{\tilde{\beta}}_j = \sum_{i=1}^n \tilde{x}_{j,i} Y_i / \sum_{i=1}^n \tilde{x}_{j,i}^2$, for $j = 1, 2, \ldots, k$.

In a linear statistical model with orthogonal explanatory variables, we have $Y = \tilde{X}_1\tilde{\beta}_1 + \ldots + \tilde{X}_k\tilde{\beta}_k + e$ where the errors e_i are assumed to be uncorrelated each of mean zero and variance σ^2. This model is used to deduce properties of the distributions of the $\hat{\tilde{\beta}}_j$. Now this fitted coefficient $\hat{\tilde{\beta}}_j$ is equal, as we have noted, to the inner product between the $\tilde{X}_j$ and Y divided by a sum of squares. Using the assumed model, we find, plugging in the terms of Y, that the orthogonality wipes out most of these, so that $\hat{\tilde{\beta}}_j = \tilde{\beta}_j + \sum_{i=1}^n \tilde{x}_{j,i} e_i / \sum_{i=1}^n \tilde{x}_{j,i}^2$. From this we find again that the least squares estimator of the coefficients is unbiased and that the coordinates of $\hat{\tilde{\beta}}$ are uncorrelated (i.e., their covariance matrix is diagonal), with each coordinate having variance (placed in diagonal entries of the covariance matrix) that are inversely proportional to the squared norm of the corresponding explanatory variable, in the same manner as for a single explanatory variable.

4 Simple Linear Regression with Intercept

The fitting of regression lines with an intercept plus an explanatory variable x is the important problem extensively studied in Rice section 14.2. Here the plot of the responses Y_i versus the inputs x_i, superimposed with the best fitting line, as an ordinary plot in 2-variables, is of even greater importance that the picture of projection we were envisioning extracted from Euclidean $n-$space (expanded upon further below). The text by Rice focuses on the model expressed in the form

$$Y_i = \beta_0 + x_i\beta_1 + e_i.$$

For this model somewhat complicated looking formulas for the least squares values of the intercept $\hat{\beta}_0$ and the slope $\hat{\beta}_1$ are given in Rice. The better way to understand what is going on is a follows. Rather than the intercept and slope formula for the line, consider expressing the line in the point and slope formula, such that our model becomes

$$Y_i = \tilde{\beta}_0 + (x_i - \bar{x})\beta_1 + e_i,$$

where $\tilde{\beta}_0$ has the interpretation as the true mean value of the response when the explanatory variable x is set to its sample mean $\bar{x}$. A key advantage of this representation is that the regressors are now orthogonal vectors $\tilde{X}_0$ and $\tilde{X}_1$ which have coordinates equal to $\tilde{x}_{0,i} = 1$ and $\tilde{x}_{1,i} = (x_i - \bar{x})$, respectively. The orthogonality follows from noting that $\sum_{i=1}^n (x_i - \bar{x}) = 0$. Now, prepared as we are to deal with this situation, we can immediately deduce that the least squares estimators are $\hat{\tilde{\beta}}_0 = \sum_{i=1}^n 1Y_i / \sum_{i=1}^n 1$, which simplifies to

$$\hat{\tilde{\beta}}_0 = \bar{Y},$$

(which means that the regression line has height $\bar{Y}$ at the input $\bar{x}$) and, again using the rule for regressing onto a variable orthogonal to the others, we have that

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n (x_i - \bar{x})Y_i}{\sum_{i=1}^n (x_i - \bar{x})^2}.$$

See also problem 8 on page 554 of Rice. As pointed out in section 14.2.3 by Rice, we may also connect this estimated slope with the correlation coefficient,

$$r = \frac{1}{n-1}\sum_{i=1}^n \frac{(x_i - \bar{x})}{s_x}\frac{(y_i - \bar{y})}{s_y},$$

where we take $s_x^2 = \frac{1}{n-1}\sum_{i=1}^n (x_i - \bar{x})^2$ and, at risk of misinterpretation, we take $s_y^2 = \frac{1}{n-1}\sum_{i=1}^n (Y_i - \bar{Y})^2$. It is seen that

$$\hat{\beta}_1 = \frac{s_y}{s_x} r.$$

The simplest explanation for the role of correlation coefficient in understanding regression is as follows. Suppose first that we "standardize" both the x and the y, by subtraction by $\bar{x}$ and $\bar{Y}$, respectively, and then division by s_x and s_y, respectively, forming $x^{stand} = (x - \bar{x})/s_x$ and $Y^{stand} = (Y - \bar{Y})/s_y$. Now we are looking at the standardized residuals of the x and the Y after having regressed out the best constant from each. Then we do the regression of the standardized responses onto the standardized explanatory variable. The result of this least squares regression has *slope equal to the correlation coefficient*, that is, the least squares fit is

$$\hat{y}^{stand} = r\hat{x}^{stand},$$

as can be seen by verifying that $r = \sum_{i=1}^{n} x_i^{stand} Y_i^{stand} / \sum (x_i^{stand})^2$. Undoing the standardization leads back to the ordinary least squares fit as follows:

$$\frac{\hat{y} - \bar{y}}{s_y} = r\frac{x - \bar{x}}{s_x}$$

and hence

$$\hat{y} = \bar{y} + (\frac{s_y}{s_x}r)(x - \bar{x}),$$

so that the slope is

$$\hat{\beta} = r\frac{s_y}{s_x}$$

as noted above.

In the simple linear statistical model, with uncorrelated errors of mean zero and variance σ^2, we can use the theory given above (for uncorrelated inputs) to conclude that the estimated slope $\hat{\beta}_1$ has mean equal to the true slope β_1 and variance

$$\sigma^2_{\hat{\beta}} = \frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}.$$

This simple expression does agree with the ugly equivalent expression at the top of page 514, but in a form more suitable for understanding and recollection.

In a customary fashion, the reexpression of the simple linear regression with intercept in to an orthogonal linear regression onto 1 and $x - \bar{x}$ follows a pattern that is be applied successively to handle the general multiple linear regression, as shall be explained in the next section.

5 Multiple Linear Regression

Let X be an n by k design matrix and Y an n by 1 response vector. These represent n observations (or cases) of k explanatory variables (the columns of X) used to model the mean of the response as a linear combination of the explanatory variables. We suppose that the explanatory variables are linearly independent, that is, none of these variables (columns) are exact linear combinations of the others. The model is $Y = X\beta + e$, where the coordinates of e are assumed to be uncorrelated with mean zero and variance σ^2.

Geometrically, we may think of Y and the columns of X as points (in Euclidean space of dimension n). If we vary the coefficients β of linear combination, the resulting points $X\beta$ span a plane $\mathcal{S}$ (also called a linear subspace or a hyperplane) of dimension k. Recalling a little linear algebra, we say that the columns of X form a basis for this linear space. If it is found to be convenient one may form other design matrices $\tilde{X}$, whose columns are linear combinations of the columns of X, to produce other bases that span the same space $\mathcal{S}$.

The least squares criterion seeks that point $\hat{Y}$ in this plane that is closest to Y (this point is also called the projection of Y onto the plane, and it is called the vector of fitted values). The text by Rice presents two perspectives on this projection. First since $\hat{Y}$ is in the plane it can be written in the form $X\hat{\beta}$. Then by calculus as on page 530-531, Rice deduces the so-called normal equations $X^T X\hat{\beta} = X^T Y$. This identity may also be written as $X^T(Y - \hat{Y}) = 0$, that is, it is the requirement that the error $Y - \hat{Y}$ of the projection be orthogonal to the columns of X and hence orthogonal to the plane they span. Indeed this orthogonality permits a Pythagorean identity which also reveals that the point $\hat{Y}$ that satisfies this orthogonality condition is the unique least squares projection. Rice then points out that the solution of this linear system of equations for $\hat{\beta}$ may be written as $\hat{\beta} = (X^T X)^{-1} X^T Y$ and hence the fitted (or projected) values are obtained as $\hat{Y} = X\hat{\beta}$. Substituting the expression for $\hat{\beta}$ gives $\hat{Y} = [X(X^T X)^{-1} X^T]Y$ so the projection gives the fitted values as linear combinations of the observed responses as expressed by $\hat{Y} = PY$ where P is the n by n projection matrix $P = X(X^T X)^{-1} X^T$.

The formula for the projection looks complicated, but as Rice points out the projection P is not usually computed in that way. To explain more, we note that there is considerable freedom in how this projection is computed. Indeed for any design matrix $\tilde{X}$ whose columns are linearly independent and span the same space as the columns of X, the result of projection will be the same and hence P is also equal to $\tilde{X}(\tilde{X}^T\tilde{X})^{-1}\tilde{X}^T$. In particular, if the design matrix is chosen to have orthogonal columns, (spanning the same space as the columns of X), each normalized to have norm 1, then coefficient estimation reduces to $\hat{\tilde{\beta}} = \tilde{X}^T Y$ (each coordinate of $\hat{\tilde{\beta}}$ is simply the inner product of Y with the corresponding column of $\tilde{X}$) and projection reduces to $\hat{Y} = \tilde{X}\hat{\tilde{\beta}} = [\tilde{X}\tilde{X}^T]Y$. The matrix $\tilde{X}^T\tilde{X}$ (and its inverse) have in this case been reduced to the identity, by orthonormality.

The way statistical packages actually compute the projections coincides with a particular choice of such $\tilde{X}$ with orthonormal columns constructed from the columns of X, in a way that is especially suitable for understanding how projections work. Specifically, for the first column $\tilde{X}_1$ we take X_1/c_1 where the constant c_1 denotes normalization to make the result have norm 1. For the second column $\tilde{X}_2$ we take the residual of the second variable X_2 in which we have subtracted its fit $r_{2,1}\tilde{X}_1$ using the first variable and then the residual is normalized, that is $\tilde{X}_2 = (X_2 - r_{1,2}\tilde{X}_1)/c_2$. We continue in this way such that for the kth column of $\tilde{X}$ we take the normalized residual of X_k in which we have subtracted its fit using the previous orthonormal variables $\tilde{X}_1$ through $\tilde{X}_{k-1}$. Thus

$$\tilde{X}_k = \frac{X_k - r_{1,k}\tilde{X}_1 - r_{2,k}\tilde{X}_2 - \ldots - r_{k-1,k}\tilde{X}_{k-1}}{c_k}.$$

Where in accordance with the coefficients of projection onto orthonormal variables each $r_{j,k}$ equals the inner product (sum of products) of $\tilde{X}_j$ and X_k for j less than k. This produces the $\tilde{X}$ variables from the X variables. While doing this we can for each explanatory variable in succession regress (project) Y onto the part of that explanatory variable that is not explained by linear combination of the previous variables. The process just described is well known in linear algebra as Gram-Schmidt orthogonalization. The orthonormal matrix $\tilde{X}$ we have created is often denoted Q. Also by slight rearrangement of the above identity, and taking $r_{k,k} = c_k$ to be the normalizing constant, we find that $X_k = \sum_{j=1}^{k} \tilde{X}_j r_{j,k}$, so that the X columns can be recovered from the $\tilde{X}$ columns, by $X = \tilde{X}R$ where R is upper triangular ($r_{j,k} = 0$ for $j > k$). Finding in succession the $Q = \tilde{X}$ columns (together with the coefficients $r_{j,k}$ for relating X to $\tilde{X}$), this is known as the QR algorithm and the result is also called the QR decomposition of X. See also Rice pages 553-554, problem 6.

Since the projection is unique, mathematically we will produce the same answer for $\hat{Y}$ whether we use the original design matrix X or the associated orthonormal design matrix $\tilde{X}$ (though as explained in problem 6, there are numerical advantages to working with $Q = \tilde{X}$).

The coefficients $\hat{\beta}$ and $\hat{\tilde{\beta}}$ yielding $\hat{Y} = X\hat{\beta} = \tilde{X}\hat{\tilde{\beta}}$ can in general be quite different, but they are related. In particular, since $X = \tilde{X}R$ it follows that $R\hat{\beta} = \hat{\tilde{\beta}}$, and by back substitution for the upper triangular matrix R we can obtain $\hat{\beta} = R^{-1}\hat{\tilde{\beta}}$. Likewise, if one does want $(X^T X)^{-1} = (R^T R)^{-1}$, it can be obtained by two sweeps of back substitution.

Using the complicated matrix formula for $\hat{\beta} = (X^T X)^{-1} X^T Y$, Rice shows on page 538 that for the linear statistical model the covariance matrix for $\hat{\beta}$ is $COV(\hat{\beta}) = (X^T X)^{-1}\sigma^2$. If one desires it, this covariance can also be derived from the orthogonalization. Indeed, from what we have said for regression on orthogonal variables we find that the coordinates of $\hat{\tilde{\beta}}$ are uncorrelated random variables of mean zero and variance σ^2. Thus $COV(\hat{\tilde{\beta}}) = I\sigma^2$ where I is the k by k identity. Consequently, using $\hat{\beta} = R^{-1}\hat{\tilde{\beta}}$, the least squares coefficients $\hat{\beta}$ for the original design X have covariance matrix $COV(\hat{\beta}) = R^{-1}R^{-T}\sigma^2 = (R^T R)^{-1}\sigma^2 = (X^T X)^{-1}\sigma^2$ as stated.

The $\hat{\tilde{\beta}}$ has an advantage of statistical stability in that its covariance matrix is $I\sigma^2$, compared to $\hat{\beta}$ which can be worse statistically, especially if $(X^T X)$ is nearly singular. This problem arises if X is such that some of its columns, though linearly independent, are very nearly equal to a linear combination of other columns.

6 Decomposition of the sum of squared errors and estimation of σ^2

First, as suggested in Rice, page 539, Lemma A, we say more about projection $\hat{Y} = PY$, from general projection properties that are not dependent on the particular matrix expression. Projection gives the closest point in $\mathcal{S}$ to a given point. In particular, if we have a point that is in the set $\mathcal{S}$, then applying the projection should leave it untouched (the closest point to it in the set is itself). Thus $PX\beta = X\beta$ for all β, i.e. $PX = X$. Likewise, if we were to project twice it wouldn't change anything: $PPY = PY$. Since this holds for all Y we have $P^2 = P$. (Also note as in Rice, Lemma A, that P is symmetric, that is $P^T = P$). In like manner one finds that $(I - P)^2 = I - P$ where I is the n by n identity. This $(I - P)$ is the matrix that produces the vector of residuals $Y - \hat{Y} = (I - P)Y$. Multiplication by $(I - P)$ yields, for any point Y, its projection on the space of vectors orthogonal $\mathcal{S}$, that is, orthogonal to the columns of X (and hence to any linear combination thereof). Indeed, $(I-P)X=0$ (from $PX=X$) and $(I-P)P=0$ (from $P=P^2$) so the residual $Y-\hat{Y}$ is orthogonal to $\hat{Y}$ and orthogonal to $X(\hat{\beta}-\beta)$. So we have a zero cross product term establishing the following Pythagorean identity

$$||Y-X\beta||^2 = ||Y-\hat{Y}||^2 + ||X(\hat{\beta}-\beta)||^2.$$

In terms of $e=Y-X\beta$ this identity may also be written as $||e||^2 = ||(I-P)e||^2 + ||Pe||^2$. The Pythagorean identity both captures the least squares optimality of $\hat{Y}$ among all choices of $X\beta$ as we have discussed before and in a statistical setting we will discuss momentarily it provides the essence of an analysis of variance. But before we get ahead of ourselves first note that this Pythagorean identity is the same no matter which basis is used to represent the given space, that is, with $X\beta = \tilde{X}\tilde{\beta}$, we have

$$||Y-\tilde{X}\tilde{\beta}||^2 = ||Y-\hat{Y}||^2 + ||\tilde{X}(\hat{\tilde{\beta}}-\tilde{\beta})||^2.$$

Also note that if the columns of $\tilde{X}$ are orthonormal, then the last squared norm above reduces to $||\hat{\tilde{\beta}}-\tilde{\beta}||^2$, which is a sum of squares of k coordinates.

Suppose again that we have the linear statistical model in which $Y = X\beta + e$ or equivalently $Y = \tilde{X}\tilde{\beta} + e$ and the coordinates of e are assumed to be uncorrelated with mean 0 and variance σ^2. In particular $\mu = X\beta$ or equivalently $\tilde{X}\tilde{\beta}$ is our model for the mean of Y. In this case, the norm squared on the left side of the Pythagorean identity is the sum of squares of the n random variables e_i each of which has $E(e_i^2)=\sigma^2$. Hence $E||Y-X\beta||^2 = n\sigma^2$. If also the distribution of the e_i is Normal$(0,\sigma^2)$, then we see that $||Y-X\beta||^2$ (or equivalently $||Y-\tilde{X}\tilde{\beta}||^2$) is distributed as a Chi-square(n) random variable times σ^2. The last term in the Pythagorean identity $||\hat{\tilde{\beta}}-\tilde{\beta}||^2$ is the sum of squares of k coordinates which we have shown above to be uncorrelated random variables of mean zero and variance σ^2. Hence $E||\hat{\tilde{\beta}}-\tilde{\beta}||^2 = k\sigma^2$. Under the normal error assumption, the $||\hat{\tilde{\beta}}-\tilde{\beta}||^2$ becomes a sum of squares of k independent normals of variance σ^2 and hence is likewise distributed as a Chi-square(k) random variable times σ^2. Moreover, the two terms on the right side of the Pythagorean inequality are the norm squares of $(I-P)e$ and Pe so their covariance is $(I-P)E[ee^T]P^T = (I-P)P\sigma^2 = 0$, that is they are uncorrelated and hence, under the normal error assumption, they are independent (likewise we find that the residuals $Y-\hat{Y}$ are independent of the estimated coefficients $\hat{\tilde{\beta}} = \tilde{X}^T Y$ and hence independent of $\hat{\beta}$). These facts enable us to deduce properties of the sum of squared residuals $||Y-\hat{Y}||^2$. First that its expected value must be $(n-k)\sigma^2$. Secondly, in a manner analogous to the proof at the top of page 182, we have (under the normal error assumption) two independent random variables which are added, one of which is Chi-square(k), and the sum is a Chi-square(n), so the other summand must be Chi-square$(n-k)$. That is, the sum of squared residuals $||Y-\hat{Y}||^2$ is distributed as a Chi-square $(n-k)$ times σ^2, independent of the fitted values. Thus we are led to

$$s^2 = \frac{1}{n-k}||Y-\hat{Y}||^2$$

as an unbiased estimator of σ^2, and likewise we are led to an F distribution with k and $n-k$ degrees of freedom for the ratio $(||X\hat{\beta}-X\beta||^2/k)/s^2$.

We can use our estimator of σ^2 in obtaining estimators of other related quantities, such as the variances of the estimated coeffiecients. For each estimated $\hat{\beta}_j$ we have found that its mean is β_j, that its variance is $C_{jj}\sigma^2$ (where C_{ij} are the entries of the matrix $C = (X^T X)^{-1}$), and, in our model with normal errors, that it has a normal distribution independent of the residuals. Thus we use $s_{\hat{\beta}_j} = C_{jj}^{(1/2)} s$ to estimate the standard deviation of the coefficient. It is used to provide the standardized estimated coefficient

$$T_j = \frac{\hat{\beta}_j - \beta_j}{s_{\hat{\beta}_j}},$$

which as a ratio of a normal and the square root of an independent Chi-square$(n-k)$, it has the T distribution with $n-k$ degrees of freedom.

We conclude this section of our statistical discussion by noting properties of the distributions of the fitted values $\hat{Y}_i$ and the residuals $Y_i - \hat{Y}_i$. The mean of $\hat{Y} = PY$ (as we have seen) is $X\beta$ and its covariance matrix is the covariance matrix of $\hat{Y} - X\beta = Pe$ which is $PP^T\sigma^2$, which simplifies to $P\sigma^2$. Likewise $Y - \hat{Y} = (I-P)e$ has mean zero and covariance matrix $(I-P)\sigma^2$. Another way to understand these properties is to recall that $\hat{Y}_i = \tilde{x}_i^T \hat{\tilde{\beta}}$ where x_i^T is the row for observation i of the matrix $\tilde{X}$, so that, for instance, the variance for $\hat{Y}_i$ is equal to $||\tilde{x}_i||^2\sigma^2$. To see what is happening in terms of the matrix P, recall that $P = \tilde{X}\tilde{X}^T$. Thus, the entries of the projection matrix P are simply $P_{i,j} = \tilde{x}_i^T\tilde{x}_j$, the inner products of the i and j rows of $\tilde{X}$. In particular, the variance of $\hat{Y}_i$ is $P_{ii}\sigma^2 = ||\tilde{x}_i||^2\sigma^2$ as we have seen. [In terms of the original variables x_i, the expression would be more complicated]. These $P_{ii} = ||\tilde{x}_i||^2$ are each less than 1 since the variance of the residual $res_i = Y_i - \hat{Y}_i$ is $(1 - P_{ii})\sigma^2$. The standard deviation of the residual may be estimated as $s_{res_i} = (1 - P_{ii})^{1/2} s$ yielding standardized residuals $(Y_i - \hat{Y}_i)/s_{res_i}$ as also explained on page 541 of Rice.

An often noted fact about the projections P and $I-P$ are that they have traces equal to the dimensions k and $n-k$ of the spaces onto which they project. Rice shows this on page 540 using the complicated matrix expression for P together with a fact about commuting products of matrices within a trace, and used these traces to determine properties of the distributions of the sums of squares. Our analysis has bypassed directly finding these traces. [Indeed, when we found the expected value of $||Pe||^2/\sigma^2 = ||\tilde{X}(\hat{\tilde{\beta}} - \beta)||^2/\sigma^2 = ||\hat{\tilde{\beta}} - \beta||^2/\sigma^2$ we simply noted that it a sum of squares of k standardized random variables, so that its mean is equal to k]. For a direct determination of the trace of P using the orthonormalized basis, simply note that it is equal to $\sum_{i=1}^n P_{ii} = \sum_{i=1}^n ||\tilde{x}_i||^2$ where each of these norms is a sum of k terms (which each sum to one from $i = 1$ to n) so the trace is equal to k.

7 Mean squared error of prediction

Suppose we have k explanatory variables used to model the mean of a response variable by linear regression. Let β denote the k-dimensional vector of true coefficients in this model. A data set consisting of n observations of the variables and associated response Y_i (with uncorrelated errors of mean zero and variance σ^2) is used to estimate the coefficients, and is accordingly called the training (or fitting) data. The least squares estimates of the coefficients are denoted $\hat{\beta}$ and the estimated means for the observed (training) data are called the *fitted* values, $\hat{Y}_i = x_i^T\hat{\beta}$, for $i = 1, 2, \ldots, n$. Thereafter, the linear combination of explanatory variables is used to predict the response in new cases. Let Y_{new} denote the unknown new response equal to the sum of the mean $x_{new}^T\beta$ plus an error which is assumed to be uncorrelated with the errors of the responses in the training data, and also to have variance σ^2. Here the input x_{new} may be the same or different from input values we have seen so far. According to the model, if the true β were known, the best prediction of the response Y_{new} would be the true mean $x_{new}^T\beta$ which would have mean squared error of σ^2. Not knowing the true β and not knowing the new response value we use the previously estimated coefficients $\hat{\beta}$ from the training data to produce what is called the *predicted* value $x_{new}^T\hat{\beta}$. What will be the somewhat larger mean squared error in this case? That is, we want to evaluate the mean squared error of prediction $E(Y_{new} - x_{new}^T\hat{\beta})^2$ at the point x_{new}. Noting that both Y_{new} and $x_{new}^T\hat{\beta}$ have mean $x_{new}^T\beta$ we recognize that this mean squared error of prediction is the variance of the sum of two uncorrelated random variables and hence is equal to the sum of the variances $\sigma^2 + VAR(x_{new}^T\hat{\beta})$, which reduces to $\sigma^2(1 + x_{new}^T C x_{new})$ where $C = (X^TX)^{-1}$ (see also page 544 for a related calculation). One may also use the transformation R which related the X to an orthonormal $\tilde{X}$ (on the training data) to create (by back substitution) a corresponding $\tilde{x}_{new} = R^{(-1)}x_{new}$ for the new observation. Then our prediction may also be written as $x_{new}^T\hat{\beta} = \tilde{x}\hat{\tilde{\beta}}$ and this mean squared error of prediction at x_{new} also equals $\sigma^2(1 + ||\tilde{x}_{new}||^2)$. The estimated standard error of prediction at x_{new} is $s_{pred|x_{new}} = s(1 + ||\tilde{x}_{new}||^2)^{1/2}$. This calculation is what underlies the so-called prediction intervals as given in the simple regression case in problem 12 pages 554-555.

Now if one has a collection of points x_{new} at which one wants to do the prediction, it is sensible to consider what will be the average mean squared error at those points, and how it can be estimated. We find that if ρ is the average of the $||\tilde{x}_{new}||^2$ then the average mean squared error of prediction is $\sigma^2(1 + \rho)$. What might be a fair assessment of ρ? Well, perhaps the training data design is representative of the sorts of new points at which the regression will be evaluated. If not, we may want to rethink the original design. For the training data the average is $\rho = (1/n)\sum_{i=1}^n ||\tilde{x}_i||^2 = k/n$. (This sum is determined as discussed at the end of the section above, note also that it is the average of the $x_i^T C x_i = P_{ii}$ and thus is equal to trace$P$/n). Thus the average mean squared error of prediction is $\sigma^2(1 + k/n)$. An unbiased estimator of this average mean squared error of prediction, called the Final Prediction Squared Error [due to Akaike (1969)], is

$$FPSE_k = (1 + \frac{k}{n})s^2 = \frac{n+k}{n-k}\frac{||Y - \hat{Y}||^2}{n}.$$

It's square root may be called the Final Prediction Error,

$$FPE_k = s(1 + k/n)^{1/2}.$$

More generally, it can be shown that if a k by k matrix M_{new} is taken to be the average of the $x_{new}x_{new}^T$ at new points and $M_{train} = X^TX/n = (1/n)\sum_{i=1}^n x_i x_i^T$ is the corresponding average on the training data, then the conclusion holds where the $\rho = k/n$

should be replaced by $\rho = \text{trace}(M_{new} M_{train}^{-1})/n$ (as it is the average of the $||\tilde{x}_{new}||^2 = x_{new}^T (X^T X)^{-1} x_{new}$). Thus the final prediction error value is justified if either the new responses are to be independently observed either at the same design points at which we obtained the training data, or if the new design is similar to the old design in the sense that the "covariance" matrix M_{new} matches M_{train} in the sense that trace$(M_{new}(M_{train})^{-1})$ is the same as the trace of the k by k identity.

Akaike (1969) introduced his FPE as a criterion for model selection. The idea is that there may be some number k^* of our variables that contains the true mean in their span. If the model under consideration does not include these variables (or variables in close proximity thereto), then there will be a bias to its regression that should produce a large value for $||Y - \hat{Y}||^2/n$, and hence is not favored in the model selection. What about the many models (some of which could be very large) that include the unknown true k^* terms as a subset? For each of these models the statistic s^2 (with its division by $n-k$) produces an unbiased estimator of σ^2, so if we selected a model by minimizing s^2 we could still wind up with a very overfit model. So minimizing s^2 does not work well. Akaike argued in favor of selecting the model to provide the best estimated mean squared errors of prediction. That is, a model with $\hat{k}$ variables is chosen to minimize FPE_k among the collection of available models. Now, for all models that contain the correct model the mean squared error of prediction is $\sigma^2(1 + k/n)$, which gets worse as k increases past k^*. If we can estimate this mean squared error of prediction well (by the FPE), then Akaike's argument is that the criterion that picks the model that has the smallest FPE should provide what we want, namely, the smallest of the models considered that contain the correct model. Building on Akaike's developments, further theory by R. Shibata, K.-C. Li and others in the 1980s justified Akaike's reasoning in part. The selected models do produce close to the optimal mean squared errors of the fits to the true mean, provided not too many models of each dimension are considered, although there is some non-vanishing probability (as n gets large) that the selected models will stay larger than the minimal correct model (even if the minimal correct model is among those considered). The fits are consistent and have certain asymptotic optimalities, but the selected model size may be inconsistent. To force consistency of the selection of the subset or to allow many models of each dimension, other criteria use somewhat larger penalities for dimension in which the k is multiplied by a penalizing factor which is often of the order of $\log n$, at the expense of losing some of the mean squared error optimality properties of the Final Prediction Error.

The FPE is closely related to Mallows C_p, a criterion with similar motivation, but with $\sigma^2 + (k/n)\sigma^2$ estimated in a somewhat different way in two terms (one of which involves the model under current consideration, and the other involves a full model with all available explanatory variables included). Further extensions to his criterion to handle models with other error distributions were later developed by Akaike in the 1970's, with the name of An Information Criterion (AIC).

In the last homework set you have opportunity to select the model both by examination of FPE and by examination, for each coefficient $\hat{\beta}_j$, of the significance of the t statistic $T_j = \hat{\beta}_j / s_{\hat{\beta}_j}$ for testing the hypothesis that the associated β_j is zero (while the others need not be). Carefully implemented one can avoid some of the difficulities with multiple comparisons. The problem with the testing methods is that in usual practice the repeated nested checking of T-values can invalidate the meaning of the reported significance levels, because the hypotheses are not fixed, but rather adapted to the results of previous decisions about which other variables are included. Criterion minimization (FPE), especially when the objective is accurate prediction, provides a helpful alternative to the testing based methods of subset selection.

8 What to walk away with

For Stat 242/542 students, you are expected to know well the simple linear regression setting, both from the vantage point in Rice, and even more from the point-slope perspective given above, and its relationship to the correlation. As for multivariate projection and associated linear model theory, you have learned at least a little about three perspectives (the matrix inversion approach in Rice, the iterative projection method onto successively determined orthogonal components, and the general intuitive properties of projection matrices), and you should be able to state the basic idea of each of these perspectives. You should know at least one of these well enough to be able to follow the development of the form of the coefficient estimators, fitted values, Pythagorean identity, and the distribution of the estimated coefficients, the distribution of the sum of squared residuals, and the T statistics for tests of coefficients. You should be able to interprete the standard outputs of regression software (coefficients, standard errors, T-values, and P-values). You should know the interpretation of the Final Prediction Error and how to compute it from the outputs of your regressions. Finally, you should be comfortable in the interpretation of residual plots.