

# SOME GEOMETRIC INTUITION IN REGRESSION AND PCA

WILLIAM MCCLOSKEY

## 1. REGRESSION

1.1. **Overview.** Consider the linear regression model  $y = X\beta + \epsilon$ , where  $X \in \mathbb{R}^{m \times n}$   $m \gg n$  is the design matrix of  $m$  data points,  $\epsilon \sim \mathcal{N}(0, \sigma^2 I)$  is the noise vector, and  $y$  is the output vector. The goal is to estimate  $\beta$ .

The usual estimate  $\hat{\beta}$  of  $\beta$  is obtained by minimizing the mean-squared error

$$\hat{\beta} = \arg \min_{\beta} \|y - X\beta\|_2^2$$

In statistics, it is usually assumed that  $X$  is full rank, and we get the estimate

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

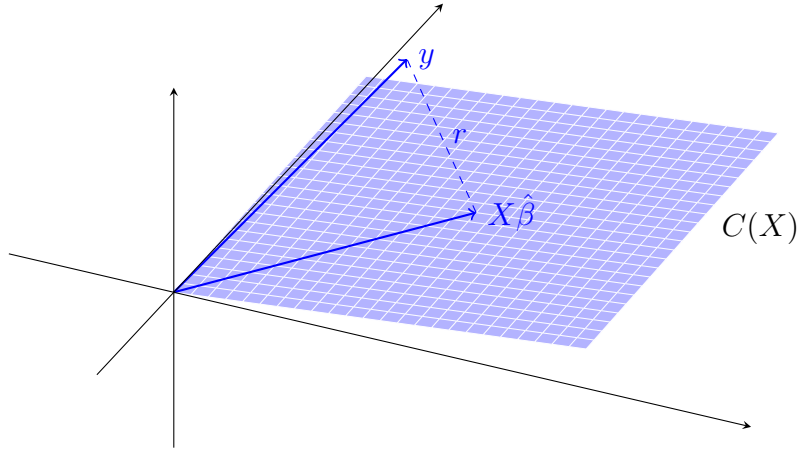
The output vector  $y \sim \mathcal{N}(X\beta, \sigma^2 I)$  is normally distributed, so  $\hat{\beta}$  is normally distributed as well with

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

Likewise the residual vector  $r = y - X\hat{\beta}$  has a normal distribution

$$r \sim \mathcal{N}(0, \sigma^2 (I - X(X^T X)^{-1} X^T)).$$

Geometrically  $X\hat{\beta}$  is the orthogonal projection of  $y$  onto the column space  $C(X)$  of  $X$ .



## 1.2. Distributional Results.

1.2.1. *Estimating  $\sigma^2$ .* The dimension of the space  $C(X)^\perp$  that contains the residual vector  $r$  is also called the degrees of freedom of  $r$ . This dimension is  $m - n$ .

Notice that  $r$  is normally distributed with  $r \sim \mathcal{N}(0, \sigma^2(I - X(X^T X)^{-1}X^T))$ . Since  $I - X(X^T X)^{-1}X^T$  is the projection matrix onto  $C(X)^\perp$ , it can be diagonalized in the form  $QI_{(m-n)}Q^T$  for some orthogonal matrix  $Q$ . Here  $I_{(m-n)}$  is a diagonal matrix with  $(m - n)$  ones and  $n$  zeros on the diagonal.

As a consequence, we see that  $Q^T r \sim \mathcal{N}(0, \sigma^2 I_{(m-n)})$ . Since orthogonal matrices are norm-preserving, we can calculate the distribution of the sum of squared residuals:

$$\begin{aligned} \sum_{i=1}^m r_i^2 &= \|r\|_2^2 \\ &= \|Q^T r\|_2^2 \\ &= \sigma^2 \sum_{i=1}^{m-n} z_i^2 \end{aligned}$$

where  $z_i \stackrel{iid}{\sim} \mathcal{N}(0, 1)$ . So we get the distributional result

$$\sum_{i=1}^m r_i^2 \sim \sigma^2 \chi_{(m-n)}^2.$$

As a result, the estimator  $\hat{\sigma}^2 = \frac{1}{m-n} \sum_{i=1}^m r_i^2$  has the distribution

$$\hat{\sigma}^2 \sim \sigma^2 \left( \frac{1}{m-n} \chi_{(m-n)}^2 \right)$$

with expectation  $\sigma^2$ .

1.2.2. *The  $F$ -test.* The vector  $y \sim \mathcal{N}(X\beta, \sigma^2 I)$  is multivariate normal. We can make use of the following property of multivariate normal variables.

**Property 1.** *Suppose  $y \in \mathbb{R}^m$  is a multivariate normal random variable with covariance matrix  $\sigma^2 I$  and  $U, V \subset \mathbb{R}^m$  are orthogonal subspaces. Then  $z_1 = \text{Proj}_U(y)$  and  $z_2 = \text{Proj}_V(y)$  are each (possibly degenerate) multivariate normal random variables. Moreover,  $z_1$  and  $z_2$  are independent.*

*Proof.* First consider the case where  $U, V$  are subspaces spanned by standard basis vectors. Then the result holds since  $y_1, \dots, y_m$  are independent random normal variables (because the covariance matrix of  $y$  is  $\sigma^2 I$ ). The general case reduces to the previous case by multiplying by an orthogonal matrix so that  $U$  and  $V$  are spanned by standard

basis vectors, and noting that the covariance matrix of  $y$  is unchanged by this transformation.  $\square$

We get the following distributional results.

**Property 2.** *Consider the setting of Property 1, and suppose additionally that  $\mathbb{E}(z_1) = \mathbb{E}(z_2) = 0$ . Then the following distributional results hold:*

$$\|z_1\|_2^2 \sim \sigma^2 \chi_{\dim(U)}^2$$

$$\|z_2\|_2^2 \sim \sigma^2 \chi_{\dim(V)}^2$$

In particular,

$$\frac{\|z_1\|_2^2 / \dim(U)}{\|z_2\|_2^2 / \dim(V)} \sim F_{\dim(U), \dim(V)}.$$

*Proof.* As in the proof of Property 1, let  $Q$  be an orthogonal matrix so that  $QU$  is spanned by the first  $\dim(U)$  standard basis vectors  $e_1, \dots, e_{\dim(U)}$ . Then by orthogonality

$$\|z_1\|_2^2 = \|Qz_1\|_2^2.$$

We additionally have the following equalities

$$\begin{aligned} Qz_1 &= Q\text{Proj}_U(y) \\ &= \text{Proj}_{QU}(Qy) \\ &= ((Qy)_1, \dots, (Qy)_{\dim(U)}, 0, \dots, 0) \end{aligned}$$

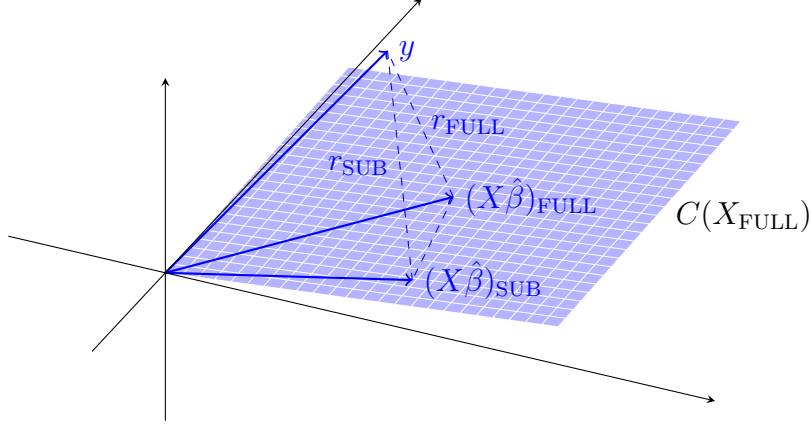
Since  $\mathbb{E}(z_1) = 0$  and  $Qy$  has covariance matrix  $\sigma^2 I$ , we obtain

$$\|z_1\|_2^2 = \|((Qy)_1, \dots, (Qy)_{\dim(U)})\|_2^2 \sim \sigma^2 \chi_{\dim(U)}^2.$$

The rest of the result follows by applying the definition of the  $F$  distribution and noting that  $z_1$  and  $z_2$  are independent.  $\square$

The  $F$  distribution in the  $F$ -test and ANOVA arises exactly from Property 2.

For the  $F$  test, we consider a model obtained by linear regression but with some of the columns of  $X$  removed. We are testing the null hypothesis that  $\beta_{i_1} = \beta_{i_2} = \dots = \beta_{i_\ell} = 0$  for the coefficients corresponding to these columns. Denote the full design matrix by  $X_{\text{FULL}}$  and the design matrix with columns removed by  $X_{\text{SUB}}$ , so that under the null hypothesis  $X_{\text{FULL}}\beta = X_{\text{SUB}}\beta_{\text{SUB}}$ . Likewise, let  $(X\hat{\beta})_{\text{FULL}}$  and  $(X\hat{\beta})_{\text{SUB}}$  be the respective projections of  $y$  onto  $C(X_{\text{FULL}})$  and  $C(X_{\text{SUB}})$ .



We want to apply Property 2 to obtain the  $F$  distribution. To do so, we need to choose suitable orthogonal subspaces  $U$  and  $V$ . Let  $U = C(X_{\text{SUB}})^\perp \cap C(X_{\text{FULL}})$  and let  $V = C(X_{\text{FULL}})^\perp$ . Then we have

$$\begin{aligned} (X\hat{\beta})_{\text{FULL}} - (X\hat{\beta})_{\text{SUB}} &= \text{Proj}_U(y) \\ r_{\text{FULL}} &= \text{Proj}_V(y) \end{aligned}$$

Next, observe that under the null hypothesis, we can write

$$y = X_{\text{FULL}}\beta + \epsilon = X_{\text{SUB}}\beta_{\text{SUB}} + \epsilon,$$

and it follows that the projected means  $\mathbb{E}(\text{Proj}_U(y))$  and  $\mathbb{E}(\text{Proj}_V(y))$  are 0. As a result, we can apply Property 2 to get the  $F$ -distribution

$$\frac{\|(X\hat{\beta})_{\text{FULL}} - (X\hat{\beta})_{\text{SUB}}\|_2^2 / \dim(U)}{\|r_{\text{FULL}}\|_2^2 / \dim(V)} \sim F_{\dim(U), \dim(V)}.$$

Finally, by orthogonality  $\|r_{\text{SUB}}\|_2^2 - \|r_{\text{FULL}}\|_2^2 = \|(X\hat{\beta})_{\text{FULL}} - (X\hat{\beta})_{\text{SUB}}\|_2^2$ , so we get the  $F$ -test

$$\frac{(\|r_{\text{SUB}}\|_2^2 - \|r_{\text{FULL}}\|_2^2) / \dim(U)}{\|r_{\text{FULL}}\|_2^2 / \dim(V)} \sim F_{\dim(U), \dim(V)}.$$

In particular, if  $X_{\text{FULL}}$  is a full rank  $m \times n$  matrix and  $X_{\text{SUB}}$  is  $m \times p$ , we get  $\dim(U) = n - p$  and  $\dim(V) = m - n$  so

$$\frac{(\|r_{\text{SUB}}\|_2^2 - \|r_{\text{FULL}}\|_2^2) / (n - p)}{\|r_{\text{FULL}}\|_2^2 / (m - n)} \sim F_{n-p, m-n},$$

which is the familiar formula for the  $F$ -test.

**1.2.3. One-way ANOVA.** The  $F$ -distribution for one-way ANOVA arises in the same manner as the  $F$  test. Suppose we have a one-way ANOVA model with  $k$  groups and let  $X_{\text{FULL}}$  be the  $m \times k$  matrix such that

$$(X_{\text{FULL}})_{i,j} = \begin{cases} 1 & \text{if measurement } i \text{ belongs to the } j\text{th group} \\ 0 & \text{otherwise} \end{cases}$$

Likewise, index  $y$  by letting  $y_{i,j}$  correspond to the  $i$ th measurement of the  $j$ th group. The null hypothesis is that  $\beta_j = \mathbb{E}(y_{\cdot,j})$  for  $j = 1, \dots, k$ , i.e. that the expected measurement for each group is the same.

Next, let  $X_{\text{SUB}} = \mathbf{1}_{m \times 1}$  and observe that under the null hypothesis  $X_{\text{FULL}}\beta = X_{\text{SUB}}\beta_1 \subseteq C(X_{\text{SUB}})$ . As a result, one can show that the arguments from Section 2 apply here as well. Writing  $\bar{y}^{(j)}$  as the mean for the  $j$ -th group and noting that  $(X\hat{\beta})_{\text{SUB}} = \mathbf{1}_{m \times 1}\bar{y}$ , we find

$$\|(X\hat{\beta})_{\text{FULL}} - (X\hat{\beta})_{\text{SUB}}\|_2^2 = \sum_{j=1}^k \sum_{i=1}^{m_j} (\bar{y}^{(j)} - \bar{y})^2$$

$$\|r_{\text{FULL}}\|_2^2 = \sum_{j=1}^k \sum_{i=1}^{m_j} (y_{i,j} - \bar{y}^{(j)})^2.$$

From the previous section,

$$\frac{\|(X\hat{\beta})_{\text{FULL}} - (X\hat{\beta})_{\text{SUB}}\|_2^2 / \dim(U)}{\|r_{\text{FULL}}\|_2^2 / \dim(V)} \sim F_{\dim(U), \dim(V)},$$

so we obtain the familiar one-way ANOVA test

$$\frac{\left( \sum_{j=1}^k \sum_{i=1}^{m_j} (\bar{y}^{(j)} - \bar{y})^2 \right) / (k-1)}{\left( \sum_{j=1}^k \sum_{i=1}^{m_j} (y_{i,j} - \bar{y}^{(j)})^2 \right) / (m-k)} \sim F_{k-1, m-k}.$$

**1.2.4. Intuition.** Intuitively, what's going on in the previous two sections is as follows. The null hypothesis is that  $X_{\text{FULL}}\beta = X_{\text{SUB}}\beta_{\text{SUB}}$ , i.e. that  $y = X_{\text{SUB}}\beta_{\text{SUB}} + \epsilon$ , where  $\epsilon \sim \mathcal{N}(0, \sigma^2 I)$ . In this scenario, Property 1 says that the component  $z_1$  of the noise  $\epsilon$  in the subspace  $U = C(X_{\text{FULL}}) \cap C(X_{\text{SUB}})^\perp$  and the component  $z_2$  in the subspace  $V = C(X_{\text{FULL}})^\perp$  are independent Gaussians whose covariance matrices are essentially  $\sigma^2 I_{\dim(U)}$  and  $\sigma^2 I_{\dim(V)}$ . As a result,

$$\frac{\|z_1\|_2^2 / \dim(U)}{\|z_2\|_2^2 / \dim(V)} \sim F_{\dim(U), \dim(V)}.$$

A helpful way to think of the distribution of  $\epsilon \sim \mathcal{N}(0, \sigma^2 I)$  is that  $\epsilon$  is a random variable that can be written as

$$\epsilon = \sigma^2 \left( \sum_{i=1}^m \alpha_i q_i \right)$$

where  $\alpha_1 \dots \alpha_m$  are independent  $\mathcal{N}(0, 1)$  random variables and  $q_1 \dots q_m$  is any arbitrary orthonormal basis. After selecting  $q_1, \dots, q_m$  by extending orthonormal bases for  $U$  and  $V$  (note  $U, V$  are orthogonal), one can see that  $\|z_1\|_2^2 / \dim(U) \sim \sigma^2 \chi_{\dim(U)}^2$  and  $\|z_2\|_2^2 / \dim(V) \sim \sigma^2 \chi_{\dim(V)}^2$  are independent so their ratio has distribution  $F_{\dim(U), \dim(V)}$ .

1.2.5. *Sample Variance.* Suppose  $Y_1, \dots, Y_m$  are *i.i.d.* random variables. The naive estimator  $\frac{1}{m} \sum_{i=1}^m (Y_i - \hat{\mu})^2$  for the variance is biased, and instead the sample variance  $\hat{\sigma}^2$  is defined as

$$\hat{\sigma}^2 = \frac{1}{m-1} \sum_{i=1}^m (Y_i - \hat{\mu})^2,$$

which is unbiased. The usual intuition is that the sample mean  $\hat{\mu}$  is correlated with each  $Y_i$ . Consequently, each term in the sum is slightly smaller than if one were to replace  $\hat{\mu}$  instead of  $\mu$ , and then some algebra is done to compute that  $m-1$  is the correct normalizing factor.

Our previous work provides another way to obtain this  $(m-1)$  factor. Consider the case where  $Y_1 \dots Y_m$  are independent  $N(\mu, \sigma)$  random variables. Letting  $X$  be the  $m \times 1$  design matrix of ones  $X = 1_{m \times 1}$  and also writing  $\beta = \mu$  and  $\epsilon \sim \mathcal{N}(0, \sigma^2 I)$ , we see that the regression model holds  $Y = X\beta + \epsilon$ . So our unbiased estimate for the variance from section 1.2.1 yields

$$\begin{aligned} \hat{\sigma}^2 &= \frac{1}{m-1} \sum_{i=1}^m r_i^2 \\ &= \frac{1}{m-1} \sum_{i=1}^m (Y_i - \hat{\mu})^2 \end{aligned}$$

To make things clear here,  $n = 1$  because  $X$  has one column. This means  $C(X)$  is one dimensional so  $r \in C(X)^\perp$  lives in an  $m-1$  dimensional subspace.

## 2. A NOTE ON PCA

Suppose we have an  $m \times n$  design matrix  $X$  consisting of  $m$  mean-centered data points  $x^{(1)}, \dots, x^{(m)}$ . PCA is a dimensionality-reduction technique which chooses a  $k$ -dimensional subspace  $k < n$  and orthogonally projects the data onto that subspace. (Here  $k$  is a parameter chosen beforehand.)

PCA is usually formulated in two equivalent ways (e.g. in Bishop). The first is finding a subspace which maximizes the variance of the projected data. The second is finding a subspace which minimizes the reconstruction error of the original data. The straightforward equivalence between these two ideas seems not often made explicit, so we do that here.

For any subspace  $U$ , we can write the variance of the projected data as

$$\sum_{i=1}^m \|\text{Proj}_U(x^{(i)})\|_2^2,$$

and we can write the reconstruction error as

$$\sum_{i=1}^m \|x^{(i)} - \text{Proj}_U(x^{(i)})\|_2^2.$$

By orthogonality

$$\sum_{i=1}^m \|x^{(i)}\|_2^2 = \sum_{i=1}^m \|\text{Proj}_U(x^{(i)})\|_2^2 + \sum_{i=1}^m \|x^{(i)} - \text{Proj}_U(x^{(i)})\|_2^2,$$

so of course

$$\arg \max_U \sum_{i=1}^m \|\text{Proj}_U(x^{(i)})\|_2^2 = \arg \min_U \sum_{i=1}^m \|x^{(i)} - \text{Proj}_U(x^{(i)})\|_2^2.$$