Regression 1: Different perspectives

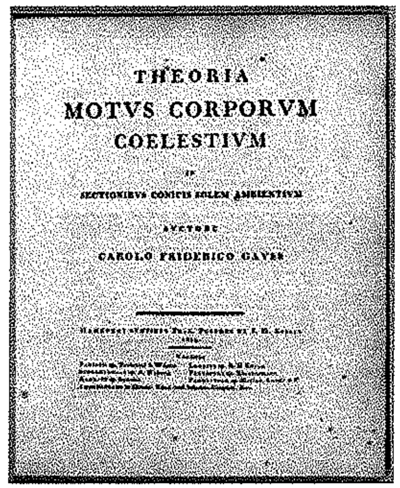
Ryan Tibshirani
Data Mining: 36-462/36-662

February 28 2013

Optional reading: ISL 3, ESL 3.2

Linear regression is an old topic

Linear regression, also called the method of least squares, is an old topic, dating back to Gauss in 1795 (he was 18!), later published in this famous book:



THEORIA
MOTVS CORPORVM
COELESTIVM
SECTIONIBVS CONICIS SOLEM AMBIENTIVM
AVCTORE
CAROLO FRIDERICO GAVSS

You have all seen linear regression before and a rigorous treatment of how to make inferences from a linear model, we won't repeat that here. The goal is to present some different perspectives on linear regression that are (hopefully) new. We'll start by reviewing the basics

Review: univariate regression

Suppose that we have observations $y = (y_1, \dots y_n) \in \mathbb{R}^n$, and we want to model these a linear function of $x = (x_1, \dots x_n) \in \mathbb{R}^n$. The univariate linear regression coefficient of y on x is

$$\hat{\beta} = \frac{\sum_{i=1}^n x_i y_i}{\sum_{i=1}^n x_i^2} = \frac{x^T y}{\|x\|_2^2}$$

This value $\hat{\beta} \in \mathbb{R}$ is optimal in the least squares sense:

$$\hat{\beta} = \underset{\beta}{\operatorname{argmin}} \sum_{i=1}^n (y_i - \beta x_i)^2 = \underset{\beta}{\operatorname{argmin}} \|y - \beta x\|_2^2.$$

We often think of the observations y as coming from the model

$$y = \beta^* x + \epsilon$$

where $x \in \mathbb{R}^n$ are fixed (nonrandom) measurements, $\beta^* \in \mathbb{R}$ is some true coefficient, and $\epsilon = (\epsilon_1, \dots \epsilon_n) \in \mathbb{R}^n$ are errors with $\mathrm{E}[\epsilon_i] = 0$, $\mathrm{Var}(\epsilon_i) = \sigma^2$, $\mathrm{Cov}(\epsilon_i, \epsilon_j) = 0$

We can also add an intercept term to the linear model:

$$y = \beta_0^* + \beta_1^* x + \epsilon$$

Again we estimate $\hat{\beta}_0, \hat{\beta}_1$ using least squares,

$$\hat{\beta}_0, \hat{\beta}_1 = \underset{\beta_0, \beta_1}{\operatorname{argmin}} \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2 = \underset{\beta_0, \hat{\beta}_1}{\operatorname{argmin}} \|y - \beta_0 \mathbb{1} - \beta_1 x\|_2^2$$

giving

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}, \quad \hat{\beta}_1 = \frac{(x - \bar{x}\mathbb{1})^T (y - \bar{y}\mathbb{1})}{\|x - \bar{x}\mathbb{1}\|_2^2} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2}$$

Notice that

$$\hat{\beta}_1 = \frac{\operatorname{cov}(x, y)}{\operatorname{var}(x)} = \operatorname{cor}(x, y) \sqrt{\frac{\operatorname{var}(y)}{\operatorname{var}(x)}}$$

Review: multivariate regression

Now suppose that we are considering $y \in \mathbb{R}^n$ as a function of multiple predictors $X_1, \ldots X_p \in \mathbb{R}^n$. We collect these predictors into columns of a predictor matrix $X \in \mathbb{R}^{n \times p}$. We assume that $X_1, \ldots X_p$ are linearly independent,[1] so that $\text{rank}(X) = p$

Our model:

$$y = X\beta^* + \epsilon$$

where $X \in \mathbb{R}^{n \times p}$ is considered fixed, $\beta^* = (\beta_1^*, \ldots \beta_p^*) \in \mathbb{R}^p$ are the true coefficients, and the errors $\epsilon = (\epsilon_1, \ldots \epsilon_n) \in \mathbb{R}^n$ are as before (i.e., satisfying $\mathrm{E}[\epsilon] = 0$ and $\mathrm{Cov}(\epsilon) = \sigma^2 I$)

$Var(\varepsilon_i) = \sigma^2$
$Cov(\varepsilon_i, \varepsilon_j) = 0 \quad i \neq j$

For an intercept term, we can just append a column $\mathbb{1} \in \mathbb{R}^n$ of all 1s to the matrix X

[1] Note that this necessarily implies that $p \leq n$

We estimate the coefficients $\hat{\beta} \in \mathbb{R}^p$ by least squares:

$$\hat{\beta} = \underset{\beta \in \mathbb{R}^p}{\operatorname{argmin}} \|y - X\beta\|_2^2$$

This gives

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

$$X = \begin{pmatrix} | & x_1 \\ | & x_2 \\ | & \vdots \\ | & x_n \end{pmatrix}$$

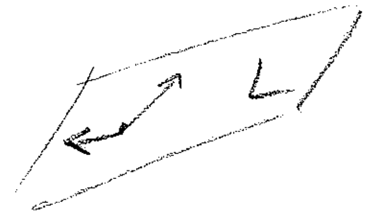
(Check: does this match the expressions for univariate regression, without and with an intercept?)

The fitted values are

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

This is a linear function of y, $\hat{y} = Hy$, where $H = X(X^T X)^{-1} X^T$ is sometimes called the hat matrix

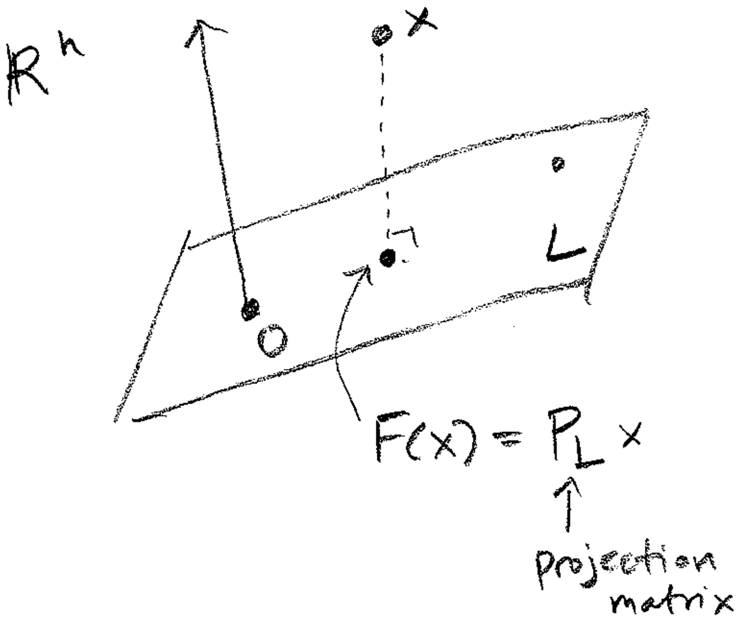
Review: projection matrices



Let $L \subseteq \mathbb{R}^n$ be a linear subspace, i.e., $L = \mathrm{span}\{v_1, \ldots v_k\}$ for some $v_1, \ldots v_k \in \mathbb{R}^n$. If $V \in \mathbb{R}^{n \times k}$ contains $v_1, \ldots v_k$ on its columns, then

$$\mathrm{span}\{v_1, \ldots v_k\} = \{a_1 v_1 + \ldots + a_k v_k : a_1, \ldots a_k \in \mathbb{R}\} = \mathrm{col}(V)$$

The function $F : \mathbb{R}^n \to \mathbb{R}^n$ that projects points onto L is called the projection map onto L. This is actually a linear function, $F(x) = P_L\, x$, where $P_L \in \mathbb{R}^{n \times n}$ is the projection matrix onto L



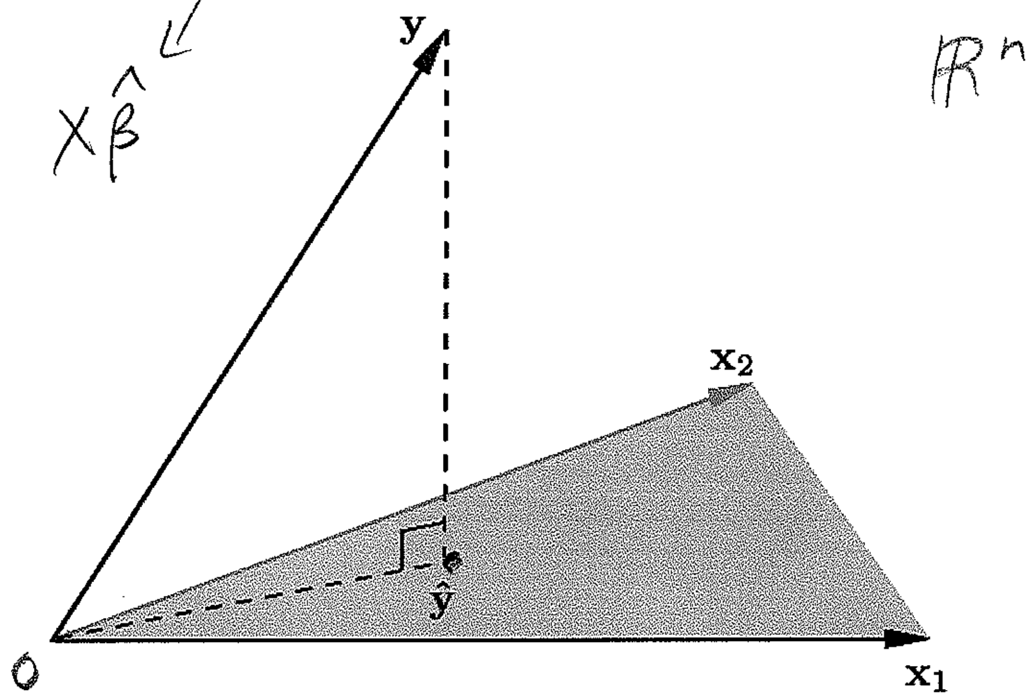
The matrix P_L is symmetric: $P_L^T = P_L$, and idempotent: $P_L^2 = P_L$. Furthermore, we have

- $P_L\, x = x$ for all $x \in L$, and
- $P_L\, x = 0$ for all $x \perp L$

Geometry of linear regression

The linear regression fit $\hat{y} \in \mathbb{R}^n$ is exactly the projection of $y \in \mathbb{R}^n$ onto the linear subspace $\mathrm{span}\{X_1, \ldots X_p\} = \mathrm{col}(X) \subseteq \mathbb{R}^n$

$X\hat{\beta}$



(Figure from ESL page 46.) Recall that $\hat{y} = \underbrace{X(X^TX)^{-1}X^T}_{H} y$, so we want to show that $X(X^TX)^{-1}X^T = P_{\mathrm{col}(X)}$

$$\hat{y} = Hy$$

First, we show that $H = X(X^TX)^{-1}X^T$ is symmetric and idempotent:

- Symmetric: $H^T = H$ ✓

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

- Idempotent: $H^2 = H$

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

Now suppose that $y \in \mathrm{col}(X)$; then $y = Xa$ for some $a \in \mathbb{R}^p$,

$= a_1X_1 + \dots a_pX_p$

$Hy = y$

$X\cancel{(X^TX)^{-1}X^TX}a = Xa$ ✓

Finally suppose that $y \perp \mathrm{col}(X)$; then $y \perp X_i$ for all $i = 1, \dots p$, so

$Hy = 0$

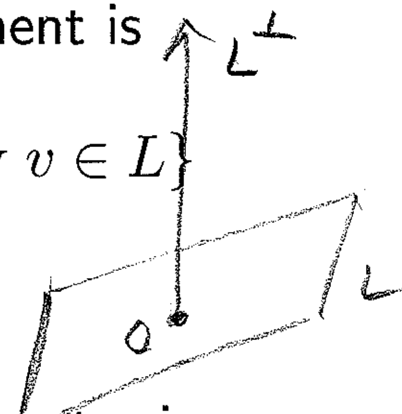
$X(X^TX)^{-1}X^Ty = 0.$ $\quad \begin{pmatrix} X_1^Ty \\ \vdots \\ X_p^Ty \end{pmatrix} = 0$ ✓

We proved that $H = X(X^TX)^{-1}X^T = P_{\mathrm{col}(X)}$, and therefore $\hat{y} = P_{\mathrm{col}(X)}\,y$

What do we gain from this geometry?

What does this geometric perspective do for us? There are some facts about projection maps that translate to useful facts about linear regression

E.g., for any subspace $L \subseteq \mathbb{R}^n$, its orthogonal complement is



$$L^\perp = \{x \in \mathbb{R}^n : x \perp L\} = \{x \in \mathbb{R}^n : x \perp v \text{ for any } v \in L\}$$

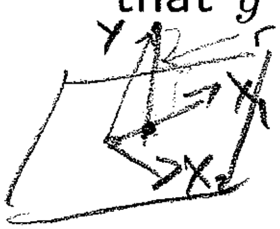
Fact: $P_L + P_{L^\perp} = I$, so that $P_{L^\perp} = I - P_L$

$P_{L^\perp} x = x - P_L x$

Hence for the linear regression of y on X, the residual vector is

$$y - \hat{y} = (I - P_{\mathrm{col}(X)})y = P_{\{\mathrm{col}(X)\}^\perp}\, y$$

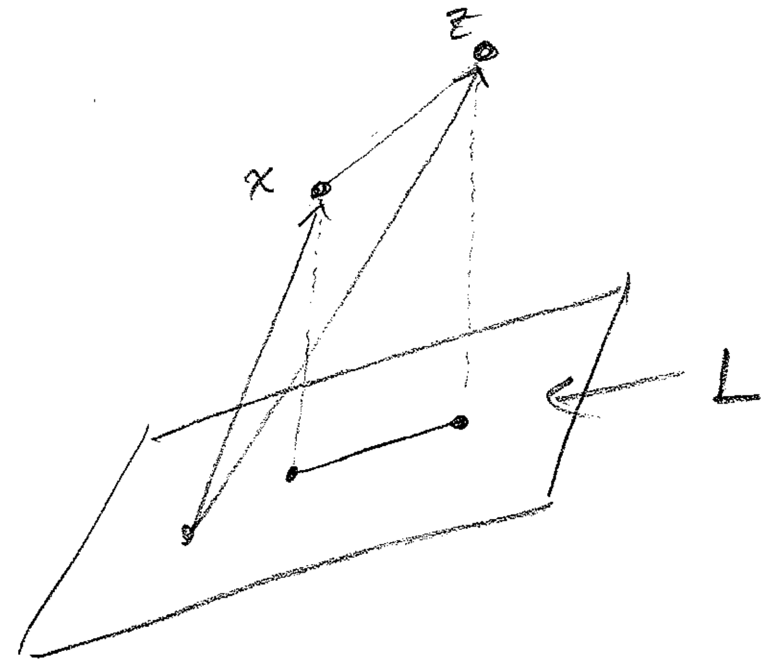
So $y - \hat{y}$ is orthogonal to any $v \in \mathrm{col}(X)$; in particular, this means that $y - \hat{y}$ is orthogonal to each of $X_1, \ldots X_p$



$X_i^T(y - \hat{y}) = 0$ for all $i = 1, \ldots, p$ because $y - \hat{y} \in (\mathrm{col}(X))^\perp$

E.g., the projection map P_L onto any linear subspace $L \subseteq \mathbb{R}^n$ is always non-expansive, that is, for any points $x, z \in \mathbb{R}^n$,

$$\|P_L\, x - P_L\, z\|_2 \leq \|x - z\|_2$$



Hence if $y_1, y_2 \in \mathbb{R}^n$ and $\hat{y}_1, \hat{y}_2 \in \mathbb{R}^n$ are their regression fits, then

$$\|\hat{y}_1 - \hat{y}_2\|_2 = \|P_{\mathrm{col}(X)}\, y_1 - P_{\mathrm{col}(X)}\, y_2\|_2 \leq \|y_1 - y_2\|_2$$

Furthermore, the geometric viewpoint is very helpful in proving more substantial results about linear regression. We'll cover two such results next

Unbiased estimates of linear functions

Note that we can write our linear regression model as

$$y_i = x_i^T \beta^* + \epsilon_i \qquad x_{i1}\beta_1^* + \dots \, x_{ip}\beta_p^* + \varepsilon_i$$

where $x_i \in \mathbb{R}^p$ is the ith measurement of predictor values (i.e., the ith row of $X \in \mathbb{R}^{n\times p}$), $\beta^* \in \mathbb{R}^p$ is the true coefficient vector, and $\epsilon_i \in \mathbb{R}$ is a random error satisfying $\mathrm{E}[\epsilon_i] = 0$, $\mathrm{Var}(\epsilon_i) = \sigma^2$, and $\mathrm{Cov}(\epsilon_i, \epsilon_j) = 0$. We expect future observations at some $x_0 \in \mathbb{R}^p$ to be of the form

$\varepsilon_i \overset{iid}{\sim} N(0, \sigma^2)$

$$y_0 = x_0^T \beta^* + \epsilon_0$$

with ϵ_0 coming from the same error distribution

Fact: if $\hat\beta$ is the linear regression estimate, then for any $a \in \mathbb{R}^p$, the estimate $\underline{a^T\hat\beta}$ is unbiased for $\underline{a^T\beta^*}$, i.e., $\mathrm{E}[a^T\hat\beta] = a^T\beta^*$

$E[a^T\hat\beta] = a^T E[\hat\beta]$

Why is this important? Because it says that our predictions $x_0^T\hat\beta$ at any $x_0 \in \mathbb{R}^p$ will be unbiased for the true mean $\underline{x_0^T\beta^*}$ at x_0

Proof of this fact:

$$\begin{aligned}
\mathrm{E}[a^T\hat{\beta}] &= \mathrm{E}[a^T(X^TX)^{-1}X^Ty] \\
&= a^T(X^TX)^{-1}X^T\,E[y] \qquad\qquad y = X\beta^* + \varepsilon \\
&= a^T\cancel{(X^TX)}^{-1}\cancel{X^TX}\beta^*) \\
&= a^T\beta^*
\end{aligned}$$

Note that the estimate $a^T\hat{\beta} = a^T(X^TX)^{-1}X^Ty = b^Ty$ is just a linear function of y, with $b = X(X^TX)^{-1}a$

(think of $a = x_0$)

We're going to consider the estimation of $\underline{a^T\beta^*}$, for an arbitrary $a \in \mathbb{R}^p$, and restrict our attention to linear functions of y, c^Ty for some $c \in \mathbb{R}^n$, that are unbiased for $a^T\beta^*$:

$$\mathrm{E}[c^Ty] = a^T\beta^*$$

Best linear unbiased estimate (BLUE)

A natural question is: what is the best linear unbiased estimate (BLUE) $c^T y$ for estimating $\underline{a^T \beta^*}$? Recall that the linear regression estimate $a^T \hat{\beta} = b^T y$ falls into this category (linear and unbiased)

By "best" here, we mean the estimate $c^T y$ that minimizes the mean squared error in estimating $a^T \beta^*$:

$$\mathrm{MSE}(c^T y) = \mathrm{E}[(c^T y - a^T \beta^*)^2]$$

Gauss-Markov theorem: the linear regression estimate $a^T \hat{\beta} = b^T y$ is the BLUE, i.e., if $c^T y$ is any other unbiased estimate of $a^T \beta^*$, then

$$\mathrm{MSE}(\underbrace{a^T \hat{\beta}}_{b^T y}) \leq \mathrm{MSE}(c^T y) \quad \Longleftarrow$$

The proof uses the facts from geometry (Homework 4)

predict $a^T \hat{\beta}$ we only estimates of form $c^T y$

not $f(y)$

Note that for an unbiased estimator $F = F(y)$, of μ

$$\begin{aligned}\mathrm{MSE}(F) &= E\left[(F(y)-\mu)^2\right] \\ &= E\left[(F(y)-E[F(y)])^2\right] \\ &= \mathrm{Var}(F)\end{aligned}$$

So the Gauss-Markov theorem equivalently says that the regression estimate $a^T\hat{\beta}$ has smallest variance compared to all linear unbiased estimates of $a^T\beta^*$

Does this mean we should always use linear regression?

Univariate regression revisted

Write $\langle a, b \rangle = a^T b = \sum_{i=1}^n a_i b_i$ as the inner-product for vectors $a, b \in \mathbb{R}^n$

In this notation, we can write the univariate linear regression coefficient of $y \in \mathbb{R}^n$ on a single predictor $x \in \mathbb{R}^n$ as

$$\hat{\beta} = \frac{\langle x, y \rangle}{\|x\|_2^2} = \frac{\sum x_i y_i}{\sum x_i^2}$$

Given p predictor variables $X_1, \ldots X_p \in \mathbb{R}^n$, the univariate linear regression coefficient of y on X_j is

$$\underline{\hat{\beta}_j} = \frac{\langle X_j, y \rangle}{\|X_j\|_2^2}$$

this not generally j^{th} multivariate coefficient

Fact: if $X_1, \ldots X_p$ are orthogonal, then this is also the coefficient of X_j in the multivariate linear regression of y on all of $X_1, \ldots X_p$

Univariate regression with intercept

For univariate linear regression with an intercept term, i.e., for regressing $y \in \mathbb{R}^n$ on predictors $\mathbb{1}, x \in \mathbb{R}^n$, we can write the coefficient of x as

$$\hat{\beta}_1 = \frac{\langle \overbrace{x - \bar{x}\mathbb{1}}^{z}, y \rangle}{\|\underbrace{x - \bar{x}\mathbb{1}}_{z}\|_2^2} = \frac{\sum (x_i - \bar{x}) y_i}{\sum (x_i - \bar{x})^2}$$

$\mathbb{1} = (1, 1, \ldots 1)$

We can alternatively view this as result of two steps:

1. Regress x on $\mathbb{1}$, yielding the coefficient

$$\frac{\langle \mathbb{1}, x \rangle}{\|\mathbb{1}\|_2^2} = \frac{\langle \mathbb{1}, x \rangle}{n} = \bar{x}$$

   and the residual $z = x - \bar{x}\mathbb{1} \in \mathbb{R}^n$

2. Regress y on z, yielding the coefficient

$$\hat{\beta}_1 = \frac{\langle z, y \rangle}{\|z\|_2^2} = \frac{\langle x - \bar{x}\mathbb{1}, y \rangle}{\|x - \bar{x}\mathbb{1}\|_2^2}$$

Gram-Schmidt

Multivariate regression by orthogonalization

This idea extends to multivariate linear regression of $y \in \mathbb{R}^n$ on predictors $X_1, \ldots X_p \in \mathbb{R}^n$. Consider the p-step procedure:

1. Let $Z_1 = X_1$
2. For $j = 2, \ldots p$:
   Regress X_j onto $Z_1, \ldots Z_{j-1}$ to get coefficients $\hat{\gamma}_{jk} = \frac{\langle Z_k, X_j \rangle}{\|Z_k\|_2^2}$ for $k = 1, \ldots j-1$, and residual vector

$$Z_j = X_j - \sum_{k=1}^{j-1} \hat{\gamma}_{jk} Z_k$$

3. Regress y on Z_p to get the coefficient $\hat{\beta}_p$ = p^{th} coefficient in multiple reg. of y on $X_1, \ldots X_p$

Claim: the output $\hat{\beta}_p$ of this algorithm is exactly the coefficient of X_p in the multivariate linear regression of y on $X_1, \ldots X_p$

Why is this true? To see this, we argue in several steps

Step 1: The vectors $Z_1, \ldots Z_p \in \mathbb{R}^n$ produced by this algorithm are orthogonal. To see this, note that at any stage, we define Z_j to be the residual from regressing X_j onto $Z_1, \ldots Z_{j-1}$. Therefore (by an earlier fact), Z_j is orthogonal to $Z_1, \ldots Z_{j-1}$

Step 2: For any $j = 1, \ldots p$, the definition $Z_j = X_j - \sum_{k=1}^{j-1} \hat{\gamma}_{jk} Z_k$ shows that each Z_j is a linear combination of $X_1, \ldots X_j$, so

$$\text{span}\{Z_1, \ldots Z_j\} \subseteq \text{span}\{X_1, \ldots X_j\}$$

But rearranging the above definition shows that each X_j is also a linear combination of $Z_1, \ldots Z_j$, so

$$\text{span}\{X_1, \ldots X_j\} \subseteq \text{span}\{Z_1, \ldots Z_j\}$$

Hence the spans are equal, $\text{span}\{X_1, \ldots X_j\} = \text{span}\{Z_1, \ldots Z_j\}$

Step 3: Using that $\mathrm{span}\{X_1, \ldots X_p\} = \mathrm{span}\{Z_1, \ldots Z_p\}$ (and using what we know about linear regression and projections), the linear regression fit y on $X_1, \ldots X_p$ is the same as the linear regression fit of y on $Z_1, \ldots Z_p$. Call this fit $\hat{y}$. Hence we can write

$$\hat{y} = c_1 Z_1 + \ldots + c_p Z_p$$

for some $c_1, \ldots c_p$

Furthermore, as $Z_1, \ldots Z_p$ are orthogonal, the coefficients $c_1, \ldots c_p$ are just given by univariate linear regression, so in particular we have

$$c_p = \frac{\langle Z_p, y \rangle}{\|Z_p\|_2^2} = \hat{\beta}_p$$

Step 4: For each Z_j in the expression

$$\hat{y} = c_1 Z_1 + \ldots + c_{p-1} Z_{p-1} + \hat{\beta}_p Z_p$$

plug in the linear representation in terms of $X_1, \ldots X_p$. Note that the variable X_p appears only through Z_p, and the coefficient of X_p is 1:

$$Z_p = X_p - \sum_{k=1}^{p-1} \hat{\gamma}_{pk} Z_k$$

Therefore we can write, for some constants $a_1, \ldots a_{p-1}$,

$$\hat{y} = a_1 X_1 + \ldots + a_{p-1} X_{p-1} + \hat{\beta}_p X_p$$

Hence $\hat{\beta}_p$ is the coefficient of X_p in the multiple regression of y on $X_1, \ldots X_p$

Closed-form expression for multiple regression coefficients

We just proved that, in the regression of $y \in \mathbb{R}^n$ onto predictors $X_1, \ldots X_p \in \mathbb{R}^n$, the coefficient of X_p is

$$\hat{\beta}_p = \frac{\langle Z_p, y \rangle}{\|Z_p\|_2^2}$$

where Z_p is the residual from regressing X_p onto $Z_1, \ldots Z_{p-1}$, i.e., the residual from regressing X_p onto $X_1, \ldots X_{p-1}$

Note that our algorithm didn't depend in any way on the order of the variables, so for any $j = 1, \ldots p$, we could have modified this order by putting X_j at the end, and we get the multiple regression coefficient of X_j:

$$\hat{\beta}_j = \frac{\langle Z_j, y \rangle}{\|Z_j\|_2^2}$$

univariate regression of y on Z_j

where $\underline{Z_j}$ is the residual from regressing $\underline{X_j}$ onto all X_i, $i \neq j$

"removing" the effect of X_i, $i \neq j$ on X_j

Recap: perspectives on linear regression

In this lecture we saw some new perspectives on linear regression

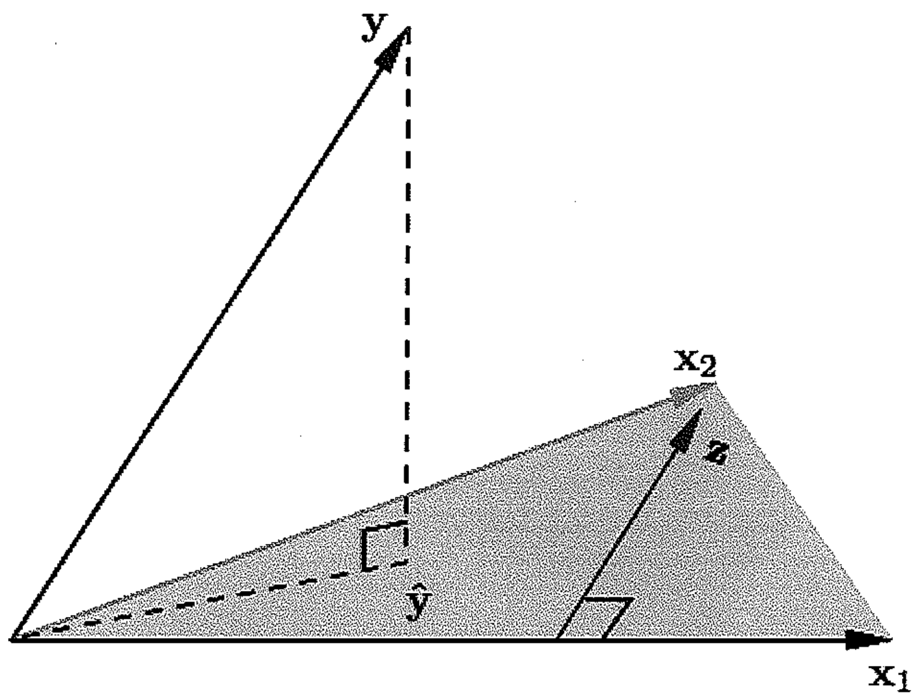
We saw that the linear regression fit of $y \in \mathbb{R}^n$ onto $X \in \mathbb{R}^{n \times p}$, whose columns are $X_1, \ldots X_p \in \mathbb{R}^n$ is the projection of y onto the linear subspace $\mathrm{col}(X) = \mathrm{span}\{X_1, \ldots X_p\}$. This immediately gives us some usual facts about regression

Given any vector $a \in \mathbb{R}^p$, if we assume that y comes from a model with true coefficients β^* (and uncorrelated errors with mean zero and constant variance), then the regression estimate $a^T \hat{\beta}$ is the best linear unbiased estimate (BLUE) of $a^T \beta^*$

Each coefficient $\hat{\beta}_j$ in multiple linear regression can be expressed explicitly in terms y and the residual from projecting X_j onto all variables X_i, $i \neq j$

Next time: more regression

A few more perspectives on regression ... moving into modern regression



(From ESL page 54)