Notes for 751-752
Sections 4-5

Martin Lindquist*

September 3, 2017

*Thanks to Brian Caffo and Ingo Ruczinski for supplying their notes, upon which much of this document is based.

4 Linear regression

Next we consider the case of simple linear regression. This entails minimizing:

$$||\mathbf{y} - (\beta_0 \mathbf{J}_n + \beta_1 \mathbf{x})||^2 \tag{1}$$

over β_1 and β_2. We can think about the problem in two ways. First, the space

$$\Gamma = \{\beta_0 \mathbf{1}_n + \beta_1 \mathbf{x} \mid \beta_0, \beta_1 \in \mathbb{R}\}$$

is a two dimensional subspace of $\mathbb{R}^n$. Therefore, the least squares equation finds the projection of the observed data point onto two dimensional subspace spanned by the two vectors $\mathbf{1}_n$ and $\mathbf{x}$.

Second, we can consider the scatterplot of points (x_i, y_i). The goal is to find the best fitting line of the form $y = \beta_0 + \beta_1 x$ by minimizing the sum of the squared vertical distances between the points and the fitted line. Here β_1 is referred to as the slope, and β_0 as the intercept. The slope β_1 has units 'y-units per x-units'. The intercept corresponds to the value of y when $x = 0$, and is not always meaningful if the 0 lies outside the range of reasonable values for x.

We begin by rewriting the least squares criterion as follows:

$$f(\beta_0, \beta_1) = ||\mathbf{y} - (\beta_0 \mathbf{J}_n + \beta_1 \mathbf{x})||^2 = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2. \tag{2}$$

The least squares estimates can be found by differentiating f with respect to β_0 and β_1 and setting the partial derivatives equal to 0, i.e.

$$\frac{\partial f(\beta_0, \beta_1)}{\partial \beta_0} = -2 \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i) = 0 \tag{3}$$

$$\frac{\partial f(\beta_0, \beta_1)}{\partial \beta_0} = -2 \sum_{i=1}^{n} x_i (y_i - \beta_0 - \beta_1 x_i) = 0. \tag{4}$$

The values of $\hat{\beta}_0$ and $\hat{\beta}_0$ that minimize f are given by the solution to the normal equations:

$$\sum_{i=1}^{n} y_i = n\hat{\beta}_0 + \hat{\beta}_1 \sum_{i-1}^{n} x_i \tag{5}$$

$$\sum_{i=1}^{n} x_i y_i = \hat{\beta}_0 \sum_{i-1}^{n} x_i + \hat{\beta}_1 \sum_{i-1}^{n} x_i^2. \tag{6}$$

Solving the normal equations gives us the following estimates:

$$\hat{\beta}_1 = \frac{\sum_{i-1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \tag{7}$$

and

$$\hat{\beta}_0 = \bar{\mathbf{y}} - \hat{\beta}_1 \bar{\mathbf{x}}. \tag{8}$$

Note we can also write $\hat{\beta}_1$ as follows:

$$\hat{\beta}_1 = \frac{(\mathbf{x} - \bar{x}\mathbf{J}_n)'(\mathbf{y} - \bar{y}\mathbf{J}_n)}{(\mathbf{x} - \bar{x}\mathbf{J}_n)'(\mathbf{x} - \bar{x}\mathbf{J}_n)}. \tag{9}$$

To check whether $\hat{\beta}_0$ and $\hat{\beta}_1$ correspond to the minimum of f, it suffices to check whether the Hessian matrix is positive definite.The Hessian matrix can be expressed as

$$\begin{pmatrix} \frac{\partial f}{\partial \beta_0^2} & \frac{\partial f}{\partial \beta_0 \beta_1} \\ \frac{\partial f}{\partial \beta_0 \beta_1} & \frac{\partial f}{\partial \beta_1^2} \end{pmatrix} = \begin{pmatrix} 2n & 2\sum x_i \\ 2\sum x_i & 2\sum x_i^2 \end{pmatrix}$$

The matrix is positive definite if $n > 0$ and the determinant is > 0. This corresponds to $\sum (x_i - \bar{x})^2 > 0$, which holds if not all values of x_i are the same. If this does not hold, simple linear regression is not meaningful, so this is a reasonable assumption.

Note that we can rewrite the estimated slope as follows:

$$\hat{\beta}_1 = \hat{\rho}_{xy} \frac{\hat{\sigma}_y}{\hat{\sigma}_x}.$$

Thus, the best fitting line has a slope equal to the correlation times the ratio of the standard deviations. If we reverse the role of $\mathbf{x}$ and $\mathbf{y}$, we simply invert the ratio of the standard deviations. Thus we also note, that if we center and scale our data first so that the resulting vectors have mean 0 and variance 1, our slope is exactly the correlation between the vectors.

Note: 4.1 These estimates generalize the single parameter models previously discussed.

4.1 Coding example

Refering back to the `diamond` example from the previous section. Let us compute β_0 and β_1 in two differnt ways.

```
> library(UsingR)
> data(diamond)
> x = diamond$carat
> y = diamond$price
```

```
> beta1 = cor(x, y) * sd(y) / sd(x)
> beta0 = mean(y) - beta1 * mean(x)
> c(beta0, beta1)
[1] -259.6259 3721.0249
> # versus estimate with lm
> coef(lm(y ~ x))
(Intercept)           x
  -259.6259   3721.0249
```

4.2 Fitted values

The term $y_i = \beta_0 + \beta_1 x_i$, for $i = 1, \ldots n$ is called the fitted value for the i^{th} observation. We define $\hat{\mathbf{y}} = (\hat{y}_1, \ldots, \hat{y}_n)'$ to be the vector of fitted values. This can be expressed as $\hat{\mathbf{y}} = \hat{\beta}_0 \mathbf{J}_n + \hat{\beta}_1 \mathbf{x}$. Whereas $\mathbf{y}$ lives in $\mathbb{R}^n$, $\hat{\mathbf{y}}$ lives in Γ, the two dimensional linear subspace of $\mathbb{R}^n$ spanned by the two vectors, $\mathbf{J}_n$ and $\mathbf{x}$. We can think of our least squares as minimizing

$$||\mathbf{y} - \hat{\mathbf{y}}||$$

over all $\hat{\mathbf{y}} \in \Gamma$. The fitted values are the orthogonal projection of the observed data onto this linear subspace.

4.3 Coding example

Compute the predicted values for the `diamond` data in two different ways.

```
> yhat = beta0 + beta1*x
> yhat
 [1]  372.9483  335.7381  372.9483  410.1586  670.6303  335.7381  298.5278
 [8]  447.3688  521.7893  298.5278  410.1586  782.2611  335.7381  484.5791
[15]  596.2098  819.4713  186.8971  707.8406  670.6303  745.0508  410.1586
[22]  335.7381  372.9483  335.7381  372.9483  410.1586  372.9483  410.1586
[29]  372.9483  298.5278  372.9483  931.1020  931.1020  298.5278  335.7381
[36]  335.7381  596.2098  596.2098  372.9483  968.3123  670.6303 1042.7328
[43]  410.1586  670.6303  670.6303  298.5278  707.8406  298.5278

> fit = lm(y~x)

> max(abs(yhat-predict(fit)))
[1] 0
```

Let us now compute the predicted value for $x = 0.20$ in two ways.

```
> beta0 + beta1 * .20
[1] 484.5791

> predict(lm(y ~ x), newdata = data.frame(x = .2))
       1
484.5791
```

4.4 Residuals

A residual, denoted e_i, is the difference between the observed and the predicted value of y_i, i.e. $e_i = y_i - \hat{y}_i$. The residuals show how far the individual data points fall from the regression function. We define $\mathbf{e} = (e_1, \ldots, e_n)'$ to be the vector of residuals, which can also be expressed as $\mathbf{e} = \mathbf{y} - \hat{\mathbf{y}}$ Geometrically speaking, the residuals are the orthogonal vector pointing to $\mathbf{y}$ from $\hat{\mathbf{y}}$. Least squares can be thought of as minimizing the sum of the squared residuals. The quantity $||\mathbf{e}||^2$ is called the sum of the squared errors while $\frac{1}{n-2}||\mathbf{e}||^2$ is called the mean squared error or the residual variance.

Note: 4.2 The regression line and the residuals have the following properties:

- The sum of the residuals is zero.
- The sum of the squared residuals is a minimum of f.
- $\mathbf{y}'\mathbf{J}_n = \hat{\mathbf{y}}'\mathbf{J}_n 0$
- $\mathbf{x}'\mathbf{e} = 0$ and $\hat{\mathbf{y}}'\mathbf{e} = 0$
- The regression line always goes through the point $(\bar{x}, \bar{y})$.

4.5 Coding example

Compute the residuals for the `diamond` data set in two different ways and compare.

```
> e = y-yhat
 [1] -17.9483176  -7.7380691 -22.9483176 -85.1585661 -28.6303057   6.2619309
 [7]  23.4721795  37.6311854 -38.7893116  24.4721795  51.8414339  40.7389488
[13]   0.2619309  13.4209369  -1.2098087  40.5287002  36.1029250 -44.8405542
[19]  79.3696943 -25.0508027  57.8414339   9.2619309 -20.9483176  -3.7380691
```

```
[25] -19.9483176  27.8414339 -54.9483176   8.8414339 -26.9483176  16.4721795
[31] -22.9483176 -13.1020453 -12.1020453  -0.5278205   3.2619309   2.2619309
[37]  -1.2098087 -43.2098087 -27.9483176 -23.3122938 -15.6303057  43.2672091
[43]  32.8414339   7.3696943   4.3696943 -11.5278205 -14.8405542  17.4721795

> max(abs(e - resid(fit)))
[1] 0
```

4.6 Connecting the pieces

Here is an alternative approach towards estimating the parameters of the simple linear regression model that links back to the single variable regression models discussed in the previous chapter.

Consider fixing β_1 and minimizing the least square criteria

$$||\mathbf{y} - \beta_1 \mathbf{x} - \beta_0 \mathbf{J}_n||^2$$

with respect to β_0. Let $\hat{\beta}_0(\beta_1)$ be the least squares minimum for β_0 for a given value of β_1. Note β_0 is now a function of β_1. Following the results from mean only regression we know that

$$\hat{\beta}_0(\beta_1) = \frac{1}{n}(\mathbf{y} - \beta_1 \mathbf{x})\mathbf{J}_n = \bar{y} - \beta_1 \bar{x}.$$

Therefore, plugging this into the least squares equation, we know that

$$||\mathbf{y} - \beta_1 \mathbf{x} - \beta_0 \mathbf{J}_n||^2 \geq ||\mathbf{y} - \bar{y}\mathbf{J}_n + \beta_1(\mathbf{x} - \bar{x}\mathbf{J}_n)||^2 = ||\tilde{\mathbf{y}} - \beta_1 \tilde{\mathbf{x}}||^2, \tag{10}$$

where $\tilde{\mathbf{y}}$ and $\tilde{\mathbf{x}}$ are the centered versions of $\mathbf{y}$ and $\mathbf{x}$, respectively. We know from previously that (10) is minimized by

$$\hat{\beta}_1 = \frac{< \tilde{\mathbf{x}}, \tilde{\mathbf{y}} >}{< \tilde{\mathbf{x}}, \tilde{\mathbf{x}} >} = \frac{(\mathbf{x} - \bar{x}\mathbf{J}_n)'(\mathbf{y} - \bar{y}\mathbf{J}_n)}{(\mathbf{x} - \bar{x}\mathbf{J}_n)'(\mathbf{x} - \bar{x}\mathbf{J}_n)}.$$

Plugging this into $\hat{\beta}_0(\hat{\beta}_1)$ we get that

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}.$$

Note, the slope estimate when including an intercept is identical to that of regression through the origin after centering the data. The intercept simply forces the line through $(\bar{x}, \bar{y})$.

4.7 Regression by Successive Orthogonalization

Consider two vectors $\mathbf{x}$ and $\mathbf{z}$ that are orthogonal, i.e. $\mathbf{x}'\mathbf{z} = 0$. Then, regressing $\mathbf{y}$ on $\mathbf{x}$ and $\mathbf{z}$ separately gives the same coefficients as regressing $\mathbf{y}$ on $\mathbf{x}$ and $\mathbf{z}$ together. This can be seen by re-expressing the least-squares criteria as follows:

$$||\mathbf{y} - (\beta\mathbf{x} + \gamma\mathbf{z})||^2 = (\mathbf{y} - \beta\mathbf{x})'(\mathbf{y} - \beta\mathbf{x})' + (\mathbf{y} - \gamma\mathbf{z})'(\mathbf{y} - \gamma\mathbf{z})' - \mathbf{y}'\mathbf{y}$$

Thus, maximizing the least-squares criteria can be perfomred for each variable seperately.

We can apply these ideas to the previosly discussed simple linear regression model. Note the vector $\mathbf{x} - \bar{x}\mathbf{J}_n$ is orthognal to the vector $\mathbf{J}_n$.

The projection of $\mathbf{y}$ onto $\mathbf{u}_1$ can be expressed as $\hat{\mathbf{y}}_1 = \hat{\alpha}_0\mathbf{J}_n$ where $\hat{\alpha}_0 = \bar{y}$. Similarly, the projection of $\mathbf{y}$ onto $\mathbf{u}_2$ can be expressed as $\hat{\mathbf{y}}_1 = \hat{\alpha}_1(\mathbf{x} - \bar{x}\mathbf{J}_n)$ where

$$\hat{\alpha}_1 = \frac{(\mathbf{x} - \bar{x}\mathbf{J}_n)'\mathbf{y}}{(\mathbf{x} - \bar{x}\mathbf{J}_n)'(\mathbf{x} - \bar{x}\mathbf{J}_n)} = \frac{(\mathbf{x} - \bar{x}\mathbf{J}_n)'(\mathbf{y} - \bar{y}\mathbf{J}_n)}{(\mathbf{x} - \bar{x}\mathbf{J}_n)'(\mathbf{x} - \bar{x}\mathbf{J}_n)}.$$

Now note that $\hat{\alpha}_1 = \hat{\beta}_1$ from before. Thus, we can write

$$\begin{aligned}
\hat{\mathbf{y}} &= \hat{\mathbf{y}}_1 + \hat{\mathbf{y}}_2 \\
&= \bar{y}\mathbf{J}_n + \hat{\beta}_1(\mathbf{x} - \bar{x}\mathbf{J}_n) \\
&= (\bar{y} - \hat{\beta}_1\bar{x})\mathbf{J}_n + \hat{\beta}_1\mathbf{x}
\end{aligned}$$

Setting $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1\bar{x}$ provides the familiar solution.

We will see later that this results holds for the more general case consisting of linear regression with multiple explanatory variables.

4.8 Extension to other spaces

It is interesting to note that nothing we've discussed is intrinsic to $\mathbb{R}^n$. Any space with a norm and inner product and absent of extraordinary mathematical pathologies would suffice. Hilbert spaces are perhaps the most directly extendable.

As an example, let's develop linear regression for a space of (Lebesgue) square integrable functions. That is, let y be in the space of functions from $[0, 1] \to \mathbb{R}$ with finite squared itegral. Define the inner product as $\langle f, g\rangle = \int_0^1 f(t)g(t)dt$. Consider finding the best approximation to y from the function x (also in that space).

Thus, we want to minimize:

$$||y - \beta_1 x||^2 = \int_0^1 \{y(t) - \beta_1 x(t)\}^2 dt.$$

You might have guessed that the solution will be $\hat{\beta} = \frac{\langle y,x \rangle}{\langle x,x \rangle} = \frac{\langle y,x \rangle}{||x||^2}$. Let's show it (knowing that this is the solution):

$$\begin{aligned}
||y - \beta_1 x||^2 &= ||y - \hat{\beta}_1 x + \hat{\beta}_1 x - \beta_1 x||^2 \\
&= ||y - \hat{\beta}_1 \bar{x}||^2 - 2\langle y - \hat{\beta}_1 x, \hat{\beta}_1 x - \beta_1 x \rangle + ||\hat{\beta}_1 x - \beta_1 x||^2 \\
&\geq ||y - \hat{\beta}_1 \bar{x}||^2 - 2\langle y - \hat{\beta}_1 x, \hat{\beta}_1 x - \beta_1 x \rangle \\
&= ||y - \hat{\beta}_1 \bar{x}||^2 - 2\hat{\beta}_1 \langle y, x \rangle + 2\beta_1 \langle y, x \rangle + 2\hat{\beta}_1^2 ||x||^2 - 2\hat{\beta}_1 \beta_1 ||X||^2 \\
&= ||y - \hat{\beta}_1 \bar{x}||^2 - 2\frac{\langle y, x \rangle^2}{||x||^2} + 2\beta_1 \langle y, x \rangle + 2\frac{\langle y, x \rangle^2}{||x||^2} - 2\beta_1 \langle y, x \rangle \\
&= ||y - \hat{\beta}_1 x||^2
\end{aligned}$$

Therefore, $\hat{\beta}_1$ is the least squares estimate.

We can extend this to include an intercept. Let j be a function that is constant at 1. Let $\bar{y} = \int_0^1 y(t)dt$ be the average of y over the domain and define $\bar{x}$ similarly. Then consider minimizing (over β_0 and β_1)

$$||y - \beta_0 j - \beta_1 x||^2$$

First, hold β_1 fixed. By our previous result, we have that the minimizer must satisfy:

$$\beta_0 = < y - \beta_1 x, j > / ||j||^2 = \bar{y} - \beta_1 \bar{x}.$$

Plugging this back into our least squares equation we obtain that:

$$\begin{aligned}
||y - \beta_0 j - \beta_1 x||^2 &\geq ||y - \bar{y} - \beta_1 (x - \bar{x})||^2 \\
&= ||\tilde{y} - \beta_1 \tilde{x}||^2
\end{aligned}$$

where $\tilde{y}$ and $\tilde{x}$ are the centered functions. We know that this is minimized

5 Least squares

In this chapter we develop least squares for the general linear model $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$.

5.1 Basics

Let $\mathbf{X}$ be a design matrix, notationally its elements and column vectors are:

$$\mathbf{X} = \begin{bmatrix} x_{11} & \dots & x_{1p} \\ \vdots & \dots & \vdots \\ x_{n1} & \dots & x_{np} \end{bmatrix} = [\mathbf{x}_1 \dots \mathbf{x}_p].$$

We are assuming that $n \geq p$ and $\mathbf{X}$ is of full (column) rank. Consider ordinary least squares

$$f(\boldsymbol{\beta}) = ||\mathbf{y} - \mathbf{X}\boldsymbol{\beta}||^2 = (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) = \mathbf{y}'\mathbf{y} - 2\boldsymbol{\beta}'\mathbf{X}'\mathbf{y} + \boldsymbol{\beta}'\mathbf{X}'\mathbf{X}\boldsymbol{\beta}. \tag{11}$$

To minimize $f(\boldsymbol{\beta})$, begin by taking the derivative with respect to $\boldsymbol{\beta}$:

$$\frac{df}{d\boldsymbol{\beta}} = -2\mathbf{X}'\mathbf{y} + 2\mathbf{X}'\mathbf{X}\boldsymbol{\beta}.$$

Solving for 0 leads to the so called normal equations:

$$\mathbf{X}'\mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{X}'\mathbf{y}.$$

The matrix $\mathbf{X}'\mathbf{X}$ retains the same rank as $\mathbf{X}$. Therefore, it is a full rank $p \times p$ matrix and hence is invertible. We can then solve the normal equations as:

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

The Hessian of (11) is simply $2\mathbf{X}'\mathbf{X}$, which is positive definite. (This is clear since for any non-zero vector, $\mathbf{a}$, we have that $\mathbf{X}'\mathbf{a}$ is non-zero since $\mathbf{X}$ is full rank and then $\mathbf{a}'\mathbf{X}'\mathbf{X}\mathbf{a} = ||\mathbf{X}\mathbf{a}||^2 > 0$.) Thus, the root of our derivative is indeed a minimum.

5.2 Coding example

```
> y = swiss$Fertility
> x = as.matrix(swiss[,-1])
> solve(t(x) %*% x, t(x) %*% y)
                         [,1]
1                66.9151817
```

```
Agriculture      -0.1721140
Examination      -0.2580082
Education        -0.8709401
Catholic          0.1041153
Infant.Mortality  1.0770481
> summary(lm(y ~ x - 1))$coef
                    Estimate  Std. Error   t value     Pr(>|t|)
x1                66.9151817 10.70603759  6.250229 1.906051e-07
xAgriculture      -0.1721140  0.07030392 -2.448142 1.872715e-02
xExamination      -0.2580082  0.25387820 -1.016268 3.154617e-01
xEducation        -0.8709401  0.18302860 -4.758492 2.430605e-05
xCatholic          0.1041153  0.03525785  2.952969 5.190079e-03
xInfant.Mortality  1.0770481  0.38171965  2.821568 7.335715e-03
```

5.3 Projections

The vector of fitted values is given by

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

and the vector of residuals is given by

$$\mathbf{e} = \mathbf{y} - \hat{\mathbf{y}} = (\mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}')\mathbf{y}.$$

Thus, multiplication by the matrix $\mathbf{H} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$ takes any vector in $\mathbb{R}^n$ and produces the fitted values. Note $\mathbf{H}$ is referred to as the 'hat matrix' since it transforms $\mathbf{y}$ into $\hat{\mathbf{y}}$. Similarly, multiplication by $(\mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}')$ produces the residuals. Notice that since the $\hat{\mathbf{y}}$ vector is a linear combination of the $\mathbf{X}$, it is orthogonal to the residuals, i.e.

$$\hat{\mathbf{y}}'\mathbf{e} = \mathbf{y}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'(\mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}')\mathbf{y} = 0.$$

It is useful to think of least squares in the terms of projections. Consider the column space of the design matrix, $\Gamma = \{\mathbf{X}\boldsymbol{\beta} \mid \boldsymbol{\beta} \in \mathbb{R}^p\}$. This p dimensional subspace lies in $\mathbb{R}^n$. Consider the vector $\mathbf{y}$ which lives in $\mathbb{R}^n$. Multiplication by the matrix $\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$ projects $\mathbf{y}$ into Γ. That is,

$$\mathbf{y} \to \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$$

is the linear projection map between $\mathbb{R}^n$ and Γ. The point $\hat{\mathbf{y}}$ is the point in Γ that is closest to $\mathbf{y}$ and $\hat{\boldsymbol{\beta}}$ is the specific linear combination of the columns of $\mathbf{X}$ that yields $\hat{\mathbf{y}}$. $\mathbf{e}$ is the vector connecting $\mathbf{y}$ and $\hat{\mathbf{y}}$, and it is orthogonal to all elements in Γ.

Thinking this helps us interpret statistical aspects of least squares. First, if $\mathbf{W}$ is any $p \times p$ invertible matrix, then the fitted values, $\hat{\mathbf{y}}$ will be the same for the design matrix $\mathbf{XW}$. This is because the spaces

$$\{\mathbf{X}\boldsymbol{\beta} \mid \boldsymbol{\beta} \in \mathbb{R}^p\}$$

and

$$\{\mathbf{X}\mathbf{W}\boldsymbol{\gamma} \mid \boldsymbol{\gamma} \in \mathbb{R}^p\}$$

are the same, since if $\mathbf{a} = \mathbf{X}\boldsymbol{\beta}$ then $\mathbf{a} = \mathbf{X}\boldsymbol{\gamma}$ via the relationship $\gamma = \mathbf{W}\boldsymbol{\beta}$ and thus any element of the first space is in the second. The same argument implies in the other direction, thus the two spaces are the same.

Therefore, any linear reorganization of the columns of $\mathbf{X}$ results in the same column space and thus the same fitted values. Furthermore, any addition of redundant columns to $\mathbf{X}$ adds nothing to the column space, and thus it's clear what the fit should be in the event that $\mathbf{X}$ is not full rank. Any full rank subset of the columns of $\mathbf{X}$ defines the same column and thus the same fitted values.

5.4 Another approach

In this section we generate yet another derivation of least squares. For vectors $\mathbf{a}$ (outcome) and $\mathbf{b}$ (predictor), let us define the coefficient function as:

$$c(\mathbf{a}, \mathbf{b}) = \frac{\langle \mathbf{a}, \mathbf{b} \rangle}{||\mathbf{b}||^2}.$$

and the residual function as

$$e(\mathbf{a}, \mathbf{b}) = \mathbf{a} - c(\mathbf{a}, \mathbf{b})\mathbf{b}$$

We argue that the least squares estimate of outcome $\mathbf{y}$ for predictor matrix $\mathbf{X} = [\mathbf{x}_1 \dots \mathbf{x}_p]$ can be obtained by taking successive residuals, in the following sense. Consider the least squares equation holding $\beta_2, \dots, \beta_p$ fixed, i.e.

$$||\mathbf{y} - \mathbf{x}_1\boldsymbol{\beta}_1 - \dots - \mathbf{x}_p\boldsymbol{\beta}_p||^2. \tag{12}$$

This is greater than or equal to if we replace β_1 by it's estimate and keep with the remainder fixed. That estimate being:

$$c(\mathbf{y} - \mathbf{x}_2\boldsymbol{\beta}_2 - \dots, \mathbf{x}_p\boldsymbol{\beta}_p, \mathbf{x}_1).$$

Plugging that back into the least squares equation we get

$$(12) \geq ||e(\mathbf{y}, \mathbf{x}_1) - e(\mathbf{x}_2, \mathbf{x}_1)\beta_2, \dots, e(\mathbf{x}_p, \mathbf{x}_1)\beta_p||^2.$$

Thus we have a new least squares equation with all residuals having "removed" $\mathbf{x}_1$ from all other regressors and the outcome. Then we can repeat this process again holding $\beta_3, \dots, \beta_p$ fixed and obtain

$$(13) \geq ||e\{e(\mathbf{y}, \mathbf{x}_1), e(\mathbf{x}_2, \mathbf{x}_1)\} - e\{e(\mathbf{x}_3, \mathbf{x}_1), e(\mathbf{x}_2, \mathbf{x}_1)\}\beta_3, \dots, e\{e(\mathbf{x}_p, \mathbf{x}_1), e(\mathbf{x}_2, \mathbf{x}_1)\}\beta_p||^2.$$

This could then be iterated to the p^{th} regressor. Moreover, because we know the same inequalities will be obtained no matter what order we get to the p^{th} regressor we can conclude that the order of taking residuals doesn't matter. Furthermore, picking the p^{th} coefficient was arbitrary as well, so the same conclusion applies to all regressors: the least squares estimate for all coefficients can be obtained by iteratively taking residuals with all of the other regressors (in any order).

This is interesting for many reasons. First, it is interesting to note that one need only regression through the origin to develop full multivariable regression. Secondly it helps us interpret our regression coefficients and how they are "adjust" for the other variables.

There was nothing particular about using vectors. Let $\mathbf{X} = [\mathbf{X}_1 \ \mathbf{X}_2]$ be two submatrices of size p_1 and p_2, and $\boldsymbol{\beta} = (\boldsymbol{\beta}_1' \ \boldsymbol{\beta}_2')'$ and consider minimizing:

$$||\mathbf{y} - \mathbf{X}_1\boldsymbol{\beta}_1 - \mathbf{X}_2\boldsymbol{\beta}_2||^2.$$

If $\boldsymbol{\beta}_2$ were held fixed, this would be maximized at

$$\hat{\boldsymbol{\beta}}_1(\boldsymbol{\beta}_2) = (\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1'(\mathbf{y} - \mathbf{X}_2\boldsymbol{\beta}_2).$$

Plugging that back in we obtain a smaller quantity

$$||\{\mathbf{I} - \mathbf{X}_1(\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1'\}\mathbf{y} - \{\mathbf{I} - \mathbf{X}_1(\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1'\}\mathbf{X}_2\boldsymbol{\beta}_2||^2$$

This is equivalent to the residual of $\mathbf{y}$ having regressed out $\mathbf{X}_1$ and the residual matrix of $\mathbf{X}_2$ having regressed $\mathbf{X}_1$ out of every column. Thus out β_2 estimate will be the regression matrix of these residuals. Again, this explains why $\boldsymbol{\beta}_2$'s estimate has been adjusted for $\mathbf{X}_1$, both the outcome and the $\mathbf{X}_2$ predictors have been orthogonalized to the space spanned by the columns of $\mathbf{X}_1$!

5.5 Full row rank case

In the case where $\mathbf{X}$ is $n \times n$ of full rank, then the columns of $\mathbf{X}$ form a basis for $\mathbb{R}^n$. In this case, $\hat{\mathbf{y}} = \mathbf{y}$, since $\mathbf{y}$ lives in the space spanned by the columns of $\mathbf{X}$. All the linear model accomplishes is a lossless linear reorganization of $\mathbf{y}$. This is perhaps surprisingly useful, especially when the columns of $\mathbf{X}$ are orthonormal ($\mathbf{X}'\mathbf{X} = \mathbf{I}$). In this case, the function that takes the outcome vector and converts it to the coefficients is called a "transform". The most well known versions of transforms are Fourier and wavelet.