STAT 540: Data Analysis and Regression

Wen Zhou

http://www.stat.colostate.edu/~riczw/
Email: riczw@stat.colostate.edu

Department of Statistics
Colorado State University

Fall 2015

Contents

Multiple Linear Regression I

- Multiple linear regression model
  1. Multiple linear regression model in matrix terms
  2. Estimation of regression coefficients
- Inference
  1. ANOVA results
  2. Inference about regression parameters
  3. Estimation of mean response and prediction of new observation
- Inference about regression parameters
- Estimation and prediction
- Geometric interpretation of linear model and regression
- Estimating estimable function of regression or linear coefficient β

Multiple Linear Regression

- Example: # of predictor variables = 2.

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \epsilon_i, \quad \epsilon_i \sim \text{iid } N(0, \sigma^2),$$

  for $i = 1, \ldots, n$.

- Response surface:

$$\mathbb{E}(Y_i) =$$

- Example:
  - $Y =$ Pine bark beetle density
  - $X_1 =$ Temperature
  - $X_2 =$ Tree species

Interpretation of Coefficients

- β_0: Intercept. When the model scope includes $X_1 = X_2 = 0$.
  - β_0 is interpreted as the mean response $E(Y)$ at $X_1 = X_2 = 0$.
- β_j: Slope in the direction of X_j (effect).
  - $\partial \mathbb{E}(Y)/\partial X_j =$
  - $\mathbb{E}_{Y|X=(X_1,X_2)}(Y) - \mathbb{E}_{Y|X=(X_1',X_2)}(Y) =$

- Interpreted as the change in the mean response $\mathbb{E}(Y)$ per unit increase in X_j, **when X_{-j} are held constant.**
- What if X_j is qualitative?

Multiple Linear Regression

- A "general" linear regression model is, for $i = 1, \cdots, n$

$$Y_i = \beta_0 + \sum_{j=1}^{p} X_{ij}\beta_j + \epsilon_i, \ \ \epsilon_i \sim \text{iid } N(0, \sigma^2).$$

- Response surface:

$$\mathbb{E}(Y_i) = \beta_0 + \sum_{j=1}^{p} X_{ij}\beta_j$$

- Regression coefficients: $\beta_0, \beta_1, \ldots, \beta_{p-1}, \beta_p$.
- Predictor variables: $X_1, \ldots, X_p$ are known constants/values.
- The model is linear in the parameters, not necessarily in the shape of the response surface.

Response Surface Examples

- Polynomial regression

$$\mathbb{E}(Y) = \beta_0 + \beta_1 X + \beta_2 X^2 + \beta_3 X^3.$$

- Transformed variables

$$\mathbb{E}(\log(Y)) = \beta_0 + \beta_1 X_1 + \beta_2 \sqrt{X_2}.$$

- Interaction effects

$$\mathbb{E}(Y) = \beta_0 + \beta_1 X_1 + \beta_2 \sqrt{X_2} + \beta_3 X_1 X_2.$$

  - The change in the mean response corresponding to a unit change in X_1 depends on X_2 and vice versa.
  - Testing whether $\beta_3 = 0$ or not is very challenging in high-dimensional $(n = o(p))$.

Qualitative Predictor Variables

- Example: Let $Y =$ length of hospital stay, $X_1 =$ age, and $X_2 =$ gender: 0 for male and 1 for female.
  - An additive model is
  - Thus the response surface for males is

    and for females is

- β_2 is _____
- This kind of model sometimes is called ANVOCA model.

Qualitative Predictor Variables

- Interaction: the relationship between X_1 and Y for a fixed value of $X_2 = x_2$ depends on x_2.
- An interaction model is

- Thus the response surface for males is

  and for females is

Notation

- n observations, 1 response variable, $p-1$ β's with predictors (i.e. β_0 is the pth).
- Response variable: $\boldsymbol{Y}_{n\times 1} = (Y_1, Y_2, \ldots, Y_n)^T$.
- The predictors are arranged in the design matrix

$$\boldsymbol{X}_{n\times p} = \begin{bmatrix} 1 & X_{11} & X_{12} & \cdots & X_{1,p-1} \\ 1 & X_{21} & X_{22} & \cdots & X_{2,p-1} \\ \vdots & \vdots & \vdots & & \\ 1 & X_{n1} & X_{n2} & \cdots & X_{n,p-1} \end{bmatrix}$$

- Random error: $\boldsymbol{\epsilon}_{n\times 1} = (\epsilon_1, \epsilon_2, \ldots, \epsilon_n)^T$.
- Regression coefficients: $\boldsymbol{\beta}_{p\times 1} = (\beta_0, \beta_1, \ldots, \beta_{p-1})^T$.

Multiple Linear Regression Model in Matrix Terms

- The multiple linear regression model can be written as

  where as we have seen before

$$\mathbb{E}(\boldsymbol{\epsilon}) = \mathbf{0}_{n\times 1}, \quad \text{Var}\{\boldsymbol{\epsilon}\} = \sigma^2 \boldsymbol{I}_{n\times n}.$$

- Thus,

  and

$$\boldsymbol{Y} \sim$$

Least Squares Estimation

- Consider the criterion:

$$Q = \sum_{i=1}^{n}(Y_i - \beta_0 - \sum_{j=1}^{p-1}\beta_j X_{ij})^2 =$$

- The least squares estimate of $\boldsymbol{\beta}$ is

assuming that $\boldsymbol{X}^T\boldsymbol{X}$ is invertible.

  - This is also the MLE.
  - What condition on $\boldsymbol{X}$ do we need to have $\boldsymbol{X}^T\boldsymbol{X}$ invertible?

  - What if $\boldsymbol{X}^T\boldsymbol{X}$ is not invertible?

Fitted Values and Residuals

- Fitted values: $\hat{\boldsymbol{Y}} =$

  where the hat matrix is

- Residuals: $\boldsymbol{e} =$

Sums of Squares

- We have sums of squares in matrix forms that

$$SSR = \sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2 =$$

$$SSE = \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 =$$

$$SSTO = \sum_{i=1}^{n}(Y_i - \bar{Y})^2 =$$

- Partitioning of total sum of squares and particularly the df are

$$\underbrace{SSTO}_{df=n-1} = \underbrace{SSR}_{df=p-1} + \underbrace{SSE}_{df=n-p} .$$

Mean Squares

- Define mean squares

$$MSR = \frac{SSR}{p-1}, \quad MSE = \frac{SSE}{n-p}.$$

- It can be shown that $\mathbb{E}(MSE) =$
- Also can be shown that

$$\mathbb{E}(MSR) \quad \begin{cases} = \sigma^2 & \text{if } \beta_j = 0 \text{ for } \forall\ j \\ > \sigma^2 & \text{otherwise} \end{cases}.$$

ANOVA Table

The ANOVA table is

Source	SS	df	MS	F
Regression	SSR		MSR	$F = MSR/MSE$
Error	SSE		MSE	
Total	$SSTO$			

- If _____ then

$$\mathbb{E}(MSE) = \mathbb{E}(MSR) = \sigma^2$$

in which case $MSR/MSE \approx 1$.

Overall F Test for Regression Relation

- Test

$$H_0 : \underline{\qquad\qquad\qquad\qquad} \text{ v.s. } H_a : \underline{\qquad\qquad\qquad\qquad}.$$

  - It can be shown that under H_0,

$$F^* = \frac{MSR}{MSE} \sim$$

  - Thus we can perform an F-test at level α by the decision rule:

- Conditional on H_0 being rejected, we may want to find

$$S = \{j \mid \beta_j \neq 0\}$$

(or a.s.)– Identification/Selection.

Coefficient of Multiple Determination, R^2

- The coefficient of multiple determination is denoted by R^2 and is defined as

$$R^2 = \frac{SSR}{SSTO} = 1 - \frac{SSE}{SSTO}$$

- Interpretation: The proportion of variation in the Y_i's explained by the regression relation.

More on R^2

- As more predictors are added to the model ($p \uparrow$), R^2 must increase. Why?
  - Recall
    $$SSTO = SSR + SSE$$
    $SSTO$ is fixed for $\boldsymbol{Y}$ while SSE is a minimum of the unconstraint convex optimization problem $\boldsymbol{\beta} = \arg\min SSE(\beta_0, \ldots, \beta_{p-1})$.
  - Suppose we consider an extra predictor and thus consider $SSE(\beta_0, \ldots, \beta_p)$. The $\boldsymbol{\beta}$ that minimizes this SSE cannot be inferior to the previous minimizer because $\beta_p = 0$ is a special case within the new minimization problem that incorporates the previous one.

Adjusted R^2

- R^2 depends on p (even for $p \ll n$), how to remove that dependence?
- The *adjusted* coefficient of multiple determination is denoted by R_a^2 and is defined as

$$R_a^2 = 1 - \frac{SSE/n-p}{SSTO/n-1} = 1 - \left(\frac{n-1}{n-p}\right)\frac{SSE}{SSTO}.$$

- The adjusted coefficient of multiple determination R_a^2 may decrease when more predictors are in the model.
- Many other statistics such as AIC, BIC, Mallow's C_p, etc. will be discussed and they are superior over R_a^2.

Estimation of Regression Coefficients

- Mean satisfies

$$\mathbb{E}(\hat{\boldsymbol{\beta}}) = \boldsymbol{\beta}.$$

  That is, the LS estimate $\hat{\boldsymbol{\beta}}$ is an unbiased estimate of $\boldsymbol{\beta}$.

- Variance-covariance matrix:

$$\boldsymbol{\Sigma}_{\boldsymbol{\beta}} := \text{Var}\{\hat{\boldsymbol{\beta}}\} = \sigma^2 \left(\boldsymbol{X}^T \boldsymbol{X}\right)^{-1}.$$

  - $(\boldsymbol{\Sigma}_{\boldsymbol{\beta}})_{kk} =$

  - $(\boldsymbol{\Sigma}_{\boldsymbol{\beta}})_{kl} =$

Inference about Regression Coefficients

- The estimated variance-covariance matrix.

$$\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\beta}} := s^2\{\hat{\boldsymbol{\beta}}\} = MSE \cdot \left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}$$

$$:= \begin{bmatrix} s^2\{\hat{\beta}_0\} & s\{\hat{\beta}_0, \hat{\beta}_1\} & \cdots & s\{\hat{\beta}_0, \hat{\beta}_{p-1}\} \\ s\{\hat{\beta}_1, \hat{\beta}_0\} & s^2\{\hat{\beta}_1\} & \cdots & s\{\hat{\beta}_1, \hat{\beta}_{p-1}\} \\ \vdots & \vdots & \vdots & \\ s\{\hat{\beta}_{p-1}, \hat{\beta}_0\} & s\{\hat{\beta}_{p-1}, \hat{\beta}_1\} & \cdots & s^2\{\hat{\beta}_{p-1}\} \end{bmatrix}$$

- Under the multiple linear regression model, we have

$$\frac{\hat{\beta}_k - \beta_k}{s\{\hat{\beta}_k\}} \sim$$

for $k = 0, 1, \ldots, p-1$.

Inference about Regression Coefficients

- Thus the $(1-\alpha)$ confidence interval for β_k is

$$\hat{\beta}_k \pm t_{1-\alpha/2;n-p} s\{\hat{\beta}_k\}.$$

- Test $H_0 : \beta_k = \beta_{k0}$ versus $H_a : \beta_k \neq \beta_{k0}$.
- Under H_0, we have

$$t^* = \frac{\hat{\beta}_k - \beta_{k0}}{s\{\hat{\beta}_k\}} \sim t_{n-p}$$

- Thus we can perform a t-test at level α by the decision rule:

Estimation of Mean Response–Hidden Extrapolation

- Define $\boldsymbol{X}_h = (1, X_{h1}, \ldots, X_{h,p-1})^T$.
- Caution about hidden extrapolations.
  - The region (with respect to $\boldsymbol{X}_0$) defined by

$$d(\boldsymbol{X}_0) = \boldsymbol{X}_0^T (\boldsymbol{X}^T \boldsymbol{X})^{-1} \boldsymbol{X}_0 \leq h_{max}$$

    where $h_{max} = \max_i h_{ii}$, is an ellipsoid enclosing all data points inside the "regressor variable hull" (RVH).
  - Predictions for any $\boldsymbol{X}_0$ outside the RVH (i.e., $d(\boldsymbol{X}_0) > h_{max}$) is hidden extrapolation, at least to some degree.

Estimation of Mean Response

- The estimated mean response corresponding to $\boldsymbol{X}_h =$
- Mean $\mathbb{E}(\hat{Y}_h) =$
- Variance $\text{Var}\{\hat{Y}_h\} =$
- Estimated variance is $s^2\{\hat{Y}_h\} =$

Confidence Intervals for Mean Response

- The $(1-\alpha)$ confidence interval for $\mathbb{E}(Y_h)$ is

$$\hat{Y}_h \pm t_{1-\alpha/2;n-p} s\{\hat{Y}_h\}$$

- The Working-Hotelling $(1-\alpha)$ confidence band for the regression surface is

$$\hat{Y}_h \pm W s\{\hat{Y}_h\}$$

  where $W^2 = pF(1-\alpha; p, n-p)$.

- The Bonferroni $(1-\alpha)$ joint confidence intervals for g mean responses are

$$\hat{Y}_h \pm B s\{\hat{Y}_h\}$$

  where $B = t_{1-\alpha/(2g);n-p}$.

Prediction of New Observation

- The predicted new observation corresponding to $\boldsymbol{X}_h$ is $\hat{Y}_h = \boldsymbol{X}_h^T \hat{\boldsymbol{\beta}}$, and
  - Mean $\mathbb{E}(\hat{Y}_h) = \boldsymbol{X}_h^T \boldsymbol{\beta} = \mathbb{E}(Y_{h(\text{new})})$.
  - Prediction error variance

    $\sigma^2_{\text{pred}} = \text{Var}(\hat{Y}_h - Y_{h(\text{new})}) =$

  - Estimated prediction error variance is

    $s^2\{\text{pred}\} =$

Prediction Intervals for New Observation

- The $(1-\alpha)$ prediction interval for $Y_{h(\text{new})}$ is

$$\hat{Y}_h \pm t_{1-\alpha/2;n-p} s\{\text{pred}\}$$

- The Scheffé $(1-\alpha)$ joint confidence intervals for g new observations are

$$\hat{Y}_h \pm S s\{\text{pred}\}$$

where $S^2 = gF(1-\alpha; g, n-p)$.

- The Bonferroni $(1-\alpha)$ joint confidence intervals for g new observations are

$$\hat{Y}_h \pm B s\{\text{pred}\}$$

where $B = t_{1-\alpha/(2g);n-p}$.

Geometric Viewpoint: The Column Space of the Design Matrix

- $\boldsymbol{X}\boldsymbol{\beta}$ is a linear combination of the columns of $\boldsymbol{X}$

$$\boldsymbol{X}\boldsymbol{\beta} = [\boldsymbol{x}_1, \ldots, \boldsymbol{x}_p] \begin{bmatrix} \beta_1 \\ \vdots \\ \beta_p \end{bmatrix} = \beta_1 \boldsymbol{x}_1 + \ldots + \beta_p \boldsymbol{x}_p$$

- The set of all possible linear combinations of the columns of $\boldsymbol{X}$ is called the column space of $\boldsymbol{X}$ and is denoted by

$$\mathcal{C}(X) = \{\boldsymbol{X}\boldsymbol{a} : \boldsymbol{a} \in \mathbb{R}^p\}$$

- The Gauss-Markov linear model says $\boldsymbol{y}$ is a random vector whose mean is in the column space of $\boldsymbol{X}$ and whose variance is $\sigma^2\boldsymbol{I}$ for some positive real number σ^2, i.e.

$$\mathbb{E}(\boldsymbol{y}) \in \mathcal{C}(\boldsymbol{X}) \quad \text{and} \quad \text{Var}(\boldsymbol{y}) = \sigma^2\boldsymbol{I}, \ \sigma^2 \in \mathbb{R}^+$$

An Example Column Space

$$\boldsymbol{X} = \begin{bmatrix} 1 \\ 1 \end{bmatrix} \Rightarrow \mathcal{C}(\boldsymbol{X}) = \{\boldsymbol{X}\boldsymbol{a} : \boldsymbol{a} \in \mathbb{R}^p\}$$

$$= \left\{ \begin{bmatrix} 1 \\ 1 \end{bmatrix} [a_1] : a_1 \in \mathbb{R} \right\}$$

$$= \left\{ a_1 \begin{bmatrix} 1 \\ 1 \end{bmatrix} : a_1 \in \mathbb{R} \right\}$$

$$= \left\{ \begin{bmatrix} a_1 \\ a_1 \end{bmatrix} : a_1 \in \mathbb{R} \right\}$$

Another Example Column Space

$$\boldsymbol{X} = \begin{bmatrix} 1 & 0 \\ 1 & 0 \\ 0 & 1 \\ 0 & 1 \end{bmatrix} \Rightarrow \mathcal{C}(\boldsymbol{X}) = \left\{ \begin{bmatrix} 1 & 0 \\ 1 & 0 \\ 0 & 1 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} a_1 \\ a_2 \end{bmatrix} : \boldsymbol{a} \in \mathbb{R}^2 \right\}$$

$$= \left\{ a_1 \begin{bmatrix} 1 \\ 1 \\ 0 \\ 0 \end{bmatrix} + a_2 \begin{bmatrix} 0 \\ 0 \\ 1 \\ 1 \end{bmatrix} : a_1, a_2 \in \mathbb{R} \right\}$$

$$= \left\{ \begin{bmatrix} a_1 \\ a_1 \\ 0 \\ 0 \end{bmatrix} + \begin{bmatrix} 0 \\ 0 \\ a_2 \\ a_2 \end{bmatrix} : a_1, a_2 \in \mathbb{R} \right\}$$

$$= \left\{ \begin{bmatrix} a_1 \\ a_1 \\ a_2 \\ a_2 \end{bmatrix} : a_1, a_2 \in \mathbb{R} \right\}$$

Another Example Column Space

$$\boldsymbol{X}_1 = \begin{bmatrix} 1 & 0 \\ 1 & 0 \\ 0 & 1 \\ 0 & 1 \end{bmatrix}, \quad \boldsymbol{X}_2 = \begin{bmatrix} 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \\ 1 & 0 & 1 \end{bmatrix}$$

$$\begin{aligned} \boldsymbol{x} \in \mathcal{C}(\boldsymbol{X}_1) &\Rightarrow \boldsymbol{x} = \boldsymbol{X}_1 \boldsymbol{a} \text{ for some } \boldsymbol{a} \in \mathbb{R}^2 \\ &\Rightarrow \boldsymbol{x} = \boldsymbol{X}_2 \begin{bmatrix} 0 \\ \boldsymbol{a} \end{bmatrix} \text{ for some } \boldsymbol{a} \in \mathbb{R}^2 \\ &\Rightarrow \boldsymbol{x} = \boldsymbol{X}_2 \boldsymbol{b} \text{ for some } \boldsymbol{b} \in \mathbb{R}^3 \\ &\Rightarrow \boldsymbol{x} \in \mathcal{C}(\boldsymbol{X}_2) \end{aligned}$$

Thus

$$\mathcal{C}(\boldsymbol{X}_1) \subset \mathcal{C}(\boldsymbol{X}_2)$$

Another Example Column Space (continued)

$$\boldsymbol{x} \in \mathcal{C}(\boldsymbol{X}_2) \Rightarrow \boldsymbol{x} = \boldsymbol{X}_2\boldsymbol{a} \text{ for some } \boldsymbol{a} \in \mathbb{R}^3$$

$$\Rightarrow \boldsymbol{x} = a_1 \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \end{bmatrix} + a_2 \begin{bmatrix} 1 \\ 1 \\ 0 \\ 0 \end{bmatrix} + a_3 \begin{bmatrix} 0 \\ 0 \\ 1 \\ 1 \end{bmatrix} \text{ for some } \boldsymbol{a} \in \mathbb{R}^3$$

$$\Rightarrow \boldsymbol{x} = \begin{bmatrix} a_1 + a_2 \\ a_1 + a_2 \\ a_1 + a_3 \\ a_1 + a_3 \end{bmatrix} \text{ for some } a_1, a_2, a_3 \in \mathbb{R}$$

$$\Rightarrow \boldsymbol{x} = \boldsymbol{X}_1 \begin{bmatrix} a_1 + a_2 \\ a_1 + a_3 \end{bmatrix} \text{ for some } a_1, a_2, a_3 \in \mathbb{R}$$

Another Example Column Space (continued)

$$\Rightarrow \boldsymbol{x} = \boldsymbol{X}_1 \begin{bmatrix} a_1 + a_2 \\ a_1 + a_3 \end{bmatrix} \text{ for some } a_1, a_2, a_3 \in \mathbb{R}$$

$$\Rightarrow \boldsymbol{x} = \boldsymbol{X}_1 \boldsymbol{b} \text{ for some } \boldsymbol{b} \in \mathbb{R}^2$$

$$\Rightarrow \boldsymbol{x} \in \mathcal{C}(\boldsymbol{X}_1)$$

Thus, $\mathcal{C}(\boldsymbol{X}_2) \subset \mathcal{C}(\boldsymbol{X}_1)$, as we have shown $\mathcal{C}(\boldsymbol{X}_1) \subset \mathcal{C}(\boldsymbol{X}_2)$. It follows that $\mathcal{C}(\boldsymbol{X}_1) = \mathcal{C}(\boldsymbol{X}_2)$.

Estimation of $\mathbb{E}(\boldsymbol{y})$

- A fundamental goal of linear model analysis is to estimate $\mathbb{E}(\boldsymbol{y})$
- We could, of course, use $\boldsymbol{y}$ to estimate $\mathbb{E}(\boldsymbol{y})$
- $\boldsymbol{y}$ is obviously an unbiased estimator of $\mathbb{E}(\boldsymbol{y})$, but it is often not a very sensible estimator.
- For example, suppose

$$\begin{bmatrix} y_1 \\ y_2 \end{bmatrix} = \begin{bmatrix} 1 \\ 1 \end{bmatrix} \mu + \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \end{bmatrix}, \quad \text{and we observe } \boldsymbol{y} = [6.1, 2.3]'$$

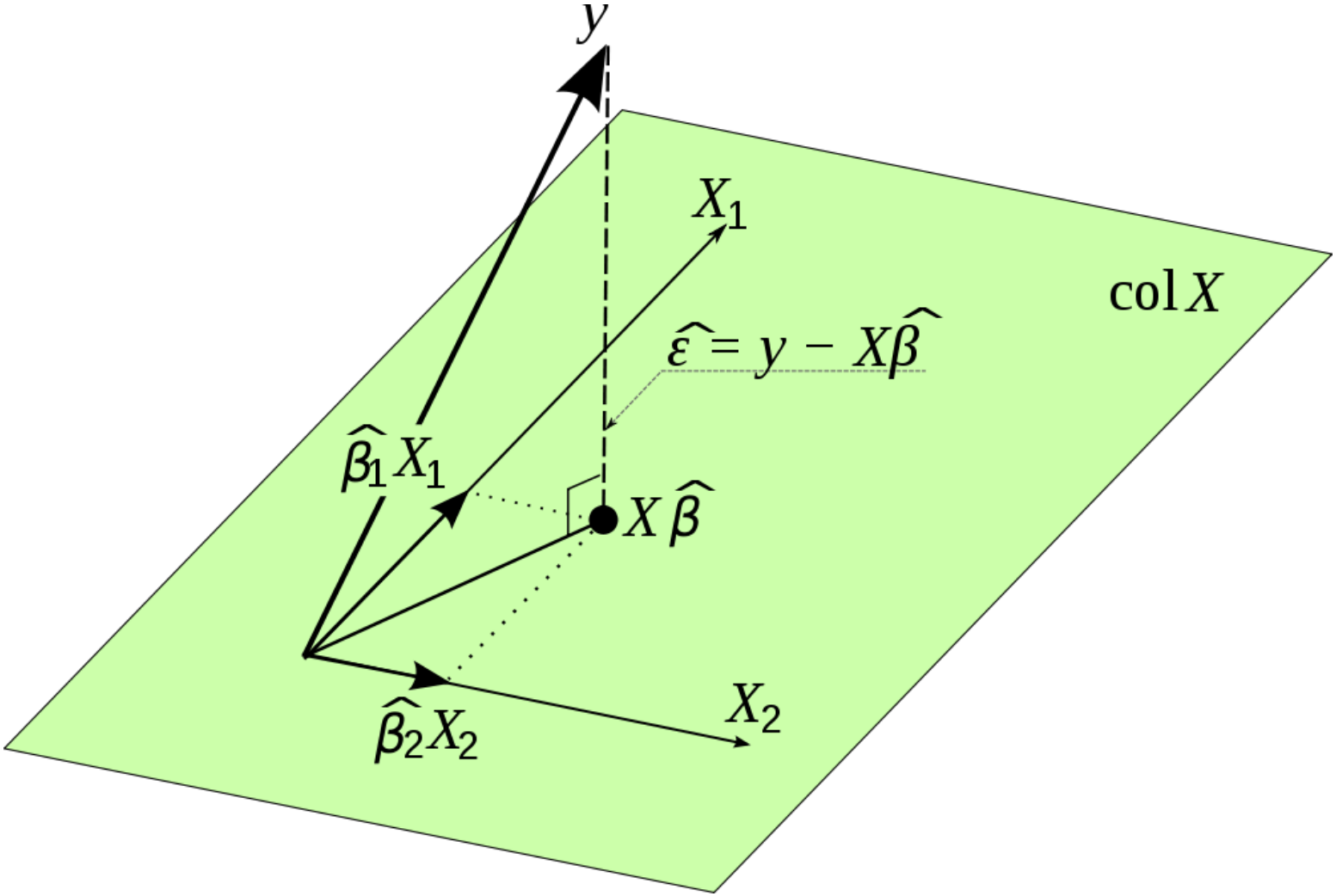
Should we estimate $\mathbb{E}(\boldsymbol{y}) = [\mu, \mu]'$ by $\boldsymbol{y} = [6.1, 2.3]'$?

Estimation of $\mathbb{E}(\boldsymbol{y})$

- The Gauss-Markov linear models says that $\mathbb{E}(\boldsymbol{y}) \in \mathcal{C}(\boldsymbol{X})$, so we should use that information when estimating $\mathbb{E}(\boldsymbol{y})$
- Consider estimating $\mathbb{E}(\boldsymbol{y})$ by the point in $\mathcal{C}(\boldsymbol{X})$ that is closest to y (as measured by the usual Euclidean distance).
- This unique point is called the orthogonal projection of $\boldsymbol{y}$ onto $\mathcal{C}(\boldsymbol{X})$ and denoted by $\hat{\boldsymbol{y}}$ (although it could be argued that $\widehat{\mathbb{E}(\boldsymbol{y})}$ might be better notation).
- By definition, $||\boldsymbol{y} - \hat{\boldsymbol{y}}|| = \min_{\boldsymbol{z} \in \mathcal{C}(\boldsymbol{X})} ||\boldsymbol{y} - \boldsymbol{z}||$ where $||\boldsymbol{a}|| = \sqrt{\sum_{i=1}^{n} a_i^2}$

Geometric Viewpoint on Multiple Regression (and LM)

- Geometrically, how to minimize the distance between $\boldsymbol{Y}$ and $\mathcal{C}(\boldsymbol{X})$?
  - That point is _____
  - The vector between $\boldsymbol{Y}$ and $\boldsymbol{X}\widehat{\boldsymbol{\beta}}$ is _____, and the distance is _____
- For R^2: if we add another predictor, $\mathcal{C}(\boldsymbol{X})$ gains 1 more dimension, so $||\boldsymbol{e}||$ can only decrease. $\mathcal{C}(\boldsymbol{X})$
  - Note: if $\dim(\mathcal{S}) = n$ then _____


y
X_1
$\mathrm{col}\, X$
$\widehat{\varepsilon} = y - X\widehat{\beta}$
$\widehat{\beta}_1 X_1$
$X\widehat{\beta}$
X_2
$\widehat{\beta}_2 X_2$

Orthogonal Projection Matrices

It can be shown that, as we did for least square estimators

- $\forall \boldsymbol{y} \in \mathbb{R}^n$, $\hat{\boldsymbol{y}} = \mathbb{P}_{\boldsymbol{X}} \boldsymbol{y}$ is the optimal one, i.e.

  $\hat{\boldsymbol{y}} = \mathbb{P}_{\boldsymbol{X}} \boldsymbol{y}$ is the best estimator of $\mathbb{E}(\boldsymbol{y})$ in the class of linear unbiased estimators

  for the unique matrix $\mathbb{P}_{\boldsymbol{X}} = \boldsymbol{H}$, the hat matrix, and is called **orthogonal projection matrix**
- $\boldsymbol{H}\boldsymbol{H} = \boldsymbol{H}$, idempotent
- $\boldsymbol{H} = \boldsymbol{H}'$, symmetric
- $\boldsymbol{H}\boldsymbol{X} = \boldsymbol{X}$ and $\boldsymbol{X}'\boldsymbol{H} = \boldsymbol{X}'$ (Why? Intuitively...)
- If $(\boldsymbol{X}'\boldsymbol{X})$ is not invertible, we use its generalized inverse $(\boldsymbol{X}'\boldsymbol{X})^-$ where $\boldsymbol{A}\boldsymbol{A}^-\boldsymbol{A} = \boldsymbol{A}$.
- The $\boldsymbol{H}$ is invariant to the choice of $(\boldsymbol{X}'\boldsymbol{X})^-$, which is itself not unique
- $\hat{\boldsymbol{y}}$ and $\boldsymbol{y} - \hat{\boldsymbol{y}}$ are orthogonal (Why?)

An Example Orthogonal Projection

Suppose $\begin{bmatrix} y_1 \\ y_2 \end{bmatrix} = \begin{bmatrix} 1 \\ 1 \end{bmatrix} \mu + \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \end{bmatrix}$, and we observe $y = [6.1, 2.3]$. Then

$$\begin{aligned}
\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}' &= \begin{bmatrix} 1 \\ 1 \end{bmatrix} \left(\begin{bmatrix} 1 \\ 1 \end{bmatrix}' \begin{bmatrix} 1 \\ 1 \end{bmatrix} \right)^{-1} \begin{bmatrix} 1 \\ 1 \end{bmatrix}' \\
&= \begin{bmatrix} 1 \\ 1 \end{bmatrix} [2]^{-1} [1 \quad 1] \\
&= \begin{bmatrix} 1 \\ 1 \end{bmatrix} [\frac{1}{2}][1 \quad 1] \\
&= \frac{1}{2} \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix} \\
&= \begin{bmatrix} 1/2 & 1/2 \\ 1/2 & 1/2 \end{bmatrix}
\end{aligned}$$

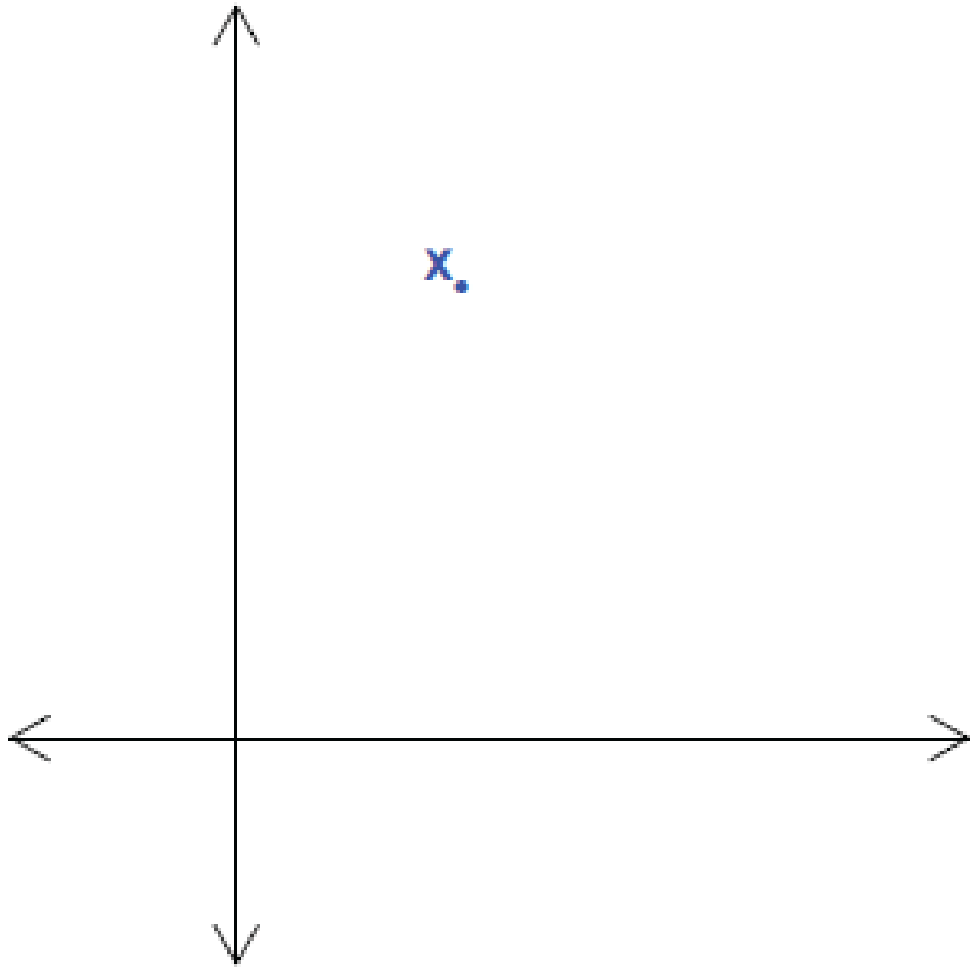
An Example Orthogonal Projection

Thus, the orthogonal projection of $y = [6.1, 2.3]$ onto the column space of $\boldsymbol{X} = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$ is

$$\mathbb{P}_{\boldsymbol{X}}\boldsymbol{y} = \boldsymbol{H}\boldsymbol{y} = \begin{bmatrix} 1/2 & 1/2 \\ 1/2 & 1/2 \end{bmatrix} \begin{bmatrix} 6.1 \\ 2.3 \end{bmatrix} = \begin{bmatrix} 4.2 \\ 4.2 \end{bmatrix}$$

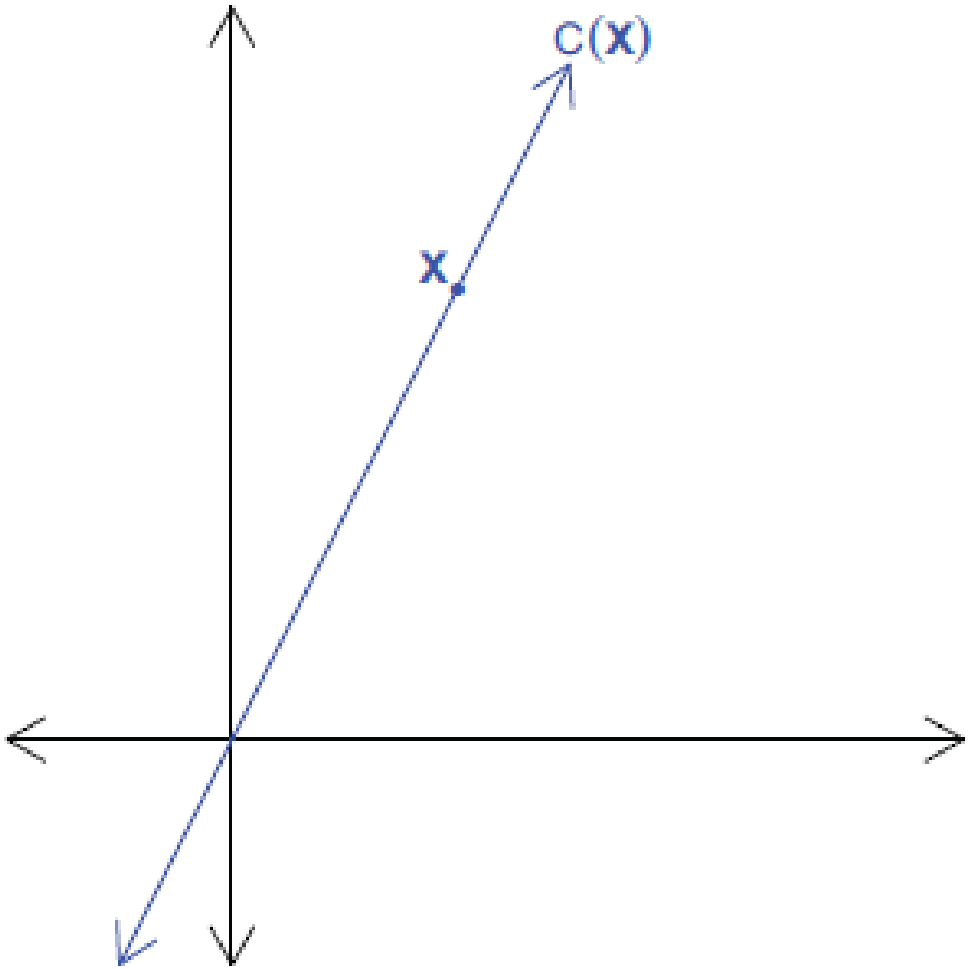
Geometric illustration

Suppose $\boldsymbol{X} = \begin{bmatrix} 1 \\ 2 \end{bmatrix}$ and $\boldsymbol{y} = \begin{bmatrix} 2 \\ 3/4 \end{bmatrix}$



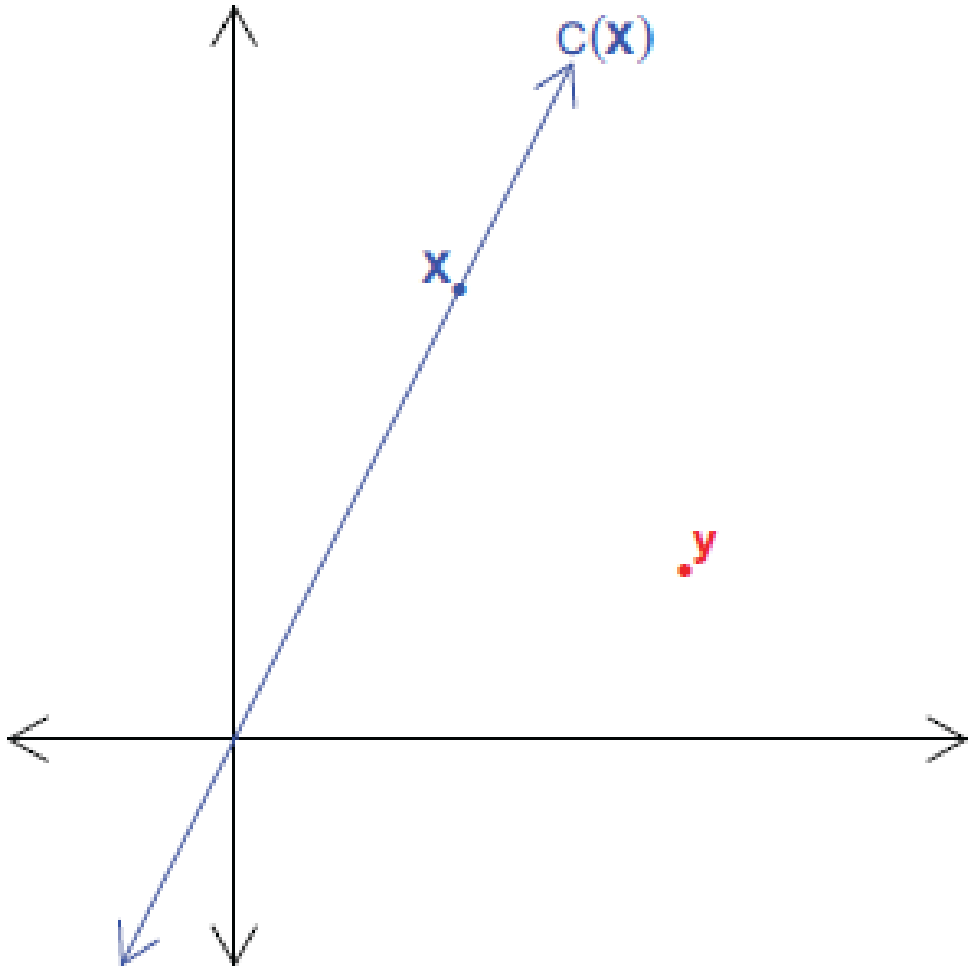
Geometric illustration

Suppose $\boldsymbol{X} = \begin{bmatrix} 1 \\ 2 \end{bmatrix}$ and $\boldsymbol{y} = \begin{bmatrix} 2 \\ 3/4 \end{bmatrix}$



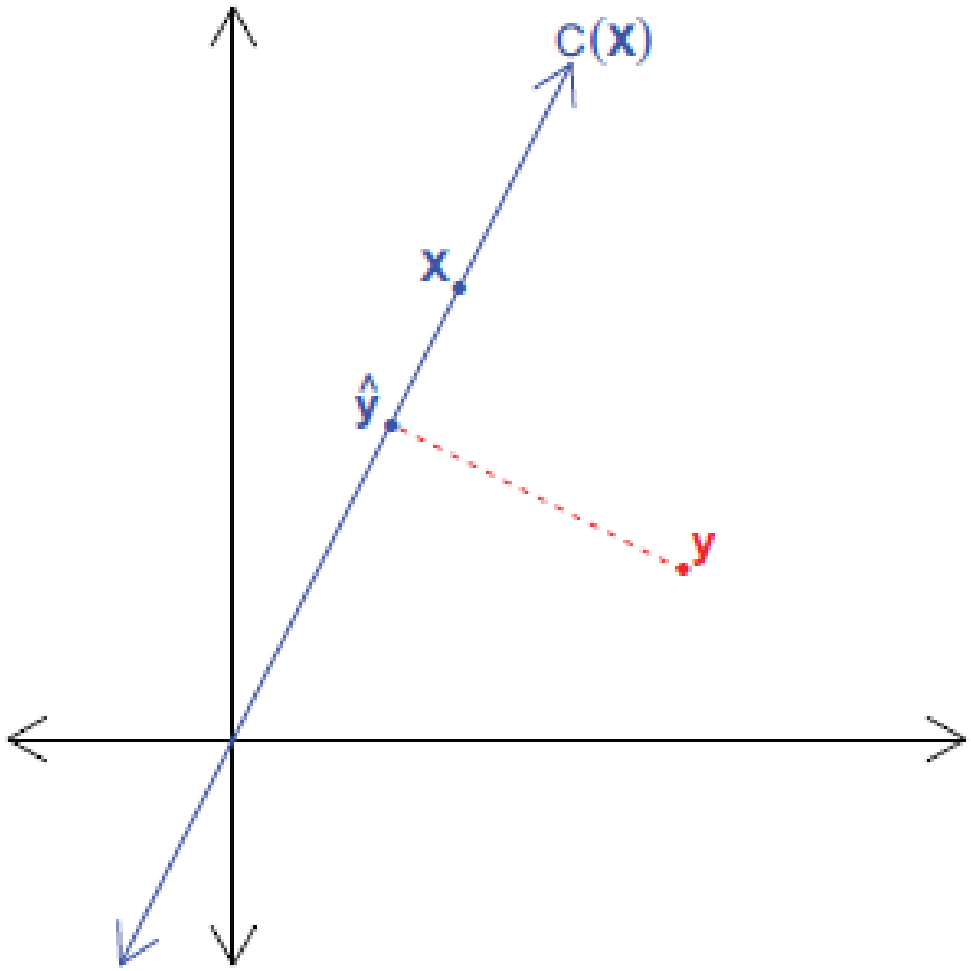
Geometric illustration

Suppose $\boldsymbol{X} = \begin{bmatrix} 1 \\ 2 \end{bmatrix}$ and $\boldsymbol{y} = \begin{bmatrix} 2 \\ 3/4 \end{bmatrix}$



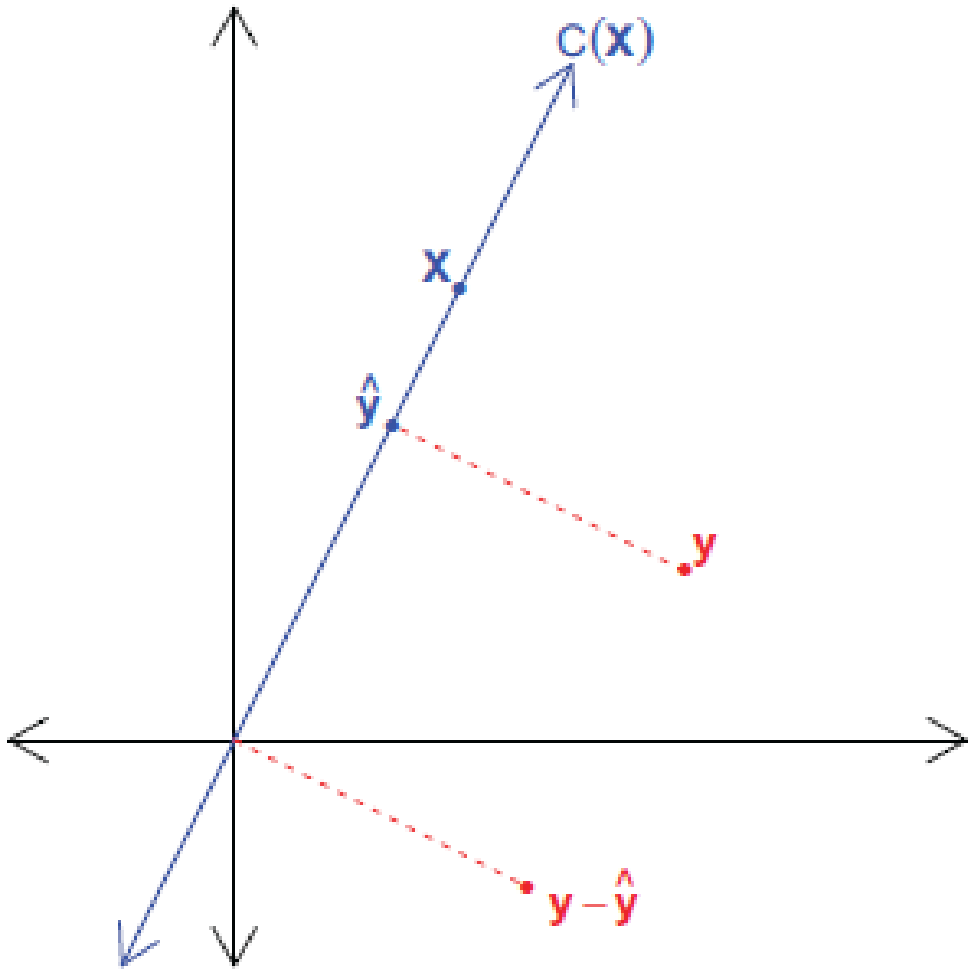
Geometric illustration

Suppose $\boldsymbol{X} = \begin{bmatrix} 1 \\ 2 \end{bmatrix}$ and $\boldsymbol{y} = \begin{bmatrix} 2 \\ 3/4 \end{bmatrix}$



Geometric illustration

The angle between $\hat{y}$ and residual $y - \hat{y}$ is 90. So, "orthogonal projection".



What if $\boldsymbol{X}$ is not full column rank?

- $\boldsymbol{X}^T\boldsymbol{X}$ is not invertible, then $(\boldsymbol{X}^T\boldsymbol{X})^{-1}$ has to be defined based on the generalized inverse matrix.
- If $\boldsymbol{X}$ is not of full column rank, then there are infinitely many vectors in the set $\{\boldsymbol{b} : \boldsymbol{X}\boldsymbol{b} = \boldsymbol{X}\boldsymbol{\beta}\}$ for any fixed value of $\boldsymbol{\beta}$.
- Thus, no matter what the value of $\mathbb{E}(\boldsymbol{y})$, there will be infinitely many vectors $\boldsymbol{b}$ such that $\boldsymbol{X}\boldsymbol{b} = \mathbb{E}(\boldsymbol{y})$ when $\boldsymbol{X}$ is not of full column rank.
- Our response vector $\boldsymbol{y}$ can help us learn about $\mathbb{E}(\boldsymbol{y}) = \boldsymbol{X}\boldsymbol{\beta}$, but when $\boldsymbol{X}$ is NOT of full column rank, there is NO hope of learning about $\boldsymbol{\beta}$ alone unless additional information about $\boldsymbol{\beta}$ is available.
- How, we could estimate **estimable** function of $\boldsymbol{\beta}$

Treatment Effects Model

Researchers randomly assigned a total of six experimental units to two treatments and measured a response of interest.

$$y_{ij} = \mu + \tau_i + \epsilon_{ij}, \ i = 1, 2; \ j = 1, 2, 3$$

$$\begin{bmatrix} y_{11} \\ y_{12} \\ y_{13} \\ y_{21} \\ y_{22} \\ y_{23} \end{bmatrix} = \begin{bmatrix} \mu + \tau_1 \\ \mu + \tau_1 \\ \mu + \tau_1 \\ \mu + \tau_2 \\ \mu + \tau_2 \\ \mu + \tau_2 \end{bmatrix} + \begin{bmatrix} \epsilon_{11} \\ \epsilon_{12} \\ \epsilon_{13} \\ \epsilon_{21} \\ \epsilon_{22} \\ \epsilon_{23} \end{bmatrix}$$

Question: what is $\boldsymbol{X}$, $\boldsymbol{\beta}$?

Treatment Effects Model (continued)

- In this case, it makes no sense to estimate $\boldsymbol{\beta} = [\mu, \tau_1, \tau_2]'$ because there are multiple (infinitely many, in fact) choices of $\boldsymbol{\beta}$ that define the same mean for $\boldsymbol{y}$.
- For example

$$\begin{bmatrix} \mu \\ \tau_1 \\ \tau_2 \end{bmatrix} = \begin{bmatrix} 5 \\ -1 \\ 1 \end{bmatrix}, \begin{bmatrix} 0 \\ 4 \\ 6 \end{bmatrix}, \begin{bmatrix} 999 \\ -995 \\ -993 \end{bmatrix}$$

  all yield same $\boldsymbol{X}\boldsymbol{\beta} = \mathbb{E}(\boldsymbol{y})$.
- When multiple values for $\boldsymbol{\beta}$ define the same $\mathbb{E}(\boldsymbol{y})$, we say that $\boldsymbol{\beta}$ is *non-estimable*.

Estimable Functions of $\boldsymbol{\beta}$

- A linear function of $\boldsymbol{\beta}$, $\boldsymbol{C\beta}$, is said to be **estimable** if there is a linear function of $\boldsymbol{y}$, say $\boldsymbol{Ay}$, that is an unbiased estimator for $\boldsymbol{C\beta}$. Otherwise, nonexistence of such linear function implies that $\boldsymbol{C\beta}$ is *non-estimable*.
- Note that $\boldsymbol{Ay}$ is an unbiased estimator of $\boldsymbol{C\beta}$ if and only if

$$\mathbb{E}(\boldsymbol{Ay}) = \boldsymbol{C\beta}, \text{ for } \forall\ \boldsymbol{\beta} \in \mathbb{R}^p \Leftrightarrow \boldsymbol{AX\beta} = \boldsymbol{C\beta} \Leftrightarrow \boldsymbol{AX} = \boldsymbol{C}$$

- This says that we can estimate $\boldsymbol{C\beta}$ as long as $\boldsymbol{C\beta} = \boldsymbol{AX\beta} = \boldsymbol{A}\mathbb{E}(\boldsymbol{y})$ for some $\boldsymbol{A}$, i.e. as long as $\boldsymbol{C\beta}$ is a linear function of $\mathbb{E}(\boldsymbol{y})$
- The bottom line is that we can always estimate $\mathbb{E}(\boldsymbol{y})$ and all linear functions of $\mathbb{E}(\boldsymbol{y})$; all other linear functions of $\boldsymbol{\beta}$ are non-estimable

Treatment Effects Model (continued)

$$\boldsymbol{X}\boldsymbol{\beta} = \begin{bmatrix} 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \\ 1 & 0 & 1 \\ 1 & 0 & 1 \\ 1 & 0 & 1 \end{bmatrix} \begin{bmatrix} \mu \\ \tau_1 \\ \tau_2 \end{bmatrix} = \begin{bmatrix} \mu + \tau_1 \\ \mu + \tau_1 \\ \mu + \tau_1 \\ \mu + \tau_2 \\ \mu + \tau_2 \\ \mu + \tau_2 \end{bmatrix}$$

so that

$$[1, 0, 0, 0, 0, 0]\boldsymbol{X}\boldsymbol{\beta} = [1, 1, 0]\boldsymbol{\beta} = \mu + \tau_1$$

$$[0, 0, 0, 1, 0, 0]\boldsymbol{X}\boldsymbol{\beta} = [1, 0, 1]\boldsymbol{\beta} = \mu + \tau_2$$

$$[1, 0, 0, -1, 0, 0]\boldsymbol{X}\boldsymbol{\beta} = [0, 1, 1]\boldsymbol{\beta} = \tau_1 - \tau_2$$

are estimable functions of $\boldsymbol{\beta}$

Estimating Estimable Functions of $\boldsymbol{\beta}$

- If $\boldsymbol{C\beta}$ is estimable, then there exists a matrix $\boldsymbol{A}$ such that $\boldsymbol{C = AX}$ and $\boldsymbol{C\beta = AX\beta = A\mathbb{E}(y)}$ for any $\boldsymbol{\beta} \in \mathbb{R}^p$
- It makes sens to estimate $\boldsymbol{C\beta}$ by

$$\begin{aligned}\widehat{\boldsymbol{A\mathbb{E}(y)}} &= \boldsymbol{A\hat{y} = AP_X y = AX(X'X)^- X'y} \\ &= \boldsymbol{AX(X'X)^- X'X\hat{\beta} = AP_X X\hat{\beta} = AX\hat{\beta} = C\beta}\end{aligned}$$

- $\boldsymbol{C\hat{\beta}}$ is called an Ordinary Least Squares (OLS) estimator of $\boldsymbol{C\beta}$
- Note that although the "hat" is on $\boldsymbol{\beta}$, it is $\boldsymbol{C\beta}$ that we are estimating
- Invariance of $\boldsymbol{C\hat{\beta}}$ to the choice of $\hat{\boldsymbol{\beta}}$: Although there are infinitely many solutions to the normal equations when $\boldsymbol{X}$ is not of full column rank, $\boldsymbol{C\hat{\beta}}$ is the same for all normal equation solutions $\hat{\boldsymbol{\beta}}$ whenever $\boldsymbol{C\beta}$ is estimable (STAT 640)

Treatment Effects Model (continued)

- Suppose our aim is to estimate $\tau_1 - \tau_2$
- As noted before

$$\boldsymbol{X}\boldsymbol{\beta} = \begin{bmatrix} 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \\ 1 & 0 & 1 \\ 1 & 0 & 1 \\ 1 & 0 & 1 \end{bmatrix} \begin{bmatrix} \mu \\ \tau_1 \\ \tau_2 \end{bmatrix} = \begin{bmatrix} \mu + \tau_1 \\ \mu + \tau_1 \\ \mu + \tau_1 \\ \mu + \tau_2 \\ \mu + \tau_2 \\ \mu + \tau_2 \end{bmatrix}, \text{ so that}$$

$$[1, 0, 0, -1, 0, 0]\boldsymbol{X}\boldsymbol{\beta} = [0, 1, 1]\boldsymbol{\beta} = \tau_1 - \tau_2$$

- Thus, we can compute the OLS estimator of $\tau_1 - \tau_2$ as

$$[1, 0, 0, -1, 0, 0]\hat{\boldsymbol{y}} = [0, 1, 1]\hat{\boldsymbol{\beta}}$$

where $\hat{\boldsymbol{\beta}}$ is any solution to the normal equations.

Treatment Effects Model (continued)

The normal equation in this case is

$$\begin{bmatrix} 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \\ 1 & 0 & 1 \\ 1 & 0 & 1 \\ 1 & 0 & 1 \end{bmatrix}' \begin{bmatrix} 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \\ 1 & 0 & 1 \\ 1 & 0 & 1 \\ 1 & 0 & 1 \end{bmatrix} \begin{bmatrix} b_1 \\ b_2 \\ b_3 \end{bmatrix} = \begin{bmatrix} 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \\ 1 & 0 & 1 \\ 1 & 0 & 1 \\ 1 & 0 & 1 \end{bmatrix}' \begin{bmatrix} y_{11} \\ y_{12} \\ y_{13} \\ y_{21} \\ y_{22} \\ y_{23} \end{bmatrix}$$

so that

$$\begin{bmatrix} 6 & 3 & 3 \\ 3 & 3 & 0 \\ 3 & 0 & 3 \end{bmatrix} \begin{bmatrix} b_1 \\ b_2 \\ b_3 \end{bmatrix} = \begin{bmatrix} y_{..} \\ y_{1.} \\ y_{2.} \end{bmatrix}$$

Treatment Effects Model (continued)

$\hat{\boldsymbol{\beta}}_1 = \begin{bmatrix} \bar{y}_{..} \\ \bar{y}_{1.} - \bar{y}_{..} \\ \bar{y}_{2.} - \bar{y}_{..} \end{bmatrix}$ and $\hat{\boldsymbol{\beta}}_2 = \begin{bmatrix} 0 \\ \bar{y}_{1.} \\ \bar{y}_{2.} \end{bmatrix}$ are both solutions to the normal equation (Check this).

Thus, the OLS estimator of $\boldsymbol{C\beta} = [0, 1, -1]\boldsymbol{\beta} = \tau_1 - \tau_2$ is

$$\boldsymbol{C}\hat{\boldsymbol{\beta}}_1 = [0, 1, -1] \begin{bmatrix} \bar{y}_{..} \\ \bar{y}_{1.} - \bar{y}_{..} \\ \bar{y}_{2.} - \bar{y}_{..} \end{bmatrix} = \bar{y}_{1.} - \bar{y}_{2.} = [0, 1, -1] \begin{bmatrix} 0 \\ \bar{y}_{1.} \\ \bar{y}_{2.} \end{bmatrix} = \boldsymbol{C}\hat{\boldsymbol{\beta}}_2$$

HW: Can you find two different generalized inverse of $(\boldsymbol{X}'\boldsymbol{X})$, $\boldsymbol{A}_1$ and $\boldsymbol{A}_2$ that $(\boldsymbol{X}'\boldsymbol{X})\boldsymbol{A}_i(\boldsymbol{X}'\boldsymbol{X}) = (\boldsymbol{X}'\boldsymbol{X})$ so that $\boldsymbol{A}_i = (\boldsymbol{X}'\boldsymbol{X})^-$ for each i, and they will give you $\hat{\boldsymbol{\beta}}_1$ and $\hat{\boldsymbol{\beta}}_2$, respectively?

The Gauss-Markov Theorem

Under the Gauss-Markov Linear Model, the OLS estimator $\boldsymbol{c}'\hat{\boldsymbol{\beta}}$ of an estimable linear function $\boldsymbol{c}'\boldsymbol{\beta}$ is the unique Best Linear Unbiased Estimator (BLUE) in the sense that $\text{Var}(\boldsymbol{c}'\hat{\boldsymbol{\beta}})$ is strictly less than the variance of any other linear unbiased estimator of $\boldsymbol{c}'\boldsymbol{\beta}$ for all $\boldsymbol{\beta} \in \mathbb{R}^p$ and all $\sigma^2 \in \mathbb{R}^+$.

- The Gauss-Markov Theorem says that if we want to estimate an estimable linear function $\boldsymbol{c}'\boldsymbol{\beta}$ using a linear estimator that is unbiased, we should always use the OLS estimator.
- In our simple example of the treatment effects model, we could have used $y_{11}y_{21}$ to estimate $\tau_1\tau_2$. It is easy to see that $y_{11}y_{21}$ is a linear estimator that is unbiased for $\tau_1\tau_2$, but its variance is clearly larger than the variance of the OLS estimator $\bar{y}_{1\cdot}\bar{y}_{2\cdot}$ (as guaranteed by the Gauss-Markov Theorem).