12. Algebra of Least Squares

Fall 2023

Matthew Blackwell

Gov 2002 (Harvard)

Where are we? Where are we going?

- We saw how the population linear projection works.

Where are we? Where are we going?

- We saw how the population linear projection works.
- How can we estimate the parameters of the linear projection or CEF?

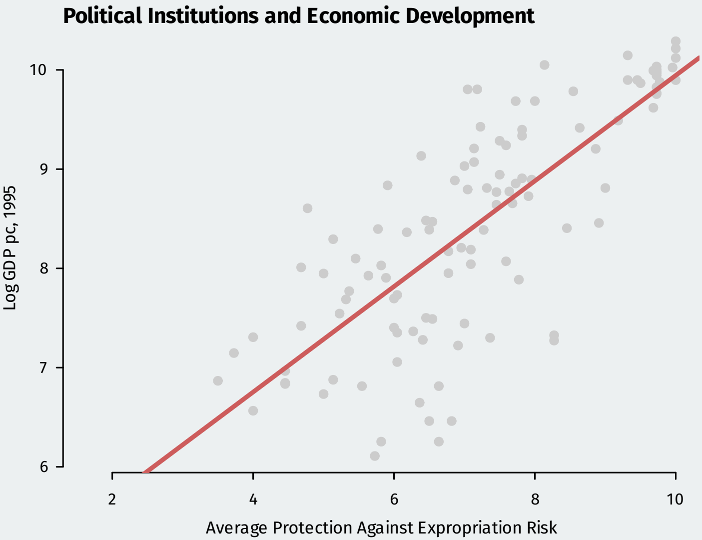
Where are we? Where are we going?

- We saw how the population linear projection works.
- How can we estimate the parameters of the linear projection or CEF?
- Now: least squares estimator and its algebraic properties.

Where are we? Where are we going?

- We saw how the population linear projection works.
- How can we estimate the parameters of the linear projection or CEF?
- Now: least squares estimator and its algebraic properties.
- After that: the statistical properties of least squares.

Acemoglu, Johnson, and Robinson (2001)



1/ Deriving the OLS estimator

Samples vs population

Assumption

The variables $\{(Y_1, \mathbf{X}_1), \ldots, (Y_i, \mathbf{X}_i), \ldots, (Y_n, \mathbf{X}_n)\}$ are i.i.d. draws from a common distribution F.

- F is the **population distribution** or **DGP**.

Samples vs population

Assumption

The variables $\{(Y_1, \mathbf{X}_1), \ldots, (Y_i, \mathbf{X}_i), \ldots, (Y_n, \mathbf{X}_n)\}$ are i.i.d. draws from a common distribution F.

- F is the **population distribution** or **DGP**.
  - Without i subscripts, $(Y, \mathbf{X})$ are r.v.s and draws from F.

Samples vs population

Assumption

The variables $\{(Y_1, \mathbf{X}_1), \dots, (Y_i, \mathbf{X}_i), \dots, (Y_n, \mathbf{X}_n)\}$ are i.i.d. draws from a common distribution F.

- F is the **population distribution** or **DGP**.
  - Without i subscripts, $(Y, \mathbf{X})$ are r.v.s and draws from F.
- $\{(Y_i, \mathbf{X}_i) : i = 1, \dots, n\}$ is the **sample** and can be seen in two ways:

Samples vs population

Assumption

The variables $\{(Y_1, \mathbf{X}_1), \dots, (Y_i, \mathbf{X}_i), \dots, (Y_n, \mathbf{X}_n)\}$ are i.i.d. draws from a common distribution F.

- F is the **population distribution** or **DGP**.
  - Without i subscripts, $(Y, \mathbf{X})$ are r.v.s and draws from F.
- $\{(Y_i, \mathbf{X}_i) : i = 1, \dots, n\}$ is the **sample** and can be seen in two ways:
  - Numbers in your data matrix, fixed to the analyst.

Samples vs population

Assumption

The variables $\{(Y_1, \mathbf{X}_1), \ldots, (Y_i, \mathbf{X}_i), \ldots, (Y_n, \mathbf{X}_n)\}$ are i.i.d. draws from a common distribution F.

- F is the **population distribution** or **DGP**.
  - Without i subscripts, $(Y, \mathbf{X})$ are r.v.s and draws from F.
- $\{(Y_i, \mathbf{X}_i) : i = 1, \ldots, n\}$ is the **sample** and can be seen in two ways:
  - Numbers in your data matrix, fixed to the analyst.
  - From a statistical POV, they are realizations of a random process.

Samples vs population

Assumption

The variables $\{(Y_1, \mathbf{X}_1), \ldots, (Y_i, \mathbf{X}_i), \ldots, (Y_n, \mathbf{X}_n)\}$ are i.i.d. draws from a common distribution F.

- F is the **population distribution** or **DGP**.
    - Without i subscripts, $(Y, \mathbf{X})$ are r.v.s and draws from F.
- $\{(Y_i, \mathbf{X}_i) : i = 1, \ldots, n\}$ is the **sample** and can be seen in two ways:
    - Numbers in your data matrix, fixed to the analyst.
    - From a statistical POV, they are realizations of a random process.
- Violations include time-series data and clustered sampling.

Samples vs population

Assumption

The variables $\{(Y_1, \mathbf{X}_1), \ldots, (Y_i, \mathbf{X}_i), \ldots, (Y_n, \mathbf{X}_n)\}$ are i.i.d. draws from a common distribution F.

- F is the **population distribution** or **DGP**.
  - Without i subscripts, $(Y, \mathbf{X})$ are r.v.s and draws from F.
- $\{(Y_i, \mathbf{X}_i) : i = 1, \ldots, n\}$ is the **sample** and can be seen in two ways:
  - Numbers in your data matrix, fixed to the analyst.
  - From a statistical POV, they are realizations of a random process.
- Violations include time-series data and clustered sampling.
  - Weakening i.i.d. usually complicates notation but can be done.

Quantity of interest

- Population linear projection model:

$$Y = \mathbf{X}'\boldsymbol{\beta} + e$$

Quantity of interest

- Population linear projection model:

$$Y = \mathbf{X}'\boldsymbol{\beta} + e$$

- Here $\boldsymbol{\beta}$ minimizes the **population** expected squared error:

$$\boldsymbol{\beta} = \arg\min_{\mathbf{b}\in\mathbb{R}^k} S(\mathbf{b}), \qquad S(\mathbf{b}) = \mathbb{E}\left[(Y - \mathbf{X}'\mathbf{b})^2\right]$$

Quantity of interest

- Population linear projection model:

$$Y = \mathbf{X}'\boldsymbol{\beta} + e$$

- Here $\boldsymbol{\beta}$ minimizes the **population** expected squared error:

$$\boldsymbol{\beta} = \underset{\mathbf{b} \in \mathbb{R}^k}{\arg\min}\, S(\mathbf{b}), \qquad S(\mathbf{b}) = \mathbb{E}\left[(Y - \mathbf{X}'\mathbf{b})^2\right]$$

- Last time we saw that this can be written:

$$\boldsymbol{\beta} = \left(\mathbb{E}[\mathbf{X}\mathbf{X}']\right)^{-1} \mathbb{E}[\mathbf{X}Y]$$

Quantity of interest

- Population linear projection model:

$$Y = \mathbf{X}'\boldsymbol{\beta} + e$$

- Here $\boldsymbol{\beta}$ minimizes the **population** expected squared error:

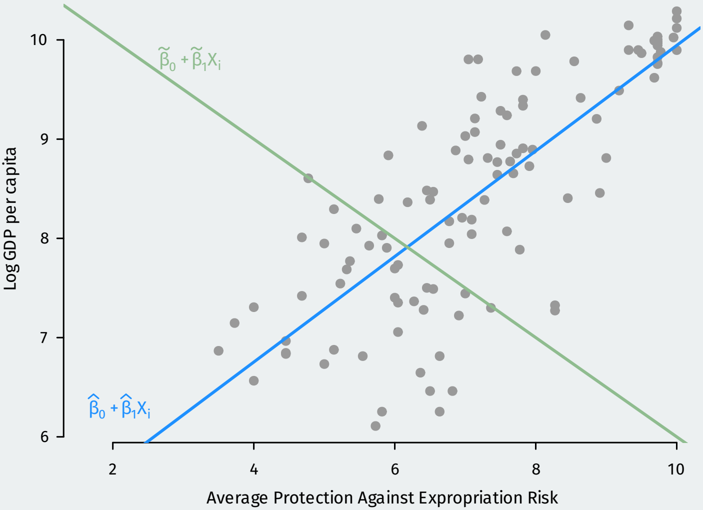
$$\boldsymbol{\beta} = \underset{\mathbf{b}\in\mathbb{R}^k}{\arg\min}\, S(\mathbf{b}), \qquad S(\mathbf{b}) = \mathbb{E}\left[(Y - \mathbf{X}'\mathbf{b})^2\right]$$

- Last time we saw that this can be written:

$$\boldsymbol{\beta} = \left(\mathbb{E}[\mathbf{X}\mathbf{X}']\right)^{-1}\mathbb{E}[\mathbf{X}Y]$$

- How do we estimate $\boldsymbol{\beta}$?

Which line is better?



Plug-in principle returns!

- **Plug-in estimator**: solve the sample version of the population goal.

Plug-in principle returns!

- **Plug-in estimator**: solve the sample version of the population goal.
- Replace projection errors with observed errors, or **residuals**: $Y_i - \mathbf{X}_i'\mathbf{b}$

Plug-in principle returns!

- **Plug-in estimator**: solve the sample version of the population goal.
- Replace projection errors with observed errors, or **residuals**: $Y_i - \mathbf{X}_i'\mathbf{b}$
  - **Sum of squared residuals,** $SSR(\mathbf{b}) = \sum_{i=1}^{n}(Y_i - \mathbf{X}_i'\mathbf{b})^2$.

Plug-in principle returns!

- **Plug-in estimator**: solve the sample version of the population goal.
- Replace projection errors with observed errors, or **residuals**: $Y_i - \mathbf{X}_i'\mathbf{b}$
    - **Sum of squared residuals,** $SSR(\mathbf{b}) = \sum_{i=1}^{n}(Y_i - \mathbf{X}_i'\mathbf{b})^2$.
    - Total prediction error using **b** as our estimated coefficient.

Plug-in principle returns!

- **Plug-in estimator**: solve the sample version of the population goal.
- Replace projection errors with observed errors, or **residuals**: $Y_i - \mathbf{X}_i'\mathbf{b}$
    - **Sum of squared residuals,** $SSR(\mathbf{b}) = \sum_{i=1}^{n}(Y_i - \mathbf{X}_i'\mathbf{b})^2$.
    - Total prediction error using **b** as our estimated coefficient.
- We can use these residuals to get a sample average prediction error:

$$\hat{S}(\mathbf{b}) = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \mathbf{X}_i'\mathbf{b})^2 = \frac{1}{n}SSR(\mathbf{b})$$

Plug-in principle returns!

- **Plug-in estimator**: solve the sample version of the population goal.
- Replace projection errors with observed errors, or **residuals**: $Y_i - \mathbf{X}_i'\mathbf{b}$
  - **Sum of squared residuals**, $SSR(\mathbf{b}) = \sum_{i=1}^{n}(Y_i - \mathbf{X}_i'\mathbf{b})^2$.
  - Total prediction error using **b** as our estimated coefficient.
- We can use these residuals to get a sample average prediction error:

$$\hat{S}(\mathbf{b}) = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \mathbf{X}_i'\mathbf{b})^2 = \frac{1}{n}SSR(\mathbf{b})$$

- $\hat{S}(\mathbf{b})$ is an estimator of the expected squared error, $S(\mathbf{b})$.

Least squares estimator

- **Ordinary least squares estimator** minimizes $\hat{S}$ in place of S.

$$\boldsymbol{\beta} = \underset{\mathbf{b}\in\mathbb{R}^k}{\arg\min}\, \mathbb{E}\left[(Y - \mathbf{X}'\mathbf{b})^2\right]$$

$$\hat{\boldsymbol{\beta}} = \underset{\mathbf{b}\in\mathbb{R}^k}{\arg\min}\, \frac{1}{n}\sum_{i=1}^{n}(Y_i - \mathbf{X}_i'\mathbf{b})^2$$

Least squares estimator

- **Ordinary least squares estimator** minimizes $\hat{S}$ in place of S.

$$\boldsymbol{\beta} = \underset{\mathbf{b}\in\mathbb{R}^k}{\arg\min}\, \mathbb{E}\left[(Y - \mathbf{X}'\mathbf{b})^2\right]$$

$$\hat{\boldsymbol{\beta}} = \underset{\mathbf{b}\in\mathbb{R}^k}{\arg\min}\, \frac{1}{n}\sum_{i=1}^{n}(Y_i - \mathbf{X}_i'\mathbf{b})^2$$

  - In words: find the coefficients that minimize the sum/average of the squared residuals.

Least squares estimator

- **Ordinary least squares estimator** minimizes $\hat{S}$ in place of S.

$$\boldsymbol{\beta} = \underset{\mathbf{b}\in\mathbb{R}^k}{\arg\min}\, \mathbb{E}\left[(Y - \mathbf{X}'\mathbf{b})^2\right]$$

$$\hat{\boldsymbol{\beta}} = \underset{\mathbf{b}\in\mathbb{R}^k}{\arg\min}\, \frac{1}{n}\sum_{i=1}^{n}(Y_i - \mathbf{X}_i'\mathbf{b})^2$$

  - In words: find the coefficients that minimize the sum/average of the squared residuals.

- After some calculus, we can write this as a plug-in estimator:

$$\hat{\boldsymbol{\beta}} = \left(\frac{1}{n}\sum_{i=1}^{n}\mathbf{X}_i\mathbf{X}_i'\right)^{-1}\left(\frac{1}{n}\sum_{i=1}^{n}\mathbf{X}_iY_i\right)$$

Least squares estimator

- **Ordinary least squares estimator** minimizes $\hat{S}$ in place of S.

$$\boldsymbol{\beta} = \underset{\mathbf{b}\in\mathbb{R}^k}{\arg\min}\, \mathbb{E}\left[(Y - \mathbf{X}'\mathbf{b})^2\right]$$

$$\hat{\boldsymbol{\beta}} = \underset{\mathbf{b}\in\mathbb{R}^k}{\arg\min}\, \frac{1}{n}\sum_{i=1}^{n}(Y_i - \mathbf{X}_i'\mathbf{b})^2$$

  - In words: find the coefficients that minimize the sum/average of the squared residuals.

- After some calculus, we can write this as a plug-in estimator:

$$\hat{\boldsymbol{\beta}} = \left(\frac{1}{n}\sum_{i=1}^{n}\mathbf{X}_i\mathbf{X}_i'\right)^{-1}\left(\frac{1}{n}\sum_{i=1}^{n}\mathbf{X}_iY_i\right)$$

  - $n^{-1}\sum_{i=1}^{n}\mathbf{X}_i\mathbf{X}_i'$ is the sample version of $\mathbb{E}[\mathbf{X}\mathbf{X}']$

Least squares estimator

- **Ordinary least squares estimator** minimizes $\hat{S}$ in place of S.

$$\boldsymbol{\beta} = \underset{\mathbf{b} \in \mathbb{R}^k}{\arg\min}\, \mathbb{E}\left[(Y - \mathbf{X}'\mathbf{b})^2\right]$$

$$\hat{\boldsymbol{\beta}} = \underset{\mathbf{b} \in \mathbb{R}^k}{\arg\min}\, \frac{1}{n} \sum_{i=1}^{n} (Y_i - \mathbf{X}_i'\mathbf{b})^2$$

  - In words: find the coefficients that minimize the sum/average of the squared residuals.

- After some calculus, we can write this as a plug-in estimator:

$$\hat{\boldsymbol{\beta}} = \left(\frac{1}{n} \sum_{i=1}^{n} \mathbf{X}_i \mathbf{X}_i'\right)^{-1} \left(\frac{1}{n} \sum_{i=1}^{n} \mathbf{X}_i Y_i\right)$$

  - $n^{-1} \sum_{i=1}^{n} \mathbf{X}_i \mathbf{X}_i'$ is the sample version of $\mathbb{E}[\mathbf{X}\mathbf{X}']$
  - $n^{-1} \sum_{i=1}^{n} \mathbf{X}_i Y_i$ is the sample version of $\mathbb{E}[\mathbf{X}Y]$

Bivariate regressions

- **Bivariate regression** is the linear projection model with $\mathbf{X} = (1, X)$:

$$Y = \beta_0 + X\beta_1 + e$$

Bivariate regressions

- **Bivariate regression** is the linear projection model with $\mathbf{X} = (1, X)$:

$$Y = \beta_0 + X\beta_1 + e$$

- Linear projection slope in the population from last times:

$$\beta_1 = \frac{\text{Cov}(X, Y)}{\mathbb{V}[X]}$$

Bivariate regressions

- **Bivariate regression** is the linear projection model with $\mathbf{X} = (1, X)$:

$$Y = \beta_0 + X\beta_1 + e$$

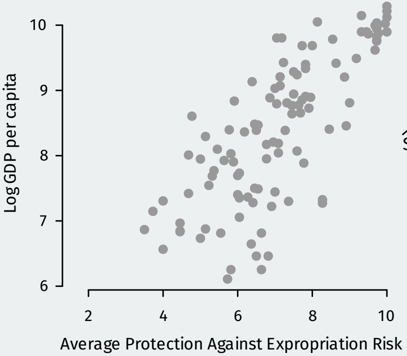
- Linear projection slope in the population from last times:

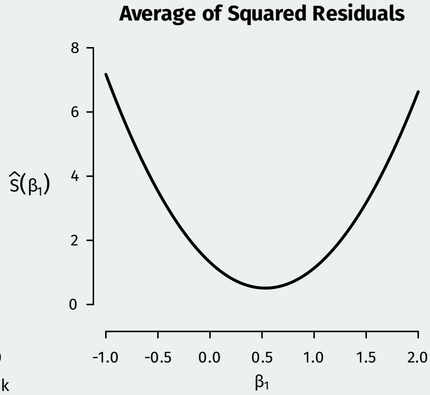
$$\beta_1 = \frac{\text{Cov}(X, Y)}{\mathbb{V}[X]}$$

- We can show the OLS estimator of the slope is:

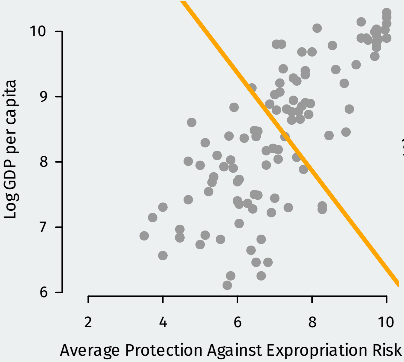
$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(Y_i - \overline{Y})(X_i - \overline{X})}{\sum_{i=1}^{n}(X_i - \overline{X})^2} = \frac{\widehat{\text{Cov}}(X, Y)}{\hat{\mathbb{V}}[X]}$$

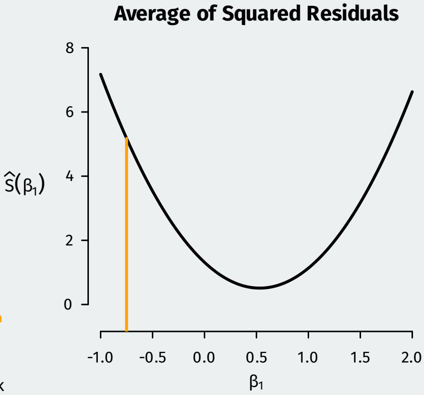
Visualizing OLS



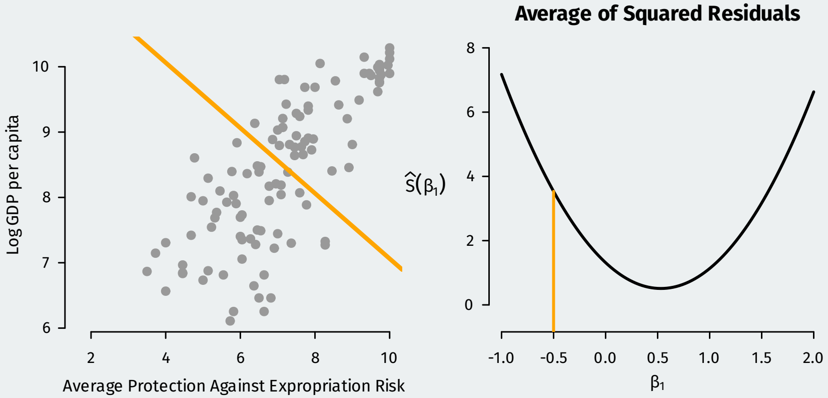


Visualizing OLS

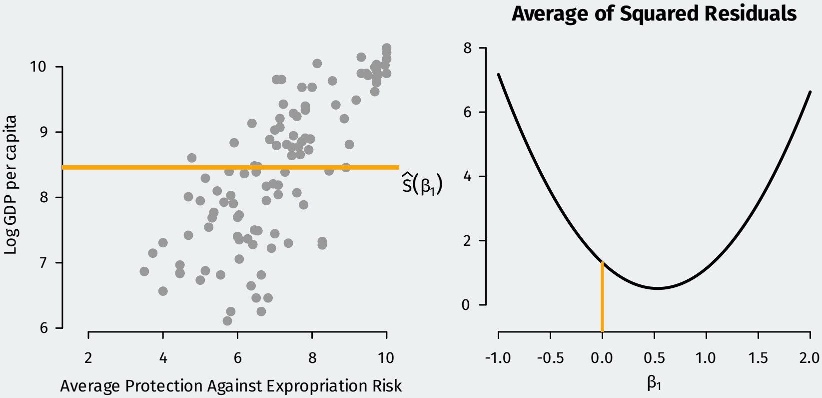




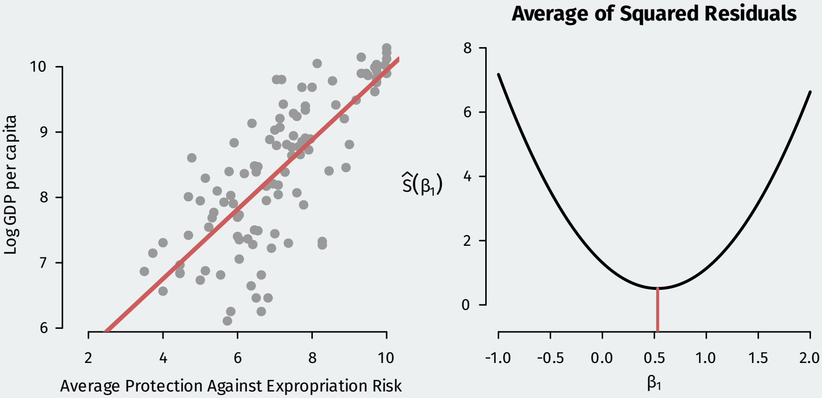
Visualizing OLS



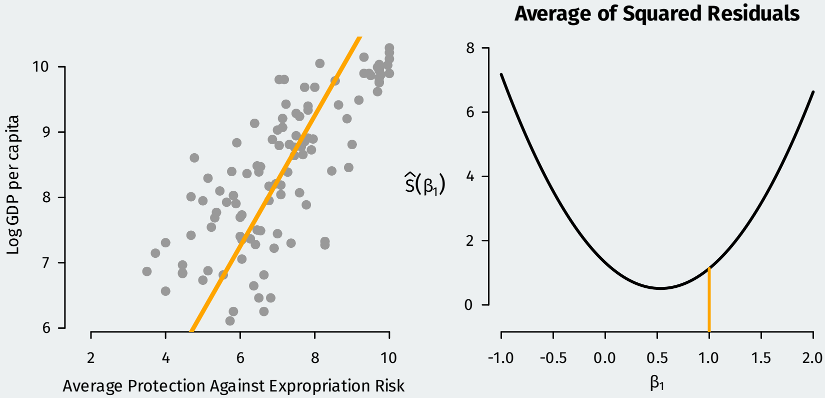
Visualizing OLS



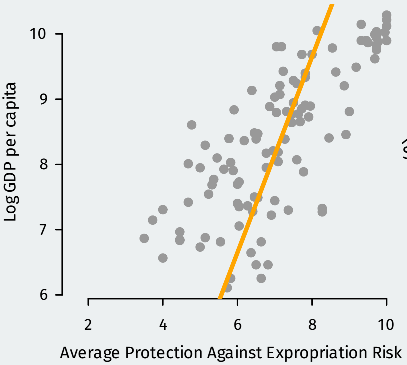
Visualizing OLS

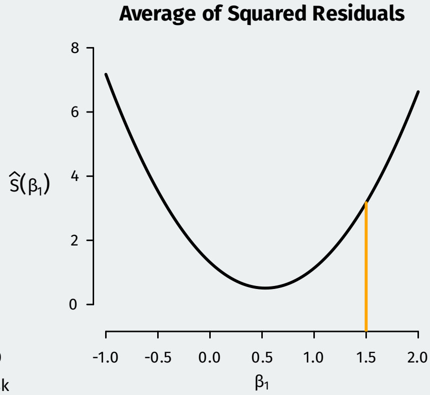


Visualizing OLS



Visualizing OLS





Residuals

- **Fitted value** $\widehat{Y}_i = \mathbf{X}_i'\hat{\boldsymbol{\beta}}$ is what the model predicts at $\mathbf{X}_i$

Residuals

- **Fitted value** $\widehat{Y}_i = \mathbf{X}_i'\hat{\boldsymbol{\beta}}$ is what the model predicts at $\mathbf{X}_i$
    - Not really a prediction for Y_i since that was used to generate $\hat{\boldsymbol{\beta}}$

Residuals

- **Fitted value** $\widehat{Y}_i = \mathbf{X}_i'\hat{\boldsymbol{\beta}}$ is what the model predicts at $\mathbf{X}_i$
    - Not really a prediction for Y_i since that was used to generate $\hat{\boldsymbol{\beta}}$
- **Residuals** are the difference between observed and fitted values:

$$\hat{e}_i = Y_i - \widehat{Y}_i = Y_i - \mathbf{X}_i'\hat{\boldsymbol{\beta}}$$

Residuals

- **Fitted value** $\widehat{Y}_i = \mathbf{X}_i'\hat{\boldsymbol{\beta}}$ is what the model predicts at $\mathbf{X}_i$
  - Not really a prediction for Y_i since that was used to generate $\hat{\boldsymbol{\beta}}$
- **Residuals** are the difference between observed and fitted values:

$$\hat{e}_i = Y_i - \widehat{Y}_i = Y_i - \mathbf{X}_i'\hat{\boldsymbol{\beta}}$$

  - We can write $Y_i = \mathbf{X}_i'\hat{\boldsymbol{\beta}} + \hat{e}_i$.

Residuals

- **Fitted value** $\widehat{Y}_i = \mathbf{X}_i'\hat{\boldsymbol{\beta}}$ is what the model predicts at $\mathbf{X}_i$
    - Not really a prediction for Y_i since that was used to generate $\hat{\boldsymbol{\beta}}$
- **Residuals** are the difference between observed and fitted values:

$$\hat{e}_i = Y_i - \widehat{Y}_i = Y_i - \mathbf{X}_i'\hat{\boldsymbol{\beta}}$$

- We can write $Y_i = \mathbf{X}_i'\hat{\boldsymbol{\beta}} + \hat{e}_i$.
- $\hat{e}_i$ are not the true errors e_i

Residuals

- **Fitted value** $\widehat{Y}_i = \mathbf{X}_i'\hat{\boldsymbol{\beta}}$ is what the model predicts at $\mathbf{X}_i$
  - Not really a prediction for Y_i since that was used to generate $\hat{\boldsymbol{\beta}}$
- **Residuals** are the difference between observed and fitted values:

$$\hat{e}_i = Y_i - \widehat{Y}_i = Y_i - \mathbf{X}_i'\hat{\boldsymbol{\beta}}$$

  - We can write $Y_i = \mathbf{X}_i'\hat{\boldsymbol{\beta}} + \hat{e}_i$.
  - $\hat{e}_i$ are not the true errors e_i
- Key **mechanical properties** of OLS residuals:

$$\sum_{i=1}^{n} \mathbf{X}_i \hat{e}_i = 0$$

Residuals

- **Fitted value** $\widehat{Y}_i = \mathbf{X}_i'\hat{\boldsymbol{\beta}}$ is what the model predicts at $\mathbf{X}_i$
    - Not really a prediction for Y_i since that was used to generate $\hat{\boldsymbol{\beta}}$
- **Residuals** are the difference between observed and fitted values:

$$\hat{e}_i = Y_i - \widehat{Y}_i = Y_i - \mathbf{X}_i'\hat{\boldsymbol{\beta}}$$

    - We can write $Y_i = \mathbf{X}_i'\hat{\boldsymbol{\beta}} + \hat{e}_i$.
    - $\hat{e}_i$ are not the true errors e_i
- Key **mechanical properties** of OLS residuals:

$$\sum_{i=1}^{n} \mathbf{X}_i \hat{e}_i = 0$$

    - Sample covariance between $\mathbf{X}_i$ and $\hat{e}_i$ is 0.

Residuals

- **Fitted value** $\widehat{Y}_i = \mathbf{X}_i'\hat{\boldsymbol{\beta}}$ is what the model predicts at $\mathbf{X}_i$
  - Not really a prediction for Y_i since that was used to generate $\hat{\boldsymbol{\beta}}$
- **Residuals** are the difference between observed and fitted values:

$$\hat{e}_i = Y_i - \widehat{Y}_i = Y_i - \mathbf{X}_i'\hat{\boldsymbol{\beta}}$$

  - We can write $Y_i = \mathbf{X}_i'\hat{\boldsymbol{\beta}} + \hat{e}_i$.
  - $\hat{e}_i$ are not the true errors e_i
- Key **mechanical properties** of OLS residuals:

$$\sum_{i=1}^{n} \mathbf{X}_i\hat{e}_i = 0$$

  - Sample covariance between $\mathbf{X}_i$ and $\hat{e}_i$ is 0.
  - If $\mathbf{X}_i$ has a constant, then $n^{-1}\sum_{i=1}^{n}\hat{e}_i = 0$

2/ Model fit

Prediction error

- How do we judge how well a regression fits the data?

Prediction error

- How do we judge how well a regression fits the data?
- How much does $\mathbf{X}_i$ help us predict Y_i?

Prediction error

- How do we judge how well a regression fits the data?
- How much does $\mathbf{X}_i$ help us predict Y_i?
- **Prediction errors without $\mathbf{X}_i$:**

Prediction error

- How do we judge how well a regression fits the data?
- How much does $\mathbf{X}_i$ help us predict Y_i?
- **Prediction errors without $\mathbf{X}_i$:**
    - Best prediction is the mean, $\overline{Y}$

Prediction error

- How do we judge how well a regression fits the data?
- How much does $\mathbf{X}_i$ help us predict Y_i?
- **Prediction errors without $\mathbf{X}_i$:**
  - Best prediction is the mean, $\overline{Y}$
  - Prediction error is called the total sum of squares (TSS) would be:

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

Prediction error

- How do we judge how well a regression fits the data?
- How much does $\mathbf{X}_i$ help us predict Y_i?
- **Prediction errors without $\mathbf{X}_i$:**
  - Best prediction is the mean, $\overline{Y}$
  - Prediction error is called the total sum of squares (TSS) would be:

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

- **Prediction errors with $\mathbf{X}_i$:**

Prediction error

- How do we judge how well a regression fits the data?
- How much does $\mathbf{X}_i$ help us predict Y_i?
- **Prediction errors without $\mathbf{X}_i$:**
    - Best prediction is the mean, $\overline{Y}$
    - Prediction error is called the total sum of squares (TSS) would be:

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

- **Prediction errors with $\mathbf{X}_i$:**
    - Best predictions are the fitted values, $\widehat{Y}_i$.

Prediction error

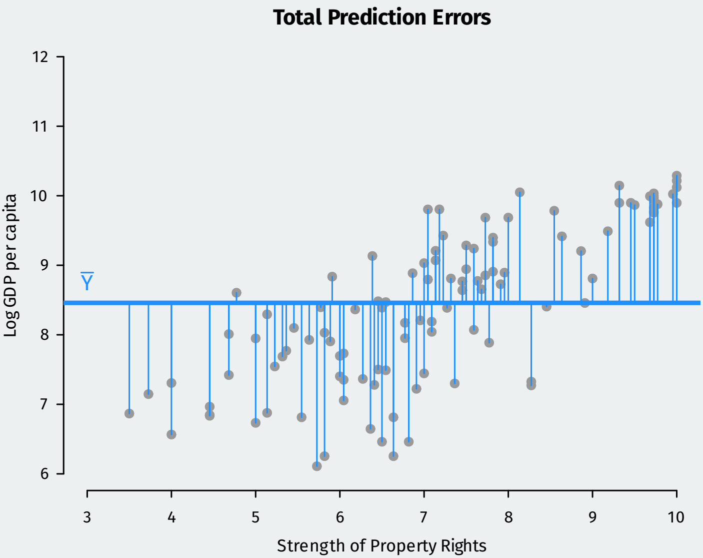
- How do we judge how well a regression fits the data?
- How much does $\mathbf{X}_i$ help us predict Y_i?
- **Prediction errors without $\mathbf{X}_i$:**
    - Best prediction is the mean, $\overline{Y}$
    - Prediction error is called the total sum of squares (TSS) would be:

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

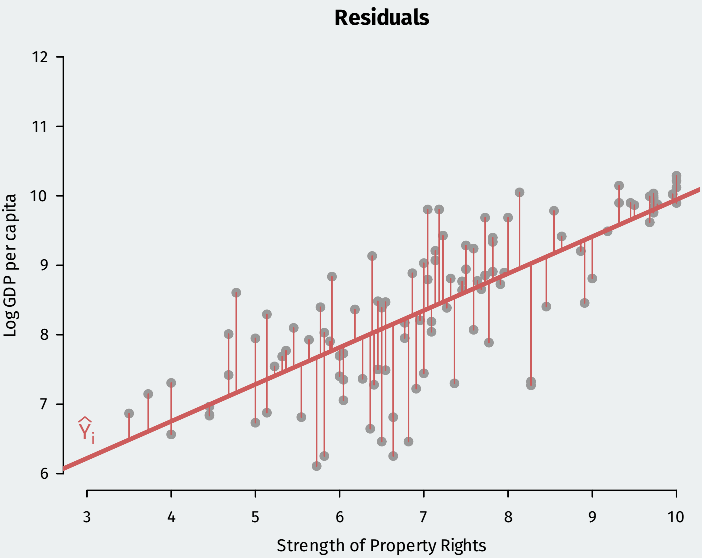
- **Prediction errors with $\mathbf{X}_i$:**
    - Best predictions are the fitted values, $\widehat{Y}_i$.
    - Prediction error is the sum of the squared residuals or SSR:

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

Total SS vs SSR



Total SS vs SSR



R-squared

- Regression will always improve in-sample fit: $TSS > SSR$

R-squared

- Regression will always improve in-sample fit: $TSS > SSR$
- How much better does using $\mathbf{X}_i$ do? **Coefficient of determination** or R^2:

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

R-squared

- Regression will always improve in-sample fit: $TSS > SSR$
- How much better does using $\mathbf{X}_i$ do? **Coefficient of determination** or R^2:

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

- $R^2 =$ fraction of the total prediction error eliminated by using $\mathbf{X}_i$.

R-squared

- Regression will always improve in-sample fit: $TSS > SSR$
- How much better does using $\mathbf{X}_i$ do? **Coefficient of determination** or R^2:

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

- $R^2 =$ fraction of the total prediction error eliminated by using $\mathbf{X}_i$.
- **Common interpretation:** R^2 is the fraction of the variation in Y_i is "explained by" $\mathbf{X}_i$.

R-squared

- Regression will always improve in-sample fit: $TSS > SSR$
- How much better does using $\mathbf{X}_i$ do? **Coefficient of determination** or R^2:

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

- $R^2 =$ fraction of the total prediction error eliminated by using $\mathbf{X}_i$.
- **Common interpretation:** R^2 is the fraction of the variation in Y_i is "explained by" $\mathbf{X}_i$.
    - $R^2 = 0$ means no relationship

R-squared

- Regression will always improve in-sample fit: $TSS > SSR$
- How much better does using $\mathbf{X}_i$ do? **Coefficient of determination** or R^2:

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

- $R^2 =$ fraction of the total prediction error eliminated by using $\mathbf{X}_i$.
- **Common interpretation:** R^2 is the fraction of the variation in Y_i is "explained by" $\mathbf{X}_i$.
  - $R^2 = 0$ means no relationship
  - $R^2 = 1$ implies perfect linear fit

R-squared

- Regression will always improve in-sample fit: $TSS > SSR$
- How much better does using $\mathbf{X}_i$ do? **Coefficient of determination** or R^2:

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

- $R^2 =$ fraction of the total prediction error eliminated by using $\mathbf{X}_i$.
- **Common interpretation:** R^2 is the fraction of the variation in Y_i is "explained by" $\mathbf{X}_i$.
  - $R^2 = 0$ means no relationship
  - $R^2 = 1$ implies perfect linear fit
- Mechanically increases with additional covariates (better fit measures exist)

3/ Geometry of OLS

Linear model in matrix form

- Linear model is a system of n linear equations:

$$
\begin{aligned}
Y_1 &= \mathbf{X}_1'\boldsymbol{\beta} + e_1 \\
Y_2 &= \mathbf{X}_2'\boldsymbol{\beta} + e_2 \\
&\vdots \\
Y_n &= \mathbf{X}_n'\boldsymbol{\beta} + e_n
\end{aligned}
$$

Linear model in matrix form

- Linear model is a system of n linear equations:

$$
\begin{aligned}
Y_1 &= \mathbf{X}_1'\boldsymbol{\beta} + e_1 \\
Y_2 &= \mathbf{X}_2'\boldsymbol{\beta} + e_2 \\
&\vdots \\
Y_n &= \mathbf{X}_n'\boldsymbol{\beta} + e_n
\end{aligned}
$$

- We can write this more compactly using matrices and vectors:

$$
\mathbf{Y} = \begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix}, \quad \mathbb{X} = \begin{pmatrix} \mathbf{X}_1' \\ \mathbf{X}_2' \\ \vdots \\ \mathbf{X}_n' \end{pmatrix} = \begin{pmatrix} 1 & X_{11} & X_{12} & \cdots & X_{1k} \\ 1 & X_{21} & X_{22} & \cdots & X_{2k} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & X_{n1} & X_{n2} & \cdots & X_{nk} \end{pmatrix}, \quad \mathbf{e} = \begin{pmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{pmatrix}
$$

Linear model in matrix form

- Linear model is a system of n linear equations:

$$
\begin{aligned}
Y_1 &= \mathbf{X}_1'\boldsymbol{\beta} + e_1 \\
Y_2 &= \mathbf{X}_2'\boldsymbol{\beta} + e_2 \\
&\vdots \\
Y_n &= \mathbf{X}_n'\boldsymbol{\beta} + e_n
\end{aligned}
$$

- We can write this more compactly using matrices and vectors:

$$
\mathbf{Y} = \begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix}, \quad \mathbb{X} = \begin{pmatrix} \mathbf{X}_1' \\ \mathbf{X}_2' \\ \vdots \\ \mathbf{X}_n' \end{pmatrix} = \begin{pmatrix} 1 & X_{11} & X_{12} & \cdots & X_{1k} \\ 1 & X_{21} & X_{22} & \cdots & X_{2k} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & X_{n1} & X_{n2} & \cdots & X_{nk} \end{pmatrix}, \quad \mathbf{e} = \begin{pmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{pmatrix}
$$

- Model is now just:

$$
\mathbf{Y} = \mathbb{X}\boldsymbol{\beta} + \mathbf{e}
$$

OLS estimator in matrix form

- Key relationship: sample sums can be written in matrix notation:

$$\sum_{i=1}^{n} \mathbf{X}_i \mathbf{X}_i' = \mathbb{X}'\mathbb{X}$$

$$\sum_{i=1}^{n} \mathbf{X}_i Y_i = \mathbb{X}'\mathbf{Y}$$

OLS estimator in matrix form

- Key relationship: sample sums can be written in matrix notation:

$$\sum_{i=1}^{n} \mathbf{X}_i \mathbf{X}_i' = \mathbb{X}'\mathbb{X}$$

$$\sum_{i=1}^{n} \mathbf{X}_i Y_i = \mathbb{X}'\mathbf{Y}$$

- Implies we can write the OLS estimator as

$$\hat{\boldsymbol{\beta}} = \left(\mathbb{X}'\mathbb{X}\right)^{-1} \mathbb{X}'\mathbf{Y}$$

OLS estimator in matrix form

- Key relationship: sample sums can be written in matrix notation:

$$\sum_{i=1}^{n} \mathbf{X}_i \mathbf{X}_i' = \mathbb{X}'\mathbb{X}$$

$$\sum_{i=1}^{n} \mathbf{X}_i Y_i = \mathbb{X}'\mathbf{Y}$$

- Implies we can write the OLS estimator as

$$\hat{\boldsymbol{\beta}} = \left(\mathbb{X}'\mathbb{X}\right)^{-1} \mathbb{X}'\mathbf{Y}$$

- Residuals:

OLS estimator in matrix form

- Key relationship: sample sums can be written in matrix notation:

$$\sum_{i=1}^{n} \mathbf{X}_i \mathbf{X}_i' = \mathbb{X}'\mathbb{X}$$

$$\sum_{i=1}^{n} \mathbf{X}_i Y_i = \mathbb{X}'\mathbf{Y}$$

- Implies we can write the OLS estimator as

$$\hat{\boldsymbol{\beta}} = \left(\mathbb{X}'\mathbb{X}\right)^{-1} \mathbb{X}'\mathbf{Y}$$

- Residuals:

$$\hat{\mathbf{e}} = \mathbf{Y} - \mathbb{X}\hat{\boldsymbol{\beta}} = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix} - \begin{bmatrix} 1\hat{\beta}_0 + X_{11}\hat{\beta}_1 + X_{12}\hat{\beta}_2 + \cdots + X_{1k}\hat{\beta}_k \\ 1\hat{\beta}_0 + X_{21}\hat{\beta}_1 + X_{22}\hat{\beta}_2 + \cdots + X_{2k}\hat{\beta}_k \\ \vdots \\ 1\hat{\beta}_0 + X_{n1}\hat{\beta}_1 + X_{n2}\hat{\beta}_2 + \cdots + X_{nk}\hat{\beta}_k \end{bmatrix}$$

Projection/hat matrix

- We can define the transformation of **Y** that does the projection.

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

Projection/hat matrix

- We can define the transformation of $\mathbf{Y}$ that does the projection.

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

- **Projection matrix**

$$\mathbf{P} = \mathbb{X}(\mathbb{X}'\mathbb{X})^{-1}\mathbb{X}'$$

Projection/hat matrix

- We can define the transformation of **Y** that does the projection.

$$\mathbb{X}\widehat{\boldsymbol{\beta}} = \mathbb{X}(\mathbb{X}'\mathbb{X})^{-1}\mathbb{X}'\mathbf{Y}$$

- **Projection matrix**

$$\mathbf{P} = \mathbb{X}(\mathbb{X}'\mathbb{X})^{-1}\mathbb{X}'$$

- Also called the **hat matrix** it puts the "hat" on **Y**:

$$\mathbf{PY} = \mathbb{X}(\mathbb{X}'\mathbb{X})^{-1}\mathbb{X}'\mathbf{Y} = \mathbb{X}\widehat{\boldsymbol{\beta}} = \widehat{\mathbf{Y}}$$

Projection/hat matrix

- We can define the transformation of $\mathbf{Y}$ that does the projection.

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

- **Projection matrix**

$$\mathbf{P} = \mathbb{X}(\mathbb{X}'\mathbb{X})^{-1}\mathbb{X}'$$

- Also called the **hat matrix** it puts the "hat" on $\mathbf{Y}$:

$$\mathbf{P}\mathbf{Y} = \mathbb{X}(\mathbb{X}'\mathbb{X})^{-1}\mathbb{X}'\mathbf{Y} = \mathbb{X}\hat{\boldsymbol{\beta}} = \widehat{\mathbf{Y}}$$

- Key properties:

Projection/hat matrix

- We can define the transformation of **Y** that does the projection.

$$\mathbb{X}\widehat{\boldsymbol{\beta}} = \mathbb{X}(\mathbb{X}'\mathbb{X})^{-1}\mathbb{X}'\mathbf{Y}$$

- **Projection matrix**

$$\mathbf{P} = \mathbb{X}(\mathbb{X}'\mathbb{X})^{-1}\mathbb{X}'$$

- Also called the **hat matrix** it puts the "hat" on **Y**:

$$\mathbf{PY} = \mathbb{X}(\mathbb{X}'\mathbb{X})^{-1}\mathbb{X}'\mathbf{Y} = \mathbb{X}\widehat{\boldsymbol{\beta}} = \widehat{\mathbf{Y}}$$

- Key properties:
    - **P** is an $n \times n$ symmetric matrix

Projection/hat matrix

- We can define the transformation of $\mathbf{Y}$ that does the projection.

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

- **Projection matrix**

$$\mathbf{P} = \mathbb{X}(\mathbb{X}'\mathbb{X})^{-1}\mathbb{X}'$$

- Also called the **hat matrix** it puts the "hat" on $\mathbf{Y}$:

$$\mathbf{P}\mathbf{Y} = \mathbb{X}(\mathbb{X}'\mathbb{X})^{-1}\mathbb{X}'\mathbf{Y} = \mathbb{X}\hat{\boldsymbol{\beta}} = \widehat{\mathbf{Y}}$$

- Key properties:
    - $\mathbf{P}$ is an $n \times n$ symmetric matrix
    - $\mathbf{P}$ is **idempotent**: $\mathbf{P}\mathbf{P} = \mathbf{P}$

Projection/hat matrix

- We can define the transformation of $\mathbf{Y}$ that does the projection.

$$\mathbb{X}\widehat{\boldsymbol{\beta}} = \mathbb{X}(\mathbb{X}'\mathbb{X})^{-1}\mathbb{X}'\mathbf{Y}$$

- **Projection matrix**

$$\mathbf{P} = \mathbb{X}(\mathbb{X}'\mathbb{X})^{-1}\mathbb{X}'$$

- Also called the **hat matrix** it puts the "hat" on $\mathbf{Y}$:

$$\mathbf{P}\mathbf{Y} = \mathbb{X}(\mathbb{X}'\mathbb{X})^{-1}\mathbb{X}'\mathbf{Y} = \mathbb{X}\widehat{\boldsymbol{\beta}} = \widehat{\mathbf{Y}}$$

- Key properties:
    - $\mathbf{P}$ is an $n \times n$ symmetric matrix
    - $\mathbf{P}$ is **idempotent**: $\mathbf{P}\mathbf{P} = \mathbf{P}$
    - Projecting $\mathbb{X}$ onto itself returns itself: $\mathbf{P}\mathbb{X} = \mathbb{X}$

Annihilator matrix

- **Annihilator matrix** projects onto the space spanned by the residual:

$$\mathbf{M} = \mathbf{I}_n - \mathbf{P} = \mathbf{I}_n - \mathbb{X}(\mathbb{X}'\mathbb{X})^{-1}\mathbb{X}'$$

Annihilator matrix

- **Annihilator matrix** projects onto the space spanned by the residual:

$$\mathbf{M} = \mathbf{I}_n - \mathbf{P} = \mathbf{I}_n - \mathbb{X}(\mathbb{X}'\mathbb{X})^{-1}\mathbb{X}'$$

- Also called the **residual maker**:

$$\mathbf{MY} = (\mathbf{I}_n - \mathbf{P})\mathbf{Y} = \mathbf{Y} - \mathbf{PY} = \mathbf{Y} - \widehat{\mathbf{Y}} = \hat{\mathbf{e}}$$

Annihilator matrix

- **Annihilator matrix** projects onto the space spanned by the residual:

$$\mathbf{M} = \mathbf{I}_n - \mathbf{P} = \mathbf{I}_n - \mathbb{X}(\mathbb{X}'\mathbb{X})^{-1}\mathbb{X}'$$

- Also called the **residual maker**:

$$\mathbf{MY} = (\mathbf{I}_n - \mathbf{P})\mathbf{Y} = \mathbf{Y} - \mathbf{PY} = \mathbf{Y} - \widehat{\mathbf{Y}} = \hat{\mathbf{e}}$$

- “Annihilates” any function in the column space of $\mathbb{X}$, $\mathcal{C}(\mathbb{X})$:

$$\mathbf{M}\mathbb{X} = (\mathbf{I}_n - \mathbf{P})\mathbb{X} = \mathbb{X} - \mathbf{P}\mathbb{X} = \mathbb{X} - \mathbb{X} = 0$$

Annihilator matrix

- **Annihilator matrix** projects onto the space spanned by the residual:

$$\mathbf{M} = \mathbf{I}_n - \mathbf{P} = \mathbf{I}_n - \mathbb{X}(\mathbb{X}'\mathbb{X})^{-1}\mathbb{X}'$$

- Also called the **residual maker**:

$$\mathbf{MY} = (\mathbf{I}_n - \mathbf{P})\mathbf{Y} = \mathbf{Y} - \mathbf{PY} = \mathbf{Y} - \widehat{\mathbf{Y}} = \hat{\mathbf{e}}$$

- "Annihilates" any function in the column space of $\mathbb{X}$, $\mathcal{C}(\mathbb{X})$:

$$\mathbf{M}\mathbb{X} = (\mathbf{I}_n - \mathbf{P})\mathbb{X} = \mathbb{X} - \mathbf{P}\mathbb{X} = \mathbb{X} - \mathbb{X} = 0$$

- Properties:

Annihilator matrix

- **Annihilator matrix** projects onto the space spanned by the residual:

$$\mathbf{M} = \mathbf{I}_n - \mathbf{P} = \mathbf{I}_n - \mathbb{X}(\mathbb{X}'\mathbb{X})^{-1}\mathbb{X}'$$

- Also called the **residual maker**:

$$\mathbf{MY} = (\mathbf{I}_n - \mathbf{P})\mathbf{Y} = \mathbf{Y} - \mathbf{PY} = \mathbf{Y} - \widehat{\mathbf{Y}} = \hat{\mathbf{e}}$$

- “Annihilates” any function in the column space of $\mathbb{X}$, $\mathcal{C}(\mathbb{X})$:

$$\mathbf{M}\mathbb{X} = (\mathbf{I}_n - \mathbf{P})\mathbb{X} = \mathbb{X} - \mathbf{P}\mathbb{X} = \mathbb{X} - \mathbb{X} = 0$$

- Properties:
    - $\mathbf{M}$ is a symmetric $n \times n$ matrix and is idempotent $\mathbf{MM} = \mathbf{M}$

Annihilator matrix

- **Annihilator matrix** projects onto the space spanned by the residual:

$$\mathbf{M} = \mathbf{I}_n - \mathbf{P} = \mathbf{I}_n - \mathbb{X}(\mathbb{X}'\mathbb{X})^{-1}\mathbb{X}'$$

- Also called the **residual maker**:

$$\mathbf{MY} = (\mathbf{I}_n - \mathbf{P})\mathbf{Y} = \mathbf{Y} - \mathbf{PY} = \mathbf{Y} - \widehat{\mathbf{Y}} = \hat{\mathbf{e}}$$

- "Annihilates" any function in the column space of $\mathbb{X}$, $\mathcal{C}(\mathbb{X})$:

$$\mathbf{M}\mathbb{X} = (\mathbf{I}_n - \mathbf{P})\mathbb{X} = \mathbb{X} - \mathbf{P}\mathbb{X} = \mathbb{X} - \mathbb{X} = 0$$

- Properties:
    - $\mathbf{M}$ is a symmetric $n \times n$ matrix and is idempotent $\mathbf{MM} = \mathbf{M}$
    - Admits a nice expression for the residual vector: $\hat{\mathbf{e}} = \mathbf{Me}$

Annihilator matrix

- **Annihilator matrix** projects onto the space spanned by the residual:

$$\mathbf{M} = \mathbf{I}_n - \mathbf{P} = \mathbf{I}_n - \mathbb{X}(\mathbb{X}'\mathbb{X})^{-1}\mathbb{X}'$$

- Also called the **residual maker**:

$$\mathbf{MY} = (\mathbf{I}_n - \mathbf{P})\mathbf{Y} = \mathbf{Y} - \mathbf{PY} = \mathbf{Y} - \widehat{\mathbf{Y}} = \hat{\mathbf{e}}$$

- "Annihilates" any function in the column space of $\mathbb{X}$, $\mathcal{C}(\mathbb{X})$:

$$\mathbf{M}\mathbb{X} = (\mathbf{I}_n - \mathbf{P})\mathbb{X} = \mathbb{X} - \mathbf{P}\mathbb{X} = \mathbb{X} - \mathbb{X} = 0$$

- Properties:
    - $\mathbf{M}$ is a symmetric $n \times n$ matrix and is idempotent $\mathbf{MM} = \mathbf{M}$
    - Admits a nice expression for the residual vector: $\hat{\mathbf{e}} = \mathbf{Me}$

- Allows the following orthogonal partition:

$$\mathbf{Y} = \mathbf{PY} + \mathbf{MY} = \text{projection} + \text{residual}$$

Geometric view of OLS

- Recall the length of a vector: $\|\hat{\mathbf{a}}\| = \sqrt{\hat{a}_1^1 + \cdots + \hat{a}_n^2}$

Geometric view of OLS

- Recall the length of a vector: $\|\hat{\mathbf{a}}\| = \sqrt{\hat{a}_1^1 + \cdots + \hat{a}_n^2}$
- Distance between two vectors: $\|\mathbf{a} - \mathbf{b}\| = \sqrt{(a_1 - b_1)^2 + \cdots + (a_n - b_n)^2}$

Geometric view of OLS

- Recall the length of a vector: $\|\hat{\mathbf{a}}\| = \sqrt{\hat{a}_1^1 + \cdots + \hat{a}_n^2}$
- Distance between two vectors: $\|\mathbf{a} - \mathbf{b}\| = \sqrt{(a_1 - b_1)^2 + \cdots + (a_n - b_n)^2}$
- We can rewrite the OLS estimator as:

$$\hat{\boldsymbol{\beta}} = \underset{\mathbf{b} \in \mathbb{R}^{k+1}}{\arg\min} \|\mathbf{Y} - \mathbb{X}\mathbf{b}\|^2 = \underset{\mathbf{b} \in \mathbb{R}^{k+1}}{\arg\min} \sum_{i=1}^{n} (Y_i - \mathbf{X}_i'\mathbf{b})^2$$

Geometric view of OLS

- Recall the length of a vector: $\|\hat{\mathbf{a}}\| = \sqrt{\hat{a}_1^1 + \cdots + \hat{a}_n^2}$
- Distance between two vectors: $\|\mathbf{a} - \mathbf{b}\| = \sqrt{(a_1 - b_1)^2 + \cdots + (a_n - b_n)^2}$
- We can rewrite the OLS estimator as:

$$\hat{\boldsymbol{\beta}} = \underset{\mathbf{b} \in \mathbb{R}^{k+1}}{\arg\min} \|\mathbf{Y} - \mathbb{X}\mathbf{b}\|^2 = \underset{\mathbf{b} \in \mathbb{R}^{k+1}}{\arg\min} \sum_{i=1}^{n} (Y_i - \mathbf{X}_i'\mathbf{b})^2$$

- Let $\mathcal{C}(\mathbb{X}) = \{\mathbb{X}\mathbf{b} : \mathbf{b} \in \mathbb{R}^{k+1}\}$ be the column space of $\mathbb{X}$

Geometric view of OLS

- Recall the length of a vector: $\|\hat{\mathbf{a}}\| = \sqrt{\hat{a}_1^1 + \cdots + \hat{a}_n^2}$
- Distance between two vectors: $\|\mathbf{a} - \mathbf{b}\| = \sqrt{(a_1 - b_1)^2 + \cdots + (a_n - b_n)^2}$
- We can rewrite the OLS estimator as:

$$\hat{\boldsymbol{\beta}} = \underset{\mathbf{b} \in \mathbb{R}^{k+1}}{\arg\min} \|\mathbf{Y} - \mathbb{X}\mathbf{b}\|^2 = \underset{\mathbf{b} \in \mathbb{R}^{k+1}}{\arg\min} \sum_{i=1}^{n} (Y_i - \mathbf{X}_i'\mathbf{b})^2$$

- Let $\mathcal{C}(\mathbb{X}) = \{\mathbb{X}\mathbf{b} : \mathbf{b} \in \mathbb{R}^{k+1}\}$ be the column space of $\mathbb{X}$
    - All n-vectors formed as a linear combination of the columns of $\mathbb{X}$.

Geometric view of OLS

- Recall the length of a vector: $\|\hat{\mathbf{a}}\| = \sqrt{\hat{a}_1^1 + \cdots + \hat{a}_n^2}$
- Distance between two vectors: $\|\mathbf{a} - \mathbf{b}\| = \sqrt{(a_1 - b_1)^2 + \cdots + (a_n - b_n)^2}$
- We can rewrite the OLS estimator as:

$$\hat{\boldsymbol{\beta}} = \underset{\mathbf{b} \in \mathbb{R}^{k+1}}{\arg\min} \|\mathbf{Y} - \mathbb{X}\mathbf{b}\|^2 = \underset{\mathbf{b} \in \mathbb{R}^{k+1}}{\arg\min} \sum_{i=1}^{n} (Y_i - \mathbf{X}_i'\mathbf{b})^2$$

- Let $\mathcal{C}(\mathbb{X}) = \{\mathbb{X}\mathbf{b} : \mathbf{b} \in \mathbb{R}^{k+1}\}$ be the column space of $\mathbb{X}$
    - All n-vectors formed as a linear combination of the columns of $\mathbb{X}$.
    - $k + 1$-dimensional subspace of $\mathbb{R}^n$

Geometric view of OLS

- Recall the length of a vector: $\|\hat{\mathbf{a}}\| = \sqrt{\hat{a}_1^1 + \cdots + \hat{a}_n^2}$
- Distance between two vectors: $\|\mathbf{a} - \mathbf{b}\| = \sqrt{(a_1 - b_1)^2 + \cdots + (a_n - b_n)^2}$
- We can rewrite the OLS estimator as:

$$\hat{\boldsymbol{\beta}} = \underset{\mathbf{b} \in \mathbb{R}^{k+1}}{\arg\min} \|\mathbf{Y} - \mathbb{X}\mathbf{b}\|^2 = \underset{\mathbf{b} \in \mathbb{R}^{k+1}}{\arg\min} \sum_{i=1}^{n} (Y_i - \mathbf{X}_i'\mathbf{b})^2$$

- Let $\mathcal{C}(\mathbb{X}) = \{\mathbb{X}\mathbf{b} : \mathbf{b} \in \mathbb{R}^{k+1}\}$ be the column space of $\mathbb{X}$
  - All n-vectors formed as a linear combination of the columns of $\mathbb{X}$.
  - $k+1$-dimensional subspace of $\mathbb{R}^n$
  - This is the space that OLS is searching over!

Geometric view of OLS

- Recall the length of a vector: $\|\hat{\mathbf{a}}\| = \sqrt{\hat{a}_1^1 + \cdots + \hat{a}_n^2}$
- Distance between two vectors: $\|\mathbf{a} - \mathbf{b}\| = \sqrt{(a_1 - b_1)^2 + \cdots + (a_n - b_n)^2}$
- We can rewrite the OLS estimator as:

$$\hat{\boldsymbol{\beta}} = \underset{\mathbf{b} \in \mathbb{R}^{k+1}}{\arg\min} \|\mathbf{Y} - \mathbb{X}\mathbf{b}\|^2 = \underset{\mathbf{b} \in \mathbb{R}^{k+1}}{\arg\min} \sum_{i=1}^{n} (Y_i - \mathbf{X}_i'\mathbf{b})^2$$

- Let $\mathcal{C}(\mathbb{X}) = \{\mathbb{X}\mathbf{b} : \mathbf{b} \in \mathbb{R}^{k+1}\}$ be the column space of $\mathbb{X}$
  - All n-vectors formed as a linear combination of the columns of $\mathbb{X}$.
  - $k + 1$-dimensional subspace of $\mathbb{R}^n$
  - This is the space that OLS is searching over!
- Geometrically OLS is:

Geometric view of OLS

- Recall the length of a vector: $\|\hat{\mathbf{a}}\| = \sqrt{\hat{a}_1^1 + \cdots + \hat{a}_n^2}$
- Distance between two vectors: $\|\mathbf{a} - \mathbf{b}\| = \sqrt{(a_1 - b_1)^2 + \cdots + (a_n - b_n)^2}$
- We can rewrite the OLS estimator as:

$$\hat{\boldsymbol{\beta}} = \underset{\mathbf{b} \in \mathbb{R}^{k+1}}{\arg\min} \|\mathbf{Y} - \mathbb{X}\mathbf{b}\|^2 = \underset{\mathbf{b} \in \mathbb{R}^{k+1}}{\arg\min} \sum_{i=1}^{n} (Y_i - \mathbf{X}_i'\mathbf{b})^2$$

- Let $\mathcal{C}(\mathbb{X}) = \{\mathbb{X}\mathbf{b} : \mathbf{b} \in \mathbb{R}^{k+1}\}$ be the column space of $\mathbb{X}$
  - All n-vectors formed as a linear combination of the columns of $\mathbb{X}$.
  - $k + 1$-dimensional subspace of $\mathbb{R}^n$
  - This is the space that OLS is searching over!
- Geometrically OLS is:
  - Find coefficients that minimize distance between the $\mathbf{Y}$ and $\mathbb{X}\mathbf{b}$.

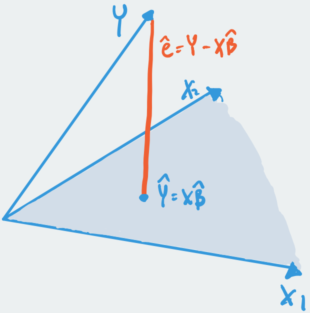
Geometric view of OLS

- Recall the length of a vector: $\|\hat{\mathbf{a}}\| = \sqrt{\hat{a}_1^1 + \cdots + \hat{a}_n^2}$
- Distance between two vectors: $\|\mathbf{a} - \mathbf{b}\| = \sqrt{(a_1 - b_1)^2 + \cdots + (a_n - b_n)^2}$
- We can rewrite the OLS estimator as:

$$\hat{\boldsymbol{\beta}} = \underset{\mathbf{b} \in \mathbb{R}^{k+1}}{\arg\min} \|\mathbf{Y} - \mathbb{X}\mathbf{b}\|^2 = \underset{\mathbf{b} \in \mathbb{R}^{k+1}}{\arg\min} \sum_{i=1}^{n} (Y_i - \mathbf{X}_i'\mathbf{b})^2$$

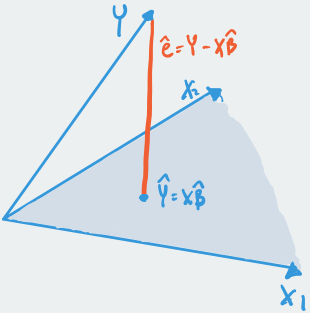
- Let $\mathcal{C}(\mathbb{X}) = \{\mathbb{X}\mathbf{b} : \mathbf{b} \in \mathbb{R}^{k+1}\}$ be the column space of $\mathbb{X}$
    - All n-vectors formed as a linear combination of the columns of $\mathbb{X}$.
    - $k + 1$-dimensional subspace of $\mathbb{R}^n$
    - This is the space that OLS is searching over!
- Geometrically OLS is:
    - Find coefficients that minimize distance between the $\mathbf{Y}$ and $\mathbb{X}\mathbf{b}$.
    - Find the point in $\mathcal{C}(\mathbb{X})$ that is closest to $\mathbf{Y}$

Projection



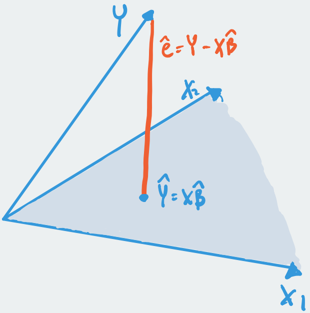
- Finding closest point in $\mathcal{C}(\mathbb{X})$ to $\mathbf{Y}$ is called **projection**

Projection



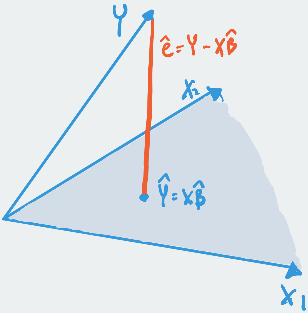
- Finding closest point in $\mathcal{C}(\mathbb{X})$ to $\mathbf{Y}$ is called **projection**
- Example: $n = 3$ and $k = 2$: points in 3D space.

Projection



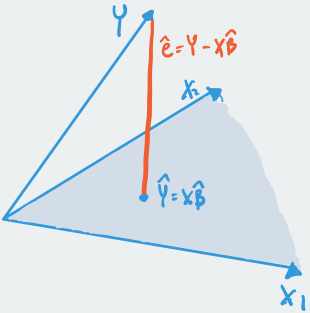
- Finding closest point in $\mathcal{C}(\mathbb{X})$ to $\mathbf{Y}$ is called **projection**
- Example: $n = 3$ and $k = 2$: points in 3D space.
  - Column space of $\mathbb{X}$ is a plane in this space.

Projection



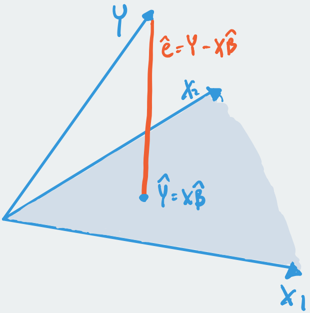
- Finding closest point in $\mathcal{C}(\mathbb{X})$ to $\mathbf{Y}$ is called **projection**
- Example: $n = 3$ and $k = 2$: points in 3D space.
    - Column space of $\mathbb{X}$ is a plane in this space.
- Residual vector $\hat{\mathbf{e}} = \mathbf{Y} - \mathbb{X}\hat{\boldsymbol{\beta}}$ is **orthogonal** to $\mathcal{C}(\mathbb{X})$

Projection



- Finding closest point in $\mathcal{C}(\mathbb{X})$ to $\mathbf{Y}$ is called **projection**
- Example: $n = 3$ and $k = 2$: points in 3D space.
  - Column space of $\mathbb{X}$ is a plane in this space.
- Residual vector $\hat{\mathbf{e}} = \mathbf{Y} - \mathbb{X}\hat{\boldsymbol{\beta}}$ is **orthogonal** to $\mathcal{C}(\mathbb{X})$
  - Shortest distance from $\mathbf{Y}$ to $\mathcal{C}(\mathbb{X})$ is a straight line to the plane, which will be perpendicular to $\mathcal{C}(\mathbb{X})$.

Projection



- Finding closest point in $\mathcal{C}(\mathbb{X})$ to $\mathbf{Y}$ is called **projection**
- Example: $n = 3$ and $k = 2$: points in 3D space.
  - Column space of $\mathbb{X}$ is a plane in this space.
- Residual vector $\hat{\mathbf{e}} = \mathbf{Y} - \mathbb{X}\hat{\boldsymbol{\beta}}$ is **orthogonal** to $\mathcal{C}(\mathbb{X})$
  - Shortest distance from $\mathbf{Y}$ to $\mathcal{C}(\mathbb{X})$ is a straight line to the plane, which will be perpendicular to $\mathcal{C}(\mathbb{X})$.
  - Implies that $\mathbb{X}'\hat{\mathbf{e}} = 0$

Multicollinearity

- Hidden assumption: $\mathbb{X}'\mathbb{X} = \sum_{i=1}^{n} \mathbf{X}_i \mathbf{X}_i'$ is invertible.

Multicollinearity

- Hidden assumption: $\mathbb{X}'\mathbb{X} = \sum_{i=1}^{n} \mathbf{X}_i \mathbf{X}_i'$ is invertible.
    - Equivalent to $\mathbb{X}$ being **full column rank**.

Multicollinearity

- Hidden assumption: $\mathbb{X}'\mathbb{X} = \sum_{i=1}^{n} \mathbf{X}_i \mathbf{X}_i'$ is invertible.
  - Equivalent to $\mathbb{X}$ being **full column rank**.
  - Equivalent to columns of $\mathbb{X}$ being **linearly independent**

Multicollinearity

- Hidden assumption: $\mathbb{X}'\mathbb{X} = \sum_{i=1}^{n} \mathbf{X}_i\mathbf{X}_i'$ is invertible.
  - Equivalent to $\mathbb{X}$ being **full column rank**.
  - Equivalent to columns of $\mathbb{X}$ being **linearly independent**
- Full column rank if $\mathbb{X}\mathbf{b} = 0$ if and only if $\mathbf{b} = \mathbf{0}$.

$$b_1\mathbb{X}_1 + b_2\mathbb{X}_2 + \cdots + b_{k+1}\mathbb{X}_{k+1} = 0 \quad \iff \quad b_1 = b_2 = \cdots = b_{k+1} = 0,$$

Multicollinearity

- Hidden assumption: $\mathbb{X}'\mathbb{X} = \sum_{i=1}^{n} \mathbf{X}_i \mathbf{X}_i'$ is invertible.
    - Equivalent to $\mathbb{X}$ being **full column rank**.
    - Equivalent to columns of $\mathbb{X}$ being **linearly independent**
- Full column rank if $\mathbb{X}\mathbf{b} = 0$ if and only if $\mathbf{b} = \mathbf{0}$.

$$b_1\mathbb{X}_1 + b_2\mathbb{X}_2 + \cdots + b_{k+1}\mathbb{X}_{k+1} = 0 \quad \iff \quad b_1 = b_2 = \cdots = b_{k+1} = 0,$$

- Typically reasonable but can be violated by user error:

Multicollinearity

- Hidden assumption: $\mathbb{X}'\mathbb{X} = \sum_{i=1}^{n} \mathbf{X}_i \mathbf{X}_i'$ is invertible.
    - Equivalent to $\mathbb{X}$ being **full column rank**.
    - Equivalent to columns of $\mathbb{X}$ being **linearly independent**
- Full column rank if $\mathbb{X}\mathbf{b} = 0$ if and only if $\mathbf{b} = \mathbf{0}$.

$$b_1\mathbb{X}_1 + b_2\mathbb{X}_2 + \cdots + b_{k+1}\mathbb{X}_{k+1} = 0 \quad \iff \quad b_1 = b_2 = \cdots = b_{k+1} = 0,$$

- Typically reasonable but can be violated by user error:
    - Accidentally adding the same variable twice.

Multicollinearity

- Hidden assumption: $\mathbb{X}'\mathbb{X} = \sum_{i=1}^{n} \mathbf{X}_i \mathbf{X}_i'$ is invertible.
    - Equivalent to $\mathbb{X}$ being **full column rank**.
    - Equivalent to columns of $\mathbb{X}$ being **linearly independent**
- Full column rank if $\mathbb{X}\mathbf{b} = 0$ if and only if $\mathbf{b} = \mathbf{0}$.

$$b_1\mathbb{X}_1 + b_2\mathbb{X}_2 + \cdots + b_{k+1}\mathbb{X}_{k+1} = 0 \quad \Longleftrightarrow \quad b_1 = b_2 = \cdots = b_{k+1} = 0,$$

- Typically reasonable but can be violated by user error:
    - Accidentally adding the same variable twice.
    - Including all dummies for a categorical variable.

Multicollinearity

- Hidden assumption: $\mathbb{X}'\mathbb{X} = \sum_{i=1}^{n} \mathbf{X}_i \mathbf{X}_i'$ is invertible.
    - Equivalent to $\mathbb{X}$ being **full column rank**.
    - Equivalent to columns of $\mathbb{X}$ being **linearly independent**
- Full column rank if $\mathbb{X}\mathbf{b} = 0$ if and only if $\mathbf{b} = \mathbf{0}$.

$$b_1\mathbb{X}_1 + b_2\mathbb{X}_2 + \cdots + b_{k+1}\mathbb{X}_{k+1} = 0 \quad \iff \quad b_1 = b_2 = \cdots = b_{k+1} = 0,$$

- Typically reasonable but can be violated by user error:
    - Accidentally adding the same variable twice.
    - Including all dummies for a categorical variable.
    - Including fixed effects for group and variables that do not vary within groups.

4/ Partitioned regression and partial regression

Partitioned regression

- Partition covariates and coefficients $\mathbb{X} = [\mathbb{X}_1 \ \mathbb{X}_2]$ and $\boldsymbol{\beta} = (\boldsymbol{\beta}_1, \boldsymbol{\beta}_2)'$:

$$\mathbf{Y} = \mathbb{X}_1\boldsymbol{\beta}_1 + \mathbb{X}_2\boldsymbol{\beta}_2 + \mathbf{e}$$

Partitioned regression

- Partition covariates and coefficients $\mathbb{X} = [\mathbb{X}_1 \ \mathbb{X}_2]$ and $\boldsymbol{\beta} = (\boldsymbol{\beta}_1, \boldsymbol{\beta}_2)'$:

$$\mathbf{Y} = \mathbb{X}_1\boldsymbol{\beta}_1 + \mathbb{X}_2\boldsymbol{\beta}_2 + \mathbf{e}$$

- Can we find expressions for $\hat{\boldsymbol{\beta}}_1$ and $\hat{\boldsymbol{\beta}}_2$?

Partitioned regression

- Partition covariates and coefficients $\mathbb{X} = [\mathbb{X}_1 \ \mathbb{X}_2]$ and $\boldsymbol{\beta} = (\boldsymbol{\beta}_1, \boldsymbol{\beta}_2)'$:

$$\mathbf{Y} = \mathbb{X}_1\boldsymbol{\beta}_1 + \mathbb{X}_2\boldsymbol{\beta}_2 + \mathbf{e}$$

- Can we find expressions for $\hat{\boldsymbol{\beta}}_1$ and $\hat{\boldsymbol{\beta}}_2$?
- **Residual regression** or Frisch-Waugh-Lovell theorem to obtain $\hat{\boldsymbol{\beta}}_1$:

Partitioned regression

- Partition covariates and coefficients $\mathbb{X} = [\mathbb{X}_1 \; \mathbb{X}_2]$ and $\boldsymbol{\beta} = (\boldsymbol{\beta}_1, \boldsymbol{\beta}_2)'$:

$$\mathbf{Y} = \mathbb{X}_1\boldsymbol{\beta}_1 + \mathbb{X}_2\boldsymbol{\beta}_2 + \mathbf{e}$$

- Can we find expressions for $\hat{\boldsymbol{\beta}}_1$ and $\hat{\boldsymbol{\beta}}_2$?
- **Residual regression** or Frisch-Waugh-Lovell theorem to obtain $\hat{\boldsymbol{\beta}}_1$:
    - Use OLS to regress $\mathbf{Y}$ on $\mathbb{X}_2$ and obtain residuals $\tilde{\mathbf{e}}_2$.

Partitioned regression

- Partition covariates and coefficients $\mathbb{X} = [\mathbb{X}_1 \; \mathbb{X}_2]$ and $\boldsymbol{\beta} = (\boldsymbol{\beta}_1, \boldsymbol{\beta}_2)'$:

$$\mathbf{Y} = \mathbb{X}_1\boldsymbol{\beta}_1 + \mathbb{X}_2\boldsymbol{\beta}_2 + \mathbf{e}$$

- Can we find expressions for $\hat{\boldsymbol{\beta}}_1$ and $\hat{\boldsymbol{\beta}}_2$?
- **Residual regression** or Frisch-Waugh-Lovell theorem to obtain $\hat{\boldsymbol{\beta}}_1$:
  - Use OLS to regress $\mathbf{Y}$ on $\mathbb{X}_2$ and obtain residuals $\tilde{\mathbf{e}}_2$.
  - Use OLS to regress each column of $\mathbb{X}_1$ on $\mathbb{X}_2$ and obtain residuals $\widetilde{\mathbb{X}}_1$.

Partitioned regression

- Partition covariates and coefficients $\mathbb{X} = [\mathbb{X}_1\ \mathbb{X}_2]$ and $\boldsymbol{\beta} = (\boldsymbol{\beta}_1, \boldsymbol{\beta}_2)'$:

$$\mathbf{Y} = \mathbb{X}_1\boldsymbol{\beta}_1 + \mathbb{X}_2\boldsymbol{\beta}_2 + \mathbf{e}$$

- Can we find expressions for $\hat{\boldsymbol{\beta}}_1$ and $\hat{\boldsymbol{\beta}}_2$?
- **Residual regression** or Frisch-Waugh-Lovell theorem to obtain $\hat{\boldsymbol{\beta}}_1$:
    - Use OLS to regress $\mathbf{Y}$ on $\mathbb{X}_2$ and obtain residuals $\tilde{\mathbf{e}}_2$.
    - Use OLS to regress each column of $\mathbb{X}_1$ on $\mathbb{X}_2$ and obtain residuals $\widetilde{\mathbb{X}}_1$.
    - Use OLS to regress $\tilde{\mathbf{e}}_2$ on $\widetilde{\mathbb{X}}_1$

Focus on simple case

- Focus on single covariate model with no intercept: $Y_i = X_i\beta + e_i$

Focus on simple case

- Focus on single covariate model with no intercept: $Y_i = X_i\beta + e_i$
- Let $\mathbf{X} = (X_1, \dots, X_n)$ and recall inner product: $\langle \mathbf{X}, \mathbf{Y} \rangle = \sum_{i=1}^{n} X_i Y_i$

Focus on simple case

- Focus on single covariate model with no intercept: $Y_i = X_i\beta + e_i$
- Let $\mathbf{X} = (X_1, \dots, X_n)$ and recall inner product: $\langle \mathbf{X}, \mathbf{Y} \rangle = \sum_{i=1}^{n} X_i Y_i$
  - Inner products measure how similar two vectors are.

Focus on simple case

- Focus on single covariate model with no intercept: $Y_i = X_i\beta + e_i$
- Let $\mathbf{X} = (X_1, \dots, X_n)$ and recall inner product: $\langle \mathbf{X}, \mathbf{Y} \rangle = \sum_{i=1}^n X_i Y_i$
    - Inner products measure how similar two vectors are.
- Slope in this case:

$$\hat{\beta} = \frac{\sum_{i=1}^n X_i Y_i}{\sum_{i=1}^n X_i^2} = \frac{\langle \mathbf{X}, \mathbf{Y} \rangle}{\langle \mathbf{X}, \mathbf{X} \rangle}$$

Focus on simple case

- Focus on single covariate model with no intercept: $Y_i = X_i\beta + e_i$
- Let $\mathbf{X} = (X_1, \dots, X_n)$ and recall inner product: $\langle \mathbf{X}, \mathbf{Y} \rangle = \sum_{i=1}^n X_i Y_i$
    - Inner products measure how similar two vectors are.
- Slope in this case:

$$\hat{\beta} = \frac{\sum_{i=1}^n X_i Y_i}{\sum_{i=1}^n X_i^2} = \frac{\langle \mathbf{X}, \mathbf{Y} \rangle}{\langle \mathbf{X}, \mathbf{X} \rangle}$$

- Suppose we add an **orthogonal covariate** $\mathbf{Y} = \mathbf{X}\beta + \mathbf{Z}\gamma + \mathbf{e}$ with $\langle \mathbf{X}, \mathbf{Z} \rangle = 0$.

$$\hat{\beta} = \frac{\langle \mathbf{X}, \mathbf{Y} \rangle}{\langle \mathbf{X}, \mathbf{X} \rangle} \quad \widehat{\gamma} = \frac{\langle \mathbf{Z}, \mathbf{Y} \rangle}{\langle \mathbf{Z}, \mathbf{Z} \rangle}$$

Focus on simple case

- Focus on single covariate model with no intercept: $Y_i = X_i\beta + e_i$
- Let $\mathbf{X} = (X_1, \dots, X_n)$ and recall inner product: $\langle \mathbf{X}, \mathbf{Y} \rangle = \sum_{i=1}^n X_i Y_i$
    - Inner products measure how similar two vectors are.
- Slope in this case:

$$\hat{\beta} = \frac{\sum_{i=1}^n X_i Y_i}{\sum_{i=1}^n X_i^2} = \frac{\langle \mathbf{X}, \mathbf{Y} \rangle}{\langle \mathbf{X}, \mathbf{X} \rangle}$$

- Suppose we add an **orthogonal covariate** $\mathbf{Y} = \mathbf{X}\beta + \mathbf{Z}\gamma + \mathbf{e}$ with $\langle \mathbf{X}, \mathbf{Z} \rangle = 0$.

$$\hat{\beta} = \frac{\langle \mathbf{X}, \mathbf{Y} \rangle}{\langle \mathbf{X}, \mathbf{X} \rangle} \quad \widehat{\gamma} = \frac{\langle \mathbf{Z}, \mathbf{Y} \rangle}{\langle \mathbf{Z}, \mathbf{Z} \rangle}$$

    - With exactly orthogonal covariates, multivariate OLS is the same as univariate OLS.

Focus on simple case

- Focus on single covariate model with no intercept: $Y_i = X_i\beta + e_i$
- Let $\mathbf{X} = (X_1, \ldots, X_n)$ and recall inner product: $\langle \mathbf{X}, \mathbf{Y} \rangle = \sum_{i=1}^n X_i Y_i$
    - Inner products measure how similar two vectors are.
- Slope in this case:

$$\hat{\beta} = \frac{\sum_{i=1}^n X_i Y_i}{\sum_{i=1}^n X_i^2} = \frac{\langle \mathbf{X}, \mathbf{Y} \rangle}{\langle \mathbf{X}, \mathbf{X} \rangle}$$

- Suppose we add an **orthogonal covariate** $\mathbf{Y} = \mathbf{X}\beta + \mathbf{Z}\gamma + \mathbf{e}$ with $\langle \mathbf{X}, \mathbf{Z} \rangle = 0$.

$$\hat{\beta} = \frac{\langle \mathbf{X}, \mathbf{Y} \rangle}{\langle \mathbf{X}, \mathbf{X} \rangle} \quad \widehat{\gamma} = \frac{\langle \mathbf{Z}, \mathbf{Y} \rangle}{\langle \mathbf{Z}, \mathbf{Z} \rangle}$$

- With exactly orthogonal covariates, multivariate OLS is the same as univariate OLS.
- Only holds in balanced, designed experiments.

Adding the intercept

- Consider the OLS slope with an intercept:

$$\hat{\beta} = \frac{\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})}{\sum_{i=1}^{n}(X_i - \overline{X})} = \frac{\langle \mathbf{X} - \overline{X}\mathbf{1}, \mathbf{Y} - \overline{Y}\mathbf{1} \rangle}{\langle \mathbf{X} - \overline{X}\mathbf{1}, \mathbf{X} - \overline{X}\mathbf{1} \rangle} = \frac{\langle \mathbf{X} - \overline{X}\mathbf{1}, \mathbf{Y} \rangle}{\langle \mathbf{X} - \overline{X}\mathbf{1}, \mathbf{X} - \overline{X}\mathbf{1} \rangle}$$

Adding the intercept

- Consider the OLS slope with an intercept:

$$\hat{\beta} = \frac{\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})}{\sum_{i=1}^{n}(X_i - \overline{X})} = \frac{\langle \mathbf{X} - \overline{X}\mathbf{1}, \mathbf{Y} - \overline{Y}\mathbf{1} \rangle}{\langle \mathbf{X} - \overline{X}\mathbf{1}, \mathbf{X} - \overline{X}\mathbf{1} \rangle} = \frac{\langle \mathbf{X} - \overline{X}\mathbf{1}, \mathbf{Y} \rangle}{\langle \mathbf{X} - \overline{X}\mathbf{1}, \mathbf{X} - \overline{X}\mathbf{1} \rangle}$$

- How can we get this?

Adding the intercept

- Consider the OLS slope with an intercept:

$$\hat{\beta} = \frac{\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})}{\sum_{i=1}^{n}(X_i - \overline{X})} = \frac{\langle \mathbf{X} - \overline{X}\mathbf{1}, \mathbf{Y} - \overline{Y}\mathbf{1} \rangle}{\langle \mathbf{X} - \overline{X}\mathbf{1}, \mathbf{X} - \overline{X}\mathbf{1} \rangle} = \frac{\langle \mathbf{X} - \overline{X}\mathbf{1}, \mathbf{Y} \rangle}{\langle \mathbf{X} - \overline{X}\mathbf{1}, \mathbf{X} - \overline{X}\mathbf{1} \rangle}$$

- How can we get this?
    1. Regress $\mathbf{X}$ on $\mathbf{1}$ to get coefficient $\overline{X}$

Adding the intercept

- Consider the OLS slope with an intercept:

$$\hat{\beta} = \frac{\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})}{\sum_{i=1}^{n}(X_i - \overline{X})} = \frac{\langle \mathbf{X} - \overline{X}\mathbf{1}, \mathbf{Y} - \overline{Y}\mathbf{1} \rangle}{\langle \mathbf{X} - \overline{X}\mathbf{1}, \mathbf{X} - \overline{X}\mathbf{1} \rangle} = \frac{\langle \mathbf{X} - \overline{X}\mathbf{1}, \mathbf{Y} \rangle}{\langle \mathbf{X} - \overline{X}\mathbf{1}, \mathbf{X} - \overline{X}\mathbf{1} \rangle}$$

- How can we get this?
    1. Regress $\mathbf{X}$ on $\mathbf{1}$ to get coefficient $\overline{X}$
    2. Regress $\mathbf{Y}$ on residuals from step 1, $\mathbf{X} - \overline{X}\mathbf{1}$

Adding the intercept

- Consider the OLS slope with an intercept:

$$\hat{\beta} = \frac{\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})}{\sum_{i=1}^{n}(X_i - \overline{X})} = \frac{\langle \mathbf{X} - \overline{X}\mathbf{1}, \mathbf{Y} - \overline{Y}\mathbf{1} \rangle}{\langle \mathbf{X} - \overline{X}\mathbf{1}, \mathbf{X} - \overline{X}\mathbf{1} \rangle} = \frac{\langle \mathbf{X} - \overline{X}\mathbf{1}, \mathbf{Y} \rangle}{\langle \mathbf{X} - \overline{X}\mathbf{1}, \mathbf{X} - \overline{X}\mathbf{1} \rangle}$$

- How can we get this?
    1. Regress $\mathbf{X}$ on $\mathbf{1}$ to get coefficient $\overline{X}$
    2. Regress $\mathbf{Y}$ on residuals from step 1, $\mathbf{X} - \overline{X}\mathbf{1}$
- If wanted to get coefficient on added variable Z_i, we could repeat this:

Adding the intercept

- Consider the OLS slope with an intercept:

$$\hat{\beta} = \frac{\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})}{\sum_{i=1}^{n}(X_i - \overline{X})} = \frac{\langle \mathbf{X} - \overline{X}\mathbf{1}, \mathbf{Y} - \overline{Y}\mathbf{1} \rangle}{\langle \mathbf{X} - \overline{X}\mathbf{1}, \mathbf{X} - \overline{X}\mathbf{1} \rangle} = \frac{\langle \mathbf{X} - \overline{X}\mathbf{1}, \mathbf{Y} \rangle}{\langle \mathbf{X} - \overline{X}\mathbf{1}, \mathbf{X} - \overline{X}\mathbf{1} \rangle}$$

- How can we get this?

    1. Regress $\mathbf{X}$ on $\mathbf{1}$ to get coefficient $\overline{X}$
    2. Regress $\mathbf{Y}$ on residuals from step 1, $\mathbf{X} - \overline{X}\mathbf{1}$

- If wanted to get coefficient on added variable Z_i, we could repeat this:

    1. Regress $\mathbf{Z}$ on $\widetilde{\mathbf{X}} = \mathbf{X} - \overline{X}\mathbf{1}$ on and obtain coefficient $\langle \mathbf{Z}, \widetilde{\mathbf{X}} \rangle / \langle \widetilde{\mathbf{X}}, \widetilde{\mathbf{X}} \rangle$

Adding the intercept

- Consider the OLS slope with an intercept:

$$\hat{\beta} = \frac{\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})}{\sum_{i=1}^{n}(X_i - \overline{X})} = \frac{\langle \mathbf{X} - \overline{X}\mathbf{1}, \mathbf{Y} - \overline{Y}\mathbf{1} \rangle}{\langle \mathbf{X} - \overline{X}\mathbf{1}, \mathbf{X} - \overline{X}\mathbf{1} \rangle} = \frac{\langle \mathbf{X} - \overline{X}\mathbf{1}, \mathbf{Y} \rangle}{\langle \mathbf{X} - \overline{X}\mathbf{1}, \mathbf{X} - \overline{X}\mathbf{1} \rangle}$$

- How can we get this?

  1. Regress $\mathbf{X}$ on $\mathbf{1}$ to get coefficient $\overline{X}$
  2. Regress $\mathbf{Y}$ on residuals from step 1, $\mathbf{X} - \overline{X}\mathbf{1}$

- If wanted to get coefficient on added variable Z_i, we could repeat this:

  1. Regress $\mathbf{Z}$ on $\widetilde{\mathbf{X}} = \mathbf{X} - \overline{X}\mathbf{1}$ on and obtain coefficient $\langle \mathbf{Z}, \widetilde{\mathbf{X}} \rangle / \langle \widetilde{\mathbf{X}}, \widetilde{\mathbf{X}} \rangle$
  2. Regress $\mathbf{Y}$ on residual from

Visualizing orthogonalization

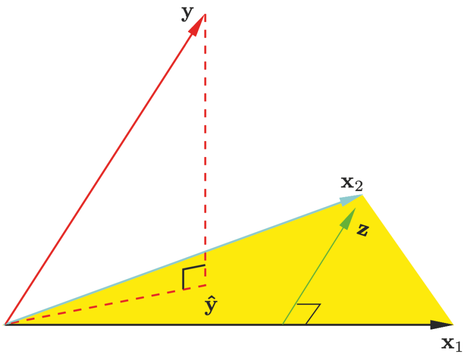


FIGURE 3.4. *Least squares regression by orthogonalization of the inputs. The vector* $\mathbf{x}_2$ *is regressed on the vector* $\mathbf{x}_1$*, leaving the residual vector* $\mathbf{z}$*. The regression of* $\mathbf{y}$ *on* $\mathbf{z}$ *gives the multiple regression coefficient of* $\mathbf{x}_2$*. Adding together the projections of* $\mathbf{y}$ *on each of* $\mathbf{x}_1$ *and* $\mathbf{z}$ *gives the least squares fit* $\hat{\mathbf{y}}$*.*

Why does residual regression work?

- We can find $\hat{\boldsymbol{\beta}}_1$ by nested minimization:

$$\hat{\boldsymbol{\beta}}_1 = \arg\min_{\boldsymbol{\beta}_1} \left(\min_{\boldsymbol{\beta}_2} \|\mathbf{Y} - \mathbb{X}_1\boldsymbol{\beta}_1 - \mathbb{X}_2\boldsymbol{\beta}_2\|^2 \right)$$

Why does residual regression work?

- We can find $\hat{\boldsymbol{\beta}}_1$ by nested minimization:

$$\hat{\boldsymbol{\beta}}_1 = \arg\min_{\boldsymbol{\beta}_1} \left(\min_{\boldsymbol{\beta}_2} \|\mathbf{Y} - \mathbb{X}_1\boldsymbol{\beta}_1 - \mathbb{X}_2\boldsymbol{\beta}_2\|^2 \right)$$

    - First find the minimum of the SSR over $\boldsymbol{\beta}_2$ fixing $\boldsymbol{\beta}_1$

Why does residual regression work?

- We can find $\hat{\boldsymbol{\beta}}_1$ by nested minimization:

$$\hat{\boldsymbol{\beta}}_1 = \arg\min_{\boldsymbol{\beta}_1} \left(\min_{\boldsymbol{\beta}_2} \|\mathbf{Y} - \mathbb{X}_1\boldsymbol{\beta}_1 - \mathbb{X}_2\boldsymbol{\beta}_2\|^2 \right)$$

  - First find the minimum of the SSR over $\boldsymbol{\beta}_2$ fixing $\boldsymbol{\beta}_1$
  - Then find $\boldsymbol{\beta}_1$ that minimizes the resulting SSR.

Why does residual regression work?

- We can find $\widehat{\boldsymbol{\beta}}_1$ by nested minimization:

$$\widehat{\boldsymbol{\beta}}_1 = \arg\min_{\boldsymbol{\beta}_1} \left(\min_{\boldsymbol{\beta}_2} \|\mathbf{Y} - \mathbb{X}_1\boldsymbol{\beta}_1 - \mathbb{X}_2\boldsymbol{\beta}_2\|^2 \right)$$

  - First find the minimum of the SSR over $\boldsymbol{\beta}_2$ fixing $\boldsymbol{\beta}_1$
  - Then find $\boldsymbol{\beta}_1$ that minimizes the resulting SSR.

- The projection and annihilator matrices are defined only by covariates.

Why does residual regression work?

- We can find $\hat{\boldsymbol{\beta}}_1$ by nested minimization:

$$\hat{\boldsymbol{\beta}}_1 = \arg\min_{\boldsymbol{\beta}_1} \left(\min_{\boldsymbol{\beta}_2} \|\mathbf{Y} - \mathbb{X}_1\boldsymbol{\beta}_1 - \mathbb{X}_2\boldsymbol{\beta}_2\|^2 \right)$$

  - First find the minimum of the SSR over $\boldsymbol{\beta}_2$ fixing $\boldsymbol{\beta}_1$
  - Then find $\boldsymbol{\beta}_1$ that minimizes the resulting SSR.

- The projection and annihilator matrices are defined only by covariates.

  - $\mathbf{M}_2 = \mathbf{I}_n - \mathbb{X}_2(\mathbb{X}_2'\mathbb{X}_2)^{-1}\mathbb{X}_2'$

Why does residual regression work?

- We can find $\hat{\boldsymbol{\beta}}_1$ by nested minimization:

$$\hat{\boldsymbol{\beta}}_1 = \underset{\boldsymbol{\beta}_1}{\arg\min} \left(\min_{\boldsymbol{\beta}_2} \|\mathbf{Y} - \mathbb{X}_1\boldsymbol{\beta}_1 - \mathbb{X}_2\boldsymbol{\beta}_2\|^2 \right)$$

    - First find the minimum of the SSR over $\boldsymbol{\beta}_2$ fixing $\boldsymbol{\beta}_1$
    - Then find $\boldsymbol{\beta}_1$ that minimizes the resulting SSR.

- The projection and annihilator matrices are defined only by covariates.

    - $\mathbf{M}_2 = \mathbf{I}_n - \mathbb{X}_2(\mathbb{X}_2'\mathbb{X}_2)^{-1}\mathbb{X}_2'$
    - Creates residuals from a regression on or $\mathbb{X}_2$

Why does residual regression work?

- We can find $\hat{\boldsymbol{\beta}}_1$ by nested minimization:

$$\hat{\boldsymbol{\beta}}_1 = \arg\min_{\boldsymbol{\beta}_1} \left(\min_{\boldsymbol{\beta}_2} \|\mathbf{Y} - \mathbb{X}_1\boldsymbol{\beta}_1 - \mathbb{X}_2\boldsymbol{\beta}_2\|^2 \right)$$

  - First find the minimum of the SSR over $\boldsymbol{\beta}_2$ fixing $\boldsymbol{\beta}_1$
  - Then find $\boldsymbol{\beta}_1$ that minimizes the resulting SSR.

- The projection and annihilator matrices are defined only by covariates.

  - $\mathbf{M}_2 = \mathbf{I}_n - \mathbb{X}_2(\mathbb{X}_2'\mathbb{X}_2)^{-1}\mathbb{X}_2'$
  - Creates residuals from a regression on or $\mathbb{X}_2$

- Solving the nested minimization gives:

$$\hat{\boldsymbol{\beta}}_1 = \left(\mathbb{X}_1'\mathbf{M}_2\mathbb{X}_1\right)^{-1}\left(\mathbb{X}_1'\mathbf{M}_2\mathbf{Y}\right)$$

Why does residual regression work?

- We can find $\widehat{\boldsymbol{\beta}}_1$ by nested minimization:

$$\widehat{\boldsymbol{\beta}}_1 = \arg\min_{\boldsymbol{\beta}_1} \left(\min_{\boldsymbol{\beta}_2} \|\mathbf{Y} - \mathbb{X}_1\boldsymbol{\beta}_1 - \mathbb{X}_2\boldsymbol{\beta}_2\|^2 \right)$$

  - First find the minimum of the SSR over $\boldsymbol{\beta}_2$ fixing $\boldsymbol{\beta}_1$
  - Then find $\boldsymbol{\beta}_1$ that minimizes the resulting SSR.
- The projection and annihilator matrices are defined only by covariates.
  - $\mathbf{M}_2 = \mathbf{I}_n - \mathbb{X}_2(\mathbb{X}_2'\mathbb{X}_2)^{-1}\mathbb{X}_2'$
  - Creates residuals from a regression on or $\mathbb{X}_2$
- Solving the nested minimization gives:

$$\widehat{\boldsymbol{\beta}}_1 = (\mathbb{X}_1'\mathbf{M}_2\mathbb{X}_1)^{-1} (\mathbb{X}_1'\mathbf{M}_2\mathbf{Y})$$

- When will $\widehat{\boldsymbol{\beta}}_1$ will be the same regardless of whether $\mathbb{X}_2$ is included?

Why does residual regression work?

- We can find $\widehat{\boldsymbol{\beta}}_1$ by nested minimization:

$$\widehat{\boldsymbol{\beta}}_1 = \arg\min_{\boldsymbol{\beta}_1} \left(\min_{\boldsymbol{\beta}_2} \|\mathbf{Y} - \mathbb{X}_1\boldsymbol{\beta}_1 - \mathbb{X}_2\boldsymbol{\beta}_2\|^2 \right)$$

  - First find the minimum of the SSR over $\boldsymbol{\beta}_2$ fixing $\boldsymbol{\beta}_1$
  - Then find $\boldsymbol{\beta}_1$ that minimizes the resulting SSR.

- The projection and annihilator matrices are defined only by covariates.
  - $\mathbf{M}_2 = \mathbf{I}_n - \mathbb{X}_2(\mathbb{X}_2'\mathbb{X}_2)^{-1}\mathbb{X}_2'$
  - Creates residuals from a regression on or $\mathbb{X}_2$

- Solving the nested minimization gives:

$$\widehat{\boldsymbol{\beta}}_1 = (\mathbb{X}_1'\mathbf{M}_2\mathbb{X}_1)^{-1} (\mathbb{X}_1'\mathbf{M}_2\mathbf{Y})$$

- When will $\widehat{\boldsymbol{\beta}}_1$ will be the same regardless of whether $\mathbb{X}_2$ is included?
  - If $\mathbb{X}_1$ and $\mathbb{X}_2$ are orthogonal so $\mathbb{X}_2'\mathbb{X}_1 = 0$ so $\mathbf{M}_2\mathbb{X}_1 = \mathbb{X}_1$

Residual regression

- Define two sets of residuals:

Residual regression

- Define two sets of residuals:
  - $\widetilde{\mathbb{X}}_2 = \mathbf{M}_1\mathbb{X}_2$ = residuals from regression of $\mathbb{X}_2$ on $\mathbb{X}_1$

Residual regression

- Define two sets of residuals:
  - $\widetilde{\mathbb{X}}_2 = \mathbf{M}_1\mathbb{X}_2$ = residuals from regression of $\mathbb{X}_2$ on $\mathbb{X}_1$
  - $\tilde{\mathbf{e}}_1 = \mathbf{M}_1\mathbf{Y}$ = residuals from regression of $\mathbf{Y}$ on $\mathbb{X}_1$.

Residual regression

- Define two sets of residuals:
  - $\widetilde{\mathbb{X}}_2 = \mathbf{M}_1 \mathbb{X}_2$ = residuals from regression of $\mathbb{X}_2$ on $\mathbb{X}_1$
  - $\tilde{\mathbf{e}}_1 = \mathbf{M}_1 \mathbf{Y}$ = residuals from regression of $\mathbf{Y}$ on $\mathbb{X}_1$.
- Then remembering that $\mathbf{M}_1$ is symmetric and idempotent:

$$\begin{aligned}
\hat{\boldsymbol{\beta}}_2 &= \left(\mathbb{X}_2' \mathbf{M}_1 \mathbb{X}_2\right)^{-1} \left(\mathbb{X}_2' \mathbf{M}_1 \mathbf{Y}\right) \\
&= \left(\mathbb{X}_2' \mathbf{M}_1 \mathbf{M}_1 \mathbb{X}_2\right)^{-1} \left(\mathbb{X}_2' \mathbf{M}_1 \mathbf{M}_1 \mathbf{Y}\right) \\
&= \left(\widetilde{\mathbb{X}}_2' \widetilde{\mathbb{X}}_2\right)^{-1} \left(\widetilde{\mathbb{X}}_2' \tilde{\mathbf{e}}_1\right)
\end{aligned}$$

Residual regression

- Define two sets of residuals:
  - $\widetilde{\mathbb{X}}_2 = \mathbf{M}_1\mathbb{X}_2$ = residuals from regression of $\mathbb{X}_2$ on $\mathbb{X}_1$
  - $\tilde{\mathbf{e}}_1 = \mathbf{M}_1\mathbf{Y}$ = residuals from regression of $\mathbf{Y}$ on $\mathbb{X}_1$.
- Then remembering that $\mathbf{M}_1$ is symmetric and idempotent:

$$
\begin{aligned}
\hat{\boldsymbol{\beta}}_2 &= \left(\mathbb{X}_2'\mathbf{M}_1\mathbb{X}_2\right)^{-1}\left(\mathbb{X}_2'\mathbf{M}_1\mathbf{Y}\right) \\
&= \left(\mathbb{X}_2'\mathbf{M}_1\mathbf{M}_1\mathbb{X}_2\right)^{-1}\left(\mathbb{X}_2'\mathbf{M}_1\mathbf{M}_1\mathbf{Y}\right) \\
&= \left(\widetilde{\mathbb{X}}_2'\widetilde{\mathbb{X}}_2\right)^{-1}\left(\widetilde{\mathbb{X}}_2'\tilde{\mathbf{e}}_1\right)
\end{aligned}
$$

- $\hat{\boldsymbol{\beta}}_2$ can be obtained from a regression of $\tilde{\mathbf{e}}_1$ on $\widetilde{\mathbb{X}}_2$.

Residual regression

- Define two sets of residuals:
  - $\widetilde{\mathbb{X}}_2 = \mathbf{M}_1\mathbb{X}_2$ = residuals from regression of $\mathbb{X}_2$ on $\mathbb{X}_1$
  - $\tilde{\mathbf{e}}_1 = \mathbf{M}_1\mathbf{Y}$ = residuals from regression of $\mathbf{Y}$ on $\mathbb{X}_1$.
- Then remembering that $\mathbf{M}_1$ is symmetric and idempotent:

$$\begin{aligned}\hat{\boldsymbol{\beta}}_2 &= \left(\mathbb{X}_2'\mathbf{M}_1\mathbb{X}_2\right)^{-1}\left(\mathbb{X}_2'\mathbf{M}_1\mathbf{Y}\right)\\ &= \left(\mathbb{X}_2'\mathbf{M}_1\mathbf{M}_1\mathbb{X}_2\right)^{-1}\left(\mathbb{X}_2'\mathbf{M}_1\mathbf{M}_1\mathbf{Y}\right)\\ &= \left(\widetilde{\mathbb{X}}_2'\widetilde{\mathbb{X}}_2\right)^{-1}\left(\widetilde{\mathbb{X}}_2'\tilde{\mathbf{e}}_1\right)\end{aligned}$$

- $\hat{\boldsymbol{\beta}}_2$ can be obtained from a regression of $\tilde{\mathbf{e}}_1$ on $\widetilde{\mathbb{X}}_2$.
  - Same result applies when using $\mathbf{Y}$ in place of $\tilde{\mathbf{e}}_1$.

Residual regression

- Define two sets of residuals:
  - $\widetilde{\mathbb{X}}_2 = \mathbf{M}_1\mathbb{X}_2$ = residuals from regression of $\mathbb{X}_2$ on $\mathbb{X}_1$
  - $\tilde{\mathbf{e}}_1 = \mathbf{M}_1\mathbf{Y}$ = residuals from regression of $\mathbf{Y}$ on $\mathbb{X}_1$.
- Then remembering that $\mathbf{M}_1$ is symmetric and idempotent:

$$\begin{aligned}\hat{\boldsymbol{\beta}}_2 &= \left(\mathbb{X}_2'\mathbf{M}_1\mathbb{X}_2\right)^{-1}\left(\mathbb{X}_2'\mathbf{M}_1\mathbf{Y}\right) \\ &= \left(\mathbb{X}_2'\mathbf{M}_1\mathbf{M}_1\mathbb{X}_2\right)^{-1}\left(\mathbb{X}_2'\mathbf{M}_1\mathbf{M}_1\mathbf{Y}\right) \\ &= \left(\widetilde{\mathbb{X}}_2'\widetilde{\mathbb{X}}_2\right)^{-1}\left(\widetilde{\mathbb{X}}_2'\tilde{\mathbf{e}}_1\right)\end{aligned}$$

- $\hat{\boldsymbol{\beta}}_2$ can be obtained from a regression of $\tilde{\mathbf{e}}_1$ on $\widetilde{\mathbb{X}}_2$.
  - Same result applies when using $\mathbf{Y}$ in place of $\tilde{\mathbf{e}}_1$.
  - Intuition: residuals are orthogonal

Residual regression

- Define two sets of residuals:
  - $\widetilde{\mathbb{X}}_2 = \mathbf{M}_1\mathbb{X}_2$ = residuals from regression of $\mathbb{X}_2$ on $\mathbb{X}_1$
  - $\tilde{\mathbf{e}}_1 = \mathbf{M}_1\mathbf{Y}$ = residuals from regression of $\mathbf{Y}$ on $\mathbb{X}_1$.
- Then remembering that $\mathbf{M}_1$ is symmetric and idempotent:

$$\begin{aligned}\hat{\boldsymbol{\beta}}_2 &= \left(\mathbb{X}_2'\mathbf{M}_1\mathbb{X}_2\right)^{-1}\left(\mathbb{X}_2'\mathbf{M}_1\mathbf{Y}\right)\\ &= \left(\mathbb{X}_2'\mathbf{M}_1\mathbf{M}_1\mathbb{X}_2\right)^{-1}\left(\mathbb{X}_2'\mathbf{M}_1\mathbf{M}_1\mathbf{Y}\right)\\ &= \left(\widetilde{\mathbb{X}}_2'\widetilde{\mathbb{X}}_2\right)^{-1}\left(\widetilde{\mathbb{X}}_2'\tilde{\mathbf{e}}_1\right)\end{aligned}$$

- $\hat{\boldsymbol{\beta}}_2$ can be obtained from a regression of $\tilde{\mathbf{e}}_1$ on $\widetilde{\mathbb{X}}_2$.
  - Same result applies when using $\mathbf{Y}$ in place of $\tilde{\mathbf{e}}_1$.
  - Intuition: residuals are orthogonal
  - Called the **Frisch-Waugh-Lovell Theorem**

Residual regression

- Define two sets of residuals:
  - $\widetilde{\mathbb{X}}_2 = \mathbf{M}_1 \mathbb{X}_2$ = residuals from regression of $\mathbb{X}_2$ on $\mathbb{X}_1$
  - $\tilde{\mathbf{e}}_1 = \mathbf{M}_1 \mathbf{Y}$ = residuals from regression of $\mathbf{Y}$ on $\mathbb{X}_1$.
- Then remembering that $\mathbf{M}_1$ is symmetric and idempotent:

$$\begin{aligned}\hat{\boldsymbol{\beta}}_2 &= \left(\mathbb{X}_2' \mathbf{M}_1 \mathbb{X}_2\right)^{-1} \left(\mathbb{X}_2' \mathbf{M}_1 \mathbf{Y}\right) \\ &= \left(\mathbb{X}_2' \mathbf{M}_1 \mathbf{M}_1 \mathbb{X}_2\right)^{-1} \left(\mathbb{X}_2' \mathbf{M}_1 \mathbf{M}_1 \mathbf{Y}\right) \\ &= \left(\widetilde{\mathbb{X}}_2' \widetilde{\mathbb{X}}_2\right)^{-1} \left(\widetilde{\mathbb{X}}_2' \tilde{\mathbf{e}}_1\right)\end{aligned}$$

- $\hat{\boldsymbol{\beta}}_2$ can be obtained from a regression of $\tilde{\mathbf{e}}_1$ on $\widetilde{\mathbb{X}}_2$.
  - Same result applies when using $\mathbf{Y}$ in place of $\tilde{\mathbf{e}}_1$.
  - Intuition: residuals are orthogonal
  - Called the **Frisch-Waugh-Lovell Theorem**
  - Sample version of the results we saw for the linear projection.

5/ Influential observations

Outliers, leverage points, and influential observations

- Least square heavily penalizes large residuals.

Outliers, leverage points, and influential observations

- Least square heavily penalizes large residuals.
- Implies a just a few unusual observations can be extremely influential.

Outliers, leverage points, and influential observations

- Least square heavily penalizes large residuals.
- Implies a just a few unusual observations can be extremely influential.
    - Dropping them leads to large changes in the estimated $\hat{\boldsymbol{\beta}}$.

Outliers, leverage points, and influential observations

- Least square heavily penalizes large residuals.
- Implies a just a few unusual observations can be extremely influential.
    - Dropping them leads to large changes in the estimated $\hat{\boldsymbol{\beta}}$.
    - Not all “unusual” observations have the same effect, though.

Outliers, leverage points, and influential observations

- Least square heavily penalizes large residuals.
- Implies a just a few unusual observations can be extremely influential.
  - Dropping them leads to large changes in the estimated $\hat{\boldsymbol{\beta}}$.
  - Not all "unusual" observations have the same effect, though.
- Useful to categorize:

Outliers, leverage points, and influential observations

- Least square heavily penalizes large residuals.
- Implies a just a few unusual observations can be extremely influential.
    - Dropping them leads to large changes in the estimated $\hat{\boldsymbol{\beta}}$.
    - Not all "unusual" observations have the same effect, though.
- Useful to categorize:

    1. **Leverage point**: extreme in one X direction

Outliers, leverage points, and influential observations

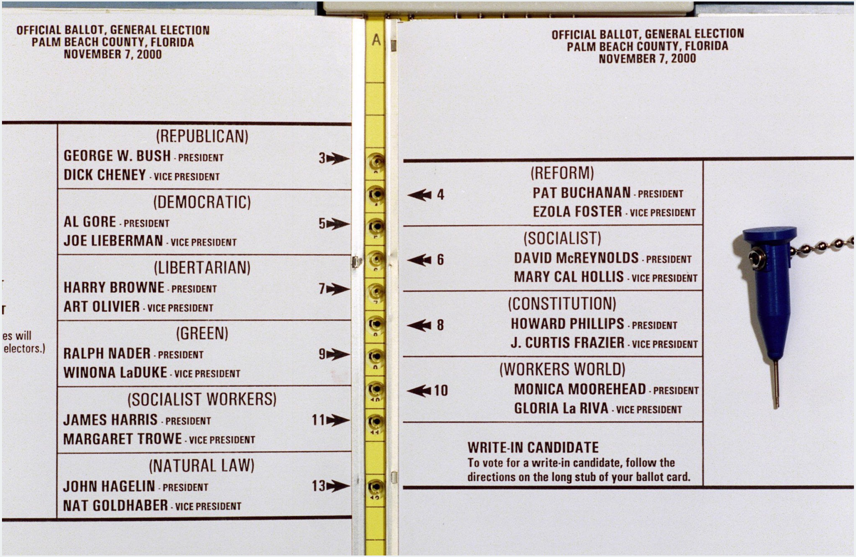
- Least square heavily penalizes large residuals.
- Implies a just a few unusual observations can be extremely influential.
  - Dropping them leads to large changes in the estimated $\hat{\boldsymbol{\beta}}$.
  - Not all "unusual" observations have the same effect, though.
- Useful to categorize:
  1. **Leverage point**: extreme in one X direction
  2. **Outlier**: extreme in the Y direction

Outliers, leverage points, and influential observations

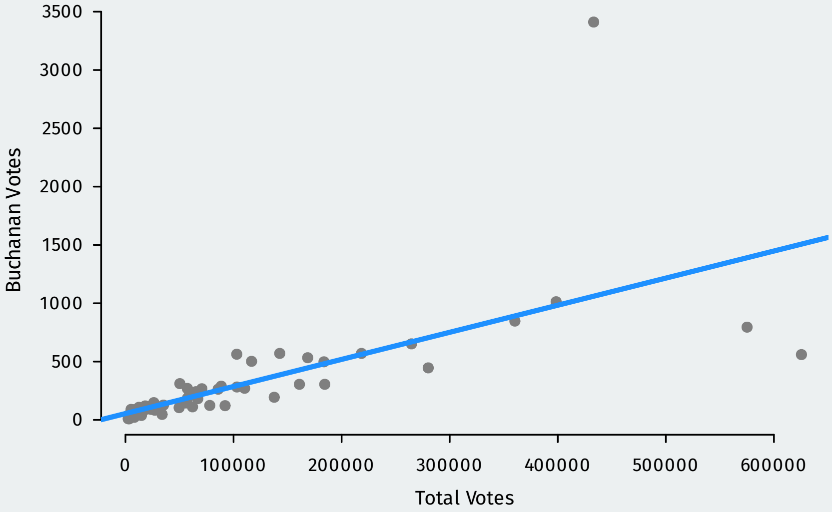
- Least square heavily penalizes large residuals.
- Implies a just a few unusual observations can be extremely influential.
  - Dropping them leads to large changes in the estimated $\hat{\boldsymbol{\beta}}$.
  - Not all "unusual" observations have the same effect, though.
- Useful to categorize:
  1. **Leverage point**: extreme in one X direction
  2. **Outlier**: extreme in the Y direction
  3. **Influence point**: extreme in both directions

Example: Buchanan votes in Florida, 2000

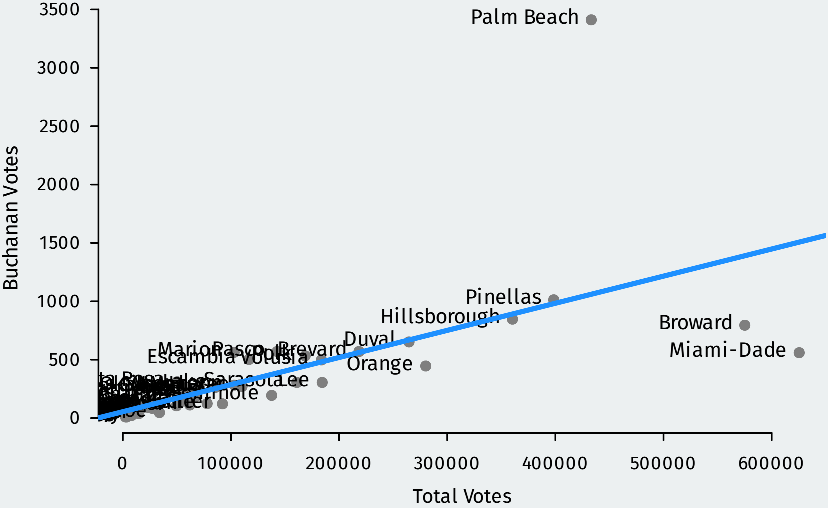
- 2000 Presidential election in FL (Wand et al., 2001, APSR)



Example: Buchanan votes in Florida, 2000



Example: Buchanan votes in Florida, 2000

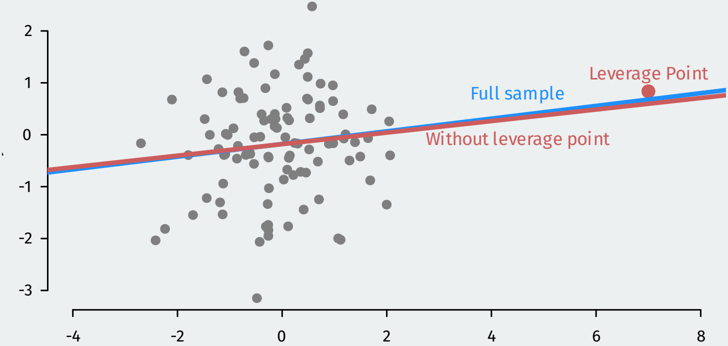


Example: Buchanan votes

```r
mod <- lm(edaybuchanan ~ edaytotal, data = flvote)
summary(mod)
```

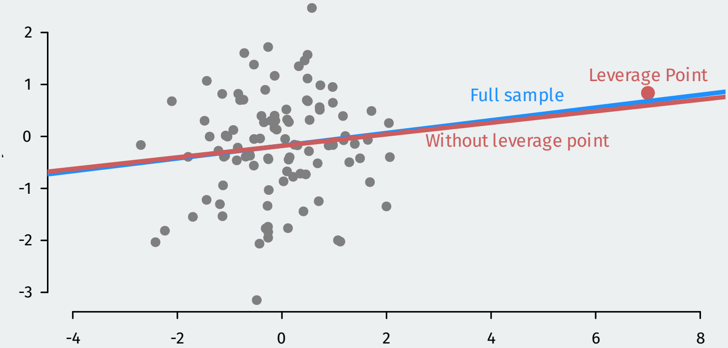
```
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 54.22945   49.14146    1.10     0.27
## edaytotal    0.00232    0.00031    7.48  2.4e-10 ***
## ---
## Signif. codes:
## 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 333 on 65 degrees of freedom
## Multiple R-squared:  0.463,  Adjusted R-squared:  0.455
## F-statistic:   56 on 1 and 65 DF,  p-value: 2.42e-10
```

Leverage point definition



- Values that are extreme in the X dimension

Leverage point definition



- Values that are extreme in the X dimension
- That is, values far from the center of the covariate distribution

Leverage values

- Let h_{ij} be the (i, j) entry of $\mathbf{P}$. Then:

$$\widehat{\mathbf{Y}} = \mathbf{PY} \qquad \Longrightarrow \qquad \widehat{Y}_i = \sum_{j=1}^{n} h_{ij} Y_j$$

Leverage values

- Let h_{ij} be the (i, j) entry of **P**. Then:

$$\widehat{\mathbf{Y}} = \mathbf{P}\mathbf{Y} \qquad \Longrightarrow \qquad \widehat{Y}_i = \sum_{j=1}^{n} h_{ij} Y_j$$

  - h_{ij} = importance of observation j is for the fitted value $\widehat{Y}_i$

Leverage values

- Let h_{ij} be the (i, j) entry of $\mathbf{P}$. Then:

$$\widehat{\mathbf{Y}} = \mathbf{P}\mathbf{Y} \qquad \Longrightarrow \qquad \widehat{Y}_i = \sum_{j=1}^{n} h_{ij} Y_j$$

  - h_{ij} = importance of observation j is for the fitted value $\widehat{Y}_i$

- **Leverage/hat values:** h_{ii} diagonal entries of the hat matrix

Leverage values

- Let h_{ij} be the (i, j) entry of $\mathbf{P}$. Then:

$$\widehat{\mathbf{Y}} = \mathbf{P}\mathbf{Y} \qquad \Longrightarrow \qquad \widehat{Y}_i = \sum_{j=1}^{n} h_{ij} Y_j$$

  - $h_{ij} =$ importance of observation j is for the fitted value $\widehat{Y}_i$
- **Leverage/hat values:** h_{ii} diagonal entries of the hat matrix
- With a simple linear regression, we have

$$h_{ii} = \frac{1}{n} + \frac{(X_i - \overline{X})^2}{\sum_{j=1}^{n}(X_j - \overline{X})^2}$$

Leverage values

- Let h_{ij} be the (i, j) entry of $\mathbf{P}$. Then:

$$\widehat{\mathbf{Y}} = \mathbf{PY} \qquad \Longrightarrow \qquad \widehat{Y}_i = \sum_{j=1}^{n} h_{ij} Y_j$$

  - h_{ij} = importance of observation j is for the fitted value $\widehat{Y}_i$
- **Leverage/hat values:** h_{ii} diagonal entries of the hat matrix
- With a simple linear regression, we have

$$h_{ii} = \frac{1}{n} + \frac{(X_i - \overline{X})^2}{\sum_{j=1}^{n}(X_j - \overline{X})^2}$$

  - $\rightsquigarrow$ how far i is from the center of the X distribution

Leverage values

- Let h_{ij} be the (i, j) entry of $\mathbf{P}$. Then:

$$\widehat{\mathbf{Y}} = \mathbf{P}\mathbf{Y} \qquad \Longrightarrow \qquad \widehat{Y}_i = \sum_{j=1}^{n} h_{ij} Y_j$$

  - h_{ij} = importance of observation j is for the fitted value $\widehat{Y}_i$
- **Leverage/hat values:** h_{ii} diagonal entries of the hat matrix
- With a simple linear regression, we have

$$h_{ii} = \frac{1}{n} + \frac{(X_i - \overline{X})^2}{\sum_{j=1}^{n}(X_j - \overline{X})^2}$$

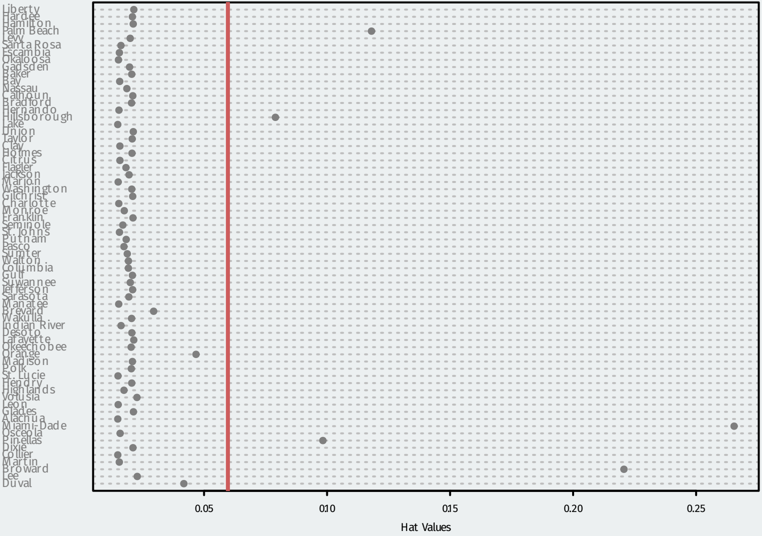
  - $\rightsquigarrow$ how far i is from the center of the X distribution
- **Rule of thumb:** examine hat values greater than $2(k+1)/n$

Buchanan hats

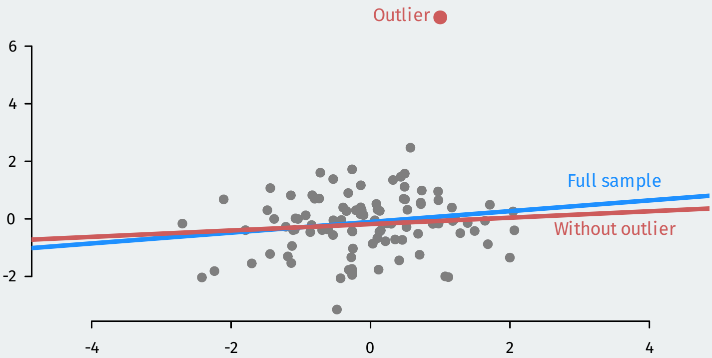
```r
head(hatvalues(mod), 5)
```

```
##      1      2      3      4      5
## 0.0418 0.0228 0.2207 0.0156 0.0149
```

Buchanan hats

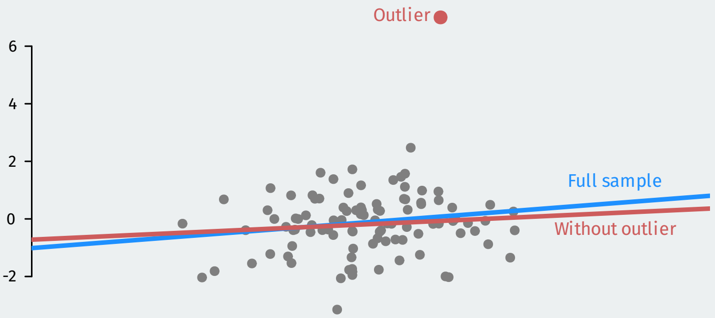


Outlier definition



- An **outlier** is far away from the center of the Y distribution.

Outlier definition



- An **outlier** is far away from the center of the Y distribution.
- Intuitively: a point that would be poorly predicted by the regression.

Detecting outliers

- Want values poorly predicted? Look for big residuals, right?

Detecting outliers

- Want values poorly predicted? Look for big residuals, right?
    - Problem: we use i to estimate $\hat{\boldsymbol{\beta}}$ so $\widehat{\mathbf{Y}}$ aren't valid predctions.

Detecting outliers

- Want values poorly predicted? Look for big residuals, right?
  - Problem: we use i to estimate $\hat{\boldsymbol{\beta}}$ so $\widehat{\mathbf{Y}}$ aren't valid predctions.
  - unit might pull the regression line toward itself $\rightsquigarrow$ small residual

Detecting outliers

- Want values poorly predicted? Look for big residuals, right?
  - Problem: we use i to estimate $\widehat{\boldsymbol{\beta}}$ so $\widehat{\mathbf{Y}}$ aren't valid predctions.
  - unit might pull the regression line toward itself $\rightsquigarrow$ small residual
- Better: **leave-one-out prediction errors**,

Detecting outliers

- Want values poorly predicted? Look for big residuals, right?
    - Problem: we use i to estimate $\hat{\boldsymbol{\beta}}$ so $\widehat{\mathbf{Y}}$ aren't valid predctions.
    - unit might pull the regression line toward itself $\rightsquigarrow$ small residual
- Better: **leave-one-out prediction errors**,
    1. Regress $\mathbf{Y}_{(-i)}$ on $\mathbb{X}_{(-i)}$, where these omit unit i:

$$\hat{\boldsymbol{\beta}}_{(-i)} = \left(\mathbb{X}'_{(-i)}\mathbb{X}_{(-i)}\right)^{-1}\mathbb{X}_{(-i)}\mathbf{Y}_{(-i)}$$

Detecting outliers

- Want values poorly predicted? Look for big residuals, right?
  - Problem: we use i to estimate $\hat{\boldsymbol{\beta}}$ so $\widehat{\mathbf{Y}}$ aren't valid predctions.
  - unit might pull the regression line toward itself $\rightsquigarrow$ small residual
- Better: **leave-one-out prediction errors**,

1. Regress $\mathbf{Y}_{(-i)}$ on $\mathbb{X}_{(-i)}$, where these omit unit i:

$$\hat{\boldsymbol{\beta}}_{(-i)} = \left(\mathbb{X}'_{(-i)}\mathbb{X}_{(-i)}\right)^{-1}\mathbb{X}_{(-i)}\mathbf{Y}_{(-i)}$$

2. Calculate predicted value of Y_i using that regression: $\widetilde{Y}_i = \mathbf{X}'_i\hat{\boldsymbol{\beta}}_{(-i)}$

Detecting outliers

- Want values poorly predicted? Look for big residuals, right?
    - Problem: we use i to estimate $\hat{\boldsymbol{\beta}}$ so $\widehat{\mathbf{Y}}$ aren't valid predctions.
    - unit might pull the regression line toward itself $\rightsquigarrow$ small residual
- Better: **leave-one-out prediction errors**,
    1. Regress $\mathbf{Y}_{(-i)}$ on $\mathbb{X}_{(-i)}$, where these omit unit i:

$$\hat{\boldsymbol{\beta}}_{(-i)} = \left(\mathbb{X}'_{(-i)}\mathbb{X}_{(-i)}\right)^{-1}\mathbb{X}_{(-i)}\mathbf{Y}_{(-i)}$$

    2. Calculate predicted value of Y_i using that regression: $\widetilde{Y}_i = \mathbf{X}_i'\hat{\boldsymbol{\beta}}_{(-i)}$
    3. Calculate prediction error: $\tilde{e}_i = Y_i - \widetilde{Y}_i$

Detecting outliers

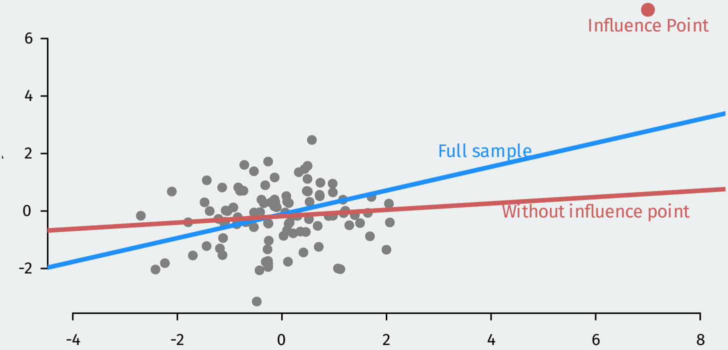
- Want values poorly predicted? Look for big residuals, right?
    - Problem: we use i to estimate $\hat{\boldsymbol{\beta}}$ so $\widehat{\mathbf{Y}}$ aren't valid predctions.
    - unit might pull the regression line toward itself $\rightsquigarrow$ small residual
- Better: **leave-one-out prediction errors**,
    1. Regress $\mathbf{Y}_{(-i)}$ on $\mathbb{X}_{(-i)}$, where these omit unit i:

$$\hat{\boldsymbol{\beta}}_{(-i)} = \left(\mathbb{X}'_{(-i)}\mathbb{X}_{(-i)}\right)^{-1}\mathbb{X}_{(-i)}\mathbf{Y}_{(-i)}$$

    2. Calculate predicted value of Y_i using that regression: $\widetilde{Y}_i = \mathbf{X}_i'\hat{\boldsymbol{\beta}}_{(-i)}$
    3. Calculate prediction error: $\tilde{e}_i = Y_i - \widetilde{Y}_i$
- Simple closed-form expressions:

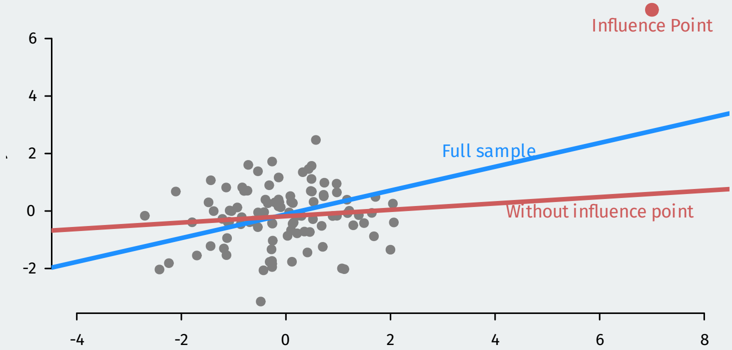
$$\hat{\boldsymbol{\beta}}_{(-i)} = \hat{\boldsymbol{\beta}} - \left(\mathbb{X}'\mathbb{X}\right)^{-1}\mathbf{X}_i\tilde{e}_i \qquad \tilde{e}_i = \frac{\hat{e}_i}{1 - h_{ii}}$$

Influence points



- An **influence point** is one that is both an outlier and a leverage point.

Influence points



- An **influence point** is one that is both an outlier and a leverage point.
- Extreme in both the X and Y dimensions

Overall measures of influence

- Influence of i can be measured by change in predictions:

$$\widehat{Y}_i - \widetilde{Y}_i = h_{ii}\tilde{e}_i$$

Overall measures of influence

- Influence of i can be measured by change in predictions:

$$\widehat{Y}_i - \widetilde{Y}_i = h_{ii}\tilde{e}_i$$

  - How much does excluding i from the regression change its predicted value?

Overall measures of influence

- Influence of i can be measured by change in predictions:

$$\widehat{Y}_i - \widetilde{Y}_i = h_{ii}\tilde{e}_i$$

  - How much does excluding i from the regression change its predicted value?
  - Equal to "leverage $\times$ outlier-ness"

Overall measures of influence

- Influence of i can be measured by change in predictions:

$$\widehat{Y}_i - \widetilde{Y}_i = h_{ii}\tilde{e}_i$$

  - How much does excluding i from the regression change its predicted value?
  - Equal to "leverage $\times$ outlier-ness"
- Lots of diagnostics exist, but are mostly heuristic.

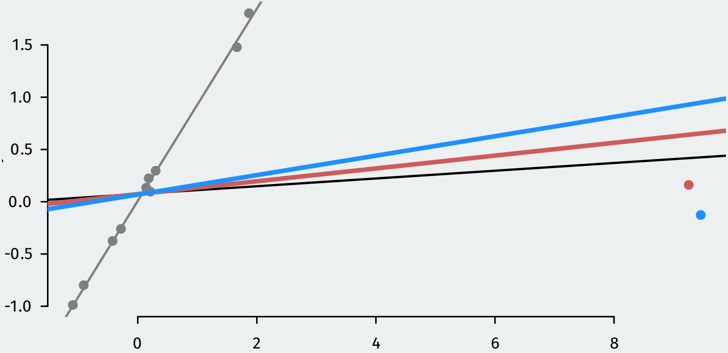
Overall measures of influence

- Influence of *i* can be measured by change in predictions:

$$\widehat{Y}_i - \widetilde{Y}_i = h_{ii}\tilde{e}_i$$

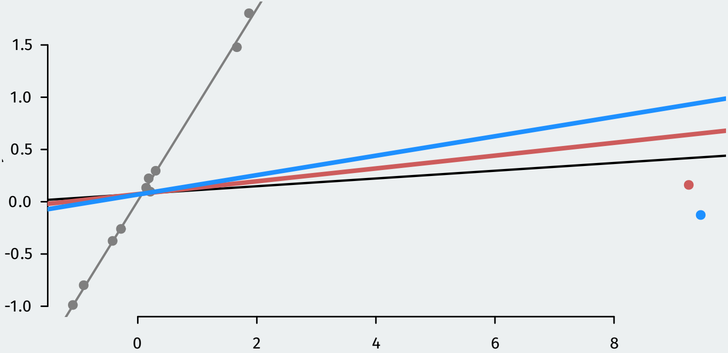
  - How much does excluding *i* from the regression change its predicted value?
  - Equal to "leverage $\times$ outlier-ness"
- Lots of diagnostics exist, but are mostly heuristic.
  - Does removing the point change a coefficient by a lot?

Limitations of the standard tools



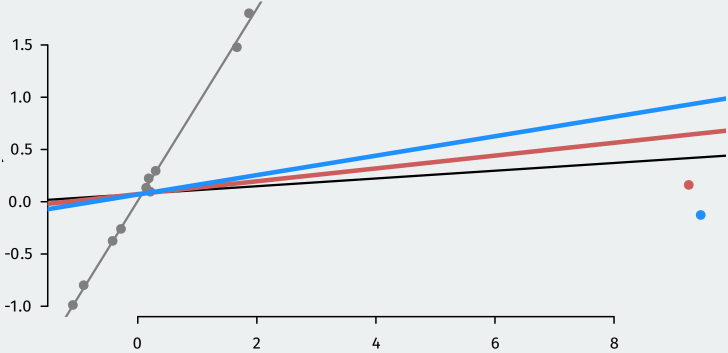
- What happens when there are two influence points?

Limitations of the standard tools



- What happens when there are two influence points?
- Red line drops the red influence point

Limitations of the standard tools



- What happens when there are two influence points?
- Red line drops the red influence point
- Blue line drops the blue influence point

What to do about outliers and influential units?

- Is the data corrupted?

What to do about outliers and influential units?

- Is the data corrupted?
    - Fix the observation (obvious data entry errors)

What to do about outliers and influential units?

- Is the data corrupted?
    - Fix the observation (obvious data entry errors)
    - Remove the observation

What to do about outliers and influential units?

- Is the data corrupted?
    - Fix the observation (obvious data entry errors)
    - Remove the observation
    - Be transparent either way

What to do about outliers and influential units?

- Is the data corrupted?
    - Fix the observation (obvious data entry errors)
    - Remove the observation
    - Be transparent either way
- Is the outlier part of the data generating process?

What to do about outliers and influential units?

- Is the data corrupted?
  - Fix the observation (obvious data entry errors)
  - Remove the observation
  - Be transparent either way
- Is the outlier part of the data generating process?
  - Transform the dependent variable (log(y))

What to do about outliers and influential units?

- Is the data corrupted?
  - Fix the observation (obvious data entry errors)
  - Remove the observation
  - Be transparent either way
- Is the outlier part of the data generating process?
  - Transform the dependent variable ($\log(y)$)
  - Use a method that is robust to outliers (robust regression, least absolute deviations)