Linear Models Lecture 2

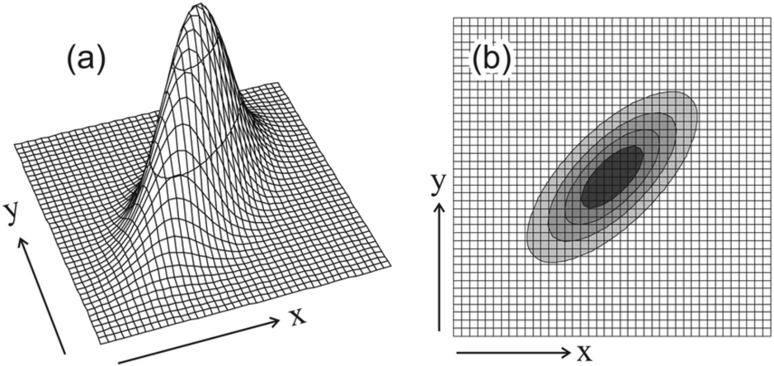
Robert Gulotty

University of Chicago

May 30, 2023

Joint Distributions in Political Science

- We saw the 'roof' distribution, but what joint distributions arise in Political Science?
- If our variables data is discrete and binary, we can have a 2x2 table of probabilities.
- If our data is continuous, we will have a 3d density.


(a)
y
x
(b)
y
x

Definition

$\boldsymbol{X} = (X_1, X_2, \ldots, X_N)$ *is a vector of random variables. If* $\boldsymbol{X}$ *has pdf*

$$f(\boldsymbol{x}) = (2\pi)^{-N/2} det\,(\boldsymbol{\Sigma})^{-1/2} \exp\left(-\frac{1}{2}(\boldsymbol{x} - \boldsymbol{\mu})' \boldsymbol{\Sigma}^{-1} (\boldsymbol{x} - \boldsymbol{\mu})\right)$$

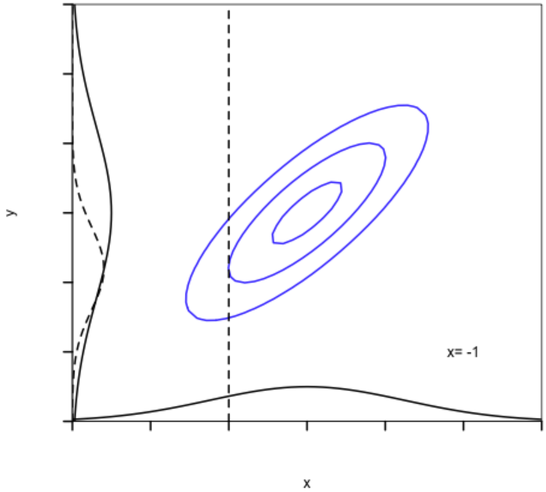
$\boldsymbol{X}$ *is a Multivariate Normal Distribution,*

$$\boldsymbol{X} \sim \textit{Multivariate Normal}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$$

$\boldsymbol{Z} = (X, Y)$ is a vector of jointly distributed random variables.

$$f(\boldsymbol{z}) = (2\pi)^{-1}\det\begin{bmatrix}\sigma_x^2 & \sigma_{xy}\\ \sigma_{yx} & \sigma_y^2\end{bmatrix}^{-1/2}\exp\left(-\frac{1}{2}(\begin{bmatrix}x\\ y\end{bmatrix}-\begin{bmatrix}\mu_x\\ \mu_y\end{bmatrix})'\begin{bmatrix}\sigma_x^2 & \sigma_{xy}\\ \sigma_{yx} & \sigma_y^2\end{bmatrix}^{-1}(\begin{bmatrix}x\\ y\end{bmatrix}-\begin{bmatrix}\mu_x\\ \mu_y\end{bmatrix})\right)$$

$$f(\boldsymbol{z}) = \frac{1}{\sqrt{(2\pi)(2\pi)(\sigma_x^2\sigma_y^2-\sigma_{xy}^2)}}\exp\left(-\frac{1}{2}\begin{bmatrix}x-\mu_x\\ y-\mu_y\end{bmatrix}'\begin{bmatrix}\sigma_x^2 & \sigma_{xy}\\ \sigma_{yx} & \sigma_y^2\end{bmatrix}^{-1}\begin{bmatrix}x-\mu_x\\ y-\mu_y\end{bmatrix}\right)$$


y
x
x= -1

Prediction

- If we want to predict y with a known pdf. The "best" prediction is μ_Y (minimum mean square error (MSE)).
- Suppose we have a bivariate distribution $f(x, y)$. We will be told X. What should we guess for Y?
- We can use any function $h(X)$. What is the best, minimizing $(Y - h(X))^2$?
- We will show that only the Conditional Expectation Function $E(Y|X)$ is the best.

Proof that CEF is the best predictor (Goldberger pg 51)

- Goal: Minimize $(Y - h(X))^2$.
- Define $U \equiv Y - h(X)$, $\epsilon \equiv Y - E(Y|X)$, $W \equiv E(Y|X) - h(X)$.

$$U = Y - h(X)$$
$$U = (\epsilon + E(Y|X)) - (E(Y|X) - W)$$
$$U = \epsilon + W$$

- W only depends on X. So for $X = x$, let's call it $W(X = x) = w$.

$$U^2 = \epsilon^2 + 2w\epsilon + w^2$$
$$\begin{aligned} E(U^2|x) &= E(\epsilon^2|x) + 2E(w\epsilon|x) + E(w^2|x) \\ &= E(\epsilon^2|x) + 2wE(\epsilon|x) + w^2 \\ &= E(\epsilon^2|x) + 2w * 0 + w^2 \end{aligned}$$

$$
\begin{aligned}
E(U^2|x) &= E(\epsilon^2|x) + 2w * 0 + w^2 && \text{from above} \\
E(U^2) &= E_X[E(U^2|X)] && \text{by Law of Iterated Expectations} \\
&= E_X[E(\epsilon^2|x) + 0 + w^2] && \text{Plugging in} \\
&= E[\sigma^2_{Y|x}] + E[W^2]
\end{aligned}
$$

$E[W^2] \geq 0$, so $E(U^2)$ is minimized if $W = 0$, and recall the definition of W.

$$
\begin{aligned}
W &= E(Y|X) - h(X) \\
0 &= E(Y|X) - h(X) \\
h(X) &= E(Y|X)
\end{aligned}
$$

The Conditional Expectation Function $E(Y|X)$ *is* the function that minimizes $E(U^2)$.

Conditional Expectation as a Prediction

Suppose Y and X are random variables.

$$\mu_{Y|X} \equiv E[Y|X]$$

$\mu_{Y|X}$ is our best guess for Y given X
ϵ is the amount we are off.

$$\epsilon \equiv Y - \mu_{Y|X}$$

Properties of Estimation error

ϵ is a random variable where:

$$
\begin{aligned}
E[\epsilon|X] &= E[Y - \mu_{Y|X}|X] \\
&= E[Y|X] - E[\mu_{Y|X}|X] \\
&= E[Y|X] - E[E[Y|X]|X] \\
&= \mu_{Y|X} - \mu_{Y|X} = 0
\end{aligned}
$$

By the law of iterated expectations:

$$E[\epsilon] = E[E[\epsilon|X]] = 0$$

The disturbances center on 0 *by construction*.

Covariance of Estimator and Disturbance

$$E[\epsilon\mu_{Y|X}] = E[E[\epsilon\mu_{Y|X}|X]] = E[\mu_{Y|X}E[\epsilon|X]] = E[\mu_{Y|X} * 0] = 0$$

because $\mu_{Y|X}$ only depends on X.

$$Cov(\epsilon, \mu_{Y|X}) = E[\epsilon\mu_{Y|X}] - E[\epsilon]E[\mu_{Y|X}] = 0 - 0 * \mu_{Y|X} = 0$$

The disturbance is uncorrelated with the conditional expectation. This is what allows us to separate the signal from the noise.

Discussion of CEF

- CEF solves the minimum mean squared error (MSE) prediction problem.
- But it depends on knowing the conditional distribution of $Y|x$, because $E[Y|X] = \int y f_{Y|x}(y|x) dy$.
- This is a general problem: MSE minimizing estimators often require knowledge about the population.
- Instead we will restrict attention to linear predictors. We seek to find the "best" linear predictor (BLP).
- Later we will show that OLS regression estimates are BLUE (best linear unbiased estimators) **if** the CEF is linear and residuals are constant.
- However, what if the CEF is not linear?
- Linear regression is the best (minimum MSE) linear approximation to the CEF.

(Best) Linear Predictors

- Suppose we want a predictor $E(Y|X)$ that is **linear**:

$$f(X) = a + bX$$

  Our standard for prediction is to minimize the mean-square error (best):

- Whereas the CEF might be infinitely complex, the BLP is characterized just by two numbers, a and b. If the CEF is linear, the BLP is the CEF.
  Choose a and b to minimize

$$M = E[(Y - (a + bX))^2]$$

Best Linear Predictor (a)

Take partial derivative with respect to a.

$$\begin{aligned}\frac{\partial E[(Y-(a+bX))^2]}{\partial a} &= E[\frac{\partial (Y-(a+bX))^2}{\partial a}]\\ &= E[-2(Y-(a+bX))]\end{aligned}$$

setting equal to 0:

$$\begin{aligned}0 &= E[-2(Y)-(a+bX)]\\ 0 &= -2E[Y]+2E(a+bX)\\ E(Y) &= a+bE(X)\\ E(Y)-bE(X) &= a\end{aligned}$$

Best Linear Predictor (b)

Take partial derivative with respect to b.

$$\frac{\partial E[(Y-(a+bX))^2]}{\partial b} = E[\partial (Y-(a+bX))^2/\partial b]$$
$$= 2E[(Y-(a+bX))(-X)]$$

setting equal to 0:

$$0 = -2E[(YX)-(a+bX)(X)]$$
$$0 = -2E(YX)+2E[(a+bX)(X)]$$

Best Linear Predictor (b continued)

$$
\begin{aligned}
0 &= -2E[YX] + 2E[(a + bX)(X)] \\
E(YX) &= aE(X) + bE(X^2) \\
E(YX) &= [E(Y) - bE(X)]E(X) + bE(X^2) \\
E(YX) &= E(Y)E(X) - bE(X)E(X) + bE(X^2) \\
E(YX) - E(Y)E(X) &= b[E(X^2) - E(X)^2] \\
b &= \frac{E(YX) - E(Y)E(X)}{[E(X^2) - E(X)^2]} = \frac{Cov(X, Y)}{Var(X)}
\end{aligned}
$$

Best Linear Predictor

$$E[Y|X] = \beta_0 + \beta_1 X$$
$$Y = \beta_0 + \beta_1 X + \epsilon$$

- Parameters (in Greek)
  - $\epsilon = Y - E[Y|X]$
  - The slope $\beta = \frac{Cov(X,Y)}{Var(X)}$
  - The intercept $\alpha = E(Y) - \beta E(X)$
- Y is called the dependent variable, X is called the independent variable.

What is "linear in parameters"?

- The Best Linear Predictor is linear in *parameters* $\theta \in \{\beta_0,\ \beta_1\ \ldots\}$.
- Examples of linear in parameters:

$$Y = \beta_0 + \beta_1 X^4 + \epsilon$$
$$Y = \beta_0 + \beta_1 e^X + \epsilon$$

- Examples of nonlinear in parameters:

$$Y = \beta_0 + \frac{1}{\beta_1} X + \epsilon$$
$$Y = \beta_0 + \beta_1^2 X + \epsilon$$
$$Y = \beta_0 + e^{\beta_1 X} + \epsilon$$

Deriving variance of BLP at minimized values

$$\beta = \frac{Cov(X,Y)}{Var(X)} \qquad \beta Var(X) = Cov(X,Y)$$

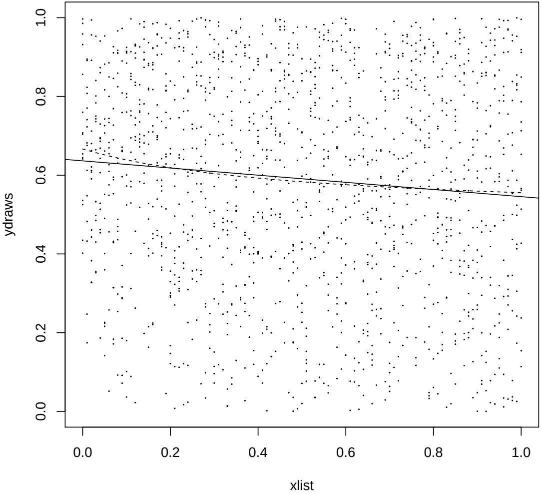
$$\begin{aligned}
E[(Y-(\alpha+\beta X))^2] - E[Y-(\alpha+\beta X)]^2 &= Var(Y-(\alpha+\beta X)) \\
&= Var(Y-\beta X) \\
&= Var(Y) + Var(\beta X) - 2Cov(\beta X, Y) \\
&= Var(Y) + \beta^2 Var(X) - 2\beta Cov(X,Y) \\
&= Var(Y) + \beta^2 Var(X) - 2\beta^2 Var(X) \\
&= Var(Y) - \beta^2 Var(X) \\
\sigma_\epsilon^2 &= \sigma_Y^2 - \beta^2 \sigma_X^2
\end{aligned}$$

Example Best Linear Predictor, roof

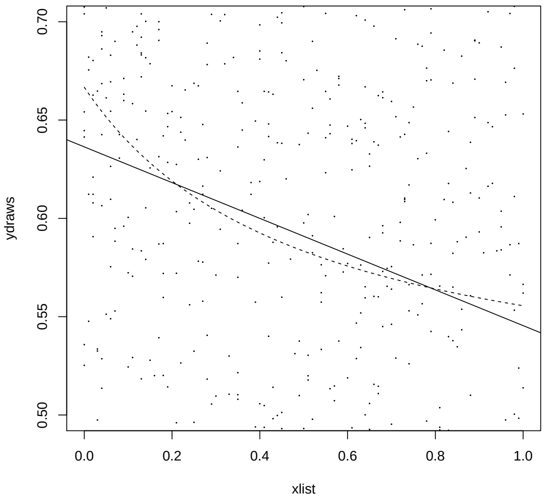
Example: $X + Y$
Suppose X and Y have pdf $x + y$ for $x, y \in [0, 1]$.

$$b = \frac{Cov(X, Y)}{Var(X)} = -1/11$$

$$a = E(Y) - bE(X) = 84/132$$



ydraws
xlist
0.0
0.2
0.4
0.6
0.8
1.0


ydraws
0.50
0.55
0.60
0.65
0.70
0.0
0.2
0.4
0.6
0.8
1.0
xlist

Sample vs Population

- Theoretical "Population" Objects (not observed): Greek Letters: α, β, μ, σ, ϵ, E(), Var(), Cov().
- Empirical Objects (observed): latin letters, a, b, $\bar{x}$, s^2, e, $\widehat{E}()$, $\widehat{Cov}()$.
- Objects of the first group are not exactly equal to their empirical analogues from the second group.
  - $\alpha = \mu_Y - \beta\mu_X$ is correct.
  - $a = \bar{y} - b\bar{x}$ is correct.
  - $E(\bar{x}) = \mu$ is correct.
  - $\bar{x} = \mu$ is not correct.
  - $Cov(x) = s_x^2$ is not correct.

Sample Linear projection

Population linear projection:

$$E(Y|X) = \alpha + \beta X$$

where

$$\beta = \frac{\sigma_{XY}}{\sigma_X^2}, \quad \alpha = \mu_Y - \beta\mu_X$$

Sample linear projection:

$$\hat{Y} = b_0 + b_1 X$$

where

$$b_1 = \frac{S_{XY}}{S_X^2}, \; b_0 = \bar{Y} - b_1\bar{X}$$

Asymptotic Distribution of Sample Slope

Call $X^* = X - \bar{X}$, $\epsilon = Y - (\alpha + \beta X)$, then the Bivariate Delta Method (Goldberg 10.5) tells us

$$b_1 \overset{A}{\sim} N(\beta, \frac{E(X^{*2}\epsilon^2)}{(\sigma_X^2)^2})$$

If $E(\epsilon^2|X)$ is a constant σ^2, then

$$\begin{aligned} E(X^{*2}\epsilon^2) &= E_X[E(X^{*2}\epsilon^2|X)] = E_X[X^{*2}E(\epsilon^2|X)] \\ &= E_X[X^{*2}\sigma^2] \\ &= \sigma^2 E_X[X^{*2}] = \sigma^2 V(X) = \sigma^2\sigma_X^2 \end{aligned}$$

$$b_1 \overset{A}{\sim} N(\beta, \frac{\sigma^2}{\sigma_X^2})$$

Ordinary Least Squares Regression Estimator

- In our effort to approximate the CEF, we simply replaced the expectations, covariances etc. in the BLP with the sample means, covariances etc.
- It turns out that we can algebraically process data with the least squares procedure (OLS) to estimate the BLP parameters.
- That is, we will solve $\min_{b_0,\ b_1} \sum_i e^2$
- In addition, **if** the CEF function is linear and $E(\epsilon^2|X)$ is a constant σ^2, then the Gauss Markov Theorem tells us that OLS is a Best Linear Unbiased Estimator (BLUE).

Brief Linear Algebra Interlude

- $\boldsymbol{x} = \begin{bmatrix} x_1 \\ x_2 \\ \dots \\ x_n \end{bmatrix}$ is $n \times 1$ column vector. $\boldsymbol{i} = \begin{bmatrix} 1 \\ 1 \\ \dots \\ 1 \end{bmatrix}$.
- $\boldsymbol{x'} = \boldsymbol{x^T} = \begin{bmatrix} x_1 & x_2 & \dots & x_n \end{bmatrix}$ is $1 \times n$ vector.
- If c is a scalar, $c\boldsymbol{x'} = \begin{bmatrix} cx_1 & cx_2 & \dots & cx_n \end{bmatrix}$.
- $\boldsymbol{x'i} = \boldsymbol{x} \cdot \boldsymbol{i} = x_1 * 1 + x_2 * 1 + x_3 * 1 + \dots + x_n * 1 = \sum x_i = n\bar{x}$
- $\boldsymbol{x'x} = x_1^2 + x_2^2 + x_3^2 + \dots + x_n^2 = \sum x_i^2 = n * \widehat{Var}(x) + n\bar{x}^2$
- $\boldsymbol{xx'} = \begin{bmatrix} x_1x_1 & x_1x_2 & \dots \\ x_2x_1 & \dots & \\ \dots & & \end{bmatrix}, \quad \boldsymbol{I} = \begin{bmatrix} 1 & 0 & \dots \\ 0 & 1 & \dots \\ \dots & & \end{bmatrix}$

Linear Algebra Rules

If $\boldsymbol{x}$, $\boldsymbol{y}$, and $\boldsymbol{z}$ are vectors of equal length:

- You can add, subtract and dot product them. No division of vectors.
- Commutative property: If $\boldsymbol{x}$ and $\boldsymbol{y}$ are vectors of equal length, $\boldsymbol{x}'\boldsymbol{y} = \boldsymbol{y}'\boldsymbol{x}$
- Distributive property: $\boldsymbol{x}'(\boldsymbol{y} + \boldsymbol{z}) = \boldsymbol{x}'\boldsymbol{y} + \boldsymbol{x}'\boldsymbol{z}$

Linear Algebra Geometry (Advanced Topic)

- $\sqrt{\boldsymbol{x}'\boldsymbol{x}} = ||\boldsymbol{x}||$ is called the Euclidean Norm, from the classic Euclidean distance $\sqrt{a^2+b^2} = c$ in Descartes' theory of coordinates.
- $\boldsymbol{u} = \frac{\boldsymbol{x}}{\sqrt{\boldsymbol{x}'\boldsymbol{x}}}$ is called the **unit vector** in the direction of $\boldsymbol{x}$.
- If $\boldsymbol{u}$ is a unit vector in the direction of $\boldsymbol{x}$, $\boldsymbol{y}'\boldsymbol{u} = \boldsymbol{u}'\boldsymbol{y}$ is the *scalar projection* of $\boldsymbol{y}$ onto $\boldsymbol{x}$.
- $(\boldsymbol{u}'\boldsymbol{y})\boldsymbol{u}$ is the *vector projection* of $\boldsymbol{y}$ onto $\boldsymbol{x}$.

$$(\boldsymbol{u}'\boldsymbol{y})\boldsymbol{u} = \left(\frac{\boldsymbol{x}'\boldsymbol{y}}{\sqrt{\boldsymbol{x}'\boldsymbol{x}}}\right)\frac{\boldsymbol{x}}{\sqrt{\boldsymbol{x}'\boldsymbol{x}}} = \left(\frac{\boldsymbol{x}'\boldsymbol{y}}{\boldsymbol{x}'\boldsymbol{x}}\right)\boldsymbol{x}$$

- If $\boldsymbol{x}'\boldsymbol{y} = 0$, $\boldsymbol{x}$ is "orthogonal" to $\boldsymbol{y}$.

Gauss-Markov Assumptions (classical fixed $\boldsymbol{x}$)

1. CEF is linear
2. $\boldsymbol{x} = \begin{bmatrix} x_1 \\ x_2 \\ \dots \\ x_n \end{bmatrix}$ is fixed.
3. $\sum_{i=1}^{n}(x_i - \bar{x})^2 > 0$ [x takes on at least two values].
4. $E(\epsilon_i) = 0 \quad \forall i$.
5. $E(\boldsymbol{\epsilon}\boldsymbol{\epsilon}^{'}) = \sigma_{\epsilon}^2 \boldsymbol{I}$.
   - $Var(\epsilon_i) = \sigma_{\epsilon}^2 \quad \forall i$ [homoskedasticity, i.i.d.]
   - $Cov(\epsilon_i, \epsilon_j) = 0$

Deriving $b_0, \ b_1$ estimator using OLS

$$\min_{b_0 \ b_1} \sum_i e^2 = \min_{b_0 \ b_1} \boldsymbol{e'e}$$

$$\begin{aligned}
\frac{\partial}{\partial b_1}\boldsymbol{e'e} &= \frac{\partial}{\partial b_1}(\boldsymbol{y} - \hat{\boldsymbol{y}})'(\boldsymbol{y} - \hat{\boldsymbol{y}}) \\
&= \frac{\partial}{\partial b_1}(\boldsymbol{y}'\boldsymbol{y} - \boldsymbol{y}'\hat{\boldsymbol{y}} - \hat{\boldsymbol{y}}'\boldsymbol{y} + \hat{\boldsymbol{y}}'\hat{\boldsymbol{y}}) \\
&= \frac{\partial}{\partial b_1}(\boldsymbol{y}'\boldsymbol{y} - 2\hat{\boldsymbol{y}}'\boldsymbol{y} + \hat{\boldsymbol{y}}'\hat{\boldsymbol{y}}) \\
&= 0 - 2\frac{\partial}{\partial b_1}\hat{\boldsymbol{y}}'\boldsymbol{y} + \frac{\partial}{\partial b_1}\hat{\boldsymbol{y}}'\hat{\boldsymbol{y}}
\end{aligned}$$

$$\begin{aligned}
\frac{\partial}{\partial b_0}\boldsymbol{e'e} &= \frac{\partial}{\partial b_0}(\boldsymbol{y} - \hat{\boldsymbol{y}})'(\boldsymbol{y} - \hat{\boldsymbol{y}}) \\
&= 0 - 2\frac{\partial}{\partial b_0}\hat{\boldsymbol{y}}'\boldsymbol{y} + \frac{\partial}{\partial b_0}\hat{\boldsymbol{y}}'\hat{\boldsymbol{y}}
\end{aligned}$$

Note on calculus with vectors (I)

$$\hat{\boldsymbol{y}} = \begin{bmatrix} b_0 + b_1 x_1 \\ b_0 + b_1 x_2 \\ b_0 + b_1 x_3 \\ \dots \end{bmatrix} \qquad \frac{\partial \hat{\boldsymbol{y}}}{\partial b_1} = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ \dots \end{bmatrix} = \boldsymbol{x} \qquad \frac{\partial \hat{\boldsymbol{y}}}{\partial b_0} = \begin{bmatrix} 1 \\ 1 \\ 1 \\ \dots \end{bmatrix} = \boldsymbol{i}$$

$$\hat{\boldsymbol{y}}' \hat{\boldsymbol{y}} = \begin{bmatrix} b_0 + b_1 x_1 & b_0 + b_1 x_2 & b_0 + b_1 x_3 & \dots \end{bmatrix} \begin{bmatrix} b_0 + b_1 x_1 \\ b_0 + b_1 x_2 \\ b_0 + b_1 x_3 \\ \dots \end{bmatrix} = \sum_i (b_0 + b_1 x_i)^2$$

Note on calculus with vectors (II)

$$
\begin{aligned}
\frac{\partial}{\partial b_1}\hat{\boldsymbol{y}}'\hat{\boldsymbol{y}} &= \frac{\partial}{\partial b_1}\sum_i (b_0 + b_1 x_i)^2 \\
&= \sum_i 2x_i(b_0 + b_1 x_i) \\
&= 2\boldsymbol{x}'[b_0 \boldsymbol{i} + b_1 \boldsymbol{x}] \\
\frac{\partial}{\partial b_0}\hat{\boldsymbol{y}}'\hat{\boldsymbol{y}} &= \sum_i 2(b_0 + b_1 x_i) \\
&= 2\boldsymbol{i}'[b_0 + b_1 \boldsymbol{x}]
\end{aligned}
$$

Deriving b_0, b_1 estimator using OLS

$$
\begin{aligned}
\frac{\partial}{\partial b_0}\boldsymbol{e}'\boldsymbol{e} &= 0 - 2\frac{\partial}{\partial b_0}\hat{\boldsymbol{y}}'\boldsymbol{y} + \frac{\partial}{\partial b_0}\hat{\boldsymbol{y}}'\hat{\boldsymbol{y}} \\
&= 0 - 2\boldsymbol{i}'\boldsymbol{y} + 2\boldsymbol{i}'[(b_0 + b_1\boldsymbol{x})] \\
\boldsymbol{i}'b_0 &= \boldsymbol{i}'\boldsymbol{y} - b_1\boldsymbol{i}'\boldsymbol{x} \\
nb_0 &= \boldsymbol{i}'\boldsymbol{y} - b_1\boldsymbol{i}'\boldsymbol{x} \\
nb_0 &= n\bar{y} - b_1 n\bar{x} \\
b_0 &= \bar{y} - b_1\bar{x}
\end{aligned}
$$

$$
\begin{aligned}
\frac{\partial}{\partial b_1}\boldsymbol{e}'\boldsymbol{e} &= 0 - 2\frac{\partial}{\partial b_1}\hat{\boldsymbol{y}}'\boldsymbol{y} + \frac{\partial}{\partial b_1}\hat{\boldsymbol{y}}'\hat{\boldsymbol{y}} \\
&= -2\boldsymbol{x}'\boldsymbol{y} + 2\boldsymbol{x}'[b_0\boldsymbol{i} + b_1\boldsymbol{x}] \\
&= -2\boldsymbol{x}'\boldsymbol{y} + 2\boldsymbol{x}'[[\bar{y} - b_1\bar{x}]\boldsymbol{i} + b_1\boldsymbol{x}] \\
&= -2\boldsymbol{x}'\boldsymbol{y} + 2\boldsymbol{x}'\bar{y}\boldsymbol{i} + 2\boldsymbol{x}'b_1(\boldsymbol{x} - \bar{x}\boldsymbol{i}) \\
&= -2\boldsymbol{x}'\boldsymbol{y} + 2n\bar{x}\bar{y} + 2b_1[\boldsymbol{x}'\boldsymbol{x} - n\bar{x}\bar{x}] \\
\boldsymbol{x}'\boldsymbol{y} - n\bar{x}\bar{y} &= b_1[\boldsymbol{x}'\boldsymbol{x} - n\bar{x}\bar{x}] \\
\widehat{Cov}(\boldsymbol{x},\ \boldsymbol{y}) &= b_1\widehat{Var}(\boldsymbol{x}) \\
\frac{\widehat{Cov}(\boldsymbol{x},\boldsymbol{y})}{\widehat{Var}(\boldsymbol{x})} &= b_1 \quad \text{By assumption 3}
\end{aligned}
$$

```
> x <- 1:50
> y <- 8 + .5 * x + rnorm(50)

> coef(lm(y~x))
(Intercept)          x
8.1679311    0.4982869

> mean(y) - cov(x,y) / var(x) * mean(x)
[1] 8.167931
> cov(x,y)/var(x)
[1] 0.4982869
```

Interpretation

- b_1 is the slope: one unit change in x is associated with a b_1 unit change in Y.

$$\hat{\boldsymbol{y}} = b_0 + \frac{\widehat{Cov}(\boldsymbol{x}, \boldsymbol{y})}{\widehat{Var}(\boldsymbol{x})}\boldsymbol{x}$$
$$\boldsymbol{y} = \hat{\boldsymbol{y}} + \boldsymbol{e}$$

- b_0 is the intercept: What is the predicted value of y when $x = 0$.
- e_i is the residual: observed y minus predicted y.

Decomposition of Variance

$$\sigma_\epsilon^2 = \sigma_Y^2 - \beta^2\sigma_X^2$$

$$\widehat{Var}(\boldsymbol{e}) = \widehat{Var}(\boldsymbol{y}) - b_1^2\widehat{Var}(\boldsymbol{x})$$

$$b_1^2\widehat{Var}(\boldsymbol{x}) + \widehat{Var}(\boldsymbol{e}) = \widehat{Var}(\boldsymbol{y})$$

$$\underbrace{b_1^2\sum(x_i-\bar{x})^2}_{\text{Explained Sum of Squares}} + \underbrace{\sum e_i^2}_{\text{Residual Sum of Squares}} = \underbrace{\sum(y_i-\bar{y})^2}_{\text{Total Sum of Squares}}$$

$$\text{ESS} + \text{RSS} = \text{TSS}$$

$$\frac{\text{ESS}}{\text{TSS}} + \frac{\text{RSS}}{\text{TSS}} = 1$$

$$R^2 + \frac{\text{RSS}}{\text{TSS}} = 1$$

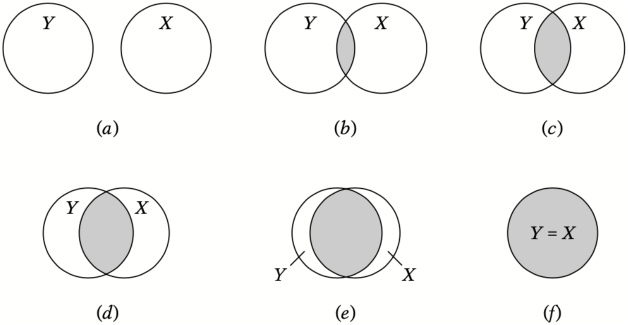


FIGURE 3.9 The Ballentine view of r^2: (*a*) $r^2 = 0$; (*f*) $r^2 = 1$.

Sample Correlation Coefficient

$$r_{xy} \equiv \frac{\widehat{Cov}(\boldsymbol{x},\ \boldsymbol{y})}{\sqrt{\widehat{Var}(\boldsymbol{x})}\sqrt{\widehat{Var}(\boldsymbol{y})}} = \frac{S_{XY}}{S_X S_Y} = \frac{\frac{1}{n-1}\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\frac{1}{n-1}\sum(x - \bar{x})^2 \frac{1}{n-1}\sum(y_i - \bar{y})^2}}$$

- The correlation coefficient (r_{xy})is the normalized linear relationship.
- The OLS slope coefficient (b_1) is the unnormalized linear relationship.

OLS coefficient vs Correlation

$$b_1 = \frac{\widehat{Cov}(\boldsymbol{x},\ \boldsymbol{y})}{\sqrt{\widehat{Var}(\boldsymbol{x})}} \frac{1}{\sqrt{\widehat{Var}(\boldsymbol{x})}} \qquad \text{(Definition of } b_1\text{)}$$

$$\frac{b_1}{\sqrt{\widehat{Var}(\boldsymbol{y})}} = \frac{\widehat{Cov}(\boldsymbol{x},\ \boldsymbol{y})}{\sqrt{\widehat{Var}(\boldsymbol{x})}\sqrt{\widehat{Var}(\boldsymbol{y})}} \frac{1}{\sqrt{\widehat{Var}(\boldsymbol{x})}}$$

$$= r_{xy} \frac{1}{\sqrt{\widehat{Var}(\boldsymbol{x})}}$$

$$b_1 = r_{xy} \frac{\sqrt{\widehat{Var}(\boldsymbol{y})}}{\sqrt{\widehat{Var}(\boldsymbol{x})}} = r_{xy} \frac{S_Y}{S_X}$$

```
>  x <- 1:50
>  y <- 8 + .5 * x + rnorm(50)

>  coef(lm(y~x))
(Intercept)          x
8.1679311    0.4982869

> cor(x,y)
[1] 0.9897854

> cor(x,y)*sd(y)/sd(x)
[1] 0.4982869
```

Obtaining the Sample Correlation Coefficient from a Regression

$$
\begin{aligned}
\hat{\boldsymbol{y}} &= b_0 + b_1 \boldsymbol{x} \\
\hat{\boldsymbol{y}} &= \bar{y} - b_1 \bar{x} + b_1 \boldsymbol{x} \\
\hat{\boldsymbol{y}} - \bar{y} &= b_1 (\boldsymbol{x} - \bar{x}) \\
\hat{\boldsymbol{y}} - \bar{y} &= r_{xy} \frac{S_Y}{S_X} (\boldsymbol{x} - \bar{x}) \\
\frac{\hat{\boldsymbol{y}} - \bar{y}}{S_Y} &= r_{xy} \frac{(\boldsymbol{x} - \bar{x})}{S_X} \\
y^* &= \rho_0 + \rho_1 x^*
\end{aligned}
$$

```
> x <- 1:50
> y <- 8+ .5 * x + rnorm(50,sd=3)
>
> cor(x,y)
[1] 0.9139259
>
> yhatnorm <- (predict(lm(y~x))-mean(y))/sd(y)
> xhatnorm <- (x-mean(x))/sd(x)
>
> coef(lm(yhatnorm ~ xhatnorm -1 ))
 xhatnorm
0.9139259
```

Proof that OLS unbiased Bias$(\hat{\theta}) \equiv E(\hat{\theta}) - \theta$

- Assume specification : $\boldsymbol{y} = [\beta_0 \boldsymbol{i} + \beta_1 \boldsymbol{x} + \boldsymbol{\epsilon}]$, $\boldsymbol{i'y} = [\boldsymbol{i'}\beta_0 \boldsymbol{i} + \boldsymbol{i'}\beta_1 \boldsymbol{x} + \boldsymbol{i'\epsilon}]$, $\bar{y} = \beta_0 + \beta_1 \bar{x} + \bar{\epsilon}$

$$
\begin{aligned}
b_1 &= \frac{\boldsymbol{x'y} - n\bar{x}\bar{y}}{\widehat{Var}(x)} \\
b_1 &= \frac{\boldsymbol{x'}[\beta_0 \boldsymbol{i} + \beta_1 \boldsymbol{x} + \boldsymbol{\epsilon}] - n\bar{x}[\beta_0 + \beta_1 \bar{x} + \bar{\epsilon}]}{\widehat{Var}(x)} \\
&= \frac{n\bar{x}\beta_0 - n\bar{x}\beta_0 + \boldsymbol{x'}\beta_1 \boldsymbol{x} - n\bar{x}\beta_1 + \boldsymbol{x'\epsilon} - n\bar{x}\bar{\epsilon}}{\widehat{Var}(x)} \\
&= \frac{\beta_1[\boldsymbol{x'x} - n\bar{x}] + \boldsymbol{x'\epsilon} - n\bar{x}\bar{\epsilon}}{\widehat{Var}(x)} \\
&= \beta_1 + \frac{\boldsymbol{x'\epsilon} - n\bar{x}\bar{\epsilon}}{\widehat{Var}(x)}
\end{aligned}
$$

Proof that OLS unbiased

- $\text{Bias}(\hat{\theta}) \equiv E(\hat{\theta}) - \theta$

$$b_1 = \beta_1 + \frac{\mathbf{x}'\epsilon - n\bar{x}\bar{\epsilon}}{\widehat{Var}(x)}$$

$$E[b_1] = E[\beta_1] + E[\frac{\mathbf{x}'\epsilon - n\bar{x}\bar{\epsilon}}{\widehat{Var}(x)}]$$

$$E[b_1] = E[\beta_1] + \frac{\mathbf{x}'E[\epsilon] - n\bar{x}E[\bar{\epsilon}]}{\widehat{Var}(x)} \quad \text{by assumption (2)}$$

$$E[b_1] = E[\beta_1] + \frac{\mathbf{x}'0 - n\bar{x}0}{\widehat{Var}(x)}$$

Change of Notation

- The notation simplifies if we center our variables $\tilde{\boldsymbol{x}} = \boldsymbol{x} - \bar{x}$, $\tilde{\boldsymbol{y}} = \boldsymbol{y} - \bar{y}$,
- $b_1 = \frac{\tilde{\boldsymbol{x}}'\tilde{\boldsymbol{y}}}{\tilde{\boldsymbol{x}}'\tilde{\boldsymbol{x}}}$
- Note: $b_1\tilde{\boldsymbol{x}} = \frac{\tilde{\boldsymbol{x}}'\tilde{\boldsymbol{y}}}{\tilde{\boldsymbol{x}}'\tilde{\boldsymbol{x}}}\tilde{\boldsymbol{x}}$ is the vector projection of $\boldsymbol{y}$ in $\tilde{\boldsymbol{x}}$.
- Define weights $\boldsymbol{w} \equiv \frac{\tilde{\boldsymbol{x}}}{\tilde{\boldsymbol{x}}'\tilde{\boldsymbol{x}}}$
- Note $\boldsymbol{w}'\boldsymbol{w} = \frac{\tilde{\boldsymbol{x}}'\tilde{\boldsymbol{x}}}{(\tilde{\boldsymbol{x}}'\tilde{\boldsymbol{x}})'\tilde{\boldsymbol{x}}'\tilde{\boldsymbol{x}}} = \frac{1}{(\tilde{\boldsymbol{x}}'\tilde{\boldsymbol{x}})'}$
- $b_1 = \boldsymbol{w}'\tilde{\boldsymbol{y}}$
- Recall under the assumptions above, $b_1 = \beta_1 + \boldsymbol{w}'\boldsymbol{\epsilon}$. $b_1 - \beta_1 = \boldsymbol{w}'\boldsymbol{\epsilon}$

Variance of the slope estimator

$$
\begin{aligned}
Var(b_1) &= E[(b_1 - E(b_1))^2] \\
&= E[(b_1 - \beta_1)^2] \\
&= E[(\boldsymbol{w}'\epsilon)^2] \\
&= E[(\boldsymbol{w}'\epsilon)'(\boldsymbol{w}'\epsilon)] \\
&= E[(\boldsymbol{w}'\epsilon\epsilon'\boldsymbol{w})] \\
&= (\boldsymbol{w}'\boldsymbol{E}[\epsilon\epsilon']\boldsymbol{w}) \rightarrow \frac{\sigma^2\boldsymbol{I}}{\tilde{\boldsymbol{x}}'\tilde{\boldsymbol{x}}} \quad \text{by assumption (4).}
\end{aligned}
$$

Gauss Markov BLUE Step 1

- The OLS weights $\boldsymbol{w}$ give us: $b_1 = \boldsymbol{w}'\tilde{\boldsymbol{y}} = \boldsymbol{w}'\boldsymbol{y}$
- Suppose we had some other unbiased linear estimator with weights $\boldsymbol{c} = \boldsymbol{w} + \boldsymbol{d}$

$$
\begin{aligned}
b_* &= \boldsymbol{c}'\tilde{\boldsymbol{y}} = \boldsymbol{c}'\boldsymbol{y} \\
b_* &= \boldsymbol{c}'[\beta_0 \boldsymbol{i} + \beta_1 \tilde{\boldsymbol{x}} + \epsilon] \\
b_* &= \beta_0 \boldsymbol{c}'\boldsymbol{i} + \beta_1 \boldsymbol{c}'\tilde{\boldsymbol{x}} + \boldsymbol{c}'\epsilon \\
E[b_*] &= \beta_0 (\boldsymbol{w} + \boldsymbol{d})'\boldsymbol{i} + \beta_1 (\boldsymbol{w} + \boldsymbol{d})'\tilde{\boldsymbol{x}} + (\boldsymbol{w} + \boldsymbol{d})'E[\epsilon] \\
&= \beta_0 (\boldsymbol{w} + \boldsymbol{d})'\boldsymbol{i} + \beta_1 \frac{\tilde{\boldsymbol{x}}'}{\tilde{\boldsymbol{x}}'\tilde{\boldsymbol{x}}}\tilde{\boldsymbol{x}} + \beta_1 \boldsymbol{d}'\tilde{\boldsymbol{x}} + (\boldsymbol{w} + \boldsymbol{d})'E[\epsilon] = \beta_1
\end{aligned}
$$

only if

$$
(\boldsymbol{w} + \boldsymbol{d})'\boldsymbol{i} = 0 \text{ and } \boldsymbol{d}'\tilde{\boldsymbol{x}} = 0
$$

Gauss Markov BLUE Step 2

- This gives us that our alternative estimator is:

$$b_* = \beta_0 \boldsymbol{c}'\boldsymbol{i} + \beta_1 \boldsymbol{c}'\tilde{\boldsymbol{x}} + \boldsymbol{c}'\boldsymbol{\epsilon} = \beta_1 + \boldsymbol{c}'\boldsymbol{\epsilon}$$

$$b_* - \beta_1 = \boldsymbol{c}'\boldsymbol{\epsilon}$$

$$\begin{aligned}
Var(b_*) &= E((b_* - \beta_1)^2) \\
&= E(\boldsymbol{c}'\boldsymbol{\epsilon}\boldsymbol{\epsilon}'\boldsymbol{c}) \\
&= \sigma_\epsilon^2(\boldsymbol{c}'\boldsymbol{c}) \\
&= \sigma_\epsilon^2(\boldsymbol{w}+\boldsymbol{d})'(\boldsymbol{w}+\boldsymbol{d})) \\
&= \sigma_\epsilon^2(\boldsymbol{w}'\boldsymbol{w} + \boldsymbol{d}'\boldsymbol{w} + \boldsymbol{w}'\boldsymbol{d} + \boldsymbol{d}'\boldsymbol{d}) \\
&= \sigma_\epsilon^2\boldsymbol{w}'\boldsymbol{w} + \sigma_\epsilon^2\boldsymbol{d}'\boldsymbol{w} + \sigma_\epsilon^2\boldsymbol{w}'\boldsymbol{d} + \sigma_\epsilon^2\boldsymbol{d}'\boldsymbol{d} \\
&= \sigma_\epsilon^2\boldsymbol{w}'\boldsymbol{w} + \sigma_\epsilon^2\boldsymbol{d}'\boldsymbol{d} > Var(b_1)
\end{aligned}$$

Discussion of Gauss Markov

- Gauss Markov shows that b_1 is the best (lowest variance) among linear unbiased estimators of β_1.
- We used all of the assumptions to get this result.
  - We can easily dispense with the assumption that $\boldsymbol{x}$ is fixed.
  - The assumption that $\widehat{Var}(x_i)$ is not zero is 1) not problematic and 2) testable.
  - If $E(\epsilon_i) = 0$ doesn't hold, we have bias,
  - If $Var(\epsilon_i)$ is not a constant, we have heteroskedasticity,
  - If $Cov(\epsilon_i, \epsilon_j)$ is not zero, we have serial dependence.

Expectation of Sum of Squared Residuals

$$
\begin{aligned}
E[\sum e_i^2] &= E[\boldsymbol{e}'\boldsymbol{e}] \\
&= E[TSS - ESS] \\
&= E[\tilde{\boldsymbol{y}}'\tilde{\boldsymbol{y}}] - E[b_1^2\tilde{\boldsymbol{x}}'\tilde{\boldsymbol{x}}] \\
&= E[(\beta\tilde{\boldsymbol{x}} + \tilde{\boldsymbol{\epsilon}})'(\beta\tilde{\boldsymbol{x}} + \tilde{\boldsymbol{\epsilon}})] - E[b_1^2\tilde{\boldsymbol{x}}'\tilde{\boldsymbol{x}}] \\
&= [\beta^2\tilde{\boldsymbol{x}}'\tilde{\boldsymbol{x}} + 2\beta E[\tilde{\boldsymbol{x}}'\tilde{\boldsymbol{\epsilon}}] + E[\tilde{\boldsymbol{\epsilon}}'\tilde{\boldsymbol{\epsilon}}]] - [Var(b_1\tilde{\boldsymbol{x}}) + E(b_1\tilde{\boldsymbol{x}})^2] \\
&= [\beta^2\tilde{\boldsymbol{x}}'\tilde{\boldsymbol{x}} + (n-1)\sigma^2] - [Var(b_1) + E(b_1)^2]\tilde{\boldsymbol{x}}'\tilde{\boldsymbol{x}} \\
&= [\beta^2\tilde{\boldsymbol{x}}'\tilde{\boldsymbol{x}} + (n-1)\sigma^2] - [\frac{\sigma^2}{\tilde{\boldsymbol{x}}'\tilde{\boldsymbol{x}}} + \beta_1^2]\tilde{\boldsymbol{x}}'\tilde{\boldsymbol{x}} \\
&= (n-2)\sigma^2
\end{aligned}
$$

Matrix Formulation

- We can condense notation even further $\boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix}$, $\boldsymbol{X} = \begin{bmatrix} \boldsymbol{i} & \boldsymbol{x} \end{bmatrix} = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \ldots & \ldots \end{bmatrix}$,

$$\begin{aligned} \boldsymbol{y} &= \beta_0 \boldsymbol{i} + \beta_1 \boldsymbol{x} + \boldsymbol{\epsilon} \\ \boldsymbol{y} &= \boldsymbol{X} \boldsymbol{\beta} + \boldsymbol{\epsilon} \end{aligned}$$

- $\boldsymbol{X}'\boldsymbol{X} = \begin{bmatrix} \boldsymbol{i}' \\ \boldsymbol{x}' \end{bmatrix} \begin{bmatrix} \boldsymbol{i} & \boldsymbol{x} \end{bmatrix} = \begin{bmatrix} \boldsymbol{i}'\boldsymbol{i} & \boldsymbol{i}'\boldsymbol{x} \\ \boldsymbol{x}'\boldsymbol{i} & \boldsymbol{x}'\boldsymbol{x} \end{bmatrix}$.
- $\boldsymbol{X}'\boldsymbol{y} = \begin{bmatrix} \boldsymbol{i}' \\ \boldsymbol{x}' \end{bmatrix} [\boldsymbol{y}] = \begin{bmatrix} \boldsymbol{i}'\boldsymbol{y} \\ \boldsymbol{x}'\boldsymbol{y} \end{bmatrix}$.

Matrix Terms

Suppose $\boldsymbol{x} = (x_1, x_2, \ldots, x_N)$

$$E[\boldsymbol{x}] = \boldsymbol{\mu}$$
$$\text{Cov}(\boldsymbol{x}) = \boldsymbol{\Sigma}$$

So that,

$$\boldsymbol{\Sigma} = \begin{pmatrix} \text{Var}(x_1) & \text{Cov}(x_1, x_2) & \ldots & \text{Cov}(x_1, x_N) \\ \text{Cov}(x_2, x_1) & \text{Var}(x_2) & \ldots & \text{Cov}(x_2, x_N) \\ \vdots & \vdots & \ddots & \vdots \\ \text{Cov}(x_N, x_1) & \text{Cov}(x_N, x_2) & \ldots & \text{Var}(x_N) \end{pmatrix}$$

Formula for Matrix Inverse

$$
\begin{aligned}
(\boldsymbol{X}'\boldsymbol{X})^{-1} &= \begin{bmatrix} \boldsymbol{i}'\boldsymbol{i} & \boldsymbol{i}'\boldsymbol{x} \\ \boldsymbol{x}'\boldsymbol{i} & \boldsymbol{x}'\boldsymbol{x} \end{bmatrix}^{-1} \\
&= \frac{1}{\boldsymbol{x}'\boldsymbol{x}\boldsymbol{i}'\boldsymbol{i} - \boldsymbol{i}'\boldsymbol{x}\boldsymbol{x}'\boldsymbol{i}} \begin{bmatrix} \boldsymbol{x}'\boldsymbol{x} & -\boldsymbol{i}'\boldsymbol{x} \\ -\boldsymbol{x}'\boldsymbol{i} & \boldsymbol{i}'\boldsymbol{i} \end{bmatrix} \\
&= \frac{1}{N(\boldsymbol{x}'\boldsymbol{x} - N\bar{x}^2)} \begin{bmatrix} \boldsymbol{x}'\boldsymbol{x} & -N\bar{x} \\ -N\bar{x} & N \end{bmatrix} \\
&= \begin{bmatrix} \frac{\frac{1}{N}\boldsymbol{x}'\boldsymbol{x}}{(\boldsymbol{x}'\boldsymbol{x} - N\bar{x}^2)} & \frac{-\bar{x}}{(\boldsymbol{x}'\boldsymbol{x} - N\bar{x}^2)} \\ \frac{-\bar{x}}{(\boldsymbol{x}'\boldsymbol{x} - N\bar{x}^2)} & \frac{1}{(\boldsymbol{x}'\boldsymbol{x} - N\bar{x}^2)} \end{bmatrix}
\end{aligned}
$$

If $\bar{x} = 0$, $\quad (\boldsymbol{X}'\boldsymbol{X})^{-1} = \begin{bmatrix} \frac{1}{N} & 0 \\ 0 & \frac{1}{(\boldsymbol{x}'\boldsymbol{x})} \end{bmatrix}$

$$\boldsymbol{y} = \beta_0 \boldsymbol{i} + \beta_1 \boldsymbol{x} + \boldsymbol{\epsilon}$$

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

$$\boldsymbol{b} = \begin{bmatrix} b_0 \\ b_1 \end{bmatrix} = \begin{bmatrix} \frac{1}{\boldsymbol{i}'\boldsymbol{i}}[\boldsymbol{i}'\boldsymbol{y} - \frac{\tilde{\boldsymbol{x}}'\boldsymbol{y}}{\tilde{\boldsymbol{x}}'\tilde{\boldsymbol{x}}}\boldsymbol{i}'\boldsymbol{x}] \\ \frac{\tilde{\boldsymbol{x}}'\boldsymbol{y}}{\tilde{\boldsymbol{x}}'\tilde{\boldsymbol{x}}} \end{bmatrix} \qquad \text{if } \bar{x} = 0, \; = \begin{bmatrix} \frac{1}{\boldsymbol{N}}[\boldsymbol{i}'\boldsymbol{y}] \\ \frac{\boldsymbol{x}'\boldsymbol{y}}{\boldsymbol{x}'\boldsymbol{x}} \end{bmatrix}$$

$$\boldsymbol{b} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}$$

$$\boldsymbol{y} = \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y} + \boldsymbol{e}$$

$$\boldsymbol{y} - \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y} = \boldsymbol{e}$$

$$(\boldsymbol{I} - \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}')\boldsymbol{y} = \boldsymbol{e}$$

$$(\boldsymbol{I} - \boldsymbol{P})\boldsymbol{y} = \boldsymbol{e} \qquad \text{Define the projection matrix}$$

$$\boldsymbol{M}\boldsymbol{y} = \boldsymbol{e} \qquad \text{Define the annihilator matrix}$$

P, M Rules

- $\boldsymbol{P} = \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'$ is called the projection matrix or hat matrix.
- $\boldsymbol{I} - \boldsymbol{P} = \boldsymbol{M}$ is called the annihilator matrix or the residual maker.
- $\boldsymbol{M}\boldsymbol{M}' = \boldsymbol{M}\boldsymbol{M} = \boldsymbol{M}$, that is, $\boldsymbol{M}$ is symmetric and idempotent.
- $\boldsymbol{P}\boldsymbol{P}' = \boldsymbol{P}\boldsymbol{P} = \boldsymbol{P}$, that is, $\boldsymbol{P}$ is symmetric and idempotent.
- $\boldsymbol{P}\boldsymbol{X} = \boldsymbol{X}$, that is, $\boldsymbol{X}$ is invariant under $\boldsymbol{P}$
- $\boldsymbol{M}\boldsymbol{y} = \boldsymbol{e}$
- $E(\boldsymbol{e}) = \boldsymbol{M}E(\boldsymbol{y})$
- $Var(\boldsymbol{e}) = \boldsymbol{M}Var(\boldsymbol{y})\boldsymbol{M}' = \sigma^2\boldsymbol{M}$

Trace Tricks

- The trace is the sum of the diagonals of a matrix.
- $tr(c) = c$, which means given a vector $\boldsymbol{y}$, $\boldsymbol{y'y} = tr(\boldsymbol{y'y})$.
- The trace of $P = \boldsymbol{X}(\boldsymbol{X'X})^{-1}\boldsymbol{X}' = rank(\boldsymbol{X})$.
- Given a vector of random variables $\boldsymbol{x}$,

$$E[\boldsymbol{x'x}] = tr[V(\boldsymbol{x})] + E(\boldsymbol{x})'E(\boldsymbol{x})$$

Expectation of Sum of Squares

$$
\begin{aligned}
E[\boldsymbol{e}'\boldsymbol{e}] &= E[tr(\boldsymbol{ee}')] \\
&= tr[E(\boldsymbol{ee}')] && \text{trace is a linear operator} \\
&= tr[Var(\boldsymbol{e}) + E(\boldsymbol{e})'E(\boldsymbol{e})] \\
&= tr[Var(\boldsymbol{e})] + E(\boldsymbol{e})'E(\boldsymbol{e}) \\
&= tr[Var(\boldsymbol{e})] + 0 \quad E(\boldsymbol{e}) = 0 \\
&= tr[\sigma^2 \boldsymbol{M}] \\
&= \sigma^2 tr[(\boldsymbol{I} - \boldsymbol{P})] \\
&= \sigma^2 [N - \text{rank}(\boldsymbol{P})] \\
&= \sigma^2 (N - 2)
\end{aligned}
$$

Estimating variance of errors

$$E[\sum e_i^2] = E[\boldsymbol{e}'\boldsymbol{e}] = \sigma^2(N-2)$$

$$s_e^2 = \frac{\sum_{i=1}^N e_i^2}{(N-2)}$$

This is an unbiased estimator:

$$E[s_e^2] = E[\frac{\sum e_i^2}{(N-2)}] = \frac{E[\boldsymbol{e}'\boldsymbol{e}]}{(N-2)} = \frac{\sigma^2(N-2)}{(N-2)}$$

```
> x <- 1:50
> y <- 8+ .5 * x + rnorm(50,sd=3)

> vcov(lm(y~x))
             (Intercept)             x
(Intercept)   0.69874415 -0.0207547766
x            -0.02075478  0.0008139128

> xtilde <- x - mean(x)

> sum(resid(lm(y~x))^2)/48 / xtilde%*%xtilde
             [,1]
[1,] 0.0008139128
```

Next time: Statistics