Lecture 14

Quantitative Political Science

Prof. Bisbee

Vanderbilt University

Lecture Date: 2023/10/26

Slides Updated: 2023-10-29

Agenda

1. Finishing up correlation
2. Linear regression
3. Sum of squares

Finishing up correlation

- Recall the correlation measure:

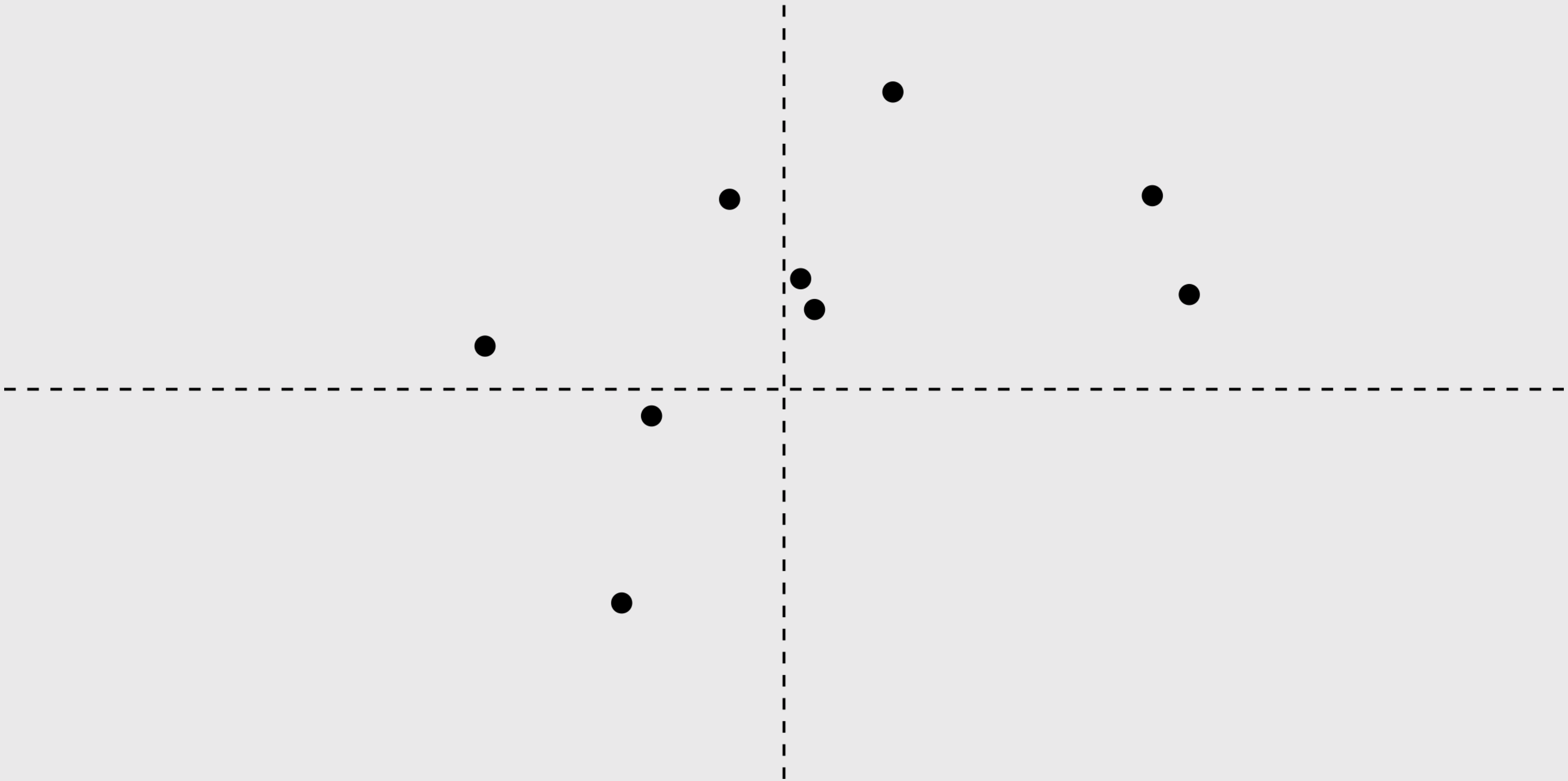
$$ r = \frac{\sum_i(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_i (X_i - \bar{X})^2 \sum_i(Y_i - \bar{Y})^2}} $$

- The quantities in this formula appear a **lot** in regression, so much that they have their own symbols
  - $S_{xy} = \sum_i(X_i - \bar{X})(Y_i - \bar{Y})$
  - $S_{xx} = \sum_i (X_i - \bar{X})^2$
  - $S_{yy} = \sum_i(Y_i - \bar{Y})^2$
- Thus we can rewrite as $r = \frac{S_{xy}}{\sqrt{S_{xx}S_{yy}}}$

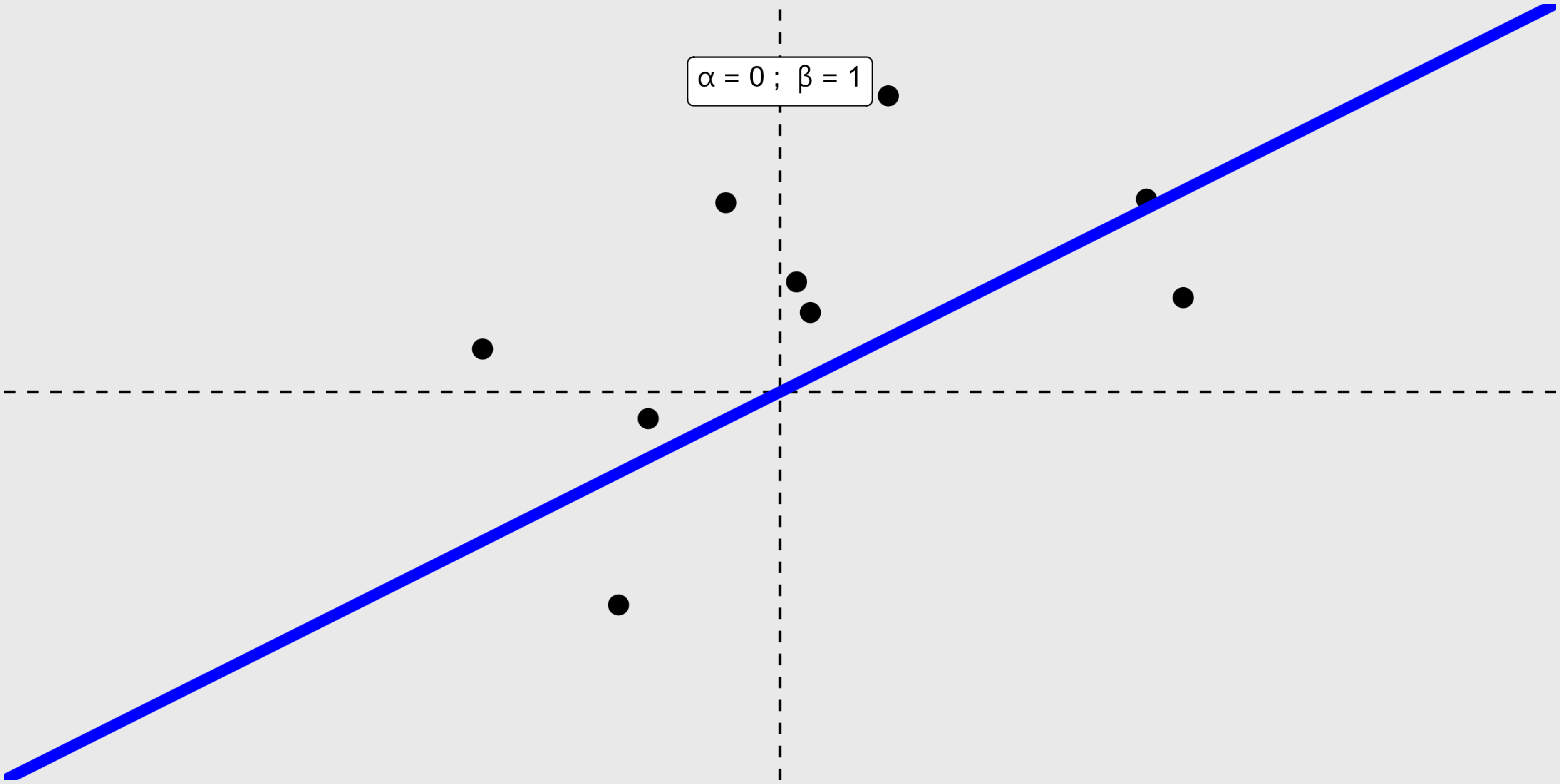
Linear Regression

- Want to say something about the line itself
- Start with **geometry**
    - $y = a + bx$ or $y = \beta_0 + \beta_1 x$ or $y = \alpha + \beta_1 x$
    - a or β_0 or α is the **intercept**
    - b or β_1 is the **slope**
- Many lines we could draw...we want the "best"

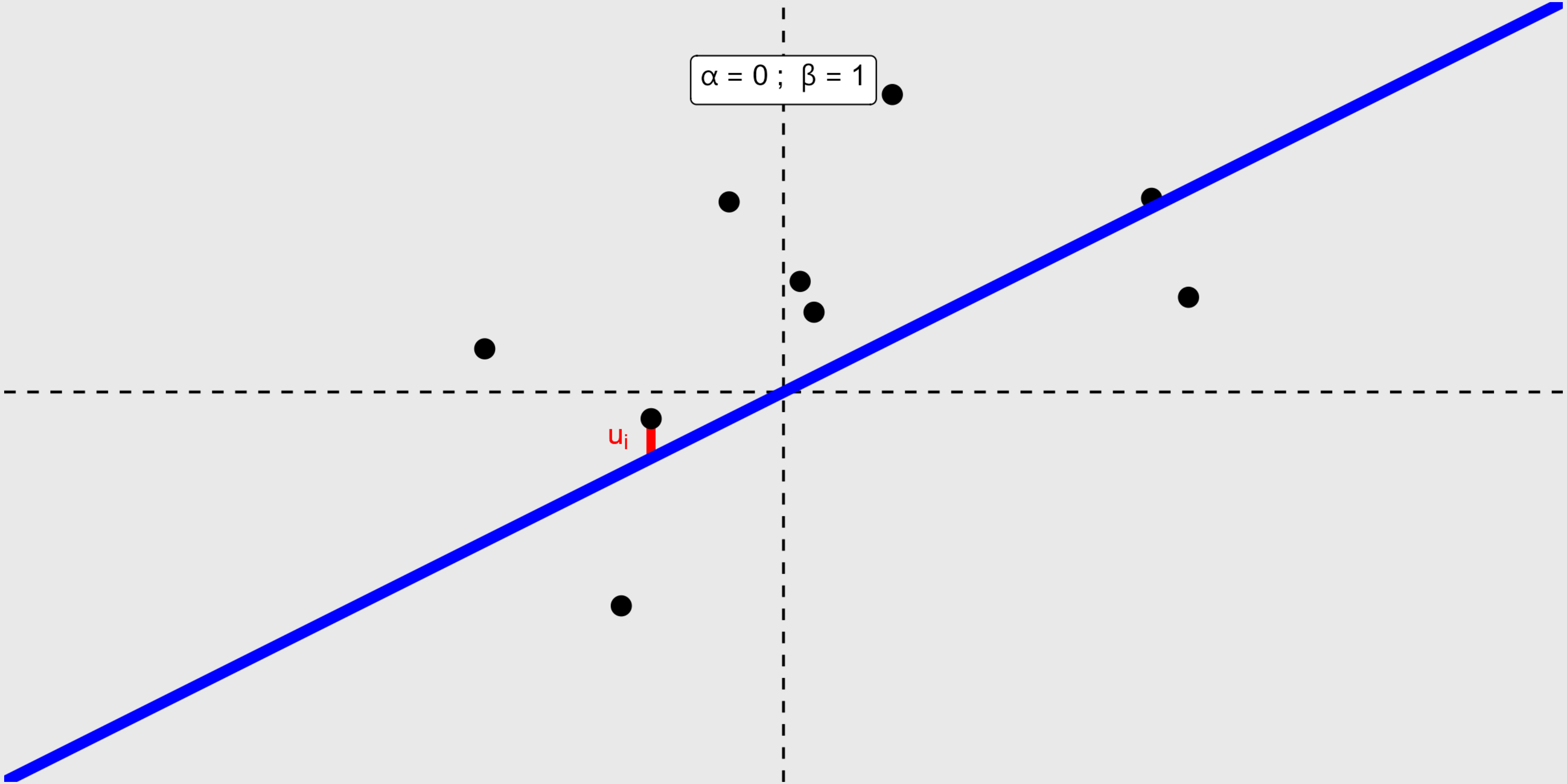
Linear Regression



Linear Regression

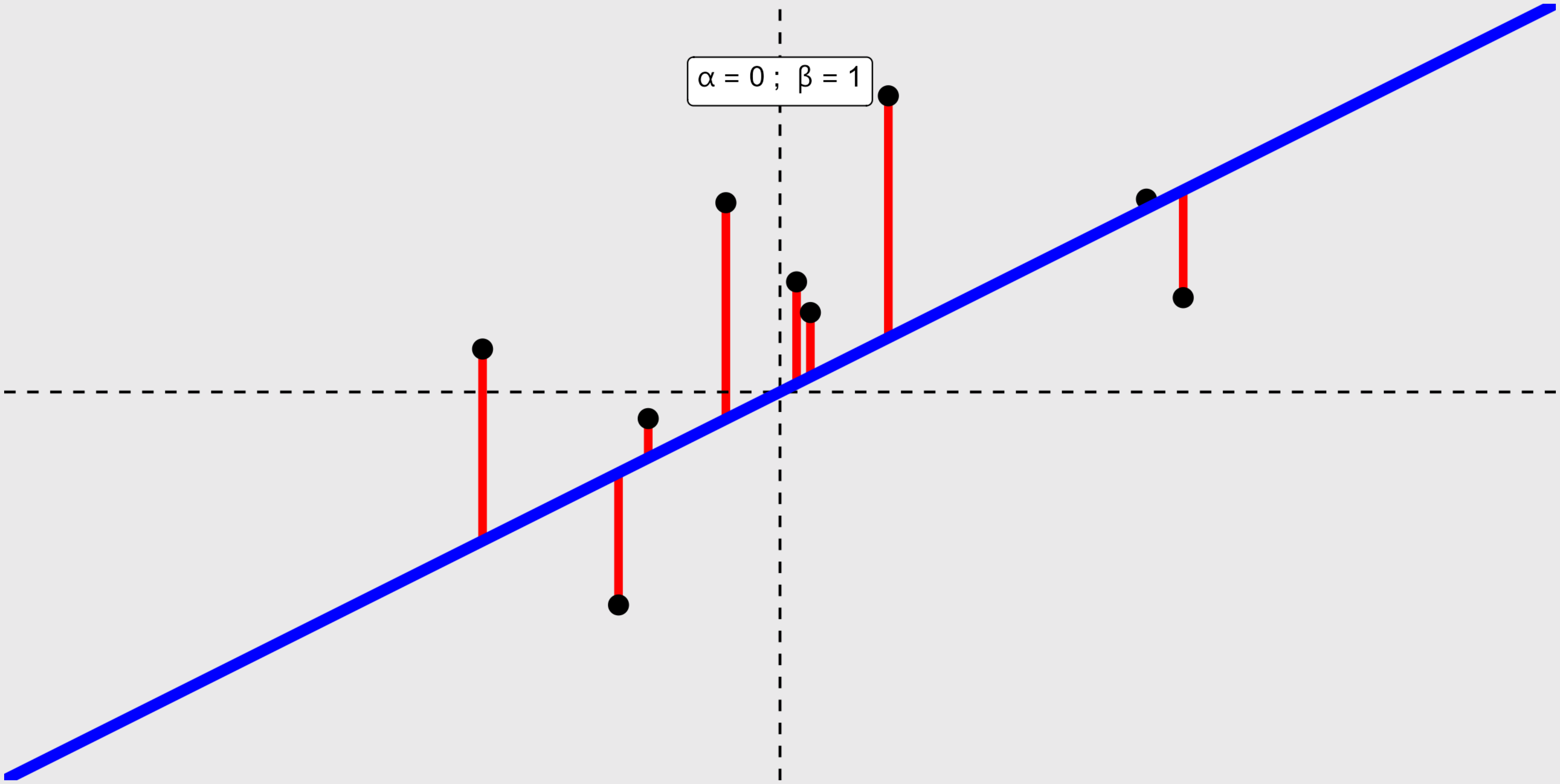


Linear Regression

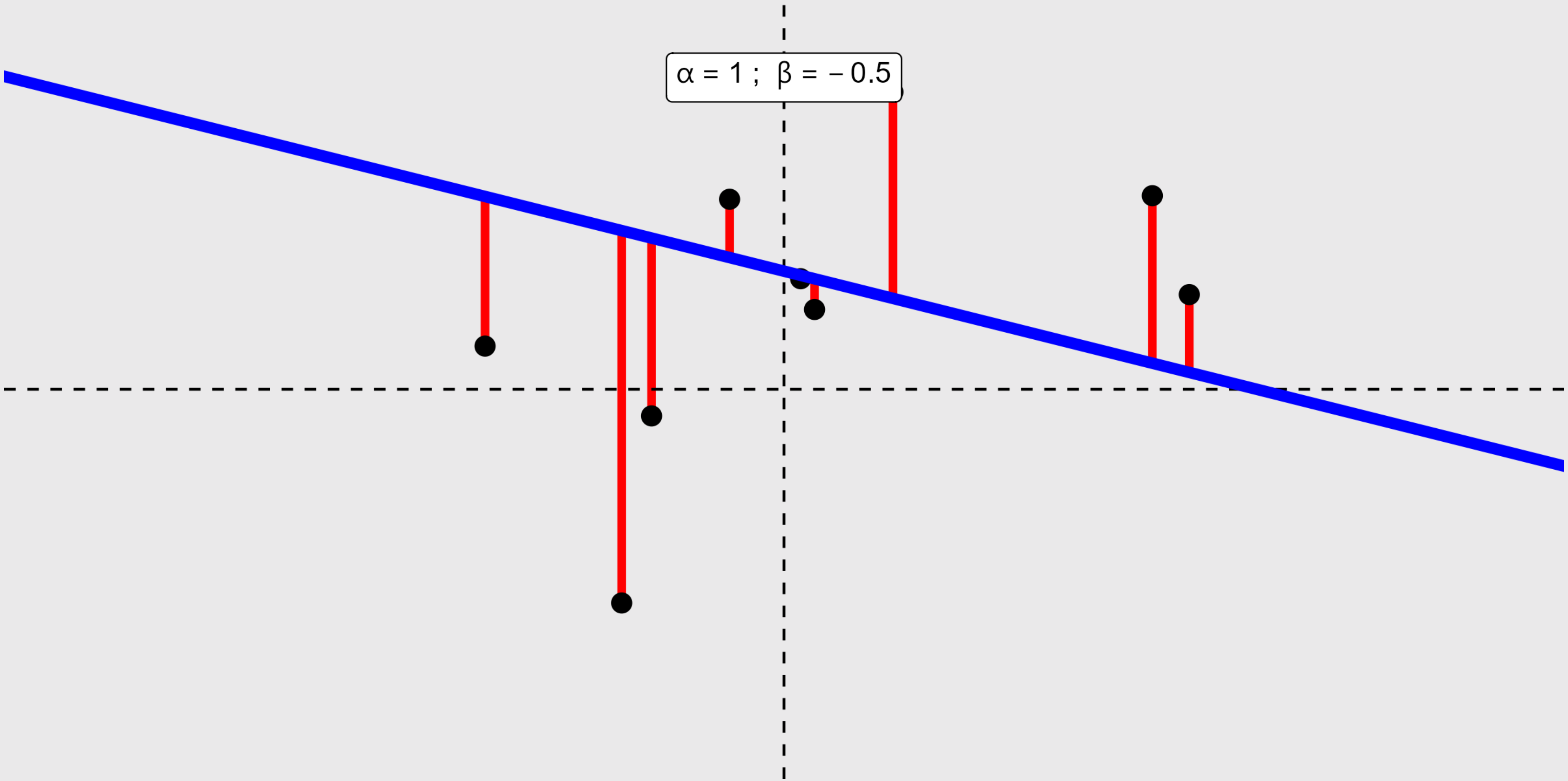


- **Residual**: mistake made by a line \(u_i = y_i - \hat{y}_i\)

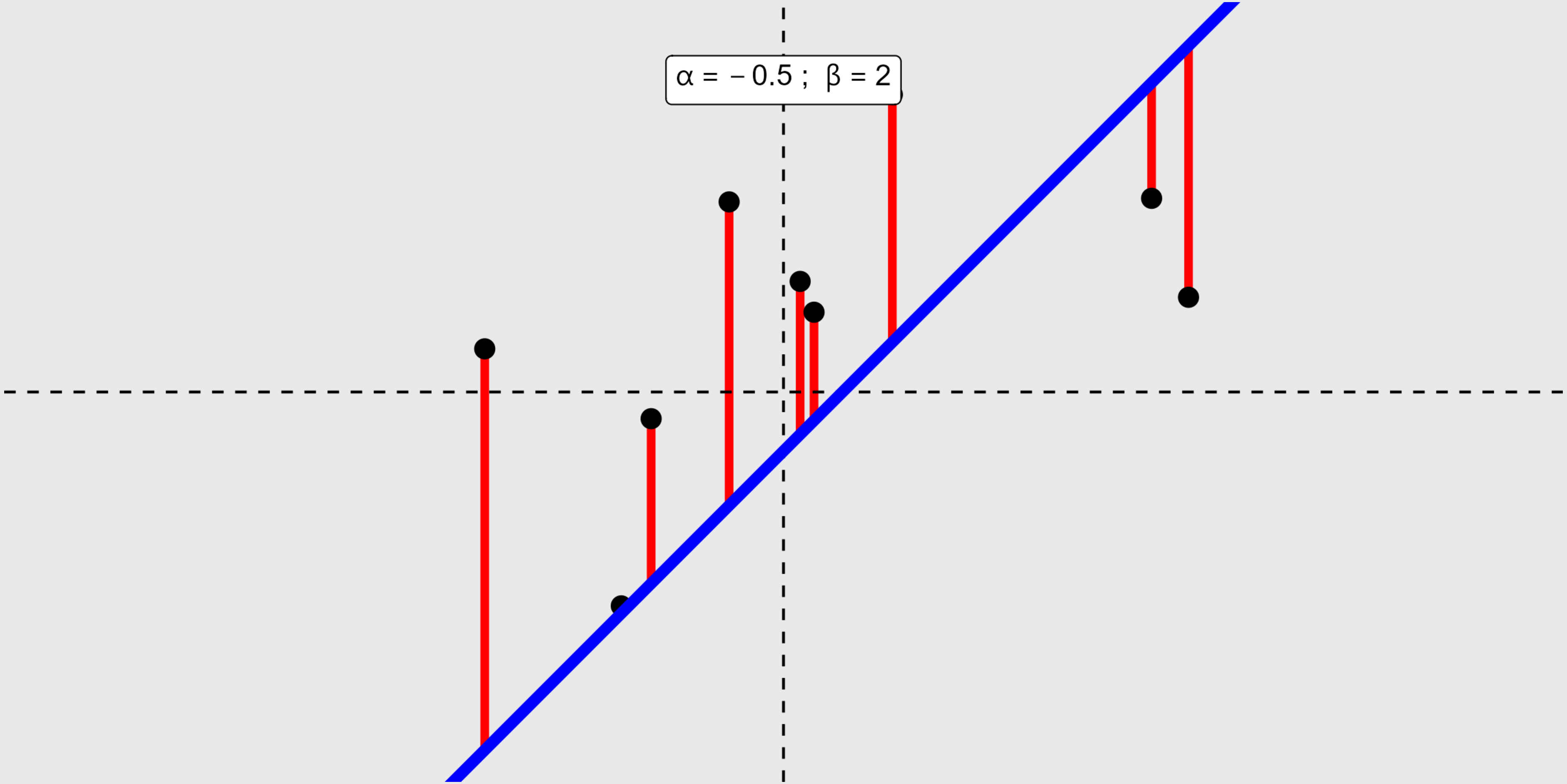
Linear Regression



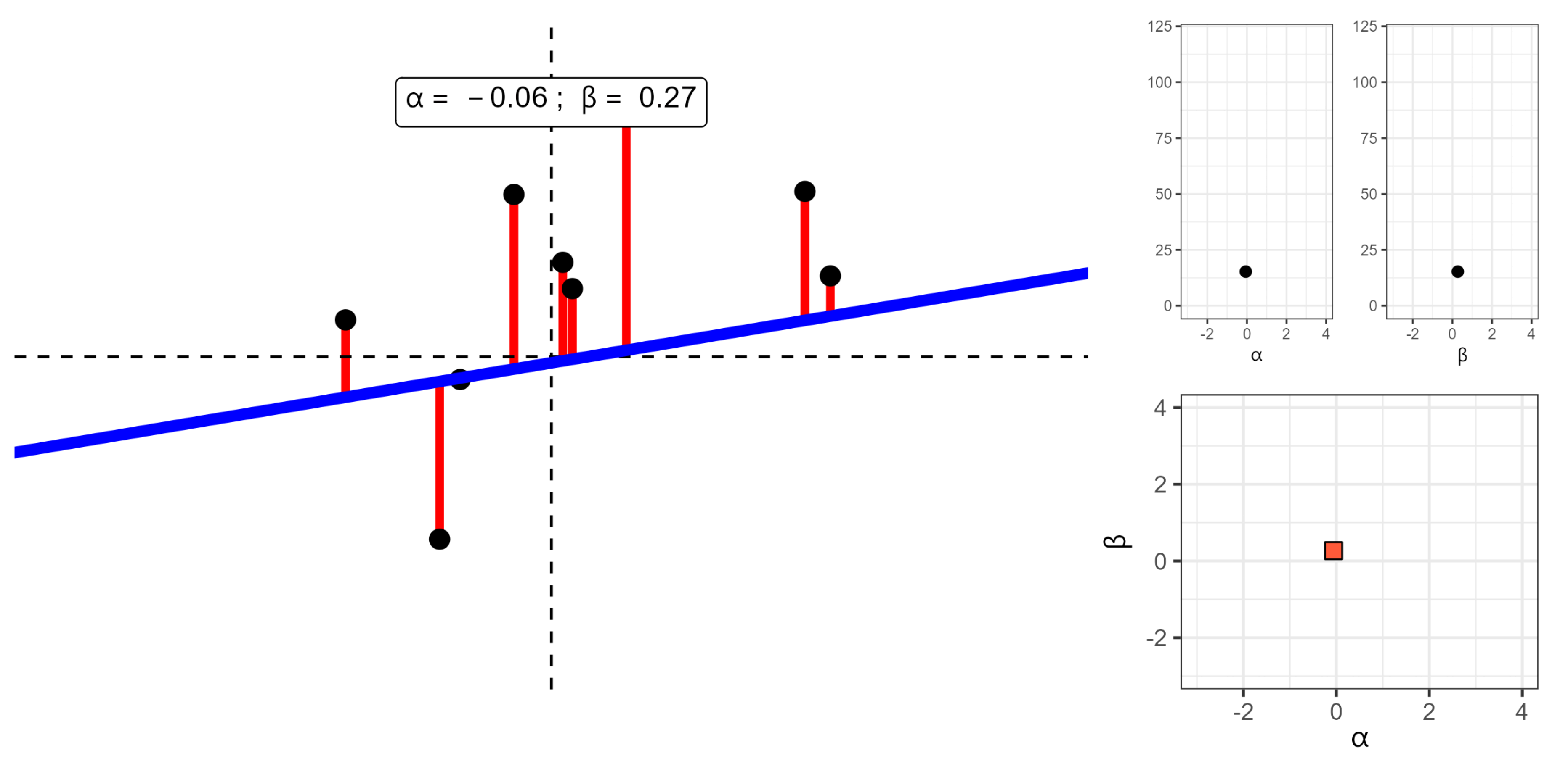
Linear Regression



Linear Regression



Linear Regression

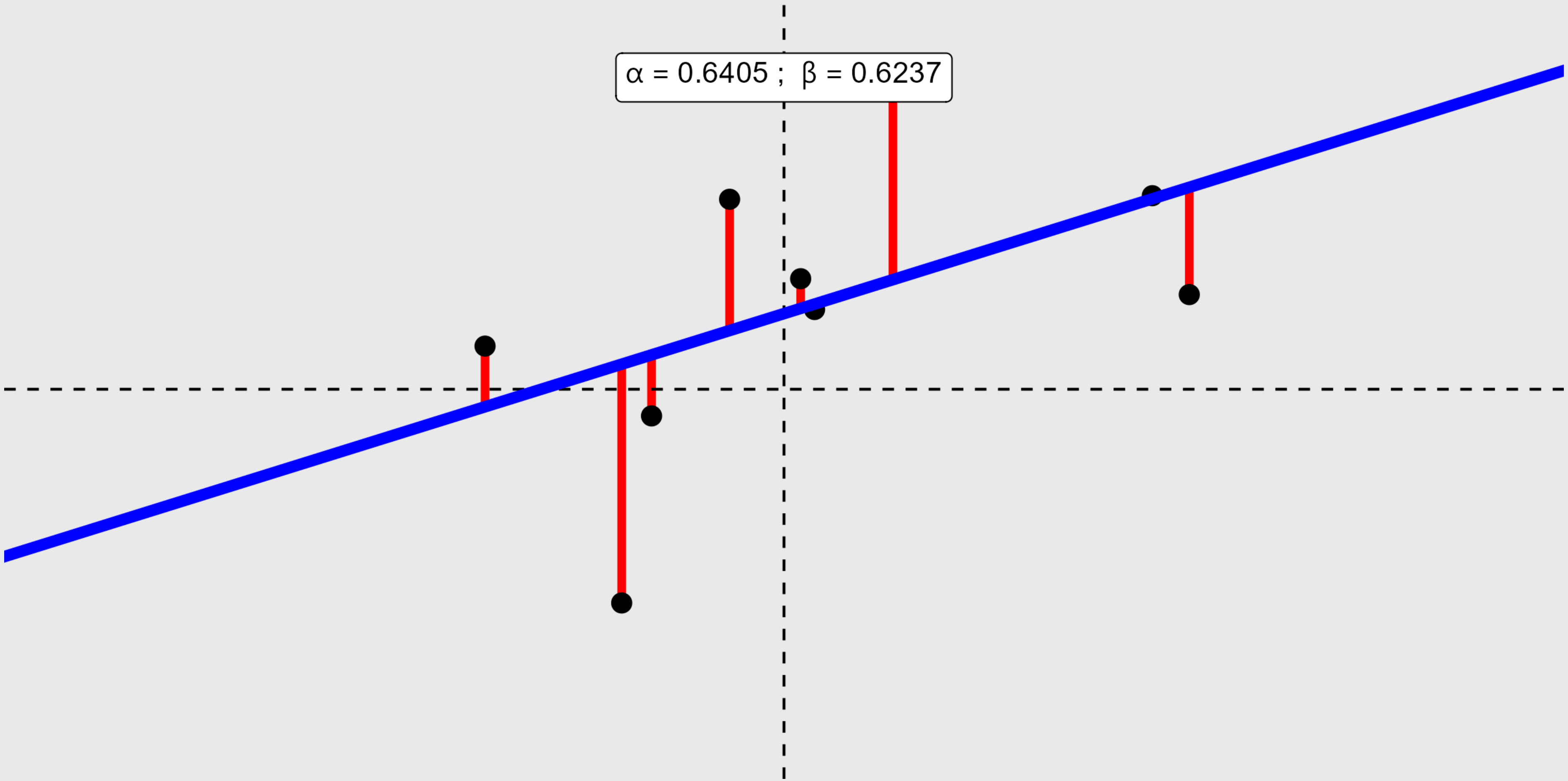


Linear Regression

```
summary(lm(Y~X))
```

```
##
## Call:
## lm(formula = Y ~ X)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.02214 -0.51682  0.02647  0.51386  1.58939
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   0.6405     0.3875   1.653    0.142
## X             0.6237     0.4099   1.522    0.172
##
## Residual standard error: 1.151 on 7 degrees of freedom
## Multiple R-squared:  0.2486,    Adjusted R-squared:  0.1412
## F-statistic: 2.316 on 1 and 7 DF,  p-value: 0.1719
```

Linear Regression



Residuals

- $u_i = y_i - \hat{y}_i$
- Line of best fit is the one that minimizes these mistakes
- Could minimize $|y_i - \hat{y}_i|$ but absolute values are an absolute pain to work with
- Instead, minimize $(y_i - \hat{y}_i)^2$
- Or more accurately, minimize all of them: $SSR = \sum_i (y_i - \hat{y}_i)^2$
- **S**um of **S**quared **R**esiduals (**SSR**)

Regression Line

- Add hats $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$ to reflect **estimates** instead of population parameters (just like θ versus $\hat{\theta}$)

- Substitute this into our definition of u_i

$$ \begin{aligned} u_i &= y_i - \hat{y}_i \\ &= y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i)\\ (u_i)^2 &= [y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i)]^2 \\ \sum_i(u_i)^2 &= \sum_i[y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i)]^2 \\ \end{aligned} $$

- Values of $\hat{\beta}_0$ and $\hat{\beta}_1$ that minimize SSR define the formula for the **least squares line**

Regression Line

- Values of $\hat{\beta}_0$ and $\hat{\beta}_1$ that minimize SSR define the formula for the **least squares line**

$$ \begin{aligned} \frac{\partial SSR}{\partial \hat{\beta}_0} &= \frac{\partial}{\partial \hat{\beta}_0} \sum_i[y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i)]^2 \\ &= -2\sum_i y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i)\\ &= -2 \bigg(\sum_i y_i - n \hat{\beta}_0 - \hat{\beta}_1 \sum_i x_i\bigg) \end{aligned} $$

- Set equal to zero to find the minimum

$$ -2 \bigg(\sum_i y_i - n \hat{\beta}_0 - \hat{\beta}_1 \sum_i x_i\bigg) = 0 $$

Regression Line

- Values of $\hat{\beta}_0$ and $\hat{\beta}_1$ that minimize SSR define the formula for the **least squares line**

$$ \begin{aligned} \frac{\partial SSR}{\partial \hat{\beta}_1} &= \frac{\partial}{\partial \hat{\beta}_1} \sum_i[y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i)]^2\\ &= -2\sum_i \bigg[y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i)\bigg]x_i\\ &= -2 \bigg(\sum_i x_iy_i - \hat{\beta}_0\sum_i x_i - \hat{\beta}_1 \sum_i x^2_i\bigg) \end{aligned} $$

- Set equal to zero to find the minimum

$$ -2 \bigg(\sum_i x_iy_i - \hat{\beta}_0\sum_i x_i - \hat{\beta}_1 \sum_i x^2_i\bigg) = 0 $$

Normal Equations

Solving for zero and rearranging yields the Normal Equations

$$ \begin{aligned} n\hat{\beta}_0 + \hat{\beta}_1 \sum_i x_i &= \sum_i y_i\\ \hat{\beta}_0 \sum_i x_i + \hat{\beta}_1 \sum_i x_i^2 &= \sum_i x_iy_i \end{aligned} $$

- In matrix notation

$$ \begin{bmatrix} n & \sum\nolimits_{i}x_{i} \\ \sum\nolimits_{i}x_{i} & \sum\nolimits_{i}x_{i}^{2} \end{bmatrix} \begin{bmatrix} \widehat{\beta }_{0} \\ \widehat{\beta }_{1} \end{bmatrix} = \begin{bmatrix} \sum\nolimits_{i}y_{i} \\ \sum\nolimits_{i}x_{i}y_{i} \end{bmatrix} $$

- Rearranging

$$ \begin{bmatrix} \widehat{\beta }_{0} \\ \widehat{\beta }_{1} \end{bmatrix} = \begin{bmatrix} n & \sum\nolimits_{i}x_{i} \\ \sum\nolimits_{i}x_{i} & \sum\nolimits_{i}x_{i}^{2} \end{bmatrix} ^{-1} \begin{bmatrix} \sum\nolimits_{i}y_{i} \\ \sum\nolimits_{i}x_{i}y_{i} \end{bmatrix} $$

Aside on Matrix Math

- Read chapter 6 in Brenton's book
- For us today, you need to understand matrix **multiplication** and **inversion**
- Multiplication: Let $\mathbf{A}$ be an $n \times m$ matrix and $\mathbf{B}$ be an $m \times n$ matrix.
    - Denote elements in $\mathbf{A}$ as a_{ij} and elements in $\mathbf{B}$ as b_{ij}, where i index rows and j indexes columns
    - Matrix multiplication creates a new matrix $\mathbf{AB}$ whose ijth element is the **dot product** of the ith row of $\mathbf{A}$ and the jth column of $\mathbf{B}$.

$$ \mathbf{A}= \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix} \text{ and } \mathbf{B=} \begin{bmatrix} b_{11} & b_{12} \\ b_{21} & b_{22} \end{bmatrix} $$

$$ \mathbf{AB}= \begin{bmatrix} a_{11}*b_{11} + a_{12}*b_{21} & a_{11}*b_{12} + a_{12}*b_{22} \\ a_{21}*b_{11} + a_{22}*b_{21} & a_{21}*b_{12} + a_{22}*b_{22} \end{bmatrix} $$

Aside on Matrix Math

- Do these:

$$ \mathbf{A}= \begin{bmatrix} 2 & 0 \\ -5 & 3 \end{bmatrix} \text{ and } \mathbf{B=} \begin{bmatrix} 4 & 10 \\ 1 & 3 \end{bmatrix} $$

$$ \mathbf{AB}= \begin{bmatrix} ? & ? \\ ? & ? \end{bmatrix} $$

$$ \mathbf{A}= \begin{bmatrix} 2 & 10 \\ 0 & 1 \\ -1 & 5 \end{bmatrix} \text{ and } \mathbf{B=} \begin{bmatrix} 1 & 4 \\ -1 & 10 \end{bmatrix} $$

$$ \mathbf{AB}= \begin{bmatrix} ? & ? \\ ? & ? \\ ? & ? \end{bmatrix} $$

Aside on Matrix Math

The inverse of a 2×2 matrix $$ \mathbf{A=} \begin{bmatrix} a & b \\ c & d \end{bmatrix} $$ is $$ \mathbf{A}^{\mathbf{-1}}=\frac{1}{\det \mathbf{A}} \begin{bmatrix} d & -b \\ -c & a \end{bmatrix} =\frac{1}{ad-bc} \begin{bmatrix} d & -b \\ -c & a \end{bmatrix} $$

so

$$ \begin{bmatrix} n & \sum_ix_i \\ \sum_ix_i & \sum_ix_i^2 \end{bmatrix} ^{-1} = \frac{1}{n\sum_ix_i^2 - (\sum_ix_i)^2} \begin{bmatrix} \sum_ix_i^2 & -\sum_ix_i \\ -\sum_ix_i & n \end{bmatrix} $$

Normal Equations

$$ \begin{bmatrix} \widehat{\beta}_{0} \\ \widehat{\beta}_{1} \end{bmatrix} = \frac{1}{n\sum_ix_i^2 - (\sum_ix_i)^2} \begin{bmatrix} \sum_ix_i^2 & -\sum_ix_i \\ -\sum_ix_i & n \end{bmatrix} \begin{bmatrix} \sum_iy_i \\ \sum_ix_iy_i \end{bmatrix} $$

- Which means

$$ \widehat{\beta }_0 = \frac{\sum_ix_i^2\sum_iy_i-\sum_ix_i\sum_ix_iy_i}{n\sum_ix_i^2-\left(\sum_ix_i\right)^2} $$

- and

$$ \widehat{\beta }_{1}=\frac{n\sum_ix_iy_i-\sum_ix_i\sum_iy_i}{n\sum_ix_i^2-\left(\sum_ix_i\right)^2} $$

Normal Equations

- Simplifying

$$ \begin{aligned} \widehat{\beta }_{0} &= \frac{n\overline{y}\sum_ix_i^2-n\overline{x}\sum_ix_iy_i}{n\sum_ix_i^2-\left(n \overline{x}\right) ^2}\\ &=\frac{\overline{y}\sum_ix_i^2-\overline{x}\sum_ix_iy_i}{\sum_ix_i^2-n\left(\overline{x}\right) ^2}\\ \widehat{\beta }_{1} &= \frac{n\sum_ix_iy_i-n^2\overline{x}\overline{y}}{n\sum_ix_i^2-n^2\left(\overline{x}\right)^2}\\ &=\frac{\sum_ix_iy_i-n\overline{x}\overline{y}}{\sum_ix_i^2-n\left(\overline{x}\right) ^2} \end{aligned} $$

Normal Equations

- Note that

$$ \begin{aligned} S_{xx} &=\sum_i\left(X_i-\bar{X}\right)^2\\ &=\sum_iX_i^2+\sum_i\bar{X}^2-\sum_i2X_i\bar{X} \\ &= \sum_iX_i^2+n\bar{X}^2-\bar{X}2n\frac{1}{n}\sum_iX_i \\ &= \sum_iX_i^2+n\bar{X}^2-\bar{X}2n\bar{X} \\ &= \sum_iX_i^2+n\bar{X}^2-2n\bar{X}^2 \\ &=\sum_iX_i^2-n\bar{X}^2 \end{aligned} $$

Normal Equations

- (Trivially, this also gives us $S_{yy} = \sum_i Y_i^2 - n\bar{Y}^2)$
- Also note that

$$ \begin{aligned} S_{xy} &=\sum_i\left(X_i-\bar{X}\right) \left(Y_i-\bar{Y}\right)\\ &=\sum_iX_iY_i-\sum_iX_i\bar{Y}-\sum_i\bar{X}Y_i+\sum_i\bar{X}\bar{Y} \\ &=\sum_iX_iY_i-n\sum_i\bar{X}\bar{Y} \end{aligned} $$

Normal Equations

- Therefore

$$ \begin{aligned} \hat{\beta}_1 &= \frac{S_{xy}}{S_{xx}} \end{aligned} $$

- Note that \(\frac{cov(x,y)}{var(x)} = \frac{\frac{S_{xy}}{n}}{\frac{S_{xx}}{n}}\)

- So

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

Normal Equations

- For $\hat{\beta}_0$, start with the derivative set to zero

$$ \begin{aligned} -2\left(\sum_iy_i-n\widehat{\beta }_0-\widehat{\beta }_{1}\sum_ix_i\right) &= 0 \\ \sum_iy_i-n\widehat{\beta }_0-\widehat{\beta }_{1}\sum_ix_i &= 0 \\ n\bar{y}-n\widehat{\beta }_0-n\widehat{\beta }_{1}\bar{x} &= 0 \\ \widehat{\beta }_0 &=\bar{y}-\widehat{\beta }_{1}\bar{x},\text{and so} \\ \widehat{\beta }_0 &=\bar{y}-\frac{S_{xy}}{S_{xx}}\bar{x} \end{aligned} $$

Properties of the Least Squares Line

- Nothing we've done yet requires assumptions about distributions of x or y
- Just straight math footwork
- Some additional properties

Prop 1. $\hat{\beta}_1 = \frac{\Delta\hat{y}}{\Delta x}$:

- Take derivative of $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$ with respect to x.
- $\frac{d\hat{y}}{dx} = \hat{\beta}_1$. A one-unit change in x is associated with a $\hat{\beta}_1$ unit change in $\hat{y}$. Or $\hat{\beta}_1 = \frac{\Delta\hat{y}}{\Delta x}$

Prop 2. $\sum_i \hat{u}_i = 0$:

- We define $\hat{u}_i = y_i - \hat{y}_i$ and substitute in $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1x_i$ to get $\hat{u}_i = y_i - (\hat{\beta}_0 + \hat{\beta}_1x_i)$.
- Sum it to see $\sum_i\hat{u}_i = \sum_iy_i - (\hat{\beta}_0 + \hat{\beta}_1x_i)$, and recall from $\frac{\partial SSR}{\partial \hat{\beta}_0} = 0$ that $\sum_i y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i) = 0$.

Properties of the Least Squares Line

Prop 3. $\bar{\hat{u}}_i = 0$:

- From previous slide, $\sum_i \hat{u}_i = 0$.
- Thus $\frac{1}{n}n\sum_i \hat{u}_i = 0$ and therefore $n\bar{\hat{u}}_i = 0$ so $\bar{\hat{u}}_i = 0$

Prop 4. $cov(x,\hat{u}) = 0$:

- We know from F.O.C. for $\hat{\beta}_1$ that $\sum_i \bigg[y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i)\bigg]x_i = 0$.
- We defined $y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i) = \hat{u}_i$ so we can rewrite as $\sum_i \hat{u}_ix_i = 0$

$$ \begin{aligned} cov(x,\hat{u}) &= \frac{\sum_i(\hat{u}_i)(x_i - \bar{x})}{n}\\ &= \frac{\sum_i \hat{u}_ix_i}{n} - \frac{\bar{x}\sum_i\hat{u}_i}{n} \\ &= 0 - 0 \end{aligned} $$

Properties of the Least Squares Line

Prop 5. $\frac{1}{n}\sum_i \hat{y}_i = \frac{1}{n}\sum_i y_i$

$$ \begin{aligned} y_i &= \hat{y}_i + \hat{u}_i\\ \sum\nolimits_i y_i &= \sum\nolimits_i \hat{y}_i + \sum\nolimits_i \hat{u}_i\\ \sum\nolimits_i y_i &= \sum\nolimits_i \hat{y}_i + 0\\ \frac{1}{n}\sum\nolimits_iy_i &= \frac{1}{n}\sum\nolimits_i \hat{y}_i \end{aligned} $$

Prop 6. The coordinate $(\bar{x},\bar{y})$ is always on the line of best fit

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

$$ \begin{aligned} \hat{y} &= \hat{\beta}_0 + \hat{\beta}_1 x_i\\ \hat{y}(\bar{x}) &= \bar{y} - \hat{\beta}_1 \bar{x} + \hat{\beta}_1 \bar{x}\\ \hat{y}(\bar{x}) &= \bar{y} \end{aligned} $$

Properties of the Least Squares Line

Prop 7. $cov(\hat{y}_i,\hat{u}_i) = 0$

$$ \begin{aligned} cov(\hat{y}_i,\hat{u}_i) &= \frac{\sum(\hat{y}_i - \bar{\hat{y}})(\hat{u}_i - \bar{\hat{u}})}{n}\\ &= \frac{\sum(\hat{y}_i - \bar{y})\hat{u}_i}{n}\\ &= \frac{\sum\hat{y}_i\hat{u}_i}{n} - \frac{\bar{y}\sum\hat{u}_i}{n}\\ &= \frac{\sum(\hat{\beta}_0 + \hat{\beta}_1 x_i)\hat{u}_i}{n} - 0\\ &= \frac{\hat{\beta}_0\sum\hat{u}_i + \hat{\beta}_1\sum x_i\hat{u}_i}{n} - 0\\\ &= 0 - 0 \end{aligned} $$

Sum of Squares

- Process of fitting least squares is **decomposing** y_i into two parts: $\hat{y}_i$ and $\hat{u}_i$
- Total sum of squares (**SST**): $\sum_i(y_i - \bar{y})^2$
- Explained sum of squares (**SSE**): $\sum_i(\hat{y}_i - \bar{y})^2$
- Residual sum of squares (**SSR**): $\sum_i \hat{u}_i^2$
- Prove: $SST = SSR + SSE$

Sum of Squares

$$ \begin{aligned} SST &= \sum_i(y_i - \bar{y})^2\\ &= \sum_i(y_i - \hat{y}_i + \hat{y}_i + \bar{y})^2\\ &= \sum_i(\hat{u}_i + \hat{y}_i + \bar{y})^2\\ &= \sum_i(\hat{u}_i)^2 + \sum_i (\hat{y}_i - \bar{y})^2 + 2\sum_i (\hat{y}_i - \bar{y})\\ &= SSR + SSE + 2\sum_i (\hat{y}_i + \bar{y})\hat{u}_i \end{aligned} $$

- We just demonstrated that $cov(\hat{y}_i,\hat{u}_i) = \frac{\sum(\hat{y}_i - \bar{y})\hat{u}_i}{n} = 0$
- Therefore $SST = SSR + SSE + 0$

R^2

- $SST = \sum_i(y_i - \bar{y})^2$ is the sample variance of y.
- If y could be perfectly explained by a straight line over values of x, the $SSE = \sum_i(\hat{y}_i - \bar{y})^2$ would be equal to the SST
- Therefore $\frac{SSE}{SST} = 1$.
- This never actually happens, but we can use this ratio to measure the "goodness of fit"
    - $R^2 = \frac{SSE}{SST}$
    - The proportion of sample variation in y that is explained by x
    - $R^2 = \frac{SSE}{SST} = 1 - \frac{SSR}{SST}$
- The name comes from the fact that, in the bivariate context, $R^2 = (r)^2$