# STA302/1001: Methods of Data Analysis

Instructor: Fang Yao

Chapter 2: Simple Linear Regression (Part I)

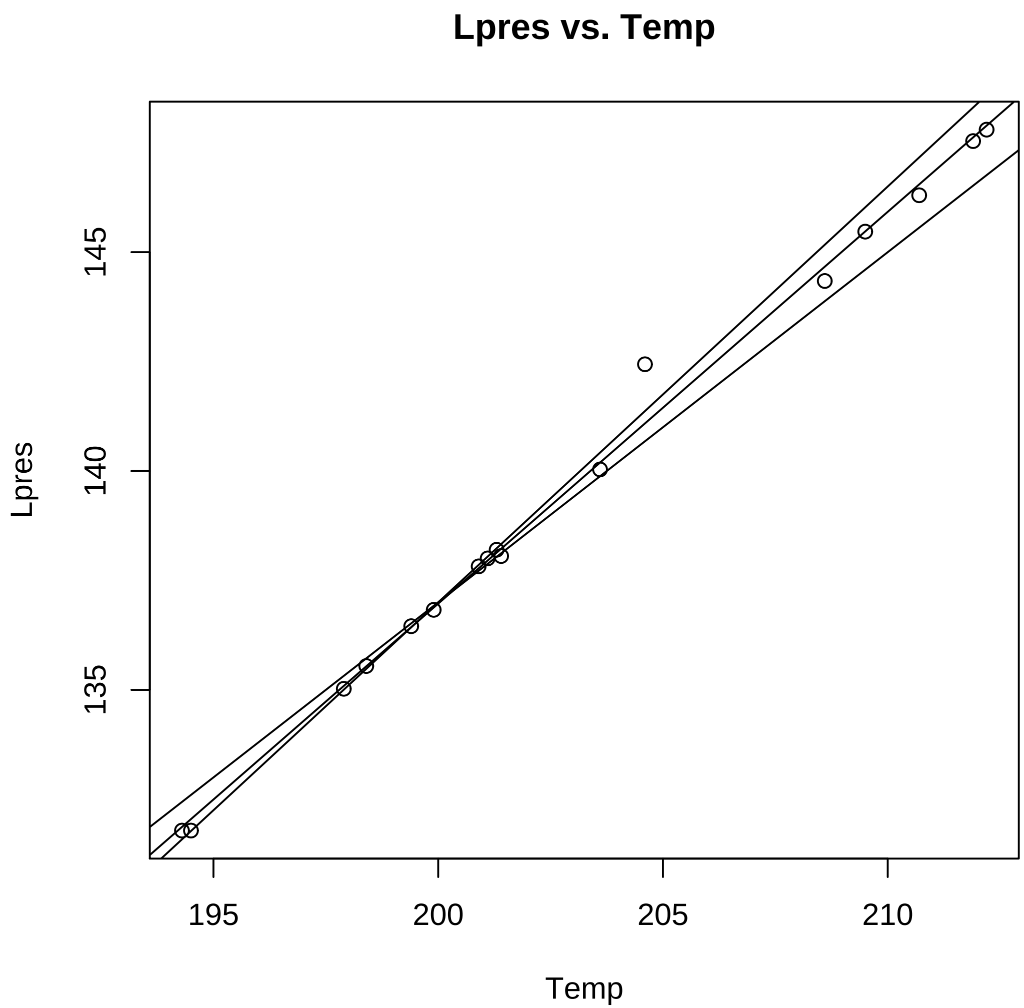
# Forbes' Data

Forbes' 1857 Data on Boiling Point and Barometric Pressure for 17 Locations in the Alps and Scotland:

| Case Number | $Temp$ (°F) | $Pressure$ (Inches Hg) | $Lpres = 100 \times \log(Pressure)$ |
|---|---|---|---|
| 1 | 194.5 | 20.79 | 131.79 |
| 2 | 194.3 | 20.79 | 131.79 |
| 3 | 197.9 | 22.40 | 135.02 |
| $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ |
| 17 | 212.2 | 30.06 | 147.80 |

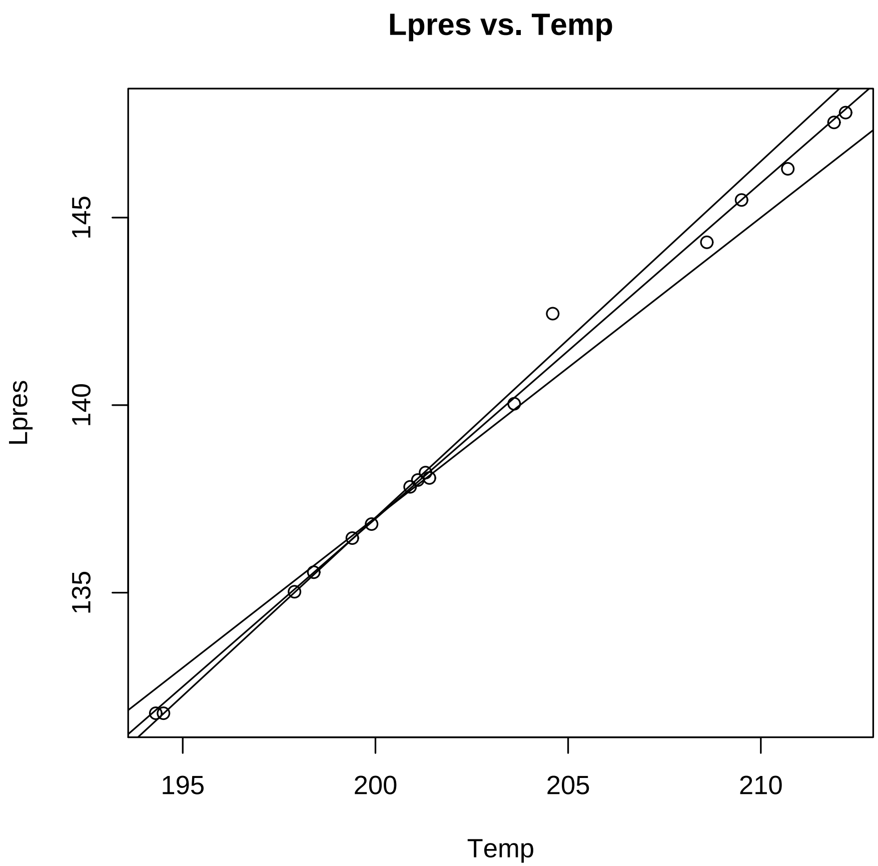
# Simple Linear Regression (SLR) Model

Plot of $Lpres$ vesus $Temp$



# Simple Linear Regression (SLR) Model

- $\mathrm{E}(Y|X=x) = \beta_0 + \beta_1 x$
  $\mathrm{Var}(Y|X=x) = \sigma^2$
- parameters to estimate: $\beta_0, \beta_1, \sigma^2$



# An Alternative Formulation

$$
\begin{aligned}
\mathrm{E}(Y|X=x) &= \beta_0 + \beta_1 x \\
\mathrm{Var}(Y|X=x) &= \sigma^2
\end{aligned}
$$

- another way to express the model

$$y_i = \beta_0 + \beta_1 x_i + e_i$$

$$\mathrm{E}(e_i) = 0, \quad \mathrm{Var}(e_i) = \sigma^2, \quad e_i's \text{ are i.i.d.}$$

- $e_i$: statistical error (no negative meaning here)
  the vertical distance between $y_i$ and the "true value"
  $\mathrm{E}(Y|X = x_i)$

# Parameter Estimation

- notation: parameters: $\alpha$, $\beta$, $\gamma$
  estimators: $\hat{\alpha}$, $\hat{\beta}$, $\hat{\gamma}$
- define: fitted value for case $i$

$$\hat{y}_i = \hat{\mathrm{E}}(Y|X = x_i) = \hat{\beta}_0 + \hat{\beta}_1 x_i$$

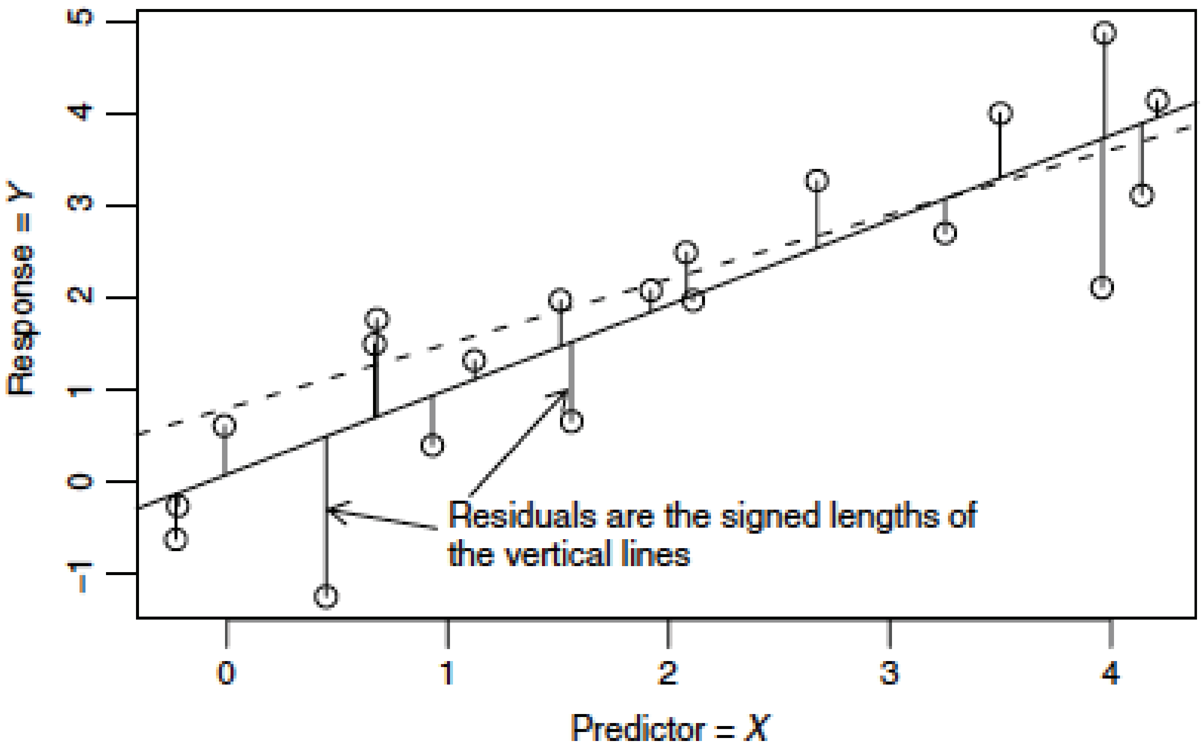
- it is a point on the fitted line
- define: residual for case $i$

$$\hat{e}_i = y_i - \hat{y}_i = y_i - \hat{\mathrm{E}}(Y|X = x_i) = y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i$$

- vertical distance between $y_i$ and its fitted value

# Ordinary Least Squares

Illustration of OLS fitting with residuals shown as the vertical distances.



# Ordinary Least Squares (cont...)

- sometimes called least squares
- a method for parameter estimation
- define residual sum of squares (RSS)

$$RSS(\beta_0, \beta_1) = \sum_{i=1}^{n} [y_i - (\beta_0 + \beta_1 x_i)]^2$$

- we estimate $(\beta_0, \beta_1)$ with the pair that minimizes $RSS(\beta_0, \beta_1)$

$$(\hat{\beta}_0, \hat{\beta}_1) = \operatorname{argmin}_{\beta_0, \beta_1} RSS(\beta_0, \beta_1)$$

# Ordinary Least Squares (cont...)

- the least squares estimates of $\beta_0$ and $\beta_1$ minimize

$$RSS(\beta_0, \beta_1) = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2$$

- differentiate w.r.t. to $\beta_0$ and $\beta_1$ and set the results to 0:

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

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

$$\Rightarrow \beta_0 n + \beta_1 \sum_i x_i = \sum_i y_i, \quad \beta_0 \sum_i x_i + \beta_1 \sum_i x_i^2 = \sum_i x_i y_i$$

# Ordinary Least Squares (cont...)

- from the previous slide:
  $\beta_0 n + \beta_1 \sum_i x_i = \sum_i y_i, \quad \beta_0 \sum_i x_i + \beta_1 \sum_i x_i^2 = \sum_i x_i y_i$
- solve these equations, denote $\bar{x} = \frac{1}{n} \sum_i x_i$, $\bar{y} = \frac{1}{n} \sum_i y_i$

$$SXY = \sum_i (x_i - \overline{x})(y_i - \overline{y}) = \sum_i x_i y_i - n\bar{x}\bar{y}$$

$$SXX = \sum_i (x_i - \overline{x})^2 = \sum_i x_i^2 - n\bar{x}^2$$

- we obtain (see Table 2.1 of text for more notation)

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}, \quad \hat{\beta}_1 = \frac{SXY}{SXX}$$

# SLR Model for Forbes' Data

For Forbes' data, $x$ is $Temp$ and $y$ is $Lpres$ and

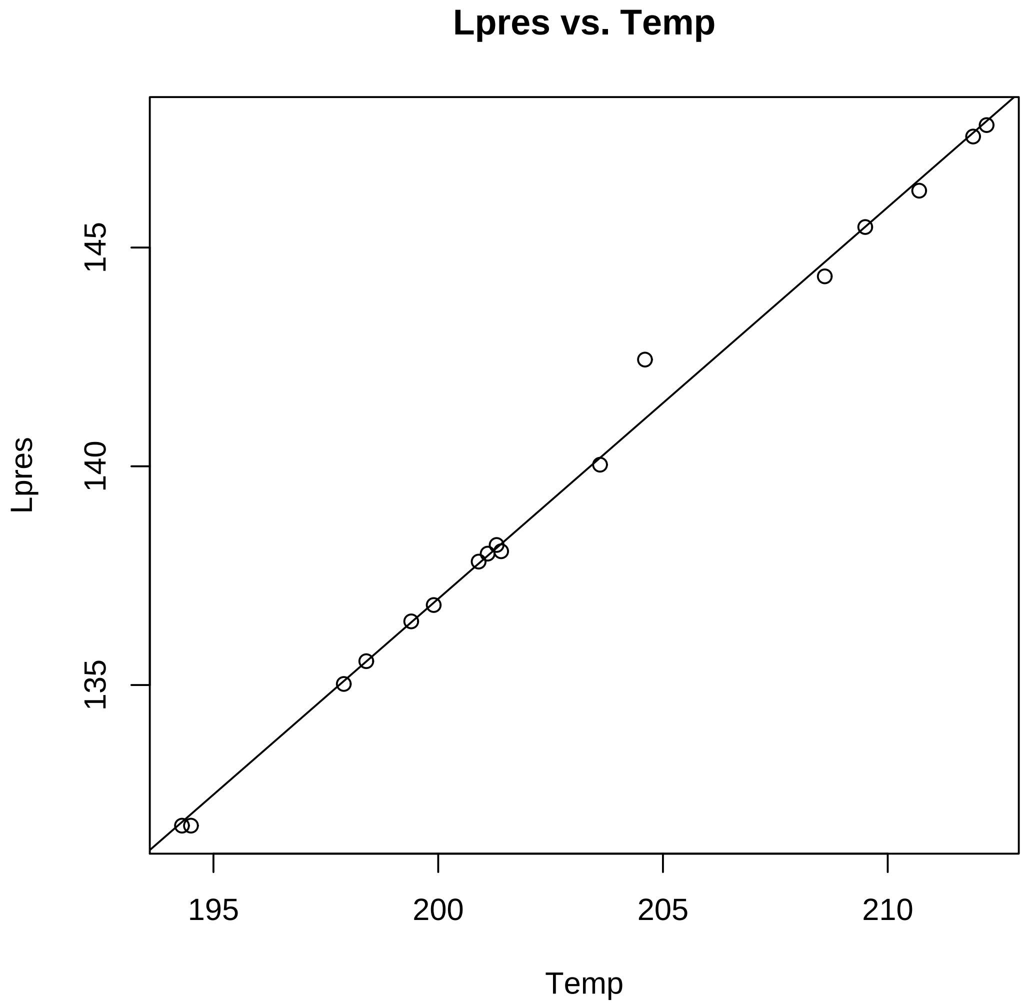
$$\overline{x} = 202.95, \quad SXX = 530.78, \quad SXY = 475.31,$$
$$\overline{y} = 139.61, \quad SYY = 427.79$$

- and therefore $\hat{\beta}_1 = \frac{SXY}{SXX} = 0.895$
  and $\hat{\beta}_0 = \overline{y} - \hat{\beta}_1 \overline{x} = -42.138$
- the fitted line, or estimated line, is

$$\hat{\mathrm{E}}(Lpres|Temp) = -42.138 + 0.895 \times Temp$$

# SLR Model for Forbes' Data

Plot of $Lpres$ vesus $Temp$ with fitted line



# Estimating $\sigma^2$

- $\hat{\sigma}^2$ is essentially the average size of $\hat{e}_i^2 = (y_i - \hat{y}_i)^2$
- $\hat{\sigma}^2$ can be obtained by dividing $RSS = \sum \hat{e}_i^2$ by its degrees of freedom ($df$)

$$\hat{\sigma}^2 = \frac{RSS}{n-2}$$

- why the $df$ is $(n-2)$? compare to $s^2 = \frac{1}{n-1} \sum_i (y_i - \bar{y})^2$
- $\hat{\sigma}^2 = \frac{RSS}{n-2}$ is called "residual mean square"
- $\hat{\sigma}$ is called "standard error of regression"
- if $e_i$ are i.i.d. from $N(0, \sigma^2)$, then $RSS/\sigma^2 \sim \chi^2_{n-2}$

# Estimating $\sigma^2$ (cont...)

- $RSS$ can be calculated by its definition

$$
\begin{aligned}
RSS &= \sum_i (y_i - \hat\beta_0 - \hat\beta_1 x_i)^2 = \sum_i [y_i - \bar y - \hat\beta_1 (x_i - \bar x)]^2 \\
&= SYY + \hat\beta_1^2 SXX - 2\hat\beta_1 SXY \\
&= SYY + \frac{SXY^2}{SXX^2} SXX - 2\frac{SXY^2}{SXX} \\
&= SYY - \frac{SXY^2}{SXX} = SYY - \hat\beta_1^2 SXX
\end{aligned}
$$

- Forbes' data:

$$
\begin{aligned}
RSS &= 427.79402 - \frac{475.31224^2}{530.78235} = 2.15493 \\
\hat\sigma^2 &= 2.15493/(17-2) = 0.14366, \text{ i.e., } \hat\sigma = 0.37903
\end{aligned}
$$

# Properties of Least Squares Estimates

- $\hat{\beta}_0$ and $\hat{\beta}_1$ can be written as a linear combination of $y_i$'s
- let $c_i = \frac{x_i - \bar{x}}{SXX}$ (free of $y_i$'s), note $\sum_i (x_i - \bar{x})\bar{y} = 0$

$$\hat{\beta}_1 = \sum_i \left(\frac{x_i - \bar{x}}{SXX}\right) y_i = \sum_i c_i y_i$$

- the fitted line passes through $(\bar{x}, \bar{y})$
- estimators are unbiased, denote $\mathbb{X} = \{x_1, \ldots, x_n\}$ (in general $\hat{\theta}$ is unbiased for $\theta$ if $\mathrm{E}(\hat{\theta}) = \theta$)

$$\mathrm{E}(\hat{\beta}_0|\mathbb{X}) = \beta_0, \quad \mathrm{E}(\hat{\beta}_1|\mathbb{X}) = \beta_1, \quad \mathrm{E}(\hat{\sigma}^2|\mathbb{X}) = \sigma^2$$

# roperties of Least Squares Estimates (cont...)

- first recall $\hat{\beta}_1 = \frac{SXY}{SXX} = \sum_i \left(\frac{x_i - \bar{x}}{SXX}\right) y_i = \sum_i c_i y_i$

$$\mathrm{E}(\hat{\beta}_1 | \mathbb{X}) = \mathrm{E}(\sum_i c_i y_i | X = x_i) = \sum_i c_i \mathrm{E}(y_i | X = x_i)$$

$$= \sum_i c_i (\beta_0 + \beta_1 x_i) = \beta_0 \sum_i c_i + \beta_1 \sum_i c_i x_i$$

- $\sum_i c_i = \sum_i (x_i - \bar{x}) = 0,\quad \sum_i c_i x_i = \frac{\sum_i (x_i - \bar{x}) x_i}{SXX} = 1$

$$\mathrm{E}(\hat{\beta}_1 | \mathbb{X}) = \beta_1$$

- Since $E(\bar{y} | \mathbb{X}) = \beta_0 + \beta_1 \bar{x}$, we have

$$E(\hat{\beta}_0 | \mathbb{X}) = E(\bar{y} | \mathbb{X}) - \beta_1 \bar{x} = \beta_0$$

# Variances of Least Square Estimates

- variances of the estimates (do we want small or big?)
- $$\mathrm{Var}(\hat{\beta}_1|\mathbb{X}) = \frac{\sigma^2}{SXX}$$
- $$\mathrm{Var}(\hat{\beta}_0|\mathbb{X}) = \sigma^2(\frac{1}{n} + \frac{\bar{x}^2}{SXX})$$
- $$\mathrm{Cov}(\hat{\beta}_0, \hat{\beta}_1|\mathbb{X}) = -\sigma^2\frac{\bar{x}}{SXX}$$
- $$\rho(\hat{\beta}_0, \hat{\beta}_1|\mathbb{X}) = \frac{-\bar{x}}{\sqrt{SXX/n + \bar{x}^2}} = \frac{-\bar{x}}{\sqrt{(n-1)SD_x^2/n + \bar{x}^2}}$$

# Variances of Least Square Estimates (cont...)

- In all previous expressions, $\sigma^2$ are unknown
- to estimate $\mathrm{Var}(\hat{\beta}_0)$ and $\mathrm{Var}(\hat{\beta}_1)$, replace $\sigma^2$ by $\hat{\sigma}^2$

$$\widehat{\mathrm{Var}}(\hat{\beta}_1|\mathbb{X}) = \hat{\sigma}^2 \frac{1}{SXX}$$

$$\widehat{\mathrm{Var}}(\hat{\beta}_0|\mathbb{X}) = \hat{\sigma}^2 (\frac{1}{n} + \frac{\bar{x}^2}{SXX})$$

- the square root of an estimated variance is called a standard error $(se)$:

$$se(\hat{\beta}_1|\mathbb{X}) = \sqrt{\widehat{\mathrm{Var}}(\hat{\beta}_1)}, \quad se(\hat{\beta}_0|\mathbb{X}) = \sqrt{\widehat{\mathrm{Var}}(\hat{\beta}_0)}$$

# Deriving Variances of LS Estimates

- recall $y_i$'s are assumed independent given $x_i$'s

$$
\begin{aligned}
\mathrm{Var}(\hat{\beta}_1|\mathbb{X}) &= \mathrm{Var}(\sum_i c_i y_i|\mathbb{X}) = \sum_i c_i^2 \mathrm{Var}(y_i|X = x_i) \\
&= \sigma^2 \sum_i c_i^2 = \sigma^2 \sum_i (x_i - \bar{x})^2 / SXX^2 \\
&= \sigma^2 / SXX
\end{aligned}
$$

$$
\begin{aligned}
\mathrm{Var}(\hat{\beta}_0|\mathbb{X}) &= \mathrm{Var}(\bar{y} - \hat{\beta}_1 \bar{x}|\mathbb{X}) \\
&= \mathrm{Var}(\bar{y}|\mathbb{X}) + \bar{x}^2 \mathrm{Var}(\hat{\beta}_1|\mathbb{X}) - 2\bar{x}\mathrm{Cov}(\bar{y}, \hat{\beta}_1|\mathbb{X})
\end{aligned}
$$

# Deriving Variances of LS Estimates (cont...)

- $\mathrm{Var}(\hat{\beta}_0|\mathbb{X}) = \mathrm{Var}(\bar{y}|\mathbb{X}) + \bar{x}^2\mathrm{Var}(\hat{\beta}_1|\mathbb{X}) - 2\bar{x}\mathrm{Cov}(\bar{y}, \hat{\beta}_1|\mathbb{X})$

$$\begin{aligned}
\mathrm{Cov}(\bar{y}, \hat{\beta}_1|\mathbb{X}) &= \mathrm{Cov}(\frac{1}{n}\sum_i y_i, \sum_i c_i y_i|\mathbb{X}) \\
&= \frac{1}{n}\sum_i c_i \mathrm{Cov}(y_i, y_i|\mathbb{X}) \\
&= \frac{\sigma^2}{n}\sum_i c_i = \frac{\sigma^2}{n}\sum_i (x_i - \bar{x}) = 0
\end{aligned}$$

- have calculated $\mathrm{Var}(\hat{\beta}_1|\mathbb{X}) = \frac{\sigma^2}{SXX}$, what is $\mathrm{Var}(\bar{y}|\mathbb{X})$?
- $\mathrm{Var}(\hat{\beta}_0) = \frac{\sigma^2}{n} + \bar{x}^2\frac{\sigma^2}{SXX} = \sigma^2(\frac{1}{n} + \frac{\bar{x}^2}{SXX})$

# Deriving Covariance of LS Estimates

- now for covariance between $\hat{\beta}_0$ and $\hat{\beta}_1$

$$\begin{aligned}
\mathrm{Cov}(\hat{\beta}_0, \hat{\beta}_1 | \mathbb{X}) &= \mathrm{Cov}(\bar{y} - \hat{\beta}_1 \bar{x}, \hat{\beta}_1 | \mathbb{X}) \\
&= \mathrm{Cov}(\bar{y}, \hat{\beta}_1 | \mathbb{X}) - \bar{x}\mathrm{Cov}(\hat{\beta}_1, \hat{\beta}_1 | \mathbb{X}) \\
&= 0 - \sigma^2 \frac{\bar{x}}{SXX} \\
&= -\sigma^2 \frac{\bar{x}}{SXX}
\end{aligned}$$

- easy to get $\rho(\hat{\beta}_0, \hat{\beta}_1 | \mathbb{X}) = \frac{\mathrm{Cov}(\hat{\beta}_0, \hat{\beta}_1 | \mathbb{X})}{\sqrt{\mathrm{Var}(\hat{\beta}_0 | \mathbb{X}) \mathrm{Var}(\beta_1 | \mathbb{X})}}$