Assignment 1

Rin Meng
Student ID: 51940633

September 25, 2024

1. (a) The linear regression model can be written as,

$$P = \beta_0 + \beta_1 C + \epsilon$$

where,

- P is the prime interest rate
- C is the core inflation rate
- β_0 is the intercept of the regression line
- β_1 is the slope of the regression line
- ϵ is the error term

Assumptions of the linear regression model are,

i. **Linearity:** The relationship between P and C is linear
ii. **Independence:** The error term ϵ is independent of C
iii. **Normality:** The error term ϵ is normally distributed
iv. **Homoscedasticity:** The error term ϵ has a constant variance across all levels of C

2. **Remark:** from ch2A.pdf slide 9 and 13,

- $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$
- $\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}}$
- $S_{xy} = \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})$
- $S_{xx} = \sum_{i=1}^{n} (x_i - \bar{x})^2$

Let e_i be the ith residual term. Then for each observation, we can see that the residual for each observation i is defined as:

$$e_i = y_i - \hat{y}_i$$

then we can say that the predicted value $\hat{y}_i$ is equivalent to

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$$

(a) $\sum_{i=1}^{n} e_i = 0$

Proof. LHS:

$$\begin{aligned}
\sum_{i=1}^{n} e_i &= \sum_{i=1}^{n} (y_i - \hat{y}_i) \\
&= \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) \\
&= \sum_{i=1}^{n} (y_i - \bar{y} + \hat{\beta}_1 \bar{x} - \hat{\beta}_1 x_i) \\
&= \sum_{i=1}^{n} y_i - n\bar{y} + \hat{\beta}_1 n\bar{x} - \hat{\beta}_1 \sum_{i=1}^{n} x_i \\
&= n\bar{y} - n\bar{y} + \hat{\beta}_1 n\bar{x} - \hat{\beta}_1 n\bar{x} \\
&= 0
\end{aligned}$$

$$LHS = 0 = RHS$$

$\therefore$ It is true that the sum of residuals e_i is zero. $\square$

(b) $\sum_{i=1}^{n} x_i e_i = 0$

Proof. LHS:

$$\begin{aligned}
\sum_{i=1}^{n} x_i e_i &= \sum_{i=1}^{n} x_i(y_i - \hat{y}_i) \\
&= \sum_{i=1}^{n} x_i(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) \\
&= \sum_{i=1}^{n}(x_i y_i - x_i(\bar{y} - \hat{\beta}_1\bar{x}) - \hat{\beta}_1 x_i^2) \\
&= \sum_{i=1}^{n}(x_i y_i - x_i\bar{y} + x_i\hat{\beta}_1\bar{x} - \hat{\beta}_1 x_i^2) \\
&= \sum_{i=1}^{n}(x_i(y_i - \bar{y}) + \hat{\beta}_1(\bar{x}x_i - x_i^2)) \\
&= \sum_{i=1}^{n} x_i y_i - \bar{y}\sum_{i=1}^{n} x_i - \hat{\beta}_1\left(\sum_{i=1}^{n} x_i^2 - \bar{x}\sum_{i=1}^{n} x_i\right) \\
&= \left(\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}\right) - \hat{\beta}_1\left(\sum_{i=1}^{n} x_i^2 - n\bar{x}^2\right) \\
&= \left(\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y} - n\bar{x}\bar{y} + n\bar{x}\bar{y}\right) - \hat{\beta}_1\left(\sum_{i=1}^{n} x_i^2 - n\bar{x}\bar{x} - n\bar{x}\bar{x} + n\bar{x}\bar{x}\right) \\
&= \left(\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y} - n\bar{x}\bar{y} + n\bar{x}\bar{y}\right) - \hat{\beta}_1\left(\sum_{i=1}^{n} x_i^2 - 2n\bar{x}\bar{x} + n\bar{x}\bar{x}\right) \\
&= \left(\sum_{i=1}^{n} x_i y_i - \bar{y}\sum_{i=1}^{n} x_i - \bar{x}\sum_{i=1}^{n} y_i + n\bar{x}\bar{y}\right) - \hat{\beta}_1\left(\sum_{i=1}^{n} x_i^2 - 2\bar{x}\sum_{i=1}^{n} x_i + n\bar{x}^2\right) \\
&= \sum_{i=1}^{n}\left(x_i y_i - \bar{y}x_i - \bar{x}y_i + \bar{x}\bar{y}\right) - \hat{\beta}_1\sum_{i=1}^{n}\left(x_i^2 - 2\bar{x}x_i + \bar{x}^2\right) \\
&= \sum_{i=1}^{n}\left(x_i(y_i - \bar{y}) + \bar{x}(y_i - \bar{y})\right) - \hat{\beta}_1\sum_{i=1}^{n}(x_i - \bar{x})^2 \\
&= \sum_{i=1}^{n}(x_i + \bar{x})(y_i - \bar{y}) - \hat{\beta}_1\sum_{i=1}^{n}(x_i - \bar{x})^2 \\
&= S_{xy} - \hat{\beta}_1 S_{xx} \\
&= S_{xy} - \frac{S_{xy}}{S_{xx}}S_{xx} \\
&= S_{xy} - S_{xy} \\
&= 0
\end{aligned}$$

$$LHS = 0 = RHS$$

□

$\therefore$ It is true that the independent variables x_i is completly uncorrelated to the residuals e_i.

(c) $\sum_{i=1}^{n} \hat{y}_i e_i = 0$

Proof. LHS:

$$\begin{aligned}
\sum_{i=1}^{n} \hat{y}_i e_i &= \sum_{i=1}^{n} \hat{y}_i (y_i - \hat{y}_i) \\
&= \sum_{i=1}^{n} (y_i \hat{y}_i - \hat{y}_i{}^2) \\
&= \sum_{i=1}^{n} (y_i(\hat{\beta}_0 + \hat{\beta}_1 x_i) - (\hat{\beta}_0 + \hat{\beta}_1 x_i)^2) \\
&= \sum_{i=1}^{n} (y_i(\bar{y} - \hat{\beta}_1 \bar{x} + \hat{\beta}_1 x_i) - (\bar{y} - \hat{\beta}_1 \bar{x} + \hat{\beta}_1 x_i)^2) \\
&= \sum_{i=1}^{n} (y_i \bar{y} - y_i \hat{\beta}_1 \bar{x} + y_i \hat{\beta}_1 x_i \\
&- (\bar{y}^2 - 2\hat{\beta}_1 \bar{y}\bar{x} + \hat{\beta}_1{}^2 \bar{x}^2 + 2\hat{\beta}_1 \bar{y} x_i - 2\hat{\beta}_1{}^2 \bar{x} x_i + \hat{\beta}_1{}^2 x_i^2)) \\
&= (\sum_{i=1}^{n} y_i \bar{y} - \sum_{i=1}^{n} y_i \hat{\beta}_1 \bar{x} + \sum_{i=1}^{n} y_i \hat{\beta}_1 x_i \\
&- \sum_{i=1}^{n} (\bar{y}^2 - 2\hat{\beta}_1 \bar{y}\bar{x} + \hat{\beta}_1{}^2 \bar{x}^2 + 2\hat{\beta}_1 \bar{y} x_i - 2\hat{\beta}_1{}^2 \bar{x} x_i + \hat{\beta}_1{}^2 x_i^2)) \\
&= (n\bar{y}^2 - \hat{\beta}_1 \bar{y}\bar{x} + \hat{\beta}_1 \bar{y} \hat{\beta}_1 x_i \\
&- (\sum_{i=1}^{n} \bar{y}^2 - \sum_{i=1}^{n} 2\hat{\beta}_1 \bar{y}\bar{x} + \sum_{i=1}^{n} \hat{\beta}_1{}^2 \bar{x}^2 + \sum_{i=1}^{n} 2\hat{\beta}_1 \bar{y} x_i - \sum_{i=1}^{n} 2\hat{\beta}_1{}^2 \bar{x} x_i + \sum_{i=1}^{n} \hat{\beta}_1{}^2 x_i^2)) \\
&= (n\bar{y}^2 - (n\bar{y}^2 - 2\hat{\beta}_1 n\bar{y}\bar{x} + \hat{\beta}_1{}^2 n\bar{x}^2 + 2\hat{\beta}_1 n\bar{y}\bar{x} - 2\hat{\beta}_1{}^2 n\bar{x}\bar{x} + \hat{\beta}_1{}^2 \sum_{i=1}^{n} x_i^2))
\end{aligned}$$

proof continued.

$$\sum_{i=1}^{n} \hat{y}_i e_i = (n\bar{y}^2 - (n\bar{y}^2 - \hat{\beta_1}^2 n\bar{x}^2 + \hat{\beta_1}^2 n\bar{x}^2))$$
$$= (n\bar{y}^2 - n\bar{y}^2)$$
$$= 0$$

$$LHS = 0 = RHS$$

□

$\therefore$ It is true that the predicted values $\hat{y}_i$ is completely orthogonal to the residuals e_i.

3. (a) It is given that:
   - $\hat{\beta}_1 = \frac{S_{xx}}{S_{xy}}$
   - $S_{xy} = \sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$
   - $S_{xx} = \sum_{i=1}^{n}(x_i - \bar{x})^2$

   We want to show that

$$\hat{\beta}_1 \sim N(\beta_1, \frac{\sigma^2}{S_{xx}})$$

Where β_1 is the mean of the distribution and $\frac{\sigma^2}{S_{xx}}$ is the variance.

Proof. LHS:

$$\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{S_{xx}}$$
$$= \frac{\sum_{i=1}^{n}(x_i - \bar{x})\epsilon_i}{S_{xx}}$$

we know that

$$\epsilon_i \sim N(0, \sigma^2) \Rightarrow Var(\epsilon_i) = \sigma^2$$

and that

$$Var(\beta_1) = \frac{\sigma^2}{S_{xx}}$$

so then it must also be true that,

$$\begin{aligned} Var(\hat{\beta}_1) &= Var\left(\frac{\sum_{i=1}^n (x_i - \bar{x})\epsilon_i}{S_{xx}}\right) \\ &= \frac{\sum_{i=1}^n (x_i - \bar{x})^2 Var(\epsilon_i)}{Var(S_{xx})} \\ &= \frac{\sum_{i=1}^n (x_i - \bar{x})^2 \sigma^2}{S_{xx}^2} = \frac{S_{xx}\sigma^2}{S_{xx}^2} \\ &= \frac{\sigma^2}{S_{xx}} \end{aligned}$$

□

and we know that, $E[\hat{\beta}_1] = \beta_1$ so then we zz can conclude that

- Variance: $Var(\hat{\beta}_1) = \frac{\sigma^2}{S_{xx}}$
- Mean: $E[\hat{\beta}_1] = \beta_1$

thus,

$$\hat{\beta}_1 \sim N(\beta_1, \frac{\sigma^2}{S_{xx}})$$

$\therefore$ $\hat{\beta}_1$ is normally distributed with mean β_1 and variance $\frac{\sigma^2}{S_{xx}}$.

4. Since $SSE = \sum_{i=1}^n (y_i - \hat{y}_i)^2$, we want to show that: $E[SSE] = \frac{\sigma^2}{n-2}$.

Proof. It is given that,

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

we recall that $y_i = \beta_0 + \beta_1 x_i + \epsilon_i$, $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$ nd $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$ so now we have,

$$
\begin{aligned}
\sum_{i=1}^{n}(y_i - \hat{y})^2 &= \sum_{i=1}^{n}((\beta_0 + \beta_1 x_i + \epsilon_i) - (\hat{\beta}_0 + \hat{\beta}_1 x_i))^2 \\
&= \sum_{i=1}^{n}(\beta_0 + \beta_1 x_i + \epsilon_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2 \\
&= SSE = \sum_{i=1}^{n} \left[(\beta_0 - \hat{\beta}_0) + (\beta_1 - \hat{\beta}_1)x_i + \epsilon_i\right]^2 \\
&= SSE = \sum_{i=1}^{n}[(\beta_0 - \hat{\beta}_0)^2 + (\beta_1 - \hat{\beta}_1)^2 x_i^2 + \epsilon_i^2 \\
&+ 2(\beta_0 - \hat{\beta}_0)(\beta_1 - \hat{\beta}_1)x_i + 2(\beta_0 - \hat{\beta}_0)\epsilon_i + 2(\beta_1 - \hat{\beta}_1)x_i\epsilon_i] \\
E[SSE] &= nE[\epsilon_i^2] + nVar(\hat{\beta}_0) + nVar(\hat{\beta}_1)E[x_i^2] + n(\beta_0 - \hat{\beta}_0)^2 + n(\beta_1 - \hat{\beta}_1)^2 E[x_i^2]
\end{aligned}
$$

$\square$