Regression Model

Guanxu Gleason WANG

University of Glasgow

版本：Past Paper of STAT5025 Regression Model, UofG

更新：December 17, 2024

目录

1 简单线性回归

1.1 最小二乘法

最小二乘法 (Ordinary Least Squares, OLS) 是一种用于估计回归模型参数的统计方法. 其核心思想是通过最小化实际观测值与模型预测值之间的残差平方和 (Residual Sum of Squares, RSS), 找到最优的回归系数, 从而使模型对数据的拟合达到最佳.

定理 1.1 (OLS (1)) 考虑简单线性回归模型:

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \quad i = 1, 2, \cdots, n,$$

其中,

- y_i – 第 i 个观测值的因变量 (被解释变量/响应变量, response variable);
- x_i – 第 i 个观测值的自变量 (解释变量, explanatory variable);
- β_0, β_1 – 待估计的回归系数;
- ε_i – 随机误差项, 假设 ε_i 独立同分布, 且 $\mathbb{E}[\varepsilon_i] = 0$, $\mathrm{Var}[\varepsilon_i] = \sigma^2$.

因此, 斜率 (slope) 估计值

$$\beta_1 = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^{n}(x_i - \overline{x})^2},$$

截距 (intercept) 估计值

$$\beta_0 = \overline{y} - \beta_1 \overline{x}.$$

证明. 最小二乘法原理: $\beta = \arg\min_{\beta} \mathrm{SSE}(\beta)$, 步骤如下:

- Objective function:

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

- The partial derivative:

$$\frac{\partial S}{\partial \beta_0} = -2 \sum_{i=1}^{n} [y_i - (\beta_0 + \beta_1 x_i)] = 0$$

$$\Rightarrow \sum_{i=1}^{n} y_i - n\beta_0 - \beta_1 \sum_{i=1}^{n} x_i = 0 \tag{1}$$

$$\frac{\partial S}{\partial \beta_1} = -2 \sum_{i=1}^{n} [y_i - (\beta_0 + \beta_1 x_i)] x_i = 0.$$

$$\Rightarrow \sum_{i=1}^{n} x_i y_i - \beta_0 \sum_{i=1}^{n} x_i - \beta_1 \sum_{i=1}^{n} x_i^2 = 0 \tag{2}$$

- Solving a system of equations:

  From equation (1),

$$\beta_0 = \frac{1}{n} \left(\sum_{i=1}^{n} y_i - \beta_1 \sum_{i=1}^{n} x_i \right) = \overline{y} - \beta_1 \overline{x} \tag{*}$$

Substituting into equation (2),

$$\begin{aligned}
&\sum_{i=1}^{n} x_i y_i - (\overline{y} - \beta_1 \overline{x}) \sum_{i=1}^{n} x_i - \beta_1 \sum_{i=1}^{n} x_i^2 = 0 \\
\Rightarrow &\sum_{i=1}^{n} x_i y_i - (\overline{y} - \beta_1 \overline{x}) \cdot n\overline{x} - \beta_1 \sum_{i=1}^{n} x_i^2 = 0 \\
\Rightarrow &\sum_{i=1}^{n} x_i y_i - n\overline{x}\,\overline{y} + \beta_1 \cdot n\overline{x}^2 - \beta_1 \sum_{i=1}^{n} x_i^2 = 0 \\
\Rightarrow &\beta_1 \left(\sum_{i=1}^{n} x_i^2 - n\overline{x}^2 \right) = \sum_{i=1}^{n} x_i y_i - n\overline{x}\,\overline{y} \\
\Rightarrow &\beta_1 = \frac{\sum_{i=1}^{n} x_i y_i - n\overline{x}\,\overline{y}}{\sum_{i=1}^{n} x_i^2 - n\overline{x}^2} = \frac{\sum_{i=1}^{n} (x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^{n} (x_i - \overline{x})^2}
\end{aligned} \tag{*}$$

that is

$$\beta_1 = \frac{\mathrm{Cov}\,[x, y]}{\mathrm{Var}\,[x]}.$$

1.2 相关系数

定义 1.1 (皮尔逊相关系数 (Pearson correlation coefficient)) 衡量自变量 x 与因变量 y 之间线性关系的强度和方向, 公式为

$$r = \frac{\operatorname{Cov}[x,y]}{\sigma_x \sigma_y} = \frac{\operatorname{Cov}[x,y]}{\sqrt{\operatorname{Var}[x]}\sqrt{\operatorname{Var}[y]}} = \frac{\sum_{i=1}^{n}(x_i-\overline{x})(y_i-\overline{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\overline{x})^2}\sqrt{\sum_{i=1}^{n}(y_i-\overline{y})^2}} \in [-1,1].$$

定义 1.2 (决定系数 (Coefficient of determination)) 表示自变量 x 与因变量 y 的解释程度, 公式为

$$R^2 = r^2.$$

1.3 假设检验

例 1.1 (回归系数的检验) 以双边检验为例, 即假设

$$H_0 : \beta = b \quad \text{vs.} \quad H_1 : \beta \neq b.$$

则构造检验统计量 (test statistics)

$$T = \left|\frac{\widehat{\beta}-b}{\sigma_{\widehat{\beta}}}\right| = \left|\frac{\widehat{\beta}-b}{\sqrt{\operatorname{Var}\left[\widehat{\beta}\right]}}\right| \sim t(n-k).$$

假设显著性水平 (significance level) 为 α, 那么

$$T < t_{1-\frac{\alpha}{2}}(n-k) \quad \Rightarrow \quad \text{Do not have evidence to reject } H_0.$$

$$T > t_{1-\frac{\alpha}{2}}(n-k) \quad \Rightarrow \quad \text{Reject } H_0 \text{ in favour of } H_1.$$

此外, β 的 $(1-\alpha)$ 置信区间 (confidence interval) 为

$$\left(\widehat{\beta} \pm t_{1-\frac{\alpha}{2}}(n-k) \times \sigma_{\widehat{\beta}}\right)$$

2 多元线性回归

2.1 最小二乘法

定理 2.1 (OLS (2)) 考虑多元线性回归模型:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik} + \varepsilon_i, \quad i = 1, 2, \cdots, n$$

矩阵形式表示为:

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon},$$

其中:

- $\boldsymbol{Y}$ – $n \times 1$ 的因变量 (response) 向量;

- $\boldsymbol{X}$ – $n \times (k+1)$ 的设计矩阵 (design matrix), 第一列全为 1 (对应截距项);
- $\boldsymbol{\beta}$ – $(k+1) \times 1$ 的回归系数向量;
- $\boldsymbol{\varepsilon}$ – $n \times 1$ 的误差项向量.

$$\boldsymbol{Y} = \begin{pmatrix} y_0 \\ y_1 \\ y_2 \\ \vdots \\ y_k \end{pmatrix}, \boldsymbol{X} = \begin{pmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1k} \\ 1 & x_{21} & x_{22} & \cdots & x_{2k} \\ 1 & x_{31} & x_{32} & \cdots & x_{3k} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{nk} \end{pmatrix}_{n\times(k+1)}, \boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_k \end{pmatrix}, \boldsymbol{\varepsilon} = \begin{pmatrix} \varepsilon_0 \\ \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_k \end{pmatrix}$$

因此, 回归系数估计值

$$\boldsymbol{\beta} = (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{Y}.$$

证明. 最小二乘法原理: $\beta = \arg\min_{\beta} \mathrm{SSE}(\beta)$, 步骤如下:

- Objective function:

$$\begin{aligned} S(\boldsymbol{\beta}) &= \sum_{i=1}^{n} (y_i - \mathbb{E}[y_i])^2 \\ &= (\boldsymbol{Y} - \mathbb{E}[\boldsymbol{Y}])^\top (\boldsymbol{Y} - \mathbb{E}[\boldsymbol{Y}]) \\ &= (\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta})^\top (\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}). \end{aligned}$$

- The partial derivative:

$$\frac{\partial S}{\partial \boldsymbol{\beta}} = -2\boldsymbol{X}^\top(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}) = 0.$$
$$\Rightarrow \ \boldsymbol{X}^\top\boldsymbol{X}\boldsymbol{\beta} = \boldsymbol{X}^\top\boldsymbol{Y}$$

- Solving the equation:

$$\boldsymbol{\beta} = (\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top\boldsymbol{Y} = \frac{\sum_{i=1}^{n} x_i y_i}{\sum_{i=1}^{n} x_i^2}$$

2.2 相关系数 (Correlation coefficient)

定义 2.1 (相关系数)

$$r_j = \frac{\mathrm{Cov}\left[\boldsymbol{X}_j, \boldsymbol{Y}\right]}{\sigma_{\boldsymbol{X}_j}\sigma_{\boldsymbol{Y}}} = \frac{\mathrm{Cov}\left[\boldsymbol{X}_j, \boldsymbol{Y}\right]}{\sqrt{\mathrm{Var}\left[\boldsymbol{X}_j\right]}\sqrt{\mathrm{Var}\left[\boldsymbol{Y}\right]}} = \frac{\sum_{i=1}^{n}(x_{ij} - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_{i=1}^{n}(x_{ij} - \overline{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \overline{y})^2}} \in [-1, 1].$$

答题模板: There is a strong/weak positive/negative correlation between [the explanatory variable] and [the response variable].

定义 2.2 (决定系数)

$$R^2 = \frac{\mathrm{ESS}}{\mathrm{TSS}} = \frac{\sum_{i=1}^{n} \left(\widehat{y}_i - \overline{y}\right)^2}{\sum_{i=1}^{n} \left(y_i - \overline{y}\right)^2} = \frac{\left(\widehat{\boldsymbol{Y}}_i - \overline{\boldsymbol{Y}}\right)^\top \left(\widehat{\boldsymbol{Y}}_i - \overline{\boldsymbol{Y}}\right)}{\left(\boldsymbol{Y}_i - \overline{\boldsymbol{Y}}\right)^\top \left(\boldsymbol{Y}_i - \overline{\boldsymbol{Y}}\right)},$$

或

$$R^2 = 1 - \frac{\mathrm{RSS}}{\mathrm{TSS}} = 1 - \frac{\sum_{i=1}^{n} \left(y_i - \widehat{y}\right)^2}{\sum_{i=1}^{n} \left(y_i - \overline{y}\right)^2} = 1 - \frac{\left(\boldsymbol{Y}_i - \widehat{\boldsymbol{Y}}\right)^\top \left(\boldsymbol{Y}_i - \widehat{\boldsymbol{Y}}\right)}{\left(\boldsymbol{Y}_i - \overline{\boldsymbol{Y}}\right)^\top \left(\boldsymbol{Y}_i - \overline{\boldsymbol{Y}}\right)},$$

其中,

- 回归平方和 (解释平方和, Explained Sum of Squares): $\mathrm{ESS} = \sum_{i=1}^{n} \left(\widehat{y}_i - \overline{y}\right)^2 = \left(\widehat{\boldsymbol{Y}}_i - \overline{\boldsymbol{Y}}\right)^\top \left(\widehat{\boldsymbol{Y}}_i - \overline{\boldsymbol{Y}}\right)$;
- 残差平方和 (Residual Sum of Squares): $\mathrm{RSS} = \sum_{i=1}^{n} \left(y_i - \widehat{y}\right)^2 = \left(\boldsymbol{Y}_i - \widehat{\boldsymbol{Y}}\right)^\top \left(\boldsymbol{Y}_i - \widehat{\boldsymbol{Y}}\right)$;
- 总离差平方和 (Total Sum of Squares): $\mathrm{TSS} = \sum_{i=1}^{n} \left(y_i - \overline{y}\right)^2 = \left(\boldsymbol{Y}_i - \overline{\boldsymbol{Y}}\right)^\top \left(\boldsymbol{Y}_i - \overline{\boldsymbol{Y}}\right)$.

注 同样的, 在多元线性回归中仍有

$$r^2 = R^2 \quad \Rightarrow \quad r = \pm\sqrt{R^2}.$$

答题模板: Therefore [the value of R^2] of the variation in [the response variable] can be explained by [the explanatory variable].

2.3 假设检验

例 2.1 (回归系数的检验) 以双边检验为例, 即假设

$$H_0 : \beta = b \quad \text{vs.} \quad H_1 : \beta \neq b.$$

则构造检验统计量 (test statistics)

$$T = \left| \frac{\widehat{\beta} - b}{\sigma_{\widehat{\beta}}} \right| = \left| \frac{\widehat{\beta} - b}{\sqrt{\text{Var}\left[\widehat{\beta}\right]}} \right| \sim t(n - (k+1)).$$

假设显著性水平 (significance level) 为 α, 那么

$$T < t_{1-\frac{\alpha}{2}}(n - (k+1)) \quad \Rightarrow \quad \text{We do not have evidence to reject the null hypothesis } (H_0).$$

$$T > t_{1-\frac{\alpha}{2}}(n - (k+1)) \quad \Rightarrow \quad \text{We can reject the null hypothesis } (H_0).$$

例 2.2 (置信区间) 此外, β 的 $(1-\alpha)$ 置信区间 (confidence interval) 为

$$\left(\widehat{\beta} \pm t_{1-\frac{\alpha}{2}}(n - (k+1)) \times \sigma_{\widehat{\beta}}\right)$$

2.4 预测及置信区间

例 2.3 (已知模型的预测) 假设已知预测数据 $\boldsymbol{b}=\begin{pmatrix}1 & b_1 & b_2 & \cdots & b_k\end{pmatrix}^\top$，那么预测值为

$$\boldsymbol{b}^\top\boldsymbol{\beta},$$

即模型给出的点预测值.

预测区间考虑了模型的误差，因此需要计算预测值的不确定性

$$\sqrt{\frac{\mathrm{RSS}}{n-(k+1)}\left(1+\boldsymbol{b}^\top\left(\boldsymbol{X}^\top\boldsymbol{X}\right)^{-1}\boldsymbol{b}\right)},$$

其中 RSS (Residual Sum of Squares, 残差平方和)

$$\mathrm{RSS}=\left(\boldsymbol{Y}_i-\widehat{\boldsymbol{Y}}\right)^\top\left(\boldsymbol{Y}_i-\widehat{\boldsymbol{Y}}\right).$$

因此得到显著性水平 α 的置信区间为

$$\left(\boldsymbol{b}^\top\boldsymbol{\beta}\pm t_{1-\frac{\alpha}{2}}(n-(k+1))\sqrt{\frac{\mathrm{RSS}}{n-(k+1)}\left(1+\boldsymbol{b}^\top\left(\boldsymbol{X}^\top\boldsymbol{X}\right)^{-1}\boldsymbol{b}\right)}\right).$$

其中,

- 总体误差

$$\sigma^2=\frac{\mathrm{RSS}}{n-(k+1)},$$

衡量回归模型中误差的整体水平(即方差估计);

- 线性组合的方差

$$\operatorname{Var}\left[\boldsymbol{b}^{\top}\boldsymbol{\beta}\right]=\sigma^{2}\cdot\boldsymbol{b}^{\top}\left(\boldsymbol{X}^{\top}\boldsymbol{X}\right)^{-1}\boldsymbol{b},$$

其中 $\varepsilon\sim N(0,\sigma^{2})$.

证明. 真实值与预测值的误差为:

$$\boldsymbol{Y}-\widehat{\boldsymbol{Y}}=\boldsymbol{b}^{\top}\left(\boldsymbol{\beta}-\widehat{\boldsymbol{\beta}}\right)+\boldsymbol{\varepsilon}.$$

因为 $\widehat{\boldsymbol{\beta}}\sim N\left(\boldsymbol{\beta},\sigma^{2}\left(\boldsymbol{X}^{\top}\boldsymbol{X}\right)^{-1}\right)$, 所以 $\boldsymbol{b}^{\top}\left(\boldsymbol{\beta}-\widehat{\boldsymbol{\beta}}\right)$ 的方差为:

$$\operatorname{Var}\left[b^{\top}\widehat{\boldsymbol{\beta}}\right]=\sigma^{2}\cdot\boldsymbol{b}^{\top}\left(\boldsymbol{X}^{\top}\boldsymbol{X}\right)^{-1}\boldsymbol{b}$$

因此,

$$\operatorname{Var}\left[\boldsymbol{Y}-\widehat{\boldsymbol{Y}}\right]=\sigma^{2}\left(1+\boldsymbol{b}^{\top}\left(\boldsymbol{X}^{\top}\boldsymbol{X}\right)^{-1}\boldsymbol{b}\right)$$

2.5 回归系数差值的置信区间

在多元线性回归中，回归系数的估计为:

$$\widehat{\boldsymbol{\beta}}=(\boldsymbol{X}^{\top}\boldsymbol{X})^{-1}\boldsymbol{X}^{\top}\boldsymbol{y}.$$

回归系数的协方差矩阵为:

$$\mathrm{Cov}(\widehat{\boldsymbol{\beta}}) = \sigma^2(\boldsymbol{X}^\top \boldsymbol{X})^{-1}.$$

假设误差方差 σ^2 用残差平方和估计为 $\widehat{\sigma}^2 = \dfrac{\mathrm{RSS}}{n-(k+1)}$, 则协方差矩阵为

$$\mathrm{Cov}(\widehat{\boldsymbol{\beta}}) = \frac{\mathrm{RSS}}{n-(k+1)}(\boldsymbol{X}^\top \boldsymbol{X})^{-1}.$$

假设 β_1 和 β_2 之间是独立的, 即 $\mathrm{Cov}\left[\beta_1, \beta_2\right] = 0$, 因此,

$$\mathrm{Var}\left[\widehat{\beta}_1 - \widehat{\beta}_2\right] = \mathrm{Var}\left[\widehat{\beta}_1\right] + \mathrm{Var}\left[\widehat{\beta}_2\right] = \frac{\mathrm{RSS}}{n-(k+1)}\left[(\boldsymbol{X}^\top \boldsymbol{X})^{-1}_{11} + (\boldsymbol{X}^\top \boldsymbol{X})^{-1}_{22}\right].$$

如果, $(\boldsymbol{X}^\top \boldsymbol{X})^{-1}$ 的对角线元素用 $S_{x_1x_1}$ 和 $S_{x_2x_2}$ 表示, 则

$$\mathrm{Var}\left[\widehat{\beta}_1 - \widehat{\beta}_2\right] = \frac{\mathrm{RSS}}{n-(k+1)}\left(\frac{1}{S_{x_1x_1}} + \frac{1}{S_{x_2x_2}}\right).$$

因此, 在显著性水平 α 下, 置信区间为:

$$\left(\widehat{\beta}_1 - \widehat{\beta}_2\right) \pm t_{1-\frac{\alpha}{2}}(n-(k+1))\sqrt{\frac{\mathrm{RSS}}{n-(k+1)}\left(\frac{1}{S_{x_1x_1}} + \frac{1}{S_{x_2x_2}}\right)},$$

其中, $\widehat{\beta}_i = \dfrac{S_{x_iy_i}}{S_{x_ix_i}}$.

3 虚拟变量 (Dummy Variable) / 分类变量 (Categorical Variable)

在回归模型中，解释变量通常表示为数值数据, 但若解释变量为分类变量, 假设只有两个类别 $(0,1)$, 那么回归方程为

$$y_i = \beta_0 + \sum_{i=1}^{n} \beta_i x_i + \sum_{i=1}^{m} \gamma_i I_i + \varepsilon_i,$$

其中,

$$I_i = \begin{cases} 1, & \text{if } i \in A, \\ 0, & \text{otherwise.} \end{cases}$$

若改写成矩阵的形式 (vector-matrix form), 即

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon},$$

其中,

$$\boldsymbol{Y}=\begin{pmatrix} y_0 \\ y_1 \\ y_2 \\ \vdots \\ y_n \\ y_{n+1} \\ y_{n+2} \\ \vdots \\ y_{n+m} \end{pmatrix}, \boldsymbol{X}=\begin{pmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1n} & 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & & \vdots & \vdots & \vdots & & \vdots \\ 1 & x_{n_1,1} & x_{n_1,2} & \cdots & x_{n_1,n} & 0 & 0 & \cdots & 0 \\ 1 & x_{n_1+1,1} & x_{n_1+1,2} & \cdots & x_{n_1+1,n} & 1 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & & \vdots & \vdots & \vdots & & \vdots \\ 1 & x_{n_2,1} & x_{n_2,2} & \cdots & x_{n_2,n} & 1 & 0 & \cdots & 0 \\ 1 & x_{n_2+1,1} & x_{n_2+1,2} & \cdots & x_{n_2+1,n} & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & & \vdots & \vdots & \vdots & & \vdots \\ 1 & x_{n_3,1} & x_{n_3,2} & \cdots & x_{n_3,n} & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & & \vdots & \vdots & \vdots & & \vdots \\ 1 & x_{n_2+1,1} & x_{n_2+1,2} & \cdots & x_{n_2+1,n} & 0 & 0 & \cdots & 1 \\ \vdots & \vdots & \vdots & & \vdots & \vdots & \vdots & & \vdots \\ 1 & x_{n_3,1} & x_{n_3,2} & \cdots & x_{n_3,n} & 0 & 0 & \cdots & 1 \end{pmatrix}_{t\times(n+m+1)}, \boldsymbol{\beta}=\begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_n \\ \gamma_1 \\ \gamma_2 \\ \vdots \\ \gamma_m \end{pmatrix}, \boldsymbol{\varepsilon}=\begin{pmatrix} \varepsilon_0 \\ \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \\ \varepsilon_{n+1} \\ \varepsilon_{n+2} \\ \vdots \\ \varepsilon_{n+m} \end{pmatrix}.$$

4 实操: 结合代码提供的输出对模型进行分析

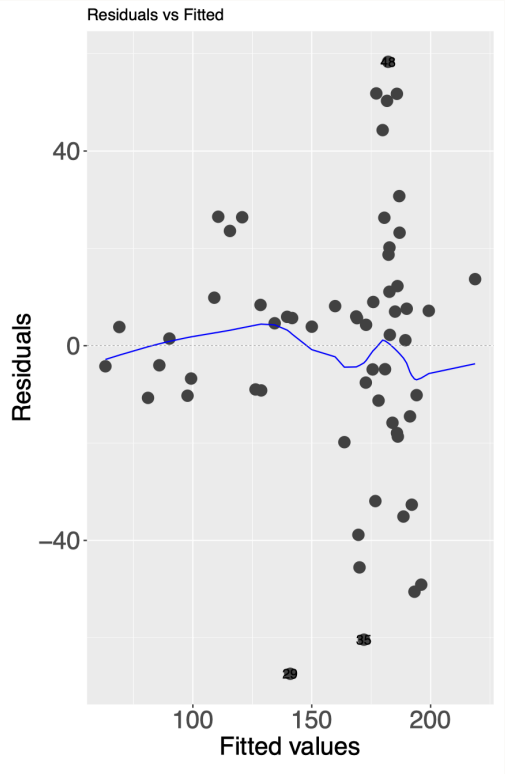
4.1 Model Assumptions

4.1.1 Residuals vs Fitted

目的 & 作用: 检验是否符合模型同方差假设 和零均值假设.

- **Assumption 1** - constant variance of residuals (同方差);
- **Assumption 2** - residual mean of zero (零均值).

例 4.1 (假设 1 不成立, 但假设 2 成立.)



Assumption 1 - constant variance of residuals.

Since the residuals against fitted values show and increase in variability as the value of fitted values increases, then this assumption appears invalid.

由于残差对拟合值的对比显示, 随着拟合值值的增加, 变异性也随之增加, 因此这个假设似乎无效。

Assumption 2 - residual mean of zero.

The residuals against fitted values indicate the the residuals are scattered around the zero line indicating this assumptions appears valid.

残差与拟合值之间的对比表明, 残差围绕着零线散布, 这表明这一假设似乎是有效的.

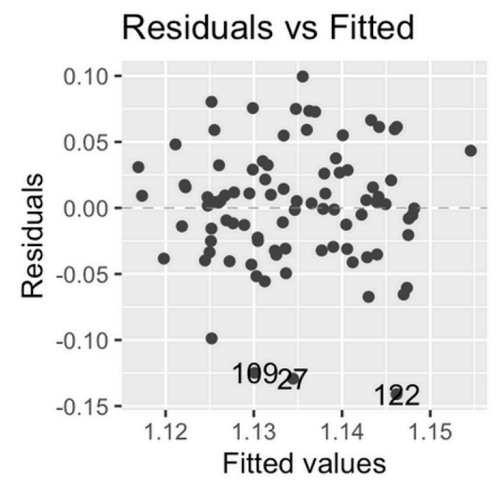
例 4.2 (假设 1 和假设 2 都成立.)

The residual against fitted values plot shows **(Assumption 2) the residuals are randomly scattered around zero** and **(Assumption 1) there are no clear patterns in the residuals**.

残差对拟合值图显示残差在零周围随机分布, 并且残差中没有明显的模式. 这表明我可以合理地假设残差具有恒定的方差和零均值.

This implies that I can reasonably assume that the residuals have constant variance and mean of zero.

Likewise, I may assume that there are no missing underlying patterns (e.g., non-linear patterns that I have missed).



4.1.2 QQ 图

目的 & 作用: 检验是否符合模型正态分布假设.

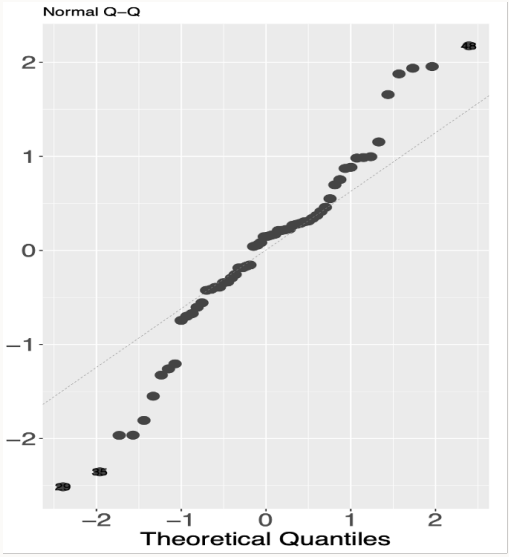
- **Assumption 3** - residuals are normally distributed.

例 4.3 (假设不成立.)

Assumption 3 - residuals are normally distributed.

The Normal QQ plot indicates that this assumption is invalid as the black dots do not appear constant with the diagonal line with deviations in the lower tail (bottom left hand side of plot).

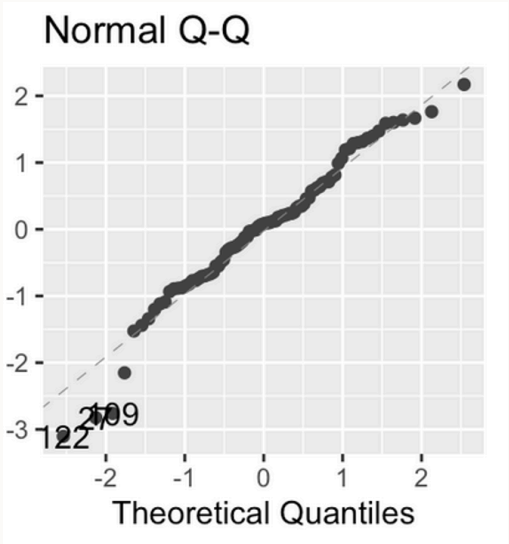
正常 QQ 图表明，这个假设无效，因为黑色点在图表的下半部分（左下角）与对角线线不保持恒定。



例 4.4 (假设成立.)

The QQ plots shows that the standardised residuals are consistent with a normal distribution and therefore I can assume that residuals are normally distributed.

QQ 图显示标准化残差与正态分布一致，因此我可以假设残差服从正态分布。



4.1.3 Assumption 3 - residuals are independent.

4.2 The Variables Included in the Analysis

4.2.1 Scatter Plot

The plot of [y] against [x] illustrated that there is (no) clear linear relationship between the two variables.

[y] 对 [x] 的图表显示，这两个变量之间有（没有）明显的线性关系。

4.2.2 Specific Interpretation of Coefficients (from Output of R)

- 显著相关: $t > t\left(n, 1-\frac{\alpha}{2}\right)$ $\Rightarrow$ Reject H_0: $\beta = 0$ i.e. 有相关性, 该解释变量不可或缺.
  Accounting for the other variables in the model, [the explanatory variable] is significantly related to [the response variable] (t-value [$t > t\left(n, 1-\frac{\alpha}{2}\right)$]) and with each unit increase in [the explanatory variable], [the response variable] is expected to increase/decrease by [the coefficients of the explanatory variable].
  考虑到模型中的其他变量，[解释变量] 与 [响应变量] 显著相关 (t 值 [t 值])，并且每当 [解释变量] 增加 1 个单位，[响应变量] 预计会减少 [解释变量的系数]。

- 不显著相关: $t < t\left(n, 1-\frac{\alpha}{2}\right)$ $\Rightarrow$ Do not reject H_0: $\beta = 0$ i.e. 无相关性, 该解释变量可有可无.
  Accounting for the other variables in the model, [the explanatory variable] is not significantly related to [the response variable] (t-value [$t < t\left(n, 1-\frac{\alpha}{2}\right)$]).
  考虑到模型中的其他变量，[解释变量] 与 [响应变量] 不显著相关 (t 值 [t 值]).

4.3 Remaining R Output (e.g. R^2)

For this multiple linear regression model, the adjusted R^2 value is [adjusted R^2]. This means that the model explains [adjusted R^2] of the variation in [the response variable] which is not very much.

对于这个多元线性回归模型，调整后的 R^2 值是 [调整后的 R^2]。这意味着该模型解释了 [调整后的 R^2] 的 [响应变量] 变化，这并不是很多。

This result is not too surprising given the amount of variability we can see from the scatterplot of [the response variable] and [the response variable].

考虑到从[响应变量]和[响应变量]的散点图中我们可以看到的变异性，这个结果并不太令人惊讶。

4.4 Further Work

In future, we would firstly want to re-fit the model with [??] as a categorical variable and re-check our model assumptions based on the new fitted model.

We would also look closer at the residual independence assumption.

Lastly, I would look closer at how [??] has changed between 2010 and 2020.