2023/02/07

藪友良

予測の統計的性質[1]

推定された回帰モデル($\hat{Y} = \hat{\alpha} + \hat{\beta}X$)の応用例として、$Y$の予測を学習しました(2.4 節参照)。ここでは、予測の統計的性質をみていきましょう。

予測量と予測誤差

説明変数が$X = X_0$のとき、予測対象Y_0の予測を**予測量**といい、$\hat{Y}_0$と表記します。予測量は、$X = X_0$とした回帰直線上の値です。

$$\hat{Y}_0 = \hat{\alpha} + \hat{\beta}X_0$$

データ(Y_1,X_1)、(Y_2,X_2)、…、(Y_n,X_n)を用いて、OLS 推定量$\hat{\alpha}$と$\hat{\beta}$が計算されますが、予測対象Y_0は推定に用いたデータに含まれていないとします[2]。また、予測対象Y_0は、

$$Y_0 = \alpha + \beta X_0 + u_0$$

となり、u_0は誤差項です。ここで、u_0は正規分布$N(0, \sigma^2)$に従い、OLS 推定に用いるデータの誤差項(u_1、u_2、…、u_n)とは独立とします。

予測誤差は、予測量$\hat{Y}_0$と真の値Y_0との差として、次式で表されます。

$$\begin{aligned}\hat{Y}_0 - Y_0 &= \left(\hat{\alpha} + \hat{\beta}X_0\right) - (\alpha + \beta X_0 + u_0) \\ &= \underbrace{(\hat{\alpha} - \alpha) + \left(\hat{\beta} - \beta\right)X_0}_{\text{パラメータの推定誤差}} - \underbrace{u_0}_{Y_0\text{の誤差項}}\end{aligned}$$

右辺第 1 項と第 2 項は、「パラメータの推定誤差($\hat{\alpha} - \alpha,\ \hat{\beta} - \beta$)から生じる部分」を示します。そして、第 3 項は、「予測対象Y_0の誤差項u_0から生じる部分」を示しています。サンプルサイズnが大きくなると、パラメータをより正確に推定でき、パラメータの推定誤差は小さくなります。一方、サンプルサイズnが大きくなっても、予測対象Y_0の誤差項u_0は小さくなりません。

予測誤差の期待値と分散

予測誤差($\hat{Y}_0 - Y_0$)の期待値は、次式で表されます(式展開では、$E[\hat{\alpha}] = \alpha$、$E\left[\hat{\beta}\right] = \beta$、$E[u_0] = 0$を用いました)。

$$E\left[\hat{Y}_0 - Y_0\right] = \underbrace{\left(E[\hat{\alpha}] - \alpha\right)}_{=0} + \underbrace{\left(E\left[\hat{\beta}\right] - \beta\right)}_{=0} X_0 + \underbrace{E[u_0]}_{=0} = 0$$

つまり、予測は平均的に正しく、予測量$\hat{Y}_0$は**不偏予測量**といえます。

次に、予測誤差の分散は、次式で表されます(証明は補足参照)。

---

[1] 藪友良『入門 実践する計量経済学』(2023 年、東洋経済新報社)の補足資料です。

[2] 現在までに利用可能な情報を用いて、将来予測をしたいといった状況を考えます。このとき、データは現在までに利用可能な情報ですが、予測対象は将来の値なのでデータに含まれていません。

$$V(\hat{Y}_0 - Y_0) = \underbrace{\sigma^2}_{Y_0\text{における誤差}} \left(1 + \underbrace{\frac{1}{n} + \frac{(X_0 - \bar{X})^2}{\sum_{i=1}^{n}(X_i - \bar{X})^2}}_{\text{パラメータの推定誤差}} \right)$$

この式から、次の 2 点がわかります。第 1 に、第 3 項$(X_0 - \bar{X})^2 / \sum_{i=1}^{n}(X_i - \bar{X})^2$は、$X_0$が$\bar{X}$に等しいと 0 となり、$X_0$が$\bar{X}$から乖離すると大きくなる点です。つまり、$X_0$が平均付近にあれば予測誤差は小さくなる一方で、$X_0$が大きすぎたり小さすぎたりすると予測誤差は大きくなります。専有面積Xと賃料Yの例でいえば、平均的な面積の部屋ならば予測精度は高くなりますが、部屋が極端に広かったり狭かったりすると予測精度は下がるといえます。

第 2 は、サンプルサイズnが大きくなると、第 2 項と第 3 項が消えて、第 1 項だけが残る点です。すなわち、nが大きい場合には、予測誤差の分散は次のようになります。

$$V(\hat{Y}_0 - Y_0) = \sigma^2$$

これは、nが大きいと、パラメータが正確に推定され、予測誤差は「予測対象Y_0の誤差項u_0から生じる部分」だけになるためです。nが大きい場合には、予測誤差を決めるうえで、Y_0の誤差項u_0が重要になるといえます。予測が目的の場合には、当てはまりの良いモデルを用いること(つまり、誤差項u_0の分散σ^2は小さくなる)が望ましい、といえます。

予測誤差の標準誤差と信頼区間

予測誤差$\hat{Y}_0 - Y_0$の標準偏差は、分散の平方根として、次式で表されます。

$$\sqrt{V(\hat{Y}_0 - Y_0)} = \sigma \sqrt{1 + \frac{1}{n} + \frac{(X_0 - \bar{X})^2}{\sum_{i=1}^{n}(X_i - \bar{X})^2}}$$

予測誤差$\hat{Y}_0 - Y_0$の標準誤差(標準偏差の推定量)は、σを推定量$s = \sqrt{\sum_{i=1}^{n} \hat{u}_i^2 / (n-2)}$で置き換えることによって、次のように計算できます。

$$s \sqrt{1 + \frac{1}{n} + \frac{(X_0 - \bar{X})^2}{\sum_{i=1}^{n}(X_i - \bar{X})^2}}$$

予測誤差$\hat{Y}_0 - Y_0$の 95%信頼区間は、予測誤差の標準誤差を用いて、次のようになります。

$$-2s \sqrt{1 + \frac{1}{n} + \frac{(X_0 - \bar{X})^2}{\sum_{i=1}^{n}(X_i - \bar{X})^2}} < \hat{Y}_0 - Y_0 < 2s \sqrt{1 + \frac{1}{n} + \frac{(X_0 - \bar{X})^2}{\sum_{i=1}^{n}(X_i - \bar{X})^2}}$$

ここでは、1.96ではなく簡便法として2を用いました(補足参照)。上式の両辺に-1を掛けてから全体に$\hat{Y}_0$を足すと、予測対象Y_0の 95%信頼区間を次のように求められます。

$$\underbrace{\hat{Y}_0 - 2s\sqrt{1+\frac{1}{n}+\frac{(X_0-\bar{X})^2}{\sum_{i=1}^{n}(X_i-\bar{X})^2}}}_{\text{信頼区間の下限}} < Y_0 < \underbrace{\hat{Y}_0 + 2s\sqrt{1+\frac{1}{n}+\frac{(X_0-\bar{X})^2}{\sum_{i=1}^{n}(X_i-\bar{X})^2}}}_{\text{信頼区間の上限}}$$

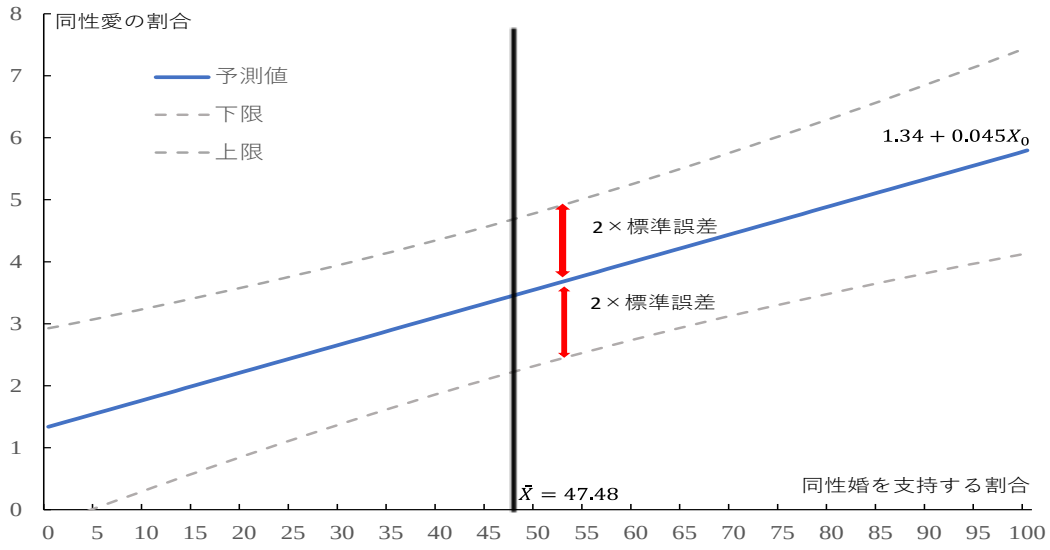
例(同性愛者割合の予測の信頼区間)

2.4 節の例 2-3 で紹介した同性愛者の割合を考えましょう。被説明変数Yは「同性愛を自己申告する割合(%)」、説明変数Xは「同性婚を支持する割合(%)」でした。OLS 推定の結果、$\hat{Y} = 1.34 + 0.045X$となります。データから、$n = 50$、$\bar{X} = 47.48$、$\sum_{i=1}^{n}(X_i - \bar{X})^2 = 3318$、$\sum_{i=1}^{n}\hat{u}_i^2 = 17.8$です。ここで、$s = \sqrt{17.8/48} = 0.61$となることから、予測誤差の標準誤差が得られます。

$$s\sqrt{1+\frac{1}{n}+\frac{(X_0-\bar{X})^2}{\sum_{i=1}^{n}(X_i-\bar{X})^2}} = 0.61\sqrt{1+\frac{1}{50}+\frac{(X_0-47.48)^2}{3318}}$$

図 1 は、予測対象Y_0の 95%信頼区間を示します。同性婚を支持する割合X_0が高くなると予測値$\hat{Y}_0(= 1.34 + 0.045X_0)$も高くなります[3][4]。95%信頼区間を計算すると、$X_0 = 47.48$(標本平均)なら(2.22、4.68)区間となり、$X_0 = 100$なら(4.14、7.15)区間となります。同性婚を支持する割合を 100%(同性愛に対する偏見や差別が小さい社会)としたとき、同性愛者の割合は、下限で評価して 4.14%、上限で評価すると 7.15%となります。

図 1 同性愛者割合の予測値と 95%信頼区間



[3] 予測量Y_0が負の値をとる可能性が許容されているため、同性婚を支持する割合が低いとき、信頼区間の下限は負の値になっています。予測量が負にならないという制約を課したいなら、同性愛の割合を対数変換するなどの工夫が必要でしょう。

[4] 図をみると、信頼区間は直線ではなく、曲線になっていることがわかります。これは信頼区間が$(X_0 - \bar{X})^2$に依存しており、X_0が変化すると、信頼区間が非線形に変化するためです。

補足

予測誤差の分散

予測誤差は、$\hat{Y}_0 - Y_0 = (\hat{\alpha} - \alpha) + (\hat{\beta} - \beta)X_0 + u_0$であるため、分散は次のようになります。

$$V(\hat{Y}_0 - Y_0) = E\left[\left((\hat{\alpha} - \alpha) + (\hat{\beta} - \beta)X_0 + u_0\right)^2\right]$$

$$= \underbrace{E[(\hat{\alpha} - \alpha)^2]}_{=V(\hat{\alpha})} + X_0^2 \underbrace{E\left[(\hat{\beta} - \beta)^2\right]}_{=V(\hat{\beta})} + 2X_0 \underbrace{E\left[(\hat{\alpha} - \alpha)(\hat{\beta} - \beta)\right]}_{=Cov(\hat{\alpha},\hat{\beta})} + E[u_0^2]$$

式展開では、$E[(\hat{\alpha} - \alpha)u_0] = E\left[(\hat{\beta} - \beta)u_0\right] = 0$を用いました(これはOLS推定量$(\hat{\alpha}, \hat{\beta})$の計算に、予測対象$Y_0$は用いられていないため、誤差項$u_0$とOLS推定量は無相関になるためです)。

上式右辺に、$V(\hat{\alpha})$、$V(\hat{\beta})$、$Cov(\hat{\alpha}, \hat{\beta})$を代入すると、

$$\sigma^2\left(\frac{1}{n} + \frac{\bar{X}^2}{\sum_{i=1}^n (X_i - \bar{X})^2}\right) + \frac{\sigma^2 X_0^2}{\sum_{i=1}^n (X_i - \bar{X})^2} - \frac{2\sigma^2 X_0 \bar{X}}{\sum_{i=1}^n (X_i - \bar{X})^2} + \sigma^2$$

となります(3章の練習問題10では$V(\hat{\alpha})$、練習問題11では$Cov(\hat{\alpha}, \hat{\beta})$が紹介されています)。
上式を整理すると、次式となります。

$$\sigma^2\left(1 + \frac{1}{n} + \frac{\bar{X}^2 + X_0^2 - 2X_0\bar{X}}{\sum_{i=1}^n (X_i - \bar{X})^2}\right) = \sigma^2\left(1 + \frac{1}{n} + \frac{(X_0 - \bar{X})^2}{\sum_{i=1}^n (X_i - \bar{X})^2}\right)$$

予測誤差の確率分布

予測誤差は、$\hat{Y}_0 - Y_0 = (\hat{\alpha} - \alpha) + (\hat{\beta} - \beta)X_0 + u_0$である。ここで、$(\hat{\alpha} - \alpha)$、$(\hat{\beta} - \beta)X_0$、$u_0$が正規分布に従うことから、その線形結合である予測誤差も正規分布に従う。つまり、

$$\hat{Y}_0 - Y_0 \sim N\left(0, \sigma^2\left(1 + \frac{1}{n} + \frac{(X_0 - \bar{X})^2}{\sum_{i=1}^n (X_i - \bar{X})^2}\right)\right)$$

となり、これを標準化(期待値0を引いて、分散の平方根で割る操作)すると、

$$\frac{\hat{Y}_0 - Y_0}{\sqrt{\sigma^2\left(1 + \frac{1}{n} + \frac{(X_0 - \bar{X})^2}{\sum_{i=1}^n (X_i - \bar{X})^2}\right)}} \sim N(0,1)$$

となります。

誤差項の分散σ^2は観察できないため、上式のσ^2を$s^2 = \frac{1}{n-2}\sum_{i=1}^n \hat{u}_i^2$で置き換えると、次のようになります。

$$\frac{\hat{Y}_0 - Y_0}{\sqrt{s^2\left(1 + \frac{1}{n} + \frac{(X_0 - \bar{X})^2}{\sum_{i=1}^n (X_i - \bar{X})^2}\right)}} = \frac{\hat{Y}_0 - Y_0}{\sqrt{\frac{1}{n-2}\sum_{i=1}^n \hat{u}_i^2\left(1 + \frac{1}{n} + \frac{(X_0 - \bar{X})^2}{\sum_{i=1}^n (X_i - \bar{X})^2}\right)}}$$

$$= \frac{\dfrac{\hat{Y}_0 - Y_0}{\sqrt{\sigma^2\left(1+\dfrac{1}{n}+\dfrac{(X_0-\bar{X})^2}{\sum_{i=1}^{n}(X_i-\bar{X})^2}\right)}}}{\sqrt{\dfrac{1}{n-2}\sum_{i=1}^{n}\left(\dfrac{\hat{u}_i}{\sigma}\right)^2}}$$

$$= \frac{N(0,1)}{\sqrt{\dfrac{1}{n-2}\chi^2(n-2)}} \sim t(n-2)$$

ここで、分子は標準正規分布$N(0,1)$、$\sum_{i=1}^{n}(\hat{u}_i/\sigma)^2$は自由度$n-2$の$\chi^2$分布です。つまり、分子は$N(0,1)$、分母は$\chi^2(n-2)$変数を自由度$n-2$で割って平方根をとったものであり、これは自由度$n-2$の$t$分布になります。

予測誤差の信頼区間

予測誤差$\hat{Y}_0 - Y_0$の 95%信頼区間は

$$0.95 = P\left\{ -t_{n-2,0.05} < \frac{\hat{Y}_0 - Y_0}{s\sqrt{1+\dfrac{1}{n}+\dfrac{(X_0-\bar{X})^2}{\sum_{i=1}^{n}(X_i-\bar{X})^2}}} < t_{n-2,0.05} \right\}$$

を用いて計算できます。{　}内は、両辺に$-s\sqrt{1+\frac{1}{n}+\frac{(X_0-\bar{X})^2}{\sum_{i=1}^{n}(X_i-\bar{X})^2}}$をかけてから$\hat{Y}_0$を足すと、

$$\hat{Y}_0 - t_{n-2,0.05}s\sqrt{1+\frac{1}{n}+\frac{(X_0-\bar{X})^2}{\sum_{i=1}^{n}(X_i-\bar{X})^2}} < Y_0 < \hat{Y}_0 + t_{n-2,0.05}s\sqrt{1+\frac{1}{n}+\frac{(X_0-\bar{X})^2}{\sum_{i=1}^{n}(X_i-\bar{X})^2}}$$

となります。ここで、$t_{n-2,0.05}$を 2 で置き換えると、95%信頼区間が得られます。

$$\hat{Y}_0 - 2s\sqrt{1+\frac{1}{n}+\frac{(X_0-\bar{X})^2}{\sum_{i=1}^{n}(X_i-\bar{X})^2}} < Y_0 < \hat{Y}_0 + 2s\sqrt{1+\frac{1}{n}+\frac{(X_0-\bar{X})^2}{\sum_{i=1}^{n}(X_i-\bar{X})^2}}$$