Лекция по эконометрике № 3

Линейная регрессионная модель для случая одной объясняющей переменной

Демидова
Ольга Анатольевна
https://www.hse.ru/staff/demidova_olga
E-mail:demidova@hse.ru
16.09.2019

План лекции № 3

- **Дисперсионный анализ**
- **R^2**
- **Теорема Гаусса-Маркова для парной регрессии**

Полезные результаты относительно регрессий

1) $\overline{Y} = \hat{\beta}_0 + \hat{\beta}_1 \overline{X}$

Доказательство

$$\hat{\beta}_0 = \overline{Y} - \hat{\beta}_1 \overline{X} \Rightarrow \quad \overline{Y} = \hat{\beta}_0 + \hat{\beta}_1 \overline{X}$$

Линия регрессии проходит через точку $(\overline{X}, \overline{Y})$

Полезные результаты относительно регрессий

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

Доказательство

$$Y_i = \hat{\beta}_0 + \hat{\beta}_1 X_i + e_i, \quad i = 1,..., n$$

$$e_i = Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i, \quad i = 1,..., n$$

$$\sum_{i=1}^{n} e_i = \sum_{i=1}^{n} Y_i - n\hat{\beta}_0 - \hat{\beta}_1 \sum_{i=1}^{n} X_i, \quad i = 1,..., n$$

$$\frac{1}{n} \sum_{i=1}^{n} e_i = \overline{Y} - \hat{\beta}_0 - \hat{\beta}_1 \overline{X} = 0$$

Отсутствует систематическая ошибка

Полезные результаты относительно регрессий

3) $\sum_{i=1}^{n} Y_i = \sum_{i=1}^{n} \hat{Y}_i$

Доказательство

$$Y_i = \hat{Y}_i + e_i, \quad i = 1,..., n$$

$$\sum_{i=1}^{n} Y_i = \sum_{i=1}^{n} \hat{Y}_i + \sum_{i=1}^{n} e_i,$$

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

Сумма всех значений Y совпадает с суммой всех выровненных Y.

Полезные результаты относительно регрессий

4) $\overline{Y} = \overline{\hat{Y}}$

Доказательство

$$Y_i = \hat{Y}_i + e_i, \quad i = 1,..., \quad n$$

$$\sum_{i=1}^{n} Y_i = \sum_{i=1}^{n} \hat{Y}_i + \sum_{i=1}^{n} e_i, \quad \sum_{i=1}^{n} e_i = 0$$

$$\sum_{i=1}^{n} Y_i = \sum_{i=1}^{n} \hat{Y}_i$$

$$\frac{1}{n}\sum_{i=1}^{n} Y_i = \frac{1}{n}\sum_{i=1}^{n} \hat{Y}_i$$

Среднее арифметическое по всем значениям Y совпадает со средним арифметическим по всех выровненным Y.

Полезные результаты относительно регрессий

5) $\sum_{i=1}^{n} X_i e_i = 0$

Доказательство

$$\frac{\partial RSS}{\partial \hat{\beta}_1} = 0 \quad \Rightarrow \quad 2\hat{\beta}_1 \sum X_i^2 - 2\sum X_i Y_i + 2\hat{\beta}_0 \sum X_i = 0$$

$$\sum X_i (Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i) = 0$$

$$\sum_{i=1}^{n} X_i e_i = 0$$

Полезные результаты относительно регрессий

6) $\sum_{i=1}^{n} \hat{Y}_i e_i = 0$

Доказательство

$$\sum_{i=1}^{n} \hat{Y}_i e_i = \sum_{i=1}^{n} (\hat{\beta}_0 + \hat{\beta}_1 X_i) e_i =$$

$$= \hat{\beta}_0 \sum_{i=1}^{n} e_i + \hat{\beta}_1 \sum_{i=1}^{n} X_i e_i = 0 + 0 = 0$$

Дисперсионный анализ

$$e_i = Y_i - \hat{Y}_i \;\Rightarrow\; Y_i = \hat{Y}_i + e_i, i = 1,\ldots,n$$

$$Y_i - \overline{Y} = (\hat{Y}_i - \overline{Y}) + (Y_i - \hat{Y}_i)$$

$$y_i = \hat{y}_i + e_i,$$

$$y_i^2 = \hat{y}_i^2 + 2\,\hat{y}_i e_i + e_i^2,$$

$$\sum_{i=1}^{n} y_i^2 = \sum_{i=1}^{n} \hat{y}_i^2 + 2\sum_{i=1}^{n} \hat{y}_i e_i + \sum_{i=1}^{n} e_i^2,$$

$$\sum_{i=1}^{n} \hat{y}_i e_i = \sum_{i=1}^{n} (\hat{Y}_i - \overline{Y}) e_i =$$

$$= \sum_{i=1}^{n} \hat{Y}_i e_i - \overline{Y} \sum_{i=1}^{n} e_i = 0 - 0 = 0,$$

Дисперсионный анализ

$$\sum_{i=1}^{n} y_i^2 = \sum_{i=1}^{n} \hat{y}_i^2 + \sum_{i=1}^{n} e_i^2 ,$$

Дисперсионный анализ

$$\sum_{i=1}^{n} y_i^2 = \sum_{i=1}^{n} \hat{y}_i^2 + \sum_{i=1}^{n} e_i^2 ,$$

$$\sum_{i=1}^{n} (Y_i - \overline{Y})^2 = \sum_{i=1}^{n} (\hat{Y}_i - \overline{Y})^2 + \sum_{i=1}^{n} e_i^2 ,$$

$$\sum_{i=1}^{n} (Y_i - \overline{Y})^2 = TSS \ (Total \ \ Sum \ \ of \ \ Squares \ )$$

$$\sum_{i=1}^{n} (\hat{Y}_i - \overline{Y})^2 = ESS \ (Explained \ \ \ Sum \ \ of \ \ Squares \ )$$

$$\sum_{i=1}^{n} e_i^2 = RSS \ (\text{Re} \ sidual \ \ Sum \ \ of \ \ Squares \ )$$

$$\boldsymbol{TSS = ESS + RSS}$$

Показатель качества подгонки регрессии R^2

$$TSS = ESS + RSS$$

$$R^2 = \frac{ESS}{TSS} = \frac{\sum (\hat{Y}_i - \bar{Y})^2}{\sum (Y_i - \bar{Y})^2} =$$

$$= \frac{\sum (\hat{Y}_i - \bar{Y})^2 /(n-1)}{\sum (Y_i - \bar{Y})^2 /(n-1)} = \frac{\hat{\text{var}}(\hat{Y})}{\hat{\text{var}}(Y)}$$

R^2 является отношением ESS к TSS, (или долей дисперсии Y, объясненной с помощью регрессии). Очевидно, это неотрицательная величина.

Показатель качества подгонки регрессии R^2

$$TSS = ESS + RSS$$

$$R^2 = \frac{ESS}{TSS} = \frac{TSS - RSS}{TSS} = 1 - \frac{RSS}{TSS} = 1 - \frac{\sum e_i^2}{\sum (Y_i - \bar{Y})^2}$$

Другое выражение для R^2, из которого следует, что R^2 не превышает 1.

Показатель качества подгонки регрессии R^2

R^2 действительно является квадратом, а именно, квадратом выборочного коэффициента корреляции X и Y.

Доказательство

$$R^2 = \frac{ESS}{TSS} = \frac{\sum\left(\hat{Y}_i - \bar{\hat{Y}}\right)^2}{\sum y_i^2} = \frac{\sum\left(\hat{\beta}_0 + \hat{\beta}_1 X_i - \hat{\beta}_0 - \hat{\beta}_1 \bar{X}\right)^2}{\sum y_i^2} =$$

$$= \frac{\sum\left(\hat{\beta}_1 (X_i - \bar{X})\right)^2}{\sum y_i^2} = \hat{\beta}_1^2 \cdot \frac{\sum (x_i)^2}{\sum y_i^2} = \frac{\left(\sum x_i y_i\right)^2}{\left(\sum x_i^2\right)^2} \cdot \frac{\sum x_i^2}{\sum y_i^2}$$

$$= \frac{\left(\sum x_i y_i\right)^2}{\left(\sum x_i^2\right)} \cdot \frac{1}{\sum y_i^2} = \hat{r}_{XY}^2$$

Оценки коэффициентов – случайные величины

$$Y = \beta_0 + \beta_1 X + \varepsilon$$

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X$$

$$\hat{\beta}_1 = \frac{\hat{\mathrm{cov}}(X, Y)}{\hat{\mathrm{var}}(X)} = \frac{\hat{\mathrm{cov}}(X, [\beta_0 + \beta_1 X + \varepsilon])}{\hat{\mathrm{var}}(X)}$$

$$= \frac{\hat{\mathrm{cov}}(X, \beta_0) + \hat{\mathrm{cov}}(X, \beta_1 X) + \hat{\mathrm{cov}}(X, \varepsilon)}{\hat{\mathrm{var}}(X)}$$

$$= \frac{0 + \beta_1 \hat{\mathrm{cov}}(X, X) + \hat{\mathrm{cov}}(X, \varepsilon)}{\hat{\mathrm{var}}(X)}$$

$$= \beta_1 + \frac{\hat{\mathrm{cov}}(X, \varepsilon)}{\hat{\mathrm{var}}(X)}$$

Оценки коэффициентов – случайные величины

$$Y = \beta_0 + \beta_1 X + \varepsilon$$

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X$$

$$\hat{\beta}_1 = \beta_1 + \frac{\hat{\mathrm{cov}}(X, \varepsilon)}{\hat{\mathrm{var}}(X)}$$

Т.к. X – детерминированный, а ε – случайный вектор, то

$\hat{\beta}_1$ **- случайная величина.**

Аналогично $\hat{\beta}_0$ **- случайная величина.**

Для того, чтобы найти их основные числовые характеристики, необходимо сделать предположения об ε.

Теорема Гаусса-Маркова для парной регрессии

1) Если модель $Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$, i = 1,…,n правильно специфицирована,

2) X_i детерминированы и не все равны между собой,

3) $E(\varepsilon_i) = 0$,

4) $var(\varepsilon_i) = \sigma_\varepsilon^2$,

5) $cov(\varepsilon_i, \varepsilon_j) = 0$ при $i \neq j$ (т.е. ошибки не коррелируют)

то оценки МНК β_0 и β_1 являются BLUE (best linear unbiased estimator).

BLUE

Estimator – оценка,

Unbiased – несмещенная,

Linear – по Y,

Best – это оценки с наименьшей дисперсией в классе всех линейных несмещенных оценок

Estimator (оценка)

$$\hat{\beta}_1 = \frac{\sum X_i Y_i - n\overline{X}\overline{Y}}{\sum (X_i - \overline{X})^2}$$

$$\hat{\beta}_1 = \frac{\sum (X_i - \overline{X})(Y_i - \overline{Y})}{\sum (X_i - \overline{X})^2},$$

$$\hat{\beta}_1 = \frac{\sum x_i y_i}{\sum x_i^2}, x_i = X_i - \overline{X}, \quad y_i = Y_i - \overline{Y}$$

$$\hat{\beta}_0 = \overline{Y} - \hat{\beta}_1 \overline{X}$$

Unbiased (несмещенная)

Доказательство несмещенности оценки $\hat{\beta}_1$

$$\hat{\beta}_1 = \beta_1 + \frac{\hat{\mathrm{cov}}(X, \varepsilon)}{\hat{\mathrm{var}}(X)}$$

$$E(\hat{\beta}_1) = E\left(\beta_1 + \frac{\hat{\mathrm{cov}}(X,\varepsilon)}{\hat{\mathrm{var}}(X)}\right) = E(\beta_1) + E\left(\frac{\hat{\mathrm{cov}}(X,\varepsilon)}{\hat{\mathrm{var}}(X)}\right)$$

$$= \beta_1 + \frac{1}{\hat{\mathrm{var}}(X)} E\left(\hat{\mathrm{cov}}(X,\varepsilon)\right) = \beta_1$$

Равенство последнего слагаемого нулю будет доказано далее.

Unbiased (несмещенная)

$$E\left(\mathrm{Cov}(X,\varepsilon)\right)=E\left(\frac{1}{n-1}\sum_{i=1}^{n}(X_i-\bar{X})(\varepsilon_i-\bar{\varepsilon})\right)$$

$$=\frac{1}{n-1}\sum_{i=1}^{n}E\left((X_i-\bar{X})(\varepsilon_i-\bar{\varepsilon})\right)$$

$$=\frac{1}{n-1}\sum_{i=1}^{n}(X_i-\bar{X})E(\varepsilon_i-\bar{\varepsilon})=\frac{1}{n-1}\sum_{i=1}^{n}(X_i-\bar{X})\times 0=0$$

Аналогично доказывается несмещенность оценки параметра β_0

Linear (линейная)

Если

$$Y = Y_1 + Y_2$$

$$Y_1 = \beta_0^{'} + \beta_1^{'} X + \varepsilon_1$$

$$Y_2 = \beta_0^{''} + \beta_1^{''} X + \varepsilon_2$$

$$Y = \beta_0 + \beta_1 X + \varepsilon, \quad \text{то}$$

$$\hat{\beta}_1 = \hat{\beta}_1^{'} + \hat{\beta}_1^{''},$$

$$\hat{\beta}_0 = \hat{\beta}_0^{'} + \hat{\beta}_0^{''},$$

Linear (линейная)

Доказательство

$$\hat{\beta}_1 = \frac{\hat{\mathrm{cov}}(X,Y)}{\hat{\mathrm{var}}(X)} = \frac{\hat{\mathrm{cov}}(X,Y_1+Y_2)}{\hat{\mathrm{var}}(X)} =$$

$$= \frac{\hat{\mathrm{cov}}(X,Y_1)}{\hat{\mathrm{var}}(X)} + \frac{\hat{\mathrm{cov}}(X,Y_2)}{\hat{\mathrm{var}}(X)} =$$

$$= \hat{\beta}_1^{'} + \hat{\beta}_1^{''},$$

$$\hat{\beta}_0 = \overline{Y} - \hat{\beta}_1 \overline{X} = \overline{Y_1} + \overline{Y_2} - (\hat{\beta}_1^{'} + \hat{\beta}_1^{''})\overline{X} =$$

$$= \hat{\beta}_0^{'} + \hat{\beta}_0^{''}$$

Linear (линейная)

Если

$$Y = \alpha Y_1$$

$$Y_1 = \beta_0^{'} + \beta_1^{'} X + \varepsilon_1$$

$$Y = \beta_0 + \beta_1 X + \varepsilon, \quad \text{то}$$

$$\hat{\beta}_1 = \alpha \hat{\beta}_1^{'},$$

$$\hat{\beta}_0 = \alpha \hat{\beta}_0^{'}$$

Linear (линейная)

Доказательство

$$\hat{\beta}_1 = \frac{\hat{\mathrm{cov}}(X,Y)}{\hat{\mathrm{var}}(X)} = \frac{\hat{\mathrm{cov}}(X,\alpha Y_1)}{\hat{\mathrm{var}}(X)} =$$

$$= \alpha\hat{\beta}_1',$$

$$\hat{\beta}_0 = \overline{Y} - \hat{\beta}_1\overline{X} = \alpha\overline{Y}_{12} - \alpha\hat{\beta}_1'\overline{X} = \alpha\hat{\beta}_0'$$

Best (наилучшая)

Best – это оценки с наименьшей дисперсией в классе всех линейных несмещенных оценок

(см доказательство в прикрепленных файлах).

$$\sigma^2_{\hat{\beta}_0} = \sigma^2_{\varepsilon} \frac{\sum X_i^2}{n\sum x_i^2} \qquad \sigma^2_{\hat{\beta}_1} = \frac{\sigma^2_{\varepsilon}}{\sum x_i^2}$$

Оценка дисперсии возмущений

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

Является несмещенной оценкой дисперсии возмущений

(см доказательство в прикрепленных файлах).

Оценка дисперсии возмущений

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

Является несмещенной оценкой дисперсии возмущений

(см доказательство в прикрепленных файлах).

Оценки дисперсии оценок коэффициентов

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

$$\sigma_{\hat{\beta}_0}^2 = \sigma_{\varepsilon}^2 \frac{\sum X_i^2}{n\sum x_i^2} \qquad \sigma_{\hat{\beta}_1}^2 = \frac{\sigma_{\varepsilon}^2}{\sum x_i^2}$$

$$\hat{\sigma}_{\hat{\beta}_0}^2 = \hat{\sigma}_{\varepsilon}^2 \frac{\sum X_i^2}{n\sum x_i^2} \qquad \hat{\sigma}_{\hat{\beta}_1}^2 = \frac{\hat{\sigma}_{\varepsilon}^2}{\sum x_i^2}$$

Стандартные ошибки коэффициентов регрессии

$$s.e.(\hat{\beta}_0) = \sqrt{\hat{\sigma}_\varepsilon^2 \frac{\sum X_i^2}{n\sum x_i^2}}$$

$$s.e.(\hat{\beta}_1) = \sqrt{\frac{\hat{\sigma}_\varepsilon^2}{\sum x_i^2}}$$

Стандартные ошибки коэффициентов регрессии

```
reg EARNINGS S
  Source |       SS       df       MS                  Number of obs =     570
---------+------------------------------                F(  1,   568) =   65.64
   Model |  3977.38016     1  3977.38016                Prob > F      =  0.0000
Residual |  34419.6569   568  60.5979875                R-squared     =  0.1036
---------+------------------------------                Adj R-squared =  0.1020
   Total |  38397.0371   569  67.4816117                Root MSE      =  7.7845
------------------------------------------------------------------------------
EARNINGS |      Coef.   Std. Err.       t     P>|t|       [95% Conf. Interval]
---------+--------------------------------------------------------------------
       S |   1.073055   .1324501      8.102   0.000        .8129028    1.333206
   _cons |  -1.391004   1.820305     -0.764   0.445       -4.966354    2.184347
------------------------------------------------------------------------------
```

Оценки стандартных отклонений (standard errors) автоматически выдаются при оценивании регрессии статистическими пакетами.