НИУ ВШЭ, ОП «Политология»
Курс «Теория вероятностей и математическая статистика», 2020 – 2021

Регрессионный анализ. Занятие 2

Выведение оценки коэффициента при предикторе в парной линейной регрессии посредством МНК

Рассмотрим частную производную по $\hat{\beta}_1$:

$$\frac{\partial \sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2}{\partial \hat{\beta}_1} = 0$$

$$(-2)\sum_{i=1}^{n}(x_i)(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0$$

$$\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n}\hat{\beta}_0 x_i - \sum_{i=1}^{n}\hat{\beta}_1 x_i^2 = 0$$

Вспомним, что ранее мы уже получили оценку константы, подставим ее в уравнение:

$$\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n}(\bar{y} - \hat{\beta}_1\bar{x})x_i - \sum_{i=1}^{n}\hat{\beta}_1 x_i^2 = 0$$

$$\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n}\bar{y}x_i + \sum_{i=1}^{n}\hat{\beta}_1\bar{x}x_i - \sum_{i=1}^{n}\hat{\beta}_1 x_i^2 = 0$$

$$\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i\bar{y} + \hat{\beta}_1\sum_{i=1}^{n} x_i\bar{x} - \hat{\beta}_1\sum_{i=1}^{n} x_i^2 = 0$$

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} x_i(y_i - \bar{y})}{\sum_{i=1}^{n} x_i(x_i - \bar{x})} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{\widehat{Cov}(x,y)}{\widehat{Var}(x)}$$

Тестирование значимости коэффициента в парной линейной регрессии

На первом шаге, как всегда, формулируем нулевую гипотезу и альтернативу. Обратите внимание, что гипотезы формулируются относительно генеральных параметров, а не об оценках, оценки нам известны по выборочным данным:

$$H_0 : \beta = 0$$

$$H_1 : \beta \neq 0$$

На втором шаге нужно обозначить статистику и ее распределение при верной нулевой гипотезе.

$$\frac{\widehat{\beta}}{st.error(\widehat{\beta})} \overset{H_0}{\sim} t(df = n - 2)$$

Надо отметить, что такое количество степеней $df = n - 2$ справедливо только для случая, когда в модели один предиктор, так как в парной регрессии оцениваются 2 коэффициента: константа и коэффициент при предикторе. Для проверки гипотезы Вы можете использовать как фиксированный уровень значимости, так и p-value.

Условия для получения идентифицируемой модели линейной регрессии и BLU-оценок (BLUE: best linear unbiased estimators)

В первую очередь, мы работаем со спецификацией линейной модели. Для того, чтобы модель была идентифицируемая (то есть, мы смогли получить оценки коэффициентов), наблюдений должно быть больше, чем количество оцениваемых параметров (в противном случае данных просто не хватает для оценивания), а также не должно быть строгой мультиколлинеарности (то есть, нет линейно зависимых предикторов).

Для получения BLUE-оценок (то есть, наиболее эффективных среди класса всех линейных несмещенных оценок) ошибки в модели должны удовлетворять ряду свойств:

- $E(e_i|x) = 0$
- $Var(e_i|x) = const$ – гомоскедастичность
- $Cov(e_i, e_j|x) = 0$ – отсутствие автокорреляции
- $Cov(e_i, x_i) = 0$ – экзогенность

Важно, что данные условия именно об ошибках, а не остатках. Стоит отметить, что в литературе нет полной согласованности относительно списка данных условий. Более подробно об этом можно прочитать в данной статье Larocca (2005).