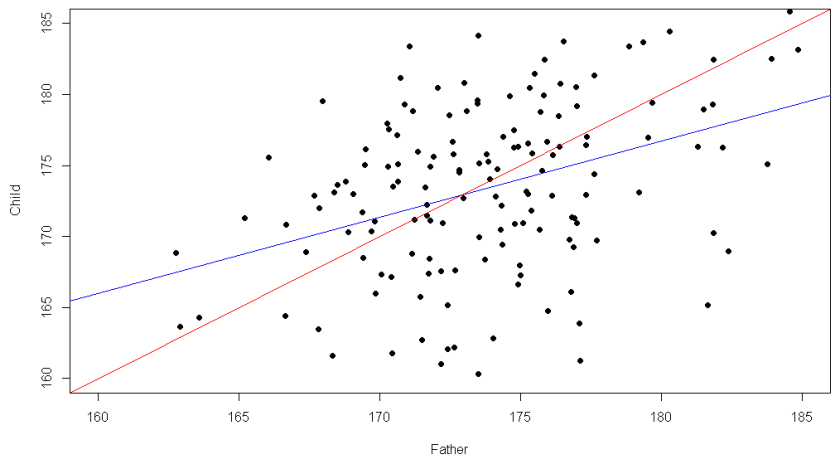
Лекция XII - Линейна регресия

Лекция XII - Линейна регресия

- Метод на най-малките квадрати
- Оценки за β_0 и β_1
- Свойства на коефицентите на регресия
- Проверка на хипотези в линейна регресия

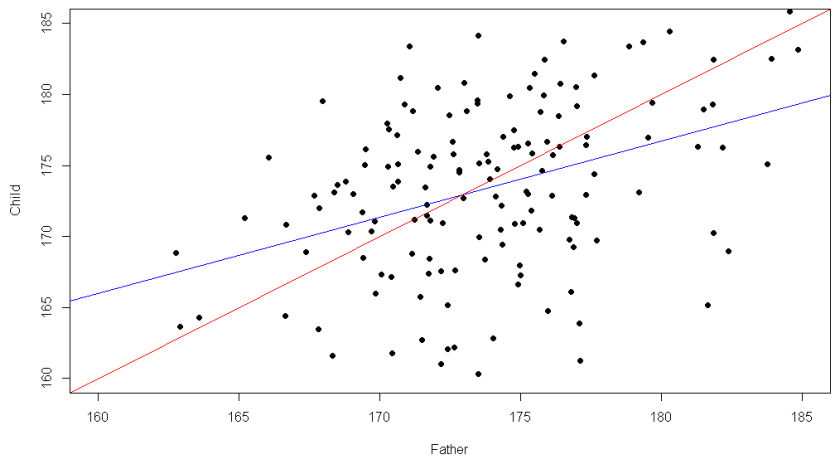
Линейна регресия

Първото ститистическо изследване за намиране на линейна връзка между две променливи, както и едно от първите статистически изследвания изобщо е проведено от Галтон в 1885г. Той е сравнявал височината на бащите и височината на синовете. Първоначалното предположение е било, че по-високите бащи имат и по-високи синове при това пропорционално така, че правата задаваща връзката е ъглополовящата с ъглов коефицент 1.



Линейна регресия

Първото ститистическо изследване за намиране на линейна връзка между две променливи, както и едно от първите статистически изследвания изобщо е проведено от Галтон в 1885г. Той е сравнявал височината на бащите и височината на синовете. Първоначалното предположение е било, че по-високите бащи имат и по-високи синове при това пропорционално така, че правата задаваща връзката е ъглополовящата с ъглов коефицент 1.



В действителност се оказва, че по високите бащи имат по високи синове, но не чак толкова, например бащи по високи с 10 см. от средното имат синове само с 6 см. по-високи и аналогично по-ниските бащи имат по-ниски синове, но не чак толкова. Като цяло ръстът на синовете е по-близо до средния, правата е с ъглов коефициент 0.6. Галтон нарича това "регрес към посредствеността". Така по исторически причини линейната връзка между случайни величини се нарича линейна регресия.

Описание на модела

Предполагаме че съществува линейна връзка между променливите X и Y, като евентуално тя е нарушена от някаква грешка ε.

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

Прието е X да се нарича независима променлива или предиктор, а Y зависима променлива или отклик. Възможно е X изобщо да не е случайна величина, а нейните стойности да бъдат предварително планувани, например дозата от някакво лекарство давана на пациента.

Описание на модела

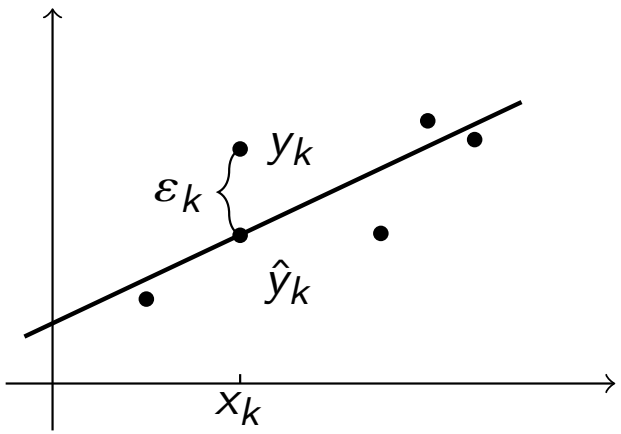
Предполагаме че съществува линейна връзка между променливите X и Y, като евентуално тя е нарушена от някаква грешка ε.

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

Прието е X да се нарича независима променлива или предиктор, а Y зависима променлива или отклик. Възможно е X изобщо да не е случайна величина, а нейните стойности да бъдат предварително планувани, например дозата от някакво лекарство давана на пациента.

Разполагаме с наблюдения над X и съответните стойности на Y, т.е. наблюденията са сдвоени (x_k, y_k) за $k = 1, \ldots, n$. Ние не знаем истинските стойности на коефициентите β_0 и β_1, целта ни е да намерим оценки за тях по данните с който разполагаме. Нека b_0 и b_1 са сответните оценки, а $\hat{y}_k$ е точката от правата която съответства на x_k. Тогава

$$y_k = b_0 + b_1 x_k + \varepsilon_k, \qquad \hat{y}_k = b_0 + b_1 x_k$$



Метод на най-малките квадрати

Нека ε_k е грешката на k-тото наблюдение. Оценките се намират по метода на най-малките квадрати, т.е. избираме такива стойности за b_0 и b_1, при които сумата от квадратите на грешките е минимална.

Прието е със SSR (Sum of Squared Residuals) да се означава сумата на грешките.

$$SSR = \sum_{k=1}^{n} \varepsilon_k^2 = \sum_{k=1}^{n} (y_k - b_0 - b_1 x_k)^2$$

Ще трябва да намерим минимума на тази функция по b_0 и b_1. Ясно е, че функцията е непрекъсната и диференцируема с една единствена особена точка, която е минимум и той ще се достига за

$$\left| \begin{array}{lll} 0 = \frac{\partial SSR}{\partial b_0} = & -2 \sum_{k=1}^{n} (y_k - b_0 - b_1 x_k) & = 2 \sum_{k=1}^{n} (\hat{y}_k - y_k) \\ 0 = \frac{\partial SSR}{\partial b_1} = & -2 \sum_{k=1}^{n} (y_k - b_0 - b_1 x_k) x_k & = 2 \sum_{k=1}^{n} (\hat{y}_k - y_k) x_k \end{array} \right.$$

Метод на най-малките квадрати

Нека ε_k е грешката на k-тото наблюдение. Оценките се намират по метода на най-малките квадрати, т.е. избираме такива стойности за b_0 и b_1, при които сумата от квадратите на грешките е минимална.

Прието е със SSR (Sum of Squared Residuals) да се означава сумата на грешките.

$$SSR = \sum_{k=1}^{n} \varepsilon_k^2 = \sum_{k=1}^{n} (y_k - b_0 - b_1 x_k)^2$$

Ще трябва да намерим минимума на тази функция по b_0 и b_1. Ясно е, че функцията е непрекъсната и диференцируема с една единствена особена точка, която е минимум и той ще се достига за

$$\left| \begin{array}{lll} 0 = \frac{\partial SSR}{\partial b_0} = & -2 \sum_{k=1}^{n} (y_k - b_0 - b_1 x_k) & = 2 \sum_{k=1}^{n} (\hat{y}_k - y_k) \\ 0 = \frac{\partial SSR}{\partial b_1} = & -2 \sum_{k=1}^{n} (y_k - b_0 - b_1 x_k) x_k & = 2 \sum_{k=1}^{n} (\hat{y}_k - y_k) x_k \end{array} \right.$$

От първото уравнение на системата получаваме

$$0 = \sum_{k=1}^{n} (y_k - b_0 - b_1 x_k) = \sum_{k=1}^{n} y_k - n b_0 - b_1 \sum_{k=1}^{n} x_k = n (\overline{y} - b_0 - b_1 \overline{x})$$

Тук с $\overline{y}$ и $\overline{x}$ сме означили средното аритметично. Следователно

$$\overline{y} = b_0 + b_1 \overline{x} \qquad (\star)$$

Метод на най-малките квадрати

Сега ще изразим b_1, за целта ще разгледаме разликата

$$\hat{y_k} - \overline{y} = b_0 + b_1 x_k - b_0 - b_1 \overline{x} = b_1 (x_k - \overline{x})$$

Метод на най-малките квадрати

Сега ще изразим b_1, за целта ще разгледаме разликата

$$\hat{y_k} - \overline{y} = b_0 + b_1 x_k - b_0 - b_1 \overline{x} = b_1 (x_k - \overline{x})$$

Ще умножим двете страни с $(x_k - \overline{x})$ и ще сумираме по k.

$$b_1 \sum_{k=1}^{n} (x_k - \overline{x})^2 = \sum_{k=1}^{n} (\hat{y_k} - \overline{y})(x_k - \overline{x}) = \sum_{k=1}^{n} \left[(\hat{y_k} - y_k) + (y_k - \overline{y}) \right] (x_k - \overline{x}) =$$

Метод на най-малките квадрати

Сега ще изразим b_1, за целта ще разгледаме разликата

$$\hat{y_k} - \overline{y} = b_0 + b_1 x_k - b_0 - b_1 \overline{x} = b_1 (x_k - \overline{x})$$

Ще умножим двете страни с $(x_k - \overline{x})$ и ще сумираме по k.

$$b_1 \sum_{k=1}^{n} (x_k - \overline{x})^2 = \sum_{k=1}^{n} (\hat{y_k} - \overline{y})(x_k - \overline{x}) = \sum_{k=1}^{n} \left[(\hat{y_k} - y_k) + (y_k - \overline{y}) \right] (x_k - \overline{x}) =$$

Ще разделим сумата на части

$$= \sum_{k=1}^{n} (\hat{y_k} - y_k) x_k - \overline{x} \sum_{k=1}^{n} (\hat{y_k} - y_k) + \sum_{k=1}^{n} (y_k - \overline{y})(x_k - \overline{x})$$

Метод на най-малките квадрати

Сега ще изразим b_1, за целта ще разгледаме разликата

$$\hat{y_k} - \overline{y} = b_0 + b_1 x_k - b_0 - b_1 \overline{x} = b_1 (x_k - \overline{x})$$

Ще умножим двете страни с $(x_k - \overline{x})$ и ще сумираме по k.

$$b_1 \sum_{k=1}^{n} (x_k - \overline{x})^2 = \sum_{k=1}^{n} (\hat{y_k} - \overline{y})(x_k - \overline{x}) = \sum_{k=1}^{n} \left[(\hat{y_k} - y_k) + (y_k - \overline{y}) \right] (x_k - \overline{x}) =$$

Ще разделим сумата на части

$$= \sum_{k=1}^{n} (\hat{y_k} - y_k) x_k - \overline{x} \sum_{k=1}^{n} (\hat{y_k} - y_k) + \sum_{k=1}^{n} (y_k - \overline{y})(x_k - \overline{x})$$

Първата сума е нула заради второто уравнение на ситемата, втората сума е нула от първото уравнение на системата. Тогава

$$b_1 \sum_{k=1}^{n} (x_k - \overline{x})^2 = \sum_{k=1}^{n} (y_k - \overline{y})(x_k - \overline{x})$$

От тук и от $(\star)$ елементарно се получават търсените оценки.

Метод на най-малките квадрати

Твърдение

Оценките по метод на най-малките квадрати за коефициентите β_0 и β_1 в линейната зависимост

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

са съответно

$$b_1 = \frac{\sum_{k=1}^{n}(y_k - \overline{y})(x_k - \overline{x})}{\sum_{k=1}^{n}(x_k - \overline{x})^2}$$

$$b_0 = \overline{y} - b_1\overline{x}$$

Метод на най-малките квадрати

Твърдение

Оценките по метод на най-малките квадрати за коефициентите β_0 и β_1 в линейната зависимост

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

са съответно

$$b_1 = \frac{\sum_{k=1}^{n}(y_k - \overline{y})(x_k - \overline{x})}{\sum_{k=1}^{n}(x_k - \overline{x})^2}$$

$$b_0 = \overline{y} - b_1\overline{x}$$

Както виждаме оценките винаги съществуват. Това все още не означава, че моделът е добър, т.е. че описва данните по подходяш начин, или че между X и Y наистина съществува линейна зависимост. Ще ни трябват критерии, по които да тестваме приложимостта на модела.

Ако разглеждаме наблюденията като случайни величини, каквито те в действителност са, то оценките b_0 и b_1 също са случайни величини. Намирането на техните характеристики ще ни позволи да строим доверителни интервали и да проверяваме хипотези свързани с тях.

Ще въведем някой ограничения на модела.

Ограничения на модела

- Грешките $\varepsilon_1, \ldots, \varepsilon_n$ са независими в съвкупност - това означава, че и наблюденията $Y_1, \ldots, Y_n$ са независими, което е нормално изискване в статистиката.

Ограничения на модела

- Грешките $\varepsilon_1, \ldots, \varepsilon_n$ са независими в съвкупност - това означава, че и наблюденията $Y_1, \ldots, Y_n$ са независими, което е нормално изискване в статистиката.
- $\mathrm{E}\varepsilon_k = 0,\ k = 1, \ldots, n$ - ако съществува някакво изместване, т.е. има не нулево очакване, то би трябвало да се отрази на коефициента b_0, а не на грешката.

Ограничения на модела

- Грешките $\varepsilon_1, \ldots, \varepsilon_n$ са независими в съвкупност - това означава, че и наблюденията $Y_1, \ldots, Y_n$ са независими, което е нормално изискване в статистиката.
- $\mathrm{E}\varepsilon_k = 0,\ k = 1, \ldots, n$ - ако съществува някакво изместване, т.е. има не нулево очакване, то би трябвало да се отрази на коефициента b_0, а не на грешката.
- $\mathrm{D}\varepsilon_k = \sigma^2,\ k = 1, \ldots, n$ - големината на грешките не зависи от $X_1, \ldots, X_n$, т.е. очакваме зависимостта между променливите да се описва от модела.

Ограничения на модела

- Грешките $\varepsilon_1, \ldots, \varepsilon_n$ са независими в съвкупност - това означава, че и наблюденията $Y_1, \ldots, Y_n$ са независими, което е нормално изискване в статистиката.
- $\mathrm{E}\varepsilon_k = 0,\ k = 1, \ldots, n$ - ако съществува някакво изместване, т.е. има не нулево очакване, то би трябвало да се отрази на коефициента b_0, а не на грешката.
- $\mathrm{D}\varepsilon_k = \sigma^2,\ k = 1, \ldots, n$ - големината на грешките не зависи от $X_1, \ldots, X_n$, т.е. очакваме зависимостта между променливите да се описва от модела.
- Грешките са нормално разпределени, т.е. $\varepsilon_k \in N(0, \sigma^2)$. Още Гаус при изследването на грешки в астрономически наблюдения е установил, че те са нормално разпределени. Така, че това изискване не е толкова ограничаващо и често се реализира на практика.

Ограничения на модела

- Грешките $\varepsilon_1, \ldots, \varepsilon_n$ са независими в съвкупност - това означава, че и наблюденията $Y_1, \ldots, Y_n$ са независими, което е нормално изискване в статистиката.
- $\mathrm{E}\varepsilon_k = 0,\ k = 1, \ldots, n$ - ако съществува някакво изместване, т.е. има не нулево очакване, то би трябвало да се отрази на коефициента b_0, а не на грешката.
- $\mathrm{D}\varepsilon_k = \sigma^2,\ k = 1, \ldots, n$ - големината на грешките не зависи от $X_1, \ldots, X_n$, т.е. очакваме зависимостта между променливите да се описва от модела.
- Грешките са нормално разпределени, т.е. $\varepsilon_k \in N(0, \sigma^2)$. Още Гаус при изследването на грешки в астрономически наблюдения е установил, че те са нормално разпределени. Така, че това изискване не е толкова ограничаващо и често се реализира на практика.

Сега ще определим математическото очакване на коефициента b_1. Разглеждаме отклиците $Y_1, \ldots, Y_n$ като случайни величини, а предикторите x_k като известни зададени стойности.

$$b_1 = \frac{\sum_{k=1}^{n}(Y_k - \overline{Y})(x_k - \overline{x})}{\sum_{k=1}^{n}(x_k - \overline{x})^2}$$

Математическо очакване на b_1

Ще разделим сумата в числителя на две суми

$$b_1 = \frac{\sum_{k=1}^{n} Y_k(x_k - \overline{x}) - \overline{Y}\sum_{k=1}^{n}(x_k - \overline{x})}{\sum_{k=1}^{n}(x_k - \overline{x})^2}$$

Математическо очакване на b_1

Ще разделим сумата в числителя на две суми

$$b_1 = \frac{\sum_{k=1}^{n} Y_k(x_k - \overline{x}) - \overline{Y}\sum_{k=1}^{n}(x_k - \overline{x})}{\sum_{k=1}^{n}(x_k - \overline{x})^2}$$

Втората сума е нула, тъй като

$$\sum_{k=1}^{n}(x_k - \overline{x}) = \sum_{k=1}^{n} x_k - n\overline{x} = \sum_{k=1}^{n} x_k - \sum_{k=1}^{n} x_k = 0$$

Математическо очакване на b_1

Ще разделим сумата в числителя на две суми

$$b_1 = \frac{\sum_{k=1}^{n} Y_k (x_k - \overline{x}) - \overline{Y} \sum_{k=1}^{n} (x_k - \overline{x})}{\sum_{k=1}^{n} (x_k - \overline{x})^2}$$

Втората сума е нула, тъй като

$$\sum_{k=1}^{n} (x_k - \overline{x}) = \sum_{k=1}^{n} x_k - n\overline{x} = \sum_{k=1}^{n} x_k - \sum_{k=1}^{n} x_k = 0$$

Тогава можем да разглеждаме b_1 като линейна комбинация на $Y_1, \ldots, Y_n$, с коефициенти v_k

$$b_1 = \sum_{k=1}^{n} \overbrace{\frac{(x_k - \overline{x})}{\sum_{m=1}^{n} (x_m - \overline{x})^2}} Y_k = \sum_{k=1}^{n} v_k \, Y_k$$

Математическо очакване на b_1

Ще разделим сумата в числителя на две суми

$$b_1 = \frac{\sum_{k=1}^{n} Y_k(x_k - \overline{x}) - \overline{Y}\sum_{k=1}^{n}(x_k - \overline{x})}{\sum_{k=1}^{n}(x_k - \overline{x})^2}$$

Втората сума е нула, тъй като

$$\sum_{k=1}^{n}(x_k - \overline{x}) = \sum_{k=1}^{n} x_k - n\overline{x} = \sum_{k=1}^{n} x_k - \sum_{k=1}^{n} x_k = 0$$

Тогава можем да разглеждаме b_1 като линейна комбинация на $Y_1, \ldots, Y_n$, с коефициенти v_k

$$b_1 = \sum_{k=1}^{n} \overbrace{\frac{(x_k - \overline{x})}{\sum_{m=1}^{n}(x_m - \overline{x})^2}}\, Y_k = \sum_{k=1}^{n} v_k\, Y_k$$

Ако сумираме коефициентите v_k, то сумата в числителя както показахме е нула. Тогава

$$\sum_{k=1}^{n} v_k = \frac{\sum_{k=1}^{n}(x_k - \overline{x})}{\sum_{m=1}^{n}(x_m - \overline{x})^2} = 0$$

и освен това

$$\sum_{k=1}^{n} v_k(x_k - \overline{x}) = \frac{\sum_{k=1}^{n}(x_k - \overline{x})^2}{\sum_{m=1}^{n}(x_m - \overline{x})^2} = 1$$

Математическо очакване на b_0 и b_1

Следователно

$$\sum_{k=1}^{n} v_k x_k = \sum_{k=1}^{n} v_k x_k - \overline{x} \overbrace{\sum_{k=1}^{n} v_k}^{=0} = \sum_{k=1}^{n} v_k (x_k - \overline{x}) = 1$$

Знаем, че $\mathrm{E}Y_k = \mathrm{E}(\beta_0 + \beta_1 x_k + \varepsilon_k) = \beta_0 + \beta_1 x_k$, така от линейното представяне на b_1 получаваме

$$\mathrm{E}b_1 = \sum_{k=1}^{n} v_k \, \mathrm{E}Y_k = \beta_0 \overbrace{\sum_{k=1}^{n} v_k}^{0} + \beta_1 \overbrace{\sum_{k=1}^{n} v_k \, x_k}^{1} = \beta_1$$

Това означава, че b_1 е неизместена оценка за β_1.

Математическо очакване на b_0 и b_1

Следователно

$$\sum_{k=1}^{n} v_k x_k = \sum_{k=1}^{n} v_k x_k - \overline{x} \overbrace{\sum_{k=1}^{n} v_k}^{=0} = \sum_{k=1}^{n} v_k (x_k - \overline{x}) = 1$$

Знаем, че $\mathrm{E}Y_k = \mathrm{E}(\beta_0 + \beta_1 x_k + \varepsilon_k) = \beta_0 + \beta_1 x_k$, така от линейното представяне на b_1 получаваме

$$\mathrm{E}b_1 = \sum_{k=1}^{n} v_k \, \mathrm{E}Y_k = \beta_0 \overbrace{\sum_{k=1}^{n} v_k}^{0} + \beta_1 \overbrace{\sum_{k=1}^{n} v_k \, x_k}^{1} = \beta_1$$

Това означава, че b_1 е неизместена оценка за β_1.
Отново ще използваме $\mathrm{E}Y_k$ за да пресметнем

$$\mathrm{E}\overline{Y} = \frac{1}{n} \sum_{k=1}^{n} \mathrm{E}Y_k = \frac{1}{n} \sum_{k=1}^{n} (\beta_0 + \beta_1 x_k) = \beta_0 + \beta_1 \overline{x}$$

За математическото очакване на b_0 получаваме

$$\mathrm{E}b_0 = \mathrm{E}(\overline{Y} - b_1 \overline{x}) = \mathrm{E}\overline{Y} - \overline{x} \, \mathrm{E}b_1 = \beta_0 + \beta_1 \overline{x} - \overline{x} \beta_1 = \beta_0$$

Това означава, че и оценката b_0 е неизместена.

Дисперсия на b_1

За да пресметнем дисперсията на b_1 отново ще се върнем към представянето му като линейна комбинация

$$b_1 = \sum_{k=1}^{n} v_k \, Y_k = \sum_{k=1}^{n} v_k \, (\beta_0 + \beta_1 x_k + \varepsilon_k) = \overbrace{\sum_{k=1}^{n} v_k \, (\beta_0 + \beta_1 x_k)}^{Const} + \sum_{k=1}^{n} v_k \, \varepsilon_k$$

Дисперсия на b_1

За да пресметнем дисперсията на b_1 отново ще се върнем към представянето му като линейна комбинация

$$b_1 = \sum_{k=1}^{n} v_k \, Y_k = \sum_{k=1}^{n} v_k \, (\beta_0 + \beta_1 x_k + \varepsilon_k) = \overbrace{\sum_{k=1}^{n} v_k \, (\beta_0 + \beta_1 x_k)}^{Const} + \sum_{k=1}^{n} v_k \, \varepsilon_k$$

Първата сума е равна на константа следователно дисперсията и е нула. Тогава от $\mathrm{D}\varepsilon_k = \sigma^2$ следва

$$\mathrm{D}b_1 = \sum_{k=1}^{n} \mathrm{D} \, (v_k \, \varepsilon_k) = \sigma^2 \sum_{k=1}^{n} v_k^2 = \sigma^2 \sum_{k=1}^{n} \left[\frac{(x_k - \overline{x})}{\sum_{m=1}^{n} (x_m - \overline{x})^2} \right]^2 =$$

$$= \sigma^2 \frac{\sum_{k=1}^{n} (x_k - \overline{x})^2}{\left[\sum_{m=1}^{n} (x_m - \overline{x})^2 \right]^2} = \frac{\sigma^2}{\sum_{k=1}^{n} (x_k - \overline{x})^2}$$

Дисперсия на b_1

За да пресметнем дисперсията на b_1 отново ще се върнем към представянето му като линейна комбинация

$$b_1 = \sum_{k=1}^{n} v_k\ Y_k = \sum_{k=1}^{n} v_k\ (\beta_0 + \beta_1 x_k + \varepsilon_k) = \overbrace{\sum_{k=1}^{n} v_k\ (\beta_0 + \beta_1 x_k)}^{Const} + \sum_{k=1}^{n} v_k\ \varepsilon_k$$

Първата сума е равна на константа следователно дисперсията и е нула. Тогава от $\mathrm{D}\varepsilon_k = \sigma^2$ следва

$$\mathrm{D}b_1 = \sum_{k=1}^{n} \mathrm{D}\left(v_k\ \varepsilon_k\right) = \sigma^2 \sum_{k=1}^{n} v_k^2 = \sigma^2 \sum_{k=1}^{n} \left[\frac{(x_k - \overline{x})}{\sum_{m=1}^{n}(x_m - \overline{x})^2}\right]^2 =$$

$$= \sigma^2 \frac{\sum_{k=1}^{n}(x_k - \overline{x})^2}{\left[\sum_{m=1}^{n}(x_m - \overline{x})^2\right]^2} = \frac{\sigma^2}{\sum_{k=1}^{n}(x_k - \overline{x})^2}$$

Както показахме по-горе b_1 е линейна комбинация на Y_k, които са нормално разпределени. Следователно b_1 също е нормално разпределена, при това ние изведохме очакването и дисперсията и.

$$b_1\ \in\ N\left(\beta_1, \frac{\sigma^2}{\sum_{k=1}^{n}(x_k - \overline{x})^2}\right) \qquad (*)$$

Дисперсия на b_0

$$b_0 = \overline{Y} - b_1\overline{x} = \frac{1}{n}\sum_{k=1}^{n} Y_k - \overline{x}\sum_{k=1}^{n} v_k\ Y_k = \sum_{k=1}^{n}\left(\frac{1}{n} - \overline{x}\ v_k\right) Y_k =$$

$$= \sum_{k=1}^{n}\left(\frac{1}{n} - \overline{x}\ v_k\right)(\beta_0 + \beta_1 x_k + \varepsilon_k) =$$

$$= \underbrace{\sum_{k=1}^{n}\left(\frac{1}{n} - \overline{x}\ v_k\right)(\beta_0 + \beta_1 x_k)}_{const} + \sum_{k=1}^{n}\left(\frac{1}{n} - \overline{x}\ v_k\right)\varepsilon_k$$

Дисперсия на b_0

$$b_0 = \overline{Y} - b_1\overline{x} = \frac{1}{n}\sum_{k=1}^{n} Y_k - \overline{x}\sum_{k=1}^{n} v_k\ Y_k = \sum_{k=1}^{n}\left(\frac{1}{n} - \overline{x}\ v_k\right) Y_k =$$

$$= \sum_{k=1}^{n}\left(\frac{1}{n} - \overline{x}\ v_k\right)(\beta_0 + \beta_1 x_k + \varepsilon_k) =$$

$$= \underbrace{\sum_{k=1}^{n}\left(\frac{1}{n} - \overline{x}\ v_k\right)(\beta_0 + \beta_1 x_k)}_{const} + \sum_{k=1}^{n}\left(\frac{1}{n} - \overline{x}\ v_k\right)\varepsilon_k$$

Първата сума е константа, тогава за дисперсията получаваме

$$\mathrm{D}b_0 = \sum_{k=1}^{n}\left(\frac{1}{n} - \overline{x}\ v_k\right)^2 \mathrm{D}\varepsilon_k = \sigma^2\sum_{k=1}^{n}\left(\frac{1}{n} - \overline{x}\ v_k\right)^2 =$$

$$= \sigma^2\left(\sum_{k=1}^{n}\frac{1}{n^2} - \frac{2\,\overline{x}}{n}\sum_{k=1}^{n} v_k + \overline{x}^2\sum_{k=1}^{n} v_k^2\right) = \sigma^2\left(\frac{1}{n} + \frac{\overline{x}^2}{\sum_{k=1}^{n}(x_k - \overline{x})^2}\right)$$

Статистиката b_0 също е линейна комбинация на Y_k, следователно и тя е нормално разпределена.

$$b_0 \ \in\ N\left(\beta_0,\ \sigma^2\left[\frac{1}{n} + \frac{\overline{x}^2}{\sum_{k=1}^{n}(x_k - \overline{x})^2}\right]\right)$$

Оценка за σ^2

Дотук намерихме дисперсията на оценките b_0 и b_1, но тя зависи от дисперсията на грешката σ^2. Ако тя е неизвестна, а в практически задачи най-често е така, ще трябва да оценим σ^2 също от данните.

Оценка за σ^2

Дотук намерихме дисперсията на оценките b_0 и b_1, но тя зависи от дисперсията на грешката σ^2. Ако тя е неизвестна, а в практически задачи най-често е така, ще трябва да оценим σ^2 също от данните.

Съгласно предположенията, които направихме за модела $\varepsilon_k \in N(0, \sigma^2)$. Тогава за Y_k също е случайна величина с нормално разпределение, доколкото β_0, β_1 и x_k са константи, т.е. $Y_k = \beta_0 + \beta_1 x_k + \varepsilon_k \in N(\beta_0 + \beta_1 x_k, \sigma^2)$. Следователно

$$\frac{Y_k - \beta_0 + \beta_1 x_k}{\sigma} \in N(0,1)$$

Както показахме в Лекция VIII сумата от квадратите на тези случайни величини ще има хи-квадрат разпределение с n степени на свобода.

$$\sum_{k=1}^{n} \frac{(Y_k - \beta_0 + \beta_1 x_k)^2}{\sigma^2} \in \chi^2(n)$$

Така можем да намерим разпределението на SSR. Оказва се, че

$$\frac{SSR}{\sigma^2} = \frac{\sum_{k=1}^{n}(Y_k - b_0 + b_1 x_k)^2}{\sigma^2} \in \chi^2(n-2)$$

В SSR два от параметрите , а именно β_0 и β_1 са оценени от данните затова и степените на свобода падат с две. Няма да доказваме този факт формално. Подобно доказателство приведохме за S^2 (Твърдение 2 от Лекция XII).

Оценка за σ^2

Знаем, че хи-квадрат разпределението е частен случай на гама

$$X \in \chi^2(n) \qquad \Longleftrightarrow \qquad X \in \Gamma\left(\frac{n}{2}, \frac{1}{2}\right)$$

Така можем да изведем формула за очакването на хи-квадрат, като знаем очакването на гама разпределението, а именно $\mathrm{E}X = \frac{n/2}{1/2} = n$. Следователно

$$\mathrm{E}\left(\frac{SSR}{\sigma^2}\right) = n - 2$$

Това означава, че SSR корегирано със съответната константа, може да се разглежда като неизместена оценка за параметъра σ^2

$$\hat{\sigma}^2 = \frac{SSR}{n-2}$$

Оценка за σ^2

Знаем, че хи-квадрат разпределението е частен случай на гама

$$X \in \chi^2(n) \qquad \Longleftrightarrow \qquad X \in \Gamma\left(\frac{n}{2}, \frac{1}{2}\right)$$

Така можем да изведем формула за очакването на хи-квадрат, като знаем очакването на гама разпределението, а именно $\mathrm{E}X = \frac{n/2}{1/2} = n$. Следователно

$$\mathrm{E}\left(\frac{SSR}{\sigma^2}\right) = n - 2$$

Това означава, че SSR корегирано със съответната константа, може да се разглежда като неизместена оценка за параметъра σ^2

$$\hat{\sigma}^2 = \frac{SSR}{n-2}$$

Ще обърнем внимание, че в SSR участват оценените параметри b_0 и b_1, тогава оценката $\hat{\sigma}^2$ всъщност е зависима от модела и при друг модел естествено ще бъде друга.

Също така е възможно да се използва метода на максимално правдоподобие за да се построи статистика за оценка на параметъра σ^2. Получената в този случай оценка е различна от дадената по-горе, при това тя е и изместена. Затова ще предпочетем $\hat{\sigma}^2$ за по-нататъшните изчисления.

Проверка на хипотези за b_1

Вече разполагаме с всичко необходимо за да пристъпим към тестване на хипотези свързани с модела.

Най често първата хипотеза, който проверяваме е

$$H_0 \ : \ \beta_1 = 0$$

Тази хипотеза е важна, защото ако $\beta_1 = 0$ то модела се изражда до $Y = \beta_0 + \varepsilon$, което означава, че изобщо не съществува линейна връзка между X и Y, т.е. линейната регресия е безмислена.

Проверка на хипотези за b_1

Вече разполагаме с всичко необходимо за да пристъпим към тестване на хипотези свързани с модела.

Най често първата хипотеза, който проверяваме е

$$H_0 \ : \ \beta_1 = 0$$

Тази хипотеза е важна, защото ако $\beta_1 = 0$ то модела се изражда до $Y = \beta_0 + \varepsilon$, което означава, че изобщо не съществува линейна връзка между X и Y, т.е. линейната регресия е безмислена.

Ние ще разгледаме тази хипотеза като частен случай на по общата

$$H_0 \ : \ \beta_1 = b$$

$$H_1 \ : \ \beta_1 \neq b$$

където b е известна константа. Нека α е нивото на значимост.

Както показахме по-горе $(*)$ b_1 е нормално разпределена, което ни позволява да конструираме критична област за проверка на хипотезата.

- Ако σ^2 е известна ще използваме статистика

$$Z = \frac{b_1 - b}{\sqrt{\frac{\sigma^2}{\sum_{k=1}^{n}(x_k - \overline{x})^2}}}$$

При изпълнена хипотеза H_0, следва $Z \in N(0,1)$.

Проверка на хипотези за b_1

Критичната област има вида

$$W = \{|Z| \geq q_{\alpha/2}\}$$

където $q_{\alpha/2}$ е квантил на $N(0,1)$.

Идеята тук е, че ако е изпълнена H_0, то стойността на b_1 ще бъде близо до b и съответно статистиката Z ще е около нулата. Тогава, ако Z е прекалено малка или пракалено голяма трябва да отхвърлим H_0 и да приемем H_1. Това обуславя критична зона от типа $|Z| \geq Const$.

Проверка на хипотези за b_1

Критичната област има вида

$$W = \left\{|Z| \geq q_{\alpha/2}\right\}$$

където $q_{\alpha/2}$ е квантил на $N(0,1)$.

Идеята тук е, че ако е изпълнена H_0, то стойността на b_1 ще бъде близо до b и съответно статистиката Z ще е около нулата. Тогава, ако Z е прекалено малка или пракалено голяма трябва да отхвърлим H_0 и да приемем H_1. Това обуславя критична зона от типа $|Z| \geq Const$.

Съответно, ако проверяваме същата хипотеза срещу едностранна алтернатива $H_1 : \beta_1 > b$, ще използваме критична област

$$W = \{Z \geq q_\alpha\}$$

А при едностранна алтернатива $H_1 : \beta_1 < b$, критичната област е

$$W = \{Z \leq q_\alpha\}$$

Проверка на хипотези за b_1

Критичната област има вида

$$W = \{|Z| \geq q_{\alpha/2}\}$$

където $q_{\alpha/2}$ е квантил на $N(0,1)$.

Идеята тук е, че ако е изпълнена H_0, то стойността на b_1 ще бъде близо до b и съответно статистиката Z ще е около нулата. Тогава, ако Z е прекалено малка или пракалено голяма трябва да отхвърлим H_0 и да приемем H_1. Това обуславя критична зона от типа $|Z| \geq Const$.

Съответно, ако проверяваме същата хипотеза срещу едностранна алтернатива $H_1 : \beta_1 > b$, ще използваме критична област

$$W = \{Z \geq q_\alpha\}$$

А при едностранна алтернатива $H_1 : \beta_1 < b$, критичната област е

$$W = \{Z \leq q_\alpha\}$$

Статистиката Z може да бъде използвана и като централна статистика (Лекция XII) за построяване на доверителен интервал за β_1. Достатъчно е в нея b да бъде заменено с β_1 и по познатия начин да се изрази β_1. Полученият интервал е

$$I = \left\{ b_1 \pm q_{\alpha/2} \sqrt{\frac{\sigma^2}{\sum_{k=1}^{n} (x_k - \overline{x})^2}} \right\}$$

Проверка на хипотези за b_1

- Ако σ^2 е неизвестна използваме подобна на Z статистика, в която заместваме дисперсията с оценката $\hat{\sigma}^2$за нея

$$T = \frac{b_1 - b}{\sqrt{\frac{\hat{\sigma}^2}{\sum_{k=1}^{n}(x_k - \overline{x})^2}}} = \frac{b_1 - b}{\sqrt{\frac{SSR}{(n-2)\sum_{k=1}^{n}(x_k - \overline{x})^2}}}$$

Проверка на хипотези за b_1

- Ако σ^2 е неизвестна използваме подобна на Z статистика, в която заместваме дисперсията с оценката $\hat{\sigma}^2$за нея

$$T = \frac{b_1 - b}{\sqrt{\frac{\hat{\sigma}^2}{\sum_{k=1}^{n}(x_k - \overline{x})^2}}} = \frac{b_1 - b}{\sqrt{\frac{SSR}{(n-2)\sum_{k=1}^{n}(x_k - \overline{x})^2}}}$$

Не е трудно да се съобрази, че

$$T = \frac{Z}{\sqrt{\frac{SSR}{(n-2)\sigma^2}}}$$

С подобни статистики работихме при конструирането на доверителен интервал за очакването на нормално разпределение (Лекция XII). Знаем, че $Z \in N(0,1)$ и $\frac{SSR}{\sigma^2} \in \chi^2(n-2)$ следователно T има разпределение на Стюдънт с $n-2$ степени на свобода, т.е. $T \in t(n-2)$. Критичните области се конструират както в случая с известна диспресия, единствената разлика е, че квантилите се взимат от таблици на Стюдънт.

$$H_1 : \beta_1 \neq b \qquad H_1 : \beta_1 > b \qquad H_1 : \beta_1 < b$$

$$W = \{|Z| \geq q_{\alpha/2}\} \qquad W = \{Z \geq q_\alpha\} \qquad W = \{Z \leq q_\alpha\}$$

Проверка на хипотези за b_1

- Ако σ^2 е неизвестна използваме подобна на Z статистика, в която заместваме дисперсията с оценката $\hat{\sigma}^2$за нея

$$T = \frac{b_1 - b}{\sqrt{\frac{\hat{\sigma}^2}{\sum_{k=1}^{n}(x_k - \overline{x})^2}}} = \frac{b_1 - b}{\sqrt{\frac{SSR}{(n-2)\sum_{k=1}^{n}(x_k - \overline{x})^2}}}$$

Не е трудно да се съобрази, че

$$T = \frac{Z}{\sqrt{\frac{SSR}{(n-2)\sigma^2}}}$$

С подобни статистики работихме при конструирането на доверителен интервал за очакването на нормално разпределение (Лекция XII). Знаем, че $Z \in N(0,1)$ и $\frac{SSR}{\sigma^2} \in \chi^2(n-2)$ следователно T има разпределение на Стюдънт с $n-2$ степени на свобода, т.е. $T \in t(n-2)$. Критичните области се конструират както в случая с известна диспресия, единствената разлика е, че квантилите се взимат от таблици на Стюдънт.

$$H_1 : \beta_1 \neq b \qquad H_1 : \beta_1 > b \qquad H_1 : \beta_1 < b$$

$$W = \{|Z| \geq q_{\alpha/2}\} \qquad W = \{Z \geq q_\alpha\} \qquad W = \{Z \leq q_\alpha\}$$

Аналогично на предходния случай T може да се използва за построяване на доверителен интервал за b_1.

Проверка на хипотези за b_0

Както доказахме по-горе

$$b_0 \in N\left(\beta_0, \sigma^2\left[\frac{1}{n} + \frac{\overline{x}^2}{\sum_{k=1}^{n}(x_k - \overline{x})^2}\right]\right)$$

Ако е изпълнена хипотеза $H_0 : \beta_0 = b$, ще знаем разпределенията на Z и T съответно за случайте на известна и неизвестна дисперсия.

$$Z = \frac{b_0 - b}{\sqrt{\sigma^2\left(\frac{1}{n} + \frac{\overline{x}^2}{\sum_{k=1}^{n}(x_k - \overline{x})^2}\right)}} \in N(0,1)$$

$$T = \frac{b_0 - b}{\sqrt{\frac{SSR}{n-2}\left(\frac{1}{n} + \frac{\overline{x}^2}{\sum_{k=1}^{n}(x_k - \overline{x})^2}\right)}} \in t(n-2)$$

Критичните области за проверка на хипотези и доверителните интервали се построяват аналогично на тези за b_1. Ще оставим любознателния читател да довърши подробностите.

Проверка на хипотези за b_0

Както доказахме по-горе

$$b_0 \in N\left(\beta_0,\ \sigma^2\left[\frac{1}{n}+\frac{\overline{x}^2}{\sum_{k=1}^{n}(x_k-\overline{x})^2}\right]\right)$$

Ако е изпълнена хипотеза $H_0 : \beta_0 = b$, ще знаем разпределенията на Z и T съответно за случайте на известна и неизвестна дисперсия.

$$Z = \frac{b_0 - b}{\sqrt{\sigma^2\left(\frac{1}{n}+\frac{\overline{x}^2}{\sum_{k=1}^{n}(x_k-\overline{x})^2}\right)}} \in N(0,1)$$

$$T = \frac{b_0 - b}{\sqrt{\frac{SSR}{n-2}\left(\frac{1}{n}+\frac{\overline{x}^2}{\sum_{k=1}^{n}(x_k-\overline{x})^2}\right)}} \in t(n-2)$$

Критичните области за проверка на хипотези и доверителните интервали се построяват аналогично на тези за b_1. Ще оставим любознателния читател да довърши подробностите.

21.6.2023 ЕК