Занятие 2

Выведение оценок коэффициентов линейной регрессии посредством метода наименьших квадратов

1. **Для начала в качестве разминки рассмотрим случай регрессии на контанту**

   Запишем исходную спецификацию:

   $$y_i = \hat{\beta}_0 + \hat{\varepsilon}_i$$

   Перепишем в терминах модельных (предсказанных) значений, то есть, отклик (зависимая переменная) в среднем равна константе (некоторому постоянному значению):

   $$\hat{y}_i = \hat{\beta}_0$$

   Руководствуясь принципом МНК, минимизируем сумму квадратов остатков:

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

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

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

   $$\hat{\beta}_0 = \frac{\sum_{i=1}^{n} y_i}{n}$$

   $$\boxed{\hat{\beta}_0 = \bar{y}}$$

2. **Случай парной регрессии (один предиктор)**

   Найдем оптимальную оценку константы ($\hat{\beta}_0$) в парной линейной регрессии, при которой сумма квадратов остатков будет минимальна. Рассмотрим частную производную по $\hat{\beta}_0$:

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

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

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

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

Мы получили оценку константы в парной регрессии:

$$\boxed{\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}}$$

Далее мы получим МНК-оценку коэффициента при предикторе в парной линейной регрессии.

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:

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

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

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

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

На основе выведения оценок коэффициентов мы получили, в частности, что

1. $\sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0$
   То есть, сумма остатков равна 0.
2. $\sum_{i=1}^{n} (x_i)(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0$
   То есть, скалярное произведение x и вектора остатков равно 0. Или иными словами, корреляция между объясняющей переменной (предиктором) и остатками равна 0.

Это утверждения верные по построению регрессионной модели!

3. Случай множественной регрессии (k предикторов)

Запишем спецификацию множественной регрессионной модели в векторно-матричном виде:

$$\begin{pmatrix} y_1 \\ ... \\ y_N \end{pmatrix} = \begin{pmatrix} 1 & x_{11} & ... & x_{k1} \\ ... & ... & ... & .... \\ 1 & x_{1N} & ... & x_{kN} \end{pmatrix} \times \begin{pmatrix} \hat{\beta}_0 \\ ... \\ \hat{\beta}_k \end{pmatrix} + \begin{pmatrix} \hat{\epsilon_1} \\ ... \\ \hat{\epsilon_N} \end{pmatrix}$$

Мы знаем, что остатки и предикторы нескоррелированны в линейной регрессионной модели по ее построению. То есть,

$$X^T(\vec{y} - X\vec{\hat{\beta}}) = 0$$

$$X^T\vec{y} = X^TX\vec{\hat{\beta}}$$

$$\vec{\hat{\beta}} = (X^TX)^{-1}X^T\vec{y}$$

4. R^2 – коэффициент детерминации (показывает, какую долю дисперсии зависимой переменной объясняют включенные в модель предикторы)

$$R^2 = \frac{ESS}{TSS} = \frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} = \hat{cor}^2(\hat{y_i}, y_i)$$

Для того, чтобы проверить гипотезу о незначимости модели в целом (или незначимости R^2), мы используем статистику Фишера:

$$H_0 : \beta_1 = ... = \beta_k = 0$$

$$F = \frac{ESS/k}{RSS/(n-k-1)} = \frac{R^2/k}{(1-R^2)/(n-k-1)}$$

Если нулевая гипотеза отвергается (p-value мало), то можем сделать вывод, что наша модель небесполезна, лучше, чем модель на константу.

Задание 1. Задача из семинарского листка

На данных по 44 городам построена модель, обясняющая динамику уровня преступности за последние 10 лет. `change_in_crime_rate` — прирост преступности в %, `change_in_pop` — прирост численности населения, %; `kids` — процент детей; `free_lunch` — процент бесплатных школьных обедов; `income_change` — прирост доходов домохозяйств.

Решение:
Ниже таблица с воостановленными пропусками:

```
Coefficients:
                 coef     std. err  t      Pr>|t|  [0.025  0.975]
Intercept     -22.3548    12.3097  -1.816  0.0771  -47.253; 2.544
change_in_pop   0.3188     0.2052   1.533  0.1333   -0.096; 0.734
kids            1.1128     0.2869   3.879  0.0004    0.532; 1.693
free_lunch     -0.3681     0.0973  -3.783  0.0005   -0.565;-0.171
income_change  -0.1944     0.3681  -0.528  0.6004   -0.939; 0.551
---

              df   sum_sq  mean_sq  f      PR(>F)
change_in_pop  1    803.2   803.2  6.248  .000
kids           1   1380.1  1380.1
free_lunch     1   3186.6  3186.6
income_change  1     60.6    60.6
Residual       39  8476.0   217.3
---
```

$$R^2 = \frac{803.2 + 1380.1 + 3186.6 + 60.6}{803.2 + 1380.1 + 3186.6 + 60.6 + 8476} = 0.39$$

Так как для F-статистики значение p-value = 0, то мы отвергаем нулевую гипотезу о незначимости всех коэффициентов при предикторах (незначимости модели в целом), значит, можем говорить, что данная модель лучше, чем модель на константу.