

Занятие 2

Разложение дисперсии

$$\sum_{i=1}^n (y_i - \bar{y})^2 = \sum_{i=1}^n (\hat{y}_i - \bar{y})^2 + \sum_{i=1}^n (y_i - \hat{y}_i)^2$$

$$TSS = ESS + RSS$$

$$R^2 = \frac{ESS}{TSS} = 1 - \frac{RSS}{TSS}$$

Общий вид таблицы разложения дисперсии:

Source	df	Sum of Squares	Mean Square	F
x	k	ESS	$\frac{ESS}{k}$	$\frac{ESS/k}{RSS/(n-k-1)}$
Residual	n - k - 1	RSS	$\frac{RSS}{n-k-1}$	
Total	n - 1	TSS		

В данном случае k – количество предикторов, n – количество наблюдений

Мы проверяем гипотезу о незначимости коэффициента детерминации, или о том, что модель на константу не хуже, чем модель с предикторами

$$H_0 : R^2 = 0 \text{ или } \beta_1 = \dots \beta_k = 0$$

$$H_1 : R^2 > 0 \text{ или } \exists j \in \{1, \dots, k\} \text{ такой, что } \beta_j \neq 0$$

При верной нулевой гипотезе $F_{obs} \sim F(df_1 = k, df_2 = n - k - 1)$

Проверяем гипотезу против односторонней, а именно, правосторонней альтернативы. Если p-value достаточно мал, делаем вывод в пользу значимого коэффициента детерминации.

Задание 1.

На данных по 44 городам построена модель, объясняющая 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. Восстановим пропуски в таблице:

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		

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

$$F = \frac{R^2/k}{(1 - R^2)/(n - k - 1)} = \frac{0.39/4}{0.61/39} \approx 6.248$$

Данная модель лучше, чем модель на константу (опираемся на малое значение p-value для соответствующей F-статистики)

Спецификация модели множественной линейной регрессии

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{1i} + \dots + \hat{\beta}_k x_{ki}$$

У предикторов появляется 2 субиндекса, в данном случае первый субиндекс обозначает номер предиктора, второй – номер наблюдения

Запишем ту же спецификацию в векторно-матричном виде:

$$\begin{pmatrix} y_1 \\ \dots \\ y_N \end{pmatrix} = \begin{pmatrix} 1 & x_{11} & \dots & x_{k1} \\ \dots & \dots & \dots & \dots \\ 1 & x_{1N} & \dots & x_{kN} \end{pmatrix} \times \begin{pmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \dots \\ \hat{\beta}_k \end{pmatrix} + \begin{pmatrix} \hat{\epsilon}_1 \\ \dots \\ \hat{\epsilon}_N \end{pmatrix}$$

$$\vec{y} = X\vec{\hat{\beta}} + \vec{\hat{\epsilon}}$$

Мы уже показывали, что в линейной регрессии вектор остатков ортогонален предикторам (столбцам матрицы X).

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

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

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

Задание 2. Ниже в таблице представлены значения переменных: X , Z , Y .

X	-2	2	1	-1	0
Z	2	0	0	-1	-1
Y	2	6	3	2	10

1. Без использования Python получите оценки коэффициентов в регрессии Y на X с помощью общей формулы получения оценок коэффициентов, подходящей как для парной, так и для множественной регрессии. Представьте промежуточные расчеты, выпишите полученный вектор оценок коэффициентов и запишите спецификацию модели, подставив эти оценки в уравнение
Для начала запишем матрицу X :

$$X = \begin{pmatrix} 1 & -2 \\ 1 & 2 \\ 1 & 1 \\ 1 & -1 \\ 1 & 0 \end{pmatrix}$$

$$X^T X = \begin{pmatrix} 5 & 0 \\ 0 & 10 \end{pmatrix}$$

$$(X^T X)^{-1} = \frac{1}{50} \times \begin{pmatrix} 10 & 0 \\ 0 & 5 \end{pmatrix} = \begin{pmatrix} 0.2 & 0 \\ 0 & 0.1 \end{pmatrix}$$

$$X^T Y = \begin{pmatrix} 1 & 1 & 1 & 1 & 1 \\ -2 & 2 & 1 & -1 & 0 \end{pmatrix} \times \begin{pmatrix} 2 \\ 6 \\ 3 \\ 2 \\ 10 \end{pmatrix} = \begin{pmatrix} 23 \\ 9 \end{pmatrix}$$

$$(X^T X)^{-1} X^T Y = \begin{pmatrix} 0.2 & 0 \\ 0 & 0.1 \end{pmatrix} \times \begin{pmatrix} 23 \\ 9 \end{pmatrix} = \begin{pmatrix} 4.6 \\ 0.9 \end{pmatrix}$$

2. Без использования Python получите оценки коэффициентов в регрессии Y на X и Z с помощью общей формулы получения оценок коэффициентов, подходящей как для парной, так и для множественной регрессии. Представьте промежуточные расчеты, выпишите полученный вектор оценок коэффициентов и запишите спецификацию модели, подставив эти оценки в уравнение

$$X = \begin{pmatrix} 1 & -2 & 2 \\ 1 & 2 & 0 \\ 1 & 1 & 0 \\ 1 & -1 & -1 \\ 1 & 0 & -1 \end{pmatrix}$$

$$X^T X = \begin{pmatrix} 5 & 0 & 0 \\ 0 & 10 & -3 \\ 0 & -3 & 6 \end{pmatrix}$$

$$\det(X^T X) = 5 \times 10 \times 6 - (-3) \times (-3) \times 5 = 255$$

$$(X^T X)^{-1} = \frac{1}{255} \times \begin{pmatrix} 51 & 0 & 0 \\ 0 & 30 & 15 \\ 0 & 15 & 50 \end{pmatrix}$$

$$X^TY = \begin{pmatrix} 1 & 1 & 1 & 1 & 1 \\ -2 & 2 & 1 & -1 & 0 \\ 2 & 0 & 0 & -1 & -1 \end{pmatrix} \times \begin{pmatrix} 2 \\ 6 \\ 3 \\ 2 \\ 10 \end{pmatrix} = \begin{pmatrix} 23 \\ 9 \\ -8 \end{pmatrix}$$

$$(X^TX)^{-1}X^TY = \frac{1}{255} \times \begin{pmatrix} 51 & 0 & 0 \\ 0 & 30 & 15 \\ 0 & 15 & 50 \end{pmatrix} \times \begin{pmatrix} 23 \\ 9 \\ -8 \end{pmatrix} = \begin{pmatrix} 4.6 \\ 0.588 \\ -1.039 \end{pmatrix}$$