Множественная регрессия

Н.П. Алексеева,

СПбГУ, математико-механический факультет

2020 г.

Пример

j	Годы	Продолж-ть жизни ξ_1	Нац.богатсво (ард дол.) ξ_2	Водка (бут.в год) ξ_3
1	1970	68.9	2.372*	25.3
2	1975	68.1	2.372*	28
3	1980	67.6	2.489	30
4	1985	69.2	3.379	23.5
5	1990	69.2	4.130	18
6	1995	64.6	1.171	38.4
7	1998	67	0.689	29.6

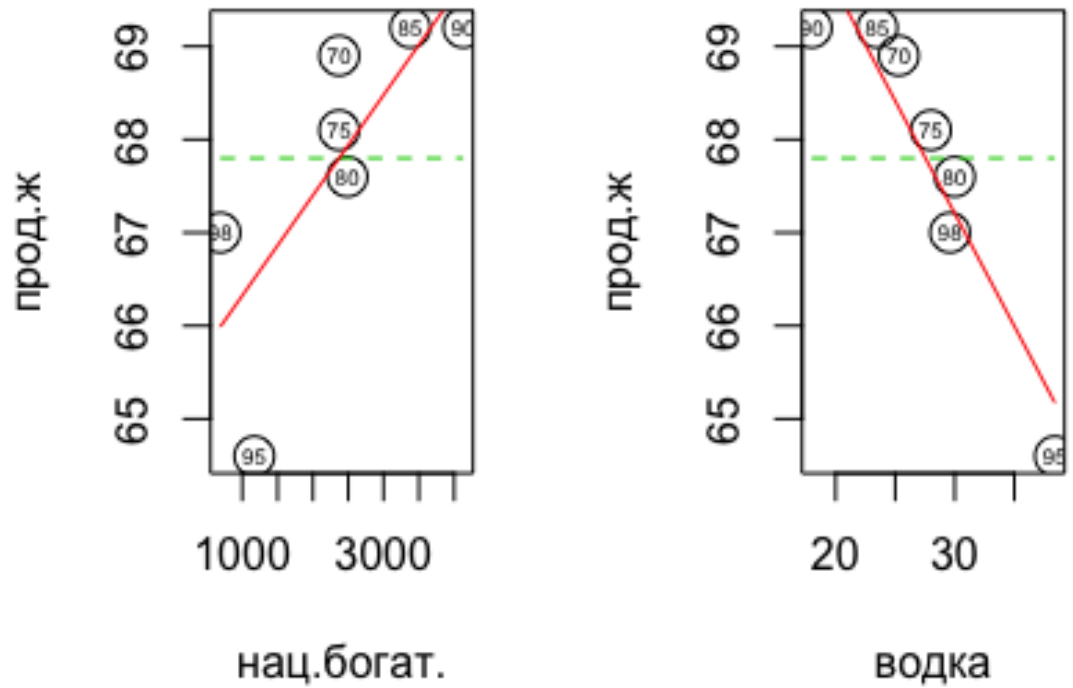


Рис.: $r_1 = 0.776, R_1^2 = 0.6, p = 0.04,$
$r_2 = -0.934, R_2^2 = 0.87, p = 0.002.$

Модель множественной регрессии

Множественной регрессией переменной Y в зависимости от значений переменных $X_1, \ldots, X_s$ называют условное среднее $E(Y|X_1, \ldots, X_s)$.
Модель множественной регрессии имеет вид уравнения

$$Y = E(Y|X_1, \ldots, X_s) + \varepsilon,$$

где ε – случайная несистематическая ошибка с нулевым математическим ожиданием и дисперсией σ^2.

Модель множественной линейной регрессии

$$Y = \beta_0 + \beta_1 X_1 + \ldots + \beta_s X_s + \varepsilon, \quad \varepsilon \sim \mathcal{N}(0, \sigma)$$

Пусть имеется выборка $\{y_j\}$, $j = 1, \ldots, n$ для зависимой переменной и многомерная выборка $\{x_{1j}, \ldots . x_{sj}\}$ независимых переменных.
Частные коэффициенты регрессии $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_s$ вычисляются при помощи специального метода из условия минимальности выражения

$$\sum_{j=1}^{n} (y_j - \beta_0 - \beta_1 x_{1j} - \ldots - \beta_s x_{sj})^2 \to \min$$

$$\hat{y}_j = \hat{\beta}_0 + \hat{\beta}_1 x_{1j} + \ldots + \hat{\beta}_s x_{sj}$$

наилучшее предсказание переменной Y по $X_1, \ldots, X_s$

Пример оценки параметров

Вычисляем частные коэффициены регрессии

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	74.276	3.170	23.430	0.000
vodka	-0.238	0.082	-2.910	0.044
rich	0.028	0.437	0.064	0.952

Уравнение регрессии имеет вид
$Y = 74.276 + 0.02X_1 - 0.238X_2$, где Y — продолжительность жизни, X_1 объем национального богатства в ардах долларов, X_2 число бутылок водки, продаваемых за год на душу населения.

Коэффициент детерминации

Источником вариации переменной Y является

$$SS_T = \sum_{j=1}^{n} (y_j - \bar{y})^2,$$

Можно показать, $SS_T = SS_R + SS_E$, где

$$SS_R = \sum_{j=1}^{n} (\hat{y}_j - \bar{y})^2, \quad SS_E = \sum_{j=1}^{n} (y_j - \hat{y}_j)^2$$

– источники вариации, обусловленные соответственно регрессией и ошибкой.

Коэффициентом детерминации называется доля дисперсии, объясняемая регрессией, равная отношению $R^2 = \frac{SS_R}{SS_T}$.

Чем ближе коэффициент детерминации к единице, тем ближе точки наблюдений к наилучшему линейному предсказанию, тем значимее прогноз.

В примере коэффициент детерминации равен $R^2 = 0.87$.

Множественный коэффициент корреляции

- Корень из коэффициента детерминации называется множественным коэффициентом корреляции.
- Множественный коэффициент корреляции равен коэффициенту корреляции между переменной Y и ее наилучшим предсказанием $\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \ldots + \hat{\beta}_s X_s$.
- При $s = 1$ множественный коэффициент корреляции равен обычному коэффициенту корреляции.

Проверка значимости коэффициентов регрессии

- Y продолжительность жизни
- X_1 национальное богатство
- X_2 объемы продаваемой водки

```
Y <- c(68.9, 68.1, 67.6, 69.2, 69.2, 64.6, 67.0)
X1 <- c(2.372, 2.372, 2.489, 3.379, 4.130, 1.171, 0.689)
X2 <- c(25.3, 28.0, 30.0, 23.5, 18.0, 38.4, 29.6)
LM <- lm(Y ~ X1 + X2)
coef(summary(LM))
```

	Estimate	Std..Error	t.value	Pr.. > t.
(Intercept)	74.28	3.17	23.43	0.00
X_2	0.03	0.44	0.06	0.95
X_3	-0.24	0.08	-2.91	0.04

```
summary(LM)$r.squared
```

0.8725756

```
predict(LM)
```

1	2	3	4	5	6	7
68.33	67.69	67.22	68.79	70.12	65.19	67.26

	Y	X_1	X_2	$\hat{Y}$	$\tilde{Y} + \bar{Y}$
1	68.90	2.37	25.30	68.33	68.37
2	68.10	2.37	28.00	67.69	68.21
3	67.60	2.49	30.00	67.22	68.18
4	69.20	3.38	23.50	68.79	68.21
5	69.20	4.13	18.00	70.12	66.88
6	64.60	1.17	38.40	65.19	67.21
7	67.00	0.69	29.60	67.26	67.54

- Y зависимая переменная
- X_1, X_2 независимые переменные
- $\hat{Y}$ наилучшее линейное предсказание
- $\tilde{Y} = Y - \hat{Y}$ остаток
- $\bar{Y}$ среднее значение

Проверка значимости прогноза

- Значимость коэффициента детерминации, т.е. гипотеза равенства нулю R^2 проверяется при помощи статистики Фишера вида:

$$F = \frac{SS_R/s}{SS_E/(n-s-1)},$$

которая при справедливости $H_0 : R^2 = 0$ имеет распределение Фишера с s и $n - s - 1$ степенью свободы, s — число независимых переменных.

- Если доверительный уровень вероятности $P = P\{F > F_*\} < 0.05$, то гипотеза о равенстве нулю R^2 отвергается, и прогнозу можно доверять с вероятностью, большей 0.95.
- Доверительный уровень вероятности $P = P\{F > F_*\}$ можно вычислить в Excel при помощи функции FPACП. Например, $F_* = 13.7, s = 2, n = 7$, $P = \text{FPACП}(13.7; 2; 4) = 0.01623 < 0.05$, поэтому можно говорить о значимости прогноза.

Частные корреляции: остаток и его свойства

Определение остатка $\tilde{Y}$

Разность $\tilde{Y} = Y - \hat{Y}$ между признаком Y и его наилучшим линейным предсказанием $\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \ldots + \hat{\beta}_s X_s$ по переменным $X_1, \ldots, X_s$ называется остатком.

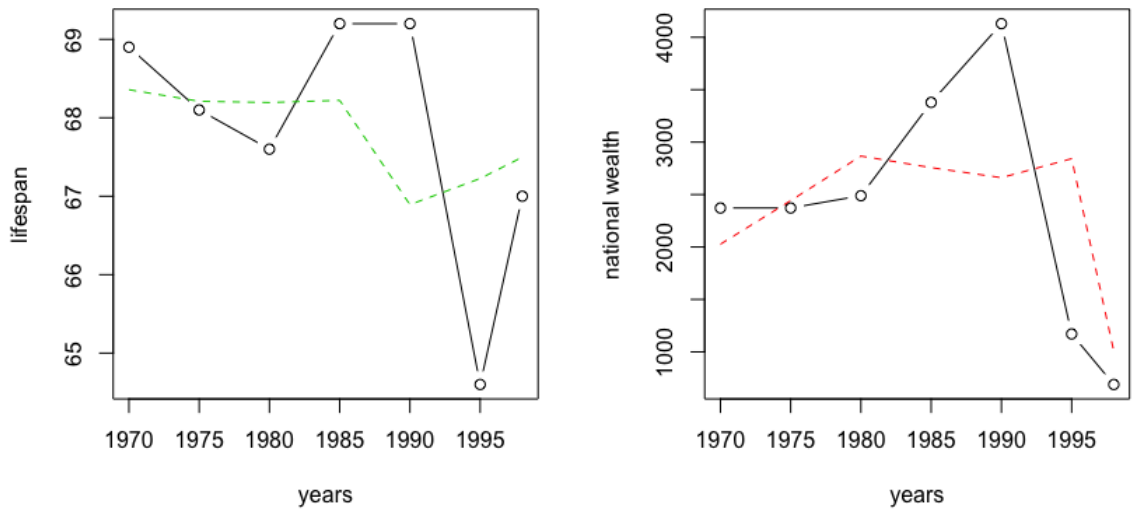
Свойства остатка $\tilde{Y}$

Нулевое математическое ожидание и некоррелированность ни с одной из переменных $X_1, \ldots, X_s$, входящих в предиктор $\hat{Y}(X_1, \ldots, X_s)$, то есть $\mathbb{E}\tilde{Y} = 0$, $\mathbb{E}\tilde{Y}(X_i - \mathbb{E}X_i) = 0$.

Инструмент для устранения влияния одних переменных на другие.

Частные корреляции

Устраним влияние переменных $X_2, \ldots, X_s$ на Y и X_1.
Рассмотрим остатки $Y - \hat{Y}(X_2, \ldots, X_s)$ и $X_1 - \hat{X}_1(X_2, \ldots, X_s)$.



Проверка значимости частных коэффициентов корреляции

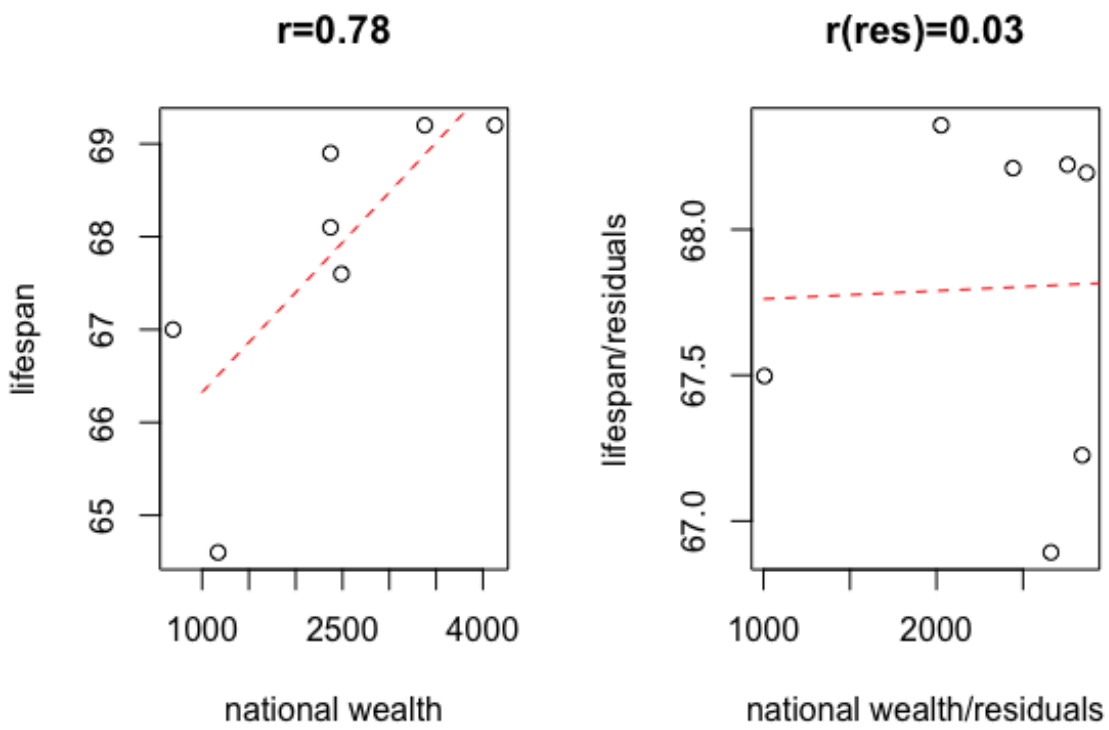
Значимость частного коэффициента корреляции проверяется также как и в случае обычного коэффициента корреляции при помощи статистики Стьюдента вида:

$$T = \frac{\rho_{yx_1|x_2 \ldots x_s}}{\sqrt{1 - \rho^2_{yx_1|x_2 \ldots x_s}}} \sqrt{n - s - 1}, \qquad (1)$$

которая в случае $H_0 : \varrho = 0$ имеет распределение Стьюдента с числом степеней свободы, равным $n - s - 1$.

Пусть T_* значение статистики T. Доверительный уровень вероятности считается как $p = P\{|T| > |T_*|\}$.

Коэффициент корреляции $\rho_{yx_1|x_2 \ldots x_s}$ между остатками $Y - \hat{Y}(X_2, \ldots, X_s)$ и $X_1 - \hat{X}_1(X_2, \ldots, X_s)$ называется частным коэффициентом корреляции.



В случае трех переменных X_1, X_2, X_3 для вычисления частных коэффициентов корреляции $r_{ij|k}$ по коэффициентам корреляции $r_{ij} = \varrho(X_i, X_j)$ можно использовать непосредственно формулы

$$r_{12|3} = \frac{r_{12} - r_{13}r_{23}}{\sqrt{1 - r_{13}^2}\sqrt{1 - r_{23}^2}}, \quad r_{13|2} = \frac{r_{13} - r_{12}r_{23}}{\sqrt{1 - r_{12}^2}\sqrt{1 - r_{23}^2}}.$$

- X_1 продолжительность жизни
- X_2 национальное богатство
- X_3 объемы продаваемой водки

r_{ij}	X_1	X_2	X_3
X_1	1.000	0.776	-0.934
X_2	0.776	1.000	-0.824
X_3	-0.934	-0.824	1.000

$$r_{12|3} = \frac{r_{12} - r_{13}r_{23}}{\sqrt{1 - r_{13}^2}\sqrt{1 - r_{23}^2}} = 0.032, \quad p = 0.95$$

$$r_{13|2} = \frac{r_{13} - r_{12}r_{23}}{\sqrt{1 - r_{12}^2}\sqrt{1 - r_{23}^2}} = -0.824, \quad p = 0.04.$$

- X_1 – детская смертность в городах США,
- X_2 процент афроамериканцев,
- X_3 – уровень безработицы.

$r_{12} = 0.84$ $(p = 2.45 \cdot 10^{-21})$, $r_{13} = 0.32$ $(p = 0.004)$,
$r_{23} = 0.37$ $(p = 0.001)$.

Не может ли одна зависимость являться следствием другой?
Вычисляем частные коэффициенты корреляции

$r_{12|3} = 0.82$ $(p < 0.0001)$, $r_{13|2} = 0.03$ $(p = 0.808)$.

Высокая детская смертность среди афроамериканцев не зависит от того, высокая или низкая безработица в городе, а наблюдаемая корреляция между безработицей и детской смертностью, скорее всего обусловлена тем, что высокая безработца имеет место в городах с высоким процентом афроамериканцев и соответственно с более высоким уровенем детской смертности.