

Модель линейной множественной регрессии

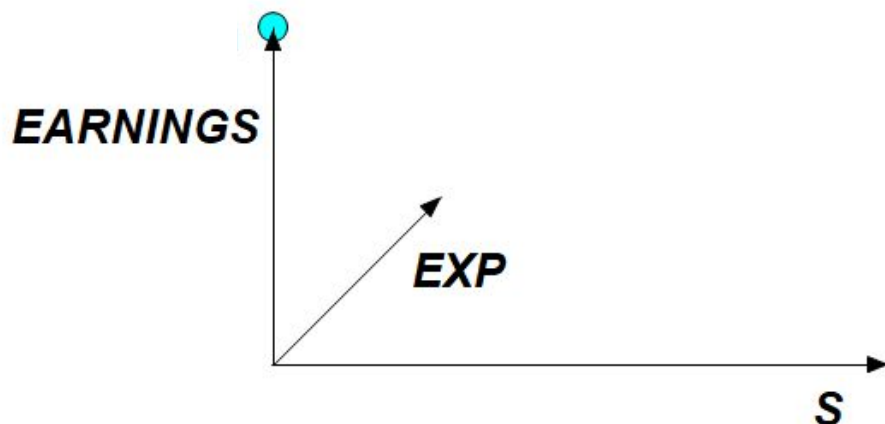
Модель линейной множественной регрессии

$$\text{EARNING} = \beta_0 + \beta_1 S + \beta_2 \text{EXP} + \varepsilon$$

Earnings – размер почасовой заработной платы, долл.

S – длительность обучения, лет

EXP – опыт работы, лет.



Пример регрессии с двумя объясняющими переменными.

Модель линейной множественной регрессии

$$\text{EARNING} = \beta_0 + \beta_1 S + \beta_2 \text{EXP} + \varepsilon$$

Почасовая заработная плата зависит от длительности обучения S и опыта работы EXP .

Формально, почасовая з/п не имеющих образования и опыта работы равна β_0

Для индивидов с одинаковым опытом работы при увеличении длительности обучения S на 1 год почасовая з/п EARNINGS увеличивается на β_1 долларов.

Аналогично, для индивидов с одинаковым образованием при увеличении опыта работы S на 1 год почасовая з/п EARNINGS увеличивается на β_2 долларов.

Модель линейной множественной регрессии

```
. reg EARNINGS S EXP
```

Source	SS	df	MS	Number of obs = 540		
Model	22513.6473	2	11256.8237	F(2, 537)	=	67.54
Residual	89496.5838	537	166.660305	Prob > F	=	0.0000
Total	112010.231	539	207.811189	R-squared	=	0.2010
				Adj R-squared	=	0.1980
				Root MSE	=	12.91

EARNINGS	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
S	2.678125	.2336497	11.46	0.000	2.219146	3.137105
EXP	.5624326	.1285136	4.38	0.000	.3099816	.8148837
_cons	-26.48501	4.27251	-6.20	0.000	-34.87789	-18.09213

$$EARNINGS = -26.49 + 2.68S + 0.56EXP$$

Каждый дополнительный год обучения увеличивает почасовую з/п на \$2.68, а каждый дополнительный год трудового стажа на \$0.56.

Модель линейной множественной регрессии

- Процедура проверки гипотезы о равенстве коэффициента регрессии конкретному числу (частный случай – проверка значимости коэффициента регрессии) для модели линейной **множественной** регрессии **аналогична** проверке гипотезы для модели **парной** линейной регрессии.

Модель линейной множественной регрессии

```
. reg EARNINGS S EXP
```

Source	SS	df	MS	Number of obs = 540		
Model	22513.6473	2	11256.8237	F(2, 537)	=	67.54
Residual	89496.5838	537	166.660305	Prob > F	=	0.0000
Total	112010.231	539	207.811189	R-squared	=	0.2010
				Adj R-squared	=	0.1980
				Root MSE	=	12.91

EARNINGS	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
S	2.678125	.2336497	11.46	0.000	2.219146	3.137105
EXP	.5624326	.1285136	4.38	0.000	.3099816	.8148837
_cons	-26.48501	4.27251	-6.20	0.000	-34.87789	-18.09213

$$EARNINGS = -26.49 + 2.68S + 0.56EXP$$

Каждый дополнительный год обучения увеличивает почасовую з/п на \$2.68, а каждый дополнительный год трудового стажа на \$0.56.

Модель линейной множественной регрессии

Общий вид модели линейной множественной

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \dots + \beta_k X_{ki} + \varepsilon_i, i = 1 \dots n$$

$X_1, \dots, X_k$ – факторы (независимые переменные);

Y – зависимая переменная;

ε_i – возмущения;

n – число наблюдений.

Выборочное регрессионное уравнение:

$$Y_i = \hat{\beta}_0 + \hat{\beta}_1 X_{1i} + \hat{\beta}_2 X_{2i} + \dots + \hat{\beta}_k X_{ki} + e_i, i = 1 \dots n$$

e_i – остатки

Модель линейной множественной регрессии

Рассмотрим двухфакторную регрессионную модель

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \varepsilon_i \quad i = 1 \dots n$$

$$Y_i = \hat{\beta}_0 + \hat{\beta}_1 X_{1i} + \hat{\beta}_2 X_{2i} + e_i, i = 1 \dots n$$

$$\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2$$

Для нахождения оценок

используется метод наименьших квадратов (МНК).

Модель линейной множественной регрессии

Выражение для i-го остатка:

$$e_i = Y_i - \hat{Y}_i = Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_{1i} - \hat{\beta}_2 X_{2i}$$

Минимизируем сумму квадратов

$$RSS = \sum_{i=1}^n e_i^2 = \sum_{i=1}^n (Y_i - \hat{Y}_i)^2 = \sum_{i=1}^n (Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_{1i} - \hat{\beta}_2 X_{2i})^2 \rightarrow \min$$

Приравниваем к 0 частные производные по всем переменным:

$$\frac{\partial RSS}{\partial \hat{\beta}_0} = 0, \quad \frac{\partial RSS}{\partial \hat{\beta}_1} = 0, \quad \frac{\partial RSS}{\partial \hat{\beta}_2} = 0.$$

Модель линейной множественной регрессии

Получаем систему уравнений:

$$\begin{cases} \sum Y_i = n \cdot \hat{\beta}_0 + \hat{\beta}_1 \cdot \sum X_{1i} + \hat{\beta}_2 \cdot \sum X_{2i}, \\ \sum Y_i \cdot X_{1i} = \hat{\beta}_0 \cdot \sum X_{1i} + \hat{\beta}_1 \cdot \sum X_{1i}^2 + \hat{\beta}_2 \cdot \sum X_{1i} \cdot X_{2i}, \\ \sum Y_i \cdot X_{2i} = \hat{\beta}_0 \cdot \sum X_{2i} + \hat{\beta}_1 \cdot \sum X_{1i} \cdot X_{2i} + \hat{\beta}_2 \cdot \sum X_{2i}^2. \end{cases}$$

Решением данной системы будут значения оценок коэф($\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2$) тов

Модель линейной

множественной регрессии

В общем виде для модели

$$Y_i = \hat{\beta}_0 + \hat{\beta}_1 X_{1i} + \hat{\beta}_2 X_{2i} + \dots \hat{\beta}_k X_{ki} + e_i, i = 1 \dots n$$

система уравнений для нахождения оценок коэффициентов регрессии

ИМЕЕТ ВИД:

[illegible]

Модель линейной множественной регрессии в векторном виде

Оцененное уравнение регрессии

$$Y_i = \hat{\beta}_0 + \hat{\beta}_1 X_{1i} + \hat{\beta}_2 X_{2i} + e_i, i = 1 \dots n$$

n – число наблюдений

Обозначим :

$$Y = \begin{pmatrix} Y_1 \\ Y_2 \\ \dots \\ Y_n \end{pmatrix}, \quad X_1 = \begin{pmatrix} X_{11} \\ X_{12} \\ \dots \\ X_{1n} \end{pmatrix}, \dots, X_k = \begin{pmatrix} X_{k1} \\ X_{k2} \\ \dots \\ X_{kn} \end{pmatrix}, e = \begin{pmatrix} e_1 \\ e_2 \\ \dots \\ e_n \end{pmatrix}$$

Получаем уравнение регрессии в векторном виде :

$$Y = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \dots \hat{\beta}_k X_k + e$$

Модель линейной множественной регрессии в матричном виде

Введем матрицу наблюдений X размера $(n \times k)$ и вектор оценок коэффициентов $\hat{\beta}$ размера $(k \times 1)$

$$X = \begin{pmatrix} 1 & X_{11} & \dots & X_{k1} \\ 1 & X_{12} & \dots & X_{k2} \\ \dots & \dots & \dots & \dots \\ 1 & X_{1n} & \dots & X_{kn} \end{pmatrix} \quad \hat{\beta} = \begin{pmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \dots \\ \hat{\beta}_k \end{pmatrix}$$

Тогда уравнение регрессии можно переписать в матричном виде:

$$Y = X\hat{\beta} + e$$

Модель линейной множественной регрессии в матричном виде

$$Y = X\hat{\beta} + e \quad (1)$$

Сумма квадратов отклонений МНК для уравнения (1) записывается следующим образом:

$$RSS = \sum_{i=1}^n e_i^2 = (e_1 \ e_2 \ \dots \ e_n) \cdot \begin{pmatrix} e_1 \\ e_2 \\ \dots \\ e_n \end{pmatrix} = e^T \cdot e = (Y - X\hat{\beta})^T \cdot (Y - X\hat{\beta}) \rightarrow \min$$

Модель линейной множественной регрессии в матричном виде

Выполним умножение:

$$(Y - X\hat{\beta})^T \cdot (Y - X\hat{\beta})$$

Сначала вспомним свойства операции транспонирования матриц:

A, B – матрицы,
 α – действительное число

$$1. (A^T)^T = A$$

$$2. (\alpha \cdot A)^T = \alpha \cdot (A^T)$$

$$3. (A + B)^T = A^T + B^T$$

$$4. (A \cdot B)^T = B^T \cdot A^T$$

Модель линейной множественной регрессии в матричном виде

$$\begin{aligned}(Y - X\hat{\beta})^T \cdot (Y - X\hat{\beta}) &= Y^T \cdot Y - (X\hat{\beta})^T Y - Y^T X\hat{\beta} + (X\hat{\beta})^T X\hat{\beta} = \\&= Y^T Y - \hat{\beta}^T X^T Y - \hat{\beta}^T X^T Y + \hat{\beta}^T X^T X\hat{\beta} = \\&= Y^T Y - 2\hat{\beta}^T X^T Y + \hat{\beta}^T X^T X\hat{\beta}\end{aligned}$$

Продифференцируем по вектору β результат перемножения:

$$\frac{\partial RSS}{\partial \hat{\beta}} = (Y^T Y - 2\hat{\beta}^T X^T Y + \hat{\beta}^T X^T X\hat{\beta})'_{\hat{\beta}} = -2X^T Y + 2X^T X\hat{\beta}$$

Приравняем полученное выражение 0:

$$-2X^T Y + 2X^T X\hat{\beta} = 0 \Leftrightarrow X^T Y = (X^T X)\hat{\beta} \Leftrightarrow \hat{\beta} = (X^T X)^{-1} X^T Y$$

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

Пример

Имеются данные по 10 предприятиям о прибыли (Y, млн.руб.), выработке продукции на 1 работника (X1, единиц) и доле продукции, производимой на экспорт (X2,%).

Найдем оценки параметров регрессионного уравнения

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$$

№ п/п	Y	X1	X2
1	2	11	3
2	1	10	2
3	3	12	4
4	8	18	10
5	7	15	11
6	5	13	6
7	4	13	5
8	6	15	7
9	7	16	10
10	7	17	12

Пример

$$Y = \begin{pmatrix} 2 \\ 1 \\ 3 \\ 8 \\ 7 \\ 5 \\ 4 \\ 6 \\ 7 \\ 7 \\ 7 \end{pmatrix}, \quad X = \begin{pmatrix} 1 & 11 & 3 \\ 1 & 10 & 2 \\ 1 & 12 & 4 \\ 1 & 18 & 10 \\ 1 & 15 & 11 \\ 1 & 13 & 6 \\ 1 & 13 & 5 \\ 1 & 15 & 7 \\ 1 & 16 & 10 \\ 1 & 17 & 12 \end{pmatrix}$$

Пример

$$X^T X = \begin{pmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 11 & 10 & 12 & 18 & 15 & 13 & 13 & 15 & 16 & 17 \\ 3 & 2 & 4 & 10 & 11 & 6 & 5 & 7 & 10 & 12 \end{pmatrix} \cdot \begin{pmatrix} 1 & 11 & 3 \\ 1 & 10 & 2 \\ 1 & 12 & 4 \\ 1 & 18 & 10 \\ 1 & 15 & 11 \\ 1 & 13 & 6 \\ 1 & 13 & 5 \\ 1 & 15 & 7 \\ 1 & 16 & 10 \\ 1 & 17 & 12 \end{pmatrix} = \begin{pmatrix} 10 & 140 & 70 \\ 140 & 2022 & 1058 \\ 70 & 1058 & 604 \end{pmatrix}$$

$$(X^T X)^{-1} = \frac{1}{9840} \cdot \begin{pmatrix} 101924 & -10500 & 6580 \\ -1050 & 1140 & -780 \\ 6580 & -780 & 620 \end{pmatrix}$$

Пример

$$X^T Y = \begin{pmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 11 & 10 & 12 & 18 & 15 & 13 & 13 & 15 & 16 & 17 \\ 3 & 2 & 4 & 10 & 11 & 6 & 5 & 7 & 10 & 12 \end{pmatrix} \cdot \begin{pmatrix} 2 \\ 1 \\ 3 \\ 8 \\ 7 \\ 5 \\ 4 \\ 6 \\ 7 \\ 7 \end{pmatrix} = \begin{pmatrix} 50 \\ 755 \\ 423 \end{pmatrix}$$

Пример

$$\hat{\beta} = (X^T X)^{-1} X^T Y = \frac{1}{9840} \begin{pmatrix} -47960 \\ 5760 \\ 2360 \end{pmatrix} = \begin{pmatrix} -4,874 \\ 0,585 \\ 0,240 \end{pmatrix}$$

$$\hat{Y} = -4,8474 + 0,585 \cdot X_1 + 0,240 \cdot X_2$$

Теорема Гаусса- Маркова для случая линейной множественной регрессии

Если модель множественной линейной регрессии

$$Y = \beta_0 + \beta_1 X_1 + \dots \beta_k X_k + \varepsilon$$

- 1) Правильно специфицирована
- 2) Не существует линейной связи между объясняющими переменными
- 3) Возмущения имеют нулевое мат. ожидание $E(\varepsilon_i) = 0$,
- 4) Дисперсии возмущений одинаковы $D(\varepsilon_i) = \sigma_\varepsilon^2$, $j = 1, \dots, n$
- 5) Возмущения с разными номерами не коррелируют
 $\text{cov}(\varepsilon_i, \varepsilon_j) = 0, i \neq j$

Тогда оценки МНК являются BLUE (Best Linear Unbiased Estimations)
(Лучшими линейными несмещенными оценками)

Дисперсионный анализ

$$\sum (Y - \bar{Y})^2 = \sum (\hat{Y} - \bar{Y})^2 + \sum e^2$$

$$TSS = ESS + RSS$$

TSS (Total Sum of Squares) - общая сумма квадратов отклонений зависимой переменной от среднего значения.

ESS (Explained Sum of Squares) – объясненная с помощью регрессии сумма квадратов отклонений.

RSS (Residual Sum of Squares) – сумма квадратов остатков.

Коэффициент множественной детерминации R^2

$$\sum (Y - \bar{Y})^2 = \sum (\hat{Y} - \bar{Y})^2 + \sum e^2$$

$$TSS = ESS + RSS$$

$$R^2 = \frac{ESS}{TSS} = \frac{\sum (\hat{Y}_i - \bar{Y})^2}{\sum (Y_i - \bar{Y})^2} = 1 - \frac{\sum e_i^2}{\sum (Y_i - \bar{Y})^2}$$

Показатель качества подгонки регрессии R^2 – коэффициент множественной детерминации, равен отношению ESS к TSS . Очевидно, эта величина неотрицательна.

Коэффициент множественной детерминации R^2

$$\sum (Y - \bar{Y})^2 = \sum (\hat{Y} - \bar{Y})^2 + \sum e^2$$

$$TSS = ESS + RSS$$

$$R^2 = \frac{ESS}{TSS} = \frac{\sum (\hat{Y}_i - \bar{Y})^2}{\sum (Y_i - \bar{Y})^2} = 1 - \frac{\sum e_i^2}{\sum (Y_i - \bar{Y})^2}$$

Показатель качества подгонки регрессии R^2 – коэффициент множественной детерминации, равен отношению ESS к TSS . Очевидно, эта величина неотрицательна.

Коэффициент множественной детерминации R^2

$$\sum (Y - \bar{Y})^2 = \sum (\hat{Y} - \bar{Y})^2 + \sum e^2$$

$$TSS = ESS + RSS$$

$$R^2 = \frac{ESS}{TSS} = \frac{\sum (\hat{Y}_i - \bar{Y})^2}{\sum (Y_i - \bar{Y})^2} = 1 - \frac{\sum e_i^2}{\sum (Y_i - \bar{Y})^2}$$

Из последней формулы очевидно, что R^2 не превышает 1.

Таким образом, $0 \leq R^2 \leq 1$.

Коэффициент множественной детерминации R^2

$$\sum (Y - \bar{Y})^2 = \sum (\hat{Y} - \bar{Y})^2 + \sum e^2$$

$$TSS = ESS + RSS$$

$$R^2 = \frac{ESS}{TSS} = \frac{\sum (\hat{Y}_i - \bar{Y})^2}{\sum (Y_i - \bar{Y})^2} = 1 - \frac{\sum e_i^2}{\sum (Y_i - \bar{Y})^2}$$

Очевидно, что, минимизируя сумму квадратов остатков, мы максимизируем R^2 . Таким образом, чем ближе R^2 к 1, тем выше качество подгонки регрессии.

Коэффициент множественной детерминации R^2

$$R^2 = \frac{ESS}{TSS} = \frac{\sum (\hat{Y}_i - \bar{Y})^2}{\sum (Y_i - \bar{Y})^2} = 1 - \frac{\sum e_i^2}{\sum (Y_i - \bar{Y})^2}$$

Коэффициент детерминации R^2
показывает долю дисперсии
зависимой переменной,
«объясненной» уравнением регрессии

R^2 имеет существенный недостаток: R^2 возрастает при добавлении любого регрессора X (т.к. при этом уменьшается сумма квадратов остатков). Но это может вести к мультиколлинеарности (следующая лекция).

R_{adj}^2 (скорректированный R^2)

$$R_{adj}^2 = 1 - \frac{RSS / (n - k - 1)}{TSS / (n - 1)}$$

Для сравнения качества подгонки двух регрессий с одной и той же зависимой переменной Y используется коэффициент множественной детерминации, скорректированный на число степеней свободы R_{adj}^2 . R_{adj}^2 не превышает 1, но может быть меньше 0.

R_{adj}^2 - это R^2 , скорректированный с учетом числа степеней свободы. Он «наказывает» за включение лишних регрессоров.

Пример

reg	EARNINGS	S	SM	SF	ASVAB02	ASVAB03	ASVAB04	ASVAB05	SIBLINGS
Source	SS	df	MS	Number of obs = 540					
				F(8, 531) = 20.06					
Model	25995.7171	8	3249.46463	Prob > F = 0.0000					
Residual	86014.514	531	161.985902	R-squared = 0.2321					
				Adj R-squared = 0.2205					
Total	112010.231	539	207.811189	Root MSE = 12.727					

EARNINGS	Coef.	Std. Err.	t	P>t	[95% Conf.Interval]	
S	1.606076	.2842578	5.65	0.000	1.047668	2.164484
SM	-.0995345	.2606452	-0.38	0.703	-.6115567	.4124876
SF	.3080043	.2059354	1.50	0.135	-.0965438	.7125523
ASVAB02	.421419	.0870657	4.84	0.000	.2503835	.5924544
ASVAB03	-.1388054	.1106814	-1.25	0.210	-.3562325	.0786216
ASVAB04	-.075362	.101286	-0.74	0.457	-.2743324	.1236083
ASVAB05	.1509049	.0736544	2.05	0.041	.0062151	.2955947
SIBLINGS	.0784485	.2701161	0.29	0.772	-.4521788	.6090758
_cons	-23.31213	4.332361	-5.38	0.000	-31.8228	-14.80146

В оцененной регрессии исключим переменные с наименьшими t – value.

Пример

reg EARNINGS S SF ASVAB02 ASVAB03 ASVAB05

Source	SS	df	MS	Number of obs	= 540
Model	25861.8812	5	5172.37623	F(5, 534)	= 32.06
Residual	86148.3499	534	161.326498	Prob > F	= 0.0000
				R-squared	= 0.2309
				Adj R-squared	= 0.2237
Total	112010.231	539	207.811189	Root MSE	= 12.701

EARNINGS	Coef.	Std. Err.	t	P>t	[95% Conf.Interval]	
S	1.580369	.2819491	5.61	0.000	1.026504	2.134235
SF	.2573391	.1739785	1.48	0.140	-.0844271	.5991053
ASVAB02	.4108779	.0857274	4.79	0.000	.2424736	.5792821
ASVAB03	-.1932668	.0900419	-2.15	0.032	-.3701465	-.016387
ASVAB05	.1394809	.0712416	1.96	0.051	-.0004672	.279429
_cons	-23.23215	3.745907	-6.20	0.000	-30.59067	-15.87363

R^2 уменьшился, R^2_{adj} увеличился. Вторая модель лучше соответствует данным.

Есть соблазн свести выбор уравнения к задаче максимизации R^2 или R^2_{adj} . Не стоит этого делать.

1. Высокий R^2 (или R^2_{adj}) **говорит** о том, что регрессоры предсказывают большую долю изменений y .
2. Высокий R^2 (или R^2_{adj}) **не говорит** о том, что вы верно выявили причинно-следственную связь между переменными
3. Высокий R^2 (или R^2_{adj}) **не гарантирует** отсутствия смещения оценок из-за некорректной спецификации

Стандартная ошибка регрессии

(Standard Error of Estimate — не путать с ESS)

$$SEE = \sqrt{S^2} = \sqrt{\frac{\sum e_i^2}{n - k - 1}}$$

Измеряет среднюю величину ошибки модели

Используется для оценки качества подгонки модели: чем меньше SEE, тем точнее модель

Можно использовать для сравнения нескольких однотипных уравнений регрессии

Тестирование гипотез

1. Тестирование значимости коэффициента
2. Тестирование гипотезы $\beta_j = A$
3. Доверительный интервал для коэффициента
4. Тестирование значимости уравнения
5. Сравнение «короткой» и «длинной» регрессий

Тестирование значимости коэффициента регрессии

P – VALUE для проверки гипотезы о значимости коэффициента

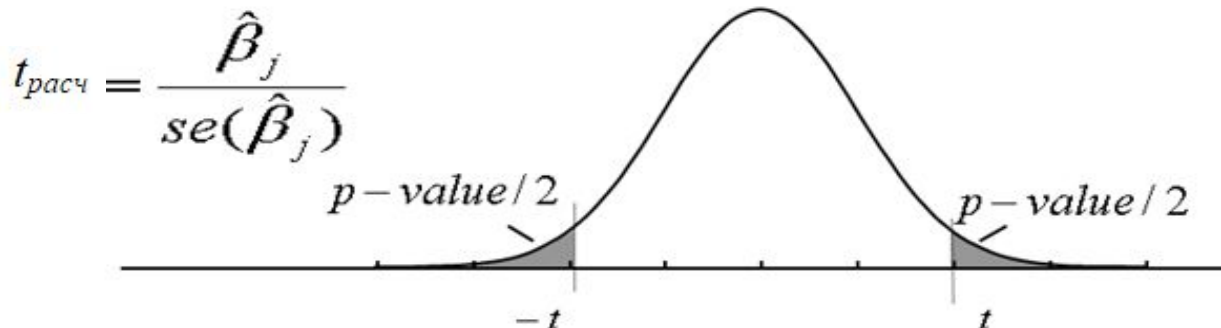
$$Y_i = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots \beta_k X_k + \varepsilon$$

$$H_0 : \beta_j = 0$$

$$H_1 : \beta_j \neq 0$$

Если $|t_{расч}| < t_{кр}$,

то гипотеза H_0 не отклоняется (принимается)



P – value – минимальный уровень значимости, при котором основная гипотеза не отвергается. Если P-value для коэффициента регрессии меньше выбранного уровня значимости α , то нулевая гипотеза отвергается и выбранный коэффициент является значимым (при выбранном уровне значимости).

Тестирование гипотезы о равенстве коэффициента конкретному значению

Тестирование гипотезы $H_0: \beta_j = A$

$$t_{\text{расч}} = \frac{\hat{\beta}_j - A}{se(\hat{\beta}_j)}$$

Доверительный интервал для коэффициента регрессии

$$(\hat{\beta}_j - t_{n-k-1} * se(\hat{\beta}_j), \hat{\beta}_j + t_{n-k-1} * se(\hat{\beta}_j))$$

Проверка гипотезы о значимости (адекватности) уравнения регрессии

$$Y = \beta_0 + \beta_1 X_1 + \dots \beta_k X_k + \varepsilon$$

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

$$H_1 : \exists \beta_i \neq 0$$

$$F_{расч} = F(k, n - k - 1) = \frac{ESS / k}{RSS / (n - k - 1)}$$

n – число наблюдений;

k – число факторов.

Гипотезу H_0 можно переформулировать следующим образом:
выбранный набор независимых переменных не оказывает
влияния на переменную Y .

Проверка гипотезы об адекватности регрессии

Формулу для F – статистики можно преобразовать, выразив эту статистику через R^2 .

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

n – число наблюдений;

k – число факторов.

Проверка гипотезы об адекватности регрессии

Сравниваем расчетное и критическое значение F-статистик

Если $F_{расч} < F_{кр}$,

то гипотеза H_0 не отклоняется (принимается),

то есть мы делаем вывод о том, что все переменные $x_1, \dots, x_k$ не оказывает значимого влияния на переменную y

В этом случае уравнение называют незначимым.

В противном случае гипотеза H_0 не принимается (отклоняется).

2 Таблица значений F -критерия Фишера при уровне значимости $\alpha = 0,05$

$n-m-1$	Число независимых факторов									
	1	2	3	4	5	6	8	12	24	∞
1	161,45	199,50	215,72	224,57	230,17	233,97	238,89	243,91	249,04	254,32
2	18,51	19,00	19,16	19,25	19,30	19,33	19,37	19,41	19,45	19,50
3	10,13	9,55	9,28	9,12	9,01	8,94	8,84	8,74	8,64	8,53
4	7,71	6,94	6,59	6,39	6,26	6,16	6,04	5,91	5,77	5,63
5	6,61	5,79	5,41	5,19	5,05	4,95	4,82	4,68	4,53	4,36
6	5,99	5,14	4,76	4,53	4,39	4,28	4,15	4,00	3,84	3,67
7	5,59	4,74	4,35	4,12	3,97	3,87	3,73	3,57	3,41	3,23
8	5,32	4,46	4,07	3,84	3,69	3,58	3,44	3,28	3,12	2,93
9	5,12	4,26	3,86	3,63	3,48	3,37	3,23	3,07	2,90	2,71
10	4,96	4,10	3,71	3,48	3,33	3,22	3,07	2,91	2,74	2,54
11	4,84	3,98	3,59	3,36	3,20	3,09	2,95	2,79	2,61	2,40
12	4,75	3,88	3,49	3,26	3,11	3,00	2,85	2,69	2,50	2,30
13	4,67	3,80	3,41	3,18	3,02	2,92	2,77	2,60	2,42	2,21
14	4,60	3,74	3,34	3,11	2,96	2,85	2,70	2,53	2,35	2,13
15	4,54	3,68	3,29	3,06	2,90	2,79	2,64	2,48	2,29	2,07
16	4,49	3,63	3,24	3,01	2,85	2,74	2,59	2,42	2,24	2,01
17	4,45	3,59	3,20	2,96	2,81	2,70	2,55	2,38	2,19	1,96
18	4,41	3,55	3,16	2,93	2,77	2,66	2,51	2,34	2,15	1,92
19	4,38	3,52	3,13	2,90	2,74	2,63	2,48	2,31	2,11	1,88
20	4,35	3,49	3,10	2,87	2,71	2,60	2,45	2,28	2,08	1,84
21	4,32	3,47	3,07	2,84	2,68	2,57	2,42	2,25	2,05	1,81
22	4,30	3,44	3,05	2,82	2,66	2,55	2,40	2,23	2,03	1,78
23	4,28	3,42	3,03	2,80	2,64	2,53	2,38	2,20	2,00	1,76
24	4,26	3,40	3,01	2,78	2,62	2,51	2,36	2,18	1,98	1,73
25	4,24	3,38	2,99	2,76	2,60	2,49	2,34	2,16	1,96	1,71
26	4,22	3,37	2,98	2,74	2,59	2,47	2,32	2,15	1,95	1,69
27	4,21	3,35	2,96	2,73	2,57	2,46	2,30	2,13	1,93	1,67
28	4,20	3,34	2,95	2,71	2,56	2,44	2,29	2,12	1,91	1,65
29	4,18	3,33	2,93	2,70	2,54	2,43	2,28	2,10	1,90	1,64
30	4,17	3,32	2,92	2,69	2,53	2,42	2,27	2,09	1,89	1,62
35	4,12	3,26	2,87	2,64	2,48	2,37	2,22	2,04	1,83	1,57
40	4,08	3,23	2,84	2,61	2,45	2,34	2,18	2,00	1,79	1,51
45	4,06	3,21	2,81	2,58	2,42	2,31	2,15	1,97	1,76	1,48
50	4,03	3,18	2,79	2,56	2,40	2,29	2,13	1,95	1,74	1,44
60	4,00	3,15	2,76	2,52	2,37	2,25	2,10	1,92	1,70	1,39
70	3,98	3,13	2,74	2,50	2,35	2,23	2,07	1,89	1,67	1,35
80	3,96	3,11	2,72	2,49	2,33	2,21	2,06	1,88	1,65	1,31
90	3,95	3,10	2,71	2,47	2,32	2,20	2,04	1,86	1,64	1,28
100	3,94	3,09	2,70	2,46	2,30	2,19	2,03	1,85	1,63	1,26
125	3,92	3,07	2,68	2,44	2,29	2,17	2,01	1,83	1,60	1,21
150	3,90	3,06	2,66	2,43	2,27	2,16	2,00	1,82	1,59	1,18
200	3,89	3,04	2,65	2,42	2,26	2,14	1,98	1,80	1,57	1,14
∞	3,84	2,99	2,60	2,37	2,21	2,09	1,94	1,75	1,52	1,00

Проверка гипотезы об адекватности регрессии

$$S = \beta_1 + \beta_2 ASVABC + \beta_3 SM + \beta_4 SF + \varepsilon$$

$$H_0: \beta_2 = \beta_3 = \beta_4 = 0$$

```
. reg S ASVABC SM SF
```

Source	SS	df	MS	Number of obs =	540
Model	1181.36981	3	393.789935	F(3, 536) =	104.30
Residual	2023.61353	536	3.77539837	Prob > F	= 0.0000
Total	3204.98333	539	5.94616574	R-squared	= 0.3686
				Adj R-squared	= 0.3651
				Root MSE	= 1.943

S	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
ASVABC	.1257087	.0098533	12.76	0.000	.1063528	.1450646
SM	.0492424	.0390901	1.26	0.208	-.027546	.1260309
SF	.1076825	.0309522	3.48	0.001	.04688	.1684851
_cons	5.370631	.4882155	11.00	0.000	4.41158	6.329681

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

$$F(3, 536) = \frac{1181/3}{2024/536} = 104.3$$

Значение F – статистики для проверки адекватности регрессии указывается в выдаваемой таблице.

$$F(3; 536) = \frac{0,3686/3}{(1-0,3686)/536} = 104,3$$

Проверка гипотезы об адекватности регрессии

Проверка гипотезы об адекватности регрессии

$$S = \beta_1 + \beta_2 ASVABC + \beta_3 SM + \beta_4 SF + \varepsilon$$

$$H_0 : \beta_2 = \beta_3 = \beta_4 = 0$$

```
. reg S ASVABC SM SF
```

Source	SS	df	MS	Number of obs =	540
Model	1181.36981	3	393.789935	F(3, 536) =	104.30
Residual	2023.61353	536	3.77539837	Prob > F	= 0.0000
Total	3204.98333	539	5.94616574	R-squared	= 0.3686
				Adj R-squared	= 0.3651
				Root MSE	= 1.943

S	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
ASVABC	.1257087	.0098533	12.76	0.000	.1063528 .1450646
SM	.0492424	.0390901	1.26	0.208	-.027546 .1260309
SF	.1076825	.0309522	3.48	0.001	.04688 .1684851
_cons	5.370631	.4882155	11.00	0.000	4.41158 6.329681

$$F_{\text{crit},0.1\%}(3,500) = 5.51 \quad F(3,536) = \frac{1181/3}{2024/536} = 104.3$$

Сравниваем рассчитанное значение F – статистики с критическим. При уровне значимости 0.1% критическое значение равно 5.51. Рассчитанное значение F – статистики превышает критическое, следовательно, нулевая гипотеза отвергается при заданном уровне значимости. Регрессия является адекватной.

Тестирование гипотезы о группе переменных

Рассмотрим регрессионную модель

(длинная регрессия, unrestricted model (UR)) $m + q = k$

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \dots + \beta_m X_{im} + \beta_{m+1} X_{i,m+1} + \dots + \beta_{m+q} X_{i,m+q} + \varepsilon_i$$

Тестируем гипотезу:

$$H_0 : \beta_{m+1} = \dots = \beta_{m+q} = 0 \quad H_1 : \text{хотя бы один из коэф-тов} \neq 0$$

Для тестирования гипотезы используется F – статистика

$$F_{\text{расч}} = \frac{(R_{UR}^2 - R_R^2) / q}{(1 - R_{UR}^2) / (n - k - 1)}$$

n – число наблюдений;
 k – число факторов;
 q – число ограничений

$$F_{\text{крит}} = F(q, n - k - 1) \quad (\text{тестируемых коэффициентов}).$$

Тестирование гипотезы о группе переменных

Длинная регрессия, unrestricted model (UR)

$$Y_i = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots \beta_m X_m + \boxed{\beta_{m+1} X_{m+1} + \dots + \beta_{m+q} X_{m+q}} + \varepsilon$$

Короткая регрессия, restricted model (R)

$$Y_i = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots \beta_m X_m + \varepsilon$$

Вытаскиваем из моделей R^2 : R_{UR}^2 и R_R^2

$$F_{расч} = \frac{(R_{UR}^2 - R_R^2) / q}{(1 - R_{UR}^2) / (n - k - 1)}$$

Если $F_{расч} < F_{кр}$,

то гипотеза H_0 не отклоняется (принимается)

Тестирование гипотезы о группе переменных

В учебниках для расчета $F_{расч}$ также можно встретить формулу:

$$F_{расч} = \frac{(RSS_R - RSS_{UR}) / q}{RSS_{UR} / (n - k - 1)}$$

Она даёт тот же результат, что и формула

$$F_{расч} = \frac{(R_{UR}^2 - R_R^2) / q}{(1 - R_{UR}^2) / (n - k - 1)}$$

Пример 1

- Unrestricted model (модель без ограничений)

```
> m1<-lm(data=df1, log(wage)~educ+exper+exper2)
> summary(m1)
```

Call:

```
lm(formula = log(wage) ~ educ + exper + exper2, data = df1)
```

Residuals:

	Min	1Q	Median	3Q	Max
	-1.96387	-0.29375	-0.04009	0.29497	1.30216

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	0.1279975	0.1059323	1.208	0.227
educ	0.0903658	0.0074680	12.100	< 2e-16 ***
exper	0.0410089	0.0051965	7.892	1.77e-14 ***
exper2	-0.0007136	0.0001158	-6.164	1.42e-09 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.4459 on 522 degrees of freedom

Multiple R-squared: 0.3003, Adjusted R-squared: 0.2963

F-statistic: 74.67 on 3 and 522 DF, p-value: < 2.2e-16

wage – почасовая заработная плата, долл.;

educ – количество лет обучения;

exper – опыт работы, лет;

exper2 – квадрат опыта работы, лет;

$$H_0 : \beta_{\text{exper}} = \beta_{\text{exper}^2} = 0$$

$$H_1 : \text{хотя бы один из коэф-тов} \neq 0$$

Пример 1

- Restricted model (модель с ограничениями)

```
> m2<-lm(data=df1, log(wage)~educ)
> summary(m2)
```

```
call:
lm(formula = log(wage) ~ educ, data = df1)
```

```
Residuals:
    Min       1Q   Median       3Q      Max
-2.21158 -0.36393 -0.07263  0.29712  1.52339
```

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.583773   0.097336   5.998 3.74e-09 ***
educ         0.082744   0.007567  10.935 < 2e-16 ***
```

```
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
Residual standard error: 0.4801 on 524 degrees of freedom
Multiple R-squared: 0.1858, Adjusted R-squared: 0.1843
F-statistic: 119.6 on 1 and 524 DF, p-value: < 2.2e-16
```

Пример 1

$$F_{расч} = \frac{(R^2_{UR} - R^2_R) / q}{(1 - R^2_{UR}) / (n - k - 1)} = \frac{(0,3003 - 0,1858) / 2}{(1 - 0,3003) / 522} = \frac{0,05725}{0,00134} = 42,72$$

$$F_{расч} > F_{крит}$$

H0 отвергается

Пример 1

- В R-studio используем команду `anova`

```
> anova(m1,m2)
Analysis of Variance Table

Model 1: log(wage) ~ educ + exper + exper2
Model 2: log(wage) ~ educ
  Res.Df    RSS Df Sum of Sq    F    Pr(>F)
1     522 103.79
2     524 120.77 -2    -16.979 42.696 < 2.2e-16 ***
---
signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Пример 2

Харис пытается понять, что лучше помогает решать задачи по эконометрике:

-поедание пирожков(штуки)

-посещение лекций(академические часы)

$$problems_t = \beta_1 + \beta_2 lecture_t + \beta_3 pie_t + u_t$$

$$H_0 : \beta_2 = \beta_3$$

Какую регрессию нужно оценить, чтобы найти RSS_R ?

Согласно H_0 должно выполняться: $\beta_2 = \beta_3$. Тогда:

$$problem_i = \beta_1 + \beta_2(lecture_i + pie_i) + u_i = \beta_1 + \beta_2 lp_i + u_i$$

где $lp_i = lecture_i + pie_i$