

Построить уравнение регрессии, проверить на наличие мультиколлинеарности; сделать вывод относительно значимости коэффициентов и уравнения в целом, проанализировать "объясняющую способность" уравнения. При наличии проблемы мультиколлинеарности - устранить ее, построить новое уравнение регрессии, оценить значимость коэффициентов и уравнения в целом, оценить R^2 (+ изменение скорректированного коэфф-та детерминации), дать экономическую интерпретацию коэффициентам уравнений.

год	потребление куриного мяса на душу в год, кг.	реальный располагаемый доход на душу, доллары США	розничная цена за 1 кг. курицы, центы	розничная цена за 1 кг. свинины, центы
	y	income	p_ch	p_pork

```
REGRESSION
  /MISSING LISTWISE
  /STATISTICS COEFF OUTS R ANOVA
  /CRITERIA=PIN(.05) POUT(.10)
  /NOORIGIN
  /DEPENDENT y

  /METHOD=ENTER God income p_ch p_pork.
```

Regression

Строим регрессию:

Y– потребление куриного мяса – зависимая переменная

Факторы:

God – год

Income- реальный доход на душу населения

p_ch – цена за 1 кг курицы

p_pork – цена за 1 кг свинины

[DataSet0]

Variables Entered/Removed^b

Model	Variables Entered	Variables Removed	Method
1	p_pork, God, income, p_ch ^a	.	Enter

a. All requested variables entered.

b. Dependent Variable: y

Model Summary				
Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	,991 ^a	,983	,979	,53267

a. Predictors: (Constant), p_pork, God, income, p_ch

Статистика

$$R^2 = R_{ss} / T_{ss} = 1 - E_{ss} / T_{ss}$$

называется ***коэффициентом детерминации***. Если $R^2 = 0$, это означает, что регрессия ничего не дает, т.е. знание x не улучшает предсказания для y по сравнению с тривиальным $\hat{y}_i = \bar{y}$. Другой крайний случай $R^2 = 1$ означает точную подгонку: все точки наблюдений лежат на регрессионной прямой. Чем ближе к 1 значение R^2 , тем лучше качество подгонки.
Так как в рассматриваемой регрессии $R^2 = 0,983$ что очень близко к 1, это значит что независимые факторы оказывают достаточно большое влияние на зависимую переменную.
Так как коэффициент корреляции $R = 0,991$ тоже достаточно большой и положительный, это говорит о том, что зависимая и независимые переменные находятся в прямой зависимости (с ростом значений по независимым переменным, будет увеличиваться и значение зависимой переменной Y)

Скорректированный R^2

$$\bar{R}^2 = R^2 - \frac{k}{n - k - 1}(1 - R^2) = 0.979$$

Показывает долю объясненной дисперсии с учетом числа переменных уравнения регрессии.
Скорректированные R^2 в разных моделях с разным числом переменных (и наблюдений) ограниченно сравнимы.

ANOVA ^b						
Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	295,547	4	73,887	260,403	,000 ^a
	Residual	5,107	18	,284		
	Total	300,655	22			

a. Predictors: (Constant), p_pork, God, income, p_ch

b. Dependent Variable: y

Проверка корректности многофакторной регрессионной модели

Завершающий этап регрессионного анализа – проверка адекватности модели по имеющимся статистическим данным. Для проверки постулированной модели используется:

А. дисперсионный анализ модели – для проверки качества уравнения в целом

Дисперсионный анализ в матричном виде

Источник	SS	Число степеней свободы	Средние квадраты
Регрессия	$B'X'Y - n\bar{Y}^2$	k	ME_{SS}
Остаток	$Y'Y - B'X'Y$	$n - k - 1$	MR_{SS}
Общий	$Y'Y - n\bar{Y}^2$	$n - 1$	

Проверка гипотезы о незначимости регрессии в целом

Выдвигаем гипотезу:

H_0 : все коэффициенты β (за исключением β_0) равны нулю;

H_1 : по крайней мере один коэффициент (за исключением β_0) не равен нулю.

F – критерий значимости регрессии

$F = ME_{SS} / MR_{SS} = 73,887 / 0,284 = 260,403$

– это отношение подчиняется F - распределению с $(k; n - k - 1)$ степенями свободы.

$F_{kr}(4,18) = 4.59$ (при уровне надежности 99%)

1. Если F – незначимый, то это показывает, что модель, по-видимому, неадекватна. => необходимо изучить, где встречается неадекватность (например, график остатков).

2. Если F – значимый, то это показывает, что нет оснований сомневаться в адекватности.

$F > F_{kr} (260.403 > 4.59)$ значит F – значимо, поэтому найденная регрессия адекватна.

Coefficients^a

		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	-887,262	125,746		-7,056	,000
	God	,461	,064	,847	7,208	,000
	income	,003	,001	,573	4,891	,000
	p_ch	-,043	,024	-,254	-1,775	,093
	p_pork	-,010	,011	-,192	-,925	,367

a. Dependent Variable: y

По данным этой таблицы записываем уравнение линейной регрессии:

$Y = 0.461 * God + 0.003 * income - 0.043 * p_ch - 0.010 * p_pork - 887,262$

Таким образом при увеличении переменной God на 1 единицу, величина потребления куриного мяса увеличивается на 0,461

При увеличении дохода на 1 единицу – увеличение потребления курицы происходит на 0,003

С увеличением цены на курицу на 1 единицу, происходит снижение потребления на -0,043

При увеличении цены на свинину, так же наблюдается снижение потребления куриного мяса на -0,01

Значимость коэффициентов регрессии проверяется с помощью критерия Стьюдента:

Так как выше мы нашли, что F -критерий значим, то следующий этап – проверка значимости каждой отдельной сопутствующей переменной.
Расчет стандартной ошибки коэффициента для случая двух факторов

$$SE(b_i) = \sqrt{\frac{s_u^2}{nVar(x_1)} \cdot \frac{1}{1 - r_{x_1, x_2}^2}}$$

Проверка гипотезы о значимости i-ой сопутствующей переменной

Выдвинем следующую гипотезу:

$H_0: b_i = 0$, т.е. нет существенного влияния x_i на Y .

$H_1: b_i \neq 0$, т.е. имеется существенное влияние.

Нас интересует вклад оценки b_i в разброс Y . Критерием служит отношение:

$$t = \frac{b_i}{\sqrt{\sigma^{2*}(b_i^*)}} \text{ , наблюдаемые статистики отображены в 4 колонке t/}$$

где $\sigma^{2*}(b_i^*)$ – стандартное отклонение оценки b_i^* , которая находится как i-ый диагональный элемент матрицы ковариаций $C^* = \sigma^{2*}(X^T X)^{-1}$

Гипотеза принимается, если критерий t попадает в интервал:

$$\left(-t_{\frac{1+\gamma}{2}}(n-k-1); t_{\frac{1+\gamma}{2}}(n-k-1) \right).$$

Дуровнем принятия гипотезы о значимости коэффициента будем считать $T_{кр}(99\%, 18)=2,55$

ОПГ= $(-2,55; 2,55)$

Для коэффициента b0- const $t=-7.056 < -2.55$ а значит не попадает в ОПГ гипотезы H_0 , этот коэффициент значим!

Для коэффициента b1- god $t=7.208 > 2.55$ не попадает в ОПГ, коэффициент значим!

Для коэффициента b2- income $t=4.891 > 2.55$ не попадает в ОПГ, коэффициент значим!

Для коэффициента b3- p_ch $t=-1.775 > -2.55$ попадает в ОПГ, коэффициент не значим!

Для коэффициента b4- p_pork $t=-0.925 > -2.55$ попадает в ОПГ, коэффициент не значим!

Если в качестве уровня надежности использовать 99%, то в построенной модели есть 2 не значимых фактора, это цена на куриное мясо и цена на свинину.

Так как модель имеет высокий коэффициент детерминации, и 2 не значимых параметра, то стоит проверить модель на наличие мультиколлинеарности.

Мультиколлинеарность - это понятие, которое используется для описания проблемы, когда нестрогая линейная зависимость между объясняющими переменными приводит к получению ненадежных оценок регрессии. Разумеется, такая зависимость совсем необязательно дает неудовлетворительные оценки. Если все другие условия благоприятствуют, т.е. если число наблюдений и выборочные дисперсии объясняющих переменных велики, а дисперсия случайного члена - мала, то в итоге можно получить вполне хорошие оценки.

Итак, мультиколлинеарность должна вызываться сочетанием нестрогой зависимости и одного (или более) неблагоприятного условия, и это- вопрос степени выраженности явления, а не его вида. Оценка любой регрессии будет страдать от нее в определенной степени, если только все независимые переменные не окажутся абсолютно некоррелированными. Рассмотрение данной проблемы начинается только тогда, когда это серьезно влияет на результаты оценки регрессии.

Строгая и нестрогая мультиколлинеарность:

1. Строгая (perfect) мультиколлинеарность - наличие линейной функциональной связи между независимыми переменными (иногда также и зависимой)
2. Нестрогая (imperfect) мультиколлинеарность - наличие сильной линейной корреляционной связи между независимыми переменными (иногда также и зависимой)

Причина мультиколлинеарности

1. Ошибочное включение в уравнение двух или более линейно зависимых переменных.
2. Две или более объясняющие переменные, в нормальной ситуации слабо коррелированные, становятся в конкретных условиях выборки сильно коррелированными.

В модель включается переменная, сильно коррелирующая с зависимой переменной (такая независимая переменная называется **доминантной**)

С помощью парной корреляции проверяем есть ли зависимость между всеми переменными:

CORRELATIONS

/VARIABLES=God y income p_ch p_pork

/PRINT=TWOTAIL NOSIG

/MISSING=PAIRWISE.

Correlations[DataSet0]

		Correlations				
		God	y	income	p_ch	p_pork
God	Pearson Correlation	1	,979**	,939**	,883**	,948**
	Sig. (2-tailed)		,000	,000	,000	,000
	N	23	23	23	23	23
y	Pearson Correlation	,979**	1	,948**	,841**	,913**
	Sig. (2-tailed)	,000		,000	,000	,000
	N	23	23	23	23	23
income	Pearson Correlation	,939**	,948**	1	,931**	,958**
	Sig. (2-tailed)	,000	,000		,000	,000
	N	23	23	23	23	23
p_ch	Pearson Correlation	,883**	,841**	,931**	1	,969**
	Sig. (2-tailed)	,000	,000	,000		,000
	N	23	23	23	23	23
p_pork	Pearson Correlation	,948**	,913**	,958**	,969**	1
	Sig. (2-tailed)	,000	,000	,000	,000	
	N	23	23	23	23	23

** . Correlation is significant at the 0.01 level (2-tailed).

Как видим из таблицы наблюдается сильная парная корреляция.
Исключим переменную p_pork так как у нее самая высокая корреляция с независимой переменной p_ch а так же ранее мы установили, что эта переменная не является значимой

После удаления переменной p_pork опять строим регрессию:

```
REGRESSION
/MISSING LISTWISE
/STATISTICS COEFF OUTS R ANOVA
/CRITERIA=PIN(.05) POUT(.10)
/NOORIGIN
/DEPENDENT y

/METHOD=ENTER God income p_ch.
```

Regression

[DataSet0]

Variables Entered/Removed ^b			
Model	Variables Entered	Variables Removed	Method
1	p_ch, God, income ^a	.	Enter

- a. All requested variables entered.
- b. Dependent Variable: y

Model Summary				
Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	,991 ^a	,982	,979	,53064

- a. Predictors: (Constant), p_ch, God, income

Коэффициент детерминации (и скорректированный R²) не изменились, что подтверждает правильность исключения не значимой переменной.

ANOVA ^b						
Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	295,305	3	98,435	349,577	,000 ^a
	Residual	5,350	19	,282		
	Total	300,655	22			

- a. Predictors: (Constant), p_ch, God, income
- b. Dependent Variable: y

Как видим из таблицы, коэффициент Фишера значительно увеличился =349,577, что указывает на адекватность построенной модели.

Coefficients^a

Model	Unstandardized Coefficients		Standardized Coefficients	t	Sig.
	B	Std. Error	Beta		
1	(Constant)	-811,817	95,336	-8,515	,000
	God	,423	,049	,776	,000
	Income	,003	,001	,554	,000
	p_ch	-,061	,014	-,360	,000

a. Dependent Variable: y

Значения t- статистики для всех коэффициентов достаточно велико и не попадает в ОПГ (-2,55; 2,55) при уровне надежности 99%

Вывод данная регрессионная модель лучше предыдущей

И так как после исключения переменной p_rok переменная p_ch стала значимой, то скорее всего мы столкнулись с мультиколлинеарностью, вызванной «Ошибочным включением в уравнение двух или более линейно зависимых переменных.»

Уравнение регрессии:

$$Y=0.423*God+0.003*income-0.061*p_ch -811.817$$

при увеличении переменной God на 1 единицу, величина потребления куриного мяса увеличивается на 0,423

При увеличении дохода на 1 единицу – увеличение потребления курицы происходит на 0,003

С увеличением цены на курицу на 1 единицу, происходит снижение потребления на -0,061