

Множественная регрессия. Задания

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.

```
> model <- lm(change_in_crime_rate ~ change_in_pop + kids + free_lunch + income_change)
```

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	-22.3548	12.3097		
change_in_pop	0.3188	0.2052		
kids	1.1128	0.2869		
free_lunch	-0.3681	0.0973		
income_change	-0.1944	0.3681		

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

```
> anova(model)
```

Response: change_in_crime_rate

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
change_in_pop	1	803.2			
kids	1	1380.1			
free_lunch	1	3186.6			
income_change	1	60.6			
Residuals	40	8476.0			

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

```
> describe(df)
```

	n	mean	sd	median	min	max	range
change_in_pop	44	7.83	12.20	4.70	-3.3	68.6	71.9
kids	44	27.27	8.86	28.85	7.5	41.5	34.0
free_lunch	44	51.51	25.88	57.55	5.7	88.5	82.8
income_change	44	26.96	6.77	26.35	11.7	50.5	38.8
change_in_crime_rate	44	-13.72	17.98	-13.45	-45.6	45.4	91.0

- Восстановите все пропуски в таблице.
- Определите статистически значимые оценки коэффициентов.
- Проинтерпретируйте оценку константы.
- Проинтерпретируйте оценку коэффициента при предикторе «процент бесплатных школьных обедов».
- Рассчитайте коэффициент детерминации и проверьте гипотезу о том, что регрессия на константу (то есть, модель без объясняющих переменных) не хуже модели с предикторами. Сформулируйте нулевую и альтернативную гипотезы для критерия, а также содержательный вывод.
- Вы подозреваете, что между переменными `income_change` и `free_lunch` имеется сильная линейная связь, как она может отразиться на вашей модели? Присутствует ли проблема мультиколлинеарности в модели (ответ обоснуйте)?

```
> vif(model)
```

	change_in_pop	kids	free_lunch	income_change
	1.241	1.278	1.256	1.230

Задание 2. Встречаются исследования, где центрирование используется для того, чтобы снизить мультиколлинеарность в модели. Почему это неверная стратегия?