Возможные косвенные показатели для улучшения мультиколлинеарности.

$$\sigma^2_{\hat{\beta}_2} = \frac{\sigma^2_u}{\sum\left(X_{2i} - \bar{X}_2\right)^2} \times \frac{1}{1 - r^2_{X_2,X_3}} = \frac{\sigma^2_u}{n\mathrm{MSD}\left(X_2\right)} \times \frac{1}{1 - r^2_{X_2,X_3}}$$

(1) Сочетание (объединение) коррелированных переменных.

В данном примере мы рассмотрим четыре возможных метода решения проблем с мультиколлинераностью. Первый: Сочетание (Объединение) коррелированных переменных.

Возможные косвенные показатели для улучшения мультиколлинеарности.

$$\sigma^2_{\hat{\beta}_2} = \frac{\sigma^2_u}{\sum\left(X_{2i} - \bar{X}_2\right)^2} \times \frac{1}{1 - r^2_{X_2,X_3}} = \frac{\sigma^2_u}{n\mathrm{MSD}\left(X_2\right)} \times \frac{1}{1 - r^2_{X_2,X_3}}$$

(1) Сочетание (объединение) коррелированных переменных.

Первый метод гласит: если коррелированные переменные одинаковы по своему принципу, то резонно было объединить их в некоторый общий (обобщённый) показатель.

Возможные косвенные показатели для улучшения мультиколлинеарности.

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

Source	SS	df	MS		
Model	1235.0519	3	411.683966	Number of obs	500
Residual	2518.9701	496	5.07856875	F(3, 496)	81.06
Total	3754.022	499	7.52309018	Prob > F	0.0000
				R-squared	0.3290
				Adj R-squared	0.3249
				Root MSE	2.2536

S	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
ASVABC	1.242527	.123587	10.05	0.000	.999708	1.485345
SM	.091353	.0459299	1.99	0.047	.0011119	.1815941
SF	.2028911	.0425117	4.77	0.000	.1193658	.2864163
_cons	10.59674	.6142778	17.25	0.000	9.389834	11.80365

Данное действие определенно можно выполнить с помощью трех (*ASVAB)* показателей. *ASVABC* считается как среднее значение подсчетов вспомогательных показателей: *ASVABAR* (арифметически обоснованный), *ASVABWK* (группа чисел), and *ASVABPC* (охват определенной группы чисел).

Возможные косвенные показатели для улучшения мультиколлинеарности.

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

Source	SS	df	MS		
Model	1235.0519	3	411.683966	Number of obs	500
Residual	2518.9701	496	5.07856875	F(3, 496)	81.06
Total	3754.022	499	7.52309018	Prob > F	0.0000
				R-squared	0.3290
				Adj R-squared	0.3249
				Root MSE	2.2536

S	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
ASVABC	1.242527	.123587	10.05	0.000	.999708	1.485345
SM	.091353	.0459299	1.99	0.047	.0011119	.1815941
SF	.2028911	.0425117	4.77	0.000	.1193658	.2864163
_cons	10.59674	.6142778	17.25	0.000	9.389834	11.80365

Объединение и подсчет среднего значения этих трех показателей поможет установить большую связь (корреляцию), нежели использование каждого из показателей отдельно, что позволит избежать потенциальных проблем с мультиколлинеарностью.

Возможные косвенные показатели для улучшения мультиколлинеарности.

$$\sigma^2_{\hat{\beta}_2} = \frac{\sigma^2_u}{\sum\left(X_{2i} - \bar{X}_2\right)^2} \times \frac{1}{1 - r^2_{X_2,X_3}} = \frac{\sigma^2_u}{n\mathrm{MSD}\left(X_2\right)} \times \frac{1}{1 - r^2_{X_2,X_3}}$$

(2) Отбрасывание (упущение) одной из коррелированных переменных.

Второй Метод: в случае если одна из коррелированных переменных имеет незначительный коэффициент, её можно отбросить (упустить), что также позволит улучшить мультиколлинеарность.

Возможные косвенные показатели для улучшения мультиколлинеарности.

$$\sigma^2_{\hat{\beta}_2} = \frac{\sigma^2_u}{\sum\left(X_{2i} - \bar{X}_2\right)^2} \times \frac{1}{1 - r^2_{X_2,X_3}} = \frac{\sigma^2_u}{n\mathrm{MSD}\left(X_2\right)} \times \frac{1}{1 - r^2_{X_2,X_3}}$$

(2) Отбрасывание (упущение) одной из коррелированных переменных.

Однако, такой подход к решению может быть опасным. Вполне возможно, что переменная с незначительным коэффициентом занимает важное место в модели, а единственная причина, почему её коэффициент незначителен, это проблема в мультиколлинеарности.

Возможные косвенные показатели для улучшения мультиколлинеарности.

$$\sigma^2_{\hat{\beta}_2} = \frac{\sigma^2_u}{\sum\left(X_{2i} - \bar{X}_2\right)^2} \times \frac{1}{1 - r^2_{X_2,X_3}} = \frac{\sigma^2_u}{n\mathrm{MSD}\left(X_2\right)} \times \frac{1}{1 - r^2_{X_2,X_3}}$$

(2) Отбрасывание (упущение) одной из коррелированных переменных.

Если такое происходит, то метод «упущения» приведет к неправильным расчетам. (Подробнее в главе 6)

Возможные косвенные показатели для улучшения мультиколлинеарности.

$$\sigma^2_{\hat{\beta}_2} = \frac{\sigma^2_u}{\sum\left(X_{2i} - \bar{X}_2\right)^2} \times \frac{1}{1 - r^2_{X_2,X_3}} = \frac{\sigma^2_u}{n\mathrm{MSD}\left(X_2\right)} \times \frac{1}{1 - r^2_{X_2,X_3}}$$

(3) Эмпирическое ограничение на основе дополнительных данных.

$$Y = \beta_1 + \beta_2 X + \beta_3 P + u$$

Третий метод решения проблем с мультиколлинеарностью это использование дополнительной информации об одной из переменных, если такая информация имеется.

Возможные косвенные показатели для улучшения мультиколлинеарности.

$$\sigma^2_{\hat{\beta}_2} = \frac{\sigma^2_u}{\sum\left(X_{2i} - \bar{X}_2\right)^2} \times \frac{1}{1 - r^2_{X_2,X_3}} = \frac{\sigma^2_u}{n\mathrm{MSD}\left(X_2\right)} \times \frac{1}{1 - r^2_{X_2,X_3}}$$

(3) Эмпирическое ограничение на основе дополнительных данных.

$$Y = \beta_1 + \beta_2 X + \beta_3 P + u$$

Предположим, что Y это количество потребительских расходов, X это количество располагаемого личного дохода, а P – ценовой индекс.

Возможные косвенные показатели для улучшения мультиколлинеарности.

$$\sigma^2_{\hat{\beta}_2} = \frac{\sigma^2_u}{\sum\left(X_{2i} - \bar{X}_2\right)^2} \times \frac{1}{1 - r^2_{X_2,X_3}} = \frac{\sigma^2_u}{n\mathrm{MSD}\left(X_2\right)} \times \frac{1}{1 - r^2_{X_2,X_3}}$$

(3) Эмпирическое ограничение на основе дополнительных данных.

$$Y = \beta_1 + \beta_2 X + \beta_3 P + u$$

Чтобы оперировать данным методом, необходимо использовать временные ряды. Если показатели X и P являются значимыми (максимально коррелированы), что является частым случаем при использовании временных рядов, то проблема с мультиколлинеарностью может быть устранена данным методом.

Возможные косвенные показатели для улучшения мультиколлинеарности.

$$\sigma^2_{\hat{\beta}_2} = \frac{\sigma^2_u}{\sum\left(X_{2i} - \bar{X}_2\right)^2} \times \frac{1}{1 - r^2_{X_2,X_3}} = \frac{\sigma^2_u}{n\mathrm{MSD}\left(X_2\right)} \times \frac{1}{1 - r^2_{X_2,X_3}}$$

(3) Эмпирическое ограничение на основе дополнительных данных.

$$Y = \beta_1 + \beta_2 X + \beta_3 P + u$$

$$Y' = \beta_1' + \beta_2' X' + u$$

$$\hat{Y}' = \hat{\beta}_1' + \hat{\beta}_2' X'$$

Полученные в ходе опроса данные о доходах и расходах. Регрессия *Y'* от *X'*. (отметка ' с буквенными обозначениями переменных, показывает, что это данные, полученные в ходе опроса, а не данные уравнения.)

Возможные косвенные показатели для улучшения мультиколлинеарности.

$$\sigma^2_{\hat{\beta}_2} = \frac{\sigma^2_u}{\sum\left(X_{2i} - \bar{X}_2\right)^2} \times \frac{1}{1 - r^2_{X_2,X_3}} = \frac{\sigma^2_u}{n\mathrm{MSD}\left(X_2\right)} \times \frac{1}{1 - r^2_{X_2,X_3}}$$

(3) Эмпирическое ограничение на основе дополнительных данных.

$$Y = \beta_1 + \beta_2 X + \beta_3 P + u$$

$$Y' = \beta_1' + \beta_2' X' + u$$

$$\hat{Y}' = \hat{\beta}_1' + \hat{\beta}_2' X'$$

Это (простая) линейная регрессия, потому что в ходе опроса был выявлен сравнительно маленький разброс цены, которую уплачивали опрашиваемые.

Возможные косвенные показатели для улучшения мультиколлинеарности.

$$\sigma^2_{\hat{\beta}_2} = \frac{\sigma^2_u}{\sum\left(X_{2i} - \bar{X}_2\right)^2} \times \frac{1}{1 - r^2_{X_2,X_3}} = \frac{\sigma^2_u}{n\mathrm{MSD}\left(X_2\right)} \times \frac{1}{1 - r^2_{X_2,X_3}}$$

(3) Эмпирическое ограничение на основе дополнительных данных.

$$Y = \beta_1 + \beta_2 X + \beta_3 P + u$$

$$Y = \beta_1 + \hat{\beta}_2' X + \beta_3 P + u$$

$$Z = Y - \hat{\beta}_2' X = \beta_1 + \beta_2 P + u$$

$$Y' = \beta_1' + \beta_2' X' + u$$

$$\hat{Y}' = \hat{\beta}_1' + \hat{\beta}_2' X'$$

Рассмотрим величину $\hat{\beta}_2'$ для β_2 во временных рядах. Сократим $\hat{\beta}_2' X$ с обеих сторон, и создадим регрессию $Z = Y - \hat{\beta}_2' X$ для цены. Это (простая) линейная регрессия, следовательно проблема с мультиколлинеарностью решена.

Возможные косвенные показатели для улучшения мультиколлинеарности.

$$\sigma^2_{\hat{\beta}_2} = \frac{\sigma^2_u}{\sum\left(X_{2i} - \bar{X}_2\right)^2} \times \frac{1}{1 - r^2_{X_2,X_3}} = \frac{\sigma^2_u}{n\mathrm{MSD}\left(X_2\right)} \times \frac{1}{1 - r^2_{X_2,X_3}}$$

(3) Эмпирическое ограничение на основе дополнительных данных.

$$Y = \beta_1 + \beta_2 X + \beta_3 P + u$$

$$Y = \beta_1 + \hat{\beta}_2^{'} X + \beta_3 P + u$$

$$Z = Y - \hat{\beta}_2^{'} X = \beta_1 + \beta_2 P + u$$

$$Y^{'} = \beta_1^{'} + \beta_2^{'} X^{'} + u$$

$$\hat{Y}^{'} = \hat{\beta}_1^{'} + \hat{\beta}_2^{'} X^{'}$$

Существует несколько проблем, связанных с данным методом. Во-первых, коэффициент β_2 во временных рядах, может отличаться от самого себя в выборке, относящейся к одному моменту времени.

Возможные косвенные показатели для улучшения мультиколлинеарности.

$$\sigma^2_{\hat{\beta}_2} = \frac{\sigma^2_u}{\sum\left(X_{2i} - \bar{X}_2\right)^2} \times \frac{1}{1 - r^2_{X_2,X_3}} = \frac{\sigma^2_u}{n\mathrm{MSD}\left(X_2\right)} \times \frac{1}{1 - r^2_{X_2,X_3}}$$

(3) Эмпирическое ограничение на основе дополнительных данных.

$$Y = \beta_1 + \beta_2 X + \beta_3 P + u$$

$$Y = \beta_1 + \hat{\beta}_2' X + \beta_3 P + u$$

$$Z = Y - \hat{\beta}_2' X = \beta_1 + \beta_2 P + u$$

$$Y' = \beta_1' + \beta_2' X' + u$$

$$\hat{Y}' = \hat{\beta}_1' + \hat{\beta}_2' X'$$

Во-вторых, Изначально мы вычисляли предполагаемую единицу $\hat{\beta}_2' X$, а не исконно верную $\beta_2 X$. При построении Z, мы, через Y нашли погрешность измерения зависимой переменной.

Возможные косвенные показатели для улучшения мультиколлинеарности.

$$\sigma^2_{\hat{\beta}_2} = \frac{\sigma^2_u}{\sum\left(X_{2i} - \bar{X}_2\right)^2} \times \frac{1}{1 - r^2_{X_2,X_3}} = \frac{\sigma^2_u}{n\mathrm{MSD}\left(X_2\right)} \times \frac{1}{1 - r^2_{X_2,X_3}}$$

(4) Теоретическое ограничение.

Последний, среди приведенных косвенных методов по улучшению мультиколлинеарности, это метод теоретического сокращения, который определяется как гипотетическое соотношение между параметрами регрессионной модели.

Возможные косвенные показатели для улучшения мультиколлинеарности.

$$\sigma^2_{\hat{\beta}_2} = \frac{\sigma^2_u}{\sum\left(X_{2i} - \bar{X}_2\right)^2} \times \frac{1}{1 - r^2_{X_2,X_3}} = \frac{\sigma^2_u}{n\mathrm{MSD}\left(X_2\right)} \times \frac{1}{1 - r^2_{X_2,X_3}}$$

(4) Теоретическое ограничение.

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

Данный метод можно объяснить с помощью простой модели на примере сверху. Предположим, что значение переменной *S*, Зависит от *ASVABC*, а само значение S построено с помощью определенных данных о маме и папе, *SM* и *SF*, соответственно.

Возможные косвенные показатели для улучшения мультиколлинеарности.

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

Source	SS	df	MS		
Model	1235.0519	3	411.683966	Number of obs	500
Residual	2518.9701	496	5.07856875	F(3, 496)	81.06
Total	3754.022	499	7.52309018	Prob > F	0.0000
				R-squared	0.3290
				Adj R-squared	0.3249
				Root MSE	2.2536

S	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
ASVABC	1.242527	.123587	10.05	0.000	.999708	1.485345
SM	.091353	.0459299	1.99	0.047	.0011119	.1815941
SF	.2028911	.0425117	4.77	0.000	.1193658	.2864163
_cons	10.59674	.6142778	17.25	0.000	9.389834	11.80365

Значение S **увеличивается на 0.09 за каждую дополнительную полученную степень образования у мамы, и на 0.20 за каждую дополнительную полученную степень образования у папы.**

Возможные косвенные показатели для улучшения мультиколлинеарности.

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

Source	SS	df	MS		
Model	1235.0519	3	411.683966	Number of obs	500
Residual	2518.9701	496	5.07856875	F(3, 496)	81.06
Total	3754.022	499	7.52309018	Prob > F	0.0000
				R-squared	0.3290
				Adj R-squared	0.3249
				Root MSE	2.2536

S	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
ASVABC	1.242527	.123587	10.05	0.000	.999708	1.485345
SM	.091353	.0459299	1.99	0.047	.0011119	.1815941
SF	.2028911	.0425117	4.77	0.000	.1193658	.2864163
_cons	10.59674	.6142778	17.25	0.000	9.389834	11.80365

Образование у мамы считается как минимум важнее чем образование, полученное папой, по меркам образовательной подготовки. Значение SM является более значимым, чем значение SF, что неожиданно.

Возможные косвенные показатели для улучшения мультиколлинеарности.

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

Source	SS	df	MS		
Model	1235.0519	3	411.683966	Number of obs	500
Residual	2518.9701	496	5.07856875	F(3, 496)	81.06
Total	3754.022	499	7.52309018	Prob > F	0.0000
				R-squared	0.3290
				Adj R-squared	0.3249
				Root MSE	2.2536

S	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
ASVABC	1.242527	.123587	10.05	0.000	.999708	1.485345
SM	.091353	.0459299	1.99	0.047	.0011119	.1815941
SF	.2028911	.0425117	4.77	0.000	.1193658	.2864163
_cons	10.59674	.6142778	17.25	0.000	9.389834	11.80365

```
. cor SM SF
(obs=500)
             |       SM       SF
-------------+------------------
          SM |   1.0000
          SF |   0.5312   1.0000
```

***Однако соединение показателей ведет к корреляции между SM и SF* и регрессия может пострадать из за мультиколлинеарности. Это может привести к неточным расчетам коэффициентов.**

$$\sigma^2_{\hat{\beta}_2} = \frac{\sigma^2_u}{\sum\left(X_{2i} - \bar{X}_2\right)^2} \times \frac{1}{1 - r^2_{X_2,X_3}} = \frac{\sigma^2_u}{n\mathrm{MSD}\left(X_2\right)} \times \frac{1}{1 - r^2_{X_2,X_3}}$$

(4) Теоретическое ограничение.

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

$$\beta_3 = \beta_4$$

Предположим, что образование (показатели образования) мамы и папы одинаково важны, в таком случае мы можем наложить ограничение $\beta_3 = \beta_4$.

Возможные косвенные показатели для улучшения мультиколлинеарности.

$$\sigma^2_{\hat{\beta}_2} = \frac{\sigma^2_u}{\sum\left(X_{2i} - \bar{X}_2\right)^2} \times \frac{1}{1 - r^2_{X_2,X_3}} = \frac{\sigma^2_u}{n\mathrm{MSD}\left(X_2\right)} \times \frac{1}{1 - r^2_{X_2,X_3}}$$

(4) Теоретическое ограничение.

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

$$\beta_3 = \beta_4$$

$$S = \beta_1 + \beta_2 ASVABC + \beta_3 \left(SM + SF\right) + u$$

Это позволит нам переформировать уравнение, как показано на экране.

Возможные косвенные показатели для улучшения мультиколлинеарности.

$$\sigma^2_{\hat{\beta}_2} = \frac{\sigma^2_u}{\sum\left(X_{2i} - \bar{X}_2\right)^2} \times \frac{1}{1 - r^2_{X_2,X_3}} = \frac{\sigma^2_u}{n\mathrm{MSD}\left(X_2\right)} \times \frac{1}{1 - r^2_{X_2,X_3}}$$

(4) Теоретическое ограничение.

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

$$\beta_3 = \beta_4$$

$$S = \beta_1 + \beta_2 ASVABC + \beta_3 \left(SM + SF\right) + u$$

$$= \beta_1 + \beta_2 ASVABC + \beta_3 SP + u$$

Определяем *SP* как сумму *SM* и *SF, переформировываем уравнение, как показано на экране.* Проблема, вызванная корреляцией между SM и SF, была устранена.

Возможные косвенные показатели для улучшения мультиколлинеарности.

```
. g SP=SM+SF

. reg S ASVABC SP
```

Source	SS	df	MS		
Model	1223.98508	2	611.992542	Number of obs	500
Residual	2530.03692	497	5.09061754	F(2, 497)	120.22
Total	3754.022	499	7.52309018	Prob > F	0.0000
				R-squared	0.3260
				Adj R-squared	0.3233
				Root MSE	2.2562

S	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
ASVABC	1.243199	.1237327	10.05	0.000	1.000095	1.486303
SP	.1500751	.0229866	6.53	0.000	.1049123	.1952379
_cons	10.50285	.6117	17.17	0.000	9.301009	11.70468

Значение β_3 теперь равняется 0.150.

Возможные косвенные показатели для улучшения мультиколлинеарности.

```
. g SP=SM+SF
. reg S ASVABC SP
```

S	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
ASVABC	1.243199	.1237327	10.05	0.000	1.000095	1.486303
SP	.1500751	.0229866	6.53	0.000	.1049123	.1952379
_cons	10.50285	.6117	17.17	0.000	9.301009	11.70468

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

S	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
ASVABC	1.242527	.123587	10.05	0.000	.999708	1.485345
SM	.091353	.0459299	1.99	0.047	.0011119	.1815941
SF	.2028911	.0425117	4.77	0.000	.1193658	.2864163
_cons	10.59674	.6142778	17.25	0.000	9.389834	11.80365

Значение SP это компромисс между значениями SM и SF. Расчет значения SP был показан на предыдущем слайде.

Возможные косвенные показатели для улучшения мультиколлинеарности.

```
. g SP=SM+SF
. reg S ASVABC SP
```

S	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
ASVABC	1.243199	.1237327	10.05	0.000	1.000095	1.486303
SP	.1500751	.0229866	6.53	0.000	.1049123	.1952379
_cons	10.50285	.6117	17.17	0.000	9.301009	11.70468

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

S	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
ASVABC	1.242527	.123587	10.05	0.000	.999708	1.485345
SM	.091353	.0459299	1.99	0.047	.0011119	.1815941
SF	.2028911	.0425117	4.77	0.000	.1193658	.2864163
_cons	10.59674	.6142778	17.25	0.000	9.389834	11.80365

Стандартная ошибка *SP* значительно меньше чем у *SM* и *SF*. Использование ограничения привело нас к увеличению эффективности решения задачи, что помогло решить и проблему с мультиколлинеарностью.

Возможные косвенные показатели для улучшения мультиколлинеарности.

```
. g SP=SM+SF
. reg S ASVABC SP
```

S	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
ASVABC	1.243199	.1237327	10.05	0.000	1.000095	1.486303
SP	.1500751	.0229866	6.53	0.000	.1049123	.1952379
_cons	10.50285	.6117	17.17	0.000	9.301009	11.70468

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

S	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
ASVABC	1.242527	.123587	10.05	0.000	.999708	1.485345
SM	.091353	.0459299	1.99	0.047	.0011119	.1815941
SF	.2028911	.0425117	4.77	0.000	.1193658	.2864163
_cons	10.59674	.6142778	17.25	0.000	9.389834	11.80365

Значение *t* достаточно велико. Это означает, что наложение ограничения улучшило результаты регрессии .Однако, возможно, что ограничение было наложено неправильно. Нам необходимо это проверить. Подробнее о проверке метода в главе 6.