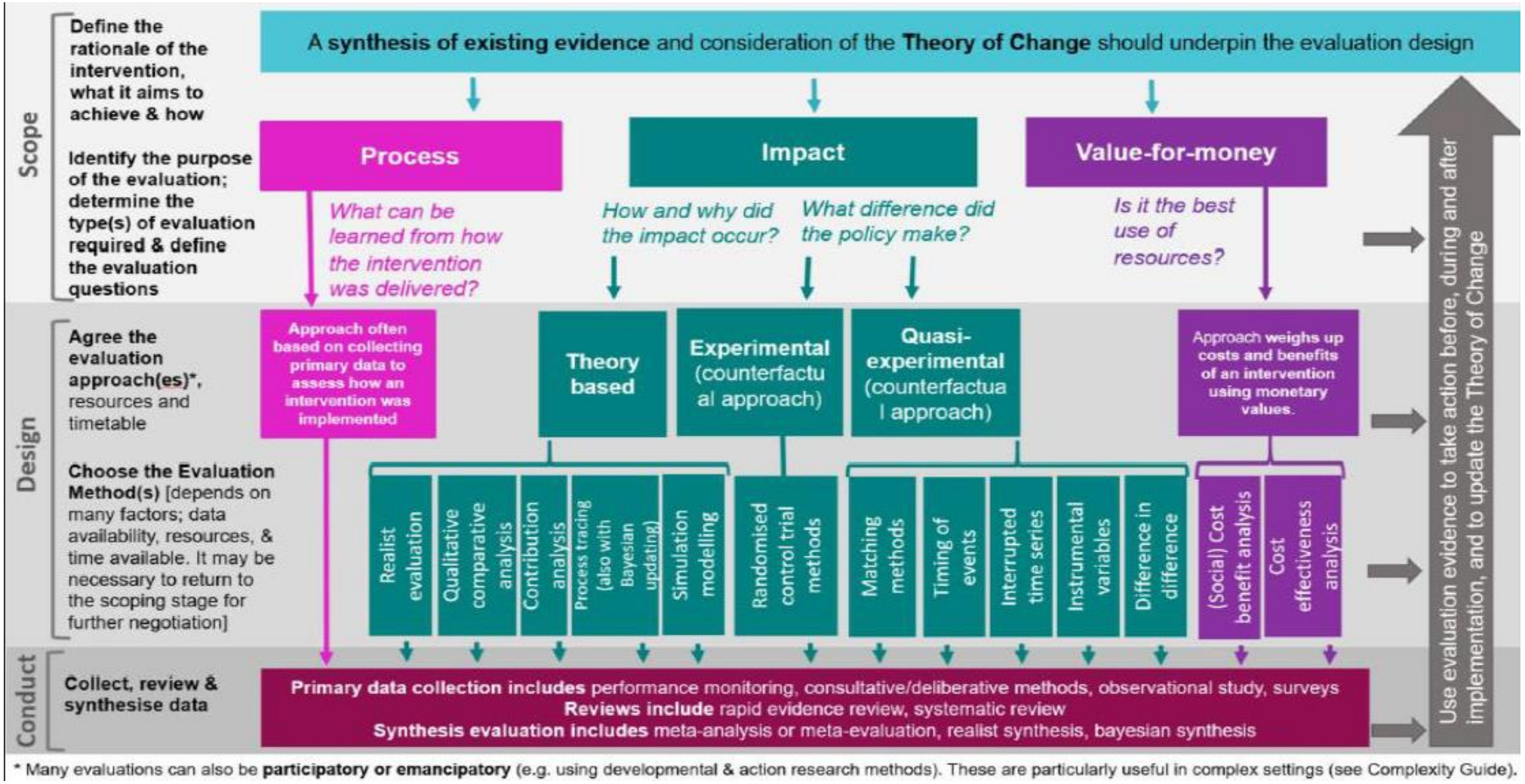
Обобщение методов и разговор про работу с данными и про оформление

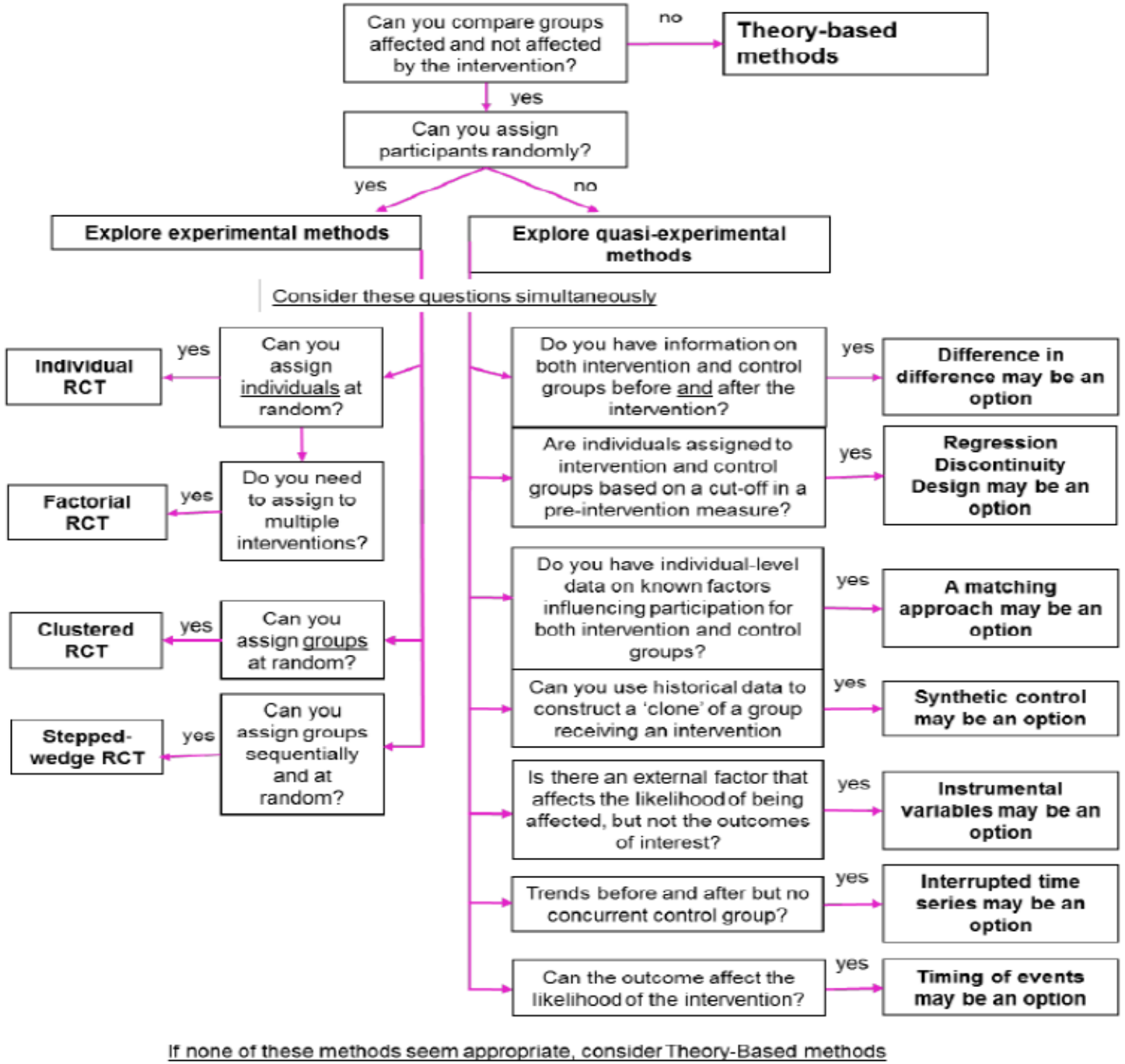
(Вторая часть будет полезна для диплома)

Про «лестницу методов»

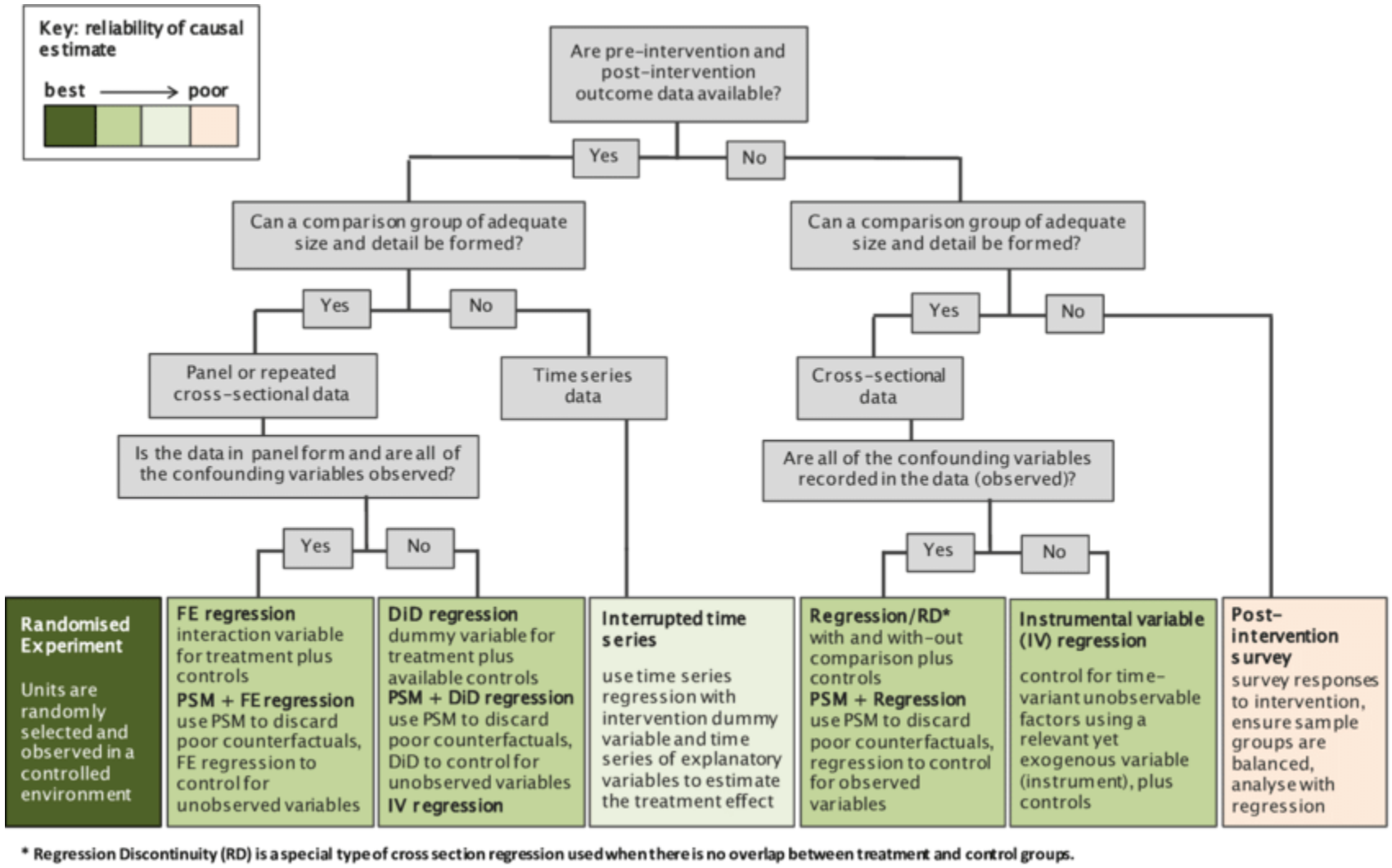
Пример – руководство по оценке госпрограмм в Великобритании



Magenta book: блок-схема выбора методов



Пример



https://www.researchgate.net/publication/320419299_Ex-_post_evaluation_of_transport_interventions_using_causal_inference_methods

- Большая схема-таблица – на доске

Про работу с данными и оформление

Основное:

- «Эмпирическая часть» = эконометрика ? Нет!
- О чём стоит подумать при работе с данными
- Примеры хорошие и плохие
- Как оформлять эмпирическую часть

«Эмпирическая часть» = эконометрика ? Нет

- В Положении о НИР и ВКР нет требования, чтобы обязательно была эконометрика https://www.econ.msu.ru/sys/raw.php?o=63865&p=attachment

Эконометрика – не обязательна, всё зависит от темы

- Авторская теоретическая модель. Пример «Конкуренция платформ на двусторонних рынках» (ВКР 2017). Эконометрика – только для проверки вывода из модели, если достаточно данных
- Вопрос требует проведения качественного, а не количественного анализа. Пример - глубинное интервью (ВКР 2020)
- Вопрос касается ограниченного числа уникальных случаев => кейс-стади.

- Но в любом случае это НЕ реферативные работы!

А если эконометрика?

- Мотивация
- Теория, обзор => гипотезы
- Обзор эмпирики => обоснование метода
- Спецификация модели
- Данные: описание и первичный анализ, чистка
- Расчёты
- Анализ устойчивости (робастности) результатов

Опросы – очень осторожно

- Где проводится опрос? Анкеты в бумажном или электронном виде?
- Насколько корректно оставлен опрос?
- Каков размер выборки?
- Репрезентативная ли выборка?

Пример (проект по одному из предметов 2021):

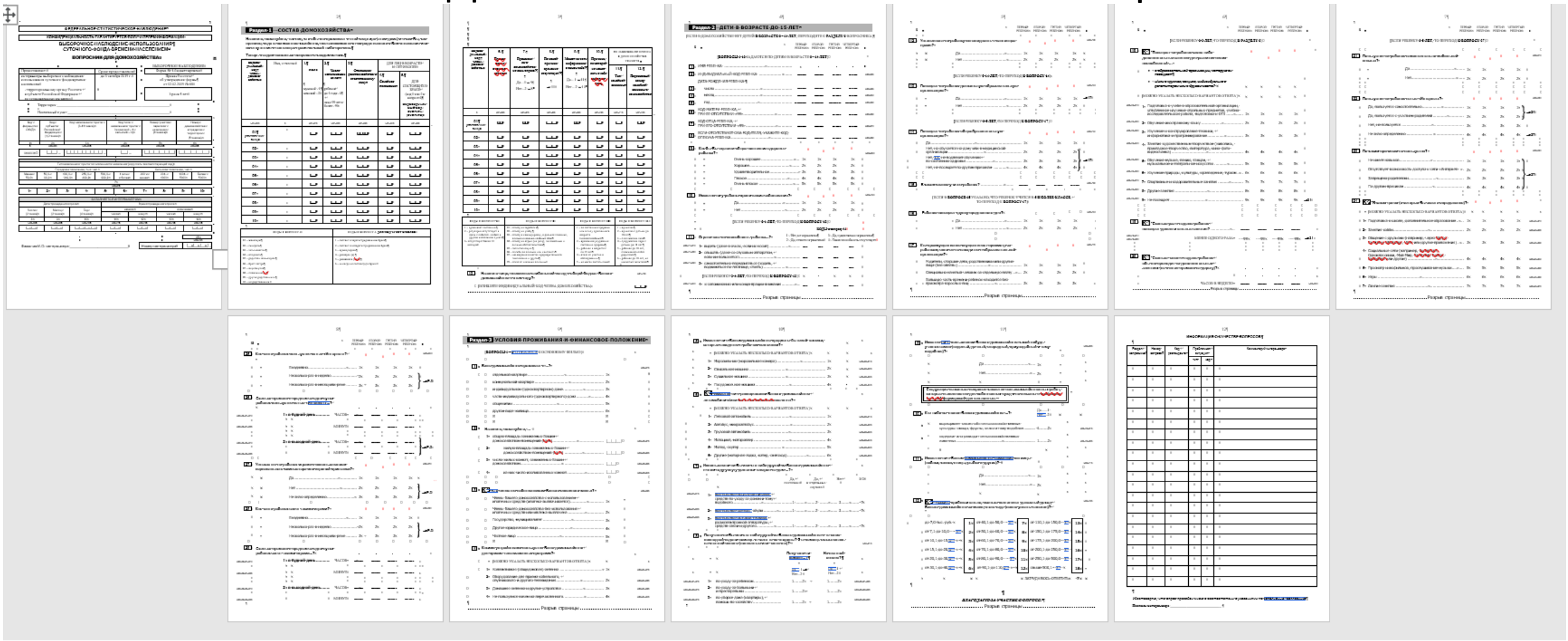
«Влияние разделения труда в семье на гендерный разрыв в зарплатах»

Авторы хотели провести опрос на Яндекс.Взгляд

Проблемы?

Проблемы?

- Одна из проблем – максимум 4 вопроса в анкете Яндекс.Взгляд,
- Вопросник «ВЫБОРОЧНОЕ НАБЛЮДЕНИЕ ИСПОЛЬЗОВАНИЯ СУТОЧНОГО ФОНДА ВРЕМЕНИ НАСЕЛЕНИЕМ» - в N раз больше



О чём стоит подумать при работе с данными

- Какова цель работы?
- Какую гипотезу проверяете?
- Отражает ли выбранный показатель именно те изменения, которые вы хотите оценить?
- Источники данных
- Методология расчёта показателей – сопоставимы ли данные?
- Предварительный анализ данных (графики, описательная статистка, корреляции)
- Если эконометрика, то важен выбор метода оценки и осторожная интерпретация оценок, проверка устойчивости результатов

Отражает ли выбранный показатель именно те изменения, которые хотите оценить?

Пример из ВКР 2019 года:

- Гипотеза: рост жилищного строительства в регионе привлекает население
- Данные: ввод (кв. м) нового жилья в 1 регионе с 2011 по 2016 гг. и численность населения этого региона с 2011 по 2016 гг.
- (Не очень удачная) идея проверки: рассчитать парный коэффициент корреляции между показателями.
- Результат: корреляция 0,89
- Автор сделал вывод, что для предотвращения оттока населения из этого региона необходимо увеличить темы роста жилищного строительства
- Проблемы?

Отражает ли выбранный показатель именно те изменения, которые хотите оценить?

- Проблема 1: На численность населения влияют не только миграционный прирост, но и естественный. Выбранный автором показатель их не разграничивает.
- Проблема 2: Парная корреляция по 6 точкам мало о чём говорит, очень мало наблюдений, поэтому нельзя даже доверять t-тесту на значимость
- Проблема 3: Не учтены прочие факторы
- Проблема 4: Корреляция не отвечает на вопрос о причинно-следственной связи.

Источники данных и методология расчёта показателей

- Обязательно – указывать источник данных по каждому показателю, ссылку на базу, дату обращения к базе
- Обязательно – расшифровка в тексте всех условных обозначений

Таблица 11. Переменные и ожидаемый знак влияния на госрасходы на образование

Переменная	Источник	Обозначение	Ож. знак
Совокупные госрасходы на образование (все ступени) в % ВВП	World Bank	ExpEdu	Зависимая переменная
Нефтяные доходы в % ВВП, лаг 1	WB	ResDep	- при условии «плохих» институтов
ВВП на душу населения, в пост. ценах 2010 г.(долл. США), скользящее среднее за 3 года, логарифм, лаг 1	WB	GDPpcMovAv	+
Экспорт услуг в % ВВП, лаг 1	WB	Services	+
Индекс ограничения политических прав (1 – права не ограничены, 7 – полностью ограничены), лаг 1	Freedom House	PolRights	-
Политический режим (-10 – автократия, 10 – демократия), лаг 1	Systemic Peace	PolityIV	+
Ожидаемая продолжительность жизни при рождении, лет, годовой прирост	WB	LifeExpect	-
Международная помощь от Комитета содействия развитию ОЭСР(DAC), в пост. ценах 2017 г. (млн. долл. США), логарифм, лаг 1	OECD/DAC	ForeignAid	-

Источник: составлено автором

Источники данных и методология расчёта показателей

- Обязательно – указывать источник данных по каждому показателю, ссылку на базу, дату обращения к базе, давать расшифровку обозначений
- Полезная привычка – записывать, как именно Вы преобразовывали данные, писать комментарии к коду
- Хороший тон – к тексту статьи прилагать код и набор данных или ссылку на репозиторий с ними

Пример – в ВКР 2020 г. «Управление талантами в современной компании» Рыбниковой Е.А. ссылка на диск с аудиозаписями и расшифровкой глубинного интервью

- Важно – посмотреть в методологию расчёта показателя

Пример: Индекс восприятия коррупции (CPI) менял методологию в 2012 году. Данные до и после несопоставимы.

Пример: сопоставление ОКВЭД-1 и ОКВЭД-2 (из ВКР 2019)

Описательная статистика – зачем? Описание выборки и проверка ошибок в данных

Так - плохо. Почему?

Statistic	N	Mean	St. Dev.	Min	Pctl(25)	Median	Pctl(75)	Max
spirits	336	1.75	0.68	0.79	1.30	1.67	2.01	4.90
unemp	336	7.35	2.53	2.40	5.48	7.00	8.90	18.00
income	336	13,880.18	2,253.05	9,513.76	12,085.85	13,763.13	15,175.12	22,193.46
emppop	336	60.81	4.72	42.99	57.69	61.36	64.41	71.27
beertax	336	0.51	0.48	0.04	0.21	0.35	0.65	2.72
baptist	336	7.16	9.76	0.00	0.63	1.75	13.13	30.36
mormon	336	2.80	9.67	0.10	0.27	0.39	0.63	65.92
drinkage	336	20.46	0.90	18	20	21	21	21
dry	336	4.27	9.50	0.00	0.00	0.09	2.42	45.79
youngdrivers	336	0.19	0.02	0.07	0.17	0.19	0.20	0.28
miles	336	7,890.75	1,475.66	4,576.35	7,182.54	7,796.22	8,504.02	26,148.27
fatal	336	928.66	934.05	79	293.8	701	1,063.5	5,504
nfatal	336	182.58	188.43	13	53.8	135	212	1,049
sfatal	336	109.95	108.54	8	35	81	131	603
fatal1517	336	62.61	55.73	3	25.8	49	77	318
nfatal1517	336	12.26	12.25	0	4	10	15.2	76
fatal1820	336	106.66	104.22	7	38	82	130.2	601
nfatal1820	336	33.53	33.24	0	11	24	44	196
fatal2124	336	126.87	131.79	12	42	97.5	150.5	770
nfatal2124	336	41.38	42.93	1	13	30	49	249
afatal	336	293.33	303.58	24.60	90.50	211.59	363.96	2,094.90
pop	336	4,930,272.00	5,073,704.00	478,999.70	1,545,251.00	3,310,503.00	5,751,735.00	28,314,028.00
pop1517	336	230,815.50	229,896.30	21,000.02	71,749.93	163,000.20	270,500.20	1,172,000.00
pop1820	336	249,090.40	249,345.60	20,999.96	76,962.12	170,982.30	308,311.30	1,321,004.00
pop2124	336	336,389.90	345,304.40	30,000.16	103,500.00	240,999.90	413,000.10	1,892,998.00
milestot	336	37,101.49	37,454.37	3,993	11,691.5	28,483.5	44,139.8	241,575
unempus	336	7.53	1.48	5.50	6.20	7.20	9.60	9.70
emppopus	336	59.97	1.59	57.80	57.90	60.10	61.50	62.30
gsp	336	0.03	0.04	-0.12	0.001	0.03	0.06	0.14
mrall	336	0.0002	0.0001	0.0001	0.0002	0.0002	0.0002	0.0004
fatality_rate	336	2.04	0.57	0.82	1.62	1.96	2.42	4.22

А так - хорошо.

Таблица 8. Описательная статистика. Совокупные государственные расходы на образование в % ВВП по регионам.

Регион	Наблюдения	Страны	Среднее	Ст. от.	Мин.	Макс.
Африка южнее Сахары	193	12	3,64	1,69	0,70	8,14
Южная Азия	61	2	2,13	0,53	0,94	3,02
Северная Америка	29	2	5,87	0,81	4,64	7,70
Ближний Восток и Северная Африка	206	13	4,24	1,81	1,42	14,20
Латинская Америка и Карибский бассейн	246	11	4,24	1,26	1,05	7,40
Европа и Центральная Азия	659	24	5,08	1,20	2,07	8,56
Восточная Азия и Тихий океан	180	8	4,34	1,56	0,87	7,66

Источник: составлено автором на основе World Bank

Из ВКР Екатерины Ерёминой, 2020 г.

Описательная статистика – зачем? Описание выборки и проверка ошибок в данных

Таблица 4. Описание используемых переменных.

Переменная	Описание	Нью-Йорк		Москва (2014-2017)		Москва (2017-2019)		Россия	
		$\overline{x}$	σ	$\overline{x}$	σ	$\overline{x}$	σ	$\overline{x}$	σ
P_{it}	*Нью-Йорк*: индекс аренды Zillow Rent Index *Москва (2014-2017):* средняя цена аренды квартиры, руб. в месяц *Москва, Россия*: средняя цена аренды м2, руб. в месяц	2619	850,4	40963	29218	885	196,9	378	144,5
$Airbnb_{it}$	*Нью-Йорк, Москва (2014-2017)*: количество активных предложений жилья целиком на Airbnb *Москва (2017-2019), Россия*: количество активных предложений на Airbnb	112,3	180,8	12,5	57,5	77	234,6	817	5231
$population_{it}$	Численность населения (*Россия:* млн чел.)	47633	26990	95313	42315	93867	40365	1,068	2,453
$income_{it}$	Медианный доход (*Нью-Йорк*: $ в год; *Россия*: руб. в месяц)	69508	30467	–	–	–	–	29983	10129
$ownrate_{it}$	Доля резидентов-собственников	0,384	0,217	–	–	–	–	–	–
$tourdensity_{i2018}$	Отношение количества туристов в 2018 году к численности населения	–	–	–	–	–	–	1,045	3,9
g_t	Значение индекса Google trends для общемирового запроса «airbnb»	56,12	14,47	47,5	22,89	76,33	11,17	76,33	11,17
$tourism_{it_0}$	*Нью-Йорк*: количество ресторанов, отелей и достопримечательностей в 2014 году *Москва (2014-2017)*: количество достопримечательностей в 2014 году *Москва (2017-2019), Россия*: количество достопримечательностей в 2015 году	131	128,9	15	77,36	18	88,57	220,5	693,4

Примечание: $\overline{x}$ означает среднее значение показателя по выборке, σ означает стандартное отклонение показателя по выборке.
Источник: составлено автором на основе zillow.com; domofond.ru; insideairbnb.com; tomslee.net; airdna.co; Росстат; American Community Survey; Google trends; Ростуризм; TripAdvisor.

А так – тоже хорошо.

Из ВКР Антона Бреннермана, 2020 г.

Способ 1 - stargazer

- table_1 <- cov_balance(data_baseline$treat, data_baseline[, 2:13])
- stargazer(table_1, type = 'text', summary = FALSE, rownames = FALSE)

Variable	Mean_treatment	Mean_control	Difference	Standard_error	Observations_treatment	Observations_control
Female	0.760	0.760	0.004	0.034	314	305
Age	12.670	12.410	0.267	0.143	230	231
SES_index	-0.030	0.040	-0.070	0.137	314	305
Grade4	0.010	0.010	-0.003	0.007	305	299
Grade5	0.010	0.020	-0.007	0.010	305	299
Grade6	0.270	0.300	-0.035	0.037	305	299
Grade7	0.260	0.260	0.005	0.036	305	299
Grade8	0.300	0.280	0.017	0.037	305	299
Grade9	0.150	0.130	0.024	0.028	305	299
Math	-0.010	0.010	-0.016	0.081	313	304
Hindi	0.050	-0.050	0.096	0.081	312	305
Percent_at_endline	0.850	0.900	-0.048	0.027	314	305

Или так:

- Data <- rio::import('ms_blel_jpal_long.dta')
- Data <- filter(Data, Data$round==1)
- colnumlist=c(5, 4, 22, 23:28, 19,21, 36)
- tablenamelist=c('Female', 'Age (years)', 'SES index', 'Grade4', 'Grade5', 'Grade6', 'Grade7','Grade8', 'Grade9', 'Math', 'Hindi', 'Present at endline')

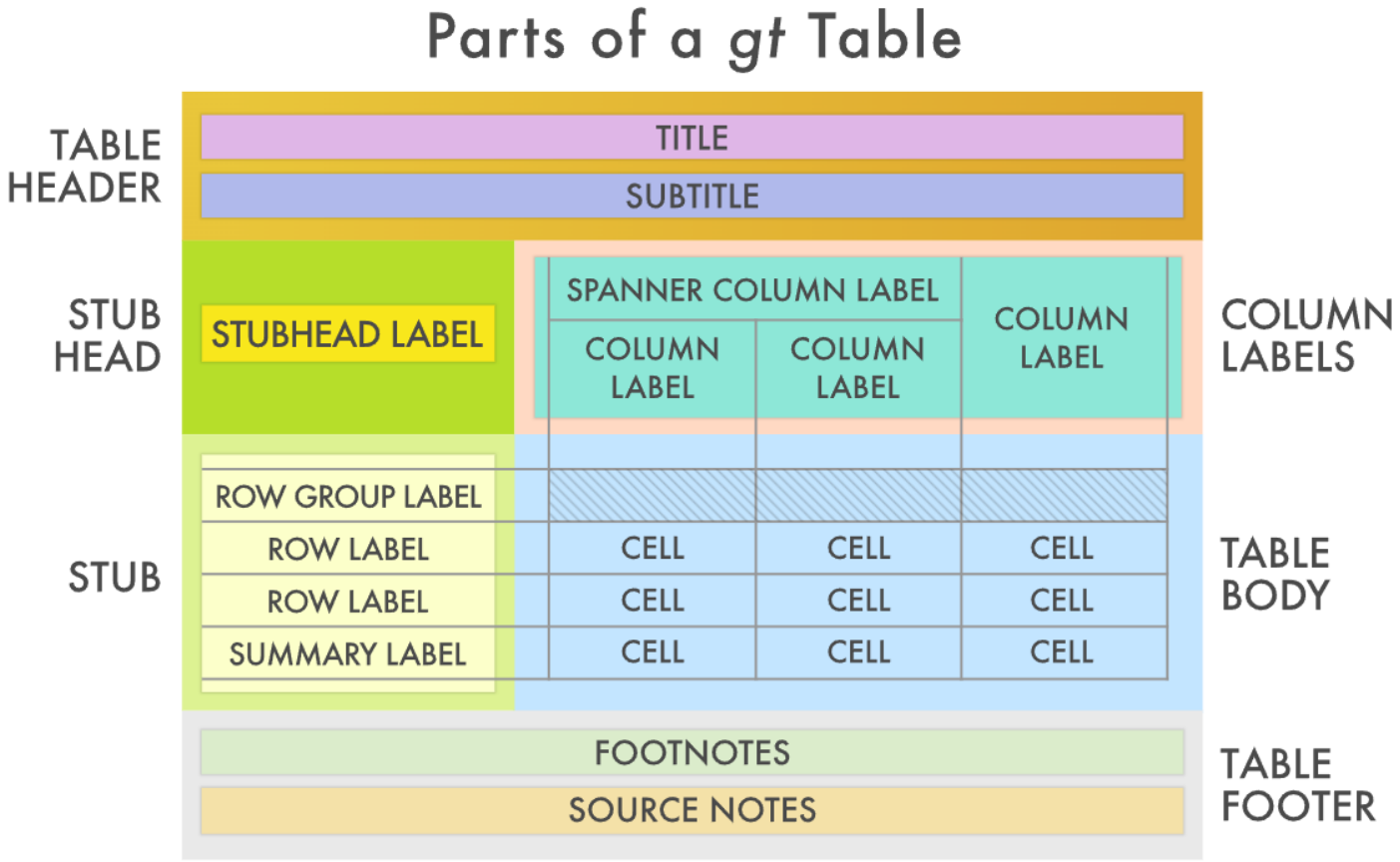
Или так:

	rows	Treatment_mean	Control_mean	Difference	Standart.error	Observations_treatment	Observations_control
1	Female	0.761	0.757	0.004	0.034	314	305
2	Age (years)	12.674	12.407	0.267	0.143	230	231
3	SES index	-0.035	0.036	-0.070	0.137	314	305
4	Grade4	0.007	0.010	-0.003	0.007	305	299
5	Grade5	0.013	0.020	-0.007	0.010	305	299
6	Grade6	0.266	0.301	-0.035	0.037	305	299
7	Grade7	0.262	0.258	0.005	0.036	305	299
8	Grade8	0.302	0.284	0.017	0.037	305	299
9	Grade9	0.151	0.127	0.024	0.028	305	299
10	Math	-0.008	0.008	-0.016	0.081	313	304
11	Hindi	0.047	-0.048	0.096	0.081	312	305
12	Present at endline	0.847	0.895	-0.048	0.027	314	305

```
table <- balance(Data, colnumlist, tablenamelist, treatcolname='treat')
stargazer(table, summary=FALSE, digits=3, type='html', out='table.html', label=c('Mean (treatment)’,
'Mean (control)', 'Difference', 'Standart error', 'Observation (treatment)', 'Observation (control)’))
```

Таблица почти такая же, как предыдущая, а код совсем другой!
Это нормально.
Люди независимо друг от друга пишут разные коды

Способ 2 – пакет gt https://gt.rstudio.com/

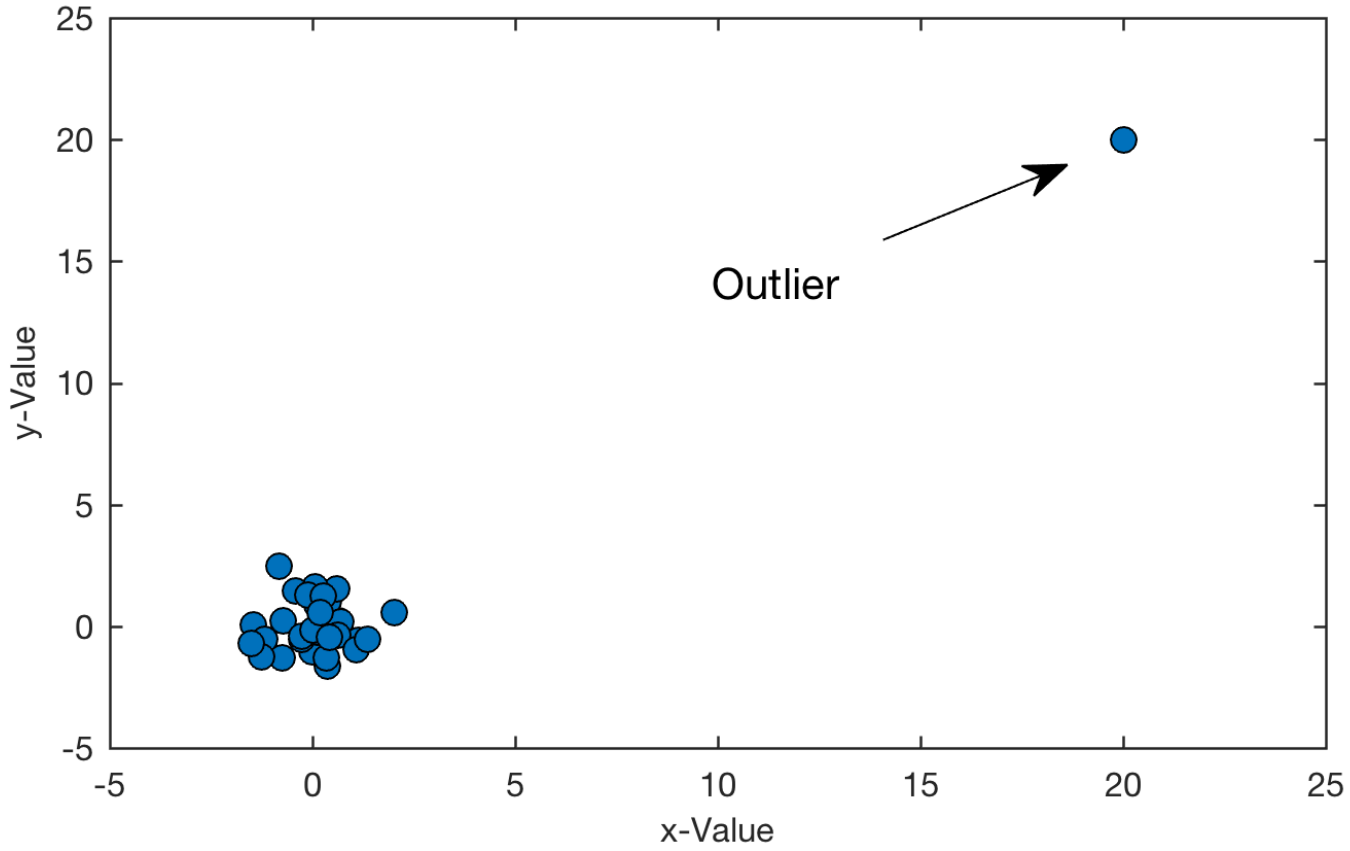


Viewer Zoom

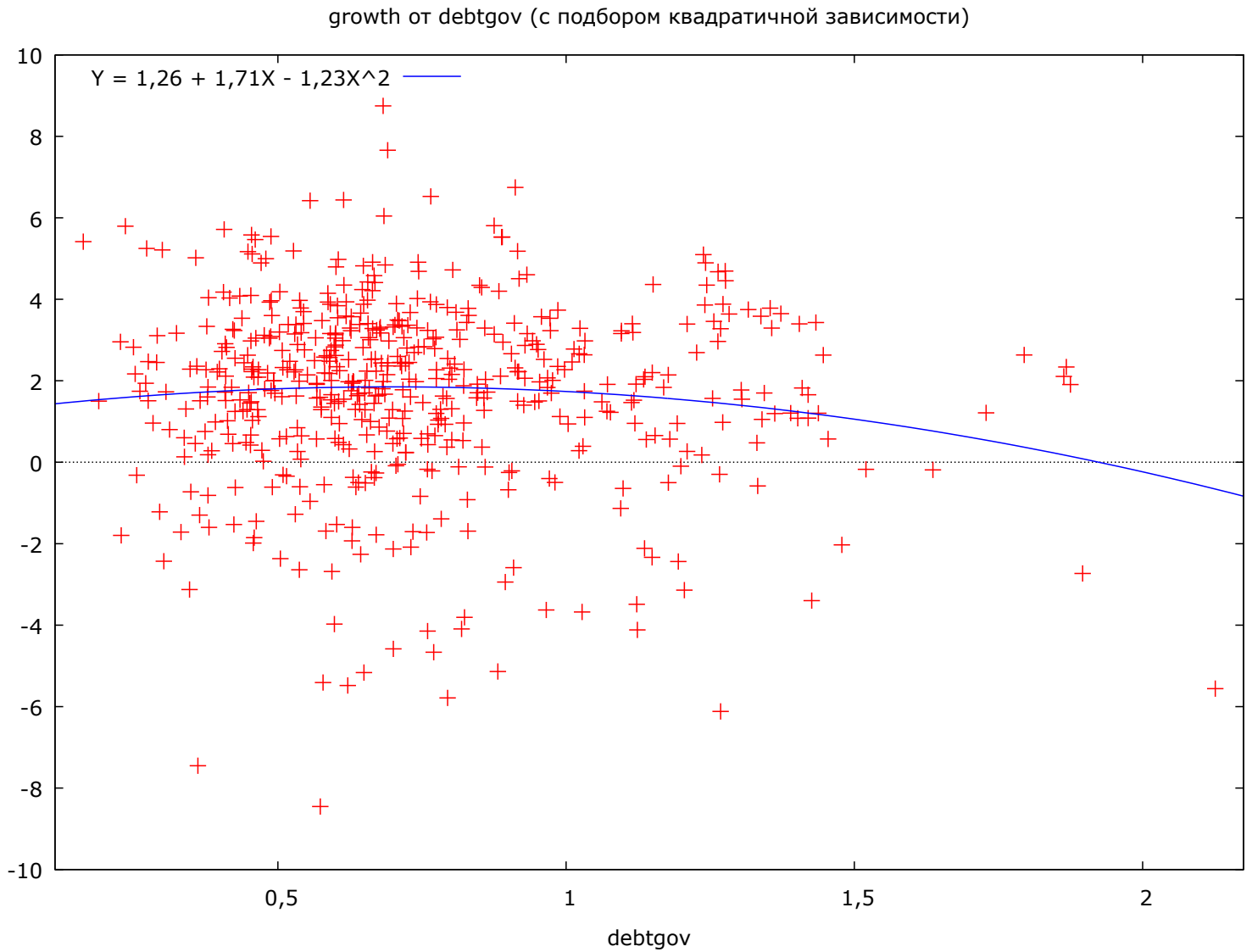
Table 1—Sample Descriptives and Balance on Observables

	Mean, (treatment)	Mean (control)	Difference	Standard error	Observations (treatment)	Observations (control)
Panel A. All students in the baseline sample						
Female	0.76	0.76	0.004	0.034	314	305
Age (years)	12.67	12.41	0.267	0.143	230	231
SES index	-0.03	0.04	-0.07	0.137	314	305
Grade 4	0.01	0.01	-0.003	0.007	305	299
Grade 5	0.01	0.02	-0.007	0.01	305	299
Grade 6	0.27	0.3	-0.035	0.037	305	299
Grade 7	0.26	0.26	0.005	0.036	305	299
Grade 8	0.3	0.28	0.017	0.037	305	299
Grade 9	0.15	0.13	0.024	0.028	305	299
Math	-0.01	0.01	-0.016	0.081	313	304
Hindi	0.05	-0.05	0.096	0.081	312	305
Present at endline	0.85	0.9	-0.048	0.027	314	305
Panel B. Only students present in endline						
Female	0.77	0.76	0.013	0.036	266	273
Age (years)	12.61	12.37	0.243	0.156	196	203

Графики – зачем?
Найти выбросы



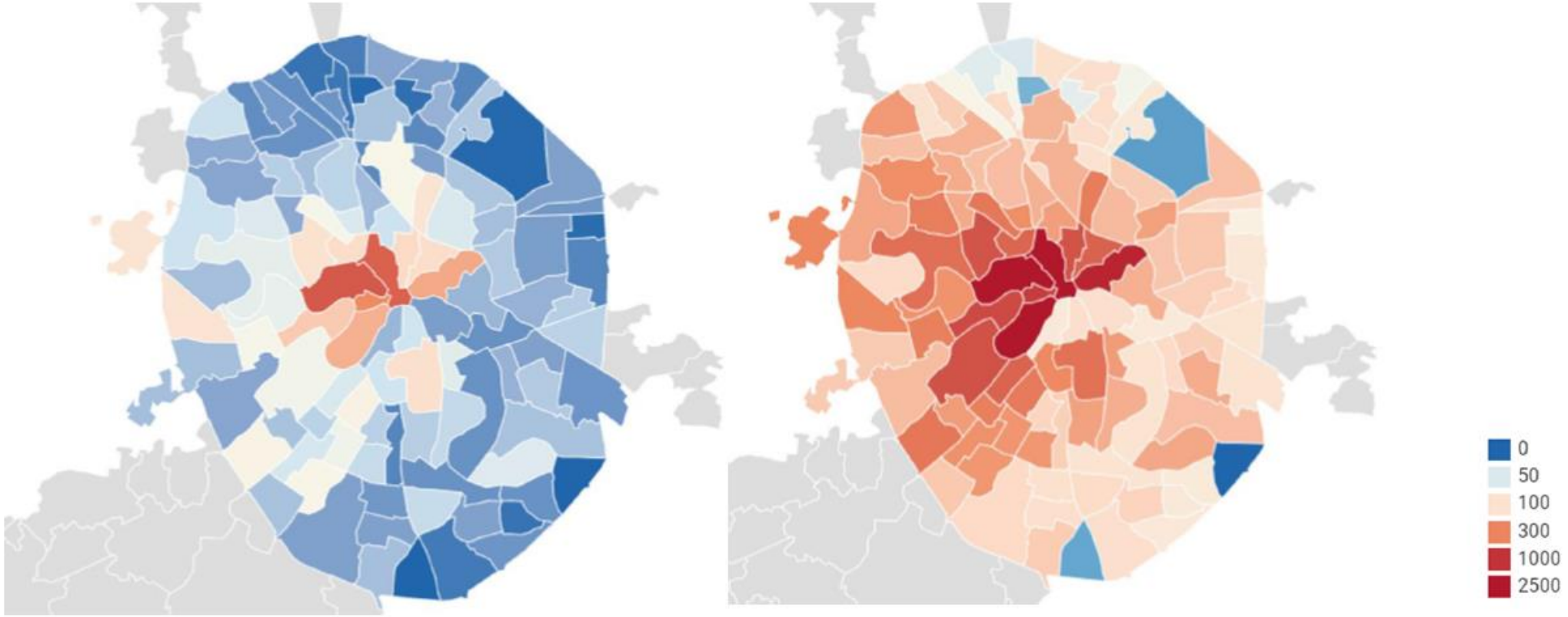
Графики – зачем? Проверка результатов



- Пример: оценка влияния государственного долга на темпы роста ВВП.
- Оценена пороговая модель, рассчитан «порог», а что с данными?
- Есть ли наблюдения выше и ниже «порога»?
- Для каких стран эти точки?
- «Знание фактуры»

Графики – зачем?
Наглядное представление данных

График 2. Карта расположения предложений на онлайн-платформе Airbnb в Москве.



1а: конец 2016

1б: середина 2018

Источник: составлено автором.

Из ВКР Антона Бреннермана, 2020 г.

Сводная таблица с результатами

- Адекватность метода
- Аккуратная интерпретация
- Устойчивость результатов

Таблица 6. Результаты оценивания моделей по Нью-Йорку.

Зависимая переменная: $log(P_{it})$ (индекс аренды Zillow Rent Index)

Количество наблюдений: 5280

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
	(Pooled OLS)	(FE)	(RE)	(FE)	(FE)	(FE+IV)	(FE+IV)
$log\ (1 + Airbnb_{it})$	0,09*** (0,008)	0,02*** (0,004)	0,02*** (0,004)	0,03*** (0,01)	0,03*** (0,01)	0,21*** (0,06)	0,21*** (0,06)
$log(1 + Airbnb_{it}) * ownrate_{i2014}$				– 0,02 (0,02)	– 0,02 (0,02)	– 0,11* (0,06)	– 0,11* (0,06)
$log\ (population_{it})$					0,09 (0,07)		0,14 (0,11)
$log\ (income_{it})$					– 0,08* (0,04)		– 0,03 (0,08)
Константа	7,53*** (0,02)		7,71*** (0,02)				
Фиксированные эффекты времени		✓	✓	✓	✓	✓	✓
Фиксированные эффекты района		✓		✓	✓	✓	✓
Инструментальные переменные						✓	✓
Скорректированный R^2	0,4						
Within R^2		0,3		0,3	0,3	0,3	0,3

Примечание: в скобках под оценками коэффициентов указаны их робастные стандартные ошибки.

Тест на наличие индивидуальных эффектов в модели (2): H_0 – группы имеют общие константы; Welch F(164, 1725,7) = 2783,72; p-значение = 0,0000.

Тест Вальда на наличие временных эффектов в модели (2): H_0 – нет временных эффектов; $\chi^2(31) = 1095{,}66$; p-значение = 0,0000.

Тест Бреуша-Пагана в модели (3): H_0 – дисперсия специфических для наблюдений ошибок отсутствует; $\chi^2(1) = 78376{,}3$; p-значение = 0,0000.

Тест Хаусмана в модели (3): H_0 – ОМНК оценки состоятельны; $\chi^2(1) = 79{,}78$; p-значение = 0,0000.

Источник: составлено автором.

*** $p < 0{,}01$, ** $p < 0{,}05$, * $p < 0{,}1$

Из ВКР Антона Бреннермана, 2020 г.

Результаты эконометрических оценок

Важно «обсудить» результаты:

- «Внешняя валидность»: насколько применим результат вне выборки
- «Робастность»: насколько устойчивы оценки к изменению метода, выборки, периода, спецификации
- «Ограниченность» результатов: насколько было доступны и «чисты» исходные данные, каково альтернативное объяснение результатов?

Эконометрические уравнения без теории и содержательного объяснения механизмов влияния показателей друг на друга – бессмысленны!

Пример: влияние цифровизации экономики на безработицу в регионе М.

Метод оценки - МНК	
Зависимая переменная: perc_unemploy	
const	-93,42**
	(1,9750)
labor_productivity	0,9708**
	(0,0182)
RD_perc_VRP	11,65**
	(0,7319)
proport_innov_prods	-0,3658**
	(0,0508)
d_internet100_access	1,288**
	(0,0409)
d_budget_RD	-0,1520**
	(0,0159)
n	10
R^2	0,9619
lnL	-3,61
Исправленный R^2	0,89

* обозначает значимость на 10-процентном уровне

** обозначает значимость на 5-процентном уровне

Зависимая переменная – уровень безработицы,

регрессоры:

- производительность труда,
- расходы на R&D в % от валового регионального продукта,
- доля инновационных продуктов,
- доступ к интернету,
- изменение бюджета на расходы на R&D.

Проблема?

Пример – внимательно читайте, перепроверяйте, что написали

Проблема?

Таблица 10. Модель №2, предельные эффекты.

Переменная	Предельный эффект	Стандартная ошибка	t-значение	p-значение	
(Intercept)	-0.018	0.253	-0.071	0.944	
Internship_practice	-0.214	1.666	-0.129	0.898	**
Own_business	0.249	5.821	0.043	0.966	
Graduate	0.309	8.037	0.038	0.969	.
Specialist	0.357	10.328	0.035	0.972	*
Full.time.education	-0.216	4.407	-0.049	0.961	*
Nothing	-0.662	10.169	-0.065	0.948	
Only.work	-0.408	2.196	-0.186	0.853	.
Study.Remote.work	0.190	3.656	0.052	0.959	*
Freelance-1	0.394	8.972	0.044	0.965	***
Own.business	0.223	4.481	0.050	0.960	**
Russian.state.company	-0.353	0.257	-1.372	0.171	***
Architecture..Design	0.522	8.713	0.060	0.952	

Пример – внимательно читайте, перепроверяйте, что написали

Проблемы

Таблица поехала

Не соответствуют р-значения и «звёздочки»

2 почти одинаковых не расшифрованных показателя

Таблица 10. Модель №2, предельные эффекты.

Переменная	Предельный эффект	Стандартная ошибка	t-значение	p-значение	
(Intercept)	-0.018	0.253	-0.071	0.944	
Internship_practice	-0.214	1.666	-0.129	0.898	**
Own_business	0.249	5.821	0.043	0.966	
Graduate	0.309	8.037	0.038	0.969	.
Specialist	0.357	10.328	0.035	0.972	*
Full.time.education	-0.216	4.407	-0.049	0.961	*
Nothing	-0.662	10.169	-0.065	0.948	
Only.work	-0.408	2.196	-0.186	0.853	.
Study.Remote.work	0.190	3.656	0.052	0.959	*
Freelance-1	0.394	8.972	0.044	0.965	***
Own.business	0.223	4.481	0.050	0.960	**
Russian.state.company	-0.353	0.257	-1.372	0.171	***
Architecture..Design	0.522	8.713	0.060	0.952	

Пример: оценка реализации потенциала торговли внутри СНГ

Куда / Откуда	ARM	BLR	KAZ	KGZ	RUS
ARM		1,77	6,30	15,04	0,54
BLR	3,57		2,94	1,34	0,92
KAZ	3,14	0,62		1,22	1,16
KGZ	4,29	0,45	1,08		0,97
RUS	0,49	1,10	1,24	0,91	

- Показатель реализации потенциала торговли - это отношение расчётного и фактического значений экспорта
- Почему результаты именно такие? Как можно объяснить «подозрительные» цифры?

1) Ограничение модели - не учтен торговый партнёр- КНР
2) Экспорт из Армении в Киргизию – выброс (см. описательную статистику!)

Описательная статистика: экспорт

- Для выборки стран СНГ: 292 нулевых значения экспорта

	Экспорт (ЕАЭС)	Экспорт (СНГ)
Минимум	854 долл. (KGZ -> ARM, 2005)	0
Среднее	2,5 млрд долл.	0,72 млрд долл.
Медиана	0,18 млрд долл.	0,07 млрд долл.
Максимум	24,93 млрд долл.	24,93 млрд долл.

Где почитать про оформление результатов работы?

- Положение о ВКР экономического факультета МГУ:

https://www.econ.msu.ru/sys/raw.php?o=63865&p=attachment

- Анна Малькова – о хорошей и плохой визуализации: https://www.econ.msu.ru/sys/raw.php?o=53608&p=attachment
- Анатольев (2008), «Оформление эконометрических отчетов», Квантиль №4 04-SA.pdf (quantile.ru)