

**Демонстрационная версия
контрольной работы**

Задание 1.

1. Выберите из нижеприведенных матриц подходящую на роль Hat-matrix (матрицы проекции) в контексте парной линейной регрессии, включающей константу. Свой ответ обоснуйте. (2 балла)

$$A = \begin{pmatrix} 0.2 & 0.2 & 0.2 & 0.2 & 0.2 \\ 0.2 & 0.45 & 0.45 & -0.05 & -0.05 \\ 0.2 & 0.45 & 0.45 & -0.05 & -0.05 \\ 0.2 & -0.05 & -0.05 & 0.45 & 0.45 \\ 0.2 & -0.05 & -0.05 & 0.45 & 0.2 \end{pmatrix}; \quad B = \begin{pmatrix} 0.2 & 0.2 & 0.2 & 0.2 & 0.2 \\ 0.2 & 0.45 & 0.45 & -0.05 & -0.05 \\ 0.2 & 0.45 & 0.45 & -0.05 & -0.05 \\ 0.2 & 0.45 & -0.05 & 0.45 & -0.05 \\ 0.2 & -0.05 & -0.05 & 0.45 & 0.45 \end{pmatrix};$$
$$C = \begin{pmatrix} 0.2 & 0.2 & 0.2 & 0.2 & 0.2 \\ 0.2 & 0.45 & 0.45 & -0.05 & -0.05 \\ 0.2 & 0.45 & 0.45 & -0.05 & -0.05 \\ 0.2 & -0.05 & -0.05 & 0.45 & 0.45 \\ 0.2 & -0.05 & -0.05 & 0.45 & 0.45 \end{pmatrix}; \quad D = \begin{pmatrix} 0.4 & 0.2 & 0.2 & 0.2 & 0.2 \\ 0.2 & 0.4 & 0.45 & -0.05 & -0.05 \\ 0.2 & 0.45 & 0.4 & -0.05 & -0.05 \\ 0.2 & -0.05 & -0.05 & 0.4 & 0.45 \\ 0.2 & -0.05 & -0.05 & 0.45 & 0.4 \end{pmatrix};$$

2. По выбранной матрице определите наблюдение(-я) с наиболее высоким потенциалом влияния. Объясните своими словами, что означает потенциал влияния

Задание 2.

1. Отметьте ВСЕ верные утверждения, если таковые имеются. Если верных утверждений нет, то напишите в ответе «НЕТ» (1 балл)

- (a) С ростом R^2 из вспомогательной регрессии, построенной для диагностики мультиколлинеарности, вариация оценки соответствующего коэффициента возрастает линейным образом
- (b) В условиях гетероскедастичности теряется эффективность МНК-оценок параметров регрессии
- (c) Выброс – нетипичное наблюдение по объясняющей переменной в регрессионной модели
- (d) У матрицы с линейно зависимыми строками обратная матрица также имеет линейно зависимые строки

2. Ответьте на вопросы ниже:

- 1) Объясните, в чем разница между мерой Кука и мерой DFBETA, что они показывают?
- 2) Выведите формулу в векторно-матричном виде для оценки дисперсии $\hat{\beta}$ в условиях гомоскедастичности и отсутствия автокорреляции: приведите пошаговое выведение и итоговую формулу. Что будет меняться в случае гетероскедастичности?

Задание 3. Ниже представлены результаты оценивания линейной регрессионной модели. В качестве зависимой переменной используется переменная `ch_schools_pc`.

Ниже представлено краткое описание переменных:

<code>ch_schools_pc</code>	Изменение в количестве сельских школ с 1860 до 1880 гг. на душу сельского населения уезда
<code>afreq</code>	Доля лет между 1851 и 1863 гг., в которые были зафиксированы крестьянские выступления
<code>nozemstvo</code>	Бинарная переменная: Единицей закодированы уезды тех губерний, в которых в результате реформы 1864 года земства созданы не были, 0 – в противном случае.
<code>distance_moscow</code>	Расстояние от Москвы до центра уезда
<code>goodsoil</code>	Показатель плодородности почвы
<code>lnurban</code>	Логарифм городского населения уезда на 1863 г.
<code>lnpopn</code>	Логарифм населения уезда на 1863 г.
<code>province_capital</code>	Бинарная переменная: принимает значение 1, если в уезде находился «столичный» город губернии, 0 – в противном случае.

	coef	std. error	t	Pr> t	[0.025; 0.975]
(Intercept)	0.676390	0.218253			
<code>afreq</code>	-0.179940	0.054391			
<code>nozemstvo</code>	0.081681	0.021824			
<code>distance_moscow</code>	-0.012284	0.031880			
<code>goodsoil</code>	-0.009406	0.024005			
<code>lnurban</code>	0.013754	0.007281			
<code>lnpopn</code>	-0.042032	0.019883			
<code>province_capital</code>	0.038771	0.030189			

ANOVA					
	sum_sq	df	mean_sq	f	PR(>F)
Regression	1.7535				
Residual	15.2350	480			

- Проверьте гипотезу о незначимости коэффициента при предикторе `afreq` против двусторонней альтернативы. Запишите нулевую гипотезу и альтернативу, рассчитайте статистику критерия, укажите примерное значение *p-value* и сделайте вывод. Так как в данном случае выборка достаточно большая, Вы можете использовать при расчете *p-value* нормальную аппроксимацию
- Рассчитайте коэффициент детерминации и проинтерпретируйте полученное значение
- Проверьте гипотезу, что регрессия на константу не хуже модели с предикторами, на фиксированном уровне значимости 0.05. Запишите значение статистики и ее промежуточные расчеты, а также выберите необходимую критическую точку – квантиль. Сделайте вывод
 - квантиль хи-квадрат распределения уровня 0.95, $df=480$: **532.075**
 - квантиль распределения Фишера уровня 0.95, $df1=8$, $df2=480$: **1.958**
 - квантиль распределения Фишера уровня 0.95, $df1=7$, $df2=480$: **2.029**
 - квантиль распределения Фишера уровня 0.975, $df1=8$, $df2=480$: **2.218**
 - квантиль распределения Фишера уровня 0.975, $df1=7$, $df2=480$: **2.314**
- В представленную спецификацию модели в качестве предиктора была добавлена дамми-переменная `zemstvo`, принимающая значение 1 при наличии земства, 0 – в противном случае. Кроме добавления этого предиктора спецификация модели осталась без изменений. Какие результаты (что произойдет с оценками коэффициентов модели) по итогам оценивания такой дополненной модели будут получены? Свой ответ объясните

Задание 4. Ниже представлены оценки регрессионной модели. Зависимая переменная – доля граждан, имеющих наиболее высокий уровень удовлетворенности жизнью. Качество институтов (исходный показатель) измеряется в непрерывной шкале от 1 до 5, где 5 соответствует максимальному значению качества институтов. В модели используется преобразованное значение качества институтов: разница между исходным и максимальным значением качества институтов по выборке (Inst_dif). Исследователь сравнивает западноевропейские и латиноамериканские страны. Для групп стран введена дамми-переменная (LA), которая принимает значение 0, если страна – западноевропейская, значение 1 – для латиноамериканской страны.

Life Satisfaction	
Inst_dif	0.48*** (5.2)
LA	0.163*** (6.23)
LA × Inst_dif	0.04*** (4.24)
Intercept	0.3*** (9.53)

t-statistics are given in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

1. Проинтерпретируйте оценку коэффициента при переменной LA
2. Проинтерпретируйте оценку коэффициента при переменной взаимодействия
3. Рассчитайте значение предельного эффекта качества политических институтов (Inst_dif) в случае, если рассматривается западноевропейская страна

Задание 5.

Ниже представлены результаты применения метода главных компонент. Исходные индикаторы: X, Y, Z.

	PC1	PC2	PC3
X	0.5884	−0.4993	0.6360
Y	0.6129	−0.2377	−0.7536
Z	0.5274	−0.8332	0.1662
Variance	2.5149	0.4305	0.0545

1. Рассчитайте информативность **первой** главной компоненты?
2. Сколько главных компонент необходимо извлечь на основании критерия Кайзера? Свой ответ поясните

Задание 6. На основании представленных ниже данных найдите веса для построения синтетического контроля и рассчитайте средний эффект воздействия, если город А – объект воздействия, остальные города составляют контрольную группу. 2018, 2019 гг. - период до воздействия.

Год	Город А	Город В	Город С
2018	26.1	25.8	26.3
2019	26.3	26.0	26.4
2020	26.0	26.2	26.6
2021	25.7	26.3	26.8
2022	25.4	26.5	27.0
2023	25.2	26.6	27.1

Задание 7. Можно ли сказать, что африканские страны, обозначенные на графике ниже, являются нетипичными наблюдениями? Если да, то конкретизируйте, являются ли данные страны нетипичными по предиктору (*average church attendance*) и / или зависимой переменной (*volunteers*)? Свой ответ поясните. Как их присутствие в выборке отразится на результатах оценивания регрессионной модели *volunteers* на *average church attendance*? (1 балл)

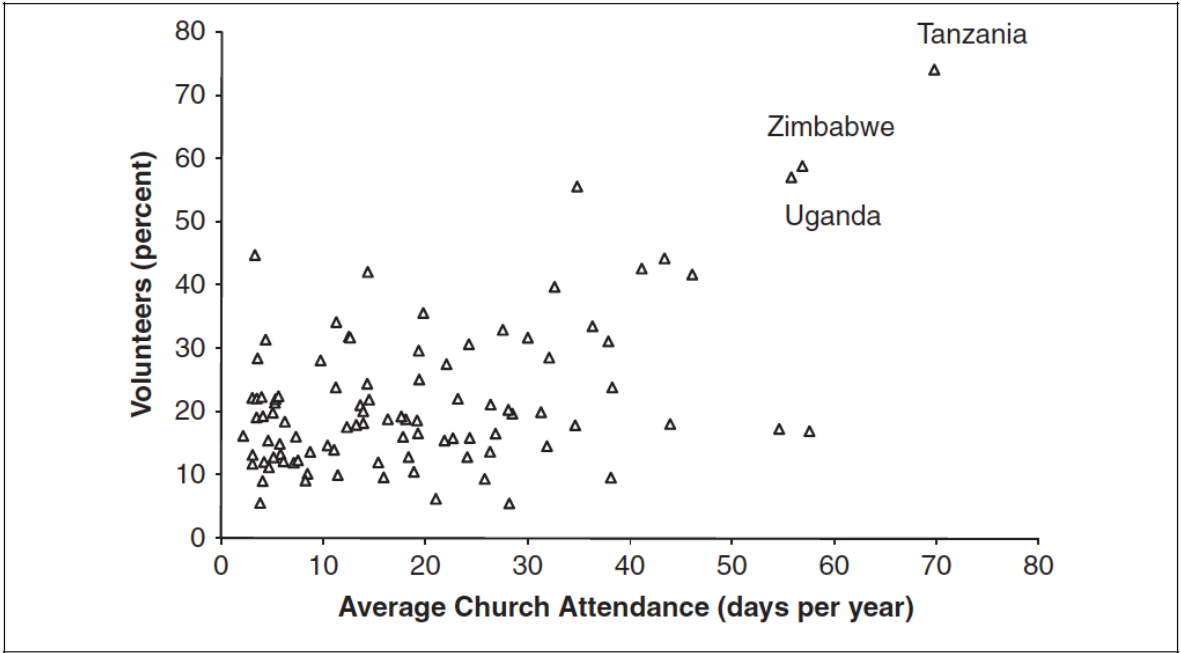


Figure 1. Scatter Plot for Average Church Attendance and Percentage Volunteers, in 96 Surveys Conducted in 53 Countries during Three Waves

Задание 8. Изучается эффект расширения налогового кредита на заработанный доход (EITC) в США в 1993 г. для семей с двумя детьми и более на занятость среди женщин. Рассматривается 3 года до и после расширения EITC. Дамми-переменная «Период» закодирована следующим образом: 1 – период после расширения EITC, 0 – период до соответствующей реформы. Семьи, в которых не было детей или только один ребенок, составляют контрольную группу. Ниже представлены результаты оценивания модели разность разностей (DiD: difference-in-differences). Большее значение зависимой переменной соответствует более высокой занятости среди женщин.

	Зависимая переменная:
	Занятость среди женщин
Группа воздействия	−0.129*** (0.012)
Период	−0.002 (0.0003)
Группа воздействия × Период	0.047** (0.017)
Константа	0.575*** (0.09)
Количество наблюдений	13,746
Примечание: В скобках приведены стандартные ошибки.	

По представленной выдаче

1. Вычислите среднее значение занятости среди женщин с двумя детьми и более в период после введения расширения EITC
2. В предположении о соблюдении допущения параллельности трендов рассчитайте значение гипотетического исхода (counterfactual outcome) в группе воздействия в период после расширения EITC. Своими словами объясните, что показывает counterfactual outcome в контексте модели DiD