Задания для дополнительной тренировки

Задание 1.

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. Сформулируйте критерий Кайзера для определения количества извлекаемых главных компонент. Сколько, согласно данному критерию, главных компонент необходимо извлечь в данном случае?

3. Чему равен след (trace) ковариационной матрицы исходных переменных X, Y, Z (то есть, сумма элементов матрицы по главной диагонали)? Проинтерпретируйте это значение.

Задание 2. Известно, что настоящая модель регрессии выглядит следующим образом: $y_i = \beta_0 + \beta_1 x_i + \beta_2 z_i + \varepsilon_i$, и ее МНК-оценки являются BLUE. Вы оцениваете модель $y_i = \theta_0 + \theta_1 x_i + \eta_i$. Чем грозит исключение из модели релевантного регрессора z? В каком случае при исключении предиктора оценки останутся BLUE?

Задание 3. Изучается влияние факта, занимается ли школьник у репетитора, на его успеваемость. У исследователя возник вопрос, нужно ли включать среднюю успеваемость в его классе в качестве контрольной переменной. Всего в классе 6 учащихся. Рассмотрите случаи как включения средней успеваемости в качестве контрольной переменной, так и невключения. Какие могут быть последствия в случае первого и второго решения? Свой ответ объясните.

Задание 4. Ознакомьтесь с постановочной частью исследования Lü, Landry, 2014.

1. Проинтерпретируйте оценки коэффициентов Table 2, Model 1.

2. Проинтерпретируйте график Figure 4, Panel 2. Что можно сказать о значимости предельного эффекта?

Задание 5. Ниже представлены результаты оценивания линейной регрессионной модели процента детей, получающих образование, (eduger), на социально-экономические показатели. *polityIV* — индекс демократии Polity IV (от -10 до 10); *gini_8090* — средний индекс имущественного неравенства в 1980 и 1990 г. (от 0 до 100); *ssw* — консолидированные бюджетные расходы на социальное обеспечение (% ВВП); *prot*80 — процент протестантов от численности населения в 1980 г.

```
               coef    std. error   t   Pr>|t|  [0.025; 0.975]
(Intercept)  59.0982     9.2578
polityIV      1.0462     0.4522
gini_8090     0.2556     0.1952
ssw           1.4918     0.3066
prot80        0.1016     0.0576
---

                                ANOVA
---------------------------------------------------------------------
               sum_sq        df    mean_sq    f    PR(>F)
---------------------------------------------------------------------
Regression
Residual       5665.644
Total         14178.832      57
```

1. Проверьте гипотезу о незначимости коэффициента при предикторе prot80 против двусторонней альтернативы на фиксированном уровне значимости 0.05. Отметьте необходимое значение критической точки из представленного ниже списка

   (a) квантиль распределения Стьюдента, 0.95, df = 53: 1.674

   (b) квантиль распределения Стьюдента, 0.95, df = 57: 1.672

   (c) квантиль распределения Стьюдента 0.975, df = 53: 2.006

   (d) квантиль распределения Стьюдента 0.975, df = 57: 2.002

   Ответ

   В данном случае статистика критерия при верной нулевой гипотезе имеет распределение Стьюдента с $df = n - k - 1$. Таким образом, df = 58 - 4 - 1 = 53. Для построения критической области выбираем пункт c: квантиль распределения Стьюдента уровня 0.975 (так как двусторонняя альтернатива), df = 53.

   Статистика критерия:

   $$\frac{\hat{\beta}}{se} = \frac{0.1016}{0.0576} \approx 1.764$$

   Значение статистики критерия попадает в доверительную область, следовательно, нулевую гипотезу не отвергаем. Данный коэффициент является незначимым.

2. Постройте 95%-ый доверительный интервал для коэффициента при предикторе prot80

   Ответ

   $$0.1016 - 0.0576 \times 2.002 \leq \beta_{prot80} \leq 0.1016 + 0.0576 \times 2.002$$
   $$-0.0137 \leq \beta_{prot80} \leq 0.2169$$

3. Проинтерпретируйте оценку коэффициента при предикторе *ssw*

   Ответ

   При прочих равных условиях в стране, в которой консолидированные бюджетные расходы на социальное обеспечение выше на 1%, процент детей, получающих образование, в среднем выше на 1.49.

4. Рассчитайте коэффициент детерминации и проинтерпретируйте полученное значение

   Ответ

$$R^2 = 1 - \frac{5665.644}{14178.832} \approx 0.6$$

   Моделью объясняется порядка 60% вариации зависимой переменной.

5. Проверьте гипотезу, что регрессия на константу не хуже модели с предикторами, на фиксированном уровне значимости 0.01. Запишите значение статистики и ее промежуточные расчеты, а также выберите необходимую критическую точку – квантиль. Сделайте вывод.

   (a) квантиль хи-квадрат распределения уровня 0.99, df= 57: **84.733**
   (b) квантиль хи-квадрат распределения уровня 0.01, df = 57: **35.131**
   (c) квантиль распределения Фишера уровня 0.99, df1 = 5, df2 = 53: **3.384**
   (d) квантиль распределения Фишера уровня 0.99, df1 = 4, df2 = 53: **3.695**
   (e) квантиль распределения Фишера уровня 0.01, df1 = 5, df2 = 53: **0.108**
   (f) квантиль распределения Фишера уровня 0.01, df1 = 4, df2 = 53: **0.073**

   Ответ

   Подходящее значение квантили - пункт d: квантиль распределения Фишера уровня 0.99, df1 = 4, df2 = 53.

   Рассчитаем статистику критерия:

$$\frac{\frac{R^2}{k}}{\frac{1-R^2}{n-k-1}} \approx 19.9$$

   Значение статистики попадает в критическую область, следовательно, данная модель лучше, чем модель на константу.

Задание 6. 1. Модель, представленная в предыдущем задании, была проверена на гетероскедастичность посредством теста Уайта. Запишите количество параметров в дополнительной (вспомогательной) модели, которую необходимо построить на промежуточном этапе реализации теста Уайта

   Ответ

   15 параметров (1 константа + 4 коэффициента при исходных предикторах + 4 коэффициента при предикторах в квадрате + 6 коэффициентов при попарных произведениях)

2. Порассуждайте, какие могут быть источники гетероскедастичности в данном случае? Приведите не менее двух, свой ответ поясните

   Ответ

   К примеру, можно предположить скошенность распределений для социально-экономических показателей, наличие нетипичных наблюдений

Задание 7. Ниже представлена таблица, представляющая результаты исследования S. Galiani, E. Schargrodsky «Effects of Land Titling on Child Health». Ознакомьтесь также с аннотацией данной статьи:

This paper examines the impact of urban land titling on child health. We hypothesize that land titling may translate into positive effects on health through its impact on housing investments and household structure. To address selection concerns, we take advantage of a natural experiment of land occupation in a suburban area of Buenos Aires, Argentina, that ensures that the allocation of property rights is exogenous to the characteristics of the squatters.

Weight-for-height Z-score

Age group	Property right offer = 1	Property right offer = 0	Difference
All (0–11 years old)	0.279 (0.065) [239]	0.065 (0.087) [132]	0.214 (0.109)
0–4 years old	0.110 (0.100) [114]	−0.028 (0.119) [60]	0.139 (0.156)
5–11 years old	0.434 (0.083) [125]	0.143 (0.126) [72]	0.291 (0.151)

Note: Weight-for-height Z-scores are available for boys through 138 months of age and less than 145 cm of height, and girls through 120 months of age and less than 137 cm of height. The sample is restricted to children born after 1991. Standard errors in parentheses. Number of observations in brackets.

Предложите спецификацию регрессионной модели, которая позволила бы оценить различия в эффекте получения прав собственности на здоровье детей между возрастными группами от 0 до 4 лет и от 5 до 11 лет.

Задание 8. Объясните нижеприведенное утверждение. А также укажите, на что в таком случае должен опираться исследователь?

«The analyst cannot infer whether X has a meaningful conditional effect on Y from the magnitude and significance of the coefficient on the interaction term.»

Задание 9. На основе представленных ниже данных оценивается парная линейная регрессия y на x.

y	x
7	2
4	4
5	3
10	6
9	7
12	9
8	3
9	6

1. Чему равен коэффициент корреляции Пирсона между зависимой переменной и предсказанным на основе модели (см. условие выше) значением отклика? Ответьте на поставленный вопрос без предварительных расчетов предсказанных значений зависимой переменной. Свое решение обоснуйте

2. Рассчитайте коэффициент детерминации для указанной регрессионной модели. Проинтерпретируйте полученное значение

3. Проверьте значимость коэффициента детерминации

Задание 10.

1. По указанным ниже значениям предиктора и отклика (x и y соответственно) найдите вектор оценок коэффициентов в регрессионной модели y на x. Используйте для этого общую формулу оценки коэффициентов в векторно-матричном виде (релевантную как для парной, так и для множественной регрессии). Запишите промежуточные расчеты. В качестве ответа запишите сам вектор и полную спецификацию модели, подставив эти оценки коэффициентов

x	2	5	2	0	1
y	10	3	1	12	5

   Ответ

   $\hat{\beta}_0 = 9.057$; $\hat{\beta}_1 = -1.429$

2. Рассчитайте предсказанное значение зависимой переменной для первого наблюдения

   Ответ

   $\hat{y_1} = 9.057 - 1.429 \times 2 = 6.199$

3. Определите, есть ли нетипичные наблюдения по X, определив потенциал влияния каждого наблюдения

Задание 11.

Ниже представлены оценки регрессионной модели. Зависимая переменная – доля граждан, имеющих наиболее высокий уровень удовлетворенности жизнью. Качество институтов (исходный показатель) измеряется в непрерывной шкале от 0 до 5, где 5 соответствует максимальному значению качества институтов. В модели используется преобразованное значение качества институтов: центрированный показатель (Inst_c). 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_c	0.48***
	(5.2)
LA	0.163***
	(6.23)
LA× Inst_c	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.

   Ответ

   В странах Латинской Америки в среднем удовлетворенность жизнью выше на 0.163 по сравнению с западноевропейскими странами при условии, что показатель качества институтов принимает среднее значение.

2. Проинтерпретируйте оценку коэффициента при переменной взаимодействия.

   Ответ

   В странах Латинской Америки взаимосвязь качества институтов и удовлетворенности жизнью более выраженная положительная по сравнению с западноевропейскими странами.

3. Рассчитайте значение предельного эффекта качества политических институтов в случае, если рассматривается латиноамериканская страна.

   Ответ

   0.48+0.04 = 0.52

Задание 12. Ниже представлена спецификация модели, отражающая взаимосвязь трат на еду и месячного дохода. Переменные измерены в единицах – «фунтиках».

$$food\_exp_i = 20 + 2.5income_i - 0.25income_i^2 + \epsilon_i$$

1. Рассчитайте предельный эффект месячного дохода на траты на еду, если месячный доход составляет 10 «фунтиков».

   Ответ

   $2.5 - 0.5 \times 10 = -2.5$

2. Проинтерпретируйте оценку коэффициента при предикторе $income^2$

   Ответ

   С ростом месячного дохода на один фунтик эффект дохода на траты на еду уменьшается в среднем на 0.5 (то есть, сначала положительный эффект ослабляется, затем эффект становится отрицательным).

3. Вычислите «переломную» точку (значение показателя *income*), после которой характер взаимосвязи *income* и зависимой переменной меняет знак. В ответе представьте необходимые расчеты.

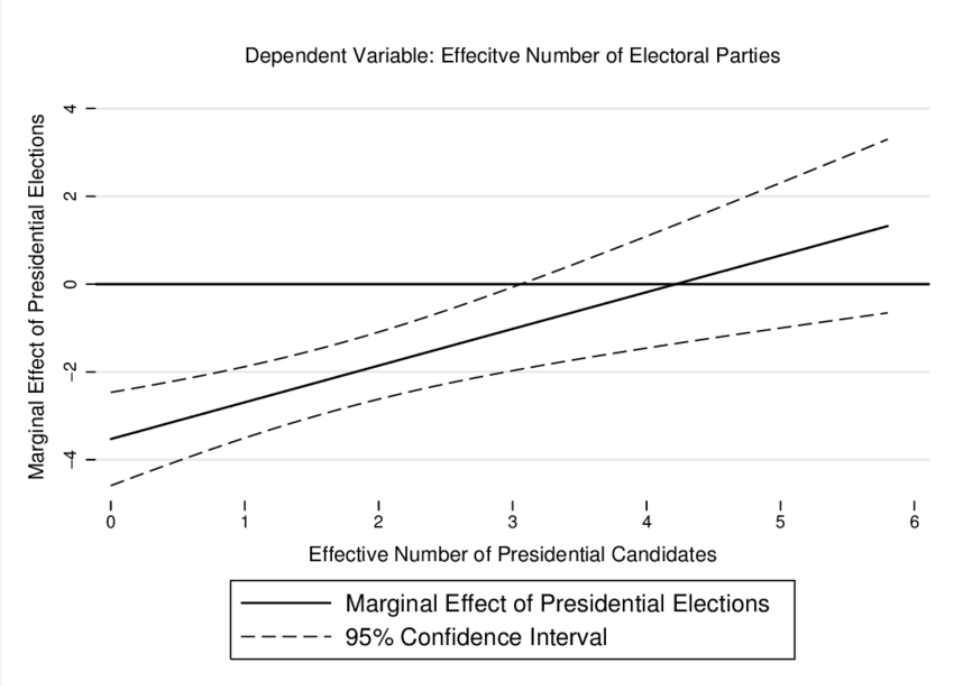
   Ответ

   $$2.5 - 0.5 \times income = 0$$

   $$income = 5$$

   Начиная с 6 фунтиков эффект становится отрицательным.

Задание 13. На основе оценивания регрессионной модели с переменной взаимодействия был построен следующий график зависимости предельного эффекта ключевого предиктора от значений модератора:



При каких значениях модератора предельный эффект является значимым (уровень доверия 0.95)? Если таковых значений на графике нет, напишите об этом, кратко пояснив свой ответ.

Ответ

Если модератор принимает значения от 0 до 3, то эффект является значимым.