Домашнее задание №3.

Домашнее задание сдается в электронном виде в SmartLMS.

Срок сдачи: 23:59, 15 декабря 2024 г.

На титульном листе обязательно указать:

Ф.И.О., номер варианта.

Задача 1. (30 баллов)

Рассмотрим модель множественной регрессии вида

$$y = X\beta + \varepsilon,$$

где X — $n \times k$ детерминированная матрица с рангом равным k, $\mathbb{E}(\varepsilon) = 0$, $\mathrm{Var}(\varepsilon) = \Omega$. Обозначим через $\hat{\beta}_{OLS}$ и $\hat{\beta}_{GLS}$ МНК-оценку и ОМНК-оценку для вектора параметров β. Покажите, что если столбцы матрицы X являются собственными векторами ковариационной матрицы Ω, то

(a) **(15 баллов)** аналитические выражения для $\hat{\beta}_{OLS}$ и $\hat{\beta}_{GLS}$ совпадут;

(b) **(15 баллов)** аналитические выражения для ковариационных матриц оценок $\hat{\beta}_{OLS}$ и $\hat{\beta}_{GLS}$ также совпадут.

Задача 2. (70 баллов)

Домашнее задание основано на результатах опроса населения РМЭЗ НИУ ВШЭ в 2020, 2022 и 2023 годах (https://www.hse.ru/rlms/spss). В файле $Homework\_2\_data.csv$ (файл CSV) содержатся следующие переменные:

- wage — заработная плата, полученная за последние 30 дней по основному месту работы после удержания налогов в рублях;
- educ — уровень образования, категориальная переменная (0 для индивидов, учившихся в школе):
  1. ПТУ, техническое училище
  2. институт, университет, академия
- female = 1, если респондент – женщина, = 0 для мужчин;
- age — возраст в годах;
- is_children = 1, если у респондента есть хотя бы 1 ребенок, = 0 иначе;
- work_hours — количество часов, которое продолжается рабочий день;

- foreign_language = 1, если респондент знает ли иностранный язык, = 0 иначе;
- internet = 1, если респонденту приходилось в течение последних 12 месяцев пользоваться Интернетом, = 0 иначе;
- alcohol = 1, если респондент употребляет алкогольные напитки (хотя бы изредка), = 0 иначе;
- health = 1, если респондент испытывал проблемы со здоровьем за последний месяц, = 0 иначе;
- weight — вес респондента в кг;
- height — рост респондента в см;
- smoke = 1, если респондент курит, = 0 иначе;
- industry — отрасль занятости:

1. ЛЕГКАЯ, ПИЩЕВАЯ ПРОМЫШЛЕННОСТЬ
2. ГРАЖДАНСКОЕ МАШИНОСТРОЕНИЕ
3. ВОЕННО-ПРОМЫШЛЕННЫЙ КОМПЛЕКС
4. НЕФТЕГАЗОВАЯ ПРОМЫШЛЕННОСТЬ
5. ДРУГАЯ ОТРАСЛЬ ТЯЖЕЛОЙ ПРОМЫШЛЕННОСТИ
6. СТРОИТЕЛЬСТВО
7. ТРАНСПОРТ, СВЯЗЬ
8. СЕЛЬСКОЕ ХОЗЯЙСТВО
9. ОРГАНЫ УПРАВЛЕНИЯ
10. ОБРАЗОВАНИЕ
11. НАУКА, КУЛЬТУРА
12. ЗДРАВООХРАНЕНИЕ
13. АРМИЯ, МВД, ОРГАНЫ БЕЗОПАСНОСТИ
14. ТОРГОВЛЯ, БЫТОВОЕ ОБСЛУЖИВАНИЕ
15. ФИНАНСЫ
16. ЭНЕРГЕТИЧЕСКАЯ ПРОМЫШЛЕННОСТЬ
17. ЖИЛИЩНО-КОММУНАЛЬНОЕ ХОЗЯЙСТВО
18. ОПЕРАЦИИ С НЕДВИЖИМОСТЬЮ
19. СОЦИАЛЬНОЕ ОБСЛУЖИВАНИЕ

20. ЮРИСПРУДЕНЦИЯ
21. ЦЕРКОВЬ
22. ХИМИЧЕСКАЯ ПРОМЫШЛЕННОСТЬ
23. ДЕРЕВООБРАБАТЫВАЮЩАЯ ПРОМЫШЛЕННОСТЬ, ЛЕС
24. СПОРТ, ТУРИЗМ, РАЗВЛЕЧЕНИЯ
25. УСЛУГИ НАСЕЛЕНИЮ
26. IT, ИНФОРМАЦИОННЫЕ ТЕХНОЛОГИИ
27. ЭКОЛОГИЯ, ЗАЩИТА ОКРУЖАЮЩЕЙ СРЕДЫ
28. ОРГАНИЗАЦИЯ ОБЩЕСТВЕННОГО ПИТАНИЯ
29. СМИ, ИЗДАТЕЛЬСТВО, ПЕЧАТЬ, ТЕЛЕКОММУНИК
30. РЕКЛАМА, МАРКЕТИНГ
31. ОБЩЕСТВЕННЫЕ ОРГАНИЗАЦИИ, СОВЕТ ВЕТЕРАН

- regions — регион проживания респондента;
- year — год проведения опроса;

Используя выбранные для Вас данные (см. год, отрасль и регион в ведомости), выполните приведенные ниже упражнения.

Для отбора варианта — исполните следующий код, вставив вместо пропусков строки с назначенными Вам годом, отраслью или регионом.

```
import pandas as pd
df_hw = pd.read_csv('Homework_2_data.csv')
year = # your text
industry = # your text
region = # your text
if industry == '       ':
    my_data = (df_hw[(df_hw.year == year) &
                     (df_hw.region == region)])
elif region == '       ':
    my_data = (df_hw[(df_hw.year == year) &
                     (df_hw.industry == industry)])
my_data
```

В случае сдачи работы не своего варианта (расчеты не будут соответствовать выборке из варианта), оценка за работу составит 0 баллов.

Результаты всех тестов должны быть проинтерпретированы (отвергается или не отвергается гипотеза, на каком уровне значимости и что это значит). Задания в тексте обязательно должны быть пронумерованы, согласно пунктам ниже.

1. **(3 балла)** Еще раз из предыдущего домашнего задания: оцените линейную модель, которая объясняет заработную плату ($wage$) возрастом (age), наличием высшего образования ($high$), полом ($female$), наличием детей ($is\_children$), курением ($smoke$) и константой. Проинтерпретируйте полученные результаты. Все ли коэффициенты оказались значимы? Выпишите уравнение оцененной модели.
2. **(3 балла)** Оцените полулогарифмическую модель, которая объясняет логарифм заработной платы ($\ln(wage)$) возрастом (age), наличием высшего образования ($high$), полом ($female$), наличием детей ($is\_children$), курением ($smoke$) и константой. Проинтерпретируйте полученные результаты. Все ли коэффициенты оказались значимы? Выпишите уравнение оцененной модели.
3. **(3 балла)** Оцените линейную в логарифмах модель, которая объясняет логарифм заработной платы ($\ln(wage)$) логарифмом возраста ($ln(age)$), наличием высшего образования ($high$), полом ($female$), наличием детей ($is\_children$), курением ($smoke$) и константой. Проинтерпретируйте полученные результаты. Все ли коэффициенты оказались значимы? Выпишите уравнение оцененной модели.
4. **(4 балла)** Сделайте выбор между моделями из пунктов (1)–(3). Обоснуйте свой выбор.
5. **(6 баллов)** Для выбранной модели из пункта (4) протестируйте есть ли различия в моделях заработных плат для мужчин и женщин? Протестируйте двумя способами (с помощью дамми переменных и теста Чоу).
6. **(6 баллов)** Для выбранной модели из пункта (4) протестируйте наличие выбросов в вашей модели. Используйте несколько критериев.
7. **(6 баллов)** Для выбранной модели из пункта (4) протестируйте наличие мультиколлинеарности несколькими способами. Сделайте вывод. Примите меры, если мультиколлинеарность обнаружена.
8. **(6 баллов)** Уместно ли применять в рассматриваемой модели метод главных компонент (МГК)? Для каких задач используется данный метод? Обсудите возможные проблемы, которые могут здесь возникнуть. Если это возможно, продемонстрируйте применение МГК для ваших данных. Проинтерпретируйте результаты.
9. **(4 балла)** Для выбранной модели из пункта (4) постройте график «остатки—прогнозы». Сделайте вывод.

10. **(6 баллов)** Для выбранной модели из пункта (4) протестируйте наличие гетероскедастичности разными способами. Сделайте вывод. Примите меры, если гетероскедастичность обнаружена.

11. **(4 балла)** Для выбранной модели из пункта (4) проведите тест Рамсея с одним вспомогательным регрессором (только с квадратом). Сделайте вывод.

12. **(5 баллов)** Оцените модель из пункта (4), оставив в ней только значимые коэффициенты. Выпишите уравнение оцененной модели. Сравните результаты с моделью из пункта (1). Какие критерии для сравнения моделей здесь стоит использовать?

13. **(7 баллов)** Предложите иные функциональные формы для уравнения заработной платы. Какие гипотезы вы проверяете таким образом? Проверьте их.

14. **(7 баллов)** Выпишите итоговую оценку модели на основании результатов тестирования в предыдущих пунктах. Обсудите, какие еще потенциальные эконометрические проблемы могут быть в этой модели. Как бы вы их стали решать?