**Y1.** Имеются данные о нескольких крупных изданиях и тарифы на размещение рекламы (цена модульного рекламного объявления).

df = pd.read\_csv(“C:/data/regressionlab\_1.csv”, index\_col=0)

а) постройте диаграмму рассеяния для зависимости тарифа на размещение рекламы от тиража газет

import numpy as np

import matplotlib.pyplot as plt

plt.scatter(df['issue'], df['adcost'])

plt.title('Scatterplot of Issue volume on Ad placement cost')

plt.xlabel(‘issue volume’)

plt.ylabel(‘ad placement cost’)

plt.show()

б) дайте оценку корреляции тарифа на размещение рекламы и тиража. Можно ли объяснить эту корреляцию с экономической точки зрения?

import scipy.stats as scp

scp.pearsonr(df['issue'], df['adcost'])

в) найдите уравнение регрессии, позволяющее прогнозировать тариф на размещение рекламы на основании тиража.

import pandas as pd

import statsmodels.formula.api as smf

lm = smf.ols("adcost~issue", df).fit()

lm.summary()

г) проверьте, является ли статистически значимой связь между тарифом на размещение рекламы и тиражом;

д) найдите прогнозируемое значение и остаток для газеты Коммерсантъ. Ответьте на вопрос, является ли тариф на размещение одной строки рекламного объявления выше или ниже той величины, которую можно было бы ожидать от газеты с таким тиражом.

print(lm.get\_prediction(df, weights=1).summary\_frame())

**2.** По данным предыдущей задачи виден один возможный выброс (Эксперт). Необходимо выяснить, принадлежит ли это значение той же генеральной совокупности, что и остальные. С этой целью рассмотрим это значение как новое наблюдение.

а) Удалите Эксперт из рассматриваемой совокупности данных и найдите уравнение регрессии, позволяющее прогнозировать тариф на размещение рекламы на основании тиража для остальных газет;

df2 = df.drop(labels='Ekspert', axis=0)

lm2 = smf.ols("adcost~issue", df2).fit()

б) Вычислите двусторонний 95% доверительный интервал для нового наблюдения, когда X0 является тарифом Эксперт;

print(lm2.get\_prediction(df[1:2], weights=1).summary\_frame())

в) Проверьте, является ли Эксперт выбросом, выяснив, попадает ли тариф на размещение рекламы в этой газете в построенный доверительный интервал.

|  |  |  |
| --- | --- | --- |
|  | Тираж (тыс. экз.) | Стоимость разовой модульной рекламы «клэссифайд», руб. |
| Коммерсантъ | 135 | 22000 |
| Эксперт | 120 | 38230 |
| Секрет фирмы | 110 | 18500 |
| Дело | 95 | 15650 |
| Профиль | 90 | 25320 |
| Налоговый вестник | 40 | 14000 |
| Деньги | 80 | 18000 |
| Вопросы экономики | 40 | 14500 |
| Российское предпринимательство | 15 | 9100 |
| Финанс | 90 | 19500 |
| Экономист | 70 | 14200 |
| THESIS | 40 | 9400 |
| Мировая экономика | 60 | 10600 |
| ЭКО | 30 | 5000 |
| Маркетолог | 80 | 18350 |
| Business Online | 69 | 17000 |

В таблице представлены результаты исследования рынка жилья города. Необходимо построить аналитическую модель оценки стоимости жилья в зависимости от предложенных факторов.

Оцените параметры линейного уравнения множественной регрессии с фиктивными переменными:

Y – цена, тыс.руб.;

X1 – площадь, м2;

X2 – этаж;

X3 - «балкон», 0 – при отсутствии балкона, 1- при наличии балкона;

X4 - «ремонт», 0 – при отсутствии ремонта, 1 – при наличии ремонта.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | Y | X1 | X2 | X3 | X4 |
| 1 | 3150 | 35 | 4 | 0 | 0 |
| 2 | 3150 | 31 | 1 | 0 | 0 |
| 3 | 6150 | 42 | 2 | 1 | 1 |
| 4 | 3750 | 31 | 2 | 1 | 1 |
| 5 | 2700 | 36 | 4 | 0 | 0 |
| 6 | 4050 | 36 | 5 | 1 | 1 |
| 7 | 3450 | 31 | 5 | 0 | 1 |
| 8 | 3450 | 32 | 5 | 1 | 0 |
| 9 | 4050 | 38 | 9 | 1 | 1 |
| 10 | 3450 | 31 | 5 | 0 | 1 |
| 11 | 2250 | 23 | 1 | 0 | 0 |
| 12 | 2850 | 33 | 5 | 0 | 0 |
| 13 | 2400 | 29 | 1 | 0 | 0 |
| 14 | 4950 | 50 | 8 | 0 | 1 |
| 15 | 4200 | 32 | 3 | 1 | 1 |
| 16 | 3750 | 31,5 | 7 | 1 | 1 |
| 17 | 2550 | 30 | 1 | 0 | 1 |
| 18 | 3450 | 31,5 | 3 | 1 | 0 |
| 19 | 2100 | 29 | 2 | 0 | 0 |
| 20 | 3450 | 35 | 3 | 1 | 0 |
| 21 | 3600 | 34 | 2 | 1 | 0 |
| 22 | 3000 | 31 | 2 | 0 | 1 |
| 23 | 3600 | 34 | 1 | 1 | 0 |
| 24 | 2850 | 33 | 5 | 0 | 0 |
| 25 | 3000 | 32 | 3 | 1 | 0 |
| 26 | 3600 | 31 | 2 | 1 | 1 |
| 27 | 3390 | 33 | 2 | 1 | 0 |
| 28 | 3300 | 30 | 5 | 1 | 0 |
| 29 | 5190 | 38 | 7 | 1 | 1 |
| 30 | 4050 | 32 | 1 | 1 | 1 |
| 31 | 3750 | 36 | 5 | 0 | 1 |
| 32 | 3600 | 36 | 3 | 0 | 1 |
| 33 | 3750 | 33 | 2 | 1 | 1 |
| 34 | 4050 | 33,7 | 5 | 1 | 1 |
| 35 | 3720 | 31 | 5 | 0 | 1 |

Подготовка:

**в системной консоли (cmd)** установить *seaborn*

pip install seaborn

**В Python:**

Ввод данных

df = pd.read\_csv("C:/data/regressionlab\_2.csv", decimal=",")

Построение попарных диаграммы рассеяния:

import seaborn as sns

sns.pairplot(df)

plt.show()

Построение модели:

import pandas as pd

import statsmodels.formula.api as smf

model = smf.ols("Y~X1+X2+X3+X4", df).fit()

model.summary()

Проверка распределения регрессионных остатков

import matplotlib.pyplot as plt

scp.probplot(model.resid, dist="norm", plot=plt)

plt.show()

**Данные**

|  |  |  |
| --- | --- | --- |
|  | **issue** | **adcost** |
| **Kommersant** | 135 | 22000 |
| **Ekspert** | 120 | 38230 |
| **Sekretfirmy** | 110 | 18500 |
| **Delo** | 95 | 15650 |
| **Profil** | 90 | 25320 |
| **Nalogovyyvestnik** | 40 | 14000 |
| **Dengi** | 80 | 18000 |
| **Voprosyekonomiki** | 40 | 14500 |
| **Rossiyskoyepredprinimatelstvo** | 15 | 9100 |
| **Finans** | 90 | 19500 |
| **Ekonomist** | 70 | 14200 |
| **THESIS** | 40 | 9400 |
| **Mirovayaekonomika** | 60 | 10600 |
| **EKO** | 30 | 5000 |
| **Marketolog** | 80 | 18350 |
| **BusinessOnline** | 69 | 17000 |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Y** | **X1** | **X2** | **X3** | **X4** |
| **3150** | 35 | 4 | 0 | 0 |
| **3150** | 31 | 1 | 0 | 0 |
| **6150** | 42 | 2 | 1 | 1 |
| **3750** | 31 | 2 | 1 | 1 |
| **2700** | 36 | 4 | 0 | 0 |
| **4050** | 36 | 5 | 1 | 1 |
| **3450** | 31 | 5 | 0 | 1 |
| **3450** | 32 | 5 | 1 | 0 |
| **4050** | 38 | 9 | 1 | 1 |
| **3450** | 31 | 5 | 0 | 1 |
| **2250** | 23 | 1 | 0 | 0 |
| **2850** | 33 | 5 | 0 | 0 |
| **2400** | 29 | 1 | 0 | 0 |
| **4950** | 50 | 8 | 0 | 1 |
| **4200** | 32 | 3 | 1 | 1 |
| **3750** | 31,5 | 7 | 1 | 1 |
| **2550** | 30 | 1 | 0 | 1 |
| **3450** | 31,5 | 3 | 1 | 0 |
| **2100** | 29 | 2 | 0 | 0 |
| **3450** | 35 | 3 | 1 | 0 |
| **3600** | 34 | 2 | 1 | 0 |
| **3000** | 31 | 2 | 0 | 1 |
| **3600** | 34 | 1 | 1 | 0 |
| **2850** | 33 | 5 | 0 | 0 |
| **3000** | 32 | 3 | 1 | 0 |
| **3600** | 31 | 2 | 1 | 1 |
| **3390** | 33 | 2 | 1 | 0 |
| **3300** | 30 | 5 | 1 | 0 |
| **5190** | 38 | 7 | 1 | 1 |
| **4050** | 32 | 1 | 1 | 1 |
| **3750** | 36 | 5 | 0 | 1 |
| **3600** | 36 | 3 | 0 | 1 |
| **3750** | 33 | 2 | 1 | 1 |
| **4050** | 33,7 | 5 | 1 | 1 |
| **3720** | 31 | 5 | 0 | 1 |