Практика построения регрессии

10 вопросов

1 point

1.
Давайте проанализируем данные опроса 4361 женщин из Ботсваны:

botswana.tsv

О каждой из них мы знаем:

- сколько детей она родила (признак ceb)
- возраст (age)
- длительность получения образования (educ)
- религиозная принадлежность (religion)
- идеальное, по её мнению, количество детей в семье (idlnchld)
- была ли она когда-нибудь замужем (evermarr)
- возраст первого замужества (agefm)
- длительность получения образования мужем (heduc)
- знает ли она о методах контрацепции (knowmeth)
- использует ли она методы контрацепции (usemeth)
- живёт ли она в городе (urban)
- есть ли у неё электричество, радио, телевизор и велосипед (electric, radio, tv, bicycle)

Давайте научимся оценивать количество детей ceb по остальным признакам.

Загрузите данные и внимательно изучите их. Сколько разных значений принимает признак religion?

Введите ответ здесь

1 point

2.
Во многих признаках есть пропущенные значения. Сколько объектов из 4361 останется, если выбросить все, содержащие пропуски?

Введите ответ здесь

1 point

3.
В разных признаках пропуски возникают по разным причинам и должны обрабатываться по-разному.

Например, в признаке agefm пропуски стоят только там, где evermarr=0, то есть, они соответствуют женщинам, никогда не выходившим замуж. Таким образом, для этого признака NaN соответствует значению "не применимо".

В подобных случаях, когда признак x_1 на части объектов в принципе не может принимать никакие значения, рекомендуется поступать так:

- создать новый бинарный признак

$$x_2 = \begin{cases} 1, & x_1 = \text{'не применимо'}, \\ 0, & \text{иначе}; \end{cases}$$

- заменить "не применимо" в x_1 на произвольную константу c, которая среди других значений x_1 не встречается.

Теперь, когда мы построим регрессию на оба признака и получим модель вида

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2,$$

на тех объектах, где x_1 было измерено, регрессионное уравнение примет вид

$$y = \beta_0 + \beta_1 x,$$

а там, где x_1 было "не применимо", получится

$$y = \beta_0 + \beta_1 c + \beta_2.$$

Выбор c влияет только на значение и интерпретацию β_2, но не β_1.

Давайте используем этот метод для обработки пропусков в agefm и heduc.

1. Создайте признак nevermarr, равный единице там, где в agefm пропуски.
2. Удалите признак evermarr — в сумме с nevermarr он даёт константу, значит, в нашей матрице X будет мультиколлинеарность.
3. Замените NaN в признаке agefm на $c_{agefm} = 0$.
4. У объектов, где nevermarr = 1, замените NaN в признаке heduc на $c_{heduc_1} = -1$ (ноль использовать нельзя, так как он уже встречается у некоторых объектов выборки).

Сколько осталось пропущенных значений в признаке heduc?

Введите ответ здесь

1 point

4.

Избавимся от оставшихся пропусков.

Для признаков idlnchld, heduc и usemeth проведите операцию, аналогичную предыдущей: создайте индикаторы пропусков по этим признакам (idlnchld_noans, heduc_noans, usemeth_noans), замените пропуски на нехарактерные значения ($c_{idlnchld} = -1$, $c_{heduc_2} = -2$ (значение -1 мы уже использовали), $c_{usemeth} = -1$).

Остались только пропуски в признаках knowmeth, electric, radio, tv и bicycle. Их очень мало, так что удалите объекты, на которых их значения пропущены.

Какого размера теперь наша матрица данных? Умножьте количество строк на количество всех столбцов (включая отклик ceb).

Введите ответ здесь

1 point

5.

Постройте регрессию количества детей ceb на все имеющиеся признаки методом smf.ols, как в разобранном до этого примере. Какой получился коэффициент детерминации R^2? Округлите до трёх знаков после десятичной точки.

Если код из примера у вас не воспроизводится:

- убедитесь, что вы сделали так:

```
import statsmodels.formula.api as smf
```

- возможно, вам нужно обновить библиотеку patsy; выполните в командной строке

```
pip install -U patsy
```

Введите ответ здесь

1 point

6.

Обратите внимание, что для признака religion в модели автоматически создалось несколько бинарных фиктивных переменных. Сколько их?

Введите ответ здесь

1 point

7.

Проверьте критерием Бройша-Пагана гомоскедастичность ошибки в построенной модели. Выполняется ли она?

Если ошибка гетероскедастична, перенастройте модель, сделав поправку Уайта типа HC1.

- Ошибка гомоскедастична, $p > 0.05$
- Ошибка гетероскедастична, $p \leq 0.05$, нужно делать поправку Уайта

---

1 point

8.

Удалите из модели незначимые признаки religion, radio и tv. Проверьте гомоскедастичность ошибки, при необходимости сделайте поправку Уайта.

Не произошло ли значимого ухудшения модели после удаления этой группы признаков? Проверьте с помощью критерия Фишера. Чему равен его достигаемый уровень значимости? Округлите до четырёх цифр после десятичной точки.

Если достигаемый уровень значимости получился маленький, верните все удалённые признаки; если он достаточно велик, оставьте модель без религии, тв и радио.

Введите ответ здесь

---

1 point

9.

Признак usemeth_noans значим по критерию Стьюдента, то есть, при его удалении модель значимо ухудшится. Но вообще-то отдельно его удалять нельзя: из-за того, что мы перекодировали пропуски в usemeth произвольно выбранным значением $c_{usemeth} = -1$, удалять usemeth_noans и usemeth можно только вместе.

Удалите из текущей модели usemeth_noans и usemeth. Проверьте критерием Фишера гипотезу о том, что качество модели не ухудшилось. Введите номер первой значащей цифры в достигаемом уровне значимости (например, если вы получили 5.5×10^{-8}, нужно ввести 8).

Если достигаемый уровень значимости получился маленький, верните удалённые признаки; если он достаточно велик, оставьте модель без usemeth и usemeth_noans.

Введите ответ здесь

---

1 point

10.

Посмотрите на доверительные интервалы для коэффициентов итоговой модели (не забудьте использовать поправку Уайта, если есть гетероскедастичность ошибки) и выберите правильные выводы.

- [ ] У женщин, знакомых с методами контрацепции, при прочих равных в среднем на 0.6 ребёнка меньше (p=0.001, 95% доверительный интервал для разницы между средними — [-0.9, -0.2])

- [ ] Итоговая модель объясняет 63% вариации отклика

- [ ] С увеличением возраста женщины на 1 год среднее количество детей возрастает на 0.17 (p<0.001, 95% доверительный интервал — [0.16, 0.18])

- [ ] У женщин, никогда не выходивших замуж, при прочих равных в среднем на 2.3 ребёнка меньше (p<0.001, 95% доверительный интервал для разницы между средними — [-2.6, -1.9])

- [ ] У женщин, не знающих, какое количество детей идеально, в среднем на 0.66 ребёнка больше (p=0.002, 95% доверительный интервал — [0.2, 1.1])

- [ ] У женщин, не знающих, какое количество детей идеально, в среднем на

$$\beta_{idlnchld\_noans} + c_{idlnchld}\beta_{idlnchld} \approx 0.58$$

детей больше

I, **Владислав Лисеев**, understand that submitting work that isn't my own may result in permanent failure of this course or deactivation of my Coursera account.
Узнайте больше о Кодексе чести Coursera

9 вопросов без ответа

Сдать тест