Задача 2

*Задача №2.1*

Необходимо загрузить данные из указанного набора и произвести следующие действия.

Набор данных: Swiss.

Объясняемая переменная: Infant.Mortality.

Регрессоры: Agriculture, Fertility, Education.

1. Проверьте, что в наборе данных нет линейной зависимости (построить зависимости между переменными, указанными в варианте, и проверить, что в каждой из них невысокий). В случае, если большой, один из таких столбцов можно исключить из рассмотрения.

Проверим линейную регрессию Catholic ~ Agriculture.

Таблица 2.1.1. Характеристики полученной модели Catholic ~ Agriculture.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Переменная | Estimate | Std. Error | t value | Pr(>|t|) | Уровень значимости |
| (Intercept) | 3.8313 | 13.8966 | 0.276 | 0.7840 |  |
| Agriculture | 0.7365 | 0.2508 | 22.71 | 0.0052 | \*\* |

= 16%. Значение p-статистики слабое - 2 звезды (см. Таблица 2.1.1). Принимая эти параметры во внимание, можно сказать, что линейная зависимость практически отсутствует. Можно использовать вместе в одной модели

Теперь проверим линейную регрессию Catholic ~ Education.

Таблица 2.1.2. Характеристики полученной модели Catholic ~ Education.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Переменная | Estimate | Std. Error | t value | Pr(>|t|) | Уровень значимости |
| (Intercept) | 48.4703 | 9.2805 | 5.223 | 4.37e-06 | \*\*\* |
| Education | -0.6673 | 0.6389 | -1.045 | 0.302 |  |

= 2%. Хоть уровень значимости у вспомогательной переменной и сильная (см. Таблица 2.1.2), но сама Education объясняет достаточно слабо. Зависимости совершенно не наблюдается, можно использовать в модели.

Наконец, проверим и линейную регрессию Agriculture ~ Education.

Таблица 2.1.3. Характеристики полученной модели Agriculture ~ Education.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Переменная | Estimate | Std. Error | t value | Pr(>|t|) | Уровень значимости |
| (Intercept) | 48.4703 | 9.2805 | 5.223 | 4.37e-06 | \*\*\* |
| Agriculture | -0.6673 | 0.6389 | -1.045 | 0.302 |  |

= 2%. Для данной зависимости наблюдаются те же значения (см. Таблица 2.1.3), что и у Catholic ~ Education. Зависимости совершенно не наблюдается, можно использовать в модели.

1. Построить линейную модель зависимой переменной (Infant.Mortality) от регрессоров (Agriculture, Fertility, Education) по методу наименьших квадратов. Оценить, насколько хороша модель, согласно: 1) , 2) p-значениям каждого коэффициента

Таблица 2.1.4. Характеристики полученной модели Infant.Mortality ~ Catholic+Agriculture+Education.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Параметр \ Характеристики | Значение | Std. Error | t value | Pr(>|t|) | Уровень значимости |
| (Intercept) | 22.36903 | 1.75389 | 12.754 | 3.29e-16 | \*\*\* |
| Catholic | 0.01904 | 0.01117 | 1.705 | 0.0954 | . |
| Agriculture | -0.04490 | 0.02636 | - 1.703 | 0.0957 | . |
| Education | -0.08520 | 0.05771 | -1.476 | 0.1472 |  |

10%. Значение низкое, а значит, нельзя сделать адекватные выводы по данной зависимости (см. Таблица 2.1.4). Также можно заметить, что p-значения у регрессоров очень высокие, т.е. Infant Mortality не зависит от них.

1. Ввести в модель логарифмы регрессоров. Сравнить модели и выбрать наилучшую.

Были построены и проверены следующие модели:

1. Infant Mortality ~ log(Catholic) + log(Agriculture) + log(Education)
2. log(drivers) ~ log(Catholic) + log(Agriculture) + log(Education) + Catholic + Agriculture + Education.

Значения для проверенных моделей:

1. 5% - p-значение у коэффициентов только увеличилось.
2. 14% - p-значение у коэффициентов только увеличилось.
3. Ввести в модель всевозможные произведения пар регрессоров, в том числе квадраты регрессоров. Найти одну или несколько наилучших моделей по доле объяснённого разброса в данных .

Для рассмотрения построим модель со всеми возможными парами регрессоров.

Таблица 2.1.5. Характеристики полученной модели Infant.Mortality ~ Catholic + Agriculture + Education + + + + Log(Catholic) + Log(Agriculture) + Log(Education) + (Catholic \* Agriculture) + (Catholic\*Education) + (Agriculture \* Education)

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Параметр \ Характеристики | Значение | Std. Error | t value | Pr(>|t|) | Уровень значимости |
| (Intercept) | 18.0117962 | 1.427825 | 12.615 | 1.06e-15 |  |
| Catholic | -0.354176 | 0.263906 | -1.342 | 0.1885 |  |
| Agriculture | -0.154620 | 0.388734 | -0.398 | 0.6933 |  |
| Education | -0.250492 | 0.664041 | -0.377 | 0.7084 |  |
|  | 0.0006190 | 0.000384 | 1.612 | 0.11474 |  |
|  | 0.001007 | 0.002786 | 0.362 | 0.7199 |  |
|  | -0.0021146 | 0.0010389 | -2.035 | 0.04831 |  |
| Log(Catholic) | 1.2905177 | 0.7173768 | 1.799 | 0.07939 | . |
| Log(Agriculture) | 4.713491 | 5.382163 | 0.876 | 0.3873 |  |
| Log(Education) | 1.149989 | 2.646806 | 0.434 | 0.6667 |  |
| Catholic \* Agriculture | -0.0013712 | 0.0004525 | -3.031 | 0.00422 | . |
| Catholic \* Education | 0.002650 | 0.003251 | 0.815 | 0.4206 |  |
| Agriculture \* Education | -0.0003028 | 0.0015981 | -0.189 | 0.85066 |  |

полученной модели (см. Таблица 2.1.5) равен 33%, но уровень значимости всех параметров очень низкий. Проверим VIF данной модели, чтобы избавиться от наиболее незначимых параметров.

Таблица 2.1.6. VIF полученной модели со всеми возможными парами регрессоров.

|  |  |
| --- | --- |
| Параметр \ Характеристики | VIF |
| Catholic | 722.52177 |
| Agriculture | 464.90517 |
| Education | 243.16702 |
|  | 460.25190 |
|  | 223.76821 |
|  | 180.14396 |
| Log(Catholic) | 60.31557 |
| Log(Agriculture) | 101.51128 |
| Log(Education) | 26.30228 |
| Catholic \* Agriculture | 61.40025 |
| Catholic \* Education | 15.57974 |
| Agriculture \* Education | 19.76244 |

Наибольший VIF имеется у Catholic (см. Таблица 2.1.6), избавимся от него. После удаления модель получила =30%, а значит, удаление не сильно привело к потери качества модели. Продолжим избавляться от регрессоров с наибольшим VIF, пока не будет меняться слишком сильно.

Таблица 2.1.7. VIF полученной модели со всеми возможными парами регрессоров после удаления всех параметров, у которых был наибольший VIF.

|  |  |
| --- | --- |
| Параметр \ Характеристики | VIF |
|  | 19.984286 |
|  | 1.456378 |
| Log(Catholic) | 9.395438 |
| Catholic \* Agriculture | 29.270946 |
| Agriculture \* Education | 1.057898 |

При дальнейшем избавлении от регрессоров с наибольшим VIF (таблица 2.1.7) модели падал более чем на 5%. В итоге полученная модель имеет следующий вид.

Таблица 2.1.8. Характеристики полученной модели Infant.Mortality ~ + + Log(Catholic)+(Catholic \* Agriculture)+(Agriculture \* Education)

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Параметр \ Характеристики | Значение | Std. Error | t value | Pr(>|t|) | Уровень значимости |
| (Intercept) | 18.0117962 | 1.427825 | 12.615 | 1.06e-15 | \*\*\* |
|  | 0.0006190 | 0.0003841 | 1.612 | 0.11474 |  |
|  | -0.0021146 | 0.0010389 | -2.035 | 0.04831 | \* |
| Log(Catholic) | 1.2905177 | 0.7173768 | 1.799 | 0.07939 | . |
| Catholic \* Agriculture | -0.0013712 | 0.0004525 | -3.031 | 0.00422 | \*\* |
| Agriculture \* Education | -0.0003028 | 0.0015981 | -0.189 | 0.85066 |  |

данной модели (см. Таблица 2.1.8) равен 26%.

Как итог, можно выделить две наилучшие модели:

А) Модель со всеми возможными парами регрессоров (см. Таблица 2.1.5) – Infant.Mortality ~ Catholic + Agriculture + Education + + + + Log(Catholic) + Log(Agriculture) + Log(Education) + (Catholic \* Agriculture) + (Catholic\*Education) + (Agriculture \* Education). Имеет неплохой, по сравнению с остальными моделями, , но все параметры совершенно не объясняют зависимость между ними и объясняемой переменной. Нельзя сделать какие-либо адекватные выводы.

Б) Infant Mortality ~ (см. Таблица 2.1.8). Хоть модели и ниже, чем у модели А, но появились регрессоры, по которым можно объяснить зависимость. Так, например, Catholic\*Agriculture действительно влияет на детскую смертность. Это можно объяснить тем, что более бедные регионы были сильнее подвержены католицизму, но если регион активно занимался фермерством, то у крестьянина было больше шансов заработать деньги и обратиться ко врачу в случае заболевания ребенка. Остальные регрессоры имеют незначительный вес, но в случае их удаления падал еще сильнее, поэтому их можно оставить.

Итог: После того, как мы убедились в том, что в наборе данных между объясняемой переменной и регрессорами нет линейной зависимости, были найдены две наиболее подходящие для работы модели. Первая такая модель является имеет все возможные пары регрессоров (см. Таблица 5) – у нее достаточно большой (по сравнению другими моделями) %, но ее регрессоры имеют очень малый уровень значимости. Тогда с помощью избавления от регрессоров с наибольшим VIF была получена вторая модель (см. Таблица 2.1.8), которая хоть и имеет меньший, чем у первой модели, но имеет более значимые переменные, которые уже возможно адекватно описать. В данной модели есть действительно значимая переменная Catholic\*Agriculture, и взаимосвязь данной переменной с Infant.Mortality можно описать следующим образом - более бедные регионы были сильнее подвержены католицизму, но если регион активно занимался фермерством, то у крестьянина было больше шансов заработать деньги и обратиться ко врачу в случае заболевания ребенка, отсюда и малая детская смертность в таком регионе.

*Задача №2.2*

1. Оцените доверительные интервалы для всех коэффициентов в модели, p=95%.

Всего имеется 40 степеней свободы в обучающей выборке. Было рассчитано 6 коэффициентов, тогда число степеней свободы в модели: 40–6=34. Отсюда можно подсчитать критерий Стьюдента, который будет равен .

Тогда доверительный интервал будет иметь следующий вид: , где:

* - значение коэффициента;
* t - значение t-критерия Стьюдента;
* σ- стандартная ошибка коэффициента в модели.

Подсчитаем доверительный интервал для каждого коэффициента и сделаем вывод о отвержении или невозможности отвергнуть статистическую гипотезу о том, что коэффициент равен 0.

Доверительный интервал для свободного коэффициента.

Середина интервала . СКО .

= =

0 не попадает в доверительный интервал. Коэффициент не может быть равен 0 на уровне значимости 5%.

Доверительный интервал для .

Доверительный интервал для свободного коэффициента.

Середина интервала . СКО .

= =

0 не попадает в доверительный интервал. Коэффициент не может быть равен 0 на уровне значимости 5%.

Доверительный интервал для .

Середина интервала . СКО .

= =

0 не попадает в доверительный интервал. Коэффициент не может быть равен 0 на уровне значимости 5%.

Доверительный интервал для .

Середина интервала . СКО .

= =

0 не попадает в доверительный интервал. Коэффициент не может быть равен 0 на уровне значимости 5%.

Доверительный интервал для .

Середина интервала . СКО .

= =

0 не попадает в доверительный интервал. Коэффициент не может быть равен 0 на уровне значимости 5%.

Доверительный интервал для .

Середина интервала . СКО .

= =

0 может попасть в доверительный интервал.

Таблица 2.2.1. Доверительные интервалы для всех регрессоров

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Регрессор | Оценка коэффициента перед регрессором | СКО | Доверительный интервал  (p = 95%, df=33, t=1.7) | Может ли = 0 |
| Свободный коэффициент | 16.6 | 2.1 | [13.03,20.17] | Нет |
|  | 0.0007 | 0.0004 | [0.00002,0.0014] | Нет |
|  | 0.0020607 | 0.0010427 | [−0.004,−0.0003] | Нет |
|  | 1.7047206 | 0.8494241 | [0.26,3.14] | Нет |
| Catholic×Argiculture | −0.0018518 | 0.0006937 | [−0.003,−0.0007] | Нет |
| Argiculture×Education | −0.0004987 | 0.0016156 | [−0.003,0.002] | Да |

Только в одном регрессоре может быть равен 0 – в Argiculture×Education (см. Таблица 2.2.1).

1. Оцените доверительный интервал для одного прогноза.

Доверительный интервал для прогноза . Построим датасет, взяв следующие данные из датасета Swiss:

1. Catholic – 10 объектов.
2. Agriculture – 20 объектов.
3. Education – 8 объектов.

Тогда прогноз модели ≈ 20.16, нижняя граница доверительного интервала ≈ 18.34, верхняя граница доверительного интервала ≈ 22.

Итог: помощью степеней свободы выборки, критерия Стьюдента и СКО для каждого регрессора нашли их доверительные интервалы (см. Таблица 1), причем у регрессора Argiculture×Education есть шанс, что 0 может попасть в доверительный интервал.

Вывод по задаче №2

*Задача №2.1*. Рассматриваемая модель была проверена на наличие линейной зависимости между регрессорами. Зависимостей обнаружено не было, и все переменные были использованы в модели. Однако при исследовании самой модели было выявлено, что объясняемая переменная (Infant.Mortality) практически не зависит от всех регрессоров.

В пункте №3 была попытка улучшить рассматриваемую модель, с помощью введения логарифмов регрессоров. Однако, не смотря на некоторый рост , точность все-еще является низкой.

В пункте №4 в модель были введены всевозможные произведения пар регрессоров, и была выявлена наилучшая модель по доле объяснённого разброса в данных . Найдена действительно значимая взаимосвязь между Catholic\*Agriculture и Infant.Mortality, которую можно описать следующим образом - более бедные регионы были сильнее подвержены католицизму, но если регион активно занимался фермерством, то у крестьянина было больше шансов заработать деньги и обратиться ко врачу в случае заболевания ребенка, отсюда и малая детская смертность в таком регионе.

*Задача №2.2*. Были найдены доверительные интервалы для всех коэффициентов в

рассматриваемой модели при p=95% и было выявлено, что из-за того, что значение коэффициента перед регрессором Argiculture×Education может быть равно 0, то данная объясняющая переменная практически не связана с объясняемой переменной Infant.Mortality.

В пункте №3 для оценивания доверительного интервала для одного прогноза, были

выбраны следующие значения:

1. Catholic – 10 объектов.
2. Agriculture – 20 объектов.
3. Education – 8 объектов.

Затем, с помощью функции predict() был вычислен прогноз ≈ 20.16 и доверительный интервал для рассматриваемой модели - нижняя граница доверительного интервала ≈ 18.34, верхняя граница доверительного интервала ≈ 22.

Приложение 1

ca\_ar = lm(Catholic~Agriculture, swiss)

summary(ca\_ar)

ca\_ed = lm(Catholic~Education, swiss)

summary(ca\_ed)

ag\_ed = lm(Agriculture~Education, swiss)

summary(ca\_ed)

#2

model = lm(Infant.Mortality ~ Catholic + Agriculture + Education , swiss)

summary(model)

# 3

model\_log = lm(Infant.Mortality ~ I(log(Catholic)) + I(log(Agriculture)) + I(log(Education)) , swiss)

summary(model\_log)

model\_log = lm(Infant.Mortality ~ Catholic + Agriculture + Education + I(log(Catholic)) + I(log(Agriculture)) + I(log(Education)) , swiss)

summary(model\_log)

# 4

model\_prod1 = lm(Infant.Mortality ~ Catholic + Agriculture + Education + I(Catholic^2) + I(Agriculture^2)

+ I(Education^2) + I(log(Catholic)) + I(log(Agriculture)) + I(log(Education)) + I(Catholic\*Agriculture)

+ I(Catholic\*Education) + I(Agriculture\*Education), swiss)

summary(model\_prod1)

vif(model\_prod1)

model\_prod1 = lm(Infant.Mortality ~  Agriculture + Education + I(Catholic^2) + I(Agriculture^2)

+ I(Education^2) + I(log(Catholic)) + I(log(Agriculture)) + I(log(Education)) + I(Catholic\*Agriculture)

+ I(Catholic\*Education) + I(Agriculture\*Education), swiss)

summary(model\_prod1)

vif(model\_prod1)

model\_prod1 = lm(Infant.Mortality ~ Education + I(Catholic^2) + I(Agriculture^2)

+ I(Education^2) + I(log(Catholic)) + I(log(Agriculture)) + I(log(Education)) + I(Catholic\*Agriculture)

+ I(Catholic\*Education) + I(Agriculture\*Education), swiss)

summary(model\_prod1)

vif(model\_prod1)

model\_prod1 = lm(Infant.Mortality ~ I(Catholic^2) + I(Agriculture^2)

+ I(Education^2) + I(log(Catholic)) + I(log(Agriculture)) + I(log(Education)) + I(Catholic\*Agriculture)

+ I(Catholic\*Education) + I(Agriculture\*Education), swiss)

summary(model\_prod1)

model\_prod1 = lm(Infant.Mortality ~ I(Catholic^2) + I(Agriculture^2)

+ I(Education^2) + I(log(Catholic)) + I(Catholic\*Agriculture)

 + I(Agriculture\*Education), swiss)

summary(model\_prod1)

vif(model\_prod1)

model\_prod1 = lm(Infant.Mortality ~ I(Catholic^2) + I(Agriculture^2)

+ I(Education^2) + I(log(Catholic)) + I(Catholic\*Agriculture)

 + I(Agriculture\*Education), swiss)

summary(model\_prod1)

Приложение 2

library("car")

require(stats)

require(graphics)

plot(swiss)

model = lm(Infant.Mortality ~ I(Catholic^2) + I(Agriculture^2) + I(Education^2) + I(log(Catholic)) + I(Catholic\*Agriculture) + I(Agriculture\*Education), swiss)

summary(model)

qt(0.95, df = 33)

data\_new = data.frame(Catholic = 10, Agriculture= 20, Education= 8)

predict(model, data\_new, interval = "confidence")