Балашов Д. С. КМБО-03-20(Вариант 1)

Задача №2

Необходимо загрузить данные из указанного набора и произвести следующие действия.

Набор данных: Swiss.

Объясняемая переменная: *Fertility.*

Регрессоры: *Agriculture*, *Examination, Infant.Mortality.*

*Задача №2.1*

***№1.*** Проверить, что в наборе данных нет линейной зависимости (построить зависимости

между переменными, и проверить, что R^2 в каждой из них не высокий). В случае если R^2 большой, один из таких столбцов можно исключить из рассмотрения.

Проверим линейную регрессию *Agriculture* ~ *Examination*, *Infant.Mortality*. R^2 = = 49.1%. VIF = 1/(1-R^2) = 1.96. Хоть VIF и меньше 5, но если учесть, что при регрессоре (*Examination*) хорошее значение p-статистики (3 звезды, *см. Таблица 1*), и что R^2 = 49.1% - это довольно неплохой показатель, то можно сказать, что существует небольшая зависимость между регрессором (*Examination*) и регрессором(Agriculture).

Зависимость *Examination ~ Agriculture, Infant.Mortality*. R^2 = 49.6%. VIF = 1/(1- - R^2) = 1.98. Хоть VIF и меньше 5, но если учесть, что при регрессоре (Agriculture) хорошее значение p-статистики (3 звезды, *см. Таблица 2*), и что R^2 = 49.1% - это довольно неплохой показатель, то можно сказать, что существует небольшая зависимость между регрессором (*Agriculture*) и регрессором(*Examination*).

В регрессии *Infant.Mortality ~ Agriculture, Examination* R^2 = 4.9%. VIF = 1/(1-R^2) = = 1.05. Т.к. VIF значительно меньше 5 , и значение p-статистики плохое (0 звезд / 0 звезд, *см. Таблица 3*), то можно с уверенностью сказать, что регрессор(*Infant.Mortality)* не зависит от регрессоров(Agriculture, *Examination).*

Хоть между регрессором(*Agriculture*) и регрессором(*Examination*) и существует небольшая зависимость, но она слишком незначительна, чтобы их исключить. Таким образом, заключаем, что для построения моделей, можно использовать все регрессоры из условия.

***№2.*** Построить линейную модель зависимой переменной (*Fertility)* от регрессоров (*Agriculture*/*Examination*/*Infant.Mortality)* по методу наименьших квадратов. Оценить, насколько хороша модель, согласно: 1) R2, 2) p-значениям каждого коэффициента.

***2.1)*** В регрессии *Fertility* ~ *Agriculture*, *Examination*, *Infant.Mortality* R^2 = 53.98%. Для трех регрессоров значение R^2 довольно низкое, следовательно, сделать какие-либо выводы по данной зависимости нельзя.

***2.2)*** Рассмотрим p-статистику у регрессоров (*Agriculture*, *Examination*, *Infant.Mortality*).

* Значение p-статистики при регрессоре (*Examination)* низкое ( \*\*\*, *см. Таблица 4*);
* Значение p-статистики при регрессоре (*Infant.Mortality)* довольно низкое ( \*\*,

*см. Таблица 4*);

* Значение p-статистики при регрессоре (*Agriculture)* высокое (0 звезд,

*см. Таблица 4*);

Обратим внимание, что регрессор (*Agriculture)* не значим. P-статистика достаточно велика (*см. Таблица 4*), так что можно провести эксперимент по его исключению:

* R^2 =53.64%, следовательно R^2 изменился всего на 0.35%, что << 5%. Таким образом, мы можем исключить из рассмотрения регрессор (Agriculture).

***№3.*** Ввести в модель логарифмы регрессоров. Сравнить модели и выбрать наилучшую.

*При решении этой задачи были проверены модели:*

1. *Fertility ~ log(Examination), log(Infant.Mortality) – Таблица 5*
2. *Log(Fertility) ~ log(Examination), log(Infant.Mortality) – Таблица 6*
3. *Fertility ~ log(Examination), Infant.Mortality – Таблица 7*
4. *Fertility ~ Examination, log(Infant.Mortality) – Таблица 8*

*Значения R^2 для проверенных моделей:*

1. 52.16% - показатели ухудшились;

2. 48.58% - показатели ухудшились;

3. 54% - показатели немного улучшились;

4. 52% - показатели ухудшились.

Наилучшей оказалась модель: *Fertility ~ log(Examination), Infant.Mortality.*

*Таблица 7*. Характеристики модели зависимости параметра *Fertility* от параметров *log(Examination),* и *Infant.Mortality*  в наборе данных Swiss.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Параметр \ Характеристики | Значение | Std. Error | t value | Pr(>|t|) | Уровень  значимости |
| (Intercept) | 67.48 | 10.45 | 6.46 | 7.11e-08 | \*\*\* |
| *log(Examination)* | -12.89 | 2.18 | -5.92 | 4.44e-07 | \*\*\* |
| *Infant.Mortality* | 1.85 | 0.44 | 4.22 | 0.000119 | \*\*\* |

*# Fertility* *= -12.89\* log(Examination) + 1.85\* Infant.Mortality + 67.48*

*Вывод:* Хоть введение в модель логарифмы регрессоров и дало прирост, но слишком незначительный.

***№4.*** Ввести в модель всевозможные произведения пар регрессоров, в том числе квадраты регрессоров. Найти одну или несколько наилучших моделей по доле объяснённого разброса в данных R^2 .

***4.1)*** При решении этой задачи была проверена модель (*Fertility*~ *Examination*, *Infant.Mortality),* в которую были добавлены параметры: *I(Examination^2) / I(Infant.Mortality^2) / I(Examination \*Infant.Mortality). #R^2 = 54.69%*

*Таблица 9*. Проверка на линейную зависимость между *регрессоров (Examination, Infant.Mortality , I(Examination^2), I(Infant.Mortality^2), I(Examination\*Infant.Mortality))* с помощью команды VIF.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Параметр \ Характеристики | *Examination* | *Infant.Mortality* | *I(Examination^2)* | *I(Infant.Mortality^2)* | *I(Examination\**  *Infant.Mortality).* |
| VIF | 57.84 | 159.2 | 15.68 | 114.72 | 64.08 |

Поскольку у многих объясняющих переменных значения VIF довольно большое, то можно сделать вывод, что в модели присутствует линейная зависимость между регрессорами.

***4.2)*** Будем избавляться от регрессоров с максимальным VIF, пока все значения VIF не будут меньше 10.

1. *Fertility ~ Examination, I(Examination^2), I(Infant.Mortality^2), I(Examination\*Infant.Mortality). – Таблица 10;*
2. *Fertility ~ I(Examination^2), I(Infant.Mortality^2), I(Examination\*Infant.Mortality). -Таблица 11;*
3. *Fertility ~ I(Examination^2), I(Infant.Mortality^2.) - Таблица 12;*

*Таблица 12*. Проверка на линейную зависимость между регрессоров *(I(Examination^2), I(Infant.Mortality^2.))* с помощью команды VIF.

|  |  |  |
| --- | --- | --- |
| Параметр \ Характеристики | *I(Examination^2)* | *I(Infant.Mortality^2* |
| VIF | 1.027 | 1.027 |

*Таблица 13*. Характеристики модели зависимости параметра *Fertility* от параметров *I(Examination^2),* и *I(Infant.Mortality^2)* в наборе данных Swiss.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Параметр \ Характеристики | Значение | Std. Error | t value | Pr(>|t|) | Уровень  значимости |
| (Intercept) | 63.49 | 5.39 | 11.78 | 3.35e-15 | \*\*\* |
| *I(Examination^2)* | -0.02 | 0.004 | 5.38 | 2.76e-06 | \*\*\* |
| *I(Infant.Mortality^2)* | 0.04 | 0.01 | 3.054 | 0.004 | \*\* |

*# Fertility* *= -0.02\* I(Examination^2) + 0.04\* I(Infant.Mortality^2) + 63.49*

#Значение R^2 для данной модели = 50.43%

*Вывод:* Наилучшая модель - это *«Fertility ~ I(Examination^2), I(Infant.Mortality^2)»,* поскольку все значения VIF регрессоров меньше 5.

*Задача №2.2*

***№1.*** Оценить доверительные интервалы для всех коэффициентов в модели «*Fertility* ~ *Agriculture*, *Examination*, *Infant.Mortality*».

Всего проводилось 47 наблюдений, в данной модели оценивалось 4 коэффициента.

Следовательно, количество свободных коэффициентов = 47 – 4 = 43.

* *Agriculture*:

СКО(se) = 0.08 (*см. Таблица 14*)

Критерий Стьюдента(t) = 2.02

Доверительный интервал: [-0.21, 0.12]

* *Examination*:

СКО(se) = 0.23 (*см. Таблица 14*)

Критерий Стьюдента(t) = 2.02

Доверительный интервал: [-1.5, -0.58]

* *Infant.Mortality*:

СКО(se) = 0.46 (*см. Таблица 14*)

Критерий Стьюдента(t) = 2.02

Доверительный интервал: [0.52, 2.36]

* *Intercept*:

СКО(se) = 12.83 (*см. Таблица 14*)

Критерий Стьюдента(t) = 2.02

Доверительный интервал: [35, 86.74]

***№2.*** Построим таблицу с доверительными интервалами для всех коэффициентов в модели и сделаем вывод о том, может ли коэффициент быть равен 0.

*Таблица 14*. Характеристики модели зависимости параметра *Fertility* от параметров *Agriculture*, *Examination* и *Infant.Mortality* в наборе данных Swiss.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Параметр \ Характеристики | Значение | СКО | Доверительный интервал | “Может ли коэффициент быть равен 0?” |
| (Intercept) | 60.87 | 12.83 | [35, 86.74] | Нет |
| *Examination* | -1.04 | 0.23 | [-1.5, -0.58] | Нет |
| *Infant.Mortality* | 1.44 | 0.46 | [0.52, 2.36] | Нет |
| *Agriculture* | -0.05 | 0.08 | [-0.21, 0.12] | Да |

*Вывод:* Поскольку 0 попадает в доверительный интервал регрессора (*Agriculture),* то значение коэффициента перед этим регрессором может быть равно 0. Следовательно, объясняющая переменная (*Agriculture)* практически не связана с объясняемой переменной (*Fertility).*

***№3.*** Оценить доверительный интервал для одного прогноза для модели «*Fertility* ~ *Agriculture*, *Examination*, *Infant.Mortality*».

Зададим следующий набор значений для регрессоров: (*Agriculture = 10*, *Examination = 30* , *Infant.Mortality = 20).*

Применим функцию *predict()* для оцениваемой модели, что вычислить прогноз модели и доверительный интервал:

*Таблица 15*. Результат выполнения функции *predict()*

|  |  |  |
| --- | --- | --- |
| Прогноз модели | Нижняя граница интервала | Верхняя граница интервала |
| 58.01866 | 52.43242 | 63.60491 |

*Вывод:* Прогноз модели «*Fertility* ~ *Agriculture*, *Examination*, *Infant.Mortality*» оценивается как 58.02. Доверительный интервал для свободного коэффициента имеет вид: [52.43, 63.6].

*Приложение1*

library("lmtest")

library("GGally")

library("car")

data = Swiss

ggpairs(Swiss)

# Проверить отсутствие линейной зависимости между регрессорами:

model\_test\_1 = lm(Agriculture ~ Examination + Infant.Mortality, data)

model\_test\_2 = lm(Examination ~ Agriculture + Infant.Mortality, data)

model\_test\_3 = lm(Infant.Mortality ~ Agriculture + Examination, data)

summary(model\_test\_1)

summary(model\_test\_2)

summary(model\_test\_3)

# Модель, которую необходимо исследовать:

model = lm(Fertility ~ Agriculture + Examination + Infant.Mortality, data)

summary(model)

# Модели с введенными логарифмами регрессоров:

model\_1 = lm(Fertility ~ log(Examination) + log(Infant.Mortality), data)

model\_2 = lm(log(Fertility) ~ log(Examination) + log(Infant.Mortality), data)

model\_3 = lm(Fertility ~ log(Examination) + Infant.Mortality, data)

model\_4 = lm(Fertility ~ Examination + log(Infant.Mortality), data)

summary(model\_1)

summary(model\_2)

summary(model\_3)

summary(model\_4)

*Приложение2*

# Модель со всевозможными произведениями пар регрессоров, включая их квадраты:

model = lm(Fertility~ Examination + Infant.Mortality + I(Examination^2) + (Infant.Mortality^2) + I(Examination \*Infant.Mortality), data)

summary(model)

# Функция для вычисления Критерия Стьюдента(t):

t = qt(0.975, df = 43)

# Формула для вычисления доверительно интервала:

[model$coefficient [] – t\*se, model$coefficient [] + t\*se]

# Модель для оценивания доверительного интервала:

model = lm(Fertility ~ Agriculture + Examination + Infant.Mortality, data)

summary(model)

# Функция *predict()* для вычисления прогноза модели и доверительного интервала:

Predict (model, new. data, interval = “confidence”)

*Таблица 1*. Характеристики модели зависимости параметра *Agriculture* от параметров *Examination* и *Infant.Mortality*  в наборе данных Swiss.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Параметр \ Характеристики | Значение | Std. Error | t value | Pr(>|t|) | Уровень  значимости |
| (Intercept) | 105.56 | 18.29 | 5.77 | 7.32e-07 | \*\*\* |
| *Examination* | -2 | 0.3 | -6.49 | 6.42e-08 | \*\*\* |
| *Infant.Mortality* | -1.1 | 0.84 | -1.3 | 0.2 |  |

*Таблица 2*. Характеристики модели зависимости параметра *Examination* от параметров *Agriculture* и *Infant.Mortality*  в наборе данных Swiss.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Параметр \ Характеристики | Значение | Std. Error | t value | Pr(>|t|) | Уровень  значимости |
| (Intercept) | 37.42 | 6.33 | 5.91 | 4.53e-07 | \*\*\* |
| *Agriculture* | -0.24 | 0.04 | -6.49 | 6.42e-08 | \*\*\* |
| *Infant.Mortality* | -0.43 | 0.29 | -1.46 | 0.15 |  |

*Приложение3*

*Таблица 3*. Характеристики модели зависимости параметра *Infant.Mortality* от параметров *Agriculture* и *Examination* в наборе данных Swiss.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Параметр \ Характеристики | Значение | Std. Error | t value | Pr(>|t|) | Уровень  значимости |
| (Intercept) | 23.43 | 2.36 | 9.92 | 8.6e-13 | \*\*\* |
| *Agriculture* | -0.03 | 0.03 | -1.3 | 0.2 |  |
| *Examination* | -0.1 | 0.07 | -1.46 | 0.15 |  |

*Таблица 4*. Характеристики модели зависимости параметра *Fertility* от параметров *Agriculture*, *Examination* и *Infant.Mortality* в наборе данных Swiss.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Параметр \ Характеристики | Значение | Std. Error | t value | Pr(>|t|) | Уровень  значимости |
| (Intercept) | 60.87 | 12.83 | 4.75 | 2.32e-05 | \*\*\* |
| *Examination* | -1.04 | 0.23 | -4.56 | 4.22e-05 | \*\*\* |
| *Infant.Mortality* | 1.44 | 0.46 | 3.16 | 0.003 | \*\* |
| *Agriculture* | -0.05 | 0.08 | -0.57 | 0.57 |  |

*Таблица 5*. Характеристики модели зависимости параметра *Fertility* от параметров *log(Examination),* и log(*Infant.Mortality)*  в наборе данных Swiss.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Параметр \ Характеристики | Значение | Std. Error | t value | Pr(>|t|) | Уровень  значимости |
| (Intercept) | 6.35 | 25.49 | 0.25 | 0.8 |  |
| *log(Examination)* | -12.76 | 2.22 | -5.75 | 7.87e-07 | \*\*\* |
| *log(Infant.Mortality)* | 32.79 | 8.33 | 3.94 | 0.0003 | \*\*\* |

*Таблица 6*. Характеристики модели зависимости параметра log(*Fertility)* от параметров *log(Examination),* и log(*Infant.Mortality)* в наборе данных Swiss.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Параметр \ Характеристики | Значение | Std. Error | t value | Pr(>|t|) | Уровень  значимости |
| (Intercept) | 3.26 | 0.41 | 7.85 | 6.85e-10 | \*\*\* |
| *log(Examination)* | -0.19 | 0.04 | -5.34 | 3.13e-06 | \*\*\* |
| *log(Infant.Mortality)* | 0.5 | 0.14 | 3.68 | 0.0006 | \*\*\* |

*Приложение4*

*Таблица 8*. Характеристики модели зависимости параметра *Fertility* от параметров *Examination,* и log(*Infant.Mortality)*  в наборе данных Swiss.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Параметр \ Характеристики | Значение | Std. Error | t value | Pr(>|t|) | Уровень  значимости |
| (Intercept) | 8.53 | 25.61 | 0.33 | 0.74 |  |
| *Examination* | -0.95 | 0.16 | -5.73 | 8.33e-07 | \*\*\* |
| *log(Infant.Mortality)* | 2.89 | 8.41 | 3.08 | 0.004 | \*\* |

*Таблица 10*. Проверка на линейную зависимость между регрессоров (Examination, I(Examination^2), I(Infant.Mortality^2), I(Examination\*Infant.Mortality)) с помощью команды VIF.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Параметр \ Характеристики | Examination | I(Examination^2) | I(Infant.Mortality^2) | I(Examination\*  Infant.Mortality) |
| VIF | 35.28 | 14 | 3.77 | 27.15 |

*Таблица 11*. Проверка на линейную зависимость между регрессоров (I(Examination^2), I(Infant.Mortality^2), I(Examination\*Infant.Mortality)) с помощью команды VIF.

|  |  |  |  |
| --- | --- | --- | --- |
| Параметр \ Характеристики | I(Examination^2) | I(Infant.Mortality^2) | (Examination\*  Infant.Mortality) |
| VIF | 10.29 | 2.42 | 10.49 |

*Приложение5*

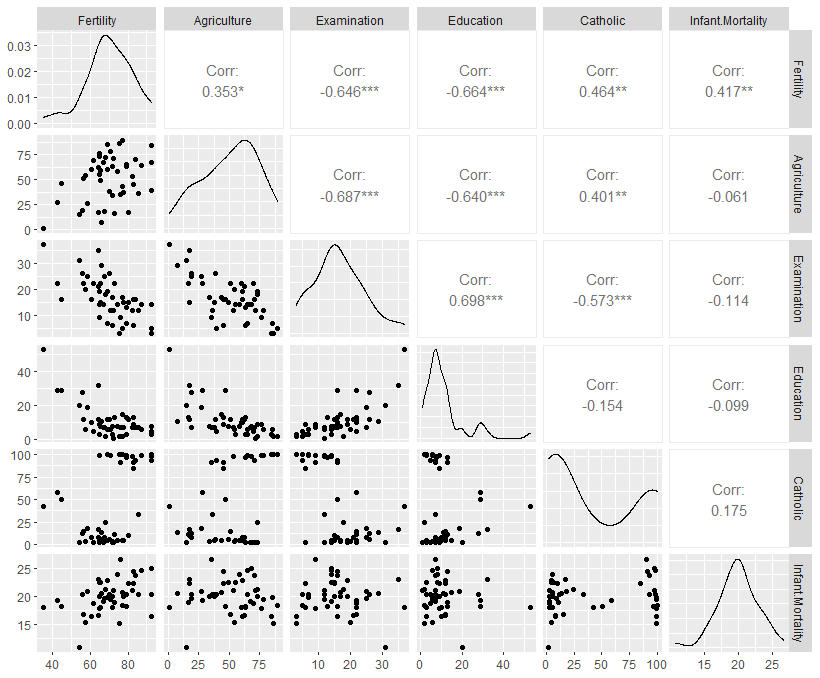


Рисунок 1. Результат работы команды ggpairs() – графики зависимостей между парами

переменных наборе данных Swiss.

Заключение

*Задача №2.1*. Рассматриваемая модель была проверена на наличие линейной зависимости между регрессорами. Хоть и была обнаружена небольшая зависимость между двумя объясняющими переменными, но в виду её незначительности, было принято решение не исключать ни одну из переменных в модели.

Однако при исследовании самой модели было выявлено, что объясняемая переменная(*Fertility*) почти не зависит от одного из регрессоров(*Agriculture),* поэтому было всё же решено исключить из рассмотрения одну из объясняющих переменных.

В пункте ***№3*** была попытка улучшить рассматриваемую модель, путём введения логарифмов регрессоров. Однако, это не дало видимых результатов.

В пункте ***№4*** в модель были введены всевозможные произведения пар регрессоров, и была выявлена одна наилучшая модель по доле объяснённого разброса в данных R^2.

*Задача №2.2*. Были найдены доверительные интервалы для всех коэффициентов в рассматриваемой модели (при p=95%), и было выявлено, что т.к. значение коэффициента перед регрессором(*Agriculture)* может быть равно 0, то объясняющая переменная (*Agriculture)* практически не связана с объясняемой переменной (*Fertility).*

В пункте ***№3*** для оценивания доверительного интервала для одного прогноза, были выбраны следующие значения: «*Agriculture = 10*, *Examination = 30* , *Infant.Mortality = 20*». Затем, с помощью функции *predict()* был вычислен прогноз и доверительный интервал для рассматриваемой модели.

*Приложение*

*Код решения всех задач:*

*Задача №2.1*.

library("lmtest")

library("GGally")

library("car")

data = swiss

help(swiss)

vif(model)

model\_test\_1 = lm(Agriculture ~ Examination + Infant.Mortality, data) model\_test\_1

summary(model\_test\_1

model\_test\_2 = lm(Examination ~ Agriculture + Infant.Mortality, data) model\_test\_2

summary(model\_test\_2)

model\_test\_3 = lm(Infant.Mortality ~ Agriculture + Examination , data)

model\_test\_3

summary(model\_test\_3)

model = lm(Fertility ~ Agriculture + Examination + Infant.Mortality, data)

model

summary(model)

model = lm(Fertility ~ Examination + Infant.Mortality, data)

model

summary(model)

model = lm(Fertility ~ log(Examination) + log(Infant.Mortality), data)

model

summary(model)

vif(model)

model = lm(log(Fertility) ~ log(Examination) + log(Infant.Mortality), data

model

summary(model)

vif(model)

model = lm(Fertility ~ log(Examination) + Infant.Mortality, data)

model

summary(model)

vif(model)

model = lm(Fertility ~ Examination + log(Infant.Mortality), data)

model

summary(model)

vif(model)

model\_1 = lm(Fertility ~ Examination + Infant.Mortality + I(Examination^2) + I(Infant.Mortality^2) + I(Examination\*Infant.Mortality), data)

model\_1

summary(model\_1)

vif(model\_1)

model\_2 = lm(Fertility ~ Examination + I(Examination^2) + I(Infant.Mortality^2) + I(Examination\*Infant.Mortality), data)

model\_2

summary(model\_2)

vif(model\_2)

model\_3 = lm(Fertility ~ I(Examination^2) + I(Infant.Mortality^2) + I(Examination\*Infant.Mortality), data)

model\_3

summary(model\_3)

vif(model\_3)

model\_4 = lm(Fertility ~ I(Examination^2) + I(Infant.Mortality^2), data)

model\_4

summary(model\_4)

vif(model\_4)

*Задача №2.2*

library("lmtest")

library("GGally")

library("car")

data = swiss

help(swiss)

model = lm(Fertility ~ Agriculture + Examination + Infant.Mortality, data)

model

summary(model)

se = 0.07975

t = qt(0.975, df = 43)

model$coefficients[2] - t \* se

model$coefficients[2] + t \* se

confint(model, level = 0.95)

se = 0.22811

t = qt(0.975, df = 43)

model$coefficients[3] - t \* se

model$coefficients[3] + t \* se

confint(model, level = 0.95)

se = 0.45513

t = qt(0.975, df = 43)

model$coefficients[4] - t \* se

model$coefficients[4] + t \* se

confint(model, level = 0.95)

se = 12.82691

t = qt(0.975, df = 43)

model$coefficients[1] - t \* se

model$coefficients[1] + t \* se

confint(model, level = 0.95)

model = lm(Fertility ~ Agriculture + Examination + Infant.Mortality, data)

model

summary(model)

new.data = data.frame(Agriculture = 10, Examination = 30, Infant.Mortality = 20)

predict(model, new.data, interval = "confidence")