Домашняя работа №1

Регрессионный анализ в R

Автор: Миронюк Даниил Никитич

Группа: ПСА-2

2019

Введение(0)

Данные. Для выполнения домашнего задания предлагается использовать набор данных «education» из пакета “robustbase”. Эти данные использовались в работе (Chatterjee and Price, 1977) и отражают расходы на образование в 50 штатах.

Набор данных содержит следующие переменные:

State-штат

Region – регион (1-Северо-восток; 2-Север; 3- Юг; 4- Запад)

X1 – количество жителей в расчете на тысячу жителей, проживающих в урбанизированных районах в

X2 – доход на душу населения

X3 – количество жителей в расчете на тысячу жителей в возрасте младше 18 лет

У – государственные расходы на образование

Предварительная работа с данными (1)

Код:

```
install.packages("robustbase")

library("robustbase")

df<-data.frame(education, package="robustbase")

summary(df)
```

Результат:

```
> library("robustbase")
> df<-data.frame(education, package="robustbase")
> summary(df)
     State        Region          X1              X2            X3              Y              package
 AK     : 1   Min.   :1.00   Min.   :322.0   Min.   :3448   Min.   :287.0   Min.   :208.0   robustbase:50
 AL     : 1   1st Qu.:2.00   1st Qu.:546.8   1st Qu.:4137   1st Qu.:310.8   1st Qu.:234.2
 AR     : 1   Median :3.00   Median :662.5   Median :4706   Median :324.5   Median :269.5
 AZ     : 1   Mean   :2.66   Mean   :657.8   Mean   :4675   Mean   :325.7   Mean   :284.6
 CA     : 1   3rd Qu.:3.75   3rd Qu.:782.2   3rd Qu.:5054   3rd Qu.:333.0   3rd Qu.:316.8
 CO     : 1   Max.   :4.00   Max.   :909.0   Max.   :5889   Max.   :386.0   Max.   :546.0
 (Other):44
>
```

Рис1. Вывод R - описательные статистики

Выводы по пункту:

Мы успешно загрузили данные и провели анализ предварительных статистик.

Важно отметить, что переменные Region и State являются фиктивными. (Качественные переменные, введенные для удобства при работе с данными.) Описательные статистики для них не несут особого смысла.

Регрессионная модель Y~X1 (2)

Код:

summary(m1<-lm(Y ~ X1,data = df))

Результат:

```
> summary(m1<-lm(Y ~ X1,data = df ))

Call:
lm(formula = Y ~ X1, data = df)

Residuals:
   Min     1Q Median     3Q    Max
-76.57 -39.10 -11.03  28.16 285.08

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 194.9741    38.9150   5.010 7.78e-06 ***
X1            0.1363     0.0578   2.357   0.0225 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 58.67 on 48 degrees of freedom
Multiple R-squared:  0.1038,    Adjusted R-squared:  0.08509
F-statistic: 5.557 on 1 and 48 DF,  p-value: 0.02253
```

Рис.2 Результат регрессии Y~X1

Выводы по пункту:

Данная модель регрессии объясняет 10,38% дисперсии результирующего признака.

Коэффициент при регрессоре X1 в данной модели 0.13.

Отметим, что данный коэффициент оценен на уровне значимости 0.01 что является довольно точной оценкой. (99%)

График регрессионной модели Y~X1 (3)

Код:

plot(df$Y~df$X1, xlab="x", ylab="y")

abline(m1)

Результат:

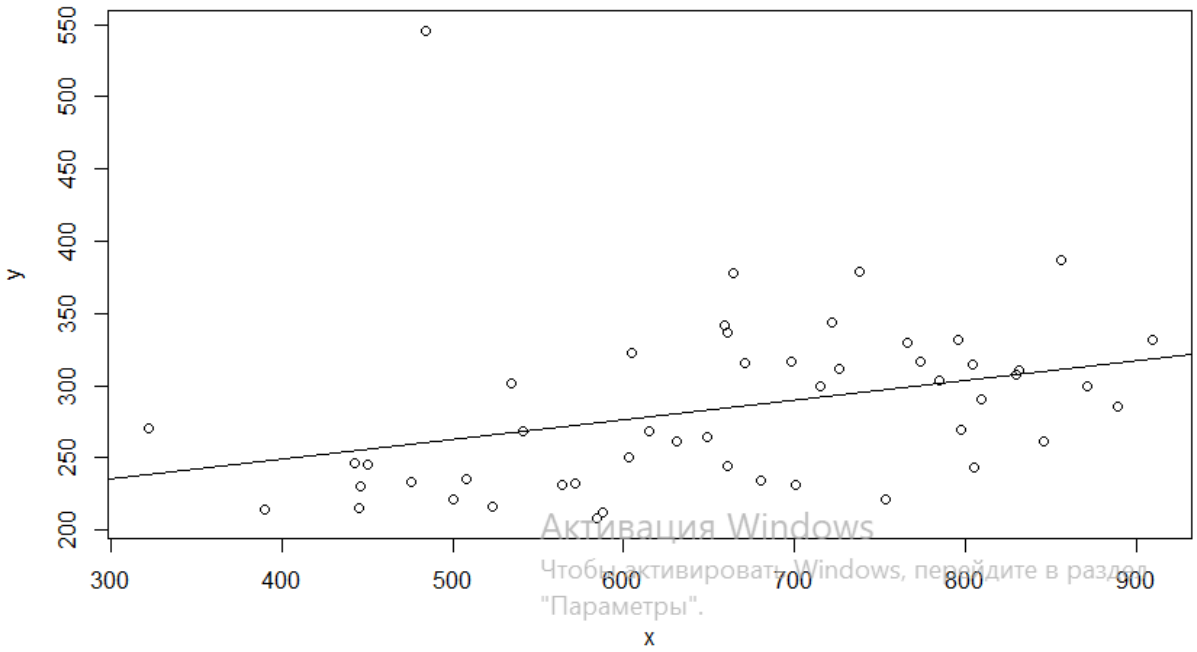

Рис.3 График регрессионной модели Y~X1

Выводы по пункту:

Как видно из графика наша модель действительно, некоторым образом, верно отображает общий характер зависимости. Одна пока ее точность слишком мала. Попробуем включить больше регрессоров.

Регрессионная модель Y~X1+X2+X3 (4)

Код:

```
summary(m2<-lm(Y ~ X1+X2+X3,data = df ))
```

Результат:

```
Call:
lm(formula = Y ~ X1 + X2 + X3, data = df)

Residuals:
    Min      1Q  Median      3Q     Max
-84.878 -26.878  -3.827  22.246  99.243

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -5.566e+02  1.232e+02  -4.518 4.34e-05 ***
X1          -4.269e-03  5.139e-02  -0.083    0.934
X2           7.239e-02  1.160e-02   6.239 1.27e-07 ***
X3           1.552e+00  3.147e-01   4.932 1.10e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 40.47 on 46 degrees of freedom
Multiple R-squared:  0.5913,    Adjusted R-squared:  0.5647
F-statistic: 22.19 on 3 and 46 DF,  p-value: 4.945e-09
```

Рис.4 Результат регрессии Y~X1+X2+X3

Выводы по пункту:

Данная модель значительно лучше объясняет дисперсию целевой переменной. R^2 в данном случае равен 0.5913. Необходимо отметить так же что при добавлении переменных в уравнение регрессии R^2 имеет свойство необоснованно расти. Скорректированый R^2 штрафует модель за добавление лишних регрессоров. В нашем случае его значение (0.5647) близко к обычному R^2.Можно сделать вывод что добавление регрессоров обосновано. В данной модели все коэффициенты регрессоров, кроме коэффициента при X1 оценены на высоком уровне значимости.

Какую дисперсию Y объясняет каждый из регрессоров в модели из п.4? (5)

Код:

anova(m2)

Результат:

```
> anova(m2)
Analysis of Variance Table

Response: Y
          Df Sum Sq Mean Sq F value    Pr(>F)
X1         1  19130   19130  11.679  0.001332 **
X2         1  50042   50042  30.551 1.471e-06 ***
X3         1  39848   39848  24.328 1.103e-05 ***
Residuals 46  75348    1638
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
>
```

Рис 5. Результат ANOVA теста модели.

Выводы по пункту:

Все источники дисперсии, включённые в уравнение оценены значимо. Как мы можем видеть остаточная дисперсия примерно равна 41%, наибольший вклад в объяснение дисперсии моделью вносят X2– доход на душу населения и X3 – количество жителей в расчете на тысячу жителей в возрасте младше 18 лет. Так X1 объясняет примерно 10%, X2-27%, а X3 – 21,6% дисперсии целевого признака соответственно.

Регрессионная модель Y~X1+X2+X3+Region (6)

Код:

```
summary(m6<-lm(Y ~ X1+X2+X3+factor(Region),data = df ))

df <- within(df, Region <- as.factor(Region))

region=relevel(df$Region,2)

summary(m6<-lm(Y ~ X1+X2+X3+region,data = df ))
```

Результат:

```
lm(formula = Y ~ X1 + X2 + X3 + factor(Region), data = df)

Residuals:
    Min      1Q  Median      3Q     Max
-77.963 -25.499  -2.214  17.618  89.106

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)     -451.67542  139.53852  -3.237 0.002329 **
X1                -0.03456    0.05319  -0.650 0.519325
X2                 0.07204    0.01305   5.520 1.82e-06 ***
X3                 1.30146    0.35717   3.644 0.000719 ***
factor(Region)2  -15.72741   18.16260  -0.866 0.391338
factor(Region)3   -8.63998   18.53938  -0.466 0.643543
factor(Region)4   18.59675   19.68837   0.945 0.350163
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 39.88 on 43 degrees of freedom
Multiple R-squared:  0.6292,    Adjusted R-squared:  0.5774
F-statistic: 12.16 on 6 and 43 DF,  p-value: 6.025e-08
```

Рис. 6 Результат регрессии Y=X1+X2+X3+Region,1 регион

Region1(северо-восток) – референтная группа.

Коэффициент 18,60 у факторной переменной Region4(запад) говорит о том, что на западе в среднем траты на образование на 18,60 тыс. долларов выше, чем на северо-востоке.

```
Call:
lm(formula = Y ~ X1 + X2 + X3 + region, data = df)

Residuals:
    Min      1Q  Median      3Q     Max
-77.963 -25.499  -2.214  17.618  89.106

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -467.40283  142.57669  -3.278 0.002073 **
X1            -0.03456    0.05319  -0.650 0.519325
X2             0.07204    0.01305   5.520 1.82e-06 ***
X3             1.30146    0.35717   3.644 0.000719 ***
region1       15.72741   18.16260   0.866 0.391338
region3        7.08742   17.29950   0.410 0.684068
region4       34.32416   17.49460   1.962 0.056258 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 39.88 on 43 degrees of freedom
Multiple R-squared:  0.6292,    Adjusted R-squared:  0.5774
F-statistic: 12.16 on 6 and 43 DF,  p-value: 6.025e-08
```

Рис. 7 Результат регрессии Y=X1+X2+X3+Region,2 регион

Теперь Region2(Север) – референтная группа.

Коэффициент у переменной region3(Юг) означает, что на юге в среднем расходы на образование составляют на 7 тыс. больше чем на севере.

Данные отличия могут быть вызваны различным уровнем экономического благосостояния в различных регионах, различными программами образования и т.д.

С учетом политической системой в США, где традиционно сложились привязанные к географическому положению различные политические течения, а также относительно высокий уровень самоуправления отдельных штатов — это довольно любопытное наблюдение.

Построить график прогноза и доверительных интервалов для него на основе модели из п.3. (7)

Код:

```
# 0. Build linear model

m1<-lm(Y ~ X1,data = df )

# 1. Add predictions

pred.int <- predict(m1, interval = "prediction")

mydata <- cbind(education, pred.int)

# 2. Regression line + confidence intervals
```

```
library("ggplot2")
p <- ggplot(mydata, aes(X1,Y)) +
 geom_point() +
 stat_smooth(method = lm)
# 3. Add prediction intervals
p + geom_line(aes(y = lwr), color = "red", linetype = "dashed")+
 geom_line(aes(y = upr), color = "red", linetype = "dashed")
```

Результат:

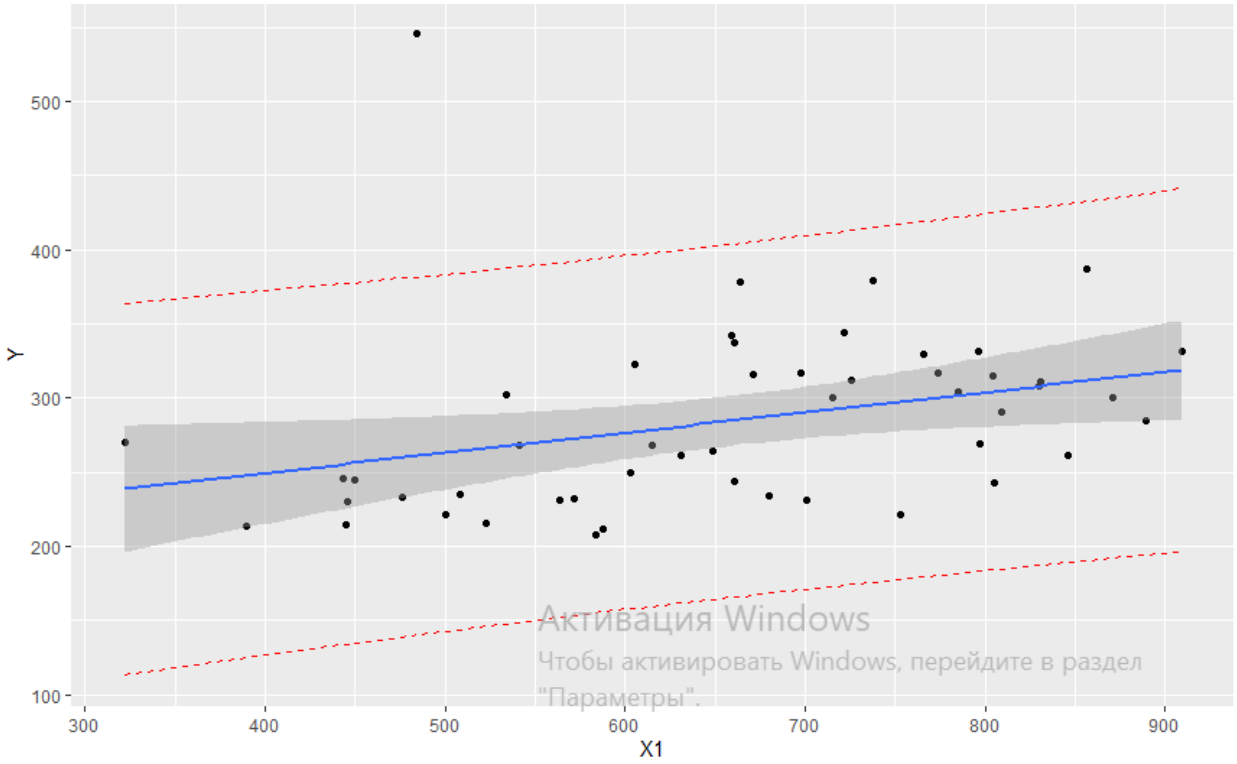

Рис.8 График Прогноза и доверительных интервалов.

Вывод по пункту:

График показывает предсказания нашей модели для расходов на образование в штате на основе количества обучающихся из городов на тысячу человек. Так же мы видим доверительные интервалы предсказаний нашей модели. Заметно что модель далеко не идеально объясняет наши наблюдения. Доверительные интервалы сужаются ближе к средним значениям предиктора и расширяются у крайних значений.

График Residuals vs Fitted для модели из п.6.(8)

Код: plot(m6)

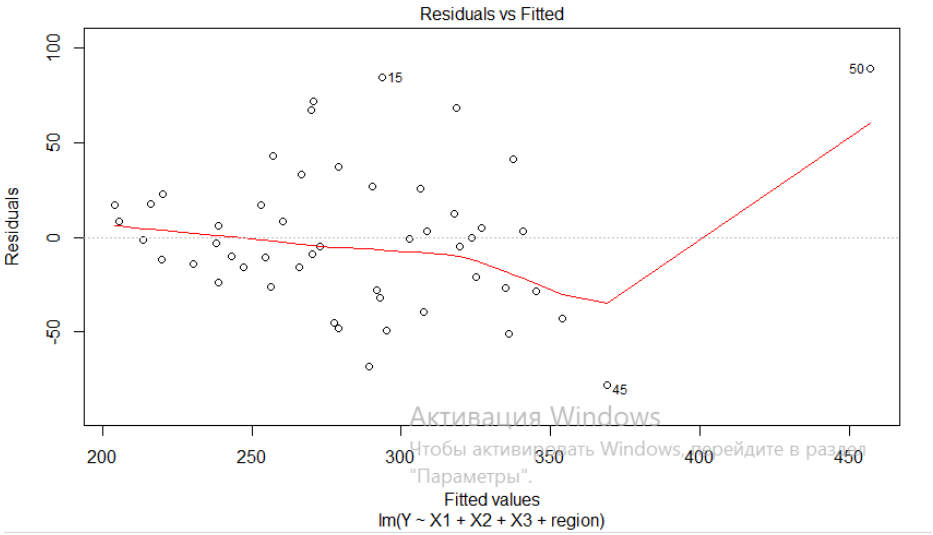

Рис.9 График Residuals vs Fitted

Результат:

Вывод по пункту: По графику видно неравномерное распределение остатков относительно координат. Возможно, присутствует гетеросекдастичность. В таком случае модель некорректна из-за того, что:

- Нарушены предпосылки теоремы Гаусса-Маркова. Получаемые β^olsβ не являются эффективными.
- Несмещенность оценок β^ols сохраняется
- Нарушены предпосылки для использования t-статистик. При использовании стандартных формул t-статистики не имеют t-распределения. Доверительные интервалы и проверка гипотез по стандартным формулам даёт неверные результаты.

Тест Бреуша-Пагана (9)

Код:

```
library("lmtest")
bptest(m1, studentize = FALSE)
```

Результат:

```
> bptest(m6, studentize = FALSE)

        Breusch-Pagan test

data:  m6
BP = 18.278, df = 6, p-value = 0.005573
```

Выводы по пункту:

Тест значим на уровне p-value 0.05. Гипотеза о гомоскедастичности отвергается. Предположение о несостоятельности модели подтверждается.

Устойчивость к гетероскедастичность. (10)

Код:

```
coeftest(m6, vcov = vcovHC(m6, "HC0"))


coeftest(m6, vcov = vcovHC(m6, "HC2"))
```

Результат:

```
> coeftest(m6, vcov = vcovHC(m6, "HC0"))

t test of coefficients:

               Estimate  Std. Error t value  Pr(>|t|)
(Intercept) -467.402827  172.577569 -2.7084  0.009666 **
X1            -0.034558    0.054145 -0.6382  0.526703
X2             0.072036    0.016638  4.3296 8.773e-05 ***
X3             1.301458    0.387743  3.3565  0.001659 **
region1       15.727405   20.488148  0.7676  0.446899
region3        7.087424   17.755889  0.3992  0.691752
region4       34.324157   19.308578  1.7777  0.082532 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

```
> coeftest(m6, vcov = vcovHC(m6, "HC2"))

t test of coefficients:

               Estimate  Std. Error t value Pr(>|t|)
(Intercept) -467.402827  229.676125 -2.0351 0.048040 *
X1            -0.034558    0.069580 -0.4967 0.621957
X2             0.072036    0.022139  3.2538 0.002221 **
X3             1.301458    0.511248  2.5456 0.014575 *
region1       15.727405   22.571885  0.6968 0.489697
region3        7.087424   20.121358  0.3522 0.726383
region4       34.324157   21.326438  1.6095 0.114833
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Рис. 10 Результаты регрессии Y=X1+X2+X3+Region, на основе устойчивых матриц

Выводы по пункту:

Элементы диагональной матрицы могут задаваться разными способами:

const	$\hat{\sigma}^2$
HC0	$\hat{\varepsilon}_i^2$
HC1	$\frac{n}{n-k}\hat{\varepsilon}_i^2$
HC2	$\frac{\hat{\varepsilon}_i^2}{1-h_i}$
HC3	$\frac{\hat{\varepsilon}_i^2}{(1-h_i)^2}$
HC4	$\frac{\hat{\varepsilon}_i^2}{(1-h_i)^{\min\left(4,\frac{nh_i}{k}\right)}}$

Рис. 11 Элементы диагональной матрицы для различных способов.

При использовании матриц устойчивых к гетероскедастичности ковариационную матрицу параметров (HC0 — оценка Уайта, HC1— модификация МакКиннона-Уайта) мы получаем почти такие же коэффициенты уравнения регрессии. Уровни значимости несколько снизились, но у изначально значимых предикторов остались в приемлемых рамках. В случае если бы мы не рассчитали робастные оценки ,использование модели с неверными оценками могло бы привести к неверным выводам.