otchet3laba

Gatiyatullin\_Lenar

13 Р°РїСЂРµР»СЏ 2018 Рі

# Тема 1. Регрессионные уравнения.

## 1.Построить регрессионную модель.

library(ggplot2)  
library(memisc)

## Warning: package 'memisc' was built under R version 3.4.4

## Loading required package: lattice

## Loading required package: MASS

##   
## Attaching package: 'memisc'

## The following objects are masked from 'package:stats':  
##   
## contr.sum, contr.treatment, contrasts

## The following object is masked from 'package:base':  
##   
## as.array

library(DescTools)

## Warning: package 'DescTools' was built under R version 3.4.4

##   
## Attaching package: 'DescTools'

## The following object is masked from 'package:memisc':  
##   
## %nin%

library(broom)  
library(caTools)

## Warning: package 'caTools' was built under R version 3.4.4

library(lmtest)

## Warning: package 'lmtest' was built under R version 3.4.4

## Loading required package: zoo

## Warning: package 'zoo' was built under R version 3.4.4

##   
## Attaching package: 'zoo'

## The following objects are masked from 'package:base':  
##   
## as.Date, as.Date.numeric

library(dplyr)

##   
## Attaching package: 'dplyr'

## The following objects are masked from 'package:memisc':  
##   
## collect, recode, rename

## The following object is masked from 'package:MASS':  
##   
## select

## The following objects are masked from 'package:stats':  
##   
## filter, lag

## The following objects are masked from 'package:base':  
##   
## intersect, setdiff, setequal, union

library(readxl)  
dmnds <- read.csv(file = "C:\\Users\\Lenar\\Desktop\\лабы\\статмоделирование\\laba3\\bikes\_rent.csv", sep = ',')  
View(dmnds)

set.seed(56)  
split <- sample.split(dmnds$temp, SplitRatio = 0.75)  
train <- subset(dmnds, split == TRUE)  
test <- subset(dmnds, split == FALSE)

model\_1 <- lm (data = train, cnt ~ temp)  
summary(model\_1)

##   
## Call:  
## lm(formula = cnt ~ temp, data = train)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -4588.4 -1144.9 -126.3 1078.1 3758.9   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 1216.499 187.055 6.503 1.78e-10 \*\*\*  
## temp 160.935 8.678 18.546 < 2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 1519 on 546 degrees of freedom  
## Multiple R-squared: 0.3865, Adjusted R-squared: 0.3854   
## F-statistic: 344 on 1 and 546 DF, p-value: < 2.2e-16

## 2.Проанализировать адекватность построенного регрессионного уравнения.

### 2.1 Оценить качества подгонки.

Нахождение остаточной дисперсии.

summary(lm(data = train, cnt ~ temp))$sigma^2

## [1] 2306584

Нахождение коэффициентов детерминации.

summary(lm(data = train, cnt ~ temp))$r.squared

## [1] 0.3864829

summary(lm(data = train, cnt ~ temp))$adj.r.squared

## [1] 0.3853593

Вывод: Так как величина остаточной дисперсии велика, то это говорит о низкой качестве подгонки. Коэффициенты детерминации далеки от единицы, что также говорит о низкой качестве подгонки.

### 2.2 Проверить различные гипотезы относительно параметров уравнения.

1. Наличие линейной связи. F-критерий. H0: a = b =  0 , (т.е. линейная связь между x и y отсутствует); H1: a^2 + b^2 = 0  , (т.е. наличие линейной связи).

F расчетное F-statistic: 344 on 1 and 546 DF, p-value: < 2.2e-16

qf(.95, 1, 546)

## [1] 3.858546

qf(.99, 1, 546)

## [1] 6.68152

F𝐹табл F< 𝐹расч ⇒ отвергаем H0 в пользу H1

1. Значимость коэффициентов. t-критерий. H0b: b = 0, (т.е. фактор Х незначим); Н1b: b != 0, (т.е. фактор Х значим).

Н0a: a = 0 (свободный член незначим) H1a: a != 0 (свободный член значим)

t value  
15.03  
18.55

qt(.95, 546)

## [1] 1.647649

qt(.99, 546)

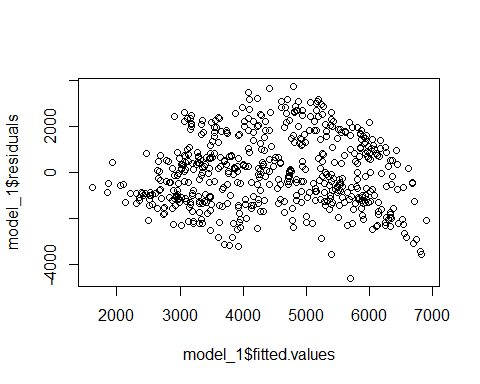
## [1] 2.333197

t расч > t таб, значит отвергаем H0 в пользу H1.

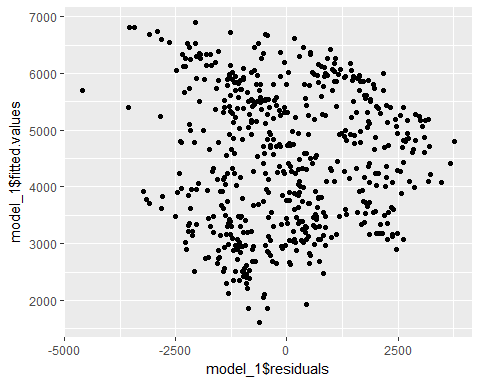
### 2.3 Проверить условия для получения состоятельных, несмещенных, эффективных оценок.

#### График зависимостри остатков от спрогнозированных значений модели.

plot(model\_1$fitted.values, model\_1$residuals)



qplot(y = model\_1$fitted.values, x = model\_1$residuals)



#### Проверка условия M(e) = 0

Н0: |t расч| < t табл => M(e) = 0 H1: |t расч| > t табл => M(e) != 0

mean(model\_1$residuals)

## [1] -1.64014e-13

Проверим!

Для вычисления расч t воспользуемся следующими возможностями R:

a<-mean(model\_1$residuals)  
b<-sd(model\_1$residuals, na.rm = FALSE)  
n <- sqrt(548)  
tm <- a/b\*n  
str(tm)

## num -2.53e-15

Для вычисления tтабл воспользуемся командой qt()

qt(0.10,42125)

## [1] -1.281572

|t расч| < t табл, не опровергаем H0 гиппотезу, а значит мат. ожидание равно нулю

#### Проверка условия D(e) = const

Н0: гетероскедастичности нет H1: гетероскедастичность есть

Проведем в R Тест Бройша-Погана

bptest(model\_1)

##   
## studentized Breusch-Pagan test  
##   
## data: model\_1  
## BP = 10.216, df = 1, p-value = 0.001392

Поскольку значение p-value меньше 0.05, нулевая гипотеза о гомоскедастичности отвергается.

#### Проверка условия cov(ei, ej) = 0.

Н0: ρ=0 (т.е. автокорреляция остатков отсутствует); H1: ρ>0 или ρ<0 (наличие положительной или отрицательной автокорреляции остатков).

Проведем в R тесты Бройша- Годфри и Дарбина-Уотсона

bgtest(model\_1)

##   
## Breusch-Godfrey test for serial correlation of order up to 1  
##   
## data: model\_1  
## LM test = 293.09, df = 1, p-value < 2.2e-16

dwtest(model\_1)

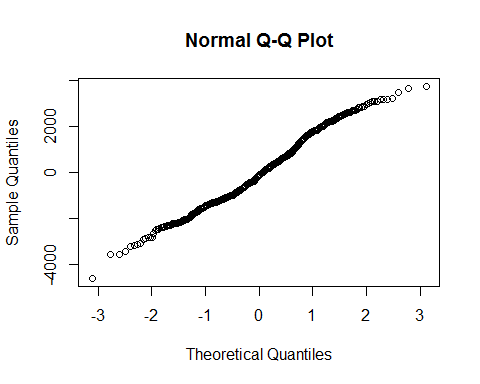
##   
## Durbin-Watson test  
##   
## data: model\_1  
## DW = 0.53329, p-value < 2.2e-16  
## alternative hypothesis: true autocorrelation is greater than 0

Поскольку значение p-value меньше 0.05 нулевая гипотеза об отсутствии автокорреляции остатков отвергается.

#### Проверка условия согласованности остатков регрессии с нормальным законом.

Н0: ρ=0 согласуется; H1: ρ>0 или ρ<0 не согласуется.

qqnorm(model\_1$residuals)



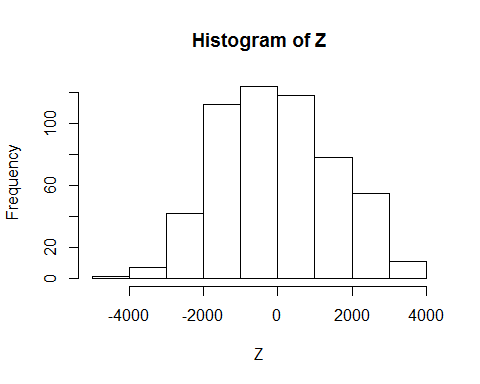
library(sm)

## Package 'sm', version 2.2-5.4: type help(sm) for summary information

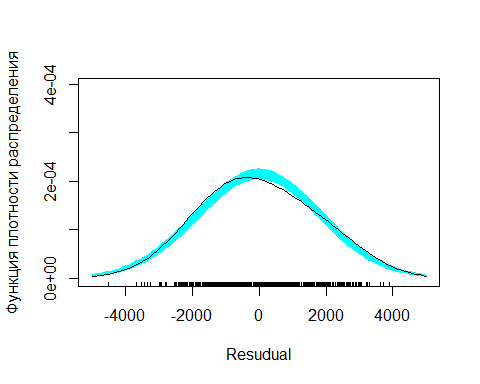
##   
## Attaching package: 'sm'

## The following object is masked from 'package:MASS':  
##   
## muscle

Z<- model\_1$residuals  
hist(Z)



sm.density(model\_1$residuals, model = "Normal",xlab = "Resudual", ylab = "Функция плотности распределения", xlim=c(-5000,5000))



Рассмотрим параметрические тесты:

library(nortest)   
lillie.test(model\_1$residuals)

##   
## Lilliefors (Kolmogorov-Smirnov) normality test  
##   
## data: model\_1$residuals  
## D = 0.056249, p-value = 0.0002874

Поскольку значение p-value меньше 0.05 нулевая гипотеза о согласие распределения остатков с нормальным законом распределения отвергается. ###2.4 Содержательный анализ модели и корректировка модели.

Так как мы получили регрессионное уравнение y = 1216.499 + 160.935\*x, то это говорит о том, что с увеличением значения температуры на единицу, прокат велосипедов увеличивается на 1377.

### 2.5 Прогнозирование значений по полученной модели на тестовой выборке

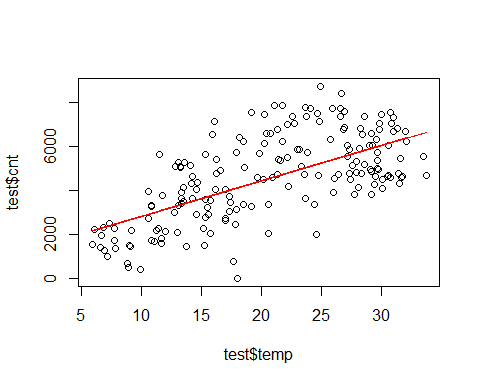
test$model\_1 <- predict(model\_1, test)  
View(test$model\_1)

Посмотрим результат модели на графике

length(test$temp)

## [1] 183

plot(test$temp, test$cnt)  
lines(test$temp, predict(model\_1, test), col = 'red')

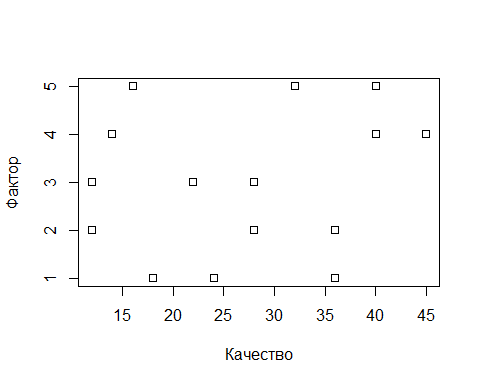
 Вывод: Наша модель обладает гетероскедастичностью, отсутствует автокорреляция, остатки не согласуются с нормальным распределением,да и сама модель плохо предсказывает, так как зависимость между температурой и прокатом не такая явная. #Тема 2. «Дисперсионный анализ».

## 2.1. Однофакторный дисперсионный анализ

№ 2. При уровне значимости =0,05 методом дисперсионного анализа проверить нулевую гипотезу о влиянии фактора на качество объекта на основании пяти измерений для трех уровней фактора Ф1-Ф3.

Z<- read.table(file = "C:\\Users\\Lenar\\Desktop\\лабы\\статмоделирование\\laba3\\StatMod\_3laba\\kd.csv", sep = ';')  
View(Z)

attach(Z)  
stripchart(V1 ~ V3, xlab = "Качество", ylab = "Фактор")



tapply(V1, V3, mean)

## 1 2 3 4 5   
## 26.00000 25.33333 20.66667 33.00000 29.33333

summary(aov(V1 ~ V3, data = Z))

## Df Sum Sq Mean Sq F value Pr(>F)  
## V3 1 61.6 61.63 0.487 0.497  
## Residuals 13 1644.1 126.47

qt(.95, 821)

## [1] 1.646712

qt(.99, 821)

## [1] 2.330898

Так как F > Ftable, то из этого следует, что возраст и стаж влияют на производительность.

## 2.2.Многофакторный дисперсионный анализ

№ 1. Определить, влияют ли возраст и стаж работы конкретных специалистов на среднюю производительность труда.

disp<- read.table(file = "C:\\Users\\Lenar\\Desktop\\лабы\\статмоделирование\\laba3\\ds.csv", sep = ';')  
View(disp)

summary(aov(V1 ~ V2 + V3, data = disp))

## Df Sum Sq Mean Sq F value Pr(>F)   
## V2 1 1568.7 1568.7 65.70 4.59e-11 \*\*\*  
## V3 1 792.1 792.1 33.18 3.54e-07 \*\*\*  
## Residuals 57 1360.8 23.9   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

df(.05, 2,12)

## [1] 0.9435635

Так как F > Ftable, возраст и стаж влияют на производительность.

№ 3 На предприятии Н. повысился уровень травматизма в связи с чем врач провел исследование отдельных факторов, среди которых изучался стаж работы работающих в цехах. Выборки сделаны на предприятии Н. из 4 цехов с близкими условиями и характером труда. Уровни травматизма рассчитаны на 100 работающих за прошлый год.

S<- read.table(file = "C:\\Users\\Lenar\\Desktop\\лабы\\статмоделирование\\laba3\\StatMod\_3laba\\d.csv", sep = ';')  
View(S)

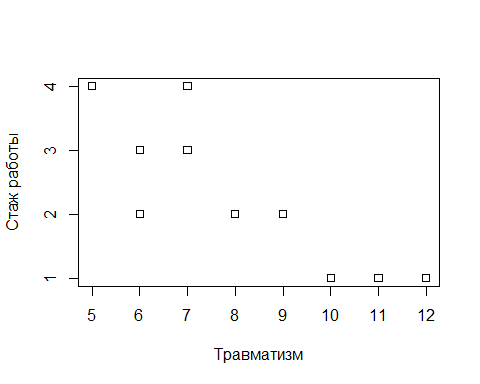
summary(aov(V1 ~ V3, data = S))

## Df Sum Sq Mean Sq F value Pr(>F)   
## V3 1 33.75 33.75 19.66 0.00127 \*\*  
## Residuals 10 17.17 1.72   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

attach(S)

## The following objects are masked from Z:  
##   
## V1, V2, V3

stripchart(V1 ~ V3, xlab = "Травматизм", ylab = "Стаж работы")



tapply(V1, V3, mean)

## 1 2 3 4   
## 11.000000 7.666667 6.666667 6.333333

summary(aov(V1 ~ V3, data = S))

## Df Sum Sq Mean Sq F value Pr(>F)   
## V3 1 33.75 33.75 19.66 0.00127 \*\*  
## Residuals 10 17.17 1.72   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

qt(.95, 821)

## [1] 1.646712

qt(.99, 821)

## [1] 2.330898

Травматизм зависит от стажа работы.