Модульное домашнее задание №1

1. Загрузить набор данных в R, построить таблицу с описательными статистиками.

Таблица 1. Описательные статистики количественных статистических данных

	vars	n	mean	sd	median	trimmed	mad	min	max	range	skew	kurtosis	se
X1	1	50	657.80	145.02	662.5	661.82	180.88	322	909	587	-0.24	-0.90	20.51
X2	2	50	4675.12	644.51	4706.0	4666.15	773.92	3448	5889	2441	0.05	-1.05	91.15
X3	3	50	325.74	19.42	324.5	323.72	15.57	287	386	99	0.92	1.21	2.75
Y	4	50	284.60	61.34	269.5	278.28	57.82	208	546	338	1.56	4.37	8.67

2. Построить регрессионную модель Y~X1, проинтерпретировать результаты.

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

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.1 Результаты построения регрессионной модели Y~X1

Интерпретация:
- Модель: Y = 194,9741 + 0,1363*X1
- P-value ≈ 0,022 < 0,05, значит можно говорить о справедливости регрессионной модели для всей генеральной совокупности
- R^2 ≈ 0,1, значит только 10% результатов наблюдений объясняются регрессионной моделью.
- $F_{1,48}$ = 5,557, p-value = 0.2253 <u>или</u> t = 2,357, df = 48, p-value = 0.2253

3. Построить график Y~X1, на котором точками будут отмечены исходные данные и на котором будет представлена линия регрессии (одновременно).

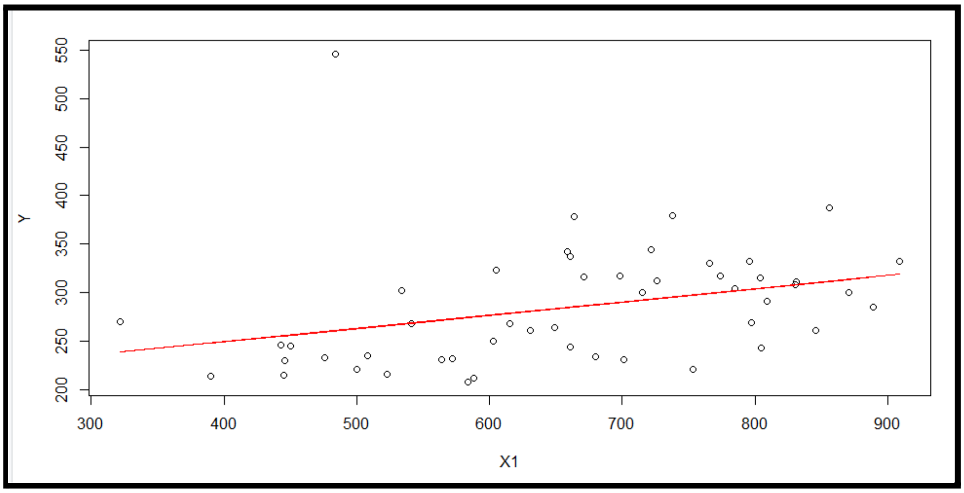

Рис. 2 График исходных данных и линии регрессии модели Y~X1

4. Построить регрессионную модель Y~X1+X2 +X3, проинтерпретировать результаты.

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

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. 3 Результаты построение регрессионной модели Y~X1+X2 +X3

Интерпретация:

- Модель: Y = -556,6 - 0,004269*X1 + 0,07239*X2 + 1,552*X3
- P-value ≈ 0,000000005 < 0,05, значит можно говорить о справедливости регрессионной модели для всей генеральной совокупности
- R^2 ≈ 0,6, значит только 60% результатов наблюдений объясняются регрессионной моделью.

5. Какую дисперсию Y объясняет каждый из регрессоров в модели Y~X1+X2 +X3?

- Найдем частные корреляции между Y и каждым из регрессоров, чтобы определить их влияние на показатели зависимой переменной.

$r_{yx1/x2,x3} \approx -0{,}012$, значит значения Y имеют слабую обратную зависимость от показателя X1

$r_{yx2/x1,x3} \approx 0{,}677$, значит значения Y имеют прямую выше средней зависимость от показателя X2

$r_{yx3/x1,x2} \approx 0{,}588$, значит значения Y имеют прямую среднюю зависимость от показателя X3

6. Построить регрессионную модель Y~X1+X2 +X3+Region, проинтерпретировать результаты. Где в среднем выше расходы на образование в расчете на душу населения?

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

Residuals:
    Min      1Q  Median      3Q     Max
-80.143 -24.595  -4.734  15.016  95.274

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -522.77813  126.45851  -4.134 0.000153 ***
X1            -0.01803    0.05269  -0.342 0.733707
X2             0.07509    0.01182   6.355 9.27e-08 ***
X3             1.37998    0.34905   3.953 0.000270 ***
Region         7.02248    6.24147   1.125 0.266499
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 40.36 on 45 degrees of freedom
Multiple R-squared:  0.6025,    Adjusted R-squared:  0.5672
F-statistic: 17.05 on 4 and 45 DF,  p-value: 1.406e-08
```

Рис. 4 Результаты построение регрессионной модели Y~X1+X2 +X3+Region

Интерпретация:

- Модель: Y = -522,77813 - 0,01803*X1 + 0,07509*X2 + 1,37998*X3 + 7,02248*Region
- P-value ≈ 0,00000001 < 0,05, значит можно говорить о справедливости регрессионной модели для всей генеральной совокупности
- $R^2 \approx 0{,}6$, значит только 60% результатов наблюдений объясняются регрессионной моделью.

Сравнение:

- Принимаем северо-восточный регион за базовый уровень, тогда коэффициент B при западном регионе будет равен 18,59675. Значит расходы на образование на душу населения больше на западе по сравнению с северо-востоком.

```
Call:
lm(formula = Y ~ X1 + X2 + X3 + relevel(Region1, ref = "1"),
    data = education)

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 ***
relevel(Region1, ref = "1")2  -15.72741   18.16260  -0.866 0.391338
relevel(Region1, ref = "1")3   -8.63998   18.53938  -0.466 0.643543
relevel(Region1, ref = "1")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. 5 Результаты построение регрессионной модели Y~X1+X2 +X3+Region с измененным базовым уровнем

- Принимаем северный регион за базовый уровень, тогда коэффициент B при южном регионе будет равен 7,08742. Значит расходы на образование на душу населения больше на юге по сравнению с севером.

```
Call:
lm(formula = Y ~ X1 + X2 + X3 + relevel(Region1, ref = "2"),
    data = education)

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 ***
relevel(Region1, ref = "2")1   15.72741   18.16260   0.866 0.391338
relevel(Region1, ref = "2")3    7.08742   17.29950   0.410 0.684068
relevel(Region1, ref = "2")4   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. 6 Результаты построение регрессионной модели Y~X1+X2 +X3+Region с измененным базовым уровнем (2)

7. Построить график прогноза и доверительных интервалов для него на основе модели Y~X1.

- Найдем доверительные интервалы коэффициентов с уровнем доверия 0,95.

```
                  2.5 %      97.5 %
(Intercept) 116.73018067 273.2179181
X1            0.02003861   0.2524635
```

Рис. 7 Результаты нахождения доверительные интервалы коэффициентов с уровнем доверия 0,95 для регрессионной модели Y~X1.

- Составим уравнения для прогноза y^ = 194,9741 + 0,1363 * X1 на основе коэффициентов полученных в п.2, а также уравнения yb = 116,73018067 + 0,020038617 * X1 для нижней границы и yt = 273,2179181 + 0,2524635 * X1 для верхней границы, используя полученные коэффициенты.
- Построим полученные графики.

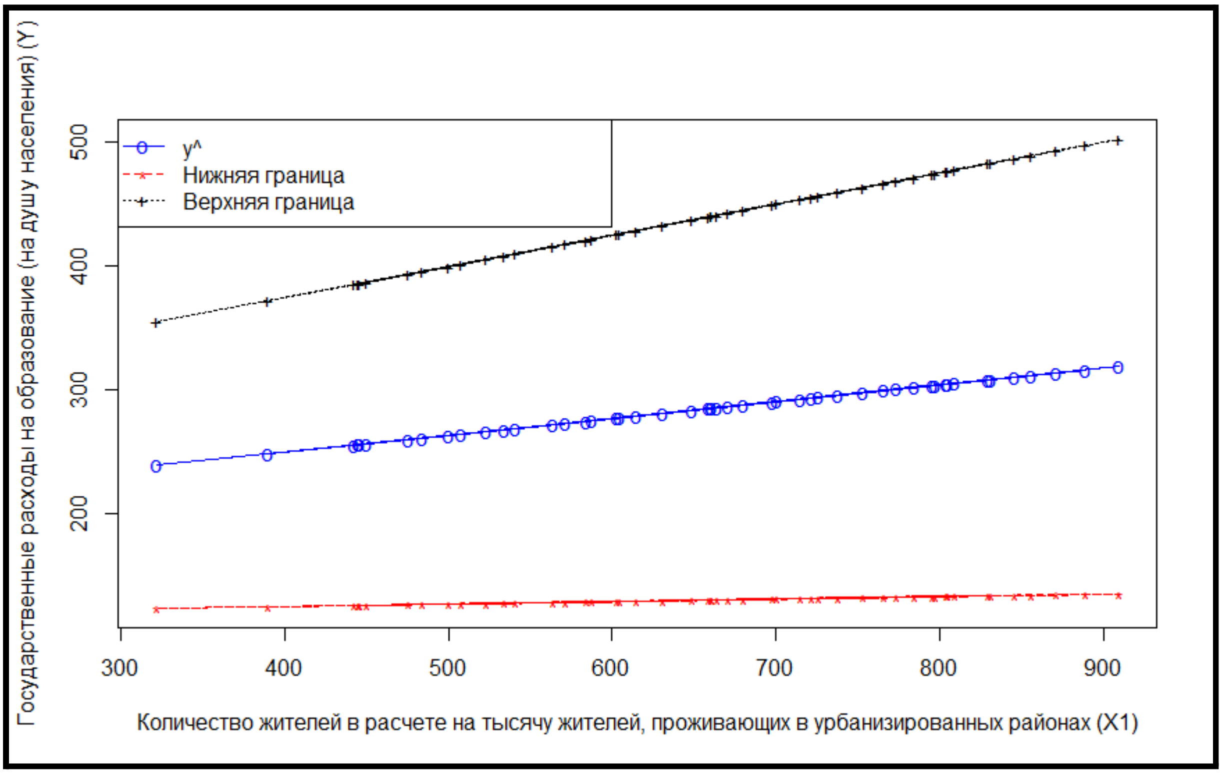

Рис. 8 График прогноза и доверительных интервалов для него на основе модели Y~X1

- Также есть возможность построить данный график с помощью встроенной функции R.

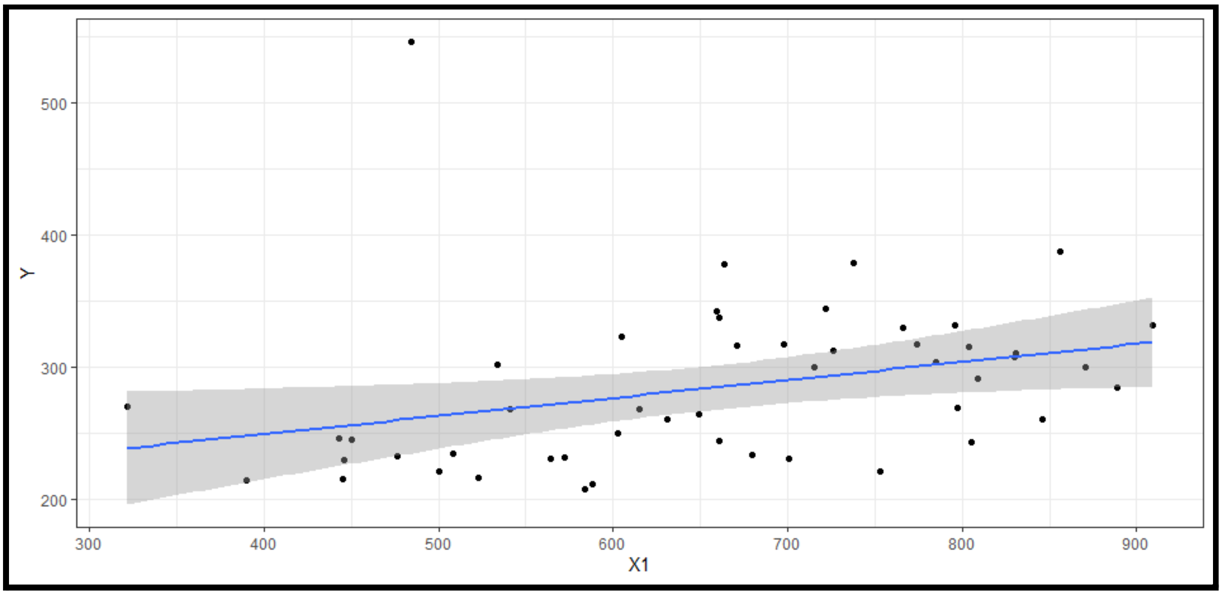

Рис. 9 График прогноза и доверительных интервалов для него на основе модели Y~X1 (2)

8. Построить и проанализировать график Residuals vs Fitted Для модели Y~X1+X2+X3.

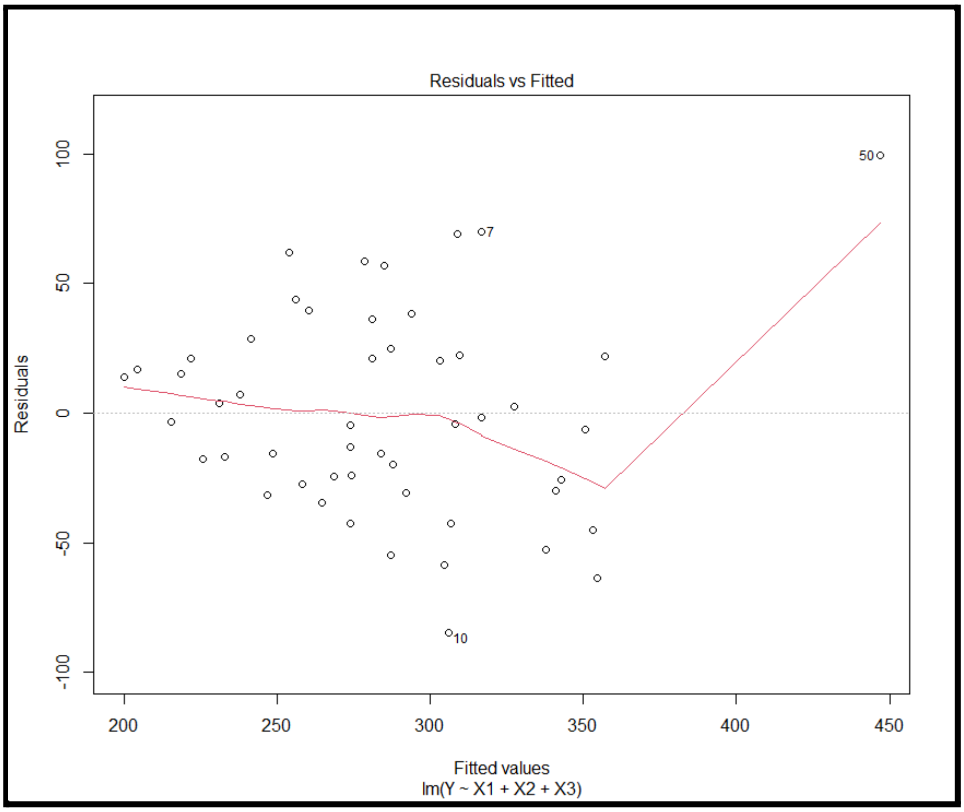

Рис. 10 График Residuals vs Fitted на основе модели Y~X1+X2+X3

Вывод: На данном графике видно, что значение остатков не меняется в зависимости от прогнозных значений y^, однако количество остатков растет вместе с увеличением рассчитанных значений. Хоть увеличение количества остатков и их рассеивание непостоянны, в модели все же может присутствовать гетероскедастичность.

9. Проверить наличие гетероскедастичности в модели Y~X1+X2 +X3+Region на основе теста Бреуша-Пагана.

```
> bptest(lm.education3)

	studentized Breusch-Pagan test

data:  lm.education3
BP = 15.99, df = 4, p-value = 0.003032

> bptest(lm.education3, studentize=FALSE)

	Breusch-Pagan test

data:  lm.education3
BP = 14.541, df = 4, p-value = 0.005753
```

Рис 11 Результаты тестов Бреуша – Пагана на наличие гетероскедастичности

Проведя классическую версию теста, а также версию Коэнкера, можно увидеть, что в обоих случаях значения p-value < 0.05, а значит отвергается нулевая гипотеза о гомоскедастичности и предполагается наличие гетероскедастичности.

10. Получить устойчивую к гетероскедастичности ковариационную матрицу параметров модели Y~X1+X2 +X3+Region. Рассчитать новые t-статистики. Сравнить результаты.

```
> vcov(lm.education3)
              (Intercept)           X1            X2            X3       Region
(Intercept) 15991.7556616 -1.181366602 -0.5120354817 -4.078975e+01 187.44370462
X1             -1.1813666  0.002775739 -0.0003772860  4.059900e-03  -0.07636417
X2             -0.5120355 -0.000377286  0.0001396150  2.075589e-04   0.01499521
X3            -40.7897529  0.004059900  0.0002075589  1.218393e-01  -0.95456103
Region        187.4437046 -0.076364174  0.0149952125 -9.545610e-01  38.95599217
> coeftest(lm.education3)

t test of coefficients:

              Estimate  Std. Error t value  Pr(>|t|)    
(Intercept) -522.778129  126.458514 -4.1340 0.0001533 ***
X1            -0.018035    0.052685 -0.3423 0.7337072    
X2             0.075088    0.011816  6.3549 9.268e-08 ***
X3             1.379979    0.349055  3.9535 0.0002695 ***
Region         7.022480    6.241474  1.1251 0.2664987    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Рис 12 Ковариационная матрица и результаты t-тестов для исходной модели Y~X1+X2 +X3+Region

```
> vcovHC(lm.education3)
              (Intercept)           X1            X2            X3       Region
(Intercept) 74567.892353 18.230100882 -6.5145757234 -170.71580835 -547.4011392
X1             18.230101  0.007662670 -0.0020279880   -0.04093173   -0.2552210
X2             -6.514576 -0.002027988  0.0006799251    0.01406643    0.0697517
X3           -170.715808 -0.040931734  0.0140664304    0.40495774    0.8010198
Region       -547.401139 -0.255221048  0.0697517023    0.80101978   46.7613921
> coeftest(lm.education3, vcov = vcovHC(lm.education3))

t test of coefficients:

               Estimate  Std. Error t value Pr(>|t|)
(Intercept) -522.778129 273.071222 -1.9144 0.061935 .
X1            -0.018035   0.087537 -0.2060 0.837699
X2             0.075088   0.026075  2.8797 0.006074 **
X3             1.379979   0.636363  2.1685 0.035443 *
Region         7.022480   6.838230  1.0269 0.309935
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Рис 13 Устойчивая к гетероскедастичности ковариационная матрица на основе параметров модели Y~X1+X2 +X3+Region и результаты t-тестов на ее основе

Вывод: оценки при проведении теста на основе новой матрицы остались такими же, однако стандартные ошибки при изменились и стали больше, а значит интервальная оценка будет другой с учетом изменившихся показателей стандартных ошибок.

Код R:

```
#Пункт 1
install.packages("robustbase")
install.packages("psych")
data(education, package="robustbase")
library(psych)
describe(data.matrix(education[, 3:6]))


#Пункт 2
lm.education1 <- lm(Y ~ X1, data=education)
summary(lm.education1)

#Пункт 3
Y <- data.matrix(education[,6])
X1 <- data.matrix(education[,3])
plot(Y ~ X1)
abline(lm.education1, col = 'red')

#Пункт 4
```

```
lm.education2 <- lm(Y ~ X1+X2+X3, data=education)
summary(lm.education2)

# Пункт 5
install.packages("ggm")
install.packages('foreign')
X2 <- data.matrix(education[,4])
X3 <- data.matrix(education[,5])
library(ggm)
pcor(c("Y","X1","X2","X3"), var(data.matrix(education[, 3:6])))
pcor(c("Y","X2","X1","X3"), var(data.matrix(education[, 3:6])))
pcor(c("Y","X3","X2","X1"), var(data.matrix(education[, 3:6])))

#Пункт 6
lm.education3 <- lm(Y ~ X1+X2+X3+Region, data=education)
summary(lm.education3)
library("vcd")
Region <- data.matrix(education[,2])
Region1 <- factor(Region, c(1,2,3,4))
lm.education31 <- lm(Y ~ X1+X2+X3+relevel(Region1,ref = "1"), data=education)
summary(lm.education31)
lm.education32 <- lm(Y ~ X1+X2+X3+relevel(Region1,ref = "2"), data=education)
summary(lm.education32)


#Пункт 7 (Первый вариант)
confint(lm.education1, level = 0.95)
install.packages("ggpubr")
library(ggpubr)
y1 <- 194.9741 + 0.1363 * X1          #y^
yb <- 116.73018067 + 0.020038617 * X1    #уравнение для нижней границы
yt <- 273.2179181 + 0.2524635 * X1       #уравнение для верхней границы
plot(X1, y1, type="o", col="blue", pch="o", lty=1, ylim=range(yb,yt),  ylab="Государственные
расходы на образование (на душу населения) (Y)", xlab = "Количество жителей в
расчете на тысячу жителей, проживающих в урбанизированных районах (X1)")
points(X1, yb, col="red", pch="*")
lines(X1, yb, col="red",lty=2)
points(X1, yt, col="black",pch="+")
lines(X1, yt, col="black", lty=3)
legend(298, 518,legend=c("y^","Нижняя граница","Верхняя граница"),
col=c("blue","red","black"),
      pch=c("o","*","+"),lty=c(1,2,3), ncol=1)

#Пукт 7 (Второй вариант, со встроенной функцией R)
ggplot(education, aes(x = X1, y = Y)) +
  geom_point() +
  geom_smooth(method = "lm", se = TRUE)
```

```
#Пукнт 8
plot(lm(Y ~ X1+X2+X3))

#Пункт 9
install.packages("bstats")
install.packages("lmtest")
install.packages("zoo")

library(ggplot2) # графики
library(sandwich) # оценка Var для гетероскедастичности
library(bstats) # тест Уайта, тест Бройша-Пагана
library(lmtest) # тест Бройша-Пагана


theme_set(theme_bw())

bptest(lm.education3)
bptest(lm.education3, studentize=FALSE)

#Пункт 10
vcov(lm.education3)
coeftest(lm.education3)
vcovHC(lm.education3)
coeftest(lm.education3, vcov = vcovHC(lm.education3))
```