Отчет по лабораторной работе
Регрессия

7 февраля 2024 г.

Шаронов Артем
Группа: 5140201/30302

Задание №1

Загрузите данные из файла reglab1.txt. Используя функцию lm, постройте регрессию (используйте разные модели). Выберите наиболее подходящую модель, объясните свой выбор.

```
# Код на R
# Загрузка данных
data <- read.delim("D:/ML/Лабы по МО/Regression/reglab1.txt",␣
 ↪stringsAsFactors = TRUE)

# Выявим зависимый параметр
reg<-lm (x~.,data)
summary(reg)$r.squared

reg<-lm (y~.,data)
summary(reg)$r.squared

reg<-lm (z~.,data)
summary(reg)$r.squared

# Подберем более походящую модель
reg<-lm (z~x+y,data)
summary(reg)$r.squared

reg<-lm (z~(x+y)^2,data)
summary(reg)$r.squared

reg<-lm (z~x*y,data)
    summary(reg)$r.squared
```

Сравним модели по коэффициенту детерминации:

$x \sim . \rightarrow 0.9186677$

$y \sim . \rightarrow 0.9505275$

$z \sim . \rightarrow 0.9686287$

$z \sim x+y \rightarrow 0.9686287$

$z \sim (x+y)^2 \rightarrow 0.9996832$

$z \sim x*y \rightarrow 0.9996832$

Для зависимого параметра z лучшей моделью будет x*y.

Задание №2

Реализуйте следующий алгоритм для уменьшения количества признаков, используемых для построения регрессии: для каждого k d 0,1,..., выбрать подмножество признаков мощности 1 k , минимизирующее остаточную сумму квадратов RSS. Используя полученный алгоритм, выберите оптимальное подможество признаков для данных из файла reglab2.txt. Объясните свой выбор. Для генерации всех возможных сочетаний по m элементов из некоторого множества x можно использовать функцию combn(x, m, ...).

```r
# Код на R
# Загрузка данных
data <- read.delim("D:/ML/Лабы по МО/Regression/reglab2.txt",␣
 ↪stringsAsFactors = TRUE)

# Создаем вектор x, который содержит названия всех столбцов в данных␣
 ↪кроме первого
x <- colnames(data)[-1]

# Инициализация таблицы для результатов
results <- data.frame(Formula = character(), RSS = numeric())

# Перебираем подмножества признаков и вычисляем для каждого RSS
for (k in 1 : length(x)) {
  combs <- combn(x, k)
  for (c in 1:dim(combs)[2]) {
    form <- as.formula(paste("y ~", paste(combs[,c], collapse = "+")))
    reg <- lm(form, data)
    results <- rbind(results, data.frame(Formula = capture.output(form),␣
 ↪RSS = sum(reg$residuals^2)))
  }
}
```

Formula	RSS
y ~ x1	157.2197758
y ~ x2	268.2457708
y ~ x3	393.4904686
y ~ x4	394.5904975
y ~ x1 + x2	0.5379617
y ~ x1 + x3	156.3540657
y ~ x1 + x4	157.2192683
y ~ x2 + x3	267.7954542
y ~ x2 + x4	267.8061361
y ~ x3 + x4	393.4587281
y ~ x1 + x2 + x3	0.3322662
y ~ x1 + x2 + x4	0.3619682
y ~ x1 + x3 + x4	156.3483397
y ~ x2 + x3 + x4	267.4415472
y ~ x1 + x2 + x3 + x4	0.1928635

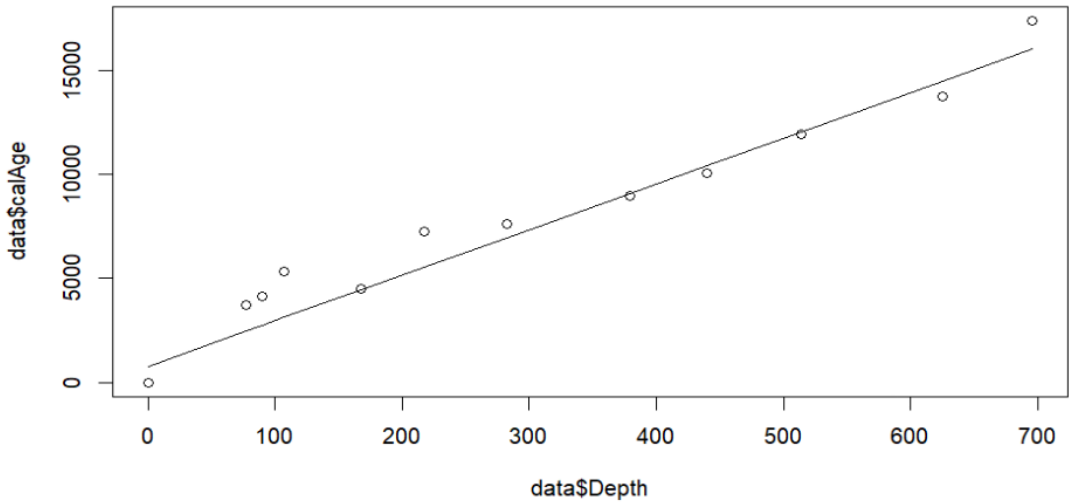
Оптимальным подмножеством параметров, по минимальной остаточной сумме квадратов, является x1, x2, x3, x4, т.е. всё множество параметров.

Задание №3

Загрузите данные из файла cygage.txt. Постройте регрессию, выражающую зависимость возраста исследуемых отложений от глубины залегания, используя веса наблюдений. Оцените качество построенной модели.

```
# Код на R
# Загрузка данных
data <- read.delim("D:/ML/Лабы по МО/Regression/cygage.txt",␣
 ↪stringsAsFactors = TRUE)

# Построение модели
reg <- lm(calAge ~ Depth, data, weights = data$Weight)

# Построение графика линейной регрессии
plot(data$Depth, data$calAge)
lines(data$Depth, predict(reg))

# Вывод коэффициента детерминации
summary(reg)$r.squared
```

График линейной регрессии:



Коэффициента детерминации равен 0.9736839.

Задание №4

Загрузите данные Longley (макроэкономические данные). Данные состоят из 7 экономических переменных, наблюдаемых с 1947 по 1962 годы (n=16):

GNP.deflator - дефлятор цен,

GNP - валовой национальный продукт,

Unemployed – число безработных

Armed.Forces – число людей в армии

Population – население, возраст которого старше 14 лет

Year - год

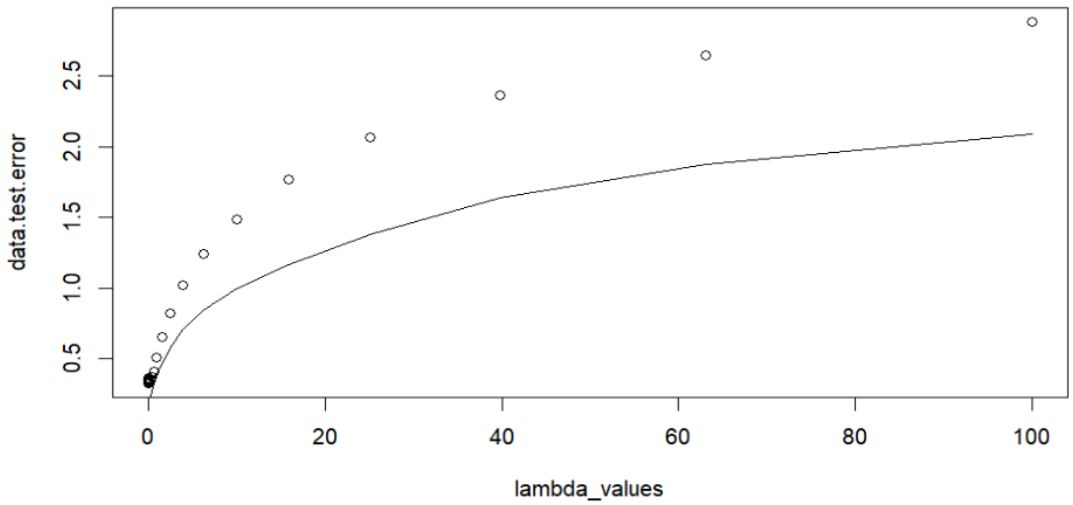
Employed – количество занятых

Построить регрессию lm(Employed .) . Исключите из набора данных longley переменную "Population". Разделите данные на тестовую и обучающую выборки равных размеров случайным образом. Постройте гребневую регрессию для значений 3 0.2 10 , 0,...,25 i i , подсчитайте ошибку на тестовой и обучающей выборке для данных значений λ, постройте графики. Объясните полученные результаты..

```
# Код на R
# Загрузка данных
data <- longley

# Исключаем переменную Population
data <- longley[,-5]

# Разделение данных на обучающую и тестовую выборку раных размеров
n_row = nrow(data)
train_sample <- sample(1: n_row, round(0.5 * n_row))
data.train <- data[train_sample, ]
data.test <- data[-train_sample, ]

# Параметр lambda
lambda_values <- 10^(-3+0.2*(0:25))

# Инициализация векторов для хранения ошибок
data.train.error <- c()
data.test.error <- c()

# Построение гребневой регрессии для разных значений lambda
for(lambda_val in lambda_values){
  ridge_model <- lm.ridge(Employed~., data.train, lambda = lambda_val)

  # Предсказание и вычисление среднеквадратичной ошибки на обучающей␣
 ↪выборке
```

```
  data.train.pred = ridge_model$ym + scale(data.train[, -6], center =␣
 ↪TRUE,scale = ridge_model$scales) %*% ridge_model$coef
  data.train.error <- c(data.train.error, mean(sqrt((data.train.pred -␣
 ↪data.train$Employed)^2)))

  # Предсказание и вычисление среднеквадратичной ошибки на тестовой␣
 ↪выборке
  data.test.pred = ridge_model$ym + scale(data.test[, -6], center =␣
 ↪TRUE,scale = ridge_model$scales) %*% ridge_model$coef
  data.test.error <- c(data.test.error, mean(sqrt((data.test.pred - data.
 ↪test$Employed)^2)))
}

# Отрисовка графика зависимости ошибки от lambda
plot(lambda_values, data.train.error)
lines(lambda_values, data.test.error)
```

График зависимости ошибок от lambda:



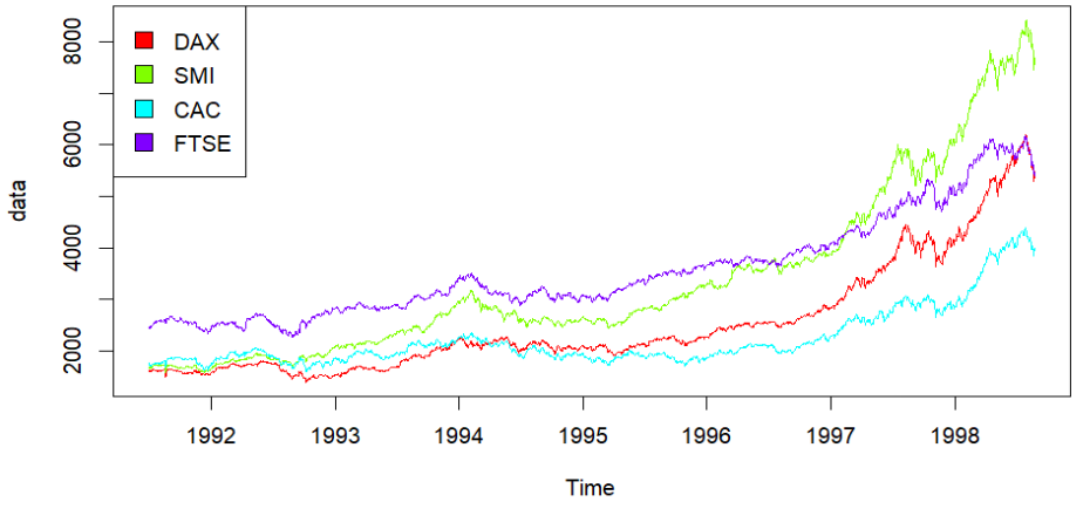
С увеличением lambda ошибка возрастает как для тестовой так и для обучающей выборки.

Задание №5

Загрузите данные EuStockMarkets из пакета « datasets». Данные содержат ежедневные котировки на момент закрытия фондовых бирж: Germany DAX (Ibis), Switzerland SMI, France CAC, и UK FTSE. Постройте на одном графике все кривые изменения котировок во времени. Постройте линейную регрессию для каждой модели в отдельности и для всех моделей вместе. Оцените, какая из бирж имеет наибольшую динамику.

```
# Код на R
# Загрузка данных
data <- EuStockMarkets

# График изменений котировок во времени
plot.ts(data, plot.type = "single", col = rainbow(4))
legend("topleft", colnames(data), fill=rainbow(4))

# Построение линейной регрессии для каждой модели в отдельности
reg_DAX <- lm(DAX ~ time(data), data)
summary(reg_DAX)$coefficients
reg_SMI <- lm(SMI ~ time(data), data)
summary(reg_SMI)$coefficients
reg_CAC <- lm(CAC ~ time(data), data)
summary(reg_CAC)$coefficients
reg_FTSE <- lm(FTSE ~ time(data), data)
summary(reg_FTSE)$coefficients

# Построение линейной регрессии для всех моделей вместе
reg_all <- lm(DAX+SMI+CAC+FTSE ~ time(data), data)
summary(reg_all)$coefficients
```

График изменений котировок во времени:



DAX $\sim$ time(data) $\rightarrow$ 449.6524

SMI $\sim$ time(data) $\rightarrow$ 717.5365

CAC $\sim$ time(data) $\rightarrow$ 204.5757

FTSE $\sim$ time(data) $\rightarrow$ 435.4562

DAX+SMI+CAC+FTSE $\sim$ time(data) $\rightarrow$ 1807.221

Наибольшую динамику имеет биржа SMI.

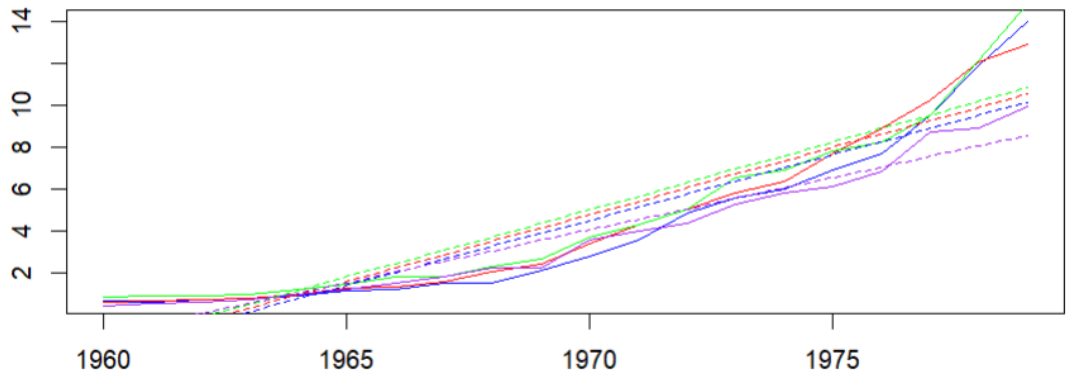
Задание №6

Загрузите данные JohnsonJohnson из пакета «datasets». Данные содержат поквартальную прибыль компании Johnson Johnson с 1960 по 1980 гг. Постройте на одном графике все кривые изменения прибыли во времени. Постройте линейную регрессию для каждого квартала в отдельности и для всех кварталов вместе. Оцените, в каком квартале компания имеет наибольшую и наименьшую динамику доходности. Сделайте прогноз по прибыли в 2016 году во всех кварталах и в среднем по году.

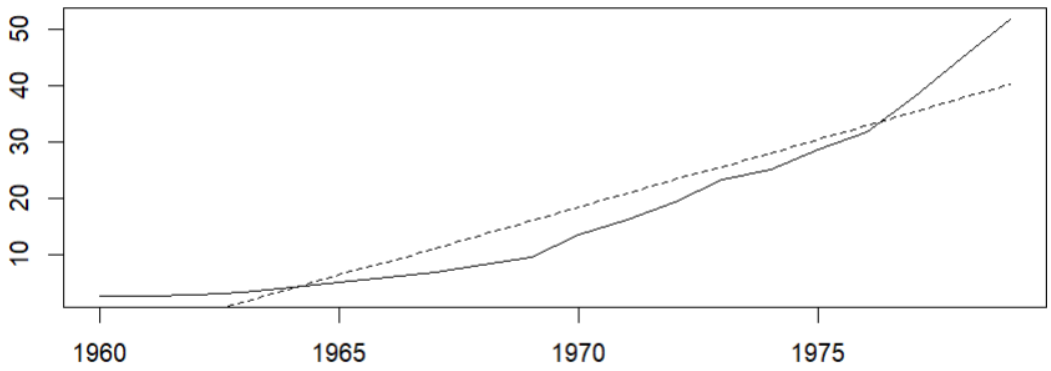
```
# Код на R
# Загрузка данных
data <- JohnsonJohnson

# Резделение данных по кварталлам
n <- 1:20
q1<-c()
q2<-c()
q3<-c()
q4<-c()
year <-c()

for (i in n){
  q1<-c(q1,data[i*4-3])
  q2<-c(q2,data[i*4-2])
  q3<-c(q3,data[i*4-1])
  q4<-c(q4,data[i*4])
  year <-c(year,1959+i)
}

data.total <- data.frame(year,q1,q2,q3,q4)

# Построение линейной регрессии для каждого квартала в отдельности, для␣
 ↪всех кварталов вместе и среднюю по кварталам
reg_q1 <- lm(q1 ~ year, data.total)
reg_q2 <- lm(q2 ~ year, data.total)
reg_q3 <- lm(q3 ~ year, data.total)
reg_q4 <- lm(q4 ~ year, data.total)
reg_all <- lm(q1+q2+q3+q4 ~ year, data.total)
reg_average <- lm((q1+q2+q3+q4)/4 ~ year, data.total)

# Отображение изменения прибыли во времени каждого квартала
plot(data.total$year, data.total$q1, type = "l", col = "blue")
lines(data.total$year, data.total$q2, type = "l", col = "red")
lines(data.total$year, data.total$q3, type = "l", col = "green")
lines(data.total$year, data.total$q4, type = "l", col = "purple")

# Отображение регрессий для каждого квартала
```

```
lines(data.total$year, predict.lm(reg_q1),lty = 2, col = "blue")
lines(data.total$year, predict.lm(reg_q2),lty = 2, col = "red")
lines(data.total$year, predict.lm(reg_q3),lty = 2, col = "green")
lines(data.total$year, predict.lm(reg_q4),lty = 2, col = "purple")

# Отображение изменения прибыли и регрессии для всех кварталов вместе
plot(data.total$year, data.total$q1+data.total$q2+data.total$q3+data.
 ↪total$q4, type = "l")
lines(data.total$year, predict.lm(reg_all),lty = 2)

# Оценка динамики доходности
reg_q1$coefficients
reg_q2$coefficients
reg_q3$coefficients
reg_q4$coefficients

# Прогноз прибыли в 2016 году
predict(reg_q1, list(year = 2016))
predict(reg_q2, list(year = 2016))
predict(reg_q3, list(year = 2016))
predict(reg_q4, list(year = 2016))
predict(reg_average, list(year = 2016))
```

Отображение изменения прибыли во времени каждого квартала:



Отображение изменения прибыли по годам во времени:



Наибольшая динамика прибыли зафиксирована в 3 квартале, а наименьшая в 4 квартале.

Прогноз по прибыли в 2016 году во всех кварталах и в среднем по году:

1 Квартал = 33.26287

2 Квартал = 34.25344

3 Квартал = 34.73457

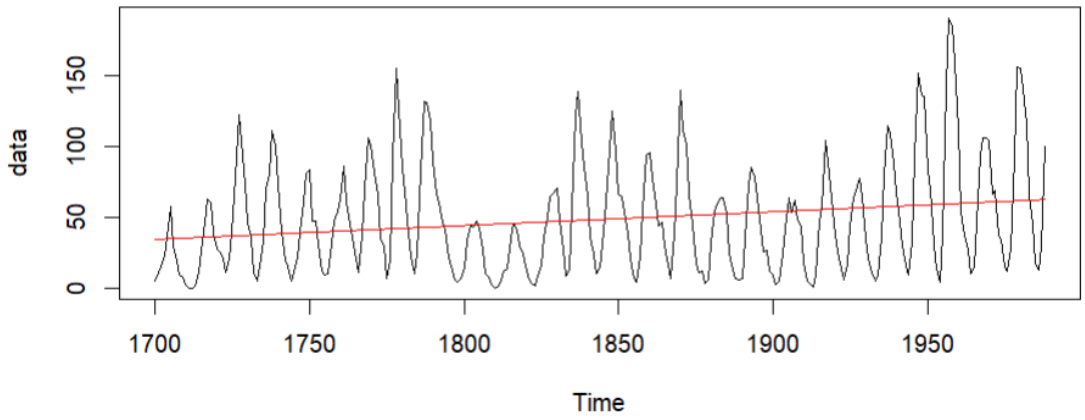
4 Квартал = 27.16839

В среднем за год = 32.35482

Задание №7

Загрузите данные sunspot.year из пакета «datasets». Данные содержат количество солнечных пятен с 1700 по 1988 гг. Постройте на графике кривую изменения числа солнечных пятен во времени. Постройте линейную регрессию для данных.

```r
# Код на R
# Загрузка данных
data <- sunspot.year

data.total <- data.frame(
  year = as.numeric(time(data)),
  sunspots = as.matrix(data)
)

# Построение регрессии
reg <- lm(sunspots ~ year, data.total)

# Отрисовка данных и регрессии
plot(data)
lines(data.total$year, predict(reg), col = "red")
```



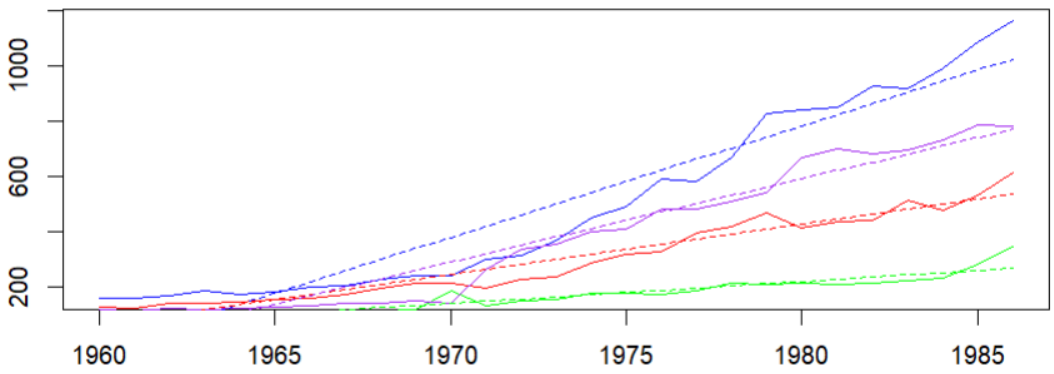
Задание №8

Загрузите данные из файла пакета «UKgas.scv». Данные содержат объемы ежеквартально потребляемого газа в Великобритании с 1960 по 1986 гг. Постройте линейную регрессию для каждого квартала в отдельности и для всех кварталов вместе. Оцените, в каком квартале потребление газа имеет наибольшую и наименьшую динамику доходности. Сделайте прогноз по потреблению газа в 2016 году во всех кварталах и в среднем по году.

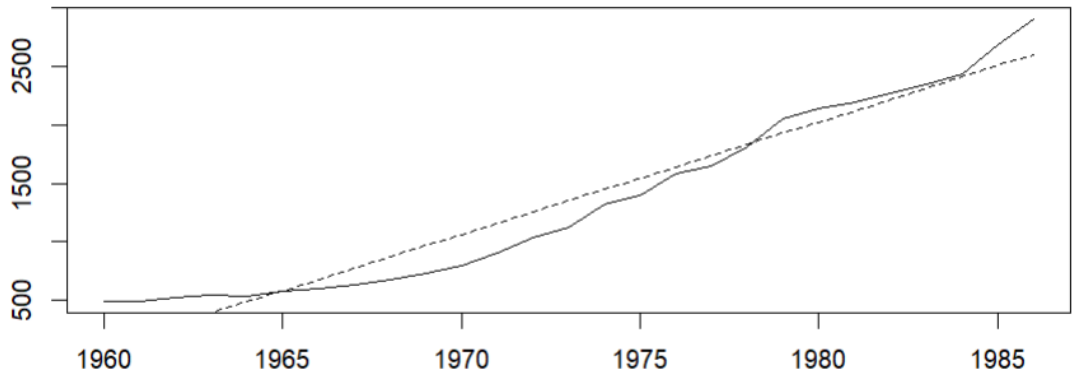
```
# Код на R
# Загрузка данных
data <- read.csv("D:/ML/Лабы по МО/Regression/UKgas.csv")

# Резделение данных по кварталлам
n <- 1:27
q1<-c()
q2<-c()
q3<-c()
q4<-c()
year <-c()

for (i in n){
  q1<-c(q1,data$UKgas[i*4-3])
  q2<-c(q2,data$UKgas[i*4-2])
  q3<-c(q3,data$UKgas[i*4-1])
  q4<-c(q4,data$UKgas[i*4])
  year <-c(year,1959+i)
}

data.total <- data.frame(year,q1,q2,q3,q4)

# Построение линейной регрессии для каждого квартала в отдельности, для␣
 ↪всех кварталов вместе и среднюю по кварталам
reg_q1 <- lm(q1 ~ year, data.total)
reg_q2 <- lm(q2 ~ year, data.total)
reg_q3 <- lm(q3 ~ year, data.total)
reg_q4 <- lm(q4 ~ year, data.total)
reg_all <- lm(q1+q2+q3+q4 ~ year, data.total)
reg_average <- lm((q1+q2+q3+q4)/4 ~ year, data.total)

# Отображение изменения прибыли во времени каждого квартала
plot(data.total$year, data.total$q1, type = "l", col = "blue")
lines(data.total$year, data.total$q2, type = "l", col = "red")
lines(data.total$year, data.total$q3, type = "l", col = "green")
lines(data.total$year, data.total$q4, type = "l", col = "purple")

# Отображение регрессий для каждого квартала
```

```
lines(data.total$year, predict.lm(reg_q1),lty = 2, col = "blue")
lines(data.total$year, predict.lm(reg_q2),lty = 2, col = "red")
lines(data.total$year, predict.lm(reg_q3),lty = 2, col = "green")
lines(data.total$year, predict.lm(reg_q4),lty = 2, col = "purple")

# Отображение изменения прибыли и регрессии для всех кварталов вместе
plot(data.total$year, data.total$q1+data.total$q2+data.total$q3+data.
 ↪total$q4, type = "l")
lines(data.total$year, predict.lm(reg_all),lty = 2)

# Оценка динамики доходности
reg_q1$coefficients
reg_q2$coefficients
reg_q3$coefficients
reg_q4$coefficients

# Прогноз прибыли в 2016 году
predict(reg_q1, list(year = 2016))
predict(reg_q2, list(year = 2016))
predict(reg_q3, list(year = 2016))
predict(reg_q4, list(year = 2016))
predict(reg_average, list(year = 2016))
```

Отображение изменения прибыли во времени каждого квартала:



Отображение изменения прибыли по годам во времени:



Наибольшая динамика прибыли зафиксирована в 1 квартале, а наименьшая в 3 квартале.

Прогноз по прибыли в 2016 году во всех кварталах и в среднем по году:

1 Квартал = 2230.936

2 Квартал = 1076.885

3 Квартал = 505.9368

4 Квартал = 1677.392

В среднем за год = 1372.787

Задание №9

Загрузите данные cars из пакета «datasets». Данные содержат зависимости тормозного пути автомобиля (футы) от его скорости (мили в час). Данные получены в 1920 г. Постройте регрессионную модель и оцените длину тормозного пути при скорости 40 миль в час.

```r
# Код на R
# Загрузка данных
data <- cars

# Построение регрессии
reg <- lm(dist ~ speed, data)

# Прогноз при скорости 40 милях
predict(reg, list(speed=40))
```

Прогнозируемый тормозной путь при скорости 40 миль/ч составляет 140 футов.