Линейная параметрическая регрессия

ЦМФ

Линейная параметрическая регрессия

$y_t = \beta_0 + \sum_{i=1}^{d} \beta_i x_{i,t} + \varepsilon_t,\ \ \varepsilon_t \sim N(0; \sigma^2),\ \ \mathrm{cor}\left(\varepsilon_{t_1}, \varepsilon_{t_2}\right) = 0$

$y = X\beta + \varepsilon$

y — зависимая (эндогенная переменная)

$x_1, \dots, x_d$ — независимые (экзогенные) переменные

$\beta_0, \dots, \beta_d$ — коэффициенты регрессии

ε — случайная ошибка

Параметрическая регрессия в R

исходные данные

```
library(datasets)
ozone <- airquality$Ozone
rad <- airquality$Solar.R

rem <- is.na(ozone) | is.na(rad)
ozone <- ozone[!rem]; rad <- rad[!rem]
```

разделим выборку на обучающую и экзаменующую

```
N <- length(ozone); E <- 20; T <- N-E
train.obs <- (1:T)
eval.obs <- (T+1):N

t.rad <- rad[train.obs]; t.ozone <- ozone[train.obs]
e.rad <- rad[eval.obs]; e.ozone <- ozone[eval.obs]
```

Параметрическая регрессия в R

регрессионная модель

```
fit.par <- lm(ozone ~ radiation,
data=data.frame(radiation=t.rad,ozone=t.ozone),
weights=NULL)
```

другой вариант

```
fit.par <- lm(t.ozone ~ t.rad)
```

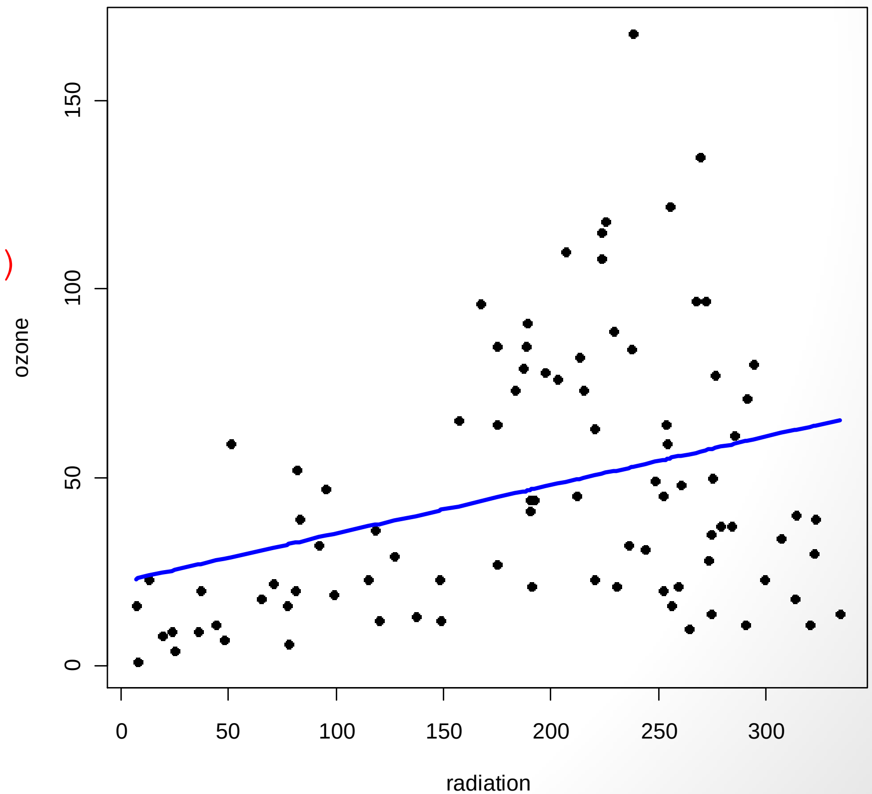
Анализ качества модели

```
summary(fit.par)
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 22.18688    8.02140   2.766  0.00690 **
radiation    0.12879    0.03816   3.375  0.00109 **
Multiple R-squared: 0.1135,     Adjusted R-squared: 0.1035
F-statistic: 11.39 on 1 and 89 DF,  p-value: 0.001094
```

```
fit.par$coefficients
fit.par$residuals
fit.par$fitted.values

plot(t.rad,t.ozone,pch=16,
xlab="radiation",ylab="ozone")

z <- order(t.rad)
lines(t.rad[z],
fit.par$fitted.values[z],
col="blue",lwd=3)
```



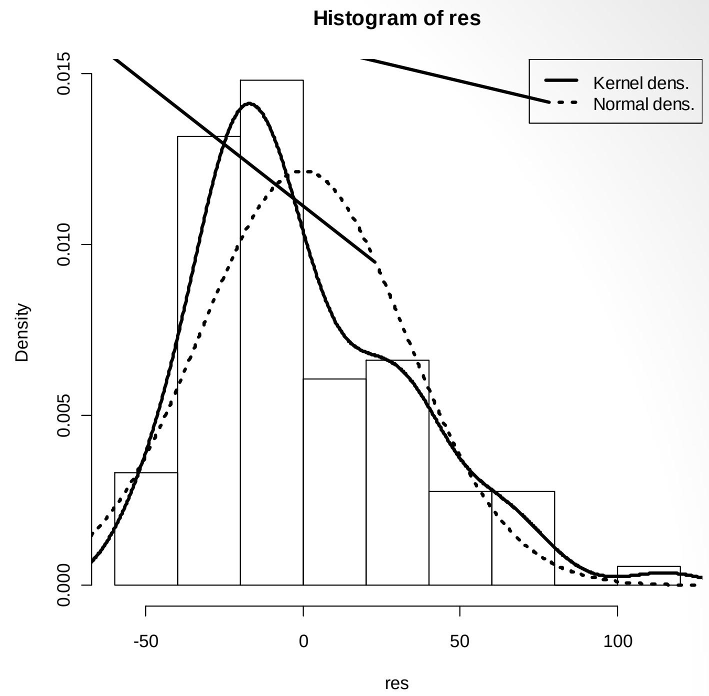
Анализ остатков модели

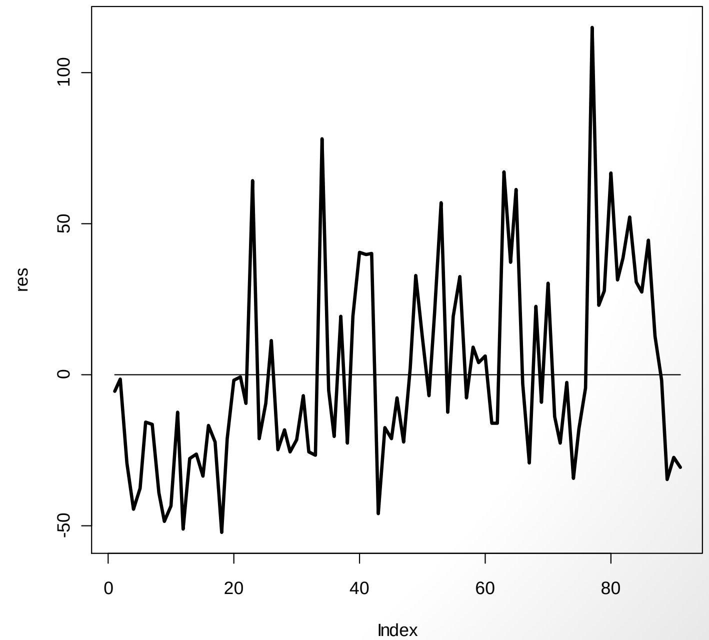
```
res <- fit.par$residuals

hist(res)
plot(res,type="l")
```

тесты на нормальность

```
library(fBasics)
shapiro.test(res)
jarqueberaTest(res)
```





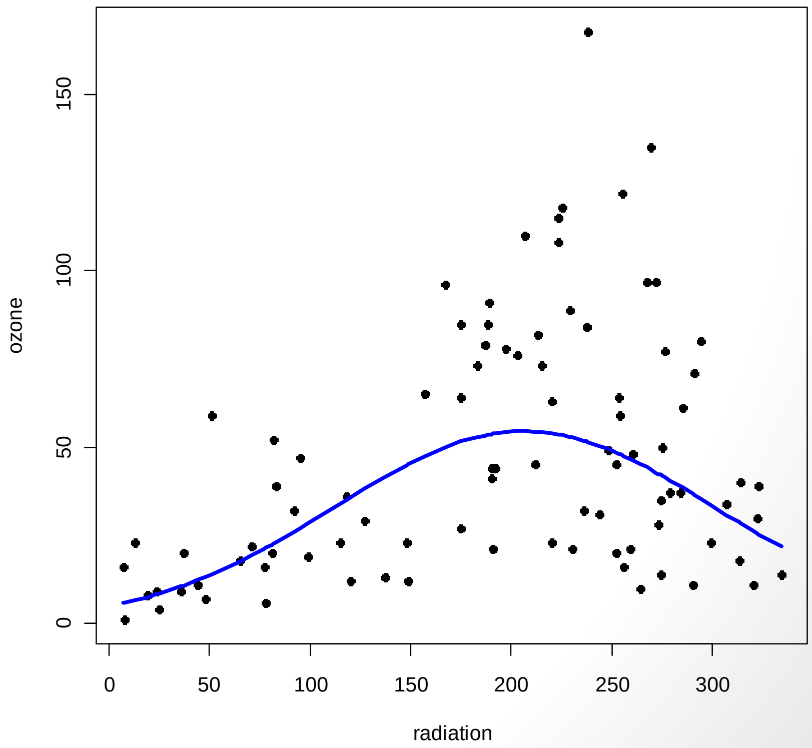
Тест	p.value
Shapiro – Wilks	0.000
Jarque – Bera	0.001

Переформулировка модели

```
fit.par <- lm(log(ozone) ~ rad + rad2,
data=data.frame(rad=t.rad,rad2=t.rad^2,ozone=t.ozone))
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.585e+00  2.514e-01   6.304 1.12e-08 ***
rad          2.337e-02  3.313e-03   7.055 3.77e-10 ***
rad2        -5.660e-05  9.564e-06  -5.919 6.11e-08 ***
Multiple R-squared: 0.4234,     Adjusted R-squared: 0.4103
```

```
plot(t.rad,t.ozone,pch=16,
xlab="radiation",ylab="ozone")

z <- order(t.rad)
lines(t.rad[z],
exp(fit.par$fitted.values[z]),
col="blue",lwd=3)
```



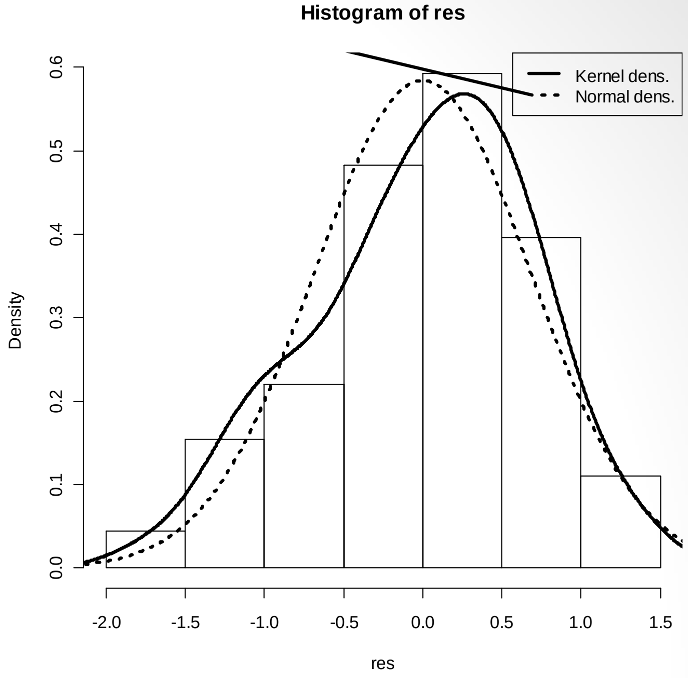
Анализ остатков модели

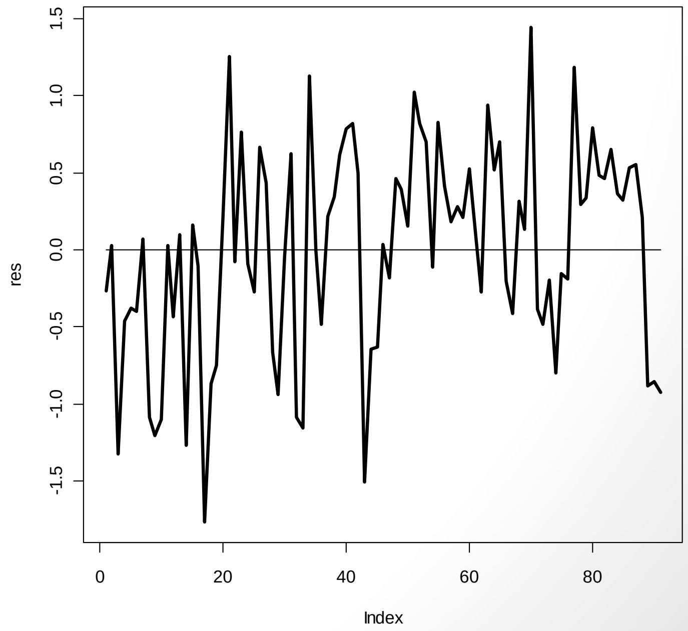
```
res <- fit.par$residuals

hist(res)
plot(res,type="l")
```

тесты на нормальность

```
shapiro.test(res)
jarqueberaTest(res)
```





Тест	p.value
Shapiro – Wilks	0.288
Jarque – Bera	0.279

Тесты на гетероскедастичность

$$y_t = \beta_0 + \sum_{i=1}^{d} \beta_i x_{i,t} + \varepsilon_t,\ \ \varepsilon_t \sim N(0;\sigma^2)$$

Тест Бреуша–Пагана

Пусть $e_t = y_t - \hat{y}_t = y_t - \hat{\beta}_0 - \sum_{i=1}^{d} \hat{\beta}_i x_{i,t}$ — остатки модели

Если $E\varepsilon_t = 0$, тогда ${\hat{\sigma}_t}^2 = e_t^2$

Проверим линейную зависимость нормированных квадратов остатков от независимых переменных:

$$\frac{\mathrm{T}}{\sum_{t=1}^{T} e_t^2} \cdot e_t^2 = \epsilon_t = \gamma_0 + \sum_{i=1}^{d} \gamma_i z_{i,t} + \eta_t$$

$H_0{:}\, \gamma_1 = \cdots = \gamma_d = 0$

$H_{alt}{:}\, \exists i \in \{1; \dots; d\} : \gamma_i \neq 0$

Статистика: $BP = \frac{1}{2}\sum_{t=1}^{T}(\hat{\epsilon}_t - \bar{\epsilon})^2 \ \sim^{H_0} \chi^2(d)$

Тесты на гетероскедастичность

$y_t = \beta_0 + \sum_{i=1}^{d} \beta_i x_{i,t} + \varepsilon_t,\ \ \varepsilon_t \sim N(0;\sigma^2)$

Тест Голфелда–Квандта

Опускаем f наблюдений, соответствующих средним величинам независимой переменной и рассматриваем две независимые регрессии

Пусть $e_{1,t}$ — остатки по меньшей модели, $e_{1,t}$ — остатки по большей модели

$H_0: \sigma^2 = const$
$H_{alt}: \sigma^2 = \sigma^2(t)$

Статистика: $GQ = \frac{\sum_{i=1}^{\frac{T}{2}-\frac{f}{2}} e_{1,i}^2}{\sum_{j=1}^{\frac{T}{2}-\frac{f}{2}} e_{2,j}^2} \sim^{H_0} F\left(\frac{T}{2}-\frac{f}{2}-d-1, \frac{T}{2}-\frac{f}{2}-d-1\right)$

Тесты на гетероскедастичность в R

```
library(lmtest)
```

тест Бреуша–Пагана

```
bptest(fit.par,varformula=NULL,data=NULL,studentize=FALSE)
Breusch-Pagan test
data:  fit.par
BP = 3.2609, df = 2, p-value = 0.1958
```

тест Голдфелда–Куандта

```
gqtest(fit.par,fraction=25,alternative="two.sided")
Goldfeld-Quandt test
data:  fit.par
GQ = 0.6948, df1 = 30, df2 = 30, p-value = 0.324
```

Учёт гетероскедастичности

Пусть имеется регрессия с гетероскедастичностью

```
oz <- lm(t.ozone ~ t.rad)
bptest(oz,varformula=NULL,data=NULL,studentize=FALSE)
BP = 9.3894, df = 1, p-value = 0.002182
```

Рассчитаем оценки стандартных отклонений $\hat{\sigma}$

```
e.sq <- oz$residuals^2
sigma.hat <- lm(e.sq ~ t.rad)$fitted.values ^ 0.5
```

Используем взвешенный МНК

```
oz.wgt <- lm(t.ozone ~ t.rad, weights = 1/sigma.hat)
```

```
oz          Estimate Std. Error t value Pr(>|t|)
(Intercept) 22.18688    8.02140   2.766  0.00690 **
t.rad        0.12879    0.03816   3.375  0.00109 **

oz.wgt      Estimate Std. Error t value Pr(>|t|)
(Intercept) 14.95107    5.69516   2.625   0.0102 *
t.rad        0.16457    0.03111   5.290 8.68e-07 ***
```

Тест на автокорреляцию

Тест Дарбина–Ватсона

$H_0: cor(e_t, e_{t-1}) = 0$

$H_{alt}: cor(e_t, e_{t-1}) \neq 0$

Статистика: $DW = \frac{\sum_{t=2}^{T}(e_t - e_{t-1})^2}{\sum_{t=1}^{T} e_t^2} \approx 2(1-\rho)$

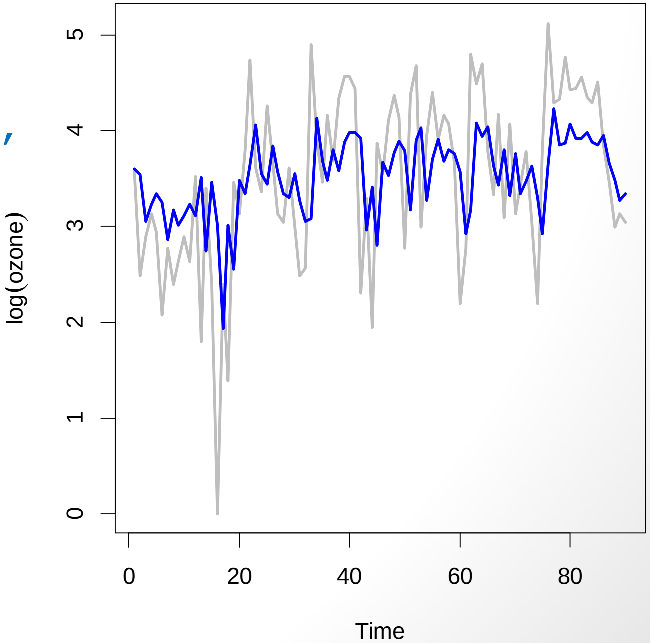
тест Дарбина–Ватсона в R

```
dwtest(fit.par,alternative="two.sided")
Durbin-Watson test
data:  fit.par
DW = 1.3138, p-value = 0.0006297
alternative hypothesis: true autocorrelation is not 0
```

Выделение авторегрессионной составляющей

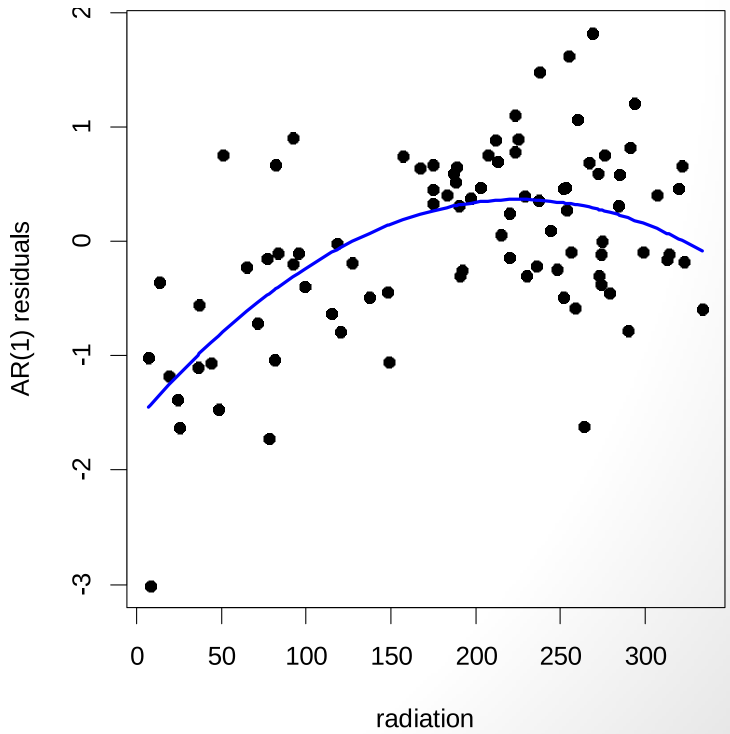
```
library(tseries)
adf.test(log(t.ozone))
Dickey-Fuller = -3.8002, Lag order = 4, p-value = 0.02235
alternative hypothesis: stationary
ar1 <- arma(log(t.ozone), order = c(1,0))
             Estimate  Std. Error  t value Pr(>|t|)
ar1            0.4475      0.0939    4.766 1.88e-06 ***
intercept     1.9386      0.3413    5.680 1.35e-08 ***

adf.test(ar1$residuals[2:T])
Dickey-Fuller = -3.922, Lag order = 4,
p-value = 0.01669
alternative hypothesis: stationary
```



Модель для остатков авторегрессии

```
fit.par <- lm(ozone ~ rad + rad2,
data = data.frame(ozone = ar1$residuals[2:T],
rad = t.rad[2:T], rad2 = t.rad[2:T]^2))
                 Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.564e+00  2.368e-01  -6.605 3.03e-09 ***
rad          1.713e-02  3.135e-03   5.465 4.36e-07 ***
rad2        -3.805e-05  9.052e-06  -4.204 6.35e-05 ***
```



Тест	p.value
Shapiro–Wilks	0.409
Jarque–Bera	0.594
Breusch–Pagan	0.338
Goldfeld–Quandt	0.354
Durbin–Watson	0.525

Построение прогноза

$y = X\beta + \varepsilon$

$\hat{\beta} = (X'X)^{-1}X'y$ — оценки коэффициентов регрессии

$\widehat{\sigma^2} = s^2 = \frac{e'e}{T-d}$ — оценка дисперсии ошибок

$\hat{y}_{T+1} = x'_{T+1}\hat{\beta}$ — прогноз

$\delta = s\sqrt{1 + x'_{T+1}(X'X)^{-1}x_{T+1}}$ — среднекваратичная ошибка прогноза

Пусть $(1-\alpha)$ — уровень значимости, тогда

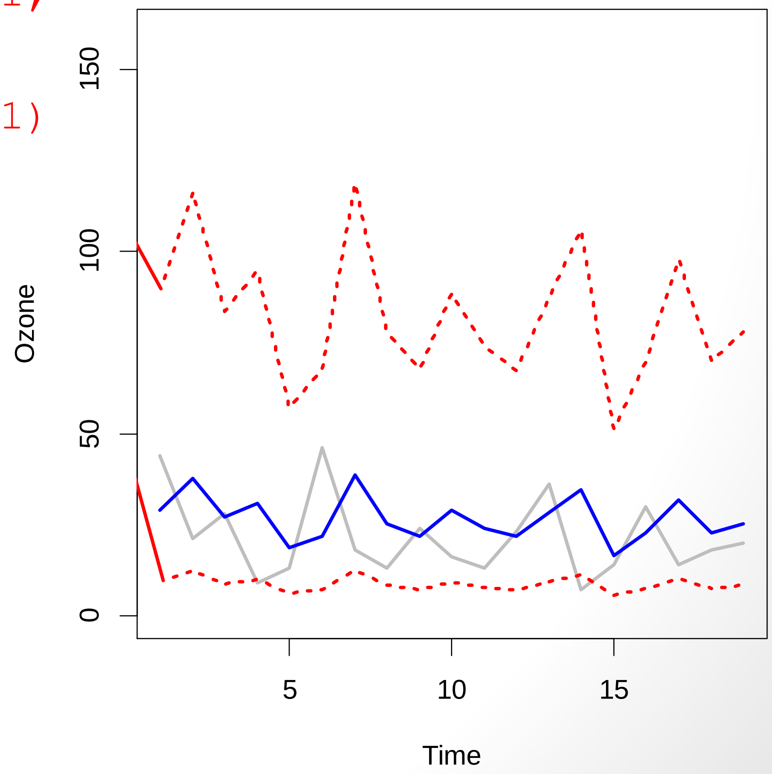
$\hat{y} \pm \delta \cdot t_{T-d}^{-1}\left(\frac{\alpha}{2}\right)$ — доверительный интервал для прогноза

$t_{T-d}^{-1}(.)$ — квантиль t-распределения с $(T-d)$ степенями свободы

Прогноз по авторегрессионной модели

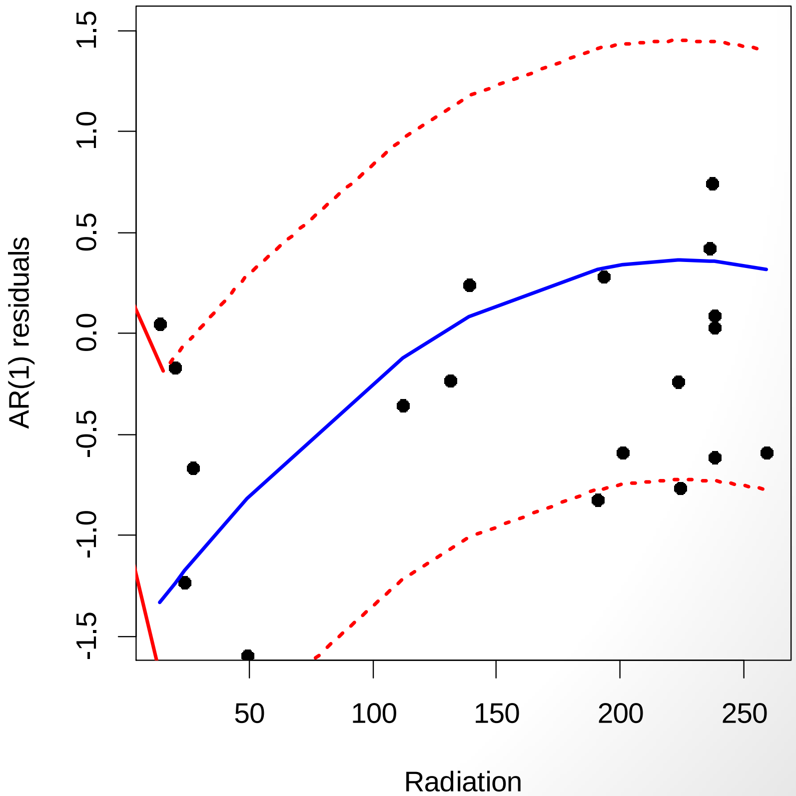
```
predict.arma <- function(model, data, newdata, alpha = 0.05) {
# your code here
list(fit = fit, lower = fit + delta*qt(alpha/2,df=df),
                upper = fit - delta*qt(alpha/2,df=df))
}

ar1.frc <- predict.arma(model = ar1,
data = log(t.ozone)[1:(T-1)],
newdata = log(e.ozone), alpha = 0.1)
```



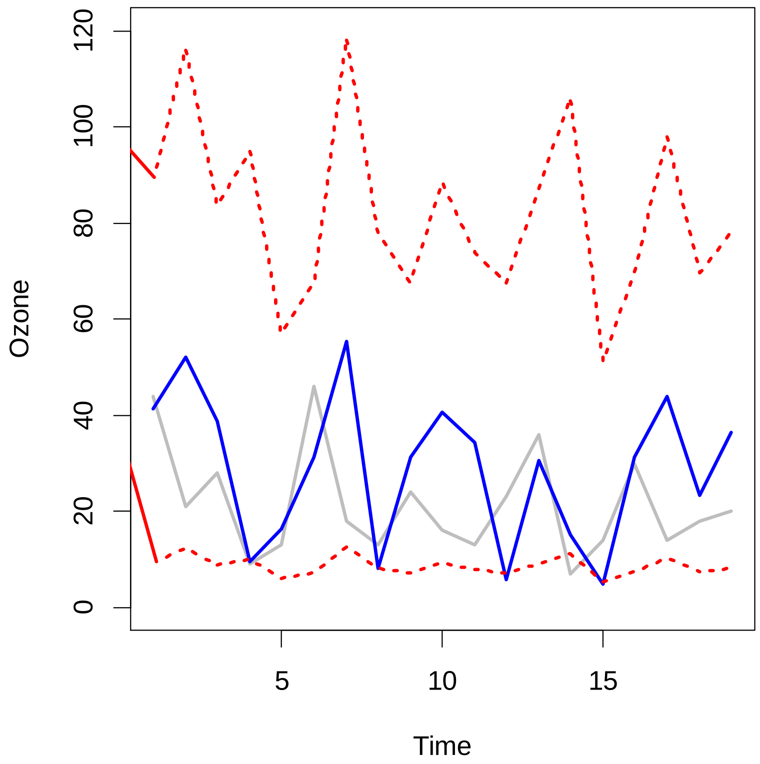
Прогноз остатков

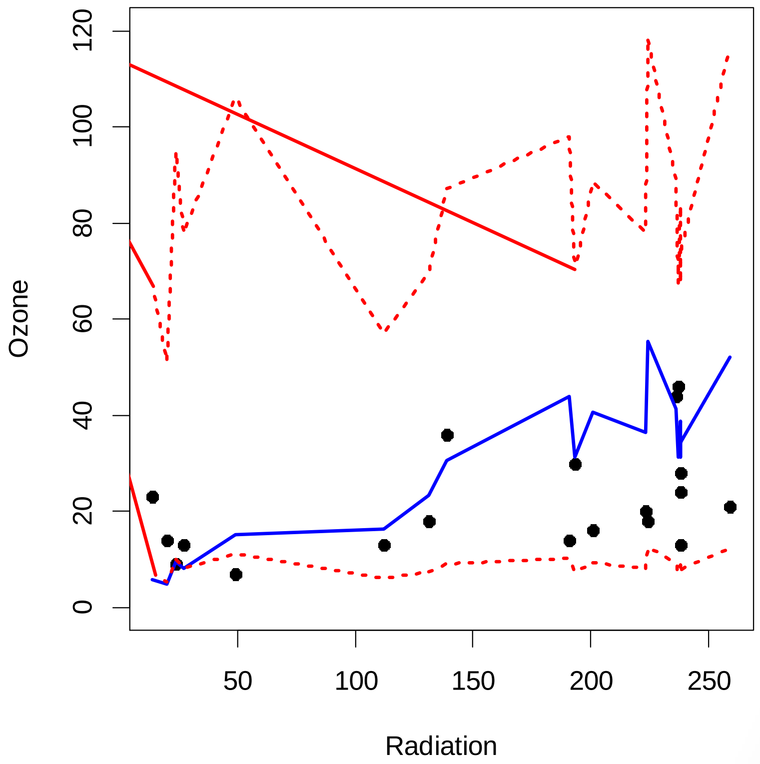
```
par.frc <- predict(fit.par,
  newdata=data.frame(rad=e.rad[2:E],rad2=e.rad[2:E]^2),
  se.fit=TRUE,interval="prediction",level=0.90)
```



Итоговый прогноз

```
frc <- exp(ar1.frc$fit[2:E] + par.frc$fit[,"fit"])
```





Домашнее задание

Создать файл «regression_func.r» и записать в него пользовательскую функцию `predict.arma`