RUDN university

5100

Программирование в среде R

Шевцов Василий Викторович,
директор ДИТ РУДН, shevtsov_vv@rudn.university

Корреляция и линейная регрессия

Определения

Корреляция (от лат. correlatio «соотношение, взаимосвязь»), или корреляционная зависимость, — статистическая взаимосвязь двух или более случайных величин.

Математической мерой корреляции двух случайных величин служит корреляционное отношение либо коэффициент корреляции R или r

cor.test()

```
> df <- mtcars
> cor.test(x = df$mpg, y = df$hp)

        Pearson's product-moment correlation

data:  df$mpg and df$hp
t = -6.7424, df = 30, p-value = 1.788e-07
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 -0.8852686 -0.5860994
sample estimates:
       cor
-0.7761684
```

p-value<<0.05 Отклоняем нулевую гипотезу об отсутствии взаимосвязи

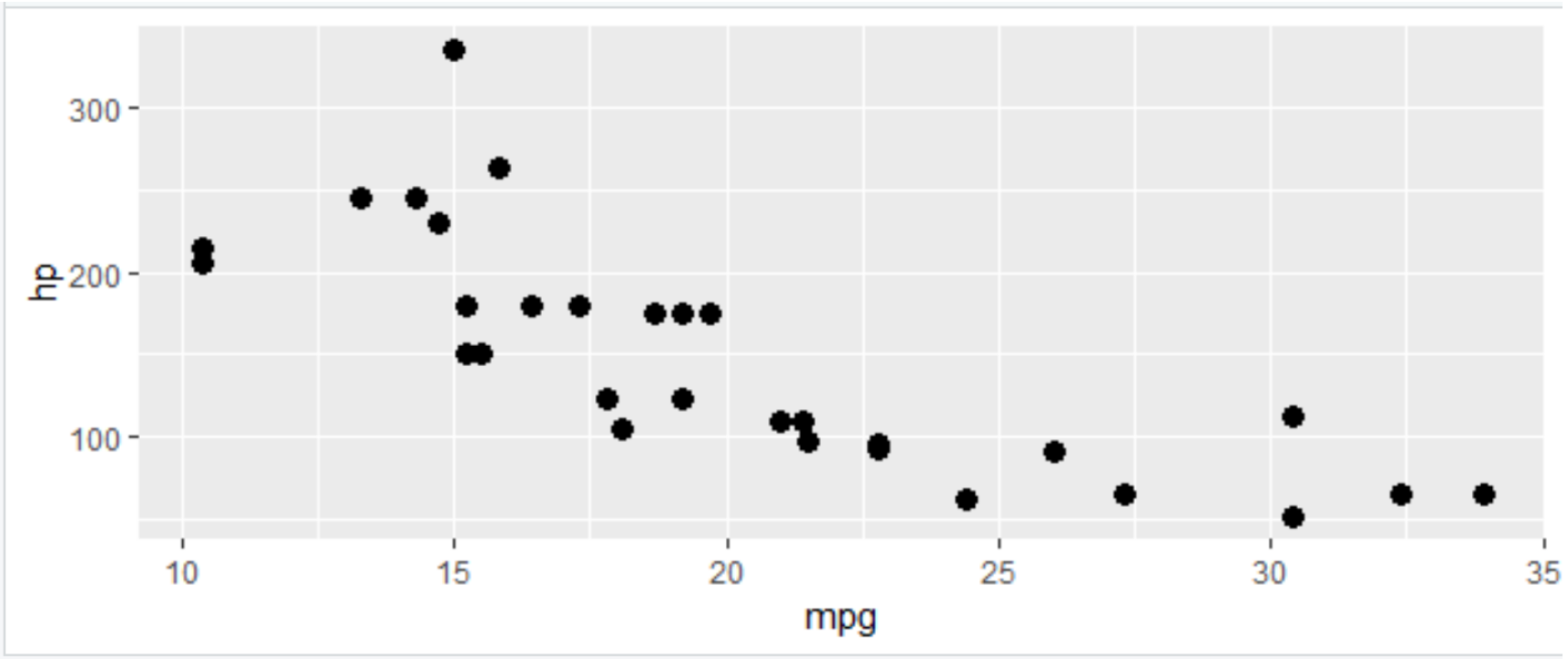
method
a character string indicating which correlation coefficient is to be used for the test. One of "pearson", "kendall", or "spearman", can be abbreviated.

cor.test()

```
> f1 <- cor.test(x = df$mpg, y = df$hp)
> str(f1)
List of 9
 $ statistic  : Named num -6.74
  ..- attr(*, "names")= chr "t"
 $ parameter  : Named int 30
  ..- attr(*, "names")= chr "df"
 $ p.value    : num 1.79e-07
 $ estimate   : Named num -0.776
  ..- attr(*, "names")= chr "cor"
 $ null.value : Named num 0
  ..- attr(*, "names")= chr "correlation"
 $ alternative: chr "two.sided"
 $ method     : chr "Pearson's product-moment correlation"
 $ data.name  : chr "df$mpg and df$hp"
 $ conf.int   : num [1:2] -0.885 -0.586
  ..- attr(*, "conf.level")= num 0.95
 - attr(*, "class")= chr "htest"
```

Построение графика

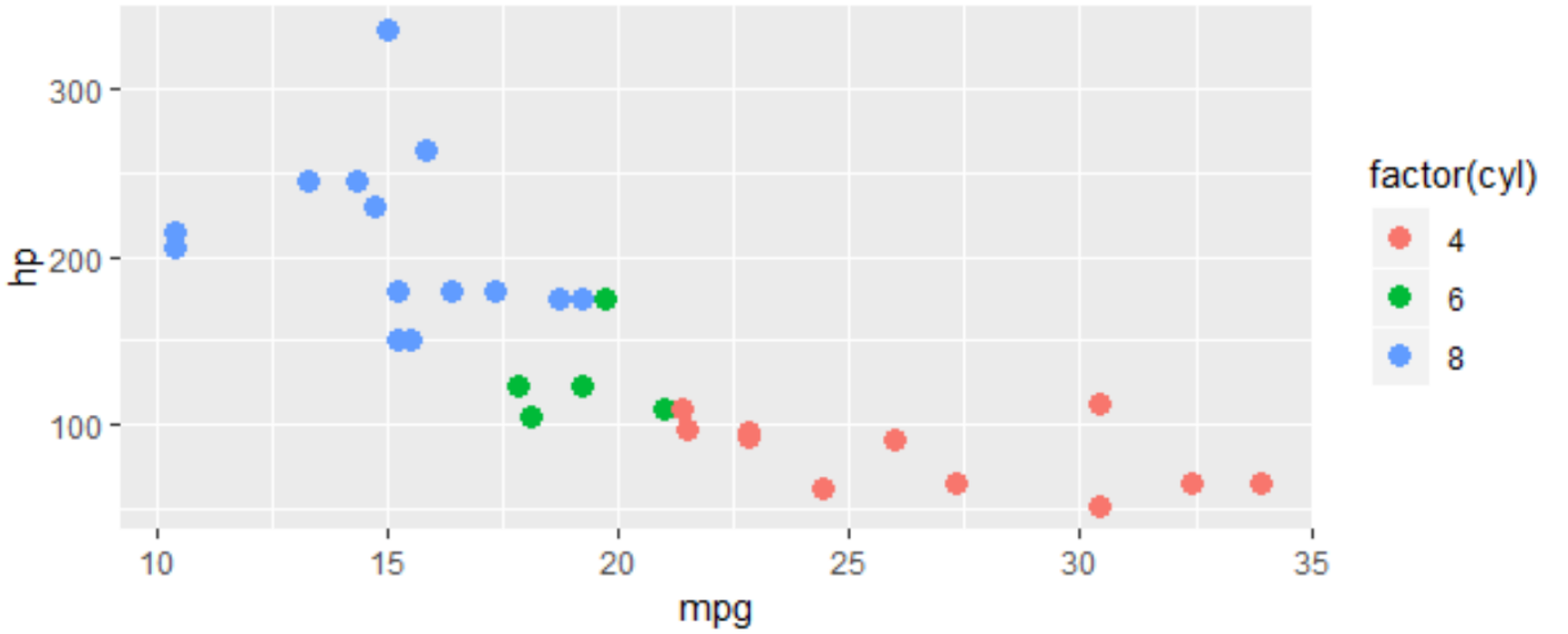
```
> ggplot(df, aes(x=mpg,y=hp))+
+   geom_point(size=3)
```



На графике взаимосвязь выражена слабо

Группировка

```
> ggplot(df, aes(x=mpg,y=hp,color=factor(cyl)))+
+   geom_point(size=3)
```

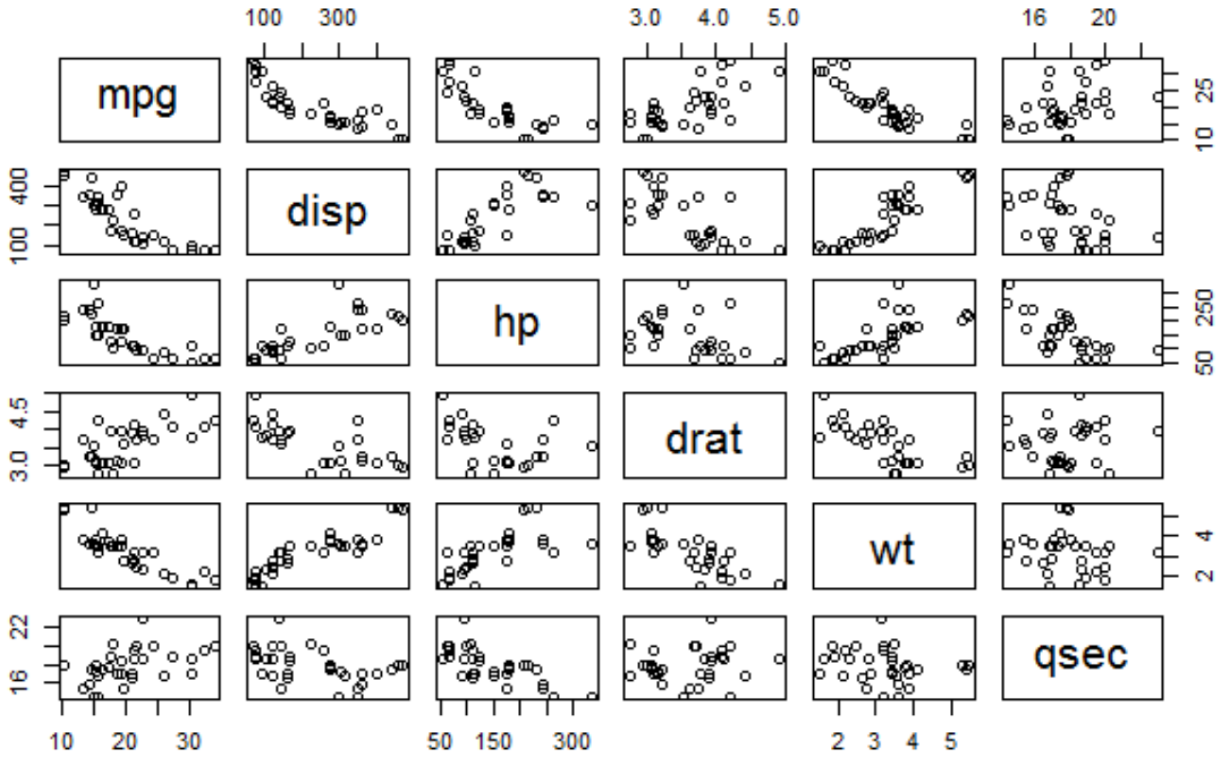


. pairs()

```
> df2 <- df[,c(1,3:7)]
```

	mpg	disp	hp	drat	wt	qsec
Mazda RX4	21.0	160.0	110	3.90	2.620	16.46
Mazda RX4 Wag	21.0	160.0	110	3.90	2.875	17.02
Datsun 710	22.8	108.0	93	3.85	2.320	18.61
Hornet 4 Drive	21.4	258.0	110	3.08	3.215	19.44

```
> pairs(df2)
```



RUDN university

Попарный анализ

```
> cor(df2)
             mpg       disp         hp        drat         wt        qsec
mpg   1.0000000 -0.8475514 -0.7761684  0.68117191 -0.8676594  0.41868403
disp -0.8475514  1.0000000  0.7909486 -0.71021393  0.8879799 -0.43369788
hp   -0.7761684  0.7909486  1.0000000 -0.44875912  0.6587479 -0.70822339
drat  0.6811719 -0.7102139 -0.4487591  1.00000000 -0.7124406  0.09120476
wt   -0.8676594  0.8879799  0.6587479 -0.71244065  1.0000000 -0.17471588
qsec  0.4186840 -0.4336979 -0.7082234  0.09120476 -0.1747159  1.00000000
```

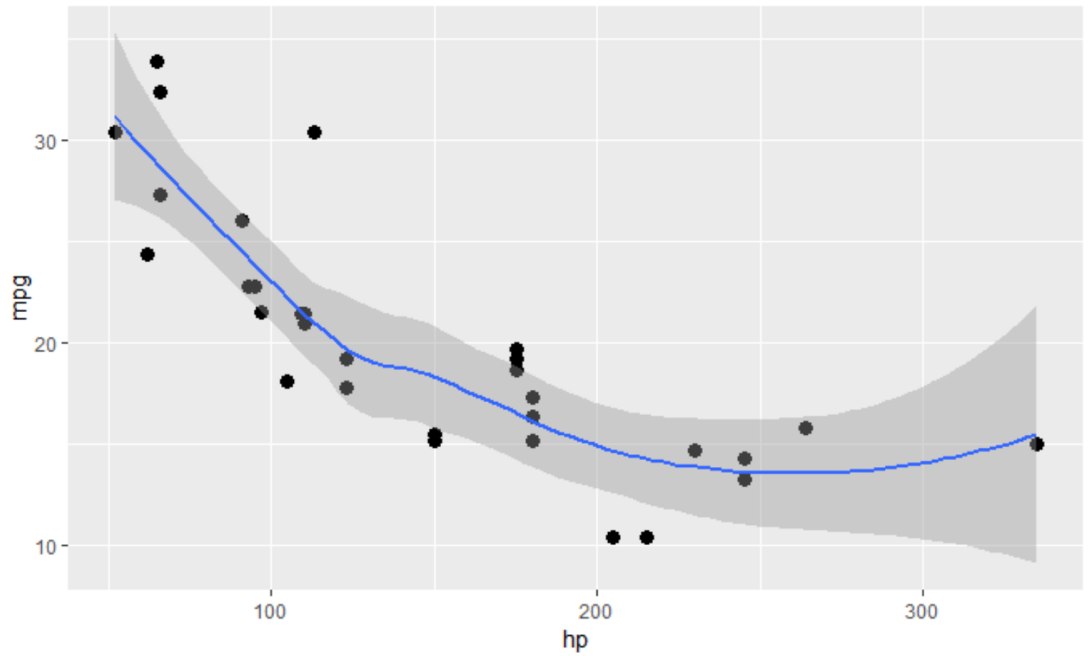
```
> library(psych)
> f3 <- corr.test(df2)
> f3$r
             mpg       disp         hp        drat         wt        qsec
mpg   1.0000000 -0.8475514 -0.7761684  0.68117191 -0.8676594  0.41868403
disp -0.8475514  1.0000000  0.7909486 -0.71021393  0.8879799 -0.43369788
hp   -0.7761684  0.7909486  1.0000000 -0.44875912  0.6587479 -0.70822339
drat  0.6811719 -0.7102139 -0.4487591  1.00000000 -0.7124406  0.09120476
wt   -0.8676594  0.8879799  0.6587479 -0.71244065  1.0000000 -0.17471588
qsec  0.4186840 -0.4336979 -0.7082234  0.09120476 -0.1747159  1.00000000
> f3$p
              mpg          disp            hp          drat            wt          qsec
mpg  0.000000e+00 1.219442e-08 1.966619e-06 1.243368e-04 1.811542e-09 0.0525761455
disp 9.380327e-10 0.000000e+00 8.571214e-07 4.784260e-05 1.833479e-10 0.0525761455
hp   1.787835e-07 7.142679e-08 0.000000e+00 4.994386e-02 2.487496e-04 0.0000478426
drat 1.776240e-05 5.282022e-06 9.988772e-03 0.000000e+00 4.784260e-05 0.6777365683
wt   1.293959e-10 1.222320e-11 4.145827e-05 4.784260e-06 0.000000e+00 0.6777365683
qsec 1.708199e-02 1.314404e-02 5.766253e-06 6.195826e-01 3.388683e-01 0.0000000000
```

lm()

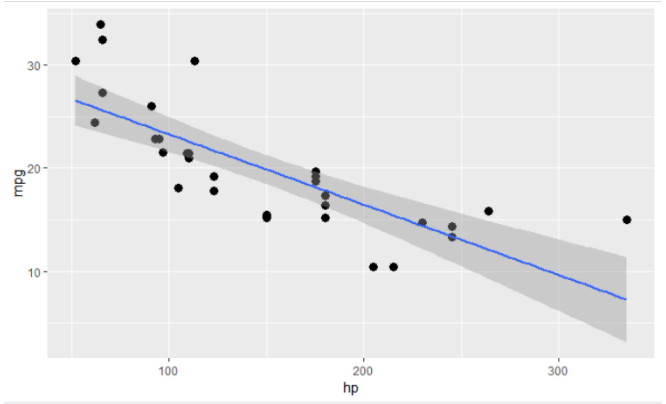
```
> f4 <- lm(mpg ~ hp, df2)
> summary(f4)

Call:
lm(formula = mpg ~ hp, data = df2)

Residuals:
    Min      1Q  Median      3Q     Max
-5.7121 -2.1122 -0.8854  1.5819  8.2360

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 30.09886    1.63392  18.421  < 2e-16 ***
hp          -0.06823    0.01012  -6.742 1.79e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.863 on 30 degrees of freedom
Multiple R-squared:  0.6024,    Adjusted R-squared:  0.5892
F-statistic: 45.46 on 1 and 30 DF,  p-value: 1.788e-07
```

Линия тренда

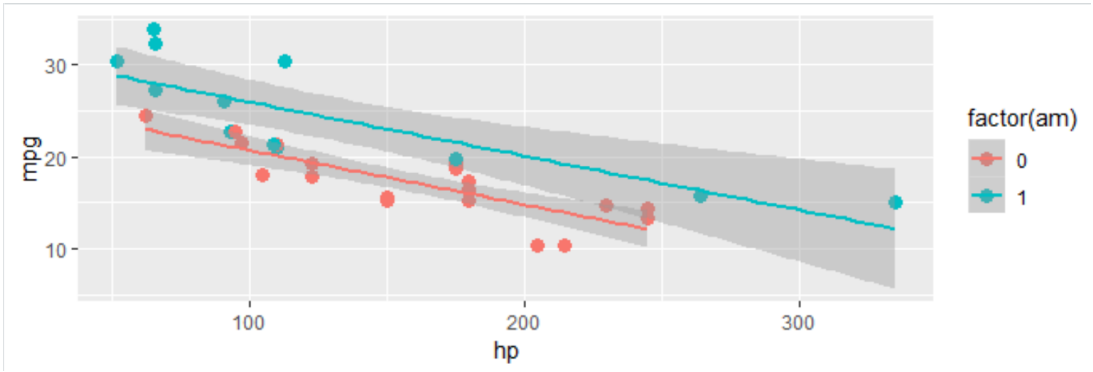
```
> ggplot(df, aes(x=hp,y=mpg))+
+    geom_point(size=3)+
+    geom_smooth()
```



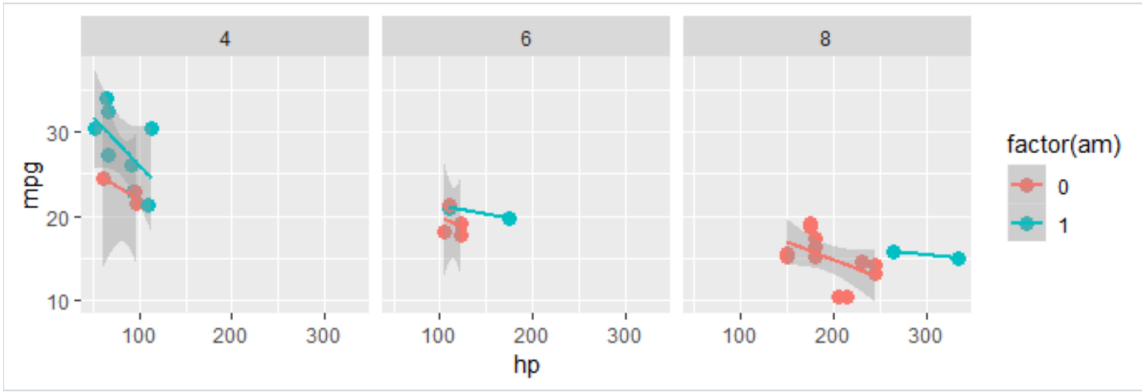

```
> ggplot(df, aes(x=hp,y=mpg))+
+    geom_point(size=3)+
+    geom_smooth(method = lm)
```



```
> ggplot(df, aes(x=hp,y=mpg,color=factor(am)))+
+   geom_point(size=3)+
+   geom_smooth(method = lm)
```




```
> ggplot(df, aes(x=hp,y=mpg,color=factor(am)))+
+   geom_point(size=3)+
+   geom_smooth(method = lm)+
+   facet_grid(.~cyl)
```



Предсказание значений

```
> df_lm <- lm(mpg ~ hp, df2)
> df_result <- data.frame(mpg=df2$mpg,fitted <- df_lm$fitted.values)
```

	mpg	fitted....df_lm.fitted.values
Mazda RX4	21.0	22.593750
Mazda RX4 Wag	21.0	22.593750
Datsun 710	22.8	23.753631
Hornet 4 Drive	21.4	22.593750
Hornet Sportabout	18.7	18.158912
Valiant	18.1	22.934891
Duster 360	14.3	13.382932
Merc 240D	24.4	25.868707
Merc 230	22.8	23.617174
Merc 280	19.2	21.706782
Merc 280C	17.8	21.706782
Merc 450SE	16.4	17.817770
Merc 450SL	17.3	17.817770
Merc 450SLC	15.2	17.817770
Cadillac Fleetwood	10.4	16.112064

Showing 1 to 17 of 32 entries, 2 total columns

```
> df_new <- data.frame(hp=c(100,150,200,300))
> predict(df_lm,df_new)
       1        2        3        4
23.276033 19.864619 16.453205  9.630377
```

	hp
1	100
2	150
3	200
4	300

```
> df_new$mpg_new <- predict(df_lm,df_new)
```

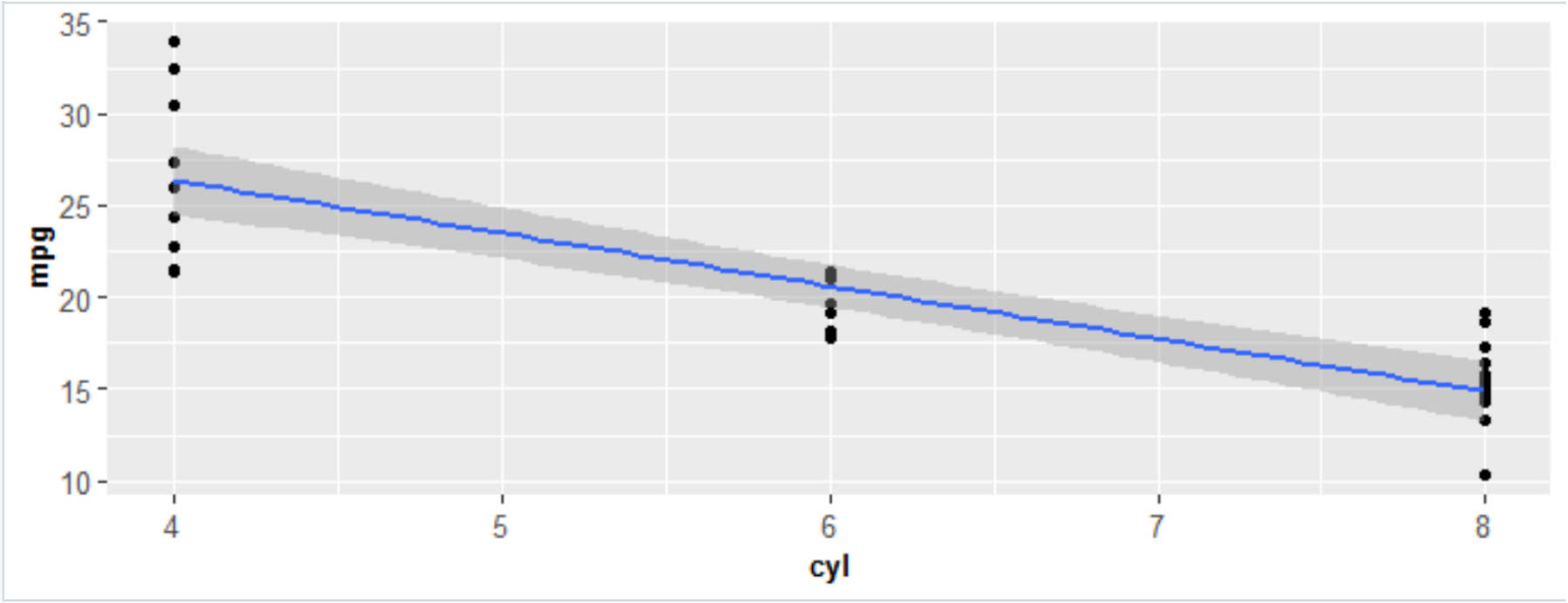
	hp	mpg_new
1	100	23.276033
2	150	19.864619
3	200	16.453205
4	300	9.630377

Линейная модель + номинативная переменная

```
> f_lm <- lm(mpg ~ cyl, df)
> summary(f_lm)

Call:
lm(formula = mpg ~ cyl, data = df)

Residuals:
    Min      1Q  Median      3Q     Max
-4.9814 -2.1185  0.2217  1.0717  7.5186

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  37.8846     2.0738   18.27  < 2e-16 ***
cyl          -2.8758     0.3224   -8.92 6.11e-10 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.206 on 30 degrees of freedom
Multiple R-squared:  0.7262,    Adjusted R-squared:  0.7171
F-statistic: 79.56 on 1 and 30 DF,  p-value: 6.113e-10
```

Линейная модель + номинативная переменная

```
> ggplot(df, aes(cyl,mpg))+
+   geom_point()+
+   geom_smooth(method = "lm")+
+   theme(axis.text=element_text(size=10),
+         axis.title = element_text(size = 10,face="bold"))
```



Линейная модель + номинативная переменная

```
> df$cyl <- factor(df$cyl,labels=c("4cyl","6cyl","8cyl"))
> f_lm <- lm(mpg ~ cyl, df)
> summary(f_lm)

Call:
lm(formula = mpg ~ cyl, data = df)

Residuals:
    Min      1Q  Median      3Q     Max
-5.2636 -1.8357  0.0286  1.3893  7.2364

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  26.6636     0.9718  27.437  < 2e-16 ***
cyl6cyl      -6.9208     1.5583  -4.441 0.000119 ***
cyl8cyl     -11.5636     1.2986  -8.905 8.57e-10 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.223 on 29 degrees of freedom
Multiple R-squared:  0.7325,    Adjusted R-squared:  0.714
F-statistic:  39.7 on 2 and 29 DF,  p-value: 4.979e-09
```

Значение intercept – в данном случае среднеарифметическое в группе базового уровня, далее идут значения уменьшения

```
> aggregate(mpg ~ cyl, df, mean)
   cyl      mpg
1 4cyl 26.66364
2 6cyl 19.74286
3 8cyl 15.10000
```

Множественная регрессия

swiss

```
> df <- swiss
> View(df)
```

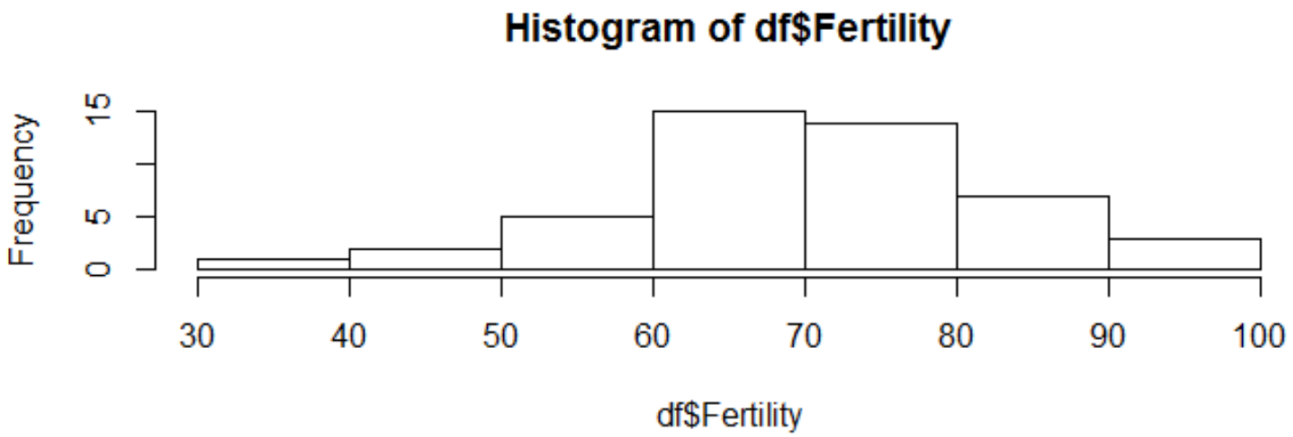
	Fertility	Agriculture	Examination	Education	Catholic	Infant.Mortality
Courtelary	80.2	17.0	15	12	9.96	22.2
Delemont	83.1	45.1	6	9	84.84	22.2
Franches-Mnt	92.5	39.7	5	5	93.40	20.2
Moutier	85.8	36.5	12	7	33.77	20.3
Neuveville	76.9	43.5	17	15	5.16	20.6
Porrentruy	76.1	35.3	9	7	90.57	26.6
Broye	83.8	70.2	16	7	92.85	23.6
Glane	92.4	67.8	14	8	97.16	24.9

- Данные 1888-го года по регионам,
- Fertility — это количество детей до пяти лет, делённое на количество женщин до 50-ти лет и отмасштабированное, на 1000 домноженное;
- Agriculture — это процент мужчин, занятых в сельском хозяйстве;
- Examination — процент тех, кто получил высокий результат оценки на призывном пункте
- Catholic — процент католиков (?)
- Infant.Mortality — Смертность младенцев

swiss

```
> str(df)
'data.frame':   47 obs. of  6 variables:
 $ Fertility       : num  80.2 83.1 92.5 85.8 76.9 76.1 83.8 92.4 82.4 82.9 ...
 $ Agriculture     : num  17 45.1 39.7 36.5 43.5 35.3 70.2 67.8 53.3 45.2 ...
 $ Examination     : int  15 6 5 12 17 9 16 14 12 16 ...
 $ Education       : int  12 9 5 7 15 7 7 8 7 13 ...
 $ Catholic        : num  9.96 84.84 93.4 33.77 5.16 ...
 $ Infant.Mortality: num  22.2 22.2 20.2 20.3 20.6 26.6 23.6 24.9 21 24.4 ...
```

```
> hist(df$Fertility)
```



swiss

```
> f1 <- lm(Fertility ~ Examination + Catholic, data = df)
> summary(f1)

Call:
lm(formula = Fertility ~ Examination + Catholic, data = df)

Residuals:
     Min       1Q   Median       3Q      Max
-26.2643  -5.6510  -0.0017   7.7268  17.7103

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 83.03566    4.97730  16.683  < 2e-16 ***
Examination -0.88619    0.21736  -4.077 0.000188 ***
Catholic     0.04179    0.04158   1.005 0.320322
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 9.641 on 44 degrees of freedom
Multiple R-squared:  0.4302,    Adjusted R-squared:  0.4043
F-statistic: 16.61 on 2 and 44 DF,  p-value: 4.218e-06
```

Чем выше значение оценки физической подготовки, тем ниже рождаемость

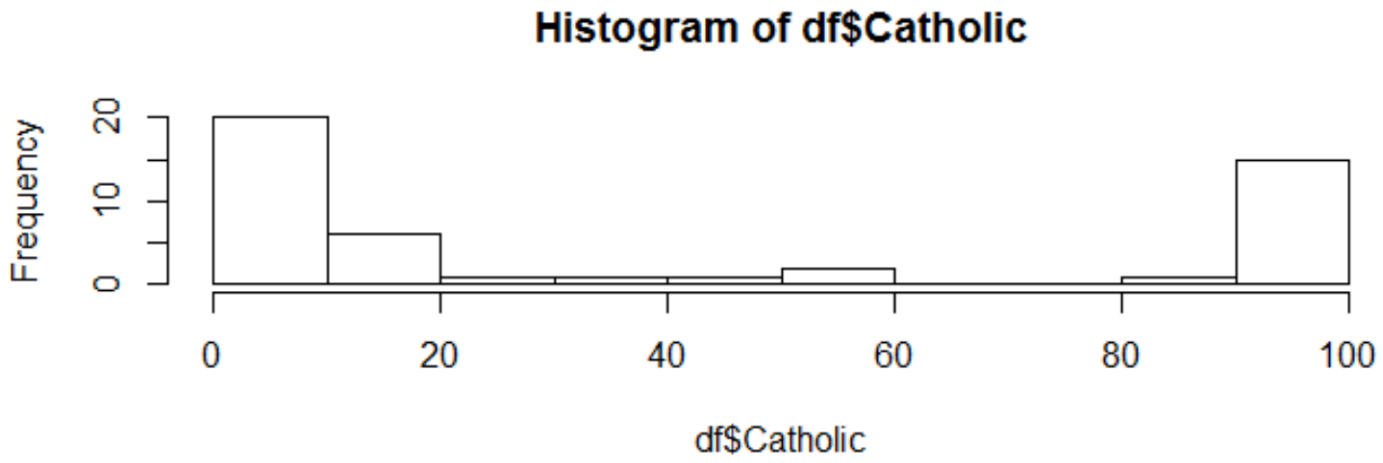
confint()

```
> f1 <- lm(Fertility ~ Examination * Catholic, data = df)
> summary(f1)

Call:
lm(formula = Fertility ~ Examination * Catholic, data = df)

Residuals:
     Min       1Q   Median       3Q      Max
-25.5446  -5.3640   0.5461   7.5383  18.5540

Coefficients:
                     Estimate Std. Error t value Pr(>|t|)
(Intercept)          80.957567   6.471732  12.509 6.37e-16 ***
Examination          -0.765480   0.323031  -2.370   0.0224 *
Catholic              0.083823   0.092648   0.905   0.3706
Examination:Catholic -0.003337   0.006559  -0.509   0.6135
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 9.723 on 43 degrees of freedom
Multiple R-squared:  0.4337,    Adjusted R-squared:  0.3941
F-statistic: 10.98 on 3 and 43 DF,  p-value: 1.77e-05
```

```
> confint(f1)
                         2.5 %       97.5 %
(Intercept)          67.90607532 94.009058379
Examination          -1.41693405 -0.114025080
Catholic             -0.10301954  0.270665084
Examination:Catholic -0.01656482  0.009890962
```

Расчет доверительных интервалов для оценки коэффициентов:
если границы доверительных интервалов пересекают 0, то они не предсказывают зависимую переменную

Добавить категориальную переменную



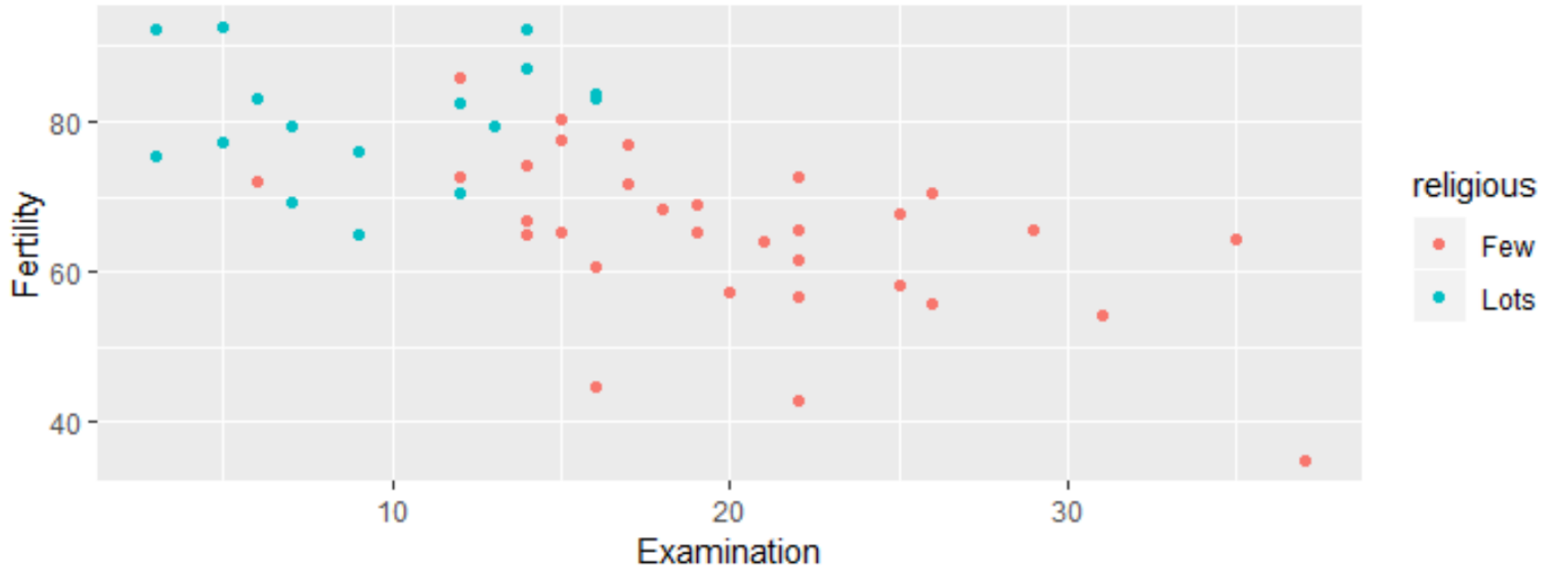

```
> df$religious <- ifelse(df$Catholic>60,"Lots","Few")
> df$religious <- factor(df$religious)
> str(df)
'data.frame':   47 obs. of  7 variables:
 $ Fertility       : num  80.2 83.1 92.5 85.8 76.9 76.1 83.8 92.4 82.4 82.9 ...
 $ Agriculture     : num  17 45.1 39.7 36.5 43.5 35.3 70.2 67.8 53.3 45.2 ...
 $ Examination     : int  15 6 5 12 17 9 16 14 12 16 ...
 $ Education       : int  12 9 5 7 15 7 7 8 7 13 ...
 $ Catholic        : num  9.96 84.84 93.4 33.77 5.16 ...
 $ Infant.Mortality: num  22.2 22.2 20.2 20.3 20.6 26.6 23.6 24.9 21 24.4 ...
 $ religious       : Factor w/ 2 levels "Few","Lots": 1 2 2 1 1 2 2 2 2 2 ...
```

```
> f1 <- lm(Fertility ~ Examination + religious, data = df)
> summary(f1)

Call:
lm(formula = Fertility ~ Examination + religious, data = df)

Residuals:
     Min       1Q   Median       3Q      Max
-22.9026  -4.8974   0.1926   7.1239  15.4542

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)    78.5753     4.7701  16.472   <2e-16 ***
Examination    -0.6858     0.2222  -3.086   0.0035 **
religiousLots   8.4469     3.7016   2.282   0.0274 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 9.221 on 44 degrees of freedom
Multiple R-squared:  0.4788,    Adjusted R-squared:  0.4552
F-statistic: 20.21 on 2 and 44 DF,  p-value: 5.934e-07
```

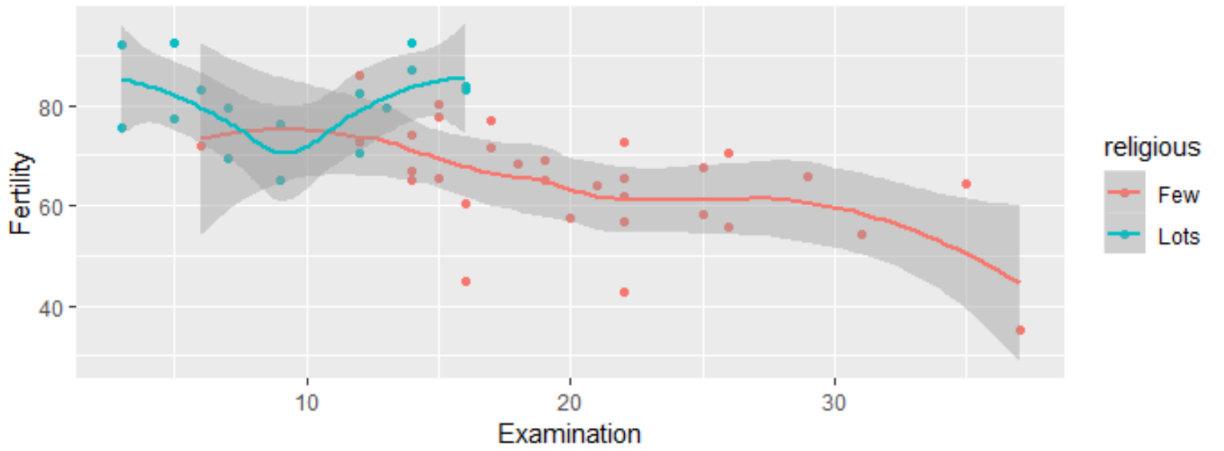
```
> f1 <- lm(Fertility ~ Examination * religious, data = df)
> summary(f1)

Call:
lm(formula = Fertility ~ Examination * religious, data = df)

Residuals:
     Min       1Q   Median       3Q      Max
-23.6289  -4.2417   0.0795   6.4508  14.0243

Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)                82.1160     5.0736  16.185  < 2e-16 ***
Examination                -0.8617     0.2389  -3.607 0.000801 ***
religiousLots              -2.9615     7.4096  -0.400 0.691366
Examination:religiousLots   1.0096     0.5723   1.764 0.084839 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 9.007 on 43 degrees of freedom
Multiple R-squared:  0.514,     Adjusted R-squared:  0.4801
F-statistic: 15.16 on 3 and 43 DF,  p-value: 7.128e-07
```

```
> f1 <- lm(Fertility ~ religious * Examination, data = df)
> summary(f1)

Call:
lm(formula = Fertility ~ religious * Examination, data = df)

Residuals:
     Min       1Q   Median       3Q      Max
-23.6289  -4.2417   0.0795   6.4508  14.0243

Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)                82.1160     5.0736  16.185  < 2e-16 ***
religiousLots              -2.9615     7.4096  -0.400 0.691366
Examination                -0.8617     0.2389  -3.607 0.000801 ***
religiousLots:Examination   1.0096     0.5723   1.764 0.084839 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 9.007 on 43 degrees of freedom
Multiple R-squared:  0.514,     Adjusted R-squared:  0.4801
F-statistic: 15.16 on 3 and 43 DF,  p-value: 7.128e-07
```

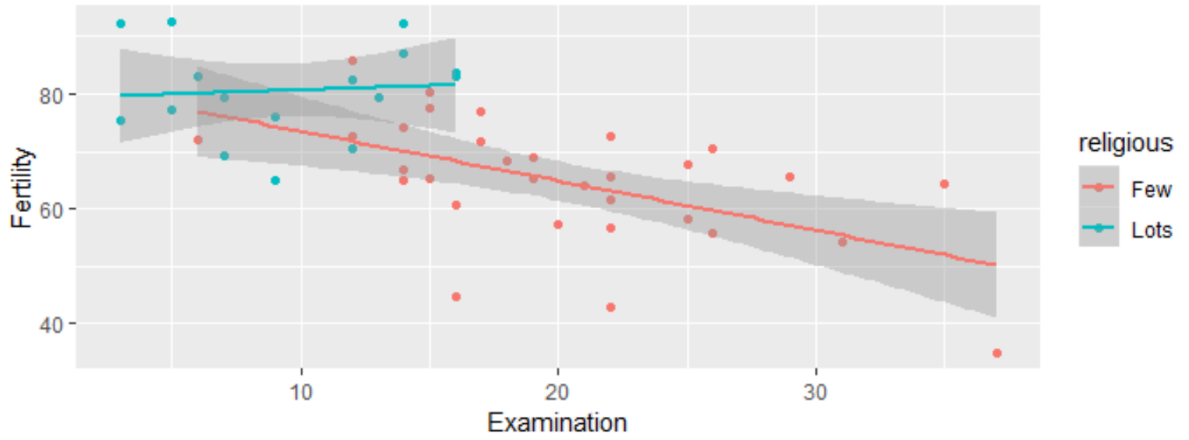
```
> ggplot(df,aes(x=Examination,y=Fertility,col=religious))+
+   geom_point()
```



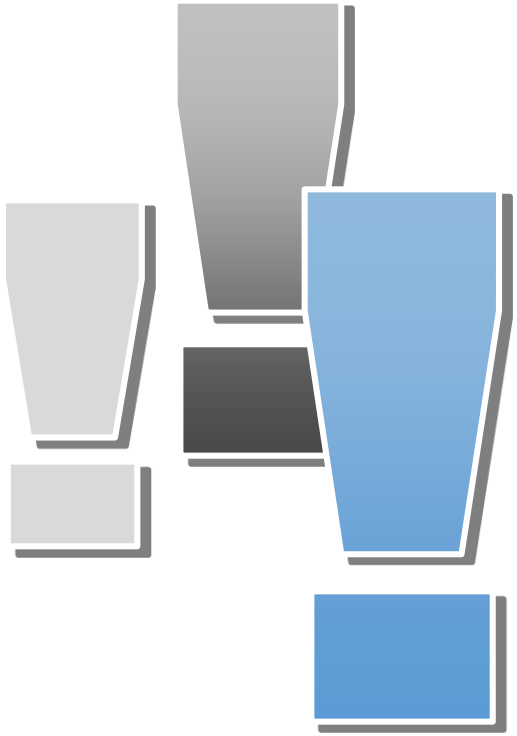
```
> ggplot(df,aes(x=Examination,y=Fertility,col=religious))+
+   geom_point()+
+   geom_smooth()
```




```
> ggplot(df,aes(x=Examination,y=Fertility,col=religious))+
+   geom_point()+
+   geom_smooth(method = "lm")
```



Спасибо за внимание!



Шевцов Василий Викторович

shevtsov_vv@rudn.university
+7(903)144-53-57