МИРЭА
Российский технологический университет

Институт технологий управления

Программные средства анализа данных

01.03.05 Статистика

Профиль «Анализ данных в бизнесе и экономике»

Квалификация «бакалавр»

Бурцева Татьяна Александровна

Профессор кафедры статистики и математических методов в управлении

Burceva_t@mirea.ru

Москва, 2022

Тема 4. Построение моделей парной и множественной линейной и нелинейной регрессии в R

План лекции

1. Команды в R для построения регрессии.
2. Проверка гипотез и обоснование статистической значимости коэффициентов.
3. Визуализации парной и множественной регрессии.
4. Регрессия с учётом факторной переменной.

МИРЭА
Российский технологический университет

1. Команды в R для построения регрессии

fm <- lm(y ~ x, data=dummy) summary(fm)	Подогнать простую линейную регрессию **y** по **x** и вывести результат анализа.
fm1 <- lm(y ~ x, data=dummy, weight=1/w^2) summary(fm1)	Поскольку мы знаем стандартные отклонения, мы можем подогнать взвешенную регрессию.
attach(dummy)	Сделать столбцы в таблице данных видимыми в качестве переменных.
lrf <- lowess(x, y)	Вычислить функцию непараметрической локальной регрессии.
plot(x, y)	Стандартный рисунок точек.
lines(x, lrf$y)	Добавить в него локальную регрессию.
abline(0, 1, lty=3)	Истинная линия регрессии: (отсекающий отрезок 0, наклон 1).
abline(coef(fm))	Невзвешенная линия регрессии.

В 2005 году Вито Риччи (Vito Ricci) составил список из 205 функций, которые используются для регрессионного анализа в R (http:// cran.r-project.org/doc/contrib/Ricci-refcard-regression.pdf).

Российский технологический университет МИРЭА

Задачи регрессии

- 1. Задача оценки с помощью одного параметра. Примеры: оценка заёмщика по критериям, оценка стоимости жилья.
- 2. Предсказание значений (прогноз).
- 3. Понижение размерности для визуализация многомерных данных и векторных представлений данных

Технология применения

Принято делить набор данных на три непересекающиеся части

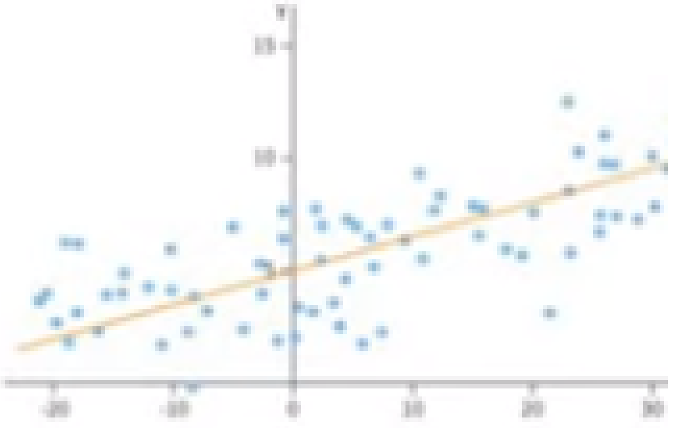
- **обучающая** (training sa – на ней происходит обучение модели
- **валидационная** (validation sample) – на ней считают метрики качества, а по ним уже подбирают гиперпараметры
- **тестовая** (test sample) – по ней оценивают качество обученной модели

Разновидности регрессионного анализа: типы регрессии

Тип регрессии	Для чего обычно используется
Простая линейная	Предсказание значений количественной зависимой переменной по значениям одной количественной независимой переменной
Полиномиальная	Предсказание значений количественной зависимой переменной по значениям количественной независимой переменной, когда взаимосвязь моделируется как полином n-ой степени
Множественная линейная	Предсказание значений количественной зависимой переменной по значениям двух и более количественных независимых переменных
Многомерная	Предсказание значений более чем одной зависимой переменной по значениям одной и более независимых переменных
Логистическая	Предсказание значений категориальной зависимой переменной по значениям одной и более независимых переменных
Пуассона	Предсказание значений зависимой счетной переменной по значениям одной или более независимых переменных
Пропорциональных рисков Кокса	Предсказание времени до наступления события (смерти, аварии, рецидива) по значениям одной или более независимых переменных
Временных рядов	Моделирование временных рядов с коррелированными ошибками
Нелинейная	Предсказание значений количественной зависимой переменной по значениям одной и более независимых переменных с использованием нелинейной модели
Непараметрическая	Предсказание значений количественной зависимой переменной по значениям одной и более независимых переменных с использованием полученной из данных и незаданной заранее модели
Устойчивая	Предсказание значений количественной зависимой переменной по значениям одной и более независимых переменных с использованием метода, устойчивого к выбросам

Российский технологический университет МИРЭА

Линейная регрессия



Описание

Линейная зависимость между переменными описывается уравнением общего вида $y = \alpha_0 + \alpha_1 x_1 + \alpha_2 x_2 + \ldots + \alpha_n x_n + \varepsilon$ где y - зависимая переменная, $\alpha_0, \alpha_1, \ldots, \alpha_n$ - неизвестные константы, $x_1, x_2, \ldots, x_n$ - известные (независимые) переменные, и ε - нормально распределенная случайная величина с нулевым матожиданием и дисперсией σ^2_{ost}. Задачей построения линейной среднеквадратической модели регрессионной зависимости переменной y от независимых переменных является получение оценки параметров $\alpha_0, \alpha_1, \ldots, \alpha_n$ и оценка адекватности построенной модели вида

$$\hat{y} = a_0 + a_1 x_1 + \ldots + a_n x_n$$

где $a_0, a_1, \ldots, a_n$ - оценки параметров $\alpha_0, \alpha_1, \ldots, \alpha_n$.

Рассмотрим простейший случай одной независимой переменной:

$\hat{y} = a + bx$ В этом уравнении модели линейной регрессии a - свободный член, а параметр b определяет наклон линии регрессии по отношению к осями координат. Параметры a и b определяются методом наименьших квадратов, который приводит к формуле:

$$b = r_{xy}\frac{S_y}{S_x},\ a = \bar{y} - b\bar{x},$$

где

$\bar{y}, \bar{x}$ - выборочные средние арифметические;

S_x, S_y - выборочные средние квадратичные отклонения;

r_{xy} - выборочный коэффициент корреляции.

Символы, которые используются в формулах

Символ	Назначение
~	Отделяет зависимые переменные (слева) от независимых (справа). Например, предсказание значений `y` по значениям `x`, `z` и `w` будет закодировано так: `y ~ x + z + w`
+	Разделяет независимые переменные
:	Обозначает взаимодействие между независимыми переменными. Предсказание значений `y` по значениям `x`, `z` и взаимодействия между `x` и `z` будет закодировано как `y ~ x + z + x:z`
*	Краткое обозначение для всех возможных взаимодействий. Код `y ~ x * z * w` в полном виде означает `y ~ x + z + w + x:z + x:w + z:w + x:z:w`
^	Обозначает взаимодействия до определенного порядка. Код `y ~ (x + z + w)^2` в полном виде будет записан как `y ~ x + z + w + x:z + x:w + z:w`
.	Символ-заполнитель для всех переменных в таблице данных, кроме зависимой. Например, если таблица данных содержит переменные `x`, `y`, `z` и `w`, то код `y ~ .` будет означать `y ~ x + z + w`
–	Знак минуса удаляет переменную из уравнения. Например, `y ~ (x + z + w)^2 – x:w` соответствует `y ~ x + z + w + x:z + z:w`
-1	Подавляет свободный член уравнения. Например, формула `y ~ x -1` позволяет подогнать такую регрессионную модель для предсказания значений `y` по `x`, чтобы ее график проходил через начало координат
I()	Элемент в скобках интерпретируется как арифметическое выражение. Например, `y ~ x + (z + w)^2` означает `y ~ x + z + w + z:w`. Для сравнения `y ~ x + I((z + w)^2)` означает `y ~ x + h`, где `h` – это новая переменная, полученная при возведении в квадрат суммы `z` и `w`
function	В формулах можно использовать математические функции. Например, `log(y) ~ x + z + w` будет предсказывать значения `log(y)` по значениям `x`, `z` и `w`

Для построения линейной модели регрессии используется функция **lm(formula=f)**, которая в простейшем случае содержит только формулу от переменных (векторов, содержащих элементы парной выборки); запись *y~x* означает, что строится модель зависимости *y* от *x*.

```
> x<-c(3.6,7.8,9.6,5.7,8.9)
> y<-c(2.7,8.9,6.5,8.8,6.4)
> p.lm<-lm(formula=x~y)
> summary(p.lm)

Residuals:
      1       2       3       4       5
-1.7151 -0.3409  2.5529 -2.3954  1.8985

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   4.0845     3.5050   1.165    0.328
y             0.4558     0.4985   0.914    0.428

Residual standard error: 2.511 on 3 degrees of freedom
Multiple R-Squared: 0.2179,    Adjusted R-squared: -
0.0428
F-statistic: 0.8358 on 1 and 3 DF,  p-value: 0.428
```

Команда **summary()** выдает полную информацию о построенной модели:

значения остатков (residuals - разность модельных и истинных значений переменной *y*). Если объем выборки большой, то печатается оценка распределения остатков (квартили).

коэфициенты модели и оценку их значимости по критерию Стьюдента (в нашем случае все коэфициенты не значимы, поскольку все вероятности (0.328 и 0.428) больше 0.05 - т.е. нельзя считать, что существует линейная зависимость между x и *y*).

Оценку значимости зависимости по критерию Фишера и квадрат коэфициента корреляции (R-squared), который показывает долю дисперсии *y*, объясненной с использованием модели (исправленное значение для R2 равно 0, статистика Фишера *F*=0.8358, уровень значимости критерия Фишера 42.8%, т.е. зависимость отсуствует).

Линейная регрессия в R

https://stepik.org/lesson/11508/step/8?unit=2531

Пример парной линейной регрессии на переменных df$mpg и df$hpиз data frame df

- fit <- lm(mpg ~ hp, df) – сохранили модель в переменную fit, переменная hp – это независимая переменная

```
> fit  <- lm(mpg ~ hp, df)
> fit

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

Coefficients:
(Intercept)           hp
   30.09886     -0.06823
```

```
fit                 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"
```

МИРЭА
Российский технологический университет

Расчетные формулы

1. Оценки коэффициентов однофакторной регрессионной модели:

$$b_1 = \frac{\overline{xy} - \bar{x}\,\bar{y}}{\overline{x^2} - \bar{x}^2}, \qquad b_0 = \bar{y} - b_1\bar{x},$$

где

$$\bar{x} = \frac{1}{N}\sum_{i=1}^{N} x_i, \qquad \bar{y} = \frac{1}{N}\sum_{i=1}^{N} y_i, \qquad \overline{xy} = \frac{1}{N}\sum_{i=1}^{N} x_i y_i, \qquad \overline{x^2} = \frac{1}{N}\sum_{i=1}^{N} x_i^2,$$

x - независимая переменная, y - зависимая переменная, N - число элементов выборочной совокупности.

2. Коэффициент корреляции:

$$r_{xy} = b_1\frac{\sigma_x}{\sigma_y} = \frac{\overline{xy} - \bar{x}\,\bar{y}}{\sigma_x \sigma_y},$$

где σ_x, σ_y - среднеквадратические ошибки, вычисляемые по формулам

$$\sigma_x = \sqrt{\frac{1}{n}\sum x_i^2 - \bar{x}^2}, \qquad \sigma_y = \sqrt{\frac{1}{n}\sum y_i^2 - \bar{y}^2}.$$

3. Коэффициент детерминации:

$$D = r^2.$$

4. Дисперсионное отношение Фишера (F-критерий):

$$F_{расч} = \frac{\sum(\hat{y}-\bar{y})^2 / m}{\sum(y-\hat{y})^2 / (n-m-1)} = \frac{r_{xy}^2}{1-r_{xy}^2}(n-2),$$

где $\hat{y}$ – расчетное значение зависимой переменной ($\hat{y} = b_0 + b_1 x$), n – число элементов выборочной совокупности, m – число факторов.

5. Стандартные ошибки параметров линейной регрессии:

$$s_{b_1} = \sqrt{\frac{\sum(y-\hat{y})^2/(n-2)}{\sum(x-\bar{x})^2}} = \sqrt{\frac{S_{ост}^2}{\sum(x-\bar{x})^2}} = \frac{S_{ост}}{\sigma_x\sqrt{n}},$$

$$s_{b_0} = \sqrt{\frac{\sum x^2}{n\sum(x-\bar{x})^2}\cdot\frac{\sum(y-\hat{y})^2}{(n-2)}} = \sqrt{S_{ост}^2\frac{\sum x^2}{n^2\sigma_x^2}} = S_{ост}\frac{\sqrt{\sum x^2}}{n\sigma_x},$$

где $S_{ост}^2$ – остаточная дисперсия, рассчитываемая по формуле

$$S_{ост}^2 = \frac{\sum(y-\hat{y})^2}{n-m-1}.$$

6. t-статистики Стьюдента:

$$t_{b_0} = \frac{b_0}{s_{b_0}}, \quad t_{b_1} = \frac{b_1}{s_{b_1}}.$$

Доверительные интервалы для коэффициентов уравнения регрессии

7. Доверительные интервалы:

$$b_0 - \Delta_{b_0} \le b_0 \le b_0 + \Delta_{b_0}, \quad b_1 - \Delta_{b_1} \le b_1 \le b_1 + \Delta_{b_1},$$

где Δ_{b_0}, Δ_{b_1} – предельные ошибки, рассчитываемые по формулам

$$\Delta_{b_0} = t_{табл} s_{b_0}, \quad \Delta_{b_1} = t_{табл} s_{b_1},$$

$t_{табл}$ – табличное значение t-статистики.

8. Индекс корреляции: $\sqrt{1 - \dfrac{\sum (y - \hat{y})^2}{\sum (y - \bar{y})^2}}$.

9. Усредненное значение коэффициента эластичности: $b_1 \cdot \dfrac{\bar{x}}{\bar{y}}$

МИРЭА
Российский технологический университет

Функции, полезные при подгонке линейных моделей

Функция	Действие
`summary()`	Показывает детальную информацию о подогнанной модели
`coefficients()`	Перечисляет параметры модели (свободный член и регрессионные коэффициенты)
`confint()`	Вычисляет доверительные интервалы для параметров модели (по умолчанию 95%)
`fitted()`	Выводит на экран предсказанные значения, согласно подогнанной модели
`residuals()`	Показывает остатки для подогнанной модели
`anova()`	Создает таблицу ANOVA (дисперсионного анализа) для подогнанной модели или таблицу ANOVA, сравнивающую две или более моделей
`vcov()`	Выводит ковариационную матрицу для параметров модели
`AIC()`	Вычисляет информационный критерий Акаике (Akaike's Information Criterion)
`plot()`	Создает диагностические диаграммы для оценки адекватности модели
`predict()`	Использует подогнанную модель для предсказания зависимой переменной для нового набора данных

МИРЭА
Российский технологический университет

2. Проверка гипотез и обоснование статистической значимости коэффициентов

ОЦЕНКА КАЧЕСТВА РЕГРЕССИИ

- Средний квадрат отклонения (MSE – Mean Squared Error) $MSE = \overline{(\hat{f}(X) - y)^2}$
- RMSE – Root Mean Squared Error $RMSE = \sqrt{MSE}$
- Средний модуль отклонения (MAE – Mean Absolute Error) $MAE = \overline{|\hat{f}(X) - y|}$
- Средний процент отклонения (MAPE – Mean Absolute Percent Error) $MAPE = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{\hat{f}(X_i)-y_i}{y_i}\right|$
- Коэффициент детерминации (R^2) $R^2 = 1 - \frac{MSE}{V(y)}$

summary(fit)

```
> df  <- mtcars
> df_numeric  <- df[,c(1,3:7)]
>
> fit  <- lm(mpg ~ hp, df)
> summary(fit)

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

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
```

*** - значим коэффициент

t-статистика

При оценке значимости коэффициента линейной регрессии на начальном этапе можно использовать следующее *"грубое" правило*, позволяющее не прибегать к таблицам.

Если стандартная ошибка коэффициента больше его модуля ($|t| < 1$), то коэффициент не может быть признан значимым, т. к. доверительная вероятность здесь при двусторонней альтернативной гипотезе составит менее чем 0.7.

Если $1 < |t| < 2$, то найденная оценка может рассматриваться как

относительно (слабо) значимая. Доверительная вероятность в этом случае лежит между значениями 0.7 и 0.95.

Если $2 < |t| < 3$, то это свидетельствует о значимой линейной связи между X и Y. В этом случае доверительная вероятность колеблется от 0.95 до 0.99.

Наконец, если $|t| > 3$, то это почти гарантия наличия линейной связи.

Визуализация результатов регрессии в R

- Для визуализации построенной модели можно использовать вспомогательные функции:

Описание функций

```
abline(a, b, untf = FALSE, ...)
abline(h=, untf = FALSE, ...)
abline(v=, untf = FALSE, ...)
```

Параметры

`a,b`	Параметры в линейном уравнении
`untf`	Если TRUE, то рисует линию в преобразованных координатах
`h,v`	Y и X значения для горизонтальной и вертикальной линии соответственно

```
plot(x, y, xlim=range(x),ylim=range(y),type="p",
main, xlab, ylab, ...)
```

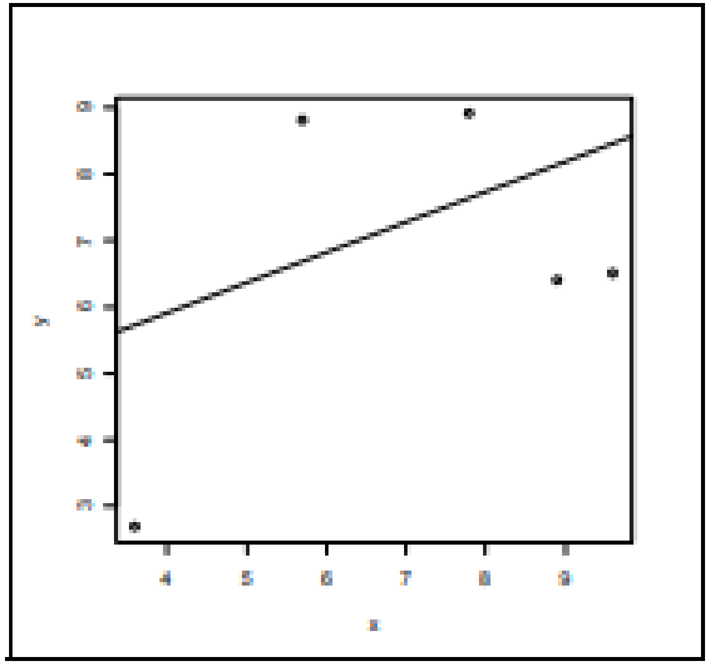
Параметры

`X,Y`	Координаты точек x и y.
`xlim, ylim`	Значения для осей x и y.
`Type`	Тип графика(" p" для точек)
`Main`	**Название графика**
`Xlab,ylab`	Название осей.

Функция `abline()` строит прямую по найденным a и b.
Функция `plot()` строит экспериментальные точки.

Пример

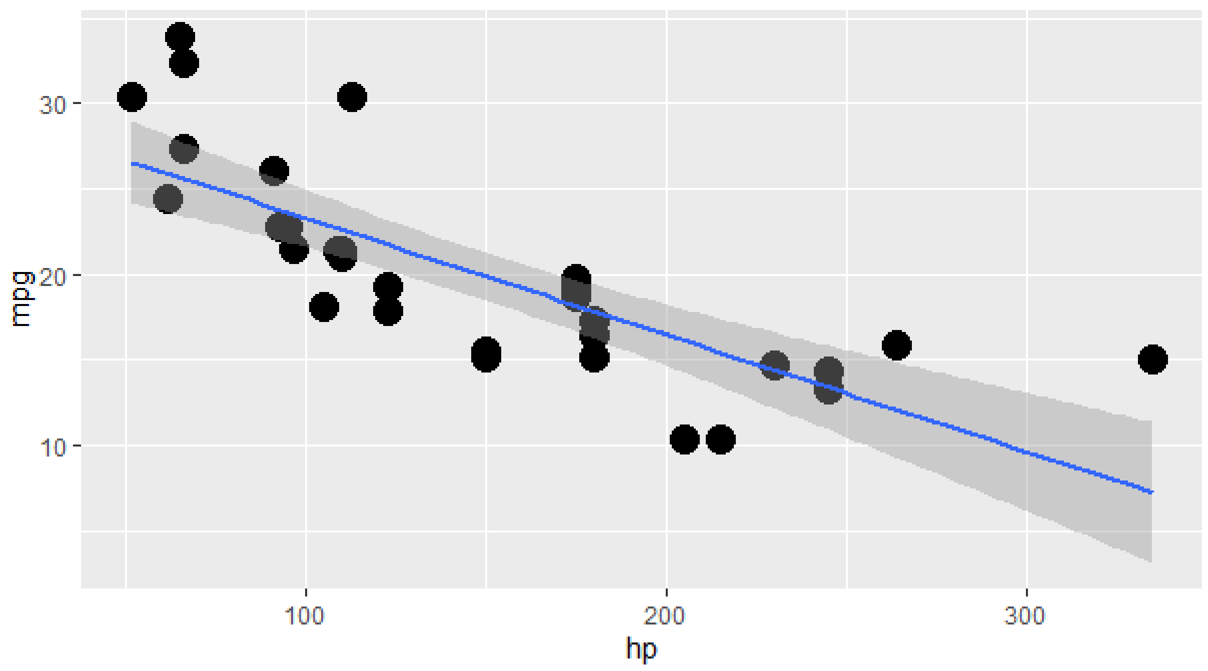
```
plot(x,y)
abline(lm(x~y))
```



Визуализация результатов и доверительного интервала в R

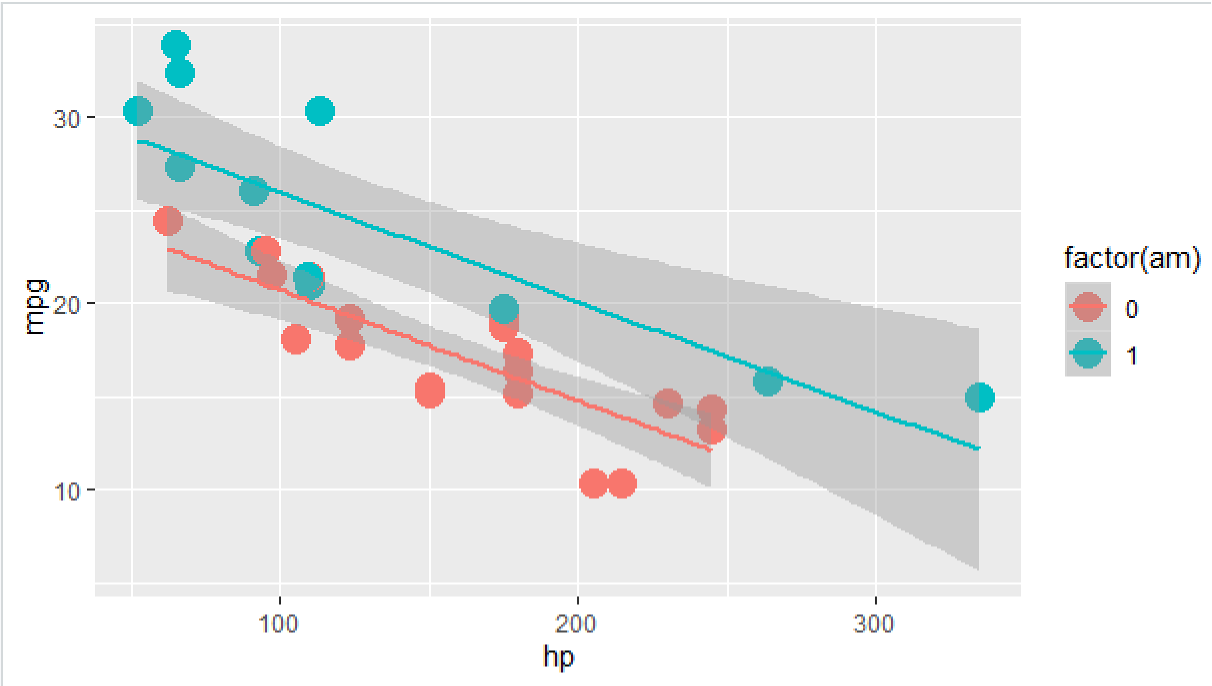
```
fit  <- lm(mpg ~ hp, df)
summary(fit)
library(ggplot2)
ggplot(df, aes(hp, mpg))+
  geom_point(size = 5)+
  geom_smooth(method = "lm")
```

указываем тип линии регрессии в параметре method



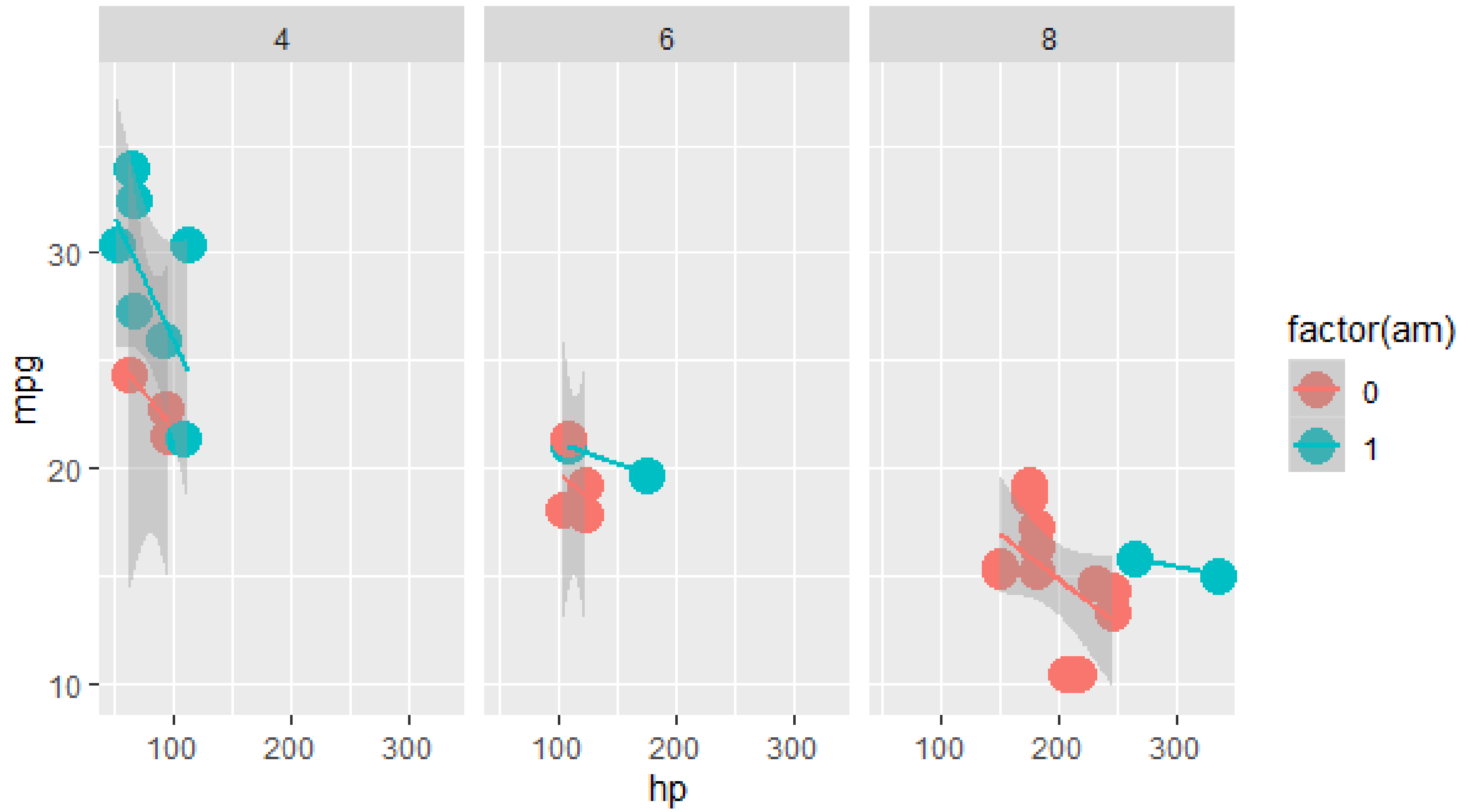
Можно наблюдения разбить по группам по фактору am и получить две линии регрессии

```
ggplot(df, aes(hp, mpg, col = factor(am)))+
  geom_point(size = 5)+
  geom_smooth(method = "lm")
```



Если еще одну переменную ввести

```
ggplot(df, aes(hp, mpg, col = factor(am)))+
  geom_point(size = 5)+
  geom_smooth(method = "lm")+
  facet_grid(.~cyl)
```



Как убрать доверительные интервалы

```
ggplot(df, aes(hp, mpg))+
  geom_smooth(method = "lm", se = F)+
  facet_grid(.~cyl)
```

Как считать с помощью модели

- new_hp <- data.frame(hp = c(100, 150, 129, 300)) – задали новые значения независимой переменной
- predict(fit, new_hp) рассчитали их с помощью модели, которая хранится в переменной fit

```
> predict(fit, new_hp)
        1         2         3         4 
23.276033 19.864619 21.297413  9.630377 
```

Если в качестве зависимой переменной выступает не числовая переменная

- my_df <- mtcars
- my_df$cyl <- factor(my_df$cyl, labels = c("four", "six", "eight"))
- fit <- lm(mpg ~ cyl, my_df)

Памятка

```
cor.test(mtcars$mpg, mtcars$disp) # Расчет корреляции Пирсона

cor.test(~ mpg + disp, mtcars) # запись через формулу

cor.test(mtcars$mpg, mtcars$disp, method = "spearman") # Расчет корреляции Спирмена
cor.test(mtcars$mpg, mtcars$disp, method = "kendall") # Расчет корреляции Кендала
cor(iris[, -5]) # построение корреляционной матрицы
fit <- lm(mpg ~ disp, mtcars) # построение линейной регрессии
fit$coefficients # коэффициенты регрессии
fit$fitted.values # предсказанные значения зависимой переменной
```

Множественная регрессия

- Если существует больше одной независимой переменной, простая линейная регрессия превращается во множественную линейную регрессию, а ход вычислений становится более сложным.
- С технической точки зрения, полиномиальная регрессия – это частный случай множественной регрессии.
- При квадратичной регрессии есть две независимые переменные (X и X^2), а при кубической регрессии – три независимые переменные (X, X^2, X^3).
- Если существует больше одной независимой переменной, то регрессионные коэффициенты показывают, на сколько увеличится значение зависимой переменной при изменении данной независимой переменной на единицу при условии, что все остальные независимые переменные останутся неизменными

Множественная регрессия в R
multiple linear regression

(в swiss из R хранятся данные по областям Швейцарии - Standardized fertility measure and socio-economic indicators for each of 47 French-speaking provinces of Switzerland at about 1888.)

- Зависимая Независимые переменные
- lm(Fertility ~ Examination + Catholic, data = swiss)

```
> str(swiss)
'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 ...
```

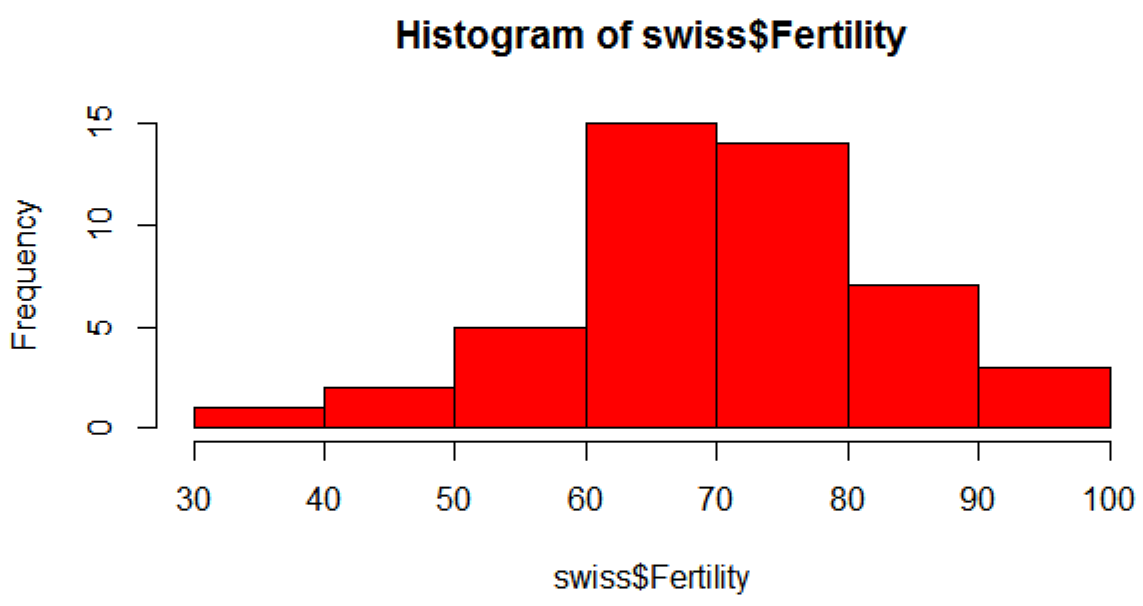
```
> swiss
```

	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
Gruyere	82.4	53.3	12	7	97.67	21.0
Sarine	82.9	45.2	16	13	91.38	24.4
Veveyse	87.1	64.5	14	6	98.61	24.5
Aigle	64.1	62.0	21	12	8.52	16.5
Aubonne	66.9	67.5	14	7	2.27	19.1
Avenches	68.9	60.7	19	12	4.43	22.7
Cossonay	61.7	69.3	22	5	2.82	18.7
Echallens	68.3	72.6	18	2	24.20	21.2
Grandson	71.7	34.0	17	8	3.30	20.0
Lausanne	55.7	19.4	26	28	12.11	20.2
La Vallee	54.3	15.2	31	20	2.15	10.8
Lavaux	65.1	73.0	19	9	2.84	20.0
Morges	65.5	59.8	22	10	5.23	18.0
Moudon	65.0	55.1	14	3	4.52	22.4

[,1] Fertility *Ig*, 'common standardized fertility measure'
[,2] Agriculture % of males involved in agriculture as occupation
[,3] Examination % draftees receiving highest mark on army examination
[,4] Education % education beyond primary school for draftees.
[,5] Catholic % 'catholic' (as opposed to 'protestant').
[,6] Infant.Mortality live births who live less than 1 year.

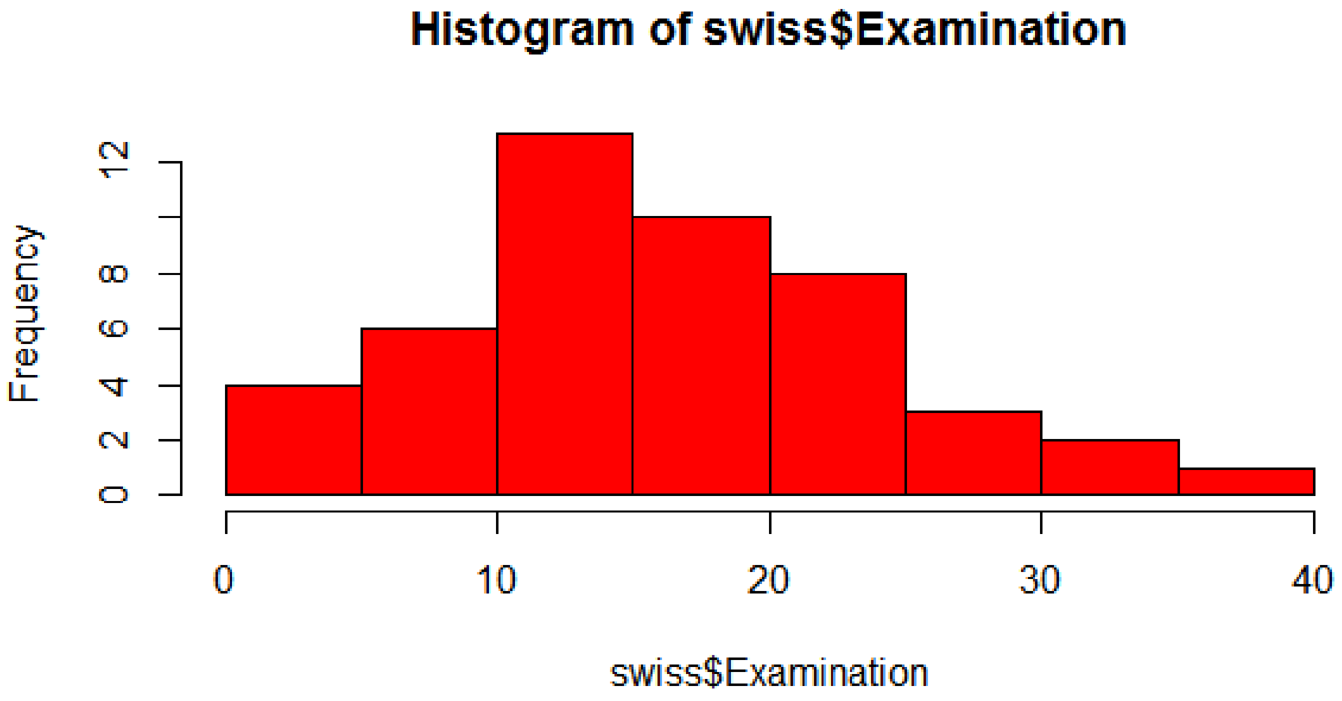
Анализ зависимой переменной Fertility (рождаемость)

- hist(swiss$Fertility, col = 'red')



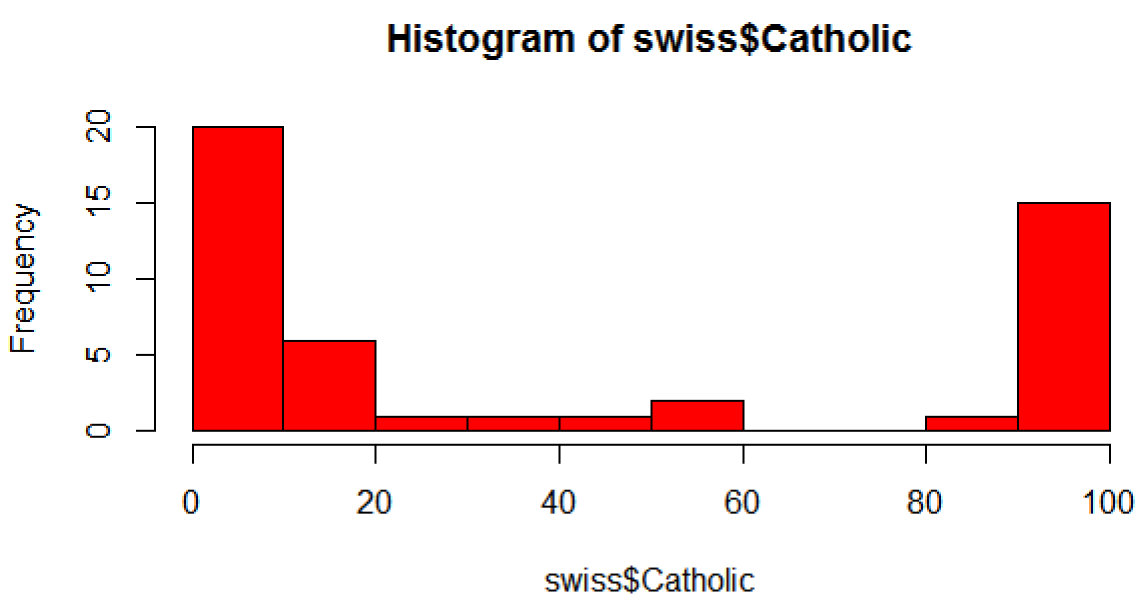
Анализ независимой переменной (оценка здоровья призывника) Examination

- hist(swiss$Examination, col = 'red')



Анализ независимой переменной Catholic (% католического населения)

- hist(swiss$Catholic, col = 'red')



РЕЗУЛЬТАТЫ (связь обратная между У и Х1)

summary(fit)

```
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
```

Catholic

Значим х1 и свободный член

- Свободный член

нормированный

$$R_{ck}^2 = 1 - \frac{n-1}{n-p-1}(1-R^2)$$

Множественная линейная регрессия со взаимодействиями

```
> fit <- lm(mpg ~ hp + wt + hp:wt, data=mtcars)
> summary(fit)

Call:
lm(formula=mpg ~ hp + wt + hp:wt, data=mtcars)

Residuals:
   Min     1Q Median     3Q    Max
-3.063 -1.649 -0.736  1.421  4.551

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 49.80842    3.60516   13.82  5.0e-14 ***
hp          -0.12010    0.02470   -4.86  4.0e-05 ***
wt          -8.21662    1.26971   -6.47  5.2e-07 ***
hp:wt        0.02785    0.00742    3.75  0.00081 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.1 on 28 degrees of freedom
Multiple R-squared: 0.885,      Adjusted R-squared: 0.872
F-statistic: 71.7 on 3 and 28 DF,  p-value: 2.98e-13
```

Из столбца Pr(>|t|) видно, что взаимодействие между мощностью двигателя и весом машины значимо. Что это означает? Значимое взаимодействие между двумя независимыми переменными свидетельствует о том, что на взаимосвязь между одной независимой переменной и зависимой влияют значения другой независимой переменной. В данном случае характер зависимости между расходом топлива и мощностью двигателя не одинаков для автомобилей разного веса.

- Можно подобрать регрессионную модель, включающую обе независимые переменные, а также взаимодействие между ними.
- Множественная линейная регрессия со значимым эффектом взаимодействия

$$\widehat{mpg} = 49.81 - 0.12 \times hp - 8.22 \times wt + 0.03 \times hp \times wt.$$

Модель с учетом взаимодействия факторов

- fit4 <- lm(Fertility ~ Catholic*Examination, data = swiss)
- summary(fit4)

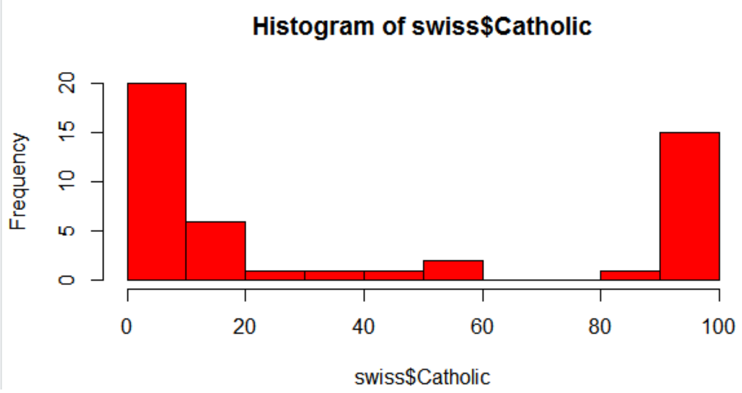
```
Residuals:
     Min       1Q   Median       3Q      Max 
-25.5446  -5.3640   0.5461   7.5383  18.5540 

Coefficients:
                      Estimate Std. Error t value Pr(>|t|)    
(Intercept)          80.957567   6.471732  12.509 6.37e-16 ***
Catholic              0.083823   0.092648   0.905   0.3706    
Examination          -0.765480   0.323031  -2.370   0.0224 *  
Catholic:Examination -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
```

Оценка доверительных интервалов коэффициентов confint(переменная)

```
> confint(fit4)
                            2.5 %       97.5 %
(Intercept)           67.90607532 94.009058379
Catholic              -0.10301954  0.270665084
Examination           -1.41693405 -0.114025080
Catholic:Examination  -0.01656482  0.009890962
>
```

Линейная регрессия с категориальными переменными

В примере есть переменная независимая, у которой такой тип



Нужно сделать эту переменную фактором

- swiss$religious <- ifelse(swiss$Catholic > 60, “Lots”, “Few”)
- swiss$religious <- as.factor(swiss$religious)

```
swiss$religious <- ifelse(swiss$Catholic > 60, "Lots", "Few")
swiss$religious <- as.factor(swiss$religious)

fit3 <- lm(Fertility ~ Examination + religious, data = swiss)
summary(fit3)
```

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

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
```

78,5753 – среднее значение рождаемости для провинций , которые типа “Few”
-0,6858 – добавка в рождаемость из-за физподготовки для этих провинций
8,4469 – рождаемость в провинциях, где много католиков

Введем взаимодействие факторов

- fit4 <- lm(Fertility ~ Examination*religions, data = swiss)
- summary(fit4)

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

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
```

- Выводы: физподготовка по-разному влияет на рождаемость в провинциях с разным уровнем фактора

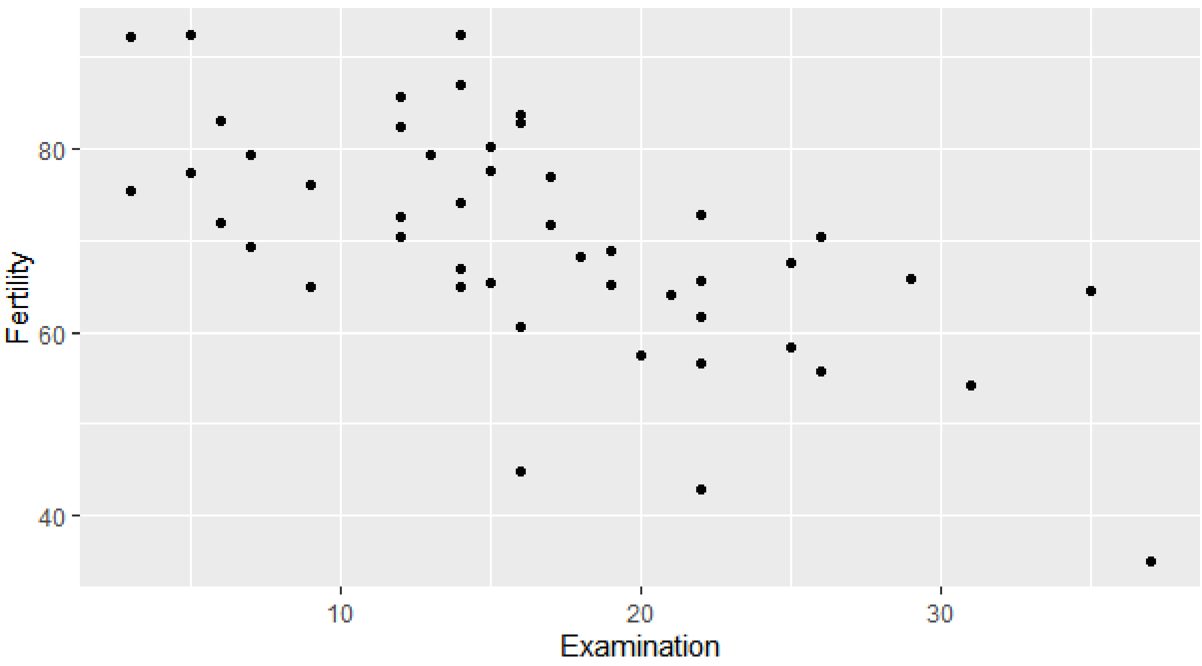
Российский
технологический
университет

3. Визуализации парной и множественной регрессии

Диаграмма рассеивания

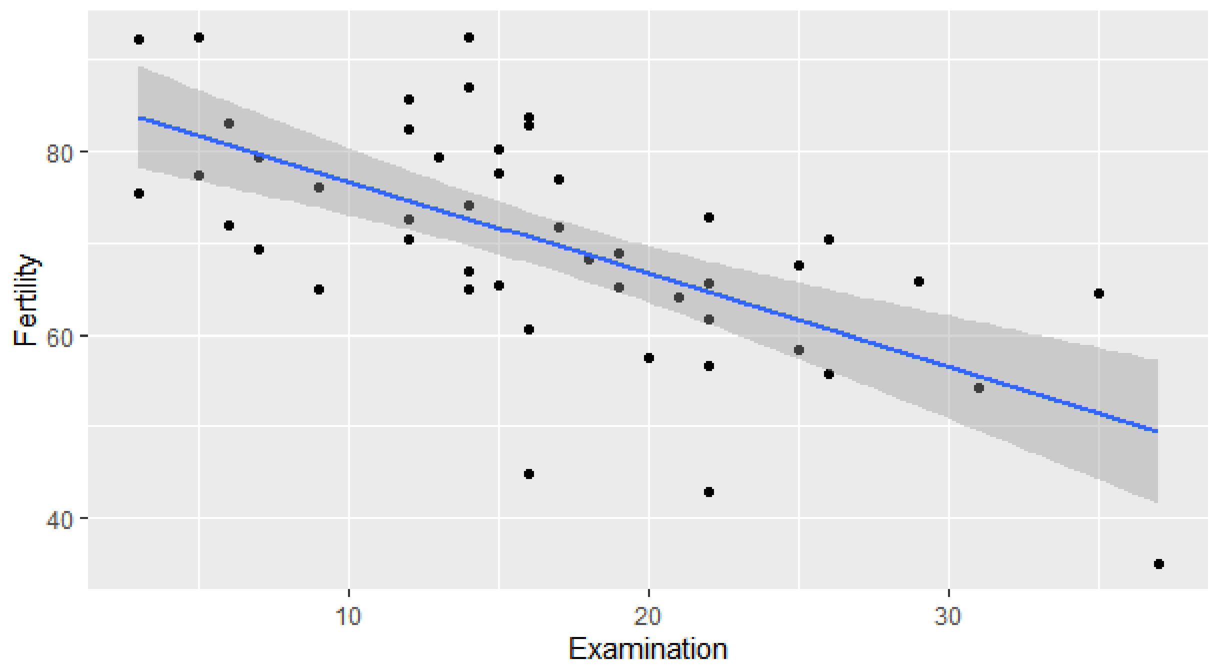
-

```
ggplot(swiss, aes(x = Examination, y = Fertility)) +   geom_point()
```



С линией регрессии

```
ggplot(swiss, aes(x = Examination, y = Fertility)) +
  geom_point() +
  geom_smooth(method = "lm")
```



Визуализация взаимодействия переменных

Взаимодействия можно визуализировать при помощи функции `effect()` из одноименного пакета. Формат ее применения таков:

```
plot(effect(term, mod, xlevels), multiline=TRUE)
```

где *term* – это член модели, который нужно отобразить на диаграмме, *mod* – подогнанная модель, выдаваемая функцией `lm()`, а *xlevels* – это список переменных, значения которых будут фиксированы, и самих этих значений. Опция `multiline=TRUE` позволяет наложить на диаграмму линии. Для нашего последнего примера это выглядит так:

```
library(effects)
plot(effect("hp:wt", fit, list(wt=c(2.2,3.2,4.2))), multiline=TRUE)
```

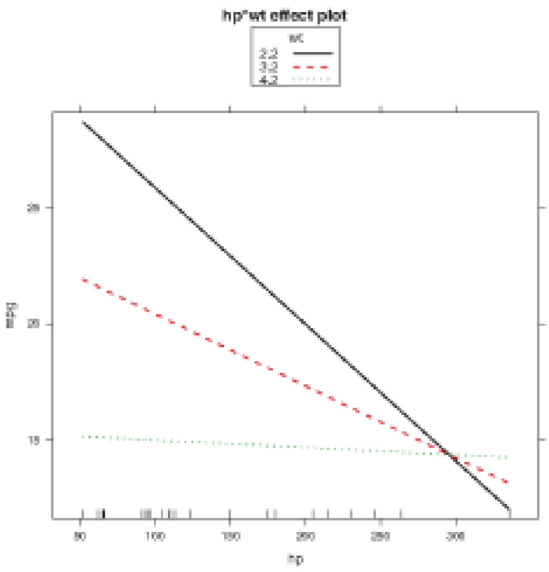


Диаграмма взаимодействий для `hp*wt`. Показана взаимосвязь между `mpg` и `hp` для трех значений `wt`

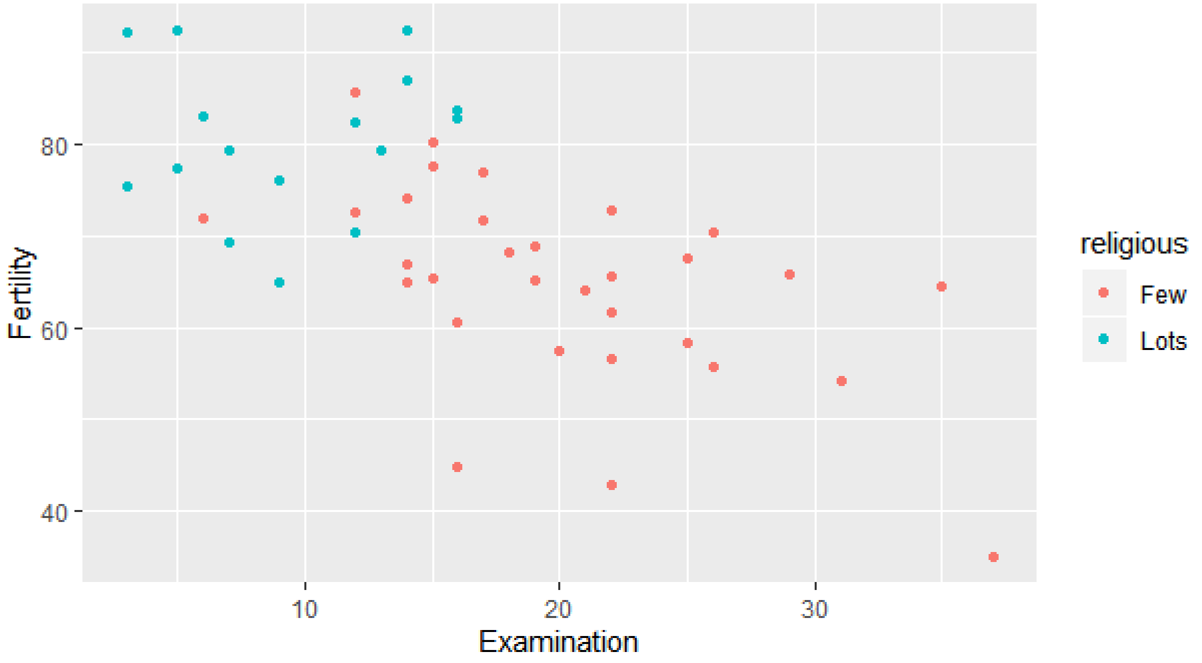
МИРЭА
Российский технологический университет

4. Регрессия с учетом факторной переменной

Диаграмма рассеивания с учетом фактора

ggplot(swiss, aes(x = Examination, y = Fertility, col = religious)) +

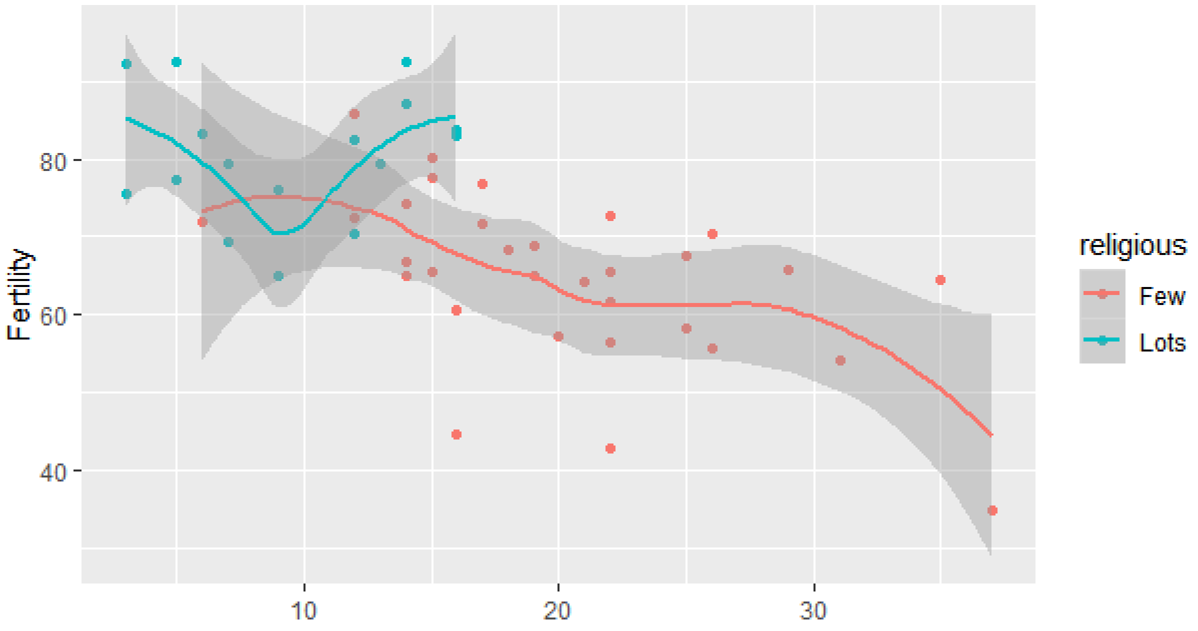
geom_point()



Линия регрессии с учётом фактора

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

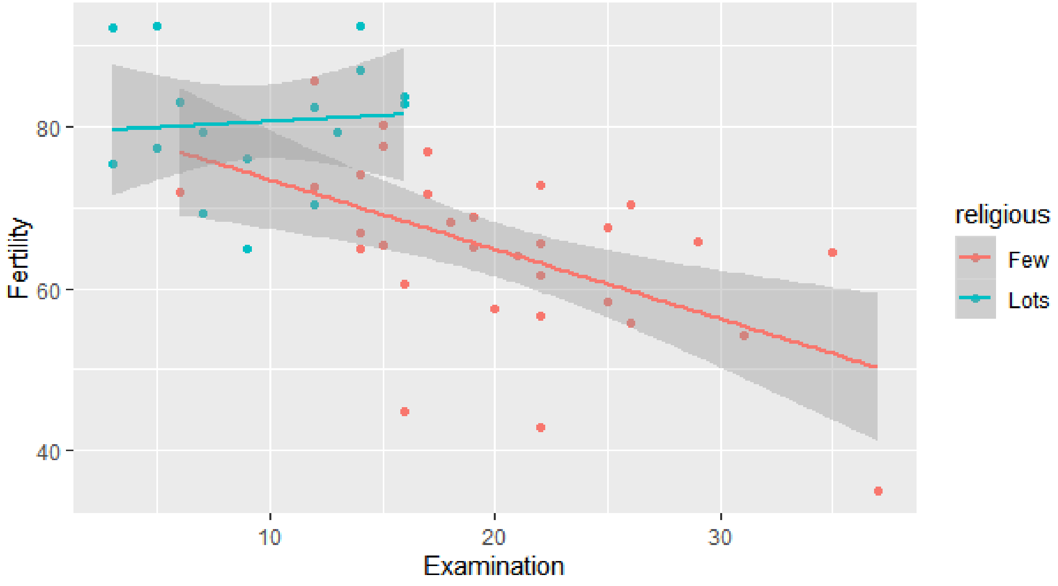
Для областей, где много католиков тренд нелинейный



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

Для областей, где мало католиков линия регрессии сохраняется как без учета фактора

Вывод: не католики обладают высокой физподготовкой и низкой рождаемость.



Памятка по интерпретации результатов регрессионного анализа с категориальными и непрерывными переменными

Модель для примера:

*DV ~ IV_numeric * IV_categorical*

IV_categorical - фактор с двумя уровнями (Level1 и Level2)

Коэффициенты:

Intercept — предсказанное значение *DV* для первого уровня *IV_categorical* с учётом того, что *IV_numeric* равна нулю.

IV_numeric — насколько изменяется предсказанное значение *DV* при увеличении *IV_numeric* на одну единицу в группе, соответствующей первому уровню *IV_categorical*

IV_categoricalLevel2 — насколько изменяется предсказанное значение *DV* при переходе от первого уровня *IV_categorical* ко второму уровню. С учётом того, что *IV_numeric* равна нулю.

IV_numeric:IV_categoricalLevel2 — насколько сильнее (или слабее) изменяется предсказанное значение *DV* при увеличении *IV_numeric* на одну единицу в группе, соответствующей второму уровню *IV_categorical*, по сравнению с первым уровнем.

Как предсказывать значения в новом датасете на основе полученных коэффициентов

1). Предположим у нас есть новый объект, про который мы знаем, что он принадлежит к группе, соответствующей *IV_categorical (Level1)* и измеренный у него *IV_numeric* составил **10**:

Предсказанное значение DV = **Intercept** + **10** * **IV_numeric**

2). Предположим у нас есть новый объект, про который мы знаем, что он принадлежит к группе, соответствующей *IV_categorical (Level2)* и измеренный у него *IV_numeric* составил **6**:

Предсказанное значение DV = **Intercept** + **IV_categoricalLevel2** + **6** * (**IV_numeric** + **IV_numeric:IV_categoricalLevel2**)

В этом примере будем работать с хорошо вам известным встроенным датасетом *mtcars*. Переменная *am* говорит о том, какая коробка передач используется в машине: 0 - автоматическая, 1 - ручная.

Сделаем эту переменную факторной.

```
mtcars$am <- factor(mtcars$am, labels = c('Automatic', 'Manual'))
```

Теперь постройте линейную модель, в которой в качестве зависимой переменной выступает расход топлива (*mpg*), а в качестве независимых - вес машины (*wt*) и коробка передач (модифицированная *am*), а также их взаимодействие. Выведите summary этой модели.

Что отражает значение intercept в данной модели?

- [ ] Расход топлива у машин со средним весом
- [ ] Средний расход топлива у машин с автоматической коробкой передач
- [ ] Средний расход топлива у машин с нулевым весом и ручной коробкой передач
- [x] Расход топлива у машин с автоматической коробкой передач и нулевым весом
- [ ] Расход топлива у машин с нулевым весом

Пошаговая регрессия в R (step)

Строим полную модель

- fit_full <- lm(Fertility ~ ., data = swiss)

Строим оптимальную модель

- optimal_fit <- step(fit_full, direction = "backward")
- summary(optimal_fit)

```
lm(formula = Fertility ~ Agriculture + Examination + Education +
    Catholic + Infant.Mortality + religious, data = swiss)

Residuals:
     Min       1Q   Median       3Q      Max 
-16.0147  -3.8274  -0.5049   4.2812  15.9119 

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)    
(Intercept)       69.4737    10.1570   6.840 3.13e-08 ***
Agriculture       -0.1656     0.0664  -2.493  0.01689 *  
Examination       -0.3465     0.2423  -1.430  0.16044    
Education         -0.5991     0.2051  -2.921  0.00571 ** 
Catholic          -0.1562     0.1111  -1.406  0.16730    
Infant.Mortality   0.9894     0.3620   2.733  0.00930 ** 
religiousLots     23.6278     9.6172   2.457  0.01845 *  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 6.762 on 40 degrees of freedom
Multiple R-squared:  0.7452,	Adjusted R-squared:  0.707 
F-statistic:  19.5 on 6 and 40 DF,  p-value: 1.765e-10
```

Проверка МНК-предпосылок в R

Функция	Назначение
`qqPlot()`	Диаграмма сравнения квантилей
`durbinWatsonTest()`	Тест Дарбина-Уотсона (Durbin-Watson test) на автокорреляцию в остатках
`crPlots()`	Диаграмма компонент и остатков
`ncvTest()`	Тест на неоднородность дисперсии остатков

Функция	Назначение
`spreadLevelPlot()`	Диаграмма для обнаружения неоднородности дисперсии остатков (spread-level plot)
`outlierTest()`	Тест Бонферрони на выбросы
`avPlots()`	Диаграммы добавленных переменных
`influencePlot()`	Диаграмма влияния наблюдений на регрессию
`scatterplot()`	Усовершенствованная диаграмма рассеяния
`scatterplotMatrix()`	Усовершенствованная матрица диаграмм рассеяния
`vif()`	Фактор инфляции дисперсии

Вопросы по теме

МИРЭА
Российский технологический университет

1. Что такое регрессия?
2. В чем отличие множественной регрессии от парной?
3. Какие команды в R реализуют регрессию?
4. Как на основе регрессии измерить влияние отдельного фактора или их совокупное влияние?
5. Как сравнить модели регрессии?

Литература

1.Роберт И. Кабаков R в действии. Анализ и визуализация данных в программе R / пер. с англ. Полины А. Волковой. – М.: ДМК Пресс, 2014. – 588 с.: ил. ISBN 978-5-947060-077-1

МИРЭА
Российский технологический университет