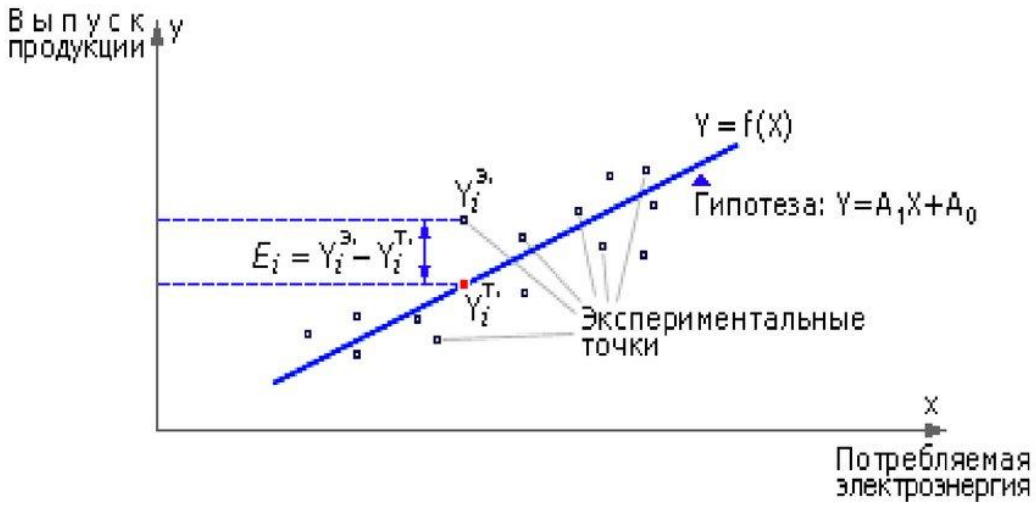
Основы регрессионного анализа в R

Регрессионный анализ применяется для нахождения линейной зависимости двух и более переменных. Допустим, имеются 2 переменные Y и X. Допустим также, что Y зависит от X. В таком случает переменную Y будем называть зависимой переменной, а переменную X - предиктором.

Регрессионный анализ позволяет построить линейную зависимость между переменными, в случае 2-х переменных – это будет прямая линия, уравнение которой представлено выше.

Прямая подбирается таким образом, чтобы наиболее точно аппроксимировать зависимость. Один из самых популярных способ аппроксимации – метод наименьших квадратов.

РЕГРЕССИОННЫЙ АНАЛИЗ



$$Q = \sum_{i=1}^{N} (\hat{y}_i - y_i)^2 \to \min .$$

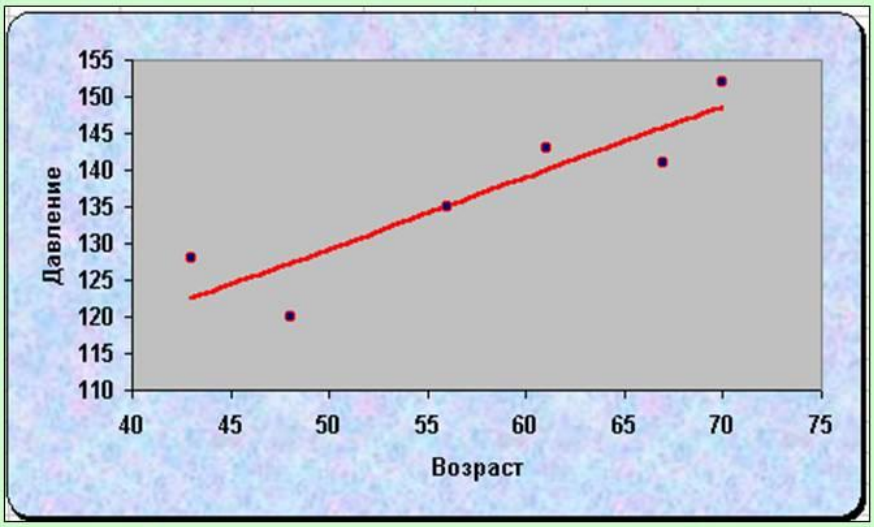
Допустим мы представили зависимость в виде синей прямой , как на рисунке выше. Почему у этой прямой именно такой наклон? Почему она расположена именно таким образом? Именно метод наименьших квадратов позволяет это определить. Каким образом он работает:

Подсчитывается сумма квадратов расстояний каждой точки до теоретической прямой (чем больше точка удалена от прямой тем больше расстояние). Наша цель – добиться чтобы сумма квадратов расстояний Q была минимальной. То есть метод наименьших квадратов перебирает все возможные варианты и ищет то положение прямой , которая наименее удалена от всех точек. Где мерой удаления считается сумма квадратов расстояний (или как их еще называют – сумма квадратов остатков).

Задача регрессионного анализа – найти уравнение прямой, значения коэффициентов прямой.

Регрессионный анализ
Парная линейная регрессия

Пример 6. Построить уравнение линейной регрессии для зависимости величин возраста и давления, приведенных в примере 1.

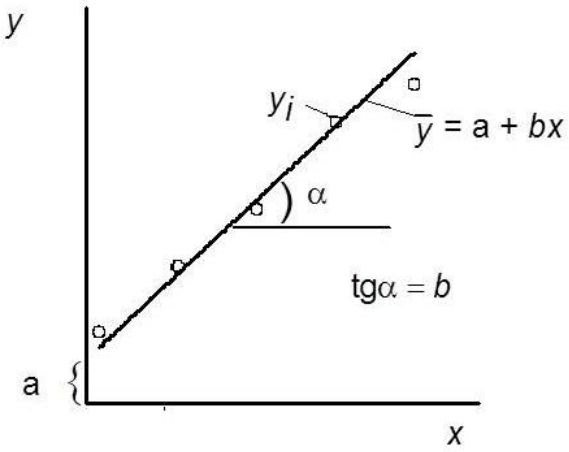


$$\hat{y} = 81{,}048 + 0{,}964\ x$$

18

В случае 2-х переменных – будут 2 коэффициента a и b, наша задача их найти:

Рассмотрим простую линейную модель: $y = a + bx$.



В R регрессионную линейную модель строит функция ***lm()*** .

Примеры использования функции, список аргументов можно увидеть в справке **help('lm')**.

В качестве аргументов указывается зависимая переменная, и список независимых:

model = lm (y ~ x)

Символ ~ разделяет зависимую и независимые переменные.

В случае, если предикторов несколько, выражение будет выглядеть следующим образом:

model = lm (y ~ x + y + z)

Рассмотрим пример, построим зависимость длины листа от ширины листа цветка Ирис, встроенный датафрейм **iris** , построим модель и выведем ее коэффициенты:

```
> model=lm(Sepal.Length ~ Sepal.Width, data=iris)
> model

Call:
lm(formula = Sepal.Length ~ Sepal.Width, data = iris)

Coefficients:
(Intercept)  Sepal.Width
     6.5262      -0.2234
```

Intercept – тот самый нулевой член уравнения прямой, второй коэффициент - тот что расположен при предикторе «Ширина лепестка».

Уравнение будет выглядеть следующим образом:

Длина_лепестка = 6.53 – 0.22 * Ширина_лепестка

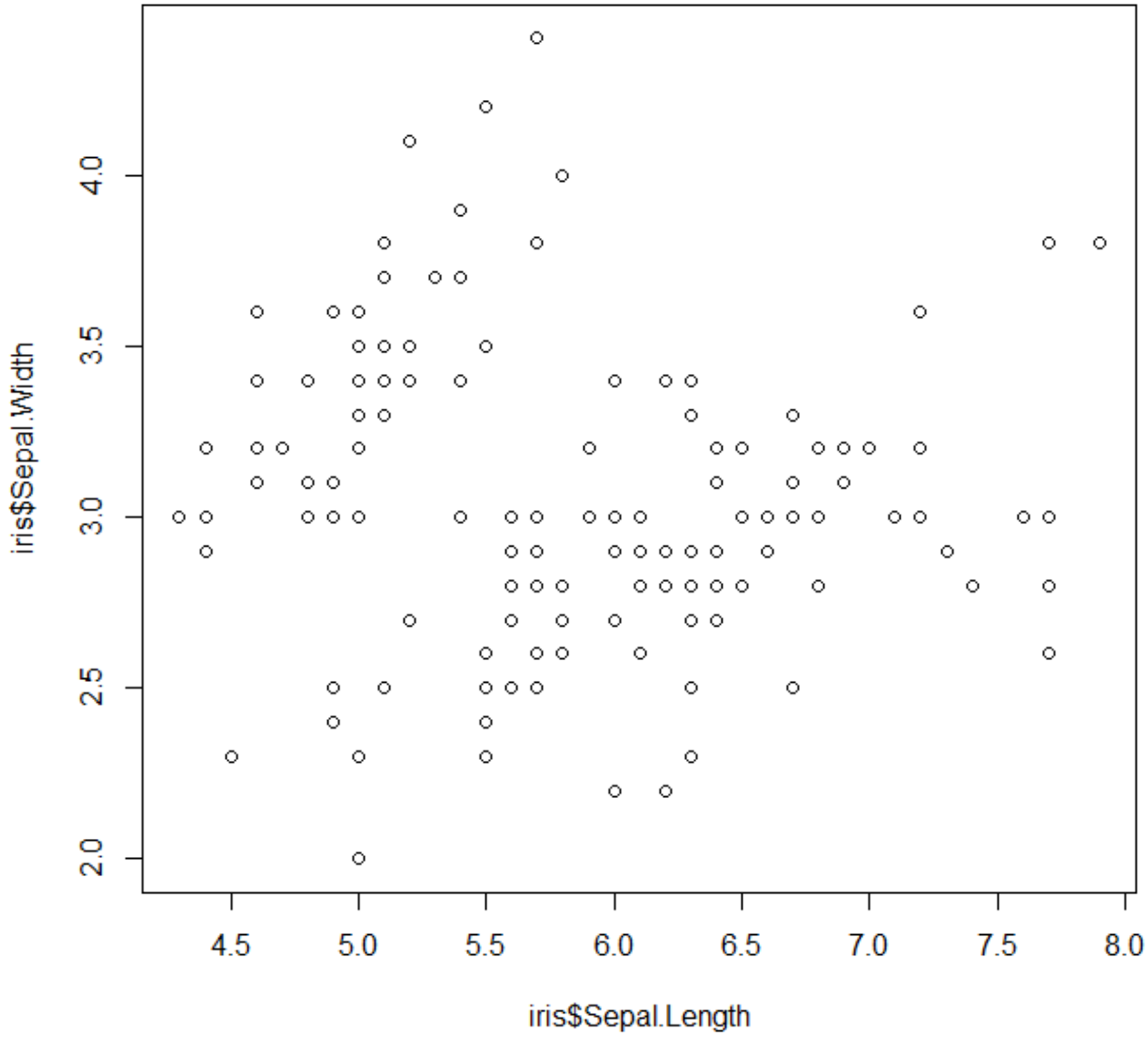
Для вывода подробной статистики используется команда **summary().**

```
> summary(model)

Call:
lm(formula = Sepal.Length ~ Sepal.Width, data = iris)

Residuals:
    Min      1Q  Median      3Q     Max
-1.5561 -0.6333 -0.1120  0.5579  2.2226

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)    6.5262     0.4789   13.63   <2e-16 ***
Sepal.Width   -0.2234     0.1551   -1.44    0.152
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.8251 on 148 degrees of freedom
Multiple R-squared:  0.01382,   Adjusted R-squared:  0.007159
F-statistic: 2.074 on 1 and 148 DF,  p-value: 0.1519
```

Здесь так же указаны 2 коэффициента: Intercept и Sepal.Width. Но что еще показывает данная таблица: Во-первых – как понять, насколько точно оценивает наша модель текущий набор точек. В первую очередь необходимо посмотреть на коэффициент R-squared. Он изменяется от 0 до 1 (от 0% до 100%). Он показывает какой процент изменчивости переменных объясняет наша модель. В данном случае он практически равен нулю – это повод задуматься, что то-то не так.

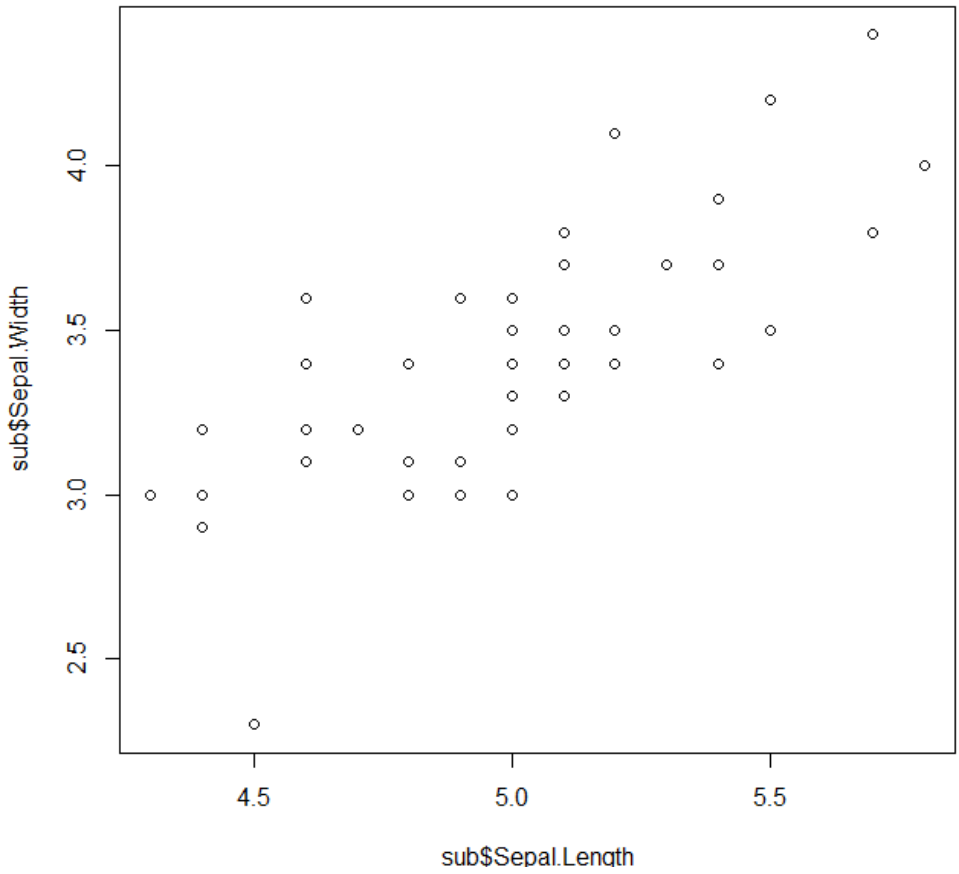
Во вторых необходимо взглянуть на величину p-value у каждого коэффициента. Если она меньше 0.05(5%), тогда значение коэффициента статистически значимо, если больше, значит данный предиктор не влияет на зависимую переменную. В нашем случае мы видим, что p-value для предиктора слишком большой (0.152)/. Второй повод задуматься, что то-то не так.

Давайте построим график зависимости длины лепестка от ширины:



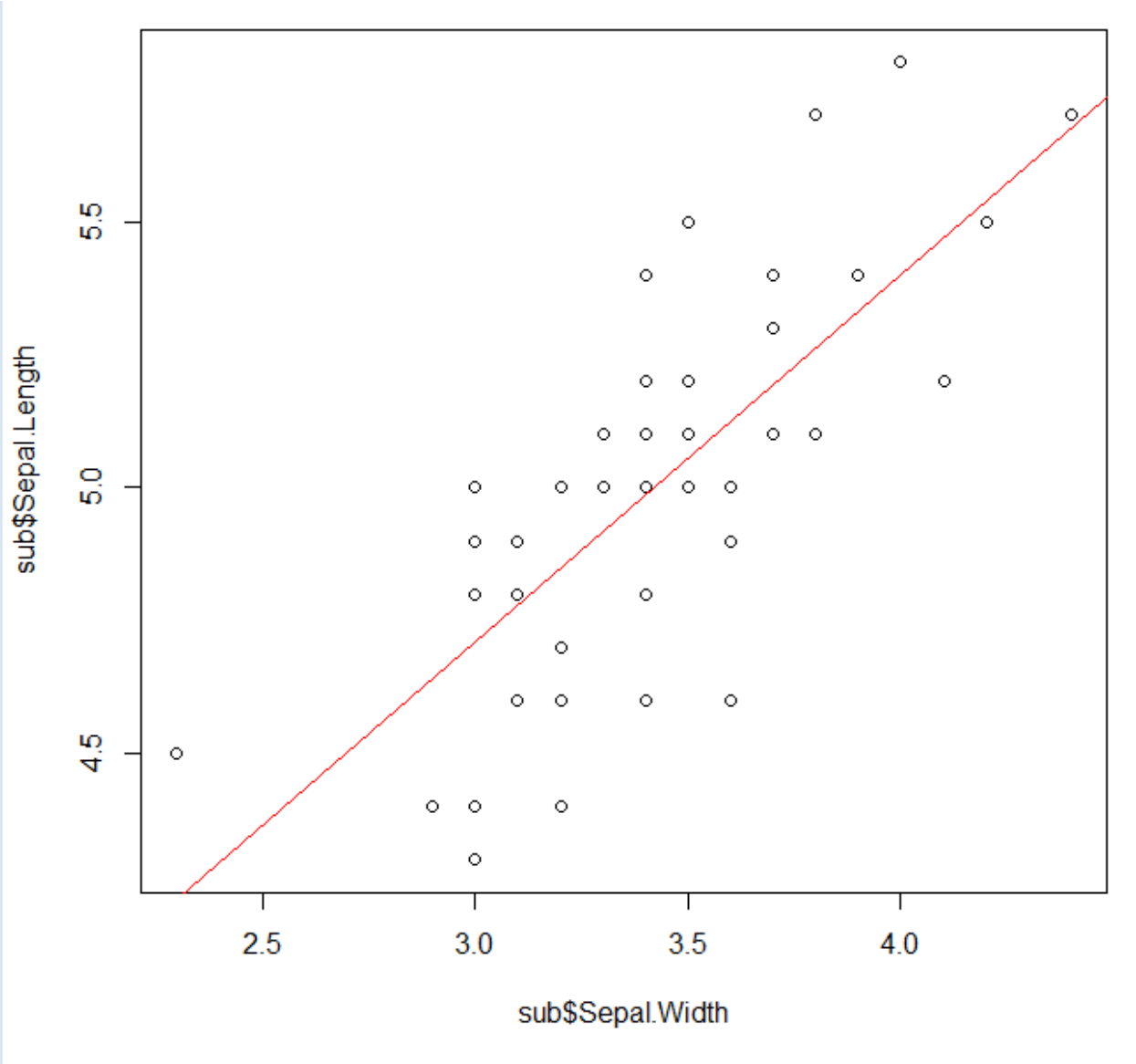
И вот теперь все встает на свои места. Данные действительно имеют линейную зависимость, но они сгруппированы по двух кластерам , как оказалось, в таблице 3 различных сорта цветка, поэтому данные таким образом разняться. Поэтому следует отфильтровать таблицу с данными и построить зависимости для каждого сорта растения, которые очевидно будут линейны.

Вот пример графика зависимости только для 1 из сортов (setosa):




```
Call:
lm(formula = Sepal.Length ~ Sepal.Width, data = sub)

Residuals:
     Min       1Q   Median       3Q      Max
-0.52476 -0.16286  0.02166  0.13833  0.44428

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   2.6390     0.3100   8.513 3.74e-11 ***
Sepal.Width   0.6905     0.0899   7.681 6.71e-10 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.2385 on 48 degrees of freedom
Multiple R-squared:  0.5514,    Adjusted R-squared:  0.542
F-statistic: 58.99 on 1 and 48 DF,  p-value: 6.71e-10
```

Теперь оба коэффициента статистически значимы, R – коэффициент равен 0.54, что гораздо лучше. И построим нашу прямую:



Данная прямая наиболее точно описывает нашу модель согласно методу наименьших квадратов.

Примечание: Если предикторов несколько, важно чтобы они были линейно независимы друг от друга в отдельности, поскольку линейно-зависимые предикторы могу неожиданно существенно исказить результаты модели. Один из таких предикторов следует удалить из модели, оставив лишь 1.

Например мы хотим построить зависимость веса ребенка до 18 лет от его роста и возраста. Зависимой переменной будет вес, предикторы: рост и возраст. Но очевидно , что чем старше ребенок тем он выше, предикторы между собой коррелируют, это можно увидеть построив их график, либо посчитав корреляционную функцию. Поэтому 1 из них нужно удалит из

модели, например удалим возраст. Тогда модель будет записана следующим образом:

$$\textbf{вес} = a_0 + a_1 * \textbf{рост}$$

Как уже было сказано, проверять зависимость предикторов можно как с помощью корреляционной функции, так и визуально графически. Кроме того можно построить линейную модели зависимости двух предикторов которая наиболее точно ее отобразит и опишет.