Regression Analysis: Telling Relationships

Abram Hindle

2012-02-14 Tue

Outline

What is regression analysis?

- relating dependant and independent variables
- relating observations and to inputs
- inferring structure, causality and functions

Dependant and Independent

- $y = b + ax$ where (a and b are constants
  - y is the response (measured)
  - x is the input or variable under scrutiny
- Independent variables are variables that we control or are controlling for (like x)
- Dependant variables are variables who's outcome might be dependant on other variables (like our independent variable) (like y)
- Essentially we want to see the response in y that is caused by x
  - This might only be a correlation
    - Remember correlation != causation

Linear Regression

- We want to investigate the linear effect of independent variables on a dependent variable using a model that is a linear combination of independent variables.
- The model:
  - $y = b + a_1 x_1 + a_2 x_2 + ... + a_n x_n$
    - Does anyone smell a matrix?
    - Essentially you're trying to solve an inexact linear system with the least amount of error

R code

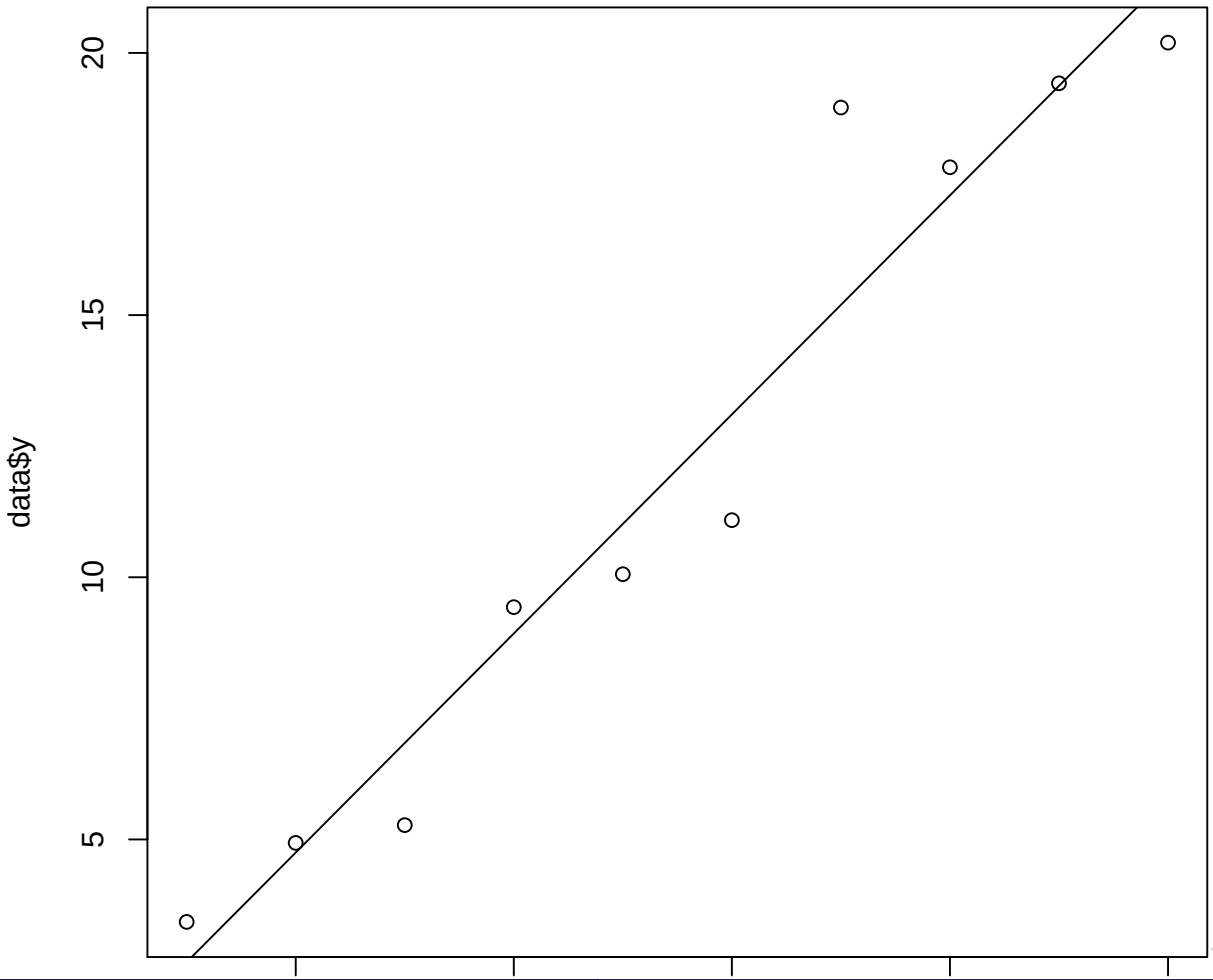
```
data <- c()
data$x <- c(1:10)
data$y <- runif(10) + 2*(data$x + rnorm(10))
lmfit <- lm( y ~ x , data = data)
plot(data)
lines(lmfit$fitted.values)
summary(lmfit)
```

The output

```
> summary(lmfit)
Call:
lm(formula = y ~ x, data = data)
Residuals:
     Min       1Q   Median       3Q      Max
-2.50315 -0.75578  0.04244  1.10641  1.81079
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   1.7481     0.9841   1.776    0.114
x             1.7450     0.1586  11.003 4.14e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1
Residual standard error: 1.441 on 8 degrees of freedom
Multiple R-squared: 0.938,      Adjusted R-squared: 0.930
F-statistic: 121.1 on 1 and 8 DF,  p-value: 4.14e-06
```

Fitted Line

- The data + model



R^2

- Explained variance
  - "How much variance is explained by the model"
  - 1.0 completely explained
  - 0.5 50% explained
  - 0.1 10% explained
- You want adjusted R-Squared
- It is a comparison of error to the result
  - Not available by all tests, and not always meaningful

AIC - Akaike information criterion

- Goodness of fit test
- Lower is better
- `AIC` in R
- Information theoretic suggestion of what is lost by the model

VIF - Variance Inflation Factor

- Goodness of fit test
- How much variance is caused by colinearity
  - If values are correlated they can be colinear.
- in `library(car)` in R
- `vif` in R

Other uses for regression

- Estimate functions
- Explore the relationships between variables
    - Ignore the results and focus on the stability of the relationship by watching how stable a variable is in multiple models.