Introduction to Geostatistics

10. Correlation and regression

Edzer Pebesma

edzer.pebesma@uni-muenster.de
Institute for Geoinformatics (**ifgi**)
University of Münster

June 22, 2010

Correlation and regression

t-tests and analysis of variance look at how a single *continuous* variable depends on a single *categorical* variable with two levels (t-test), more levels (one-way anova), or on more than one categorical variable (two-way, more-way anova).
The focus now shifts to the relation between two (or more) continuous variables. We start with the relationship between two continuous variables, and how one continuous variable depends on another dependent variable.

sample and population correlation

We can compute sample correlation,

```
> cor(Length, Weight, use = "complete.obs")

[1] 0.6818044
```

but also test whether the population correlation (ρ) has a certain value. Typically, $H_0 : \rho = 0$.

```
> cor.test(Length, Weight)

        Pearson's product-moment correlation

data:  Length and Weight
t = 11.223, df = 145, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.5844191 0.7598282
sample estimates:
      cor
0.6818044
```

correlation: symmetry

As can be glanced from the equation how to compute correlation,

$$r(X, Y) = \frac{\text{Cov}(X, Y)}{\sqrt{\text{Var}(X)\text{Var}(Y)}}$$

it is true that $r(x, y) = r(y, x)$. Indeed,

```
> cor(Length, Weight, use = "complete.obs")

[1] 0.6818044

> cor(Weight, Length, use = "complete.obs")

[1] 0.6818044
```

Linear regression

Regression looks at asymmetric problems, where one variable depends on another. E.g. in simple linear regression, for n observations y_i, $i = 1, \ldots, n$:

$$y_i = \beta_0 + \beta_1 x_i + e_i$$

with e a zero-mean random variable, β_0 and β_1 unknown but non-random population parameters, and X known. So,

$$\mathrm{E}(y_i) = \beta_0 + \beta_1 x_i$$

As e is random, it means that y is random as well, whereas x is not.

A test the regression slope

The typical problem in looking at linear relationships between two continuous variables, is to ask oneself *whether* one variable *depends* on the other. Dependence is a rather broad concept, and can have many forms. We usually first look at whether one variable linearly depends on the other, as in

$$y_i = \beta_0 + \beta_1 x_i + e_i$$

If this dependence is not the case, then $\beta_1 = 0$. So, this is the typical H_0 for this kind of test.

How to estimate the parameters?

Under the assumptions that

(i) the observations are independent (and consequently the e_i are independent) and

(ii) that the variance of e_i is constant,

the best estimates for β_0 and β_1 are obtained by minimizing the sum of squared regression residuals, $\sum_{i=1}^{n} e_i^2$: and are

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

Regression output from R – 1

```
> lm(Weight ~ Length)

Call:
lm(formula = Weight ~ Length)

Coefficients:
(Intercept)       Length
   -120.311        1.073
```

The intercept refers to the value of y when x is zero, the value called `Length` to the regression coefficient that belongs to variable `Length`. Thus, the equation for the regression line is:

$$\mathrm{E}(\texttt{Weight}) = -120.311 + 1.073 \times \texttt{Length}$$

Under the *additional* assumptions of normaly distributed residuals:

Regression output from R – 2

```
> summary(lm(Weight ~ Length))

Call:
lm(formula = Weight ~ Length)

Residuals:
    Min      1Q  Median      3Q     Max 
-18.054  -6.950  -2.297   3.369  84.350 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept) -120.31118   17.06402  -7.051 6.72e-11 ***
Length         1.07255    0.09557  11.223  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 12.59 on 145 degrees of freedom
  (2 observations deleted due to missingness)
Multiple R-squared: 0.4649,     Adjusted R-squared: 0.4612 
F-statistic:   126 on 1 and 145 DF,  p-value: < 2.2e-16
```

ifgi

A model for the data

For each data point y_i, we can decompose the difference from the mean of y, $\bar{y}$ as

$$y_i - \bar{y} = (y_i - \hat{y}) + (\hat{y} - \bar{y})$$

As the two right-hand side terms are independent, we can write this as

$$(y_i - \bar{y})^2 = (y_i - \hat{y})^2 + (\hat{y} - \bar{y})^2$$

and summed over all measurements:

$$SS_{tot} = SS_{resid} + SS_{reg}$$

```
> summary(aov(Weight ~ Length))

             Df Sum Sq Mean Sq F value    Pr(>F)    
Length        1  19956 19956.0  125.96 < 2.2e-16 ***
Residuals   145  22973   158.4                      
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
2 observations deleted due to missingness
```

ifgi

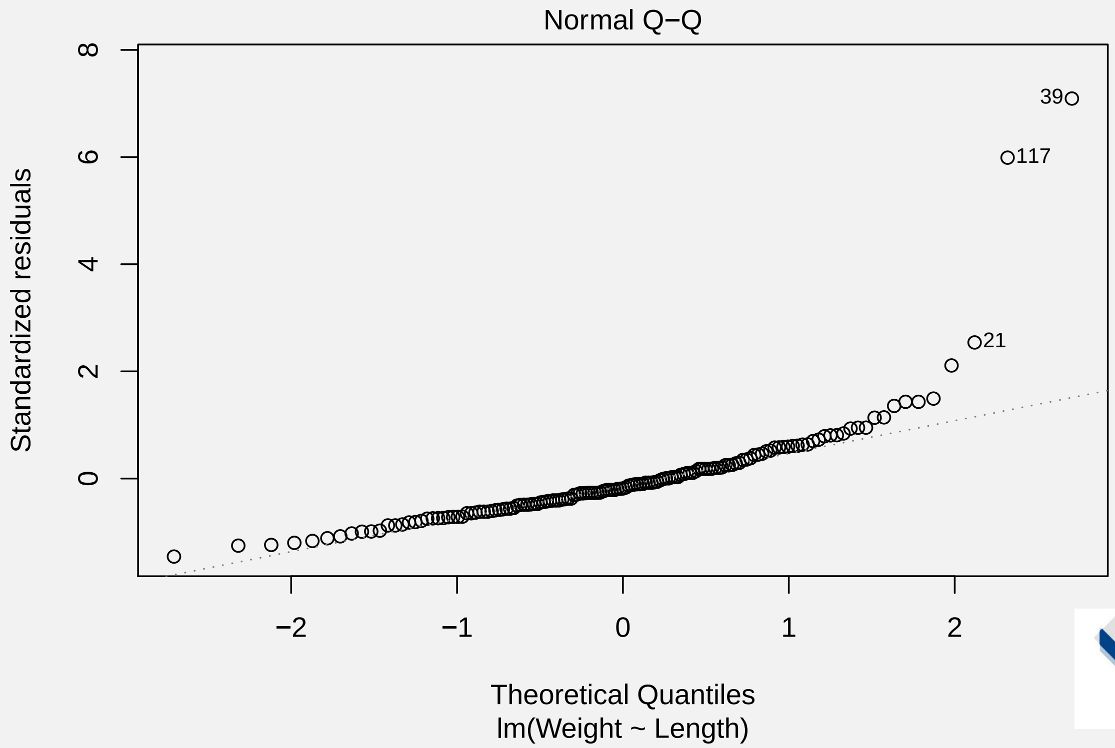
- `Residual standard error: 12.59`: this is the square-root of `MS Residuals` (158.4)
- `on 145 degrees of freedom`: $n-2$ (two coefficients were estimated: β_0 and β_1, to obtain residuals)
- `Multiple R-squared: 0.4649` this is SS_{reg}/SS_{tot}, a measure between 0 and 1, where 1 indicates a perfect fit, 0 absence of fit
- `Adjusted R-squared: 0.4512` (next week)
- `F-statistic: 126 on 1 and 145 DF` the ratio of the mean squares (MS_{reg}/MS_{resid})
- `p-value: < 2.2e-16` the p-value of the test for the slope, on $H_0 : \beta_1 = 0$

Diagnostic plots, 1

```
> plot(lm(Weight ~ Length), which = 1)
```

Diagnostic plots, 2

```
> plot(lm(Weight ~ Length), which = 2)
```



Diagnostic plots, 3

```
> plot(lm(Weight ~ Length), which = 3)
```