

Introduction to Geostatistics

10. Correlation and regression

Edzer J. Pebesma

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

summer semester 2007/8,
June 12, 2008

Correlation and regression

t-tests and analysis of variance look at how a single continuous variable depends on a categorical variable with two levels (t-test), more levels (one-way anova), or on more categorical variables (more-way anova).

The focus now shifts to the relation between two (or more) **continuous** variables. We start with how one continuous variable depends on another dependent variable.

sample and population correlation

We can compute sample correlation,

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

```
[1] 0.6797413
```

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 = 8.494, df = 84, p-value = 6.19e-13
```

```
alternative hypothesis: true correlation is not equal to 0
```

```
95 percent confidence interval:
```

```
0.5465855 0.7793713
```

```
sample estimates:
```

```
cor
```

```
0.6797413
```

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)
```

```
[1] 0.6797413
```

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

```
[1] 0.6797413
```

Linear regression

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

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

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

$$E(y_i) = \beta_0 + \beta_1 x_i$$

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

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2}$$
$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

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

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2}$$
$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

Let's do it (with R) – 1

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

Call:

```
lm(formula = Weight ~ Length)
```

Coefficients:

| (Intercept) | Length |
|-------------|--------|
| -145.998 | 1.222 |

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:

$$E(\text{Weight}) = -145.998 + 1.222 \text{ Length}$$

```
>
```

Let's do it (with R) – 2

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

```
Call:
```

```
lm(formula = Weight ~ Length)
```

```
Residuals:
```

| Min | 1Q | Median | 3Q | Max |
|---------|--------|--------|-------|--------|
| -18.555 | -9.143 | -2.674 | 5.003 | 77.239 |

```
Coefficients:
```

| | Estimate | Std. Error | t value | Pr(> t) |
|-------------|-----------|------------|---------|--------------|
| (Intercept) | -145.9984 | 25.6926 | -5.683 | 1.87e-07 *** |
| Length | 1.2216 | 0.1438 | 8.494 | 6.19e-13 *** |

```
---
```

```
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
Residual standard error: 15.38 on 84 degrees of freedom
```

```
Multiple R-squared: 0.462, Adjusted R-squared: 0.4556
```

```
F-statistic: 72.15 on 1 and 84 DF, p-value: 6.19e-13
```

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 | 17071.2 | 17071.2 | 72.148 | 6.19e-13 *** |
| Residuals | 84 | 19875.6 | 236.6 | | |

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

- ▶ Residual standard error: 15.38: this is the square-root of MS Residuals (236.6)
- ▶ on 84 degrees of freedom: $n - 2$ (two coefficients were estimated: β_0 and β_1 , to obtain residuals)
- ▶ Multiple R-squared: 0.462 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: 0.4556 forget for now
- ▶ F-statistic: 72.15 on 1 and 84 DF the ration of the mean squares (MS_{reg}/MS_{resid})
- ▶ p-value: 6.19e-1 the p-value of the test for the slope, on $H_0: \beta_1 = 0$

- ▶ Residual standard error: 15.38: this is the square-root of MS Residuals (236.6)
- ▶ on 84 degrees of freedom: $n - 2$ (two coefficients were estimated: β_0 and β_1 , to obtain residuals)
- ▶ Multiple R-squared: 0.462 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: 0.4556 forget for now
- ▶ F-statistic: 72.15 on 1 and 84 DF the ration of the mean squares (MS_{reg}/MS_{resid})
- ▶ p-value: 6.19e-1 the p-value of the test for the slope, on $H_0 : \beta_1 = 0$

- ▶ Residual standard error: 15.38: this is the square-root of MS Residuals (236.6)
- ▶ on 84 degrees of freedom: $n - 2$ (two coefficients were estimated: β_0 and β_1 , to obtain residuals)
- ▶ Multiple R-squared: 0.462 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: 0.4556 forget for now
- ▶ F-statistic: 72.15 on 1 and 84 DF the ration of the mean squares (MS_{reg}/MS_{resid})
- ▶ p-value: 6.19e-1 the p-value of the test for the slope, on $H_0: \beta_1 = 0$

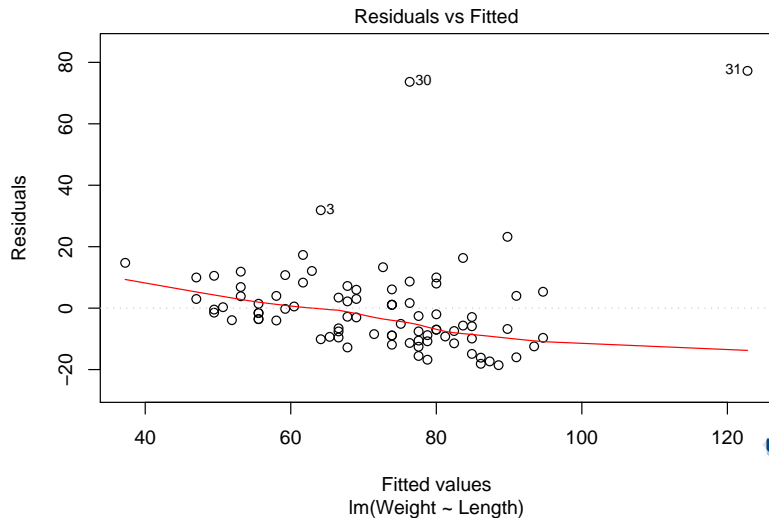
- ▶ Residual standard error: 15.38: this is the square-root of MS Residuals (236.6)
- ▶ on 84 degrees of freedom: $n - 2$ (two coefficients were estimated: β_0 and β_1 , to obtain residuals)
- ▶ Multiple R-squared: 0.462 this is SS_{reg}/SS_{tot} , a measure between 0 and 1, where 1 indicates a perfect fit, 0 absence of fit
- ▶ Adjusted R-squared: 0.4556 forget for now
- ▶ F-statistic: 72.15 on 1 and 84 DF the ration of the mean squares (MS_{reg}/MS_{resid})
- ▶ p-value: 6.19e-1 the p-value of the test for the slope, on $H_0: \beta_1 = 0$

- ▶ Residual standard error: 15.38: this is the square-root of MS Residuals (236.6)
- ▶ on 84 degrees of freedom: $n - 2$ (two coefficients were estimated: β_0 and β_1 , to obtain residuals)
- ▶ Multiple R-squared: 0.462 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: 0.4556 forget for now
- ▶ F-statistic: 72.15 on 1 and 84 DF the ration of the mean squares (MS_{reg}/MS_{resid})
- ▶ p-value: 6.19e-1 the p-value of the test for the slope, on $H_0: \beta_1 = 0$

- ▶ Residual standard error: 15.38: this is the square-root of MS Residuals (236.6)
- ▶ on 84 degrees of freedom: $n - 2$ (two coefficients were estimated: β_0 and β_1 , to obtain residuals)
- ▶ Multiple R-squared: 0.462 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: 0.4556 forget for now
- ▶ F-statistic: 72.15 on 1 and 84 DF the ration of the mean squares (MS_{reg}/MS_{resid})
- ▶ p-value: 6.19e-1 the p-value of the test for the slope, on $H_0: \beta_1 = 0$

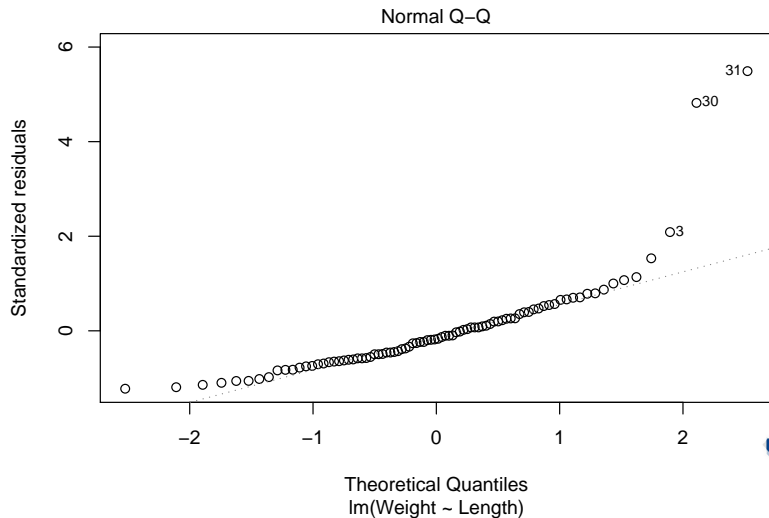
Diagnostic plots, 1

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



Diagnostic plots, 2

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



Diagnostic plots, 3

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

