Introduction to Geostatistics

11. Multiple regression, regression extensions

Edzer J. Pebesma

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

summer semester 2007/8,
June 22, 2010

The multiple linear regression model

The multiple linear regression model extends the simple regression model with one single predictor

$$y_i = \beta_0 + \beta_1 X_{i,1} + e_i$$

to two predictors

$$y_i = \beta_0 + \beta_1 X_{i,1} + \beta_2 X_{i,2} + e_i$$

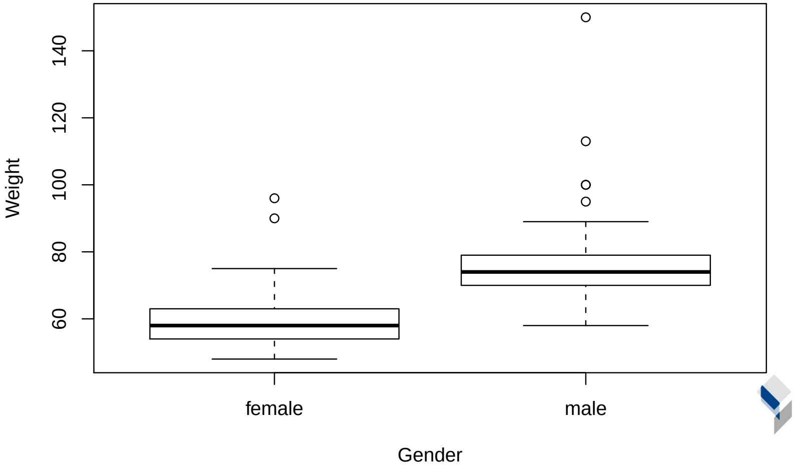
or p predictors:

$$y_i = \beta_0 + \beta_1 X_{i,1} + ... + \beta_p X_{i,p} + e_i$$

Example: two groups

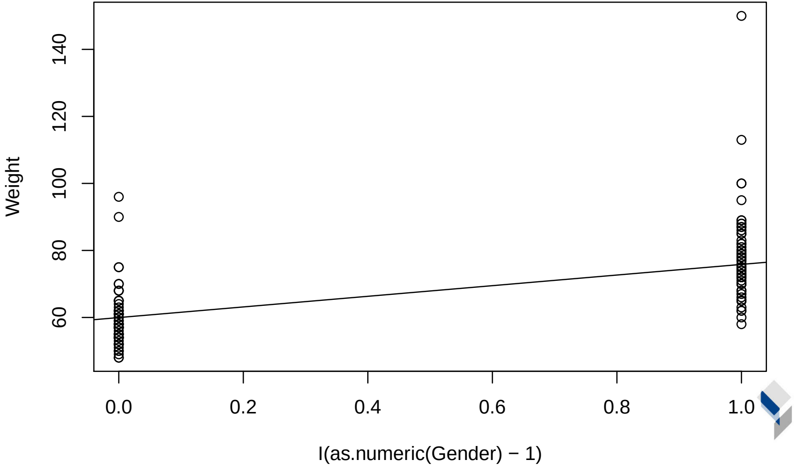
(Ignoring the outlier)

```
> plot(Weight ~ Gender)
```



Example: ... seen through a linear regression glasses

```
> plot(Weight ~ I(as.numeric(Gender) - 1))
> abline(lm(Weight ~ Gender))
```



Example: simple

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

Call:
lm(formula = Weight ~ Gender)

Residuals:
    Min      1Q  Median      3Q     Max 
-17.826  -5.826  -1.895   3.174  74.174 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)   59.963      1.503  39.890  < 2e-16 ***
Gendermale    15.863      1.894   8.377 4.42e-14 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1 

Residual standard error: 11.05 on 144 degrees of freedom
  (2 observations deleted due to missingness)
Multiple R-squared: 0.3276,	Adjusted R-squared: 0.323 
F-statistic: 70.17 on 1 and 144 DF,  p-value: 4.420e-14 
```

Interpretation

So, weight depends on Gender.
But, there's also a length effect. Longer people are usually heavier, and men are usually taller than women.
Questions we could ask:

1. is there, besides a Length effect still an effect of Gender on Weight? (testing)
2. how large is the effect of Length on Weight? (estimation)
3. Does this effect depend on Gender? (testing)

Interpretation

So, weight depends on Gender.
But, there's also a length effect. Longer people are usually heavier, and men are usually taller than women.
Questions we could ask:

1. is there, besides a Length effect still an effect of Gender on Weight? (testing)
2. how large is the effect of Length on Weight? (estimation)
3. Does this effect depend on Gender? (testing)

Interpretation

So, weight depends on Gender.
But, there's also a length effect. Longer people are usually heavier, and men are usually taller than women.
Questions we could ask:

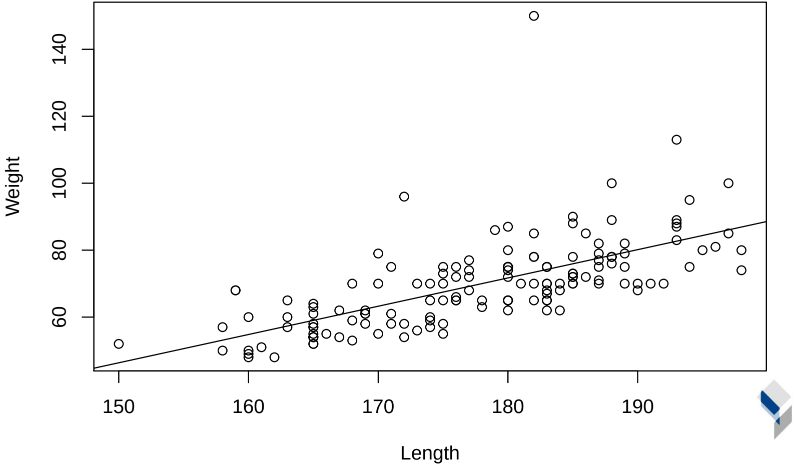
1. is there, besides a Length effect still an effect of Gender on Weight? (testing)
2. how large is the effect of Length on Weight? (estimation)
3. Does this effect depend on Gender? (testing)

Example: simple

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

Call:
lm(formula = Weight ~ Length)

Residuals:
    Min      1Q  Median      3Q     Max
-13.081  -6.521  -2.174   3.952  76.609

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -80.39563   14.57031  -5.518 1.55e-07 ***
Length        0.84498    0.08175  10.337  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 10.21 on 144 degrees of freedom
  (2 observations deleted due to missingness)
Multiple R-squared: 0.4259,        Adjusted R-squared: 0.422
F-statistic: 106.8 on 1 and 144 DF,  p-value: < 2.2e-16
```

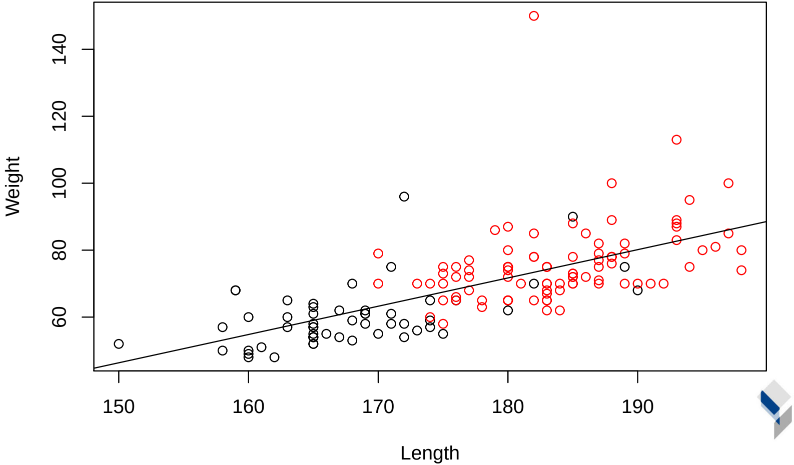
Example: simple

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

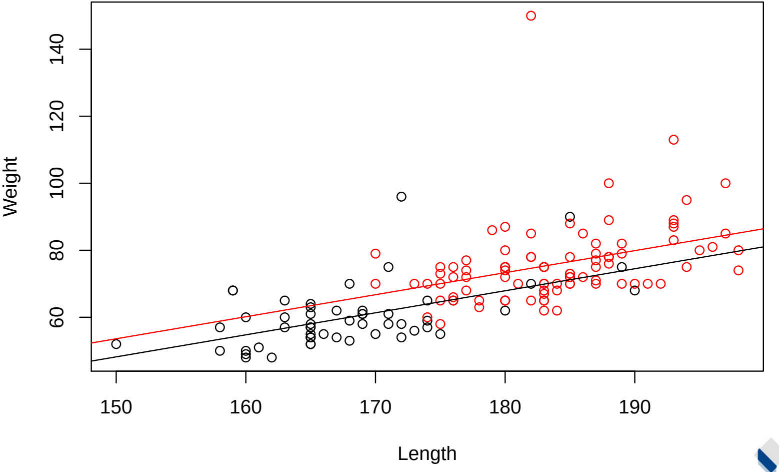


Example: simple

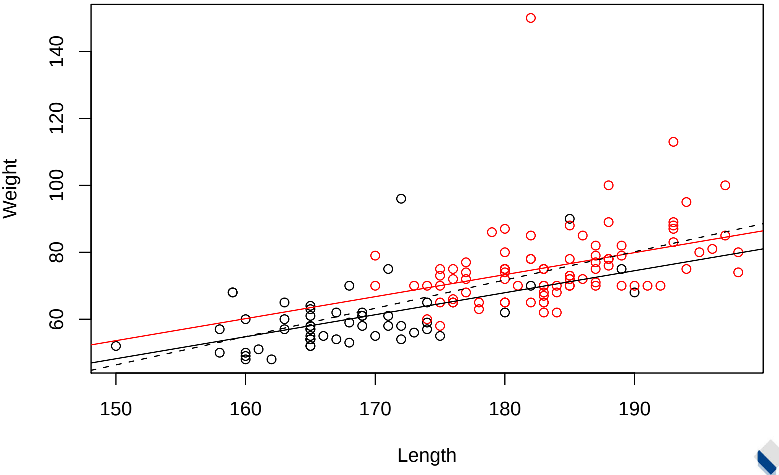
```
> plot(Weight ~ Length, col = Gender)
> abline(lm(Weight ~ Length))
```



Example: the two parallel lines



Example: the two parallel lines added



Example: corresponding model

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

Call:
lm(formula = Weight ~ Length + Gender)

Residuals:
    Min      1Q  Median      3Q     Max 
-13.940  -5.934  -1.010   2.970  75.374 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) -50.4297    20.3431  -2.479   0.0143 *  
Length        0.6575     0.1209   5.439 2.26e-07 ***
Gendermale    5.3972     2.5874   2.086   0.0388 *  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1 

Residual standard error: 10.09 on 143 degrees of freedom
  (2 observations deleted due to missingness)
Multiple R-squared: 0.4429,        Adjusted R-squared: 0.4351 
F-statistic: 56.84 on 2 and 143 DF,  p-value: < 2.2e-16 
```

3 Questions

1. is there, besides a Length effect still an effect of Gender on Weight? No, it is not significant; it can be there, but based on our data we cannot say whether it is positive or negative
2. how large is the effect of Length on Weight? Is it 0.88 or 0.71? Despite the fact that gender is not significant, assuming H_0 that the effect is zero is not very realistic. We may therefor give a preference to the 0.71 estimate .
3. Does this effect depend on Gender? See next slide.

3 Questions

1. is there, besides a Length effect still an effect of Gender on Weight? No, it is not significant; it can be there, but based on our data we cannot say whether it is positive or negative
2. how large is the effect of Length on Weight? Is it 0.88 or 0.71? Despite the fact that gender is not significant, assuming H_0 that the effect is zero is not very realistic. We may therefor give a preference to the 0.71 estimate .
3. Does this effect depend on Gender? See next slide.

3 Questions

1. is there, besides a Length effect still an effect of Gender on Weight? No, it is not significant; it can be there, but based on our data we cannot say whether it is positive or negative
2. how large is the effect of Length on Weight? Is it 0.88 or 0.71? Despite the fact that gender is not significant, assuming H_0 that the effect is zero is not very realistic. We may therefor give a preference to the 0.71 estimate .
3. Does this effect depend on Gender? See next slide.

Does the effect depend on Gender?

- Both models (simple linear, and multiple linear) give a *single* dependence (slope of the line) for Weight on Length.
- The question whether this effect (the slope) depends on Gender, is the following: does the slope (Weight ~ Length) differ for male persons from that of female persons?

Let $X_{i,1}$ be Length, and let $X_{i,2}$ be zero for female, and one for male persons. Then

$$y_i = \beta_0 + \beta_1 X_{i,1} + \beta_2 X_{i,2} + \beta_3 X_{i,1} X_{i,2} + e_i$$

is a single regression model that reduces for female persons to

$$y_i = \beta_0 + \beta_1 X_{i,1} + e_i$$

and for male persons to

$$y_i = (\beta_0 + \beta_2) + (\beta_1 + \beta_3) X_{i,1} + e_i$$

so, we have two completely free regression lines, each with a unique slope and intercept.

Does the effect depend on Gender?

- Both models (simple linear, and multiple linear) give a *single* dependence (slope of the line) for Weight on Length.
- The question whether this effect (the slope) depends on Gender, is the following: does the slope (Weight ~ Length) differ for male persons from that of female persons?

Let $X_{i,1}$ be Length, and let $X_{i,2}$ be zero for female, and one for male persons. Then

$$y_i = \beta_0 + \beta_1 X_{i,1} + \beta_2 X_{i,2} + \beta_3 X_{i,1} X_{i,2} + e_i$$

is a single regression model that reduces for female persons to

$$y_i = \beta_0 + \beta_1 X_{i,1} + e_i$$

and for male persons to

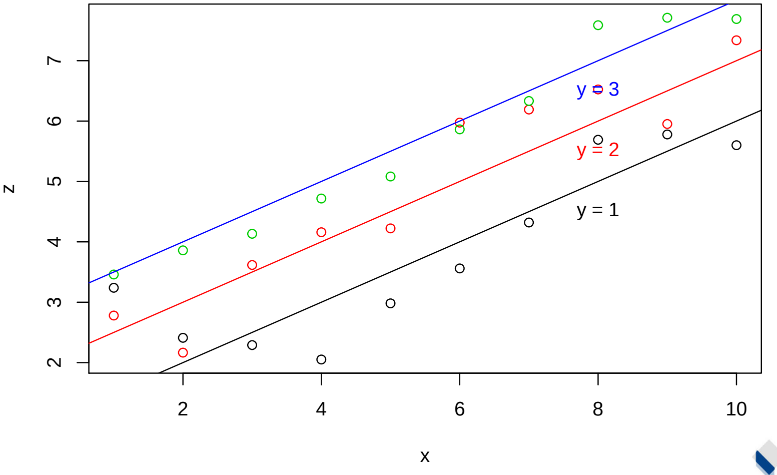
$$y_i = (\beta_0 + \beta_2) + (\beta_1 + \beta_3) X_{i,1} + e_i$$

so, we have two completely free regression lines, each with a unique slope and intercept.

The R model

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

Call:
lm(formula = Weight ~ Length * Gender)

Residuals:
    Min      1Q  Median      3Q     Max
-13.944  -5.928  -1.002   3.063  75.415

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)       -45.80442   30.57002  -1.498 0.136264
Length              0.62991    0.18188   3.463 0.000705 ***
Gendermale         -3.28331   42.78824  -0.077 0.938943
Length:Gendermale   0.04961    0.24407   0.203 0.839234
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 10.12 on 142 degrees of freedom
  (2 observations deleted due to missingness)
Multiple R-squared: 0.4431,        Adjusted R-squared: 0.4313
F-statistic: 37.65 on 3 and 142 DF,  p-value: < 2.2e-16
```

Multiple linear regression with two variables.



```
> summary(lm(z ~ x))

Call:
lm(formula = z ~ x)

Residuals:
    Min      1Q  Median      3Q     Max
-1.9759 -0.7131  0.2829  0.7043  1.5639

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.03122    0.38227   5.314 1.18e-05 ***
x            0.49912    0.06161   8.101 8.05e-09 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.9692 on 28 degrees of freedom
Multiple R-squared: 0.701,       Adjusted R-squared: 0.6903
F-statistic: 65.63 on 1 and 28 DF,  p-value: 8.05e-09
```

```
> summary(lm(z ~ x + y))

Call:
lm(formula = z ~ x + y)

Residuals:
     Min       1Q   Median       3Q      Max
-1.05017 -0.31838 -0.09206  0.29609  1.63476

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.17983    0.34740   0.518    0.609
x            0.49912    0.03703  13.477 1.67e-13 ***
y            0.92570    0.13028   7.106 1.22e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.5826 on 27 degrees of freedom
Multiple R-squared: 0.8958,        Adjusted R-squared: 0.8881
F-statistic: 116.1 on 2 and 27 DF,  p-value: 5.507e-14
```

Why using multiple regression?

1. There is a difference in interpretation for slopes
    - when (some of) the predictors X are correlated, the slopes differ from eachother.
    - the slope for the model $y = \beta_0 + \beta_1 X_1 + e$ is simply *the expected change in y as a function of X_1, ignoring everything else*
    - the slope for the model $y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + e$ is simply *the expected change in y as a function of X_1, everything else (meaning: X_2) held constant.*
    - in the first model, the slope may be partly due to X_2.
2. Their power is often larger (smaller residual standard error).

Why using multiple regression?

1. There is a difference in interpretation for slopes
   - when (some of) the predictors X are correlated, the slopes differ from eachother.
   - the slope for the model $y = \beta_0 + \beta_1 X_1 + e$ is simply *the expected change in y as a function of* X_1*, ignoring everything else*
   - the slope for the model $y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + e$ is simply *the expected change in y as a function of* X_1*, everything else (meaning:* X_2*) held constant.*
   - in the first model, the slope may be partly due to X_2.
2. Their power is often larger (smaller residual standard error).

Why using multiple regression?

1. There is a difference in interpretation for slopes
    - when (some of) the predictors X are correlated, the slopes differ from eachother.
    - the slope for the model $y = \beta_0 + \beta_1 X_1 + e$ is simply *the expected change in y as a function of X_1, ignoring everything else*
    - the slope for the model $y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + e$ is simply *the expected change in y as a function of X_1, everything else (meaning: X_2) held constant.*
    - in the first model, the slope may be partly due to X_2.
2. Their power is often larger (smaller residual standard error).

Why using multiple regression?

1. There is a difference in interpretation for slopes
    - when (some of) the predictors X are correlated, the slopes differ from eachother.
    - the slope for the model $y = \beta_0 + \beta_1 X_1 + e$ is simply *the expected change in y as a function of X_1, ignoring everything else*
    - the slope for the model $y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + e$ is simply *the expected change in y as a function of X_1, everything else (meaning: X_2) held constant.*
    - in the first model, the slope may be partly due to X_2.
2. Their power is often larger (smaller residual standard error).

Correlated errors

When observations are correlated, and cannot be considered independent (e.g. by the random sampling argument), regression can be applied under a more general model that addresses these correlations.

- the structure of the correlation needs to be assessed
  - correlation in space: a function of spatial distance?
  - correlation over time: a function of time separation?
  - within-item correlation: e.g. longitudinal studies.
- the magnitude of the correlations needs to be assessed

Correlated errors

When observations are correlated, and cannot be considered independent (e.g. by the random sampling argument), regression can be applied under a more general model that addresses these correlations.

- the structure of the correlation needs to be assessed
  - correlation in space: a function of spatial distance?
  - correlation over time: a function of time separation?
  - within-item correlation: e.g. longitudinal studies.
- the magnitude of the correlations needs to be assessed

Correlated errors

When observations are correlated, and cannot be considered independent (e.g. by the random sampling argument), regression can be applied under a more general model that addresses these correlations.

- the structure of the correlation needs to be assessed
  - correlation in space: a function of spatial distance?
  - correlation over time: a function of time separation?
  - within-item correlation: e.g. longitudinal studies.
- the magnitude of the correlations needs to be assessed

Correlated errors

When observations are correlated, and cannot be considered independent (e.g. by the random sampling argument), regression can be applied under a more general model that addresses these correlations.

- the structure of the correlation needs to be assessed
  - correlation in space: a function of spatial distance?
  - correlation over time: a function of time separation?
  - within-item correlation: e.g. longitudinal studies.
- the magnitude of the correlations needs to be assessed

Correlated errors

When observations are correlated, and cannot be considered independent (e.g. by the random sampling argument), regression can be applied under a more general model that addresses these correlations.

- the structure of the correlation needs to be assessed
  - correlation in space: a function of spatial distance?
  - correlation over time: a function of time separation?
  - within-item correlation: e.g. longitudinal studies.
- the magnitude of the correlations needs to be assessed

Generalized linear models

Generalized linear models extend the (multiple) linear regression models by

- not assuming a (free) continuous variable as dependent
- not assuming a Gaussian distribution for the residuals

Examples:

- logistic regression: dependent variable is 0/1 (absence/presence)
- log-linear models: dependent variable is a count (Poisson)
- regression on log-transforms: the logarithm of y is taken instead of y

These models are very common in ecology.

Generalized linear models

Generalized linear models extend the (multiple) linear regression models by

- not assuming a (free) continuous variable as dependent
- not assuming a Gaussian distribution for the residuals

Examples:

- logistic regression: dependent variable is 0/1 (absence/presence)
- log-linear models: dependent variable is a count (Poisson)
- regression on log-transforms: the logarithm of y is taken instead of y

These models are very common in ecology.

Generalized linear models

Generalized linear models extend the (multiple) linear regression models by

- not assuming a (free) continuous variable as dependent
- not assuming a Gaussian distribution for the residuals

Examples:

- logistic regression: dependent variable is 0/1 (absence/presence)
- log-linear models: dependent variable is a count (Poisson)
- regression on log-transforms: the logarithm of y is taken instead of y

These models are very common in ecology.

Generalized linear models

Generalized linear models extend the (multiple) linear regression models by

- not assuming a (free) continuous variable as dependent
- not assuming a Gaussian distribution for the residuals

Examples:

- logistic regression: dependent variable is 0/1 (absence/presence)
- log-linear models: dependent variable is a count (Poisson)
- regression on log-transforms: the logarithm of y is taken instead of y

These models are very common in ecology.

Generalized linear models

Generalized linear models extend the (multiple) linear regression models by

- not assuming a (free) continuous variable as dependent
- not assuming a Gaussian distribution for the residuals

Examples:

- logistic regression: dependent variable is 0/1 (absence/presence)
- log-linear models: dependent variable is a count (Poisson)
- regression on log-transforms: the logarithm of y is taken instead of y

These models are very common in ecology.

R-squared and adjusted R-squared

Coefficient of multiple correlation:

$$R^2 = \frac{\sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2}{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}$$

Adjusted R^2:

$$\bar{R}^2 = \frac{(n-1)R^2 - k}{n-k-1}$$

with n the number of observations, and k the number of parameters fitted.

Regression prediction, prediction SE

We can make regression predictions, for specific conditions of the X variables, e.g. for simple linear regression at x_0:

$$\hat{y}_0 = \hat{\beta}_0 + \hat{\beta}_1 x_0$$

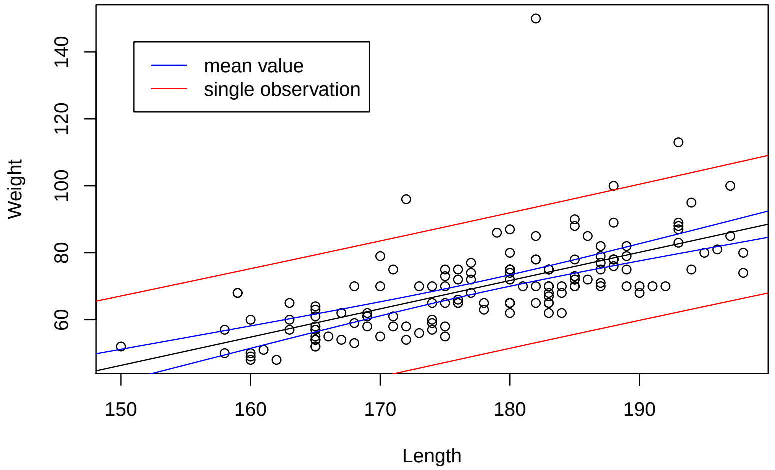
This prediction has an error, the prediction standard error that we can retrieve from the analysis; for the mean value it is:

$$\mathsf{SE}_{\bar{y}_0} = \hat{\sigma}\sqrt{\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}}$$

for a single observation it is:

$$\mathsf{SE}_{\hat{y}_0} = \hat{\sigma}\sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}}$$

two types of 95% prediction intervals


mean value
single observation
Weight
Length