Multiple regression

Dr. Margriet A. Groen

Outline

- Regression line with multiple predictors
- Example
- Standardized coefficients
- Assumptions
- Adjusted R^2

Linear regression

… is a statistical method used to create a linear model

… there are different types:

- **Simple linear regression:** models using only one predictor
- **Multiple linear regression:** models using multiple predictors
- **Logistic regression**: models a categorical response variable
- **Multivariate linear regression:** models for multiple response variables

Regression line

weight

$$y = b_0 + b_1 * x$$

dependent variable

intercept

predictor

The linear model

… can apply to any research question where you are trying to predict a continuous variable of interest (the response or dependent variable) on the basis of one or more other variables (the predictor or independent variables)

… takes the following form:

weights

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots \beta_n x_n \quad + \text{ error}$$

dependent variable

intercept

predictors

Example

```
head(diamonds)
```

##	weight	clarity	color	value	value.lm	weight.c	clarity.c	v
## 1	9.3	0.88	4	182	186	-0.55	-0.12	
## 2	11.1	1.05	5	191	193	1.20	0.05	
## 3	8.7	0.85	6	176	183	-1.25	-0.15	
## 4	10.4	1.15	5	195	194	0.53	0.15	
## 5	10.6	0.92	5	182	189	0.72	-0.08	
## 6	12.3	0.44	4	183	183	2.45	-0.56	

$$\beta_{Int} + \beta_{weight} \times weight + \beta_{clarity} \times clarity + \beta_{color} \times color$$

Nathanial D. Phillips
YaRrr! The Pirate's Guide to R
https://bookdown.org/ndphillips/YaRrr/regression.html

Example: linear regression with `lm()`

Argument	Description
`formula`	A formula in the form `y ~ x1 + x2 + ...` where y is the dependent variable, and x1, x2, … are the independent variables. If you want to include all columns (excluding y) as independent variables, just enter `y ~ .`
`data`	The dataframe containing the columns specified in the formula.

```r
# Create a linear model of diamond values
#   DV = value, IVs = weight, clarity, color

diamonds.lm <- lm(formula = value ~ weight + clarity + color,
                  data = diamonds)
```

Example: output

```
# Print summary statistics from diamond model
summary(diamonds.lm)
##
## Call:
## lm(formula = value ~ weight + clarity + color, data = diamonds)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -10.405  -3.547  -0.113   3.255  11.046
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  148.335      3.625   40.92   <2e-16 ***
## weight         2.189      0.200   10.95   <2e-16 ***
## clarity       21.692      2.143   10.12   <2e-16 ***
## color         -0.455      0.365   -1.25     0.21
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.7 on 146 degrees of freedom
## Multiple R-squared:  0.637,  Adjusted R-squared:  0.63
## F-statistic: 85.5 on 3 and 146 DF,  p-value: <2e-16
```

Example: output (2)

```
# Print summary statistics from diamond model
summary(diamonds.lm)
##
## Call:
## lm(formula = value ~ weight + clarity + color, data = diamonds)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -10.405  -3.547  -0.113   3.255  11.046
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  148.335      3.625   40.92   <2e-16 ***
## weight         2.189      0.200   10.95   <2e-16 ***
## clarity       21.692      2.143   10.12   <2e-16 ***
## color         -0.455      0.365   -1.25     0.21
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.7 on 146 degrees of freedom
## Multiple R-squared:  0.637,  Adjusted R-squared:  0.63
## F-statistic: 85.5 on 3 and 146 DF,  p-value: <2e-16
```

Example: output (3)

```r
# Print summary statistics from diamond model
summary(diamonds.lm)
##
## Call:
## lm(formula = value ~ weight + clarity + color, data = diamonds)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -10.405  -3.547  -0.113   3.255  11.046
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  148.335      3.625   40.92   <2e-16 ***
## weight         2.189      0.200   10.95   <2e-16 ***
## clarity       21.692      2.143   10.12   <2e-16 ***
## color         -0.455      0.365   -1.25     0.21
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.7 on 146 degrees of freedom
## Multiple R-squared:  0.637, Adjusted R-squared:  0.63
## F-statistic: 85.5 on 3 and 146 DF,  p-value: <2e-16
```

Example: output (4)

```
# Print summary statistics from diamond model
summary(diamonds.lm)
##
## Call:
## lm(formula = value ~ weight + clarity + color, data = diamonds)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -10.405  -3.547  -0.113   3.255  11.046
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  148.335      3.625   40.92   <2e-16 ***
## weight         2.189      0.200   10.95   <2e-16 ***
## clarity       21.692      2.143   10.12   <2e-16 ***
## color         -0.455      0.365   -1.25     0.21
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.7 on 146 degrees of freedom
## Multiple R-squared:  0.637,  Adjusted R-squared:  0.63
## F-statistic: 85.5 on 3 and 146 DF,  p-value: <2e-16
```

Standardized coefficients

It is important to keep the metric of each variable in mind when performing multiple regression.

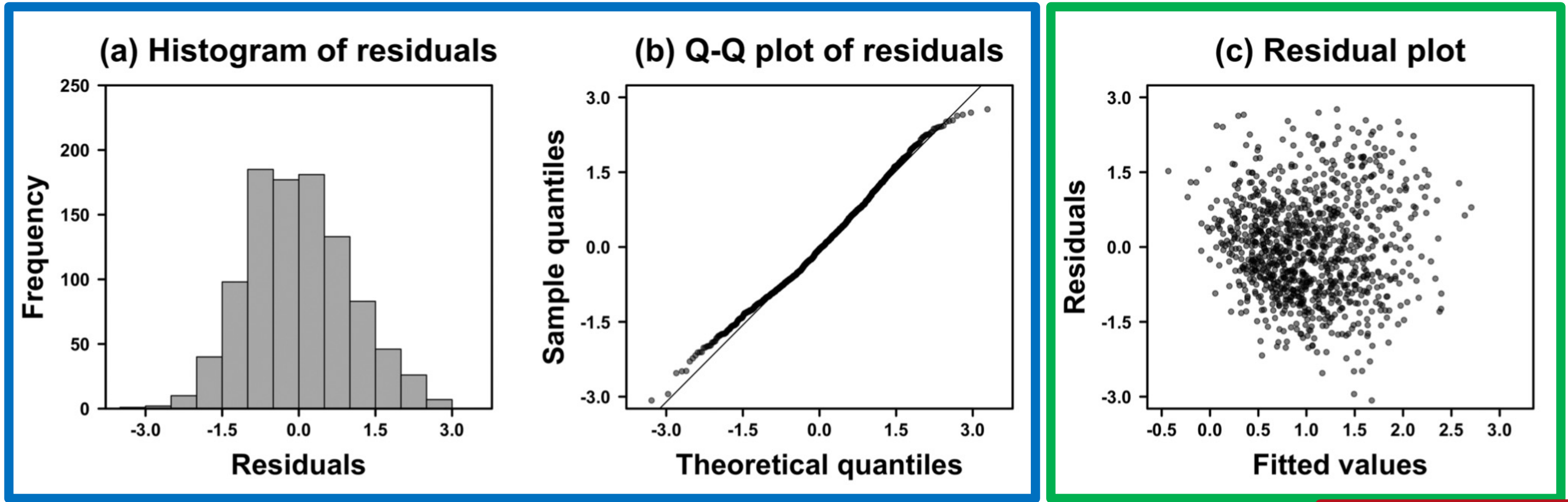
Iconicity study by Winter et al. (2017)

Systematicity
Min -0.000481104
Max 0.000630891


```
Iconicity ~ Sensory + Imageability + Systematicity + Word frequency
   = 1.5 + 0.5*SER + (-0.3)*Imag + 401.5*Systema + (-0.3)*Word freq    unstandardised
   = 1.3 + 0.5*SER + (-0.4)*Imag + 0.0*Systema + (-0.3)*Word freq      standardised
```

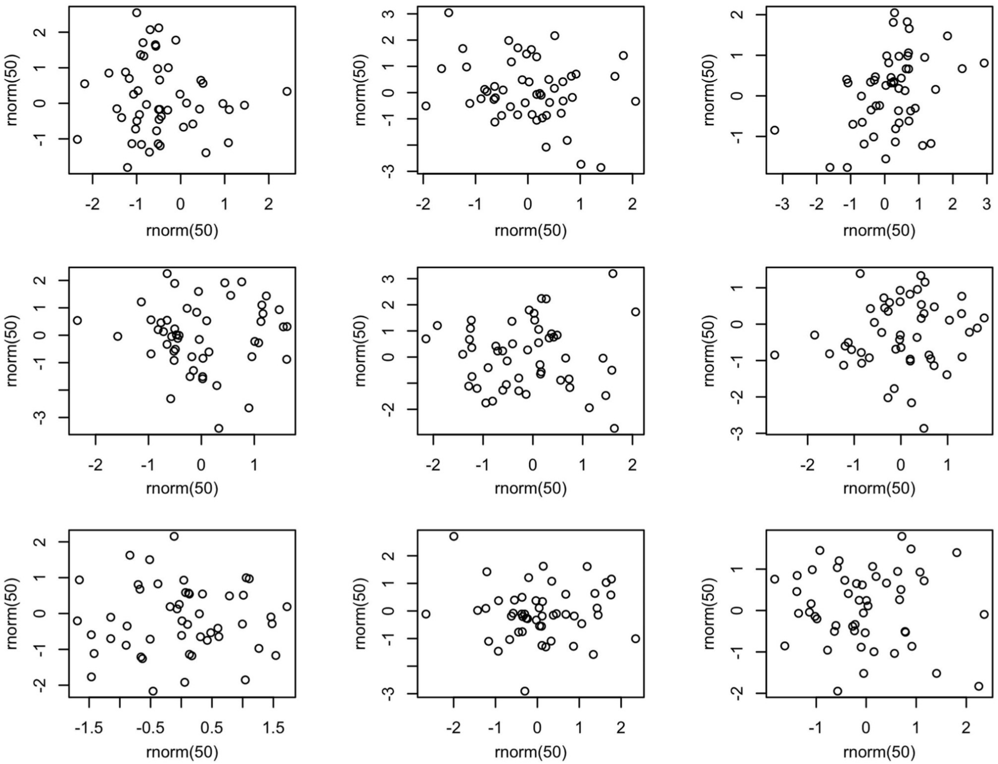
For the standardised coefficients, a one-unit change always corresponds to a change of 1 standard deviation.

Assumptions

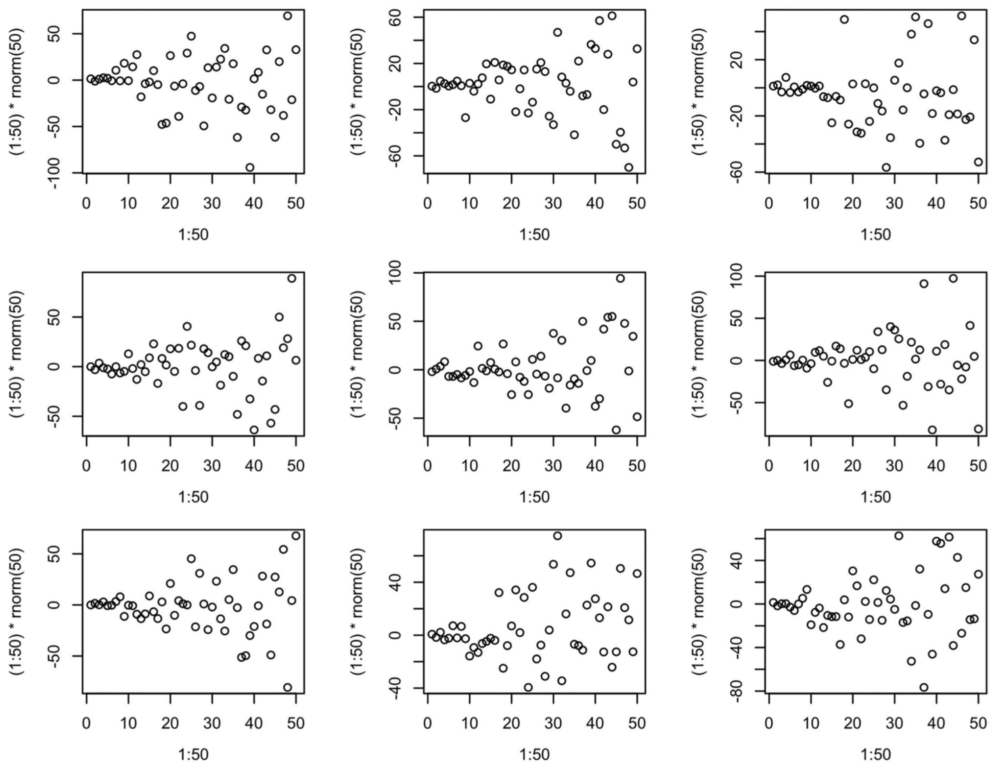


- Normality of residuals
- Homoscedasticity of residuals
- Collinearity

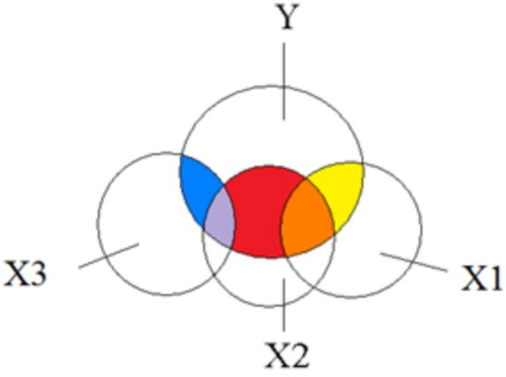
‘Good’ residual plots



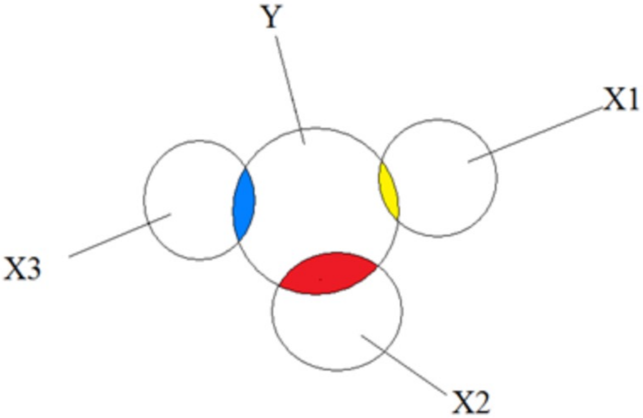
‘Bad’ residual plots



Assumptions: Collinearity (1)



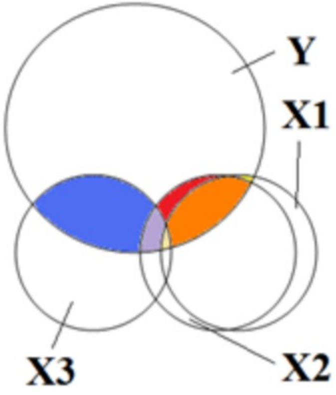
Moderate collinearity



No collinearity

Assumptions: Collinearity (2)

Extreme collinearity




```
x <– rnorm(50)
y <– 10 + 3 * x + rnorm(50)
```

```
x2 <– x
x2 [50] <– -1
```

r(48) = .98

	term	estimate	std.error	statistic	p.value
1	(Intercept)	10.093852	0.1283994	78.61294	2.218721e-52
2	x	2.807983	0.1125854	24.94091	3.960627e-29

	term	estimate	std.error	statistic	p.value
1	(Intercept)	10.181083	0.1669154	60.99548	3.833676e-47
2	x2	2.724396	0.1457204	18.69605	1.123051e-23

	term	estimate	std.error	statistic	p.value
1	(Intercept)	10.0794940	0.1303314	77.3374070	3.352074e-51
2	x	3.2260125	0.5599087	5.7616761	6.164895e-07
3	x2	-0.4255257	0.5582050	-0.7623108	4.496834e-01

Assumptions: Collinearity (3)

- 'Variance inflation factors' (VIFs) can be used to assess whether you have to worry about collinearity.
- VIFs > 3 or 4 are deemed problematic by some (Zuur et al., 2010). Others suggest VIFs > 10 indicate collinearity issues (Montgomery & Peck, 1992).

```
library(car)

vif(xmdl_both)
```

```
       x       x2
24.51677 24.51677
```

```
vif(icon_mdl_z)
```

```
   SER_z CorteseImag_z       Syst_z       Freq_z
1.148597      1.143599     1.015054     1.020376
```

Assumptions: Collinearity (4)

Solutions:

- Remove one of the predictor variables with a high VIF (use you subject knowledge to decide and justify which one).
- Collect more data as that will allow you to estimate the regression coefficients more precisely.
- Use an approach other than regression (e.g., random forests) or first do a principle component analysis to combine predictor variables before doing regression.
- Consider this issue at the planning stage of your study and make theoretically motivated choices as to which one of possibly highly correlated measures to include.

Adjusted R^2

```
glance(icon_mdl_z)
```

```
   r.squared adj.r.squared     sigma  statistic      p.value
1 0.2124559     0.2092545 1.001714 66.36346   9.786184e-50
    df   logLik      AIC      BIC deviance df.residual
1 5 -1402.517 2817.035 2846.415 987.3758         984
```

- Like R^2, it measures how much of the variance in the outcome variable is described by all the predictors in the model together.
- Adjusted R^2 takes the number of predictors in the model into account.
- You should report 'adjusted R^2' .

Summary

Regression line with multiple predictors $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots \beta_n x_n$ + error

Standardized coefficients Make predictors more comparable by converting them to standard units, helps with interpreting coefficients

Assumptions Normality and homoscedasticity of residuals + check collinearity

Adjusted R^2 Takes number of predictors in model into account, therefore more conservative, alerts you to ‘overfitting.