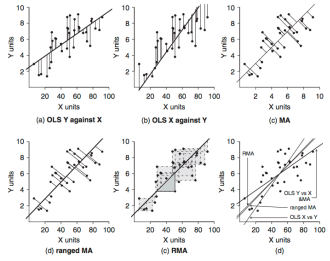
Advancing in R Regression



Outline

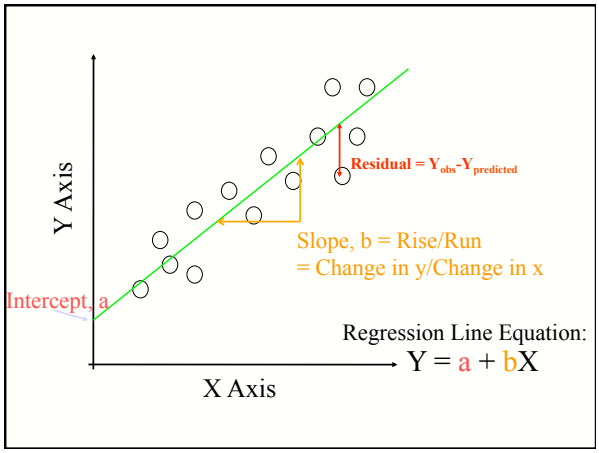
- Brief review: theory of linear regression
- Linear regression in R
  1) Data exploration (plotting!)
  2) Model construction
  3) Diagnostics
  4) Model evaluation

What is a regression?

- Regression analysis establishes a mathematical relationship between 2 or more variables by estimating a best-fit line (linear regression)
- How to find the best-fit line?
  - Minimizing the unexplained variation (the *residual* variation), through minimizing the *sums of squared deviations*

Assumptions of Regression

- Independent variable X, should be fixed
  - *Rarely the case, but should have negligible error*
- Variance of the dependent variable Y is constant for all values of the independent variable X
- Residual values must be normally distributed, *i.e., residuals should be randomly distributed about the regression line*

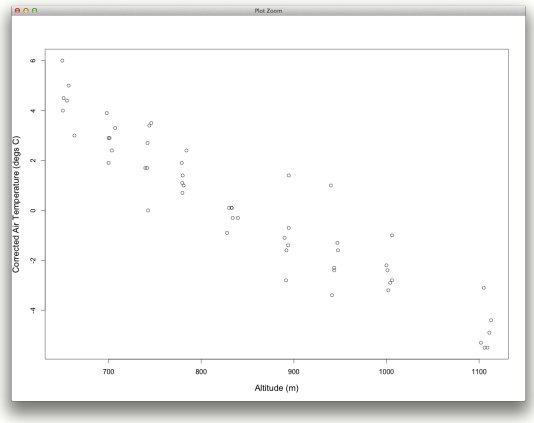


How does air temperature change with altitude in the Cairngorms?

Steps in univariate linear regression

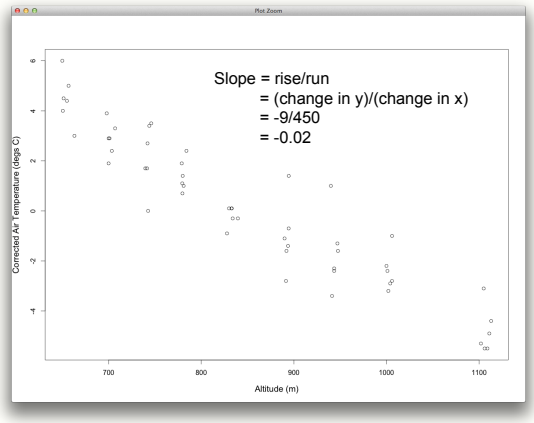
1) Data exploration (plotting!)
2) Model construction
3) Diagnostics
4) Model evaluation

What is the predicted slope for the line?



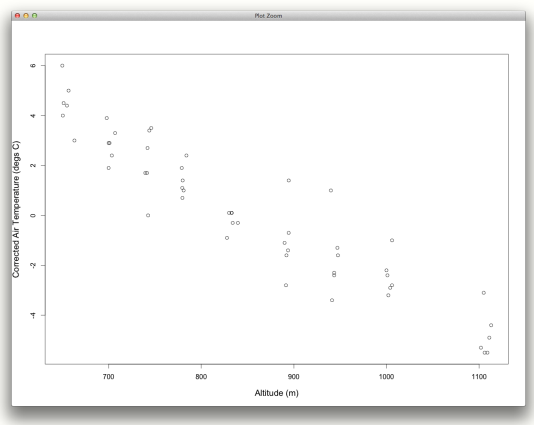
What is the predicted slope for the line?

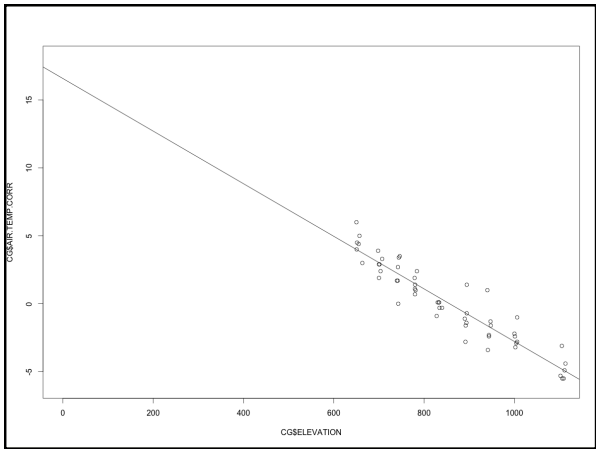
A. -0.2
B. -0.1
C. -0.02
D. -0.01
E. 0.01
F. 0.02
G. 0.1
H. 2



What is the predicted temperature at sea level?

A. 2°C
B. 8°C
C. 12°C
D. 17°C
E. 22°C
F. Not enough information



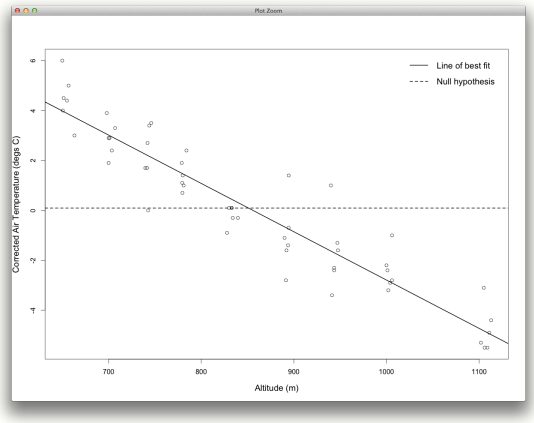


What is the predicted temperature at sea level?

A. 2°C
B. 8°C
C. 12°C
D. 17°C
E. 22°C
F. Not enough information

Which of the following is an appropriate null hypothesis for a regression of air temperature on elevation?

A. Y intercept is positive
B. Y intercept is negative
C. Y-intercept = 0
D. Slope is positive
E. Slope is negative
F. Slope = 0
G. None of the above choices



Steps in univariate linear regression

1) Data exploration (plotting!)
2) Model construction
3) Diagnostics
4) Model evaluation

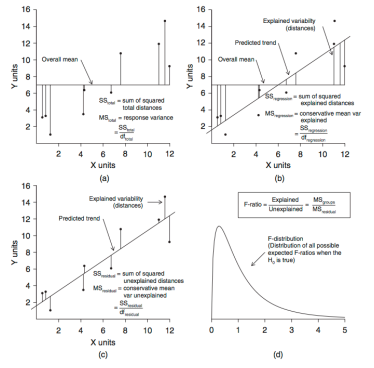


Fig 8.3 Fictitious data illustrating the partitioning of (a) total variation into components (b) explained ($MS_{regression}$) and (c) unexplained ($MS_{residual}$) by the linear trend. The probability of collecting our sample, and thus generating the sample ratio of explained to unexplained variation (or one more extreme), when the null hypothesis is true (and there is no relationship between *X* and *Y*) is the area under the *F*-distribution (d) beyond the sample *F*-ratio.

The Variance

Value of an individual observation

Mean of N observations

$$Var = \frac{\sum_{i=1}^{N}(x_i - \bar{x})^2}{N} = s^2$$

Number of observations

Calling a linear regression in R

➤ `MODEL.NAME <- lm(Y ~ X)`

➤ `AIR.MOD <- lm(AIR.TEMP.CORR~ELEVATION,data=CG)`

OR

➤ `AIR.MOD <- lm(CG$AIR.TEMP.CORR~CG$ELEVATION)`

Steps in univariate linear regression

1) Data exploration (plotting!)
2) Model construction
3) Diagnostics
4) Model evaluation

Residual diagnostics: to evaluate model

Residuals must be examined to test for normal errors and randomness in the distribution.

- The **residuals vs fits** plot: should show no pattern between the residuals and the fitted values
- The **quantile-quantile (QQ) normal plot**: if the theoretical and observed quantiles are similar, the dots should make a straight line on y=x
- The **scale-location plot:** as for the fits vs. resids, should show no pattern
- The **Cook's distance plot:** illustrates data points with high influence on parameter estimates

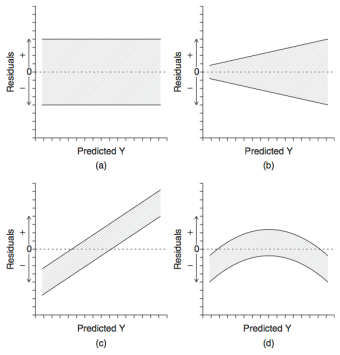
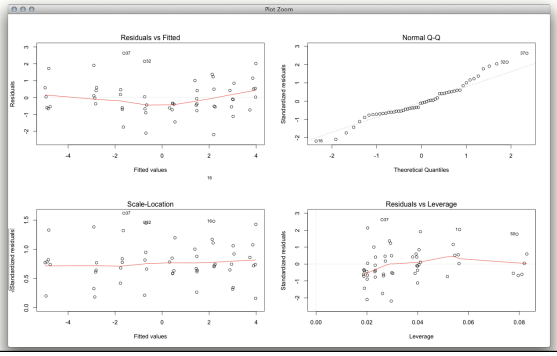
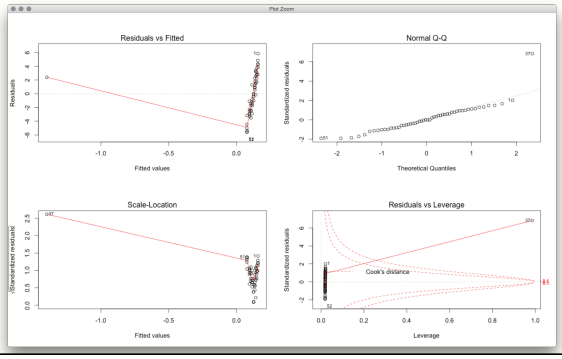


Fig 8.5 Stylised residual plots depicting characteristic patterns of residuals (a) random scatter of points - homogeneity of variance and linearity met (b) "wedge-shaped" - homogeneity of variance not met (c) linear pattern remaining - erroneously calculated residuals or additional variable(s) required and (d) curved pattern remaining - linear function applied to a curvilinear relationship. Modified from Zar (1999).

```
par(mfrow=c(2,2))
plot(AIR.MOD)
par(mfrow=c(1,1))
```



```
par(mfrow=c(2,2))
plot(AIR.MOD.BAD)
par(mfrow=c(1,1))
```



Steps in univariate linear regression

1) Data exploration (plotting!)
2) Model construction
3) Diagnostics
4) Model evaluation

```
> summary(AIR.MOD)

Call:
lm(formula = AIR.TEMP.CORR ~ ELEVATION, data = CG)

Residuals:
     Min       1Q   Median       3Q      Max
-2.1941 -0.6031 -0.1092  0.5059  2.6280

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 16.5723075  0.8498285   19.50   <2e-16 ***
ELEVATION   -0.0193620  0.0009853  -19.65   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’
1

Residual standard error: 1.017 on 52 degrees of freedom
Multiple R-squared:  0.8813,     Adjusted R-squared:  0.879
F-statistic: 386.2 on 1 and 52 DF,  p-value: < 2.2e-16
```

The ANOVA table is much less informative than the summary call:

```
➤ summary.aov(AIR.MOD

            Df Sum Sq Mean Sq F value Pr(>F)
ELEVATION    1  399.1   399.1   386.2 <2e-16 ***
Residuals   52   53.7     1.0
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

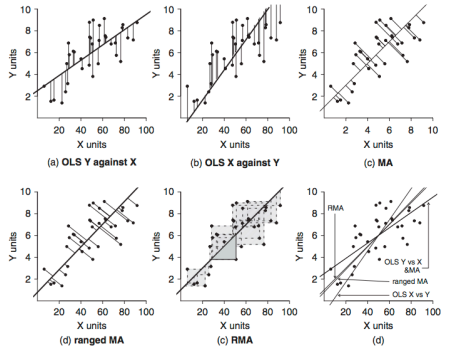
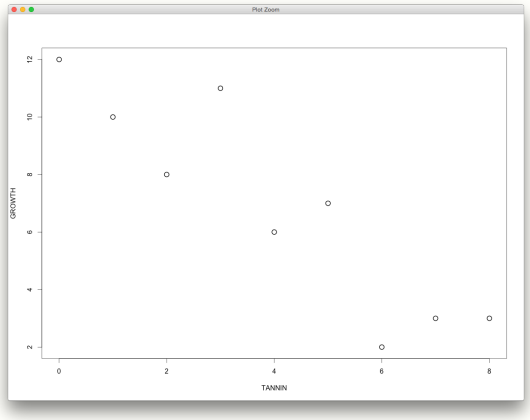


Fig 8.4 Fictitious data illustrating the differences between (a) ordinary least squares, (b) major axis and (c) reduced major axis regression. Each are also contrasted in (d) along with a depiction of ordinary least squares regression for X against Y. Note that the fitted line for all techniques passes through the center mean of the data cloud. When the X and Y are measured on the same scale, MA and RMA are the same.

Practical exercise

- Tannin is a plant compound that can interfere with insect digestion
- N = 9 caterpillars, all given a different concentration of tannin
- How does diet affect caterpillar growth?
- Predictor variable?
- Response variable?
- Plot?



Suggested reading:

- Ch. 8 in Logan (excellent!)
- Or Ch. 8 in Crawley *Statistics*…
- Or Ch. 10 in Crawley *The R Book*