Problems addressed using regression

- Mapping the locations of children at high risk of lead exposure (Y) as a function of the year the home was built (X_1), % African American residents in census block (X_2), median income in the census blockgroup (X_3) and county (X_4).
- Estimate the average time to extinction (Y) for bird species over several decades as a function of the number of nesting pairs (X_1), size of the species (X_2) and migratory status (X_3).
- Does foreign debt (X_1) cause deforestation (Y) after accounting for population (X_2) in a set of countries?
- What site characteristics, such as percent forest canopy cover (X_1) and age of a forest (X_2), best predict the abundance of a particular species?

Course Overview

- Multivariate regression
  - Estimation, inference, prediction, critique
  - Transformations of data
  - Robustness of regression procedures
  - Different types of X's: categorical/nominal, ordinal, continuous
  - Influential points
  - Model selection from classical and Bayesian perspectives
- Temporally correlated data, Logistic regresssion, Poisson regression

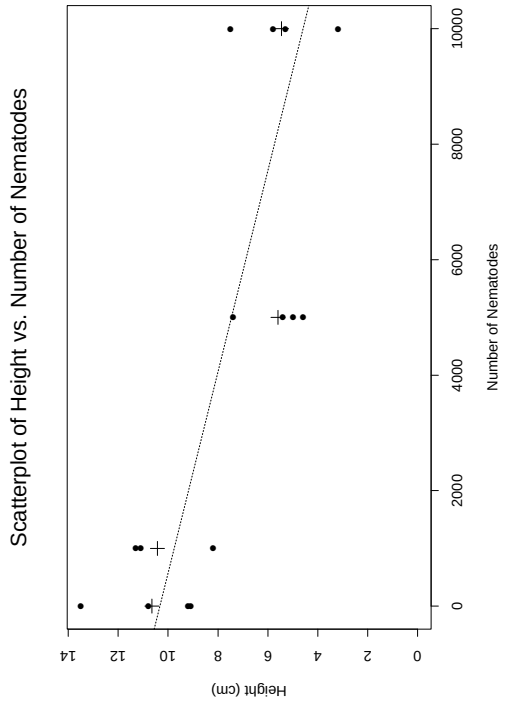
To start, Statistical Sleuth, Ch. 1, 7.1-7.4. (Note M&M 2.1-2.5, 6.1-6.3, 10.1 p. 657-670)

Simple Linear Regression Example

How do nematodes (microscopic worms) affect plant growth? A botanist prepares 16 identical planting pots and randomly assigns each pot to receive different numbers of nematodes. A tomato seedling is transplanted into each plot.

Table 1: Increase in the height of tomato seedlings 16 days after planting

Nematodes	Seedling Growth (cm)			
0	10.8	9.1	13.5	9.2
1000	11.1	11.1	8.2	11.3
5000	5.4	4.6	7.4	5.0
10000	5.8	5.3	3.2	7.5



Regression estimation

- Estimated mean function:

$$\hat{\mu}\{Y|X\} = \hat{\beta}_0 + \hat{\beta}_1 X \quad (1)$$

Let X be measured in 1000's of nematodes.

$$\hat{\mu}\{\text{Height}|\text{Number}\} = 10.33 - 0.57 \times \text{Number} \quad (2)$$


```
*** Linear Model ***

Coefficients:
                      Value Std. Error  t value Pr(>|t|)
        (Intercept)  10.3264   0.6890    14.9876   0.0000
Number.of.Nematodes  -0.5738   0.1228    -4.6740   0.0004

Residual standard error: 1.933 on 14 df
Multiple R-Squared: 0.6094
F-statistic: 21.85 on 1 and 14 df, the p-value is 0.000358
```


Fitted values and residuals

	number	height	fitted	residual	sq.resid
1	0	10.8	10.33	0.47	0.22
2	0	9.1	10.33	-1.23	1.50
3	0	13.5	10.33	3.17	10.07
4	0	9.2	10.33	-1.13	1.27
5	1	11.1	9.75	1.35	1.82
6	1	11.1	9.75	1.35	1.82
7	1	8.2	9.75	-1.55	2.41
8	1	11.3	9.75	1.55	2.39
9	5	5.4	7.46	-2.06	4.23
10	5	4.6	7.46	-2.86	8.17
....					
14	10	5.3	4.59	0.71	0.51
15	10	3.2	4.59	-1.39	1.93
16	10	7.5	4.59	2.91	8.48

Presentation of Regression Results

(bottom page 185)

Estimated plant height	=	10.33	+	-0.57	(Number of nematodes)
		(0.69)		(0.12)	

Estimated SD of plant heights = 1.93 (14 df)

R^2=0.61

What is the difference between “Estimated SD of plant heights” above and

$SD(Y)$?