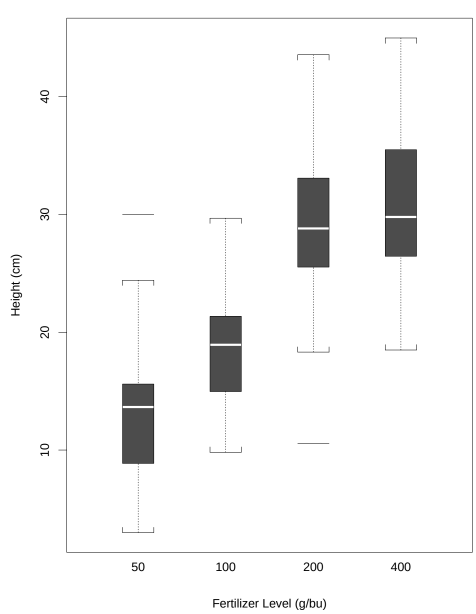
Week 3: Simple Linear Regression

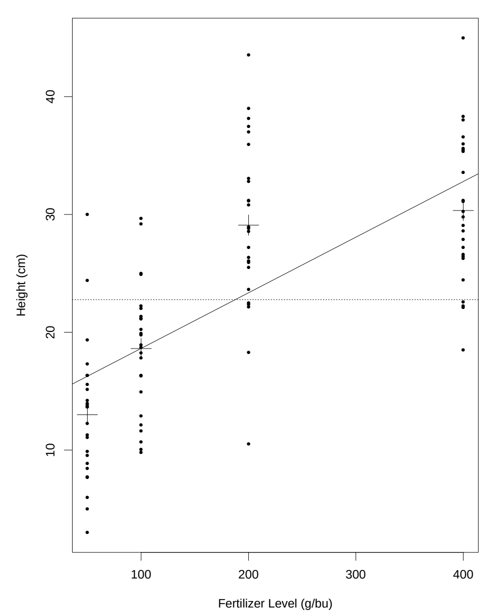
- Last class: One-way ANOVA models
- Week 3 Reading: Chapter 7, *Statistical Sleuth*, plus all conceptual exercises.
  - Suggested: Moore and McCabe, Sec. 2.2, 2.3, 2.4, 10.1, 10.2
- Topics
  - Simple linear regression model
  - Estimation methods
  - Properties of least squares estimates
  - Inference
  - Correlation

1-way ANOVA: Week 2 Plant Example

- Let Y_{ij}= height of plant i in treatment group j, $j = 1, 2, 3, 4$ fertilizer levels.
- Let $\bar{Y}_j$= mean of plant heights in group j.
- **ANOVA model**: $Y_{ij} = \mu_j + \varepsilon_{ij}$
- *Unknown parameters*: group means $\mu_1, \mu_2, \mu_3, \mu_4$ and single standard deviation σ.
- ANOVA hypothesis and inference
- Splus output:

	Df	Sum.Sq	M.Sq	F	Pr(F)
LEVEL	3	5257.2	1752.4	44.3	0
Resid	96	3499.4	39.6		





Correlation

Regression Model

- Let Y = random height of plant
- Let X = fixed amount of fertilizer given to each plant. X takes on *fixed* values of 50, 100, 200, 400 g/bu
- **Linear Regression Model**:
  - statistical model relating Y to X
  - Let $\mu\{Y|X\}$ be the population mean (or *expected value*) of Y given X.
  - $\mu\{Y|X\} = \beta_0 + \beta_1 X$
  - *Unknown parameters*: $\beta_0, \beta_1, \sigma = \sigma\{Y|X\}$
  - Individual Y_i are modeled as:
    * $Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$
- Assumptions: $\varepsilon_i \sim$ IID N$(0, \sigma)$

Estimation of Regression Parameters

- **Least Squares:** Find $\hat{\beta}_0$ and $\hat{\beta}_1$ to minimize

$$\text{SS}(\beta_0, \beta_1) = \sum_{i=1}^{n} [Y_i - (\beta_0 + \beta_1 X_1)]^2 \qquad (1)$$

- Solutions to the LS equations:

$$\begin{aligned} \hat{\beta}_1 &= \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^n (X_i - \bar{X})^2} \\ &= r \frac{s_y}{s_x} \\ \hat{\beta}_0 &= \bar{Y} - \hat{\beta}_1 \bar{X} \end{aligned}$$

- How well does the model fit?

$$\hat{\sigma} = \sqrt{\frac{\text{RSS}}{N-2}} = \sqrt{\frac{\sum_{i=1}^n (Y_i - \hat{Y}_i)^2}{N-2}}$$

d.f.= N - number of parameters = $N - 2$

Estimation: Plant Data

$\hat{Y}$	=	13.900	+	0.047	X
		(1.233)		(0.005)	

$\hat{\sigma}$ = 7.174 (98 df)

#	level	ht	fit	resid	sq.resid
1	1	14.2	16.3	-2.0	4.1
2	1	13.7	16.3	-2.6	6.8
3	1	24.4	16.3	8.2	66.5
4	1	8.4	16.3	-7.8	61.3
5	1	16.3	16.3	0.1	0.0
6	1	12.3	16.3	-4.0	16.1
7	1	9.5	16.3	-6.7	45.1
8	1	5.0	16.3	-11.3	126.9
9	1	9.9	16.3	-6.4	40.5
10	1	16.4	16.3	0.1	0.0

Splus output for plant regression

```
Coefficients:
               Value Std. Error t value Pr(>|t|)
(Intercept) 13.9004  1.2335    11.2691  0.0000
      level  0.0473  0.0054     8.8310  0.0000

Residual standard error: 7.174 on 98 df
Multiple R-Squared: 0.4431
F-stat: 77.99 on 1 and 98 df, p-val=4.152e-14
```


Properties of Least Squares Estimators

Hypothesis tests for β_o and β_1

- Does X contribute any information for prediction of Y?
  Test: $H_o : \beta_1 = 0$ vs. $H_A : \beta_1 \neq 0$
  (or $H_A : \beta_1 > 0$)

$$\frac{\hat{\beta}_1 - \beta_1}{SE(\hat{\beta}_1)} \sim t_{n-2} \qquad (2)$$

- Tests regarding β_o:

$$\frac{\hat{\beta}_o - \beta_o}{SE(\hat{\beta}_o)} \sim t_{n-2} \qquad (3)$$

- 95% confidence intervals for β_o and β_1

- Distribution of the mean of Y at a given X
- Prediction interval for Y_o at $X = X_o$