L21: Chapter 12: Linear regression

Department of Statistics, University of South Carolina

Stat 205: Elementary Statistics for the Biological and Life Sciences

So far...

- One sample continuous data (Chapters 6 and 8).
- Two sample continuous data (Chapter 7).
- One sample categorical data (Chapter 9).
- Two sample categorical data (Chapter 10).
- More than two sample continuous data (Chapter 11).
- Now: continuous predictor X instead of group.

Two continuous variables

- Instead of relating an outcome Y to "group" (e.g. 1, 2, or 3), we will relate Y to another continuous variable X.
- First we will measure how linearly related Y and X are using the correlation.
- Then we will model Y vs. X using a line.
- The data arrive as n *pairs* $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$.
- Each pair (x_i, y_i) can be listed in a table and is a point on a scatterplot.

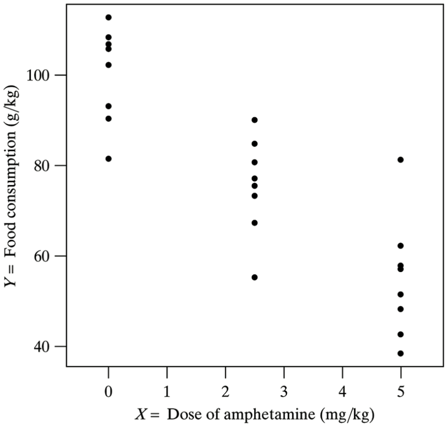
Example 12.1.1 Amphetamine and consumption

Amphetamines suppress appetite. A pharmacologist randomly allocated $n = 24$ rats to three amphetamine dosage levels: 0, 2.5, and 5 mg/kg. She measured the amount of food consumed (gm/kg) by each rat in the 3 hours following.

Table 12.1.1 Food consumption (Y) of rats (gm/kg)

	X = Dose of amphetamine (mg/kg)		
	0	2.5	5.0
	112.6	73.3	38.5
	102.1	84.8	81.3
	90.2	67.3	57.1
	81.5	55.3	62.3
	105.6	80.7	51.5
	93.0	90.0	48.3
	106.6	75.5	42.7
	108.3	77.1	57.9
Mean	100.0	75.5	55.0
SD	10.7	10.7	13.3
No. of animals	8	8	8

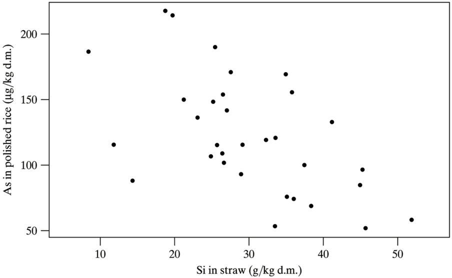
Example 12.1.1 Amphetamine and consumption



How does Y change with X? Linear? How strong is linear relationship?

Example 12.1.2 Arsenic in rice

Environmental pollutants can contaminate food via the growing soil. Naturally occurring silicon in rice may inhibit the absorption of some pollutants. Researchers measured Y, amount of arsenic in polished rice (μg/kg rice), & X, silicon concentration in the straw (g/kg straw), of $n = 32$ rice plants.



Example 12.2.1 Length and weight of snakes

In a study of a free-living population of the snake Vipera bertis, researchers caught and measured nine adult females.

Table 12.2.1

	Length X (cm)	Weight Y (g)
	60	136
	69	198
	66	194
	64	140
	54	93
	67	172
	59	116
	65	174
	63	145
Mean	63	152
SD	4.6	35.3

Example 12.2.1 Length and weight of snakes

How strong is linear relationship?

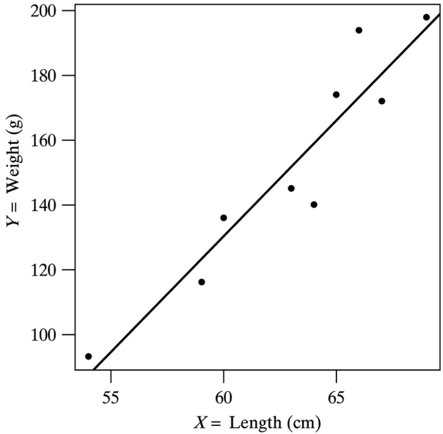


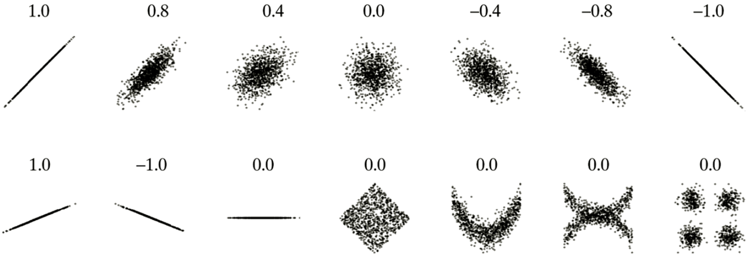
Figure 12.2.1 Body length and weight of nine snakes with fitted regression line

12.2 The correlation coefficient r

$$r = \frac{1}{n-1} \sum_{i=1}^{n} \left(\frac{x_i - \bar{x}}{s_x} \right) \left(\frac{y_i - \bar{y}}{s_y} \right).$$

- r measures the strength and direction (positive or negative) of how *linearly* related Y is with X.
- $-1 \leq r \leq 1$.
- If $r = 1$ then Y increases with X according to a perfect line.
- If $r = -1$ then Y decreases with X according to a perfect line.
- If $r = 0$ then X and Y are not *linearly* associated.
- The closer r is to 1 or -1, the more the points lay on a straight line.

Examples of r for 14 different data sets



Population correlation ρ

- Just like $\bar{y}$ estimates μ and s_y estimates σ, r estimates the unknown *population correlation* ρ.
- If $\rho = 1$ or $\rho = -1$ then *all points in the population* lie on a line.
- Sometimes people want to test $H_0 : \rho = 0$ vs. $H_A : \rho \neq 0$, or they want a 95% confidence interval for ρ.
- These are easy to get in R with the `cor.test(sample1,sample2)` command.

R code for amphetamine data

```
> cons=c(112.6,102.1,90.2,81.5,105.6,93.0,106.6,108.3,73.3,84.8,67.3,55.3,
+        80.7,90.0,75.5,77.1,38.5,81.3,57.1,62.3,51.5,48.3,42.7,57.9)
> amph=c(0,0,0,0,0,0,0,0,2.5,2.5,2.5,2.5,2.5,2.5,2.5,2.5,5.0,5.0,5.0,5.0,5.0,5.0,5.0,5.0)
> cor.test(amph,cons)

        Pearson's product-moment correlation

data:  amph and cons
t = -7.9003, df = 22, p-value = 7.265e-08
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 -0.9379300 -0.6989057
sample estimates:
      cor
-0.859873
```

$r = -0.86$, a strong, negative relationship.
P-value$= 0.000000073 < 0.05$ so reject $H_0 : \rho = 0$ at the 5% level.
There is a signficant, negative linear association between amphetamine intake and food consumption. We are 95% confident that the true population correlation is between -0.94 and -0.70.

R code for snake data

```
> length=c(60,69,66,64,54,67,59,65,63)
> weight=c(136,198,194,140,93,172,116,174,145)
> cor.test(length,weight)

        Pearson's product-moment correlation

data:  length and weight
t = 7.5459, df = 7, p-value = 0.0001321
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.7489030 0.9883703
sample estimates:
      cor
0.9436756
```

$r = 0.94$, a strong, positive relationship. What else do we conclude?

Comments

- Order doesn't matter, either (X, Y) or (Y, X) gives the same correlation and conclusions. Correlation is "symmetric."
- Significant correlation, rejecting $H_0 : \rho = 0$ doesn't mean ρ is close to 1 or -1; it can be small, yet significant.
- Rejecting $H_0 : \rho = 0$ doesn't mean X causes Y or Y causes X, just that they are linearly associated.

12.3 Fitting a line to scatterplot data

We will fit the line

$$Y = b_0 + b_1 X$$

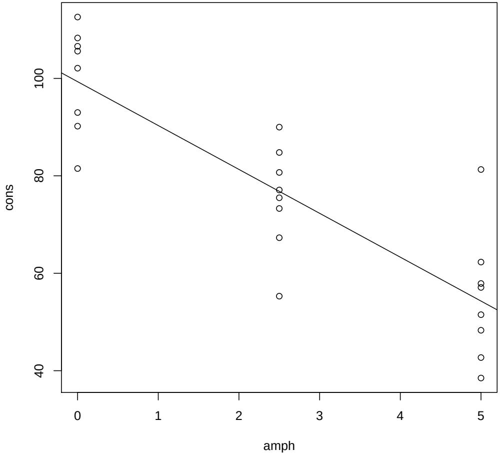
to the data pairs.

- b_0 is the **intercept**, how high the line is on the Y-axis.
- b_1 is the **slope**, how much the line changes when X is increase by one unit.
- The values for b_0 and b_1 we use gives the **least squares** line.
- These are the values that make $\sum_{i=1}^{n}[y_i - (b_0 + b_1 x_i)]^2$ as small as possible.
- They are

$$b_1 = r\left(\frac{s_y}{s_x}\right) \text{ and } b_0 = \bar{y} - b_1\bar{x}.$$

```
> fit=lm(cons~amph)
> plot(amph,cons)
> abline(fit)
> summary(fit)

Call:
lm(formula = cons ~ amph)

Residuals:
    Min      1Q  Median      3Q     Max
-21.512  -7.031   1.528   7.448  27.006

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   99.331      3.680   26.99  < 2e-16 ***
amph          -9.007      1.140   -7.90 7.27e-08 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 11.4 on 22 degrees of freedom
Multiple R-squared: 0.7394,     Adjusted R-squared: 0.7275
F-statistic: 62.41 on 1 and 22 DF,  p-value: 7.265e-08
```

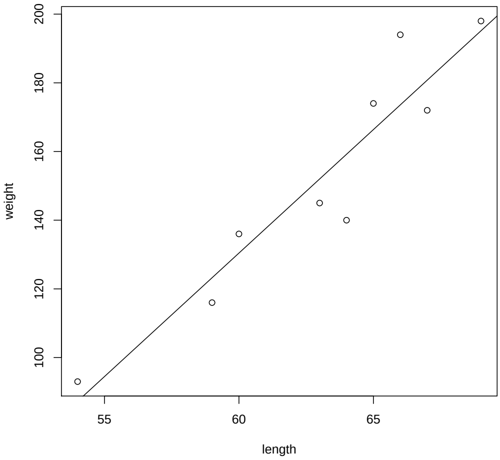
For now, just pluck out $b_0 = 99.331$ and $b_1 = -9.007$



$$\text{cons} = 99.33 - 9.01 \text{ amph}.$$

```
> fit=lm(weight~length)
> plot(length,weight)
> abline(fit)
> summary(fit)

Call:
lm(formula = weight ~ length)

Residuals:
    Min      1Q  Median      3Q     Max
-19.192  -7.233   2.849   5.727  20.424

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -301.0872    60.1885  -5.002 0.001561 **
length         7.1919     0.9531   7.546 0.000132 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 12.5 on 7 degrees of freedom
Multiple R-squared: 0.8905,     Adjusted R-squared: 0.8749
F-statistic: 56.94 on 1 and 7 DF,  p-value: 0.0001321
```

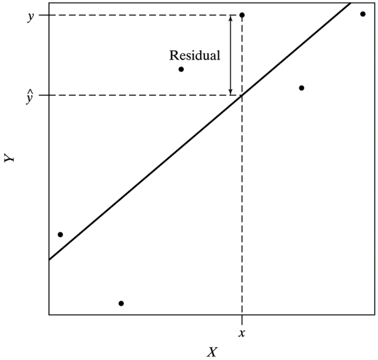
Here, $b_0 = -301.1$ and $b_1 = 7.19$



$$\text{weight} = -301.1 + 7.19 \text{ length}.$$

Residuals

- The ith fitted value is $\hat{y}_i = b_0 + b_1 x_i$, the point on the line above x_i.
- The ith residual is $e_i = y_i - \hat{y}_i$. This gives the vertical amount that the line missed y_i by.



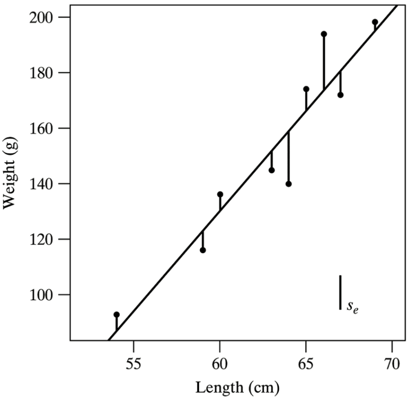
Residual sum of squares and s_e

- SS(resid)$= \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} e_i^2$.
- (b_0, b_1) make SS(resid) as small as possible.
- $s_y = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2}$ is sample standard deviation of the Y's. Measures the "total variability" in the data.

s_e, s_y, and r^2

- $s_e = \sqrt{\frac{1}{n-2}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} = \sqrt{\mathrm{SS(resid)}/(n-2)}$ is "residual standard deviation" of the Ys. Measures *variability around the regression line*.
- If $s_e \approx s_y$ then the regression line isn't doing anything!
- If $s_e < s_y$ then the line is doing something.
- $r^2 \approx 1 - \frac{s_e^2}{s_y^2}$ is called the **multiple R-squared**, and is the percentage of variability in Y explained by X through the regression line.
- R calls s_e the *residual standard error*.

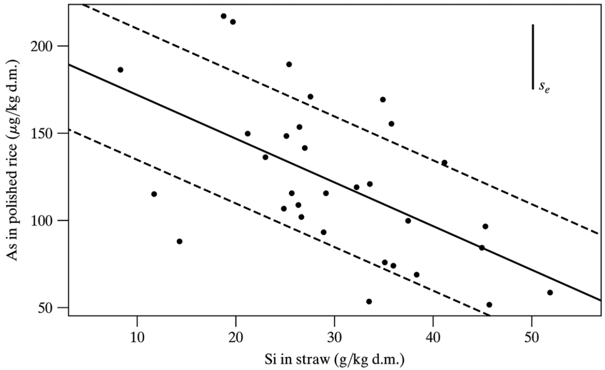
s_e is just average length of residuals




```
> sd(weight)
[1] 35.33766
```

$s_e = 12.5$ and $s_y = 35.3$. $r^2 = 0.89$ so 89% of the variability in

68%-95% rule for regression lines



Roughly 68% of observations are within s_e *of the regression line* (shown above); 95% are within 2 s_e.

12.4 The regression model

- We assume the underlying model with Greek letters (as usual)

$$y = \beta_0 + \beta_1 x + \epsilon$$

- For each subject i we see x_i and $y_i = \beta_0 + \beta_1 x_i + \epsilon_i$.
- β_0 is the population intercept.
- β_1 is the population slope.
- ϵ_i is the ith error, we assume these are $N(0, \sigma_e)$.
- We don't know any of β_0, β_1, or σ_e.

Visualizing the model

- $\mu_{y|x} = \beta_0 + \beta_1 x$ is mean response for <u>everyone</u> with covariate x.
- σ_e is constant variance. Variance doesn't change with x.
- Example 12.4.4, pretend *we know* that the mean weight $\mu_{y|x}$ given height x is

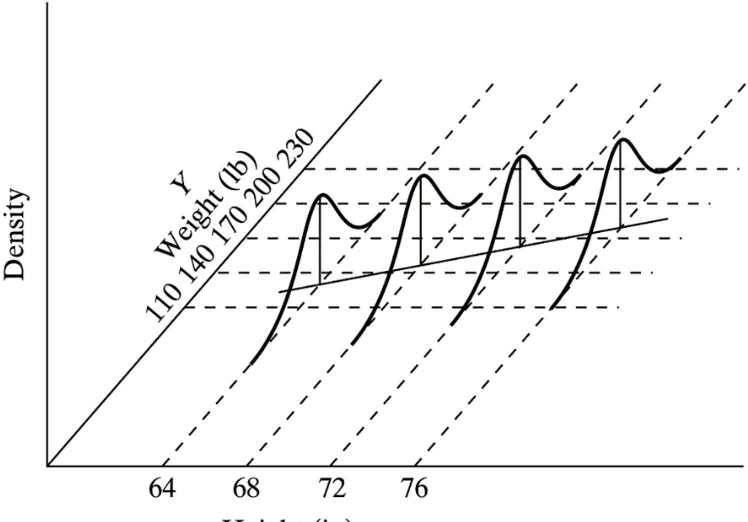
$$\mu_{y|x} = -145 + 4.25x \text{ and } \sigma_e = 20.$$

Table 12.4.1 Conditional means and SDs of weight given height in a population of young men*

Height (in) X	Mean weight (lb) $\mu_{Y\|X}$	Standard deviation of weights (lb) $\sigma_{Y\|X}$
64	127	20
68	144	20
72	161	20
76	178	20

*Note that all values of $\sigma_{Y|X}$ are the same; they equal $\sigma_\varepsilon = 20$.

Weight vs. height



Estimating β_0, β_1, and σ_ϵ

- b_0 estimates β_0.
- b_1 estimates β_1.
- s_e estimates σ_e.
- Example 12.4.5. For the snake data, $b_0 = -301$ estimates β_0, $b_1 = 7.19$ estimates β_1, and $s_e = 12.5$ estimates σ_e.
- We estimate the the mean weight $\hat{y}$ of snakes with length x as

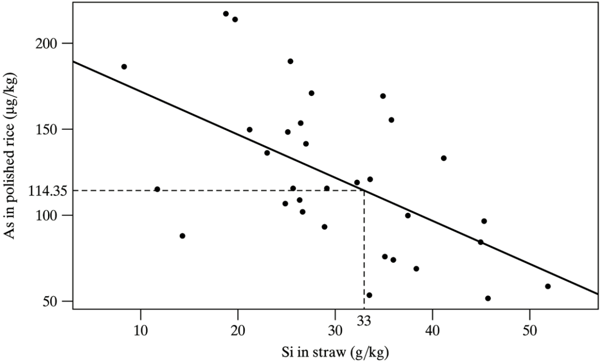
$$\hat{y} = -301 + 7.19x$$

Example 12.4.6 Arsenic in rice

- If we believe the data follow a line, we can estimate the mean for any x we want.
- $b_0 = 197.17$ estimates β_0, $b_1 = 2.51$ estimates β_1, and $s_e = 37.30$ estimates σ_e.
- For straw silicon concentration of $x = 33$ g/kg we estimate a mean arsenic level of

$$\hat{y} = 197.17 - 2.51(33) = 114.35 \ \mu\text{gm/kg with } s_e = 37.30 \ \mu\text{gm/kg}.$$

Arsenic in rice at $X = 33$ g/kg



$$\hat{y} = 197.17 - 2.51x$$

$$114.35 = 197.17 - 2.51(33)$$

12.5 Inference for β_1

- Often people want a 95% confidence interval for β_1 and want to test $H_0 : \beta_1 = 0$.
- If we reject $H_0 : \beta_1 = 0$, then y is significantly linearly assocatied with x. Same as testing $H_0 : \rho = 0$.
- A 95% confidence interval for β_1 gives us a range for how the mean changes when x is increased by one unit.
- Everything comes from

$$\frac{b_1 - \beta_0}{SE_{b_1}} \sim t_{n-2}, \quad SE_{b_1} = \frac{s_e}{s_x\sqrt{n-1}}.$$

- R automatically gives a P-value for testing $H_0 : \beta_1 = 0$.
- Need to ask R for 95% confidence interval for β_1.

R code

```
> amph=c(0,0,0,0,0,0,0,0,2.5,2.5,2.5,2.5,2.5,2.5,2.5,2.5,5.0,5.0,5.0,5.0,5.0,5.0,5.0,5.0)
> cons=c(112.6,102.1,90.2,81.5,105.6,93.0,106.6,108.3,73.3,84.8,67.3,55.3,
+        80.7,90.0,75.5,77.1,38.5,81.3,57.1,62.3,51.5,48.3,42.7,57.9)
> fit=lm(cons~amph)
> summary(fit)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   99.331      3.680   26.99  < 2e-16 ***
amph          -9.007      1.140   -7.90 7.27e-08 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

> confint(fit)
                2.5 %     97.5 %
(Intercept)  91.69979 106.962710
amph        -11.37202  -6.642979
```

P-value for testing $H_0 : \beta_1 = 0$ vs. $H_A : \beta_1 \neq 0$ is 0.0000000727, we reject at the 5% level. We are 95% confidence that true mean consumption is reduced by 6.6 to 11.4 g/kg for every mg/kg increase in amphetamine dose.

Multiple regression

- Often there are more than one predictors we are interested in, say we have two x_1 and x_2.
- The model is easily extended to

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \epsilon$$

- Example: Dwayne Portrait Studio is doing a sales analysis based on data from $n = 21$ cities.
  - y = sales (thousands of dollars) for a city
  - x_1 = number of people 16 years or younger (thousands)
  - x_2 = per capita disposable income (thousands of dollars)

The data

x_1	x_2	y	x_1	x_2	y
68.5	16.7	174.4	45.2	16.8	164.4
91.3	18.2	244.2	47.8	16.3	154.6
46.9	17.3	181.6	66.1	18.2	207.5
49.5	15.9	152.8	52.0	17.2	163.2
48.9	16.6	145.4	38.4	16.0	137.2
87.9	18.3	241.9	72.8	17.1	191.1
88.4	17.4	232.0	42.9	15.8	145.3
52.5	17.8	161.1	85.7	18.4	209.7
41.3	16.5	146.4	51.7	16.3	144.0
89.6	18.1	232.6	82.7	19.1	224.1
52.3	16.0	166.5			

R code for multiple regression

```
> under16=c(68.5,45.2,91.3,47.8,46.9,66.1,49.5,52.0,48.9,38.4,87.9,72.8,88.4,42.9,52.5,
+           85.7,41.3,51.7,89.6,82.7,52.3)
>
> income=c(16.7,16.8,18.2,16.3,17.3,18.2,15.9,17.2,16.6,16.0,18.3,17.1,17.4,15.8,17.8,
+          18.4,16.5,16.3,18.1,19.1,16.0)
>
> sales=c(174.4,164.4,244.2,154.6,181.6,207.5,152.8,163.2,145.4,137.2,241.9,191.1,232.0,
+         145.3,161.1,209.7,146.4,144.0,232.6,224.1,166.5)
> fit=lm(sales~under16+income)
> summary(fit)

Call:
lm(formula = sales ~ under16 + income)

Residuals:
     Min       1Q   Median       3Q      Max
-18.4239  -6.2161   0.7449   9.4356  20.2151

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -68.8571    60.0170  -1.147   0.2663
under16       1.4546     0.2118   6.868    2e-06 ***
income        9.3655     4.0640   2.305   0.0333 *
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 11.01 on 18 degrees of freedom
Multiple R-squared: 0.9167,     Adjusted R-squared: 0.9075
```

Interpretation...

- The fitted regression *surface* is

$$\text{sales} = -68.857 + 1.455\ (\text{under } 16) + 9.366\ \text{income}.$$

- For every unit increase (1000 people) in those under 16, average sales go up 1.455 thousand, \$1,455.
- For every unit increase (\$1000) in disposable income, average sales go up 9.366 thousand, \$9,366.
- 91.67% of the variability in sales is explained by those under 16 and disposable income.
- σ_e is estimated to be 11.01.

Regression homework

- 12.2.5, 12.2.7, 12.3.1, 12.3.3, 12.3.5, 12.3.7, 12.3.8. Use R for all problems; i.e. don't do anything by hand.
- 12.4.3, 12.4.6, 12.4.8, 12.4.9, 12.5.1, 12.5.3, 12.5.5, 12.5.9(a). Use R for all problems; don't do anything by hand.