CHAPTER 10:

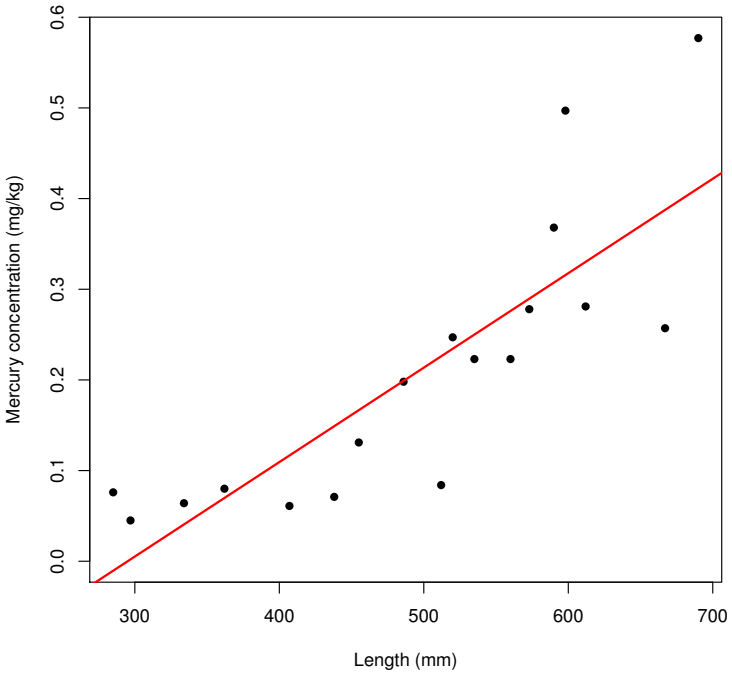
10.1. Mercury can accumulate in fish tissue over time which, in turn, can pose a public health risk to humans who consume fish. Researchers at the Florida Fish and Wildlife Conservation Commission recently sampled $n = 18$ scamp grouper fish from the Gulf of Mexico and measured the following two variables on each fish:

$$\begin{aligned} Y &= \text{mercury concentration (mg/kg)} \\ x &= \text{length (mm).} \end{aligned}$$

One goal was to model Y as a function of x using simple linear regression; i.e.,

$$Y = \beta_0 + \beta_1 x + \epsilon,$$

where $\epsilon \sim \mathcal{N}(0, \sigma^2)$. Below is a scatterplot of the data for the 18 fish caught; superimposed on the scatterplot is the least-squares regression line.



(a) Describe what one might consider "the population" to be in this example.
(b) The least squares estimates of β_0 and β_1 are given in the output below:

```
> fit = lm(mercury ~ length)
Coefficients:
(Intercept)       length
   -0.30733      0.00104
```

Ninety-five percent (95%) confidence intervals for β_0 and β_1 are given below:

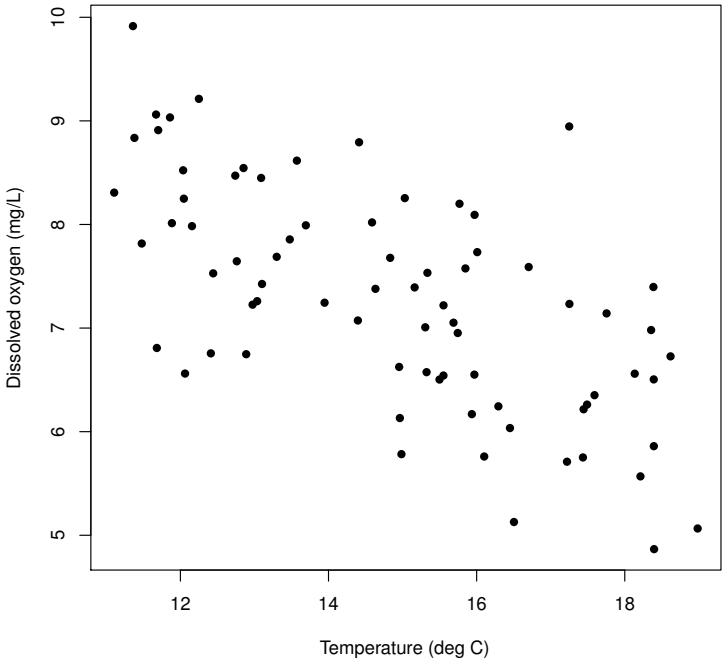
```
> confint(fit,conf.level=0.95)
                2.5 %   97.5 %
(Intercept) -0.49921 -0.11546
length       0.00067  0.00142
```

Does this analysis demonstrate mercury concentration and length are linearly related in the population? Explain.
(c) I used R to calculate $R^2 \approx 0.68$ for the simple linear regression fit. Explain what this means.
(d) The researchers would like to consider the subpopulation of scamp grouper fish whose length is $x = 500$ mm. Calculate a point estimate of the mean of this subpopulation.

10.2. Researchers recorded the following variables on $n = 75$ water specimens taken from a large lake in northern California:

$$\begin{aligned} Y &= \text{level of dissolved oxygen (mg/L)} \\ x &= \text{water temperature (deg C).} \end{aligned}$$

A scatterplot of the data is shown below:



Consider the population-level model

$$Y = \beta_0 + \beta_1 x + \epsilon \iff E(Y) = \beta_0 + \beta_1 x$$

for these data. I fit this model in R and obtained the following output:

```
> fit = lm(dissolved.o2 ~ temp)
> summary(fit)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 11.90532    0.64108  18.571  < 2e-16 ***
temp        -0.31143    0.04261  -7.309 2.78e-10 ***

Residual standard error: 0.827 on 73 degrees of freedom
Multiple R-squared:  0.422
```

I also asked R to calculate 95% confidence intervals for β_0 and β_1:

```
> confint(fit)
              2.5 %   97.5 %
(Intercept) 10.6276  13.1829
temp        -0.3963  -0.2265
```

(a) Are temperature and the level of dissolved oxygen linearly related in the population? Explain with statistical evidence.
(b) From the `summary` output, we see the value of $R^2 \approx 0.422$. Interpret what this means.
(c) I calculated a 95% confidence interval when temperature = 15 deg C:

```
> predict(fit,data.frame(temp=15),level=0.95,interval="confidence")
       fit      lwr     upr
    7.2338   7.0431  7.4245
```

Interpret what this interval means.

10.3. Zoologist researchers at the Wildlife Research Institute are studying adult black bears in Alaska. The researchers were interested in the relationship between the two variables:

$$\begin{aligned} Y &= \text{chest girth (in inches)} \\ x &= \text{weight (in lbs).} \end{aligned}$$

A sample of $n = 48$ black bears was caught. Each bear in the sample was measured and released back into the wild. The researchers assumed a simple linear regression model

$$Y = \beta_0 + \beta_1 x + \epsilon \quad \Longleftrightarrow \quad E(Y) = \beta_0 + \beta_1 x$$

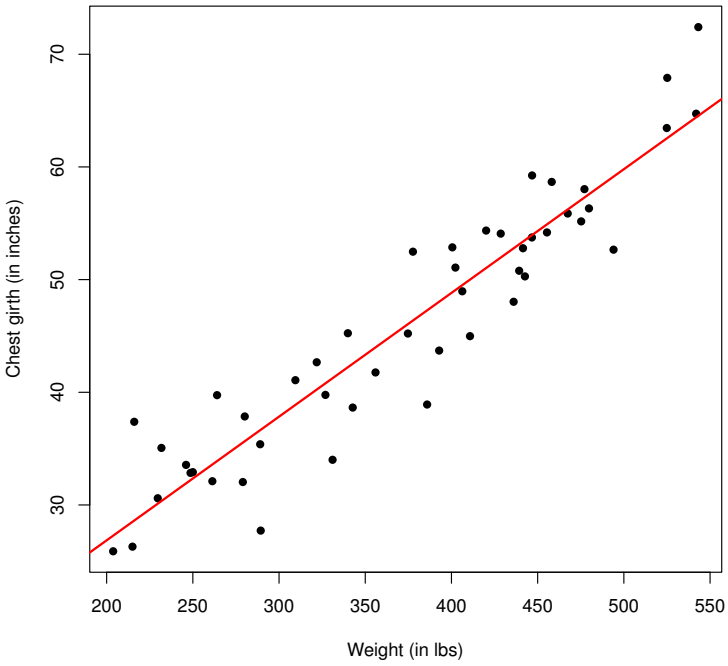
to describe the relationship between Y and x for all black bears in Alaska. Here are the least-squares estimates of β_0 and β_1 and 95% confidence intervals for each parameter:

```
> fit = lm(chest.girth ~ weight)
> fit

Coefficients:
(Intercept)       weight
      4.898        0.109

> confint(fit,conf.level=0.95)
              2.5%  97.5%
(Intercept)  0.227  9.570
weight       0.097  0.121
```

A scatterplot of the sample data (with the least-squares regression line superimposed) is shown at the top of the next page.
(a) The R output reports a 95% confidence interval for β_1 to be $(0.097, 0.121)$. Interpret what this interval means in terms of the population mean chest girth $E(Y)$ when weight is increased by one pound.



(b) Will the confidence interval for β_0 be useful for the researchers? Explain.
(c) The zoologists would like to estimate the population mean chest girth for all black bears in Alaska whose weight is $x_0 = 400$ lbs. Would a confidence interval or prediction interval be appropriate here? Explain.
(d) (Chapter 11) In addition to weight, the researchers also measured two additional predictor variables: body fat percentage and the age of the bear. Write a multiple linear regression model for the population that includes Y and the three predictors:

$$\begin{aligned} x_1 &= \text{weight (in lbs)} \\ x_2 &= \text{body fat (percentage of weight)} \\ x_3 &= \text{age (in years).} \end{aligned}$$

How many regression parameters are in your model?

CHAPTER 11:

11.1. This problem deals with an extrusion process used in soybeans. "Extrusion" refers to the process by which certain materials are extracted from the soybeans (e.g., fiber, oil, etc.) to be used in other products (e.g., cattle feed, flour, etc.). An experiment was performed to investigate the relationship between

$$Y = \text{soluble dietary fiber percentage (}\texttt{SDFP}\text{) in soybean residue}$$

and three independent variables

- x_1 = extrusion temperature, (`temp`, in deg C)
- x_2 = feed moisture (`moisture`, in %)

- x_3 = extrusion screw speed (`speed`, in rpm).

Here are the data recorded in the experiment:

Observation	x_1	x_2	x_3	Y
1	35	110	160	11.13
2	25	130	180	10.98
3	30	110	180	12.56
4	30	130	200	11.46
5	30	110	180	12.38
6	30	110	180	12.43
7	30	110	180	12.55
8	25	110	160	10.59
9	30	130	160	11.15
10	30	90	200	10.55
11	30	90	160	9.25
12	25	90	180	9.58
13	35	110	200	11.59
14	35	90	180	10.68
15	35	130	180	11.73
16	25	110	200	10.81
17	30	110	180	12.68

Experimenters initially considered the multiple linear regression model to relate `SDFP` to the three independent variables:

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \epsilon_i,$$

for $i = 1, 2, ..., 17$.
(a) This regression model can be written out in the form $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$. What is the dimension of the matrix $\mathbf{X}$? Give me the first two rows of the matrix.
(b) The ANOVA table (with sequential sums of squares) for the multiple linear regression model fit is shown on the next page.

```
> fit = lm(SDFP~temp+moisture+speed)
> anova(fit)

Analysis of Variance Table
          Df  Sum Sq Mean Sq F value  Pr(>F)
temp       1  1.2561  1.2561  1.4289 0.25330
moisture   1  3.4585  3.4585  3.9341 0.06885 .
speed      1  0.6555  0.6555  0.7457 0.40350
Residuals 13 11.4281  0.8791

Residual standard error: 0.9376 on 13 degrees of freedom
Multiple R-squared:  0.3197,    Adjusted R-squared:  0.1627
F-statistic: 2.036 on 3 and 13 DF,  p-value: 0.1585
```

What is the conclusion from this analysis? In your answer, cite the value of the overall F statistic above (and corresponding p-value). Make sure you say precisely what hypotheses the overall F statistic is testing.

(c) The experimenters also considered a multiple linear regression model with quadratic terms:

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_4 x_{i1}^2 + \beta_5 x_{i2}^2 + \beta_6 x_{i3}^2 + \epsilon_i,$$

for $i = 1, 2, ..., 17$. The three extra independent variables are the squared versions of x_1, x_2, and x_3, respectively. Here is the ANOVA table (with sequential sums of squares) for this multiple linear regression model fit:

```
> fit.quad = lm(SDFP~temp+moisture+speed+temp.sq+moisture.sq+speed.sq)
> anova(fit)

Analysis of Variance Table
            Df Sum Sq Mean Sq F value    Pr(>F)
temp         1 1.2561  1.2561  24.860 0.0005486 ***
moisture     1 3.4585  3.4585  68.447 8.763e-06 ***
speed        1 0.6555  0.6555  12.973 0.0048330 **
temp.sq      1 2.5869  2.5869  51.197 3.086e-05 ***
moisture.sq  1 5.5393  5.5393 109.629 1.041e-06 ***
speed.sq     1 2.7967  2.7967  55.351 2.211e-05 ***
Residuals   10 0.5053  0.0505

Residual standard error: 0.2248 on 10 degrees of freedom
Multiple R-squared:  0.9699,    Adjusted R-squared:  0.9519
F-statistic: 53.74 on 6 and 10 DF,  p-value: 4.915e-07
```

Which model seems to fit the data better−the model without the quadratic terms or the model with the quadratic terms? Justify your choice with evidence.

11.2. Rayon whiteness is an important factor for fabric quality. A random sample of $n = 16$ rayon specimens is available. On each specimen, the response Y (a numerical measure of rayon whiteness) is measured. The following six independent variables are also measured:

$$\begin{aligned}
x_1 &= \text{acid bath temperature (deg C)} \\
x_2 &= \text{cascade acid concentration (percentage)} \\
x_3 &= \text{water temperature (deg C)} \\
x_4 &= \text{sulfide concentration (percentage)} \\
x_5 &= \text{amount of chlorine bleach (lb/min)} \\
x_6 &= \text{blanket finish temperature (deg C).}
\end{aligned}$$

Chemists would like to model the relationship between Y and the six independent variables using the multiple linear regression model

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_4 x_{i4} + \beta_5 x_{i5} + \beta_6 x_{i6} + \epsilon_i,$$

for $i = 1, 2, ..., 16$. I fit this model to the data in R, and here is the output:

```
> fit = lm(whiteness ~ acid.temp + cascade.conc + water.temp + sulfide.conc
    + chlorine + blanket.temp)
```

```
> summary(fit)

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  -172.6744   179.2521  -0.963    0.361
acid.temp      -1.2619     1.5706  -0.803    0.442
cascade.conc  -27.9731    47.1617  -0.593    0.568
water.temp      2.1838     1.5819   1.381    0.201
sulfide.conc    1.3639   236.4142   0.006    0.996
chlorine      136.1482   119.0395   1.144    0.282
blanket.temp    1.0401     0.7725   1.346    0.211
Residual standard error: 19.23 on 9 degrees of freedom
Multiple R-squared:  0.3015
F-statistic: 0.6475 on 6 and 9 DF,  p-value: 0.6927
```

(a) The F statistic (above) is used to test two hypotheses: H_0 and H_1. Write out what these hypotheses are. You can do this using notation (that you clearly define) or you can write this out in words. Which hypothesis is more supported by the data? Use a significance level of $\alpha = 0.05$ to make your decision.
(b) Here is the ANOVA table for the least-squares fit with sequential sums of squares for each independent variable:

```
> anova(fit)
Analysis of Variance Table

Response: whiteness
             Df Sum Sq Mean Sq F value Pr(>F)
acid.temp     1  155.5  155.48  0.4206 0.5328
cascade.conc  1    6.0    5.98  0.0162 0.9016
water.temp    1  537.7  537.72  1.4547 0.2585
sulfide.conc  1    0.8    0.80  0.0022 0.9639
chlorine      1   65.8   65.84  0.1781 0.6829
blanket.temp  1  670.1  670.15  1.8130 0.2111
Residuals     9 3326.7  369.63
```

We know the following partition of the variability for any least-squares regression fit:

$$\text{SS}_{\text{TOT}} = \text{SS}_{\text{R}} + \text{SS}_{\text{E}}.$$

What are the values of these sums of squares for this least-squares fit? What are the degrees of freedom associated with each one?
(c) This analysis makes four statistical assumptions on the error terms in the model. List what these assumptions are. Also, describe how you could detect violations of each one using the residuals. *Hint:* Think of using a qq plot and a residual plot.

11.3. A turf grower who provides high quality turf for sporting events performs a study to investigate how

$$Y \quad = \quad \text{turf quality score}$$

is related to the two independent variables

$$\begin{aligned} x_1 &= \text{amount of water applied (in hundreds of gallons)} \\ x_2 &= \text{amount of fertilizer applied (in lbs).} \end{aligned}$$

The study is carried out using $n = 14$ plots of land, and same variety of turf is planted on each plot. The quality score Y is based on the appearance of the turf, the health of the roots, the health of the grass blades, and the density of the grass. High turf quality scores mean better quality turf. The data from the study are shown below:

Plot	Quality score (Y)	Water (x_1)	Fertilizer (x_2)
1	72	9	26
2	71	12	20
3	62	6	20
4	32	15	24
5	48	15	16
6	37	13	28
7	68	13	12
8	27	11	32
9	46	11	8
10	48	7	32
11	46	5	28
12	41	5	12
13	38	3	24
14	34	3	16

The R output from estimating the multiple linear regression model

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \epsilon \quad \Longleftrightarrow \quad \underbrace{E(Y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2}_{\text{a plane in } \mathbb{R}^3}.$$

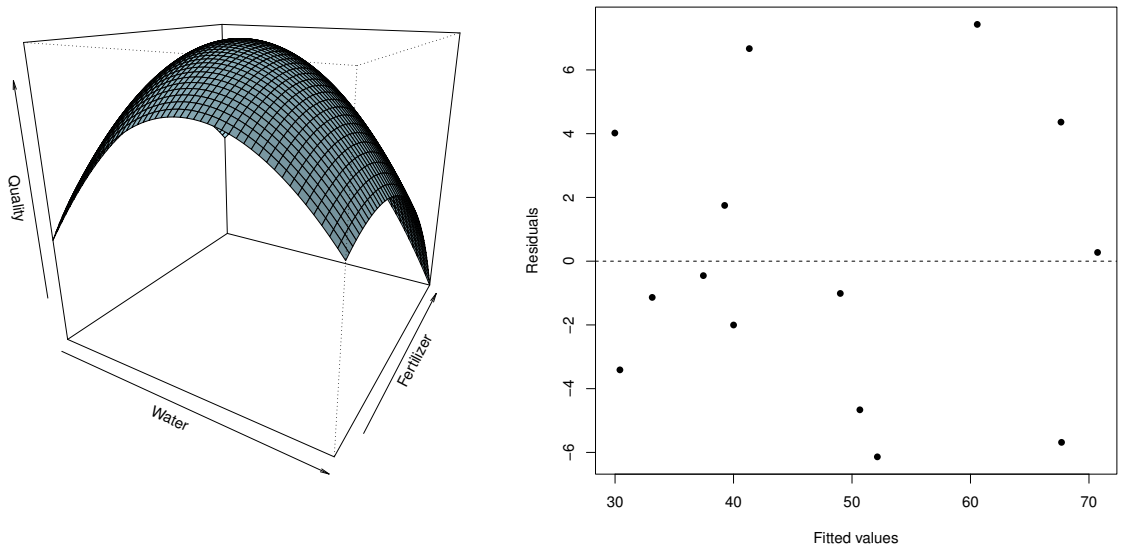
is shown below:

```
> fit = lm(Quality ~ Water + Fertilizer)
> summary(fit)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   53.510     16.220    3.30   0.0071 **
Water          0.378      1.016    0.37   0.7168
Fertilizer    -0.428      0.566   -0.76   0.4657

Residual standard error: 15.6 on 11 degrees of freedom
Multiple R-squared:  0.0636
F-statistic: 0.374 on 2 and 11 DF,  p-value: 0.697
```

(a) The analysis above suggests that `Water` and `Fertilizer` are not linearly related to the population mean turf quality score $E(Y)$ in the population of all plots. Cite statistical evidence of this from the output.



(b) The turf grower knows that both `Water` and `Fertilizer` should influence turf quality, so he tries estimating another model which incorporates quadratic effects and an interaction between the two predictors:

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1^2 + \beta_4 x_2^2 + \beta_5 x_1 x_2 + \epsilon$$

$$\iff \quad \underbrace{E(Y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1^2 + \beta_4 x_2^2 + \beta_5 x_1 x_2}_{\text{a curvilinear surface in } \mathbb{R}^3}.$$

The R output from estimating this model is shown below:

```
> fit.2 = lm(Quality ~ Water + Fertilizer + Water.sq + Fertilizer.sq + Water*Fertilizer)
> summary(fit.2)

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)     -144.140     23.935   -6.02  0.00032 ***
Water             24.164      2.818    8.58  2.6e-05 ***
Fertilizer        11.511      1.643    7.01  0.00011 ***
Water.sq          -1.014      0.139   -7.27  8.6e-05 ***
Fertilizer.sq     -0.236      0.035   -6.61  0.00017 ***
Water:Fertilizer  -0.269      0.063   -4.27  0.00274 **

Residual standard error: 5.57 on 8 degrees of freedom
Multiple R-squared:  0.914
F-statistic: 16.9 on 5 and 8 DF,  p-value: 0.00045
```

Cite statistical evidence explaining how this model is a better fit than the one in part (a).
(c) From the `fit.2` output (above), the estimated curvilinear regression model is

$$\widehat{Y} = -144.140 + 24.164 x_1 + 11.511 x_2 - 1.014 x_1^2 - 0.236 x_2^2 - 0.269 x_1 x_2.$$

I used R to plot this function in three dimensions (see above, top left). The turf grower would like to know the value of `Water` (x_1) and `Fertilizer` (x_2) that maximizes the predicted turf quality score. Using the estimated model above, write out a system of two equations and two

unknowns which, when solved, will produce the answer to this question. You don't have to solve the system (just write it out). *Hint:* Think partial derivatives.
(d) The residual plot from the curvilinear regression model fit in part (b) is shown on the last page (top right). What does this plot suggest about the quality of the model fit? Do you detect any violations in the underlying assumptions? Explain.

CHAPTER 12:

12.1. Researchers performed a 2×2 factorial experiment to examine how much carbon dioxide (CO_2) is produced when pine wood is burned (Y, measured as a percent) and how this is related to two factors:

Factor A: Moisture
Factor B: Temperature.

The levels of Factor A are $a_1 = 0$ (no moisture present) and $a_2 = 22$ percent (moisture present). The levels of Factor B are $b_1 = 1100$ deg K and $b_2 = 1500$ deg K. There were 2 replicates at each treatment combination, resulting in the following 8 CO_2 measurements.

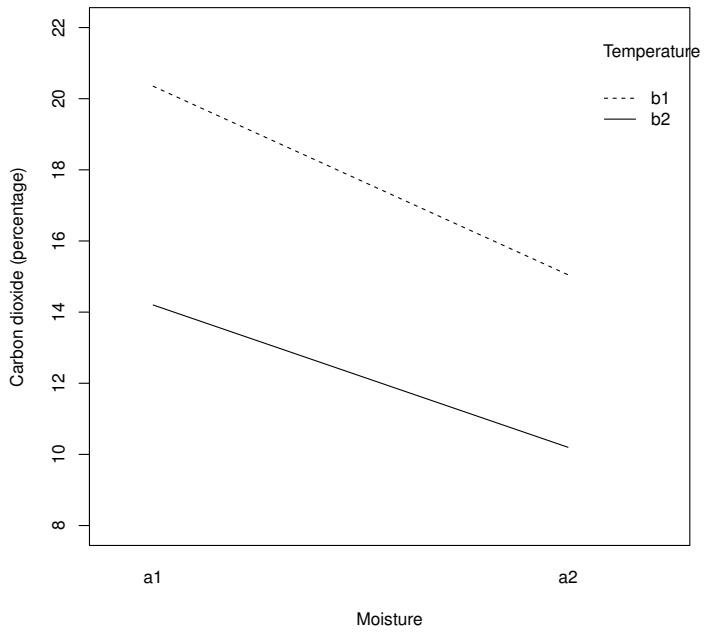
Moisture (A)	Temperature (B)	Replication 1	Replication 2
0	1100	20.3	20.4
22	1100	13.6	14.8
0	1500	15.0	15.1
22	1500	9.7	10.7

I used R to construct the moisture-temperature interaction plot. This is shown on the next page (top). I also calculated the ANOVA table appropriate for the 2×2 factorial treatment structure:

```
> fit = lm(CO.2 ~ Moisture + Temperature + Moisture*Temperature)
> anova(fit)

                     Df  Sum Sq  Mean Sq   F value    Pr(>F)
Moisture              1  43.245  43.2450  140.6342  0.000289 ***
Temperature           1  60.500  60.5000  196.7480  0.000149 ***
Moisture:Temperature  1   0.845   0.8450    2.7480  0.172718
Residuals             4   1.230   0.3075
```

(a) One researcher suggests there should be a population-level interaction between moisture and temperature. Provide evidence to support or refute this assertion.
(b) Which of the main effects are significant in explaining CO_2 percentage: moisture, temperature, or both? Explain.
(c) Instead of analyzing the data as we did above, suppose we analyzed the data as data from a one-way classification with $t = 4$ treatments: a_1b_1, a_1b_2, a_2b_1, and a_2b_2. In a one-way analysis, what would the treatment sum of squares SS_T equal?
(d) In the analysis described in part (c), calculate the F statistic to test the equality of population means across the four treatment groups (a_1b_1, a_1b_2, a_2b_1, and a_2b_2). Is this a value of F you would expect to see if the four population mean CO_2 percentages were equal? Explain.



12.2. Civil engineers performed a 2^2 factorial experiment to investigate how the fracture toughness of an asphalt specimen (Y, measured in MPa) depends on two factors: mixture type (Factor A) and temperature (Factor B). The data from the experiment are below.

Mixture Type (A)	Temperature (B)	Replicate 1	Replicate 2	Replicate 3
Normal	-35 deg C	15.8	15.6	14.9
Polymer added	-35 deg C	13.7	13.8	13.2
Normal	-10 deg C	13.3	13.9	12.8
Polymer added	-10 deg C	15.6	15.9	16.6

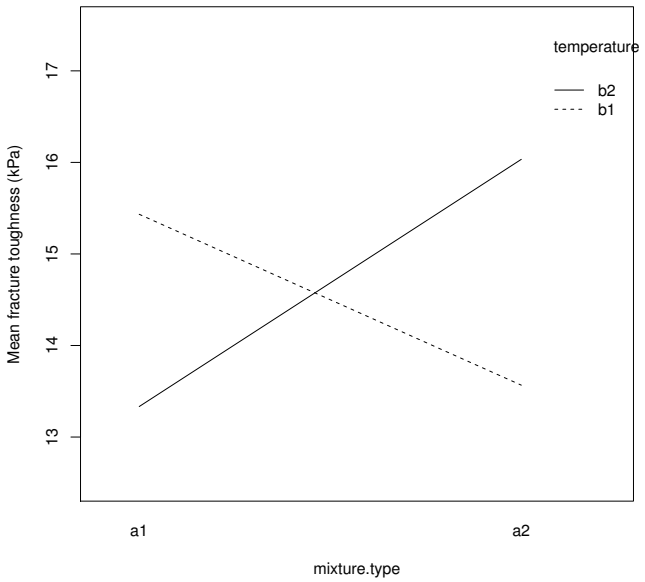
(a) Ignoring the factorial treatment structure, I analyzed the data as data from a one-way classification with four treatment groups like we did in Chapter 9:

```
> fit = lm(fracture.toughness ~ treatment)
> anova(fit)

Analysis of Variance Table
          Df  Sum Sq Mean Sq F value    Pr(>F)
treatment  3 16.2625  5.4208  24.272 0.0002267
Residuals  8  1.7867  0.2233
```

The F statistic (here $F \approx 24.3$) tests which two hypotheses? You can write your answer out in words, or you can use statistical symbols. If you use symbols, define what the symbols mean.
(b) Acknowledging the factorial treatment structure, let SS_A, SS_B, and SS_AB denote the sums of squares for the main effect of mixture type, the main effect of temperature, and the interaction effect between mixture type and temperature, respectively. What is $\text{SS}_\text{A} + \text{SS}_\text{B} + \text{SS}_\text{AB}$?

(c) Here is the interaction plot between mixture type (A) and temperature (B):



Consider the following two hypotheses:

H_0: mixture type and temperature do not interact in the population
H_1: mixture type and temperature do interact in the population.

The interaction plot above makes a strong argument for which hypothesis? Explain.
(d) In addition to mixture type (A) and temperature (B), suppose the engineers wanted to include two additional factors in the experiment:

C: manufacturer (M1 and M2)
D: air void percentage (4 percent and 6 percent).

With four factors now (each with two levels), how many asphalt specimens would be needed to complete three full replications?

12.3. An engineer is interested in the effects of cutting speed (A), tool geometry (B), and cutting angle (C) on the lifetime (Y, in hours) of a machine tool. Two levels of each factor are chosen, and three replications of a 2^3 factorial experiment are run. Here are the data:

			Treatment	Replicate		
A	B	C	Combination	I	II	III
$-$	$-$	$-$	$a_1b_1c_1$	22	31	25
$+$	$-$	$-$	$a_2b_1c_1$	32	43	29
$-$	$+$	$-$	$a_1b_2c_1$	35	34	50
$+$	$+$	$-$	$a_2b_2c_1$	55	47	46
$-$	$-$	$+$	$a_1b_1c_2$	44	45	38
$+$	$-$	$+$	$a_2b_1c_2$	40	37	36
$-$	$+$	$+$	$a_1b_2c_2$	60	50	54
$+$	$+$	$+$	$a_2b_2c_2$	39	41	47

I used R to analyze these data as a 2^3 factorial experiment. The output is below. Recall R's convention; for example, `speed` denotes the main effect of speed; `speed:geometry` denotes the two-way interaction between speed and geometry, etc.

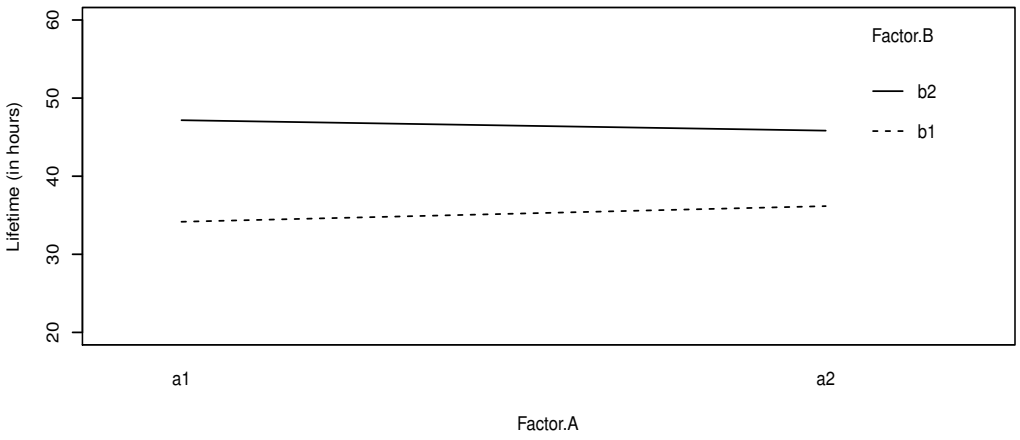
```
> fit = lm(lifetime ~ speed*geometry*angle)
> anova(fit)
Analysis of Variance Table
Response: lifetime
                     Df Sum Sq Mean Sq F value    Pr(>F)
speed                 1   0.67    0.67  0.0221 0.8836803
geometry              1 770.67  770.67 25.5470 0.0001173 ***
angle                 1 280.17  280.17  9.2873 0.0076787 **
speed:geometry        1  16.67   16.67  0.5525 0.4680784
speed:angle           1 468.17  468.17 15.5193 0.0011722 **
geometry:angle        1  48.17   48.17  1.5967 0.2244753
speed:geometry:angle  1  28.17   28.17  0.9337 0.3482825
Residuals            16 482.67   30.17
```

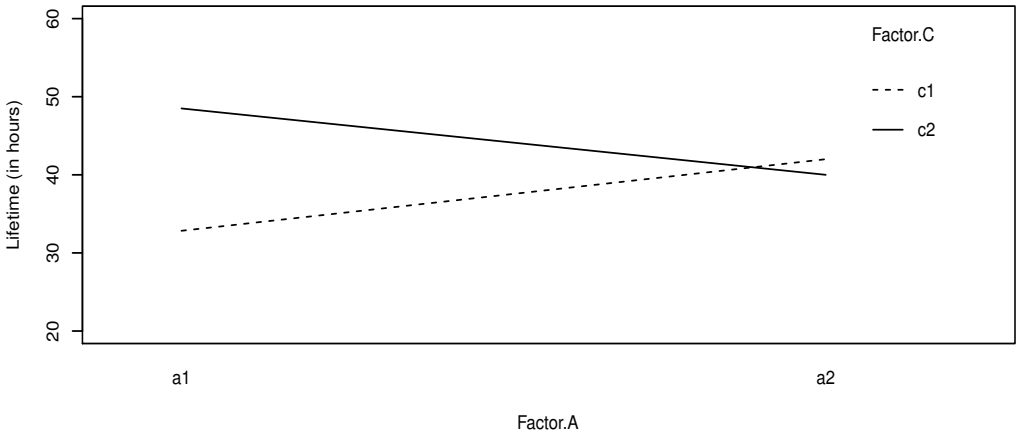
(a) Ignoring two- and three-way interactions, which main effects are significant in the population? Explain.
(b) Examine the two-way interaction plots on the next page. Are all two-way interactions significant in the population? Explain why or why not.
(c) I refit the model using only the main effects `speed` (A), `geometry` (B), and `angle` (C), and the two-way interaction effect `speed:angle` (AC). Here is the output:

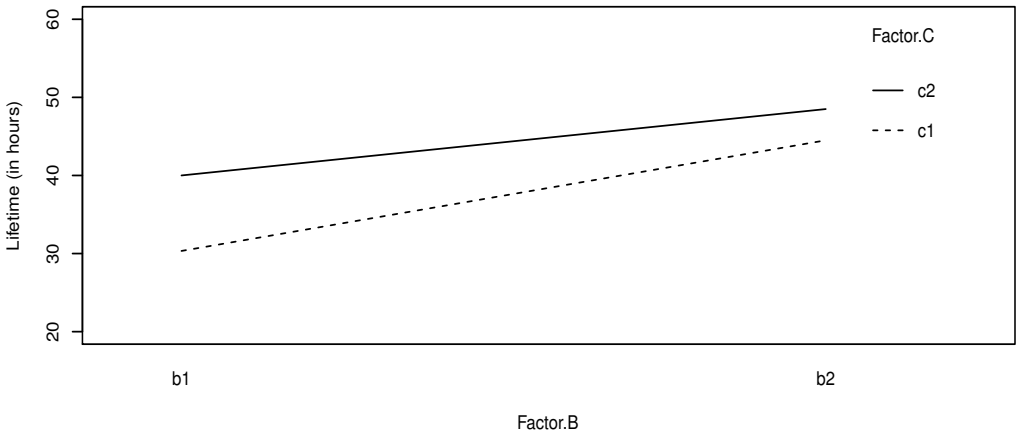
```
> anova(fit)
Analysis of Variance Table
Response: lifetime
            Df Sum Sq Mean Sq F value    Pr(>F)
speed        1   0.67    0.67   0.022 0.8836408
geometry     1 770.67  770.67  25.436 7.216e-05 ***
angle        1 280.17  280.17   9.247 0.0067238 **
speed:angle  1 468.17  468.17  15.452 0.0008972 ***
Residuals   19 575.67   30.30
```

Comparing this ANOVA table to the ANOVA table presented earlier, we see that the residual sum of squares has increased from 482.67 to 575.67. Explain where this increase comes from.

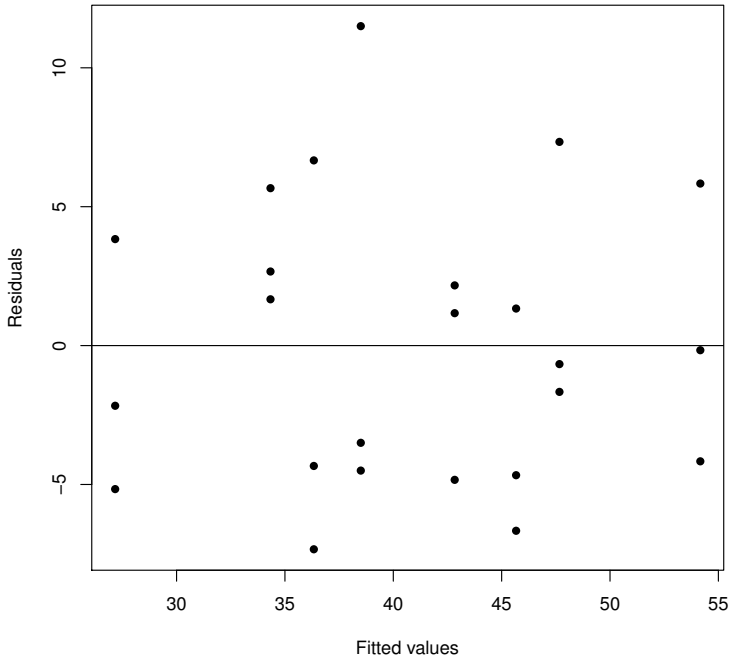
Part (d) is on Page 15.


Lifetime (in hours)
20
30
40
50
60
Factor.B
b2
b1
a1
a2
Factor.A


Lifetime (in hours)
20
30
40
50
60
Factor.C
c1
c2
a1
a2
Factor.A


Lifetime (in hours)
20
30
40
50
60
Factor.C
c2
c1
b1
b2
Factor.B

(d) After using regression to estimate the model in part (c), I examined the residual plot (i.e., a plot of the residuals versus the fitted values):



What does this plot say about the quality of the model fit in part (c)?