```
Regression Analysis

The regression equation is
P/E = 6.70 + 0.183 Profit + 0.213 Growth + 0.84 Oil + 3.82 Drug

Predictor        Coef       StDev        T        P
Constant        6.704       2.178     3.08    0.008
Profit         0.1827      0.2092     0.87    0.397
Growth         0.2128      0.1311     1.62    0.127
Oil             0.835       1.509     0.55    0.589
Drug            3.819       1.779     2.15    0.050

S = 2.309       R-Sq = 69.5%     R-Sq(adj) = 60.8%

Analysis of Variance

Source            DF         SS         MS        F        P
Regression         4    169.887     42.472     7.97    0.001
Residual Error    14     74.650      5.332
Total             18    244.537

Source       DF      Seq SS
Profit        1     131.398
Growth        1      13.194
Oil           1       0.724
Drug          1      24.570
```

Only the DRUG coefficient is significant at $\alpha = 0.05$.

(c) If we choose the drug/healthcare industry as the baseline, then we would instead have an indicator COMPUTER. There is no need to refit the model. The new coefficients, denoted with a subscript of N, depend on the previously fitted coefficients, denoted with a subscript of O, as below:

$$\text{Constant}_N = \text{Constant}_O + \text{Drug}_O = 6.704 + 3.819 = 10.523.$$

Since now the constant term represents the baseline industry, DRUG, the PROFIT and GROWTH coefficients will not change.

$$\text{OIL}_N = \text{OIL}_O - \text{DRUG}_O = 0.835 - 3.819 = -2.984.$$

$$\text{COMPUTER}_N = \text{COMPUTER}_O - \text{DRUG}_O = 0 - 3.819 = -3.819.$$

Therefore the new model would be

$$\hat{y} = 10.523 + 0.183 \text{ PROFIT } + 0.213 \text{ GROWTH } - 2.984 \text{ OIL } - 3.819 \text{ COMPUTER}.$$

11.40 (a) The partial correlation coefficients are

$$r_{yx_1|x_2} = \sqrt{\frac{\text{SSE}(x_2) - \text{SSE}(x_1, x_2)}{\text{SSE}(x_2)}} = \sqrt{\frac{12.606 - 5.988}{12.606}} = 0.725,$$

$$r_{yx_2|x_1} = \sqrt{\frac{\text{SSE}(x_1) - \text{SSE}(x_1, x_2)}{\text{SSE}(x_1)}} = \sqrt{\frac{13.519 - 5.988}{13.519}} = 0.746.$$

(b) The F statistics for testing the significance of these partial correlation coeficients are

$$F_1 = \frac{\text{SSE}(x_2) - \text{SSE}(x_1, x_2)}{\text{SSE}(x_2)/(n-3)} = \frac{12.606 - 5.988}{5.988/(40-3)} = 40.90,$$

$$F_2 = \frac{\text{SSE}(x_1) - \text{SSE}(x_1, x_2)}{\text{SSE}(x_1, x_2)/(n-3)} = \frac{13.519 - 5.988}{5.988/(40-3)} = 46.54.$$

Then

$$t_1 = \sqrt{40.90} = 6.395,$$
$$t_2 = \sqrt{46.54} = 6.822.$$

These the match the t statistics obtained in Exercise 11.2.

11.41 (a) $r_{yx_1} = 0.378$, $r_{yx_2} = -0.093$, and $r_{yx_3} = 0.003$. x_1 would enter first.

(b) The partial correlation coefficients are

$$r_{yx_2|x_1} = \sqrt{\frac{\text{SSE}(x_1) - \text{SSE}(x_1, x_2)}{\text{SSE}(x_1)}} = \sqrt{\frac{16198 - 13322}{16198}} = 0.421,$$

$$r_{yx_3|x_1} = \sqrt{\frac{\text{SSE}(x_1) - \text{SSE}(x_1, x_3)}{\text{SSE}(x_1)}} = \sqrt{\frac{16198 - 15258}{16198}} = 0.241.$$

(c) The F statistics for testing the significance of these partial correlation coeficients are

$$F_2 = \frac{\text{SSE}(x_1) - \text{SSE}(x_1, x_2)}{\text{SSE}(x_1, x_2)/(n-3)} = \frac{16198 - 13322}{13322/(38-3)} = 7.556,$$

$$F_3 = \frac{\text{SSE}(x_1) - \text{SSE}(x_1, x_3)}{\text{SSE}(x_1, x_3)/(n-3)} = \frac{16198 - 15258}{15258/(38-3)} = 2.156.$$

Height (x_2) is the better predictor given that brain size is included in the model.

11.42 (a) $r_{\log y, \log x_1} = -0.761$, $r_{\log y, \log x_2} = -0.549$, and $r_{\log y, x_3} = -0.644$. $\log x_1$ would enter first.

(b) The partial correlation coefficients are

$$r_{\log y, \log x_2|\log x_1} = \sqrt{\frac{\text{SSE}(\log x_1) - \text{SSE}(\log x_1, \log x_2)}{\text{SSE}(\log x_1)}} = \sqrt{\frac{6.112 - 5.815}{6.112}} = 0.220,$$

$$r_{\log y, x_3|\log x_1} = \sqrt{\frac{\text{SSE}(\log x_1) - \text{SSE}(\log x_1, x_3)}{\text{SSE}(\log x_1)}} = \sqrt{\frac{6.112 - 6.053}{6.112}} = 0.098.$$

(c) The F statistics for testing the significance of these partial correlation coeficients are

$$F_{\log x_2} = \frac{\text{SSE}(\log x_1) - \text{SSE}(\log x_1, \log x_2)}{\text{SSE}(\log x_1, \log x_2)/(n-3)} = \frac{6.112 - 5.815}{5.815/(38-3)} = 1.788,$$

$$F_{x_3} = \frac{\text{SSE}(\log x_1) - \text{SSE}(\log x_1, x_3)}{\text{SSE}(\log x_1, x_3)/(n-3)} = \frac{6.112 - 6.053}{6.053/(38-3)} = 0.341.$$

Log(calcium) is the better predictor, given that log(Alkalinity) is included in the model.

11.43

$$\begin{aligned} F_p &= \frac{(\text{SSE}_{p-1} - \text{SSE}_p)/1}{\text{SSE}_p/[n-(p+1)]} \\ &= \frac{(r^2_{yx_p|x_1,\ldots,x_{p-1}})(\text{SSE}_{p-1})}{\text{SSE}_p/[n-(k+1)]} \\ &= \frac{r^2_{yx_p|x_1,\ldots,x_{p-1}}[n-(k+1)]}{\text{SSE}_p/\text{SSE}_{p-1}} \\ &= \frac{r^2_{yx_p|x_1,\ldots,x_{p-1}}[n-(k+1)]}{1-r^2_{yx_p|x_1,\ldots,x_{p-1}}}. \end{aligned}$$

As $r^2_{yx_p|x_1,\ldots,x_{p-1}}$ increases, the numerator increases and the denominator decreases, so that F_p is an increasing function in $r^2_{yx_p|x_1,\ldots,x_{p-1}}$.

.44 (a) The stepwise regression output is shown below:

```
Stepwise Regression


 F-to-Enter:        2.00    F-to-Remove:       2.00

 Response is    P/E     on  4 predictors, with N =    19

     Step          1        2
Constant      10.309    8.319

Drug             5.6      4.7
T-Value         5.02     4.19

Growth                   0.23
T-Value                  1.97

S               2.41     2.23
R-Sq           59.73    67.60
```

Drug enters at step 1, and Growth enters at step 2. The final model is $\hat{y} = 8.319 + 4.7$ Drug $+ 0.23$ Growth . To determine if these factors are significant at $\alpha = 0.05$, compare the t statistics to $t_{19-3,0.025} = 2.120$. Only Drug is significant.

(b) The best subsets regression output is shown below:

```
Best Subsets Regression


Response is P/E Ratio

                                P G
                                r r
                                o o   D
```

Vars	R-Sq	Adj. R-Sq	C-p	s	fit	wth	0il	rug
1	59.7	57.4	3.5	2.4067				X
1	53.7	51.0	6.2	2.5798	X			
1	32.1	28.1	16.1	3.1247		X		
1	17.3	12.4	22.9	3.4500			X	
2	67.6	63.5	1.9	2.2254		X		X
2	63.5	59.0	3.7	2.3611	X			X
2	59.8	54.8	5.4	2.4788			X	X
2	59.1	54.0	5.7	2.4993	X	X		
2	53.8	48.0	8.2	2.6582	X		X	
3	68.8	62.6	3.3	2.2551	X	X		X
3	67.8	61.4	3.8	2.2908		X	X	X
3	63.7	56.5	5.6	2.4316	X		X	X
3	59.4	51.3	7.6	2.5719	X	X	X	
4	69.5	60.8	5.0	2.3091	X	X	X	X

According to the C_p criterion, the best subset includes the Growth and Drug factors, with a C_p of 1.9. This is identical to the model found in part (a).

11.45 The best subsets regression output is given below:

```
Best Subsets Regression

Response is log(y)
```

Vars	R-Sq	Adj. R-Sq	C-p	s	log(x1)	log(x2)	x3
1	57.9	56.8	2.4	0.41203	X		
1	41.4	39.8	16.7	0.48625			X
1	30.2	28.2	26.4	0.53099		X	
2	60.0	57.7	2.6	0.40759	X	X	
2	58.3	56.0	4.0	0.41588	X		X
2	43.2	40.0	17.1	0.48547		X	X
3	60.7	57.2	4.0	0.40988	X	X	X

Using the C_p criterion, the model with only $\log x_1$ has the lowest C_p. This is the same model that was selected in Exercise 11.19. The fitted model is $\log \hat{y} = 7.21 - 0.398 \log x_1$, where $\hat{\beta}_{\log x_1}$ was highly significant (P-value ≈ 0.000).

11.46 (a)

Variables in Model	SSE_p	p	Error d.f.	MSE_p	Adj. r_p^2	C_p
None	950	0	19	50	0	20
x_1	720	1	18	40	0.2	12.8
x_2	630	1	18	35	0.3	9.2
x_3	540	1	18	30	0.4	5.6
x_1, x_2	595	2	17	35	0.3	9.8
x_1, x_3	425	2	17	25	0.5	3
x_2, x_3	510	2	17	30	0.4	6.4
x_1, x_2, x_3	400	3	16	25	0.5	4

(b) Subsets (x_1, x_3) and (x_1, x_2, x_3) have the maximum adjusted r_p^2. Subset (x_1, x_3) has the minimum C_p. Choose (x_1, x_3) since it has less variables and the minimum C_p.

(c) x_3 gives the biggest reduction in SSE_p (no need to calculate partial F's for x_1, x_2, x_3). So it will be the first variable to enter the model. The F to enter for x_3 is

$$F_3 = \frac{950 - 540}{540/18} = 13.67.$$

Since $F_3 > F_{\text{IN}} = 4.0$, x_3 will enter the model.

(d) (x_1, x_3) gives the biggest reduction in $SSE_p(x_3)$ (no need to calculate partial F for x_2. The F to enter for x_1 is

$$F_{1|3} = \frac{SSE(x_3) - SSE(x_1, x_3)}{SSE(x_1, x_3)/(20-3)} = \frac{540 - 425}{425/17} = 4.6.$$

Since $F_{1|3} > F_{\text{IN}}$, x_1 will enter the model next. Its partial correlation is

$$r^2_{yx_1|x_3} = \frac{SSE(x_3) - SSE(x_1, x_3)}{SSE(x_3)} = \frac{540 - 425}{540} = 0.213.$$

(e) The F to remove for x_3 is

$$F_{3|1} = \frac{SSE(x_1) - SSE(x_1, x_3)}{SSE(x_1, x_3)/(20-3)} = \frac{720 - 425}{425/17} = 11.8.$$

Since $F_{3|1} > F_{\text{OUT}} = 4.0$, x_3 will not be removed.

(f) The F to enter for x_2 is

$$F_{2|1,3} = \frac{SSE(x_1, x_3) - SSE(x_1, x_2, x_3)}{SSE(x_1, x_2, x_3)/(20-4)} = \frac{425 - 500}{400/16} = 1.0.$$

Since $F_{2|1,3} < F_{\text{IN}} = 4.0$, x_2 will not enter, and the full model will not be chosen.

Chapter 12 Solutions

Solutions to Section 12.1

12.1 (a) **Sugar:**

$$s^2 = \frac{(2.138)^2 + (1.985)^2 + (1.865)^2}{3} = 3.996,$$

so that $s = \sqrt{3.996} = 1.999$ with $3 \times 19 = 57$ d.f. Using the critical value $t_{57,0.025} \approx 2.000$, the 95% CI's are :

$$\begin{aligned}
\text{Shelf 1} &: \ 4.80 \pm (2.000) \times \frac{1.999}{\sqrt{20}} = [3.906, 5.694] \\
\text{Shelf 2} &: \ 9.85 \pm (2.000) \times \frac{1.999}{\sqrt{20}} = [8.956, 10.744] \\
\text{Shelf 3} &: \ 6.10 \pm (2.000) \times \frac{1.999}{\sqrt{20}} = [5.206, 6.994].
\end{aligned}$$

Fiber:

$$s^2 = \frac{(1.166)^2 + (1.162)^2 + (1.277)^2}{3} = 1.447,$$

so that $s = \sqrt{1.447} = 1.203$ with $3 \times 19 = 57$ d.f. Using the critical value $t_{57,0.025} \approx 2.000$, the 95% CI's are :

$$\begin{aligned}
\text{Shelf 1} &: \ 1.68 \pm (2.000) \times \frac{1.203}{\sqrt{20}} = [1.142, 2.218] \\
\text{Shelf 2} &: \ 0.95 \pm (2.000) \times \frac{1.203}{\sqrt{20}} = [0.412, 1.488] \\
\text{Shelf 3} &: \ 2.17 \pm (2.000) \times \frac{1.203}{\sqrt{20}} = [1.632, 2.708].
\end{aligned}$$

Shelf 2 cereals are higher in sugar content than shelves 1 and 3, since the CI for shelf 2 is above those of shelves 1 and 3. Similarly, the shelf 2 fiber content CI is below that of shelf 3. So in general, shelf 2 cereals are higher in sugar and lower in fiber.

(b) **Sugar:**

$$\begin{aligned}
SSE &= 57 \times 3.996 = 227.80, \\
SSA &= n \sum \bar{y}_i^2 - N \bar{\bar{y}}^2 \\
&= 20[(4.80)^2 + (9.85)^2 + (6.10)^2] - 60 \times (6.92)^2 \\
&= 275.03.
\end{aligned}$$

Then the ANOVA table is below:

Analysis of Variance				
Source	SS	d.f.	MS	F
Shelves	275.03	2	137.5	34.41
Error	227.80	57	3.996	
Total	502.83	59		

Since $F > f_{2,57,0.05} = 3.15$, there do appear to be significant differences among the shelves in terms of sugar content.
Fiber:

$$
\begin{aligned}
SSE &= 57 \times 1.447 = 82.47, \\
SSA &= n\sum \bar{y}_i^2 - N\bar{\bar{y}}^2 \\
&= 20[(1.68)^2 + (0.95)^2 + (2.17)^2] - 60 \times (1.60)^2 \\
&= 15.08.
\end{aligned}
$$

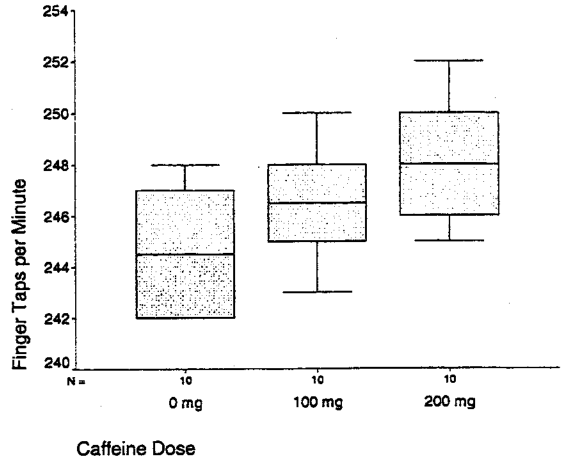
Then the ANOVA table is below:

Analysis of Variance				
Source	SS	d.f.	MS	F
Shelves	15.08	2	7.54	5.21
Error	82.47	57	1.447	
Total	97.55	59		

Since $F > f_{2,57,0.05} = 3.15$, there do appear to be significant differences among the shelves in terms of fiber content, as well as sugar content.

(c) The grocery store strategy is to place high sugar/low fiber cereals at the eye height of school children where they can easily see them.

12.2 (a)



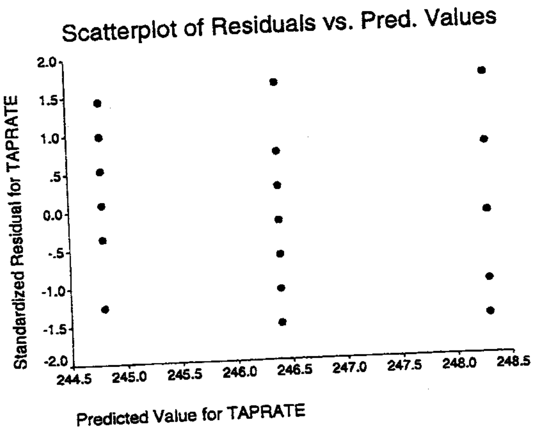
The boxplot indicates that the number of finger taps increases with higher doses of caffeine.

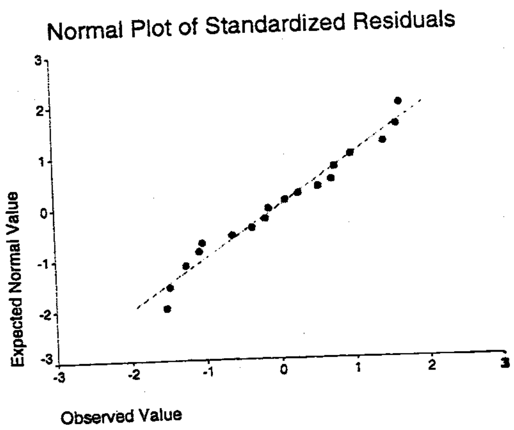
(b)

Analysis of Variance				
Source	SS	d.f.	MS	F
Dose	61.400	2	30.700	6.181
Error	134.100	27	4.967	
Total	195.500	29		

Since $F > f_{2,27,0.05} = 3.35$, there do appear to be significant differences in the numbers of finger taps for different doses of caffeine.

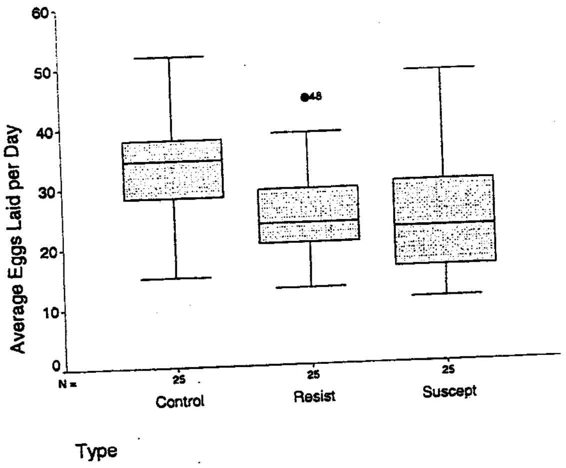
(c)





From the plot of the residuals against the predicted values, the constant variance assumption appears satisfied. From the normal plot of the residuals, the residuals appear to follow the normal distribution.

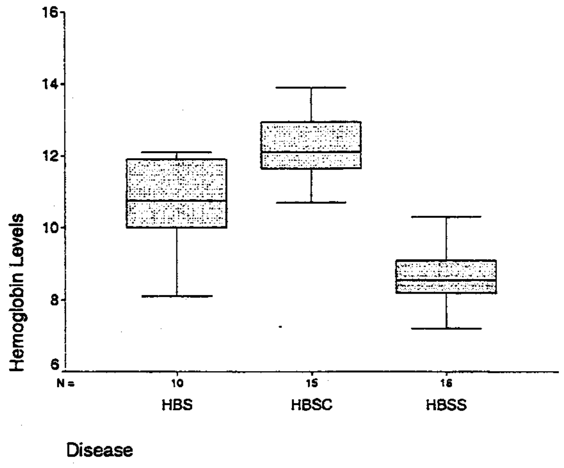
12.3 (a)



The boxplot indicates that the control is higher than the two treatments in the average number of eggs laid.

(b)

12.57(a)



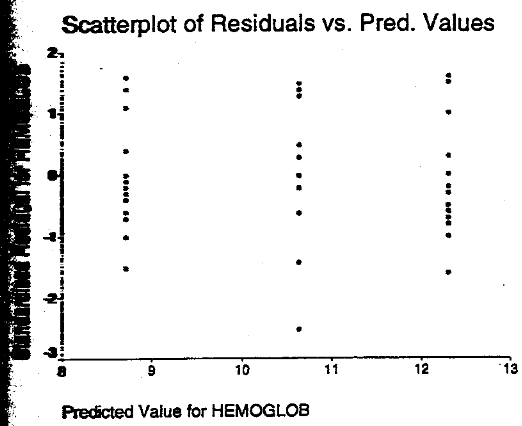
The boxplot indicates that HBSC has the highest average hemoglobin level, followed by HBS, and then HBSS.

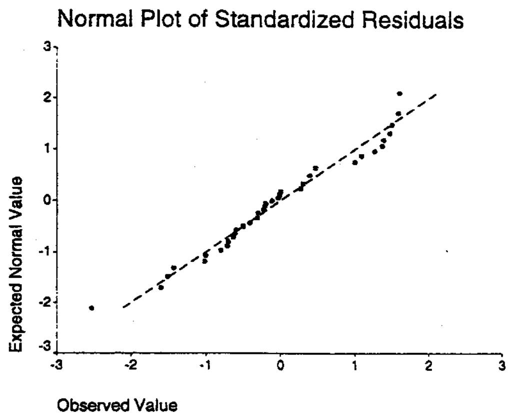
(b)

Analysis of Variance				
Source	SS	d.f.	MS	F
Disease type	99.889	2	49.945	49.999
Error	37.959	38	0.999	
Total	137.848	40		

Since $F > f_{2,38,0.05} = 3.23$, there do appear to be significant differences in the hemoglobin levels between patients with different types of sickle cell disease.

(c)

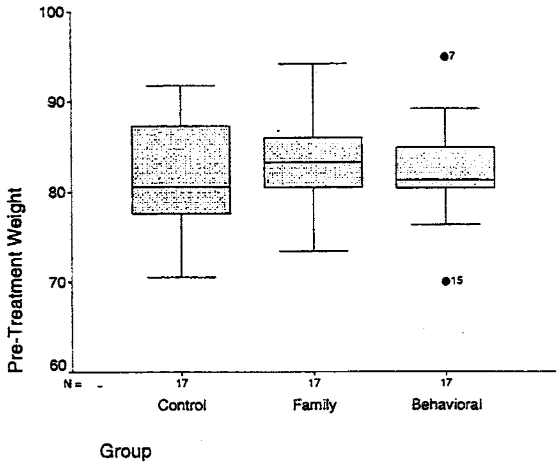




From the plot of the residuals against the predicted values, the constant variance assumption appears satisfied. From the normal plot of the residuals, the residuals appear to follow the normal distribution.

From the plot of the residuals against the predicted values, the constant variance assumption appears satisfied. From the normal plot of the residuals, the residuals appear to follow the normal distribution.

7 (a)



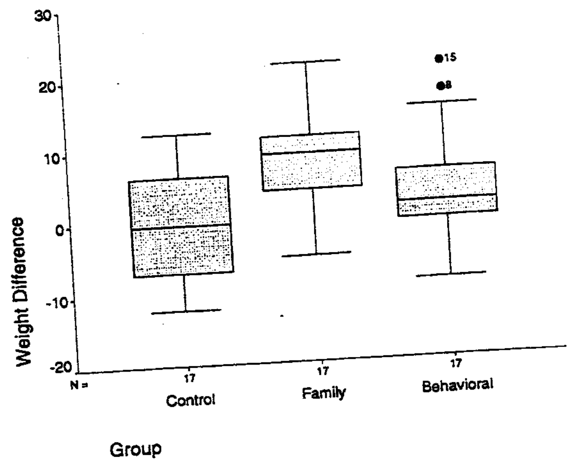
The side-by-side boxplots of the pre-treatment weights overlap quite a bit, and don't indicate large differences among the groups.

(b)

Analysis of Variance				
Source	SS	d.f.	MS	F
Treatment	31.2	2	15.6	0.50
Error	1503.5	48	31.0	
Total	1534.7	50		

Since $F < f_{2,48,0.05} = 3.19$, the pre-treatment weights are not significantly different among the different treatment groups.

(c)



Here the differences among the treatment groups are more pronounced. The family group appears to be different than the other two, but it is hard to tell if the differences are significant.

(d)

Analysis of Variance				
Source	SS	d.f.	MS	F
Treatment	479.3	2	239.7	3.85
Error	2990.6	48	62.3	
Total	3469.9	50		

Since $F > f_{2,48,0.05} = 3.19$, the weight differences are significantly different among the different treatment groups.

Solutions to Section 12.2

12.8 Sugar: The number of comparisons is

$$\binom{a}{2} = \binom{3}{2} = 3,$$

Then

$$t_{57,\frac{0.05}{2\times 3}} = t_{57,0.0083} = 2.468,$$

and the Bonferroni critical value is

$$t_{57,0.0083}s\sqrt{\frac{2}{n}} = 2.468 \times 1.999\sqrt{\frac{2}{20}} = 1.56.$$

For the Tukey method,

$$q_{3,57,0.05} \approx 3.40,$$

and the Tukey critical value is

$$q_{3,57,0.05}s\sqrt{\frac{1}{n}} = 3.40 \times 1.999\sqrt{\frac{1}{20}} = 1.52.$$

Since $t_{13} > 2.051$, conclude that sites 1 and 3 are significantly different. Similarly, since $t_{23} > 2.051$, conclude that sites 2 and 3 are significantly different. So the conclusion is the same as with the Tukey method, namely that all three sites are significantly different from one another.

12.10 The number of comparisons is

$$\binom{a}{2} = \binom{3}{2} = 3,$$

Then

$$t_{38,\frac{0.01}{2\times 3}} = t_{38,0.0017} = 3.136,$$

and, since $s = \sqrt{0.999} = 1$, the form of the Bonferroni confidence interval is

$$\bar{y}_i - \bar{y}_j \pm (3.136)(1)\sqrt{\frac{1}{n_i} + \frac{1}{n_j}}.$$

For the Tukey method,

$$\frac{q_{3,38,0.01}}{\sqrt{2}} \approx \frac{4.39}{\sqrt{2}} = 3.104,$$

and the form of the Tukey confidence interval is

$$\bar{y}_i - \bar{y}_j \pm (3.104)(1)\sqrt{\frac{1}{n_i} + \frac{1}{n_j}}.$$

For the LSD method, $t_{38,0.01/2} = 2.712$, and the form of the LSD confidence interval is

$$\bar{y}_i - \bar{y}_j \pm (2.712)(1)\sqrt{\frac{1}{n_i} + \frac{1}{n_j}}.$$

The 99% confidence intervals are summarized in the table below:

		Bonferroni		Tukey		LSD	
Comparison	$\lvert\bar{y}_i - \bar{y}_j\rvert$	Lower	Upper	Lower	Upper	Lower	Upper
HBSC(3) vs. HBS(2)	1.670	0.392	2.948	0.406	2.934	0.564	2.776
HBS(2) vs. HBSS(1)	1.918	0.656	3.179	0.669	3.166	0.825	3.010
HBSC(3) vs. HBSS(1)	3.588	2.462	4.713	2.475	4.700	2.614	4.561

Since all of these intervals are entirely above 0, all of the types of disease have significantly different hemoglobin levels from one another. Note that the Bonferroni method has the widest intervals, followed by the Tukey. The LSD intervals are narrowest because there is no adjustment for multiplicity.

12.11 The ANOVA output, including Tukey 90% confidence intervals, is given below:

```
One-way Analysis of Variance

Analysis of Variance for Height
Source     DF        SS         MS         F        P
Part        3     81.11      27.04      3.95    0.010
```

```
Error      103    705.67       6.85
Total      106    786.79
                                       Individual 95% CIs For Mean
                                       Based on Pooled StDev
Level        N      Mean     StDev  ---------+---------+---------+-------
Bass1       39    70.718     2.361              (----*-----)
Bass2       26    71.385     2.729                 (------*------)
Tenor1      21    68.905     3.330   (------*-------)
Tenor2      21    69.905     2.071         (-------*-------)
                                    ---------+---------+---------+-------
Pooled StDev =     2.617                    69.0      70.5      72.0

Tukey's pairwise comparisons

    Family error rate = 0.100
Individual error rate = 0.0224

Critical value = 3.28

Intervals for (column level mean) - (row level mean)

             Bass1       Bass2       Tenor1

  Bass2     -2.204
             0.870

  Tenor1     0.170       0.699
             3.456       4.261

  Tenor2    -0.830      -0.301      -2.873
             2.456       3.261       0.873
```

The only significant difference is that Tenor 1 men are shorter than both Bass 1 and Bass 2 men on average, since those confidence intervals are entirely above 0.

12.12 Since we suspect that caffeine will increase the rate of finger taps, the one-sided Dunnett critical value is

$$t_{a-1,\nu,\alpha} = t_{2,27,0.10} \approx 1.625.$$

Then the Dunnett lower confidence bounds, using $s = \sqrt{4.97} = 2.23$, are

$$\begin{aligned}
\mu_i - \mu_1 &\geq \bar{y}_i - \bar{y}_1 - t_{2,27,0.10}s\sqrt{2/n} \\
\mu_2 - \mu_1 &\geq 246.40 - 244.80 - (1.625)(2.23)\sqrt{2/10} = -0.021 \\
\mu_3 - \mu_1 &\geq 248.30 - 244.80 - (1.625)(2.23)\sqrt{2/10} = 1.879.
\end{aligned}$$

Only the 200 mg dose is significantly higher than the control, since the confidence interval is entirely above 0.

12.13

Player	l_i	u_i
Jordan	$\lvert 3.54-5.82-1.668\rvert^- = -3.948$	$\lvert 3.54-5.82+1.668\rvert^+ = 0$
Rodman	$\lvert 2.55-5.82-1.668\rvert^- = -4.938$	$\lvert 2.55-5.82+1.668\rvert^+ = 0$
Kukoc	$\lvert 5.82-3.54-1.668\rvert^- = 0$	$\lvert 5.82-3.54+1.668\rvert^+ = 3.948$
Longley	$\lvert 3.09-5.82-1.668\rvert^- = -4.398$	$\lvert 3.09-5.82+1.668\rvert^+ = 0$
Harper	$\lvert 2.82-5.82-1.668\rvert^- = -4.668$	$\lvert 2.82-5.82+1.668\rvert^+ = 0$

Since Kukoc is the only player with $l_i = 0$, he does appear to have the most assists.

Solutions to Section 12.3

12.18 (a) Since

$$MSE = \frac{s_1^2 + \ldots + s_{10}^2}{10} = 10.533,$$

then

$$\begin{aligned} SSE &= MSE \times (N-a) = 10.533 \times (120-10) = 1158.63, \\ SSA &= n\sum \bar{y}_i^2 - N\bar{\bar{y}}^2 \\ &= 12[(36.76)^2 + \ldots + (32.98)^2] - 120 \times 34.454 \\ &= 279.932, \text{ and} \\ SST &= SSE + SSA = 1158.63 + 279.932 = 1438.562. \end{aligned}$$

Then the ANOVA table is given below:

Analysis of Variance				
Source	SS	d.f.	MS	F
Batch	279.932	9	31.104	2.953
Error	1158.630	110	10.533	
Total	1438.562	119		

(b) The variance components estimates are

$$\begin{aligned} \hat{\sigma}^2_{\text{Error}} &= MSE = 10.533 \text{ and} \\ \hat{\sigma}^2_{\text{Batch}} &= \frac{MSA - MSE}{n} \\ &= \frac{31.104 - 10.533}{12} = 1.714. \end{aligned}$$

Batch to batch variation accounts for about

$$\frac{1.714}{1.714 + 10.533} = 14\%$$

of the total error variation.

12.19 (a)

$$\text{Var}(\bar{Y}) = \text{Var}\left(\frac{1}{n}\sum_{j=1}^{n} Y_{ij}\right)$$

$$= \operatorname{Var}\left(\frac{1}{n}\sum_{j=1}^{n}(\mu+\tau_i+\epsilon_{ij})\right)$$

$$= \operatorname{Var}\left(\mu+\tau_i+\frac{1}{n}\sum_{j=1}^{n}\epsilon_{ij}\right)$$

$$= \operatorname{Var}(\tau_i)+\frac{1}{n^2}\sum_{j=1}^{n}\operatorname{Var}(\epsilon_{ij})$$

$$= \sigma^2_{\text{Batch}}+\frac{\sigma^2_{\text{Error}}}{n}.$$

(b) Using

$$\hat{\sigma}_Y=\sqrt{\hat{\sigma}^2_{\text{Batch}}+\frac{\hat{\sigma}^2_{\text{Error}}}{12}}=\sqrt{1.714+\frac{10.533}{12}}=\sqrt{2.592}=1.610,$$

the three sigma control limits are

$$34.454 \pm 3(1.610) = [29.624, 39.284].$$

Since all the batch means fall within these limits, the process is under control.

12.20 From the ANOVA table given in Exercise 12.16, the variance components estimates are

$$\hat{\sigma}^2_{\text{Error}} = MSE = 26.53 \text{ and}$$

$$\hat{\sigma}^2_{\text{Cable}} = \frac{MSA-MSE}{n} = \frac{240.54-26.53}{12} = 17.834.$$

Cable to cable variation accounts for about

$$\frac{17.834}{17.834+26.53} = 40.2\%$$

of the total error variation.

Solutions to Section 12.4

12.21 (a) Since the overall mean is 86, the effect estimates are

Quantity	Estimate	Quantity	Estimate
τ_A	$84-86=-2$	β_1	$92-86=6$
τ_B	$85-86=-1$	β_2	$83-86=-3$
τ_C	$89-86=3$	β_3	$85-86=-1$
τ_D	$86-86=0$	β_4	$88-86=2$
		β_5	$82-86=-4$

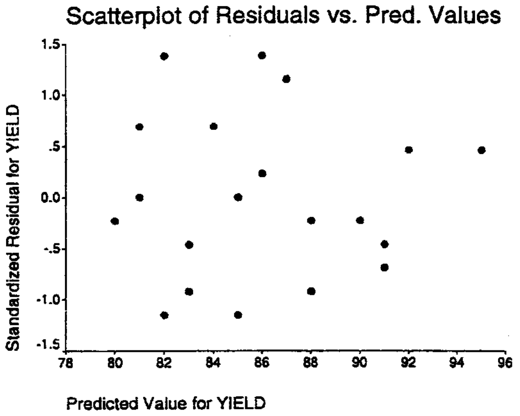
(b)

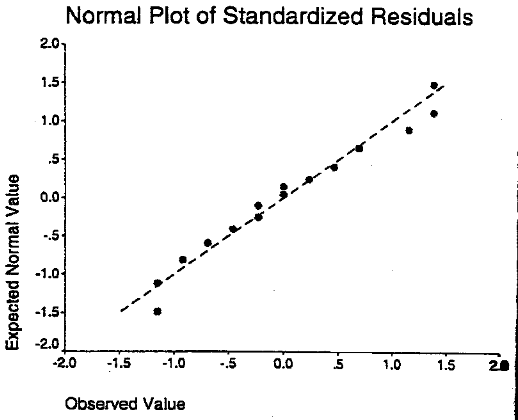
Analysis of Variance

Source	SS	d.f.	MS	F
Blend	264	4	66.000	3.504
Method	70	3	23.333	1.239
Error	226	12	18.833	
Total	560	19		

For method, $F < f_{3,12,0.05} = 3.49$, so there do not appear to be any significant differences between methods. For blend, $F > f_{4,12,0.05} = 3.26$, so there do appear to be significant differences between blends.

(c)





There is possibly a decreasing variance with increasing yield, but this is unclear because there are also less observations at higher yields, so the distribution is less likely to be filled out. From the normal plot of the residuals, the residuals appear to follow the normal distribution.

12.22 (a) The ANOVA output is given below:

```
General Linear Model


Factor      Type Levels Values
Player     fixed      6 Becker  Cole    Cordova Mack     Munoz   Puckett
Stadium    fixed      3 Home     Other    Outdoors

Analysis of Variance for Zone Rat, using Adjusted SS for Tests

Source      DF       Seq SS      Adj SS      Adj MS        F       P
Player       5  0.0148578  0.0148578  0.0029716   23.79  0.000
Stadium      2  0.0016591  0.0016591  0.0008296    6.64  0.015
Error       10  0.0012489  0.0012489  0.0001249
Total       17  0.0177658

Least Squares Means for Zone Rat

 Stadium       Mean     StDev
Home         0.8007  0.004562
Other        0.8020  0.004562
```

```
Outdoors     0.8217  0.004562

Tukey 95.0% Simultaneous Confidence Intervals
Response Variable Zone Rat
All Pairwise Comparisons among Levels of Stadium

Stadium = Home subtracted from:

Stadium       Lower     Center       Upper   -+---------+---------+---------+-----
Other      -0.01637  0.001333    0.01904   (-----------*-----------)
Outdoors    0.00330  0.021000    0.03870              (-----------*-----------)
                                            -+---------+---------+---------+-----
                                          -0.015     0.000     0.015     0.030

Stadium = Other subtracted from:

Stadium       Lower     Center       Upper   -+---------+---------+---------+-----
Outdoors   0.001965    0.01967     0.03737              (-----------*-----------)
                                            -+---------+---------+---------+-----
                                          -0.015     0.000     0.015     0.030
```

Since the P-value for Stadium is $0.015 < \alpha = 0.05$, conclude that there are significant differences between the stadiums.

(b) The Tukey confidence intervals are given in the computer output from (a). From whether these confidence intervals contain 0 or not, we can see that Outdoor stadiums are significantly different from both Home Domes and Other Domes, but Home Domes are not significantly different from Other Domes.

23 The estimated contrast is

$$c = \frac{1}{2}(\bar{y}_1 + \bar{y}_2) - \bar{y}_3 = \frac{1}{2}(0.8007 + 0.802) - 0.8217 = -0.020,$$

with a standard error of

$$\text{s.e.}(c) = \sqrt{s^2 \sum_i \left(\frac{c_i^2}{n_i}\right)} = \sqrt{0.000125\left(\frac{(0.5)^2 + (0.5)^2 + (-1)^2}{6}\right)} = 0.0056.$$

Then the test statistic is

$$t = \frac{-0.020}{0.0056} = -3.578.$$

Since $|t| > t_{10,0.025} = 2.228$, conclude that there is a significant difference between domed stadiums and outdoor stadiums.

24 (a) The ANOVA output is given below:

Analysis of Variance				
Source	SS	d.f.	MS	F
Park	0.170	1	0.170	14.130
Year	1.064	7	0.152	12.627
Error	0.084	7	0.012	
Total	1.319	15		

Since $F_{\text{Park}} > f_{1,7,0.05} = 5.59$, there do appear to be significant differences between new and unchanged parks in their HR/G ratios.

12.28 (a) If batches were ignored as a blocking factor, the SS and d.f. for batches would both be included in the error SS and d.f.

(b) The new ANOVA table would be

Analysis of Variance				
Source	SS	d.f.	MS	F
Position	40.396	7	5.771	3.614
Error	25.549	16	1.597	
Total	65.945	23		

Since $F < f_{7,16,0.01} = 4.03$, the positions are not significantly different from one another, when the batches are ignored as a blocking factor.

(c) A nonsignificant result is obtained in (b) because the variance associated with blocks was no longer removed and was included in the error variance. This raised the denominator, reduced the F statistic, and made it more difficult to detect differences among the positions. However, this will not always happen. If the MS_{Batches} is smaller than MSE, including the batch to batch variation in the error term will decrease the overall MSE and have the opposite effect. Also, including the batch d.f. in the error d.f. will make the critical F value smaller, and make it easier to reject H_0.

12.29 Since there are 6 blocks, $n = 6$ observations per variety. Also, since

$$s = \sqrt{79.64} = 8.924,$$

then

$$d = t_{7-1,30,0.10}\, s\sqrt{\frac{2}{n}} = 2.046 \times 8.924 \times \sqrt{\frac{2}{6}} = 10.542.$$

Using

$$l_i = |\bar{y}_i - \max_{j \neq i} \bar{y}_j - d|^-,$$

$$u_i = |\bar{y}_i - \max_{j \neq i} \bar{y}_j + d|^+,$$

the results are summarized in the table below:

Variety	l_i	u_i
A	$\|49.6 - 71.3 - 10.542\|^- = -32.242$	$\|49.6 - 71.3 + 10.542\|^+ = 0$
B	$\|71.2 - 71.3 - 10.542\|^- = -10.642$	$\|71.2 - 71.3 + 10.542\|^+ = 10.442$
C	$\|67.6 - 71.3 - 10.542\|^- = -14.242$	$\|67.6 - 71.3 + 10.542\|^+ = 6.842$
D	$\|61.5 - 71.3 - 10.542\|^- = -20.342$	$\|61.5 - 71.3 + 10.542\|^+ = 0.742$
E	$\|71.3 - 71.2 - 10.542\|^- = -10.442$	$\|71.3 - 71.2 + 10.542\|^+ = 10.642$
F	$\|58.1 - 71.3 - 10.542\|^- = -23.742$	$\|58.1 - 71.3 + 10.542\|^+ = 0$
G	$\|61.0 - 71.3 - 10.542\|^- = -20.842$	$\|61.0 - 71.3 + 10.542\|^+ = 0.242$