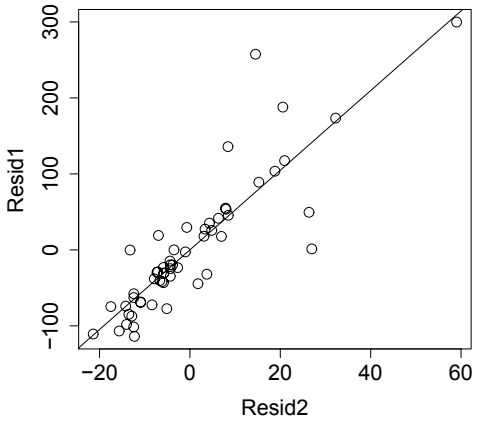
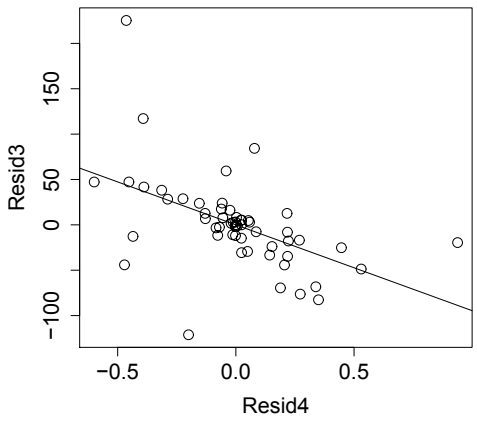
Chapter 4 Solutions

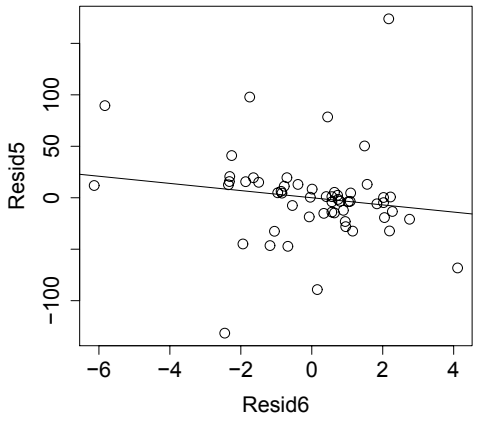
4.1 a. We start with a model to predict $Weight$ based on $Length$ and $Width$ alone and save the residuals in a new variable called $Resid1$. We expect that the interaction term $Length \cdot Width$ is strongly related to these two predictors, so we build a model to predict $Length \cdot Width$ using $Length$ and $Width$. We save the residuals from this second model in $Resid2$. To see what is unique in the interaction, $Length \cdot Width$, that can help predict what $Length$ and $Width$ haven't predicted about weight, we plot $Resid1$ versus $Resid2$. The result is the added variable plot shown as follows. We see a strong, positive association, indicating that the $Length \cdot Width$ interaction would be a valuable addition to this model to predict perch weights.



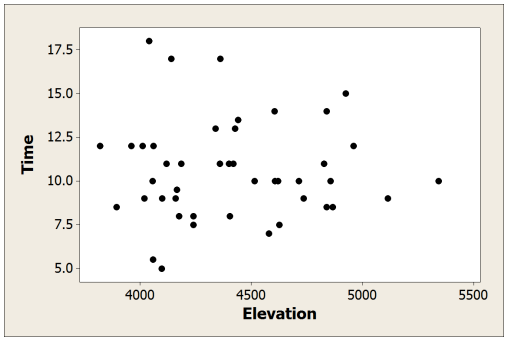
b. We start with a model to predict $Weight$ based on the other two variables, $Length$ and $Length \cdot Width$, and save the residuals in a new variable called $Resid3$. We build a second model to predict $Width$ using $Length$ and $Length \cdot Width$. We save the residuals from this second model in $Resid4$. To see what is unique in $Width$, that can help predict what $Length$ and $Length \cdot Width$ haven't predicted about weight, we plot $Resid3$ versus $Resid4$. The result is the added variable plot shown below. We see a fairly strong, negative association, which is consistent with a negative coefficient for $Width$ in the three-predictor model to predict $Weight$ and the fact that $Width$ is a important predictor in that model.



c. We start with a model to predict $Weight$ based on the other two variables, $Width$ and $Length \cdot Width$, and save the residuals in a new variable called $Resid5$. We build a second model to predict $Length$ using $Width$ and $Length \cdot Width$. We save the residuals from this second model in $Resid6$. To see what is unique in $Length$, that can help predict what $Width$ and $Length \cdot Width$ haven't predicted about weight, we plot $Resid5$ versus $Resid6$. The result is the added variable plot shown below. We see a very weak, negative trend. This is consistent with the fact that $Length$ has a negative coefficient, but is not a very effective predictor in the three-predictor model to predict $Weight$.



4.2 a. A scatterplot of $Time$ versus $Elevation$ is shown below and the correlation between these variables is $r = -0.016$. This is a very small correlation and we see no association in the plot, so $Elevation$ would be a very poor predictor of hike $Time$ on its own.

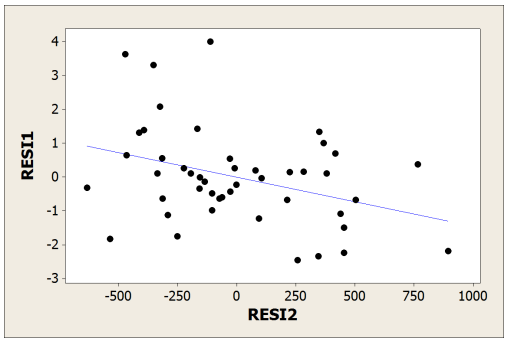


b. Some regression output for fitting the two-predictor model is shown below. The P-values for *Elevation* (0.017) and *Length* (0.000) are both small, indicating that both predictors are helpful in this model to predict hike *Time*.

```
The regression equation is
Time = 8.08 - 0.00145 Elevation + 0.712 Length

Predictor        Coef    SE Coef      T      P
Constant        8.075      2.533   3.19  0.003
Elevation  -0.0014483  0.0005805  -2.49  0.017
Length        0.71233    0.05933  12.01  0.000

S = 1.37020   R-Sq = 77.0%   R-Sq(adj) = 76.0%
```

The value of R^2 for this two-predictor model (77.0%) is somewhat better than for *Length* alone ($R^2 = 73.7\%$) and much better than for *Elevation* alone ($R^2 = 0.0\%$).

c. We start with a model to predict *Time* based on *Length* and save the residuals in a new variable called $RESI1$. Next, we fit a model to predict *Elevation* based on *Length* and save the residuals in $RESI2$. To see what is unique in *Elevation* that can help predict variability in *Time* that *Length* doesn't explain, we plot $RESI1$ versus $RESI2$. The result is the following added variable plot. We see some negative association in this plot, which indicates that *Elevation* is a useful predictor of *Time* after accounting for trip *Length*.



4.3 a. After four steps of forward selection, the four-predictor model for predicting *WinPct* includes *ERA*, *Runs*, *Saves*, and *WHIP*. The R^2 for this model is 88.63.%. Partial software output is shown as follows.

	Step 1		Step 2		Step 3		Step 4	
	Coef	P	Coef	P	Coef	P	Coef	P
Constant	0.9845		0.6039		0.378		0.516	
ERA	-0.1158	0.000	-0.1059	0.000	-0.0856	0.000	-0.0364	0.178
Runs			0.000468	0.000	0.000528	0.000	0.000519	0.000
Saves					0.002307	0.003	0.002643	0.001
WHIP							-0.266	0.047
S	0.0406051		0.0298312		0.0255379		0.0240377	
R-sq	63.65%		81.08%		86.65%		88.63	
R-sq(adj)	62.35%		79.68%		85.11%		86.81%	
R-sq(pred)	59.36%		77.52%		81.26%		83.07%	
Mallows Cp	73.59		27.83		14.58		11.16	

b. Using backward elimination the predictors are dropped from the original set of all 16 predictors in this order: *Doubles*, *SB*, *SLG*, *Triples*, *Strikeouts*, *HR*, *RBI*, *ERA*, *OBP*, *HitsAllowed*, *Walks*, and *Hits*. This leaves a four-predictor model for predicting *WinPct* with *BattingAvg*, *Runs*, *Saves*, and *WHIP*. The R^2 for this model is 88.36%.

c. Using best subsets, a four-predictor model for predicting *WinPct* includes *Runs*, *Doubles*, *ERA*, and *WHIP*. The R^2 for this model is 88.9%. Partial software output is shown below.

Vars	R-Sq	R-Sq(adj)	Mallows Cp	S	BattingAvg	Runs	Hits	Doubles	Triples	HR	RBI	SB	Walks	ERA	Strikeouts	OBP	SLG	HitsAllowed	WHIP	Saves
1	63.6	62.4	73.6	0.040605										X						
2	81.1	79.7	27.8	0.029831		X								X						
3	87.8	86.3	11.6	0.024458		X								X						X
4	88.9	87.1	10.5	0.023796		X		X						X						X

d. Mallow's Cp is 11.16, 11.88, and 10.54 for the models in parts (a), (b), and (c), respectively.

e. The model identified in part (c) is the best four-variable model because it has the lowest Cp and the highest R^2. However, it does have a fairly weak predictor (*WHIP*, P-value = 0.181), so a three-predictor model might be preferable. Note: It is reasonable to compare R^2 values for these three models because they all contain four variables. When comparing models with different numbers of predictor variables, adjusted R^2 should be used.

4.4 a. Here are the correlations (with P-values following) for $MeanAFC$ with each of the potential predictors:

	Age	FSH	E2	MaxE2	MaxDailyGn	TotalGn	Oocytes	Embryos
MeanAFC	-0.230	-0.296	-0.127	0.246	-0.397	-0.384	0.417	0.346
	0.000	0.000	0.020	0.000	0.000	0.000	0.000	0.000

The strongest correlation with $MeanAFC$ is *Oocytes* ($r = 0.417$), the weakest correlation is $E2$ ($r = -0.127$).

b. We test $H_0 : \rho = 0$ versus $H_a : \rho \neq 0$, where ρ is the correlation for the population between $MeanAFC$ and the weakest predictor $E2$. From the output in part (a), we see that the P-value for this test is 0.020, which gives significant evidence (at a 5% level) for some relationship between average antral follicle counts and fertility levels ($E2$). All of the other potential predictors are significant with P-values that are essentially zero (to three decimal places).

c. Here is some computer output from a best subsets regression procedure using this set of predictors for $MeanAFC$.

Response is MeanAFC

									M						
									a						
									x						
									D	T	O	E			
									a	o	o	m			
								M	i	t	c	b			
								a	l	a	y	r			
							A	F	x	y	l	t	y		
			Mallows				g	S	E	E	G	G	e	o	
Vars	R-Sq	R-Sq(adj)	Cp	S			e	H	2	2	n	n	s	s	
1	17.4	17.2	42.3	6.7602								X			
1	15.7	15.5	49.8	6.8284							X				
2	26.0	25.5	5.8	6.4105							X		X		
2	25.4	25.0	8.1	6.4325								X	X		
3	26.9	26.2	3.7	6.3798					X		X		X		
3	26.5	25.9	5.2	6.3953				X			X		X		

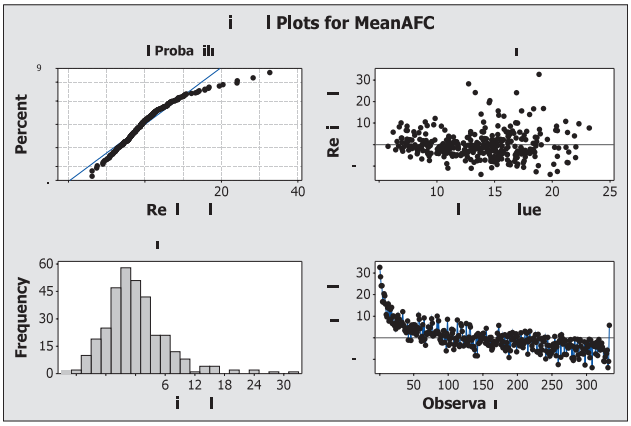
```
4  27.6       26.7      2.4  6.3577    X X   X   X
4  27.1       26.2      4.8  6.3812      X   X X X
5  27.8       26.7      3.6  6.3596    X X   X   X X
5  27.7       26.6      4.0  6.3639    X X   X X X
6  27.9       26.5      5.2  6.3659    X X   X X X X
6  27.8       26.5      5.5  6.3681  X X X   X   X X
7  27.9       26.3      7.1  6.3744  X X X   X X X X
7  27.9       26.3      7.1  6.3745    X X X X X X X
8  27.9       26.1      9.0  6.3832  X X X X X X X X
```

Looking at the first model with three predictors, we find that they are $E2$, $MaxDailyGn$, and $Oocytes$. Together, these three predictors explain 26.9% of the variability in $MeanAFC$.

d. Here is some output for fitting the model for $MeanAFC$ based on $E2$, $MaxDailyGn$, and $Oocytes$.

```
The regression equation is
MeanAFC = 16.9 - 0.0474 E2 - 0.0199 MaxDailyGn + 0.402 Oocytes

Predictor        Coef   SE Coef      T      P
Constant       16.917     1.800   9.40  0.000
E2           -0.04744   0.02320  -2.05  0.042
MaxDailyGn  -0.019904  0.003154  -6.31  0.000
Oocytes        0.40170   0.06216   6.46  0.000

S = 6.37983   R-Sq = 26.9%   R-Sq(adj) = 26.2%
```

We see that the coefficients for each of the predictors are significant at a 5% level (although just barely so for $E2$ with a P-value = 0.042). However, residual plots for fitting this model show some mild concerns.



There are several large residuals that deviate from a normal pattern in both the normal probability plot and somewhat right-skewed histogram of the residuals. These large positive residuals also appear in the residuals versus fits plots and seem to be associated with above-average fitted values, showing slightly increased variability. There is an interesting decreasing pattern when residuals are plotted versus the observations order, but this is explained by the fact that the data (except for the last case) have been ordered by decreasing $MeanAFC$ value in the original dataset.

4.5 a. Here are the correlations (with P-values below) for $LowAFC$ with each of the potential predictors:

	Age	FSH	E2	MaxE2	MaxDailyGn	TotalGn	Oocytes	Embryos
LowAFC	-0.238	-0.325	-0.125	0.247	-0.432	-0.416	0.381	0.338
	0.000	0.000	0.023	0.000	0.000	0.000	0.000	0.000

The strongest correlation with $LowAFC$ is $MaxDailyGn$ ($r = -0.432$); the weakest correlation is $E2$ ($r = -0.125$).

b. We test $H_0 : \rho = 0$ versus $H_a : \rho \neq 0$, where ρ is the correlation for the population between $LowAFC$ and the weakest predictor $E2$. From the output in part (a), we see that the P-value for this test is 0.023, which gives significant evidence (at a 5% level) for some relationship between smallest antral follicle counts and fertility levels ($E2$). All of the other potential predictors are significant with P-values that are essentially zero (to three decimal places).

c. Here is some computer output from a best subsets regression procedure using this set of predictors for $LowAFC$.

Response is LowAFC

										M					
										a					
										x					
										D	T	O	E		
										a	o	o	m		
									M	i	t	c	b		
									a	l	a	y	r		
								A	F	x	y	l	t	y	
			Mallows					g	S	E	E	G	G	e	o
Vars	R-Sq	R-Sq(adj)	Cp	S				e	H	2	2	n	n	s	s
1	18.7	18.4	41.3	6.2510								X			
1	17.3	17.0	47.6	6.3039									X		
2	26.1	25.6	9.6	5.9688								X			X
2	25.2	24.7	13.6	6.0042									X	X	
3	27.1	26.4	7.0	5.9366						X		X			X
3	27.1	26.4	7.1	5.9379					X			X			X

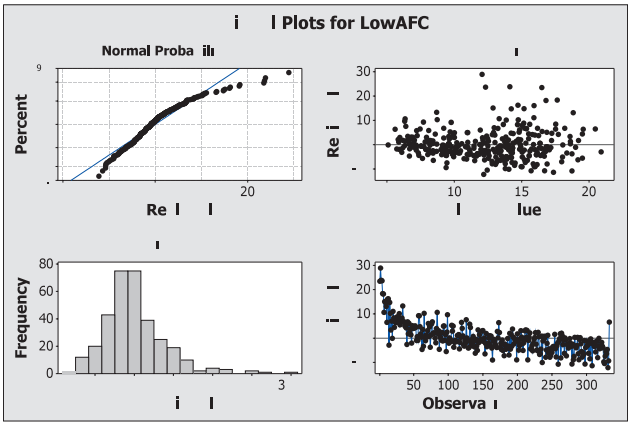
```
4  28.3      27.4      3.5  5.8967    X X   X   X
4  27.7      26.8      6.3  5.9213    X X   X     X
5  28.8      27.7      3.3  5.8853    X X   X   X X
5  28.3      27.2      5.3  5.9033    X X   X X X
6  28.8      27.5      5.0  5.8919    X X   X X X X
6  28.8      27.5      5.3  5.8942    X X X X   X X
7  28.8      27.3      7.0  5.9009    X X X X X X X
7  28.8      27.3      7.0  5.9010  X X X   X X X X
8  28.8      27.1      9.0  5.9099  X X X X X X X X
```

Looking at the first model with three predictors, we find that they are $E2$, $MaxDailyGn$, and $Oocytes$. Together, these three predictors explain 27.1% of the variability in $LowAFC$.

d. Here is some output for fitting the model for $LowAFC$ based on $E2$, $MaxDailyGn$, and $Oocytes$.

```
The regression equation is
LowAFC = 17.1 - 0.0463 E2 - 0.0215 MaxDailyGn + 0.316 Oocytes

Predictor        Coef   SE Coef      T      P
Constant       17.143     1.675  10.24  0.000
E2           -0.04625   0.02159  -2.14  0.033
MaxDailyGn  -0.021501  0.002934  -7.33  0.000
Oocytes       0.31551   0.05784   5.45  0.000

S = 5.93661   R-Sq = 27.1%   R-Sq(adj) = 26.4%
```

We see that the coefficients for each of the predictors are significant at a 5% level (although not strongly so for $E2$ with a P-value = 0.033). However, residual plots for fitting this model show some mild concerns.



There are several large residuals that deviate from a normal pattern in both the normal probability plot and somewhat right-skewed histogram of the residuals. These large positive residuals also appear in the residuals versus fits plots and seem to be associated with above average fitted values, showing slightly increased variability. There is an interesting decreasing pattern when residuals are plotted versus the observations order, but this is explained by the fact that the data (except for the last case) have been ordered by decreasing $MeanAFC$ (which is very closely related to $LowAFC$, $r = 0.953$) in the original dataset.

4.6 a. Here is some output from a stepwise regression procedure (in Minitab) to predict *Embryos* based on the other nine available predictors.

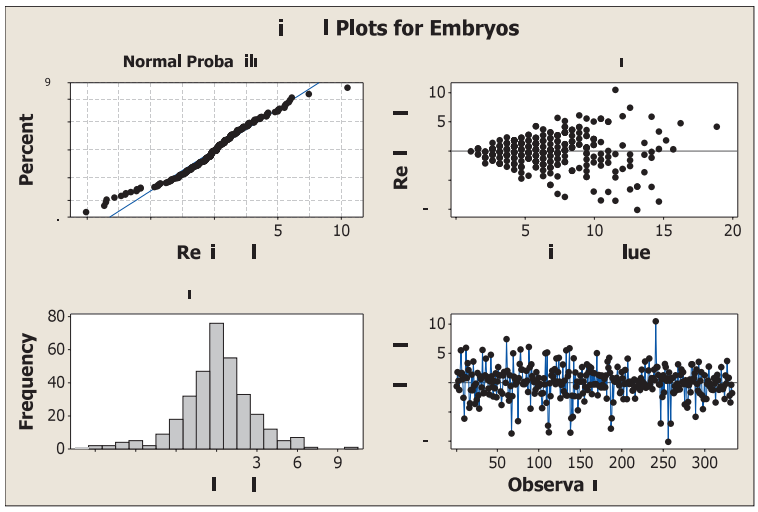
```
Stepwise Regression: Embryos versus Age, LowAFC, ...
  Alpha-to-Enter: 0.15  Alpha-to-Remove: 0.15
Response is Embryos on 9 predictors, with N = 333

Step             1        2
Constant    0.5318   0.2568

Oocytes      0.523    0.499
T-Value      21.15    17.48
P-Value      0.000    0.000

MaxE2                0.00036
T-Value                 1.68
P-Value                0.095

S             2.67     2.66
R-Sq         57.47    57.83
R-Sq(adj)    57.34    57.57
Mallows Cp     3.2      2.4
```

Oocytes is the initial predictor chosen at the first step and its P-value is essentially zero, so we can consider it a significant predictor of *Embryos*.

The next predictor chosen is $MaxE2$, but the P-value for its coefficient is 0.095, which is not significant at a 5% level. Unless we wanted to apply a more generous significance level (like 10% or the default 15% that Minitab uses to enter the model), we would probably stop at just the simple linear model based on *Oocytes*.

b. The first step of the stepwise regression output shows that *Oocytes* alone is an effective predictor of *Embryos* (P-value = 0.000) and it explains 57.47% of the variability in this response variable. Some residual plots for this simple linear model are shown below. Normality looks okay except for a few slight outliers, which should not be an issue for a sample this large

($n = 333$). The biggest concern is the clear fan pattern in the residual versus fits plot that shows the variability in the residuals increasing consistently for larger predicted values.



Note: If we had reported the two-predictor model from the original Minitab stepwise regression output in part (a), we would have some concern in this part about the effectiveness of the $MaxE2$ predictor in that model (with a P-value of 0.095).

c. If we omit *Oocytes* from the pool of potential predictors for *Embryos* and re-run the stepwise regression, we obtain the following output.

```
Response is Embryos on 8 predictors, with N = 333

Step              1        2
Constant      3.223    1.832

MaxE2       0.00227  0.00194
T-Value        8.75     7.54
P-Value       0.000    0.000

MeanAFC                0.140
T-Value                 5.19
P-Value                0.000

S              3.68     3.55
R-Sq          18.79    24.92
R-Sq(adj)     18.55    24.46
Mallows Cp     25.5      0.8
```

Without the *Oocytes* predictor, the stepwise regression procedure admits two predictors to the model, first $MaxE2$, with P-value = 0.000, and then $MeanAFC$, also with a P-value that is also essentially zero. Together, these two predictors explain 24.92% of the variability in *Embryos*. This is not nearly as effective as *Oocytes* alone ($R^2 = 57.5\%$), but perhaps it is easier to measure variables such as $MaxE2$ and $MeanAFC$ than to count *Oocytes*.

4.7 Software output from best subsets is shown as follows:

```
Response is Time
```

										A
										t
									P	t
									i	e
								M	t	n
								a	c	d
							R	r	h	a
							u	g	e	n
		R-sq	R-sq	Mallows			n	i	r	c
Vars	R-sq	(adj)	(pred)	Cp		S	s	n	s	e
1	55.5	51.8	46.3	1.3		15.342	X			
1	42.0	37.1	25.8	4.7		17.517			X	
2	60.7	53.5	40.9	2.0		15.065	X			X
2	57.9	50.2	36.2	2.7		15.587	X		X	
3	64.5	53.8	18.6	3.0		15.017	X		X	X
3	60.7	49.0	35.9	4.0		15.784	X	X		X
4	64.5	48.7	12.7	5.0		15.818	X	X	X	X

a. Based on the output, the model with the highest R^2 ($R^2 = 64.5\%$) includes either all four predictors or *Runs*, *Pitchers*, and *Attendance*.

b. Based on the output, the model with the highest adjusted R^2 is the first three-predictor model, which includes *Runs*, *Pitchers*, and *Attendance*, with an adjusted $R^2 = 53.8\%$.

c. Based on the output, the model with the lowest C_p (1.3) is the first single-predictor model, which includes *Runs*.

d. The simple linear regression model identified in part (c) is preferred. This simple model has the lowest C_p, only one predictor variable, and has a value of adjusted R^2 only slightly lower than the maximum adjusted R^2. A quick check of the residuals reveals one potential influential point, but otherwise the conditions appear to be met.

4.8 a. Computer output for fitting the model to the training sample:

```
Coefficients
Term         Coef  SE Coef  T-Value  P-Value
Constant    -3.17     2.69    -1.18    0.247
Hospitals   6.785    0.528    12.84    0.000
---
Model Summary
S          R-sq   R-sq(adj)    R-sq(pred)
9.62659  83.32%      82.82%        80.19%
```

the least squares line is $\widehat{\sqrt{MDs}} = -3.17 + 6.785 Hospitals$

b. Here are the predictions for $\sqrt{MDs}$ for the first few cases of the holdout sample:

64.680, 23.970, 17.185, 17.185, 17.185, 30.755, 10.400, ...

The cross-validation correlation between the predicted and actual values for the holdout sample is $r = 0.9531$.

c. Shrinkage is $0.8332 - 0.9531^2 = 0.8332 - 0.9084 = -0.0752$. The squared cross-validation correlation is remarkably close to the R^2 value for the training sample, so we can conclude that the model to predict $\sqrt{MDs}$ based on *Hospitals* works as well for the holdout sample as it did for the training sample.

4.9 a. Partial software output for fitting to the training data follows.

```
The regression equation is
GPA = 1.15 + 0.466 HSGPA + 0.0153 HU + 0.199 White

Predictor      Coef   SE Coef     T      P
Constant     1.1475    0.3115  3.68  0.000
HSGPA       0.46605   0.08839  5.27  0.000
HU         0.015328  0.004091  3.75  0.000
White       0.19917   0.07615  2.62  0.010

S = 0.377334   R-Sq = 28.4%   R-Sq(adj) = 26.9%
```

The prediction equation is $\widehat{GPA} = 1.1475 + 0.46605 HSGPA + 0.015328 HU + 0.19917 White$. Each of the three predictor variables is significant at the 0.05 level because the P-values for *HSGPA* (0.000), *HU* (0.000), and *White* (0.010) are all below 0.05. The estimated standard deviation of the errors is 0.377, and the overall R^2 is 28.4%. Thus, even though the predictor variables are significant, this model only explains 28.4% of the variation in *GPA*.

b. The fitted values ($\widehat{GPA}$), obtained using the equation in part (a) to predict the holdout sample, and the first few prediction errors are shown below.

Row	GPA-hat	Resid
1	3.15256	0.177436
2	3.04061	0.709392
3	2.96003	0.719967
4	2.73032	-0.570322
5	2.94408	-0.414084
6	3.26825	-0.038247
7	3.60474	-0.204738
8	3.24505	-0.655049
9	3.41013	0.379866
10	3.06194	0.458057
...		
69	2.81017	-0.190171

c. The mean of the residuals for the holdout cases is -0.06, which is reasonably close to zero. The standard deviation of the residuals (0.407) is reasonably close to 0.377, the estimated standard deviation of the error.

d. The cross-validation correlation is $r = 0.596$.

e. The square of the cross-validation correlation is $R^2 = 0.596^2 = 0.355$. Surprisingly this is even higher than $R^2 = 0.284$ for the model fit to the training data. The shrinkage, $0.284 - 0.355 = -0.071$, is negative! This means that the training model explains a higher percentage of variability in GPAs of the holdout sample than it does for the training model. The training model appears to work very well for the holdout sample.

4.10 The fitted regression model is $\widehat{Calories} = 109 + 1.00Sugar - 3.74Fiber$. The standardized residuals (SRES1), studentized residuals (TRES1), leverages (HI1), and values of Cook's D (COOK1) follow.

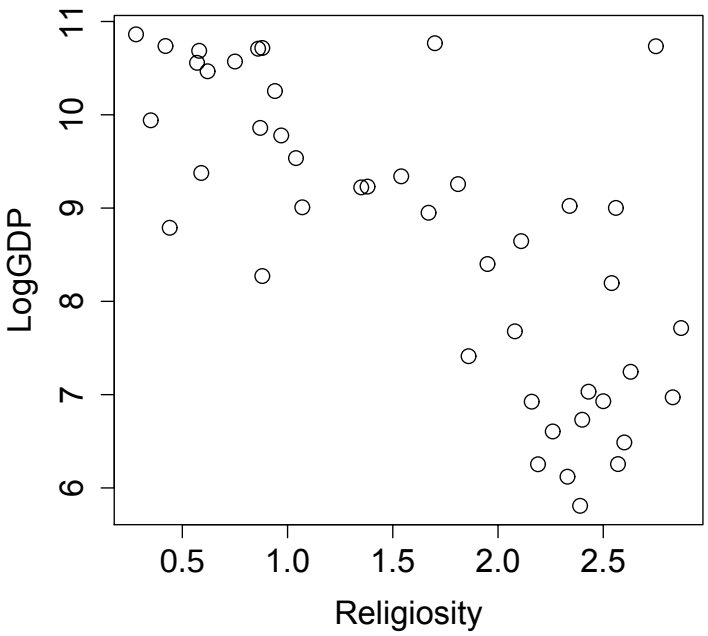
Row	Cereal	SRES1	TRES1	HI1	COOK1
1	Common Sense Oat Bran	-0.27020	-0.26637	0.028560	0.000715
2	Product 19	-0.57922	-0.57330	0.076995	0.009329
3	All Bran Xtra Fiber	-0.52192	-0.51609	0.266744	0.033032
4	Just Right	1.93143	2.01951	0.042586	0.055311
5	Original Oat Bran	-0.48161	-0.47594	0.138491	0.012429
6	Heartwise	-0.12379	-0.12193	0.041724	0.000222
7	Special K	-0.15896	-0.15660	0.101388	0.000950
8	Oatbake Raisin Nut	0.00927	0.00913	0.036218	0.000001
9	Kenmei Rice Bran	2.59435	2.86338	0.042586	0.099795
10	Nutri Grain	-0.00573	-0.00564	0.058749	0.000001
11	Shredded Wheat Squares	-0.86312	-0.85970	0.030699	0.007865
12	Oatmeal raisin Crisp	-0.10585	-0.10425	0.045898	0.000180
13	hole Wheat Total	-0.07242	-0.07132	0.045793	0.000084

14	Cheerios	0.48859	0.48288	0.092601	0.008120
15	Total Raisin Bran	-1.30347	-1.31795	0.048796	0.029053
16	Wheaties	-0.07242	-0.07132	0.045793	0.000084
17	Raisin Nut Bran	0.25651	0.25284	0.035101	0.000798
18	Wheat Chex	-0.32320	-0.31877	0.058463	0.002162
19	Batman	-0.62712	-0.62126	0.063138	0.008835
20	Ninja Turtles	-0.84637	-0.84264	0.100832	0.026777
21	Capt. Crunch	-0.09274	-0.09134	0.084662	0.000265
22	Trix	-0.77073	-0.76588	0.084662	0.018314
23	Frosted Flakes	-0.69782	-0.69229	0.072097	0.012612
24	Honey Smacks	-0.74732	-0.74222	0.147324	0.032165
25	Froot Loops	-0.58983	-0.58391	0.099545	0.012820
26	Puffed Rice	-0.66348	-0.65775	0.171868	0.030453
27	Mueslix Crispy Blend	3.36805	4.09414	0.114555	0.489201
28	Uncle Sam	1.81496	1.88374	0.075690	0.089915
29	100% Bran	-0.55640	-0.55049	0.152526	0.018572
30	Fruit & Fiber	-0.31368	-0.30936	0.029763	0.001006
31	Bran Flakes	-0.37005	-0.36516	0.032187	0.001518
32	Bran Buds	-0.69939	-0.69387	0.076783	0.013560
33	Fruit'n Oat Bran	1.06060	1.06267	0.081738	0.033376
34	Fruit'n Oat Bran Crunch	-0.04664	-0.04593	0.221865	0.000207
35	Hodgson's Mill Wheat	0.92522	0.92314	0.081738	0.025399
36	Hodgson's Mill Oat Bran	-0.66274	-0.65700	0.071844	0.011333

Kenmei Rice Bran (case 9) has a moderately large standardized residual of 2.59435, which is greater than 2. Mueslix Crispy Blend (case 27) has a very large standardized residual of 3.36805, which is greater than 3. All of the other standardized residuals are between -2 and 2. The studentized residuals show two moderately large values, Just Right (case 4) 2.01951 and Kenmei Rice Bran (case 9) 2.86338, and one very large value, Mueslix Crispy Blend (case 27) 4.09414. All of the other studentized residuals are between -2 and 2.

There is one very unusual leverage value beyond $3(2+1)/36 = 0.25$, $h_3 = 0.2667$ for All Bran Xtra Fiber (case 3). Moderately unusual leverage values are above $1/6 = 0.167$. The two moderately unusual leverages are 0.171868 for Puffed Rice (case 26) and 0.221865 for Fruit'n Oat Bran Crunch (case 34). All values of Cook's D are below 0.5, so none of the cereals are considered unusual with this measure.

4.11 a. We compute the natural logarithms of the *GDP* data and store the results in *LogGDP*. A scatterplot of *LogGDP* versus *Religiosity* is shown below. We see a negative association so that countries with higher scores on the *Religiosity* index tend to have lower *LogGDP*.

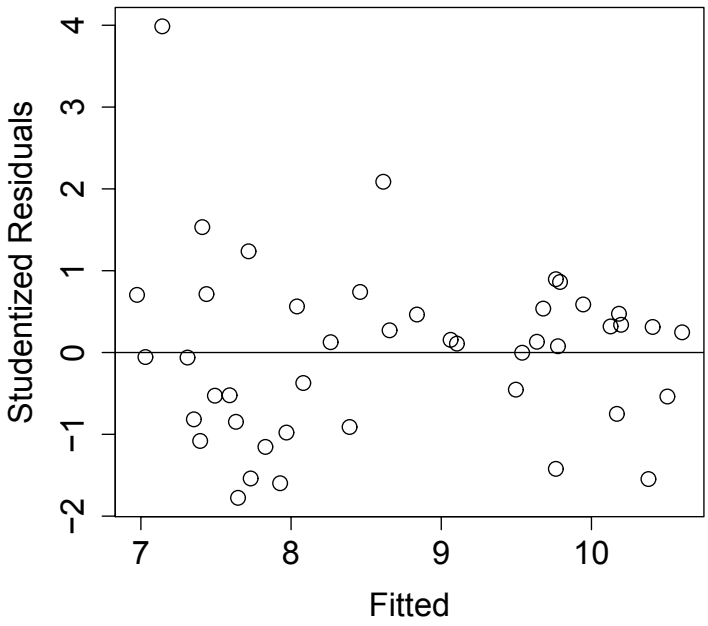


b. Some output for fitting a regression model to predict *LogGDP* based on *Religiosity* is shown below. From the value of R^2, we see that 53.9% of the variability in *LogGDP* for these countries can be explained by the *Religiosity* score.

```
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  10.9961     0.3656  30.079  < 2e-16 ***
Religiosity  -1.4013     0.2001  -7.005 1.43e-08 ***

Residual standard error: 1.085 on 42 degrees of freedom
Multiple R-squared: 0.5388,     Adjusted R-squared: 0.5278
F-statistic: 49.06 on 1 and 42 DF,  $P$-value: 1.432e-08
```

c. The estimated slope is $\hat{\beta}_1 = -1.40$. This means that, for every one unit increase in *Religiosity* for countries, we expect the *LogGDP* to decrease by about 1.40.

d. Using technology, we compute and save the studentized residuals for this model. The plot below shows the values of the studentized residuals plotted versus the *Religiosity* scores.



Kuwait is the point at the top of the plot with a studentized residual near 4 (actual value is 3.99).

e. Here is some output for predicting $LogGDP$ using $Religiosity$ and indicators for $EastEurope$, $MiddleEast$, $Asia$, $WestEurope$, and $Americas$ (leaving out $Africa$). Based on R^2, this model explains 72.4% of the variability in $LogGDP$ for these countries.

```
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   9.2019     0.7452  12.348 1.09e-14 ***
Religiosity  -0.9979     0.2852  -3.498  0.00124 **
EastEurope    0.7901     0.6709   1.178  0.24639
MiddleEast    1.9374     0.4797   4.039  0.00026 ***
Asia          0.9856     0.4556   2.163  0.03706 *
WestEurope    2.0538     0.6975   2.944  0.00556 **
Americas      1.5937     0.4778   3.336  0.00195 **

Residual standard error: 0.8947 on 37 degrees of freedom
Multiple R-squared: 0.7235,     Adjusted R-squared: 0.6787
F-statistic: 16.14 on 6 and 37 DF,  p-value: 5.095e-09
```

f. The coefficient of $Religiosity$ in part (e) is -0.9979. After accounting for region of the

world, we expect the log(GDP) to go down by about one (0.9979) for every increase of one in *Religiosity*.

g. We use a nested F-test to test for the contribution of the five indicator terms, $H_0 : \beta_2 = \beta_3 = \beta_4 = \beta_5 = \beta_6 = 0$ versus H_a : Some $\beta_i \neq 0$ for $i = 2, \ldots, 6$. Here is some computer output that compares the full model to a reduced model that omits the five indicator variables.

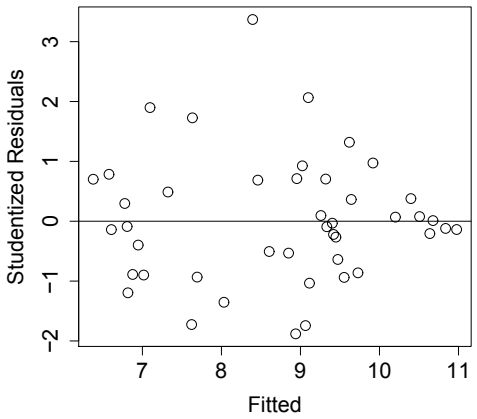
```
Model 1: LogGDP ~ Religiosity
Model 2: LogGDP ~ Religiosity + EastEurope + MiddleEast + Asia + WestEurope +
    Americas
  Res.Df    RSS Df Sum of Sq      F   Pr(>F)
1     42 49.405
2     37 29.615  5     19.79 4.9449 0.001448 **
```

We see that the sum of squared errors (RSS in this output) decreases from 49.405 to 29.615 when we add the five indicator variables to the model. The nested F-test statistic (also shown in the output) is

$$F = \frac{(49.405 - 29.615)/5}{29.615/37} = \frac{3.958}{0.800} = 4.95$$

Comparing this to the upper tail of a F-distribution with 5 and 37 degrees of freedom gives the P-value = 0.001438. Since this is a small P-value, we have strong evidence that including the indicators for the regions helps improve the predictions for *LogGDP*.

h. Following is the plot of studentized residuals versus fitted values for the model that includes the indicators for regions.



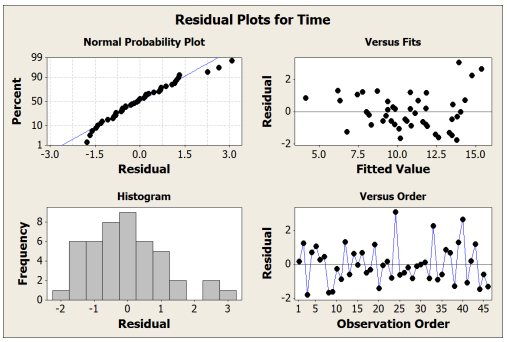
Kuwait is still an outlier (studentized residual = 3.37), but not as extreme as in the model that used *Religiosity* alone.

4.12 a. Fitting a model with all four predictors produces the output below.

```
Predictor          Coef     SE Coef       T      P
Constant          5.957       2.231    2.67  0.011
Elevation   -0.0016703   0.0005183   -3.22  0.002
Difficulty       0.8655      0.2285    3.79  0.000
Ascent       0.0006011   0.0003310    1.82  0.077
Length         0.44401     0.08125    5.46  0.000

S = 1.17085   R-Sq = 84.0%   R-Sq(adj) = 82.4%
```

We see that the t-tests for all of the predictors are easily significant at a 5% level, with the exception of *Ascent*, where the P-value is 0.077. However, if we drop this term from the model, the value of R^2 drops to 82.7% and the adjusted R^2 drops to 81.5%. For this reason, we will keep *Ascent* in the model.

b. Several plots of residuals for the four-predictor model follow.



The residuals versus fitted values plot shows a fairly uniform scatter on either side of the zero line, so conditions of linearity and equal variance look reasonable. The normal probability plot and histogram of the residuals look fairly normal, with the exception of three possible outliers on the high end.

c. The studentized residuals show no values beyond ± 3, but three mountains exceed ± 2. These are Seward Mt. (2.96), Mt. Donaldson (2.10), and Mt. Emmons (2.56). Each of these mountains take quite a bit longer to climb than the model predicts.

d. The largest leverage is for Nye Mt., $h_{45} = 0.276$, and the largest value of Cook's D is for Mt. Emmons, $D_{40} = 0.156$. Since the sample size is 46 and the model has four predictors, the lower threshold for leverage is $2(4+1)/46 = 0.2174$. This would flag Nye Mt. as a moderately influential point, although its leverage is not beyond $3(4+1)/46 = 0.326$. Mt. Marcy (the highest peak and $h_1 = 0.2231$), Cascade Mt. ($h_{36} = 0.2177$), and Cliff Mt. ($h_{44} = 0.2178$) would also barely qualify as moderately influential by the $2(k+1)/n$ criteria. None of the mountains is beyond the 0.5 threshold to be viewed as unusual by Cook's D.

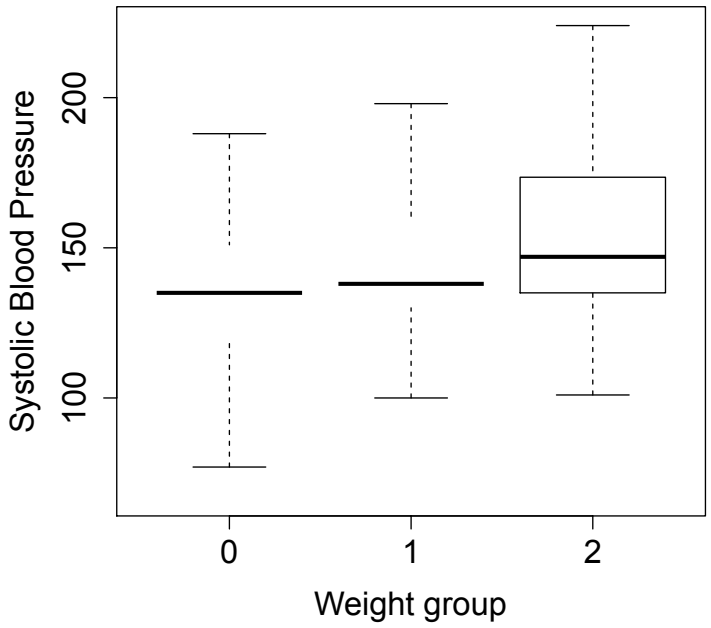
4.13 The intercept, 117.87, is the mean birth weight (in ounces) for the group that was was not included as an indicator in the model, $Whites$. The coefficient of each indicator variable tells how much the mean birth weight for that group differs from the mean of the reference group ($Whites$). So the mean birth weight for $Blacks$ is $117.87 - 7.31 = 110.56$, $Hispanics$ is $117.87 + 0.65 = 118.52$, and $Other$ is $117.87 - 0.73 = 117.14$.

4.14 a. The P-value for $Black$ is 0.000, indicating that the mean birth weight for $Blacks$ is different from $Whites$. The P-values for $Hispanic$ (0.731) and $Other$ (0.825) are not small, so we find no significant evidence for a difference in mean birth weights between either of these groups and Whites.

b. The value of R^2 indicates that 1.9% of the variability in birth weights for this sample can be explained by race.

c. The ANOVA table is used to test $H_0 : \beta_1 = \beta_2 = \beta_3 = 0$. The F-statistic (9.53) and P-value (0.000) indicate that this null hypothesis should be rejected. This means that at least one of the racial groups is useful for predicting birth weights.

4.15 Boxplots comparing the systolic blood pressure readings for the three weight groups (0 = Normal, 1 = Overweight, 2 = Obese) are given. This shows that all three distributions are relatively symmetric and the systolic blood pressures tend to be higher as weight increases.



Here is some computer output for predicting *SystolicBP* based on *Overwt* as a quantitative variable.

```
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  136.229      1.859  73.272  < 2e-16 ***
Overwt         8.437      1.367   6.172  1.4e-09 ***

Residual standard error: 27.01 on 498 degrees of freedom
Multiple R-squared: 0.07106,    Adjusted R-squared: 0.06919
$F$-statistic: 38.09 on 1 and 498 DF,  p-value: 1.399e-09
```

Here is some computer output for predicting *SystolicBP* using indicators for the *Overweight* and *Obese* groups (leaving *Normal* as the reference group). Note that we could choose any of the three categories to be the reference group.

```
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  136.316      1.977  68.949  < 2e-16 ***
OverWeight     8.051      3.258   2.471   0.0138 *
Obese         16.866      2.737   6.162 1.49e-09 ***

Residual standard error: 27.04 on 497 degrees of freedom
Multiple R-squared: 0.07109,    Adjusted R-squared: 0.06735
F-statistic: 19.02 on 2 and 497 DF,  p-value: 1.101e-08
```

Although the R^2 values are fairly low (7.1% for both models), the individual t-tests and overall ANOVA indicate that the weight groups are useful for predicting systolic blood pressure. The predicted values under the two models are very similar for the two models.

Model	Normal	Overweight	Obese
Single predictor	136.23	144.67	153.10
Indicators	136.32	144.37	153.18

For these data, it would appear that we don't gain much by using the separate indicator variables. The means for each group, as seen by the predicted values for the indicator model, are close to falling along the same line (given by the single-predictor model). Note that the adjusted R^2 actually goes down when we go from one to two predictors. In this case, the two models do equally well at describing the relationship between systolic blood pressure and weight categories.

4.16 a. Here is some output for a model to predict *LogNassim* based on *LogMass*.

```
The regression equation is LogNassim = - 1.89 + 0.371 LogMass
```

```
Predictor      Coef  SE Coef        T      P
Constant   -1.88738  0.01841  -102.53  0.000
LogMass     0.37096  0.01332    27.85  0.000

S = 0.250145   R-Sq = 75.5%   R-Sq(adj) = 75.5%

Analysis of Variance
Source           DF      SS      MS       F      P
Regression        1  48.528  48.528  775.55  0.000
Residual Error  251  15.706   0.063
Total           252  64.234
```

We see that `R-Sq = 75.5%`, so 75.5% of the variability in log nitrogen assimilation for these caterpillars is explained by $LogMass$.

b. We create indicators for the $Instar$ categories and fit a model to predict $LogNassim$, using four of the five indicators. The following output arises from using all but the indicator for $Instar = 5$.

```
The regression equation is
LogNassim = - 1.47 - 1.25 Instar1 - 1.10 Instar2 - 0.943 Instar3 - 0.558 Instar4

Predictor      Coef  SE Coef       T      P
Constant   -1.47367  0.02585  -57.00  0.000
Instar1    -1.24924  0.04169  -29.97  0.000
Instar2    -1.09848  0.04169  -26.35  0.000
Instar3    -0.94310  0.04024  -23.44  0.000
Instar4    -0.55822  0.03656  -15.27  0.000

S = 0.206835   R-Sq = 83.5%   R-Sq(adj) = 83.2%

Analysis of Variance
Source           DF      SS      MS       F      P
Regression        4  53.624  13.406  313.37  0.000
Residual Error  248  10.610   0.043
Total           252  64.234
```

The indicators for $Instar$ stages explain about 83.5% of the variability in log nitrogen assimilation in these caterpillars.

c. Since the indicator for $Instar = 5$ is omitted from the model in part (b), the constant term, $\hat{\beta}_0 = -1.474$, estimates the mean $LogNassim$ for caterpillars with $Instar = 5$.

The coefficient of $Instar1$, $\hat{\beta}_1 = -1.249$, indicates that the mean $LogNassim$ for caterpillars with $Instar = 1$ is about 1.249 less than the mean for $Instar5$ caterpillars, or about $-1.474 - 1.249 = -2.723$.

Note: Answers will differ if a different one of the $Instar$ categories is omitted from the model.

d. If we use $LogMass$ together with four of the five $Instar$ indicators, we obtain the output below.

```
The regression equation is
LogNassim = - 1.52 + 0.0615 LogMass - 1.05 Instar1 - 0.947 Instar2
            - 0.834 Instar3 - 0.503 Instar4

Predictor      Coef  SE Coef       T      P
Constant   -1.51868  0.04152  -36.57  0.000
LogMass     0.06148  0.04444    1.38  0.168
Instar1     -1.0488   0.1508   -6.96  0.000
Instar2     -0.9467   0.1173   -8.07  0.000
Instar3    -0.83396  0.08852   -9.42  0.000
Instar4    -0.50265  0.05427   -9.26  0.000

S = 0.206455   R-Sq = 83.6%   R-Sq(adj) = 83.3%

Analysis of Variance
Source           DF      SS      MS       F      P
Regression        5  53.706  10.741  252.00  0.000
Residual Error  247  10.528   0.043
Total           252  64.234
```

Together, $LogMass$ and the $Instar$ indicators explain 83.6% of the variability in $LogNassim$ for these caterpillars. This is only slightly larger than the 83.5% explained by just the $Instar$ indicators.

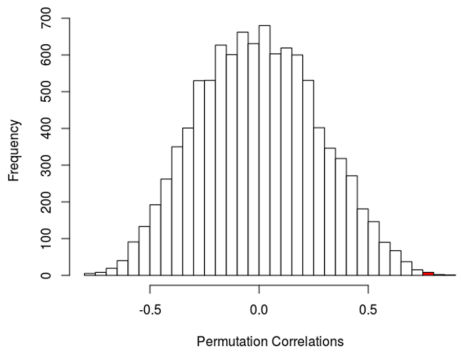
e. To see if $LogMass$ is a useful predictor for the model in part (d), we test $H_0 : \beta_1 = 0$ versus $H_a : \beta_1 \neq 0$, using the t-statistic ($t = 1.38$) and P-value (0.168) from the test for the individual coefficient in the regression output. This is not a small P-value, so we do not find sufficient evidence to show that $LogMass$ is an effective predictor in this model.

f. To test the $Instar$ indicators, we do a nested F-test, where $H_0 : \beta_2 = \beta_3 = \beta_4 = \beta_5 = 0$ versus $H_a :$ One of these $\beta_i \neq 0$. The reduced model (without these indicators) is shown in part (a) and the full model is in part (d). We compute the F-statistic with

$$F = \frac{(53.706 - 48.528)/4}{10.528/247} = \frac{1.2945}{0.0426} = 30.4$$

Comparing this to an F-distribution with 4 and 247 degrees of freedom, we see a P-value that is essentially zero, giving strong evidence that the information based on the $Instar$ indicators is important in this model for log nitrogen assimilation.

4.17 The correlation between $Runs$ and $Time$ is $r = 0.745$. We test $H_0 : \rho = 0$ versus $H_a : \rho > 0$, where ρ represents the correlation between runs and game time for all MLB games. To compute the randomization samples, we randomly scramble the $Time$ values and find the correlation with the original $Runs$. Repeating this 10,000 times, we get the following histogram of randomization correlations under a null hypothesis of no association.

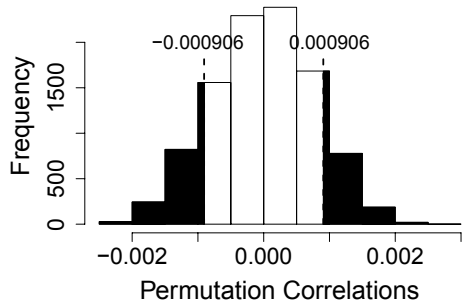


The correlation from the original sample, $r = 0.745$, lies fairly far out in the upper tail. Only 14 of the 10,000 simulated correlations are beyond 0.745 (the shaded area in the figure). This gives an estimated P-value $= 14/10{,}000 = 0.0014$ (note that the results will change slightly for different sets of simulations). This is a very small P-value, providing solid evidence that there is a positive correlation between the number of runs scored in MLB games and the times of the games.

4.18 Here is some output for fitting a model to predict GPA based in $VerbalSAT$. The sample slope is $\hat{\beta}_1 = 0.000906$.

```
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 2.6042036  0.4377919   5.948  5.5e-06 ***
VerbalSAT   0.0009056  0.0007659   1.182     0.25
```

To test $H_0 : \beta_1 = 0$ versus $H_a : \beta_1 \neq 0$, we create randomization samples by randomly scrambling the GPAs and assigning them to the verbal SAT scores from the sample. For each of these simulated samples, we fit a regression model to predict the scrambled GPA based on $VerbalSAT$ and record the slope. The following histogram shows the distribution of such slopes for 10,000 randomizations.



There are 1242 of the 10,000 simulations that give slopes above the 0.0009006 of the original sample. Doubling to account for two tails gives an estimated P-value of 0.2484. We could also count simulation slopes beyond ± 0.000906 in either tail to handle two tails. For the 10,000 randomizations shown above, this would give the P-value $(1324 + 1242)/10000 = 0.2566$. In either case, this is not a small P-value, so we lack evidence to conclude that the slope, when using verbal SAT scores to predict GPA, differs from zero.

We see in the output shown above to fit this model that a t-test for the slope has a P-value of 0.25. The randomization test for correlation between these same variables that is done as an example in the text yields an estimated P-value of 0.239. These results are all consistent with each other.

4.19 a. Output for a model to predict *Time* based on *Pitchers* and *Attendance* using the original data in **BaseballTimes2017** is shown below.

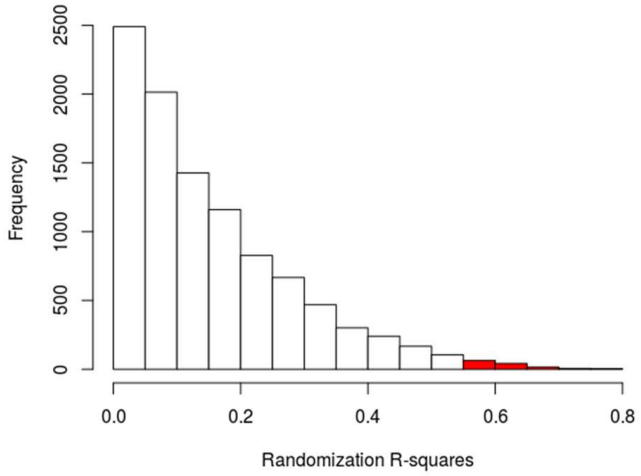
```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 1.031e+02  2.556e+01   4.032  0.00198 **
Pitchers    8.109e+00  2.557e+00   3.171  0.00890 **
Attendance  6.312e-04  3.913e-04   1.613  0.13502
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 16.45 on 11 degrees of freedom
Multiple R-squared:  0.5307,Adjusted R-squared:  0.4454
F-statistic: 6.219 on 2 and 11 DF,  p-value: 0.0156
```

b. We choose $R^2 = 0.531$ as a measure of effectiveness of this model, although $F = 6.219$ or $\hat{\sigma}_\epsilon = 16.45$ would also work (or one could find $SSModel$ or SSE from the ANOVA table).

c. Here is the output for one of the randomization samples. We see that $R^2 = 0.1282$, even though the game times were randomly assigned to the cases.

```
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.585e+02  3.484e+01   4.549 0.000831 ***
Pitchers     4.356e+00  3.485e+00   1.250 0.237210
Attendance  -1.107e-04  5.333e-04  -0.208 0.839312
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 22.42 on 11 degrees of freedom
Multiple R-squared:  0.1282,Adjusted R-squared:  -0.03026
F-statistic: 0.8091 on 2 and 11 DF,  p-value: 0.4701
```

d. Repeating the process of (b) for 10,000 randomizations gives the plot below for the R^2 values from each model.



e. To find the P-value, we count the number of randomization R^2 values that exceed the $R^2 = 0.531$ of the original sample (see the shaded region in the tail above). There are 168 values of R^2 in this region, so the estimated P-value is $168/10{,}000 = 0.0168$. This is a very small P-value, so we have evidence that the R^2 value from the original sample is unusually large compared to values when there is no relationship between game times and the two predictors. From this, we can conclude that there probably is some relationship between $Time$ and at least one of $Pitchers$ and $Attendance$.

f. From the output for the original model in part (a), we see the P-value for the F-statistic is 0.0156. This is consistent with the randomization P-value and leads to the same conclusion about the effectiveness of the model.

4.20 a. Some output for predicting *Length* based on *Time* for the original **HighPeaks** data is shown below.

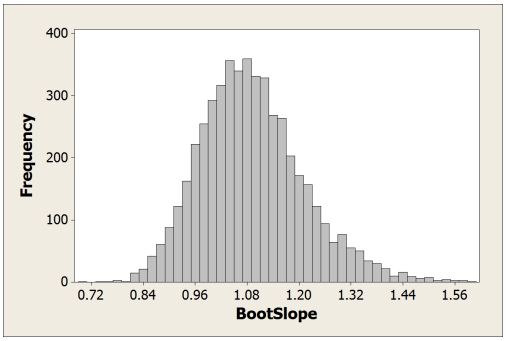
```
Predictor      Coef  SE Coef      T      P
Constant      1.100    1.067   1.03  0.308
Time        1.07711  0.09699  11.11  0.000
```

From this output, the original sample slope is $\hat{\beta}_1 = 1.077$ and the standard error for the slope is $s_{\hat{\beta}_1} = 0.097$. If we use a t-distribution with $n - 2 = 44$ degrees of freedom, the t-value for a 90% confidence interval is $t^* = 1.68$. To compute the confidence interval for the slope, we use

$$\hat{\beta}_1 \pm t^* \cdot s_{\hat{\beta}_1} = 1.077 \pm 1.68(0.097) = 1.077 \pm 0.163 = (0.914, 1.240)$$

Based on these data, we are 90% sure that the slope of the model to predict hike length (in miles) based on hike time (in hours) is between 0.914 miles per hour and 1.240 miles per hour.

b. We sample 46 cases at a time (with replacement) from the **HighPeaks** data and fit the regression model to predict *Time* based on *Length* of hikes. The slopes for 5000 such bootstrap samples are shown below. The distribution is reasonably bell-shaped (with just a bit of right skew) and centered around the sample slope of 1.08.



c. The mean of the 5000 slopes in the bootstrap distribution is 1.091 (similar to the estimated slope from the original sample). The standard deviation of the slopes is $SE = 0.1207$. This is a bit larger, but otherwise similar to the estimate of 0.097 from the computer output.

d. For 90% confidence, the standard normal value is $z^* = 1.645$, which gives the following interval.

$$\hat{\beta}_1 \pm z^* \cdot SE = 1.077 \pm 1.645(0.1207) = 1.077 \pm 0.199 = (0.878, 1.276)$$

e. For the 5000 bootstrap slopes in part (b), the 5^{th} percentile is 0.913 and the 95^{th} percentile is 1.311, which gives a 90% confidence interval for the slope of (0.913, 1.311).

f. The upper percentile is $1.311 - 1.077 = 0.234$ units above the original estimate and the lower percentile is $1.077 - 0.913 = 0.164$ below. Reversing these, we expect the "true" slope is probably somewhere between $1.077 - 0.234 = 0.843$ and $1.077 + 0.164 = 1.241$ or (0.843, 1.241).

g. Here are the four intervals from the earlier parts.

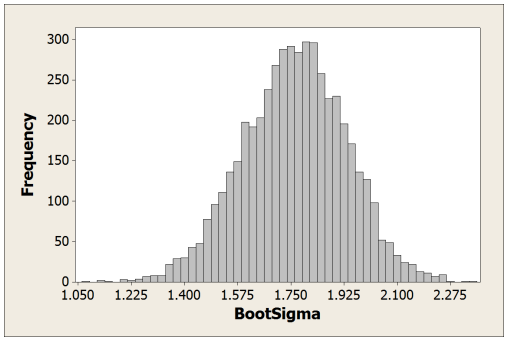
t-interval from original data	(0.914, 1.240)
z-interval using bootstrap SE	(0.878, 1.276)
Bootstrap percentiles	(0.913, 1.311)
Bootstrap reverse percentiles	(0.843, 1.241)

The intervals are similar to each other, although the t-interval is a bit narrower.

4.21 From the output for fitting *Time* based on *Length*, we see the estimated standard error of the regression model is $\hat{\sigma}_\epsilon = 1.82$.

```
Predictor      Coef  SE Coef      T      P
Constant      1.100    1.067   1.03  0.308
Time        1.07711  0.09699  11.11  0.000

S = 1.81827   R-Sq = 73.7%   R-Sq(adj) = 73.1%
```

To form a bootstrap distribution, we sample with replacement from the original **HighPeaks** data, fit the model for each bootstrap sample, and record the estimated standard error of the regression. The following histogram shows the results for 5000 such bootstrap samples.



To construct a 90% confidence interval for the standard error of regression from the bootstrap distribution:

Method #1 (SE): The standard deviation of the 5000 bootstrap regression standard errors is SE = 0.171. Using $z^* = 1.645$, we get a 90% confidence interval with

$$\hat{\sigma}_\epsilon \pm z^* \cdot SE = 1.82 \pm 1.645(0.171) = 1.82 \pm 0.28 = (1.54, 2.10)$$

Based on this analysis, we are 90% sure that the standard error for a regression model to predict lengths of hikes by the times for the hikes is somewhere between 1.54 and 2.10 miles.

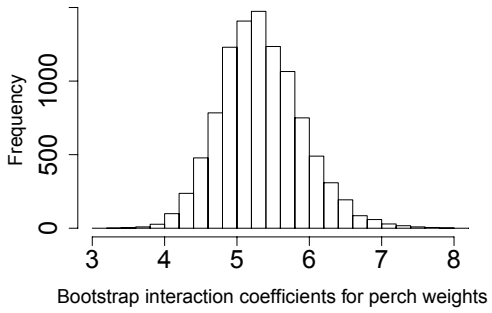
Method #2 (Percentile): Finding the $q_{0.05}$ and the $q_{0.95}$ percentiles of the 5000 bootstrap regression standard errors, we get a 90% confidence interval from 1.48 to 2.03.

Method #3 (Reverse percentile): The upper percentile is $2.03 - 1.82 = 0.21$ units above the original estimate and the lower percentile is $1.82 - 1.48 = 0.34$ below. Reversing these, we expect the "true" standard error of this regression model is probably somewhere between $1.82 - 0.21 = 1.61$ and $1.82 + 0.34 = 2.16$. This gives a 90% confidence interval of (1.61, 2.16).

4.22 Here is some output for a model to predict $Weight$ of perch using $Length$, $Width$, and an interaction term, $Length \cdot Width$. The coefficient of the interaction term is $\hat{\beta}_3 = 5.24$ and the standard error is $s_{\hat{\beta}_3} = 0.413$.

```
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)       113.9349    58.7844   1.938    0.058 .
Length             -3.4827     3.1521  -1.105    0.274
Width             -94.6309    22.2954  -4.244 9.06e-05 ***
I(Length * Width)   5.2412     0.4131  12.687  < 2e-16 ***
```

We construct a bootstrap sample by randomly choosing 54 cases with replacement from the **Perch** data. For each bootstrap sample, we fit the three-predictor model to predict $Weight$ and keep track of the coefficient of the interaction term. We repeat this 10,000 times to get the histogram shown below for the bootstrap distribution of interaction coefficients, $(\hat{\beta}_3)$.



We use three different ways to construct a 90% confidence interval using this bootstrap distribution for the coefficient of the interaction in the three-predictor model.

Method #1 (SE): The standard deviation of the 10,000 bootstrap coefficients is SE = 0.569. Using $z^* = 1.96$, we get a 95% confidence interval with

$$\hat{\beta}_3 \pm z^* \cdot SE = 5.24 \pm 1.96(0.569) = 5.24 \pm 1.12 = (4.12, 6.36)$$

Based on this analysis, we are 95% sure that the coefficient of the interaction term in the three-predictor regression model to predict perch weights is somewhere between 4.12 and 6.36.

Method #2 (Percentile): Finding the $q_{0.025}$ and the $q_{0.975}$ percentiles of the 10,000 bootstrap interaction coefficients, we get a 95% confidence interval from 4.31 to 6.53.

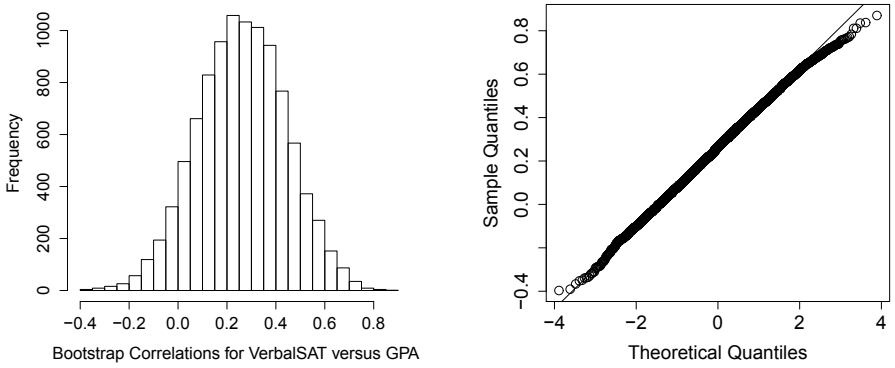
Method #3 (Reverse percentile): The upper percentile is $6.53 - 5.24 = 1.29$ units above the original estimate and the lower percentile is $5.24 - 4.31 = 0.93$ below. Reversing these, we expect the coefficient of the interaction term for this three-predictor regression model is somewhere between $5.24 - 1.29 = 3.95$ and $5.24 + 0.93 = 6.17$. This gives a 95% confidence interval of (3.95, 6.17).

Using the original regression output and $t^* = 2.005$ from a t-distribution with 54 degrees of freedom, we find a 95% confidence interval for the coefficient of $Length \cdot Width$ with

$$\hat{\beta}_3 \pm t^* \cdot s_{\hat{\beta}_3} = 5.24 \pm 2.00(0.413) = 5.24 \pm 0.83 = (4.41, 6.07)$$

The t-interval from the original regression fit is somewhat narrower than the three intervals based on the bootstrap distribution.

4.23 a. The correlation between GPA and $VerbalGPA$ in the original **SATGPA** data is $r = 0.244$. We generate a bootstrap distribution by sampling 24 cases at a time (with replacement) from the original data and calculating the correlation between GPA and $VerbalSAT$ for each sample. The histogram below shows 10,000 such bootstrap correlations.



A normal quantile plot of the bootstrap correlations is shown in the graph to the right above. Except for some small departure in the upper (right) tail, the agreement with normality looks good for this bootstrap distribution.

b. To construct a 95% confidence interval using this bootstrap distribution for the correlation between GPA and $VerbalSAT$:

Method #1 (SE): The standard deviation of the 10,000 bootstrap correlations is SE = 0.181. Using $z^* = 1.96$, we get a 95% confidence interval with

$$r \pm z^* \cdot SE = 0.244 \pm 1.96(0.181) = 0.244 \pm 0.355 = (-0.111, 0.599)$$

Based on this analysis, we are 95% sure that the correlation between GPA and $VerbalSAT$ scores is somewhere between $= -0.111$ and 0.599.

Method #2 (Percentile): Finding the $q_{0.025}$ and the $q_{0.975}$ percentiles of the 10,000 bootstrap correlations, we get a 95% confidence interval from -0.094 to 0.610.

Method #3 (Reverse percentile): We find the deviation from $r = 0.244$ to the upper percentile, $0.610 - 0.244 = 0.366$, and subtract this from the estimate to get the lower bound of the interval, $0.244 - 0.366 = -0.122$. For the other direction, find the deviation to the lower percentile, $0.244 - (-0.094) = 0.338$, and add this to $r = 0.244$ to get the upper bound of the interval, $0.244 + 0.338 = 0.582$. This gives a 95% confidence interval for the correlation that is (−0.122, 0.582).

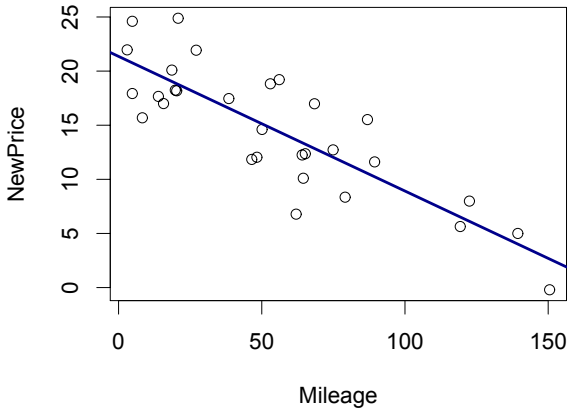
The results of these three confidence intervals for the correlation between $VerbalSAT$ and GPA are similar.

c. In fact, all three intervals have a lower bound near −0.10 and an upper bound around +0.60. The fact that each of the intervals includes both positive and negative correlations, as well as a correlation of zero, among the plausible values, tells us that we don't have much evidence for a consistent association between verbal SAT scores and GPA.

4.24 a. Here is some output from fitting a model to predict *Price* from *Mileage* for the data in **AccordPrice**. The estimated slope for the original data is $\hat{\beta}_1 = -0.1198$.

```
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  20.8096     0.9529   21.84  < 2e-16
Mileage      -0.1198     0.0141   -8.50 3.06e-09
```

b. We save the fits and residuals from the original model. To make a bootstrap sample, we sample (with replacement) from the residuals and add those residuals to the original fitted values to create new prices. This keeps the same values for the predictors and matches the model assumption that residuals are randomly assigned to the line to get the response variable (price) for each data pair. Here is one scatterplot where the residuals have been randomly sampled and added to the regression line to produce new prices.

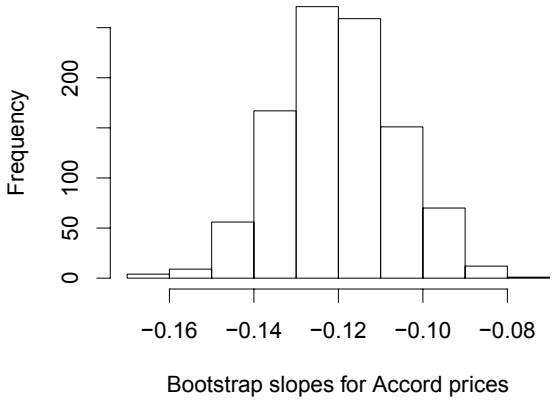


c. Here is output for fitting a regression line to the bootstrap sample in the graph above.

```
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 21.35883    1.01546  21.034  < 2e-16
Mileage     -0.12440    0.01502  -8.282 5.18e-09
```

The slope of the regression line fitted to this bootstrap sample is -0.1244 and the intercept is 21.36. These are close to the original estimates, but not exactly the same.

d. A bootstrap distribution with slopes for 1000 samples simulated by the method in (b) produce the following histogram. The mean of these 1000 bootstrap slopes is -0.1199, which is very close to the original slope. The standard deviation of the bootstrap slopes is 0.0138, which is close to the standard error of the slope in the original regression output.



e. To construct a 95% confidence interval using this bootstrap distribution for the slope between *Price* and *Mileage*:

Method #1 (Percentile): Finding the $q_{0.025}$ and the $q_{0.975}$ percentiles of the 1000 bootstrap slopes, we get a 95% confidence interval for the slope of the model to predict Accord prices from -0.1471 to -0.0936.

Method #2 (SE): The standard deviation of the 1000 bootstrap slopes is SE = 0.0138. Using $z^* = 1.96$, we get a 95% confidence interval with

$$\hat{\beta}_1 \pm z^* \cdot SE = -0.1198 \pm 1.96(0.0138) = -0.1198 \pm 0.0270 = (-0.1468, -0.0928)$$

Based on this analysis, we are 95% sure that the slope for predicting a used Accord price based on its mileage is somewhere between -0.1468 and -0.0928.

Method #3 (Bootstrap t): If we find the distribution of

$$T = (\text{Bootstrap slope} - \text{Original slope})/\text{Bootstrap SE of slope}$$

the 0.025 and 0.975 quantiles are

$$qt_{0.025} = -2.034 \qquad \text{and} \qquad qt_{0.975} = 2.080$$

This gives a 95% confidence interval with

$$-0.1198 - 2.034(0.0138) \text{ to } -0.1198 + 2.080(0.0138) = (-0.1477, 0.0913)$$

The results of these three confidence intervals for the slope between *Price* and *Mileage* for used Accords are quite close to each other and close to the three intervals that were produced in the corresponding text example that sampled the cases themselves, rather than the residuals. Those intervals are

percentile interval: $(-0.1517, -0.0879)$
z-interval from SE: $(-0.1579, -0.0944)$
reverse percentile: $(-0.1565, -0.0839)$