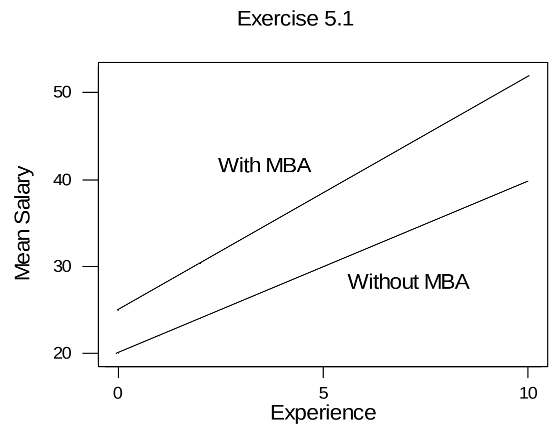
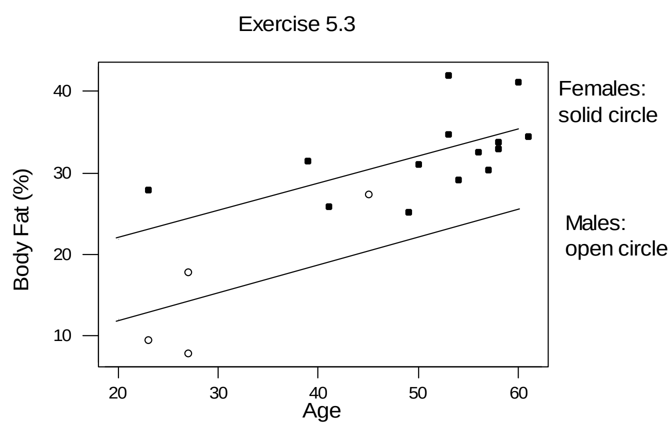
CHAPTER 5

5.1 Interaction; bonus for having a MBA; furthermore, salary increases faster for MBAs.



5.2 (a) $ 3,000; (b) $ 900

5.3



Minitab regression output. Significant age and gender effects; body fat of males is 9.79 percent lower than that of females. However, very few data for males.

```
The regression equation is
bodyfat = 15.1 + 0.339 age - 9.79 gender

Predictor        Coef     SE Coef          T        P
Constant       15.071       6.224       2.42    0.029
age            0.3392      0.1196       2.84    0.013
gender         -9.791       3.697      -2.65    0.018

S = 4.905       R-Sq = 74.6%     R-Sq(adj) = 71.2%

Analysis of Variance

Source            DF          SS          MS         F        P
Regression         2     1060.66      530.33     22.04    0.000
Residual Error    15      360.88       24.06
Total             17     1421.54
```

Regression with an interaction component: Interaction component is not needed.

```
The regression equation is
bodyfat = 20.1 + 0.240 age - 29.3 gender + 0.572 age*gen

Predictor        Coef     SE Coef          T        P
Constant       20.112       6.239       3.22    0.006
age            0.2401      0.1204       1.99    0.066
gender         -29.27       10.41      -2.81    0.014
age*gen        0.5725      0.2893       1.98    0.068

S = 4.488       R-Sq = 80.2%     R-Sq(adj) = 75.9%

Analysis of Variance

Source            DF          SS          MS         F        P
Regression         3     1139.51      379.84     18.86    0.000
Residual Error    14      282.02       20.14
Total             17     1421.54
```

5.4 $\text{VIF}_1 = 1/(1-R_1^2) = 2.5$; $\text{VIF}_2 = 1/(1-R_2^2) = 5$; $\text{VIF}_3 = 1/(1-R_3^2) = 10$; evidence of multicollinearity since variance inflation factors are large (10 or larger).

5.5 (e)

5.6 Define two indicator variables x_1 and x_2 such that $x_1 = 0$ and $x_2 = 0$ represent the group Sparrow, $x_1 = 1$, $x_2 = 0$ represent Robin, and $x_1 = 0$ and $x_2 = 1$ represent Wren. Then the model can be expressed as $E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2$ in which $\beta_1 = \mu(\text{Robin}) - \mu(\text{Sparrow})$ and $\beta_2 = \mu(\text{Wren}) - \mu(\text{Sparrow})$.

```
Analysis of Variance
                                    Sum of          Mean
 Source                  DF        Squares        Square   F Value   Pr > F
 Model                    2       31.11193      15.55596     22.33   <.0001
 Error                   42       29.26052       0.69668
 Corrected Total         44       60.37244
```

F-statistic = 22.33 tests whether there are differences among the three group means; p-value < 0.0001; reject H_0: $\mu_1 = \mu_2 = \mu_3$ (or $\beta_1 = \beta_2 = 0$)

5.7 Minitab output for regression with averages

```
The regression equation is
yield = 78.4 - 3.55 fac1 - 1.45 fac2 + 3.20 fac3

Predictor        Coef     SE Coef          T        P
Constant       78.375       1.022      76.65    0.000
fac1           -3.550       1.022      -3.47    0.026
fac2           -1.450       1.022      -1.42    0.229
fac3            3.200       1.022       3.13    0.035

S = 2.892       R-Sq = 85.6%     R-Sq(adj) = 74.9%

Analysis of Variance

Source            DF          SS          MS         F        P
Regression         3     199.560      66.520      7.95    0.037
Residual Error     4      33.455       8.364
Total              7     233.015
```

$V(\bar{y}_i) = s^2/5 = 40/5 = 8$; $s(\bar{y}_i) = \sqrt{8} = 2.83$ (calculated from the pure error sum of squares) is very similar to s = 2.892 that is calculated from the residuals. Hence there is no lack of fit. However, in general this must not be the same, and should be checked.

$$V(\hat{\boldsymbol{\beta}}) = (X'X)^{-1}X'\boldsymbol{y} = (s^2/5)(X'X)^{-1} = 8\begin{bmatrix} 0.125 & 0 & 0 & 0 \\ 0 & 0.125 & 0 & 0 \\ 0 & 0 & 0.125 & 0 \\ 0 & 0 & 0 & 0.125 \end{bmatrix}$$

$s.e.(\hat{\beta}_i) = 1$; $t(\hat{\beta}_1) = -3.55$; $t(\hat{\beta}_2) = -1.45$; $t(\hat{\beta}_3) = 3.20$; the effect of factor 2 is not significant.

5.8
(a) Expected difference in systolic blood pressure for females versus males who drink the same number of cups of coffee, excercise the same, and are of the same age
(b) Represents variation due to measurement error and omitted factors
(c) Association, but not causation
(d) Represents interaction between gender and coffee consumption

5.9
(a) $E(y_t) = \begin{cases} \beta_0 + \beta_1 t, \ t = 1,2,...,7 \\ \beta_2 + \beta_3 t, \ t = 8,9,...,14 \end{cases}$

Intersecting lines at t = 8: $\beta_2 = \beta_0 + 8(\beta_1 - \beta_3)$, and

E (y_t) = $\begin{cases} \beta_0 + \beta_1 t, \ t = 1,2,...,7 \\ \beta_0 + \beta_1 8 + \beta_3 (t-8), \ t = 8,9,...,14 \end{cases}$

In matrix form, $E(\boldsymbol{y}) = X\boldsymbol{\beta}$ where

$$X = \begin{bmatrix} 1 & 1 & 0 \\ 1 & 2 & 0 \\ . & . & . \\ 1 & 7 & 0 \\ 1 & 8 & 0 \\ 1 & 8 & 1 \\ . & . & . \\ 1 & 8 & 6 \end{bmatrix} \quad \text{and} \quad \boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_3 \end{bmatrix}$$

(b) $E(y_t) = \beta_0 + \beta_1 t, \ t = 1,2,...,14$
(c) F = 55.95; p-value = P(F(1,11) > 55.95) = 0.0000; model in (a) is preferable.

5.10
(a) $E(y_t) = \beta_0 + \beta_1 t, \ t = 1,2,...,12$
(b) $E(y_t) = \beta_0 + \beta_1 t + \beta_2 t^2, \ t = 1,2,...,12$
(c) $E(y_t) = \beta_0 + \beta_1 t + \beta_2 x_t, \ t = 1,2,...,12$ where $x_t = 0$ for $t = 1,2...,6$, and $x_t = 1$ for t = 7,8,...,12
(d) $E(y_t) = \begin{cases} \beta_0 + \beta_1 t, \ t = 1,2,...,7 \\ \beta_2 + \beta_3 t, \ t = 8,9,...,14 \end{cases}$

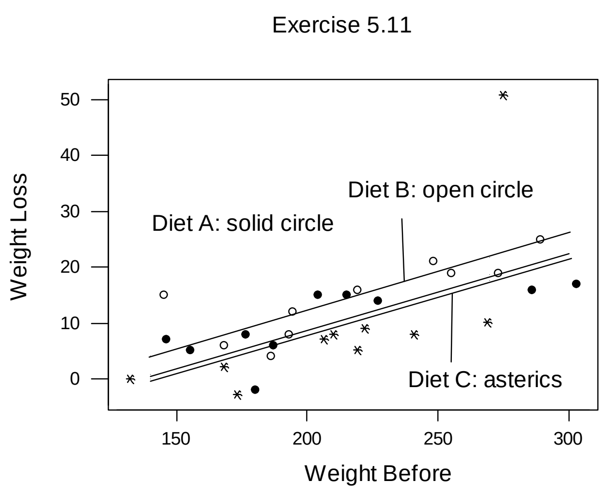
Intersecting lines at t = 7: $\beta_2 = \beta_0 + 7(\beta_1 - \beta_3)$, and

$$E(y_t) = \begin{cases} \beta_0 + \beta_1 t, \ t = 1,2,\ldots,6 \\ \beta_0 + \beta_1 7 + \beta_3 (t-7), \ t = 7,8,\ldots,12 \end{cases}$$

In matrix form, $E(\boldsymbol{y}) = X\boldsymbol{\beta}$ where

$$X = \begin{bmatrix} 1 & 1 & 0 \\ 1 & 2 & 0 \\ . & . & . \\ 1 & 6 & 0 \\ 1 & 7 & 0 \\ 1 & 7 & 1 \\ . & . & . \\ 1 & 7 & 5 \end{bmatrix} \quad \text{and} \quad \boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_3 \end{bmatrix}$$

5.11



Note the unusual observation for one subject on diet C (x = 275, y = 51). We define indicators for the three diets: IndA = 1 if diet A and = 0 otherwise; IndB = 1 if diet B and = 0 otherwise; IndC = 1 if diet C and = 0 otherwise.
Minitab output from the estimation of the model $y = \beta_0 + \beta_1 x + \beta_2 \text{IndB} + \beta_3 \text{IndC} + \varepsilon$ is shown below.

Using all n = 30 cases we find not much difference between the three diets. F-statistic for testing $\beta_2 = \beta_3 = 0$: F = (1740.1 - 1650.12)/2] / (1650.12/26) = 0.71; p-value = P(F(2,26) > 0.71) = 0.50; conclude $\beta_2 = \beta_3 = 0$.

Models with all 30 cases:

```
The regression equation is
y = - 18.4 + 0.137 x + 3.15 IndB - 0.89 IndC

Predictor        Coef     SE Coef          T        P
Constant      -18.388       7.067      -2.60    0.015
x             0.13703     0.03176       4.31    0.000
IndB            3.153       3.574       0.88    0.386
IndC           -0.893       3.565      -0.25    0.804

S = 7.967       R-Sq = 44.5%     R-Sq(adj) = 38.1%

Analysis of Variance

Source            DF           SS          MS         F        P
Regression         3      1323.25      441.08      6.95    0.001
Residual Error    26      1650.12       63.47
Total             29      2973.37


The regression equation is
y = - 18.2 + 0.140 x

Predictor     Coef  SE Coef      T      P
Constant   -18.167    6.799  -2.67  0.012
x          0.13954  0.03132   4.45  0.000

S = 7.88328   R-Sq = 41.5%   R-Sq(adj) = 39.4%

Analysis of Variance

Source          DF      SS      MS      F      P
Regression       1  1233.3  1233.3  19.84  0.000
Residual Error  28  1740.1    62.1
Total           29  2973.4
```

The observation (diet C; x = 275, y = 51) is highly unusual. Omitting this case, leads to the results given below. In the next chapter (Chapter 6) you will learn about diagnostic measures that allow you to quantify the effects of outliers. After reading Chapter 6, you may want to confirm that this case leads to the standardized residual = 4.48 and Cook's distance = 0.98.

Models with outlying case omitted:

```
The regression equation is
y = - 10.2 + 0.0977 x + 3.51 IndB - 4.65 IndC

Predictor     Coef  SE Coef      T      P
Constant   -10.205    3.567  -2.86  0.008
x          0.09767  0.01610   6.07  0.000
IndB         3.511    1.747   2.01  0.055
IndC        -4.651    1.789  -2.60  0.015

S = 3.89272   R-Sq = 72.0%   R-Sq(adj) = 68.7%

Analysis of Variance

Source          DF       SS      MS      F      P
Regression       3   975.03  325.01  21.45  0.000
Residual Error  25   378.83   15.15
Total           28  1353.86


The regression equation is
y = - 12.1 + 0.106 x

Predictor     Coef  SE Coef      T      P
Constant   -12.132    4.465  -2.72  0.011
x          0.10574  0.02079   5.09  0.000

S = 5.06040   R-Sq = 48.9%   R-Sq(adj) = 47.0%

Analysis of Variance

Source          DF       SS      MS      F      P
Regression       1   662.45  662.45  25.87  0.000
Residual Error  27   691.41   25.61
Total           28  1353.86
```

F-statistic for testing $\beta_2 = \beta_3 = 0$: F = (691.41 – 378.83)/2] / (378.83/25) = 10.31; p-value = P(F(2,25) > 10.31) = 0.001; reject $\beta_2 = \beta_3 = 0$.

(b) There are differences among the three diets in terms of their effectiveness on weight reduction. Diet C has the largest benefit.

5.12

```
Analysis of Variance
                             Sum of          Mean
Source              DF      Squares       Squares    F Value  Pr > F
Model                4     39.37694       9.84423      14.07  <.0001
Error               25     17.49506       0.69980
Corrected Total     29     56.87200
```

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	-0.91221	0.87548	-1.04	0.3074
x1	1	0.16073	0.06617	2.43	0.0227
x2	1	0.21978	0.03406	6.45	<.0001
x3	1	0.01123	0.00497	2.26	0.0330
x4	1	0.10197	0.05874	1.74	0.0948

(b) $\hat{\mu} = -0.9122 + 0.1607x_1 + 0.2198x_2 + 0.0112x_3 + 0.1020x_4$; R^2 = 0.692; s = 0.8365;

(i) $t(\hat{\beta}_1) = 2.43$; p-value = 0.023; reject $\beta_1 = 0$

(ii) F = (5.45747/2)/(0.69980) = 3.90 (use of additional SS); p-value = 0.034; reject the null hypothesis $\beta_3 = \beta_4 = 0$

(iii) F=14.07; p-value <.0001; reject hypothesis $\beta_1 = \beta_2 = \beta_3 = \beta_4 = 0$.

(c)

$\hat{\mu} = -1.462 + 0.1536x_1 + 0.3221x_2 + 0.0166x_3 + 0.0571x_4 - 0.00087x_2x_3 + 0.00599x_2x_4$

$H_0 : \beta_5 = \beta_6 = 0$: F = 0.40; p-value = 0.67; interactions not important.

(d) (i) Since all coefficients are positive: Lower wrinkle resistance for lower x_1, x_2, x_3, and x_4.

(ii) Increased wrinkle resistance for higher x_1, x_2, x_3, and x_4.

(e) It is difficult to generalize the conclusions from this study since the values of x_1, x_2, x_3, and x_4 were not controlled. One suggestion for improvement is to conduct an experiment in which the values of x_1, x_2, x_3, and x_4 are controlled and the resulting response *y* measured.

5.13

(b) z = 0 (protein-rich); z = 1 (protein-poor): $\hat{\mu} = 50.324 + 16.009x + 0.918z - 7.329xz$

H_0: $\beta_2 = \beta_3 = 0$. Test whether the linear relationship between height (y) and age (x) is the same for the two diets. Additional SS = ResidualSS (reduced model) – ResidualSS (full model) = 1120.22, and F = (1120.22/2)/(5.22290) = 107.24; p-value < 0.0001; reject $\beta_2 = \beta_3 = 0$; linear relationships between height and age not the same for the two diets.

5.14

(a) Since the columns of X are orthogonal, X'X is a diagonal matrix. Let $X'X = \Lambda = \text{diag}(\lambda_1, \lambda_2, ..., \lambda_{p+1})$. We have seen that $\hat{\boldsymbol{\beta}} = (X'X)^{-1}X'\boldsymbol{y}$. Also $V(\hat{\boldsymbol{\beta}}) = (X'X)^{-1}\sigma^2 = \Lambda^{-1}\sigma^2 = \sigma^2\text{diag}(\lambda_1^{-1}, \lambda_2^{-1}, ..., \lambda_{p+1}^{-1})$. Since the off diagonal elements are zero, $\text{Cov}(\hat{\beta}_i, \hat{\beta}_j) = 0$, for all $i \neq j$. In addition, $\hat{\beta}_i$ and $\hat{\beta}_j$ are normally distributed. Hence $\hat{\beta}_i$ and $\hat{\beta}_j$ are statistically independent.

(b) $\boldsymbol{y} = X\boldsymbol{\beta} + \gamma \boldsymbol{z} + \boldsymbol{\varepsilon}$, where $\boldsymbol{z}$ is orthogonal to the columns of X; that is, $X'\boldsymbol{z} = \boldsymbol{0}$ and $\boldsymbol{z}'X = \boldsymbol{0}'$. Let $X_1 = [X\ \boldsymbol{z}]$ be a new matrix containing the columns of X and $\boldsymbol{z}$. Then

$$\begin{pmatrix} \tilde{\boldsymbol{\beta}} \\ \tilde{\gamma} \end{pmatrix} = (X_1'X_1)^{-1}X_1'\boldsymbol{y} = \left[\begin{pmatrix} X' \\ \boldsymbol{z}' \end{pmatrix}(X\ \ \boldsymbol{z})\right]^{-1}\begin{bmatrix} X' \\ \boldsymbol{z}' \end{bmatrix}\boldsymbol{y} = \begin{bmatrix} X'X & X'\boldsymbol{z} \\ \boldsymbol{z}'X & \boldsymbol{z}'\boldsymbol{z} \end{bmatrix}^{-1}\begin{bmatrix} X' \\ \boldsymbol{z}' \end{bmatrix}\boldsymbol{y}$$

$$= \begin{bmatrix} X'X & \boldsymbol{0} \\ \boldsymbol{0}' & \boldsymbol{z}'\boldsymbol{z} \end{bmatrix}^{-1}\begin{bmatrix} X'\boldsymbol{y} \\ \boldsymbol{z}'\boldsymbol{y} \end{bmatrix} = \begin{bmatrix} (X'X)^{-1} & \boldsymbol{0} \\ \boldsymbol{0}' & (\boldsymbol{z}'\boldsymbol{z})^{-1} \end{bmatrix}\begin{bmatrix} X'\boldsymbol{y} \\ \boldsymbol{z}'\boldsymbol{y} \end{bmatrix} = \begin{bmatrix} (X'X)^{-1}X'\boldsymbol{y} \\ (\boldsymbol{z}'\boldsymbol{z})^{-1}\boldsymbol{z}'\boldsymbol{y} \end{bmatrix} = \begin{pmatrix} \hat{\boldsymbol{\beta}} \\ \tilde{\gamma} \end{pmatrix}.$$

Note that $\tilde{\boldsymbol{\beta}}$ is exactly the same as $\hat{\boldsymbol{\beta}}$, and hence they have the same distribution.

(c) Let us first explain the phrase "columns are centered about their means". Let $\boldsymbol{w}_1$, $\boldsymbol{w}_2$, …, $\boldsymbol{w}_p$ be column vectors of the matrix $W = [\boldsymbol{w}_1, \boldsymbol{w}_2, ..., \boldsymbol{w}_p]$. Let $\bar{w}_i$ be the average of column vector $\boldsymbol{w}_i$. Define $\boldsymbol{x}_i = \boldsymbol{w}_i - \boldsymbol{1}\bar{w}_i$ where $\boldsymbol{1}$ is a column vector with n ones. Then $X_1 = [\boldsymbol{x}_1, \boldsymbol{x}_2, ..., \boldsymbol{x}_p]$ has columns that are centered about their means. This implies that the sum of the elements in each column of the matrix X_1 is zero; that is, $\boldsymbol{1}'\boldsymbol{x}_i = 0$, for each i.
Defining the matrix $X = [\boldsymbol{1}, X_1]$ leads to the estimates

$$\hat{\boldsymbol{\beta}} = \begin{bmatrix} \hat{\beta}_0 \\ \hat{\boldsymbol{\beta}}_* \end{bmatrix} = (X'X)^{-1}X'\boldsymbol{y}$$

$$= \begin{bmatrix} n & \boldsymbol{0}' \\ \boldsymbol{0} & X_1'X_1 \end{bmatrix}^{-1}\begin{pmatrix} \boldsymbol{1}' \\ X_1' \end{pmatrix}\boldsymbol{y} = \begin{bmatrix} n^{-1} & \boldsymbol{0}' \\ \boldsymbol{0} & (X_1'X_1)^{-1} \end{bmatrix}\begin{pmatrix} \boldsymbol{1}'\boldsymbol{y} \\ X_1'\boldsymbol{y} \end{pmatrix} = \begin{pmatrix} \bar{y} \\ (X_1'X_1)^{-1}X_1'\boldsymbol{y} \end{pmatrix}$$

This shows that $\hat{\beta}_0 = \bar{y}$.

Furthermore, $V(\hat{\boldsymbol{\beta}}) = (X'X)^{-1}\sigma^2 = \begin{bmatrix} n^{-1} & \boldsymbol{0}' \\ \boldsymbol{0} & (X_1'X_1)^{-1} \end{bmatrix}\sigma^2$ implies that the covariance between $\hat{\beta}_0$ and $\hat{\beta}_j$, for j = 1, 2, …, p, is zero. In addition, $\hat{\boldsymbol{\beta}}$ is normally distributed. Hence $\hat{\beta}_0$ is distributed independently of all other $\hat{\beta}_j$, for j = 1, 2, …, p.

5.15 Weight (x_1); $x_2 = 0$ (type A engine); $x_2 = 1$ (type B engine);
(a) $\mu = \beta_0 + \beta_1 x_1 + \beta_2 x_2$; (b) $\mu = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2$

5.16

(a) β_3 represents the change in expected yield of catalyst 2 over catalyst 1 when temperature is held fixed.

(b) Test of $\beta_3 = 0$: $t(\hat{\beta}_3) = -0.32/0.36 = -0.89$; p-value = $2P(t(26) \le -0.89) = 0.38$; conclude $\beta_3 = 0$; no evidence to suggest a difference in catalysts.

95% confidence interval for β_2: $\hat{\beta}_2 \pm (0.975;26)\text{s.e.}(\hat{\beta}_2)$, $0.41 \pm (2.065)(0.11)$ or (0.18, 0.64).

(c) (i) $\text{Cov}(\hat{\beta}_1, \hat{\beta}_3) = 0$. Since $\hat{\boldsymbol{\beta}}$ is normally distributed, $\text{Cov}(\hat{\beta}_1, \hat{\beta}_3) = 0$ implies that $\hat{\beta}_1$ and $\hat{\beta}_3$ are independent.

(ii) 95% confidence interval for E(y) when x = 0 and z =1. Let $\theta = E(y) = \beta_0 + \beta_3$.

Estimate: $\hat{\theta} = \hat{\beta}_0 + \hat{\beta}_3 = 29.51$

$$V(\hat{\theta}) = V(\hat{\beta}_0) + V(\hat{\beta}_3) + 2\text{Cov}(\hat{\beta}_0, \hat{\beta}_3) = s^2[0.114 + 0.133 + 2(-0.0671)]$$
$$= (25.05/26)[0.114 + 0.133 + 2(-0.0671)] = 0.1087$$

$\hat{\theta} \pm (0.975;26)\sqrt{V(\hat{\theta})}$, $29.51 \pm (2.065)\sqrt{0.1087}$, or (28.83, 30.19).

(iii) 95% prediction interval

$\hat{\theta} \pm (0.975;26)\sqrt{s^2 + V(\hat{\theta})}$, $29.51 \pm (2.065)\sqrt{(25.05/26) + (0.1087)}$, or (27.37, 31.65)

(d) Model equation for catalyst 1: $E(y) = \beta_0 + \beta_1 x + \beta_2 x^2$

Model equation for catalyst 2: $E(y) = (\beta_0 + \beta_3) + (\beta_1 + \beta_4)x + (\beta_2 + \beta_5)x^2$

Test $\beta_3 = \beta_4 = 0$: Additional SS = 25.05-19.70 = 5.35. Thus F = (5.35/2)/(19.70/24) = 3.26; p-value = 0.056. There is some weak evidence that the effect of temperature changes with the catalysts.

5.17

(a) Minitab output is given below. It helps to include the square of poverty as an explanatory variable (t-ratio = 2.72 and p-value = 0.007).

On Poverty only:

```
The regression equation is
test = 74.6 - 0.536 pov

Predictor       Coef  SE Coef       T      P
Constant      74.606    1.613   46.25  0.000
pov         -0.53578  0.03262  -16.43  0.000

S = 8.76595   R-Sq = 67.3%   R-Sq(adj) = 67.1%
```

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	1	20731	20731	269.79	0.000
Residual Error	131	10066	77		
Total	132	30798			

On Poverty and (Poverty)^2:

The regression equation is
test = 79.9 - 0.850 pov + 0.00343 pov**2

Predictor	Coef	SE Coef	T	P
Constant	79.950	2.520	31.72	0.000
pov	-0.8504	0.1201	-7.08	0.000
pov**2	0.003427	0.001261	2.72	0.007

S = 8.56001   R-Sq = 69.1%   R-Sq(adj) = 68.6%

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	2	21272	10636	145.16	0.000
Residual Error	130	9526	73		
Total	132	30798			

(c) It is not necessary to include an indicator for students in the college community Iowa City (t-ratio = 0.73 and p-value = 0.467).

On Poverty, (Poverty)^2, and Indicator for Iowa City:

The regression equation is
test = 79.2 - 0.832 pov + 0.00332 pov**2 + 1.73 IowaCity

Predictor	Coef	SE Coef	T	P
Constant	79.197	2.728	29.03	0.000
pov	-0.8322	0.1229	-6.77	0.000
pov**2	0.003319	0.001272	2.61	0.010
IowaCity	1.735	2.380	0.73	0.467

S = 8.57548   R-Sq = 69.2%   R-Sq(adj) = 68.5%

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	3	21311.3	7103.8	96.60	0.000
Residual Error	129	9486.5	73.5		
Total	132	30797.8			