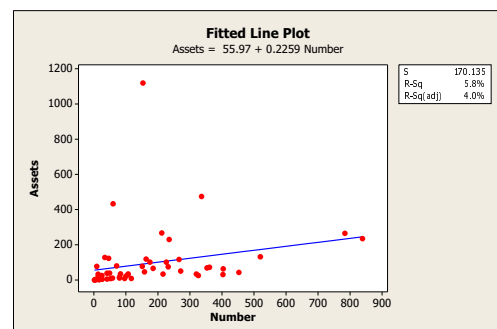
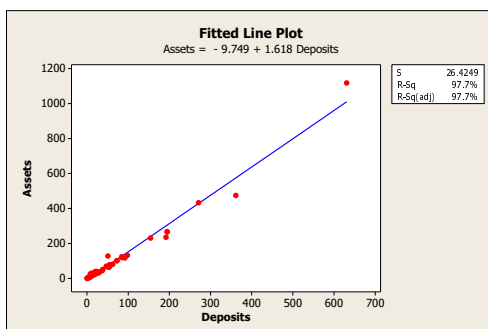
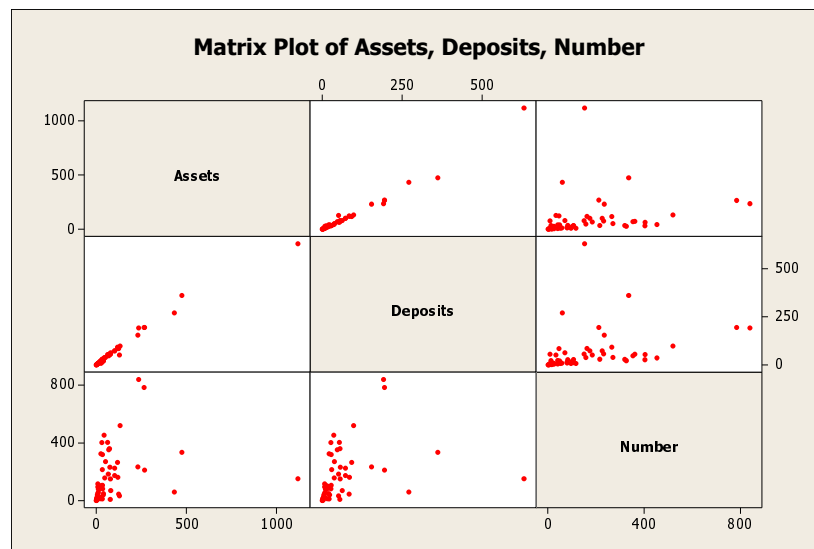


A.

• 11.3

- (a) The response variable is Bank Assets.
- (b) The explanatory (predictor) variables are Number of Banks and Deposits.
- (c) There are two predictor variables: $p = \boxed{2}$
- (d) The sample size $n = \boxed{54}$, the 50 states plus Guam, Puerto Rico, the Virgin Islands and the District of Columbia.

• 11.4



(continued)

Regression Analysis: Assets versus Deposits, Number

The regression equation is

$$\text{Assets} = 1.58 + 1.67 \text{ Deposits} - 0.0853 \text{ Number}$$

Predictor	Coef	SE Coef	T	P
Constant	1.580	4.183	0.38	0.707
Deposits	1.66642	0.03005	55.46	0.000
Number	-0.08526	0.01725	-4.94	0.000

$$S = 21.9393 \quad R\text{-Sq} = 98.5\% \quad R\text{-Sq}(\text{adj}) = 98.4\%$$

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	2	1572988	786494	1633.99	0.000
Residual Error	51	24548	481		
Total	53	1597536			

- (a) Deposits appears to be the stronger predictor.
- (b) New York State is an outlier. New York has an unusually large amount of both assets (\$1119.2 billion) and deposits (\$630.7 billion) compared to other states. This is presumably due to the concentration of investment banks near Wall Street in New York City.

(c)

$$\text{Assets} = 1.58 + (1.67)(\text{Deposits}) - (0.0853)(\text{Number of Banks})$$

- (d) Assets increase by \$1.67 billion on average for every one-billion dollar increase in deposits, when number of banks is held constant.
- (e) Assets decrease by \$85.3 million on average for every additional insured commercial bank, when deposits are held constant.
- (f) The second state is expected to have more Assets, based on the interpretation in (e).
- (g) Based on simple regression:

$$\text{Assets} = 55.97 + (0.2259)(\text{Number of Banks})$$

we estimate that an extra bank in Iowa increases Iowa Assets by 0.2259 billion dollars, or \$225.9 million

- (h) i. Total Effect = 0.2259 (from simple regression)
 ii. Direct Effect = -0.0853 (from multiple regression)
 iii.

$$\text{Total Effect} = \text{Direct Effect} + \text{Indirect Effect}$$

$$\text{Indirect Effect} = \text{Total Effect} - \text{Direct Effect} = 0.2259 - (-0.0853) = \span style="border: 1px solid black; padding: 2px;">0.3112$$

iv.

First: Number $\uparrow \implies$ Deposits \uparrow (by positive correlation 0.325)

Then: Deposits $\uparrow \implies$ Assets \uparrow (by positive correlation 0.989)

B.

- **11.8** The coefficients from the sample multiple regression equation

$$\hat{y} = b_0 + b_1x_1 + b_2x_2$$

OR

$$\text{(Estimated) Profit} = b_0 + b_1 \times \text{Assets} + b_2 \times \text{Sales}$$

are

$$b_0 = 2.3405$$

$$b_1 = 0.007406$$

$$b_2 = 0.02610$$

so the sample regression equation is

$$\text{Profit} = 2.3405 + (0.007406)(\text{Assets}) + (0.02610)(\text{Sales})$$

- **11.13**

$$s = \boxed{2.44958} \quad s^2 = \text{MSE} = \boxed{6.000}$$

- **11.14**

From the MINITAB Display Descriptive Statistics procedure,

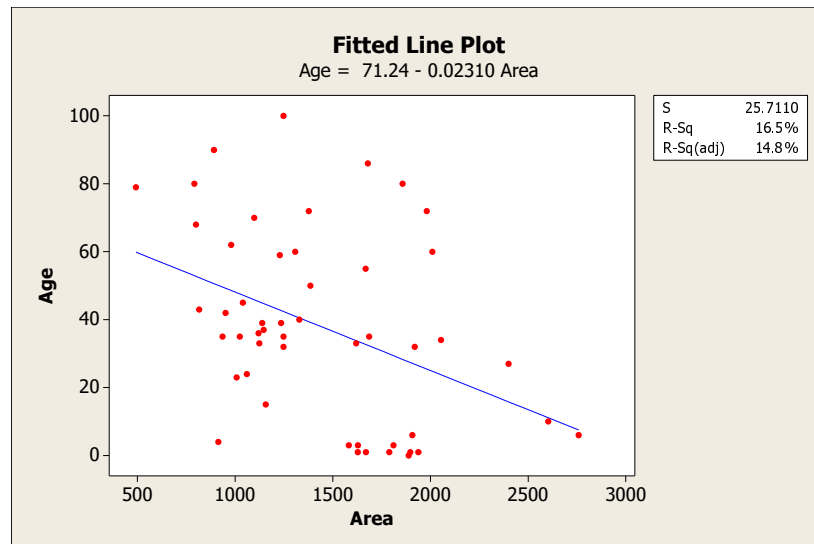
$$s_y = \boxed{3.037}$$

The regression standard deviation (sometimes called “standard error”) s is smaller than the standard deviation s_y for the variable $y = \text{Profits}$:

$$s = 2.450 < 3.037 = s_y$$

We expect this to happen if the predictor variables $x_1 = \text{Assets}$ and $x_2 = \text{Sales}$ are successful in predicting y . If it’s easier to predict y , the variability of the prediction will be reduced.

C.



(a) Indirect Effect of Area upon Sales Price (through Age) is

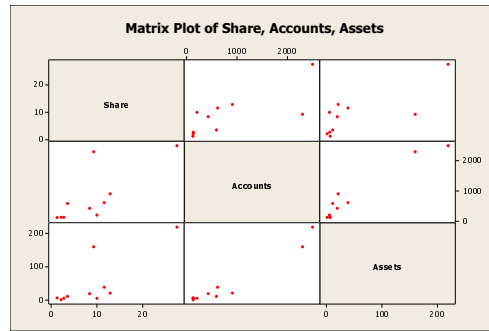
$$(-0.02310 \text{ years}) \times (-\$803/\text{year}) = \boxed{\$18.55}$$

(b) **Logic:** (Area \uparrow by one square foot) \implies (Age \downarrow by 0.02310 years)
and each extra year costs \$803 in Sales Price.

(continued)

D.

• 11.25



(a)

$$\text{Share} = 5.16 - 0.00031 \text{ Accounts} + 0.0828 \text{ Assets}$$

(b) $s = 5.488$ (rounded to three decimal places)

(c) Market Share seems positively related to both Assets and Number of Accounts.

(d) The F test indicates that at least one of the two predictor variables is linearly related to Market Share ($F = 5.45$, P -value = 0.037.)

(e) Neither predictor is significant in the multiple regression model.
(P -value for Accounts = 0.971, P -value for Assets = 0.418)

(f) The negative slope $b_1 = -0.00031$ for Accounts implies that market share decreases as the number of accounts increases, when assets are held constant. Conversely, market share increases with fewer accounts.

No. The implication holds only if assets are held constant, and closing customer accounts would necessarily reduce assets held by online brokerages, too.

(g) The full model is not a possible choice for best conservative model since its variables are not significant. The regression models which use only a single predictor are both significant (P -value for Accounts alone is 0.012, P -value for Assets alone is 0.008), so both models are possible choices. The best model is the one that uses Assets alone since it has the highest R^2 (60.9% vs. 56.7%).

If we let x_1 be the variable Accounts and x_2 be the variable Assets, then the prediction equation is

$$\hat{y} = 5.08 + 0.0793 x_2$$

Or using the common English from the MINITAB output,

$$\text{Share} = 5.08 + 0.0793 \text{ Assets}$$

(continued)

(h)

$$\begin{aligned}\hat{y} &= 5.08 + 0.0793 x_2 \\ &= 5.08 + (0.0793)(20) \\ &= \boxed{6.666\%}\end{aligned}$$

(i) Predicting in MINITAB with $x_2^* = 20$ results in the output

Predicted Values for New Observations				
New	Fit	SE Fit	95% CI	95% PI
Obs				
1	6.67	1.75	(2.64, 10.70)	(-5.84, 19.18)

Values of Predictors for New Observations	
New	Assets
Obs	
1	20.0

Since the question asks about a *particular* brokerage, we need the prediction interval, not the confidence interval. The 95% PI from MINITAB is

$$(-5.84\%, 19.18\%)$$

Of course, market share cannot be *negative*, so an improved answer is $\boxed{(0\%, 19.18\%)}$ which truncates the original prediction interval at the value 0.

Side Note to Students:

Caution is advised if a prediction or confidence interval contains impossible values for the response variable (such as negative market share) and thus requires truncation.

The fact that the interval contains impossible values shows that the mathematical assumptions behind regression are not completely satisfied for these particular values of the predictor variables.

The interval may still provide a useful estimate but should be taken by business managers “with a grain of salt.”

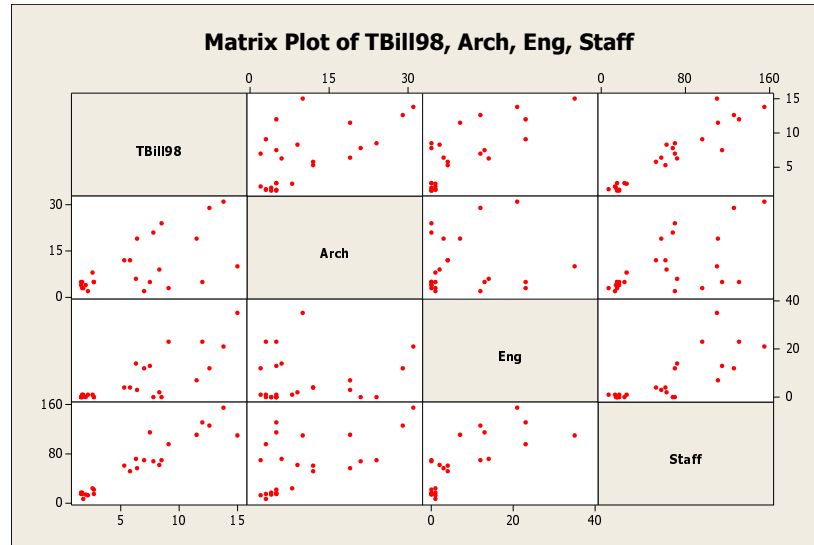
Sometimes *data transformations* (similar to the log transform which we applied in the Topic 8 Notes and homework) can solve the truncation problem. In an actual business application, the manager may wish to consult with a professional statistician.

(j) The two predictor variables are highly correlated. This means both variables provide *similar* information so neither variable is significant if used *in addition* to the other.

(continued)

E.

• 11.35



- (a) The variables Architects, Engineers, and Staff are to be used to predict Total Billings in 1998. Multiple regression in MINITAB produces the following output:

```
Regression Analysis: TBill98 versus Arch, Eng, Staff

The regression equation is
TBill98 = 0.883 + 0.138 Arch + 0.160 Eng + 0.0478 Staff

Predictor    Coef    SE Coef    T    P
Constant    0.8832   0.4165    2.12  0.046
Arch        0.13788  0.04369    3.16  0.005
Eng         0.16009  0.05114    3.13  0.005
Staff       0.04784  0.01341    3.57  0.002

S = 1.16170   R-Sq = 93.5%   R-Sq(adj) = 92.6%

Analysis of Variance

Source      DF      SS      MS      F      P
Regression    3    406.60   135.53  100.43  0.000
Residual Error 21    28.34    1.35
Total        24   434.94
```

Therefore the regression equation from this model is

$$\text{TBill98} = 0.883 + 0.138 \text{ Arch} + 0.160 \text{ Eng} + 0.0478 \text{ Staff}$$

- (b) $s = 1.162$
- (c) 1998 Total Billings appears to be positively related to all three predictors variables.
- (d) All three predictors are significant (P -value = 0.005 for Arch, P -value = 0.005 for Eng, P -value = 0.002 for Staff.)

(continued)

- (e) There are 7 possible models which you can run in MINITAB and compare, but it is a mathematical fact that using more predictors never decreases R^2 (i.e., using more variables “provides a closer fit” to the sample data.)

Since all three variables are significant, the *full* model (the one that uses all available predictors) is a possible choice for best conservative model. Since the full model also has highest R^2 , it is the best model. Therefore the prediction equation is

$$TBill98 = 0.883 + 0.138 \text{ Arch} + 0.160 \text{ Eng} + 0.0478 \text{ Staff}$$

- (f) If you enter the three predictor values 10, 20, 100 (in that order, with a space and no commas) for the “new observations” option in Regression, MINITAB provides answers to (e), (f), and (g) with the output

Predicted Values for New Observations					
New					
Obs	Fit	SE Fit	90% CI	90% PI	
1	10.248	0.399	(9.561, 10.935)	(8.134, 12.361)	
Values of Predictors for New Observations					
New					
Obs	Arch	Eng	Staff		
1	10.0	20.0	100		

From the output,

$$\hat{y} = \boxed{\$10,248,000}$$

- (g) We are 90% certain that the firm has 1998 Total Billings between \$8,134,000 and \$12,361,000.)
- (h) We are 90% certain that all such firms have mean 1998 Total Billings between \$9,561,000 and \$10,935,000.)

(continued)

F.

• **11.37**

If we let x_1 , x_2 , x_3 be HSM, HSS, and HSE, respectively, then testing HSS in the multiple regression model amounts to testing

$$H_0: \beta_2 = 0$$

$$H_A: \beta_2 \neq 0$$

in the full model

$$y = \beta_0 + \beta_1x_1 + \beta_2x_2 + \beta_3x_3 + \varepsilon$$

We Fail to Reject H_0 since $t = 0.91$ and $P\text{-value} = 0.362 > .05 = \alpha$.

Conclusion: HSS is not a significant predictor for GPA, after accounting for HSM and HSE.

• **11.38**

Testing HSS in a model which uses HSS as the only predictor means testing

$$H_0: \beta_2 = 0$$

$$H_A: \beta_2 \neq 0$$

in the model

$$y = \beta_0 + \beta_2x_2 + \varepsilon$$

We Reject H_0 ($t = 5.20$, $P\text{-value} = 0.000$).

Conclusion: HSS is a significant predictor for GPA.

Note to students: The conclusions in 11.37 and 11.38 do not contradict each other:

Though science grades are not significant when math and English grades are also in the model, when regression is performed on HSS alone it is highly-significant in predicting college GPA.

This shows that conclusions for any predictor variable depend upon what other predictor variables are also in the model. Alone, high school science grades are useful in predicting GPA. However, science grades don't contribute much if used *in addition* to math and English grades.

(continued)

• 11.43

(a)

Model	Variables	R^2
(a)	HSM, HSS, HSE	20.5%
(b)	HSM, HSE	20.2%
(c)	HSM, HSS	20.0%
(d)	HSS, HSE	12.3%
(e)	HSM	19.1%

The table above shows R^2 for the different models requested by the textbook. Notice that HSM is a necessary variable, as without it R^2 is quite low. However, once HSM is in the model, addition of other variables seems to be of limited use. The conclusion is that almost all of the information for predicting GPA is contained in HSM.

(b)

Model	Predictors	All predictors significant?	R^2
1	HSM	Yes	19.1
2	HSS	Yes	10.9
3	HSE	Yes	8.4
4	HSM, HSS	No	20.0
5	HSM, HSE	No	20.2
6	HSS, HSE	No	12.3
7	HSM, HSS, HSE	No	20.5

Any of Models 1, 2, 3 are possible choices. (Models 4–7 are disqualified from consideration since they each contain at least one nonsignificant predictor.) The best conservative model is Model 1 (highest R^2), with prediction equation

$$\text{GPA} = 0.908 + 0.208 \text{ HSM}$$

(c) Interpret slopes in the chosen model:

- $\beta_1 = 0.208$:

GPA increases on average by 0.208 points for every one-point increase in high-school math grades.

- $\beta_2 = 0$:

GPA is not related to high-school science grades, after accounting for high-school math grades.

- $\beta_3 = 0$:

GPA is not related to high-school English grades, after accounting for high-school math grades.

(continued)

• 11.59

The textbook is using the full multiple regression model (with all three predictors HSM, HSS, HSE) in this exercise.

- (a) The regression model for the population of all computer science students is

$$\mu_{\text{GPA}} = \beta_0 + \beta_1\text{HSM} + \beta_2\text{HSS} + \beta_3\text{HSE}$$

For the subpopulation of students who get certain grades (HSM = 8, HSS = 9, HSE = 7), the regression model becomes

$$\mu_{\text{GPA}} = \beta_0 + 8\beta_1 + 9\beta_2 + 7\beta_3$$

All that we need to make an actual GPA prediction for such students are the slope estimates $b_1 = \hat{\beta}_1$, $b_2 = \hat{\beta}_2$, and $b_3 = \hat{\beta}_3$ calculated from the CSDATA database.

- (b) The sample regression model is

$$\widehat{\text{GPA}} = 0.590 + (0.169)(\text{HSM}) + (0.0343)(\text{HSS}) + (0.0451)(\text{HSE})$$

so the estimated mean GPA for this group of students is

$$\widehat{\text{GPA}} = 0.590 + (0.169)(8) + (0.0343)(9) + (0.0451)(7) = \boxed{2.566}$$

• 11.60

- (a)

$$\mu_{\text{GPA}} = \beta_0 + 7\beta_1 + 6\beta_2 + 9\beta_3$$

- (b) Estimated mean GPA = $\boxed{2.385}$

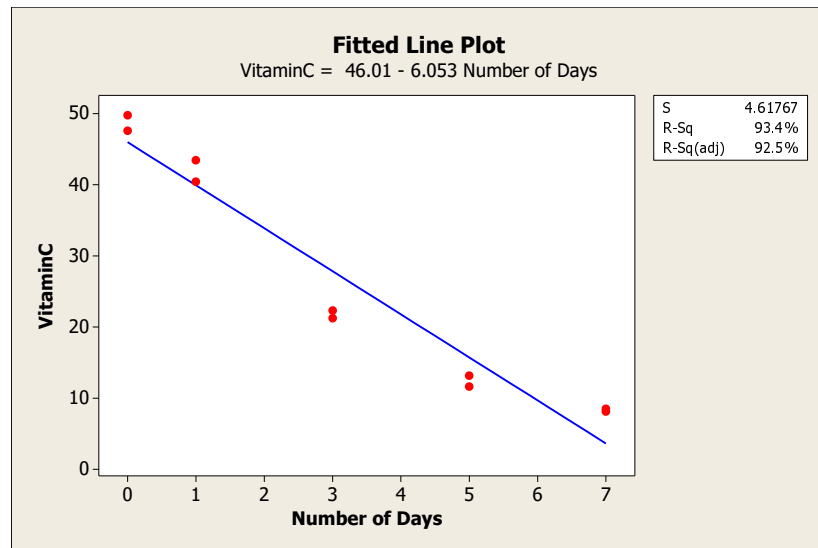
(continued)

G.

• 11.109

(a) and (b)

Here is the fitted-line plot for linear regression:

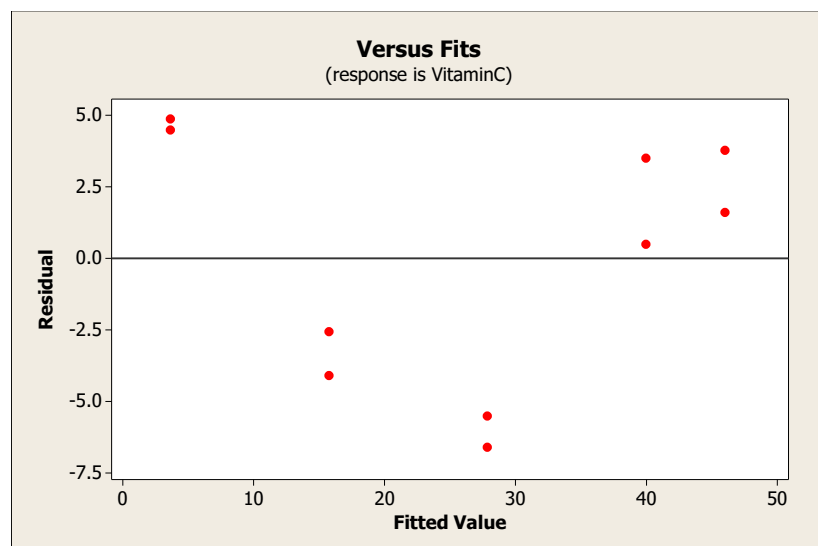


The sample linear regression equation is

$$\text{Vitamin C Loss} = 46.0 - 6.05 \text{ Number of Days}$$

The significance test for Number of Days shows ($t = -10.62$ $P\text{-value} = 0$) so the predictor is significant.

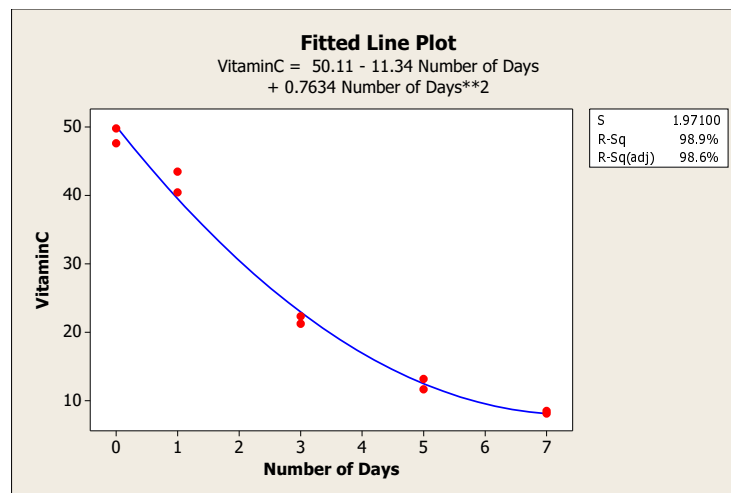
Unfortunately the residuals plot shows a clear curved pattern so regression assumptions appear to be violated. This means that the regression output is invalid. In particular, we can't trust the numbers from the significant test, nor can we use the regression equation!



(continued)

(c) and (d)

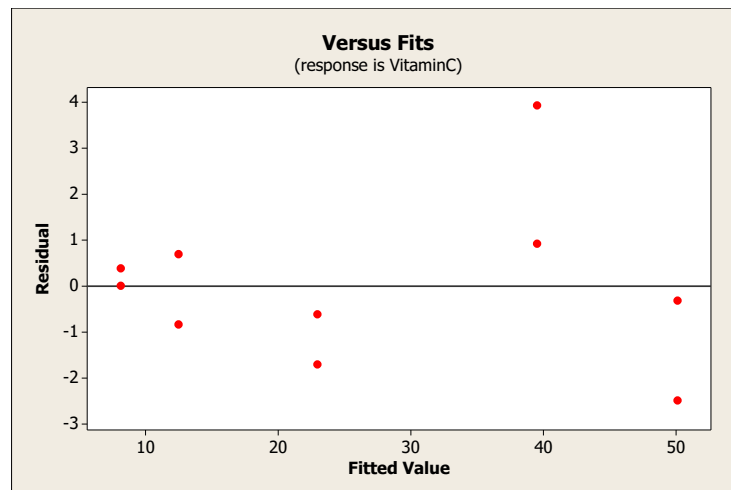
The *quadratic* fitted-line plot below shows a closer fit to the points.



More importantly, the residuals plot after quadratic regression (below) appears to show a fairly random pattern so regression assumptions appear to be satisfied for the quadratic model. Therefore the regression output is valid.

The *t* tests for quadratic regression show

- Number of Days: ($t = -12.55$ P -value = 0) (significant)
- (Number of Days)² : ($t = 6.08$ P -value = 0.001) (significant)



Therefore quadratic regression can be used to predict loss of Vitamin C over time. The most accurate regression equation available from MINITAB is

$$\text{Vitamin C Content} = 50.105 - 11.3413 (\text{Number of Days}) + 0.7634 (\text{Days Squared})$$

(e) Predicted Vitamin C content after 4 days is 16.9542 (mg/100g)

• 11.110

(a) Vitamin A

- Linear regression is significant for Vitamin A loss ($t = -28.4$ P -value = 0.022). Also the residuals plot shows no strong patterns so assumptions appear to be satisfied for linear regression.
- Quadratic regression is not significant for Vitamin A loss:
 - * Number of Days: ($t = 0.49$ P -value = 0.640) (not significant)
 - * (Number of Days)² : ($t = -1.34$ P -value = 0.222) (not significant)
- Therefore linear regression should be used to model Vitamin A loss. The most-accurate regression model is

$$\text{Vitamin A Content} = 3.33829 - 0.03884 (\text{Number of Days})$$

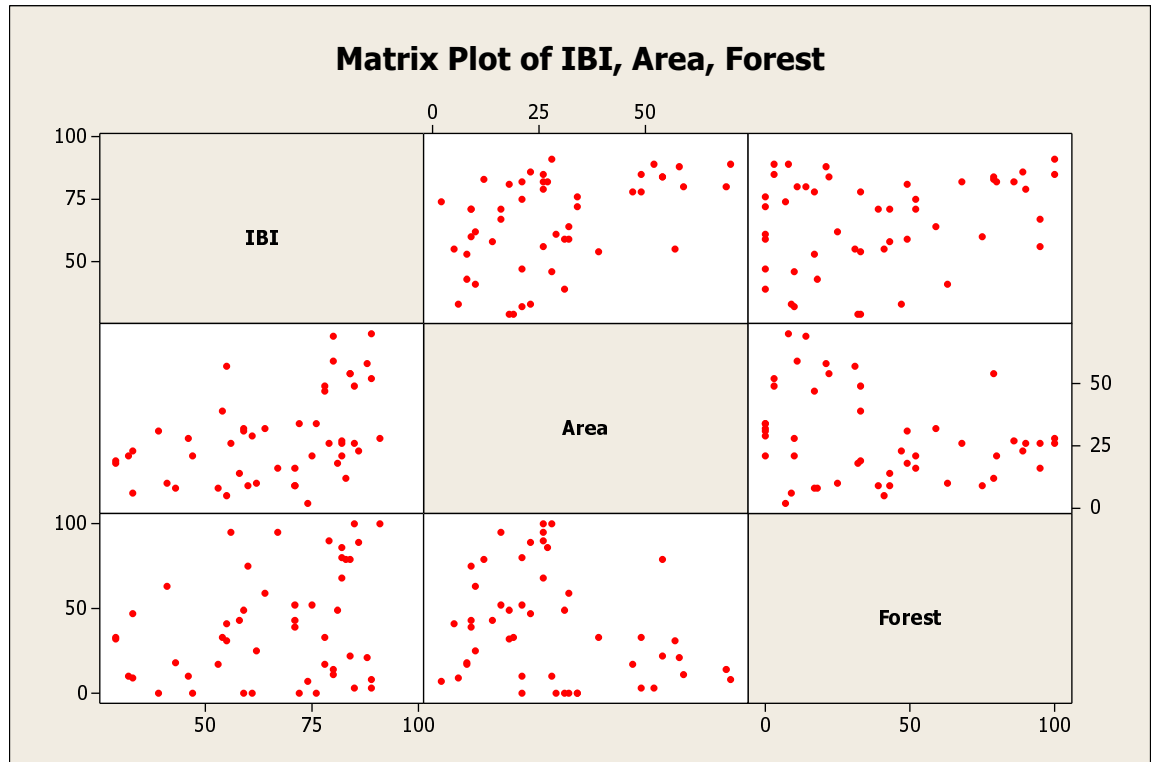
- Predicted Vitamin A content after 4 days is 3.18293 (mg/100g)

(b) Vitamin E

- Linear regression is not significant for Vitamin E content ($t = -0.34$, P -value = 0.744). In other words the data do not show that Vitamin E is lost over time.
- Stop! No regression model should be used to predict Vitamin E content after 4 days.

(continued)

H.



(a) The F test ($F = 12.78$, $P\text{-value} = 0$) is significant. This shows that at least one of the two predictor variables Area and Forest is an important predictor for IBI.

(b) • t test for Area: ($t = 4.51$ $P\text{-value} = 0$) (significant)
 Area is an important predictor of IBI, after accounting for Forest.

• t test for Forest: ($t = 3.37$ $P\text{-value} = 0.002$) (significant)
 Forest is an important predictor of IBI, after accounting for Area.

(c) The predictors Area and Forest appear to be negatively correlated. This is confirmed by the correlation $r = -0.257$. If a watershed has greater area, it tends to have a smaller percentage of forested land.

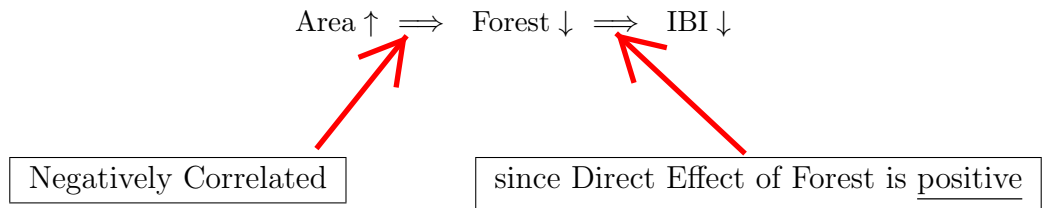
(d) •
$$\text{IBI} = 52.9 + 0.460 \text{ Area}$$

•
$$\text{IBI} = 59.9 + 0.153 \text{ Forest}$$

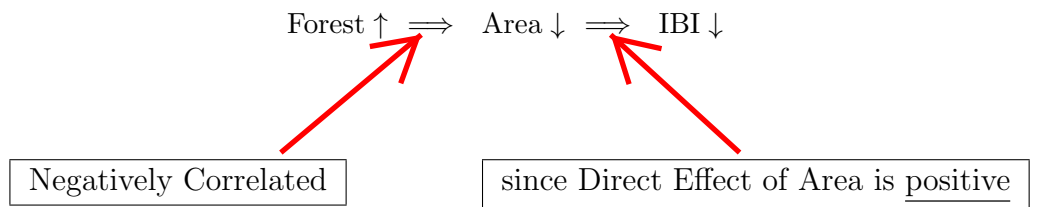
•
$$\text{IBI} = 40.6 + 0.569 \text{ Area} + 0.234 \text{ Forest}$$

(continued)

- (e)
- Total Effect of Area on IBI = 0.460
 - Direct Effect on IBI = 0.569
 - Indirect Effect on IBI = (Total Effect) – (Direct Effect) = $0.460 - 0.569 = -0.109$
 - Reason for negative sign:



- (f)
- Total Effect of Forest on IBI = 0.153
 - Direct Effect on IBI = 0.234
 - Indirect Effect on IBI = (Total Effect) – (Direct Effect) = $0.153 - 0.234 = -0.081$
 - Reason for negative sign:



(g) The IBI for a stream increases by 0.234 points on average for every one-percent increase in forested land of the stream's watershed, when watershed area is held constant.

(h) The best conservative model is the full model.

A 90% prediction interval for IBI is (50.33, 101.88) points

(i) (35.69, 91.78) points

(j) The single best estimate for (Stream C IBI) – (Stream D IBI) is

$$(90\% - 57\%) \times 0.15313 = (33\%) \times 0.15313 = \span style="border: 1px solid black; padding: 2px;">5.05329 \text{ points}$$

(end of solution)