Other Regression Models

1. **Multiple Linear Regression**: More than one predictor variables
2. **Categorical Predictors**: Predictor variables are categories such as CPU type, disk type, and so on.
3. **Curvilinear Regression**: Relationship is nonlinear
4. **Transformations**: Errors are not normally distributed or the variance is not homogeneous
5. **Outliers**
6. **Common mistakes** in regression

Multiple Linear Regression Models

$$y = b_0 + b_1 x_1 + b_2 x_2 + \cdots + b_k x_k + e$$

- Given a sample of *n* observations with *k* predictors

$$\{(x_{11}, x_{21}, \ldots, x_{k1}, y_1), \ldots, (x_{1n}, x_{2n}, \ldots, x_{kn}, y_n)\}$$

$$\begin{aligned}
y_1 &= b_0 - b_1 x_{11} - b_2 x_{21} - \cdots - b_k x_{k1} + e_1 \\
y_2 &= b_0 - b_1 x_{12} - b_2 x_{22} - \cdots - b_k x_{k2} + e_2 \\
&\;. \\
&\;. \\
&\;. \\
y_n &= -b_0 - b_1 x_{1n} - b_2 x_{2n} - \cdots - b_k x_{kn} + e_n
\end{aligned}$$

Vector Notation

In vector notation, we have:

- $$\begin{bmatrix} y_1 \\ y_2 \\ . \\ . \\ . \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & x_{11} & x_{21} & \cdots & x_{k1} \\ 1 & x_{12} & x_{22} & \cdots & x_{k2} \\ . & . & . & . & . \\ . & . & . & . & . \\ . & . & . & . & . \\ 1 & x_{1n} & x_{2n} & \cdots & x_{kn} \end{bmatrix} \begin{bmatrix} b_0 \\ b_1 \\ . \\ . \\ . \\ b_k \end{bmatrix} + \begin{bmatrix} e_1 \\ e_2 \\ . \\ . \\ . \\ e_n \end{bmatrix}$$
- or $$\mathbf{y} = \mathbf{X}\mathbf{b} + \mathbf{e}$$
- All elements in the first column of **X** are 1.
  See Box 15.1 for regression formulas.

Example 15.1

- Seven programs were monitored to observe their resource demands. In particular, the number of disk I/O's, memory size (in kBytes), and CPU time (in milliseconds) were observed.

CPU Time y_i	Disk I/O's x_{1i}	Memory Size x_{2i}
2	14	70
5	16	75
7	27	144
9	42	190
10	39	210
13	50	235
20	83	400

Example 15.1 (Cont)

$$\text{CPU time} = b_0 + b_1(\text{number of disk I/O's}) + b_2(\text{memory size})$$

- In this case:

$$\mathbf{X} = \begin{bmatrix} 1 & 14 & 70 \\ 1 & 16 & 75 \\ 1 & 27 & 144 \\ 1 & 42 & 190 \\ 1 & 39 & 210 \\ 1 & 50 & 235 \\ 1 & 83 & 400 \end{bmatrix}$$

$$\mathbf{X}^{\mathrm{T}}\mathbf{X} = \begin{bmatrix} 7 & 271 & 1324 \\ 271 & 13,855 & 67,188 \\ 1324 & 67,188 & 326,686 \end{bmatrix}$$

Example 15.1 (Cont)

$$\mathbf{C} = (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1} = \begin{bmatrix} 0.6297 & 0.0223 & -0.0071 \\ 0.0223 & 0.0280 & -0.0058 \\ -0.0071 & -0.0058 & 0.0012 \end{bmatrix}$$

$$\mathbf{X}^{\mathrm{T}}\mathbf{y} = \begin{bmatrix} 66 \\ 3375 \\ 16,388 \end{bmatrix}$$

- The regression parameters are:

$$\mathbf{b} = (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{y} = (-0.1614, 0.1182, 0.0265)^{\mathrm{T}}$$

- The regression equation is:

$$\begin{aligned} \text{CPU time} \quad &= \quad -0.1614 + 0.1182(\text{number of disk I/O's}) + \\ &\quad\quad 0.0265(\text{memory size}) \end{aligned}$$

Example 15.1 (Cont)

	CPU Time y_i	Disk I/O's x_{1i}	Memory Size x_{2i}	Est. CPU time $\hat{y}_i$	Error e_i	Error^2 e_i^2
	2	14	70	3.3490	-1.3490	1.8198
	5	16	75	3.7180	1.2820	1.6436
	7	27	144	6.8472	0.1528	0.0233
	9	42	190	9.8400	-0.8400	0.7053
	10	39	210	10.0151	-0.0151	0.0002
	13	50	235	11.9783	1.0217	1.0439
	20	83	400	20.2529	-0.2529	0.0639
Σ	66	271	1324	66.0000	-0.0003	5.3000

- From the table we see that SSE is:

$$\text{SSE} = \Sigma e_i^2 = 5.3$$

Example 15.1 (Cont)

- An alternate method to compute SSE is to use:

$$\text{SSE} = \{\mathbf{y}^{\mathrm{T}}\mathbf{y} - \mathbf{b}^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{y}\}$$

- For this data, SSY and SS0 are:

$$\text{SSY} = \Sigma y_i^2 = 828$$

$$\text{SS0} = n\bar{y}^2 = 622.29$$

- Therefore, SST and SSR are:

$$\text{SST} = \text{SSY} - \text{SS0} = 828 - 622.29 = 205.71$$

$$\text{SSR} = \text{SST} - \text{SSE} = 205.71 - 5.3 = 200.41$$

Example 15.1 (Cont)

- The coefficient of determination R^2 is:

$$R^2 = \frac{\text{SSR}}{\text{SST}} = \frac{200.41}{205.71} = 0.97$$

- Thus, the regression explains 97% of the variation of y.
- Coefficient of multiple correlation:

$$R = \sqrt{0.97} = 0.99$$

- Standard deviation of errors is:

$$s_e = \sqrt{\frac{\text{SSE}}{n-3}} = \sqrt{5.3/4} = 1.2$$

Example 15.1 (Cont)

- Standard deviations of the regression parameters are:

$$\text{Estimated std. dev. of } b_0 = s_e\sqrt{c_{00}} = 1.2\sqrt{0.6297} = 0.9131$$

$$\text{Estimated std. dev. of } b_1 = s_e\sqrt{c_{11}} = 1.2\sqrt{0.0280} = 0.1925$$

$$\text{Estimated std. dev. of } b_2 = s_e\sqrt{c_{22}} = 1.2\sqrt{0.0012} = 0.0404$$

- The 90% t-value at 4 degrees of freedom is 2.132.

$$90\% \text{ Conf. interval of } b_0 = -0.1614 \mp (2.132)(0.9131) = (-2.11,\ 1.79)$$

$$90\% \text{ Conf. interval of } b_1 = 0.1182 \mp (2.132)(0.1925) = (-0.29,\ 0.53)$$

$$90\% \text{ Conf. interval of } b_2 = 0.0265 \mp (2.132)(0.0404) = (-0.06,\ 0.11)$$

None of the three parameters is significant at a 90% confidence level.

Example 15.1 (Cont)

- A single future observation for programs with 100 disk I/O's and a memory size of 550:

$$\begin{aligned} y_{1p} &= b_0 + b_1 x_1 + b_2 x_2 \\ &= -0.1614 + 0.1182(100) + 0.0265(550) = 26.2375 \end{aligned}$$

- Standard deviation of the predicted observation is:

$$s_{y_{1p}} = s_e \sqrt{\{1 + \mathbf{x}^{\mathrm{T}}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{x}\}} = 1.2\sqrt{1 + 7.4118} = 3.3435$$

- 90% confidence interval using the t value of 2.132 is:

$$26.2375 \mp (2.132)(3.3435) = (19.1096,\ 33.3363)$$

Example 15.1 (Cont)

- Standard deviation for a mean of large number of observations is:

$$s_{\hat{y}_p} = s_e\sqrt{\{\mathbf{x}^{\mathrm{T}}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{x}\}} = 1.2\sqrt{7.4118} = 3.1385$$

- 90% confidence interval is:

$$26.2375 \mp (2.132)(3.1385) = (19.5467,\ 32.9292)$$

Analysis of Variance

- Test the hypothesis that SSR is less than or equal to SSE.

$$\mathrm{SST} = \mathrm{SSY} - \mathrm{SS0} = \mathrm{SSR} + \mathrm{SSE}$$

- Degrees of freedom = Number of independent values required to compute

$$\begin{array}{ccccccccc} \mathrm{SST} & = & \mathrm{SSY} & - & \mathrm{SS0} & = & \mathrm{SSR} & + & \mathrm{SSE} \\ n-1 & = & n & - & 1 & = & k & + & (n-k-1) \end{array}$$

- Assuming
  - Errors are i.i.d. Normal ⇒ y's are also normally distributed,
  - x's are nonstochastic ⇒ Can be measured without errors
- Various sums of squares have a chi-square distribution with the degrees of freedom as given above.

F-Test

- Given SSi and SSj with ν_i and ν_j degrees of freedom, the ratio $(SSi/\nu_i)/(SSj/\nu_j)$ has an F distribution with ν_i numerator degrees of freedom and ν_j denominator degrees of freedom.
- Hypothesis that the sum SSi is less than or equal to SSj is rejected at α significance level, if the ratio $(SSi/\nu_i)/(SSj/\nu_j)$ is greater than the $1-\alpha$ quantile of the F-variate.
- This procedure is also known as **F-test**.
- The F-test can be used to check:
  Is SSR is significantly higher than SSE?
  $\Rightarrow$ Use F-test $\Rightarrow$ Compute $(SSR/\nu_R)/(SSE/\nu_e) = MSR/MSE$

F-Test (Cont)

$$\mathrm{MSR} = \frac{\mathrm{SSR}}{k} \qquad \text{and} \qquad \mathrm{MSE} = \frac{\mathrm{SSE}}{n-k-1}$$

- MSE = Variance of Error
- MSR/MSE has F[k, n-k-1] distribution
- F-test = Null hypothesis that y doesn't depend upon any x_j: against an alternate hypothesis that y depends upon at least one x_j and therefore, at least one $b_j \neq 0$. $b_1 = b_2 = \cdots = b_k = 0$
- If the computed ratio is less than the value read from the table, the null hypothesis cannot be rejected at the stated significance level.
- In simple regression models,
  If the confidence interval of b_1 does not include zero
  ⇒ Parameter is nonzero
  ⇒ Regression explains a significant part of the response variation
  ⇒ F-test is not required.

ANOVA Table for Multiple Linear Regression

- See Table 15.3 on page 252

Component	Sum of Squares	%Variation	DF	Mean Square	F-Comp.	F-Table
y	SSY= $\Sigma\ y^2$		n			
$\bar{y}$	SS0= $n\bar{y}^2$		1			
y-$\hat{y}$	SST=SSY-SS0	100	n-1			
Regression	SSR = SST-SSE	$100\left(\frac{\mathrm{SSR}}{\mathrm{SST}}\right)$	k	MSR = $\frac{\mathrm{SSR}}{k}$	$\frac{\mathrm{MSR}}{\mathrm{MSE}}$	$F_{[1-\alpha;k,n-k-1]}$
Errors	SSE=$\mathbf{y}^{\mathrm{T}}\mathbf{y}$-$\mathbf{b}^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{y}$	$100\left(\frac{\mathrm{SSE}}{\mathrm{SST}}\right)$	n-k-1	MSE=$\frac{\mathrm{SSE}}{n-k-1}$		

$s_e = \sqrt{\mathrm{MSE}}$

Example 15.2

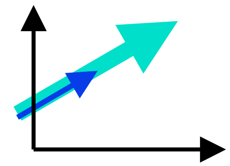
- For the Disk-Memory-CPU data of Example15.1
- Computed F ratio > F value from the table
  ⇒ Regression does explain a significant part of the variation

Component	Sum of Squares	%Variation	DF	Mean Square	F-Comp.	F-Table
y	828.					
$\bar{y}$	622.					
y-$\bar{y}$	206.	100.0%	6			
Regression	200.	97.4%	2	100.20	75.40	4.32
Errors	5.32	2.6%	4	1.33		

$$s_e=\sqrt{\text{MSE}}=\sqrt{1.33}=1.15$$

- Note: Regression passed the F test ⇒ Hypothesis of all parameters being zero cannot be accepted. However, none of the regression parameters are significantly different from zero. This contradiction ⇒ Problem of **multicollinearity**

Problem of Multicollinearity



- Two lines are said to be collinear if they have the same slope and same intercept.
- These two lines can be represented in just one dimension instead of the two dimensions required for lines which are not collinear.
- Two collinear lines are not independent.
- When two predictor variables are linearly dependent, they are called collinear.
- Collinear predictors $\Rightarrow$ Problem of multicollinearity $\Rightarrow$ Contradictory results from various significance tests.
- High Correlation $\Rightarrow$ Eliminate one variable and check if significance improves

Example 15.3

- For the data of Example 15.2, n=7, $\Sigma x_{1i}=271$, $\Sigma x_{2i}=1324$, $\Sigma x_{1i}^2=1385$, $\Sigma x_{2i}^2=326{,}686$, $\Sigma x_{1i}x_{2i}=67{,}188$.

$$\text{Correlation}(x_1, x_2) = R_{x_1 x_2}$$
$$= \frac{\sum x_{1i}x_{2i} - \frac{1}{n}\left(\sum x_{1i}\right)\left(\sum x_{2i}\right)}{\left[\sum x_{1i}^2 - \frac{1}{n}\left(\sum x_{1i}\right)\left(\sum x_{1i}\right)\right]^{1/2}\left[\sum x_{2i}^2 - \frac{1}{n}\left(\sum x_{2i}\right)\left(\sum x_{2i}\right)\right]^{1/2}}$$

$$= \frac{67,188 - \frac{1}{7}(271)(1324)}{\left[1385 - \frac{1}{7}(271)(271)\right]^{1/2}\left[326,686 - \frac{1}{7}(1324)(1324)\right]^{1/2}} = 0.9947$$

- Correlation is high
  $\Rightarrow$ Programs with large memory sizes have more I/O's
- In Example14.1, CPU time on number of disk I/O's regression was found significant.

Example 15.3 (Cont)

- Similarly, in Exercise 14.3, CPU time is regressed on the memory size and the resulting regression parameters are found to be significant.
- Thus, either the number of I/O's or the memory size can be used to estimate CPU time, but not both.
- **Lesson**:
  - *Adding a predictor variable does not always improve a regression.*
  - *If the variable is correlated to other predictors, it may reduce the statistical accuracy of the regression.*
- Try all 2^k possible subsets and choose the one that gives the best results with small number of variables.
- Correlation matrix for the subset chosen should be checked

Regression with Categorical Predictors

- Note: If all predictor variables are categorical, use one of the experimental design and analysis techniques for statistically more precise (less variant) results Use regression if most predictors are quantitative and only a few predictors are categorical.
- Two Categories: $$x_j = \begin{cases} 0 & \Rightarrow \text{First value} \\ 1 & \Rightarrow \text{Second value} \end{cases}$$
- b_j = difference in the effect of the two alternatives
  b_j = Insignificant $\Rightarrow$ Two alternatives have similar performance
- Alternatively: $$x_j = \begin{cases} -1 & \Rightarrow \text{First value} \\ +1 & \Rightarrow \text{Second value} \end{cases}$$

  b_j = Difference from the average response Difference of the effects of the two levels is $2b_j$

Categorical Predictors (Cont)

- Three Categories: Incorrect:

$$x_1 = \begin{cases} 1 & \Rightarrow \text{Type A} \\ 2 & \Rightarrow \text{Type B} \\ 3 & \Rightarrow \text{Type C} \end{cases}$$

  This coding implies an order ⇒ B is half way between A and C. This may not be true.

- Recommended: Use two predictor variables

$$x_1 = \begin{cases} 1, & \text{If type A} \\ 0, & \text{Otherwise} \end{cases}$$

$$x_2 = \begin{cases} 1, & \text{If type B} \\ 0, & \text{Otherwise} \end{cases}$$

Categorical Predictors (Cont)

Thus, $(x_1, x_2) = (1, 0) \Rightarrow \text{Type A}$

$$(x_1, x_2) = (0, 1) \Rightarrow \text{Type B}$$

$$(x_1, x_2) = (0, 0) \Rightarrow \text{Type C}$$

- This coding does not imply any ordering among the types. Provides an easy way to interpret the regression parameters.

$$y = b_0 + b_1 x_1 + b_2 x_2 + e$$

Categorical Predictors (Cont)

- The average responses for the three types are:

$$\bar{y}_A = b_0 + b_1$$

$$\bar{y}_B = b_0 + b_2$$

$$\bar{y}_C = b_0$$

- Thus, b_1 represents the difference between type A and C.
  b_2 represents the difference between type B and C.
  b_0 represents type C.

Categorical Predictors (Cont)

- Level = Number of values that a categorical variable can take
- To represent a categorical variable with k levels, define k-1 binary variables:

$$x_j = \begin{cases} 1, & \text{If jth value} \\ 0, & \text{otherwise} \end{cases}$$

- *k*th (last) value is defined by $x_1 = x_2 = \cdots = x_{k-1} = 0$.
- b_0 = Average response with the *k*th alternative.
- b_j = Difference between alternatives *j* and *k*.
- If one of the alternatives represents the status quo or a standard against which other alternatives have to be measured, that alternative should be coded as the *k*th alternative.

Case Study 15.1: RPC performance

- RPC performance on Unix and Argus

$$y = b_0 + b_1 x_1 + b_2 x_2$$

where, y is the elapsed time, x_1 is the data size and

$$x_2 = \begin{cases} 1 & \Rightarrow \text{UNIX} \\ 0 & \Rightarrow \text{ARGUS} \end{cases}$$

UNIX		ARGUS	
Data Bytes	Time	Data Bytes	Time
64	26.4	92	32.8
64	26.4	92	34.2
64	26.4	92	32.4
64	26.2	92	34.4
234	33.8	348	41.4
590	41.6	604	51.2
846	50.0	860	76.0
1060	48.4	1074	80.8
1082	49.0	1074	79.8
1088	42.0	1088	58.6
1088	41.8	1088	57.6
1088	41.8	1088	59.8
1088	42.0	1088	57.4

Case Study 15.1 (Cont)

Parameter	Mean	Std. Dev.	Confidence Interval
b_0	36.739	3.251	(31.1676, 42.3104)
b_1	0.025	0.004	(0.0192, 0.0313)
b_2	-14.927	3.165	(-20.3509, -9.5024)

- All three parameters are significant. The regression explains 76.5% of the variation.
- Per byte processing cost (time) for both operating systems is 0.025 millisecond.
- Set up cost is 36.73 milliseconds on ARGUS which is 14.927 milliseconds more than that with UNIX.

Differing Conclusions

- Case Study 14.1 concluded that there was no significant difference in the set up costs. The per byte costs were different.
- Case Study 15.1 concluded that per byte cost is same but the set up costs are different.
- Which conclusion is correct?
  - Need system (domain) knowledge. Statistical techniques applied without understanding the system can lead to a misleading result.
- Case Study 14.1 was based on the assumption that the processing as well as set up in the two operating systems are different $\Rightarrow$ Four parameters
- The data showed that the setup costs were numerically indistinguishable.

Differing Conclusions (Cont)

- The model used in Case Study 15.1 is based on the assumption that the operating systems have no effect on per byte processing.
- This will be true if the processing is identical on the two systems and does not involve the operating systems.
- Only set up requires operating system calls. If this is, in fact, true then the regression coefficients estimated in the joint model of this case study 15.1 are more realistic estimates of the real world.
- On the other hand, if system programmers can show that the processing follows a different code path in the two systems, then the model of Case Study 14.1 would be more realistic.

Curvilinear Regression

- If the relationship between response and predictors is nonlinear but it can be converted into a linear form
  $\Rightarrow$ curvilinear regression.

  Example:

  $$y = bx^a$$

  Taking a logarithm of both sides we get:

  $$\ln y = \ln b + a \ln x$$

  Thus, ln *x* and ln *y* are linearly related. The values of ln b and a can be found by a linear regression of ln *y* on ln *x*.

Curvilinear Regression: Other Examples

Nonlinear	Linear
y = a + b/x	y = a + b (1/x)
y=x/(a+bx)	(1/y)=a+bx
y = x/(a+bx)	(x/y) = a + bx
y = abx	$\ln(y) = \ln(a) + (\ln(b))x$
y=a+bx_n	y=a+b(x^n)

- If a predictor variable appears in more than one transformed predictor variables, the transformed variables are likely to be correlated ⇒ multicollinearity.
- Try various possible subsets of the predictor variables to find a subset that gives significant parameters and explains a high percentage of the observed variation.

Example 15.4

- Amdahl's law: I/O rate is proportional to the processor speed. For each instruction executed there is one bit of I/O on the average.

System No.	MIPS Used	I/O Rate
1	19.63	288.60
2	5.45	117.30
3	2.63	64.60
4	8.24	356.40
5	14.00	373.20
6	9.87	281.10
7	11.27	149.60
8	10.13	120.60
9	1.01	31.10
10	1.26	23.70

Example 15.4 (Cont)

- Let us fit the following curvilinear model to this data:

$$\text{I/O Rate} = \alpha(\text{MIPS Rate})^{b_1}$$

- Taking a log of both sides we get:

$$\log(\text{I/O Rate}) = \log(\alpha) + b_1 \log(\text{MIPS Rate})$$

$$b_0 = \log(\alpha)$$

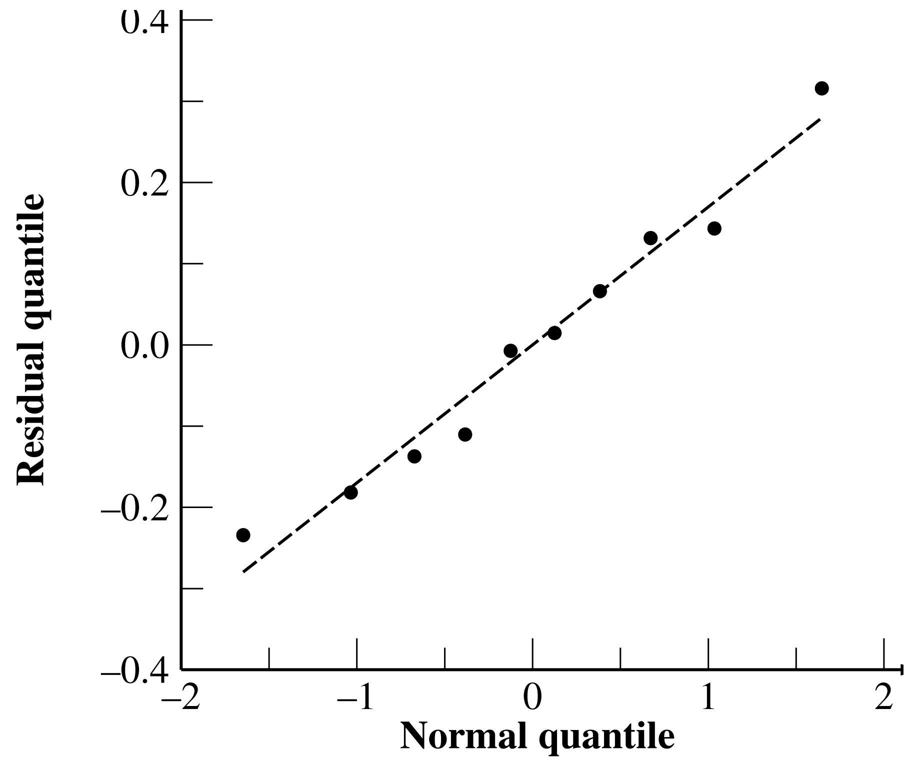
Example 15.4 (Cont)

Obs. No.	x_1	y
1	1.293	2.460
2	0.736	2.069
3	0.420	1.810
4	0.916	2.552
5	1.146	2.572
6	0.994	2.449
7	1.052	2.175
8	1.006	2.081
9	0.004	1.493
10	0.100	1.375

Parameter	Mean	Std. Dev.	Confidence Interval
b_0	1.423	0.119	(1.20, 1.64)
b_1	0.888	0.135	(0.64, 1.14)

- Both coefficients are significant at 90% confidence level.
- The regression explains 84% of the variation.
- At this confidence level, we can accept the hypothesis that the relationship is linear since the confidence interval for b_1 includes 1.

Example 15.4 (Cont)



- Errors in log I/O rate do seem to be normally distributed.

Transformations

- Transformation: Some function of the measured response variable y is used. For example,

$$\sqrt{y} = b_0 + b_1 x_1 + b_2 x_2 + \cdots + b_k x_k + e$$

  Transformation is a subset of the curvilinear regression. However, the ideas apply to non-regression model as well.

1. Physical considerations $\Rightarrow$ Transformation
   For example, if response = inter-arrival times y and it is known that the number of requests per unit time (1/y) has a linear relationship to a predictor
2. If the range of the data covers several orders of magnitude and the sample size is small. That is, if $y_{\max}/y_{\min}$ is large.
3. If the homogeneous variance (**homoscedasticity**) assumption of the residuals is violated.

Transformations (Cont)

- scatter plot shows non-homogeneous spread ⇒ Residuals are still functions of the predictors
- Plot the standard deviation of residuals at each value of $\hat{y}$ as a function of the mean $\hat{y}$.
- If s and the mean $\bar{y}$:

$$s = g(\bar{y})$$

- Then a transformation of the form:

$$w = h(y)$$

$$h(y) = \int \frac{1}{g(y)} dy$$

may help solve the problem

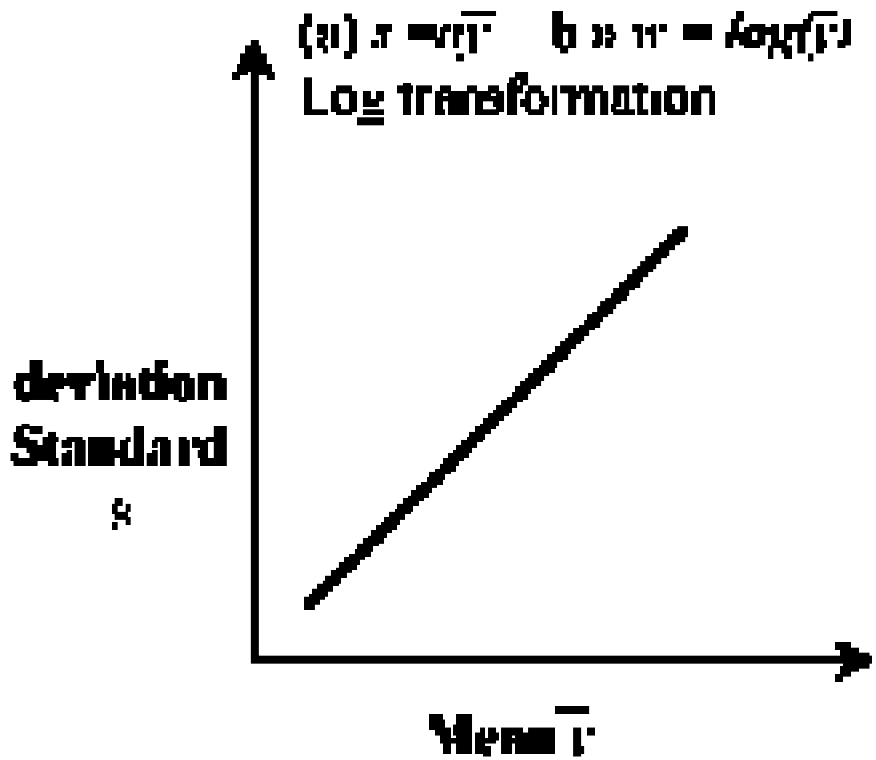
Useful Transformations

- Log Transformation: Standard deviation s is a linear function of the mean (s = a $\bar{y}$)

  w = ln *y*

  and, therefore:

$$h(y) = \int \frac{1}{ay} dy = a \ln y$$



Useful Transformations (Cont)

- Logarithmic transformation is useful only if the ratio $y_{\max}/y_{\min}$ is large.

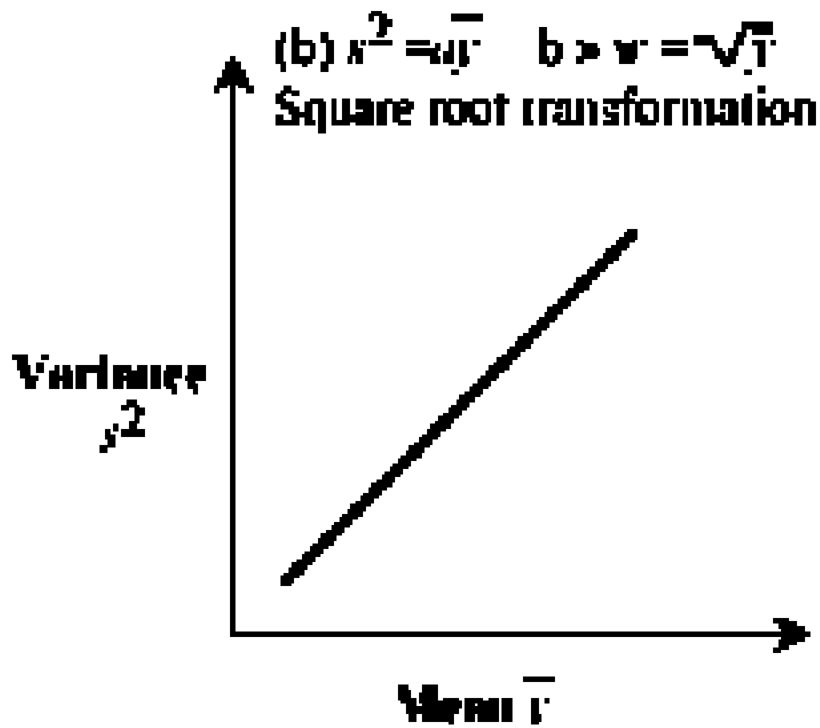
  For a small range the log function is almost linear.

- Square Root Transformation: For a Poisson distributed variable:

  $$s = \sqrt{\bar{y}}$$

  Variance versus mean will be a straight line

  $w = \sqrt{y}$ helps stabilize the variance.



Useful Transformations (Cont)

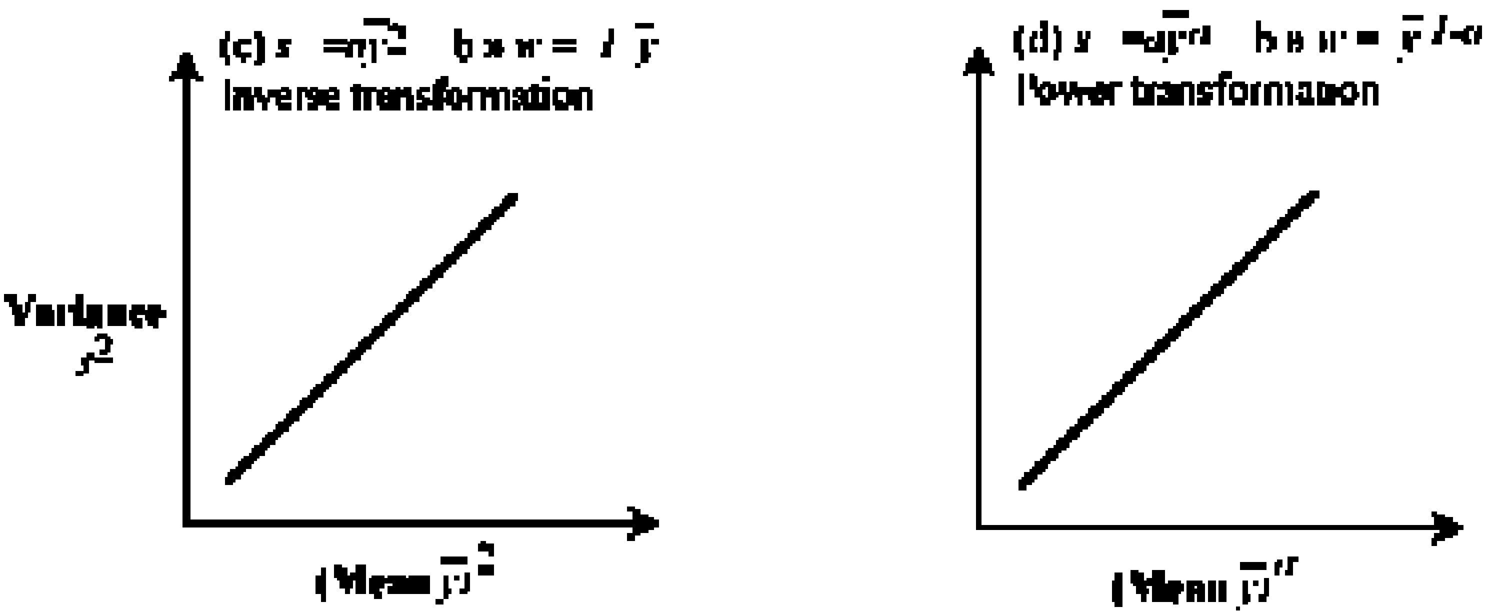
- Arc Sine Transformation: If y is a proportion or percentage, $\sin^{-1}\sqrt{y}$ may be helpful.
- Omega Transformation: This transformation is popularly used when the response y is a proportion.

$$w = 10\log_{10}\left(\frac{y}{1-y}\right)$$

  - The transformed values w's are said to be in units of *deci-Bells*. The term comes from signaling theory where the ratio of output power to input power is measured in dBs.
  - Omega transformation converts fractions between 0 and 1 to values between $-\infty$ to $+\infty$.
  - This transformation is particularly helpful if the fractions are very small or very large.
  - If the fractions are close to 0.5, a transformation may not be required.

Useful Transformations (Cont)

- **Power Transformation**: y^a is regressed on the predictor variables.
  - Standard deviation of residuals s_e is proportional to $\hat{y}^{1-a}$ a=-1 and general a, respectively.



Useful Transformations (Cont)

- **Shifting**: *y+c* (with some suitable c) may be used in place of y.
  - Useful if there are negative or zero values and if the transformation function is not defined for these values.

Relationship between s and $\bar{y}$	Transformation
s $\propto \bar{y}$	w $= \ln(y)$ or w$=\ln(y+c)$
s $\propto \bar{y}^{1/2}$	w$= y^{1/2}$
s $\propto \bar{y}^{a}$	w$= y^{1-a}$ or w$=(y+c)^{1-a}$
s $\propto \bar{y}^{2}$	w$= \frac{1}{y}$
s $\propto$ 1-$\bar{y}^{2}$	w$=\ln\left(\frac{1+y}{1-y}\right)$
s $\propto \bar{y}(1-\bar{y})$	w$=\ln\left(\frac{y}{1-y}\right)$
s $\propto (1+\bar{y})\sqrt{\bar{y}}$	w$=\sin^{-1}\sqrt{y}$

Box-Cox Transformations

- If the value of the exponent a in a power transformation is not known, Box-Cox family of transformations can be used:

$$w = \begin{cases} \frac{y^a - 1}{a g^{a-1}}, & a \neq 0 \\ (\ln y) g, & a = 0 \end{cases}$$

Where g is the geometric mean of the responses:

$$g = (y_1 y_2 \cdots y_n)^{1/n}$$

- The Box-Cox transformation has the property that w has the same units as the response y for all values of the exponent a.
- All real values of a, positive or negative can be tried. The transformation is continuous even at zero, since:

$$\lim_{a \leftarrow 0} \frac{y^a - 1}{a g^{a-1}} = (\ln y) g$$

Box-Cox Transformations (Cont)

- Use a that gives the smallest SSE.
- Use simple values for a. If if a=0.52 is found to give the minimum SSE and the SSE at a=0.5 is not significantly higher, the latter value may be preferable.
- 100(1-α) confidence interval for a:

$$\mathrm{SSE}_{\mathrm{min}} \left(1 + \frac{t^2_{[1-\alpha/2;\nu]}}{\nu}\right)$$

Where, $\mathrm{SSE}_{\mathrm{min}}$ is the minimum SSE, and ν is the number of degrees of freedom for the errors.

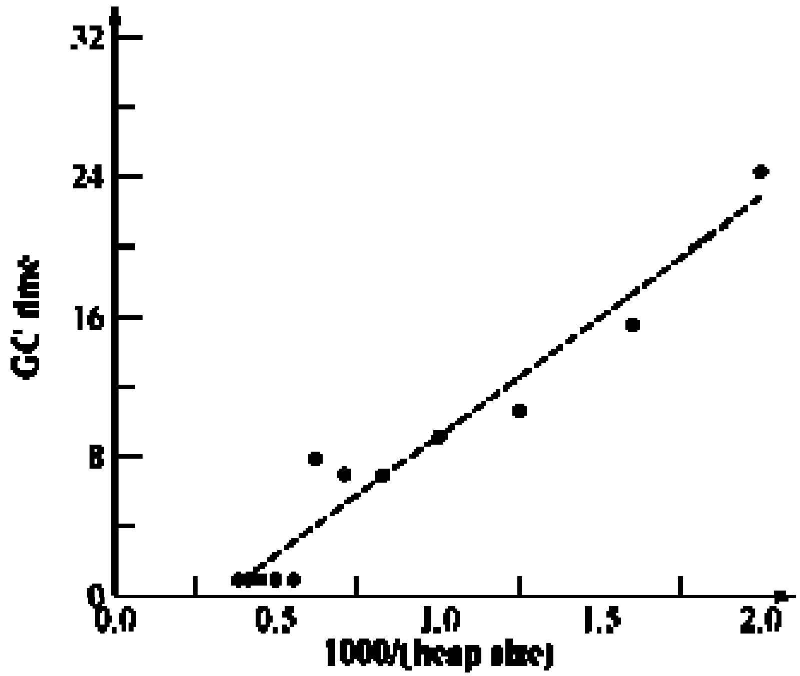
- If the confidence interval for a includes a = 1, then the hypothesis that the relationship is linear cannot be rejected $\Rightarrow$ No need for the transformation.

Case Study 15.2: Garbage collection

- The garbage collection time for various values of heap sizes.

Heap Size	Garbage Collection Time	Heap Size	Garbage Collection Time
500	594.34	1600	63.64
600	247.42	1800	1.00
800	114.24	2000	1.00
1000	85.64	2200	1.00
1200	49.60	2400	1.00
1400	50.30	2600	1.00

Case Study 15.2: Garbage collection

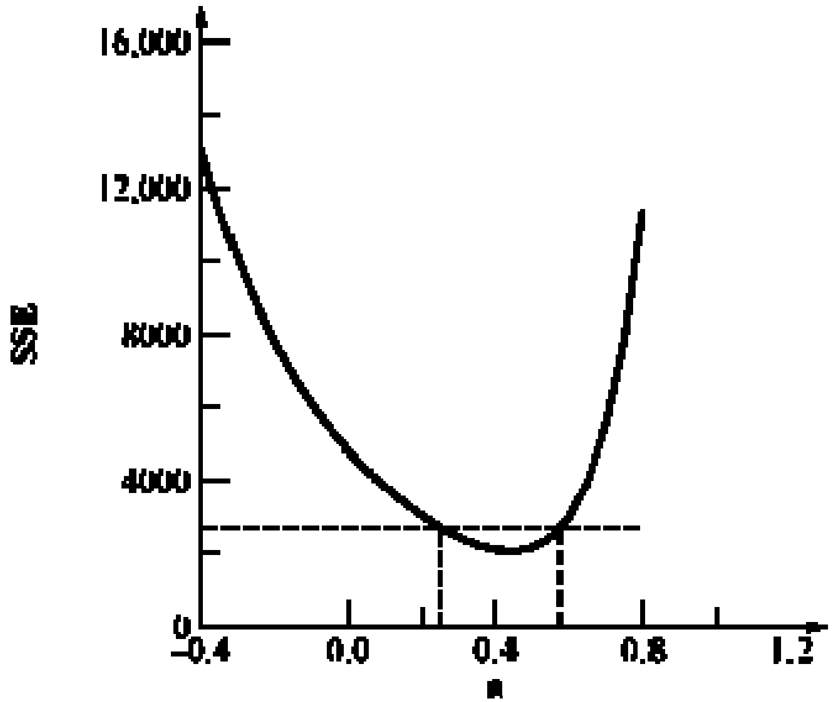


- The points do not appear to be close to the straight line.
- The analyst hypothesizes

$$(\text{Time})^{1/2} = b_0 + \frac{b_1}{\text{Heap Size}}$$

Case Study 15.2 (Cont)

- Is exponent on time is different than a half?
  ⇒ Use Box-Cox transformations with "a" ranging from -0.4 to 0.8



- The minimum SSE of 2049 occurs at a = 0.45.

Case Study 15.2 (Cont)

- Since 0.95-quantile of a t variate with 10 degrees of freedom is 1.812

$$\begin{aligned} SSE &= 2049\left(1+\frac{(1.812)^2}{10}\right) \\ &= 2721.8 \end{aligned}$$

- The SSE = 2271 line intersects the curve at a = 0.2465 and a = 0.5726.
- 90% confidence interval for a is (0.2465, 0.5726). Since the interval includes 0.5, we cannot reject the hypothesis that the exponent is 0.5.

Outliers

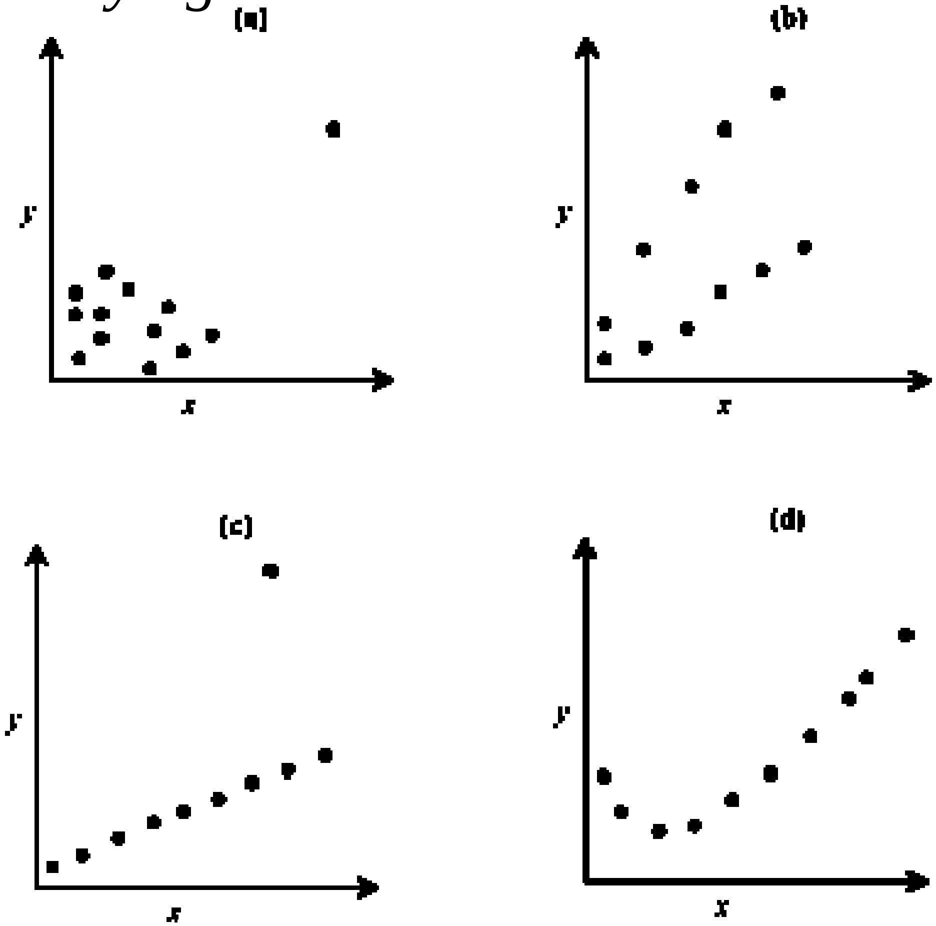
- Any observation that is atypical of the remaining observations *may* be considered an outlier.
- Including the outlier in the analysis may change the conclusions significantly.
- Excluding the outlier from the analysis may lead to a misleading conclusion, if the outlier in fact represents a correct observation of the system behavior.
- A number of statistical tests have been proposed to test if a particular value is an outlier. Most of these tests assume a certain distribution for the observations. If the observations do not satisfy the assumed distribution, the results of the statistical test would be misleading.
- Easiest way to identify outliers is to look at the scatter plot of the data.

Outliers (Cont)

- Any value significantly away from the remaining observations should be investigated for possible experimental errors.
- Other experiments in the neighborhood of the outlying observation may be conducted to verify that the response is typical of the system behavior in that operating region.
- Once the possibility of errors in the experiment has been eliminated, the analyst may decide to include or exclude the suspected outlier based on the intuition.
- One alternative is to repeat the analysis with and without the outlier and state the results separately.
- Another alternative is to divide the operating region into two (or more) sub-regions and obtain a separate model for each sub-region.

Common Mistakes in Regression

1. *Not verifying that the relationship is linear.*
2. *Relying on automated results without visual verification*



• In all these cases, R2 = High

• High R^2 is necessary but not sufficient for a good model.

Common Mistakes in Regression (Cont)

3. *Attaching importance to numerical values of regression parameters.*

CPU time in seconds = 0.01 (Number of disk I/O's) + 0.001 (Memory size in kilobytes)

0.001 is too small ≠> memory size can be ignored

CPU time in milliseconds = 10 (Number of disk I/O's) + 1 (Memory size in kilobytes)

CPU time in seconds = 0.01 (Number of disk I/O's) + 1 (Memory size in Mbytes)

4. *Not specifying confidence intervals for the regression parameters.*
5. *Not specifying the coefficient of determination.*

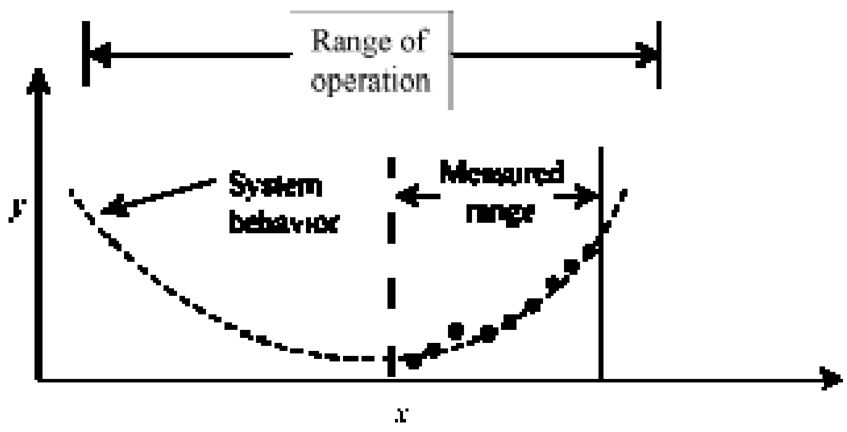
Common Mistakes in Regression (Cont)

6. *Confusing the coefficient of determination and the coefficient of correlation*
   R=Coefficient of correlation, R^2= Coefficient of determination
   $R=0.8$, $R^2=0.64$
   $\Rightarrow$ Regression explains only 64% of variation and not 80%.
7. *Using highly correlated variables as predictor variables*.
   Analysts often start a multi-linear regression with as many predictor variables as possible
   $\Rightarrow$ severe multicollinearity problems.
8. *Using regression to predict far beyond the measured range*.
   Predictions should be specified along with their confidence intervals
9. *Using too many predictor variables*.
   k predictors $\Rightarrow$ 2^k-1 subsets

Common Mistakes in Regression (Cont)

- Subset giving the minimum R^2 is the *best*. But, other subsets that are close may be used instead for practical or engineering reasons. For example, if the second best has only one variable compared to five in the best, the second best may the preferred model.

10. *Measuring only a small subset of the complete range of operation*, e.g., 10 or 20 users on a 100 user system.



Common Mistakes in Regression (Cont)

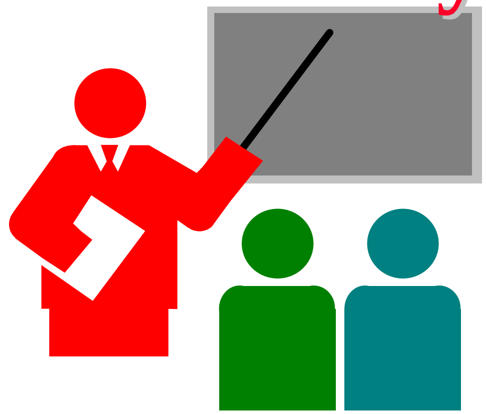
11. *Assuming that a good predictor variable is also a good control variable.*

   - Correlation ⇒ Can predict with a high precision
     ≠> Can control response with predictor
   - For example, the disk I/O versus CPU time regression model can be used to predict the number of disk I/O's for a program given its CPU time.
   - However, reducing the CPU time by installing a faster CPU will not reduce the number of disk I/O's.
   - w and y both controlled by x
     ⇒ w and y highly correlated and would be good predictors for each other.

Common Mistakes in Regression (Cont)

- The prediction works both ways:
  w can be used to predict y and vice versa.
- The control often works only one way:
  x controls y but y may not control x.

Summary



- Too many predictors may make the model weak.
- Categorical predictors are modeled using binary predictors
- Curvilinear regression can be used if a transformation gives linear relationship.
- Transformation: s = g(y) ⇒ $w = h(y) = \int \frac{1}{g(y)} dy$
- Outliers: Use your system knowledge. Check measurements.
- Common mistakes:No visual verification, control vs correlation

Exercise 15.1

- The results of a multiple regression based on 9 observations are shown in the following table.

j	b_j	s_{b_j}
1	1.3	3.6
2	2.7	1.8
3	0.5	0.6
4	5.0	8.3

Intercept $= 75.3$
Coefficient of multiple correlation $= 0.95$
Standard deviation of errors $= 12.0$
F-value $= 14.1$

Exercise 15.1 (Cont)

- Based on these results answer the following questions:
  1. What percent of variance is explained by the regression?
  2. Is the regression significant at 90% confidence level?
  3. Which variable has the highest coefficient?
  4. Which variable is most significant?
  5. Which parameters are not significant at 90%?
  6. What is the problem with this regression?
  7. What would you try next?

Exercise 15.2

- Time to encrypt or decrypt a k-bit record was measured on a uniprocessor as well as on a multi-processor. The times in milliseconds are shown below. Using a log transformation and the method for categorical predictors fit a regression model and interpret the results.

k	Uniprocessor	Multiprocessor
128	93	67
256	478	355
512	3408	2351
1024	25,410	17,022

Homework

- (Updated Exercise 15.2) Time to encrypt or decrypt a k-bit record was measured on a uniprocessor as well as on a multi-processor. The times in milliseconds are shown below. Using a log transformation and the method for categorical predictors fit a regression model and interpret the results.

k	Uniprocessor	Multiprocessor
128	93	67
256	478	355
512	3408	2351
1024	25,410	19,022