Week 3: Prediction and Confidence Intervals at specified x. Testing lack of fit with replicates at some x's. Inference for the correlation. Introduction to regression with several explanatory variables. (ch10) [pp34-60]

- ANOVA table and the F-test

Confidence Interval for the mean response

A level C CI for the mean response μ_y when x takes value x^* is

$$\hat{\mu}_y \pm t^* SE_{\hat{\mu}}$$

where $SE_{\hat{\mu}} = s\sqrt{\frac{1}{n} + \frac{(x^* - \bar{x})^2}{\sum (x_i - \bar{x})^2}}$

and t^* is the value for the $t(n-2)$ density curve with area C between $-t^*$ and t^*.

Regression Analysis: Wages versus LOS

The regression equation is
Wages = 44.2 + 0.0731 LOS

Predictor	Coef	SE Coef	T	P
Constant	44.213	2.628	16.82	0.000
LOS	0.07310	0.03015	2.42	0.018

S = 11.9791 R-Sq = 9.2% R-Sq(adj) = 7.6%

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	1	843.5	843.5	5.88	0.018
Residual Error	58	8322.9	143.5		
Total	59	9166.4			

Unusual Observations

Obs	LOS	Wages	Fit	SE Fit	Residual	St Resid
15	70	97.68	49.33	1.55	48.35	4.07R
22	222	54.95	60.44	4.82	-5.50	-0.50 X
30	150	80.59	55.18	2.85	25.41	2.18R
42	228	67.91	60.88	4.99	7.03	0.65 X
47	204	50.17	59.13	4.31	-8.95	-0.80 X

R denotes an observation with a large standardized residual.
X denotes an observation whose X value gives it large influence.

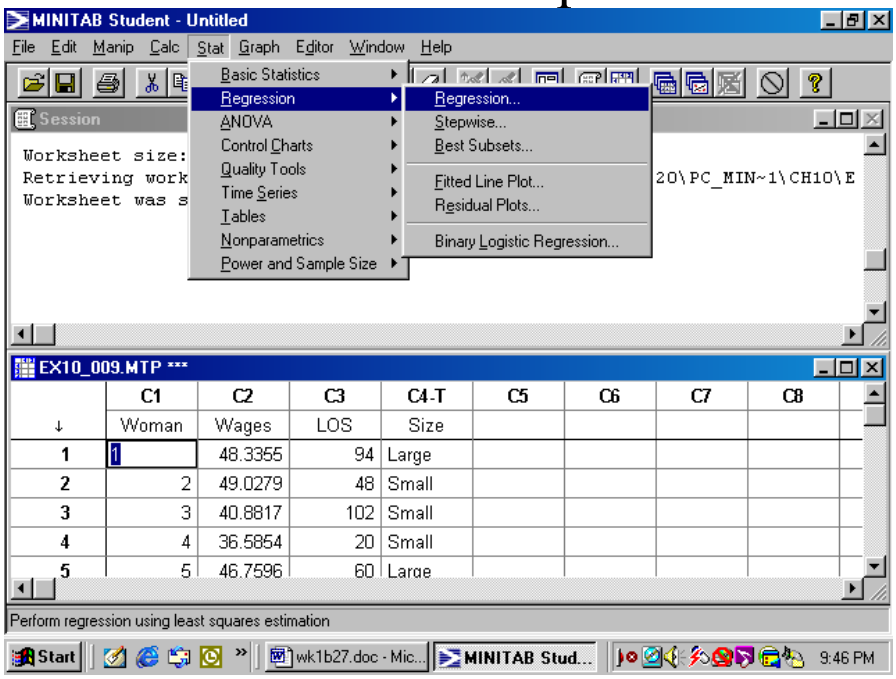
Predicted Values for New Observations

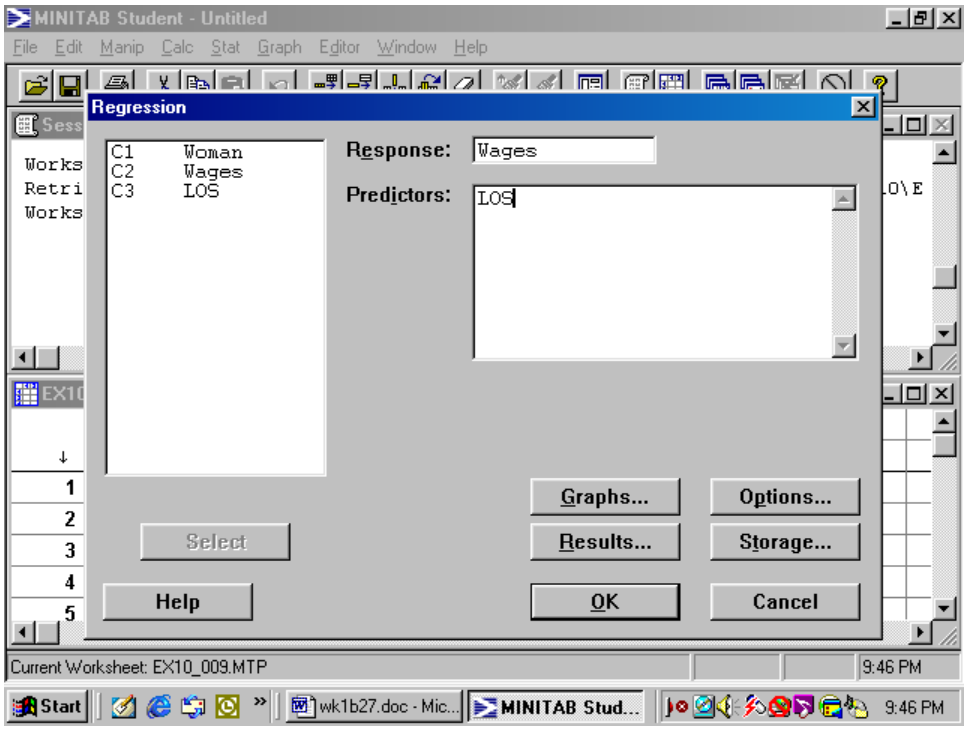
New Obs	Fit	SE Fit	95% CI	95% PI
1	46.84	1.86	(43.11, 50.57)	(22.58, 71.11)

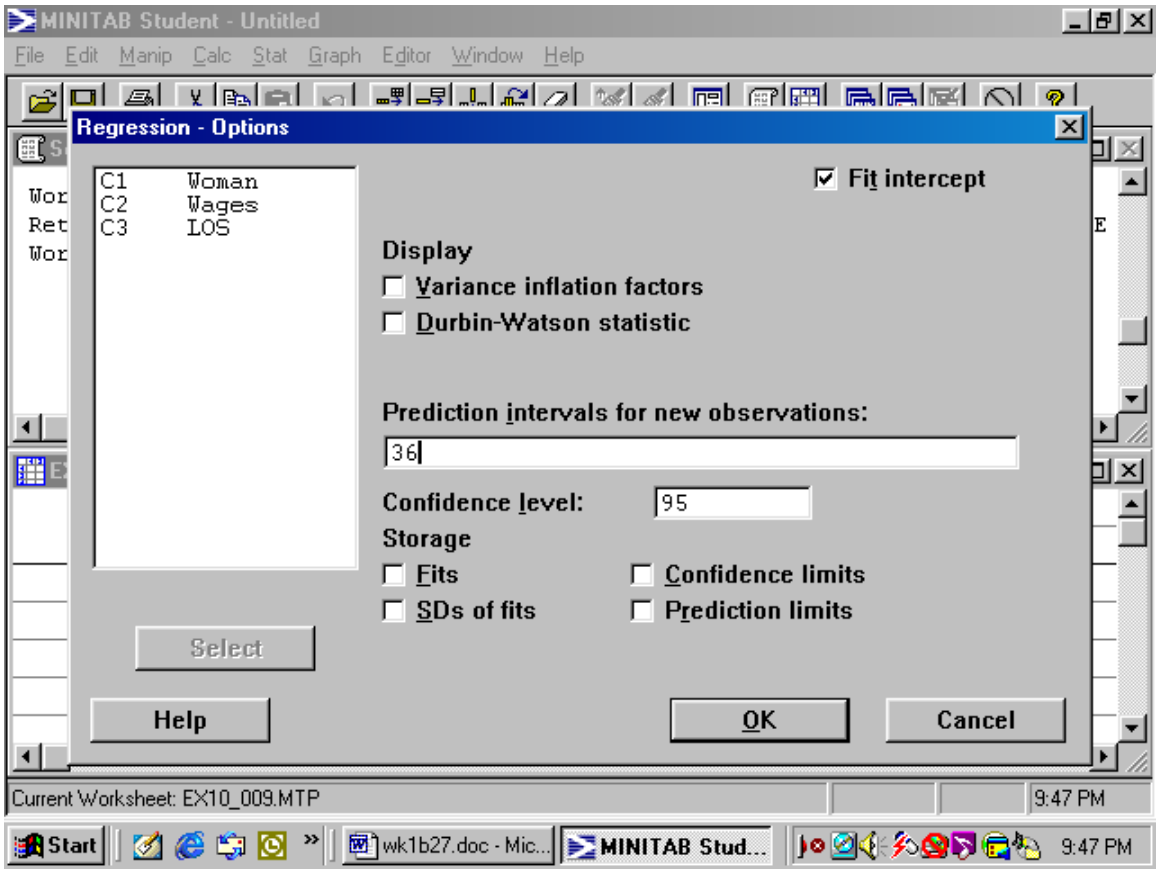
Values of Predictors for New Observations

New Obs	LOS
1	36.0

Minitab commands for prediction






MINITAB Student - Untitled
File Edit Manip Calc Stat Graph Editor Window Help
Regression - Options
C1 Woman
C2 Wages
C3 LOS
Fit intercept
Display
Variance inflation factors
Durbin-Watson statistic
Prediction intervals for new observations:
36
Confidence level:
95
Storage
Fits
Confidence limits
SDs of fits
Prediction limits
Select
Help
OK
Cancel
Current Worksheet: EX10_009.MTP
9:47 PM
Start
wk1b27.doc - Mic...
MINITAB Stud...
9:47 PM

Prediction Interval for a new observation

A level C PI for a new observation with $x = x^*$ is

$$\hat{y} \pm t^* SE_{\hat{y}}$$

where where

$$SE_{\hat{y}} = s\sqrt{1+\frac{1}{n}+\frac{(x^*-\bar{x})^2}{\sum(x_i-\bar{x})^2}}$$ and

t^* is the value for the $t(n-2)$ density curve with area C between $-t^*$ and t^*.

Note $SE_{\hat{y}} = \sqrt{SE_{\hat{\mu}}^2 + MSE}$

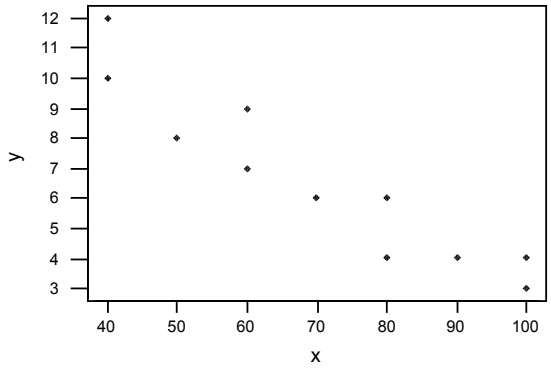
Example: Give a 95% PI for the wage of an employee with 3 years experience. (i.e LOS=36)

Example: Give a 90% PI for the wage of an employee with 3 years experience. (i.e LOS=36)

Lack of Fit Test when there are replicated x-settings

Ex Let x = amount calcium in diet, y = change in blood pressure over specified time period for each of 11 experimental subjects.

Row	x	y
1	40	10
2	40	12
3	50	8
4	60	9
5	60	7
6	70	6
7	80	6
8	80	4
9	90	4
10	100	3
11	100	4



Regression Analysis

```
The regression equation is
y = 15.2 - 0.123 x

Predictor        Coef       StDev          T        P
Constant       15.241       1.113      13.70    0.000
x            -0.12292     0.01523      -8.07    0.000

S = 1.055       R-Sq = 87.9%     R-Sq(adj) = 86.5%

Analysis of Variance

Source            DF          SS          MS         F        P
Regression         1      72.521      72.521     65.11    0.000
Residual Error     9      10.025       1.114
Total             10      82.545
```

Do you think you should try amending your model? (quadratic model? Transformation?)

Testing for the lack of fit of a linear model

Step 1 Calculate Pure Error SS =6.5
and d.f pure error = 4

Step 2 Calculate Lack of fit SS = SSE – P. E SS
=10.025 – 6.5
=3.525
and d.f. LOF = d.f Error – d.f P.E.
= 9 – 4 = 5

Step 3 Calculate MSLOF = SSLOF/d.f LOF
=3.525/5 = 0.705

and MSPE = SSPE./d.f P.E.
= 6.5/4 = 1.625

Step 4 Calculate the test statistic

F = MS LOF/MS P.E = 0.705/1.625
= 0.43

Step 5 Compare this value with the critical value from F(5,4).

Regression Analysis: y versus x

```
The regression equation is
y = 15.2 - 0.123 x


Predictor      Coef  SE Coef      T      P
Constant     15.241    1.113  13.70  0.000
x          -0.12292  0.01523  -8.07  0.000


S = 1.05539   R-Sq = 87.9%   R-Sq(adj) = 86.5%


Analysis of Variance

Source          DF      SS      MS      F      P
Regression       1  72.521  72.521  65.11  0.000
Residual Error   9  10.025   1.114
  Lack of Fit    5   3.525   0.705   0.43  0.808
  Pure Error     4   6.500   1.625
Total           10  82.545


 3 rows with no replicates
```

Multiple regression

The statistical model for multiple linear regression is

$$y_i = \beta_0 + \beta_1 x_{i,1} + \beta_2 x_{i,2} + \cdots \beta_p x_{i,p} + \varepsilon_i$$

for $i = 1,2,\ldots,n$.

The errors ε_i are independent and normally distributed with mean 0 and std dev σ.

-Interpretation of the regression coefficients
-tests and CI's for β's
-ANOVA table
-ANOVA F-test
-R-square = $\frac{SSR}{SST} = 1 - \frac{SSE}{SST}$ and

R-sq(adj)= $1 - \frac{MSE}{MST} = 1 - \left[\frac{(n-1)}{n-(k+1)}\right](1-R^2)$

R-sq(adj) ≤ R-sq
-estimate of $\sigma^2 = MSE$

Example (CS data in data appendix in IPS CD)

Regression Analysis

The regression equation is
gpa = 0.327 +0.000944 satm -0.000408 satv + 0.146 hsm + 0.0359 hss + 0.0553 hse

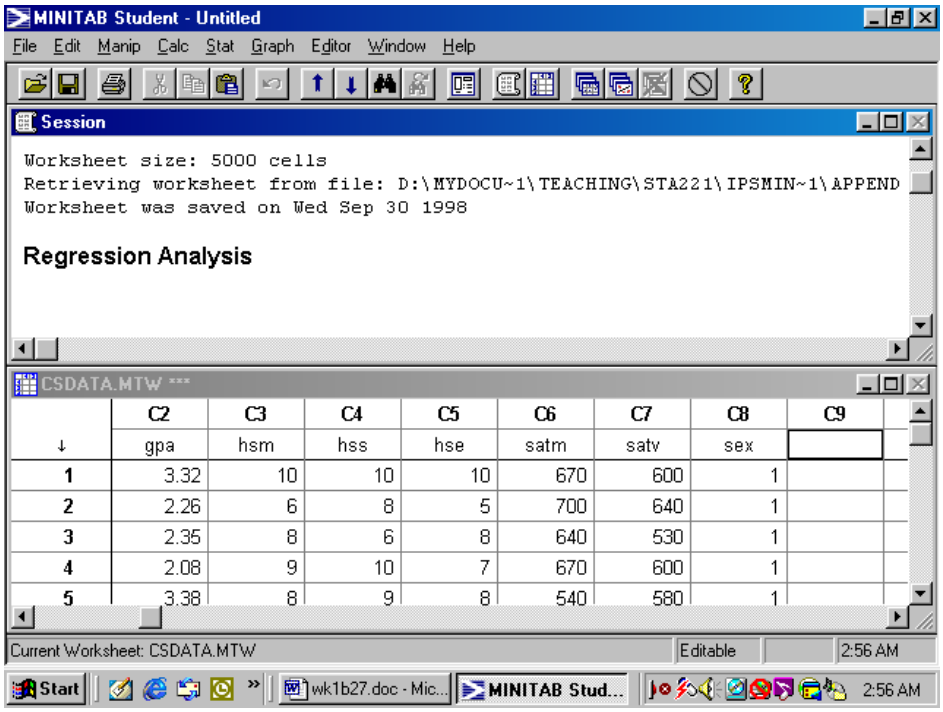
Predictor	Coef	StDev	T	P
Constant	0.3267	0.4000	0.82	0.415
satm	0.0009436	0.0006857	1.38	0.170
satv	-0.0004078	0.0005919	-0.69	0.492
hsm	0.14596	0.03926	3.72	0.000
hss	0.03591	0.03780	0.95	0.343
hse	0.05529	0.03957	1.40	0.164

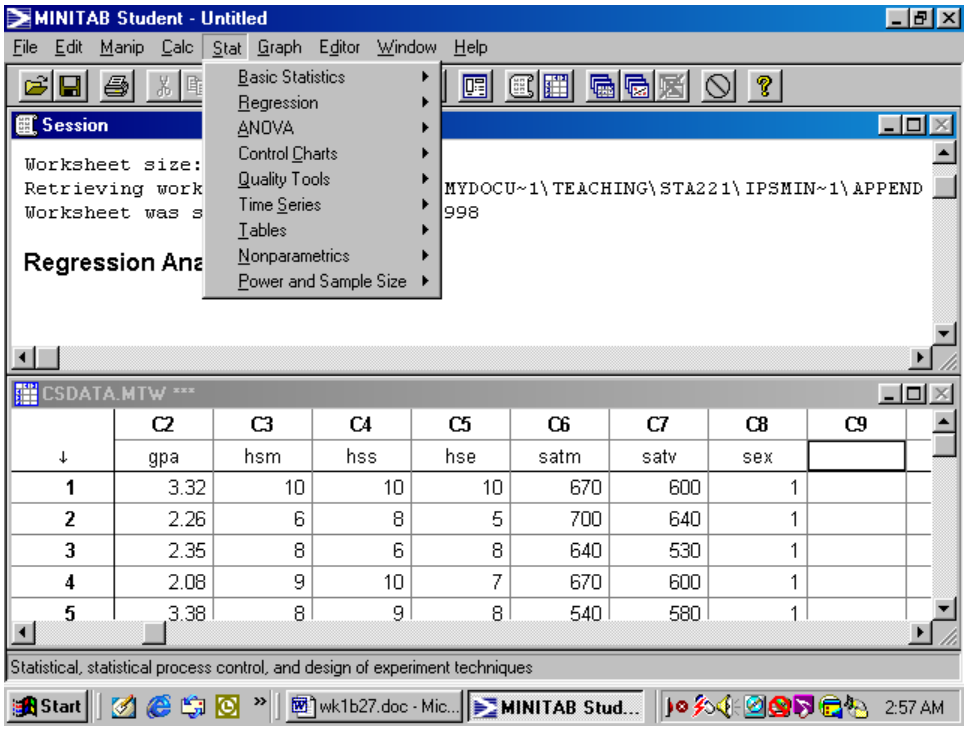
S = 0.7000 R-Sq = 21.1% R-Sq(adj) = 19.3%

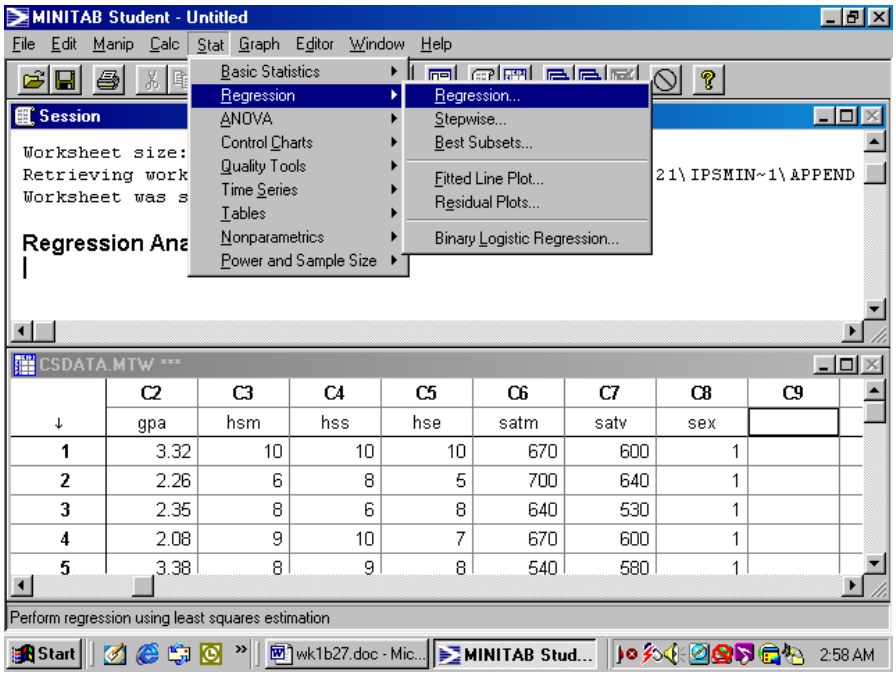
Analysis of Variance

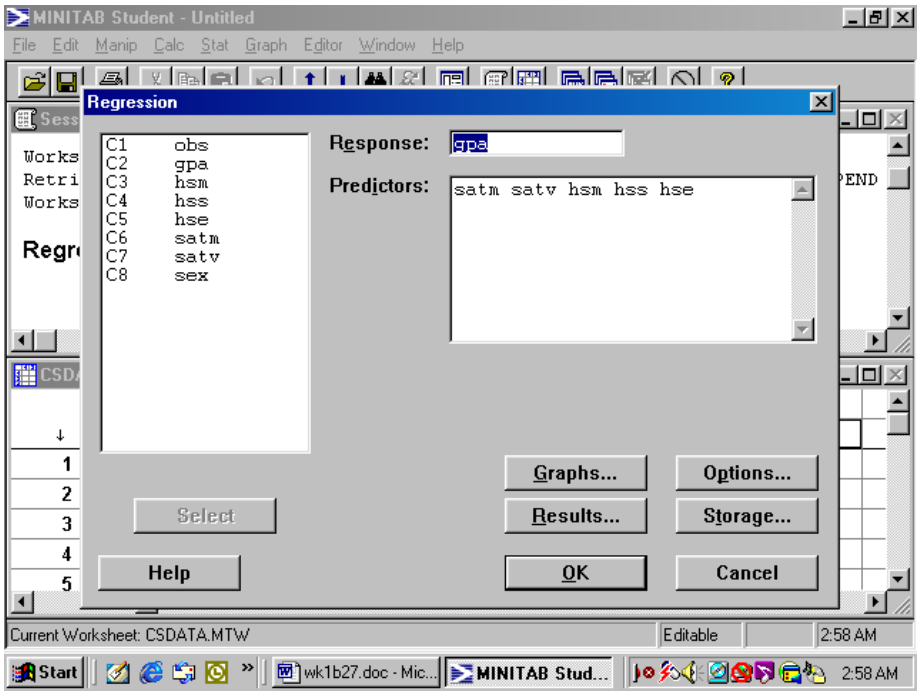
Source	DF	SS	MS	F	P
Regression	5	28.6436	5.7287	11.69	0.000
Residual Error	218	106.8191	0.4900		
Total	223	135.4628			

Minitab commands for multiple regression






MINITAB Student - Untitled
File Edit Manip Calc Stat Graph Editor Window Help
Basic Statistics
Regression
ANOVA
Control Charts
Quality Tools
Time Series
Tables
Nonparametrics
Power and Sample Size
Regression...
Stepwise...
Best Subsets...
Fitted Line Plot...
Residual Plots...
Binary Logistic Regression...
Session
Worksheet size:
Retrieving work
Worksheet was s
21\IPSMIN~1\APPEND
Regression Ana
CSDATA.MTW ***
C2 C3 C4 C5 C6 C7 C8 C9
gpa hsm hss hse satm satv sex
1 3.32 10 10 10 670 600 1
2 2.26 6 8 5 700 640 1
3 2.35 8 6 8 640 530 1
4 2.08 9 10 7 670 600 1
5 3.38 8 9 8 540 580 1
Perform regression using least squares estimation
Start
wk1b27.doc - Mic...
MINITAB Stud...
2:58 AM


MINITAB Student - Untitled
File Edit Manip Calc Stat Graph Editor Window Help
Regression
C1 obs
C2 gpa
C3 hsm
C4 hss
C5 hse
C6 satm
C7 satv
C8 sex
Response: gpa
Predictors: satm satv hsm hss hse
Graphs...
Options...
Results...
Storage...
Select
Help
OK
Cancel
Current Worksheet: CSDATA.MTW
Editable
2:58 AM
Start
wk1b27.doc - Mic...
MINITAB Stud...
2:58 AM

-Prediction

Residual Analysis

- We will use residuals for examining the following six types of departures from the model.

- The regression is nonlinear
- The error terms do not have constant variance
- The error terms are not independent
- The model fits but some outliers
- The error terms are not normally distributed
- One or more important variables have been omitted from the model

Residual plots

Residuals vs X or fitted values

Residuals vs time (when the data are obtained in a time sequence) or other variables

Residuals vs normal scores

Stemplots, boxplots of residuals

Plots of absolute values of the residuals (or squared residuals) against X or against

Fitted values are also helpful in diagnosing nonconstancy of error variance.

Example. Residual analysis for the above example