APS 425
Fall 2015

Simple Linear Regression: Prediction

Instructor: G. William Schwert
585-275-2470
schwert@schwert.ssb.rochester.edu

Ciba-Geigy Ritalin Experiment

- Ritalin is tested to see if it helps with Central Auditory Processing Disorder (CAPD)
  - Similar symptoms to ADD/ADHD
- Experiment:
  - "Randomly" select 64 children
  - All receive auditory test
  - 32 (control group) receive no drug (or placebo?)
  - 32 (treatment group) receive varying doses of Ritalin
  - All children are tested a second time

Ciba-Geigy Ritalin Experiment

- $DOSAGE_i$ = amount of Ritalin received by child i
  - Measured as Mg of Ritalin per Kg of body weight
- $IMPROVE_i$ = child's 2^{nd} test score – 1^{st} test score
  - Dataset A425_ritalin.wf1 also contains:
    - AGE of child in months
    - Gender (FEMALE = 1, for girls)

Predictions

Dependent Variable: IMPROVE
Method: Least Squares
Sample: 1 64
Included observations: 64

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	0.225872	2.097651	0.107678	0.9146
DOSAGE	12.17855	5.723027	2.127991	0.0373

R-squared	0.068066	Mean dependent var	3.359375
Adjusted R-squared	0.053035	S.D. dependent var	12.28157
S.E. of regression	11.95146	Akaike info criterion	7.830335
Sum squared resid	8855.917	Schwarz criterion	7.897800
Log likelihood	-248.5707	Hannan-Quinn criter.	7.856913
F-statistic	4.528347	Durbin-Watson stat	1.975687
Prob(F-statistic)	0.037321		

Predictions

- Predictive model:

$IMPROVE_i = 0.226 + 12.18\ DOSAGE_i$
$= \text{the estimate of } E[IMPROVE_i \mid DOSAGE_i, b_0, b_1]$

- What is your estimate of the <u>average</u> *IMPROVE* score for all children who receive a dosage of 0.35 mg/kg?

$IMPROVE_i = 0.226 + 12.18\ DOSAGE_i = 4.488$

- This question asks about the <u>average or expected value for all children</u> who get a *DOSAGE* of 0.35mg/kg.

Predictions

- A given child has been administered a *DOSAGE* of 0.35mg/kg. What <u>value</u> do you predict for the child's *IMPROVE* score?

$\hat{IMPROVE}_i = 0.226 + 12.18 \times 0.35 = 4.488$

- This question asks you to predict the <u>value for an individual child</u> who gets a *DOSAGE* of 0.35mg/kg

Predictions

- The prediction of the value for an individual child = the expected value for the population (of all children with *DOSAGE* = 0.35mg/kg) (see previous two slides)
- However, standard errors are different!
- Let's derive them next

Std Error of Predictions

- Standard error for predicting an individual value
  - The linear model is:

$$Y_i = \beta_0 + \beta_1 X_i + e_i$$

  - Our prediction is:

$$\hat{Y}_i = b_0 + b_1 X_i$$

  - Three sources of error:
    - Error in estimating β_0
    - Error in estimating β_1
    - Error in estimating e_i

Std Error of Predictions

- Standard error of Y_i:

$$SDP = [\, s^2 + s^2 / n + s^2 (X_i - \bar{X})^2 / \Sigma x_i^2 \,]^{1/2}$$

Error term uncertainty — Intercept uncertainty — Slope uncertainty

where $s^2 = \Sigma \hat{e}_i^2 / (n-2)$ is the residual variance
and $s^2 / \Sigma x_i^2$ is the variance of b_1

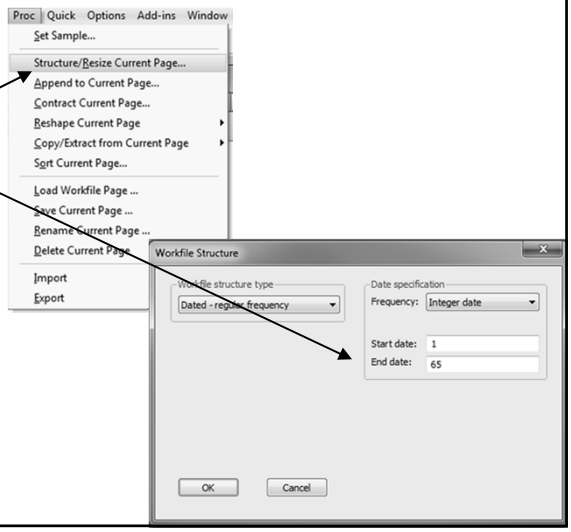
Prediction Intervals

- A $100(1-\alpha)\%$ confidence interval for Y_i is:
  $[Y_i - t_{\alpha/2}\, SDP,\ Y_i + t_{\alpha/2}\, SDP]$
  where $\Pr\{t_{n-2} > t_{\alpha/2}\} = \alpha/2$

- In the Ritalin case with one child getting a *DOSAGE* of 0.35mg/kg, we have SDP = 12.056, so a 95% confidence interval for $IMPROVE_i$ is
  $[4.488 - 2.00(12.056),\ 4.488 + 2.00(12.056)]$
  $= [-19.61,\ 28.59]$

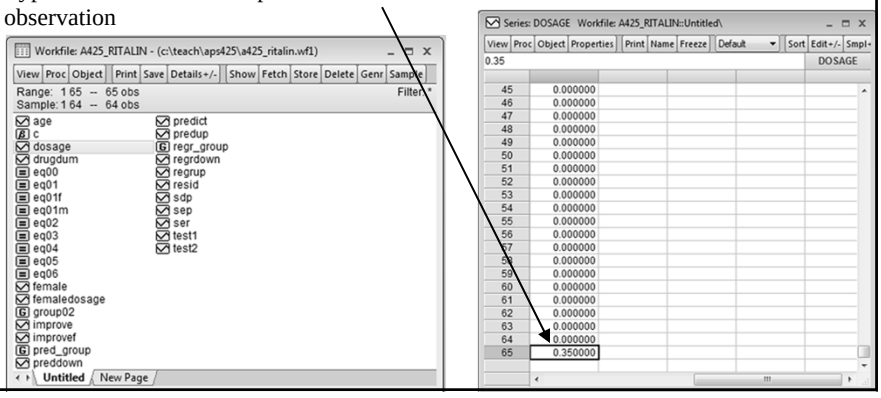
Using Eviews to Get Prediction Intervals

- Redefine the workfile range so that you can generate an "out-of-sample" prediction

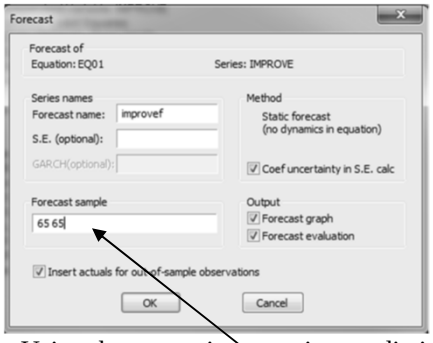


Using Eviews to Get Prediction Intervals

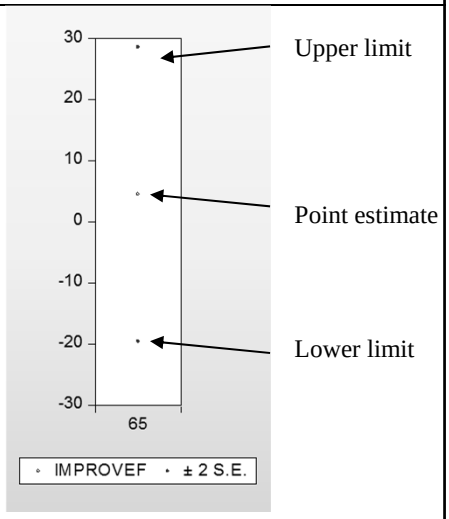
- Double-click dosage to open the spreadsheet
- Then click Edit+/-
- Move to observation 65 (which will say "NA")
- Type in .35 into the workspace bar to enter this value for the 65th observation



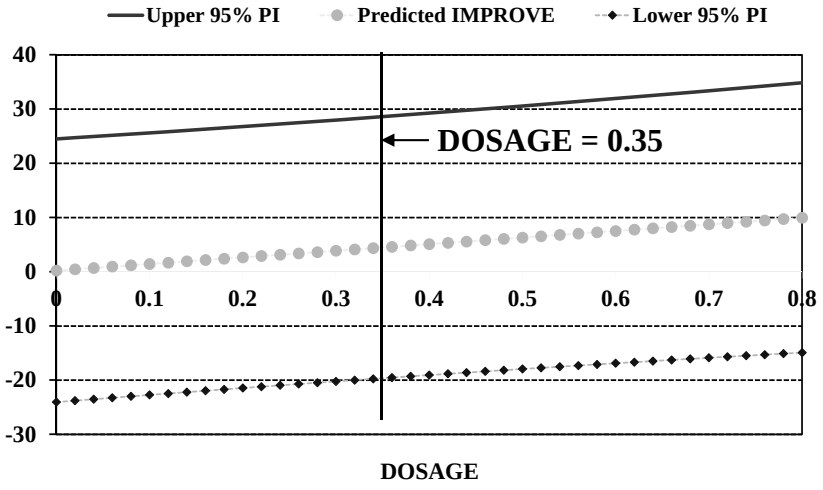
Using Eviews to Get Prediction Intervals



- Using the regression equation predicting IMPROVE as a function of DOSAGE, click FORECAST
- Then specify the forecast sample as 65 65
  - the value of DOSAGE = 0.35 you just entered



Prediction Intervals for Improvement from DOSAGE



Std Error of Prediction for the Average

- Standard error for predicting the expected or average value
  - A group of children have been administered a *DOSAGE* of 0.35mg/kg. What value do you predict for the average *IMPROVE* score of these children on the test?

$$\hat{IMPROVE}_i = 0.226 + 12.18 \times 0.35 = 4.488$$

Std Error of Prediction for the Average

- Now let's consider the standard error
  - The linear model is: $Y_i = \beta_0 + \beta_1 X_i + e_i$
  - We are predicting: $E[Y_0] = \beta_0 + \beta_1 X_0$
  - Our prediction is: $Y_i = b_0 + b_1 X_i$
  - Two sources of uncertainty:
    - Error in estimating β_0
    - Error in estimating β_1

Std Error of Prediction for the Average

- Standard error of E(Y_i):

$$SEP = [s^2 / n + s^2 (X_i - \bar{X})^2 / \Sigma x_i^2]^{1/2}$$

Intercept uncertainty

Slope uncertainty

- A 100(1–α)% confidence interval for E(Y_i) is:

$[Y_i - t_{\alpha/2} SEP, Y_i + t_{\alpha/2} SEP]$

where $Pr\{t_{n-2} > t_{\alpha/2}\} = \alpha/2$

Std Error of Prediction for the Average

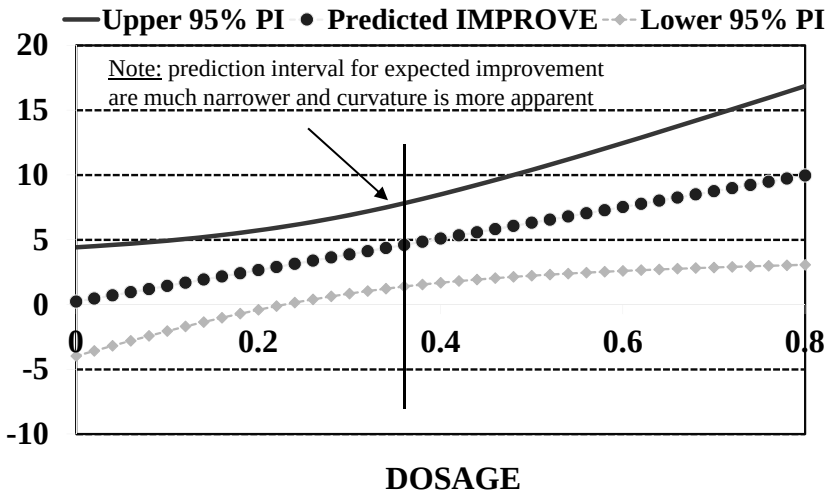
- In the Ritalin case with one child getting a *DOSAGE* of 0.35mg/kg, we have SEP = 1.585, so a 95% confidence interval for

E(*IMPROVE*$_i$ | *DOSAGE* = 0.35, b_0, b_1) is

[4.488 – 2.00(1.585), 4.488 + 2.00(1.585)]

= [1.319, 7.657]

Prediction Intervals for Expected Improvement from DOSAGE



Predictions and Eviews for DOSAGE = 0.35

$(SEP)^2 = s^2 / n + (SE(b_1))^2 (X_i - \bar{X})^2$

$= 11.951^2/ 64 + (5.723)^2 (0.35 - 0.257)^2$

$= 2.513 \Rightarrow SEP = 1.585$

$(SEP)^2 + (\text{SE of Regression})^2 = (SDP)^2$

$[1.585^2+11.951^2]^{1/2} = 12.056$

Note that SEP and SDP depend on $X_i = 0.35$

Dependent Variable: IMPROVE
Method: Least Squares
Sample: 1 64
Included observations: 64

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	0.225872	2.097651	0.107678	0.9146
DOSAGE	12.17855	5.723027	2.127991	0.0373

R-squared	0.068066	Mean dependent var	3.359375
Adjusted R-squared	0.053035	S.D. dependent var	12.28157
S.E. of regression	11.95146	Akaike info criterion	7.830335
Sum squared resid	8855.917	Schwarz criterion	7.897800
Log likelihood	-248.5707	Hannan-Quinn criter.	7.856913
F-statistic	4.528347	Durbin-Watson stat	1.975687
Prob(F-statistic)	0.037321		

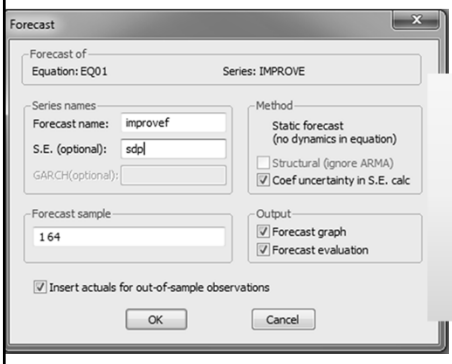
Predictions and Eviews

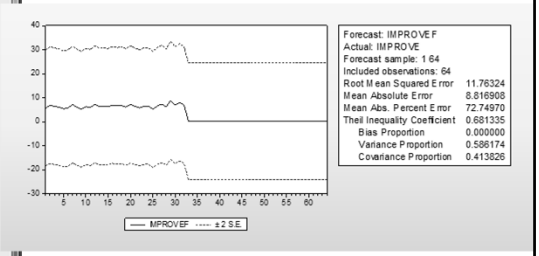
- The range of *DOSAGE* values in the data is 0 to 0.71
- Given this sample, the predictive model is:
  $\hat{IMPROVE}_i = 0.226 + 12.18 \times 0.35 = 4.488$
- We have no support from the data whether this relation extends outside of the sample range (e.g., to dosages > 0.71 mg/kg)
- To predict outside the sample range is called ***extrapolation***
- Extrapolation is ill-advised and subject to much criticism

	DOSAGE	IMPROVE
Mean	0.257297	3.359375
Median	0.203500	0.000000
Maximum	0.710000	50.00000
Minimum	0.000000	-15.00000
Std. Dev.	0.263102	12.28157
Skewness	0.089623	1.284004
Kurtosis	1.138520	5.596263
Jarque-Bera	9.325967	35.56066
Probability	0.009438	0.000000
Sum	16.46700	215.0000
Sum Sq. Dev.	4.361041	9502.734
Observations	64	64

Predictions of IMPROVE from Eviews

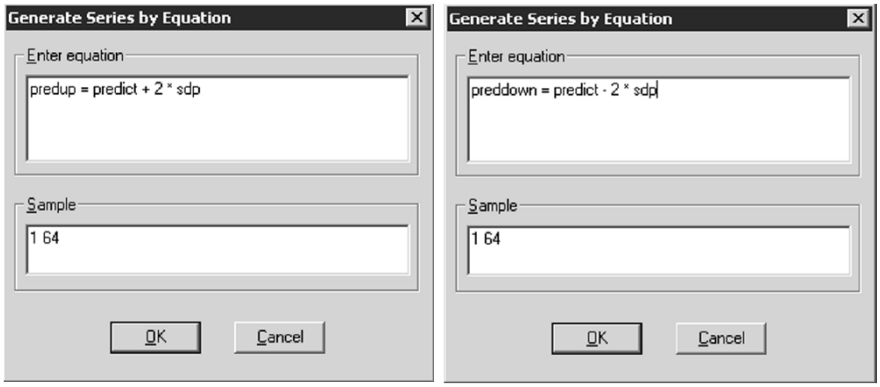
Generate predicted values, "predict", and the standard deviation of the prediction, "sdp"





Prediction Interval for IMPROVE from Eviews

Generate upper and lower limits for 95% prediction interval, "predup" and "preddown"

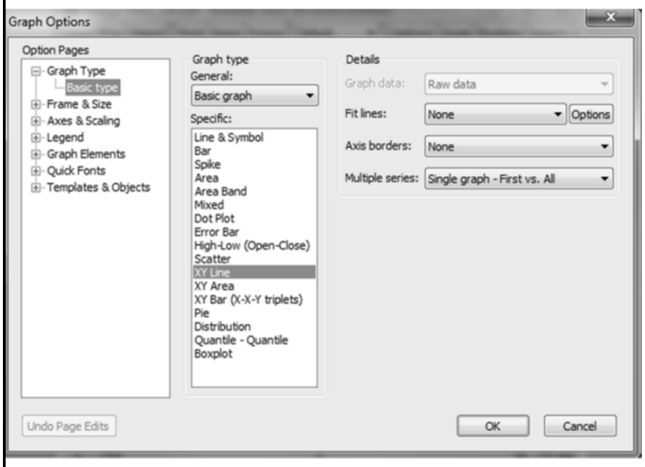


Prediction Interval for IMPROVE from Eviews

Create a "group" of "dosage", "predict", "predup", and "preddown"

obs	DOSAGE	PREDDOWN	PREDICT	PREDUP
1	0.452000	-18.46123	5.730577	29.92238
2	0.550000	-17.39672	6.924076	31.24488
3	0.508000	-17.84667	6.412576	30.67183
4	0.478000	-18.17381	6.047219	30.26825
5	0.423000	-18.78609	5.377399	29.54089
6	0.452000	-18.46123	5.730577	29.92238
7	0.569000	-17.19624	7.155468	31.50718
8	0.481000	-18.14088	6.083755	30.30839
9	0.407000	-18.96726	5.182542	29.33234
10	0.496000	-17.97695	6.266433	30.50982
11	0.463000	-18.33919	5.864541	30.06827
12	0.571000	-17.17525	7.179825	31.53490
13	0.508000	-17.84667	6.412576	30.67183
14	0.506000	-17.86833	6.388219	30.64477
15	0.496000	-17.97695	6.266433	30.50982
16	0.550000	-17.39672	6.924076	31.24488
17	0.524000	-17.67416	6.607433	30.88903
18	0.537000	-17.53500	6.765754	31.06650

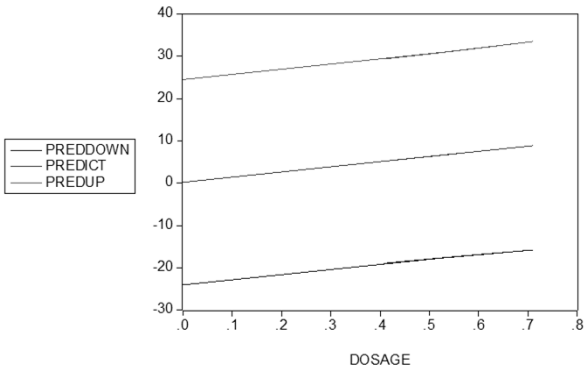
Prediction Interval for IMPROVE from Eviews



Graph XY line
One X against all Y's

Make sure that the predictor variable, DOSAGE, is the first one in the set

Prediction Interval for IMPROVE from Eviews

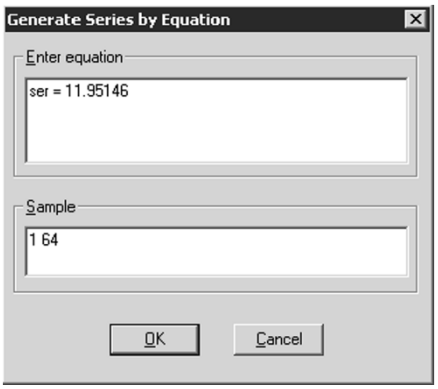


Note that the 95% prediction interval for IMPROVE covers a wide range of outcomes for individual students

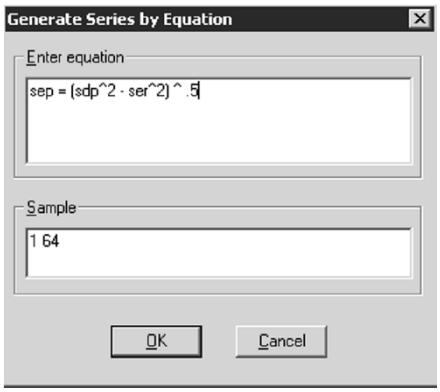
No assurance that any one child will improve if given Ritalin

Standard Error of the Regression Line for IMPROVE from Eviews

We will calculate the standard error of the regression line (SEP)



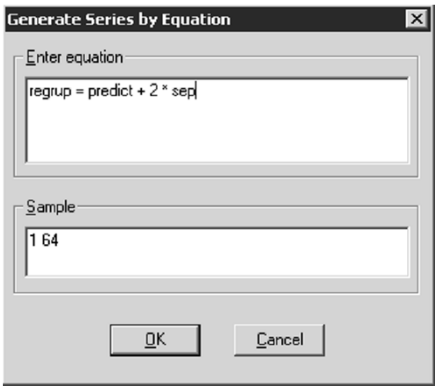
First, save the standard error of the regression, SER

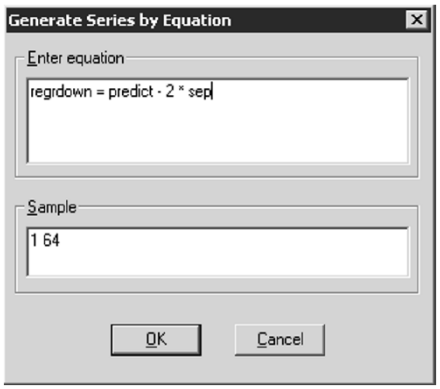


Next, derive SEP from SDP and SER

Standard Errors Around the Regression Line for IMPROVE

Generate upper and lower limits for 95% regression line, "regrup" and "regrdown"





Standard Errors Around the Regression Line for IMPROVE

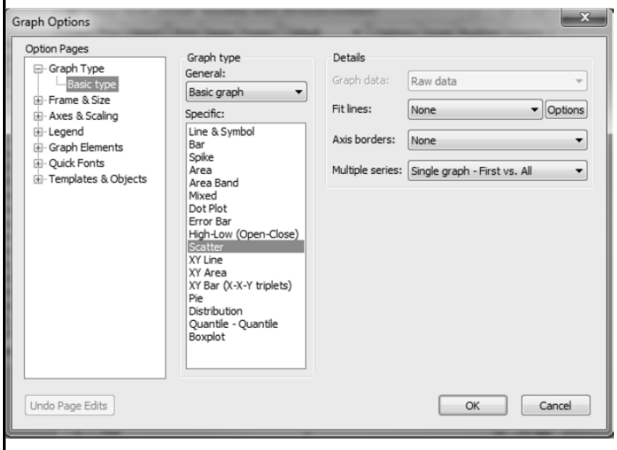
Create a "group" of "dosage", "predict", "regrup", and "regrdown"

Range: 1 64 Filter: *
Sample: 1 64

- age
- c
- dosage
- drugdum
- eq00
- eq01
- eq02
- eq03
- eq04
- female
- group02
- improve
- pred_group
- preddown
- predict
- predup
- regrdown
- regrup
- resid
- sdp
- sep
- ser

obs	DOSAGE	PREDICT	REGRDOWN	REGRUP
1	0.452000	5.730577	2.003135	9.458020
2	0.550000	6.924076	2.435007	11.41314
3	0.508000	6.412576	2.269916	10.55524
4	0.478000	6.047219	2.134569	9.959870
5	0.423000	5.377399	1.838400	8.916399
6	0.452000	5.730577	2.003135	9.458020
7	0.569000	7.155468	2.501842	11.80909
8	0.481000	6.083755	2.148847	10.01866
9	0.407000	5.182542	1.738219	8.626866
10	0.496000	6.266433	2.217704	10.31516
11	0.463000	5.864541	2.060482	9.668601
12	0.571000	7.179825	2.508625	11.85102
13	0.508000	6.412576	2.269916	10.55524
14	0.506000	6.388219	2.261383	10.51505
15	0.496000	6.266433	2.217704	10.31516
16	0.550000	6.924076	2.435007	11.41314

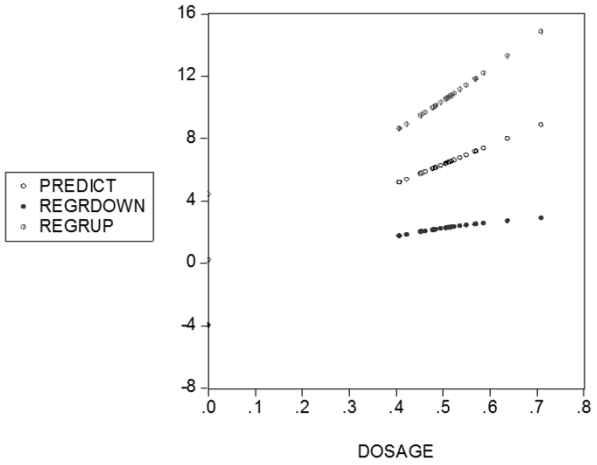
Standard Errors Around the Regression Line for IMPROVE



Graph simple scatter

Make sure that the predictor variable, DOSAGE, is the first one in the set

Standard Errors Around the Regression Line for IMPROVE



Note that the 95% confidence interval for the regression line is much narrower

For the dosages used in this experiment the entire interval covers positive improvement

Links

Ritalin Data
http://schwert.ssb.rochester.edu/a425/a425_ritalin.wf1

Return to APS 425 Home Page
http://schwert.ssb.rochester.edu/a425/a425main.htm