UCLA STAT 251

Statistical Methods for the Life and Health Sciences

•**Instructor**: **Ivo Dinov,**

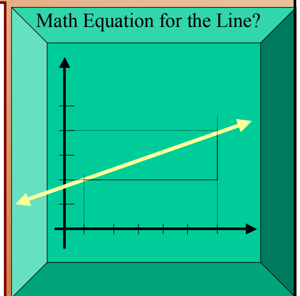
Asst. Prof. In Statistics and Neurology

University of California, Los Angeles, Winter 2002

http://www.stat.ucla.edu/~dinov/

Lines in 2D
(Regression and Correlation)

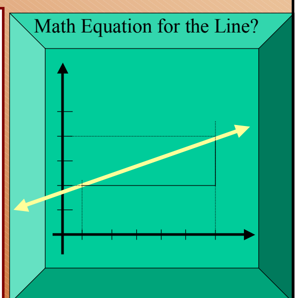
- Vertical Lines
- Horizontal Lines
- Oblique lines
- Increasing/Decreasing
- Slope of a line
- Intercept
- $Y=\alpha X + \beta$, in general.



Lines in 2D
(Regression and Correlation)

- Draw the following lines:
- Y=2X+1
- Y=-3X-5
- Line through (X_1,Y_1) and (X_2,Y_2).
- $(Y-Y_1)/(Y_2-Y_1)=$

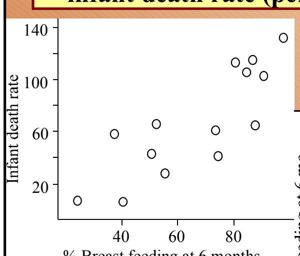
  $(X-X_1)/(X_2-X_1)$.



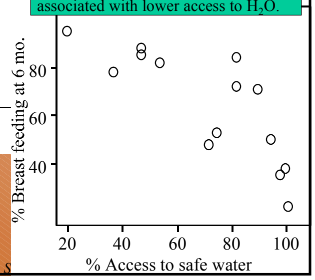
Approaches for modeling data relationships
Regression and Correlation

- There are random and nonrandom variables
- Correlation applies if both variables (X/Y) are random (e.g., We saw a previous example, systolic vs. diastolic blood pressure SISVOL/DIAVOL) and are treated symmetrically.
- Regression applies in the case when you want to single out one of the variables (response variable, Y) and use the other variable as predictor (explanatory variable, X), which explains the behavior of the response variable, Y.

Causal relationship?
– infant death rate (per 1,000) in 14 countries

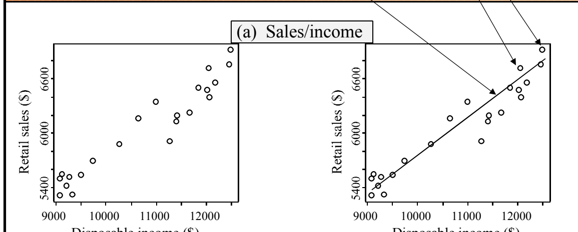


Strong evidence (linear pattern) of death rate increase with increasing level of breastfeeding (BF)? Naïve conclusion breast feeding is bad? But high rates of BF is associated with lower access to H_2O.



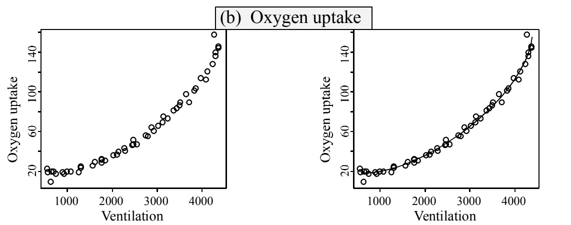
Predict behavior of Y (response) Based on the values of X (explanatory var.) Strategies for uncovering the reasons (causes) for an observed effect.

Regression relationship = trend + residual scatter



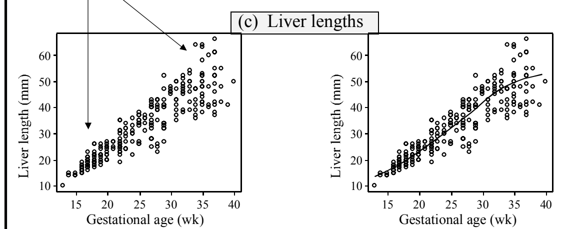
- Regression is a way of studying relationships between variables (random/nonrandom) for predicting or explaining behavior of 1 variable (response) in terms of others (explanatory variables or predictors).

Trend (does not have to be linear) + scatter (could be of any type/distribution)



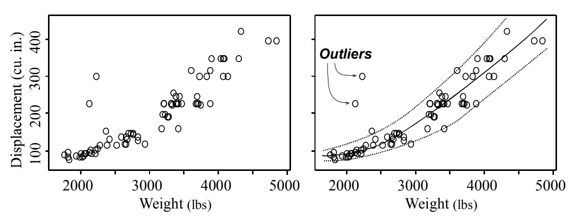
Trend + scatter (fetus liver length in *mm*)

Change of scatter with age



Trend + scatter

Dotted curves (confidence intervals) represent the extend of the scatter.

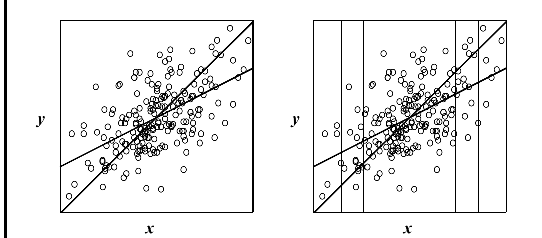


(a) Scatter plot (b) With trend plus scatter

Displacement versus weight for 74 models of automobile.

Looking vertically

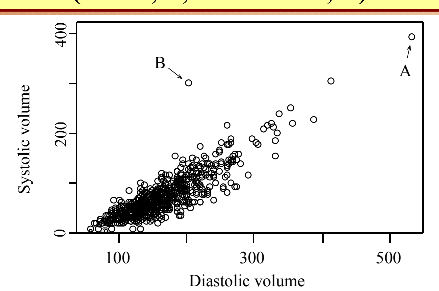
Flatter line gives better prediction, since it approx. goes through the



(a) Which line? (b) Flatter line gives better predictions.

Educating the eye to look vertically.

Outliers – odd, atypical, observations (errors, B, or real data, A)



Scatter plot from the heart attack data.

A note of caution!

In observational data, strong relationships are ***not*** necessarily causal. It is virtually **impossible** to conclude a cause-and-effect relationship between variables using observational data!

Essential Points

1. What essential difference is there between the correlation and regression approaches to a relationship between two variables? (In correlation independent variables; regression response var depends on explanatory variable.)
2. What are the most common reasons why people fit regression models to data? (predict Y or unravel reasons/causes of behavior.)
3. Can you conclude that changes in X caused the changes in Y seen in a scatter plot if you have data from an observational study? (No, there could be lurking variables, hidden effects/predictors, also associated with the predictor X, itself, e.g., time is often a lurking variable, or may be that changes in Y cause changes in X, instead of the other way around).

Essential Points

5. When can you reliably conclude that changes in X cause the changes in Y? (Only when controlled randomized experiments are used – levels of X are randomly distributed to available experimental units, or experimental conditions need to be identical for different levels of X, this includes time.

Correlation Coefficient

Correlation coefficient ($-1 \le R \le 1$): a measure of linear association, or clustering around a line of multivariate data.

Relationship between two variables (X, Y) can be summarized by: (μ_X, σ_X), (μ_Y, σ_Y) and the correlation coefficient, R. $R=1$, perfect positive correlation (straight line relationship), $R=0$, no correlation (random cloud scatter), $R=-1$, perfect negative correlation.

Computing $R(X,Y)$: (standardize, multiply, average)

$$R(X,Y) = \frac{1}{N-1}\sum_{k=1}^{N}\left(\frac{x_k-\mu_x}{\sigma_x}\right)\left(\frac{y_k-\mu_y}{\sigma_y}\right)$$

$X=\{x_1, x_2, \ldots, x_N\}$
$Y=\{y_1, y_2, \ldots, y_N\}$
(μ_X, σ_X), (μ_Y, σ_Y)
sample mean / SD.

Correlation Coefficient

Example:

$$R(X,Y) = \frac{1}{N-1}\sum_{k=1}^{N}\left(\frac{x_k-\mu_x}{\sigma_x}\right)\left(\frac{y_k-\mu_y}{\sigma_y}\right)$$

Student i	Height x_i	Weight y_i	$x_i-\bar{x}$	$y_i-\bar{y}$	$(x_i-\bar{x})^2$	$(y_i-\bar{y})^2$	$(x_i-\bar{x})(y_i-\bar{y})$
1	167	60	6	4.67	36	21.8089	28.02
2	170	64	9	8.67	81	75.1689	78.03
3	160	57	-1	1.67	1	2.7889	-1.67
4	152	46	-9	-9.33	81	87.0489	83.97
5	157	55	-4	-0.33	16	0.1089	1.32
6	160	50	-1	-5.33	1	28.4089	5.33
Total	966	332	0	≈ 0	216	215.3334	195.0

Correlation Coefficient

Example:

$$R(X,Y) = \frac{1}{N-1}\sum_{k=1}^{N}\left(\frac{x_k-\mu_x}{\sigma_x}\right)\left(\frac{y_k-\mu_y}{\sigma_y}\right)$$

$$\mu_x = \frac{966}{6} = 161\,\text{cm}, \qquad \mu_y = \frac{332}{6} = 55\,\text{kg},$$

$$\sigma_x = \sqrt{\frac{216}{5}} = 6.573, \qquad \sigma_y = \sqrt{\frac{215.3}{5}} = 6.563,$$

$$Corr(X,Y) = R(X,Y) = 0.904$$

Correlation Coefficient - Properties

Correlation is invariant w.r.t. linear transformations of X or Y

$$R(X,Y) = \frac{1}{N-1}\sum_{k=1}^{N}\left(\frac{x_k-\mu_x}{\sigma_x}\right)\left(\frac{y_k-\mu_y}{\sigma_y}\right) =$$

$$R(aX+b, cY+d), \qquad \text{since}$$

$$\left(\frac{ax_k+b-\mu_{ax+b}}{\sigma_{ax+b}}\right) = \left(\frac{ax_k+b-(a\mu_x+b)}{a\times\sigma_x}\right) =$$

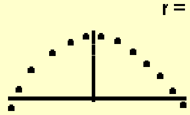
$$\left(\frac{a(x_k-\mu)+b-b}{a\times\sigma_x}\right) = \left(\frac{x_k-\mu_x}{\sigma_x}\right)$$

Correlation Coefficient - Properties

Correlation is Associative

$$R(X,Y) = \frac{1}{N}\sum_{k=1}^{N}\left(\frac{x_k-\mu_x}{\sigma_x}\right)\left(\frac{y_k-\mu_y}{\sigma_y}\right) = R(Y,X)$$

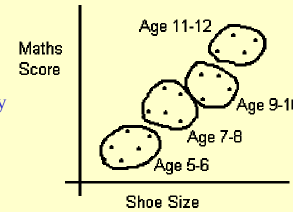
Correlation measures linear association, NOT an association in general!!! So, Corr(X,Y) could be misleading for X & Y related in a non-linear fashion.



Correlation Coefficient - Properties

$$R(X,Y) = \frac{1}{N}\sum_{k=1}^{N}\left(\frac{x_k-\mu_x}{\sigma_x}\right)\left(\frac{y_k-\mu_y}{\sigma_y}\right) = R(Y,X)$$

1. R measures the extent of linear association between two continuous variables.
2. Association does not imply causation - both variables may be affected by a third variable – age was a confounding variable.



Essential Points

6. If the experimenter has control of the levels of *X* used, how should these levels be allocated to the available experimental units?

At random! Example, testing hardness of concrete, Y, based on levels of cement, X, incorporated. Factors effecting Y: amount of H_2O, ratio stone-chips to sand, drying conditions, etc. To prevent uncontrolled differences in batches of concrete in confounding our impression of cement effects, we should choose which batch (H_20 levels, sand, dry-conditions) gets what amount of cement at random! Then investigate for X-effects in Y observations. If some significance test indicates observed trend is significantly different from a random pattern → we have evidence of causal relationship, which may strengthen even further if the results are replicable.

Essential Points

7. What theories can you explore using regression methods?

Prediction, explanation/causation, testing a scientific hypothesis/mathematical model:

a. Hooke's spring law: amount of stretch in a spring, Y, is related to the applied weight X by $Y=\alpha+\beta X$, a, b are spring constants.

b. Theory of gravity: force of gravity F between 2 objects is given by $F = \alpha/D^\beta$, where D=distance between objects, a is a constant related to the masses of the objects and $\beta=2$, according to the inverse square law.

c. Economic production function: $Q=\alpha L^\beta K^\gamma$, Q=production, L=quantity of labor, K=capital, α,β,γ are constants specific to the market studied.

Essential Points

8. People fit theoretical models to data for three main purposes.

a. To test the model, itself, by checking if the data is reasonably close agreement with the relationship predicted by the model.

b. Assuming the model is correct, to test if theoretically specified values of a parameter are consistent with the data (y=2x+1 vs. y=2.1x-0.9).

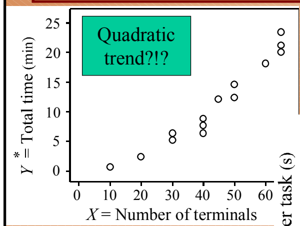
c. Assuming the model is correct, to estimate unknown constants in the model so that the relationship is completely specified (y=ax+5, a=?)

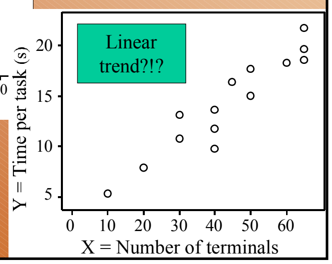
Trend and Scatter - Computer timing data

- The major components of a regression relationship are trend and scatter around the trend.
- To investigate a trend – fit a math function to data, or smooth the data.
- Computer timing data: a mainframe computer has X users, each running jobs taking Y min time. The main CPU swaps between all tasks. Y* is the total time to finish all tasks. Both Y and Y* increase with increase of tasks/users, but how?

X	=	Number of terminals:	40	50	60	45	40	10	30	20
Y*	=	Total Time (mins):	6.6	14.9	18.4	12.4	7.9	0.9	5.5	2.7
Y	=	Time Per Task (secs):	9.9	17.8	18.4	16.5	11.9	5.5	11	8.1
X	=	Number of terminals:	50	30	65	40	65	65		
Y*	=	Total Time (mins):	12.6	6.7	23.6	9.2	20.2	21.4		
Y	=	Time Per Task (secs):	15.1	13.3	21.8	13.8	18.6	19.8		

Trend and Scatter - Computer timing data

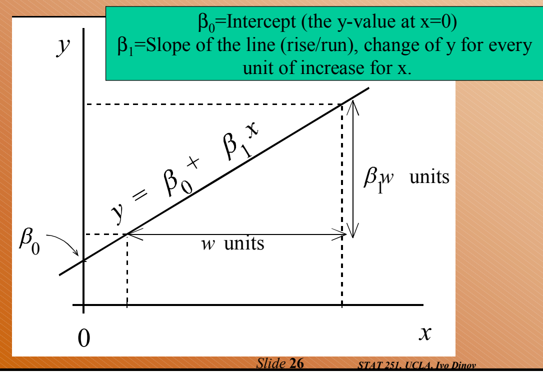




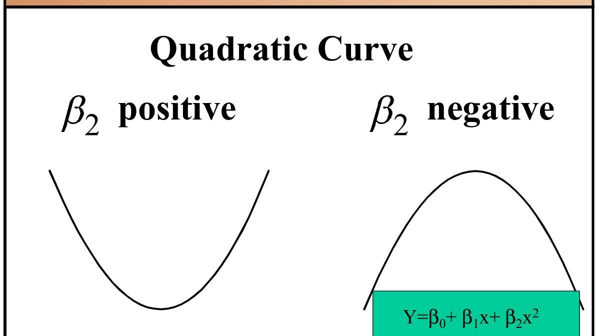
We want to find reasonable models (descriptions) for these data!

Equation for the straight line – linear/affine function

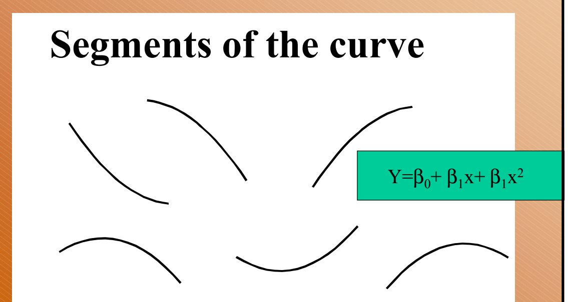
β_0=Intercept (the y-value at x=0)
β_1=Slope of the line (rise/run), change of y for every unit of increase for x.



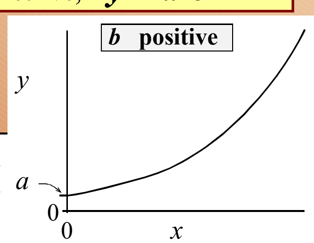
The quadratic curve

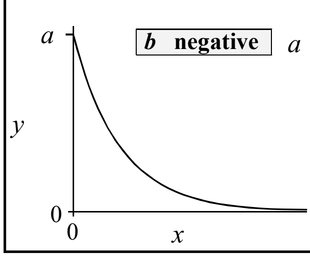


The quadratic curve



The exponential curve, $y = a\, e^{bx}$





Used in population growth/decay models.

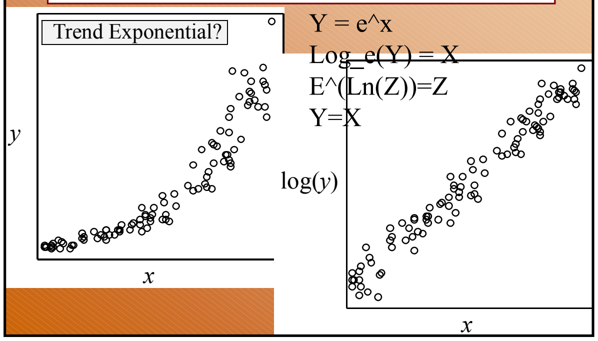
Effects of changing x for different functions/curves

A straight ***line*** changes by a fixed ***amount*** with each unit change in x.

An ***exponential*** changes by a fixed ***percentage*** with each unit change in x.

To tell whether a trend is exponential ….

check whether a plot of **log(y) versus x** has a <u>linear trend</u>.

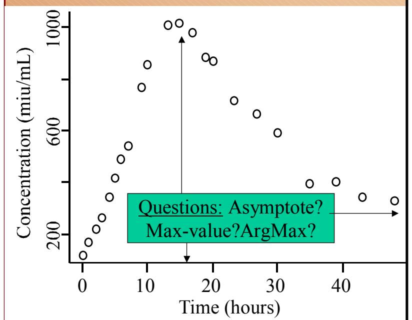


Creatine kinase concentration in patient's blood

You should not let the questions you want to ask be dictated by the tools you know how to use.

Here Y=creatine kinase concentration in blood for a set of heart attack patients vs. the time, X.

No symmetry so X^2 models won't work!

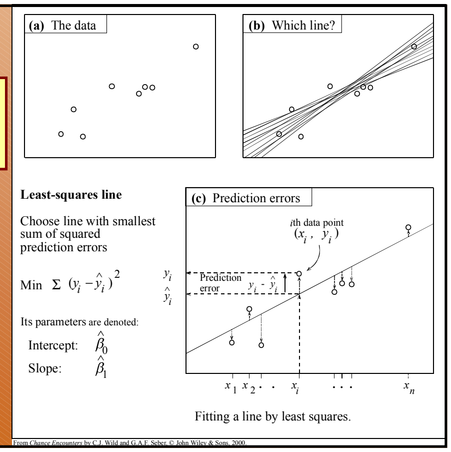


Comments

1. In statistics what are the two main approaches to summarizing trends in data? (model fitting; smoothing – done by the eye!)
2. In $y = 5x + 2$, what information do the 5 and the 2 convey? (slope, y-intercept)
3. In $y = 7 + 5x$, what change in y is associated with a 1-unit increase in x? with a 10-unit increase? (5; 50)

   How about for $y = 7- 5x$. (-5; -50)
5. How can we tell whether a trend in a scatter plot is exponential? (plot ***log***(Y) vs. X, should be linear)

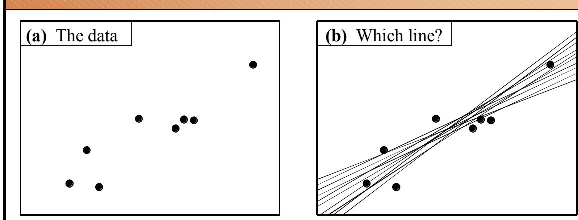
Choosing the "best-fitting" line



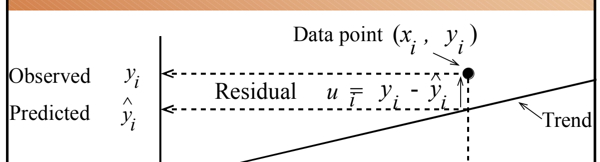
Fitting a line through the data

Show the Regression-Line Simulation Applet

[RegressionApplet.html]



The idea of a residual or prediction error



Least squares criterion

Least squares criterion: Choose the values of the parameters to *minimize the sum of squared prediction errors* (or sum of squared residuals),

$$\sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

The least squares line

Least-squares line

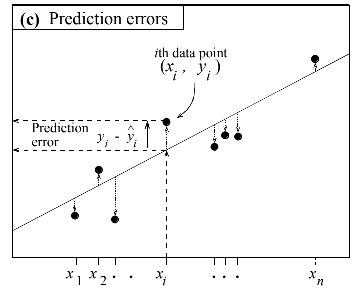
Choose line with smallest sum of squared prediction errors

Min $\Sigma\ (y_i - \hat{y}_i)^2$

Its parameters are denoted:

Intercept: $\hat{\beta}_0$

Slope: $\hat{\beta}_1$



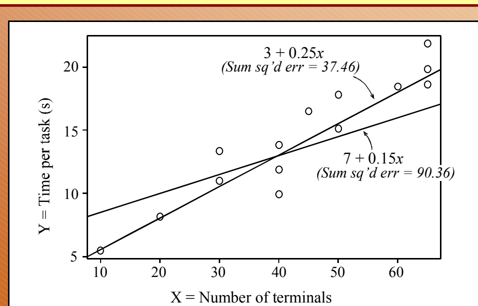
Least-squares line: $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$

The least squares line

Least-squares line: $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$

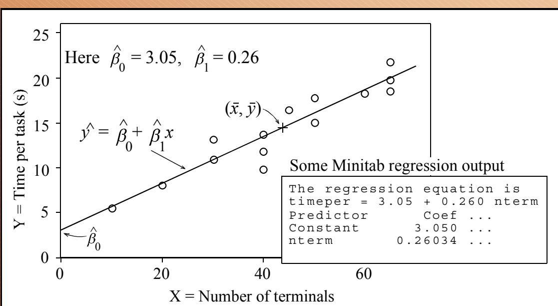
$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} \left[(x_i - \bar{x})(y_i - \bar{y})\right]}{\sum_{i=1}^{n} (x_i - \bar{x})^2}; \qquad \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

Computer timings data – linear fit



Two lines on the computer-timings data.

Adding the least squares line



Computer-timings data with least-squares line.

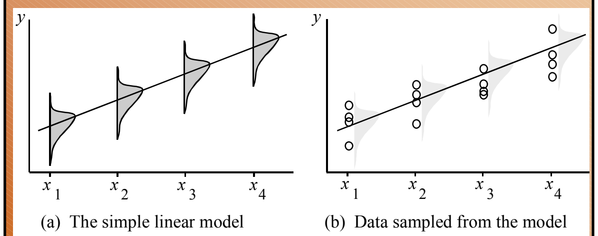
Example – Method/Hemi/Tissue/Value

1. C:\Ivo.dir\Research\Data.dir\WM_GM_CSF_tissueMaps.dir
2. SYSTAT: → regression Value = $c_0 + c_1 M + c_2 H + c_3 T$
3. Results:

Effect	Coefficient	SE	t	P(2 Tail)	
CONSTANT	1.02231E+05	9087	11.24911	0.00000	
METHOD	-3703.77667	3635	-1.01887	0.31038	← Insignif
TISSUE	-22623.47875	2226	-1.01E01	0.00000	
HEMISPH	-2.13667	3635	-0.00059	0.99953	

Effect	Coeff.	Lower < 95%>	Upper
CONSTANT	1.02231E+05	84231.33157	1.20231E+05
METHOD	-3703.77667	-10903.69304	3496.13971
TISSUE	-22623.47875	-27032.50908	-18214.44842
HEMISPH	-2.13667	-7202.05304	7197.77971

The simple linear model



(a) The simple linear model (b) Data sampled from the model

When $X = x$, $Y \sim \text{Normal}(\mu_Y, \sigma)$ where $\mu_Y = \beta_0 + \beta_1 x$, **OR**

when $X = x$, $\boldsymbol{Y = \beta_0 + \beta_1 x + U}$, where $U \sim \text{Normal}(0, \sigma)$

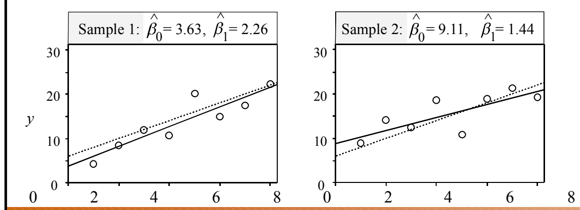
Random error

Data generated from $Y = 6 + 2x + \text{error (U)}$

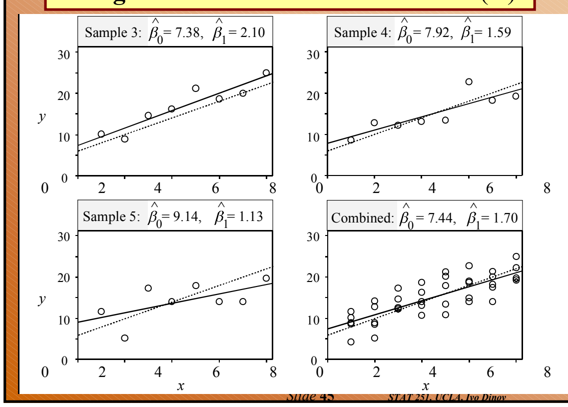
Dotted line ·········· is true line and

solid line ——— is the data-estimated LS line.

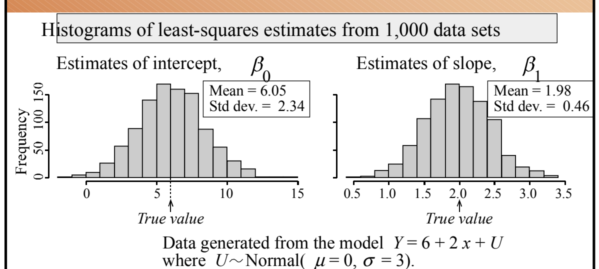
Note differences between true β_0=6, β_1=2 and their estimates β_0^ & β_1^.



Data generated from $Y = 6 + 2x + \text{error(U)}$



Data generated from $Y = 6 + 2x + \text{error(U)}$



Summary

For the simple linear model, ***least-squares estimates are unbiased [E(β^)= β] and Normally distributed.***

Noisier data produce *more-variable* least-squares estimates.

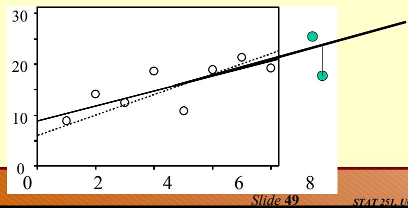
Summary

1. Before considering using the simple linear model, what sort of pattern would you be looking for in the scatter plot? (linear trend with constant scatter spread across the range of X)
2. What assumptions are made by the simple linear model, **SLM**? (X is linearly related to the mean value of the Y obs's at each X, $\mu_Y = \beta_0 + \beta_1 x$; where β_0 & β_1 are the true values of the intercept and slope of the SLM; The LS estimates β_0^ & β_1^ estimate the true values of β_0 & β_1; and the random errors $U = Y - \mu_Y \sim N(\mu, \sigma)$.)
3. If the simple linear model holds, what do you know about the sampling distributions of the least-squares estimates? (Unbiased and Normally distributed)

Summary

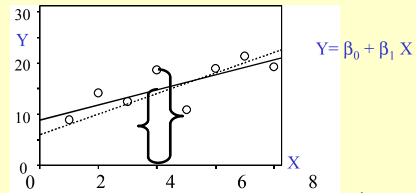
4. In the simple linear model, what behavior is governed by σ? (the spread of scatter of the data around trend)

5. Our estimate of σ can be thought of as a sample standard deviation for the set of prediction errors from the least-squares line.



RMS Error for regression

- Error = Actual value – Predicted value



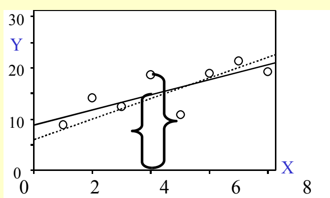
- The RMS Error for the regression line $Y = \beta_0 + \beta_1 X$ is

$$\sqrt{\frac{(y_1-\hat{y}_1)^2+(y_2-\hat{y}_2)^2+(y_3-\hat{y}_3)^2+(y_4-\hat{y}_4)^2+(y_5-\hat{y}_5)^2}{5-1}}$$

where $\hat{y}_k = \hat{\beta}_0 + \hat{\beta}_1 x_k, \quad 1 \le k \le 5$

Compute the RMS Error for this regression line

- Error = Actual value – Predicted value



X	Y
1	9
2	15
3	12
4	19
5	11
6	20
7	22
8	18

- The RMS Error for the regression line $Y = \beta_0 + \beta_1 X$ is

$$\sqrt{\frac{(y_1-\hat{y}_1)^2+(y_2-\hat{y}_2)^2+(y_3-\hat{y}_3)^2+(y_4-\hat{y}_4)^2+(y_5-\hat{y}_5)^2}{5-1}}$$

where $\hat{y}_k = \hat{\beta}_0 + \hat{\beta}_1 x_k, \quad 1 \le k \le 5$

Compute the RMS Error for this regression line

- Error = Actual value – Predicted value
- The RMS Error for the regression line $Y = \beta_0 + \beta_1 X$ is

$$\sqrt{\frac{(y_1-\hat{y}_1)^2+(y_2-\hat{y}_2)^2+(y_3-\hat{y}_3)^2+(y_4-\hat{y}_4)^2+(y_5-\hat{y}_5)^2}{5-1}}$$

where $\hat{y}_k = \hat{\beta}_0 + \hat{\beta}_1 x_k, \quad 1 \le k \le 5$

- First compute the LS linear fit (estimate β_0^ + β_1^)
- Then Compute the individual errors
- Finally compute the cumulative RMS measure.

X	Y
1	9
2	15
3	12
4	19
5	11
6	20
7	22
8	18

Compute the RMS Error for this regression line

- Then Compute the individual errors

$$(y_k-\hat{y}_k)^2, \text{where} \quad \hat{y}_k = \hat{\beta}_0 + \hat{\beta}_1 x_k, \quad 1 \le k \le 8$$

- Finally compute the cumulative RMS measure.

$$\sqrt{\frac{(y_1-\hat{y}_1)^2+(y_2-\hat{y}_2)^2+(y_3-\hat{y}_3)^2+(y_4-\hat{y}_4)^2+(y_5-\hat{y}_5)^2}{5-1}}$$

where $\hat{y}_k = \hat{\beta}_0 + \hat{\beta}_1 x_k, \quad 1 \le k \le 5$

- Note on the Correlation coefficient formula,

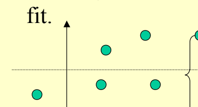
$$R(X,Y) = \frac{1}{N-1}\sum_{k=1}^{N}\left(\frac{x_k-\mu_x}{\sigma_x}\right)\left(\frac{y_k-\mu_y}{\sigma_y}\right)$$

$X=\{x_1, x_2, \ldots, x_N\}$
$Y=\{y_1, y_2, \ldots, y_N\}$
(μ_X, σ_X), (μ_Y, σ_Y)
sample mean / SD.

X	Y
1	9
2	15
3	12
4	19
5	11
6	20
7	22
8	18

Compute the RMS Error for this regression line

- The RMS Error for the regression line $Y = \beta_0 + \beta_1 X$ says how far away from the (model/predicting) regression line is each observation.
- Observe that the SD(Y) is also a RMS Error measure of another specific line – horizontal line through the average of the Y values. This line may also be taken for a regression line, but often it's not the best linear fit.



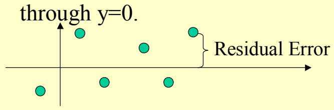
$$SD(Y) = \sqrt{\frac{1}{N-1}\sum_{k=1}^{N}(Y_k-\bar{Y})^2} \quad \text{vs.}$$

$$RMSE(Y,\hat{Y}_k = \hat{\beta}_0 + \hat{\beta}_1 X) = \sqrt{\frac{1}{N-1}\sum_{k=1}^{N}(Y_k-\hat{Y}_k)^2}$$

Predicted vs. Observed

Plotting the Residuals

- The Residuals=Observed –Predicted for the regression line Y= $\beta_0 + \beta_1$ X (just like the error).
- Residuals average to zero, mathematically, and the regression line for the residuals is a horizontal line through y=0.



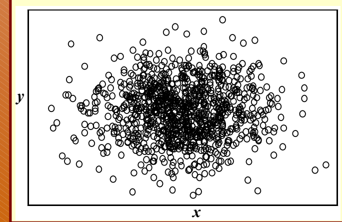
When $X = x$, $Y \sim \text{Normal}(\mu_Y, \sigma)$ where $\mu_Y = \beta_0 + \beta_1 x$, **OR**

when $X = x$, $\boldsymbol{Y = \beta_0 + \beta_1 x + U}$, where $U \sim \text{Normal}(0, \sigma)$

Random error

Plotting the Residuals – patterns?

- The Residuals=Observed –Predicted for the regression line Y= $\beta_0 + \beta_1$ X + U should show no clear trend or pattern, for our linear model to be a good and useful approximation to the unknown process.



Inference – just a glance at statistical inference

- The regression intercept β_0 and slope β_1 are usually called *regression coefficients*
  - The least squares estimates of their values are found in the coefficients column of program printouts
- **Confidence intervals** for a true regression coefficient (whether intercept or slope) is given by

  estimated coefficient ± t std errors

- **t-test statistic** $\boxed{df = n - 2}$

$$t_0 = \frac{\textit{estimated coefficient} - \textit{hypothesized value}}{\textit{standard error}}$$

Inferences

- **Confidence intervals** for a true regression coefficient (whether intercept or slope) is given by

  estimated coefficient ± t std errors

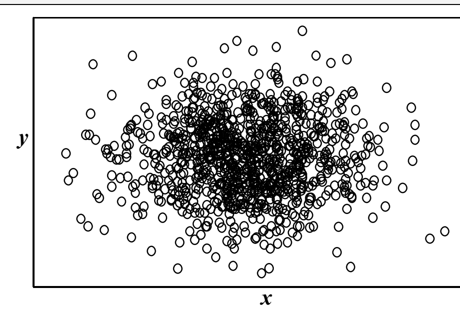
  β_1^ ± t SE(β_1^)

- **t-test statistic** $\boxed{df = n - 2}$ **Ho: β_1 =c**

$$t_0 = \frac{\hat{\beta}_1 - c}{SE(\hat{\beta}_1)}$$

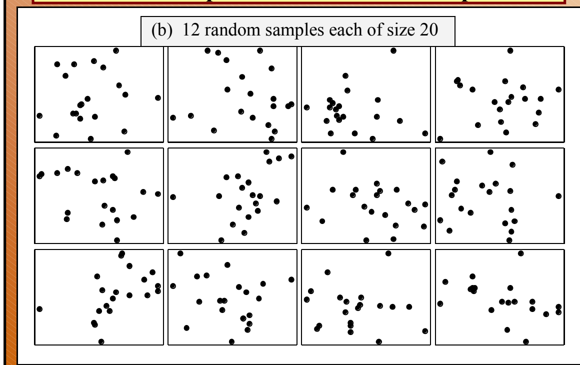
Is there always an X Y relationship? Linear Relationship ?

(a) 1000 data points with no relationship between *X* and *Y*



From *Chance Encounters* by C.J. Wild and G.A.F. Seber, © John Wiley & Sons, 1999.

Random samples from these 1000 data points



From *Chance Encounters* by C.J. Wild and G.A.F. Seber, © John Wiley & Sons, 2000.

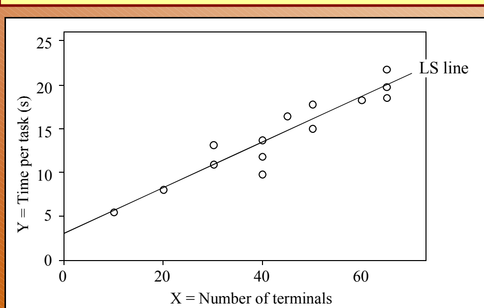
Testing for no linear relationship – trend of Y w.r.t. X is trivial!

H_0: *true slope* = 0

OR

H_0: $\beta_1 = 0$

Computer timings
How does the job completion timing depend on the number of computer tasks?



Computer timings
How does the job completion timing depend on the number of computer tasks?

Regression Analysis *Standard errors* *t-statistics* *P-values*

```
The regression equation is
timeper = 3.05 + 0.260 nterm
```

Predictor	Coef	StDev	T	P
Constant	3.050	1.260	2.42	0.032
nterm	0.26034	0.02705	9.62	0.000

testing $H_0: \beta_i = 0$

CI for true slope

Regression Analysis *Standard errors* *t-statistics* *P-values*

```
The regression equation is
timeper = 3.05 + 0.260 nterm
```

Predictor	Coef	StDev	T	P
Constant	3.050	1.260	2.42	0.032
nterm	0.26034	0.02705	9.62	0.000

testing $H_0: \beta_i = 0$

For a 95% CI with $df = n{-}2 = 12$, $t = 2.179$

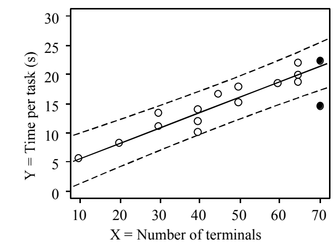
CI: estimate $\pm$ t std errors
$= 0.26034 \pm 2.179 \times 0.02705 = [0.20,\ 0.32]$

Prediction

Predicting at $X = x_p$

- The **confidence interval for the mean** estimates the **average** Y-value at $X = x_p$.
  - (averaged over many repetitions of the experiment.)
- The **prediction interval** (PI) tries to predict the next **actual** Y-value at x_p, in the future.

Predicting *time-per-task* for 70 terminals

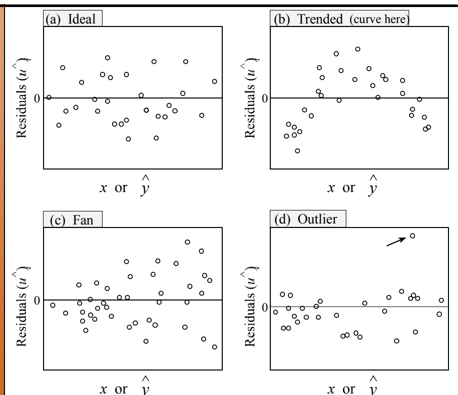


Time per Task versus Number of Terminals
(with the least-squares line and 95% PI's superimposed).

From *Chance Encounters* by C.J. Wild and G.A.F. Seber, © John Wiley & Sons, 2000.

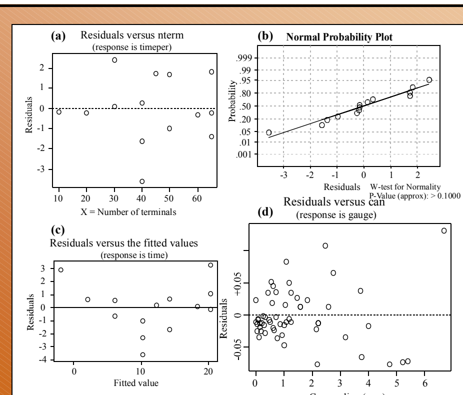
Review

1. What is the difference between a confidence interval for the mean and a prediction interval?
2. Prediction intervals make allowances for two sources of uncertainty. What are they? How does a confidence interval for the mean differ in this regard?
3. At what point along the *X*-axis are these intervals narrowest?
4. We gave some general warnings about prediction earlier. They are relevant here as well. What were those warnings?



Patterns in residual plots.

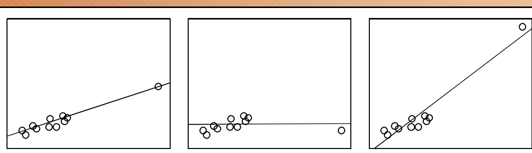
From *Chance Encounters* by C.J. Wild and G.A.F. Seber, © John Wiley & Sons, 2000.



Examples of residual plots.

From *Chance Encounters* by C.J. Wild and G.A.F. Seber, © John Wiley & Sons, 2000.

Effect of an outlier in *X* on the LS line

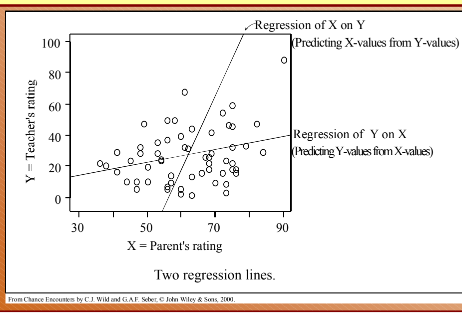


The effect of an *X*-outlier on the least-squares line.

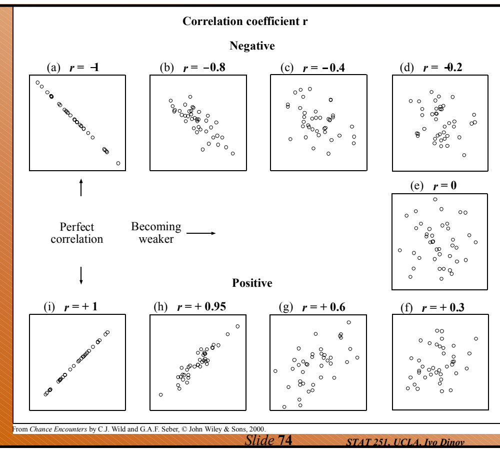
From *Chance Encounters* by C.J. Wild and G.A.F. Seber, © John Wiley & Sons, 2000.

Review

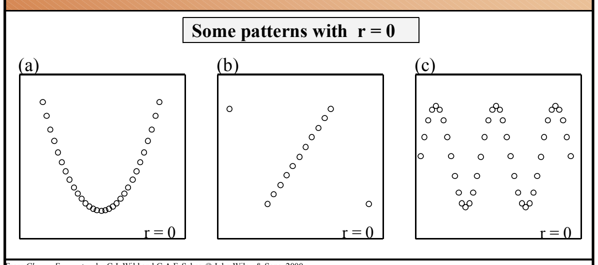
1. What assumptions are made by the simple linear model?
2. Which assumptions are critical for all types of inference?
3. What types of inference are relatively robust against departures from the Normality assumption?
4. Four types of residual plot were described. What were they, and what can we learn from each?
5. What is an outlier in *X*, and why do we have to be on the lookout for such observations?



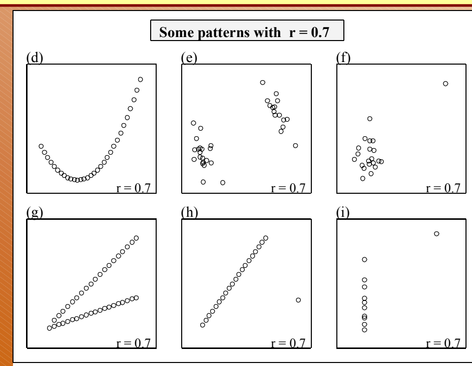
Correlation of parent and teacher = 0.297, P-value = 0.024



Misuse of the correlation coefficient



Misuse of the correlation coefficient



Correlation does not necessarily imply causation.

Review

1. Describe a fundamental difference between the way regression treats data and the way correlation treats data.
2. What is the correlation coefficient intended to measure?
3. For what shape(s) of trend in a scatter plot does it make sense to calculate a correlation coefficient?
4. What is the meaning of a correlation coefficient of $r = +1$? $r = -1$? $r = 0$?

Summary

Concepts

- Relationships between quantitative variables should be explored using **scatter plots**.
  - Usually the *Y* variable is continuous
    (or behaves like one in that there are few repeated values)
  - and the *X* variable is discrete or continuous.
- **Regression** singles out one variable (*Y*) as the response and uses the explanatory variable (*X*) to explain or predict its behavior.
- **Correlation** treats both variables symmetrically.

Concepts cont'd

In practical problems, regression models may be fitted for any of the following reasons:

- To understand a **causal relationship** better.
- To find relationships which may be **causal**.
- To make **predictions**.
  - But be cautious about predicting outside the range of the data
- To **test theories**.
- To **estimate parameters** in a theoretical model.

Concepts cont'd

- In observational data, strong relationships are not necessarily causal.
- We can only have reliable evidence of causation from controlled experiments.
- Be aware of the possibility of **lurking** variables which may effect both *X* and *Y*.

Concepts cont'd

- Two important trend curves are the **straight line** and the **exponential curve**.
  - A straight line changes by a *fixed amount* with each unit change in *x*.
  - An exponential curve changes by a *fixed percentage* with each unit change in *x*.
- You should not let the questions you want to ask of your data be dictated by the tools you know how to use. You can always ask for help.

Concepts cont'd

- The two main approaches to summarizing trends in data are using *smoothers* and *fitting mathematical curves*.
- The ***least-squares criterion*** for fitting a mathematical curve is to choose the values of the parameters (e.g. β_0 and β_1) to minimize the sum of squared prediction errors, $\sum(y_i - \hat{y}_i)^2$.

Linear Relationship

- We fit the linear relationship $\hat{y} = \beta_0 + \beta_1 x$.
- The slope β_1 is the change in $\hat{y}$ associated with a one-unit increase in x.

Least-squares estimates

- The least-squares estimates, $\hat{\beta}_0$ and $\hat{\beta}_1$ are chosen to minimize $\sum (y_i - \hat{y}_i)^2$.
- The **least-squares regression line** is $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$.

Model for statistical inference

- Our theory assumes the model $Y_i = \beta_0 + \beta_1 x_i + U_i$,
- where the random errors, $U_1, U_2, \ldots, U_n$, are a random sample from a Normal(0, σ) distribution.
- This means that the random errors ….
  - are Normally distributed (each with mean 0),
  - all have the same standard deviation σ regardless of the value of x, and
  - are all independent.

Residuals and outliers

- These assumptions should be checked using residual plots (Section 12.4.4). The *i*th *residual* (or *prediction error*) is

$$y_i - \hat{y}_i = \text{observed - predicted.}$$

- An ***outlier*** is a data point with an unexpectedly large residual (positive or negative).

Inference

- **Inferences for the intercept and slope** are just as in Chapters 8 and 9, with confidence intervals being of the form *estimate* ± *t std errors* and test statistics of the form

  t_0 = *(estimate - hypothesized value)/ standard error.*

  - We use $df = n - 2$.
  - To test for ***no linear association***, we test H_0: $\beta_1 = 0$.

*Prediction

- The predicted value for a new Y at $X = x_p$ is

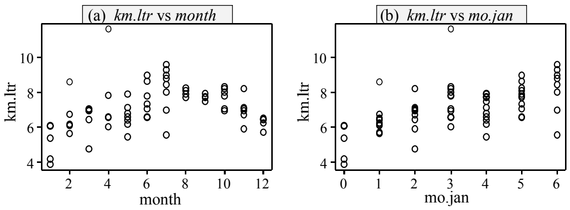
$$\hat{y}_p = \hat{\beta}_0 + \hat{\beta}_1 x_p$$

- The *confidence interval for the mean* estimates the average Y-value at $X = x_p$.
  - averaged over many repetitions of the experiment.
- The *prediction interval* tries to predict the next *actual* Y-value at $X = x_p$.
- The prediction interval is wider than the corresponding confidence interval for the mean.

Correlation coefficient

The correlation coefficient *r* is a measure of linear association with $-1 \le r \le 1$.

- If $r = 1$, then X and Y have a perfect positive linear relationship.
- If r = –1, then X and Y have a perfect negative linear relationship.
- If r = 0, then there is no linear relationship between X and Y.
- Correlation does not necessarily imply causation.



(c) Regression of *km.ltr* on *mo.jan*

	Coef	Std Err	t-value	p-value	CI lower	CI upper
Intercept	5.889	0.2617	22.506	<1.0e0-6	5.367	6.411
mo.jan	0.386	0.0721	5.361	1.134e-06	0.243	0.530

Percentage of variation explained: 30.34
Estimate of error Std dev: 1.075366
Error df: 66

Fuel consumption data.

Regression of Log(price) on Age

	Coef	Std Err	t-value	p-value	CI lower	CI upper
Intercept	3.8511	0.0494	78.02	0	---	---
age	-0.2164	0.0095	-22.67	0	-0.24	-0.20

Percent of variation explained: 90.02
Estimate of error Std dev: 0.2433205
Error df: 57

Age	0	1	2	3	4	5	6	7	8	9	10
Predicted	3.85	3.63	3.42	3.20	2.99	2.77	----	2.34	2.12	1.90	1.69
Pred lower	3.35	3.14	2.93	2.71	2.49	2.28	----	1.84	1.62	1.40	1.18
Pred upper	4.35	4.13	3.91	3.69	3.48	3.26	----	2.83	2.62	2.40	2.19