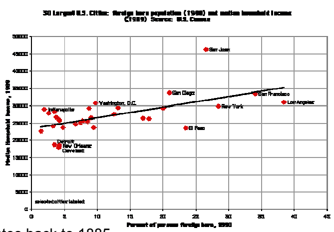
Bivariate regression



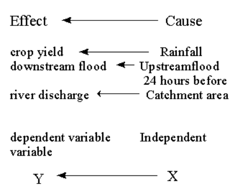
Probably dates back to 1885

- the correlation coefficient measures the association between 2 sets of paired variates, but it does not
  - tell us the way the two variables are related
  - does not allow us to predict the value of one variable with knowledge of the value of the other variable
  - doesn't signal anomalies in the relationship between individual pairs
- bivariate regression lets us do all of these things

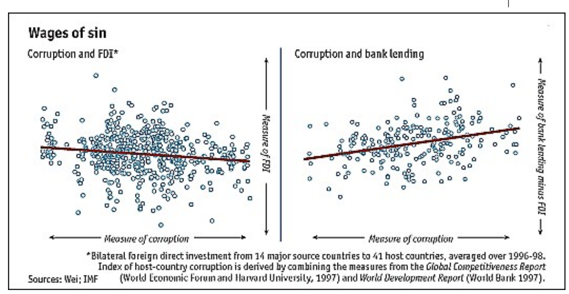
---

dependent and independent variables

- regression allows us to suggest (hypothesize) <u>causal</u> relationships and their direction - substantiated by previous research and common sense



Which way is the relationship?

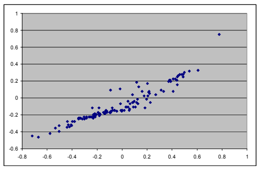


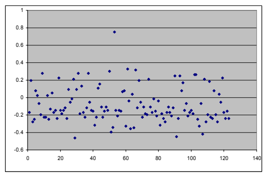
- scattergram - used to plot dependent along y axis, independent along x axis
- regression involves plotting a 'best-fit' line between the points on a scattergram
- convention is to treat the dependent variable as PREDICTED and the independent variable as the PREDICTOR

- prediction/interpolation is one of the main uses because x and y are <u>sampled</u>.
- As we don't have complete information on values for a given x we want to interpolate intermediate values from the best fit line on the scattergram

Scattergram or scatterplots

- Interpretation: how close to a single line?





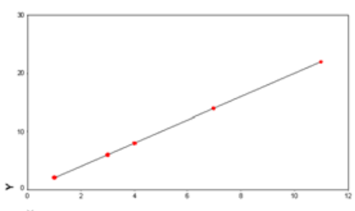
- Describing relationships (scatterplots) video
- To see the unedited version go to:
- http://www.learner.org/resources/series65.html,
  - Episode 8. Describing Relationships

---

derivation of best fit line

example 1:

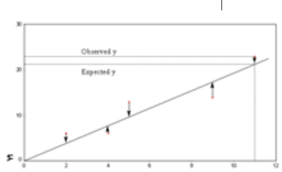
y	2	6	8	14	22
x	1	3	4	7	11



- easy to place 'best' line through these points as the association is perfect
- correlation coefficient =1
- there are no residuals/anomalies/no deviations of points from general relationship since every point is on the regression line
- however variables are rarely perfectly correlated because of 1) poor/theory/understanding or 2) measurement error

---

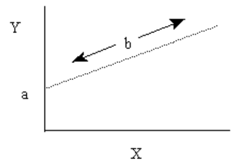
- can place 'best-fit' line through points although r<1 and so points representing variates do form a straight line
- deviations/anomalies /residuals from regression are shown as ↓↑



- residuals: why plot them vertically rather than perpendicular to the regression line?
- Because residuals are the difference between the actual/observed values of the dependent variable (y values) and the expected/predicted value of the dependent variable ($\hat{y}$) for a particular value of x

fitting the regression line by least square method

- any straight line drawn on an x y coordinate system can be represented by an equation of the form



Equation review

- Slope/Y-intercept equation

Independent Variable

Slope, or coefficient

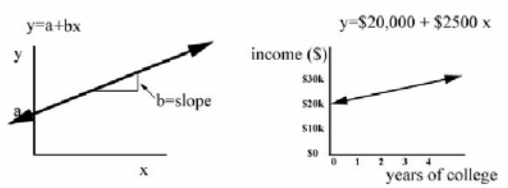
$$Y = a + bx$$

Dependent variable

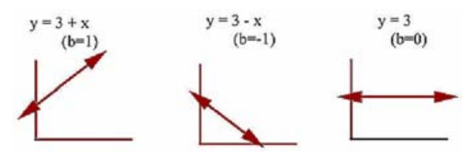
Y-intercept

Equation review

- Slope
  - rise/run
  - $\Delta Y/\Delta X$



Slopes



- least square approach - objective is to find the combination of intercept and slope values which minimize the sum of squares of the residual values, that is, minimize the difference between the actual and predicted values at particular values of x

The intercept

- There is no intercept if the data values are standardized before they're used
- The intercept is only meaningful if it makes sense for the independent variable take on a value of 0
- There should be some values recorded near zero if it is to be interpreted

Sample regression function

- We've been talking about population characteristics; we will measure <u>sample</u>:

"Hats" always indicate sample

$$\hat{Y}_i = \hat{\beta}_1 + \hat{\beta}_2 X_i$$

Y-hat: conditional mean value of Y in the sample: $E(Y|X_i)$

Our estimate of real (population) value of Y

Sample regression function

- We've been talking about population characteristics; we will measure sample:

$$\hat{Y}_i = \hat{\beta}_1 + \hat{\beta}_2 X_i$$

$$\hat{u}_i = Y_i - \hat{Y}_i$$

This is the definition of a residual

Sample Regression function

- We've been talking about population characteristics; we will measure sample:

This gives us the stochastic sample regression function. We will use this to infer things about the relationship between X and Y in population

$$\hat{Y}_i = \hat{\beta}_1 + \hat{\beta}_2 X_i$$

$$\hat{u}_i = Y_i - \hat{Y}_i$$

$$Y_i = \hat{\beta}_1 + \hat{\beta}_2 X_i + \hat{u}_i$$

Sample Regression function

- We've been talking about population characteristics; we will measure sample:

Residual or error term

$$Y_i = \hat{\beta}_1 + \hat{\beta}_2 X_i + \hat{u}_i$$

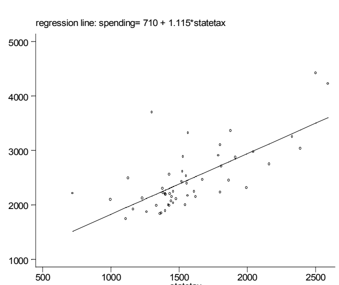
Actual Y's in our sample; correspond with X_i's

Sample coefficients: define the shape of the line between X and Y in our sample

Actual X's in our sample; correspond with Y_i's

How we do the math

- We choose ordinary least squares: minimize the squared vertical difference between line and points



How we do the math

- Ordinary Least Squares (OLS): Mathematically minimize the squared *vertical* distance between all the points and the line
- Why squared?

Why we like OLS

- Mathematical properties
  - Point estimators
  - Pass through sample means
  - Mean of residuals = 0
  - Residuals uncorrelated with X and Y

Why we like OLS

- Mathematical properties
- Statistical properties:
  - There is some real relationship (+, -, or zero) between X's and Y's in the population
  - We evaluate the relationship between X_i and Y_i in our sample

---

- The relationship between our sample parameters and the population parameters is similar to relationship between sample statistics and population parameters
- That is: our sample beta-hats are drawn from a sampling distribution around the real population parameters

OLS assumptions and Gauss-Markov

- For population beta sampling distribution:
- GM theorem proves that $\hat{\beta}$'s are BLUE:
  - Best: least variance
  - Linear (descriptive of process)
  - Unbiased: centered around real β
  - Consistent: as N → ∞, $\hat{\beta} \to \beta$

Interpretation of Coefficients

- Recall our basic equation:
- β_1 $\quad Y_i = \hat{\beta}_1 + \hat{\beta}_2 X_i + \hat{u}_i = E(Y \mid X_i)$
  - The "constant," or Y-intercept
  - Predicted value of Y when X = zero (why?)

Interpretation of Coefficients

- Recall our basic equation:
- β_1 $\quad Y_i = \hat{\beta}_1 + \hat{\beta}_2 X_i + \hat{u}_i = E(Y \mid X_i)$
  - The "constant," or Y-intercept
  - Predicted value of Y when X = zero (why?)
    - This may or may not be logically useful concept
    - Always exercise caution paying too much attention to constant predictions anyway

Interpretation of Coefficients

- Recall our basic equation:
- β_2 $\quad Y_i = \hat{\beta}_1 + \hat{\beta}_2 X_i + \hat{u}_i = E(Y \mid X_i)$
  - Slope coefficient; recall

$$\beta_2 = slope = \frac{rise}{run} = \frac{\Delta Y}{\Delta X}$$

Interpretation of Coefficients

- Recall our basic equation:
- β2 $\quad Y_i = \hat{\beta}_1 + \hat{\beta}_2 X_i + \hat{u}_i = E(Y \mid X_i)$
  - So what does that mean for a slope of 0.44 (say)?

$$\beta_2 = slope = \frac{rise}{run} = \frac{\Delta Y}{\Delta X}$$

Interpretation of Coefficients

- Units matter
  - Always interpret equations in light of units on both sides
  - Should always use logical units, but may choose the logical unit which makes regression most tractable:
    - Aim for coefficients between zero and 10
    - Avoid (non-zero) coefficients <.05 or so

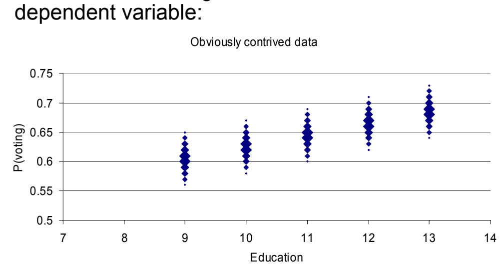
- Things we cannot draw with a Y=a+bX equation:
  - Infinite slope
  - Interrupted line (discontinuous function)
  - Non-functions
- Functions may be non-linear though

Conceptual overview

- Regression line shows relationship between fixed values of X and average values of Y
- Regression assumes relationship contains stochastic element
  - Outcomes follow probability distribution

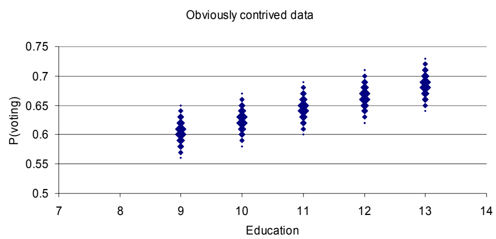
Regression: conceptual overview

- Relationship between fixed independent variables and average values of stochastic dependent variable:



Regression: conceptual overview

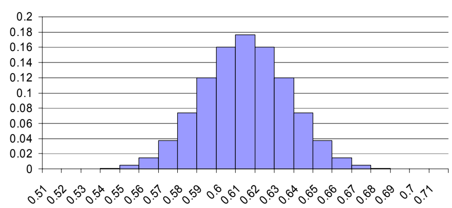
- For each value of X, Y follows a probability distribution



Regression: conceptual overview

- For each value of X, Y follows a probability distribution:
- The probability distribution of a discrete random variable is a list of probabilities associated with each of its possible values
- We would like these distributions to be normal

Regression: conceptual overview



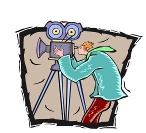
Regression: conceptual overview

- Why is the dependent variable stochastic?
  - Incomplete model on theoretical level
  - Not possible to collect quantitative data on everything
  - Measurement errors
  - Parsimony: not worth trying to develop perfect model
  - Wrong functional form

Regression: conceptual overview

- Regression vs. Correlation
  - Correlation is symmetric
    - Linear association
    - No “dependent” and “independent” variables
    - Both variables are assumed to be random
  - Regression analysis is asymmetric
    - Independent variable not random, but fixed in repeated samples
    - We assume only dependent variable is random, or that it follows probability function

- Inference for relationships video
- To view the video on your own go to:
  - http://www.learner.org/resources/series65.html
  - Episode 25



How we do the math: OLS

- The whole exercise is about fitting a line to points in our sample scatterplot, using the equation Y = a +bX
  - By changing values of a and b, this equation will give us any straight line which exists in a plane
  - We will use OLS to figure out which values of a and b provide the best fit

The slope

b = the amount of change in y with a change in 1 unit of x
b = gradient of the regression line

$$b = \frac{\sum_{i=1}^{n} x_i y_i - \frac{\sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{n}}{\sum_{i=1}^{n} x_i^2 - \frac{(\sum_{i=1}^{n} x_i)^2}{n}} = \frac{\textbf{covariation}}{\textbf{variation in x}}$$

The intercept

$$a = \frac{\sum_{i=1}^{n} y_i}{n} - b\left(\frac{\sum_{i=1}^{n} x_i}{n}\right) = \textbf{intercept with y axis when x = 0}$$

- The intercept is normally of little interest
- Often the y range doesn't even include the intercept
- There are occasions when the intercept is relevant
  - Example would be a regression of crop yield versus fertilizer use, when a=0 would denote no use of fertilizer

How we do the math: OLS

- We will use OLS to figure out which values of a and b best fit the dots in our picture
- Note that depending on how we define "best," we'll end up with different lines for same picture

Example

River	basin sq km (x)	Discharge km^3 (y)		
Nile	3031700	324		
Amazon	7050000	6630		
Chang Jiang	1800000	900		
Huang Ho Yellow)	445000	50		
Mackenzie	1805200	11		
Mississippi	3226300	620		
Indus	1138800	146		
Nelson-Saskatchewan	1109400	87		

Σy_i=8768	Σx_i= 19606.4	$\sum(y_i - \frac{\sqrt{\sum(y_i-\hat{y})^2/(n-2)}}{\sqrt{\sum x^2-(\sum x)^2/n}})^2 = 4095.8$
	$(\Sigma x_i)^2$=384410921	Σx_i^2=78527122
	$\Sigma x_i y_i$ =51648967	

SPSS output

Coefficients

	Unstandardized Coefficients		Standardized Coefficients	t	Sig.
	B	Std. Error	Beta		
(Constant)	-1329.44	559.4231		-2.37644	0.055035
BASIN	0.00099	0.000179	0.914658	5.542508	0.001456

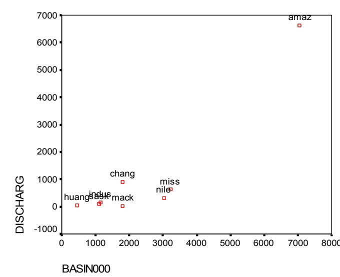
Dependent Variable: DISCHARG

Revised output with basin rescaled

Coefficients

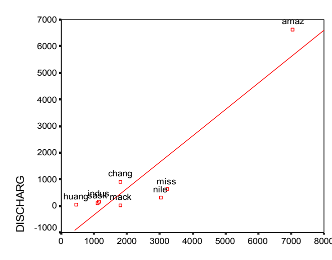
	Unstandardized Coefficients		Standardized Coefficients	t	Sig.
	B	Std. Error	Beta		
(Constant)	-1329.44	559.4231		-2.37644	0.055035
BASIN000	0.989651	0.178556	0.914658	5.542508	0.001456

Dependent Variable: DISCHARG



$$b = \frac{51648967 - \left[\frac{19606.4(876)}{8}\right]}{78527122 - \frac{384410921^2}{8}} = \frac{30160352.2}{304757523} = 0.99$$

$$a = \frac{8768}{8} - 0.99\left(\frac{19606.4}{8}\right) = -1330.3$$



Predicted values

$\hat{y}=a+bx$	i=1	-1330.3+[0.99*3031.7]	=	1671.1
	i=2	-1330.3+[0.99*7050]	=	5649.2
	i=3	-1330.3+[0.99*1800]	=	451.7

- the standard method of measuring the goodness of fit of a regression is to calculate the extent to which the regression accounts for the variation in the observed values of the dependent variable
- this is done by calculating the variance of the observed value of y

$$s^2 = \frac{\sum \hat{y}^2}{n} - \bar{y}^2 = \textbf{regression variance}$$

$$s_y^2 = \frac{\sum y^2}{n} - \bar{y}^2 = \textbf{total variance}$$

$$r^2 = \frac{s^2}{s_y^2} = \textbf{coefficient of determination}$$

- tests on the residuals
- a complementary test of goodness of fit involves looking at the residuals
- there should be no systematic variation in the residuals

Coefficient of determination

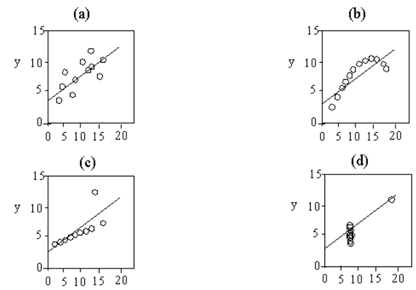
River	Basin km^2 (000s) x	km^3 y	y^2	y hat	resid
Nile	3031.7	324	104976	1671.1	-1347.1
Amazon	7050	6630	43956900	5649.2	980.8
Chang Jiang (Yangtze)	1800	900	810000	451.7	448.3
Huang Ho (Yellow)	445	50	2500	-889.8	939.8
Mackenzie	1805.2	11	121	456.8	-445.8
Mississippi	3226.3	620	384400	1863.7	-1243.7
Indus	1138.8	146	21316	-202.9	348.9
Nelson-Saskatchewan	1109.4	87	7569	-232.0	319.0
Total	19606.4	8768	45287782		

$$s^2 = \frac{\sum \hat{y}^2}{n} - \bar{y}^2 = \frac{39478877}{8} - 1201216 = 3733644$$

$$s_y^2 = \frac{\sum y^2}{n} - \bar{y}^2 = \frac{45287782}{8} - 1201216 = 4459757$$

$$r^2 = \frac{s^2}{s_y^2} = \frac{3733644}{4459757} = 0.83$$

Regression diagnostics



Regression diagnostics

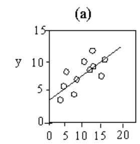


Figure (a) - is a reasonable description of y and x.

Regression diagnostics

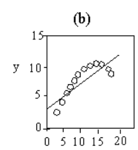
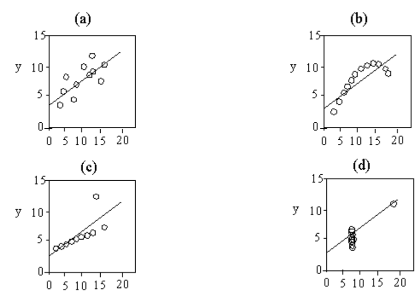


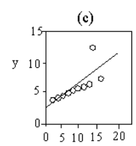
Figure (b) - is obviously curvilinear.

Regression diagnostics



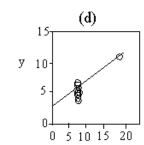
Regression diagnostics

Figure (c) - one data point has undue influence.

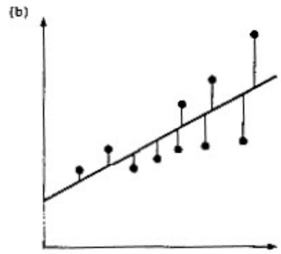


Regression diagnostics

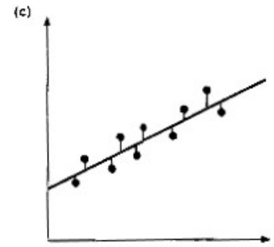
Figure (d) - can only fit a line to last point.



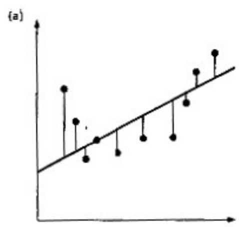
residuals tend to increase as we increase the value of x



there is no systematic variation apparent in the residuals



a curved line would be a better fit



autocorrelation

- test if there is no correlation between the absolute values of the residuals
  - this is known as <u>serial correlation</u> or <u>autocorrelation</u>
- the calculation for autocorrelation makes use of pairs of values
  - the first residual paired with the second, the second with the third and so on

absolute values of residuals

1st 7	last 7	a²	b²	ab
1347.1	980.8	1814632.6	961968.6	1321219.0
980.8	448.3	961968.6	200972.9	439692.6
448.3	939.8	200972.9	883130.1	421289.9
939.8	445.8	883130.1	198780.4	418985.7
445.8	1243.7	198780.4	1546881.7	554517.7
1243.7	348.9	1546881.7	121722.8	433924.9
348.9	319.0	121722.8	101757.2	111293.2
5754.4	4726.3	5728089.2	4015213.8	3700923.0

n=7

$$\bar{a} = \frac{5754.4}{7} = 822.06$$

$$\bar{b} = \frac{4726.3}{7} = 675.19$$

$$s_b = \sqrt{\sum \frac{b^2}{n} - \bar{b}^2} = \sqrt{\frac{4015213.8}{7} - (675.19)^2} = 343.1$$

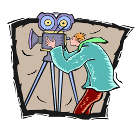
$$s_a = \sqrt{\sum \frac{a^2}{n} - \bar{a}^2} = \sqrt{\frac{5728089.2}{7} - (822.06)^2} = 377.5$$

$$r = \frac{\frac{\sum ab}{n} - \overline{a}\,\overline{b}}{s_a s_b} = \frac{\frac{3700923.0}{7} - 822.06(675.2)}{377.5(343.1)} = -.20$$

a value close to 1 or -1 suggests a relationship between successive residuals
a complete absence of a relationship would give a value of 0.0

- A **Confidence Interval** is the estimation of a **mean** response for a given X_i.
- **Confidence Bands** show an interval estimate for the entire regression line.
- The **Prediction Interval** is the prediction of a response of a single new observation of a given X_i.

- Video clip on confidence intervals
- Web video available at:
  - http://www.learner.org/resources/series65.html
  - Episode 19, Confidence Intervals



Confidence Interval

- How "wide" you have to cast your "net" to be sure of capturing the true population parameter.
  - I might say that my 95% Confidence Interval is plus or minus 2%, meaning that odds are 95 out of 100 hundred that the true population parameter is somewhere between 8 and 12%.

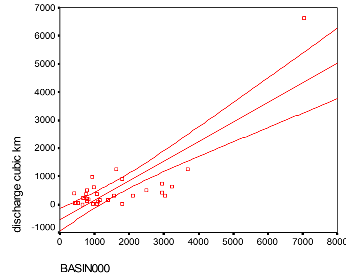
Confidence band

- Measurement of the certainty of the shape of the fitted regression line. A 95% confidence band implies a 95% chance that the true regression line fits within the confidence bands. It's a measurement of uncertainty.

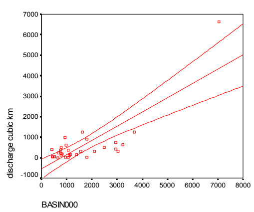
- The sample regression equation is an estimate of the population regression equation. Like any other estimate, there is an uncertainty associated with it.
- The uncertainty is expressed in confidence bands about the regression line. They have the same interpretation as the standard error of the mean, except that the uncertainty varies according to the location along the line.

- The uncertainty is least at the sample mean of the Xs and gets larger as the distance from the mean increases. The regression line is like a stick nailed to a wall with some wiggle to it.

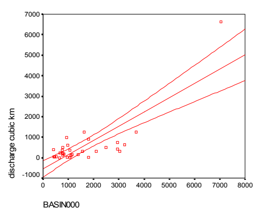
Confidence band for river discharge (95%)



Confidence band for river discharge



At 99%



At 95%

Confidence band formula

$$ts_e\sqrt{1+\frac{(x^*-\bar{x})^2}{\sum x^2-\frac{(\sum x)^2}{n}}}$$

S_e is standard error of the estimate
X* is the is the location along the X-axis where the distance is being calculated
The distance is smallest when x* = mean of x

Prediction Band (or Prediction Interval)

- Measurement of the certainty of the scatter about a certain regression line. A 95% prediction band indicates that, in general, 95% of the points will be contained within the bands.
- Used to estimate for a single value, not the mean of Y

Prediction interval for an individual y (based on existing data)

$$\hat{y}\pm t\sqrt{1+\frac{1}{n}+\frac{(x-\bar{x})^2}{\sum x^2-\frac{(\sum x)^2}{n}}}$$

t Distribution Critical Values

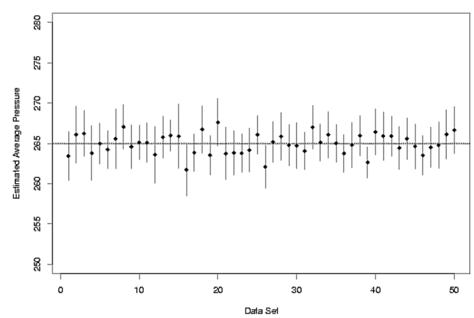
(Table entry is the point t^* with given probability p lying above it.)

df	.25	.20	.15	.10	.05	.025	.02	.01	.005	.0025	.001	.0005
1	1.000	1.376	1.963	3.078	6.314	12.71	15.89	31.82	63.66	127.3	318.3	636.6
2	.816	1.061	1.386	1.886	2.920	4.303	4.849	6.965	9.925	14.09	22.33	31.60
3	.765	.978	1.250	1.638	2.353	3.182	3.482	4.541	5.841	7.453	10.21	12.92
4	.741	.941	1.190	1.533	2.132	2.776	2.999	3.747	4.604	5.598	7.173	8.610
5	.727	.920	1.156	1.476	2.015	2.571	2.757	3.365	4.032	4.773	5.893	6.869
6	.718	.906	1.134	[illegible]	[illegible]	2.447	2.612	3.143	3.707	4.317	5.208	5.959
7	.711	.896	1.119	1.415	1.895	2.365	2.517	2.998	3.499	4.029	4.785	5.408
8	.706	.889	1.108	1.397	1.860	2.306	2.449	2.896	3.355	3.833	4.501	5.041
9	.703	.883	1.100	1.383	1.833	2.262	2.398	2.821	3.250	3.690	4.297	4.781
10	.700	.879	1.093	1.372	1.812	2.228	2.359	2.764	3.169	3.581	4.144	4.587
11	.697	.876	1.088	1.363	1.796	2.201	2.328	2.718	3.106	3.497	4.025	4.437
12	.695	.873	1.083	1.356	1.782	2.179	2.303	2.681	3.055	3.428	3.930	4.318
13	.694	.870	1.079	1.350	1.771	2.160	2.282	2.650	3.012	3.372	3.852	4.221
14	.692	.868	1.076	1.345	1.761	2.145	2.264	2.624	2.977	3.326	3.787	4.140
15	.691	.866	1.074	1.341	1.753	2.131	2.249	2.602	2.947	3.286	3.733	4.073
16	.690	.865	1.071	1.337	1.746	2.120	2.235	2.583	2.921	3.252	3.686	4.015

$$y \pm 2.45\sqrt{1+\tfrac{1}{8}+\frac{(445-2450.8)^2}{2450.8-\dfrac{384410921}{8}}} = y \pm 2.59$$

prediction band or prediction interval

- Prediction intervals estimate a random value where confidence limits estimate population parameters
- it is possible to establish limits within which predictions are made from regression equations
- when the regression equation is used to predict a confidence interval for the expected value can be calculated



Summary of prediction issues

- We cannot be certain of the mean of the distribution of Y.
- Prediction limits for $Y_{(new)}$ must take into account:
  - variation in the possible mean of the distribution of Y
  - variation in the responses Y within the probability distribution

Prediction interval for a new response

- the prediction interval is given by:

$$t\sqrt{\frac{\sum e^2}{n-1}\left[\frac{1}{n}+\frac{(x_o-\bar{x})^2}{\sum x^2-n\bar{x}^2}\right]}$$

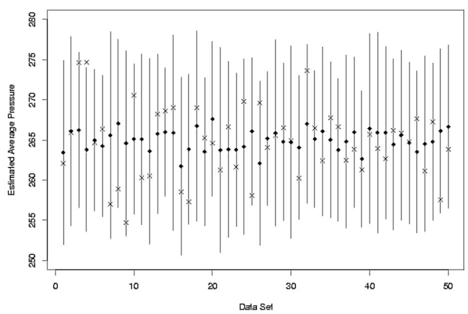
where Σe^2 is the sum of squares of the residuals from the regression

$\bar{x}$ is the mean of the values of the independent variable x

X_0 is the particular value for which $\hat{y}$ is being predicted

n is the number of pairs of measurements

t is the particular value taken from the t table



Implications on precision

- The greater the spread in the x values, the narrower the confidence interval, the more precise the prediction of $E(Y_o)$.
- Given the same set of x_i values, the further x_o is from the (sample) mean of the x, the wider the confidence interval, the less precise the prediction of $E(Y_o)$.

Comments on assumptions

- x_h is a value within scope of model, but it is not necessary that it is one of the x values in the data set.
- The confidence interval formula for $E(Y_h)$ works okay even if the error terms are only approximately normally distributed.
- If you have a large sample, the error terms can even deviate substantially from normality without greatly affecting appropriateness of the confidence interval.

- Confidence band most applicable for causal modeling
- Prediction interval most applicable for predictive uses

significance of b

- is the sample coefficient (an estimate) significantly different from the population coefficient
- b=0.99 is an estimate of the population parameter
- H_0: Y and X (in the population) are not related, i.e. b is not significantly different from 0
- H_1: Y and X (in the population) are related, b is significantly different than 0

T-test for b

$$t = \frac{b - B}{s.e.b} \quad where\ B = population\ parameter\ H_o: B = 0$$

$$so\ t = \frac{b - 0}{s.e.b}\ with\ df = n - 2$$

$$s.e.b = \frac{\sqrt{\dfrac{\sum_{i=1}^{n}(y_i - \hat{y})^2}{n-2}}}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}}$$

the denominator can be calculated via the formula

$$\sqrt{\sum x^2 - \frac{(\sum x)^2}{n}}$$

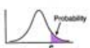
To give

$$s.e.b = \frac{\sqrt{\sum (y_i - \hat{y})^2 / n-2}}{\sqrt{\sum x^2 - \frac{(\sum x)^2}{n}}}$$

$$s.e.b = \frac{985.7}{5520.5} = 0.18$$

$$t = \frac{0.99}{0.18} = 5.5$$

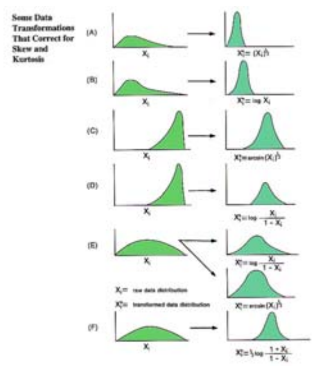
we can reject H_0, b is significantly different than 0



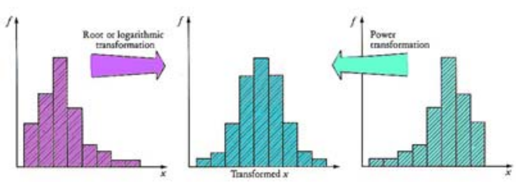
t Distribution Critical Values
(Table entry is the point t^* with given probability p lying above it.)

df	.25	.20	.15	.10	.05	.025	.02	.01	.005	.0025	.001	.0005
1	1.000	1.376	1.963	3.078	6.314	12.71	15.89	31.82	63.66	127.3	318.3	636.6
2	.816	1.061	1.386	1.886	2.920	4.303	4.849	6.965	9.925	14.09	22.33	31.60
3	.765	.978	1.250	1.638	2.353	3.182	3.482	4.541	5.841	7.453	10.21	12.92
4	.741	.941	1.190	1.533	2.132	2.776	2.999	3.747	4.604	5.598	7.173	8.610
5	.727	.920	1.156	1.476	2.015	2.571	2.757	3.365	4.032	4.773	5.893	6.869
6	.718	.906	1.134	1.440	[illegible]	2.447	2.612	3.143	3.707	4.317	5.208	5.959
7	.711	.896	1.119	1.415	[illegible]	2.365	2.517	2.998	3.499	4.029	4.785	5.408
8	.706	.889	1.108	1.397	1.860	2.306	2.449	2.896	3.355	3.833	4.501	5.041
9	.703	.883	1.100	1.383	1.833	2.262	2.398	2.821	3.250	3.690	4.297	4.781
10	.700	.879	1.093	1.372	1.812	2.228	2.359	2.764	3.169	3.581	4.144	4.587
11	.697	.876	1.088	1.363	1.796	2.201	2.328	2.718	3.106	3.497	4.025	4.437
12	.695	.873	1.083	1.356	1.782	2.179	2.303	2.681	3.055	3.428	3.930	4.318
13	.694	.870	1.079	1.350	1.771	2.160	2.282	2.650	3.012	3.372	3.852	4.221
14	.692	.868	1.076	1.345	1.761	2.145	2.264	2.624	2.977	3.326	3.787	4.140
15	.691	.866	1.074	1.341	1.753	2.131	2.249	2.602	2.947	3.286	3.733	4.073
16	.690	.865	1.071	1.337	1.746	2.120	2.235	2.583	2.921	3.252	3.686	4.015

Transforms



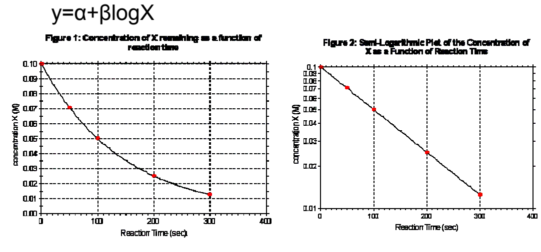
transforms



Data transformation.

nonlinear regression

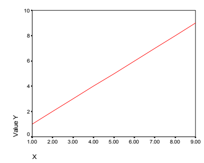
- semilog transform: y=α+βlogX

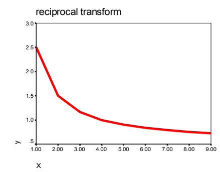


- double log
  - log y=α + βlogX or $Y=AX^B$
  - log y=α - βlogX or $Y=AX^{-B}$
    - α=0.5, β=2

x	y
2	2
4	8
6	18
8	32
10	50
12	72
14	98
16	128

- reciprocal transform
  - Y=α + β/X
  - Y=α - β/X
    - α=0.5, β=2





why transform?

- To approach normality in the data
  - a) if data is positively skewed (that is a long tail to the right)
    - a square root transform might be the answer
    - if more extreme, a logarithm transform might be necessary
    - if even more extreme higher roots might be necessary
  - b) if data is negatively skewed a power transform might be the answer

Confidence Intervals

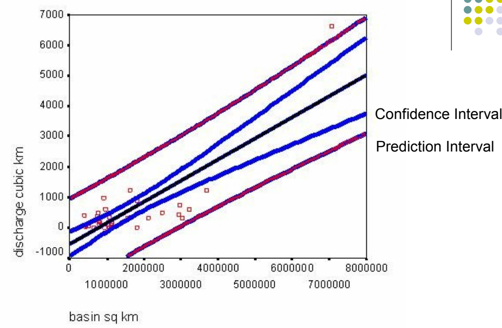
- If the correlation is perfect, the predictions are completely accurate; if the correlation is *not* perfect, what is our level of confidence in our prediction?
- Calculate the *Standard Error of the Estimate*, it expresses the degree of spread of the observations (y_i values) around the regression line in units of y.

$$s.e.y = \sqrt{\frac{\sum_{i=1}^{n}(y-\hat{y})^2}{n-2}} = \sqrt{\frac{5829846}{8-2}} = 985.7$$

68 % of the observations within 1 S.E.'s.
95 % of the observations within 2 S.E.'s.
y = a + bx + (2 SE) @ x = 2000
y = (-1330.3) + (0.99*2000) = 649.7 ± 985.7

- This expression is essentially an average error for the regression.
- The standard error of the estimate is useful in determining the range of potential Y values for a particular X value.

- 95 % of y values (19 out of 20) will lie within (-336 to 1635.4 range.
- This is high as this example, i.e. If you predicted a discharge at 2000 units of catchment area, the prediction would be 649.7, but could be *much* higher or lower.



Dummy variables

- $Y = \alpha + \beta D_i + u_i$
- How does this work if $D_i = \{0,1\}$?

- If hemisphere =n, dummy =1
- If hemisphere =s, dummy=0
- A simple regression using a dummy variable is similar to a one-way analysis of variance

River	basin sq km (x)	Discharge km³ (y)	d	hemisphere
Nile	3031700	324	0	S
Amazon	7050000	6630	0	S
Chang Jiang	1800000	900	1	N
Huang Ho	445000	50	1	N
Mackenzie	1805200	11	1	N
Mississippi	3226300	620	1	N
Indus	1138800	146	1	N
Nelson-Saskatchewan	1109400	87	1	N

Dummy variables

- $Y = \alpha + \beta D_i + e_i$
- How does this work if $D_i = \{0,1\}$?
  - When D = 0:
    - $Y = \alpha + \beta*0 + e_i$
    - $Y = \alpha + e_i$
  - When D = 1:
    - $Y = \alpha + \beta*1 + e_i$
    - $Y = (\alpha + \beta) + e_i$

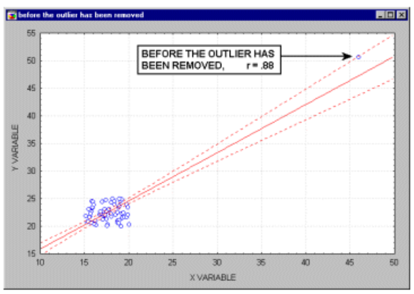
Dummy variables

- $Y = \alpha + \beta D_i + u_i$
- How does this work if $D_i = \{0,1\}$?
- Thus:
  - α is average value of Y when D_i=0
  - $\alpha + \beta$ is average value when D_i=1
  - Statistical significance of β is t-test for difference of means between the two categories

Outliers

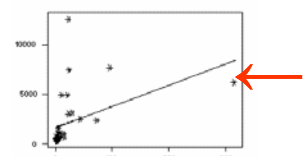
- In particular outliers (i.e., extreme cases) can seriously bias the results by "pulling" or "pushing" the regression line in a particular direction, thereby leading to biased regression coefficients.
- Often, excluding just a single extreme case can yield a completely different set of results.

Outliers



Influential observation

- If a point lies far from the other data in the horizontal direction, it is known as an ***influential observation***.
- Their removal may substantially change the regression equation



Lurking variables

- A ***lurking variable*** exists when the relationship between two variables is significantly affected by the presence of a third variable which has not been included in the modeling effort.
- Such a variable might be a factor of time (for example, the effect of political or economic cycles)

- Sometimes the lurking variable is a 'grouping' variable of sort. This is often examined by using a different plotting symbol to distinguish between the values of the third variables. For example, consider the following plot of the relationship between salary and years of experience for nurses.

  The individual lines show a positive relationship, but the overall pattern when the data are pooled, shows a negative relationship.

Lurking variable

Extrapolation

- Whenever a linear regression model is fit to a group of data, the range of the data should be carefully observed. Attempting to use a regression equation to predict values outside of this range is often inappropriate, and may yield incredible answers.
  - For example, a linear model which relates weight gain to age for young children. Applying such a model to adults, or even teenagers, would be absurd, since the relationship between age and weight gain is not consistent for all age groups.