

Overview of Statistical Concepts

Lecture 4
SIE 512 Spatial Analysis
Fall 2008

Linear Regression

Linear regression models

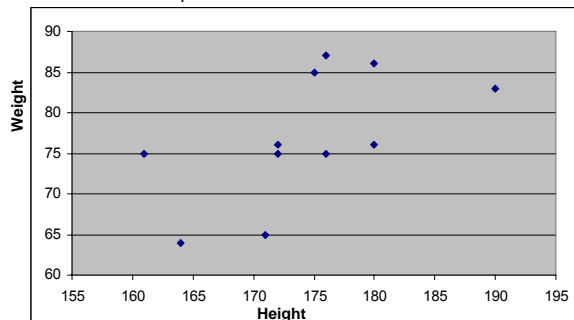
Models provide means to summarize observed relationships and are essential for making predictions and inferences.

- 1. Identification**
Evaluate data, start with simplest models first
- 2. Estimation**
Fit model to sample data to estimate parameters
- 3. Evaluation**
Assess fit – typically involves analysis of residuals
- 4. Prediction**
Test predictive quality of the model – an independent sample from that used to generate the model

Simple linear regression model

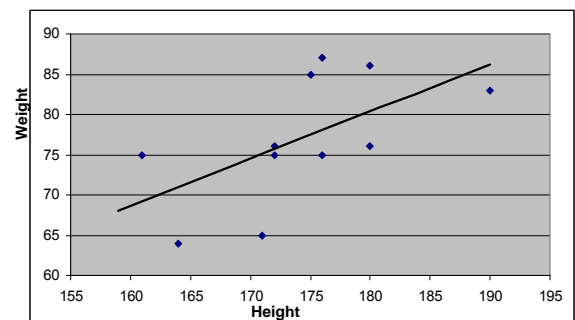
models the change in a variable compared to change in another

Start with a scatter plot



Simple linear regression model

A model of a response variable to changes in an explanatory variable $Y = aX + b$



Linear regression models

Start with an initial model which is then critically examined

Simplest model is best for first approximation

$$Y = \beta_0 + \beta_1 X + \varepsilon \quad \text{Linear first order}$$

$$Y = \beta_0 + \beta_1 X + \beta_{11} X^2 + \varepsilon \quad \text{Linear second order}$$

Y is the response (dependent) variable

X is the explanatory (independent) variable

β are parameters of the model ε are independent normally distributed residuals

Linear regression models

$$Y = \beta_0 + \beta_1 X + \varepsilon$$

Can be expressed as a probability distribution

$$Y \sim N(\beta_0 + \beta_1 X, \sigma_\varepsilon)$$

Y is normally distributed with a mean that is linearly dependent on X

Model Parameters

β_0 Y intercept

β_1 Slope

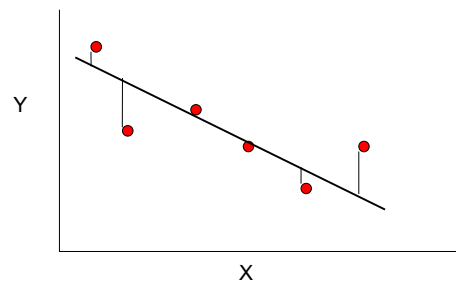
Estimated by least squares - minimize sum of squared residuals

$$SS = \sum_{i=1}^n \varepsilon_i^2 = \sum_{i=1}^n (y_i - \beta_0 - \beta_1 x_i)^2$$

Estimates $\hat{\beta}_0$ and $\hat{\beta}_1$ are values that produce the least possible value of SS

Model Parameter Estimation

The least squares method calculates the best-fitting line for the observed data by minimizing the sum of the squares of the vertical deviations from each data point to the line



Model Parameter Estimation

$$SS = \sum_{i=1}^n \epsilon_i^2 = \sum_{i=1}^n (y_i - \beta_0 - \beta_1 x_i)^2$$

Differentiate with respect to β_0 and β_1

$$\frac{\partial SS}{\partial \beta_0} = -2 \sum_{i=1}^n (y_i - \beta_0 - \beta_1 x_i) = 0$$

$$\frac{\partial SS}{\partial \beta_1} = -2 \sum_{i=1}^n (y_i - \beta_0 - \beta_1 x_i) x_i = 0$$

Model Parameter Estimation

$$\sum_{i=1}^n Y_i - nb_0 - b_1 \sum_{i=1}^n X_i = 0$$

$$\sum_{i=1}^n X_i Y_i - X_i b_0 - b_1 \sum_{i=1}^n X_i^2 = 0$$

$$nb_0 + b_1 \sum_{i=1}^n X_i = \sum_{i=1}^n Y_i$$

$$b_0 \sum_{i=1}^n X_i + b_1 \sum_{i=1}^n X_i^2 = \sum_{i=1}^n X_i Y_i$$

Normal equations

Model Parameter Estimation

$$b_1 = \frac{\sum_{i=1}^n X_i Y_i - [(\sum_{i=1}^n X_i)(\sum_{i=1}^n Y_i)]/n}{\sum_{i=1}^n X_i^2 - (\sum_{i=1}^n X_i)^2/n}$$

$$S_{XY} = \sum_{i=1}^n X_i Y_i - [(\sum_{i=1}^n X_i)(\sum_{i=1}^n Y_i)]/n \quad (\text{corrected sum of products of X and Y})$$

$$S_{XX} = \sum_{i=1}^n X_i^2 - (\sum_{i=1}^n X_i)^2/n \quad (\text{corrected sum of squares of X})$$

$$\hat{\beta}_1 = \frac{S_{XY}}{S_{XX}} \quad \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

The hat symbol ^ indicates estimated values

Regression Results

SUMMARY OUTPUT					
Regression Statistics					
Multiple R	0.59507354				
R Square	0.35411251				
Adjusted R Square	0.28234724				
Standard Error	6.60554978				
Observations	11				
ANOVA					
	df	SS	MS	F	Significance F
Regression	1	215.3004091	215.3004	4.934316	0.05344773
Residual	9	392.6995909	43.63329		
Total	10	608			
Coefficients					
	Coefficients	Standard Error	t Stat	P-value	Lower 95%
Intercept	-25.5163647	46.19380873	-0.55238	0.594141	-130.0141
Height	0.58825248	0.264819714	2.221332	0.053448	-0.01081179

Regression Results

$$Y = \beta_0 + \beta_1 X + \varepsilon$$

Regression equation

Weight = -25.5 + 0.588 Height

R^2 estimates the goodness of fit of the line. It is the percent of the response (dependent) variable explained by the explanatory (independent variables).

In this case 35.4 percent of variance in weight is explained by height

Regression Results

The more variables you have, the higher the amount of variance you can explain. Even if each variable doesn't explain much, adding a large number of variables can result in higher values of R^2 .

Adjusted R^2 is a standard downward adjustment to penalize for the possibility that with many explanatory variables some of the variance may be due to chance

$$R^2 = 1 - SSE / SST$$

$$R^2_{adj} = 1 - MSE / MST$$

$$Adjusted R^2 = 1 - (1 - R^2) \frac{n-1}{n-p-1}$$

Regression Results

The Standard Error of the Estimate

Is the square root of the Residual Mean Square.

It is the standard deviation of the data about the regression line, rather than about the sample mean.

$$\sqrt{\frac{\sum (y_i - \hat{y}_i)^2}{n-2}} \quad \text{Rather than} \quad \sqrt{\frac{\sum (y_i - \bar{y})^2}{n-1}}$$

Standard error of estimate

Standard error of sample mean

Significance of Parameters

The coefficients column reports the best estimates of the model parameters

The Standard Error column gives an estimate of the standard error of the coefficients

The t-stat column gives the ratio of the coefficient and its standard error

The p-value (probability value) gives the value for rejection of the null hypothesis that the parameter is zero

A p-value less than .05 means there is a 5 percent chance of finding data less consistent with the hypothesis

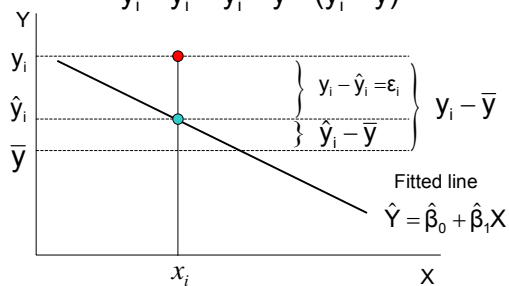
Significance of Parameters

The p_values measure the probability that the values for β are not derived by chance. The p_values are not a measure of 'goodness of fit' *per se*, rather they state the confidence that one can have in the estimated values being correct.

Precision of the Estimated Regression

$\epsilon_i = y_i - \hat{y}_i$ Residuals are the differences between the observed value and the predicted value

$$y_i - \hat{y}_i = y_i - \bar{y} - (\hat{y}_i - \bar{y})$$



Precision of the Estimated Regression

$$y_i - \hat{y}_i = (y_i - \bar{y}) - (\hat{y}_i - \bar{y})$$

$$(y_i - \bar{y}) = (\hat{y}_i - \bar{y}) + (y_i - \hat{y}_i)$$

Deviation of ith
observation
from overall
mean

Deviation of ith
predicted value
from the overall
mean

Deviation of ith
observation
from the ith
predicted value

$$\sum (y_i - \bar{y})^2 = \sum (\hat{y}_i - \bar{y})^2 + \sum (y_i - \hat{y}_i)^2$$

Sum of squares
of deviation of
observations
from the mean

Sum of squares
of deviation of
predicted
values from the
mean

Sum of squares of
deviation of
observations from
the predicted
values

Precision of the Estimated Regression

$$\sum (y_i - \bar{y})^2 = \sum (\hat{y}_i - \bar{y})^2 + \sum (y_i - \hat{y}_i)^2$$

Sum of squares about the mean Sum of squares due to regression Sum of squares about regression

Says some of the variation in y 's about the mean is due to the regression line and some to the residuals

Can assess usefulness of the regression line by how much SS about the mean is due to SS due to regression compared to SS about regression

SS due to regression should be greater than SS about regression

Analysis of Variance Table (ANOVA)

Source of variation	Degrees of freedom	Sum of Squares (SS)	Mean Square (MS)
Due to regression	1	$\sum_{i=1}^n (\hat{Y}_i - \bar{Y})^2$	MS_{reg}
About regression (residual)	n-2	$\sum_{i=1}^n (Y_i - \hat{Y}_i)^2$	$s^2 = \frac{SS}{n-2}$
Total, corrected for mean	n-1	$\sum_{i=1}^n (Y_i - \bar{Y})^2$	

Analysis of Variance Table (ANOVA)

F value is MS_{reg}/MS_{res}

A ratio of the variances

Can be used to test the significance of the fit

ANOVA	df	SS	MS	F	Significance F
Regression	1	220.3223684	220.3224	4.934316	0.053447734
Residual	9	401.8594498	44.65105		
Total	10	622.1818182			

Overall Goodness of Fit

The goodness of fit is most commonly expressed by R^2 and describes the proportion of total variation about the mean explained by the fitted trend.

$$R^2 = \frac{\sum_{i=1}^n (\hat{y}_i - \bar{y})^2}{\sum_{i=1}^n (y_i - \bar{y}_i)^2} \quad R^2 = SSR / SST$$

$$R^2 = 1 - SSE / SST$$

$R^2 = (SS \text{ due to regression}) / (\text{Total sum of squares corrected for the mean})$

Measure proportion of total variation about the mean explained by the regression

Residual diagnostics

Residuals contain information on the way in which the model fails to properly explain observed variation in Y

A linear model assumes the residuals are independent, normally distributed and have a constant variance – is this so?

Test residuals for:

Structure – should be identically distributed with no obvious outliers – randomly distributed about 0

Independence – residuals should be independent of one another – no runs of similar values in the plot - $cov(\varepsilon_i, \varepsilon_j) = 0$

Outliers – standardized residuals should not have values greater than 3

Residual diagnostics

Normality – should be normally distributed – check by plotting a histogram or normal probability plot

Linearity - the residuals should be independent of the fitted values

Residual Analysis

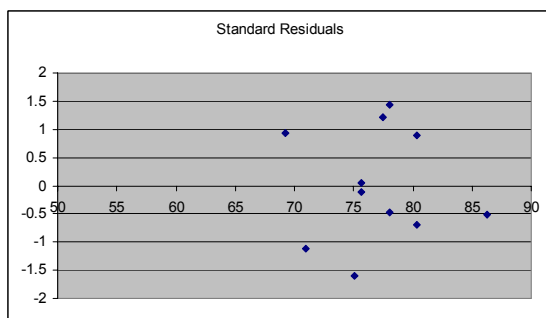
Graphical analysis of the residuals is the single most important technique for determining the need for model refinement or for verifying that the underlying assumptions of the analysis are met.

Residual plots of interest include:

- 1.residuals versus the explanatory variables
- 2.residuals versus the regression function (fitted) values
- 3.residual run order plot
- 4.residual lag plot
- 5.histogram of the residuals
- 6.normal probability plot

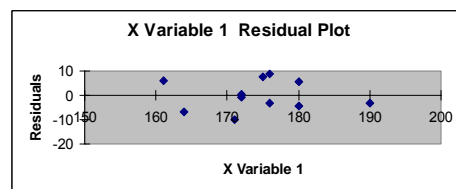
Residual diagnostics

Check for Outliers



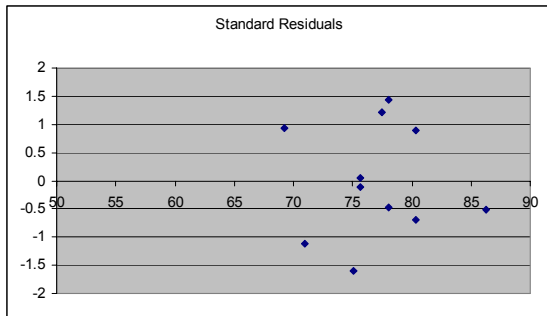
Residual diagnostics

Check against explanatory variable for structure



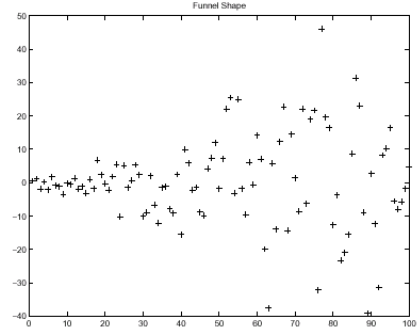
Residual diagnostics

Plot residuals against fitted values to check for constant variance



Residual diagnostics

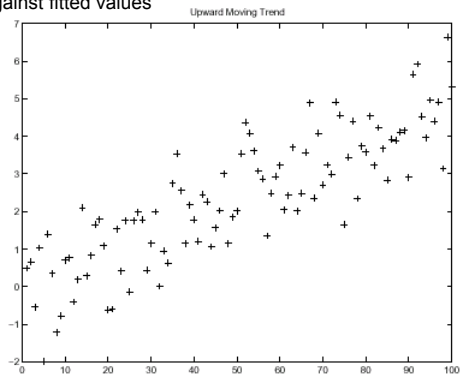
Residuals plotted against fitted values
Funnel shape indicates non-constant variance



Residual diagnostics

Residuals against fitted values

Plot is indicative of incorrect model

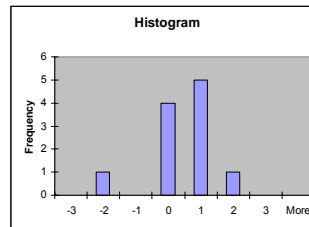


Residual diagnostics

Check normal distribution assumption

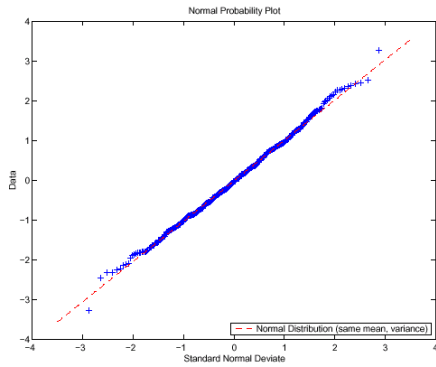
Histogram of the residuals

Normal probability plot of residuals

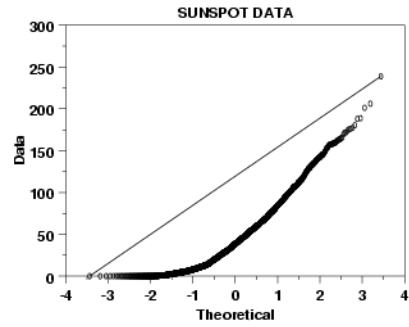


Residual diagnostics

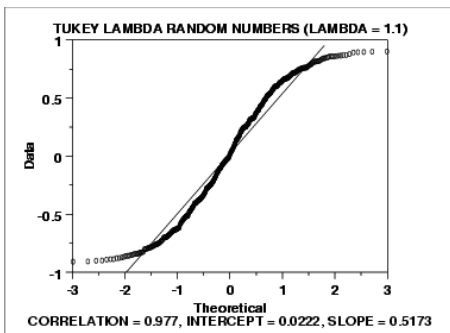
Normal probability plot of residuals



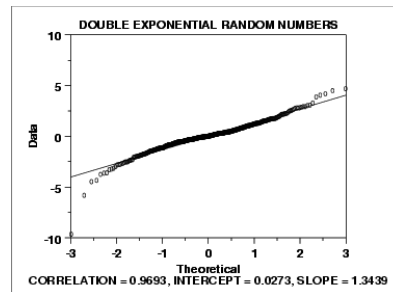
Normal Probability Plot: Data are Skewed Right



Normal Probability Plot: Data Have Short Tails



Normal Probability Plot: Data Have Long Tails



The double exponential distribution is symmetric, but relative to the normal it declines rapidly and has longer tails.







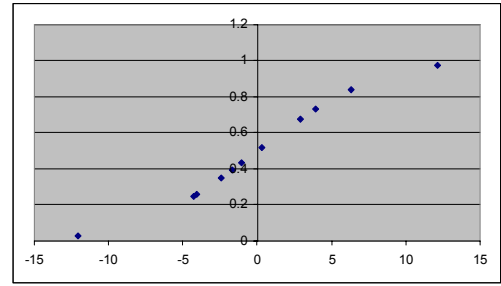
Pattern	Interpretation
	All but a few points fall on a line Outliers in the data
	Left end of the pattern is below the line while the right end of the pattern is above the line Symmetric, long tails at both ends
	Left end of the pattern is above the line while the right end of the pattern is below the line Symmetric, short tails at both ends
	Curved pattern with slope increasing from left to right Skewed to right
	Curved pattern with slope decreasing from left to right Skewed to left
	Staircase pattern Data have been rounded or may be discrete

Table 1: Plot Diagnostics

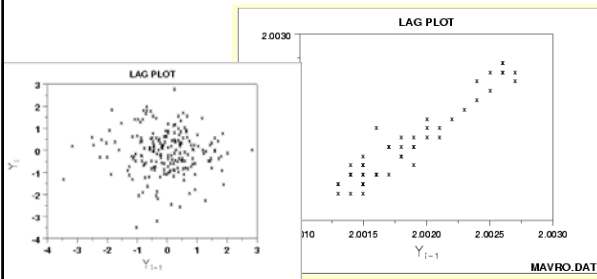
Residual diagnostics

Normal probability plot of residuals for height-weight regression



Residual diagnostics

Check for autocorrelation



This sample lag plot exhibits a linear pattern. This shows that the data are strongly non-random and further suggests that an autoregressive model might be appropriate.