Statistical Modelling in Stata 5: Linear Models

Mark Lunt

Centre for Epidemiology Versus Arthritis
University of Manchester

08/11/2022

Structure

- This Week
  - What is a linear model ?
  - How good is my model ?
  - Does a linear model fit this data ?
- Next Week
  - Categorical Variables
  - Interactions
  - Confounding
  - Other Considerations
    - Variable Selection
    - Polynomial Regression

Statistical Models

All models are wrong, but some are useful.

(G.E.P. Box)

A model should be as simple as possible, but no simpler. *(attr. Albert Einstein)*

What is a Linear Model ?

- Describes the relationship between variables
- Assumes that relationship can be described by straight lines
- Tells you the expected value of an *outcome* or *y* variable, given the values of one or more *predictor* or *x* variables

Variable Names

Outcome	Predictor
Dependent variable	Independent variables
Y-variable	x-variables
Response variable	Regressors
Output variable	Input variables
	Explanatory variables
	Carriers
	Covariates

CENTRE FOR EPIDEMIOLOGY VERSUS ARTHRITIS

The Equation of a Linear Model

The equation of a linear model, with outcome Y and predictors $x_1, \ldots x_p$

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p + \varepsilon$$

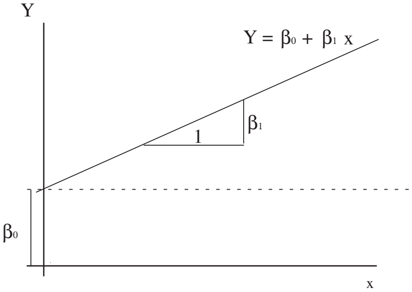
- $\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p$ is the *Linear Predictor*
- $\hat{Y} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p$ is the predictable part of Y.
- ε is the *error term*, the unpredictable part of Y.
- We assume that ε is normally distributed with mean 0 and variance σ^2.

CENTRE FOR EPIDEMIOLOGY VERSUS ARTHRITIS

Linear Model Assumptions

- Mean of $Y \mid x$ is a linear function of x
- Variables $Y_1, Y_2 \ldots Y_n$ are independent.
- The variance of $Y \mid x$ is constant.
- Distribution of $Y \mid x$ is normal.

Parameter Interpretation



- β_1 is the amount by which Y increases if x_1 increases by 1, and none of the other x variables change.
- β_0 is the value of Y when all of the x variables are equal to 0.

Estimating Parameters

- β_j in the previous equation are referred to as *parameters* or *coefficients*
- Don't use the expression "beta coefficients": it is ambiguous
- We need to obtain estimates of them from the data we have collected.
- Estimates normally given roman letters $b_0, b_1, \ldots, b_n$.
- Values given to b_j are those which minimise $\sum(Y - \hat{Y})^2$: hence "Least squares estimates"

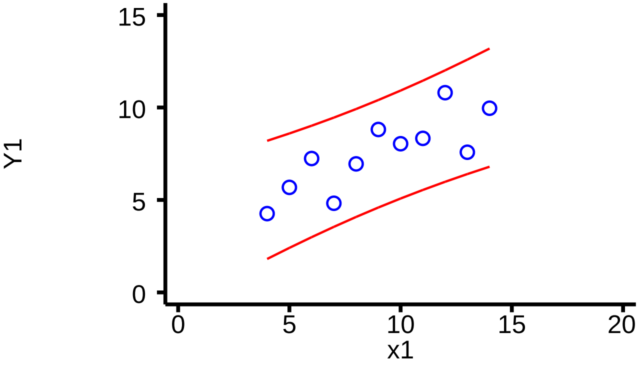
Inference on Parameters

- If assumptions hold, sampling distribution of b_j is normal with mean β_j and variance σ^2/ns_x^2 (for sufficiently large n), where :
  - σ^2 is the variance of the error terms ε,
  - s_x^2 is the variance of x_j and
  - n is the number of observations
- Can perform t-tests of hypotheses about β_j (e.g. $\beta_j = 0$).
- Can also produce a confidence interval for β_j.
- Inference in β_0 (intercept) is usually not interesting.

Inference on the Predicted Value

- $Y = \beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p + \varepsilon$
- Predicted Value $\hat{Y} = b_0 + b_1 x_1 + \ldots + b_p x_p$
- Observed values will differ from predicted values because of
  - Random error (ε)
  - Uncertainty about parameters β_j.
- We can calculate a 95% prediction interval, within which we would expect 95% of observations to lie.
- Reference Range for Y

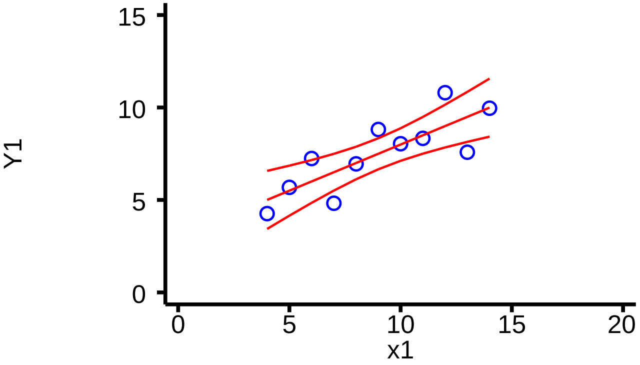
Prediction Interval



Inference on the Mean

- The *mean* value of Y at a given value of x does not depend on ε.
- The standard error of $\hat{Y}$ is called the standard error of the prediction (by stata).
- We can calculate a 95% confidence interval for $\hat{Y}$.
- This can be thought of as a confidence region for the regression line.

Confidence Interval



Analysis of Variance (ANOVA)

- Variance of Y is $\frac{\sum\left(Y-\bar{Y}\right)^2}{n-1} = \frac{\sum\left(Y-\hat{Y}\right)^2+\sum\left(\hat{Y}-\bar{Y}\right)^2}{n-1}$

CENTRE FOR EPIDEMIOLOGY VERSUS ARTHRITIS

Analysis of Variance (ANOVA)

- Variance of Y is $\frac{\sum(Y-\bar{Y})^2}{n-1} = \frac{\sum(Y-\hat{Y})^2+\sum(\hat{Y}-\bar{Y})^2}{n-1}$
- $SS_{reg} = \sum\left(\hat{Y} - \bar{Y}\right)^2$ (regression sum of squares)

Analysis of Variance (ANOVA)

- Variance of Y is $\frac{\sum(Y-\bar{Y})^2}{n-1} = \frac{\sum(Y-\hat{Y})^2+\sum(\hat{Y}-\bar{Y})^2}{n-1}$
- $SS_{reg} = \sum\left(\hat{Y} - \bar{Y}\right)^2$ (regression sum of squares)
- $SS_{res} = \sum\left(Y - \hat{Y}\right)^2$ (residual sum of squares)

Analysis of Variance (ANOVA)

- Variance of Y is $\frac{\sum(Y-\bar{Y})^2}{n-1} = \frac{\sum(Y-\hat{Y})^2+\sum(\hat{Y}-\bar{Y})^2}{n-1}$
- $SS_{reg} = \sum\left(\hat{Y} - \bar{Y}\right)^2$ (regression sum of squares)
- $SS_{res} = \sum\left(Y - \hat{Y}\right)^2$ (residual sum of squares)
- Each part has associated *degrees of freedom*: p d.f for the regression, $n - p - 1$ for the residual.

CENTRE FOR EPIDEMIOLOGY VERSUS ARTHRITIS

Analysis of Variance (ANOVA)

- Variance of Y is $\frac{\sum(Y-\bar{Y})^2}{n-1} = \frac{\sum(Y-\hat{Y})^2 + \sum(\hat{Y}-\bar{Y})^2}{n-1}$
- $SS_{reg} = \sum\left(\hat{Y} - \bar{Y}\right)^2$ (regression sum of squares)
- $SS_{res} = \sum\left(Y - \hat{Y}\right)^2$ (residual sum of squares)
- Each part has associated *degrees of freedom*: p d.f for the regression, $n - p - 1$ for the residual.
- The *mean square* $MS = SS/df$.

Analysis of Variance (ANOVA)

- Variance of Y is $\frac{\sum(Y-\bar{Y})^2}{n-1} = \frac{\sum(Y-\hat{Y})^2+\sum(\hat{Y}-\bar{Y})^2}{n-1}$
- $SS_{reg} = \sum\left(\hat{Y} - \bar{Y}\right)^2$ (regression sum of squares)
- $SS_{res} = \sum\left(Y - \hat{Y}\right)^2$ (residual sum of squares)
- Each part has associated *degrees of freedom*: p d.f for the regression, $n - p - 1$ for the residual.
- The *mean square* $MS = SS/df$.
- MS_{reg} should be similar to MS_{res} if no association between Y and x

Analysis of Variance (ANOVA)

- Variance of Y is $\frac{\sum(Y-\bar{Y})^2}{n-1} = \frac{\sum(Y-\hat{Y})^2+\sum(\hat{Y}-\bar{Y})^2}{n-1}$
- $SS_{reg} = \sum\left(\hat{Y} - \bar{Y}\right)^2$ (regression sum of squares)
- $SS_{res} = \sum\left(Y - \hat{Y}\right)^2$ (residual sum of squares)
- Each part has associated *degrees of freedom*: p d.f for the regression, $n - p - 1$ for the residual.
- The *mean square* $MS = SS/df$.
- MS_{reg} should be similar to MS_{res} if no association between Y and x
- $F = \frac{MS_{reg}}{MS_{res}}$ gives a measure of the strength of the association between Y and x.

CENTRE FOR EPIDEMIOLOGY VERSUS ARTHRITIS

ANOVA Table

Source	df	Sum of Squares	Mean Square	F
Regression	p	SS_{reg}	$MS_{reg} = \frac{SS_{reg}}{p}$	$\frac{MS_{reg}}{MS_{res}}$
Residual	n-p-1	SS_{res}	$MS_{res} = \frac{SS_{res}}{(n-p-1)}$	
Total	n-1	SS_{tot}	$MS_{tot} = \frac{SS_{tot}}{(n-1)}$	

Goodness of Fit

- Predictive value of a model depends on how much of the variance can be explained.
- R^2 is the proportion of the variance explained by the model
- $R^2 = \frac{SS_{reg}}{SS_{tot}}$
- R^2 always increases when a predictor variable is added
- Adjusted R^2 is better for comparing models.

Stata Commands for Linear Models

- The basic command for linear regression is `regress` *y-var x-vars*
- Can use *by* and *if* to select subgroups.
- The command `predict` can produce
  - predicted values
  - standard errors
  - residuals
  - etc.

CENTRE FOR EPIDEMIOLOGY VERSUS ARTHRITIS

Stata Output 1: ANOVA Table

F()	F Statistic for the Hypothesis $\beta_j = 0$ for all j
Prob $>$ F	p-value for above hypothesis test
R-squared	Proportion of variance explained by regression $= \frac{SS_{Model}}{SS_{Total}}$
Adj R-squared	$\frac{(n-1)R^2-p}{n-p-1}$
Root MSE	$\sqrt{MS_{Residual}}$ $= \hat{\sigma}$

CENTRE FOR EPIDEMIOLOGY VERSUS ARTHRITIS

Stata Output 1: Example

```
   Source |       SS       df       MS              Number of obs =      11
----------+------------------------------           F(  1,     9) =   17.99
    Model |  27.5100011     1  27.5100011           Prob > F      =  0.0022
 Residual |  13.7626904     9  1.52918783           R-squared     =  0.6665
----------+------------------------------           Adj R-squared =  0.6295
    Total |  41.2726916    10  4.12726916           Root MSE      =  1.2366
```

CENTRE FOR EPIDEMIOLOGY VERSUS ARTHRITIS

Stata Output 2: Coefficients

Coef. Estimate of parameter β for the variable in the left-hand column. (β_0 is labelled "_cons" for "constant")

Std. Err. Standard error of *b*.

t The value of $\frac{b-0}{s.e.(b)}$, to test the hypothesis that $\beta = 0$.

$P > |t|$ P-value resulting from the above hypothesis test.

95% Conf. Interval A 95% confidence interval for β.

Stata Output 2: Example

```
------------------------------------------------------------------------------
       Y |      Coef.   Std. Err.       t     P>|t|       [95% Conf. Interval]
---------+--------------------------------------------------------------------
       x |   .5000909   .1179055      4.241   0.002       .2333701    .7668117
   _cons |   3.000091   1.124747      2.667   0.026       .4557369    5.544445
------------------------------------------------------------------------------
```

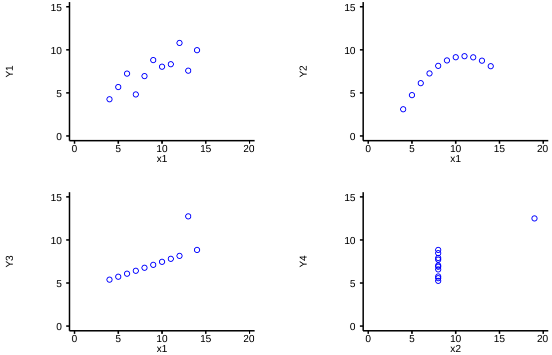
Is a linear model appropriate ?

- Does it provide adequate predictions ?
  - Goodness of fit or RMSE
- Do my data satisfy the assumptions of the linear model ?
- Are there any individual points having an inordinate influence on the model ?

Is a linear model appropriate ?

- Does it provide adequate predictions ?
  - Goodness of fit or RMSE
  - Not a statistical question: how close is “adequate”
- Do my data satisfy the assumptions of the linear model ?
- Are there any individual points having an inordinate influence on the model ?

Anscombe's Data



Linear Model Assumptions

- Linear models are based on 4 assumptions
  - Variables $Y_1, Y_2 \ldots Y_n$ are independent.
  - The variance of $Y_i \mid x$ is constant.
  - Mean of Y_i is a linear function of x_i.
  - Distribution of $Y_i \mid x$ is normal.
- If any of these are incorrect, inference from regression model is unreliable
- We may know about assumptions from experimental design (e.g. repeated measures on an individual are unlikely to be independent).
- Should test all 4 assumptions.

CENTRE FOR EPIDEMIOLOGY VERSUS ARTHRITIS

Distribution of Residuals

- Error term $\varepsilon_i = Y_i - \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \ldots + \beta_p x_{pi}$
- Residual term
  $e_i = Y_i - b_0 + b_1 x_{1i} + b_2 x_{2i} + \ldots + b_p x_{pi} = Y_i - \hat{Y}_i$
- Nearly but not quite the same, since our estimates of β_j are imperfect.
- $\hat{Y}$ varies more at extremes of x-range
- Y does not
- Hence residuals vary less at extremes of the x-range
- If error terms have constant variance, residuals don't.

Standardised Residuals

- Variation in variance of residuals as x changes is predictable.
- Can therefore correct for it.
- *Standardised Residuals* have mean 0 and standard deviation 1.
- Can use standardised residuals to test assumptions of linear model
- `predict Yhat, xb` will generate predicted values
- `predict sres, rstand` will generate standardised residuals
- `scatter sres Yhat` will produce a plot of the standardised residuals against the fitted values.

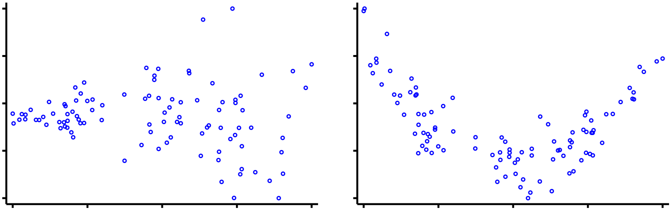
CENTRE FOR EPIDEMIOLOGY VERSUS ARTHRITIS

Testing Constant Variance:

- Residuals should be independent of predicted values
- There should be no pattern in this plot
- Common patterns
  - Spread of residuals increases with fitted values
    - This is called heteroskedasticity
    - May be removed by transforming Y
    - Can be formally tested for with `hettest`
  - There is curvature
    - The association between x and Y variables is not linear
    - May need to transform Y or x
    - Alternatively, fit x^2, x^3 etc. terms
    - Can be formally tested for with `ovtest`

CENTRE FOR EPIDEMIOLOGY VERSUS ARTHRITIS

Residual vs Fitted Value Plot Examples

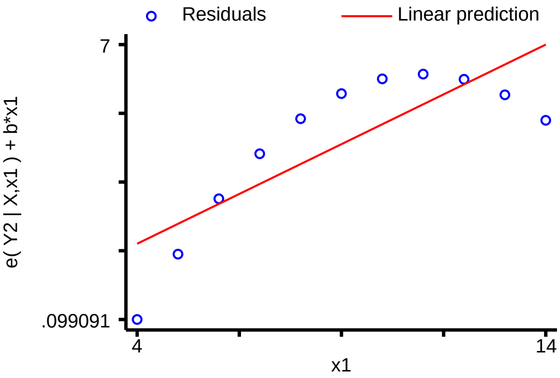


(a) Non-constant variance (b) Non-linear association

Testing Linearity: Partial Residual Plots

- Partial residual $p_j = e + b_j x_j = Y - \beta_0 - \sum_{l \neq j} b_l x_l$
- Formed by subtracting that part of the predicted value that does not depend on x_j from the observed value of Y.
- Plot of p_j against x_j shows the association between Y and x_j after adjusting for the other predictors.
- Can be obtained from stata by typing `cprplot` *`xvar`* after performing a regression.

Example Partial Residual Plot



CENTRE FOR EPIDEMIOLOGY VERSUS ARTHRITIS

Identifying Outliers

- Points which have a marked effect on the regression equation are called *influential* points.
- Points with unusual x-values are said to have high leverage.
- Points with high leverage may or may not be influential, depending on their Y values.
- Plot of *studentised residual* (residual from regression excluding that point) against leverage can show influential points.

Statistics to Identify Influential Points

DFBETA Measures influence of individual point on a single coefficient β_j.

DFFITS Measures influence of an individual point on its predicted value.

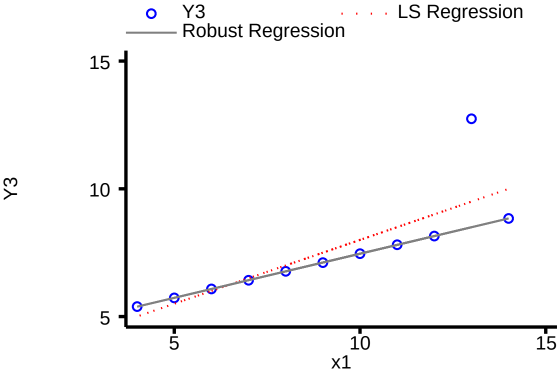
Cook's Distance Measured the influence of an individual point on *all* predicted values.

- All can be produced by `predict`.
- There are suggested cut-offs to determine influential observations.
- May be better to simply look for outliers.

Y-outliers

- A point with normal x-values and abnormal Y-value may be influential.
- Robust regression can be used in this case.
  - Observations repeatedly reweighted, weight decreases as magnitude of residual increases
- Methods robust to x-outliers are very computationally intensive.

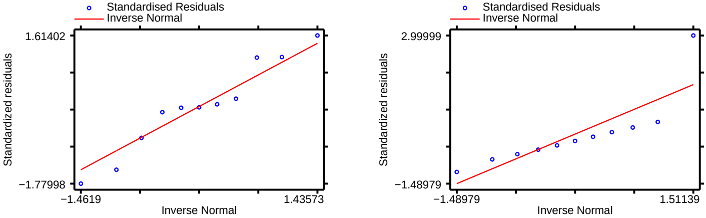
Robust Regression



Testing Normality

- Standardised residuals should follow a normal distribution.
- Can test formally with `swilk` *`varname`*.
- Can test graphically with `qnorm` *`varname`*.

Normal Plot: Example



CENTRE FOR EPIDEMIOLOGY VERSUS ARTHRITIS

Graphical Assessment & Formal Testing

- Can test assumptions both formally and informally
- Both approaches have advantages and disadvantages
  - Tests are *always* significant in sufficiently large samples.
  - Differences may be slight and unimportant.
  - Differences may be marked but non-significant in small samples.
- Best to use both