# aic_model_selection: Model Selection with AIC

Matthias Schonlau, Ph.D.

University of Waterloo

# Summary

- Introduce a new stata command, *aic_model_selection*, for forward model selection

# Model selection

- When a model has many variables, it is often harder to interpret
- Many of the variables may just represent “noise”
- You may prefer a smaller, parsimonious model that is more interpretable

# All possible subset regression

- The gold standard for model selection is “all possible subsets”
  - looks at all possible combinations of models
- The “best model” is chosen based on a criterion
  - Adjusted $R^2$, AIC, …
- There are $2^p$ models, where p is the number of x-variables
- Computing these becomes quickly unfeasible
  - Stata’s user-contributed command “*allpossible*” allows up to p=6

# Forward/stepwise/backward selection

- To reduce the model search space, sequential algorithms add/remove one variable at a time
  - Forward selection adds the best variable (given the model so far)
  - Backward selection removes the worst variable
- As stopping criterion usually a p-value is used , e.g. $p<0.05$.
- However, the p-values are not correct because they don’t take selection into account.
- Even though the stopping criterion is wrong/not interpretable, the sequence of models is sensible

# Forward selection with AIC

- One solution:
  - keep the sequence of models generated
  - Compute the AIC for each model
  - Among the sequence of models generated, choose the model with the lowest AIC
- E.g. James et al, section 6.1, “An introduction to statistical learning”

- *aic_model_selection* facilitates this approach for Stata

# Munich housing data

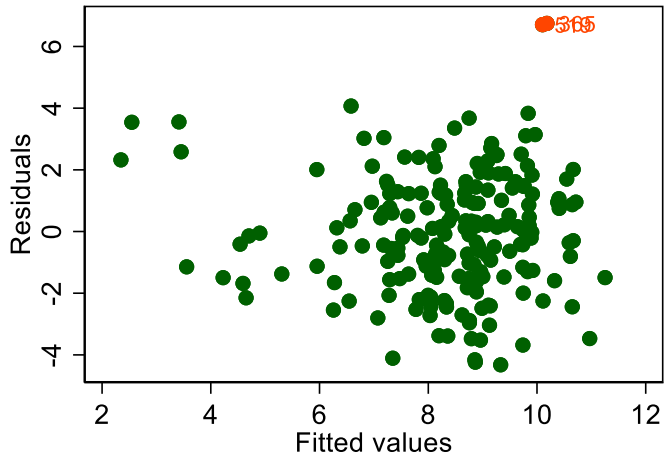
- Apartment rental cost in Munich

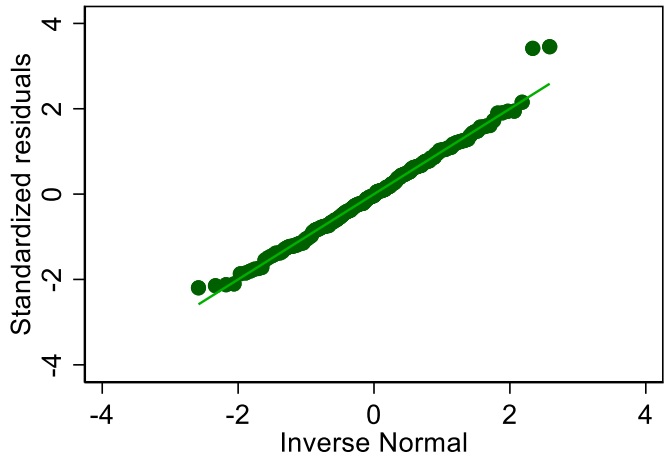
- Data believed to be from Fahrmeyer's book
  - Multivariate Statistical Modelling Based on Generalized Linear Models (Springer Series in Statistics) by Ludwig Fahrmeir, Gerhard Tutz, 2001

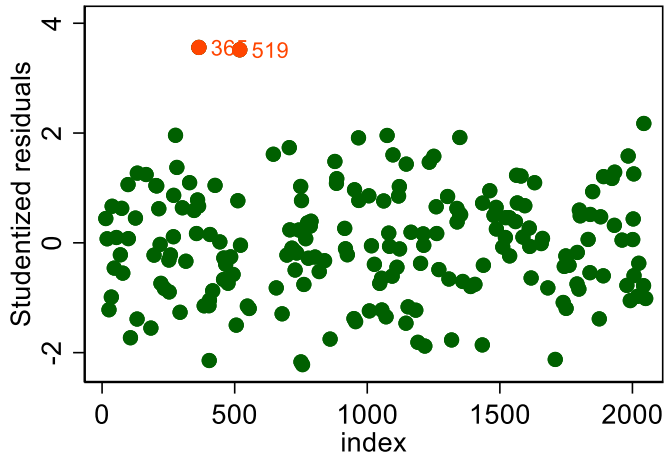
# Munich housing data

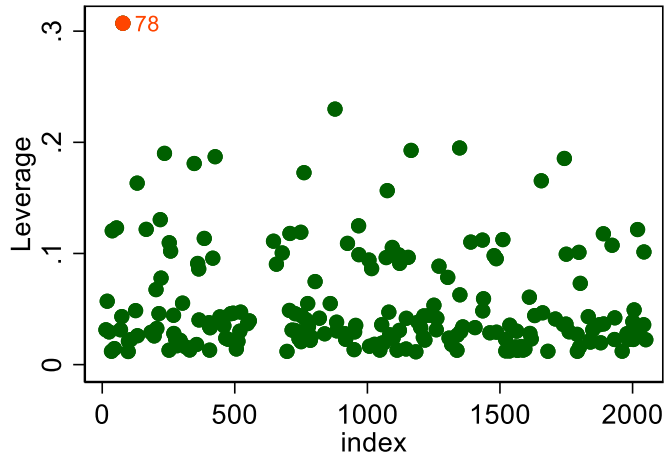
```
.  describe numrooms age size best good extrabath tiled warmwater central

              storage   display    value
variable name   type    format     label      variable label
------------------------------------------------------------------------------
numrooms        byte    %8.0g                 Number of rooms
age             float   %9.0g                 Age of apartment
size            int     %8.0g                 size in square meters
bestneighborh~_ float   %9.0g      yesno      Indicator: Great neighborhood
goodneighborh~_ float   %9.0g      yesno      Indicator: Good neighborhood
extrabath_      float   %9.0g      yesno      More than one bathroom
tiledbath_      float   %9.0g      yesno
warmwater_      float   %9.0g      yesno
centralheating_ float   %9.0g      yesno
```

# Diagnostic plots




Diagnostic plots can identify violations in assumptions.

Upper left: constant variance is ok
Upper right: Normality is ok (2 outliers)
Lower left: Almost all values are with +-2 std deviations. There are two points that may be too high (outliers).
Lower right: One point has a high leverage (unusual x-values).

# Forward selection

- Pe(.8) was chosen quite generously to make sure I won't miss a good AIC model.
- Forward selection with p=0.05 would stop after *extrabath*
- Forward selection with p=0.10 would stop after *warmwater*

```
. sw , pe(.8) : regress rent numrooms age size best good extrabath tiled warmwater central

begin with empty model
p = 0.0000 <  0.8000  adding  age
p = 0.0003 <  0.8000  adding  size
p = 0.0004 <  0.8000  adding  centralheating_
p = 0.0013 <  0.8000  adding  extrabath_
p = 0.0748 <  0.8000  adding  goodneighborhood_
p = 0.0423 <  0.8000  adding  bestneighborhood_
p = 0.0914 <  0.8000  adding  warmwater_
p = 0.6716 <  0.8000  adding  numrooms
p = 0.6938 <  0.8000  adding  tiledbath_
```

# Forward Model Selection based on AIC

- *aic_model_selection:* Specify variables in the order entered by forward selection

```
. aic_model_selection regress age size central extrabath good  best warmwater numrooms tiled
> bath
      AIC Model
 1847.355 age size
 1811.075 age size central
 1812.553 age size central extrabath
 1808.399 age size central extrabath good
 1810.272 age size central extrabath good best
 1811.206 age size central extrabath good best warmwater
 1810.016 age size central extrabath good best warmwater numrooms
 1810.838 age size central extrabath good best warmwater numrooms tiledbath
```

Forward selection with p<0.05

Lowest AIC model

Forward selection with p<0.10

- AIC finds a different subset than traditional forward selection

# Conclusion

- Because p-values in forward selection are wrong, it is not clear when to stop
- Using AIC solves this problem
- While the model selection procedure is preferred, there is nothing magic about the resulting model

- Software:

*ssc install aic_model_selection*