Problems with Predictors

ME104: Linear Regression Analysis
Kenneth Benoit

August 17, 2012

Quadratic $\beta_1 X + \beta_2 X^2$ v. $\beta \log(X)$

- Quadratic allows change in relationship (parabolas), whereas logarithmic transformation is monotone
- Log transformations are for capturing multiplicative effects of increases
- May be very similar in some contexts

Model selection and evaluation

- Using a fitted regression model, we can
  - **Interpret** the implications of the model using estimated regression coefficients, their confidence intervals and fitted values
  - Use the model to **predict** future values of the response
- However, both of these are likely to be misleading if the model is not (approximately) correct, i.e. if it is **misspecified**
- Need to have tools for evaluating and comparing models, in order to identify correctly specified ones

Tasks of model evaluation

- Finding a model with correct specification for the expected value $E(Y)$ of the response
  - i.e. selecting an appropriate set of explanatory variables
- Examining the adequacy of the other model assumptions: homoscedasticity and normality of error terms, and independence of observations
  - ... and ways of improving the model if these are not satisfied

Model selection

- Suppose we start with a set of potential explanatory variables $X_1, X_2, \ldots, X_K$ for a response Y
  - These also include any interaction (product) variables and nonlinear transformations we want to consider
- The aim of selection of explanatory variables is to identify a model which
  - includes all the variables which need to be included
  - leaves out all the variables which do not need to be included
- Here the decisions are made using significance testing:
  - All the variables in the selected model should be significant (at a stated significance level α)
  - None of the omitted variables should be significant (at level α) if they were included

General principles for specification

- *Theory* is our best guide
- If the residuals from a model are not sigificantly different from what might have occurred by chance, then conclude that the model is "mis-specified" (that nothing is going on)
- Tests for misspecification are OK when used judiciously
- We can set aside a subset of observations to be used for testing by making out-of-sample predictions
- Some authors advocate reporting the results of other specifications (a form of "sensitivity analysis") although this is done rarely, if at all in social science statistics

Common tests for misspecification

- Tests for omitted variables. This include F tests and t tests for whether coefficients are individually or jointly zero
- RESET: Regression specification error tests. Tests whether unknown variables have been omitted from a regression specification
- Tests for functional form. These include tests for recursive residuals, the rainbow test, and others (below)
- Tests for structural change. To test whether parameters change, such as the Chow test, cumsum, and cumsum-of-squares tests

Common tests for misspecification (continued)

- Tests for outliers. Cook outlier tests for instance, although there are many others
- Tests for non-spherical errors. Example: Durbin-Watson test
- Tests for exogeneity. Hausman tests.
- Others (see Kennedy)

Correlations of explanatory variables

- Multiple regression models estimate partial effects of each explanatory variable, allowing for correlations between these variables
- However, these correlations also cause some apparent complications in analysis and model selection:
  - Estimated coefficient of a variable depends on what other variables are in the model (as it should)
  - Results of tests and confidence intervals depend on what other variables are in the model
  - Conclusions for model selection may thus depend on the *order* in which variables were added to the model
- This is not the case if the explanatory variables are uncorrelated, but that is rarely true
- Particular problems if some explanatory variables are *very* strongly correlated (see notes at the end of these slides)

Sequential testing

- Such a model can be found using a series of significance tests
  - Usual t or F tests of the coefficients, all using the same significance level (e.g. 5%)
- Two basic versions are:
  - Forward selection: start with a model with no explanatory variables, and add new ones one at a time, until none of the omitted ones are significant
  - Backward selection: start with a model with all the variables included, and remove nonsignificant ones, one at a time, until the remaining ones are significant
- But always better to start with theory – what follows applies only if you are doing truly exploratory work

Example from HIE data

- Response variable: General Health Index at entry, $n = 1113$
- Potential explanatory variables: sex (dummy for men), age, log of family income, weight, blood pressure and smoking (as two dummy variables, for current and ex smokers)
  - A haphazard collection of variables with no theoretical motivation, purely for illustration of the stepwise procedure
  - For simplicity, no interactions or nonlinear effects considered
- F-tests are used for the smoking variable (with two dummies), t-tests for the rest
- Start backwards, i.e. from a full model with all candidate variables included

Example from HIE data

1. In the full model, Blood pressure ($P = 0.97$), Smoking ($P = 0.29$) and Sex ($P = 0.18$) are not significant at the 5% level
   - Remove Blood pressure
2. Now smoking is significant ($p < 0.05$) although Sex ($P = 0.17$) still not significant
3. In this model, Sex ($p = 0.21$) is the only nonsignificant variable, so remove it
4. If added to this model, Blood pressure is not be significant ($p = 0.90$), so it can stay out

Example from HIE data

- So the final model includes Age, Log-income and Weight, all of which are significant at the 5% level
- Here the nonsignificant variables were clear and unchanging throughout, but this is definitely not always the case
- Example was smoking variable in this case

Comments and caveats on stepwise model selection

- Often some variables are central to the research hypothesis, and treated differently from other control variables
  - e.g. in the Health Insurance Experiment, the insurance plan was the variable of main interest
  - Such variables are not dropped during a stepwise search, but tested separately at the end
- Variables are added or removed one at a time, not several at once
  - For categorical variables with more than two categories, this means adding or dropping all the corresponding dummy variables at once
  - Individual dummy variables (i.e. differences between particular categories) may be tested separately (e.g. at the end)

Comments and caveats on stepwise model selection

- The models should always be hierarchical:
  - if an interaction (e.g. coefficient of $X_1 X_2$) is significant, main effects (X_1 and X_2) may not be dropped
  - if coefficient of X^2 is significant, X may not be dropped
- In practice, the possible interactions and nonlinear terms are often not all considered in model selection
- Not guaranteed to find a single "best" model, because it may not exist: there may be several models satisfying the conditions stated earlier
- Theoretically motivated models are always better, when theory is available

Example from Computer class 4

Only P-values shown:

Response variable: Measure of fear of crime		
Variable		
Age	0.462	0.012
Female	< 0.001	< 0.001
Age $\times$ Female	0.358	< 0.001
Age^2	0.097	< 0.001
Age$^2 \times$ Female	0.225	—

Diagnostics from sample residuals

- Another key tool of assessment of linear models are the sample residuals
$$e_i = Y_i - \hat{Y}_i$$
for all observations $i = 1, \ldots, n$ in the sample, where $\hat{Y}_i$ are the fitted values
- "Estimates" of the error terms (model residuals) ϵ_i
- We will actually use "studentised" residuals: e_i standardised to have standard deviation of 1
- Can be used for diagnostics: examination of the assumptions of the model
- Here, in particular, examination of the assumption of homoscedasticity that the residual standard deviation σ (conditional standard deviation of Y) is the same at all values of the Xs

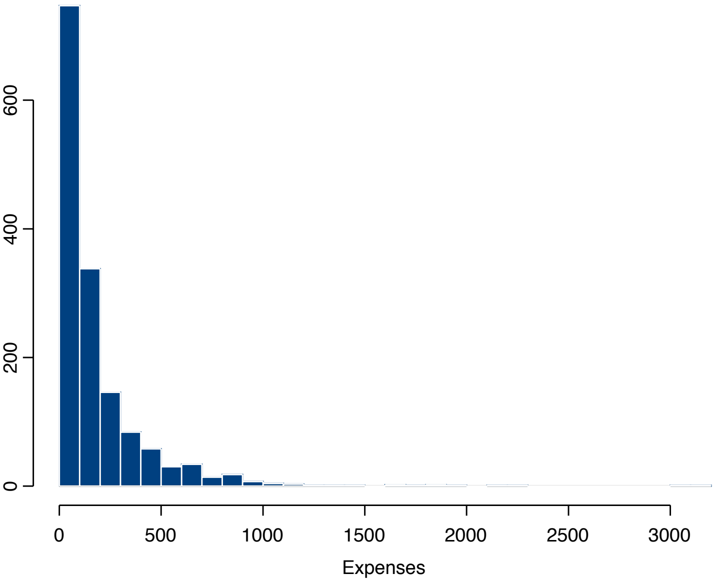
Residual plots

- Homoscedasticity may be examined using a plot of
  - residuals e_i (on the Y-axis) against fitted values $\hat{Y}_i$ (on the X-axis)
- This plot should show roughly equal level of variation of the residuals for all values of $\hat{Y}_i$
- A plot with a funnel shape (variability of residuals increasing or decreasing as $\hat{Y}_i$ increase) indicates heteroscedasticity (i.e. failure of homoscedasticity)

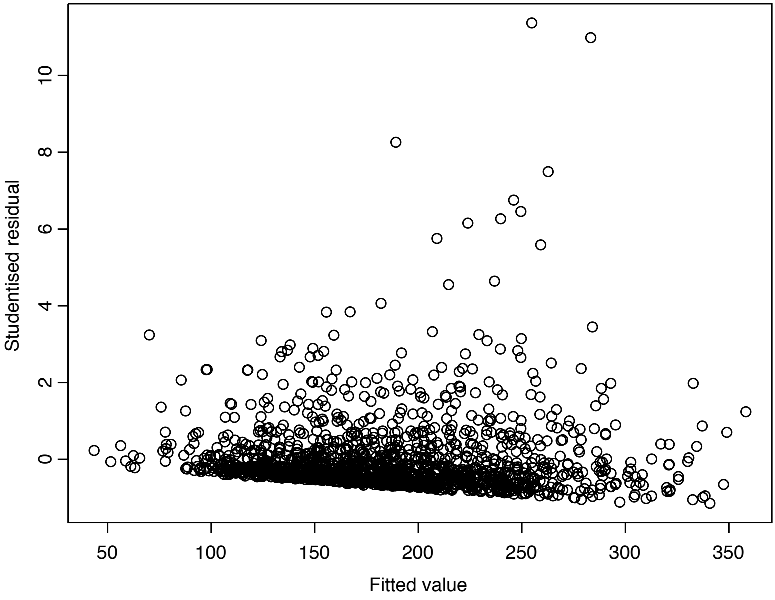
Example from HIE data

- Response variable: respondent's annual expenses on outpatient medical services
  - Here consider only those with non-zero expenses (c.f. Computer class 9 for the rest of the story)
- Explanatory variables: Age, GHI, log of family income and dummy for free health care
- The residual plot shows clear evidence of heteroscedasticity
  - Funnel opening to the right: variability of residuals is larger when fitted values are large
  - Essentially a consequence of the skewness of the distribution of the response variable

Histogram of expenses



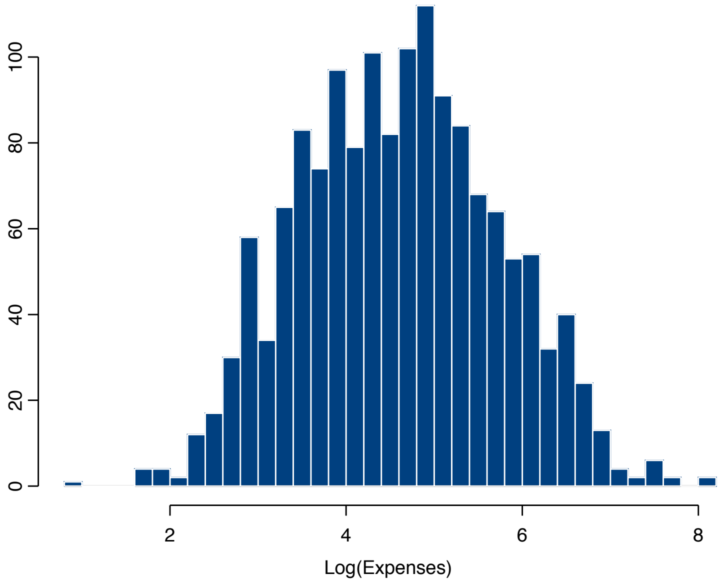
Residual plot: model for expenses



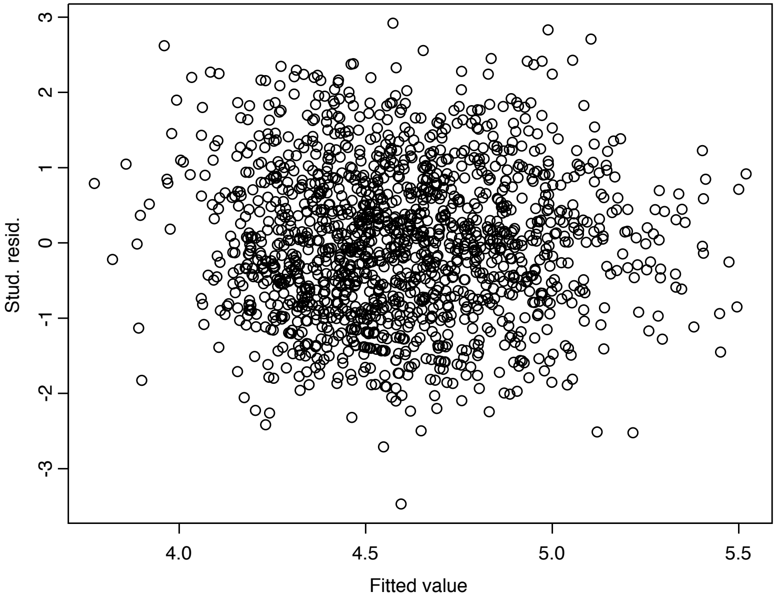
How to remove heteroscedasticity

- The only way discussed today: fit the model using some transformation of of Y as the response variable
- Today, consider only $\log(Y)$
  - Often works well when the response variable has a skewed distribution
- In the example, use log of expenses as the response
  - Residual plot now shows no heteroscedasticity
- Other ways of dealing with heteroscedastic residuals (not discussed here):
  - Other transformations of the response
  - Using "robust" standard errors which are valid even there *is* heteroscedasticity
  - Fitting a more flexible model for the variance of Y

Histogram of log-expenses



Residual plot: model for log-expenses

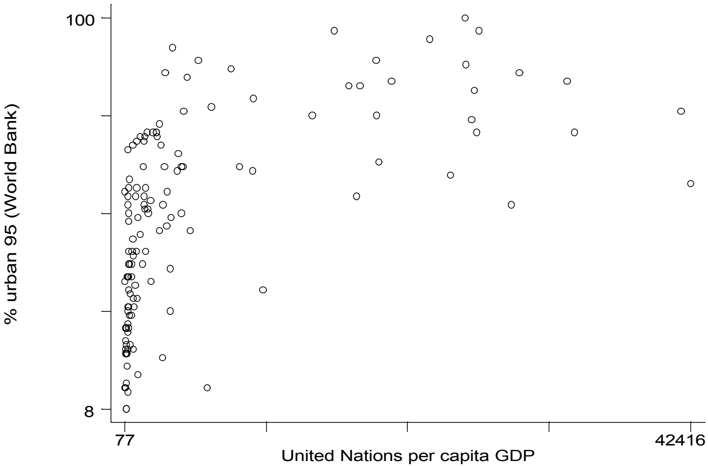


Interpreting coefficients on $\log(X)$, level Y

- $\hat{\beta}$ is the absolute change in Y when X is multiplied by e (2.718)
- You can work out the expected change in Y for a p% increase in X by multiplying $\hat{\beta}$ by $\log([100+\text{p}]/100)$
- To work out the expected change associated with a 10% increase in the independent variable, therefore, multiply by $\log(110/100) = \log(1.1) = 0.09531$
- Alternatively, $\frac{\beta}{100}$ can be interpreted as the increase in Y from a 1% increase in X

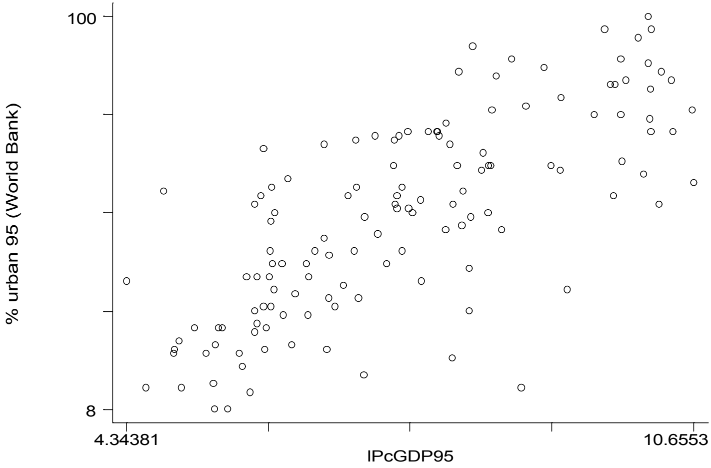
Interpreting coefficients on log(X), level Y

Consider the regression of % urban population (1995) on per capita GNP:



Interpreting coefficients on $\log(X)$, level Y

To control the skew and counter problems in heteroskedasticity, we log GNP/cap:



Interpreting coefficients on $\log(X)$, level Y

```
. regress urb95 lPcGDP95
  Source |       SS       df       MS                  Number of obs =     132
---------+------------------------------               F(  1,   130) =  158.73
   Model |  38856.2103     1  38856.2103               Prob > F      =  0.0000
Residual |  31822.7215   130  244.790165               R-squared     =  0.5498
---------+------------------------------               Adj R-squared =  0.5463
   Total |  70678.9318   131  539.533831               Root MSE      =  15.646
------------------------------------------------------------------------------
   urb95 |      Coef.   Std. Err.       t     P>|t|       [95% Conf. Interval]
---------+--------------------------------------------------------------------
lPcGDP95 |   10.43004   .8278521     12.599   0.000       8.792235    12.06785
   _cons |  -24.42095   6.295892     -3.879   0.000      -36.87662   -11.96528
------------------------------------------------------------------------------
```

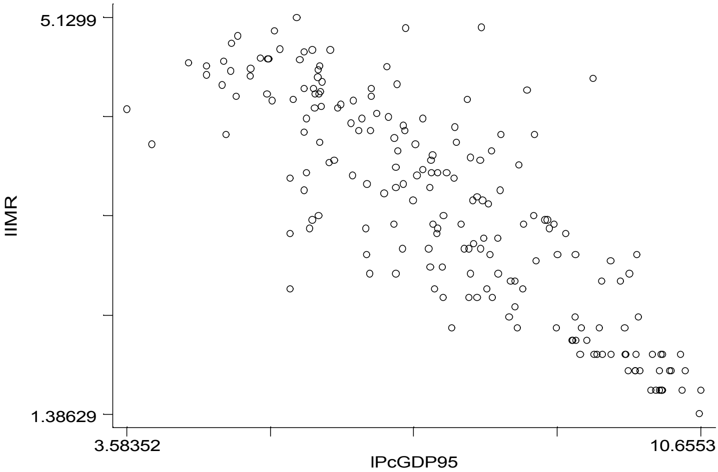
- Multiplying GNP/cap by e (2.718) will increase Y by 10.43
- A 1% increase in GNP/cap will increase Y by 10.43/100=.1043
- A 10% increase in GNP/cap will increase Y by 10.43*.09531=0.994

Interpreting coefficients on log(X) with log(Y)

- Multiplying X by e will increase Y by $e^{b\hat{e}ta}$
- You can work out the expected proportional change in Y for a p% increase in X by computing $e^{log([100+p]/100)\hat{\beta}}$
- The predicted proportional change can be converted to a predicted % change by subtracting 1 and multiplying by 100

Interpreting coefficients on $\log(X)$ with $\log(Y)$

Example: infant mortality Y on GNP/cap as X



Interpreting coefficients on log(X) with log(Y)

```
. regress lIMR lPcGDP95
  Source |       SS       df       MS              Number of obs =     194
---------+------------------------------           F(  1,   192) =  404.52
   Model |  131.035233     1  131.035233           Prob > F      =  0.0000
Residual |  62.1945021   192  .323929698           R-squared     =  0.6781
---------+------------------------------           Adj R-squared =  0.6765
   Total |  193.229735   193  1.00119034           Root MSE      =  .56915
------------------------------------------------------------------------------
    lIMR |      Coef.   Std. Err.       t     P>|t|       [95% Conf. Interval]
---------+--------------------------------------------------------------------
lPcGDP95 |  -.4984531   .0247831    -20.113   0.000      -.5473352    -.449571
   _cons |   7.088676   .1908519     37.142   0.000        6.71224    7.465111
------------------------------------------------------------------------------
```

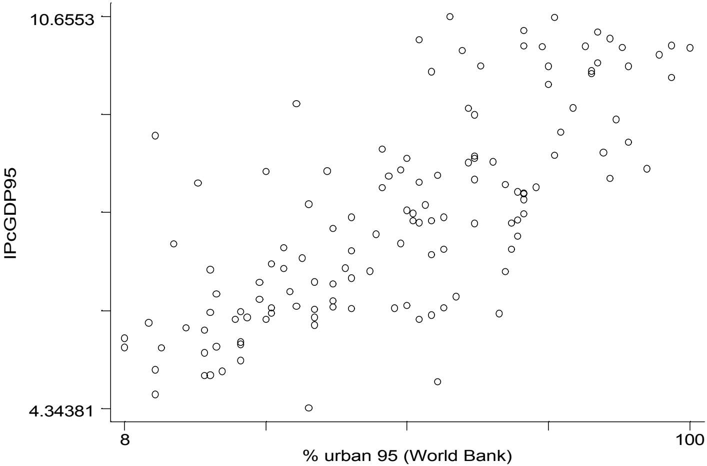
- Multiplying X (GNP/cap) by e multiplies Y by $e^{-.4984531}$
- A 10% increase in GNP/cap multiplies IMR $e^{-.4984531*\log(1.1)} = .954$
- So a 10% increase in GNP/cap reduces IMR by 4.6%

Interpreting coefficients on level X, $\log(Y)$

- Each 1 unit increase in X multiplies X by $e^{\hat{\beta}}$
- Means that very approximately, $\hat{\beta}$ is the percentage increase in Y from a one-unit increase in X

Interpreting coefficients on level X, $\log(Y)$

What if we reverse the X and Y and log urbanization as the $\log(X)$?



Interpreting coefficients on level X, $\log(Y)$

```
. regress lPcGDP95 urb95
  Source |       SS       df       MS              Number of obs =     132
---------+------------------------------           F(  1,   130) =  158.73
   Model |  196.362646     1  196.362646           Prob > F      =  0.0000
Residual |  160.818406   130  1.23706466           R-squared     =  0.5498
---------+------------------------------           Adj R-squared =  0.5463
   Total |  357.181052   131  2.72657291           Root MSE      =  1.1122
------------------------------------------------------------------------------
lPcGDP95 |      Coef.   Std. Err.       t     P>|t|       [95% Conf. Interval]
---------+--------------------------------------------------------------------
   urb95 |    .052709   .0041836     12.599   0.000       .0444322    .0609857
   _cons |   4.630287   .2420303     19.131   0.000       4.151459    5.109115
------------------------------------------------------------------------------
```

- Each one unit increase in urbanization now increases GNP/cap by a multiple of $e^{0.052709} = 1.054$ – or a 5.4% increase

Other uses of the residuals

- Residuals can also be plotted against individual explanatory variables
  - ones already included in the model: looking for evidence of nonlinear effects
  - ones not in the model: looking for evidence of linear or nonlinear effects
  - both are easier with significance tests
- Examining the adequacy of the assumption of normality: *normal probability plots*
  - If the error terms are clearly non-normal, a transformation of the response variable often helps
  - But nonnormality does not matter much, especially in large samples
- Detection of outliers: Individual observations with extreme values of Y (relative to their predicted value)

Assumption of independence

- The remaining model assumption is that the observations Y_i are statistically independent
- For some data structures (e.g. clustered or longitudinal data) it is clear that they are not
- Solution: extend the model to allow for the dependence
  - For that, take St416 (Models for multilevel and longitudinal data) in LT
  - This also provides ways of testing whether the dependence need to be taken into account in the first place

Multicollinearity of explanatory variables

- Multicollinearity occurs when some explanatory variables are exactly or nearly linearly related
  - i.e. the R^2 for any one of them given the others is high
  - for two variables, this is the same as high correlation between them
- When there is *perfect* multicollinearity, some coefficients cannot be estimated at all
  - e.g. if we try to include height in both cm and inches in the same model

Multicollinearity of explanatory variables

- When there is *approximate* multicollinearity, estimates of some coefficients will be unstable
  - e.g. in example below, respondent's income 1 and 2 years before are both included in the model, with a correlation $r = 0.887$
  - In effect, the model has difficulty assigning *separate* effects to them
- What to do about (approximate) multicollinearity?
  - Drop one of the variables causing it, or
  - Transform the variables so that they are less dependent: e.g. *average* and *difference* of the two incomes below, instead of the incomes themselves

Multicollinearity of explanatory variables

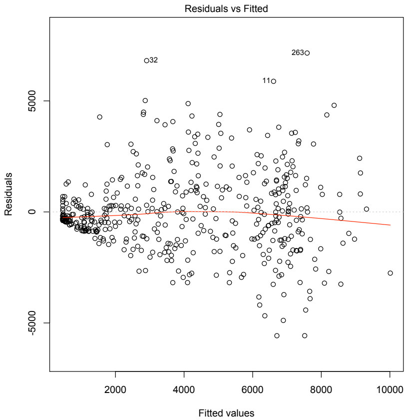
	Response variable: General Health Index				
	Model				
Variable	(1)	(2)	(3)	(4)	(5)
Income 1 year before	0.274	—	0.170	—	—
	[0.067]		[0.146]		
Income 2 years before	—	0.254	0.111	—	—
		[0.064]	[0.138]		
Average of incomes	—	—	—	0.281	0.279
1 and 2 years before				[0.068]	[0.068]
Difference of incomes	—	—	—	0.029	—
1 and 2 years before				[0.138]	
R^2	0.013	0.012	0.013	0.013	0.013

(standard errors in brackets)

Diagnosing problems in residuals (`regress postestimation`

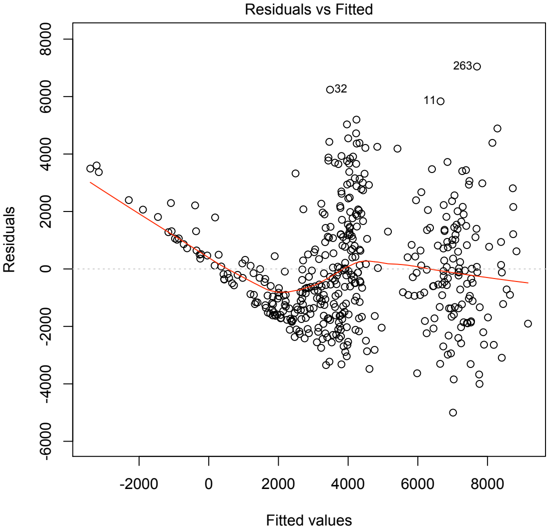
- A very easy set of diagnostic plots can be accessed following a regression, using regression post-estimation commands
- This produces, in order:
  1. residuals against fitted values
  2. Normal Q-Q plot
  3. scale-location plot of $\sqrt{|e_i|}$ against fitted values
  4. Cook's distances versus row labels
  5. residuals against leverages
  6. Cook's distances against leverage/(1-leverage)

Residuals v. fitted plots



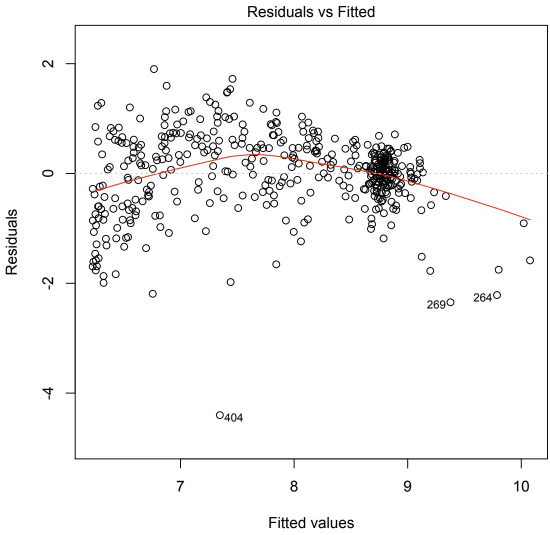
- `rvfplot` (Stata)
- `plot(lm(votes1st~spend_total*incumb, data=dail), which=1)` (R)
- If constant variance assumption holds, then residuals would not show a pattern against fitted values — this pattern suggests a transformation is needed

Residuals v. fitted plots: log(spending)



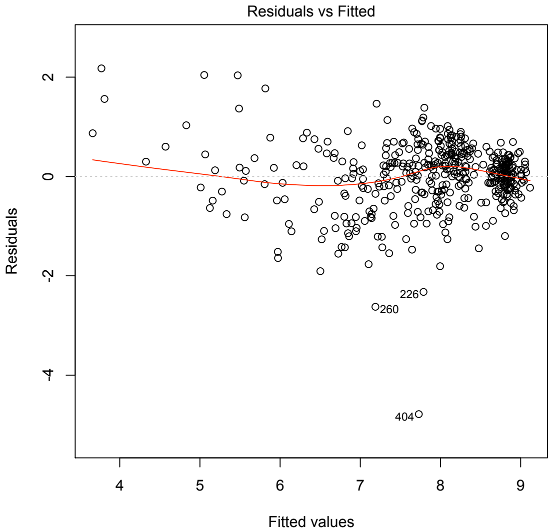

```
plot(lm(votes1st~log(spend_total)*incumb, data=dail), which=1)
```

Residuals v. fitted plots: log(votes)



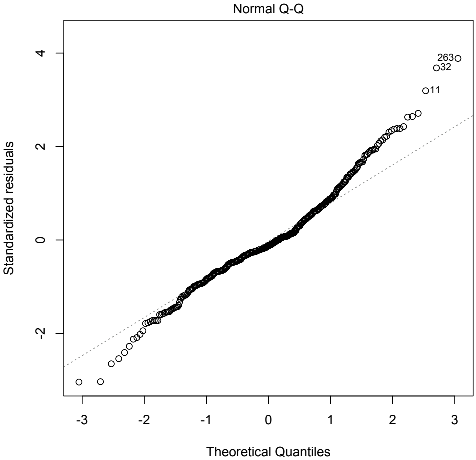

```
plot(lm(log(votes1st)~spend_total*incumb, data=dail), which=1)
```

Residuals v. fitted plots: log(votes) and log(spending)




```
plot(lm(log(votes1st)~log(spend_total)*incumb, data=dail),
which=1)
```

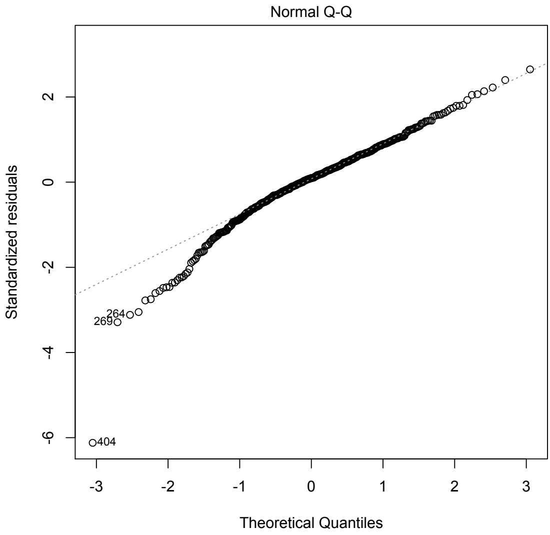
Normal Q-Q plot




```
regress votes1st c.spend_total##incumb
```
(Stata)

```
predict e, residuals
qnorm e
plot(lm(votes1st~spend_total*incumb, data=dail), which=2)
```
(R)

Normal Q-Q plot: logged(votes)




```
plot(lm(log(votes1st)~spend_total*incumb, data=dail), which=2)
```

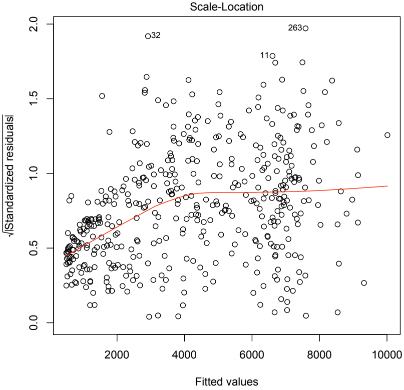
Examine the outliers!

- We can examine the points with row labels 264, 269, 404
- Note: these are not the row numbers any longer, since we removed some with missing values
- Let's see what is strange about these cases:

```
> dail[c("264","269","404"), c("district", "wholename", "party",
         "votes1st", "incumb", "spend_total")]

          district       wholename party votes1st incumb spend_total
264 Cavan Monaghan  Vincent Martin   ind     1943      0    34542.73
269 Cavan Monaghan Gerry McCaughey    pd     1131      0    30573.12
404  Limerick East      Aidan Ryan   ind       19      0    10890.19
```

Scale-Location plot



- Looks at the square root of the absolute (standardized) residuals instead of just residuals, since $\sqrt{|e|}$ is less skewed
- Note the use of *standardized* or *studentized* residuals

```
predict estud, rstudent (Stata)
predict yhat
gen rstscale = sqrt(abs(estud))
graph twoway (scatter estud yhat)
```

```
plot(lm(votes1st~spend_total*incumb, data=dail), which=3) (R)
```