**Description**

boxcox finds the maximum likelihood estimates of the parameters of the Box–Cox transform, the coefficients on the independent variables, and the standard deviation of the normally distributed errors. Any **`depvar`** or **`indepvars`** to be transformed must be strictly positive. Options can be used to control which variables remain untransformed.

**Quick start**

Box–Cox transform of `y` in a model of `y` as a function of `x1`
```
boxcox y x1
```

Same as above
```
boxcox y x1, model(lhsonly)
```

Likelihood-ratio test for each scale-variant parameter
```
boxcox y x1, lrtest
```

Different transform for each side and adding covariates `x2` and `x3`
```
boxcox y x1 x2 x3, model(theta)
```

Same transform for both sides, and include `x3` as an untransformed variable transformation
```
boxcox y x1 x2, model(lambda) notrans(x3)
```

**Menu**

Statistics > Linear models and related > Box-Cox regression
Syntax

```
boxcox depvar [ indepvars ] [ if ] [ in ] [ weight ] [ , options ]
```

### options

<table>
<thead>
<tr>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>suppress constant term</td>
</tr>
<tr>
<td>left-hand-side Box–Cox model; the default</td>
</tr>
<tr>
<td>right-hand-side Box–Cox model</td>
</tr>
<tr>
<td>both sides Box–Cox model with same parameter</td>
</tr>
<tr>
<td>both sides Box–Cox model with different parameters</td>
</tr>
<tr>
<td>do not transform specified independent variables</td>
</tr>
<tr>
<td>set confidence level; default is level(95)</td>
</tr>
<tr>
<td>perform likelihood-ratio test</td>
</tr>
<tr>
<td>suppress all iteration logs</td>
</tr>
<tr>
<td>suppress restricted-model lrtest iteration log</td>
</tr>
<tr>
<td>control the maximization process; seldom used</td>
</tr>
</tbody>
</table>

**Model**

- `noconstant`
- `model(lhsonly)`
- `model(rhsonly)`
- `model(lambda)`
- `model(theta)`
- `notrans(varlist)`

**Reporting**

- `level(#)`
- `lrtest`

**Maximization**

- `[ no ]log`
- `nologlr`  
- `maximize_options`

`depvar` and `indepvars` may contain time-series operators; see [U] 11.4.4 Time-series varlists.

`bootstrap`, `by`, `jackknife`, `rolling`, `statsby`, and `xi` are allowed; see [U] 11.1.10 Prefix commands.

Weights are not allowed with the `bootstrap` prefix; see [R] bootstrap.

`fweight`s and `iweight`s are allowed; see [U] 11.1.6 weight.

See [U] 20 Estimation and postestimation commands for more capabilities of estimation commands.

### Options

- `Model`
  - `noconstant`; see [R] Estimation options.
  - `model(lhsonly)` specifies which of the four models to fit.
    - `model(lhsonly)` applies the Box–Cox transform to `depvar` only. `model(lhsonly)` is the default.
    - `model(rhsonly)` applies the transform to the `indepvars` only.
    - `model(lambda)` applies the transform to both `depvar` and `indepvars`, and they are transformed by the same parameter.
    - `model(theta)` applies the transform to both `depvar` and `indepvars`, but this time, each side is transformed by a separate parameter.
  - `notrans(varlist)` specifies that the variables in `varlist` not be transformed when included in the model. You can specify `notrans(varlist)` with `model(lhsonly)`, but the results will be the same as specifying the variables in `varlist` in `indepvars`.

- `Reporting`
  - `level(#)`; see [R] Estimation options.
lrtest specifies that a likelihood-ratio test of significance be performed and reported for each independent variable.

log and nolog specify whether to display the iteration log. The iteration log is displayed by default unless you used set iterlog off to suppress it; see set iterlog in [R] set iter. These options control the iteration log produced by the full model and, if option lrtest is specified, by the fitted restricted models.

nologlr suppresses the iteration log when fitting the restricted models required by the lrtest option.

**maximize_options:** iterate(#) and from(init_specs); see [R] Maximize.

<table>
<thead>
<tr>
<th>Model</th>
<th>Initial value specification</th>
</tr>
</thead>
<tbody>
<tr>
<td>lhsonly</td>
<td>from(θ₀, copy)</td>
</tr>
<tr>
<td>rhsonly</td>
<td>from(λ₀, copy)</td>
</tr>
<tr>
<td>lambda</td>
<td>from(λ₀, copy)</td>
</tr>
<tr>
<td>theta</td>
<td>from(λ₀θ₀, copy)</td>
</tr>
</tbody>
</table>

**Remarks and examples**

Remarks are presented under the following headings:

- Introduction
- Theta model
- Lambda model
- Left-hand-side-only model
- Right-hand-side-only model

**Introduction**

The Box–Cox transform

\[
y'(λ) = \frac{y^λ - 1}{λ}
\]

has been widely used in applied data analysis. Box and Cox (1964) developed the transformation and argued that the transformation could make the residuals more closely normal and less heteroskedastic. Cook and Weisberg (1982) discuss the transform in this light. Because the transform embeds several popular functional forms, it has received some attention as a method for testing functional forms, in particular,

\[
y'(λ) = \begin{cases} 
  y - 1 & \text{if } λ = 1 \\
  \ln(y) & \text{if } λ = 0 \\
  1 - 1/y & \text{if } λ = -1
\end{cases}
\]

Theta model

boxcox obtains the maximum likelihood estimates of the parameters for four different models. The most general of the models, the theta model, is

$$y_j^{(\theta)} = \beta_0 + \beta_1 x_{1j}^{(\lambda)} + \beta_2 x_{2j}^{(\lambda)} + \cdots + \beta_k x_{kj}^{(\lambda)} + \gamma_1 z_{1j} + \gamma_2 z_{2j} + \cdots + \gamma_l z_{lj} + \epsilon_j$$

where $\epsilon \sim N(0,\sigma^2)$. Here the dependent variable, $y$, is subject to a Box–Cox transform with parameter $\theta$. Each of the indepvars, $x_1, x_2, \ldots, x_k$, is transformed by a Box–Cox transform with parameter $\lambda$. The $z_1, z_2, \ldots, z_l$ specified in the notrans() option are independent variables that are not transformed.

Box and Cox (1964) argued that this transformation would leave behind residuals that more closely follow a normal distribution than those produced by a simple linear regression model. Bear in mind that the normality of $\epsilon$ is assumed and that boxcox obtains maximum likelihood estimates of the $k + l + 4$ parameters under this assumption. boxcox does not choose $\lambda$ and $\theta$ so that the residuals are approximately normally distributed. If you are interested in this type of transformation to normality, see the official Stata commands lnskew0 and bcskew0 in [R] lnskew0. However, those commands work on a more restrictive model in which none of the independent variables is transformed.

Example 1

Below, we fit a theta model to a nonrepresentative extract of the Second National Health and Nutrition Examination Survey (NHANES II) dataset discussed in McDowell et al. (1981).

We model individual-level diastolic blood pressure (bpdiast) as a function of the transformed variables body mass index (bmi) and cholesterol level (tcresult) and of the untransformed variables age (age) and sex (sex).
. use https://www.stata-press.com/data/r16/nhanes2
. boxcox bpdiast bmi tcrresult, notrans(age sex) model(theta) lrtest
Fitting comparison model
Iteration 0:  log likelihood = -41178.61
Iteration 1:  log likelihood = -41032.51
Iteration 2:  log likelihood = -41032.488
Iteration 3:  log likelihood = -41032.488

Fitting full model
Iteration 0:  log likelihood = -39928.606
Iteration 1:  log likelihood = -39775.026
Iteration 2:  log likelihood = -39774.987
Iteration 3:  log likelihood = -39774.987

Fitting comparison models for LR tests
Iteration 0:  log likelihood = -39947.144
Iteration 1:  log likelihood = -39934.55
Iteration 2:  log likelihood = -39934.516
Iteration 3:  log likelihood = -39934.516
Iteration 0:  log likelihood = -39829.859
Iteration 1:  log likelihood = -39815.576
Iteration 2:  log likelihood = -39815.575

Number of obs = 10,351
LR chi2(5) = 2515.00
Log likelihood = -39774.987  Prob > chi2 = 0.000

bpdiast | Coef. Std. Err.  z  P>|z|  [95% Conf. Interval]
---------|------------------|----------|--------|-----------------|-----------------|
 /lambda | .6383286  .1577601  4.05  0.000  .3291245  .9475327
 /theta  | .1988197  .0454088  4.38  0.000  .1098201  .2878193

Estimates of scale-variant parameters

<table>
<thead>
<tr>
<th>Coef.</th>
<th>chi2(df)</th>
<th>P&gt;chi2(df)</th>
<th>df of chi2</th>
</tr>
</thead>
</table>
Notrans | age  | .003811  | 319.060    | 0.000      | 1            |
  sex   | -.1054887| 243.284   | 0.000      | 1            |
  _cons | 5.835555 |            |            |              |
Trans   | bmi  | .0872041 | 1369.235   | 0.000      | 1            |
  tcrresult | .004734 | 81.177    | 0.000      | 1            |
 /sigma | .3348267 |            |            |              |

Test
HO: theta=lambda = -1, 0, 1

<table>
<thead>
<tr>
<th>theta=lambda = -1</th>
<th>Restricted log likelihood</th>
<th>chi2</th>
<th>Prob &gt; chi2</th>
</tr>
</thead>
<tbody>
<tr>
<td>-40162.898</td>
<td>775.82</td>
<td>0.000</td>
<td></td>
</tr>
<tr>
<td>-39790.945</td>
<td>31.92</td>
<td>0.000</td>
<td></td>
</tr>
<tr>
<td>-39928.606</td>
<td>307.24</td>
<td>0.000</td>
<td></td>
</tr>
</tbody>
</table>
The output is composed of the iteration logs and three distinct tables. The first table contains a standard header for a maximum likelihood estimator and a standard output table for the Box–Cox transform parameters. The second table contains the estimates of the scale-variant parameters. The third table contains the output from likelihood-ratio tests on three standard functional form specifications.

The right-hand-side and the left-hand-side transformations each add to the regression fit at the 1% significance level and are both positive but less than 1. All the variables have significant impacts on diastolic blood pressure, \( \text{bpdiast} \). As expected, the transformed variables—the body mass index, \( \text{bmi} \), and cholesterol level, \( \text{tresult} \)—contribute to higher blood pressure. The last output table shows that the linear, multiplicative inverse, and log specifications are strongly rejected.

Technical note

Spitzer (1984) showed that the Wald tests of the joint significance of the coefficients of the right-hand-side variables, either transformed or untransformed, are not invariant to changes in the scale of the transformed dependent variable. Davidson and MacKinnon (1993) also discuss this point. This problem demonstrates that Wald statistics can be manipulated in nonlinear models. Lafontaine and White (1986) analyze this problem numerically, and Phillips and Park (1988) analyze it by using Edgeworth expansions. See Drukker (2000) for a more detailed discussion of this issue. Because the parameter estimates and their Wald tests are not scale invariant, no Wald tests or confidence intervals are reported for these parameters. However, when the `lrtest` option is specified, likelihood-ratio tests are performed and reported. Schlesselman (1971) showed that, if a constant is included in the model, the parameter estimates of the Box–Cox transforms are scale invariant. For this reason, we strongly recommend that you not use the `noconstant` option.

The `lrtest` option does not perform a likelihood-ratio test on the constant, so no value for this statistic is reported. Unless the data are properly scaled, the restricted model does not often converge. For this reason, no likelihood-ratio test on the constant is performed by the `lrtest` option. However, if you have a special interest in performing this test, you can do so by fitting the constrained model separately. If problems with convergence are encountered, rescaling the data by their means may help.

Lambda model

A less general model than the one above is called the `lambda` model. It specifies that the same parameter be used in both the left-hand-side and right-hand-side transformations. Specifically,

\[
y_j^{(\lambda)} = \beta_0 + \beta_1 x_1^{(\lambda)} + \beta_2 x_2^{(\lambda)} + \cdots + \beta_k x_k^{(\lambda)} + \gamma_1 z_1 + \gamma_2 z_2 + \cdots + \gamma_l z_l + \epsilon_j
\]

where \( \epsilon \sim N(0, \sigma^2) \). Here the `depvar` variable, \( y \), and each of the `indepvars`, \( x_1, x_2, \ldots, x_k \), is transformed by a Box–Cox transform with the common parameter \( \lambda \). Again, the \( z_1, z_2, \ldots, z_l \) are independent variables that are not transformed.

Left-hand-side-only model

Even more restrictive than a common transformation parameter is transforming the dependent variable only. Because the dependent variable is on the left-hand side of the equation, this model is known as the `lhsonly` model. Here you are estimating the parameters of the model
where \( \epsilon \sim N(0, \sigma^2) \). Here only the \textit{depvar}, \( y \), is transformed by a Box–Cox transform with the parameter \( \theta \).

\[ y_j^{(\theta)} = \beta_0 + \beta_1 x_{1j} + \beta_2 x_{2j} + \cdots + \beta_k x_{kj} + \epsilon_j \]

\[ \theta = -1 \]  
\[ \theta = 0 \]  
\[ \theta = 1 \]

The maximum likelihood estimate of the transformation parameter for this model is positive and significant. Once again, all the scale-variant parameters are significant, and we find a positive impact of body mass index (\texttt{bmi}) and cholesterol levels (\texttt{tcresult}) on the transformed diastolic blood pressure (\texttt{bpdiast}). This model rejects the linear, multiplicative inverse, and log specifications.
Right-hand-side-only model

The fourth model leaves the `depvar` alone and transforms a subset of the `indepvars` using the parameter \( \lambda \). This is the `rhsonly` model. In this model, the `depvar`, \( y \), is given by

\[
y_j = \beta_0 + \beta_1 x_1^{(\lambda)} + \beta_2 x_2^{(\lambda)} + \cdots + \beta_k x_k^{(\lambda)} + \gamma_1 z_1 + \cdots + \gamma_l z_l + \epsilon_j
\]

where \( \epsilon \sim N(0, \sigma^2) \). Here each of the `indepvars`, \( x_1, x_2, \ldots, x_k \), is transformed by a Box–Cox transform with the parameter \( \lambda \). Again, the \( z_1, z_2, \ldots, z_l \) are independent variables that are not transformed.

Example 3

Now, we consider a `rhsonly` model in which the regressors `sex` and `age` are not transformed.

```
. boxcox bpdiastric bmi tcresult, notrans(sex age) model(rhsonly) lrtest nolog
> nologlr
```

```
Fitting full model
Fitting comparison models for LR tests

Number of obs = 10,351
LR chi2(5) = 2500.79
Log likelihood = -39928.212 Prob > chi2 = 0.000

bpdiast       Coef.  Std. Err.   z  P>|z|     [95% Conf. Interval]
/lambda         .8658841   .1522387   5.69  0.000     .5675018   1.164266

Estimates of scale-variant parameters

<table>
<thead>
<tr>
<th></th>
<th>Coef.</th>
<th>chi2(df)</th>
<th>P&gt;chi2(df)</th>
<th>df of chi2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Notrans</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>sex</td>
<td>-3.544042</td>
<td>235.020</td>
<td>0.000</td>
<td>1</td>
</tr>
<tr>
<td>age</td>
<td>.128809</td>
<td>311.754</td>
<td>0.000</td>
<td>1</td>
</tr>
<tr>
<td>_cons</td>
<td>50.01498</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Trans</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>bmi</td>
<td>1.418215</td>
<td>1396.709</td>
<td>0.000</td>
<td>1</td>
</tr>
<tr>
<td>tcresult</td>
<td>0.0462964</td>
<td>78.500</td>
<td>0.000</td>
<td>1</td>
</tr>
<tr>
<td>/sigma</td>
<td>11.4557</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Test HO:</th>
<th>Restricted log likelihood</th>
<th>LR statistic chi2</th>
<th>P-value Prob &gt; chi2</th>
</tr>
</thead>
<tbody>
<tr>
<td>lambda = -1</td>
<td>-39989.331</td>
<td>122.24</td>
<td>0.000</td>
</tr>
<tr>
<td>lambda = 0</td>
<td>-39942.945</td>
<td>29.47</td>
<td>0.000</td>
</tr>
<tr>
<td>lambda = 1</td>
<td>-39928.606</td>
<td>0.79</td>
<td>0.375</td>
</tr>
</tbody>
</table>
```

The maximum likelihood estimate of the transformation parameter in this model is positive and significant at the 1% level. The transformed `bmi` coefficient behaves as expected, and the remaining scale-variant parameters are significant at the 1% level. This model rejects the multiplicative inverse and log specifications strongly. However, we cannot reject the hypothesis that the model is linear.
Stored results

boxcox stores the following in e():

Scalars
- e(N)  number of observations
- e(ll) log likelihood
- e(chi2) LR statistic of full vs. comparison
- e(df_m) full model degrees of freedom
- e(ll0) log likelihood of the restricted model
- e(df_r) restricted model degrees of freedom
- e(ll_t1) log likelihood of model \( \lambda = \theta = 1 \)
- e(chi2_t1) LR of \( \lambda = \theta = 1 \) vs. full model
- e(p_t1) p-value of \( \lambda = \theta = 1 \) vs. full model
- e(ll_tm1) log likelihood of model \( \lambda = \theta = -1 \)
- e(chi2_tm1) LR of \( \lambda = \theta = -1 \) vs. full model
- e(p_tm1) p-value of \( \lambda = \theta = -1 \) vs. full model
- e(ll_t0) log likelihood of model \( \lambda = \theta = 0 \)
- e(chi2_t0) LR of \( \lambda = \theta = 0 \) vs. full model
- e(p_t0) p-value of \( \lambda = \theta = 0 \) vs. full model
- e(rank) rank of e(V)
- e(ic) number of iterations
- e(rc) return code

Macros
- e(cmd) boxcox
- e(cmdline) command as typed
- e(depvar) name of dependent variable
- e(model) lhsonly, rhsonly, lambda, or theta
- e(wtype) weight type
- e(wexp) weight expression
- e(ntrans) yes if untransformed indepvars
- e(chi2type) LR; type of model \( \chi^2 \) test
- e(lrtest) lrtest, if requested
- e(properties) b V
- e(predict) program used to implement predict
- e(marginsnotok) predictions disallowed by margins

Matrices
- e(b) coefficient vector
- e(V) variance–covariance matrix of the estimators (see note below)
- e(pm) p-values for LR tests on indepvars
- e(df) degrees of freedom of LR tests on indepvars
- e(chi2m) LR statistics for tests on indepvars

Functions
- e(sample) marks estimation sample

e(V) contains all zeros, except for the elements that correspond to the parameters of the Box–Cox transform.
Methods and formulas

In the internal computations,

\[ y^{(\lambda)} = \begin{cases} \frac{y^{\lambda-1}}{\lambda} & \text{if } |\lambda| > 10^{-10} \\ \ln(y) & \text{otherwise} \end{cases} \]

The unconcentrated log likelihood for the theta model is

\[ \ln L = \left( -\frac{N}{2} \right) \left\{ \ln(2\pi) + \ln(\sigma^2) \right\} + (\theta - 1) \sum_{i=1}^{N} \ln(y_i) - \left( \frac{1}{2\sigma^2} \right) \text{SSR} \]

where

\[ \text{SSR} = \sum_{i=1}^{N} (y_i^{(\theta)} - \beta_0 + \beta_1 x_{i1}^{(\lambda)} + \beta_2 x_{i2}^{(\lambda)} + \cdots + \beta_k x_{ik}^{(\lambda)} + \gamma_1 z_{i1} + \gamma_2 z_{i2} + \cdots + \gamma_l z_{il})^2 \]

Writing the SSR in matrix form,

\[ \text{SSR} = (y^{(\theta)} - X^{(\lambda)}d' - Zg')'(y^{(\theta)} - X^{(\lambda)}d' - Zg') \]

where \( y^{(\theta)} \) is an \( N \times 1 \) vector of elementwise transformed data, \( X^{(\lambda)} \) is an \( N \times k \) matrix of elementwise transformed data, \( Z \) is an \( N \times l \) matrix of untransformed data, \( b \) is a \( 1 \times k \) vector of coefficients, and \( g \) is a \( 1 \times l \) vector of coefficients. Letting

\[ W_{\lambda} = \begin{pmatrix} X^{(\lambda)} \\ Z \end{pmatrix} \]

be the horizontal concatenation of \( X^{(\lambda)} \) and \( Z \) and

\[ d' = \begin{pmatrix} b' \\ g' \end{pmatrix} \]

be the vertical concatenation of the coefficients yields

\[ \text{SSR} = (y^{(\theta)} - W_{\lambda}d')'(y^{(\theta)} - W_{\lambda}d') \]

For given values of \( \lambda \) and \( \theta \), the solutions for \( d' \) and \( \sigma^2 \) are

\[ \hat{d}' = (W'_\lambda W_\lambda)^{-1} W'_\lambda y^{(\theta)} \]

and

\[ \hat{\sigma}^2 = \frac{1}{\hat{d}} \left( y^{(\theta)} - W_\lambda \hat{d}' \right)' \left( y^{(\theta)} - W_\lambda \hat{d}' \right) \]

Substituting these solutions into the log-likelihood function yields the concentrated log-likelihood function

\[ \ln L_c = \left( -\frac{N}{2} \right) \left\{ \ln(2\pi) + 1 + \ln(\hat{\sigma}^2) \right\} + (\theta - 1) \sum_{i=1}^{N} \ln(y_i) \]
Similar calculations yield the concentrated log-likelihood function for the \( \lambda \) model,

\[
\ln L_c = \left( -\frac{N}{2} \right) \{ \ln(2\pi) + 1 + \ln(\hat{\sigma}^2) \} + (\lambda - 1) \sum_{i=1}^{N} \ln(y_i)
\]

the \( lhsonly \) model,

\[
\ln L_c = \left( -\frac{N}{2} \right) \{ \ln(2\pi) + 1 + \ln(\hat{\sigma}^2) \} + (\theta - 1) \sum_{i=1}^{N} \ln(y_i)
\]

and the \( rhsonly \) model,

\[
\ln L_c = \left( -\frac{N}{2} \right) \{ \ln(2\pi) + 1 + \ln(\hat{\sigma}^2) \}
\]

where \( \hat{\sigma}^2 \) is specific to each model and is defined analogously to that in the \( \theta \) model.

References


Also see

[R] boxcox postestimation — Postestimation tools for boxcox
[R] lnskew0 — Find zero-skewness log or Box–Cox transform
[R] regress — Linear regression
[U] 20 Estimation and postestimation commands