Using and interpreting restricted cubic splines

Maarten L. Buis

Institut für Soziologie
Eberhard Karls Universität Tübingen
maarten.buis@ifsoz.uni-tuebingen.de

Outline

The default is linear

- A large part of daily statistical practice consists of estimating the relationship between two or more variables.

The default is linear

- A large part of daily statistical practice consists of estimating the relationship between two or more variables.
- The default is often to assume the relationships are linear.

The default is linear

- A large part of daily statistical practice consists of estimating the relationship between two or more variables.
- The default is often to assume the relationships are linear.
- This assumption is (almost) always wrong

The default is linear

- A large part of daily statistical practice consists of estimating the relationship between two or more variables.
- The default is often to assume the relationships are linear.
- This assumption is (almost) always wrong but is still a very good thing:

The default is linear

- A large part of daily statistical practice consists of estimating the relationship between two or more variables.
- The default is often to assume the relationships are linear.
- This assumption is (almost) always wrong but is still a very good thing:
  - The aim of a model is to simplify the situation such that mere mortals can understand the patterns present in the data.

The default is linear

- A large part of daily statistical practice consists of estimating the relationship between two or more variables.
- The default is often to assume the relationships are linear.
- This assumption is (almost) always wrong but is still a very good thing:
  - The aim of a model is to simplify the situation such that mere mortals can understand the patterns present in the data.
  - Assuming that a relationship is linear is a very natural and useful simplification.

The default is linear

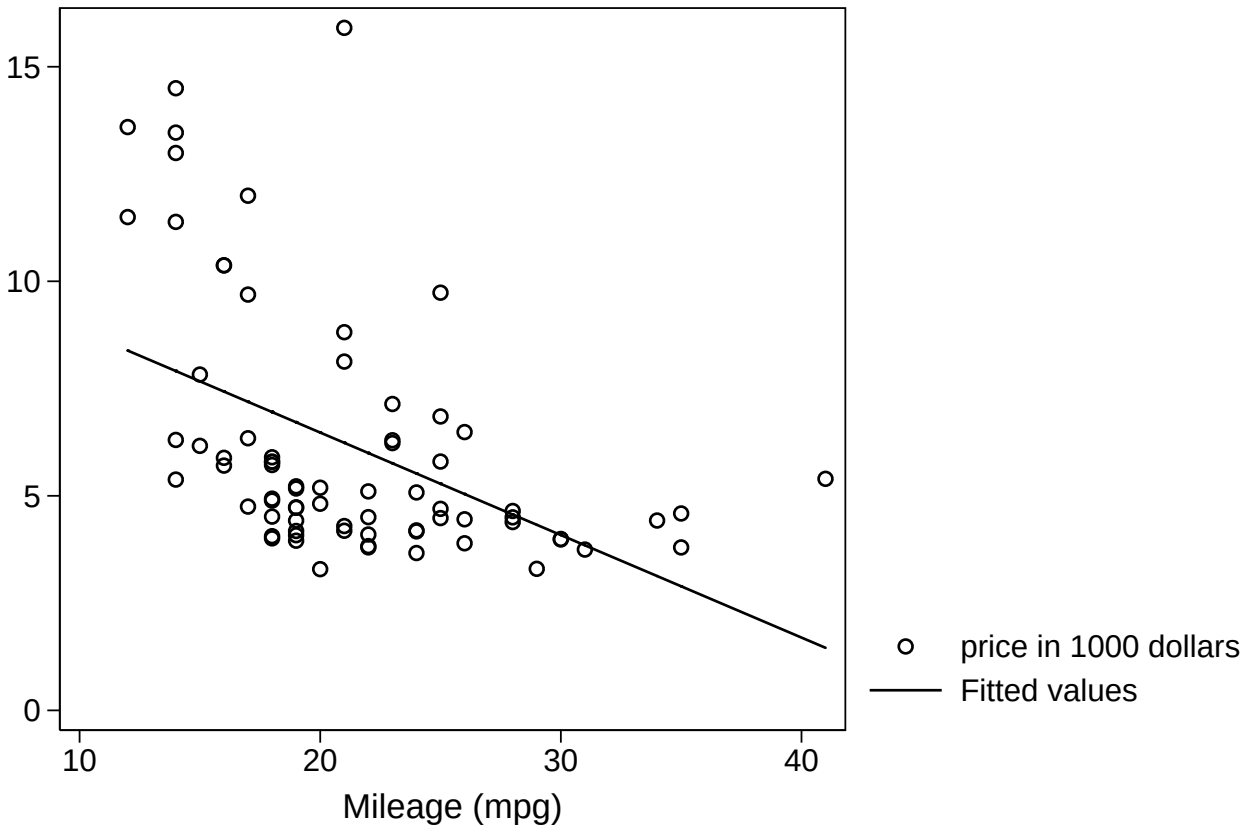
- A large part of daily statistical practice consists of estimating the relationship between two or more variables.
- The default is often to assume the relationships are linear.
- This assumption is (almost) always wrong but is still a very good thing:
  - The aim of a model is to simplify the situation such that mere mortals can understand the patterns present in the data.
  - Assuming that a relationship is linear is a very natural and useful simplification.
- This talk deals with the rare situation where we want to consider non-linear effect.

The default is linear

- A large part of daily statistical practice consists of estimating the relationship between two or more variables.
- The default is often to assume the relationships are linear.
- This assumption is (almost) always wrong but is still a very good thing:
  - The aim of a model is to simplify the situation such that mere mortals can understand the patterns present in the data.
  - Assuming that a relationship is linear is a very natural and useful simplification.
- This talk deals with the rare situation where we want to consider non-linear effect.
- This could for example occur because:
  - the relationship is too non-linear to be meaningfully summarized by a linear relationship, or
  - we are substantively interested in the non-linearity.

Outline

A linear association



How did I do that?

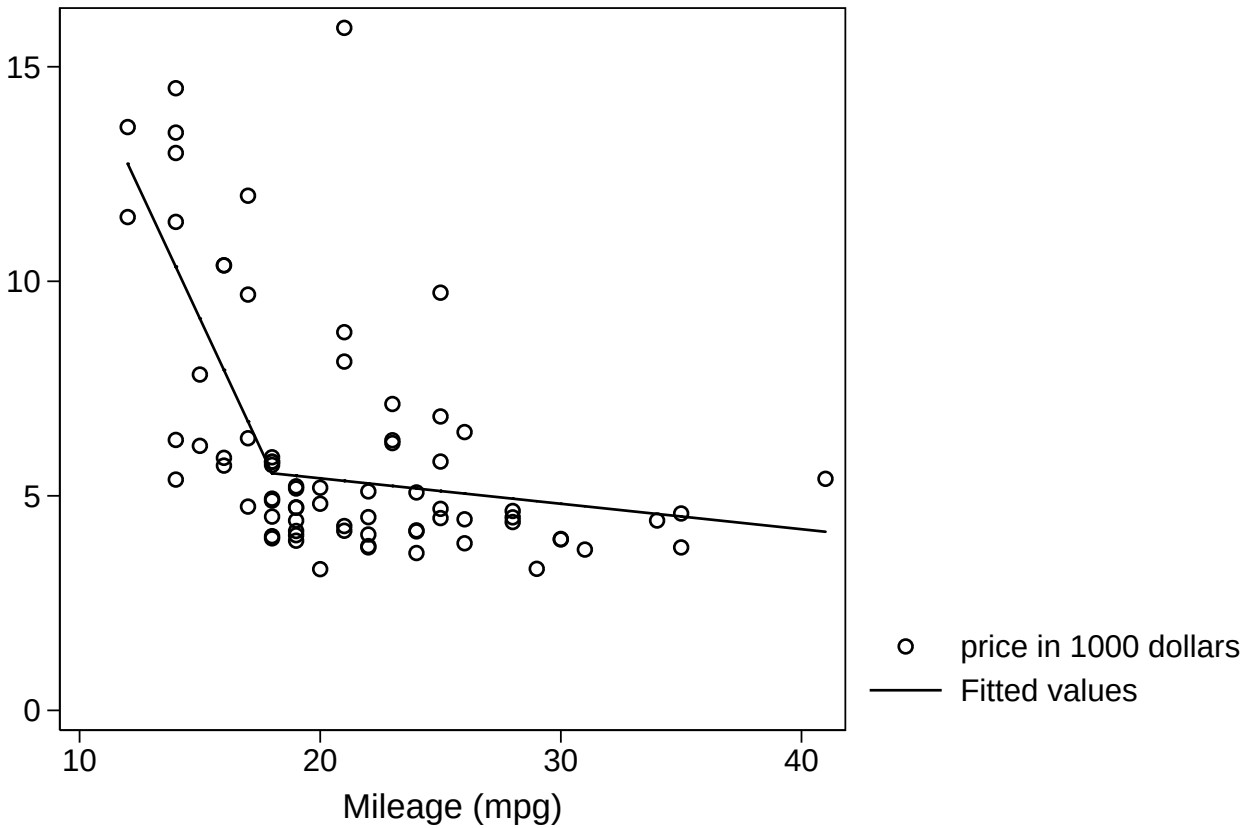
```
. sysuse auto, clear
(1978 Automobile Data)
. replace price = price / 1000
price was int now float
(74 real changes made)
. label variable price "price in 1000 dollars"
.
. reg price mpg
```

Source	SS	df	MS		
Model	139.44947	1	139.44947	Number of obs	= 74
Residual	495.615911	72	6.88355432	F(1, 72)	= 20.26
				Prob > F	= 0.0000
				R-squared	= 0.2196
				Adj R-squared	= 0.2087
Total	635.065382	73	8.69952578	Root MSE	= 2.6237

price	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
mpg	-.2388943	.0530767	-4.50	0.000	-.3447008	-.1330879
_cons	11.25306	1.170813	9.61	0.000	8.919088	13.58703

```
. predict y_lin
(option xb assumed; fitted values)
. twoway scatter price mpg || ///
>        line y_lin mpg,      ///
>        sort clstyle(solid)
```

A linear spline



How did I do that?

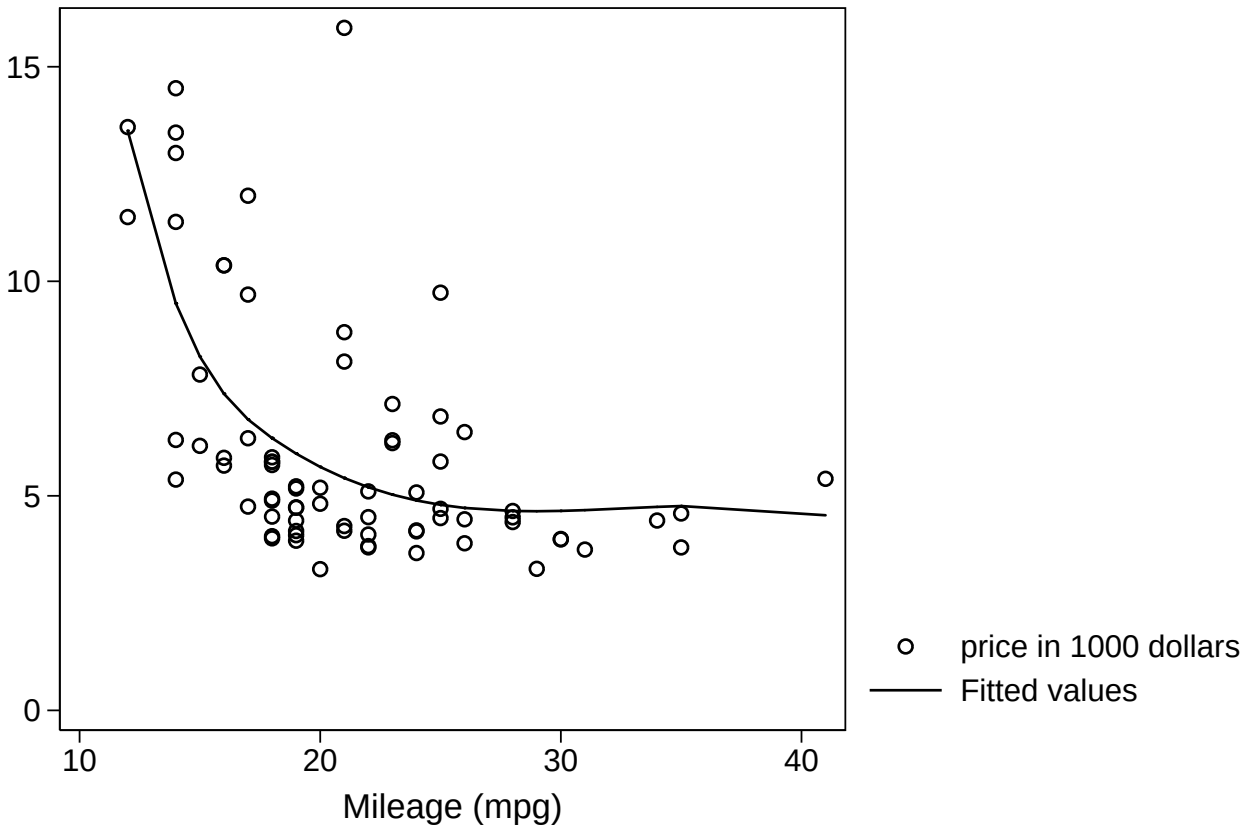
```
. mkspline linsp_mpg1 18 linsp_mpg2= mpg
. reg price linsp*
```

Source	SS	df	MS		
Model	278.152833	2	139.076416	Number of obs	= 74
Residual	356.912549	71	5.02693731	F(2, 71)	= 27.67
				Prob > F	= 0.0000
Total	635.065382	73	8.69952578	R-squared	= 0.4380
				Adj R-squared	= 0.4222
				Root MSE	= 2.2421

price	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
linsp_mpg1	-1.20196	.1888701	-6.36	0.000	-1.578556	-.8253636
linsp_mpg2	-.0592943	.0568009	-1.04	0.300	-.1725521	.0539635
_cons	27.16221	3.189679	8.52	0.000	20.80217	33.52225

```
. test linsp_mpg1 = linsp_mpg2
 ( 1)  linsp_mpg1 - linsp_mpg2 = 0
       F(  1,    71) =   27.59
            Prob > F =    0.0000
. predict y_linsp
(option xb assumed; fitted values)
. twoway scatter price mpg || line y_linsp mpg, sort clstyle(solid)
```

A cubic spline



How did I do that?

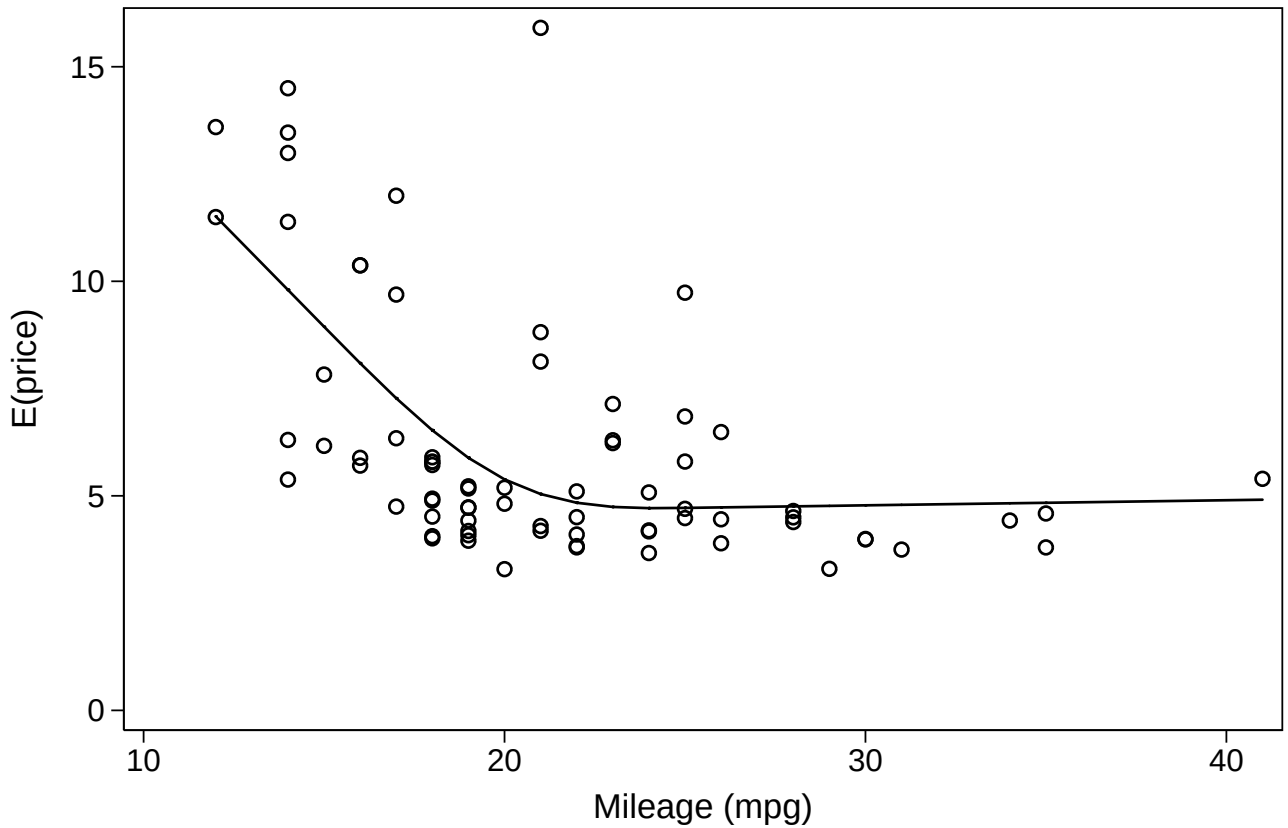
```
. mkspline cubsp_mpg1 18 cubsp_mpg2 = mpg, marginal
. foreach var of varlist cubsp* {
  2.         qui replace `var´ = `var´^3
  3. }
. gen cubsp_sq = mpg^2
. gen cubsp_lin = mpg
. reg price cubsp*
```

Source	SS	df	MS		
Model	249.529494	4	62.3823734	Number of obs	= 74
Residual	385.535888	69	5.58747664	F(4, 69)	= 11.16
				Prob > F	= 0.0000
Total	635.065382	73	8.69952578	R-squared	= 0.3929
				Adj R-squared	= 0.3577
				Root MSE	= 2.3638

price	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
cubsp_mpg1	-.0175977	.0136154	-1.29	0.201	-.0447597	.0095643
cubsp_mpg2	.0169481	.0143188	1.18	0.241	-.0116172	.0455134
cubsp_sq	.9787628	.7142946	1.37	0.175	-.446216	2.403742
cubsp_lin	-18.52005	12.34361	-1.50	0.138	-43.14487	6.104773
_cons	125.2162	70.09313	1.79	0.078	-14.61577	265.0482

```
. predict y_cubsp
(option xb assumed; fitted values)
. twoway scatter price mpg || line y_cubsp mpg, sort clstyle(solid)
```

A restricted cubic spline



How did I do that?

```
. mkspline2 rc = mpg, cubic knots(15 20 25)
. reg price rc*
```

Source	SS	df	MS		
Model	242.090418	2	121.045209	Number of obs =	74
Residual	392.974964	71	5.53485864	F(2, 71) =	21.87
				Prob > F =	0.0000
				R-squared =	0.3812
				Adj R-squared =	0.3638
Total	635.065382	73	8.69952578	Root MSE =	2.3526

price	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
rc1	-.8567267	.151159	-5.67	0.000	-1.158129	-.5553242
rc2	.5791311	.1344838	4.31	0.000	.3109781	.8472842
_cons	21.79347	2.663314	8.18	0.000	16.48297	27.10397

```
. adjustrcspline , noci addplot(scatter price mpg, msymbol(Oh))
```

Outline

The `postrcspline` package

- Available from SSC
- consists of three programs:

`mkspline2`	The same as `mkspline` except that it leaves information behind that can be used by the other commands.
`adjustrcspline`	Displays the adjusted predictions.
`mfxrcspline`	Displays marginal effects.

Adjusted predictions

- Show the predicted outcome against the spline variable.

Adjusted predictions

- Show the predicted outcome against the spline variable.
- What if we have other covariates?

Adjusted predictions

- Show the predicted outcome against the spline variable.
- What if we have other covariates?
- Predicted outcome for an observation with typical values on the other covariates

Adjusted predictions

- Show the predicted outcome against the spline variable.
- What if we have other covariates?
- Predicted outcome for an observation with typical values on the other covariates

```
. reg price rc* rep78 foreign
```

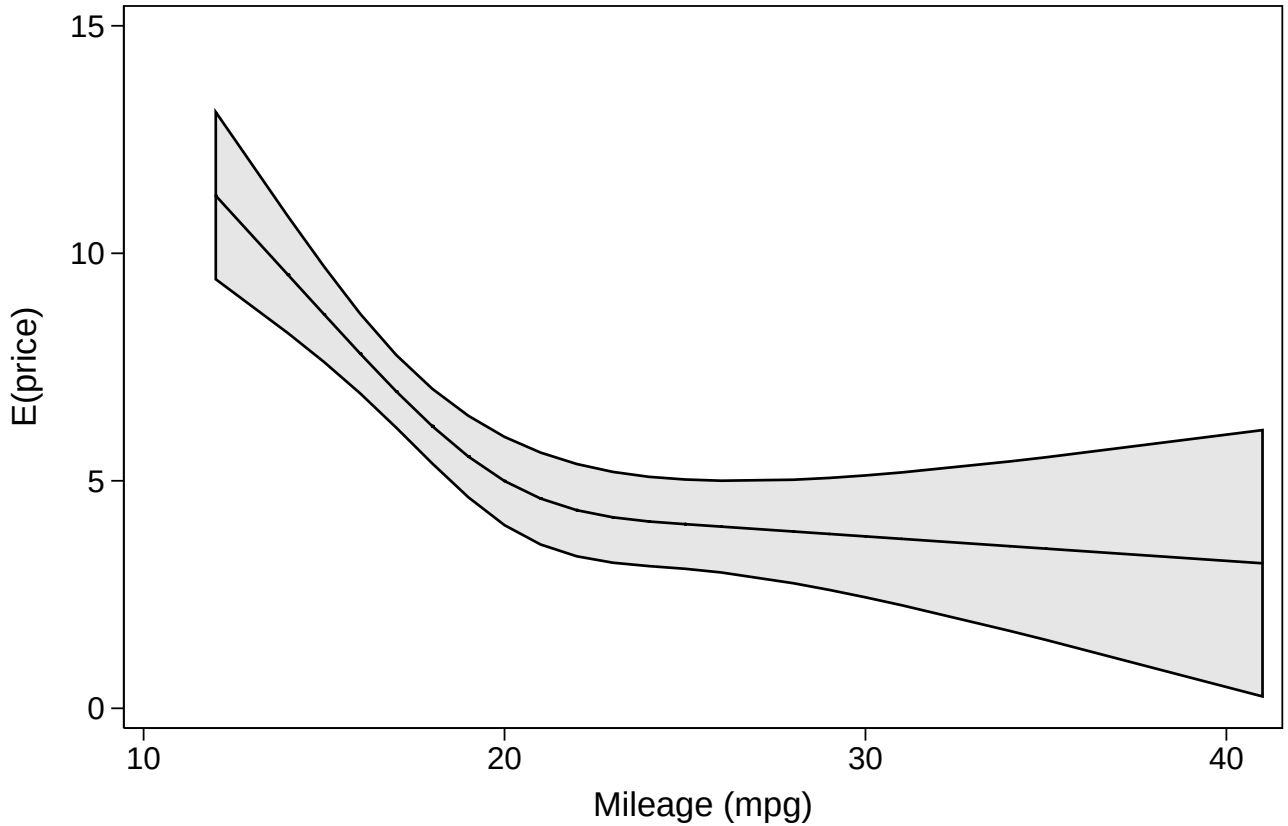
Source	SS	df	MS		
Model	230.445919	4	57.6114798		
Residual	346.351028	64	5.41173481		
Total	576.796947	68	8.48230805		

Number of obs	=	69			
F(4, 64)	=	10.65			
Prob > F	=	0.0000			
R-squared	=	0.3995			
Adj R-squared	=	0.3620			
Root MSE	=	2.3263			

price	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
rc1	-.8688077	.1627389	-5.34	0.000	-1.193916	-.5436995
rc2	.543387	.1444228	3.76	0.000	.2548693	.8319048
rep78	-.0172764	.379311	-0.05	0.964	-.7750371	.7404844
foreign	1.607754	.8049689	2.00	0.050	-.0003563	3.215864
_cons	21.75074	3.289008	6.61	0.000	15.18019	28.32128

```
. adjustrcspline, at(foreign=0)
```

Predicted price for domestic cars with average repair status



Marginal effects

- Effect is how much does the predicted outcome change for a unit change in the explanatory variable.

Marginal effects

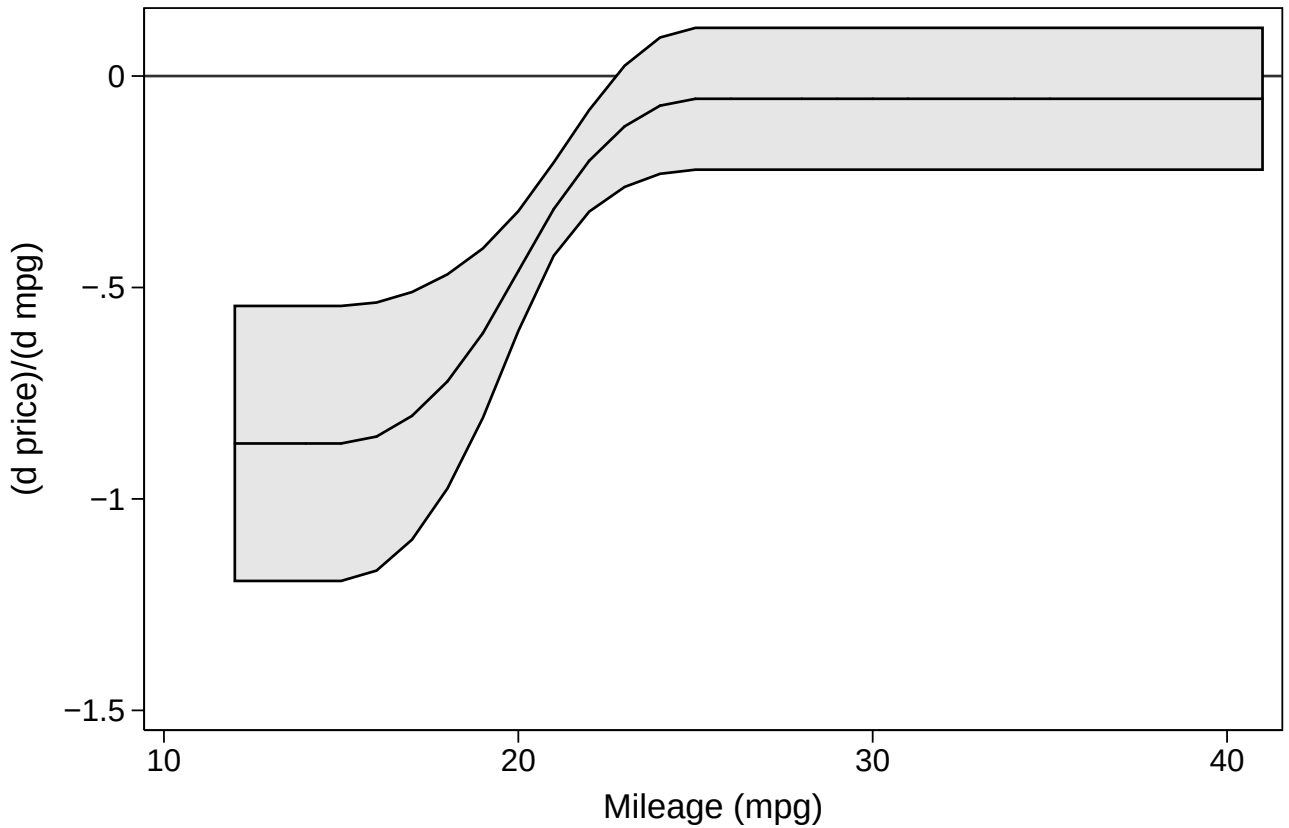
- Effect is how much does the predicted outcome change for a unit change in the explanatory variable.
- This is the first derivative.

Marginal effects

- Effect is how much does the predicted outcome change for a unit change in the explanatory variable.
- This is the first derivative.

```
. mfxrcspline, yline(0)
```

Change in predicted price for a unit change in mpg



Only `regress`?

- No, restricted cubic splines are just a transformation of an explanatory variable.

Only `regress`?

- No, restricted cubic splines are just a transformation of an explanatory variable.
- This transformed variable can be entered in any regression command like `logit` or `glm`.

Only `regress`?

- No, restricted cubic splines are just a transformation of an explanatory variable.
- This transformed variable can be entered in any regression command like `logit` or `glm`.
- This does influence how the adjusted prediction and marginal effects should be computed.

Only `regress`?

- No, restricted cubic splines are just a transformation of an explanatory variable.
- This transformed variable can be entered in any regression command like `logit` or `glm`.
- This does influence how the adjusted prediction and marginal effects should be computed.
- The `postrcspline` package will automatically recognize `regress`, `logit`, `logistic`, `betafit`, `probit`, `poisson`, `cloglog`, and `glm`.

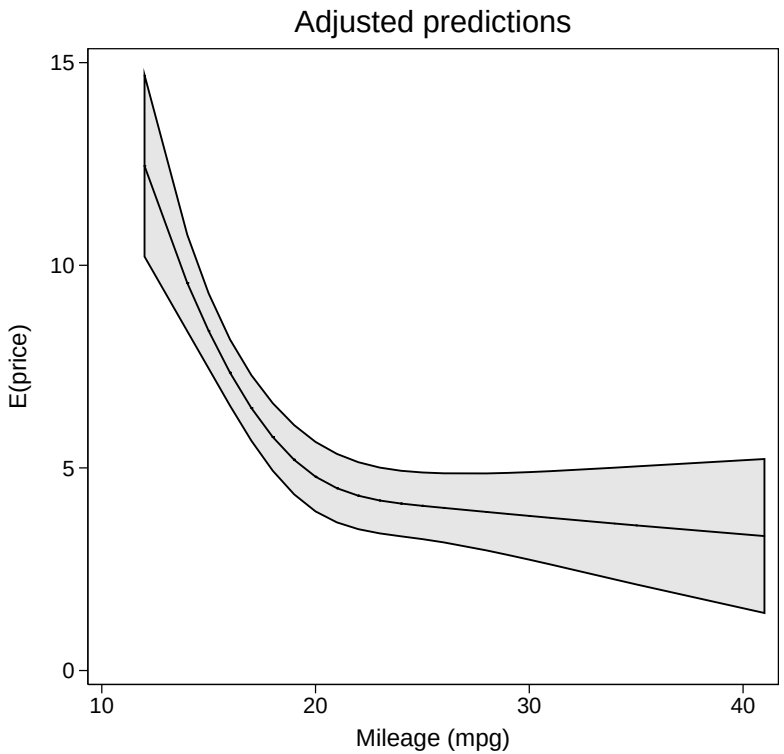
Example of a non-linear model (1)

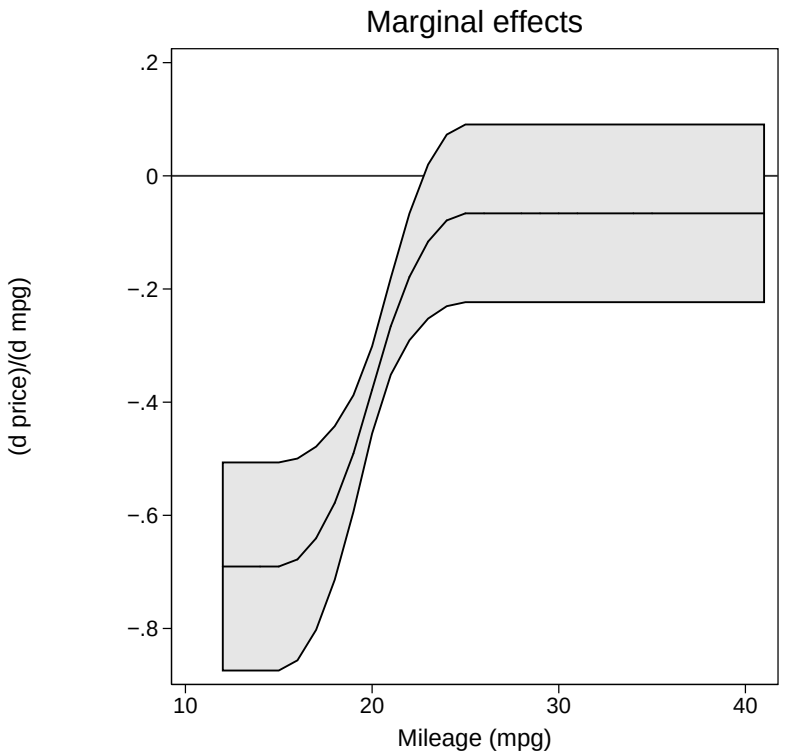
```
. glm price rc* rep78 foreign, link(log) eform
Iteration 0:   log likelihood = -154.66296
Iteration 1:   log likelihood = -151.66685
Iteration 2:   log likelihood = -151.50983
Iteration 3:   log likelihood = -151.50982
Generalized linear models                         No. of obs      =         69
Optimization     : ML                             Residual df     =         64
                                                  Scale parameter =   5.098444
Deviance         =  326.3004275                   (1/df) Deviance =   5.098444
Pearson          =  326.3004275                   (1/df) Pearson  =   5.098444
Variance function: V(u) = 1                       [Gaussian]
Link function    : g(u) = ln(u)                   [Log]
                                                  AIC             =   4.536517
Log likelihood   = -151.5098231                   BIC             =   55.31761
```

price	exp(b)	OIM Std. Err.	z	P>|z|	[95% Conf.	Interval]
rc1	.8763127	.0185517	-6.24	0.000	.8406961	.9134383
rc2	1.082826	.0224177	3.84	0.000	1.039767	1.127667
rep78	.9569288	.0559123	-0.75	0.451	.8533848	1.073036
foreign	1.445238	.2078801	2.56	0.010	1.090195	1.915907

```
. adjustrcspline, at(foreign=0) name(a) title(Adjusted predictions)
. mfxrcspline, at(foreign=0) yline(0) name(b) title(Marginal effects)
. graph combine a b, ysize(3)
```

Example of a non-linear model (2)





Syntax `adjustrcspline`

```
adjustrcspline [if] [in] , [ at(var = #[var =
#[...]])  link(linkname)
custominvlink(inv_link_specification)
ciopts(rarea_options) noci level(#)
lineopts(line_options) addplot(plot)
generate(newvar1 [newvar2 newvar3]) ]
```

Syntax `mfxrcspline`

```
mfxrcspline [if] [in] , [ at(var = #[var = #[...]])
link(linkname) customdydxb(dydxb_specification)
showknots ciopts(rarea_options) noci level(#)
lineopts(line_options) addplot(plot)
generate(newvar1 [newvar2 newvar3]) ]
```

conclusion

- Restricted cubic spline are an easy way of including an explanatory variable in a smooth non-linear way in a wide variety of models.

conclusion

- Restricted cubic spline are an easy way of including an explanatory variable in a smooth non-linear way in a wide variety of models.
- The `postrcspline` package provides tools for interpreting the results:
  - `adjustrcspline` graphs the adjusted predictions
  - `mfxrcspline` graphs the marginal effects

conclusion

- Restricted cubic spline are an easy way of including an explanatory variable in a smooth non-linear way in a wide variety of models.
- The `postrcspline` package provides tools for interpreting the results:
  - `adjustrcspline` graphs the adjusted predictions
  - `mfxrcspline` graphs the marginal effects
- These commands will work after `regress`, `logit`, `logistic`, `betafit`, `probit`, `poisson`, `cloglog`, and `glm`.