kinkyreg: Instrument-free inference for linear regression models with endogenous regressors

Sebastian Kripfganz[1] Jan F. Kiviet[2]

[1]University of Exeter Business School, Department of Economics, Exeter, UK

[2]University of Amsterdam, Amsterdam School of Economics, The Netherlands
&
Stellenbosch University, Department of Economics, Stellenbosch, South Africa

UK Stata Conference

September 11, 2020

```
ssc install kinkyreg
net install kinkyreg, from(http://www.kripfganz.de/stata/)
```

Instrument-based versus instrument-free inference

- Instrumental variables dominate the empirical literature on causal inference in linear regression models with endogenous regressors.
  - For valid inference under conventional asymptotics, instruments must be relevant and exogenous.
  - Weak instruments can lead to severe coefficient biases, poor approximations of the finite-sample distributions, and large size distortions of statistical tests.
  - Robust statistical inference in the presence of weak instruments is possible but usually leads to wide and often not very informative confidence intervals.
  - Literature overview: Stock, Wright, and Yogo (2002), Andrews and Stock (2007), and Andrews, Stock, and Sun (2019).

Instrument-based versus instrument-free inference

- Community-contributed Stata commands for weak-instruments tests and weak-instruments robust inference:
  - `ivreg2` (Baum, Schaffer, and Stillman, 2003, 2007),
  - `condivreg` (Moreira and Poi, 2003; Mikusheva and Poi, 2006),
  - `rivtest` (Finlay and Magnusson, 2009),
  - `weakivtest` (Pflueger and Wang, 2015),
  - `twostepweakiv` (Sun, 2018).
- The same features that make an instrument relevant can also be a source of a violation of the exogeneity condition (Hall, Rudebusch, and Wilcox, 1996).

Instrument-based versus instrument-free inference

- Exogeneity of an instrument necessitates that it is validly excluded from the structural model.
  - In just-identified models, this exclusion restriction is untestable in the standard instrumental variables framework.
  - Even in overidentified models, routinely used overidentification tests rely on the maintained assumption that at least as many instruments are validly excluded as there are endogenous regressors (Parente and Santos Silva, 2012).
- Alternative assumptions can be imposed to enable testing of the exclusion restrictions.

Instrument-based versus instrument-free inference

- We present the new `kinkyreg` Stata command for kinky least squares (KLS) estimation (Kiviet, 2020a,b) that does not rely on instrumental variables:
  - KLS analytically corrects the bias of OLS for all values of the endogeneity correlations on a specified grid.
  - Set identification is achieved by confining the admissible degree of regressor endogeneity within plausible bounds.
  - For a reasonably narrow range of endogeneity correlations, KLS confidence intervals are often narrower than those from 2SLS, in particular if instruments are weak, or other instrument-based methods (e.g. the approach of "plausibly exogenous" instruments by Conley, Hansen, and Rossi, 2012).
  - Exclusion restrictions are testable within the KLS framework.

Linear regression model

- Linear regression model with an endogenous regressor x_{1i} and exogenous regressors $\mathbf{x}_{2i}$ (all variables in deviations from their mean):

$$y_i = \beta_1 x_{1i} + \mathbf{x}_{2i}'\boldsymbol{\beta}_2 + \varepsilon_i,$$

with $\varepsilon_i \sim (0, \sigma_\varepsilon^2)$ and

$$\begin{pmatrix} x_{1i} \\ \mathbf{x}_{2i} \end{pmatrix} \sim \left(\begin{pmatrix} 0 \\ \mathbf{0} \end{pmatrix}, \begin{pmatrix} \sigma_1^2 & \boldsymbol{\sigma}_{12}' \\ \boldsymbol{\sigma}_{12} & \boldsymbol{\Sigma}_2 \end{pmatrix} \right).$$

  - The model can be generalized for multiple endogenous regressors (Kiviet, 2020a,b).
- OLS is inconsistent because $E[x_{1i}\varepsilon_i] = \rho\,\sigma_1\sigma_\varepsilon \neq 0$ for nonzero endogeneity correlations $Corr(x_{1i}, \varepsilon_i) = \rho \neq 0$.

Kinky least squares estimation

- While 2SLS uses orthogonality conditions $E[\mathbf{z}_i \varepsilon_i] = \mathbf{0}$, KLS utilizes the non-orthogonality condition $E[x_{1i} \varepsilon_i] = \rho \, \sigma_1 \sigma_\varepsilon$ (in addition to the orthogonality conditions for the exogenous regressors $\mathbf{x}_{2i}$).
- For a given correlation ρ, σ_ε can be consistently estimated as the square root of

$$\hat{\sigma}^2_\varepsilon(\rho) = \hat{\sigma}^2_{\varepsilon,OLS} \left(1 - \rho^2 \frac{\hat{\sigma}^2_1}{\hat{\sigma}^2_1 - \hat{\boldsymbol{\sigma}}'_{12} \hat{\boldsymbol{\Sigma}}_2^{-1} \hat{\boldsymbol{\sigma}}_{12}} \right)^{-1},$$

where $\hat{\sigma}^2_{\varepsilon,OLS} = N^{-1} \sum_{i=1}^N \hat{\varepsilon}^2_{i,OLS}$, with OLS residuals $\hat{\varepsilon}_{i,OLS}$.

  - The variance estimates $\hat{\sigma}^2_1$, $\hat{\boldsymbol{\sigma}}_{12}$, and $\hat{\boldsymbol{\Sigma}}_2$ are readily obtained from the data.

Kinky least squares estimation

- The KLS estimator corrects the inconsistency of the OLS estimator:

$$\begin{pmatrix} \hat{\beta}_1(\rho) \\ \hat{\boldsymbol{\beta}}_2(\rho) \end{pmatrix} = \begin{pmatrix} \hat{\beta}_{1,OLS} \\ \hat{\boldsymbol{\beta}}_{2,OLS} \end{pmatrix} - \frac{\rho\,\hat{\sigma}_1\hat{\sigma}_\varepsilon(\rho)}{\hat{\sigma}_1^2 - \hat{\boldsymbol{\sigma}}_{12}'\hat{\boldsymbol{\Sigma}}_2^{-1}\hat{\boldsymbol{\sigma}}_{12}} \begin{pmatrix} 1 \\ -\hat{\boldsymbol{\Sigma}}_2^{-1}\hat{\boldsymbol{\sigma}}_{12} \end{pmatrix}.$$

- Kiviet (2020a,b) derives an analytical expression for the variance-covariance matrix of the KLS estimator, $\sigma_\varepsilon^2\mathbf{V}(\rho, \kappa_x, \kappa_\varepsilon)$, as a function of the kurtosis of the regressors, κ_x, and the kurtosis of the error term, κ_ε.
  - Estimates of κ_x can be obtained from the data, and $\hat{\kappa}_\varepsilon(\rho) = N^{-1}\sum_{i=1}^{N}[\hat{\varepsilon}_i(\rho)/\hat{\sigma}_\varepsilon(\rho)]^4$, with KLS residuals $\hat{\varepsilon}_i(\rho)$.
  - For a tractable expression of $\mathbf{V}(\rho, \kappa_x, \kappa_\varepsilon)$, Kiviet (2020a,b) assumes an identical kurtosis κ_x for all regressors. By choosing $\hat{\kappa}_x$ as the maximum of the individual kurtosis estimates, we obtain (asymptotically) conservative confidence intervals.

Kinky least squares estimation

- The endogeneity correlation ρ is unknown but assumed to lie in the interval $\rho \in [r_l, r_u]$.
- The KLS estimator $\hat{\beta}(r)$ is computed for a range of values $r \in [r_l, r_u]$, subject to the feasibility bounds

$$|r| < \sqrt{1 - \frac{\hat{\boldsymbol{\sigma}}_{12}' \hat{\boldsymbol{\Sigma}}_2^{-1} \hat{\boldsymbol{\sigma}}_{12}}{\hat{\sigma}_1^2}} \leq 1.$$

- For a significance level α, the union of KLS confidence intervals over the range $r \in [r_l, r_u]$ has asymptotic coverage of at least $1 - \alpha$.

kinkyreg command syntax

`kinkyreg` *depvar* [*varlist1*] (*varlist2* [= *varlist_iv*]) [*if*] [*in*], [*options*]

- Basic command syntax similar to `ivregress`, but instrumental variables are optional:
- Main options (see the Stata help file for a full list):
  - `endogeneity`(*numlist*) to specify values for the fixed endogeneity correlations of the endogenous regressors *varlist2*,
  - `range`($\#_1$ $\#_2$) to specify the admissible endogeneity range,
  - `stepsize`(#) to specify the step size,
  - `small` to report small-sample statistics,
  - `inference`(*varlist*) to specify the variables for graphical KLS inference,
  - `lincom`(#: *exp*) to specify linear combinations for graphical KLS inference,
  - options to modify the appearance of the KLS graphs.

Specification tests

- Linear hypotheses tests (Wald/F tests) for $H_0 : \mathbf{R}\beta = \mathbf{c}$:

$$\hat{W}(r) = \left[\mathbf{R}\hat{\beta}(r) - \mathbf{c}\right]' \left[\hat{\sigma}^2_\varepsilon(r)\mathbf{V}(r, \hat{\kappa}_x, \hat{\kappa}_\varepsilon(r))\right]^{-1} \left[\mathbf{R}\hat{\beta}(r) - \mathbf{c}\right],$$

  with postestimation command `estat test`.

- Exclusion restriction tests for instrumental variables (or other variables) $\mathbf{x}_3$, i.e. $H_0 : \beta_3 = \mathbf{0}$ in the auxiliary model

$$y_i = \beta_1 x_{1i} + \mathbf{x}'_{2i}\beta_2 + \mathbf{x}'_{3i}\beta_3 + \varepsilon_i,$$

  with postestimation command `estat exclusion`.

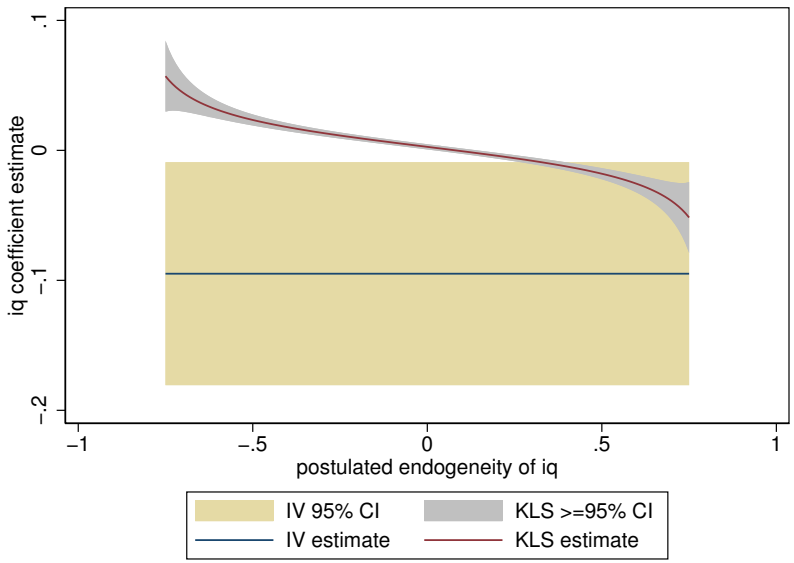
Specification tests

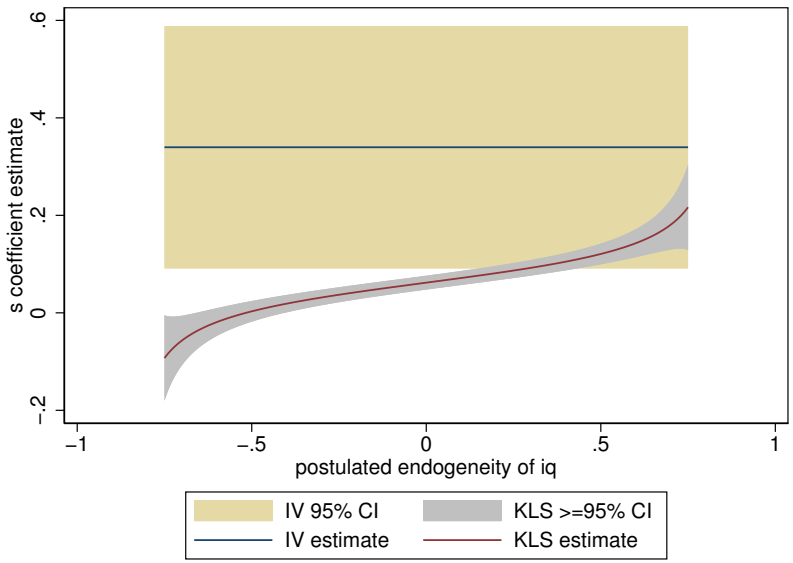
- Ramsey's RESET test, i.e. an exclusion restrictions test for higher-order polynomials of the fitted values or right-hand side variables, with postestimation command `estat reset`.
- Breusch-Pagan heteroskedasticity tests, with postestimation command `estat hettest`.
- Durbin's "alternative test" for serial correlation, with postestimation command `estat durbinalt`.
- All specification tests are computed over the same range of endogeneity correlations $r \in [r_l, r_u]$.

KLS versus 2SLS inference

Coefficient estimates and confidence intervals

```
. use http://www.stata-press.com/data/imeus/griliches
(Wages of Very Young Men, Zvi Griliches, J.Pol.Ec. 1976)

. set scheme s1color

. kinkyreg lw s expr tenure rns smsa _I* (iq = age mrt), range(-0.75 0.75) small inference(iq s)

Kinky least squares estimation                  Number of obs   =       758
```





- Assuming mild to moderate measurement error as the source of endogeneity, a plausible choice might be $r \in [-0.4, 0]$.

KLS inference

Coefficient estimates and confidence intervals

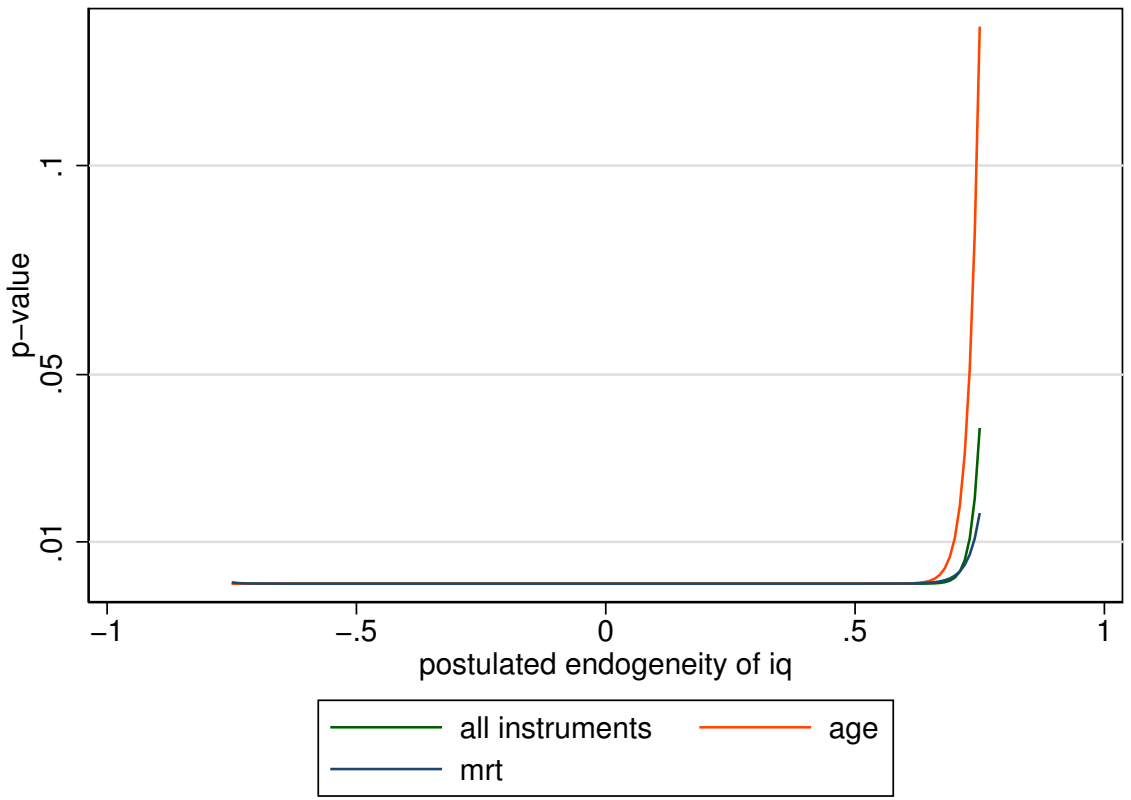
- KLS regression output for specific endogeneity correlations can be easily obtained using the *replay* syntax with the `correlation(#)` option.

```
. kinkyreg, correlation(-0.4)

Kinky least squares estimation                    Number of obs   =        758

Postulated endogeneity of iq = -0.4000
------------------------------------------------------------------------------
          lw |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          iq |   .0178505   .0015908    11.22   0.000     .0147275    .0209735
           s |    .018874   .0090115     2.09   0.037      .001183     .036565
        expr |    .036647   .0073454     4.99   0.000     .0222269    .0510672
      tenure |   .0355367   .0084409     4.21   0.000      .018966    .0521074
         rns |  -.0527647   .0312384    -1.69   0.092    -.1140905    .0085611
        smsa |   .1196815   .0299368     4.00   0.000      .060911     .178452
   _Iyear_67 |  -.0638234   .0538705    -1.18   0.236    -.1695794    .0419327
   _Iyear_68 |   .0872164   .0505387     1.73   0.085    -.0119988    .1864316
   _Iyear_69 |   .1878763   .0494006     3.80   0.000     .0908953    .2848573
   _Iyear_70 |   .1661179    .055196     3.01   0.003     .0577597    .2744761
   _Iyear_71 |   .1882715    .048602     3.87   0.000     .0928583    .2836846
   _Iyear_73 |   .3048592   .0457922     6.66   0.000      .214962    .3947564
       _cons |   3.255792   .1407933    23.12   0.000     2.979394    3.532191
------------------------------------------------------------------------------
```

KLS inference

Exclusion restriction tests for the instrumental variables

```
. estat exclusion, twoway(, ylabel(0.01 0.05 0.1, grid)) notable
```

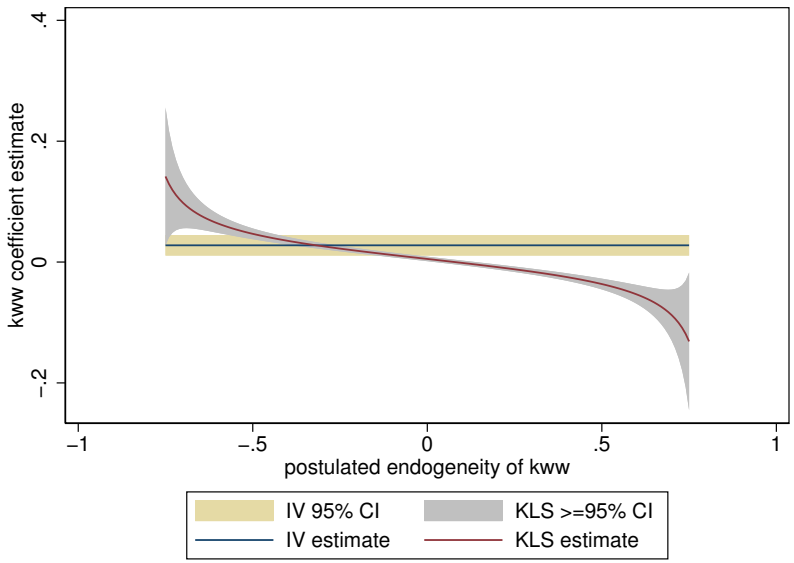


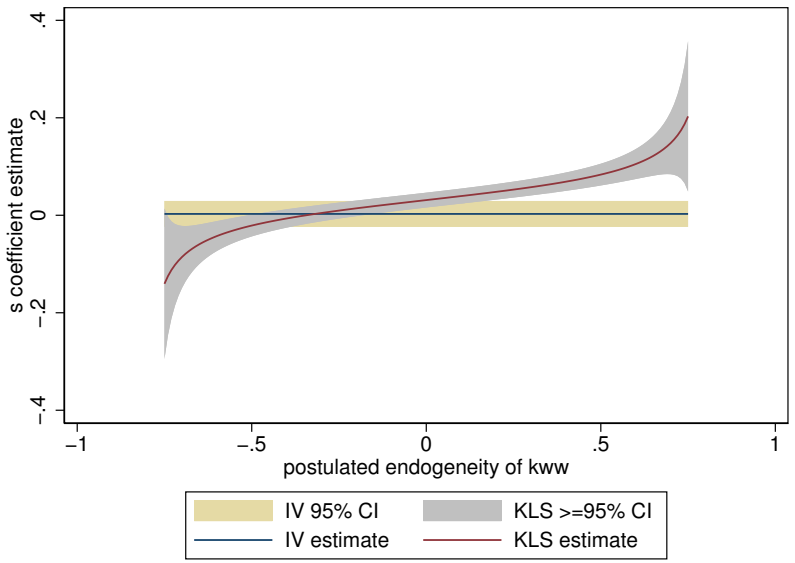
- The null hypothesis that instruments are validly excluded is rejected for plausible values of ρ.

KLS versus 2SLS inference

Coefficient estimates and confidence intervals

```
. kinkyreg lw s expr tenure rns smsa _I* age mrt (kww = iq), range(-0.75 0.75) small inference(kww s)

Kinky least squares estimation                  Number of obs   =       758
```



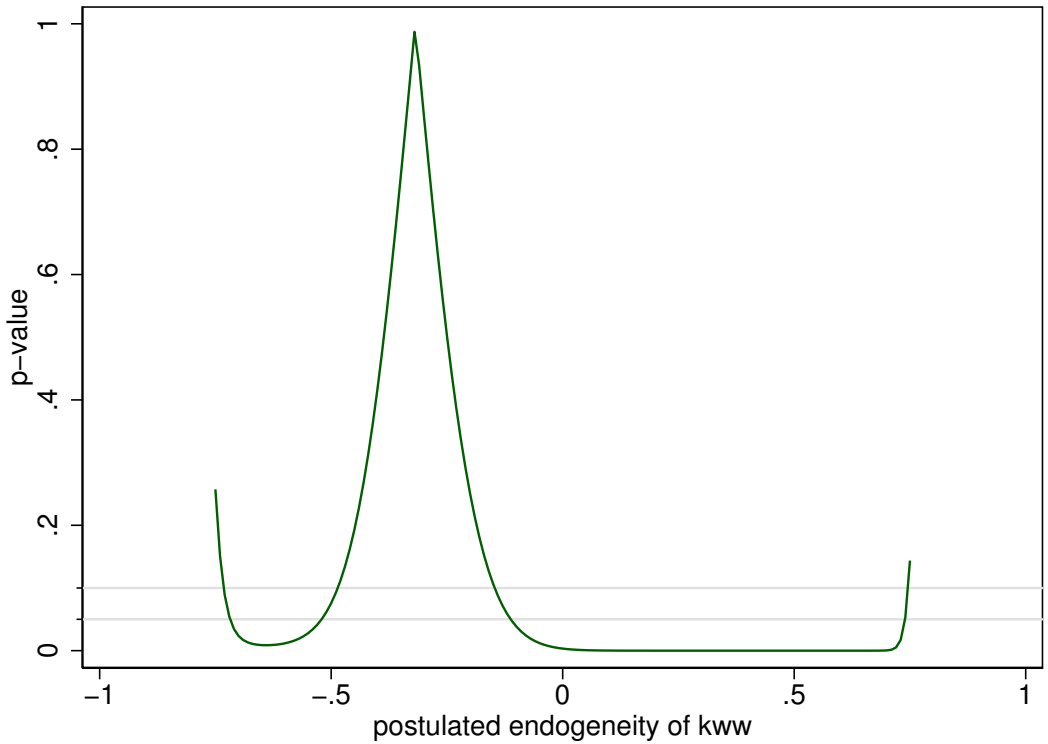


- Modified model yields more reasonable 2SLS results.

KLS inference

Exclusion restriction test for the instrumental variable

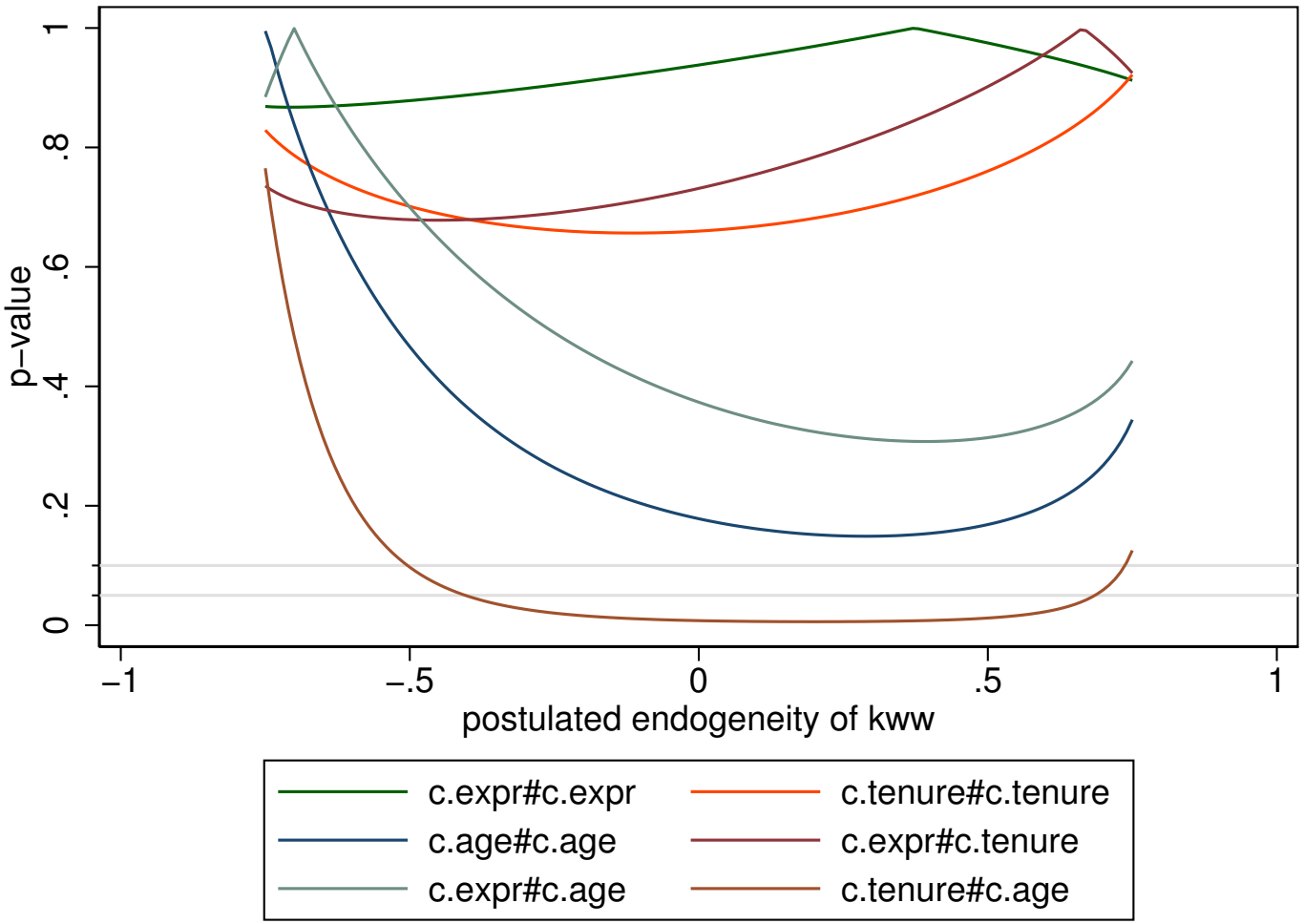
```
. estat exclusion, twoway(, ymtick(0.05 0.1, grid))

Endogeneity of kww compatible with valid exclusion
------------------------------------------------
             |     Corr.  [95% Confid. Bounds]
-------------+----------------------------------
          iq | -.3183786  -.5207143  -.1120693
------------------------------------------------
```



KLS inference

Exclusion restriction tests for interaction terms

```
. estat exclusion c.expr#c.expr c.tenure#c.tenure c.age#c.age c.expr#c.tenure c.expr#c.age
> c.tenure#c.age, twoway(, ymtick(0.05 0.1, grid)) nojoint notable
```

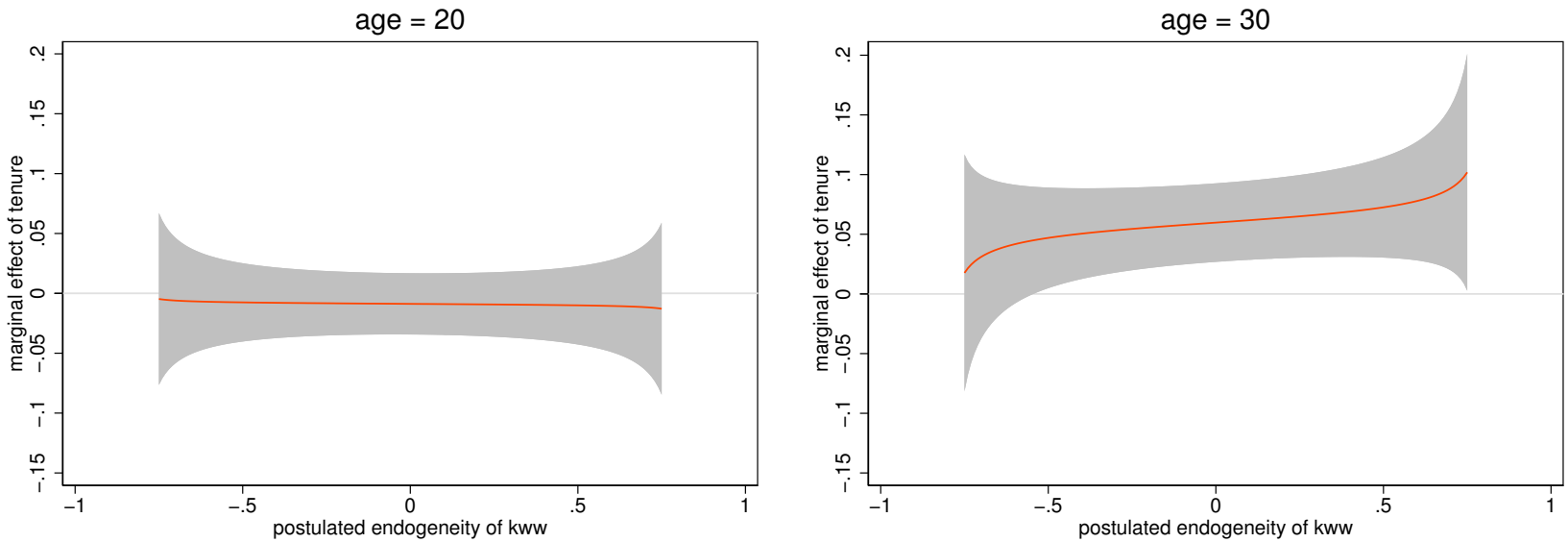


KLS inference

Marginal effects

- KLS inference for the marginal effects of `tenure`, $\beta_{\texttt{tenure}} + \beta_{\texttt{c.tenure\#c.age}} \cdot \text{age}$, for selected values of age can be produced with the `lincom()` option.

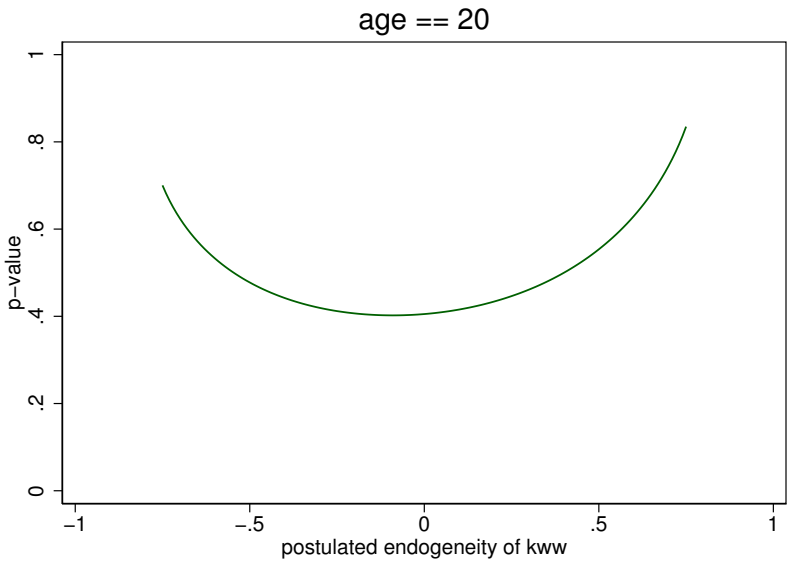
```
. kinkyreg lw s expr tenure rns smsa _I* age mrt c.tenure#c.age (kww), range(-0.75 0.75) small
> lincom(1: tenure+c.tenure#c.age*20) lincom(2: tenure+c.tenure#c.age*30)
> twoway(, ytitle("marginal effect of tenure") ylabel(-0.15(0.05)0.2) ytick(0, grid) legend(off))
> twoway(1, title("age = 20")) twoway(2, title("age = 30"))
```

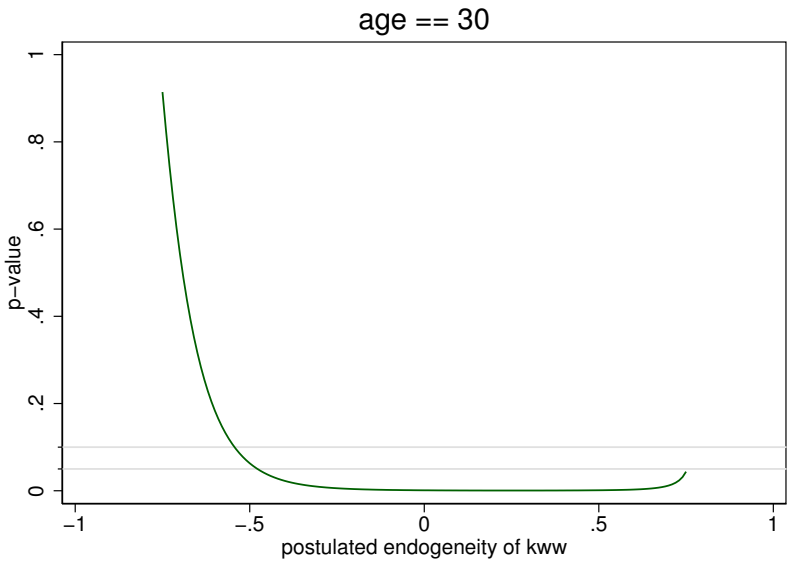


KLS inference

Linear hypothesis tests

```
. estat test tenure+c.tenure#c.age*20=expr, twoway(, title("age == 20") ylabel(0(0.2)1)
> name(kinkyreg_test1))

. estat test tenure+c.tenure#c.age*30=expr, twoway(, title("age == 30") ymtick(0.05 0.1, grid)
> name(kinkyreg_test2))
```



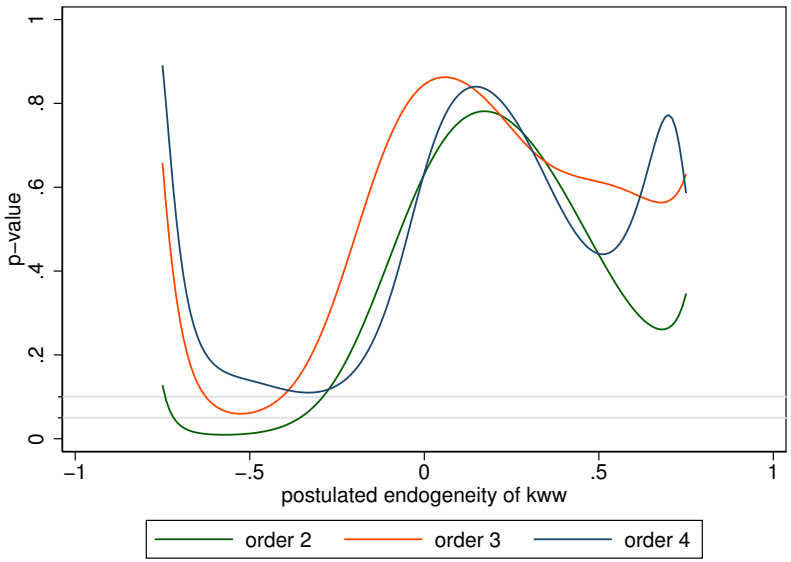


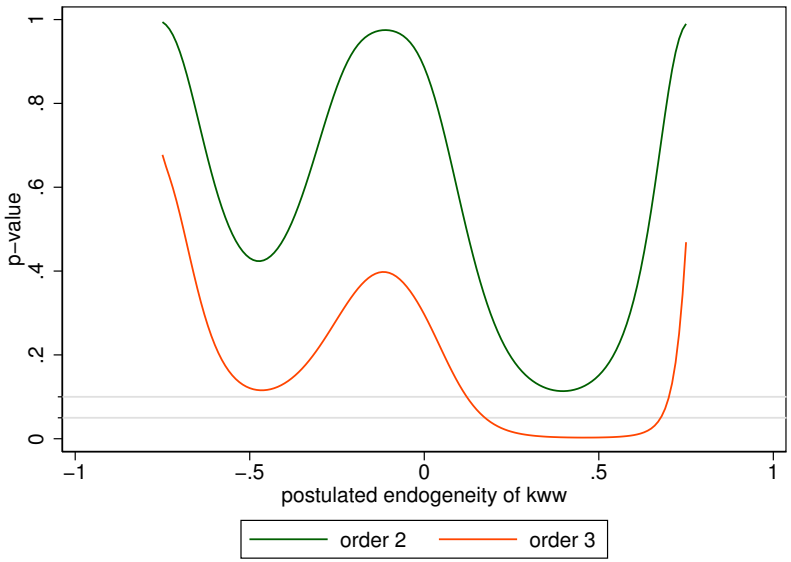
- The null hypothesis that marginal effects of `tenure` and `expr` are equal is rejected for higher ages.

KLS specification tests

Ramsey's regression equation specification error tests (RESET)

```
. estat reset, twoway(, ymtick(0.05 0.1, grid) legend(rows(1)) name(kinkyreg_reset_xb))

. estat reset, rhs order(2 3) twoway(, ymtick(0.05 0.1, grid) name(kinkyreg_reset_rhs))
```



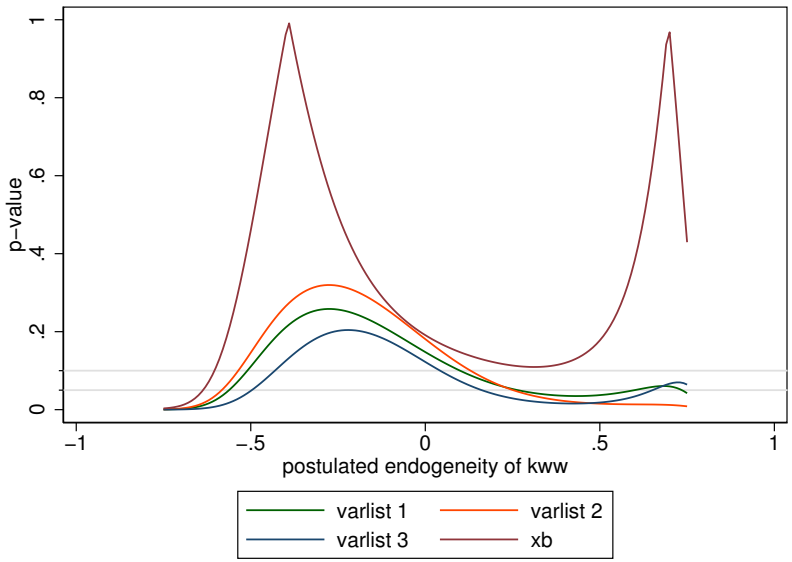


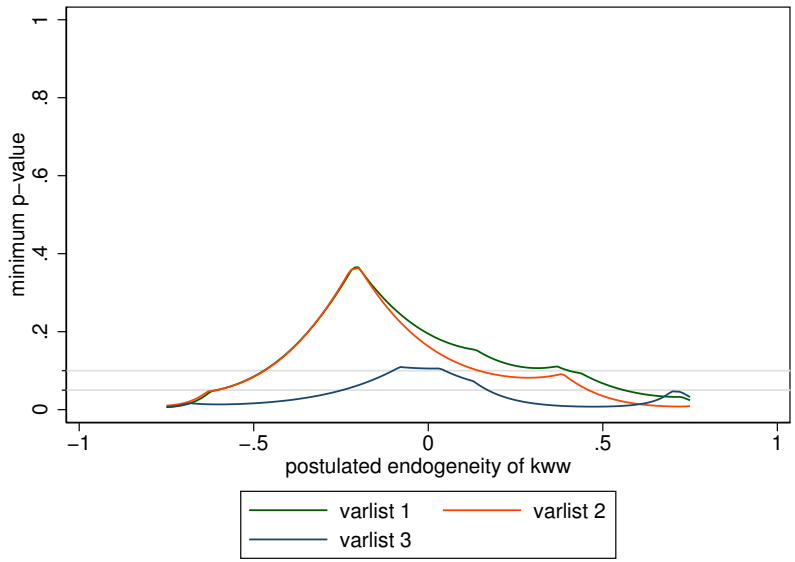
- RESET is an exclusion restrictions test for powers of the endogeneity-corrected fitted values (default option `xb`) or right-hand side variables (option `rhs`).

KLS specification tests

Breusch-Pagan heteroskedasticity tests

```
. estat hettest () (iq) (c.expr#c.expr c.tenure#c.tenure c.age#c.age c.expr#c.tenure c.expr#c.age),
> xb rhs twoway(, ymtick(0.05 0.1, grid) name(kinkyreg_hett))

. estat hettest () (iq) (c.expr#c.expr c.tenure#c.tenure c.age#c.age c.expr#c.tenure c.expr#c.age),
> rhs minp twoway(, ylabel(0(0.2)1) ymtick(0.05 0.1, grid) name(kinkyreg_hett_minp))
```



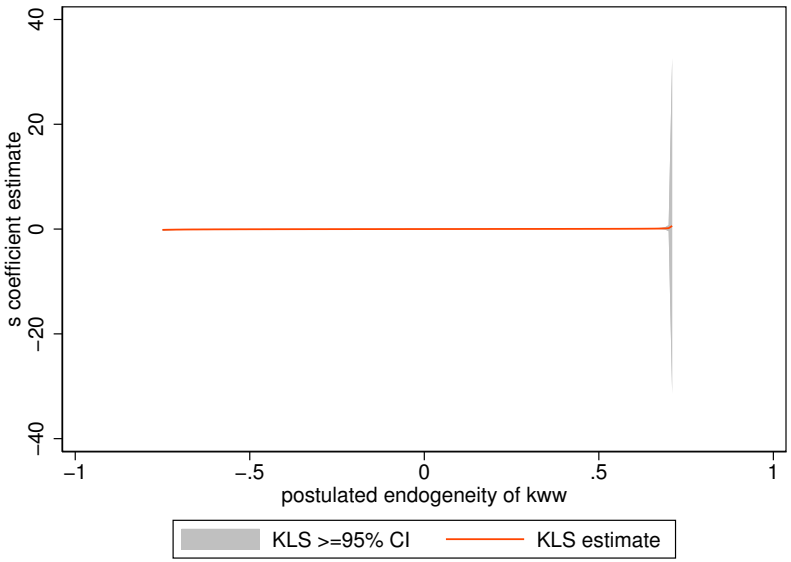


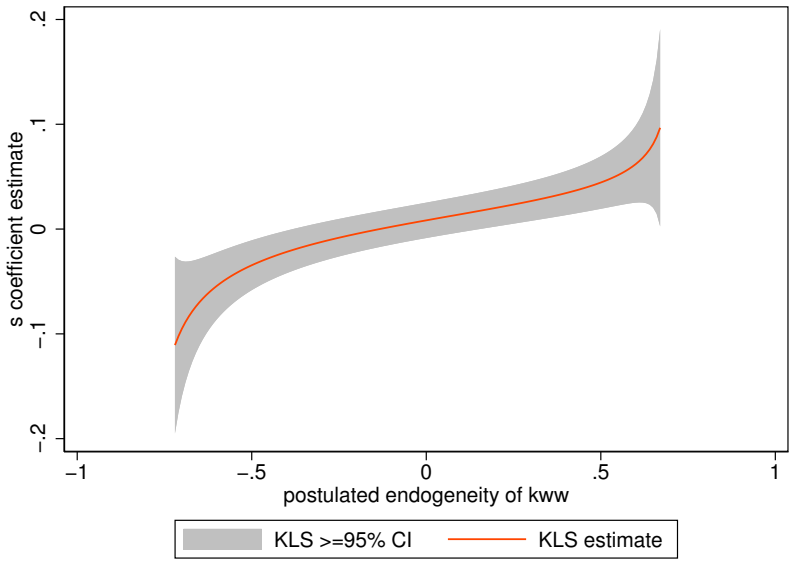
- Null hypothesis: There is no conditional heteroskedasticity.
- The `minp` option displays the minimum p-value among individual significance tests.

KLS inference with 2 endogenous regressor

Two-dimensional slice from a three-dimensional surface

```
. kinkyreg lw s expr tenure rns smsa _I* age mrt c.tenure#c.age (iq kww), endogeneity(-0.2 .)
> range(-0.75 0.75) small inference(s) twoway(s, name(kinkyreg_s1))

. kinkyreg lw s expr tenure rns smsa _I* age mrt c.tenure#c.age (iq kww), endogeneity(-0.2 .)
> range(-0.75 0.75) small inference(s) twoway(s, yrange(-0.2 0.2) name(kinkyreg_s2))
```





- The `endogeneity()` option must be used to fix all but one endogeneity correlations.
- The displayed range can be restricted with the `yrange()` suboption.

KLS inference with 2 endogenous regressors

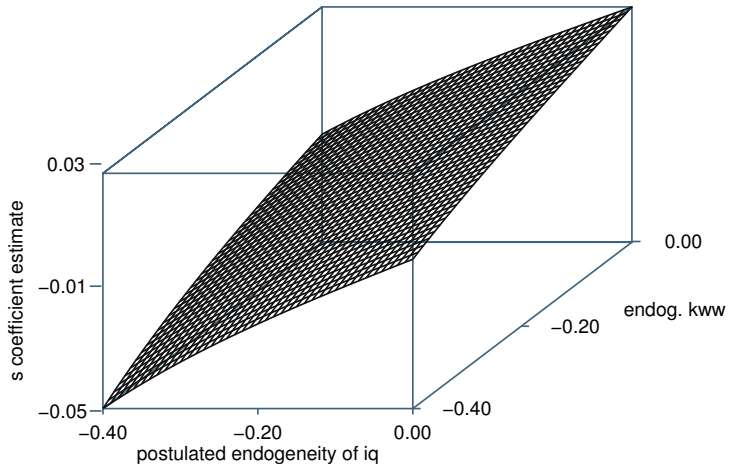
Surface and contour plots

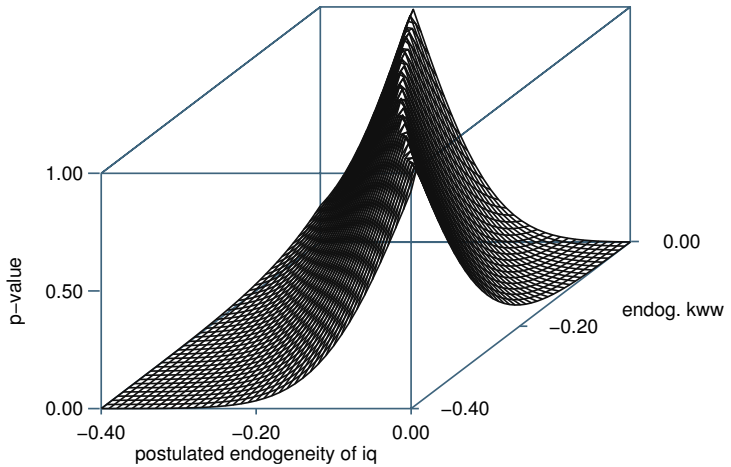
```
. set scheme s1mono

. forvalues endo = -40 / 0 {
  2. quietly kinkyreg lw s expr tenure rns smsa _I* age mrt c.tenure#c.age (iq kww),
> endogeneity(‘=‘endo’/100’ .) range(-0.4 0) small nograph
  3. matrix b = (nullmat(b), e(b_kls)[., "s"])
  4. estat test s, nograph
  5. matrix p = (nullmat(p), r(p))
  6. }

. frame create kinkyreg
. frame change kinkyreg
. quietly svmat double b, name(s)
. svmat double p, name(p)
. generate double endo_kww = -0.4 + 0.01 * (_n - 1)
. quietly reshape long s p, i(endo_kww) j(endo_iq)
. recast double endo_iq
. quietly replace endo_iq = (endo_iq - 41) / 100
. label var s "s coefficient estimate"
. label var p "p-value"
. label var endo_kww "postulated endogeneity of kww"
. label var endo_iq "postulated endogeneity of iq"
. surface endo_iq endo_kww s, plotregion(lpattern(blank)) ytitle(endog. kww) nodraw name(surface_s)
. surface endo_iq endo_kww p, plotregion(lpattern(blank)) ytitle(endog. kww) nodraw name(surface_p)
. twoway contour s endo_kww endo_iq, ccuts(-0.06(0.01)0.03) nodraw name(contour_s)
. twoway contour p endo_kww endo_iq, ccuts(0.01 0.05 0.1) nodraw name(contour_p)
. graph combine surface_s surface_p contour_s contour_p, altshrink
```

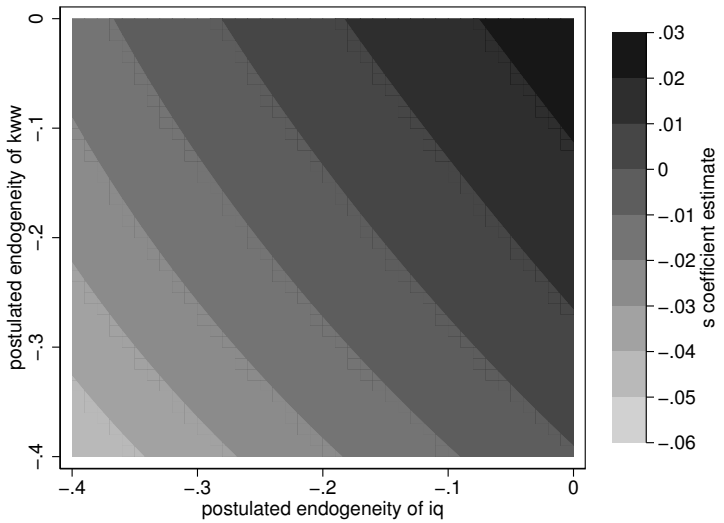
- `surface` is community contributed (Mander, 1999).

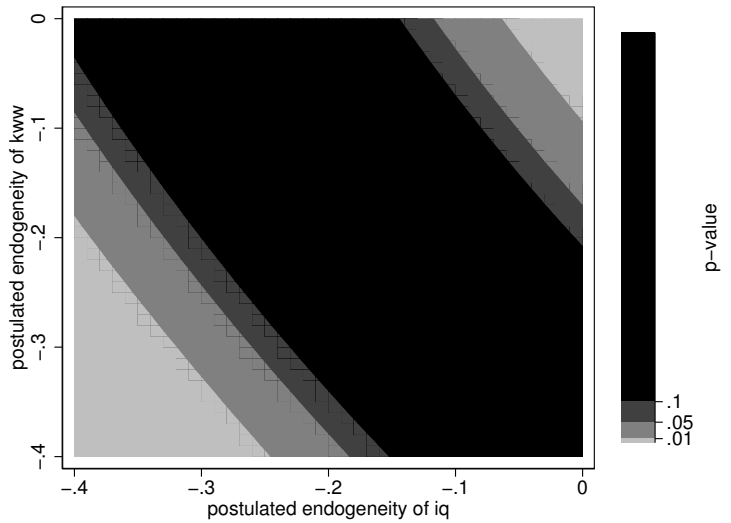
KLS inference with 2 endogenous regressors

Surface and contour plots for the return to schooling and its *p*-value









Conclusion

- The `kinkyreg` package provides new tools for (primarily graphical) instrument-free inference in linear regression models with an arbitrary number of endogenous regressors.
- KLS inference can facilitate sensitivity checks for instrument-based procedures.
- The KLS approach often yields narrower confidence intervals than instrument-based approaches, but is only as good as the chosen bounds for the admissible endogeneity correlations. It is thus often reasonable to consider KLS as a complement rather than a substitute to instrument-based procedures.

```
ssc install kinkyreg
net install kinkyreg, from(http://www.kripfganz.de/stata/)

help kinkyreg
help kinkyreg postestimation
```

References

- Andrews, D. W. K., and J. H. Stock (2007). Inference with weak instruments. In *Advances in Economics and Econometrics: Theory and Applications, Ninth World Congress*, ed. R. Blundell, W. K. Newey, and T. Persson, chap. 6, 122–173, Vol. 3, Cambridge University Press.
- Andrews, I., J. H. Stock, and L. Sun (2019). Weak instruments in instrumental variables regression: Theory and practice. *Annual Review of Economics 11*: 727–753.
- Baum, C. F., M. E. Schaffer, and S. Stillman (2003). Instrumental variables and GMM: Estimation and testing. *Stata Journal 3* (1): 1–31.
- Baum, C. F., M. E. Schaffer, and S. Stillman (2007). Enhanced routines for instrumental variables/generalized method of moments estimation and testing. *Stata Journal 7* (4): 465–506.
- Conley, T. G., C. B. Hansen, and P. E. Rossi (2012). Plausibly exogenous. *Review of Economics and Statistics 94* (1): 260–272.
- Finlay, K., and L. M. Magnusson (2009). Implementing weak-instrument robust tests for a general class of instrumental-variables models. *Stata Journal 9* (3): 398–421.
- Hall, A. R., G. D. Rudebusch, and D. W. Wilcox (1996). Judging instrument relevance in instrumental variables estimation. *International Economic Review 37* (2): 283–298.
- Kiviet, J. F. (2020a). Testing the impossible: Identifying exclusion restrictions. *Journal of Econometrics 218* (2): 294–316.
- Kiviet, J. F. (2020b). Instrument-free inference under confined regressor endogeneity; derivations and applications. Stellenbosch Economic Working Papers WP09/2020, Department of Economics, University of Stellenbosch.

References

- Mander, A. (1999). 3D surface plots. *Stata Technical Bulletin 51*: 7–10.
- Mikusheva, A., and B. P. Poi (2006). Tests and confidence sets with correct size when instruments are potentially weak. *Stata Journal 6* (3): 335–347.
- Moreira, M. J., and B. P. Poi (2003). Implementing tests with correct size in the simultaneous equations model. *Stata Journal 3* (1): 57–70.
- Parente, P. M. D. C., and J. M. C. Santos Silva (2012). A cautionary note on tests of overidentifying restrictions. *Economics Letters 115* (2): 314–317.
- Pflueger, C. E., and S. Wang (2015). A robust test for weak instruments in Stata. *Stata Journal 15* (1): 216–225.
- Stock, J. H., J. H. Wright, and M. Yogo (2002). A survey of weak instruments and weak identification in generalized method of moments. *Journal of Business & Economic Statistics 20* (4): 518–529.
- Sun, L. (2018). Implementing valid two-step identification-robust confidence sets for linear instrumental-variables models. *Stata Journal 18* (4): 803–825.