The Stata Journal (2018)
18, Number 1, pp. 174–183

heckroccurve: ROC curves for selected samples

Jonathan A. Cook
Public Company Accounting Oversight Board (PCAOB)
Washington, DC
jacook@uci.edu

Ashish Rajbhandari
Investment Strategy Group
The Vanguard Group
Malvern, PA
ashish_rajbhandari@vanguard.com

Abstract. Receiver operating characteristic (ROC) curves can be misleading when they are constructed with selected samples. In this article, we describe `heckroccurve`, which implements a recently developed procedure for plotting ROC curves with selected samples. The command estimates the area under the ROC curve and a graphical display of the curve. A variety of plot options are available, including the ability to add confidence bands to the plot.

Keywords: st0518, heckroccurve, receiver operating characteristic curves, ROC curves, classifier evaluation, sample-selection bias

1 Introduction

Receiver operating characteristic (ROC) curves are widely used in many fields to measure the performance of ratings. An advantage of ROC curves over metrics like accuracy (defined as the portion of cases correctly predicted) is that ROC curves provide the full range of tradeoffs between true positives and false negatives. Despite their widespread use, the effects of sample selection on ROC curves was not explored until recently.

Sample selection is common in many areas. Consider a medical test administered only to patients that are referred by their physicians. We want to know how well the test correctly diagnoses illness, but we observe test results only for referred patients. A different but related problem arises in commercial banking. The Basel Accords require banks to estimate the probability of default for their loans. To assess the predictive performance of their probability of default models, banks could construct a ROC curve with the sample of loan applicants that were granted loans.

Hand and Adams (2014) and Kraft, Kroisandt, and Müller (2014) appear to have been the first to discuss selection bias for ROC curves. Cook (2017) presents a procedure to plot a ROC curve that is a consistent estimate of the ROC curve that would be obtained with a random sample. The `heckroccurve` command implements Cook's procedure and provides confidence intervals for the area under the curve and confidence bands for the ROC curve.

 st0518

There are many existing Stata commands for plotting ROC curves, including `rocreg`, `roctab`, and `roccomp`, but none of these commands correct for the effects of sample selection. The syntax of `heckroccurve` was kept close to existing Stata commands for sample-selection problems, that is, `heckman`, `heckprobit`, and `heckoprobit`. Like `heckprobit` and `heckoprobit`, `heckroccurve` is based on assumptions similar to those of Heckman (1976). The output from `heckroccurve` was designed to be similar to that of Stata's built-in commands for ROC curves.

The next section describes the procedure performed by `heckroccurve`. Sections 3 and 4 provide the command's syntax and examples: The first example in section 4 illustrates the syntax. The second example in section 4 shows how selection can affect ROC curves. The dataset used for this second example is provided with `heckroccurve`. Section 5 concludes.

2 ROC curves for selected samples

We assume that each observation belongs to one of two classes (for example, positive and negative). Our task is to evaluate how well our ordinal rating predicts class. Given a threshold, we could predict that all observations with a rating value above the threshold are positive, and all observations below the threshold are negative. To see how well the rating with the threshold predicts class, we define sensitivity and specificity as

$$\text{Sensitivity} = \frac{\text{TP}}{P}, \quad \text{and} \tag{1}$$

$$\text{Specificity} = \frac{\text{TN}}{N} \tag{2}$$

where the confusion matrix in table 1 defines true positives (TP), true negatives (TN), positives (P), and negatives (N).

Table 1. Confusion matrix

		truth	
		positive	**negative**
prediction	**positive**	True Positives (TP)	False Positives (FP)
	negative	False Negatives (FN)	True Negatives (TN)
	total	Positives (P)	Negatives (N)

ROC curves, which plot sensitivity as a function of specificity for all possible thresholds, illustrate a rating's tradeoff between true positives and false negatives. A higher value of sensitivity for a given value of specificity indicates better performance. The area under the ROC curve (AUC) is a common metric for evaluating a rating's performance. If the rating has no connection to the true class, the expected AUC would be 0.5. An excellent introduction to ROC curves is provided by Fawcett (2006).

2.1 Notation and setup

We denote the rating's output as a_i for each observation i. The unobserved propensity to be a positive case is denoted as p_i. The true outcome is

$$\text{outcome}_i = \begin{cases} \text{positive} & \text{if } p_i > p^* \\ \text{negative} & \text{otherwise} \end{cases}$$

where p^* is the threshold for an instance to be a positive case. We assume that p_i follows a standard normal distribution. The modeler never observes p_i, only outcome$_i$. For a given threshold c, we can give probabilistic definitions of sensitivity and specificity:

$$\text{Sensitivity} = \text{Prob}(a_i > c \,|\, p_i > p^*), \quad \text{and} \tag{3}$$
$$\text{Specificity} = \text{Prob}(a_i \le c \,|\, p_i \le p^*) \tag{4}$$

Evaluating (1) and (2) with the sample at hand provides estimates of these probabilities.

The selection rule is

$$\begin{cases} \text{Selected} & \text{if } b_i \equiv \boldsymbol{\delta}\, X_i + \gamma\, a_i + \varepsilon_i > s \\ \text{Not selected} & \text{otherwise} \end{cases} \tag{5}$$

where s is a constant, X_i is a vector of variables, and ε_i is a standard normal random variable. The parameter $\boldsymbol{\delta}$ is a vector of coefficients, and γ indicates the degree to which the rating was incorporated into the selection process. These parameters can be estimated from a probit regression of selection on X and a. If the vector X does not contain a constant, then the intercept from the probit regression would provide an estimate of $-s$. The procedure that we describe here does not require an estimate of s.

We denote sensitivity and specificity conditional on selection as

$$\text{Sensitivity} \,|\, \text{Selection} = \text{Prob}(a_i > c \,|\, p_i > p^*, b_i > s), \quad \text{and} \tag{6}$$
$$\text{Specificity} \,|\, \text{Selection} = \text{Prob}(a_i \le c \,|\, p_i \le p^*, b_i > s) \tag{7}$$

When data are chosen according to (5), the values in (1) and (2) provide estimates of (6) and (7) instead of (3) and (4). It is possible that the ROC curve implied by (6) and (7) differs greatly from the curve implied by (3) and (4).

2.2 Procedure for creating ROC curves with selected samples

Cook's (2017) procedure for creating ROC curves with selected samples infers the predictive power of the classifier (taking selection into consideration), then draws the implied

ROC curve. After standardizing the classifier's output, the likelihood for the data can be expressed as

$$L = \prod_i \Phi_2\{\boldsymbol{\delta}\, X_i + \gamma\, a_i - s,\, -(\beta_0 + \beta_1 a_i)\ ; \rho_{\varepsilon p}\}^{\mathbb{1}(\text{outcome}_i=\text{positive})}$$
$$\times\, \Phi_2(\boldsymbol{\delta}\, X_i + \gamma\, a_i - s,\, \beta_0 + \beta_1 a_i\ ; -\rho_{\varepsilon p})^{\mathbb{1}(\text{outcome}_i=\text{negative})}$$
$$\times\, \Phi\{-(\boldsymbol{\delta}\, X_i + \gamma\, a_i - s)\}^{\mathbb{1}(\text{outcome}_i=\text{NA})} \tag{8}$$

where $\mathbb{1}(\cdot)$ is the indicator function,

$$\beta_0 \equiv \frac{p^*}{\sqrt{1-\rho_{ap}^2}}, \quad \text{and}$$
$$\beta_1 \equiv -\frac{\rho_{ap}}{\sqrt{1-\rho_{ap}^2}}$$

This likelihood function contains two correlations: ρ_{ap} and $\rho_{\varepsilon p}$. The correlation between the rating's output and p_i, denoted ρ_{ap}, is crucial for determining the strength of the classifier. The likelihood also contains $\rho_{\varepsilon p}$, which is the correlation between the unobserved component of the selection rule (that is, ε_i) and p_i.

The likelihood function in (8) is the same likelihood derived by Van de Ven and Van Praag (1981), which `heckprobit` maximizes. To take advantage of `heckprobit`'s many built-in features, `heckroccurve`'s maximum likelihood estimation is performed by calling `heckprobit`. Estimates of p^* and ρ_{ap} are found by applying the appropriate transformations to the estimates of β_0 and β_1.

To draw the ROC implied by the estimates of p^* and ρ_{ap} (denoted here as $\widehat{p^*}$ and $\widehat{\rho_{ap}}$), we begin with a set of cutoffs with a sufficiently large range (`heckroccurve` uses -4 to 4). For each cutoff $c \in [-4, 4]$, we find the corresponding value of sensitivity as

$$\text{Prob}(a_i > c \,|\, p_i > p^*) \approx \left\{1 - \Phi\left(\widehat{p^*}\right)\right\}^{-1} \int_c^\infty \phi(a)$$
$$\left[1 - \Phi\left\{\left(\widehat{p^*} - \widehat{\rho_{ap}}\, a\right) / \sqrt{1 - \widehat{\rho_{ap}}^{\,2}}\right\}\right] da$$

and specificity as

$$\text{Prob}(a_i < c \,|\, p_i < p^*) \approx \Phi\left(\widehat{p^*}\right)^{-1} \int_{-\infty}^{c} \phi(a)\, \Phi\left\{\left(\widehat{p^*} - \widehat{\rho_{ap}}\, a\right) / \sqrt{1 - \widehat{\rho_{ap}}^{\,2}}\right\} da$$

Confidence intervals and bands are obtained from confidence intervals for the maximum likelihood estimates of β_0 and β_1 and the functional invariance property of maximum likelihood.

3 The heckroccurve command

3.1 Syntax

`heckroccurve` *refvar classvar* [*if*] [*in*] [*weight*] `,`
`select(`[*depvar_s* `=`] *varlist_s*`)` [`collinear table level(`*#*`) noci cbands`
`noempirical nograph norefline irocopts(`*cline_options*`)`
`erocopts(`*cline_options*`) rlopts(`*cline_options*`) cbands(`*cline_options*`)`
twoway_options `vce(`*vcetype*`) robust` *maximize_options*]

3.2 Options

`select(`[*depvar_s* `=`] *varlist_s*`)` specifies the selection equation, dependent and independent variables, and whether to have a constant term and offset variable. `select()` is required.

`collinear` keeps collinear variables.

`table` displays the raw data in a $2 \times k$ contingency table.

`level(`*#*`)` specifies the confidence level, as a percentage, for confidence intervals. The default is `level(95)` or as set by `set level`; see [U] **20.8 Specifying the width of confidence intervals**.

`noci` does not display confidence intervals for the inferred AUC.

`cbands` displays confidence bands for the inferred ROC curve.

`noempirical` does not include the empirical ROC curve in the plot.

`nograph` suppresses graphical output.

`norefline` does not include a reference line in the plot.

`irocopts(`*cline_options*`)` affects rendition of the inferred ROC curve; see [G-3] ***cline_options***.

`erocopts(`*cline_options*`)` affects rendition of the empirical ROC curve; see [G-3] ***cline_options***.

`rlopts(`*cline_options*`)` affects rendition of the reference line; see [G-3] ***cline_options***.

`cbands(`*cline_options*`)` affects rendition of the confidence bands.

twoway_options are any of the options documented in [G-3] ***twoway_options***, excluding `by()`.

`vce(`*vcetype*`)` specifies the type of standard error reported, which includes types that are derived from asymptotic theory (`oim`, `opg`), that are robust to some kinds of misspecification (`robust`), that allow for intragroup correlation (`cluster` *clustvar*), and that use bootstrap or jackknife methods (`bootstrap`, `jackknife`); see [R] ***vce_option***.

`robust` is the synonym for `vce(robust)`.

maximize_options: `difficult`, `technique(`*algorithm_spec*`)`, `iterate(#)`, `tolerance(#)`, `ltolerance(#)`, `nrtolerance(#)`, `nonrtolerance`, `from(`*init_specs*`)`; see [R] **maximize**. These options are seldom used.

4 Examples

Example 1: Illustration of syntax

Our first example illustrates the command's syntax. We begin by loading Mroz's (1987) well-known dataset on women's wages and creating a binary variable:

```
. use http://fmwww.bc.edu/ec-p/data/wooldridge/mroz
. * Creating a binary variable to demonstrate procedure
. generate high_wage = 0 if inlf
(325 missing values generated)
. replace high_wage = 1 if wage > 2.37 & inlf
(311 real changes made)
```

If we want to see how years of education, `educ`, predicts "high wage", we can type the syntax that follows. Note that `inlf` is an indicator variable for whether a woman is in the labor force. For women not in the labor force, their wage is not observed.

```
. heckroccurve high_wage educ, select(inlf = educ kidslt6 kidsge6 nwifeinc)
Estimating inferred ROC curve...
 Empirical     Inferred         Inferred AUC
 ROC area      ROC area      95% Conf. Interval
--------------------------------------------------

  0.6472        0.6606         0.5782       0.7310
--------------------------------------------------
```

A more common use for ROC curves is constructing them after estimating a probit or logit. Here we provide an example of calling `heckroccurve` after a logit.

```
. quietly logit high_wage educ age exper if inlf
. predict predicted_xb, xb
. heckroccurve high_wage predicted_xb,
> select(inlf = predicted_xb educ kidslt6 kidsge6 nwifeinc)
Estimating inferred ROC curve...
 Empirical     Inferred         Inferred AUC
 ROC area      ROC area      95% Conf. Interval
--------------------------------------------------

  0.7211        0.7329         0.6377       0.8044
--------------------------------------------------
```

Note that we used the fitted values option `xb` rather than predicted probabilities, because Cook's (2017) assumption that the classifier's output is normally distributed is more likely to hold for the fitted values. The following syntax illustrates the command's plot options:

```
. heckroccurve high_wage predicted_xb,
> select(inlf = predicted_xb educ kidslt6 kidsge6 nwifeinc)
> noempirical cbands irocopts(lcolor(black) lwidth(medthick))
> rlopts(lcolor(gray))
Estimating inferred ROC curve...
 Empirical    Inferred        Inferred AUC
 ROC area     ROC area      95% Conf. Interval

  0.7211       0.7329        0.6377       0.8044
```

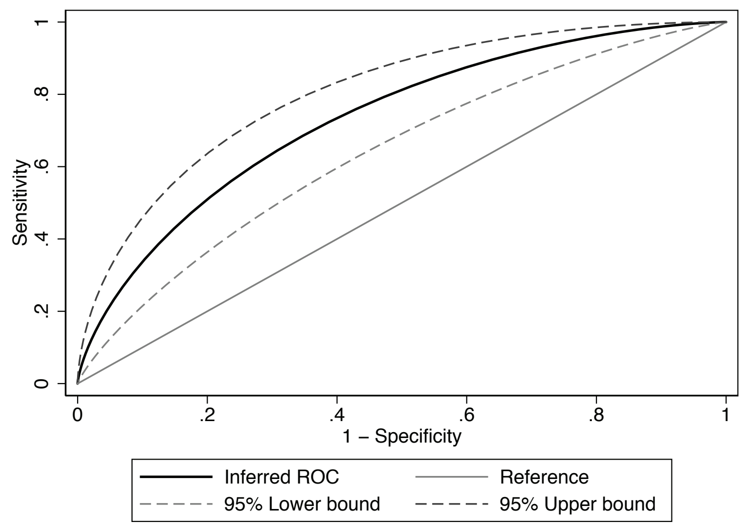


Figure 1. Plot created using `heckroccurve`'s plot options

◁

▷ Example 2: Correcting bias in ROC curves

This example uses a dataset that contains outcomes, two ratings, a selection indicator, and an independent variable x. We first compare the performance of the two ratings with the full dataset, and then we remove outcomes for the nonselected data:

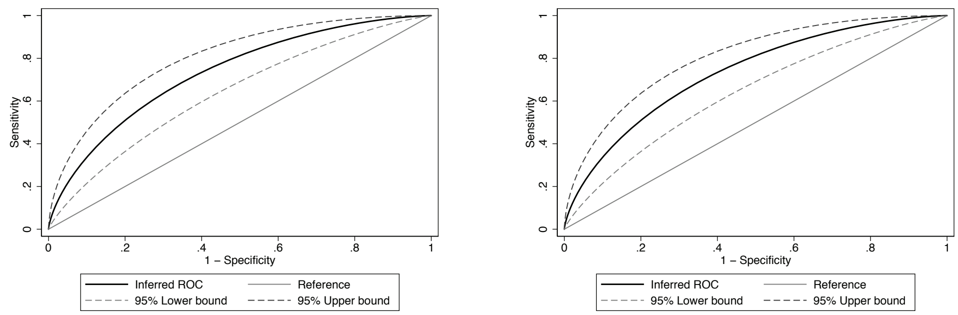
```
. sysuse heckroccurve_example, clear
. * Compare ratings using full dataset
. roccomp outcome rating_a rating_b
                              ROC                      -Asymptotic Normal--
                   Obs       Area     Std. Err.       [95% Conf. Interval]

rating_a         1,000     0.8815       0.0103         0.86120     0.90171
rating_b         1,000     0.7557       0.0151         0.72616     0.78515

Ho: area(rating_a) = area(rating_b)
    chi2(1) =    46.92       Prob>chi2 =   0.0000
```

While the outcome is observed for all 1,000 observations, this dataset contains a variable `selected` that is equal to 1 for half the observations and 0 for the rest. We set the outcome to missing when `selected` equals 0 to see the effect of selection on the classifiers' ROC curves:

```
. * Remove nonselected outcomes and compare ratings again
. replace outcome=. if !selected
(500 real changes made, 500 to missing)

. roccomp outcome rating_a rating_b

                              ROC                    -Asymptotic Normal--
                 Obs        Area     Std. Err.      [95% Conf. Interval]

rating_a         500      0.7493       0.0238        0.70273     0.79593
rating_b         500      0.7767       0.0256        0.72650     0.82685

Ho: area(rating_a) = area(rating_b)
    chi2(1) =     0.63       Prob>chi2 =   0.4262
```

Rating A performs better than rating B with the full dataset, but with the selected sample, performance is similar. Notice that the effect of selection differs for the two ratings. The AUC for rating A has decreased from 0.8815 to 0.7493, while the AUC for rating B is much less affected by selection. Calling `heckroccurve` allows us to recover the AUCs that are obtained with the full sample. Figure 2 provides the graphical output from the syntax below:

```
. heckroccurve outcome rating_a, select(x rating_a rating_b) cbands

Estimating inferred ROC curve...

 Empirical    Inferred         Inferred AUC
 ROC area     ROC area      95% Conf. Interval

  0.7493       0.8804         0.8253      0.9149

. heckroccurve outcome rating_b, select(x rating_a rating_b) cbands

Estimating inferred ROC curve...

 Empirical    Inferred         Inferred AUC
 ROC area     ROC area      95% Conf. Interval

  0.7767       0.7781         0.7283      0.8192
```



(a) Rating A (b) Rating B

Figure 2. Plots from `heckroccurve` show an example of the bias that can result from constructing ROC curves with a selected sample. Selection caused the ROC curve for rating A to cave in but left the ROC curve for rating B largely unaffected.

The inferred AUC for rating A is 0.8804. This is quite close to the value of 0.8815 that we obtained with the full dataset. The confidence interval for the inferred AUC does not contain the AUC that is obtained when only the selected sample is used (that is, 0.7493).

◁

5 Discussion and conclusion

Hand and Adams (2014) suggest an alternative approach for comparing ROC curves that are constructed with selected samples. Realizing that truncation of a rating leads to biased ROC curves, Hand and Adams explore the effects of reducing the data so that both ratings are truncated to a similar extent. The goal of truncating both ratings is to create ROC curves that are biased to a similar extent for both ratings and thus better facilitate a comparison of the ratings. While an advantage of this procedure is that it does not make any parametric assumptions, it does not remove the bias induced by selection. The procedure performed by `heckroccurve` provides a consistent estimate when the assumptions are met.

`heckroccurve` implements Cook's (2017) procedure for plotting ROC curves with selected samples and provides the AUC along with a confidence interval. The command's maximum likelihood estimation is performed by calling `heckprobit`. There are situations for which `heckprobit` will fail to converge. Changing the specification for the selection equation may allow for convergence. The inferred ROC curve is based on parametric assumptions (just as `heckman` and `heckprobit`'s estimations are based on parametric assumptions). Cook (2017) provides an example with wine data for which distributional assumptions are not met, yet the procedure recovers the AUC that is obtained with the full sample. While this one example is encouraging, the performance of the procedure when its distributional assumptions do not hold has not been thoroughly explored.

6 References

Cook, J. A. 2017. ROC curves and nonrandom data. *Pattern Recognition Letters* 85: 35–41.

Fawcett, T. 2006. An introduction to ROC analysis. *Pattern Recognition Letters* 27: 861–874.

Hand, D. J., and N. M. Adams. 2014. Selection bias in credit scorecard evaluation. *Journal of the Operational Research Society* 65: 408–415.

Heckman, J. J. 1976. The common structure of statistical models of truncation, sample selection and limited dependent variables and a simple estimator for such models. *Annals of Economic and Social Measurement* 5: 475–492.

Kraft, H., G. Kroisandt, and M. Müller. 2014. Redesigning ratings: Assessing the discriminatory power of credit scores under censoring. *Journal of Credit Risk* 10: 71–94.

Mroz, T. A. 1987. The sensitivity of an empirical model of married women's hours of work to economic and statistical assumptions. *Econometrica* 55: 765–799.

Van de Ven, W. P. M. M., and B. M. S. Van Praag. 1981. The demand for deductibles in private health insurance: A probit model with sample selection. *Journal of Econometrics* 17: 229–252.

About the authors

Jonathan A. Cook is a financial economist at PCAOB. The PCAOB, as a matter of policy, disclaims responsibility for any private publication or statement by any of its Economic Research Fellows and employees. The views expressed in this article are the views of the author and do not necessarily reflect the views of the Board, individual Board members, or staff of the PCAOB.

Ashish Rajbhandari is a quantitative investment analyst at The Vanguard Group. His research interests are applied econometrics and macroeconomics. The views expressed in this article are those of the author and do not necessarily reflect those of The Vanguard Group or its staff.