Verification Package: examples using weather forecasts.

Matthew Pocernich

October 1, 2012

The National Center for Atmospheric Sciences (NCAR) develops and implements weather and climate models. Verification statistics play a key role in this cycle. Functions in the **verification** library contains functions that have been developed in this process.

While the examples in this library focus on atmospheric topics, they are written to be applicable to any situations in which there is a prediction or forecast and an observation of the outcome. The statistics used to verify and study weather and climate forecasts are shared by many fields. Most notably, these include the fields of medicine (where the name misclassification statistics is favored) and signal detection theory. Some useful references are listed below.

The type of predictions and observations determine which methods are appropriate for verification. The following types of predictions are currently supported: binary, categorical, continuous, probabilistic and distributions. The presence or absence of fog is an example of binary data. A forecast for turbulence expressed in terms of low, moderate and extreme is a categorical forecast. The chance of precipitation is an example of a probabilistic forecast. A temperature forecast as a single value is (essentially) a continuous variable. People are finding it increasingly useful to express the uncertainly of a point forecast. In this case, the forecast may be expressed as a distribution.

1 Finley's Tornado

Any discussion of verification must begin in the beginning and for weather, that means John Finley and tornado forecasts. In 1884, John Finley using the telegraph, created yes/no tornado forecasts for 18 regions of the US [3]. Citing the results in Table 1, he explained that his method was 96.6 accurate.

Table 1: Finley Tornado Data

	Forecasts	
Observation	Yes	No
Yes	28	72
No	23	2680

```
[1] " Assume data entered as c(n11, n01, n10, n00) Obs*Forecast"

The forecasts are binary, the observations are binary.
The contingency table for the forecast
     [,1] [,2]
[1,]   28   72
[2,]   23 2680

PODy =  0.549
Std. Err. for POD =  0.06968
TS   =  0.2276
Std. Err. for TS =  0.03278
ETS  =  0.216
```

```
Std. Err. for ETS =  0.03476
FAR  =   0.72
Std. Err. for FAR =  0.0347
HSS  =  0.3553
Std. Err. for HSS =  0.04702
PC   =  0.9661
Std. Err. for PC =  0.003245
BIAS =  1.961
Odds Ratio =  45.31
Log Odds Ratio =  3.814
Std. Err. for log Odds Ratio =  0.3057
Odds Ratio Skill Score =  0.9568
Std. Err. for Odds Ratio Skill Score =
Extreme Dependency Score (EDS) =  0.7396
Std. Err. for EDS =  0.04793
Symmetric Extreme Dependency Score (SEDS) =  0.5935
Std. Err. for SEDS =  0.0439
Extremal Dependence Index (EDI) =  0.7174
Std. Err. for EDI =  0.06166
Symmetric Extremal Dependence Index (SEDI) =  0.7528
Std. Err. for SEDI =  0.06196
```

It was quickly pointed out that since tornados are so rare, if one always forecasted no tornado, the percent correct would be 98.2the time. The downside to this that the probability of detecting a tornado (PODy) drops to 0.

```
[1] " Assume data entered as c(n11, n01, n10, n00) Obs*Forecast"

The forecasts are binary, the observations are binary.
The contingency table for the forecast
     [,1] [,2]
[1,]    0    0
[2,]   51 2752

PODy =      0
Std. Err. for POD =  0
TS   =      0
Std. Err. for TS =  NaN
ETS  =      0
Std. Err. for ETS =  NaN
FAR  =    NaN
Std. Err. for FAR =  NaN
HSS  =      0
Std. Err. for HSS =  NaN
PC   =  0.9818
Std. Err. for PC =  0
BIAS =      0
Odds Ratio =    NaN
Log Odds Ratio =    NaN
Std. Err. for log Odds Ratio =  Inf
Odds Ratio Skill Score =    NaN
Std. Err. for Odds Ratio Skill Score =
Extreme Dependency Score (EDS) =     -1
Std. Err. for EDS =    NaN
Symmetric Extreme Dependency Score (SEDS) =    NaN
```

```
Std. Err. for SEDS =    NaN
Extremal Dependence Index (EDI) =    NaN
Std. Err. for EDI =    NaN
Symmetric Extremal Dependence Index (SEDI) =    NaN
Std. Err. for SEDI =    NaN
```

Note: **verify** is an overloaded function whose behavior is dictated by the types of forecasts and observations. By default, **verify** assumes that the forecast is probabilistic and the observation is binary. In the preceding example, since both the forecast and observation are binary, the forecast type needs to be described.

2 Verifying a precipitation forecast

While variables such as temperature, humidity and wind speed are traditionally forecast as a point forecast, precipitation has historically been forecast as a probability. The following example use precipitation forecast made by the Finnish Meteorological Institute [2]. This data is included as a sample data set in the verification package.

```
If baseline is not included, baseline values  will be calculated from the  sample obs.

The forecasts are probabilistic, the observations are binary.
Sample baseline calculated from observations.
Brier Score (BS)           =  0.1445
Brier Score - Baseline     =  0.1793
Skill Score                =  0.1942
Reliability                =  0.02536
Resolution                 =  0.06017
Uncertainty              =  0.1793
```

Typically, the probability of rain is expressed as one of a finite number of probabilities such as 10%, 20%, etc. Its not typical to see a forecast saying there is a 34.7% chance of rain. For automated forecasts or one's which aren't rounded a continuous range of values between 0 and 1 are possible. The *bins* option addresses this distinction. If bins = TRUE, forecast are placed into bins and assigned the center values. By default these bins are described by the *threshold* parameter and are (0, 0.1, ..., 0.9, 1). If FALSE, as in the case for precipitation forecasts, each forecast is considered individually. This becomes important when calculating statistics such as the Brier statistic.

```
> plot (A)
```

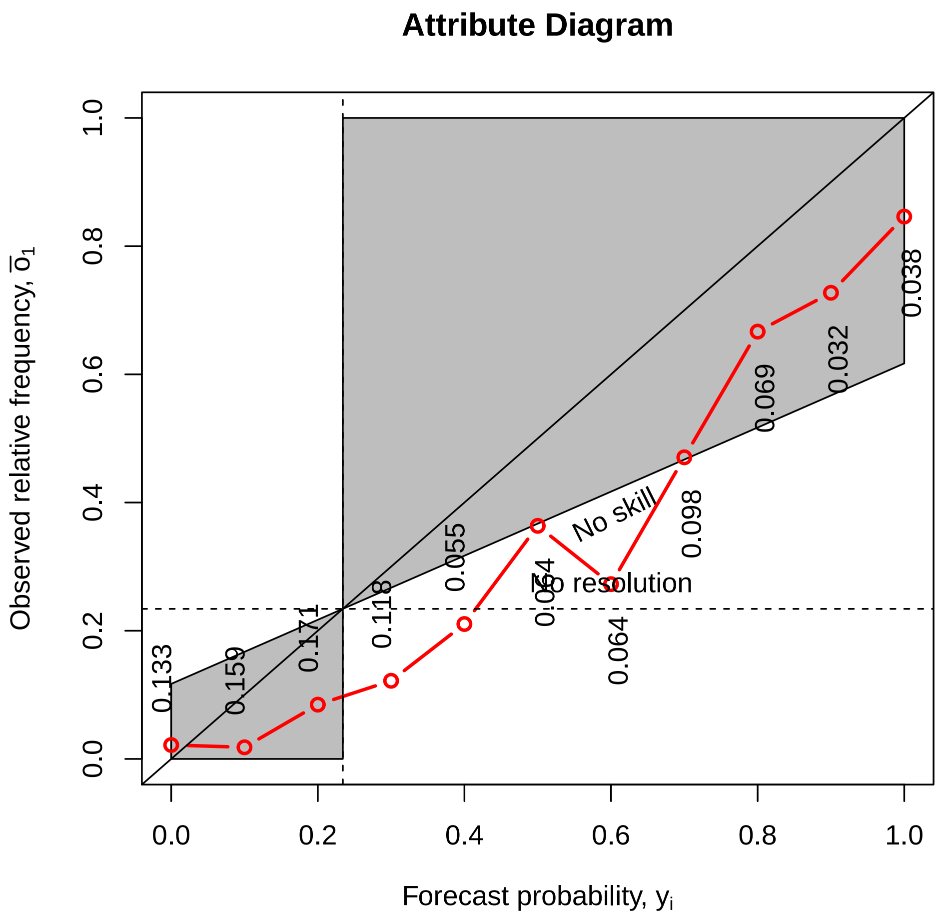

Figure 1: Attribute Diagram for light precipitation forecast

While the attribute diagram (Figure 1) is the default diagram for a probabilistic forecast, there are other very useful diagrams. The receiver operating characteristics (ROC) plot is also commonly used. A ROC plot displays the relation between false alarms and hits (successfully forecasted events) across a range of thresholds. Figure 2 shows a ROC plot for the probability of precipitation forecast. Since one wants a high ratio of hits to false alarms, the better the forecast, the further into the upper left hand corner the plot extends. This plot displays two lines. The black, un-smooth line is the empirical ROC plot. At each threshold, points are plotted. The smoother line is the result of fitted a binormal distribution to the points. For a perfect forecast, the area under the ROC curve would equal 1. In this example, the area under the curve is shown in the legend box. First the area under empirical curve is shown followed by the area under the bi-normal curve.

```
> mod24 <- verify(d$obs_norain, d$p24_norain, bins = FALSE)

If baseline is not included, baseline values  will be calculated from the  sample obs.

>  mod48 <- verify(d$obs_norain, d$p48_norain, bins = FALSE)

If baseline is not included, baseline values  will be calculated from the  sample obs.

> roc.plot(mod24, plot.thres = NULL)
> lines.roc(mod48, col = 2, lwd = 2)
> leg.txt <- c("24 hour forecast", "48 hour forecast")
> legend( 0.6, 0.4, leg.txt, col = c(1,2), lwd = 2)
```

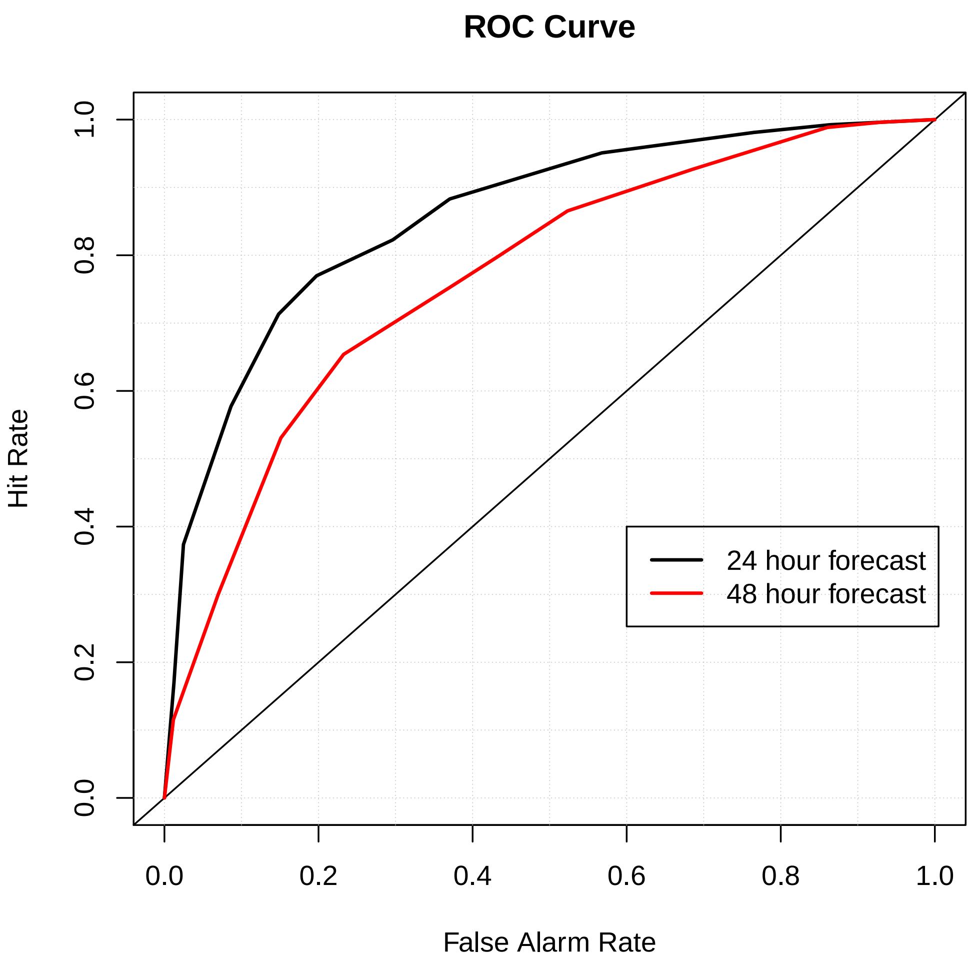

Figure 2: Receiver operating characteristic curve for chance of rain forecast at 24 and 48 hour lead times.

Unfortunately, estimating and expressing the uncertainty in these curves is seldom done. The verification packages offers a couple options for this. The data can be bootstrapped, to estimate the variance at the set thresholds (Figure 3).

```
> B<- roc.plot(A, CI = TRUE, n.boot = 100)
```

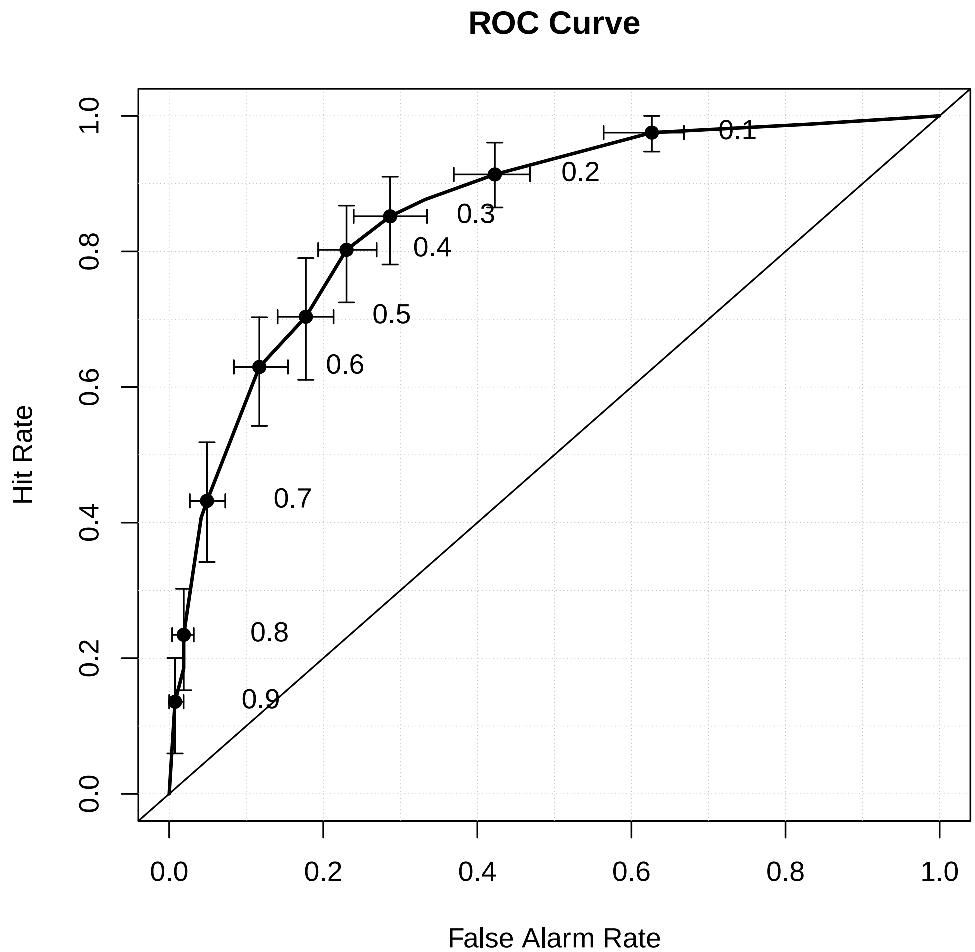

Figure 3: ROC curve with bootstrapped confidence intervals.

3 Value Diagrams

The utility of a forecast varies based upon the needs and concerns of an individual user. Value diagrams can be used to determine over what range of cost-lost (cl) ratios a forecast will provide value. The cost-lost ratio is the ratio of the cost of preparing for an event that doesn't occur over the losses that will occur if one is not prepared. Small values indicate that the costs to prepare are small in relation to the losses. The peaks of this graph occurs at the baseline average of an event. Figure 4 is an illustration of a value diagram for the Finnish precipitation data.

```
> value(d$obs_rain, d$p24_rain, main = "Rain-No Rain Forecast", cl = seq(0.01, 0.99, 0.05), all = TRUE)
```

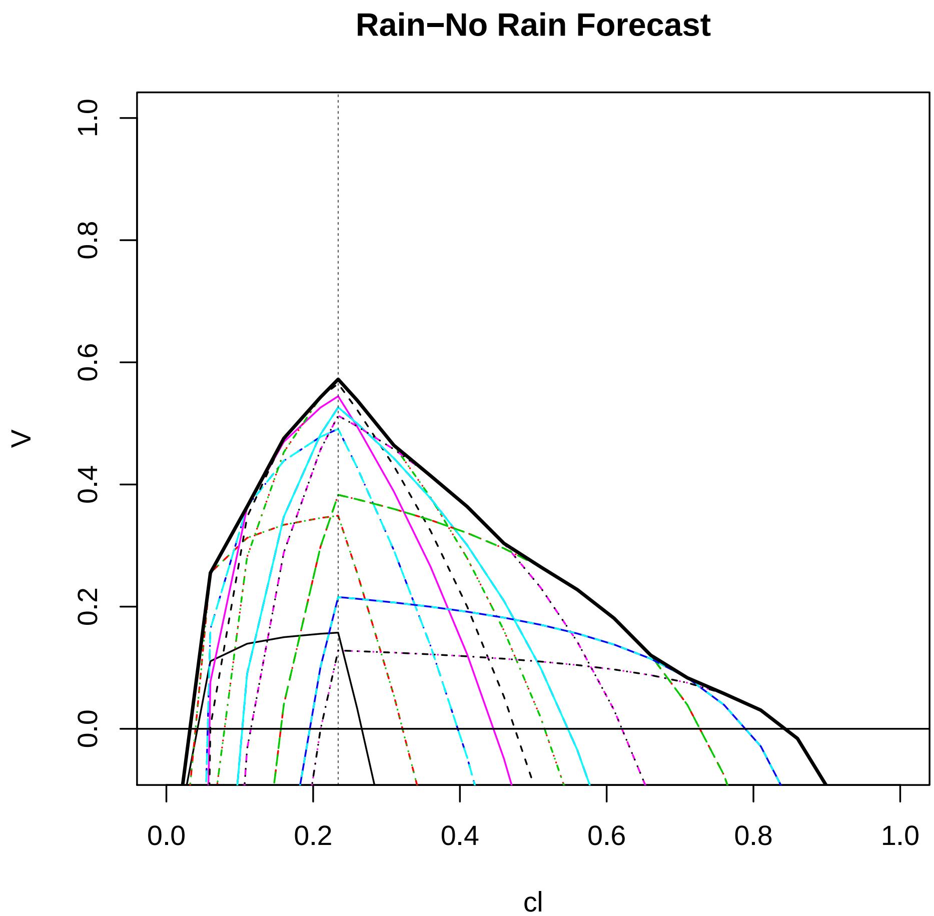

Figure 4: Value diagram of light precipitation forecast.

4 Discrimination Plot

A discrimination plot illustrates the the distributions of forecasts grouped by different types of distributions. Ideally, one would see a distinct histograms (Figure 5) .

```
> discrimination.plot(disc.dat$group.id, disc.dat$frcst, main =
+ "Sample Plot")
```

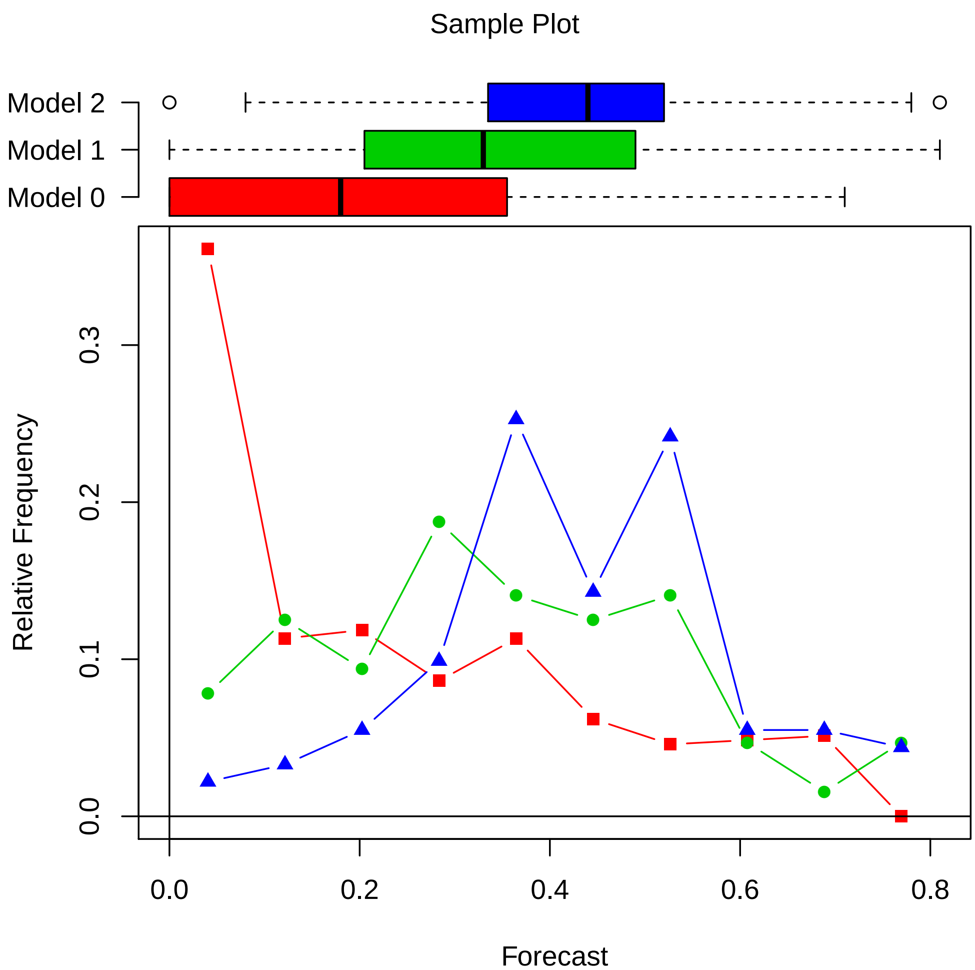

Figure 5: Discrimination plot using aviation forecast.

5 Reliability Diagram

Related to the attribute diagram is the reliability diagram. A reliability diagram can be used to compare multiple forecasts. Figure 6 is an example using data from [10].

```
> ## Data from Wilks, table 7.3 page 246.
>  y.i   <- c(0,0.05, seq(0.1, 1, 0.1))
>  obar.i <- c(0.006, 0.019, 0.059, 0.15, 0.277, 0.377, 0.511, 0.587, 0.723, 0.779, 0.934, 0.933)
>  prob.y<- c(0.4112, 0.0671, 0.1833, 0.0986, 0.0616, 0.0366, 0.0303,  0.0275, 0.245, 0.022, 0.017, 0.2
>  obar<- 0.162
> reliability.plot(y.i, obar.i, prob.y, titl = " Wilks Data", legend.names =
+ c("Model A") )
```

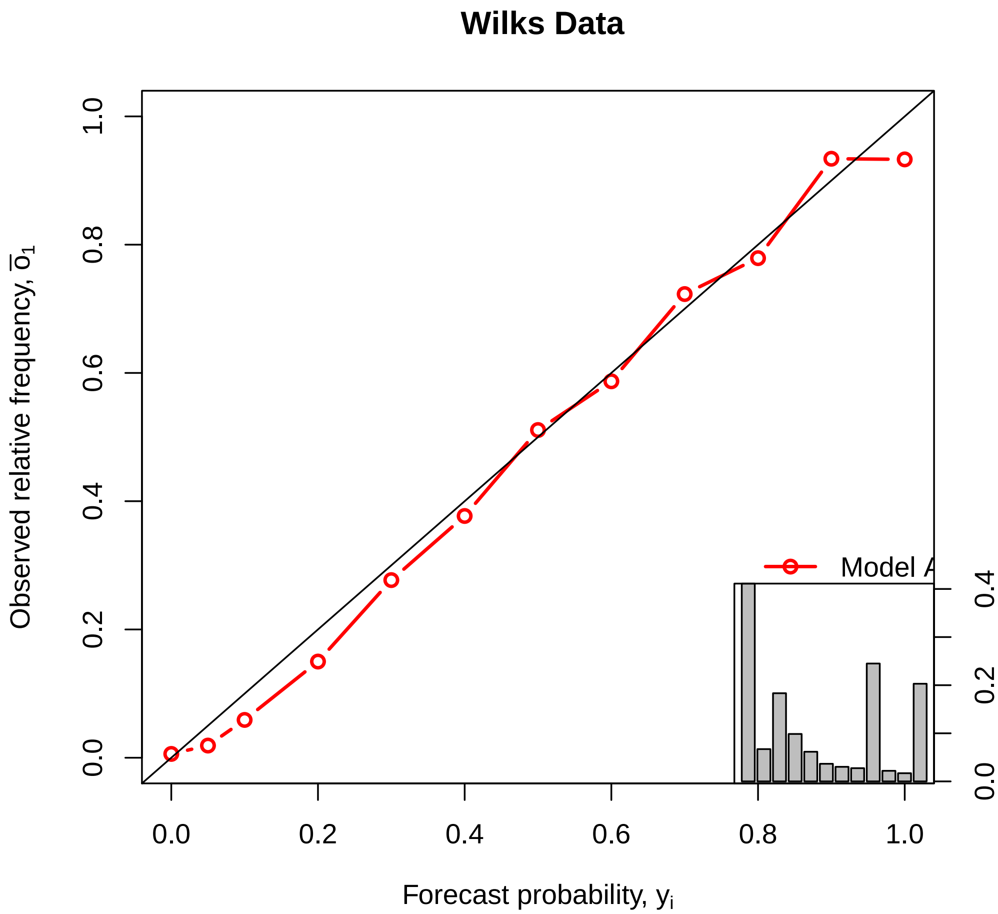

Figure 6: Reliability diagram of Wilks example.

References

[1] B.G. Brown A.H. Murphy and Y. Chen. Diagnostic verification of temperature forecasts. *Weather and Forecasting*, 1989.

[2] Beth Ebert, editor. *Forecast Verification - Issues, Methods and FAQ*. World Weather Research Programme Joint Working Group on Verification, 2006.

[3] J.P. Finley. Tornado prediction. *Amer Meteor J.*, 1884.

[4] Ian T. Jolliffe and David B. Stephenson, editors. *Forecast Verification: A Practioner's Guide in Atmospheric Science*. Wiley, West Sussex, England, first edition, 2003.

[5] I. Mason. A model for assessment of weather forecasts. *Aust. Met. Mag*, 30:291–303, 1982.

[6] S.J. Mason and N.E. Graham. Areas beneath the relative operating characteristics (roc) and relative operating levels (rol) curves: Statistical significance and interpretation. *Q. J. R. Meteorol. Soc.*, 128:2145–2166, 2002.

[7] R Development Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2005. ISBN 3-900051-07-0.

[8] D.S Richardson. Skill and relative economic value of the ecmwf ensemble prediction system. *Q. J. Roy. Met. Soc.*, 127:2473–2489, 2000.

[9] John A. Swets. *Signal Detection Theory and ROC Analysis in Psychology and Diagnostics*. Lawrence Erlbaum Associates, Inc., Mahwah, New Jersey, 1996.

[10] Daniel S. Wilks. *Statistical Methods in the Atmospheric Sciences*. Academic Press, San Diego, CA, first edition, 1995.