Biostatistics 140.654: Applied Regression Analysis
Fourth Term, 2019

Prediction Using Logistic Regression
Scott L. Zeger

Predictions from logistic regression or other models for binary responses (e.g. neural network; classification and regression tree (CART), random forests) can be used to classify subjects. For example, we may seek to classify a person as being at high or low risk of a large medical expenditure in a year. We may want to classify patients at a community clinic as having HIV infection or not. In fact, all medical diagnosis is an application of classification methodology, whether qualitative or quantitative.

The basic idea is to build a prediction model combining background scientific knowledge with evidence from "training "data that comprises response binary (or more generally categorical) Y and predictor variables X. We focus on only binary responses coded 1 or 0. One output of the model is a prediction $phat_i$ for each person, i=1,…,n in the training set. Note $phat_i$ is the expected value of Y_i so is the probability Y_i =1.

Suppose we classify a person as "positive" if her $phat_i > c$ and negative otherwise. That is, we create a dichotomous "prediction", $d_i(c)$, from the continuous probability $phat_i$: $d_i(c)$= 1 if $phat_i$>c; 0 otherwise. Note, the classification is a function of the person's X values, so we can write it phat(X_i) and d(X_i;c).

Having classified each person, we can ask how well the classification system works using two measures of accuracy: sensitivity and specificity. For a given threshold, c, *sensitivity(c)* = Pr(d(c)=1|Y=1) is the probability a person is classified positive (d=1) when they are (Y=1) and *specificity(c)* = Pr(d(c)=0|Y=0) is the probability of correct classification for negative outcomes. Because the sensitivity and specificity are obviously functions of the threshold, c, we will represent them by *sens(c)* and *spec(c)*. The goal is to find a set of predictor variables and model that have sensitivity and specificity values as close to 1.0 as nature will allow.

To illustrate, we use the National Medical Expenditure Survey (NMES) study data. We want to identify persons who are likely to spend more than $1,000 on medical services in a year, using their age, gender and whether they have a major smoking caused disease. We will distinguish lung cancer/COPD from coronary heart disease and stroke in the set of predictors. We can also use their poverty level, education and whether they regularly use a seat belt as a proxy for adversity to risk.

Below find results of a logistic regression fit to the NMES “training data” with the binary indicator of whether or not a person spent more than $1,000 on medical services.

```
                                                  Number of obs   =      11684
                                                  LR chi2(18)     =    1280.14
                                                  Prob > chi2     =     0.0000
Log likelihood = -7065.4867                       Pseudo R2       =     0.0831

------------------------------------------------------------------------------
      bigexp |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         lc5 |   1.574867   .1546359    10.18   0.000     1.271786    1.877948
        chd5 |   1.648393   .0751758    21.93   0.000     1.501051    1.795735
        MALE |  -.2726878    .043441    -6.28   0.000    -.3578305    -.187545
      agem65 |   .0315392   .0028935    10.90   0.000      .025868    .0372104
    age_sp65 |  -.0076651    .006095    -1.26   0.209    -.0196111     .004281
 _Imarital_2 |   -1.57205   .3563759    -4.41   0.000    -2.270534   -.8735662
 _Imarital_3 |   -1.58788   .3578531    -4.44   0.000    -2.289259   -.8865008
 _Imarital_4 |  -1.320355   .3615568    -3.65   0.000    -2.028993   -.6117168
 _Imarital_5 |  -1.200991   .3760931    -3.19   0.001     -1.93812   -.4638623
 _Imarital_6 |  -1.638974   .3671268    -4.46   0.000     -2.35853   -.9194192
     educate |  -.1213596   .0241988    -5.02   0.000    -.1687884   -.0739308
 _Ipoverty_2 |   .6753874   .4184442     1.61   0.107    -.1447482    1.495523
 _Ipoverty_3 |   .6400859   .4234295     1.51   0.131    -.1898207    1.469992
 _Ipoverty_4 |   .4857085   .4173876     1.16   0.245    -.3323561    1.303773
 _Ipoverty_5 |   .5490936   .4151799     1.32   0.186    -.2646441    1.362831
 _Ipoverty_6 |   .6628754   .4147359     1.60   0.110    -.1499921    1.475743
 _Ibeltuse_2 |   .0458161   .0620423     0.74   0.460    -.0757845    .1674168
 _Ibeltuse_3 |   .0731761   .0512809     1.43   0.154    -.0273327    .1736849
       _cons |    .782715    .553407     1.41   0.157    -.3019428    1.867373
------------------------------------------------------------------------------
```

To use the predicted values from this model, a continuous variable on [0,1], to classify each person, we can arbitrarily start with a threshold of 0.5. That is, we classify a person as likely to have a large expenditure if their predicted probability from the logistic model exceeds 0.5. The table below is a cross-tabulation of this prediction with the actual value.

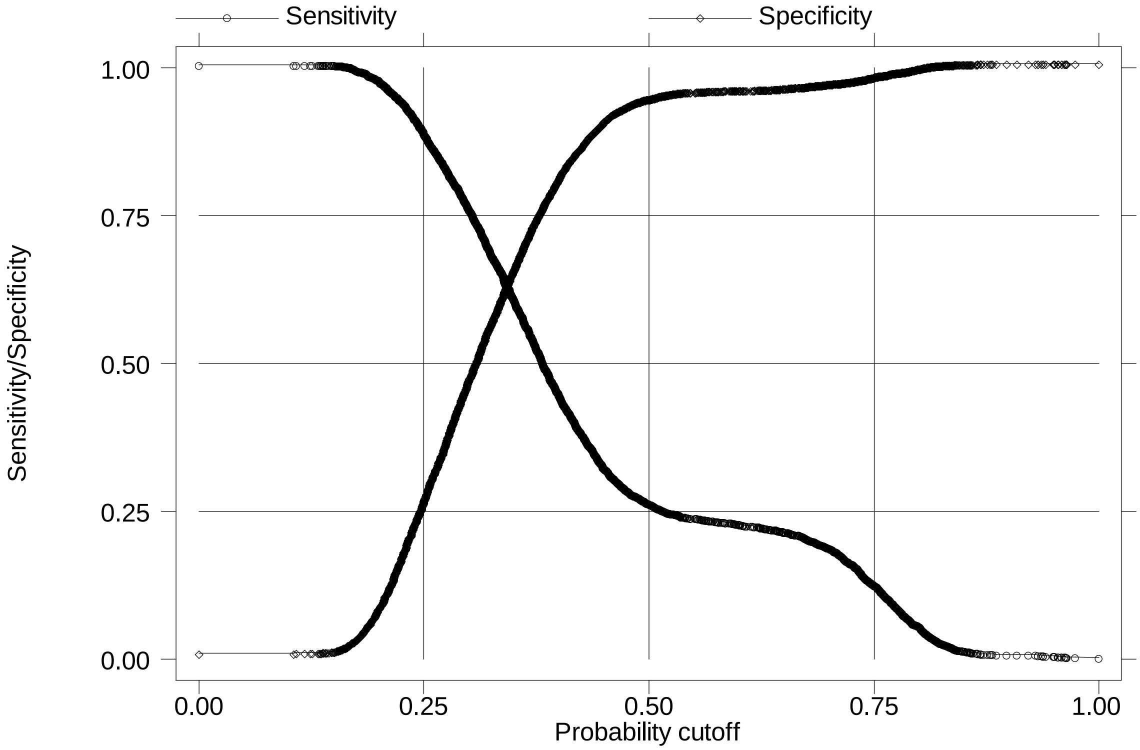
```
|                     bigexp (truth)
d1 (prediction
given X    |       0          1 |     Total
-----------+----------------------+----------
         0 |      6904       3208 |     10112
           |     93.94      74.00 |     86.55
-----------+----------------------+----------
         1 |       445       1127 |      1572
           |      6.06      26.00 |     13.45
-----------+----------------------+----------
     Total |      7349       4335 |     11684
           |    100.00     100.00 |    100.00
```

Among persons whose expenditure exceeded $1,000 (bigexp=1), the model has a predicted value greater than 0.5 for only 26% of them. That is, we estimate the sens(.5) = 0.26; similarly spec(.5) = .94.

We can improve the sensitivity by decreasing the threshold, c. Unfortunately, we will pay for this improvement with a decrease in specificity. But the trade-off may be worth it. Below we repeat the process for c=0.25.

```
|                   bigexp (truth)
d1 (prediction
given X    |       0          1 |     Total

-----------+---------------------+----------
         0 |     1897        497 |      2394
           |    25.81      11.46 |     20.49
-----------+---------------------+----------
         1 |     5452       3838 |      9290
           |    74.19      88.54 |     79.51
-----------+---------------------+----------
     Total |     7349       4335 |     11684
           |   100.00     100.00 |    100.0025)=
```

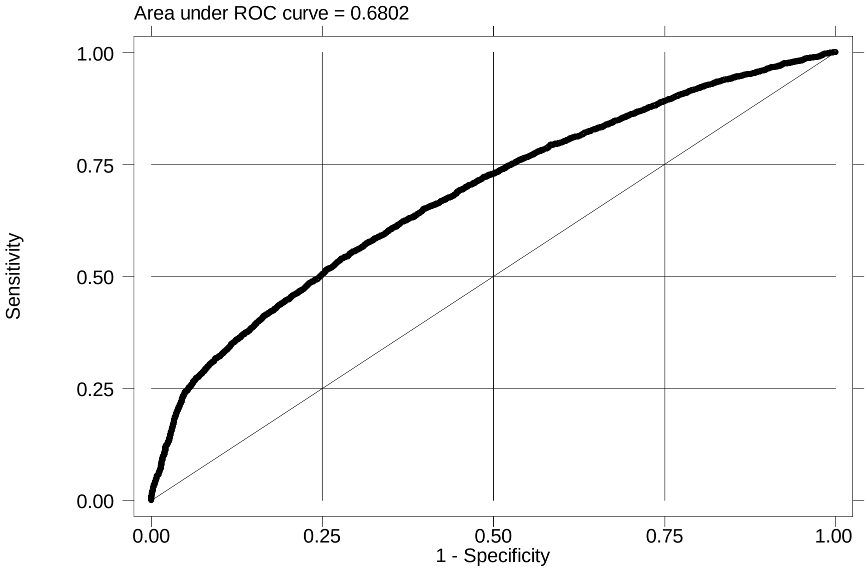
Now the estimated sensitivity and specificity are sens(.25) = .89 and spec(.25) = .26. Rather than trying a couple of threshold values, we can calculate the whole functions sens(c) and spec(c) for all c in (0,1) and plot them against c. The figure below shows the results.



The figure makes clear that to achieve a meaningful rise in sensitivity, there is a big price to pay in specificity. This is another way of saying that it is not easy to distinguish those who will have a large expenditure from those who will not using the set of predictor variables available. This is a rationale for, but also the conundrum of the medical insurance industry.

It is attractive to summarize a model's predictive ability with one statistic rather than two functions sens(c) and spec(c) of c. Toward this end, it is useful to plot the true positive rate, sens(c) against the false positive rate, 1-spec(c) for all values of c between 0 and 1. This curve is called the "receiver-operator characteristic" or ROC curve from of its origins in communication theory.

```
Logistic model for bigexp

number of observations =     11684
area under ROC curve   =    0.6802
```

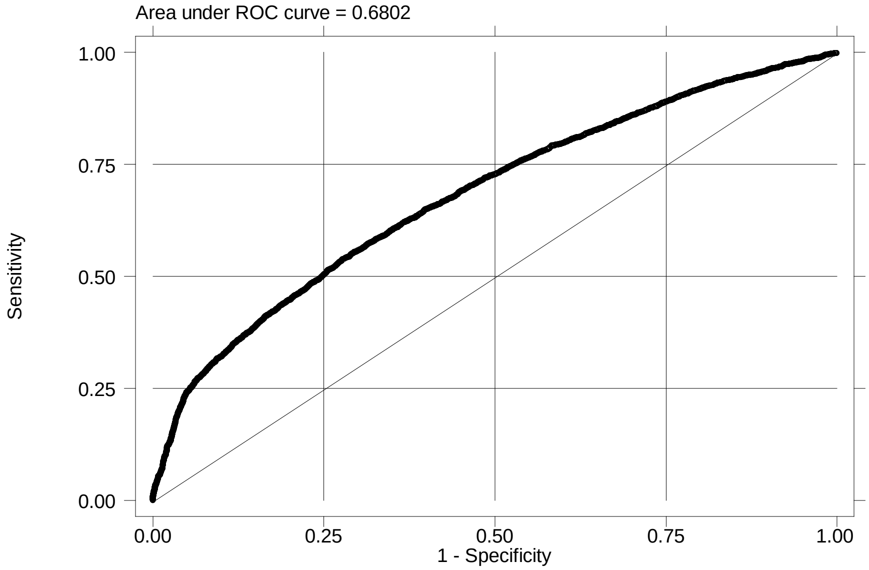


The ROC curve for a perfect predictor, that is, one that perfectly discriminates the persons with Y=1 from those with Y=0 for some value c, will start at (0,0) when c=1, rise up the y axis from (0,0) to (0,1), arriving at (0,1) for the value of c that perfectly separates the cases from controls. Then as c decreases further, the curve will travel along the sens=1 line, reaching (1,1) at c=0. If we classify by flipping a coin, the ROC curve will not depend on the data and hence the true and false positive rates will be equal for all c so that the ROC curve will be a 45-degree line from (0,0) to (1,1).

The area under the ROC curve is 1.0 for a perfect classifier and 0.5 for a coin toss or other irrelevant classifier; we use this area as a scalar measure of a model's predictive ability. Of course, the ability to classify varies by scientific problem, in particular, the degree of association of the Xs with the Y. In this way, the area under the ROC curve, called by some the C-statistic, is like R^2 for linear regression.

The area under the ROC has a second interpretation. In the case with a single X variable, the area under the ROC curve is the probability that a randomly chosen "case" (person with Y=1) has X value greater than a randomly chosen "control" (with Y=0). With multiple predictors, the area under the ROC curve is the probability that the predicted probability for the random case exceeds that for the random control.

```
Logistic model for bigexp

number of observations =    11684
area under ROC curve   =   0.6802
```



In this NMES example, the area under the ROC for the study data is 0.68. If we drop the *lc5* and *chd5* variables from the model, it decreases to 0.63. The reduction is not too great even though the diseases are very strong predictors. This is because only 1,319 out of 11,684 have a major smoking caused disease in this sample.

Cross-validation - the assessment of prediction error above can be overly optimistic, particularly in small samples. We typically select a model and estimate its regression coefficients to optimize the observed Y's likelihood. It is a mistake to then use the same Ys to ask how well we can predict them.

John Tukey, a famous statistician/scientist, used to say: "Beware: optimization capitalizes on chance". He meant that the estimated coefficients deviate from the true ones in ways that reflect chance events (particular Ys) for this one data set. This is because we have chosen those coefficient values that maximize the likelihood of the given data set. If we use that same set of Ys to evaluate the quality of prediction, the optimization will pay off. But if we predict a new set of Ys, we would not do as well since the random deviations in the optimized coefficients would hurt, not help us. Hence, we will predict the same Ys used to fit the model better than we would a new set of Ys with the same Xs.

Cross-validation is a method to obtain less biased estimates of prediction error for a new set of Ys at the same Xs. The idea is simple. In a large sample, we simply split the dataset into random halves. We fit the model with one half and then measure the quality of prediction with the other half, for example using the ROC curve. In this way, we are assessing the quality of prediction with a "new set of similar data".

In smaller samples, we cannot afford to set aside half the data. So, there is a clever alternative. We set aside a small fraction of the data set (say 10% or even a single observation), fit the model with the remainder and then predict the Ys for the fraction set aside. We compare the predictions with the actual values to measure the quality of prediction. Note, the data used to make the predictions are not used to assess their quality. We now repeat this process for all possible fractions of the data being left out. In this way, we can create a vector of predictions for each Y where that Y was not used to make its own prediction. A less biased estimate of the ROC curve can then be calculated.

To demonstrate, we have generated 20 predictor variables x1-x20. Each comprises 100 independent random Gaussian values. We have generated a set of 100 binary Ys independent of x1-x20. That is, the true logistic regression coefficients for x1-x20 are all 0.0. There are 44 values of 1 for Y. After a bit of preliminary analysis, we chose to focus on x5, a variable for which we had strong prior interest (hah) and to control for x11-x20 as possible confounders. The logistic regression results and ROC curve are shown below.

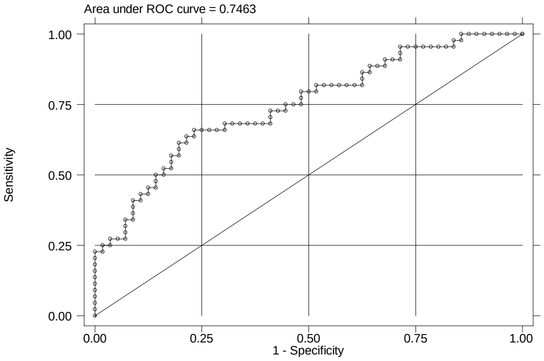
summary.glm(glm y ~ x5 + x11 + ….+ x20, family=binomial)

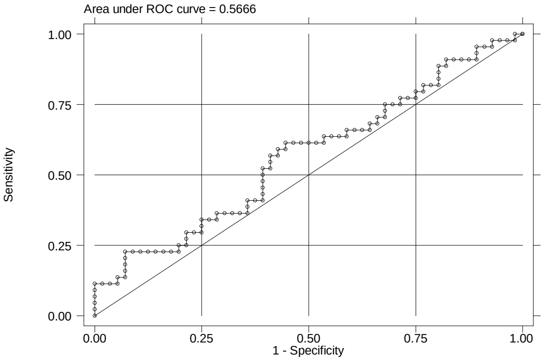
```
Log likelihood = -57.760634
------------------------------------------------------------------------------------------
---------
           y |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          x5 |   .8011291   .2653602     3.02   0.003     .2810326    1.321225
         x11 |  -.2749456   .2465971    -1.11   0.265     -.758267    .2083759
         x12 |   .3520122   .2389378     1.47   0.141    -.1162974    .8203217
         x13 |  -.1309065   .2483643    -0.53   0.598    -.6176916    .3558787
         x14 |   .2435796   .2594323     0.94   0.348    -.2648985    .7520576
         x15 |  -.2623732   .2534554    -1.04   0.301    -.7591367    .2343903
         x16 |   .0603459   .2197717     0.27   0.784    -.3703986    .4910905
         x17 |  -.1569996   .2586766    -0.61   0.544    -.6639964    .3499972
         x18 |   .0226012   .2409044     0.09   0.925    -.4495627    .4947652
         x19 |   .1932167    .264135     0.73   0.464    -.3244783    .7109117
         x20 |   .3832417   .2406217     1.59   0.111    -.0883682    .8548516
       _cons |  -.4418422   .2437237    -1.81   0.070    -.9195318    .0358473
------------------------------------------------------------------------------

number of observations =      100
area under ROC curve   =   0.7463
```

You can see that the x5 coefficient is highly statistically significant (p=.003), even after adjusting for x11-x20. The area under the ROC curve on the left is 0.74, bigger than the corresponding value for mammography for breast cancer detection.

However, if we use cross-validation to calculate the ROC curve, the prognosis is less rosy. The picture on the left below is the original ROC curve with area 0.74. The picture on the right is the cross-validated ROC whose area is now 0.56 with confidence interval 0.45 to 0.68. That is, our exciting prediction equation using 11 Xs cannot be distinguished from a predictor that is the flip of a fair coin. The bitter truth prevails.

The moral: in smaller samples (all samples to be certain), only use cross-validated measures of prediction error such as sensitivity, specificity or area under the ROC curve.





```
                 ROC                          -Asymptotic Normal--
    Obs         Area     Std. Err.         [95% Conf. Interval]
  ----------------------------------------------------------------
    100       0.5666        0.0586          0.45174      0.68138
```