Lecture 7: Model Diagnostics and Prediction Model Evaluation

Lin Chen

Department of Public Health Sciences
The University of Chicago

Logistic Regression Assumptions

Correlations among predictors.

- low r^2 is acceptable and unavoidable, since predictors are all related to Y, so it's possible for predictors to be somewhat related.
  ex) $r^2 = 0.5$ is ok.

The logistic regression method assumes that:

- Model form: There is a **linear relationship** between the logit of the probability (i.e., log of odds) of event and the linear predictors.
- Predictor:
  - There is no important predictor being omitted;
  - There is no high multicollinearity among the predictors.
- Observation: There is no outlier (influential observations).

Regression Diagnostics for Binary Outcome Models

- Similar as in linear regression analysis, post-model diagnostics for logistic regression can be used to
  - check model form / assumptions
    - check the form of predictors (original, log-transformed, etc)
    - check the adequacy of the link function (logit, probit, cloglog, etc)
  - check the predictors (multicollinearity)
  - identify unusual data observations - that do not fit the model well
  - other aspects including prediction performance
- For logistic models, we already learn to plot empirical and predicted logit of probability versus the linear combination of predictors (or individual predictors) to check linearity.
- Another useful quantity is the *residual* - difference between observed and predicted value.

Residuals – two types of residuals

First, **Pearson residuals**

- Pearson residuals are defined to be the standardized difference between the observed frequency/count and the predicted frequency/count. These measure the relative deviations between the observed and fitted values.

$$X_{Pi} = \frac{y_i - n_i \hat{p}_i}{\sqrt{n_i \hat{p}_i (1 - \hat{p}_i)}}, \tag{1}$$

- Note that the goodness of fit statistic, Pearson's χ^2 statistic, equals to the sum of squared Pearson residuals (Lecture 5, Slide 15) – – a small sum of squared residuals implies a good fit

Residuals – two types of residuals

Second, **deviance residuals**

- Deviance residuals are components of the deviance statistic, which measure the group i's contribution to the disagreement between the maxima of the observed (full model) and the fitted (current model) log likelihoods:

  Log Likelihood Function

  $$d_i = \text{sign}(y_i - \hat{y}_i)\sqrt{2y_i \log\left(\frac{y_i}{\hat{y}_i}\right) + 2(n_i - y_i)\log\left(\frac{n_i - y_i}{n_i - \hat{y}_i}\right)}, \qquad (2)$$

  where $\hat{y}_i = n_i \hat{p}_i$.

- Squaring these residuals and summing over all observations yields the deviance statistic (Lecture 5, Slide 5).
- Analogous to residuals in OLS regression, where the goal is to minimize the sum of squared residuals. Logistic regression estimation minimizes the sum of the deviance residuals (using maximum likelihood to solve). Large deviance implies a poor fit.

Standardized residuals

- The *standardised Pearson residual* is given by

$$\boxed{r_{Pi} = \frac{X_{Pi}}{\sqrt{1-h_i}},}$$

where h_i is the leverage score for generalized linear models.

- The *standardised deviance residual* is given by

$$\boxed{r_{di} = \frac{d_i}{\sqrt{1-h_i}},}$$

Residuals

- These quantities can be listed, plotted to identify unusual values
- In grouped data, those different types of (standardized or not) residuals can approximately normally distributed, zero-centered random variable (i.e., Z-statistic)
- Thus, values greater than +/- 2 may be of interest (‘extreme’ values on the Z scale)

Example 1: *Evaluation of an Anti-pneumococcus Serum*

This dataset summarizes effects of an anti-pneumococcus serum over increasing doses, exposing batches of infected mice

Table 1: Number of deaths from pneumonia amongst batches of 40 mice exposed to different doses of a serum

Dose of serum	Number of deaths out of 40
0.0028	35
0.0056	21
0.0112	9
0.0225	6
0.0450	1

Example 1: Anti-pneumococcus serum – Residuals

Example 1: Run the log dose model and examine the predicted counts ($\hat{y}_i$) and residuals. The option ‘`, d`’ produces deviance residuals. The option ‘`, p`’ produces pearson residuals. The option ‘`stan`’ after produces standardized residuals.

```
. gen ldose=log(dose)
. glm y ldose, family (binomial n) nolog
...
. predict pl
. predict dlres, d
. predict plres, p
. predict dlsr, d stan
. predict plsr, p stan

. list dose y pl dlres plres dlsr plsr
     +----------------------------------------------------------------------------+
     |  dose     y          pl       dlres       plres        dlsr         plsr |
     |----------------------------------------------------------------------------|
  1. | .0028    35    33.08491     .8344311    .8007663    1.283124    1.231357 |
  2. | .0056    21    22.95006    -.6209562   -.6234831   -.8082727   -.8115619 |
  3. | .0112     9    10.98707    -.7185567   -.7038911   -.8980333   -.8797047 |
  4. | .0225     6    3.823065     1.090101    1.170722    1.389797    1.492583 |
  5. |  .045     1    1.154902    -.1496326   -.1462671   -.1749602   -.1710249 |
     +----------------------------------------------------------------------------+
```

The predicted $\hat{y}$ is close to observed. The residuals are all small $|r|$<2.

Example 2: Grad Admission – Examining Residuals with Multiple Continuous Predictors

Example 2: A researcher is interested in how variables, such as GRE (Graduate Record Exam scores), GPA (grade point average) and prestige of the undergraduate institution (categorical, rated 1-4), effect admission into graduate school. The response variable, admit/not admit, is a binary variable.

```
. import delim using https://stats.idre.ucla.edu/stat/data/binary.csv, clear
(4 vars, 400 obs)

. list in 1/5

     +----------------------------+
     | admit   gre    gpa   rank |
     |----------------------------|
  1. |     0   380   3.61      3 |
  2. |     1   660   3.67      3 |
  3. |     1   800      4      1 |
  4. |     1   640   3.19      4 |
  5. |     0   520   2.93      4 |
     +----------------------------+
```

Example 2: Grad Admission – Examining Residuals with Multiple Continuous Predictors

```
. tab admit rank

           |                     rank
     admit |         1          2          3          4 |     Total
-----------+--------------------------------------------+----------
         0 |        28         97         93         55 |       273
         1 |        33         54         28         12 |       127
-----------+--------------------------------------------+----------
     Total |        61        151        121         67 |       400
```

The better ranked (1) undergrad institution, the higher chance of being admitted.

Examining Residuals - Continuous Predictors

First, we build a model with `rank` as an ordinal model.

```
. logistic  admit gre gpa rank, coef

Logistic regression                             Number of obs     =        400
                                                LR chi2(3)        =      40.53
                                                Prob > chi2       =     0.0000
Log likelihood = -229.72088                     Pseudo R2         =     0.0811

------------------------------------------------------------------------------
       admit |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         gre |    .002294   .0010918     2.10   0.036      .000154    .0044339
         gpa |   .7770137   .3274839     2.37   0.018     .1351571     1.41887
        rank |  -.5600314    .127137    -4.40   0.000    -.8092153   -.3108475
       _cons |  -3.449549   1.132846    -3.05   0.002    -5.669886   -1.229211
------------------------------------------------------------------------------

. estat gof, group(10)

Logistic model for admit, goodness-of-fit test

  (Table collapsed on quantiles of estimated probabilities)

       number of observations =       400
             number of groups =        10
      Hosmer-Lemeshow chi2(8) =ungrouped  3.22
                  Prob > chi2 =         0.9199 ☺
```

Examining Residuals - Continuous Predictors

```
. quiet glm  admit gre gpa rank, family(binomial)
. predict p
(option mu assumed; predicted mean admit)
. predict dres, d
. predict pres, p
. predict dsr, d stan
. predict psr, p stan

. list admit gre gpa rank p dres pres dsr psr if dres > 2.0 | dres < -2 & dres ~=., noobs clean

     admit   gre    gpa   rank          p       dres       pres        dsr        psr
         1   580   2.86      4     .10556   2.120602   2.910891   2.130251   2.924137
         1   400   3.23      4   .0942907   2.173188   3.099275   2.182424   3.112446
```

Both cases had low predicted probability of admission yet had being admitted - may be other predictors/factors being omitted.

A lack of fit for a logistic model could be due to the omission of important predictors or the inappropriate choice of link function (more later).

More Modeling

We model `i.rank` as a nominal categorical variable.

```
. logistic admit gre gpa i.rank
Logistic regression                             Number of obs     =        400
                                                LR chi2(5)        =      41.46
                                                Prob > chi2       =     0.0000
Log likelihood = -229.25875                     Pseudo R2         =     0.0829
------------------------------------------------------------------------------
       admit | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         gre |   1.002267   .0010965     2.07   0.038      1.00012    1.004418
         gpa |   2.234545   .7414652     2.42   0.015     1.166122    4.281877
        rank |
          2  |   .5089309   .1610714    -2.13   0.033     .2736922    .9463578
          3  |   .2617923   .0903986    -3.88   0.000     .1330551    .5150889
          4  |   .2119375   .0885542    -3.71   0.000     .0934435    .4806919
       _cons |   .0185001   .0210892    -3.50   0.000     .0019808    .1727834
------------------------------------------------------------------------------
Note: _cons estimates baseline odds.

. estat gof, group(10)
Logistic model for admit, goodness-of-fit test
  (Table collapsed on quantiles of estimated probabilities)
       number of observations =       400
             number of groups =        10
      Hosmer-Lemeshow chi2(8) =        11.09
                  Prob > chi2 =         0.1969
```

There is a linear decreasing trend on the log odds of admission when rank increases.

HL statistic is smaller when modeling `rank` as an ordinal variable. The ordinal `rank` model has a good fit.

Multicollinearity – strong correlations among predictors

- If collinearity is strong and we ignore it, the variance estimates will be larger than it should be (inflated), the confidence intervals become larger and the estimated β's can be very unstable, i.e., both point estimation and inferences are biased.
- In cases where prediction is of principal interest, we may be less concerned about collinearity.
- Variance Inflation Factor (VIF) is a measure of multicollinearity

  Predictors only

  Not related to Y.
- VIF = 1/tolerance.

  $$\boxed{\mathrm{VIF}_j = \frac{1}{1 - R_j^2},} \qquad j = 1, \ldots, p,$$

  where R_j^2 is the variation of the j-variable that can be explained by other predictors.
- A rule of thumb: A VIF $>$ 10 suggests strong multicollinearity – a severe violation of model assumption. Some predictor(s) is redundant and could be dropped.

Install external packages to check multicollinearity

One may install external Stata packages to run functions that were not offered by standard Stata software. For example, the `collin` function. VIF ranges from 1 to infinity. In this example, VIFs are low, implying no multicollinearity.

```
. net describe collin, from(https://stats.idre.ucla.edu/stat/stata/ado/analysis)
. net install collin (package)
. collin gre gpa rank
(obs=400)
  Collinearity Diagnostics
                       SQRT                    R-
  Variable      VIF    VIF    Tolerance    Squared
----------------------------------------------------
       gre      1.19   1.09    0.8420      0.1580
       gpa      1.17   1.08    0.8522      0.1478
      rank      1.02   1.01    0.9846      0.0154
----------------------------------------------------
  Mean VIF      1.13
                           Cond
         Eigenval          Index
---------------------------------
    1     3.8654          1.0000
    2     0.1092          5.9502
    3     0.0195         14.0760
    4     0.0059         25.5637
---------------------------------
 Condition Number         25.5637
 Eigenvalues & Cond Index computed from scaled raw sscp (w/ intercept)
 Det(correlation matrix)    0.8393
```

Additional Model Assessment - Variability Explained

In linear regression, an important concept relates to the proportion of variability explained. For the fitted regression model $\widehat{Y}_i = \widehat{\beta}_0 + \widehat{\beta}_1 X_i$, we have

- The sum of squared errors for the regression model: $SSR = \sum(\hat{Y}_i - \bar{Y})^2$ - variation in a predicted Y around the mean of Y
- The sum of squared errors for residuals: $SSE = \sum(\hat{Y}_i - Y_i)^2$ - variation between predicted and observed Y_i's
- The sum of squared errors total $SST = \sum(Y_i - \bar{Y})^2$ - variation of individual Y_is around the mean of Y

Then

$$SST = SSR + SSE$$

Additional Model Assessment - Coefficient of Determination

The coefficient of determination or more commonly the R^2 measures the fraction of SSR or 'model' variation relative to total variation (SST or simple variation of Y around its mean, ignoring X).

$$R^2 = \frac{SSR}{SST}$$

sometimes more usefully written as

$$R^2 = 1 - \frac{SSE}{SST}$$

Heuristically, if X is a good predictor, $\hat{Y}_i$'s *should* be close to Y_i and SSR would be relatively large and R^2 would be close to 1. If X is not a good predictor, $\hat{Y}_i$'s evaluated at different X_i's would not differ much from the overall mean ($\bar{Y}$), and R^2 would be close to zero. Note that $0 \leq R^2 \leq 1$. ✓

Additional Model Assessment - Variability Explained

In logistic regression, several analogues were proposed. One simple one is

$$\text{pseudo}R^2 = 1 - \frac{\log L(\hat{\beta})}{\log \hat{L}_0}$$

where $\log L(\hat{\beta})$ is the log likelihood for the current model and $\log \hat{L}_0$ is the null model with only an intercept.

- The idea is similar as before. A likelihood falls between 0 and 1, so the log of a likelihood is less than or equal to zero. The ratio of the likelihoods of the current model to the intercept-only model suggests the level of improvement in fit.
- Pseudo-R^2's do not range from 0 to 1. The lowest value is 0 but often do not reach 1. Values from 0.2-0.4 indicate excellent model fit.
- Adjusted pseudo-R^2 adjusting for the number of parameters is available
- Generally, these measures are not as reliable as fits measures as in the linear regression and the word R^2 can be misleading here too

Additional Model Output: Classification and ROC

- Several other post-model outputs relating to prediction are available. These approach the model from a classification of outcomes perspective:
    - Cross-classifying to compare agreement between predicted and observed outcomes under some assignment rule such as - ‘any case with $> .50$ predicted probability of being an event (1) will be classified as a success’ - Use familiar measures of sensitivity, specificity, etc to summarize rule
    - The receiver operating characteristic (ROC) curve - plot of sensitivity vs (1 - specificity). It evaluates prediction rule over all possible cut-points of probability. This can be shown to equal to probability of correctly determining among two random cases, which will be an event.
- Stata facilities these analyses. Results can look encouraging, but true test of model performance must be assessed on *independent* data - not the data use to build the model

Additional Model Output: Classification

```
. estat classification

Logistic model for admit

              -------- True --------
Classified |         D            ~D  |      Total
-----------+--------------------------+-----------
     +     |        29 TP         20 FP|         49
     -     |        98 FN        253 TN|        351
-----------+--------------------------+-----------
   Total   |       127           273  |        400

Classified + if predicted Pr(D) >= .5
True D defined as admit != 0
--------------------------------------------------
Sensitivity                     Pr( +| D)   22.83%
Specificity                     Pr( -|~D)   92.67%
Positive predictive value       Pr( D| +)   59.18%
Negative predictive value       Pr(~D| -)   72.08%
--------------------------------------------------
False + rate for true ~D        Pr( +|~D)    7.33%
False - rate for true D         Pr( -| D)   77.17%
False + rate for classified +   Pr(~D| +)   40.82%
False - rate for classified -   Pr( D| -)   27.92%
--------------------------------------------------
Correctly classified                        70.50%
--------------------------------------------------
```

Additional Model Output: Classification

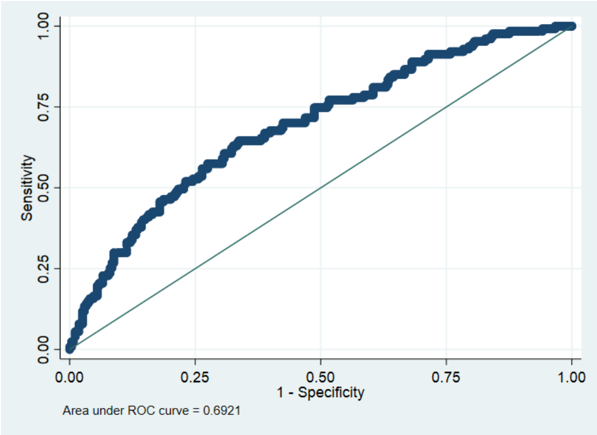
Changing the cutpoint will change the performance:

```
. estat classification, cutoff(0.35)

Logistic model for admit

              -------- True --------
Classified |         D            ~D  |      Total
-----------+--------------------------+-----------
     +     |        74            83  |        157
     -     |        53           190  |        243
-----------+--------------------------+-----------
   Total   |       127           273  |        400

Classified + if predicted Pr(D) >= .35
True D defined as admit != 0
--------------------------------------------------
Sensitivity                     Pr( +| D)   58.27%
Specificity                     Pr( -|~D)   69.60%
Positive predictive value       Pr( D| +)   47.13%
Negative predictive value       Pr(~D| -)   78.19%
--------------------------------------------------
False + rate for true ~D        Pr( +|~D)   30.40%
False - rate for true D         Pr( -| D)   41.73%
False + rate for classified +   Pr(~D| +)   52.87%
False - rate for classified -   Pr( D| -)   21.81%
--------------------------------------------------
Correctly classified                        66.00%
--------------------------------------------------
```

Additional Model Output: ROC Curve and AUC

- A Receiver Operating Characteristic (ROC) curve plots sensitivity (true positive rate) against 1-specificity (false positive rate) at various predicted probability thresholds. It can be used for decision making to select optimal threshold for prediction.
- AUC stands for **A**rea **U**nder the ROC **C**urve. Generally, a high AUC implies a better prediction.
- Below is the ROC curve Based on 3-predictor model for admission.

```
. lroc
Logistic model for admit
number of observations =       400
area under ROC curve   =    0.6921
```



Cross-validation for prediction evaluation

- You build a model to capture interesting patterns in the data, but those patterns do not generalize to other data —- this is called overfitting the data.
- Think about this: you use all your data to build a prediction model and the model is chosen to fit the data best. Now you are using the best fitted data to predict the data — possible overfitting.
- Cross-validation is a resampling method that uses some proportion of the data to build/train the model and uses the rest data to test the chosen model. Repeat the procedure to get the averaged prediction evaluation.
- For example, a 10-fold cross validation randomly splits the data into 10 parts. Each time, treat one part as the testing data and the rest 9 parts as the training data. Build the model on the training data and evaluate of prediction performance on the testing data. Repeat it for all 10 parts and obtain the averaged prediction performance.

Summary

Model Assessment

- In this lecture, we discussed 1) residuals, 2) VIF, 3) pseudo R^2, and 4) ROC as additional tools for model diagnostic and evaluation.
- The diagnostic tools follow those of linear regression models, but may be a bit more difficult to interpret. These are useful in practice.
- There are many more diagnostics covered in Collett ch. 6, including identifying influential observations, outliers, assumptions about the distribution (ie link function), etc.
- Post-model classification tools assess the utility of the model in prediction. Developing reliable prediction models further requires independent validation.