Lab 5 Logistic Regression
Fitting, Assumptions, and Diagnostics
Oct. 7, 2018

Lab Aims

Although you will quickly find the specification and fitting of logistic regression to be syntactically very similar to linear regression, it is important to remember that logistic regression represents a fundamentally different model paradigm. From the assumptions of the model, to the interpretation and inference of coefficients, the shift from regression (prediction of a continuous target variable) to classification (prediction of a discrete outcome) necessitates a number of special considerations during analysis.

Thus, building on our prior knowledge of the Statsmodels package, the primary focus on this lab will be on the interpretation, and diagnostics associated with the logistic regression model. In particular, binary logistic regression, where outcomes can assume only two possible values.

To begin: download the lab notebook (Sakai, Webpage) and upload it to your notebook directory. It contains code to load in and create the data needed to answer the various questions below. In this lab, the data is comprised of lab and vital signs recorded for a patient over their first 24 hours post-admission, as well as some demographic information merged in from other MIMIC tables. We will also look to the IAC dataset.

---

Logistic Regression Models - Specification and Fitting

Thus far we have utilized the OLS (ordinary least squares) function in Statsmodels to fit and evaluate linear regression. However, there are of course alternative regression models implemented in the package. It should then come as little surprise that logistic regression has also been implemented in the Statsmodels package. It is specified using the Logit function.

• Q1: Using the same syntax and code from OLS models specify and fit a logistic regression model with the following parameters:

  Y: HOSPITAL_EXPIRE_FLAG
  X:
    * sapsii
    * LOS

---

\(^1\) However we will now be shifting our focus to those discrete features we have previously ignored (death in the ICU, and use of an indwelling arterial catheter)

\(^2\) For linear regression these include weighted least squares, and elastic net, each of which provides some utility in addressing data challenges from multicollinearity to heteroscedasticity

\(^3\) I recommend R-style formulas, but you may use any method you are comfortable with.

\(^4\) All column names are listed exactly as they appear in the data.
* GENDER
* ADMISSION_TYPE

Model Inference

As logistic regression represents another additive linear model we begin by focusing on the three major questions about linear models discussed in class.

- Inference around the model:
  - Do any of the features provide value in explaining the target? (i.e. is the model useful)
- Inference around the coefficients:
  - Does a coefficient provide significant value in predicting the target variable?
  - As these fitted coefficients are known to be estimates, what range can we expect their values to take? (at 95% confidence)

Using the fitted model from Q1, use the model.summary() command to display the summary.

- Q2: Reading The Model Summary
  - Is the global likelihood test significant?
  - What is the pseudo $R^2$?
  - Which features are statistically significant at 95%?
    - Are any features not significant? Does this make sense? Would want to remove the feature from the model?
- Q3: Interpretation of Model parameters
  - Interpret the intercept coefficient
  - Interpret the coefficient of the GENDER feature
  - Interpret the confidence interval of the ADMISSION_TYPE feature

Although the model coefficients can be interpreted directly as log-odds, it is often far more interpretable to convert these values to odds ratios. To do so, we need only to exponentiate the coefficient (assuming the natural log is used, which is true for the Statsmodels package). However, manually doing this transformation for large models is tedious and error prone. Rather we typically convert then as a whole.

1 Note GENDER and ADMISSION_TYPE are both nominal and must be specified with the C() in the model formula.

6 The summary function is one of the only times print should be used over display() in the notebooks, as it is overridden to provide cleaner formatting.

7 Use the np.exp() function as it can operate across an entire list at once.
• Q4: Utilize the fitted model params attribute\(^8\). To compute the odds ratio for each model coefficient.

• Q5: It is often useful (especially if you convert the coefficient’s themselves), to convert the CI into odds ratios. Thankfully we do not need to do anything fancy here. We access the confidence interval for each feature outside of the summary\(^9\), and the same function used for coefficients can be applied here again.

• Q6: Interpret the transformed coefficient and confidence interval of the GENDER feature

---

**Assumptions and Plots**

Although logistic regression has fewer assumptions than linear regression, it is still important to check that each is met in order reassure ourselves the results are meaningful and interpretable. The two primary assumptions fall to the linear relation between X and \text{logit} of Y, independence of data instances\(^10\).

For this section we will shift to the IAC dataset, as the hospital mortality presents a number of data challenges outside the scope of this lab. As mortality is a product of all care the lab values in the first 24 hours are not always a great indication alone. Further, the problem is highly imbalanced (many more people survive than die in the hospital. Although this is a good thing, it presents a challenge to building models we will address during the validation week.

• Q7: Assess the relation between X and \text{logit} Y. This process is not exactly straightforward as (due to another assumption) Y must be binomial. Use the code provided to go through the process of fitting a smoothed curve to Y, taking the logit of the smoothed values, and finally plotting against two of the X values.
  - Do these satisfy the assumption?

As logistic models do not compute a Sum of Squares, it may seem counterintuitive to discuss the idea of a residual in terms of error. However, through the use of deviance residuals we can quickly evaluate the model for outliers in the fit or patterns such as temporal trends.

• Q8: Using the provided formula string, fit the logistic model. Next, plot the deviance residuals for the fitted model against the index in the dataframe\(^11\).

\(^8\) Remember back before we found the pretty Summary function, you can access the coefficients with fitted_model.params

\(^9\) fitted_model.conf_int()

\(^10\) For independence we often must know how the data was collected. However some residual plots can help us assess if there are other violations of independence such as auto-correlation

\(^11\) As we are not looking for relation X or Fitted values, we can simply plot these against the case number or index
**Interactions**

While dummy variables offer tremendous value in their ability to provide unique intercept shifts for different levels of various nominal data, their primary downside remains that the slope of each line remains the same. However various research questions may lead to the belief that slopes should be is non-constant between categories. I.e. coefficients an independent variable should vary between categories. To achieve this, we can specify interaction terms within a model.

- **Q9**: Specify and fit the following model using the `iac_data`:
  - Y: hosp_exp_flg
  - X
    - gender_num
d  - Interaction between age and the nominal value of resp_flg

- **Q10**: Print the model summary.
  - Is the interaction significant at 95%, what does this mean?
- **Q11**: Use the provided function to plot the interaction. What does the plot tell you?

---

12 All column names are listed exactly as they appear in the data.
13 nominal
14 Use the * operator to specify the interaction. Note adding the interaction like this also adds the features in univariate manner as well, so they do not need to each be explicitly added to the model.