5.1: Logistic Regression

Dr. Bean - Stat 5100

1 Why Logistic Regression?

Recall the linear regression model

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_{p-1} X_{p-1} + \epsilon \qquad (\epsilon \sim N(0, \sigma^2)).$$

(Individual) What are some properties of the variable Y that are required for ϵ to be normally distributed.

- Y must be linearly related to $X_1, \ldots X_{p-1}$.
- Y must be a **continuous, quantitative** variable

1.1 Why not regression on categorical data?

Consider fitting a regression model where we use age to try and predict whether or not a person has a disease (a 0-1 variable).

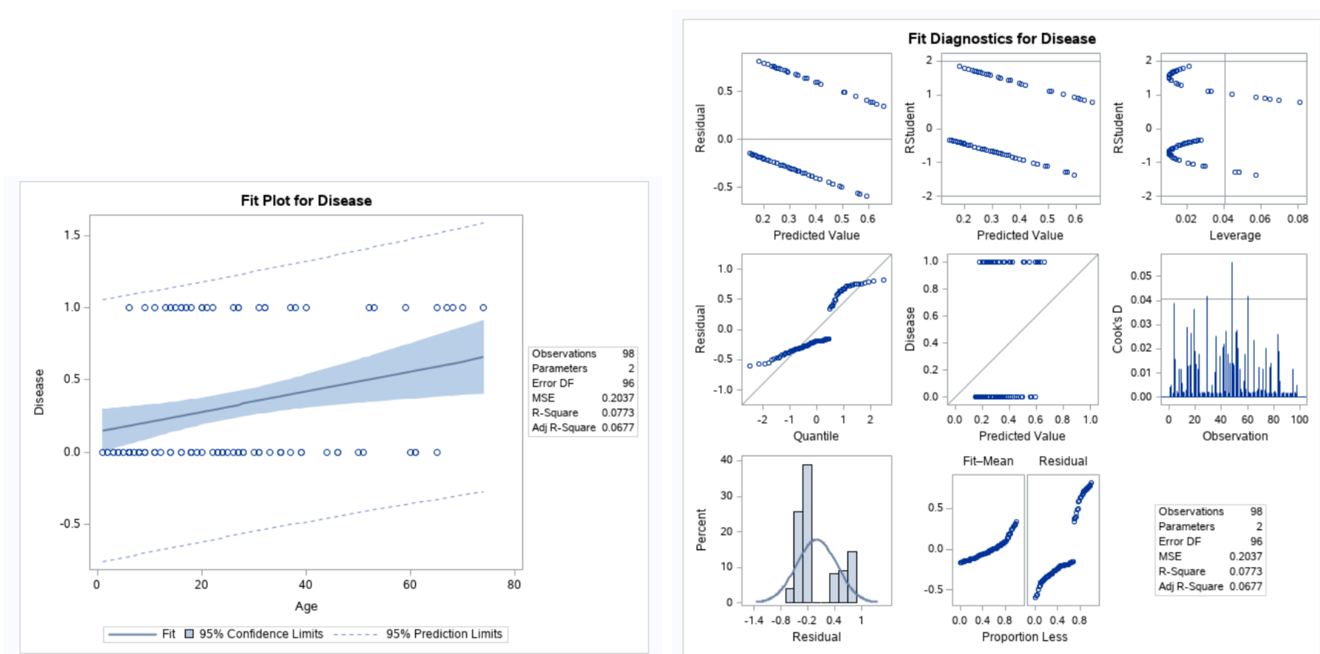

Figure 1: Fit plot and residual diagnostics for regression model that uses age to predict the presence/absence of a disease.

It is for this reason that instead of trying to predict the **value** of a categorical predictor, we should rather try to predict the **probability** of occurrence π_i,

$$\pi_i = \beta_0 + \beta_1 X_{i,1} + \cdots + \beta_{p-1} X_{i,p-1} + \epsilon_i \qquad (\epsilon \sim N(0, \sigma^2)). \tag{1}$$

(Individual) However, based on the previous example, what are some of the issues with trying to predict the probability using (1)?

- We don't actually know π_i.
- Model can predict negative probabilities or probabilities above 1.
- Residual assumptions never satisfied (impossible for residuals to be normally distributed).

2 Transforming Probabilities

Because regression works best with **unconstrained** variables (i.e. variables that can theoretically take on any value). We need to find a transformation that maps $\pi \in [0, 1]$ to $f(\pi) \in (-\infty, \infty)$.

Solution: log-odds ratio.

- $\pi \rightarrow [0, 1]$
- $\frac{\pi}{1-\pi} \rightarrow [0, \infty)$
- $L = \log\left(\frac{\pi}{1-\pi}\right) \rightarrow (-\infty, \infty)$

The **probit** function is another common transformation that achieves similar results.

- Probit: $Q_i = Z_{\pi_i} \rightarrow$ Z score (of a standard normal distribution) associated with the percentile π_i.

Other "S" shape curves exist, which tend to reach similar conclusions.

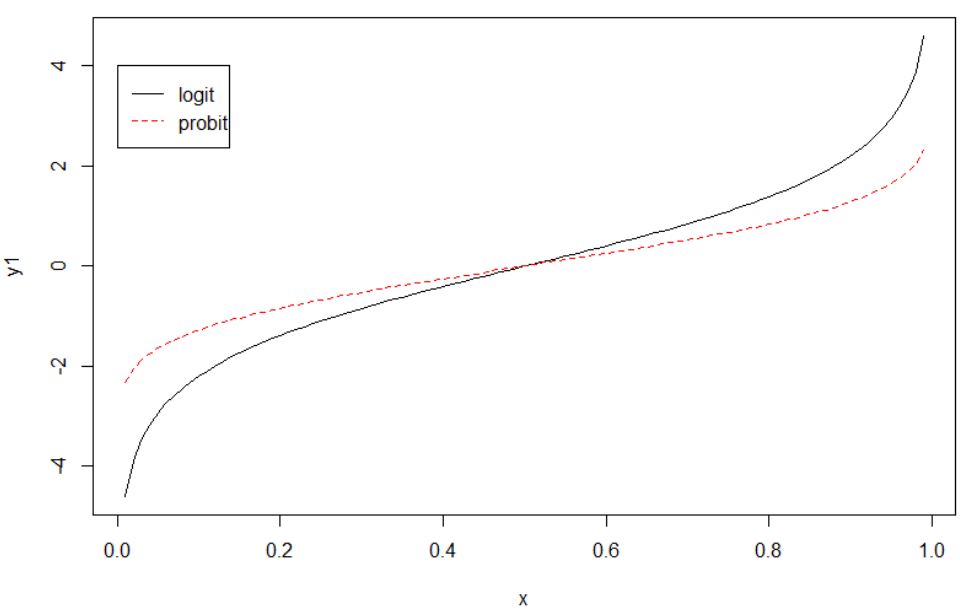

Figure 2: Visualization of logit and probit function for various probabilities.

3 Logistic Regression

$$L_i = \beta_0 + \beta_1 X_{i,1} + \cdots + \beta_{p-1} X_{i,p-1} + \epsilon_i$$

- b_k estimates obtained from MLE-based iterative procedure (Newton-Raphson, Fisher)
- Transform estimates $\hat{L}_i = b_0 + b_1 X_{i,1} + \cdots + b_{p-1} X_{i,p-1}$ back to probability scale.

$$\hat{\pi}_i = \frac{1}{1 + e^{-\hat{L}_i}} \qquad\qquad O\hat{d}ds_i = e^{\hat{L}_i}$$

3.1 Interpretation of Estimates

- $X_{i,1} = \cdots = X_{i,p-1} = 0 \implies \hat{L}_i = b_0 \implies O\hat{d}ds_i = e^{b_0}$
- Hold $X_{i,2} = \cdots = X_{i,p-1} = 0$, increase $X_{i,1}$ from 0 to 1

$$\implies \hat{L}_i = b_0 + b_1 \implies O\hat{d}ds_i = e^{b_0 + b_1} = e^{b_0} e^{b_1}$$

- Thus, an increase in one unit in X_j *multiplies the odds* (in favor of $Y = 1$) by a factor of e^{b_j}.
  - Note that it is the *odds* that are multiplied, **not** the probability.
- Alternative Interpretation: the odds of $Y = 1$ change by $100(e^{b_j} - 1)\%$ per unit increase in X_j while holding other predictors constant.
  - Example (Handout 5.1.1): b_j for sector is 1.57 $\implies e^{1.57} = 4.83$.
  - "Holding all other predictors constant, the odds of having disease are $100(4.83 - 1) = 383\%$ greater in Sector 2 than in Sector 1.

(Groups) How would you interpret the coefficient associated with Age in the Handout 5.1.1 logistic model?

"Holding all other predictors constant, the odds of having disease are $100(e^{0.297} - 1) = 3.01\%$ greater for each year increase in age."

- The "Odds Ratio" for X_j (odds of $Y = 1$ when $X_j + 1$ vs odds of $Y = 1$ when X_j)

$$\frac{e^{b_0 + b_1 X_1 + \cdots + \mathbf{b_j(X_j+1)} + \cdots + b_{p-1} X_{p-1}}}{e^{b_0 + b_1 X_1 + \cdots + \mathbf{b_j(X_j)} + \cdots + b_{p-1} X_{p-1}}} = e^{b_j}$$

3.2 Inference with Estimates

- Single Variable Test:
  - $H_0 : \beta_j = 0$ (X_j has no effect on $P(Y = 1)$).
  - Test statistic: $t = \frac{b_j}{SE\{b_j\}}$ (standard normal for "large" N).
  - $\implies t^2 \sim \chi^2_1$ (obtain confidence intervals from here)
    * This approach is called the "Wald Test"
- Subset variables test:

  - $H_0 : \beta_{p-H} = \cdots = \beta_{p-1} = 0$
    * reorder the X variables so that the subset we are checking for comes last
  - Let L_{full} be the likelihood associated with the full model
  - Test statistics: $\chi^2 = -2 \log \frac{L_{red}}{L_{full}}$
  - Under $H_0 : \chi^2 \sim \chi^2_H$

- Overall model test:

$$\text{Model}\chi^2 = -2 \log L_{intercept} + 2 \log L_{int\&covariates}$$

  - Often called the **deviance**, DEV or $DEV(X_0, X_1, X_{p-1})$

- Conditional Effect plot: predicted $\hat{\pi}$ vs one predictor X_j
  - While holding all other predictors at some constant level. The default level in SAS is the mean (average) of each variable.

4 Goodness of Fit Measures:

- Pseudo R-square: $\frac{\chi^2}{\chi^2+n}$ (χ^2 from model test)
- Hosmer-Lemeshow Goodness of Fit Test
  - H_0 : logistic regression response function is appropriate
  - Based on sorted $\hat{\pi}$ values, group observations into 5-10 roughly equal sized groups.
  - Within each group, look at the total observed numbers of $Y = 1$ and $Y = 0$
  - Based on the model fit, calculate the total *expected* numbers of $Y = 1$ and $Y = 0$.
  - Test statistic χ^2 is sum (across groups) of $\frac{(observed - expected)^2}{expected}$
- "Concordance" - look at all pairs of observations with different Y
  - Let n_c be the # of "concordant" pairs (observed $Y = 1$ has larger $\hat{\pi}$)
  - Let n_d be the # of "discordant" pairs (observed $Y = 1$ has smaller $\hat{\pi}$)
  - Let n_t be the # of "tied" paired (observed $Y = 1$ and $Y = 0$ have same $\hat{\pi}$ (likely due to identical X-profiles)
  - Define rank correlation indices (larger is better):

$$\text{Somers' } D = \frac{n_c - n_d}{n_c + n_d + n_t}$$

$$\gamma = \frac{n_c - n_d}{n_c + n_d}$$

$$\text{Tau-a} = \frac{n_c - n_d}{0.5(n-1)n}$$

$$\text{AUC} = \frac{n_c + 0.5 n_t}{n_c + n_d + n_t}$$

- ROC (Receiver Operating Characteristic) Curve
  - Sort all observations from the smallest to biggest $\hat{\pi}$.
  - At each position in the list:
    * Use $\hat{\pi}$ as threshold for $\hat{Y} = 1$, moving cutoff from the standard 0.5 threshold.
    * Calculate sensitivity: (proportion $Y_i = 1$ values with $\hat{Y}_i = 1$).
    * Calculate specificity: (proportion $Y = 0$ values with $\hat{Y} = 0$).
      · Sensitivity and Specificity - think smoke alarms and pregnancy tests.
    * False positive rate (prop $Y = 0$ values with $\hat{Y} = 1$) = 1 - specificity
    * Plot false positives against true positive rates (sensitivity)
    * Calculate the area under the curve.

Given the three ROC curves in Figure 3, which model has the best predictive power and why?

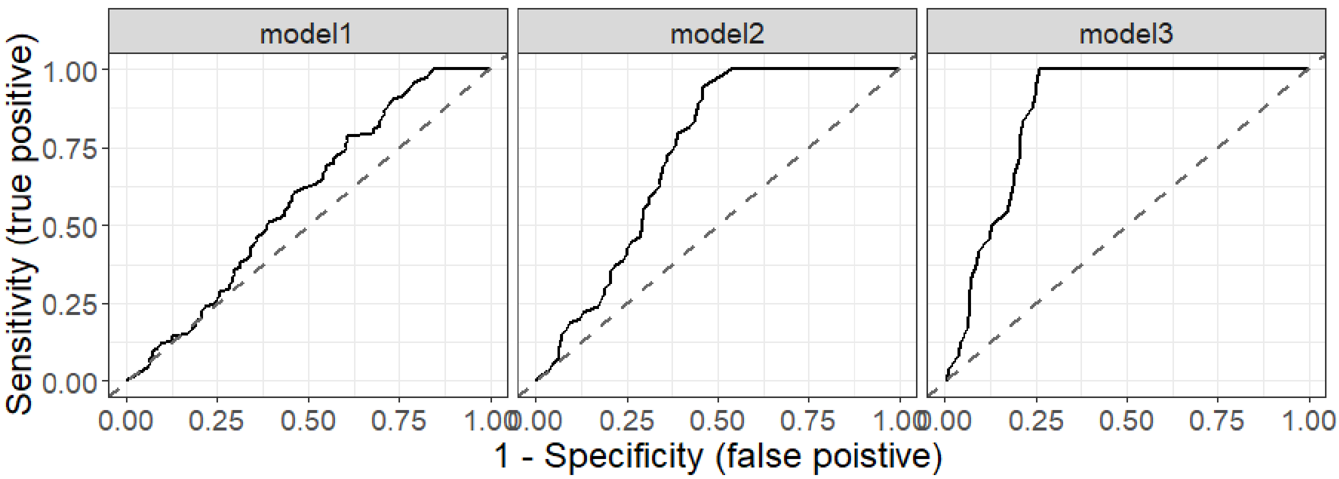

Figure 3: Comparison of three ROC curves.

Model 3 is the most accurate. The model sensitivity increases much faster than the false positive rate.

5 Multicollinearity

Recall that multicollinearity occurs when X variables are highly correlated with each other. It has **nothing** to do with the response variable Y.

As in OLS, multicollinearity inflates the variance of the b_k estimates, making them hard to interpret/test for significance.

As in OLS, stepwise selection and all possible regression methods exist to "score" each combination of explanatory variables and select a best model.

6 Outliers in Logistic Regression

(Individual) If Y can only take on two values (0 or 1), how are outlier values possible?

An outlier is a point for which the observation strongly disagrees with the predicted probability.

- Define "deviance residual" as

$$dev_i = \text{sign}(Y_i - \hat{\pi}_i)\sqrt{-2\left(Y_i \log \hat{\pi}_i + (1 - Y_i)\log(1 - \hat{\pi}_i)\right)}$$

  - The more certain we are (probability near 0 or 1), the more potential we have to be very wrong.
- $DEV(X_0, \cdots X_{p-1}) = \sum_i dev_i^2$
- "Outliers" are values not well represented by the model
- "Half-normal probability plot - observed $|dev_i|$ vs expected value under normality
  - **However,** since the residuals are not normally distributed, we asses differences from our expectation using simulations based on $\hat{\pi}_i$.
    - ∗ Create 19 simulations by generating a "new" response variable where the values of $Y_{new,i} \sim Bernoulli(\hat{\pi}_i)$
  - Simulated envelop (SEE 5.1.1 MACRO ON CANVAS) plots the minimum, maximum, and mean of the 19 simulations
    - ∗ Why 19 simulations? - Since our observed deviances represent the 20th observation, the probability that our deviances will fall outside the envelope is less than 5% IF the fitted model is appropriate.
    - ∗ Points falling outside in the envelop in the upper right corner of the plot are evidence of outliers/bad fits.

7 Influential Observations

Influential observations have the same effect on model coefficients as they did in OLS.

Diagnostics (similar to Leverage and DFBETAS):

- $\Delta D_i : DEV - DEV_{(i)}$
  - Measures decrease in "misfit" when obs. i is ignored. (essentially measures the "poorness of fit for observation i).
  - "large" $\Delta D_i \implies$ obs. i overly influences model fit
  - SAS: DIFDEV - on step difference in deviance
- ΔB_i
  - Similar to Cook's distance, measures influence of obs. i on the estimates b_j
  - SAS: C - confidence interval displacement C

- $\Delta\chi_i^2$
  - Similar to ΔD_i: "poorness of fit" for obs i
  - SAS: DIFCHISQ - one step difference in Pearson χ^2

Unlike in OLS, there is no consistent numerical rule of thumbs to determine thresholds for the Δ measures.

Instead, we will simply rely on graphical diagnostics.

- $\Delta D_i, \Delta B_i, \Delta X_i^2$ vs Observation Number - look for extreme values
- ΔD_i vs $\hat{\pi}_i$ (or ΔX_i^2 vs $\hat{\pi}_i$)
  - Look for points with low $\hat{\pi}$ but $Y_i = 1$ (upper left corner) OR high $\hat{\pi}$ but $Y = 0$ (upper right corner) which are much different than the overall pattern
  - (Optional) plot different size points where point size is determined by ΔB_i

8 Remedial Measures

Similar to OLS:

- Look for typos in the data
- Consider transformations of the X variables
- Consider dropping problematic points (only if you have a good argument for removing them).

9 Final Thought

If you have a lot of explanatory variables, you should strongly consider classification trees and random forest for classification.