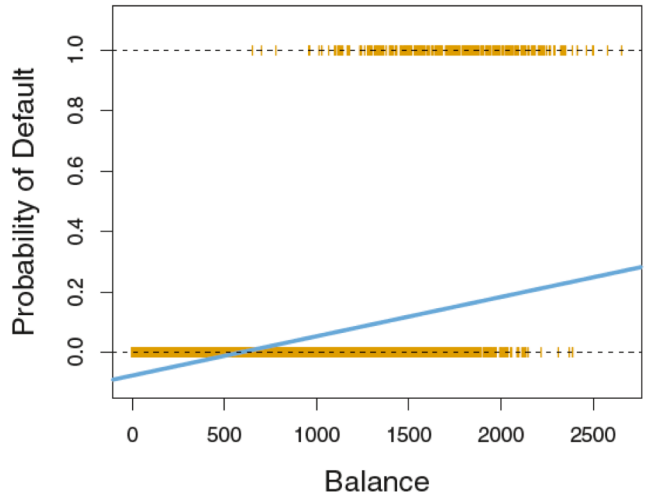
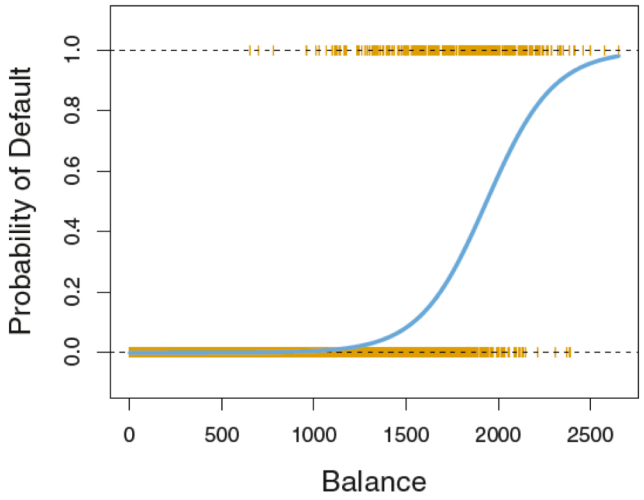
UNIVERSIDAD CATÓLICA BOLIVIANA “SAN PABLO”
FACULTAD DE INGENIERÍA
CARRERA INGENIERÍA DE SISTEMAS
MAESTRÍA EN CIENCIA DE DATOS

REGRESIÓN LOGÍSTICA

Logistic Regression

- La Regresión Logística modela la *probabilidad* que Y pertenezca a una categoría particular.





El modelo logístico

- En el caso del dataset Default se puede tener:
    - Pr(default = Yes | balance)
- Entonces para un valor dado de *balance* una predicción para *default* puede ser hecha.
- Por ejemplo, uno podría predecir *default = Yes* para un individuo para quien el *balance* > 0.5 o 0.3 o etc. Es decir, un threshold de acuerdo al caso.
- Para tener una respuesta entre cero y uno para los valores de X se utiliza la función logística,

$$p(X) = \frac{e^{\beta_0+\beta_1 X}}{1 + e^{\beta_0+\beta_1 X}}$$

- Ajustando esta formula tenemos,

$$\frac{p(X)}{1-p(X)} = e^{\beta_0+\beta_1 X}$$

- La cantidad *p(X)/[1-p(X)]* is llamada odds (i.e. posibilidades), y puede tomar cualquier valor entre 0 e ∞ que indica una probabilidad baja o muy alta para *default*.
- Por ejemplo, en promedio 1 en 5 personas con un odds de ¼ no pagará su deuda, p(X) = 0.2 lo que implica que (0.2/1-0.2) = ¼.
- Los odds son tradicionalmente utilizados en vez de posibilidades en, por ejemplo, en carreras de caballos.

- Tomando el logaritmos de ambos lados de la fórmula anterior,

$$\log\left(\frac{p(X)}{1-p(X)}\right) = \beta_0 + \beta_1 X$$

- El lado izquierdo de esta fórmula se denomina *log-odds* o *logit*

Maximun Likelihood

- En la Regresión Logística se utiliza el Maximun Likelihood para estimar los parámetros del modelo.

$$\ell(\beta_0, \beta_1) = \prod_{i:y_i=1} p(x_i) \prod_{i':y_{i'}=0} (1 - p(x_{i'}))$$

- El *likelihood* da la probabilidad de observar ceros y unos en los datos.
- Las estimaciones elegidas $\hat{\beta}_0$ y $\hat{\beta}_1$ para maximizar la función de probabilidades.

- Por ejemplo, cual sería la probabilidad de *default* para alguien con un balance de $1000.

$$\hat{p}(X) = \frac{e^{\hat{\beta}_0+\hat{\beta}_1 X}}{1+e^{\hat{\beta}_0+\hat{\beta}_1 X}} = \frac{e^{-10.6513+0.0055\times 1000}}{1+e^{-10.6513+0.0055\times 1000}} = 0.006$$

- Utilizando *student* como predictor,

$$\widehat{\Pr}(\texttt{default=Yes}|\texttt{student=Yes}) = \frac{e^{-3.5041+0.4049\times 1}}{1+e^{-3.5041+0.4049\times 1}} = 0.0431$$

Regresión Logística Múltiple

- Se puede generalizar el caso cuando se tiene un solo predictor, a cuando se tienen varios

$$\log\left(\frac{p(X)}{1-p(X)}\right) = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p$$

- Donde X = (X_1, … X_p) son *p* predictores.

$$p(X) = \frac{e^{\beta_0+\beta_1 X_1+\cdots+\beta_p X_p}}{1+e^{\beta_0+\beta_1 X_1+\cdots+\beta_p X_p}}$$

- Hay una función lineal por cada clase.

Características iniciales

- En la regresión lineal logística la variable dependiente puede ser binaria o multicategórica.
- Para una regressión binaria, el nivel “1” de la variable dependiente debiera representar el resultado deseado.
- Solamente las varibles más significativas deben ser incluidas.
- Las variables independientes deben ser independientes entre ellas. Es decir, el modelo no debiera tener multicolinearidad o muy poca.
- Las variables independientes son linealmente relacionadas al log(odds).
- La regresión logística require muestras grandes.

Bibliografía

- James G., Witten D., Hastie T., Tibshirani R.: “An Introduction to Statistical Learning”, Ed. Springer, 2013.

- Hastie T., Tibshirani R.: “The Elements of Statistical Learning”, Ed. Springer, 2013.