Ch4 Reading Notes

Alex Socarras

3/31/2022

Classification

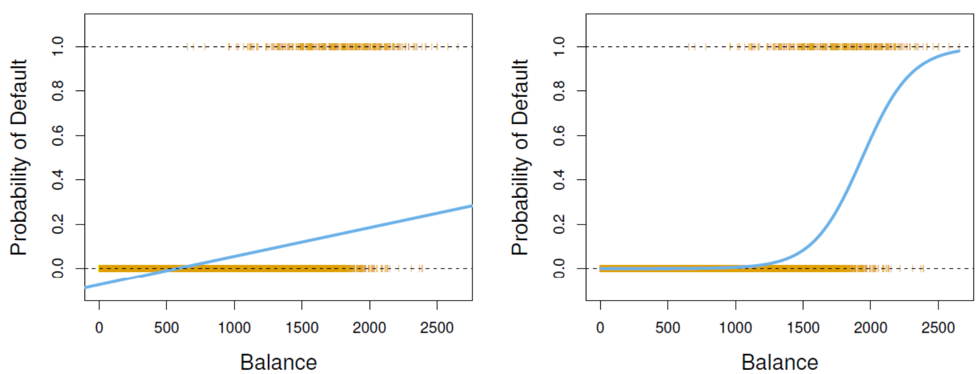

FIGURE 4.2. *Classification using the* `Default` *data.* Left: *Estimated probability of* `default` *using linear regression. Some estimated probabilities are negative! The orange ticks indicate the 0/1 values coded for* `default`*(*`No` *or* `Yes`*).* Right: *Predicted probabilities of* `default` *using logistic regression. All probabilities lie between* 0 *and* 1.

Why not linear regrssion?

a.) LR can't accommodate qualitative response with more than two classes (unless there's some natural, even-spaced ordering to the classes)

b.) LR can't provide meaningful estimates of $\Pr(Y|X)$ (some predictions might land outside [0,1] interval)

Logistic Regression

$$p(X) = \frac{e^{\beta_0+\beta_1 X}}{1+e^{\beta_0+\beta_1 X}} \qquad (4.2)$$

Odds: $$\frac{p(X)}{1-p(X)} = e^{\beta_0+\beta_1 X} \qquad (4.3)$$

Logit: $$\log\left(\frac{p(X)}{1-p(X)}\right) = \beta_0+\beta_1 X \qquad (4.4)$$

Note that unlike linear regression, the amount $p(X)$ changes from a one-unit change in X depends on the current value of X (i.e. it's derivative is not constant). But the sign of β_1 still determines effect of X on $p(X)$.

Likelihood function: $\ell(\beta_0, \beta_1) = \prod_{i:y_i=1} p(x_i) \prod_{i':y_{i'}=0} (1 - p(x_{i'}))$ (4.5)

$\hat{\beta}_0, \hat{\beta}_1$ chosen to maximize likelihood function. The function is constructed so as to generate $\hat{p}(x_i)$ for each observation as close to 0/1 based on their classification.

Multiple Logistic Regression

$$\log\left(\frac{p(X)}{1-p(X)}\right) = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p \qquad (4.6)$$

$$p(X) = \frac{e^{\beta_0+\beta_1 X_1+\cdots+\beta_p X_p}}{1+e^{\beta_0+\beta_1 X_1+\cdots+\beta_p X_p}} \qquad (4.7)$$

	Coefficient	Std. error	z-statistic	p-value
Intercept	−3.5041	0.0707	−49.55	<0.0001
student[Yes]	0.4049	0.1150	3.52	0.0004

TABLE 4.2. *For the* Default *data, estimated coefficients of the logistic regression model that predicts the probability of* default *using student status. Student status is encoded as a dummy variable, with a value of 1 for a student and a value of 0 for a non-student, and represented by the variable* student[Yes] *in the table.*

	Coefficient	Std. error	z-statistic	p-value
Intercept	−10.8690	0.4923	−22.08	<0.0001
balance	0.0057	0.0002	24.74	<0.0001
income	0.0030	0.0082	0.37	0.7115
student[Yes]	−0.6468	0.2362	−2.74	0.0062

TABLE 4.3. *For the* Default *data, estimated coefficients of the logistic regression model that predicts the probability of* default *using* balance*,* income*, and student status. Student status is encoded as a dummy variable* student[Yes]*, with a value of 1 for a student and a value of 0 for a non-student. In fitting this model,* income *was measured in thousands of dollars.*

Why does the sign of the coefficient for student change from simple to multiple logistic regression? Students *tend to have higher credit balance* and thus, on average, have a higher default risk. But at a fixed credit balance, students have a lower default rate than non-students! This is called **confounding.**

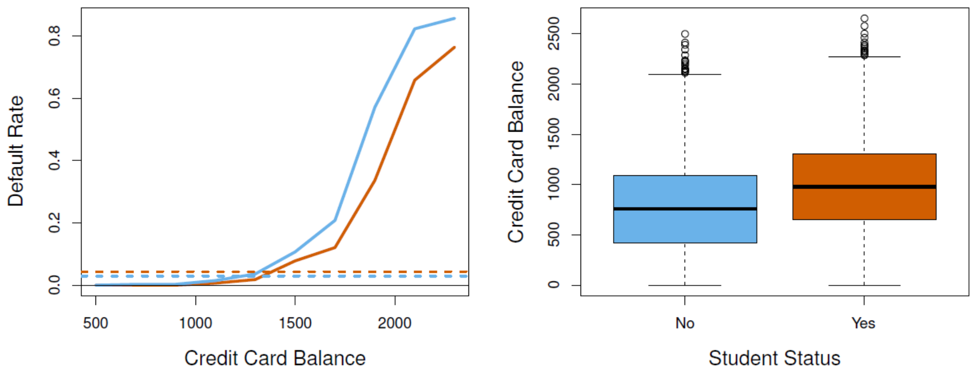

FIGURE 4.3. *Confounding in the* `Default` *data.* Left: *Default rates are shown for students (orange) and non-students (blue). The solid lines display default rate as a function of* `balance`*, while the horizontal broken lines display the overall default rates.* Right: *Boxplots of* `balance` *for students (orange) and non-students (blue) are shown.*

Multinomial Logistic Regression: K $>$ 2 classes

WLOG, choose K-th class as the *baseline*. Then, for $k = 1, \ldots, K-1$:

$$\Pr(Y = k|X = x) = \frac{e^{\beta_{k0}+\beta_{k1}x_1+\cdots+\beta_{kp}x_p}}{1 + \sum_{l=1}^{K-1} e^{\beta_{l0}+\beta_{l1}x_1+\cdots+\beta_{lp}x_p}} \qquad (4.10)$$

$$\Pr(Y = K|X = x) = \frac{1}{1 + \sum_{l=1}^{K-1} e^{\beta_{l0}+\beta_{l1}x_1+\cdots+\beta_{lp}x_p}} \qquad (4.11)$$

$$\log\left(\frac{\Pr(Y = k|X = x)}{\Pr(Y = K|X = x)}\right) = \beta_{k0} + \beta_{k1}x_1 + \cdots + \beta_{kp}x_p$$