Exercise solutions for exercise class 5b in MMS075, Feb 18, 2020

As we did not finish all exercises in exercise class 5a, we start with the three remaining ones.

NOTE: there was a mistake in the specification of exercise 6 from yesterday: all plus signs in the formulas in parts 6a and 6b should have been minus signs. (The formulas with plus signs are also correct and can be proved similarly, but they do not correspond to the desired properties of the odds and logit functions). This mistake has been corrected in the version provided below.

1. For a probability value $0 < p < 1$, the **odds** is defined as $p/(1-p)$, and the **logit** (or **log-odds**) is defined as the logarithm of the odds, i.e. as $\text{logit}(p)=\log(p/(1-p))$. Compute the odds and logit values for the following probabilities:
   a. p=0.1; Odds: 0.1/0.9 = 0.111. Log-odds: log(0.111) = -2.198
   b. p=0.3; Odds: 0.3/0.7 = 0.429. Log-odds: log(0.429) = -0.846
   c. p=0.5; Odds: 0.5/0.5 = 1. Log-odds: log(1) = 0
   d. p=0.7; Odds: 0.7/0.3 = 2.333. Log-odds: log(2.333) = 0.847
   e. p=0.9. Odds: 0.9/0.1 = 9. Log-odds: log(9) = 2.197

   Note that the log-odds are negative for p<0.5 and positive for p>0.5. Note also that logit(0.5-x)= (-1)*logit(0.5+x). This is because the odds for these values are reciprocals of each other.

   NOTE2: When using your calculator to compute the log-odds values, you need to use the ln button! As mentioned in the lecture and in the ISL book, log will be used to denote logarithm with base e (as it is most common in the scientific literature); however, on calculators, this corresponds to the ln button, while the log button is used for logarithm with base 10. This is unfortunate and confusing, but you still need to remember this, otherwise you get wrong results.

2. Show that the odds and logit functions are increasing in p, i.e., show that:
   a. if $p_2>p_1$, then $p_2/(1-p_2) > p_1/(1-p_1)$;
      Showing that $p_2/(1-p_2) > p_1/(1-p_1)$ is equivalent to showing the same inequality with both sides multiplied by $(1-p_1)(1-p_2)$. Doing this multiplication on both sides gives that we need to show $p_2(1-p_1) > p_1(1-p_2)$. Getting rid of the parentheses, we need to show that $p_2-p_2p_1 > p_1-p_1p_2$. This inequality indeed holds whenever $p_2>p_1$ (because the product of p_1 and p_2 cancels).

   b. if $p_2>p_1$, then $\log(p_2/(1-p_2)) > \log(p_1/(1-p_1))$!
      If $p_2>p_1$, then we know from part a) that $p_2/(1-p_2) > p_1/(1-p_1)$. We also know that log is an increasing function, so the log of the greater value $p_2/(1-p_2)$ will indeed be greater than the log of the smaller value $p_1/(1-p_1)$.

3. Express the value of p from the value of the logit function; that is, assuming that we have $\log(p/(1-p)) = y$, express p as a function of y!

$$\log\left(\frac{p}{1-p}\right) = y \iff \exp\left(\log\left(\frac{p}{1-p}\right)\right) = \exp(y), \text{ i.e. } \frac{p}{1-p} = e^y$$

   Multiplying the latest equation by 1-p, we get that:

$$\frac{p}{1-p} = e^y \iff p = e^y(1-p) = e^y - pe^y \iff p + pe^y = e^y.$$

We can express p from the latest equation by noting that the left side is p(1+exp(y)):

$$p = \frac{e^y}{1+e^y}$$

4. Consider the logistic regression model predicting the probability of credit card default based on balance, as discussed in the lecture:

$$\log\left(\frac{\hat{p}(X)}{1-\hat{p}(X)}\right) = \hat{\beta}_0 + \hat{\beta}_1 \times \text{balance}$$

Instead of the observed value of 0.0055, describe a correct interpretation of the following coefficient estimates in terms of log-odds, odds and probability:

   a. 1.1; This coefficient value would mean that $1 increase in balance would increase the log-odds of default by 1.1. As exp(1.1) = 3.004, a $1 increase of balance would correspond to a threefold increase in the odds of default. (Note that this would be an extreme consequence of a $1 change in balance, so it is not surprising that the actual coefficient estimate is much smaller than 1.1.) The probability of default would increase when increasing the balance value.
   b. -1.1; A $1 increase in balance would decrease the log-odds of default by 1.1. As we have exp(-1.1) = 0.333, a unit increase of balance would decrease the odds of default to a third of its original value. The probability of default would decrease when increasing the balance value.
   c. 0. This coefficient value indicates no change in the log-odds of default when changing the balance value. As exp(0)=1, increasing the balance would not have any effect on the odds or probability of default.

5. Consider the same logistic regression model as in part 4, this time with the correct coefficient estimates. We have seen in the lecture that the probability of default can be estimated from this model as follows:

$$\hat{p}(X) = \frac{e^{-10.65+0.0055\times\text{balance}}}{1+e^{-10.65+0.0055\times\text{balance}}}$$

Using this formula, estimate the probability of default for the following balance values:

   a. $1500; Predicted probability of default: 0.083, i.e. 8.3%. This follows from the following computation:

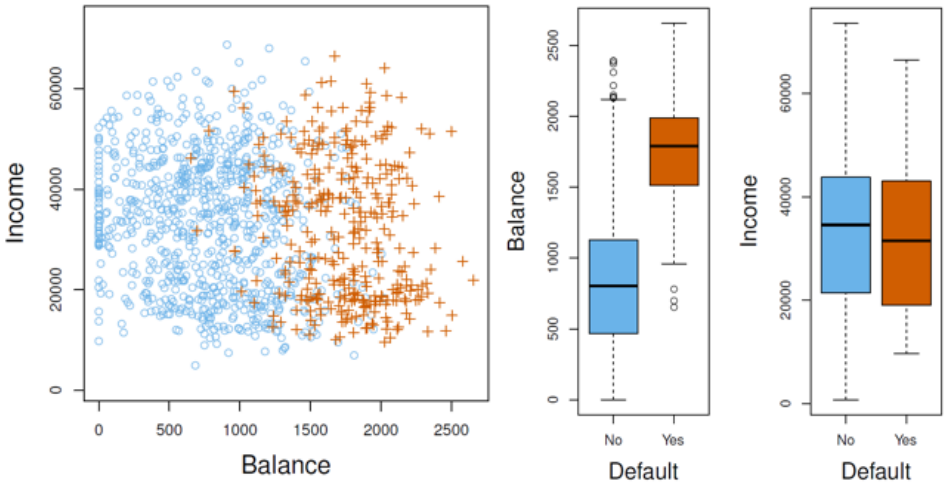
$$\hat{p}(X) = \frac{e^{-10.65+0.0055\times1500}}{1+e^{-10.65+0.0055\times1500}} = \frac{e^{-2.4}}{1+e^{-2.4}} = 0.0831 = 8.31\%$$

   b. $2000; Predicted probability of default: 0.587, i.e. 58.7%.
   c. $2500. Predicted probability of default: 0.957, i.e. 95.7%.

Which balance value gives the highest probability of default? Could you answer this question before computing the estimates?

The balance of $2500, i.e. the highest balance considered gives the highest probability of default. This is clear in advance from the interpretation of coefficients: as the coefficient of balance is positive, we have that each increase of balance is associated with an increased probability of default.

6. Relate the results of exercise 5 to Figure 4.1 in ISL, see below. Which plot(s) would help most in correctly anticipating the results before doing the computation?



The Income vs Default box plots do not help with exercise 5, because those predictions are concerned with balance, and not income, as a predictor.
The Balance vs Default box plots are somewhat informative: up to Balance=1000, there are hardly any individuals defaulted, hence the probability of default should be very low for such cases. For values of Balance>=2500, there are only defaulted cases, so the probability of default must be very high for such values. However, the implications of the box plots for Balance values between 1000 and 2500 strongly depend on the case count in each category. The sample contains way more people with Default="No" that with Default="Yes"; this is why we have only about 60% default probability for Balance=2000 even though it looks like there are very few not defaulted cases at this balance value while about 25% of defaulted individuals have balance values >=2000. We will look more at this together in Lecture 6.
When looking at the scatter plot, we would expect that at any given value of balance (i.e. along any vertical line in the plot), the share of defaulted cases (orange crosses) among all points with the same balance should be close to the probability of default. Therefore, based on the scatter plot, we could anticipate approximately 50% default probability at balance=1500, around 90% default probability at balance=2000 and very close to 100% default probability at balance=2500. The predicted probabilities in exercise 5 are much lower than that. In fact, the scatter plot presented here is misleading, and the reason will be discussed at the next lecture. (If you are curious about the reason, read page 128 carefully and you may find it somewhere hidden in the text.)

7. Feedback quiz (optional): Go to www.menti.com and use the code 32 21 69.