Contexts in which probability cut-offs can be easily interpreted

Imagine we want to see if a new medical screening test is able to predict whether a tumor is malignant based on the size of the tumor. So the dependent variable *y* is going to be Malignant (1=Yes, 0=No) and the predictor *x* is going to be the length of a tumor (measured in cm).

We can use logistic regression in this case. Imagine we have the following output:

|  |  |  |
| --- | --- | --- |
|  | **Beta Coefficient** | **P-Value** |
| Intercept | -5.234 | <0.0001 |
| Tumor Size (cm) | 0.827 | <0.0001 |

Recall from the slides that = . This means that for various values of our predictor variable *x*, we can estimate the probability that y = 1. Said differently, for various tumor sizes, we can estimate the probability that the tumor is malignant. This is exactly what we do in the table below for various values of *x*. For example, using the formula and values of and above, the probability that Y=1 (i.e., that the tumor is malignant) when tumor size is 2.4 cm may be calculated as = = .2996 ≈ .3.

Imagine that for the different probability cut-offs, we calculate the misclassification rate sensitivity, specificity, and based on that, plot the ROC curve. Using the methods we discussed in class, we can choose a cut-off value that will somehow optimize the sensitivity and specificity Because this is a medical test, where a false negative – i.e., failure to detect a malignancy, is generally more problematic than a false positive – i.e., detection of a malignancy even though the tumor is actually benign, we would probably want to weigh sensitivity much more than specificity. Imagine that this optimal cut-off is somehow determined to be .3 (e.g., we can use the Youden Index, or look at the cut-off value that minimizes the distance from the ROC Curve to the upper left corner of the curve, or use alternative methods where we weigh sensitivity more than specificity). In that case, from the table below, we can see that this cut-off corresponds to a tumor size of roughly 2.4 cm. This value is a clinically meaningful cut-off that tells us that tumors which are 2.4 cm or more in size have a high(er) probability of being malignant, and the value of 2.4 cm is used *because* this is the value of *x* which corresponds to the probability cut-off at which the sensitivity & specificity were determined to be optimal.

|  |  |  |  |
| --- | --- | --- | --- |
| **Tumor Size (cm)** |  | **Sensitivity** | **Specificity** |
| 1.2 | 0.045582 |  |  |
| 1.6 | 0.090232 |  |  |
| 2.4 | 0.299601 | 0.9841 | 0.8401 |
| 3 | 0.561438 |  |  |
| 4.4 | 0.942935 |  |  |
| 4.6 | 0.959697 |  |  |