Given a population (or random sample) *P* of positive (*YP* = 1) and negative (*YP* = 0) controls, a selected (i.e., non-random) sample *S* of positive (*YS* = 1) and negative (*YS* = 0) controls from *P*, and a vector *X* of predictor variables (i.e., prediction scores), the logit function for *X* in *S* (*XS*) is,

|  |  |
| --- | --- |
| $$ln\left[\frac{P(Y\_{S}=1|X\_{S})}{P(Y\_{S}=0|X\_{S})}\right]=α\_{S}+β\_{S}X\_{S}$$ | (1) |

where *αS* and *βS* are, respectively, the constant and vector of coefficients of *X* from logistic regression on *S*. Similarly, the logit function for *X* in *P* (*XP*) is,

|  |  |
| --- | --- |
| $$ln\left[\frac{P(Y\_{P}=1|X\_{P})}{P(Y\_{P}=0|X\_{P})}\right]=α\_{P}+β\_{P}X\_{P}$$ | (2) |

where *αP* and *βP*are, respectively, the constant and vector of coefficients of *X* from logistic regression on *P*.

Using the Bayes' theorem, the odds of being a positive control in *S* is

(3)

Similarly,

|  |  |
| --- | --- |
|  | (4) |

Since the conditional distribution of *X* given *Y* is unaffected by the sampling on *Y*, i.e.,

 and 

From Eq.3 and 4, we have,

|  |  |
| --- | --- |
|  | (5) |

Substituting Eq. 1 and 5 into Eq. 2, we get,

|  |  |
| --- | --- |
| $$ln\left[\frac{P(Y\_{P}=1|X\_{P})}{P(Y\_{P}=0|X\_{P})}\right]=ln\left[\frac{P(Y\_{S}=1)}{P(Y\_{S}=0)}\right]-ln\left[\frac{P(Y\_{P}=1)}{P(Y\_{P}=0)}\right]+α\_{S}+β\_{S}X\_{S}$$ | (6) |

Since $P\left(X\_{P}\right)=1-P(Y\_{P}=1|X\_{P})$, we obtain

|  |  |
| --- | --- |
| $$P\left(X\_{P}\right)={1}/{\left[1+Re^{-(α\_{S}+β\_{S}X\_{S})}\right]}$$ | (7) |

where$ R={\frac{P(Y\_{S}=1)}{P(Y\_{S}=0)}}/{\frac{P(Y\_{P}=1)}{P(Y\_{P}=0)}}$