

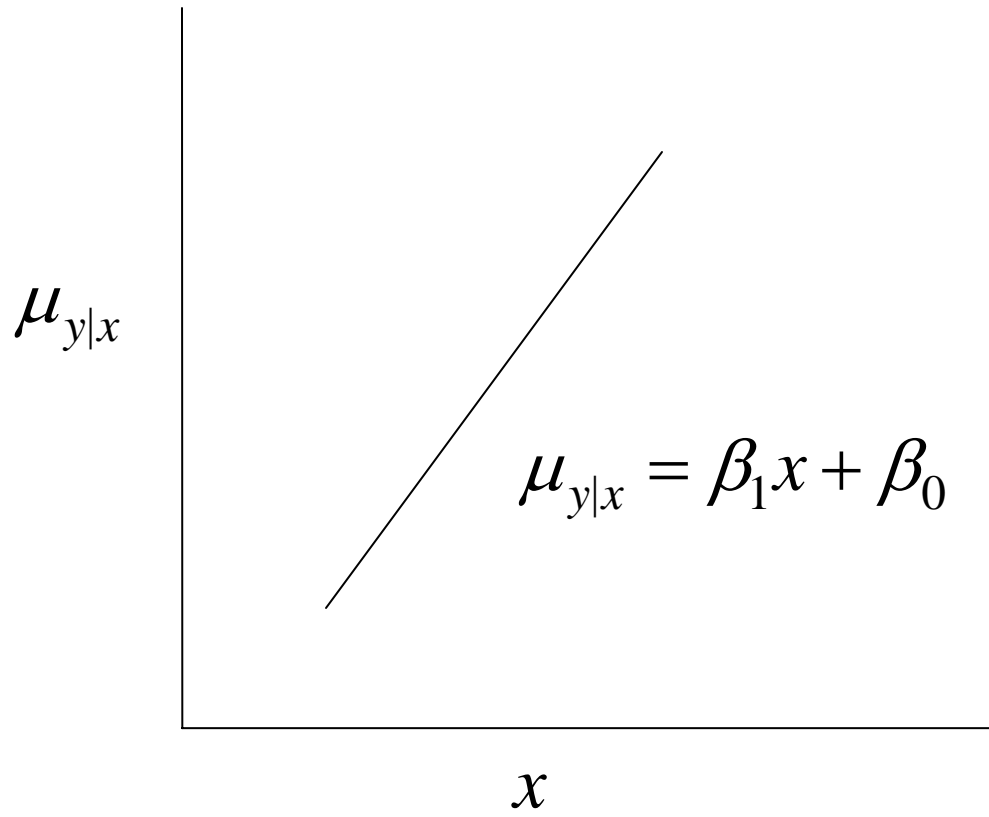
Binary Logistic Regression

In ordinary linear regression with continuous variables, we fit a straight line to a scatterplot of the X and Y data. The regression line is

$$\hat{y}_i = \beta_1 x_i + \beta_0 \quad (1)$$

It is important to remember that, when fitting the scatterplot, we estimate an equation for (a) predicting a Y score from the X score, but also for (b) estimating the *conditional mean* $\mu_{y|x=a}$ for a given value a .

In Psychology 310, you may recall that we exploited this fact to perform hypothetical distribution calculations for weight given height.



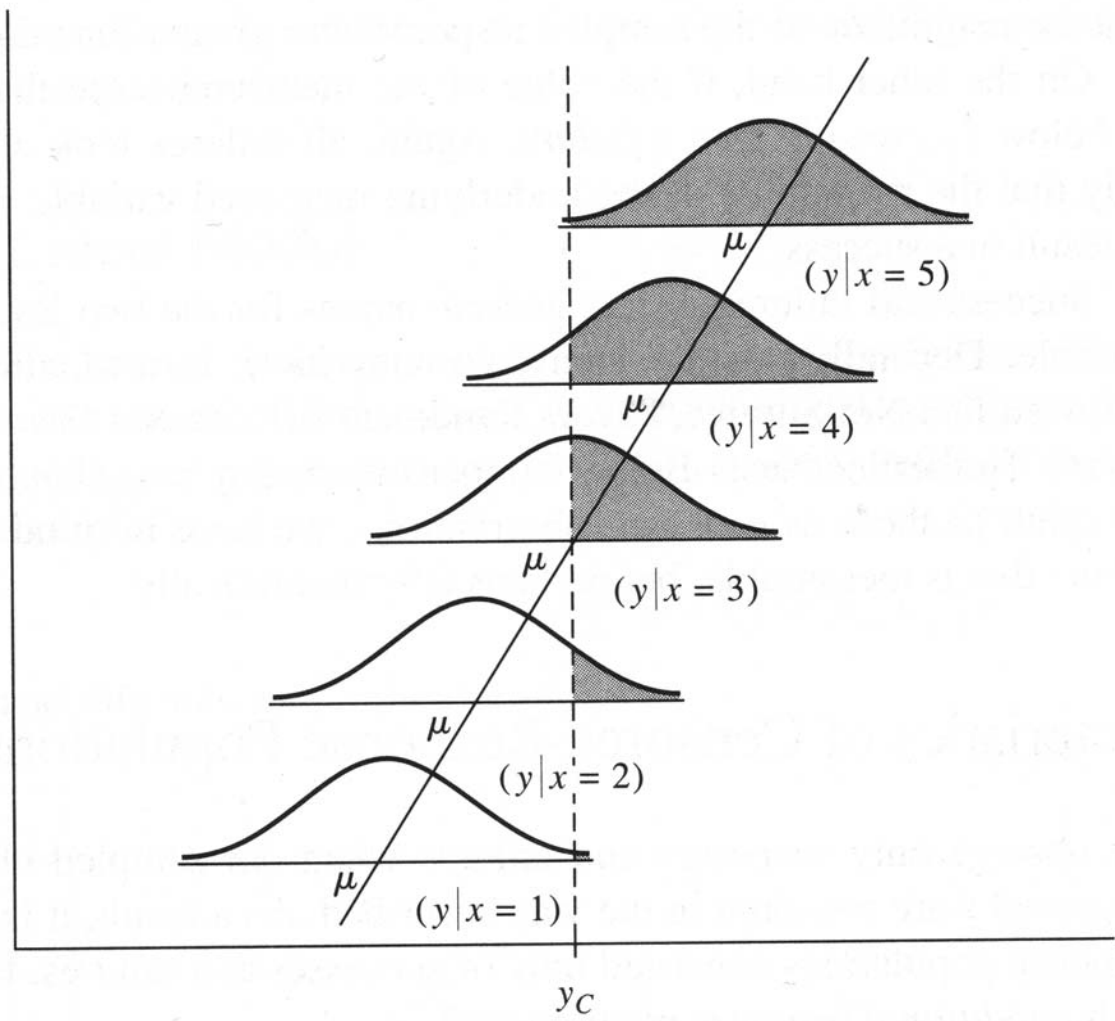
Binary Data as “Censored” Information

Suppose that the X - Y data are bivariate normal, but that the scores on Y are censored in the following way. There is a *threshold* or *critical* value y_C , and if an underlying y value exceeds the threshold, the observed score is $y = 1$, otherwise $y = 0$.

How will this censoring affect the conditional mean?

The Mean of a Binary Censored Normal Variable

Recall that, with a binary variable, the mean is equal to the probability that $y = 1$. So the conditional mean is simply the probability that $y > y_C | x$.



In this plot, we have reversed the usual positions of x and y . Notice that, as x increases, the predicted values of y increase in a straight line, and so does the conditional mean of y for that value of x . The conditional normal distributions are represented. The area above y_c is shaded in. This area is not only the probability that the binary “censored” version of y will be equal to 1, it is *also the conditional mean of the censored (binary) version of y* . If you examine the size of the shaded areas, you see the key fact. The relationship between the conditional mean of the censored y and x is not linear!

We can easily plot the relationship, as in the following example. Suppose the original data are in standard score form, and the population correlation is 0.60. Then the regression line is

$$\hat{y} = \mu_{y|x} = .6x \quad (2)$$

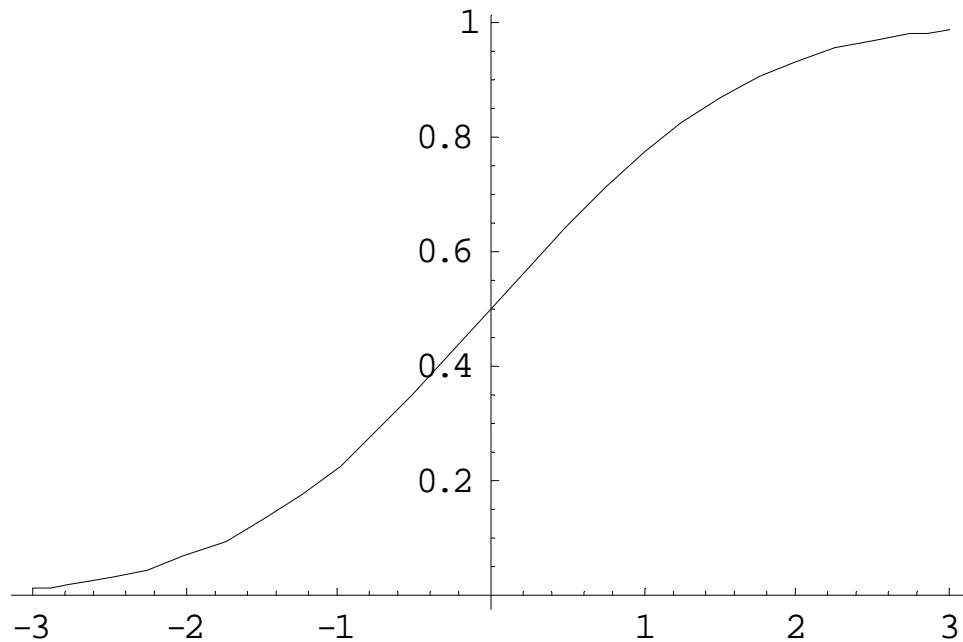
The formula for the conditional mean of the *censored version of y* (y^*) is

$$\begin{aligned}
\Pr(y > y_c | x) &= \pi(x) = 1 - \Phi\left(\frac{y_c - .6x}{\sqrt{1 - .6^2}}\right) \\
&= 1 - \Phi\left(\frac{y_c - .6x}{.8}\right) \\
&= \Phi\left(\frac{.6x - y_c}{.8}\right)
\end{aligned} \tag{3}$$

Note that if the cutoff point is at 0, the equation becomes

$$\Pr(y > y_c | x) = \mu_{y^*|x} = \Phi(.75x) \tag{4}$$

So the plot of the conditional mean of y^* versus x will have the same shape as the cumulative distribution function of the normal curve.



So there are several reasons why we would not want to use simple linear regression to predict the conditional mean $\pi(x)$, say as

$$\pi(x) = \beta_1 x + \beta_0 \quad (5)$$

First, we realize that the relationship is almost certainly not going to be linear over the whole range of x , although it may well be quite linear over a significant middle portion of the graph.

Second, Equation (5) can generate improper values, i.e., values greater than 1 or less than 0.

Third, the standard assumption of equality of variance of conditional distributions is clearly not true, since, as you recall from our study of the binomial,

$$\text{Var}(y^* | x) = \pi(x)[1 - \pi(x)] \quad (6)$$

which varies as a function of x .

So rather than fitting a linear function to $\pi(x)$, we should fit a nonlinear function. Examining Equation (3) again, we see that it can be written in the form

$$\pi(x) = \Phi(\alpha + \beta x) \quad (7)$$

Since Φ is invertible, we can write

$$\Phi^{-1}[\pi(x)] = \alpha + \beta x \quad (8)$$

This is known as a *probit* model. It is a special case of a *Generalized Linear Model* (GLM), which, broadly speaking, is a linear model for a transformed mean of a variable that has a distribution in the *natural exponential family*.

Binomial Logit Models

Suppose we simply assume that the response variable has a binary distribution, with probabilities π and $1 - \pi$ for 1 and 0, respectively. Then the probability density can be written

$$\begin{aligned} f(y; \pi) &= \pi^y (1 - \pi)^{1-y} \\ &= (1 - \pi) [\pi / (1 - \pi)]^y \\ &= (1 - \pi) \exp\left(y \log \frac{\pi}{1 - \pi}\right) \end{aligned} \tag{9}$$

Now, suppose the log-odds of $y = 1$ given x are a linear function of x , i.e.,

$$\text{logit}[\pi(x)] = \log \frac{\pi(x)}{1 - \pi(x)} = \alpha + \beta x \quad (10)$$

The logit function is invertible, and so

$$\pi(x) = \frac{\exp(\alpha + \beta x)}{1 + \exp(\alpha + \beta x)} \quad (11)$$

Interpreting Logistic Parameters

In the above simple case, if we fit this model to data, how would we interpret the estimates of the model parameters?

Exponentiating both sides of Equation (10) shows that the odds are an exponential function of x . The odds increase multiplicatively by e^{β} for every unit increase in x .

So, for example, if $\beta = .5$, the odds are multiplied by 1.64 for every unit increase in x .

Also, if we take the derivative of $\pi(x)$ with respect to x , we find that it is equal to $\beta\pi(x)[1 - \pi(x)]$. So locally, the probability of x is increasing by $\beta\pi(x)[1 - \pi(x)]$ for each unit increase in x .

This in turn implies that the steepest slope is at $\pi(x) = 1/2$, at which $x = -\alpha / \beta$. In toxicology, this is called LD_{50} , because it is the dose at which the probability of death is $1/2$.

The intercept parameter is of less interest.

Example. Agresti (2002, Chapter 5) presents a simple example, predicting whether a female crab has a “satellite,” i.e., a male living within a defined short distance, on the basis of biological characteristics of the female.

1. Load data into SPSS, create new variables “has_sat” by computing $\text{sat} > 0$.

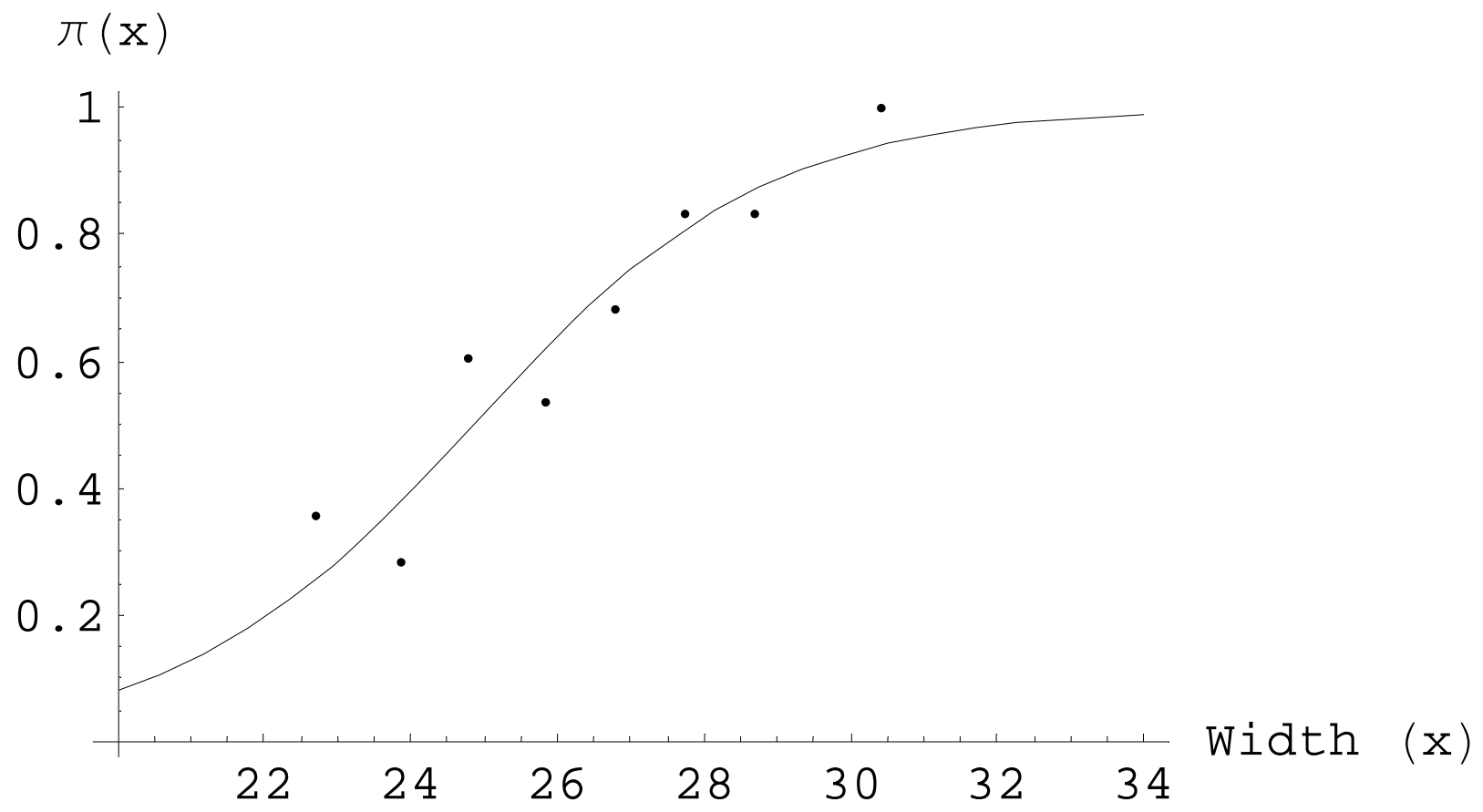
2. Analyze → Regression → Binary Logistic.

Results.

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step _a	W	.497	.102	23.887	1	.000	1.644
1	Constant	-12.351	2.629	22.075	1	.000	.000

a. Variable(s) entered on step 1: W.



Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)	95.0% C.I. for EXP(B)	
								Lower	Upper
Step 1 ^a	RACE	.055	.289	.037	1	.848	1.057	.600	1.861
	AZT	-.719	.279	6.651	1	.010	.487	.282	.841
	Constant	-1.074	.263	16.670	1	.000	.342		

a. Variable(s) entered on step 1: RACE, AZT.

Casewise List

Case	Selected Status ^a	Observed	Predicted	Predicted Group	Temporary Variable	
		SYMPTOMS			Resid	ZResid
1	S	1**	.150	0	.850	2.384
2	S	0	.150	0	-.150	-.419
3	S	1**	.265	0	.735	1.664
4	S	0	.265	0	-.265	-.601
5	S	1**	.143	0	.857	2.451
6	S	0	.143	0	-.143	-.408
7	S	1**	.255	0	.745	1.711
8	S	0	.255	0	-.255	-.585

a. S = Selected, U = Unselected cases, and ** = Misclassified cases.