Data Analysis In-Class Worksheet #14: Logistic Regression

Part #1: Logistic Regression Fundamentals

- **Probability** = favorable outcomes ÷ total outcomes
- **Odds** = favorable outcomes ÷ unfavorable outcomes

What does the “odds” measure? the likelihood of a particular outcome

If the probability of winning a 30% coupon at Kohl’s is 10%, then what’s the odds of winning a 30% coupon? ___=(0.1/(1-0.1))= (.1/0.9)=1/9=0.111

$\log_{10}(10000)$ = _4___

Why do we need to use log(odds), not probabilities, in estimating logistic regression?

Reason #1: We use **log-odds** in logistic regression because probabilities are bounded between 0 and 1, but odds can be transformed (with logs) to cover the whole number line. This lets us apply linear regression techniques while still predicting valid probabilities in the end.

Reason #2: Logistic regression doesn’t model probabilities directly because they’re bounded; it models **log-odds**, which extend infinitely in both directions. This way, even with near-constant predictors, estimation is still possible.

Suppose we have this logistic regression model.

$\hat{Y}$ = 2 + 1 * BED

If BED = 3, then $\hat{Y}$ = __5___

$\hat{Y}$ is □ probability □ odds □ ln(odds)

Convert $\hat{Y}$ back into probability using the _sigmoid_ function = $\frac{1}{1+e^{-\hat{y}}}$

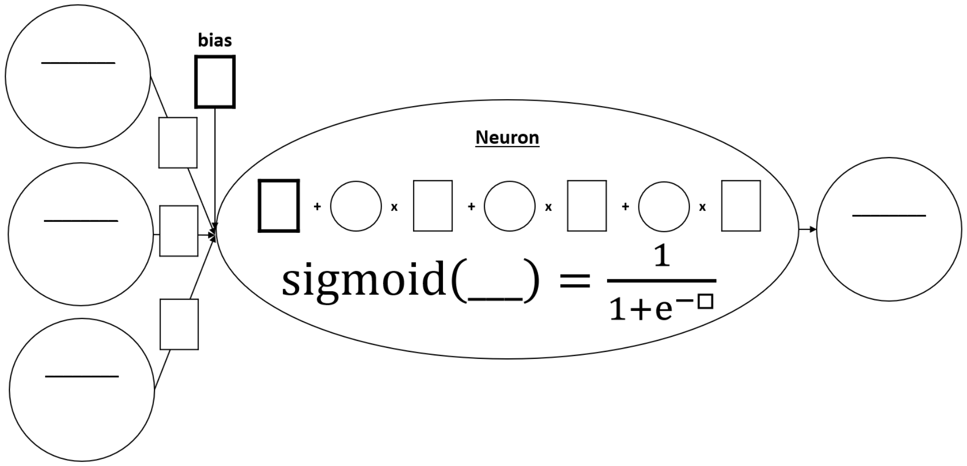
=0.9933_________

(Reminder: You can use the formula set up in Google Sheets to compute the value.)

Why do we need to convert $\hat{Y}$ back into probability?
Because probability is easier for the general audience to understand than logit. We can also easily convert probability into binary decisions.

Make a binary decision using the probability:
Convert any probability > .5 into YES, and any probability <= .5 into NO. The decision is ___YES__.

Illustrate this logistic regression model as a deep learning neuron with a sigmoid unit:



Part #2: Logistic Regression Application

Does Property Age predict the Property's Zipcode Location?

To answer this question, we can rephrase it as a pair of hypotheses:

H0: Age does not predict Zipcode
H1: Age predicts Zipcode

Express these hypotheses in terms of the regression β_1 slope

H_0: $\beta_1 = 0$__

H_1: _$\beta_1 \neq 0$__

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

For H_0: $\hat{Y} = \beta_0$

$$p = \frac{1}{1 + e^{-\beta_0}}$$

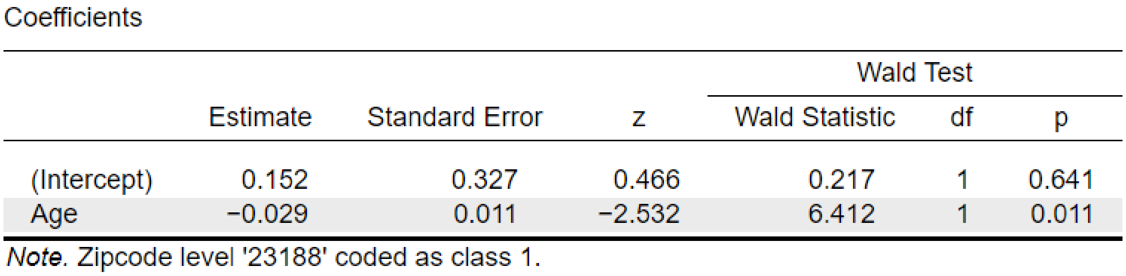
For H_1: $\hat{Y} = \beta_0 + \beta_1 * X$

In our class dataset, Zipcode has only two possible outcomes: 23185 and 23188. Therefore, Zipcode is a _________ variable.

- Binary (categorical)
- Numeric

If the probability (p) of being in 23185 is 80%, then the probability of being in 23188 would be _20_%. The odds of being in 23185 would be _80%_ / _20%_ = _4_. The natural log of this odds is _ln(4) = 1.386__

If the probability (p) of being in 23185 is 50%, then the probability of being in 23188 would be _50_% The odds of being in 23185 would be _50%_ / _50%_ = _1_. The natural log of this odds (ln(odds)) is _ln(1) = 0_



Coefficients

				Wald Test		
	Estimate	Standard Error	z	Wald Statistic	df	p
(Intercept)	0.152	0.327	0.466	0.217	1	0.641
Age	−0.029	0.011	−2.532	6.412	1	0.011

Note. Zipcode level '23188' coded as class 1.

Ŷ = ____0.152______ + __(-0.029)_______ * Age

The predicted value (Ŷ) from a logistic regression is

⬜ Probabilities

⬜ Binary categories

⬜ Odds

⬜ Natural log of odds (ln(odds) or logit)

The intercept is __0.152___, which tells us the log(odds) of zipcode 23188 when the property is new construction (age = 0)
The slope is __(-0.029)__, which tells us change in log(odds) of zipcode 23188 given an additional year of age

When a property is a year older, its odds of being in 23188 (Class 1) would exp(-.029) or 97.1%

Abdullah's Ŷ = 0.429. Convert it into odds: exp(.429) = 1.536

Convert odds into probability value: exp(0.429)/ (1 + exp(0.429)) = 0.606

Use the probability value to predict if the property is in 23185 or 23188:

We predict the property to be in class 1 (23188), because 0.606 is > .5 which is our cut-off threshold._

Abdullah's property is located in 23188.
This is a □ correct □ incorrect prediction.

If we get a confusion matrix like this one, calculate the overall accuracy rate: (4 + 24) / (4 + 6 + 16 + 24) = 56%

		Predicted	
		23185	23188
Actual	23185	4	6
	23188	16	24

APA Writeup

A logistic regression was performed to ascertain the effect of property age on the likelihood of being in 23188. The logistic regression model was statistically (in)significant, x^2 (_93_) = __4.824____, p _0.028___, McFadden's R^2 = __0.042___. The model correctly classified ___0__% of cases.