# LOGISTIC REGRESSION W/ UNGROUPED DATA.

IN OLS, GROUPED VS. UNGROUPED DIDN'T REALLY MATTER.

UNGROUPED DATA: EITHER HAVING AT LEAST ONE CONTINUOUS OR INTERVAL COVARIATE OR HAVING A LARGE # OF CATEGORICAL COVARIATES (SUCH THAT EACH LEVEL COMBINATION HAS ALMOST NO DATA).

EXAMPLE: AGE IS COMMONLY UNGROUPED

↳ ARTIFICIALLY GROUP DATA. THIS IS OFTEN UNSATISFACTORY.

1) ANALYSIS HINGES DRAMATICALLY ON GROUPING.

2) CAN HAVE REALLY SMALL COUNTS.

ANOTHER EXAMPLE: LOGISTIC REGRESSION IS COMMONLY USED TO ANALYZE SURVEY DATA. TYPICAL SURVEY MIGHT HAVE 50 QUESTIONS, EACH Q HAS 5 LEVELS. EVEN W/ LARGE SAMPLE, MANY COMBINATIONS HAVE ZEROS.

KEY DIFFERENCE IN ANALYZING GROUPED VS. UNGROUPED DATA (IN LOGISTIC REG.) IS THAT THE DEVIANCE MEANS DIFF. THINGS. IN PARTICULAR, IN UNGROUPED DATA, THE DEVIANCE BY ITSELF DOESN'T MEAN ANYTHING.

↳ THIS MEANS WE HAVE TO DO GOODNESS-OF-FIT IN A DIFFERENT WAY.

ONE WAY: GROUP DATA IN SOME MANNER & USE PEARSON $\chi^2$ OR DEVIANCE GOODNESS-OF-FIT.

ANOTHER WAY: GROUP DATA ACCORDING TO ESTIMATED PROBABILITIES. KNOWN AS "HOSMER-LEMESHOW"

PROCEDURE: 1) FIT SENSIBLE MODEL USING CONTINUOUS COVARIATES

EXAMPLE: $K=10$

2) FORM $K$ GROUPS OF (APPROX.) EQUAL SIZE. IF THERE $n$ OBSERVATIONS, $n/K$ IN GROUP 1 THAT CORRESPOND TO HIGHEST FITTED PROBABILITY. THEN GROUP 2 WOULD HAVE NEXT $n/K$ HIGHEST, AND SO ON.

FORM
$$H = \sum_{k=1}^{K} \frac{(O_k - E_k)^2}{n_k \bar{\pi}_k (1-\bar{\pi}_k)}$$

WHERE: $O_K$ = # OF OBSERVED COUNTS (TOTAL # OF OBS IN $K^{TH}$ GROUP)

$E_K$ = # OF EXPECTED COUNTS $\left(\frac{n}{K} + 0.5\right)$

$n_K$ = # OBS IN GROUP $K$

$\bar{\pi}_K$ = AVERAGE OF $\hat{\pi}_i$ IN GROUP $K$.

$H \sim \chi^2_{K-2}$ → LARGE $H$ (SMALL P-VAL) INDICATE MODEL MISFIT.

# LOGISTIC REGRESSION FOR MULTI LEVEL RESPONSES

MULTINOMIAL LOGISTIC REGRESSION.

EXAMPLE SUPPOSE EACH SUBJECT FALLS INTO ONE OF THREE LEVELS

$$\begin{cases} (1) & \text{HIV NEG.} \\ (2) & \text{HIV POS. / NO AIDS} \\ (3) & \text{HIV POS. / AIDS} \end{cases}$$

LET $\pi_1, \pi_2, \pi_3$ BE PROBABILITIES OF BEING ON EACH LEVEL. WE WOULD LIKE TO KNOW THE EFFECT OF THESE EXPLANATORY VARIABLES ON THESE PROBABILITIES.

↳ NOTE: $\pi_1 + \pi_2 + \pi_3 = 1$, SO ANY TWO DETERMINE THIRD ONE.

FORM: $$\theta_2 = \log\left(\frac{\pi_2}{\pi_1}\right), \quad \theta_3 = \log\left(\pi_3/\pi_1\right)$$

WHERE: $\theta_2$ IS LOG ODDS OF BEING HIV POS/NO AIDS RELATIVE TO HIV NEG.

IDEA: FIT SEPARATE $\overset{\text{log}}{\vee}$ ODDS-RATIOS W/ DIFFERENT ~~LOGISTIC~~ LINEAR REGRESSIONS.

EXAMPLE: SUPPOSE WE HAVE EXPLANATORY VAR. # OF SEXUAL PARTNERS IN LAST 6 MONTHS ($X$)

$$\log\left(\frac{\pi_2(x)}{\pi_1(x)}\right) = \beta_0 + \beta_1 X$$

$$\log\left(\frac{\pi_3(x)}{\pi_1(x)}\right) = \alpha_0 + \alpha_1 X$$

↳ FROM HERE, EVERYTHING IS THE SAME AS BEFORE.

CAUTION: BE SURE YOU KNOW WHAT THE REFERENCE LEVEL IS.

## ORDERED LEVELS:

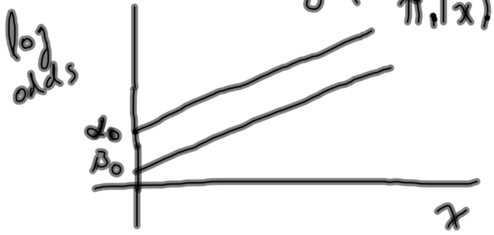
"NO" DISEASE, "SOME" DISEASE, "SEVERE" DISEASE

THERE ARE A FEW WAYS OF GOING ABOUT THIS.

1) "COMMON SLOPES"

$$\log\left(\frac{\pi_2(x)}{\pi_1(x)}\right) = \underset{\updownarrow \text{ DIFF}}{\beta_0} + \underset{\updownarrow \text{ SAME}}{\beta_1 x} \quad \left(\begin{array}{c}\text{WE EXPECT}\\ \text{TO SEE}\\ \hat{\beta}_1 > 0\end{array}\right)$$

$$\log\left(\frac{\pi_3(x)}{\pi_1(x)}\right) = \alpha_0 + \beta_1 x$$



"log ODOS OF HIV POS/AIDS NO RELATIVE TO HIV NEG IS A CONSTANT SHIFT OF HIV POS/AIDS YES RELATIVE TO HIV NEG."

"PROPORTIONAL ODDS MODEL"

HERE, WE LOOK AT CUMMULATIVE PROBABILITY AS THE RESPONSE. DEFINE $\gamma_1 = \pi_1$, $\gamma_2 = \pi_1 + \pi_2$, $\gamma_3 = \pi_1 + \pi_2 + \pi_3 = 1$. DO INFERENCE ON $\gamma_1, \gamma_2$.

NOW,
$$\log\left(\frac{\gamma_1(x)}{1-\gamma_1(x)}\right) = \underset{\updownarrow \text{ DIFF}}{\beta_0} + \underset{\updownarrow \text{ SAME}}{\beta_1 X} \quad \left(\begin{array}{c}\text{EXPECT}\\ \hat{\beta}_1 < 0\end{array}\right)$$

$$\log\left(\frac{\gamma_2(x)}{1-\gamma_2(x)}\right) = \alpha_0 + \beta_1 X$$

WE CAN USE OTHER LINK FUNCTIONS THAT LOOK/ACT A LOT LIKE THE LOGIT.

ANY $F(\pi)$ THAT IS

1) DIFFERENTIABLE
2) GOES FROM $-\infty, \infty$ AS $\pi$ GOES $0 \to 1$
3) ALWAYS INCREASING

CAN BE USED AS A LINK.

EXAMPLES

- LOGISTIC REGRESSION W/ CANONICAL LINK.
- USE THE INVERSE CDF OF A STANDARD NORMAL AS THE LINK. GIVES "PROBIT" MODEL.