TODAY: LOGISTIC REGRESSION

IDEA: MULTIPLE REGRESSION WITH DUMMY (0/1) DEPENDENT VARIABLE

- MANY APPLICATIONS
- REGULAR L-S REGRESSION DOESN'T MAKE SENSE...

EX: COMMUNITY COLLEGE DATA (RAW)

CONCERNED WITH STUDENT SUCCESS

GPA [GPA < 2.0]

EX: FORECAST GPA = 2.3

$$P(GPA < 2.0) = P\left(\frac{GPA - E[GPA]}{\sigma_{GPA}} < \frac{2.0 - 2.3}{.775}\right)$$

STD ERROR FROM REGRESSION OUTPUT

$$= P\left(Z < \frac{-.3}{.775}\right)$$

$$= P(Z < -.387)$$

$$\approx 35\%$$

[From z-table]

NOTE: σ_{GPA} ACTUALLY VARIES WITH VALUES OF INDEP. VARIABLES, BUT FOR TYPICAL VALUES IS DOMINATED BY STD. ERROR s_e

LOGISTIC REGRESSION

IDEA: LET $p = P(Y=1)$
(FOR SET OF VALUES OF INPUT VARIABLES)

- INSTEAD OF PREDICTING Y (0/1), PREDICT $\underline{p}$
- TURNS OUT TO BE BETTER TO WORK WITH

ODDS RATIO

$$OR = \frac{p}{1-p}$$

EX: $p = .5$, $OR = \frac{.5}{.5} = 1$ "EVEN"

$p = .9$ $OR = \frac{.9}{.1} = 9$ "9 TO 1 IN FAVOR"

$p = .2$ $OR = \frac{.2}{.8} = \frac{1}{4}$ "4 TO 1 AGAINST"

$$\frac{1}{OR} = \frac{1-p}{p} = \frac{1}{p} - 1$$

$$\frac{1}{OR} + 1 = \frac{1}{p} \qquad \frac{OR+1}{OR} = \frac{1}{p} \qquad \boxed{p = \frac{OR}{1+OR}}$$

- TURNS OUT EVEN BETTER TO WORK WITH LN OF OR

$$\underline{LOR} = Ln(OR) = Ln\left(\frac{p}{1-p}\right)$$

$$e^{LOR} = e^{Ln(OR)} = e^{Ln\left(\frac{p}{1-p}\right)} = \frac{p}{1-p}$$

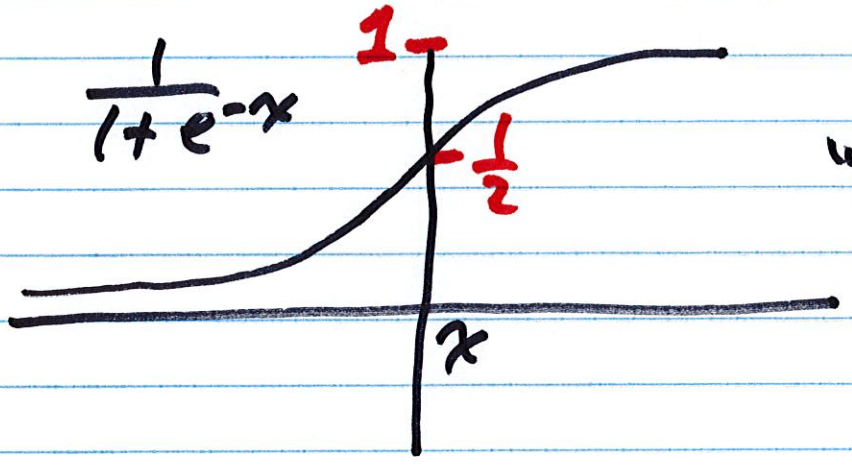
$$e^{LOR} = \frac{p}{1-p}$$

$$e^{-LOR} = \frac{1-p}{p} = \frac{1}{p} - 1$$

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

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

$f(x) = \frac{1}{1+e^{-x}}$ IS CALLED THE "LOGISTIC FUNCTION"



"S CURVE"

"MODEL" IS THAT

$$P = P(Y=1) = \frac{1}{1+e^{-LOR}}$$

$$LOR = \beta_0 + \beta_1 X_1 + \dots + \beta_K X_K$$

GIVEN DATA, COMPUTE

$$\widehat{LOR}_i = b_0 + b_1 X_{i1} + \dots + b_K X_{iK}$$

"LOR-HAT" WHERE X_{ij} IS VALUE OF j^{TH} INDEP VAR FOR OBSERVATION i.

$$\hat{P}_i = \frac{1}{1+e^{-\widehat{LOR}_i}} = \frac{\widehat{OR}_i}{1+\widehat{OR}_i}$$

NOTE: e^{b_i} GIVES FACTOR CHANGE IN OR FOR UNIT INCREASE IN X_i

EX: $e^{.05} = 1.0513$ $e^{.40} = 1.492$ [$\approx +50\%$]
$e^{.70} = 2.013$ [$\approx 100\%$]

MIN SSE DOES NOT MAKE AT A CRITERION FOR CHOOSING EQUATION $[b_0, b_1, \ldots b_k]$

INSTEAD, CHOOSE $b_0, b_1 \ldots b_k$ TO MAXIMIZE <u>LIKELIHOOD</u>

$$\left[\prod_{i|Y_i=1} \hat{p}_i\right] \cdot \left[\prod_{i|Y_i=0} (1-\hat{p}_i)\right]$$

PRODUCT SUCH THAT

- [redacted] THIS CAN BE DONE! BY ANY STAT PACKAGE (AND EXCEL)
- IN ADDITION TO POINT ESTIMATES $b_0, b_1 .. b_k$ CAN GET <u>CONF INTERVALS</u> $\Rightarrow$ EVALUATE <u>SIGNIFICANCE</u> OF VARIABLES

WHAT ABOUT RESIDUALS?

PROBLEM: GET NOTHING LOOKING AT INDIVIDUAL OBSERVATIONS

SOLUTION: AGGREGATE DATA POINTS INTO GROUPS. COMPARE OBSERVED # WITH $Y_i = 1$ TO EXPECTED #.

→ LEADS NATURALLY TO CHI-SQUARE TESTS ON RESIDUALS. SEVERAL VERSIONS USED IN STD. SOFTWARE.

WOULD LIKE TO VISUALIZE RESIDUALS FOR A CONTINUOUS X_j (AGE, AVE_ACT)

- GROUP VALUES OF X_j INTO RANGES
- FOR EACH RANGE, COMPARE OBS. # OF $Y_i = 1$ TO EXPECTED #

USING REGRESSION, CONSIDER EACH Y_i TO BE OUTCOME OF A BERNOULLI TRIAL WITH $P(Y_i = 1) = \hat{p}_i$

$\Rightarrow$ VARIANCE $= \hat{p}_i \cdot (1 - \hat{p}_i)$

SO FOR A GROUP, COMPARE $\sum Y_i$ WITH $\sum \hat{p}_i$

STANDARDIZE USING VARIANCE...

$\Rightarrow$ "PEARSON RESIDUAL"

$$\frac{\sum Y_i - \sum \hat{p}_i}{\sqrt{\sum \hat{p}_i \cdot (1 - \hat{p}_i)}}$$

[SUMS ARE FOR OBSERVATIONS IN THE GROUP]

EX: LOOK AT PEARSON RESIDUALS AGAINST AGE AND AVE_ACT IN CC DATA.