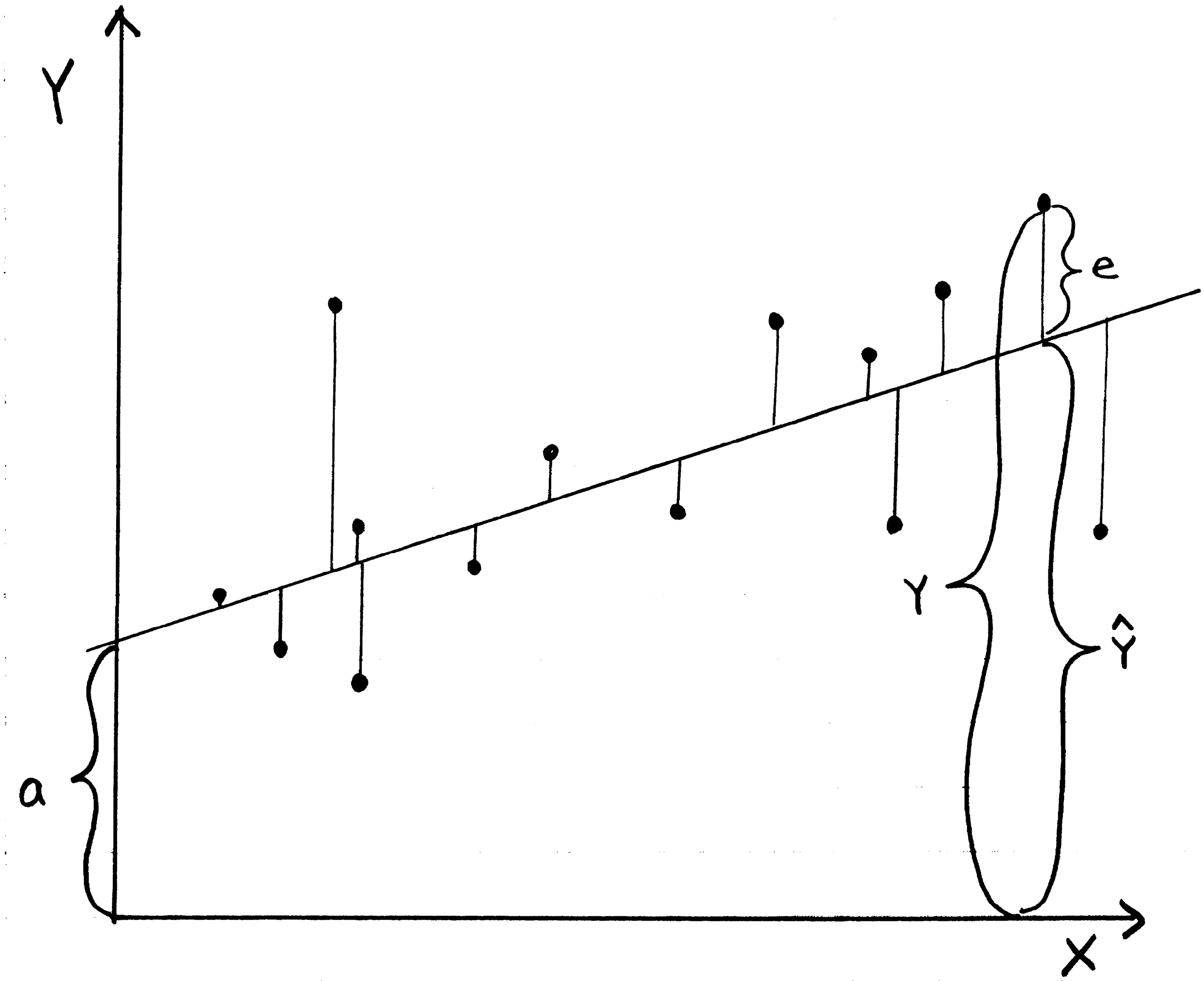
Neuendorf

Linear Regression

- The linear model predicts with a straight line one (or more) DVs from one (or more) IVs. To the extent that this line "misses" we have error, or unexplained variance.
- Bivariate regression attempts to predict Y from X. (e.g., If I know how old you are, can I predict how dogmatic you are?)
- The regression line is the line of best (vertical) fit... it minimizes:

$$[SS]\ \text{Error} = \sum (Y_i - \hat{Y}_i)^2$$

(the sum of the squared vertical distances, the "e"s)



$$\hat{Y} = a + bX$$

Regression Coefficients
- a = intercept on Y axis
- b = slope $\left(\frac{\Delta Y}{\Delta X}\right)$ in original units

$\hat{Y}$ = predicted Y
Y = actual (observed) Y
X = actual (observed) X
e = error of prediction

- The position of any given case can be described by:

$$Y_i = a + bX_i + e_i$$

so

$$Y_i = a + bX_i + (Y_i - \hat{Y}_i)$$
$$Y_i - (Y_i - \hat{Y}_i) = a + bX_i$$
$$\hat{Y}_i = a + bX_i$$

(shown on p. 2)

- Various components of a regression analysis may be subject to a statistical test of significance. Remember that:

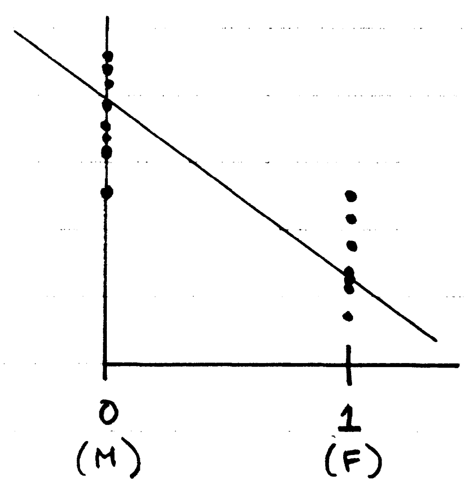
$$F = \frac{\text{variance explained}}{\text{variance unexplained}}$$

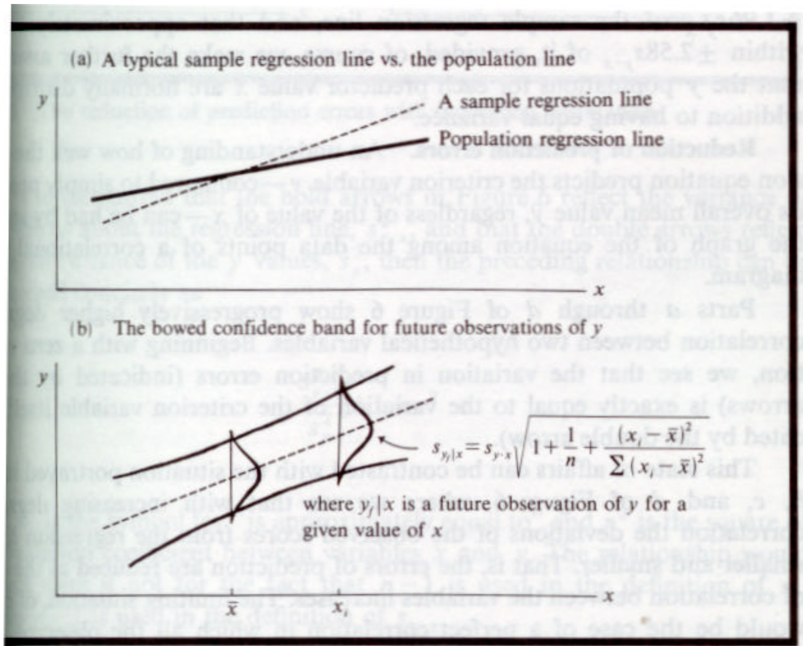
So, the total equation, or the variance contributed by a given predictor (X), or the variance contributed by a set or "block" of predictors (X_1, X_2, X_3, etc.) may be tested.

$$F = \frac{SS_{REG} / df_{REG}}{SS_{RESID} / df_{RESID}}$$

- SEE, the standard error of the estimate, allows us to gauge how close our sample regression line is to a true population regression line: it's analogous to SE of the mean. See attached page from the Kachigan book for a picture of what a regression line "confidence interval" looks like (or "confidence band").

- Using "dummy" variables . . .





5 Errors of prediction resulting from a sample regression equation (See text).