Indicator Variables (continued yet again)

prostate.lm3

Model: $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 . X_2 + \varepsilon$;

where $Y =$ lcavol $\leftarrow$ continuous responce

$X_1 =$ lpsa $\leftarrow$ continuous covariate

$$X_2 = \text{svi} = \begin{cases} 1 & \text{if svi} \\ 0 & \text{otherwise} \end{cases}$$

If svi $= 0$, $X_2 = 0$

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X_1 + 0 + 0$$

$$= \hat{\beta}_0 + \hat{\beta}_1 X_1 \quad \text{"base" model}$$

If svi $= 1$, $X_2 = 1$

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \hat{\beta}_2 \cdot 1 + \hat{\beta}_3 X_1 \cdot 1$$

$$= \underbrace{(\hat{\beta}_0 + \hat{\beta}_2)}_{\text{new intercept}} + \underbrace{(\hat{\beta}_1 + \hat{\beta}_3)}_{\text{new slope}} X_1$$

from STAT3015/4030/7030 GLMs
(ie not examinable in STAT2008/4038/6038 Regression Modelling)

17/5/2016
(2)

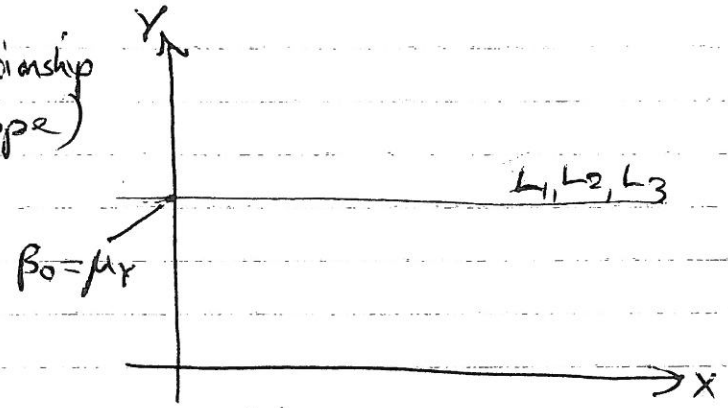
ANALYSIS OF COVARIANCE (ANCOVA) MODELS

(eg one 3 level factor variable (L1, L2, L3) & 1 covariate, X)

Intercepts

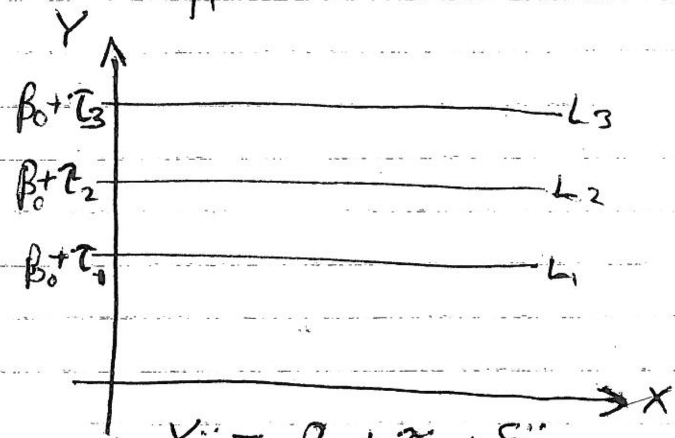
Same | Different

No relationship (0 slope)



$$Y_{ij} = \beta_0 + \varepsilon_{ij}$$

"Null" model

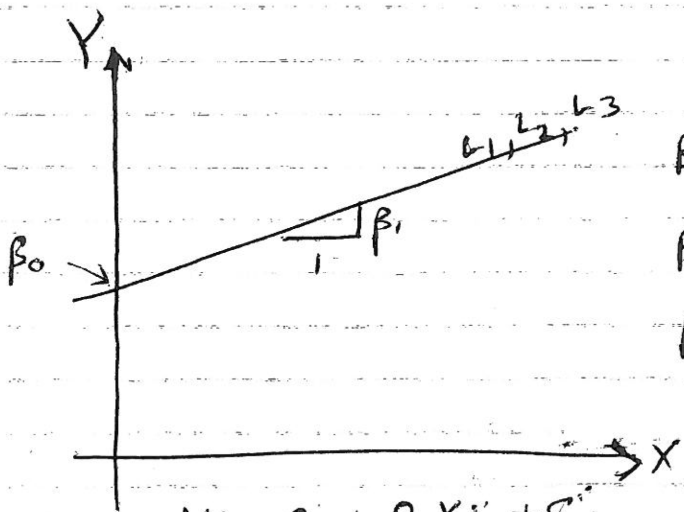


$$Y_{ij} = \beta_0 + \tau_j + \varepsilon_{ij}$$

"Different means" model

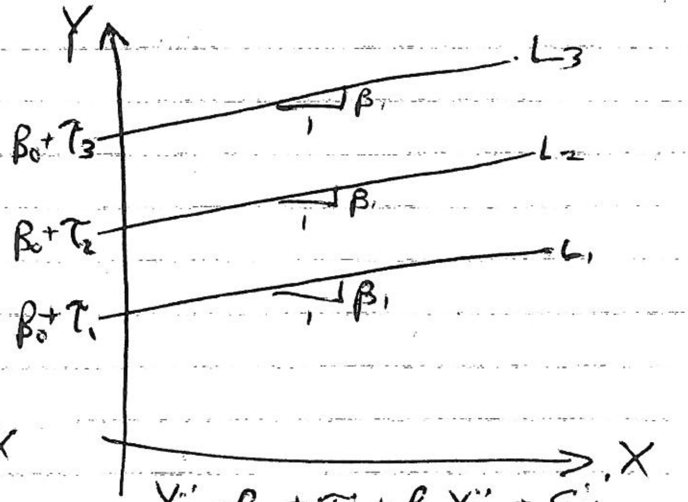
Slopes

Same Slopes



$$Y_{ij} = \beta_0 + \beta_1 X_{ij} + \varepsilon_{ij}$$

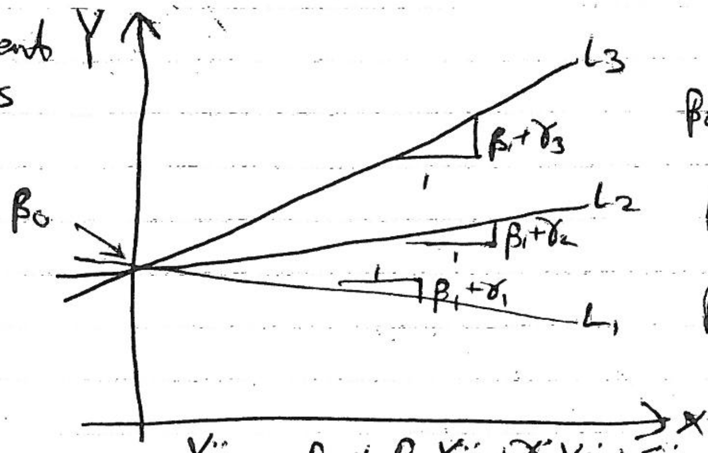
"Single (regression) line" model



$$Y_{ij} = \beta_0 + \tau_j + \beta_1 X_{ij} + \varepsilon_{ij}$$

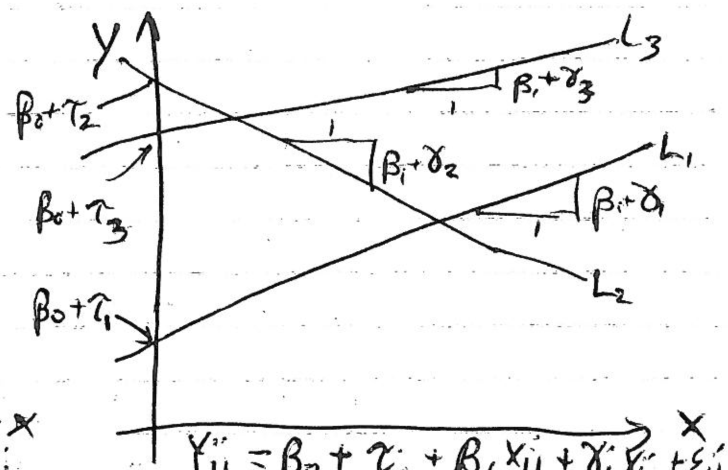
"Parallel lines" model

Different Slopes



$$Y_{ij} = \beta_0 + \beta_1 X_{ij} + \gamma_j X_{ij} + \varepsilon_{ij}$$

"Common intercepts" model



$$Y_{ij} = \beta_0 + \tau_j + \beta_1 X_{ij} + \gamma_j X_{ij} + \varepsilon_{ij}$$

"Seperate lines" model

Model Selection

A "good" model is one which we can use to address the research question, which may:

- involve certain variables, which we must include in the model, so we can observe and/or "control for" the effects of these variables
- other variables (included in the data) may also be included in the model, if they help to explain some of variation (ie they turn out to be "significant")
- ultimately the research question may require some predictions; preferably predictions that hold general validity

Note if have already chosen some scale for the variables in the model and a particular form for the model, we can then experiment with models that include other X variables in the data as predictors, as well as derived variables (X^2, $\log X$, interaction terms including X, ...)

If we have k possible predictors, then the number of candidate models is $O(2^k)$ as a minimum (as we can also allow for different orders of the predictors)

ie $k=1$, 2 possible models ; $k=10 \Rightarrow 1{,}024$ models
$k=20$, $2^{20} = 1{,}048{,}576$

For observational covariates (optional X's) we use:

Principle of Parsimony (Occam's Razor)

Of two similar models, we will tend to prefer the simpler one (esp. if there is no significant difference between them)