Mean structure

- idea: data are generated from an underlying system, which is assumed to have the form: $y = f(x_1, \ldots, x_m) + \varepsilon$, where f is ***unknown***.
- regression *approximates* the mean structure f by a linear combination of (known) *base functions* $g_i(x_1, \ldots, x_m)$'s, $i=1, \ldots, p$, i.e.,

$$f \longleftarrow \sum_{i=1}^{p} \beta_i \cdot g_i(x_1, \ldots, x_m)$$

  - ➢ when the structure of f is simple and almost linear, it can be approximated by a simple structure with fewer terms, e.g.,

$$E(y) = f \approxeq \beta_0 + \beta_1 x_1 + \cdots + \beta_m x_m$$

    - **Q**: nature is simple?
    - **Q**: are there sufficient data to support/fit a complex model?
  - ➢ when f is complex and non-linear ⇒ need more terms to get a good approximation
    - more parameters, need more degrees of freedom, i.e., more data
    - e.g., 2 levels, only linear effects; 3 levels, linear and quadratic effects
    - **Q**: what other complex models?
- base functions for quantitative and qualitative predictors x_i's are defined in different ways

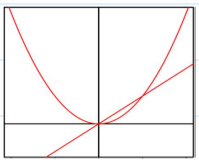
location and/or scale change

- $x_i \to (x_i+a)/b$ or $y \to (y+a)/b$, where a and b are given constants.
  a: change of location, b: change of scale
- **Q**: why we might want to do this?
  - ➢ predictors of similar magnitude are easier to compare
  - ➢ rescaling may make $\hat{\boldsymbol{\beta}}$ easier to read and may aid interpretability
  - ➢ numerical stability is enhanced when all predictors are on a similar scale
  - ➢ for experimental data, it's often that we code two levels (say, 20°C, 30°C) → (−1, 1); three levels (say, 20°C, 30°C, 40°C) → (−1, 0, 1)
- influence caused by location/scale change on x_i (i.e., $x_i \to (x_i+a)/b$)
  - ➢ (under a model with intercept) overall F-test, t-test, R^2, $\hat{\sigma}$ all unchanged
  - ➢ $\hat{\boldsymbol{\beta}}$ change: $\hat{\beta}_i \to b\hat{\beta}_i$, $\hat{\beta}_0 \to \hat{\beta}_0 - a\hat{\beta}_i$
- influence caused by location/scale change on y (i.e., $y \to (y+a)/b$)
  - ➢ (under a model with intercept) overall F-test, t-test, R^2 unchanged
  - ➢ $\hat{\boldsymbol{\beta}}$ and $\hat{\sigma}$ change: $\hat{\sigma} \to \hat{\sigma}/b$, $\hat{\beta}_i \to \hat{\beta}_i/b$ for each i, $\hat{\beta}_0 \to (\hat{\beta}_0 + a)/b$

❖ **Reading**: Faraway (2005, 1st ed.), 5.2 ❖ **Further reading**: D&S, 16.2, 16.3

Polynomial regression

- **Q**: when to use polynomial regression? ⇒ the relationship between response and *quantitative* predictors is **smooth**, but not a straight line.
- **idea supports the approach** ⇒ any smooth function (mean structure of the underlying system) can be approximated by a polynomial of high enough degree
- one predictor case: $y = \beta_0 + \beta_1 x + \beta_2 x^2 + \ldots + \beta_d x^d + \varepsilon$
  - ➢ choice of d
    - start with $y=\beta_0+\beta_1 x$, keep adding polynomial terms until last term added is not significant. ⇒ danger: stop too soon
    - start with a large d and recursively remove insignificant largest term
    - use added variable plot/partial residual plot to gain information about d
  - ➢ **Q**: Consider the model $y=\beta_0+\beta_1 x+\beta_2 x^2$. what if β_1 not significant, but β_2 is significant? should x be removed from the model?



    - x and x^2 could be highly correlated
    - location shift: $x \to x+c$ ⇒ $\hat{\beta}_2$ unchanged, but $\hat{\beta}_1$ may become significant
    - recommendation: do not remove insignificant lower-order terms from model when the highest-order term is significant

- two predictors x_1, x_2 case:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_{11} x_1^2 + \beta_{22} x_2^2 + \beta_{12} x_1 x_2 + \varepsilon \quad (d=2,\ 2^{\text{nd}}\text{-order model})$$

  - ➢ the cross-product term x_1x_2 can be interpreted as an "interaction" effect, e.g.,

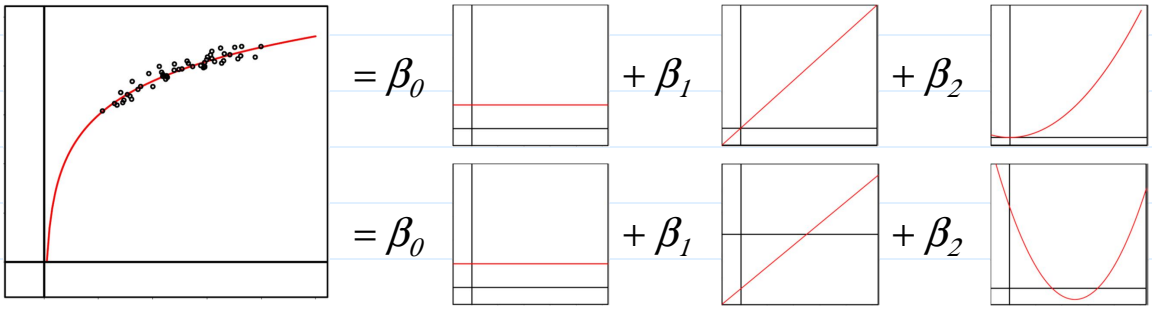
$$E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2, \text{ where } x_1, x_2 \in \{-1, 1\}$$
$$x_1 = +1 \Longrightarrow E(y) = (\beta_0 + \beta_1) + (\beta_2 + \beta_3) x_2$$
$$x_1 = -1 \Longrightarrow E(y) = (\beta_0 - \beta_1) + (\beta_2 - \beta_3) x_2$$

  - ➢ models for more predictors can be similarly extended

$$y = \beta_0 + \sum_{i=1}^{m} \beta_{1,i} x_i + \sum_{i=1}^{m} \beta_{2,i} x_i^2 + \sum_{1 \le i < j \le m} \beta_{3,ij} x_i x_j + \epsilon$$

# of x_i's	d=2	d=3
2	6	10
3	10	20
4	15	35

  - ➢ increasing degree d ⇒ model may have too many parameters
- orthogonal polynomials
  - ➢ polynomial terms can cause numerical instability (especially when d large) and collinearity
  - ➢ example: 2^{nd}-order model



- define $z_0=1$, $z_1=a_1+b_1x_1$, $z_{11}=a_2+b_2x_1+c_2x_1^2$, $z_{111}=a_3+b_3x_1+c_3x_1^2+d_3x_1^3$, ...
  Find a_i, b_i, c_i, ..., that make $z_j^T z_k=0$ if $j\neq k$ (and $\|z_i\|=1$ sometimes)
  - can apply regression to obtain z_0, z_1, z_{11},... (note: $\hat{y}^T\hat{\varepsilon}=0$), e.g., regress x_1 on z_0, then the residuals is proportional to z_1; regress x_1^2 on z_0, z_1 and the residuals is proportional to z_{11}. In R, built-in function is provided to construct orthogonal polynomials.
  - cross-product terms (i.e., interactions) can be defined in a similar manner (e.g., $z_{12}= a+b x_1+c x_2+d x_1x_2$, regress x_1x_2 on z_0, z_1, z_2, and the residuals are proportional to z_{12})
- change model based on polynomial terms to model based on z's, e.g.,

$$y=\beta_0+\beta_1 x_1+\beta_2 x_2+\beta_{11}x_1^2+\beta_{22}x_2^2+\beta_{12}x_1x_2+\epsilon$$
$$\longrightarrow y=\beta'_0+\beta'_1 z_1+\beta'_2 z_2+\beta'_{11}z_{11}+\beta'_{22}z_{22}+\beta'_{12}z_{12}+\epsilon$$

  the two models have same column space Ω (i.e., same R^2, $\hat{\sigma}$, overall F), but interpretation of β's and β''s are different (i.e., different estimates, t-tests)
- orthogonality can save works when selecting model (do not have to refit after deleting term), it's more convenient for fitting and testing

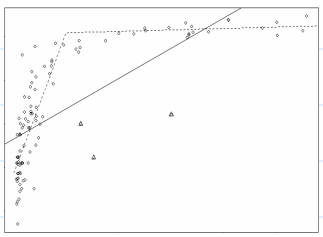
• properties of polynomial model
- offer more flexible relationship
- remember that it's an approximation, we usually do not believe it exactly represents the underlying reality

- polynomials have the advantage of smoothness
- but, have the disadvantage that each data point affects the fit **globally**
- For larger values of d, the fitted polynomial curves may become wiggly. reason: the curve may capture the random variation, rather than the overall shape of the relationship between predictors and response.
- polynomial model is hard to fit "jump function"

❖ **Reading**: Faraway(2005, 1st ed.), 7.2.2 ❖ **Further reading**: D&S, 12.1, 12.3, 22.2

broken stick (line) regression (segmented regression)

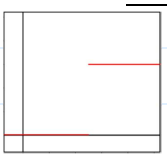
• Recall. polynomial regression: suitable for smooth mean structure, but cannot capture local abrupt change (example?)
Q: how to relax the smoothness restriction?
⇒ one solution: broken line regression.



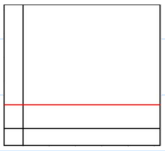
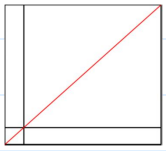
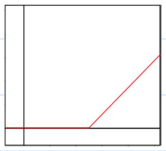
• **Q**: when to use broken line regression?
⇒ believe that different regression models apply in different regions of data, and the fit should be continuous at the broken points

• suppose the break occurs at the *known* value c, define the base function (where c is called a *knot*):

$$d_c(x)=\begin{cases}1, & \text{if } x>c,\\ 0, & \text{if } x\leq c.\end{cases}$$

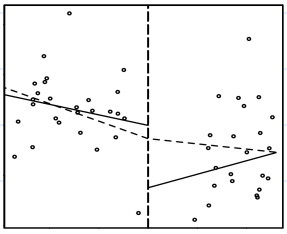


- model: $y = \beta_0 + \beta_1 x + \beta_2 (x-c) d_c(x) + \varepsilon$

$= \beta_0$  $+ \beta_1$  $+ \beta_2$  $+ \varepsilon$

$$E(y) = \begin{cases} \beta_0 + \beta_1 x, & \text{if } x \le c, \\ (\beta_0 - \beta_2 c) + (\beta_1 + \beta_2) x, & \text{if } x > c, \end{cases}$$

  - ➢ the two lines meet at c ⇒ continuous fit
  - ➢ notice only 3 parameters in the model ⇒ one degree of freedom is saved because of the continuity restriction



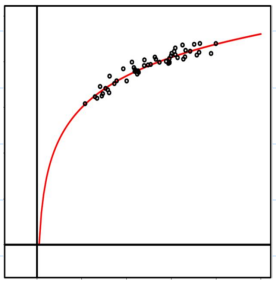
- when c is unknown
  ⇒ can regard c as a parameter ⇒ not a linear model any more
  ⇒ can be estimated by nonlinear regression
- generalization: more knot points or more predictors ⇒ define more base functions

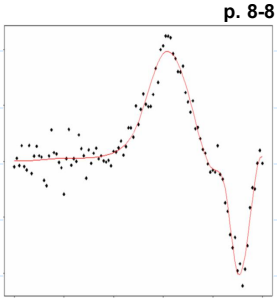
- broken curve regression: $y = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 (x-c) d_c(x) + \beta_4 (x-c)^2 d_c(x) + \varepsilon$

❖ **Reading**: Faraway (2005, 1st ed.), 7.2.1 ❖ **Further reading**: D&S, 14.3

regression spline and LOWESS

- **concept**: fitting using local points ⇔ better fit
  using global points ⇔ smoothness
  Q: which should you choose for your data?
  using local points or global points?



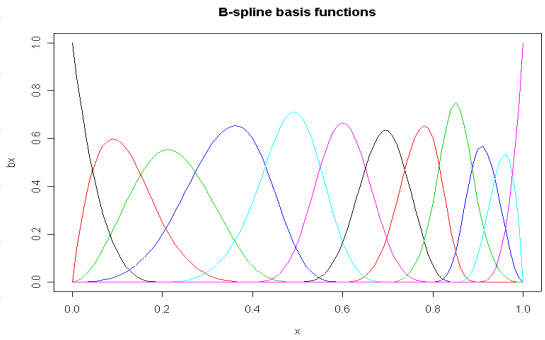


- regression spline
  - ➢ **concept**: define different *base functions* to fit data
    - power function: smoothness, but non-zero across the whole range
    - broken line: lesser smooth, but localizing the influence of data point
    - B-spline: compromise between smoothness and local influence
  - ➢ cubic B-spline base functions: $g_1, \ldots, g_{k-4}$ defined on an interval $[a, b]$ with knot-points $t_1 \le \ldots \le t_k$ (no need to be equally spaced) satisfying:
    1. non-zero on interval defined by 5 successive knots and zero elsewhere ⇒ local influence
    2. a cubic polynomial for each sub-interval between successive knots
    3. continuous, and continuous in its 1st and 2nd derivatives at each knot ⇒ smoothness
    4. integrate to one over its support

- base function at the ends of the interval are defined differently to ensure continuity
- regress y on these B-spline base functions, i.e.,

$$y = \Sigma_i \beta_i g_i(x) + \varepsilon$$

(Note: g_i's are known functions for given t's $\Rightarrow$ it's still a linear model)



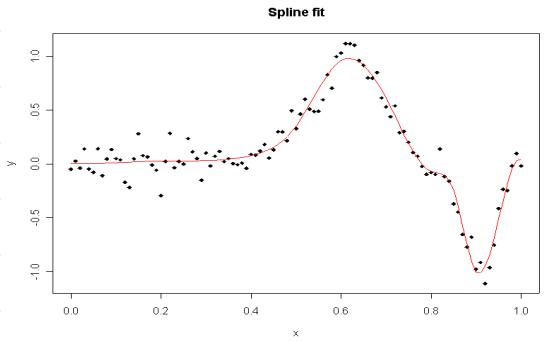
knot points: 0, 0, 0, 0, 0.2, 0.4, 0.5, 0.6, 0.7, 0.8, 0.85, 0.9, 1, 1, 1, 1

Q: why the knots are dense in some region?

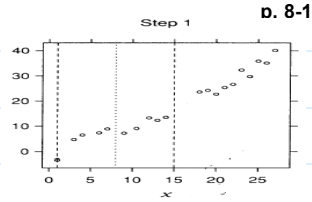
- LOWESS (LOcally WEighted Scatterplot Smoothing)
  - Recall: in previous models, # of parameters are finite
  - nonparametric regressions:

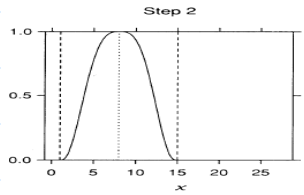
    model: $y = f(x_1, \ldots, x_m) + \varepsilon$

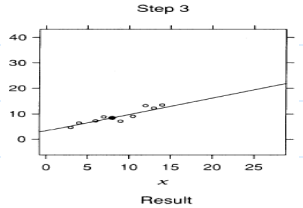
    - parametric regression: assume f is from a family of functions, in which # of parameters is finite
    - nonparametric regression: assume f is smooth only (# of parameters = ∞)

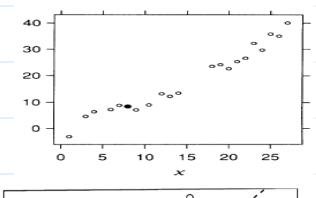


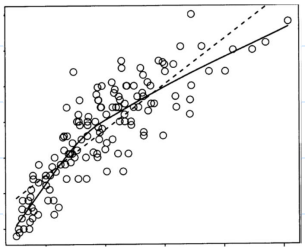
- method: (see example)
  1. select a window ($\Rightarrow$ local information), and a weighting function ($\Rightarrow$ closer points, more contribution)
  2. use weighted (closer points, higher weights) average of y_i's in the window to estimate fitted value
  3. repeat as the window moves
     - width of window is an issue (larger window, smoother curve)
     - can plot fitted value for a variety of widths and pick best result
     - different width of window to estimate f along the range of x
     - sensitive to outliers: use median, not average
- LOESS: change step 2 to locally weighted (1st or 2nd order) polynomial regression (see example)
- difficult if extrapolation is required (same difficulty in regression spline)
- nonparametric regressions are useful for fitting a curve for residual plots, added variable plots, partial residual plots











❖ **Reading**: Faraway (2005, 1st ed.), 7.2.3; Faraway (2006), chapter 11

dummy variable (indicator variable, coding)

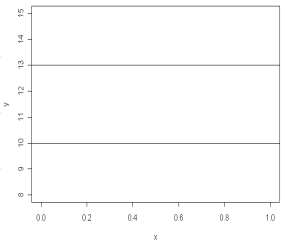
- categorical (qualitative) predictors
  - nominal v.s. ordinal
  - examples: male/female, treatment/control, eye colors, blocks, ...
  - qualitative in nature:
    - values are symbols, no quantitative meaning
    - no value exist between categories
  - Q: what properties can we explore for qualitative predictor?

    category $i \to y_{ij}$, $\mu_i = \mathrm{E}(y_{ij}) \Rightarrow$ can only study *difference* between μ_i's (cf., quantitative predictor)
  - Q: how to fit these predictors into the format of linear regression model $\boldsymbol{Y} = \boldsymbol{X\beta}+\boldsymbol{\varepsilon}$? $\Rightarrow$ Ans: dummy variables
- one dichotomous predictor: two categories
  - for a dichotomous predictor C with two categories c_1 and c_2, define a dummy variable d:
$$d(C) = \begin{cases} 0, & \text{if } C = c_1, \\ 1, & \text{if } C = c_2. \end{cases}$$
  - for a data set with response y, one quantitative predictor x, and one qualitative predictor C (dummy variable d), possible models are:

    model 1: $y = \beta_0 + \beta_1 d + \varepsilon$, model 2: $y = \beta_0 + \beta_1 x + \varepsilon$,
    model 3: $y = \beta_0 + \beta_1 d + \beta_2 x + \varepsilon$, model 4: $y = \beta_0 + \beta_1 x + \beta_2 xd + \varepsilon$,
    model 5: $y = \beta_0 + \beta_1 d + \beta_2 x + \beta_3 xd + \varepsilon$

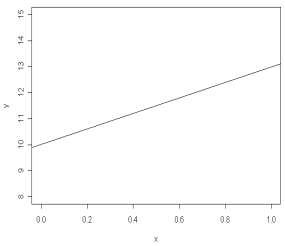
- Q: how to interpret β_i's in models 1~5?
  - model 1: $y = \beta_0 + \beta_1 d + \epsilon$

$$\begin{aligned} C = c_1: &\quad \mu_1 = E(y|d=0) = \beta_0 \\ C = c_2: &\quad \mu_2 = E(y|d=1) = \beta_0 + \beta_1 \end{aligned} \Rightarrow \begin{aligned} \beta_0 &= \mu_1 \\ \beta_1 &= \mu_2 - \mu_1 \end{aligned}$$

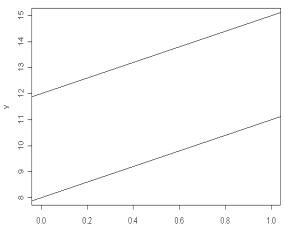


  - model 2: $y = \beta_0 + \beta_1 x + \epsilon$
  - model 3: $y = \beta_0 + \beta_1 d + \beta_2 x + \epsilon$

$$\begin{aligned} C = c_1: &\quad \mu_{1,x} = E(y|d=0,x) = \beta_0 + \beta_2 x \\ C = c_2: &\quad \mu_{2,x} = E(y|d=1,x) = (\beta_0 + \beta_1) + \beta_2 x \end{aligned}$$

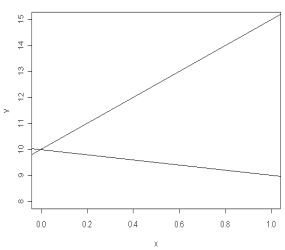


$$\begin{aligned} \beta_0 &= \mu_{1,0} \text{ (intercept in } c_1 \text{ group)} \\ \Rightarrow \beta_1 &= \mu_{2,x} - \mu_{1,x} \text{ (difference of intercepts)} \\ \beta_2 &= \text{slope (same slope in two categories)} \end{aligned}$$



  - model 4: $y = \beta_0 + \beta_1 x + \beta_2 (d \cdot x) + \epsilon$

$$\begin{aligned} C = c_1: &\quad \mu_{1,x} = E(y|d=0,x) = \beta_0 + \beta_1 x \\ C = c_2: &\quad \mu_{2,x} = E(y|d=1,x) = \beta_0 + (\beta_1 + \beta_2) x \end{aligned}$$



$$\begin{aligned} \beta_0 &= \mu_{1,0} = \mu_{2,0} \text{ (same intercept in two categories)} \\ \Rightarrow \beta_1 &= \text{slope of category } c_1 \\ \beta_2 &= \text{difference in slopes} \end{aligned}$$

- model 5: $y = \beta_0 + \beta_1 d + \beta_2 x + \beta_3 (d \cdot x) + \epsilon$

$C = c_1$: $\mu_{1,x} = E(y|d=0, x) = \beta_0 + \beta_2 x$

$C = c_2$: $\mu_{2,x} = E(y|d=1, x) = (\beta_0 + \beta_1) + (\beta_2 + \beta_3)x$

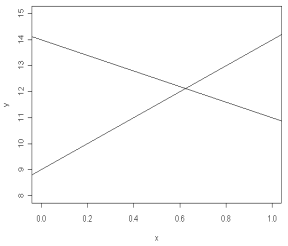
$\Rightarrow$

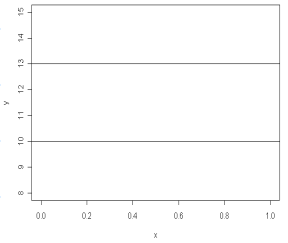
β_0 = $\mu_{1,0}$ (intercept of category c_1)

β_2 = slope of category c_1

β_1 = difference in intercepts

β_3 = difference in slopes





- alternative coding of dummy variable (better orthogonality)

$$d(C) = \begin{cases} -1, & \text{if } C = c_1, \\ 1, & \text{if } C = c_2. \end{cases}$$

Q: how to interpret β_i's in models 1~5 under this coding?

- model 1: $y = \beta_0 + \beta_1 d + \epsilon$

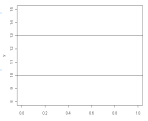
$C = c_1$: $\mu_1 = E(y|d=-1) = \beta_0 - \beta_1$

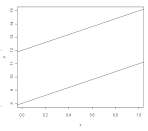
$C = c_2$: $\mu_2 = E(y|d=1) = \beta_0 + \beta_1$

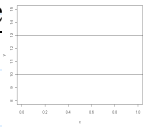
$\Rightarrow$ $\beta_0 = (\mu_1 + \mu_2)/2$, $\beta_1 = (\mu_2 - \mu_1)/2$

- analysis strategy: start from the full model (model 5) if there are enough degrees of freedom, and then test if some terms can be eliminated
- identical methodology applies for more than 2 categories and more quantitative predictors
- **Q**: what if data in the two categories have different variance?

- *ANalysis of COVAriance*: testing model 3 (Ω) against model 2 (ω) (more than 2 categories and more quantitative predictors is allowed). The quantitative predictor is called *covariate* and is expected to have the same effect in all categories. The difference between categories is assumed to be an additive effect.







• one polytomous predictor: more than two categories

- for k categories, $k-1$ dummy variables are needed to depict the difference between categories (one parameter is used to represent constant term)
- various coding of dummy variables: 4 categories c_1, c_2, c_3, c_4 example

treatment coding

	d_1	d_2	d_3
c_1	0	0	0
c_2	1	0	0
c_3	0	1	0
c_4	0	0	1

Helmert coding

	d_1	d_2	d_3
c_1	–1	–1	–1
c_2	1	–1	–1
c_3	0	2	–1
c_4	0	0	3

sum coding

	d_1	d_2	d_3
c_1	–1	–1	–1
c_2	1	0	0
c_3	0	1	0
c_4	0	0	1

- consider the model: $y = \beta_0 + \beta_1 d_1 + \beta_2 d_2 + \beta_3 d_3 + \epsilon$
- properties of treatment coding:

$C = c_1$: $\mu_1 = E(y|d_1=0, d_2=0, d_3=0) = \beta_0$

$C = c_2$: $\mu_2 = E(y|d_1=1, d_2=0, d_3=0) = \beta_0 + \beta_1$

$C = c_3$: $\mu_3 = E(y|d_1=0, d_2=1, d_3=0) = \beta_0 + \beta_2$

$C = c_4$: $\mu_4 = E(y|d_1=0, d_2=0, d_3=1) = \beta_0 + \beta_3$

$\Rightarrow$ $\beta_0 = \mu_1$, $\beta_1 = \mu_2 - \mu_1$, $\beta_2 = \mu_3 - \mu_1$, $\beta_3 = \mu_4 - \mu_1$

- treats c_1 as a reference
- it is convenient if a "standard" categories exists
- d_1, d_2, and d_3 are mutually orthogonal, but not orthogonal to constant term

- properties of Helmert coding: $y = \beta_0 + \beta_1 d_1 + \beta_2 d_2 + \beta_3 d_3 + \epsilon$

$$
\begin{aligned}
C = c_1: &\quad \mu_1 = E(y|d_1 = -1, d_2 = -1, d_3 = -1) = \beta_0 - \beta_1 - \beta_2 - \beta_3 \\
C = c_2: &\quad \mu_2 = E(y|d_1 = 1, d_2 = -1, d_3 = -1) = \beta_0 + \beta_1 - \beta_2 - \beta_3 \\
C = c_3: &\quad \mu_3 = E(y|d_1 = 0, d_2 = 2, d_3 = -1) = \beta_0 + 2\beta_2 - \beta_3 \\
C = c_4: &\quad \mu_4 = E(y|d_1 = 0, d_2 = 0, d_3 = 3) = \beta_0 + 3\beta_3
\end{aligned}
$$

$$
\Rightarrow \quad
\begin{aligned}
\beta_0 &= \frac{\mu_1 + \mu_2 + \mu_3 + \mu_4}{4} \equiv \bar{\mu} \\
\beta_1 &= \frac{\mu_2 - \mu_1}{2} \\
\beta_2 &= \frac{\mu_3 - ((\mu_1 + \mu_2)/2)}{3} \\
\beta_3 &= \frac{\mu_4 - ((\mu_1 + \mu_2 + \mu_3)/3)}{4}
\end{aligned}
$$

- constant term, d_1, d_2, and d_3 are orthogonal when there are equal # of observations in each categories
- hard to interpret parameters
- may suitable for ordinal qualitative predictor

- properties of sum coding: $y = \beta_0 + \beta_1 d_1 + \beta_2 d_2 + \beta_3 d_3 + \epsilon$

$$
\begin{aligned}
C = c_1: &\quad \mu_1 = E(y|d_1 = -1, d_2 = -1, d_3 = -1) = \beta_0 - \beta_1 - \beta_2 - \beta_3 \\
C = c_2: &\quad \mu_2 = E(y|d_1 = 1, d_2 = 0, d_3 = 0) = \beta_0 + \beta_1 \\
C = c_3: &\quad \mu_3 = E(y|d_1 = 0, d_2 = 1, d_3 = 0) = \beta_0 + \beta_2 \\
C = c_4: &\quad \mu_4 = E(y|d_1 = 0, d_2 = 0, d_3 = 1) = \beta_0 + \beta_3
\end{aligned}
$$

$$
\Rightarrow \quad
\begin{aligned}
\beta_0 &= \frac{\mu_1 + \mu_2 + \mu_3 + \mu_4}{4} \equiv \bar{\mu} \\
\beta_1 &= \mu_2 - \bar{\mu} \\
\beta_2 &= \mu_3 - \bar{\mu} \\
\beta_3 &= \mu_4 - \bar{\mu}
\end{aligned}
$$

- β_0 represent overall mean
- compare each category with the overall mean
- lesser orthogonal

➢ Note: the choice of coding does not affect the R^2, $\hat{\sigma}$ and overall F-test (to test H_0: $\beta_1=\beta_2=\beta_3= 0$, the three codings have same ω and Ω)

➢ the overall F-test is one-way ANOVA (ANalysis Of VAriance)

➢ **Q**: how to work with quantitative predictors? ⇒ identical methodology as in 2 categories case. **Q**: how to interpret parameters in the case?

• two qualitative predictors
(say, A: I=3 categories a_1, a_2, a_3; B: J=4 categories, b_1, b_2, b_3, b_4)

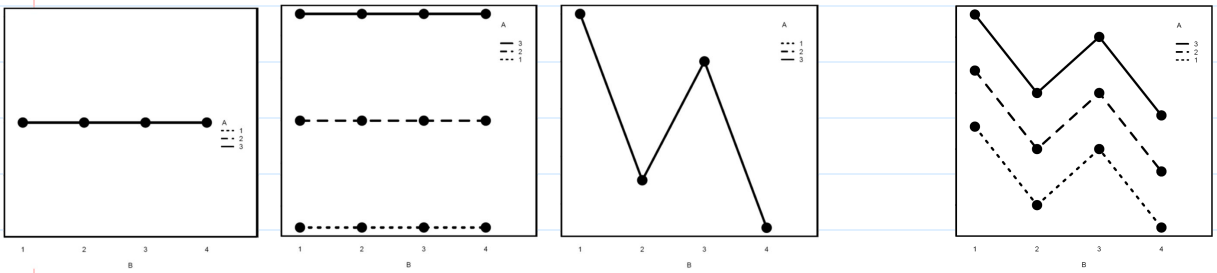
➢ number of different category combinations = 3×4 =12,
denote their means as μ_{ij}, i=1, 2, 3 and j=1, 2, 3, 4, i.e.,

$$y_{ijk} = \mu_{ij} + \epsilon_{ijk}, \quad k = 1, 2, \ldots, n_{ij},$$

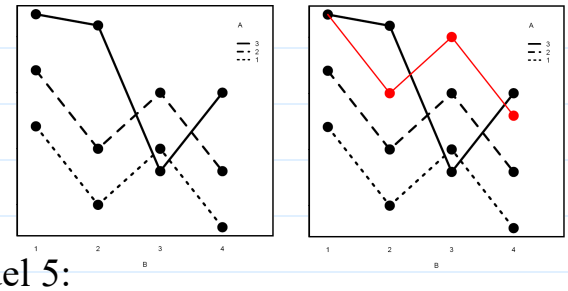
n_{ij} = number of observations in category $A=a_i$ and $B=b_j$

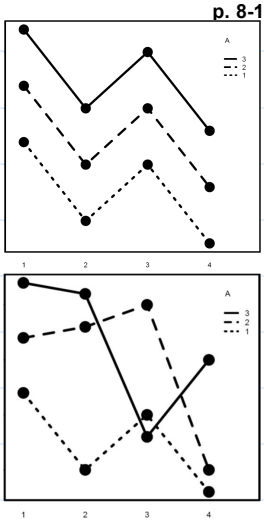
➢ **Q**: how to depict the difference between μ_{ij}'s?
consider the following linear models:

- model 1: $E(y_{ijk}) = \beta_0$
- model 2: $E(y_{ijk}) = \beta_0 + \beta_1 d_1^A + \beta_2 d_2^A$
- model 3: $E(y_{ijk}) = \beta_0 + \beta_1 d_1^B + \beta_2 d_2^B + \beta_3 d_3^B$
- model 4: $E(y_{ijk}) = \beta_0 + \beta_1 d_1^A + \beta_2 d_2^A + \beta_3 d_1^B + \beta_4 d_2^B + \beta_5 d_3^B$



Q: how to perform interaction coding? what is interaction?





- model 5:

$$E(y_{ijk}) = \beta_0 + \beta_1 d_1^A + \beta_2 d_2^A + \beta_3 d_1^B + \beta_4 d_2^B + \beta_5 d_3^B + \sum_{i=1}^{2}\sum_{j=1}^{3} \beta_{ij} \boxed{d_{ij}} \longrightarrow \boxed{\text{2-factor interaction}}$$

# of parameters: 1 + 2 + 3 + 6 = 12

interaction plot: replace μ_{ij}'s by cell means
$\bar{y}_{ij} = \sum_{k=1}^{n_{ij}} y_{ijk}/n_{ij}$

➢ Two-way (sequential) ANOVA

- anova(y~1+A+B+A:B)

1) test ω_1:model 1 (y~1) against Ω_1:model 2 (y~1+A) [$df_\omega - df_\Omega$ =2]

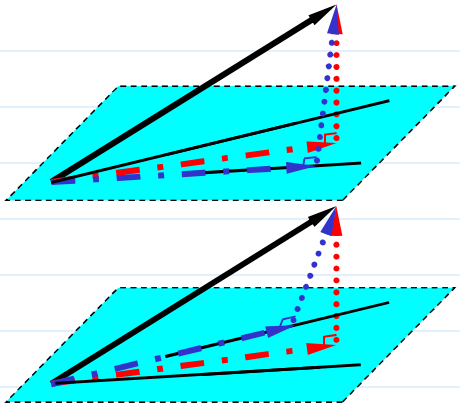
2) test ω_2:model 2 (y~1+A) against Ω_2:model 4 (y~1+A+B) [$df_\omega - df_\Omega$ =3]

3) test ω_3:model 4 (y~1+A+B) against Ω_3:model 5 (y~1+A+B+A:B) [$df_\omega - df_\Omega$ =6]

- $F = \dfrac{(RSS_\omega - RSS_\Omega)/(df_\omega - df_\Omega)}{RSS_{\text{model 5}}/df_{\text{model 5}}} \sim F_{df_\omega - df_\Omega, df_{\text{model 5}}}$
  - invariant to the choice of dummy variables if they generate same ω and Ω
- ANOVA could have different results when the order of effect sequence is changed, e.g., anova(y~*1*+*B*+*A*+*A:B*):

  α) test ω_1:model 1 (y~*1*) against Ω_1:model 3 (y~*1*+*B*) [$df_\omega - df_\Omega$=3]
  β) test ω_2:model 3 (y~*1*+*B*) against Ω_2:model 4 (y~*1*+*B*+*A*) [$df_\omega - df_\Omega$=2]
  χ) test ω_3:model 4 (y~*1*+*B*+*A*) against Ω_3:model 5 (y~*1*+*B*+*A*+*A:B*) [$df_\omega - df_\Omega$=6]

- anova(y~*1*+*A*+*B*+*A:B*) and anova(y~*1*+*B*+*A*+*A:B*) will have **identical** results when orthogonality exists between the 3 groups of effects: span{d_i^A's}, span{d_j^B's}, span{d_{ij}'s}, because in the case, $RSS_\omega - RSS_\Omega$ would equal for 1) and β), 2) and α), 3) and χ)



- identical methodology applies for more qualitative (3-factor interaction, 4-factor interaction, …) and quantitative predictors (similar modeling to what in LNp.8-12~13)

❖ **Reading**: Faraway(2015, 1st ed.), chapter 13, 14.1, 15.1 ❖ **Further reading**: D&S, chapters 14, 23

Transformation

- Recall:
  - objective: for some data, data after transformation can better fit a linear model
  - Q: how to choose an appropriate transformation?
  - transformation can be applied on response and on predictors
- transformation of response
  - Box-Cox transformation family: $t_\lambda(y) = \begin{cases} (y^\lambda - 1)/\lambda, & \text{if } \lambda \neq 0, \\ \log(y), & \text{if } \lambda = 0. \end{cases}$
    - $t_\lambda(y)$ is continuous in λ: for fixed $y>0$,
      $$\lim_{\lambda\to 0} t_\lambda(y) = \lim_{\lambda\to 0} (y^\lambda - 1)/\lambda = \lim_{\lambda\to 0} (y^\lambda \log(y))/1 = \log(y)$$
    - λ=1 ⇒ no transformation, λ=0 ⇒ log, $\lambda \neq 0$ or 1 ⇒ power transformation
    - model: $t_\lambda(y) = \boldsymbol{X\beta} + \boldsymbol{\varepsilon}$, $\boldsymbol{\varepsilon} \sim N(\boldsymbol{0}, \sigma^2 \boldsymbol{I})$
      - parameters: λ, $\boldsymbol{\beta}$, σ
      - can write down likelihood for estimation and testing of λ
      - choice of transformation becomes a estimation/test problem
    - the log-likelihood is
      $$L(\lambda) = \underbrace{(-n/2)\log(RSS_\lambda/n)}_{\text{goodness of fit}} + \underbrace{(\lambda - 1)\Sigma \log(y_i)}_{\text{adjustment}},$$
      where RSS_λ = residual sum of square when using $t_\lambda(y)$ as response, i.e.,
      $$RSS_\lambda = [t_\lambda(y)]^T (I - H) t_\lambda(y)$$

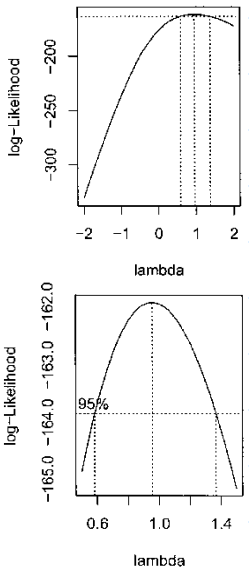
- estimation of λ: choose λ to fit data well using *maximum likelihood*.
  - can compute $L(\lambda)$ for various values of λ and compute $\hat{\lambda}$ exactly to maximize $L(\lambda)$
  - but usually $\hat{\lambda}$ is not a nice round number, e.g., $\hat{\lambda} = -0.17$. It would be hard to explain what this new response means.
  - to avoid this, maximize $L(\lambda)$ over a grid of values, such as $\{-2, -1, -1/2, 0, 1/2, 1, 2\}$. This helps with interpretation.
  - for $\hat{\lambda}$ outside $[-2, 2]$, pay more attention on whether such transformation is required
  - **Q**: why not just minimize RSS_λ to estimate λ?
- test of λ: is the transformation really necessary?
  - we can answer the question form a C.I. for λ
  - *likelihood ratio test* (H_0: $\lambda = \lambda_0$ vs. H_A: $\lambda \neq \lambda_0$):

$$-2[L(\lambda_0) - L(\hat{\lambda})] \sim \chi_1^2 \text{ under } H_0$$

  - a $100(1-\alpha)\%$ C.I. for λ can be formed by:

$$\{\lambda \mid L(\lambda) > L(\hat{\lambda}) - (1/2)\chi_1^2(1-\alpha)\}$$

  - is $\lambda=1$ in the C.I.? if so, may as well stay with no transformation.
  - if rounding $\hat{\lambda}$, check that rounded value is in the C.I.



➢ some notes:

- the Box-Cox method gets upset by outliers --- e.g., if see $\hat{\lambda}=5$, this is probably the reason (**Q**: why?)
- what if some $y_i<0$? adding a constant
- if $\max_i y_i / \min_i y_i$ (>1) is small, Box-Cox won't do anything --- power transforms are well approximated by linear transformations over short intervals
- should the estimation of λ count as an extra parameter to be taken account of in the degrees of freedom? --- difficult question
- Box and Cox (1964) formulate the problem of choosing λ to *make the errors as nearly like a normal sample as possible*

• transformation of predictors

➢ Recall: can use some graphical methods, such as added variable plots and partial residual plot, to offer suggestions for transforming the predictors

➢ could consider Box-Cox family of transformation for each predictor as follows:

$y=\beta_0+\beta_i x_i+\Sigma_{j\neq i}\beta_j x_j+\varepsilon \Rightarrow y=\beta_0+\beta_i^* t_\lambda(x_i)+\Sigma_{j\neq i}\beta_j x_j+\varepsilon$, where t_λ can be x^λ or $\log(x)$

- pick λ that minimizes RSS_λ (**Q**: why only RSS_λ here?) to transform x_i to $t_\lambda(x_i)$
- repeat the procedure for each i $\Rightarrow$ lot of works
- correct transformation for each predictor may depend on getting the others right $\Rightarrow$ may need to perform the procedure for all i's several rounds

- a simpler method
  - approximate x_i^{λ} by $x_i + (\lambda - 1)\, x_i \log(x_i)$ (i.e., first 2 terms in Taylor's expansion of x^{λ} w.r.t. λ) to determine the best $\lambda \Rightarrow$ add the terms $x_i \log(x_i)$ to this model
  - suppose $x_i \log(x_i)$ has regression coefficient $\eta \Rightarrow$ test $H_0: \eta = 0$. If accept, no transformation; if rejected, do transformation
  - $\beta_i^* x_i^{\lambda} \approx \beta_i^* [\, x_i + (\lambda - 1)\, x_i \log(x_i)\,] \Rightarrow \hat{\eta} = \hat{\beta}_i^* (\lambda - 1) \Rightarrow \hat{\lambda} = (\hat{\eta} / \hat{\beta}_i^*) + 1$

• Some issues in transformation
- transformation can be used to
  - stabilize variance
  - improve fitting
  - make errors nearly normally distributed
  - a transformation of scale may also allow use of a simpler model
- these four goals for transformation will not always be met by the same transformation, and compromises may be required
- transformation of $\boldsymbol{Y}$ can alter the error structure, e.g., additive ↔ multiplicative in exp/log. In practice, try different transformation and check if the residuals satisfy the conditions required for linear regression
- prediction in $\boldsymbol{Y}$-space ⇒ back-transforming, same for C.I. for the prediction of $\boldsymbol{Y}$
- It may be difficult to relate the parameters of the untransformed model to the parameters of transformed model. After transforming, regression coefficients will need to interpreted w.r.t. the transformed scale.

❖ **Reading**: Faraway (1st ed.), 7.1 ❖ **Further reading**: D&S, chapters 13