# 6.2 Factors

- Not so simple case: Factor with >2 levels
  - e.g. blood type, eye color, college
  - e.g. GPA = $\beta_0+\beta_1 C+e$

$$C = College = \begin{cases} 0 -- & \text{CC} \\ 1 -- & \text{UC} \\ 2 -- & \text{NA} \\ 3 -- & \text{NC (New Colleges)} \end{cases}$$

- Is it reasonable? Not : problem

if $\beta_1>0 \Rightarrow CC < UC < NA < NC$

if $\beta_1<0 \Rightarrow CC > UC > NA > NC$

# 6.2 Factors

- Factor with d levels
  - The factor rule:
    - A factor with d levels can be represented by (d-1) dummy variables and the intercept.
  - How to use dummy variable?
  - e.g. $GPA = \beta_0+\beta_1 C_{UC}+ \beta_2 C_{NA} + \beta_3 C_{NC} + e$

$$C_{UC} = \begin{cases} 0 -- \text{not UC} \\ 1 -- \text{UC} \end{cases}, \quad C_{NA} = \begin{cases} 0 -- \text{not NA} \\ 1 -- \text{NA} \end{cases}, \quad C_{NC} = \begin{cases} 0 -- \text{not NC} \\ 1 -- \text{NC} \end{cases}$$

- Essentially, we model each group with a separate mean.
- $\beta_0=E(GPA|CC)$, $\beta_0+\beta_1=E(GPA|UC)$, $\beta_0+\beta_2=E(GPA|NA)$, $\beta_0+\beta_3=E(GPA|NC)$

# 6.2 Factors

- Example (Sleep data, sleep1.txt)
  - Y = total hours of sleep
  - X = D = danger of species = 1 to 5.

Constant variance assumption ok

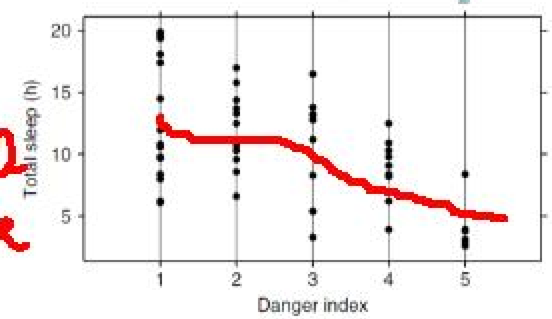



FIG. 6.5 Total sleep versus danger index for the sleep data.

- Model: Total Sleep $\quad U_2 = \begin{cases} 1 & \text{species 2} \\ 0 & \text{otherwise} \end{cases}$
  - $E(TS|D) = \eta_0 + \eta_2 U_2 + \eta_3 U_3 + \eta_4 U_4 + \eta_5 U_5$
- Result:

| | Estimate | Std. Error | $t$-value | Pr(>\|t\|) |
|---|---|---|---|---|
| (b) Mean function (6.16) | | | | |
| Intercept | 13.0833 | 0.8881 | 14.73 | 0.0000 |
| $U_2$ | −1.3333 | 1.3427 | −0.99 | 0.3252 |
| $U_3$ | −2.7733 | 1.4860 | −1.87 | 0.0675 |
| $U_4$ | −4.2722 | 1.5382 | −2.78 | 0.0076 |
| $U_5$ | −9.0119 | 1.6783 | −5.37 | 0.0000 |

- total hours of sleep
= D = danger of species = 1 to 5.

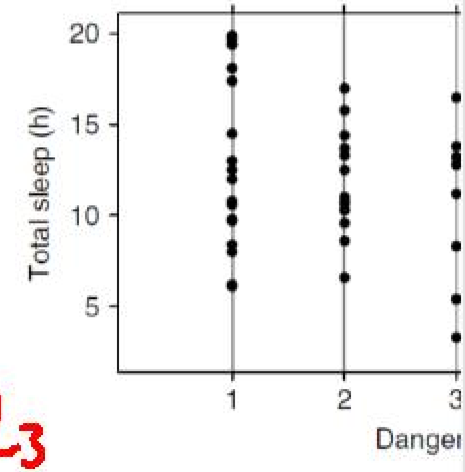



FIG. 6.5 Total sleep versus dange

E(TS | species 1) = $\eta_0$

E(TS | species 2) = $\eta_0 + \eta_2$

E(TS | species 3) = $\eta_0 + \eta_3$

del:

$$\mathrm{E}(TS|D) = \eta_0 + \eta_2 U_2 + \eta_3 U_3 + \eta_4 U_4 + \eta_5 U_5$$

sult:

| | Estimate | Std. Error | $t$-value | Pr(>\|t\|) |
|---|---|---|---|---|
| (b) Mean function (6.16) | | | | |
| Intercept | 13.0833 | 0.8881 | 14.73 | $\boxed{0.0000}$ |
| $U_2$ | −1.3333 | 1.3427 | −0.99 | 0.3252 |
| $U_3$ | −2.7733 | 1.4860 | −1.87 | 0.0675 |
| $U_4$ | −4.2722 | 1.5382 | −2.78 | 0.0076 |
| $U_5$ | −9.0119 | 1.6783 | −5.37 | 0.0000 |
| | Df | Sum Sq | Mean Sq | $F$-value | Pr(>F) |

$\hat{\eta}_0$ ← TS of sleep for species 1

$\hat{\eta}_2$ → −1.3333 ← TS difference b/w 1 & 2, but not significant

$\hat{\eta}_3$ → −2.7733

3 sleep 2.77 hr LESS than 1

= total hours of sleep

= D = danger of species = 1 to 5.

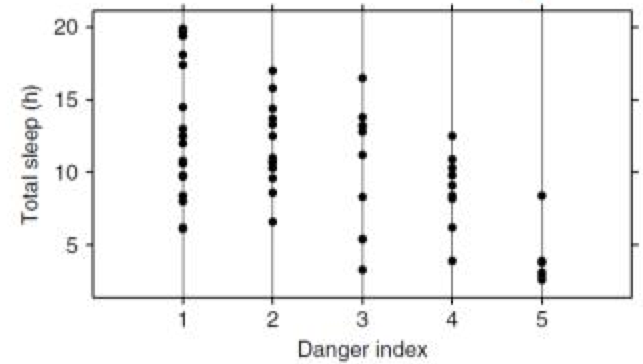



FIG. 6.5 Total sleep versus danger index for the sleep data.

del:

$$E(TS|D) = \eta_0 + \eta_2 U_2 + \eta_3 U_3 + \eta_4 U_4 + \eta_5 U_5$$

sult:

| | Estimate | Std. Error | $t$-value | Pr(>\|t\|) |
|---|---|---|---|---|
| (b) Mean function (6.16) | | | | |
| Intercept | 13.0833 | 0.8881 | 14.73 | 0.0000 |
| $U_2$ | −1.3333 | 1.3427 | −0.99 | 0.3252 |
| $U_3$ | −2.7733 | 1.4860 | −1.87 | 0.0675 |
| $U_4$ | −4.2722 | 1.5382 | −2.78 | 0.0076 |
| $U_5$ | −9.0119 | 1.6783 | −5.37 | 0.0000 |

| Overall | Df | Sum Sq | Mean Sq | $F$-value | Pr(>F) |
|---|---|---|---|---|---|
| $D$ | 4 | 457.26 | 114.31 | 8.05 | 0.0000 |
| Residuals | 53 | 752.41 | 14.20 | | |

$H_0 : TS = \beta_0 + e$

$H_A : TS = \beta_0 + \eta_2 U_2 + \dots + \eta_5 U_5 + e$

Reject $H_0$ : species is significant

# 6.2 Factors

- Remark 1 $E(TS|1)=\beta_1 \quad E(TS|i)=\beta_i$
  - We may use the model
    - $E(TS|D) = \beta_1 U_1 + \beta_2 U_2 + \beta_3 U_3 + \beta_4 U_4 + \beta_5 U_5$

      Convenient for interpretable parameters

    Instead of
    - $E(TS|D) = \eta_0 + \eta_2 U_2 + \eta_3 U_3 + \eta_4 U_4 + \eta_5 U_5$

      Convenient for comparing mean functions

    The major difference is in the interpretation of estimates:

    $\beta_k$= the mean of group k

    $\eta_k$= the mean of group k over group 1
- Remark 2
  - This is an example of One way ANOVA = overall ANOVA with categorical variable
    - Fit a separate mean for each level of a factor.
      - Null model -- $H_o$: All groups have the same mean, i.e. $\eta_s=0$, for s>0
      - Alternative -- $H_A$: Not all groups have the same mean, i.e. some $\eta_s \neq 0$

# 6.2 General Cases – add more predictors

- Example (Sleep data, sleep1.txt)
  - Y = total hours of sleep
  - $X_1$ = D = danger of species = 1 to 5.
  - $X_2$ = Body weight

| | Model 2 | Model 3 | Speci |
|---|---|---|---|
| $E(TS\|1)$ | $\eta_0+\eta_1 x$ | $\eta_0+\eta_1 x$ | 1 |
| $E(TS\|2)$ | $\eta_0+\eta_1 x + \eta_{02} = (\eta_0+\eta_{02})+\eta_1 x$ | $\eta_0+\eta_1 x+\eta_{12}x = \eta_0+(\eta_1+\eta_{12})x$ | 2, 3 |
| $E(TS\|3)$ | $=(\eta_0+\eta_{03})+\eta_1 x$ | $\eta_0+(\eta_1+\eta_{13})x$ | 4, 5 |

- Model 1 (Full model)

$$E(TS|\log(BodyWt)=x,\, D=j) = \eta_0+\eta_1 x+\sum_{i=2}^{d}(\eta_{0j}U_j+\eta_{1j}U_jx)$$

- ✓ Model 2 (Parallel regression) (slope are same)

$$E(TS|\log(BodyWt)=x,\, D=j) = \eta_0+\eta_1 x+\sum_{j=2}^{d}(\eta_{0j}U_j)$$

different species have different mean TS

- Model 3 (Common intercept) (slope are different)

$$E(TS|\log(BodyWt)=x,\, D=j) = \eta_0+\eta_1 x+\sum_{j=2}^{d}(\eta_{1j}U_jx)$$

different species have different rate of change unit weight

- ✓ Model 4 (simple regression, ignore effect of D)

$$E(TS|\log(BodyWt)=x,\, D=j) = \eta_0+\eta_1 x$$

# 6.2 General Cases – add more predictors

- Implementation
- Model 1 (Full model)

$$E(TS|\log(BodyWt) = x, D = j) = \eta_0 + \eta_1 x + \sum_{j=2}^{d} (\eta_{0j} U_j + \eta_{1j} U_j x)$$

← Raw data

| Animal | TS | D | $U_2$ | $U_3$ | $U_4$ | $U_5$ | X (BodyWt) | $U_2X$ | $U_3X$ | $U_4X$ | $U_5X$ |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 4.1 | 2 | 1 | 0 | 0 | 0 | 51 | 51 | 0 | 0 | 0 |
| 2 | 5.3 | 1 | 0 | 0 | 0 | 0 | 62 | 0 | 0 | 0 | 0 |
| 3 | 3.0 | 3 | 0 | 1 | 0 | 0 | 23 | 0 | 23 | 0 | 0 |
| 4 | 8.7 | 5 | 0 | 0 | 0 | 1 | 70 | 0 | 0 | 0 | 70 |
| … | … | … | … | | | | | | | … | … |
| | … | … | … | | | | | | | … | … |

# 6.2 General Cases – add more predictors

Use this

- Model 1 (Full model)

$$E(TS|\log(BodyWt)=x, D=j)=\eta_0+\eta_1 x+\sum_{j=2}^{d}(\eta_{0j}U_j+\eta_{1j}U_j x)$$

- Model 2 (Parallel regression)

$$E(TS|\log(BodyWt)=x, D=j)=\eta_0+\eta_1 x+\sum_{j=2}^{d}(\eta_{0j}U_j)$$

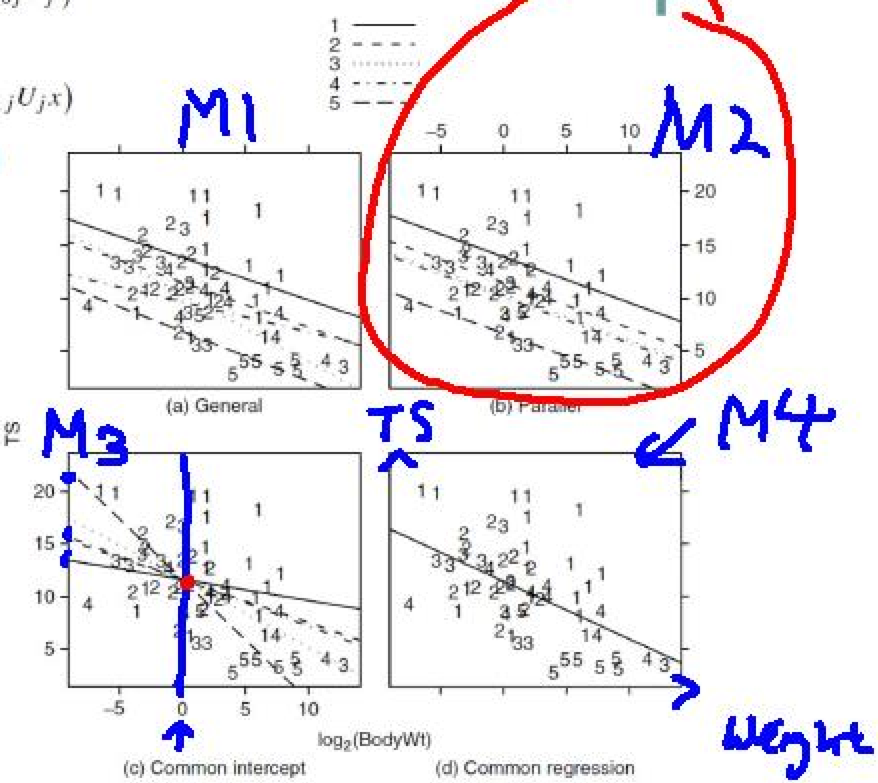
- Model 3 (Common intercept)

$$E(TS|\log(BodyWt)=x, D=j)=\eta_0+\eta_1 x+\sum_{j=2}^{d}(\eta_{1j}U_j x)$$

- Model 4 (simple regression, ignore effect of D)

$$E(TS|\log(BodyWt)=x, D=j)=\eta_0+\eta_1 x$$

- Model Selection using F- test:



(a) General
(b) Parallel
(c) Common intercept
(d) Common regression

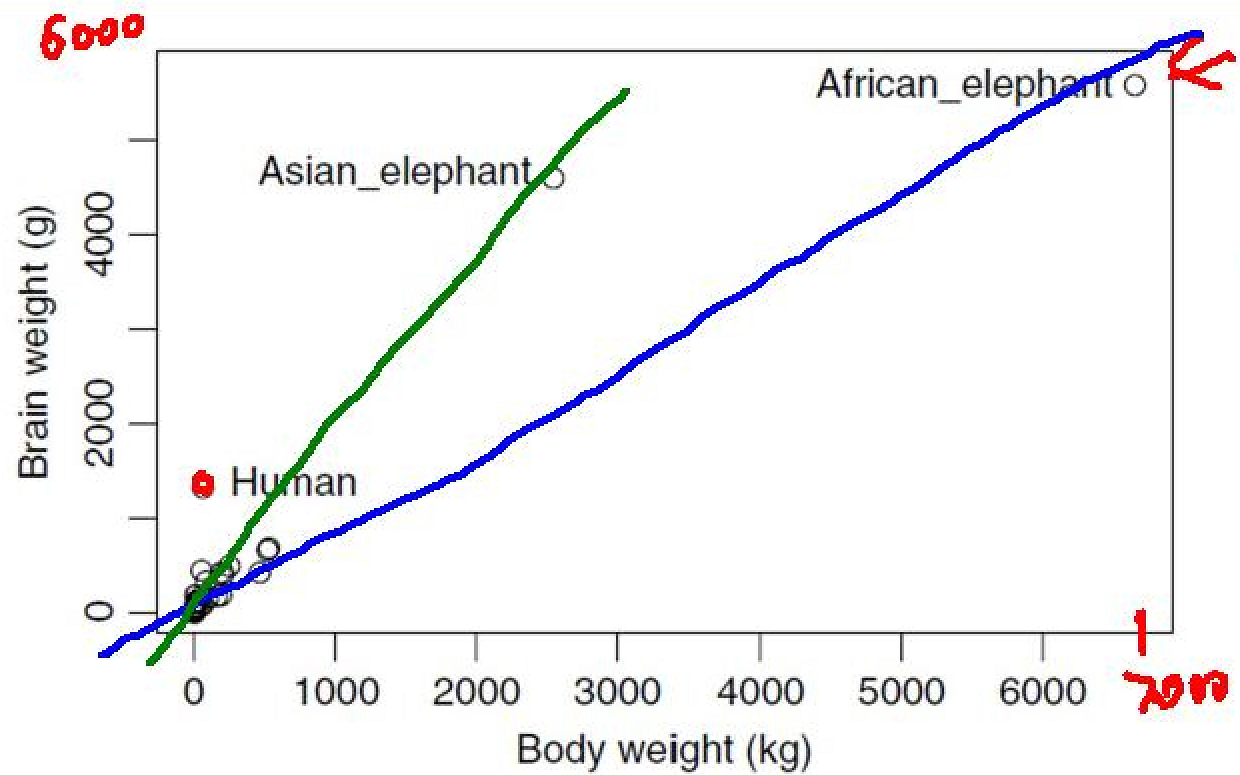
**TABLE 6.2 Residual Sum of Squares and df for the Four Mean Functions for the Sleep Data**

| | df | RSS | F | P(>F) |
|---|---|---|---|---|
| Model 1, most general | 48 | 565.46 | | |
| Model 2, parallel | 52 | 581.22 | 0.33 | 0.853 |
| Model 3, common intercept | 52 | 709.49 | 3.06 | 0.025 |
| Model 4, all the same | 56 | 866.23 | 3.19 | 0.006 |

$H_0$: Model $\ell$
$H_A$: Model 1

$$\boxed{F_\ell=\frac{(RSS_\ell-RSS_1)(df_\ell-df_1)}{RSS_1/df_1}}$$

$\ell = 2/3/4$




X - sep
y - sep
influential
Not outlier

**FIG. 7.1** Plot of *BrainWt* versus *BodyWt* for 62 mammal species.

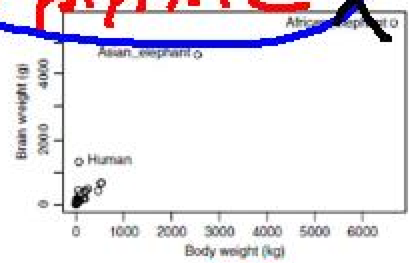
50 kg
× 140
7000

- Is linear regression appropriate?

# 7.1. Transformation

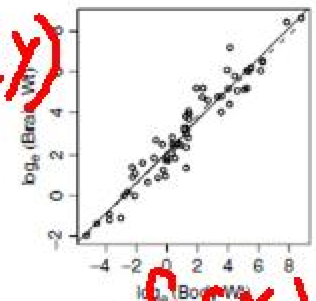
$Y = \beta_0 + \beta_1 X + e$

Y




X

$f_2(y)$




$f_1(x)$

$\Rightarrow f_2(y) = \beta_0 + \beta_1 f_1(x) + e$

- The assumption of linear relationship does not always hold
- We can transform
  - The predictor
  - The response
  - Both

  to achieve the linear relationship

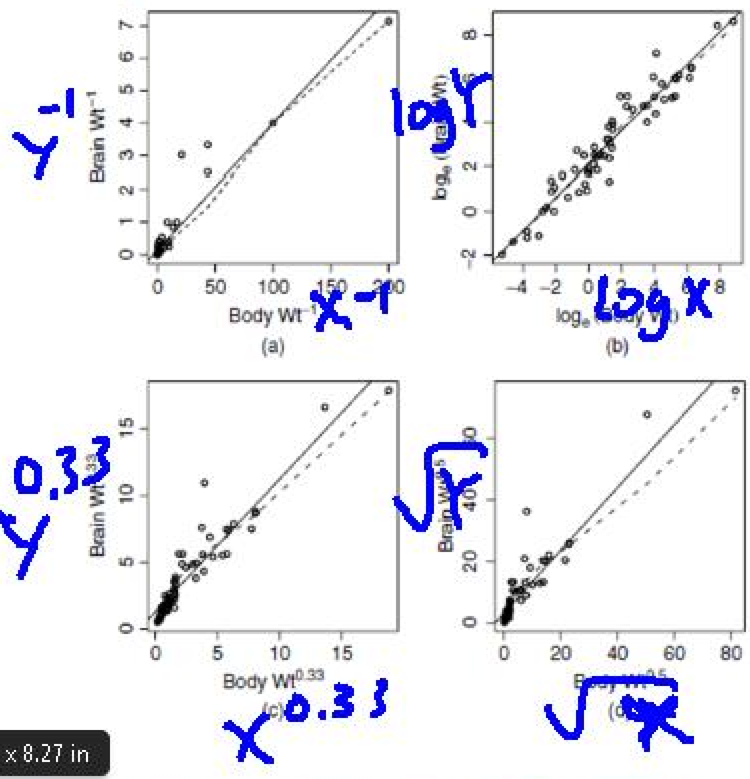
# Power transformation

- Power transformation

x-variable

$\lambda \in \mathbb{R}$ (+ve/-ve)

$$\boxed{\psi(U,\lambda) = U^{\lambda}}$$




- Want a linear relationship

$$\underline{\psi(BrainWt,\lambda)} = \alpha + \beta \underline{\psi(BodyWt,\lambda)} + e$$

- $\lambda =$
  - a) -1 ← $Y^{-1} = \alpha + \beta X^{-1} + e$
  - b) 0 (i.e. log U) ← new definition   $\log Y = \alpha + \beta \log X + e$
  - c) 0.33
  - d) 0.5 ← $\sqrt{Y} = \alpha + \beta\sqrt{X} + e$
- Which λ will you choose?

$\lambda = 0$

$X^0 = 1$

# Interpretation

$$\boxed{\psi(BrainWt, \lambda) = \alpha + \beta\psi(BodyWt, \lambda) + e}$$

- $\lambda > 0$

$$(BrainWt)^{\lambda} = \alpha + \beta(BodyWt)^{\lambda} + e$$

  - Artificial , usually has no physical meaning

- $\lambda = 0$ : log transformation
  - Corresponding to a physical model – allometric model

$$\log(BrainWt) = \alpha + \beta\log(BodyWt) + e$$

$$e^{\alpha}$$

$$\Rightarrow BrainWt = \cancel{\alpha}(BodyWt)^{\beta}\delta \leftarrow e^{e}$$

← error
← exponential function

Multiplicative error