Linear and Generalized Linear Models (4433LGLM6Y)

Statistical theory linear models

Meeting 4

Vahe Avagyan

Biometris, Wageningen University and Research

WAGENINGEN UNIVERSITY & RESEARCH

Statistical theory linear models (Fox, Chapters 9.1-9.3)

- Linear models in matrix form
- Linear contrasts
- Least squares estimation (quadratic form)
- Variance covariance matrix

Statistical theory linear models

- Linear models in matrix form
- Linear contrasts
- Least squares estimation (quadratic form)
- Variance covariance matrix

Linear Models in Matrix Form

- Examples of Linear Models for quantitative response variables:
  - Linear regression models (see Meeting 1)
  - Analysis of variance models (see Meeting 2)
  - Analysis of covariance models (see Meeting 3)

- These models have a lot in common:
  - Normality
  - Constant variance
  - Independence
  - Linearity in parameters, expected error zero.

We will see them in a moment.

Multiple regression model: Revisited

$$Y = \underbrace{\beta_0 + \beta_1 x_1 + \beta_2 x_2}_{} + \epsilon$$

response variable

e.g., observed weight loss of an individual experimental unit

systematic part:

e.g.,

- *population mean of weight loss for exposure time x_1*
- *relative humidity x_2*

random part:

error term is departure of observed weight loss from the mean, represents variation around the mean

Linear Models in Matrix Form

- General linear model for observation $i = 1, \dots, n$ given by:

$$\mathbf{Y}_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik} + \epsilon_i =$$

$$= [1, x_{i1}, x_{i2}, \dots, x_{ik}] \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_k \end{bmatrix} + \epsilon_i =$$

Collect the regressor values of the observation i into a row vector

Collect the regression coefficients into a column vector

$$= \mathrm{x}_i' \boldsymbol{\beta} + \epsilon_i$$

Linear Models in Matrix Form

- Let's collecting all n observations into matrix equation

$$\begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix} = \begin{bmatrix} 1 & x_{11} & \cdots & x_{1k} \\ 1 & x_{21} & \cdots & x_{2k} \\ \vdots & \vdots & & \vdots \\ 1 & x_{n1} & \cdots & x_{nk} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

- Linear models in Matrix Form:

$$\mathbf{y}_{n\times 1} = \mathbf{X}_{n\times(k+1)}\boldsymbol{\beta}_{(k+1)\times 1} + \boldsymbol{\epsilon}_{n\times 1}\,.$$

- $\mathbf{X}$ is called the model matrix or design matrix.

Example: Occupational Prestige

- Occupational prestige (or job prestige) is a way for sociologists to describe the relative social class positions people have.

Example: Duncan data

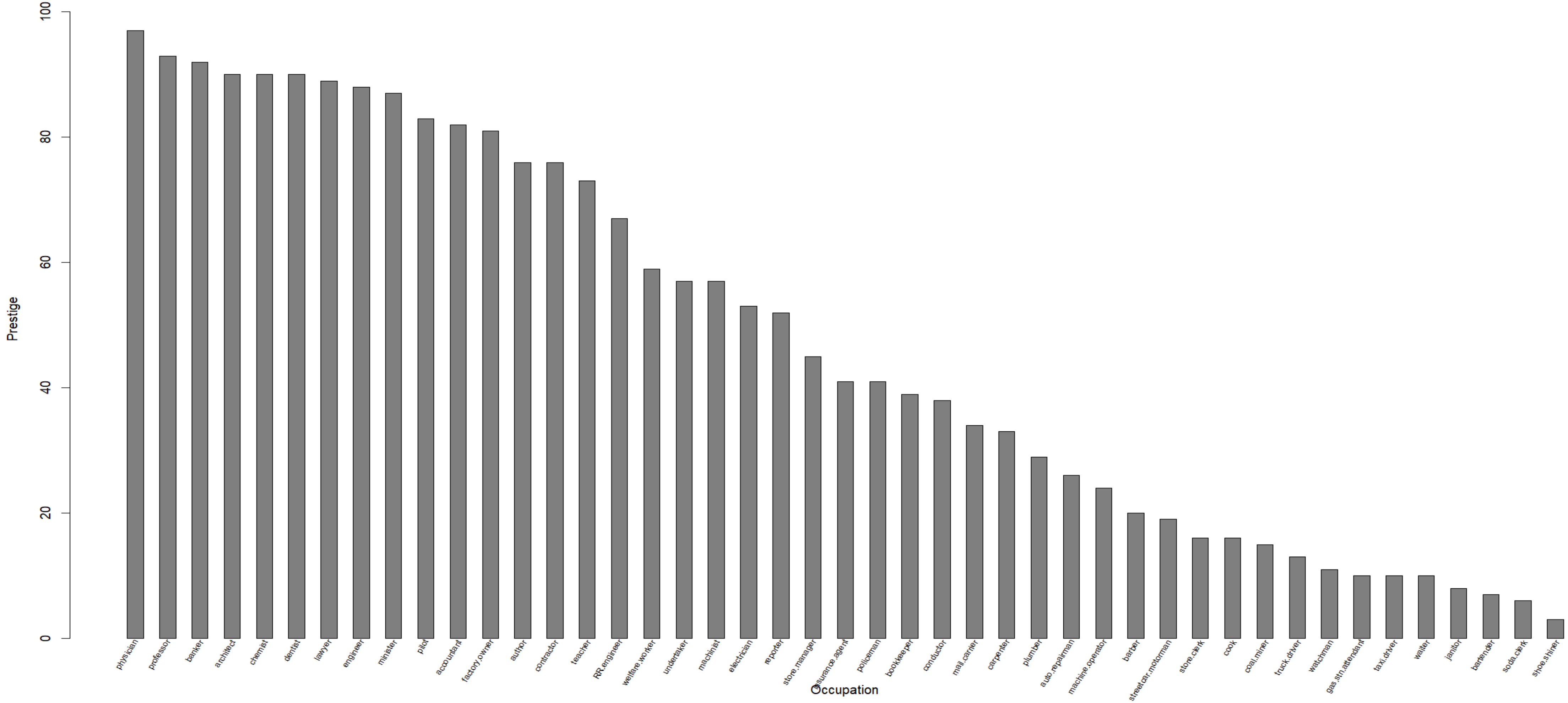
- Duncan's Occupational Prestige Data (from R library `car` or `bcgam`).

```
> library(car)
> head(Duncan)
           type income education prestige
accountant prof     62        86       82
pilot      prof     72        76       83
architect  prof     75        92       90
author     prof     55        90       76
chemist    prof     64        86       90
minister   prof     21        84       87
```

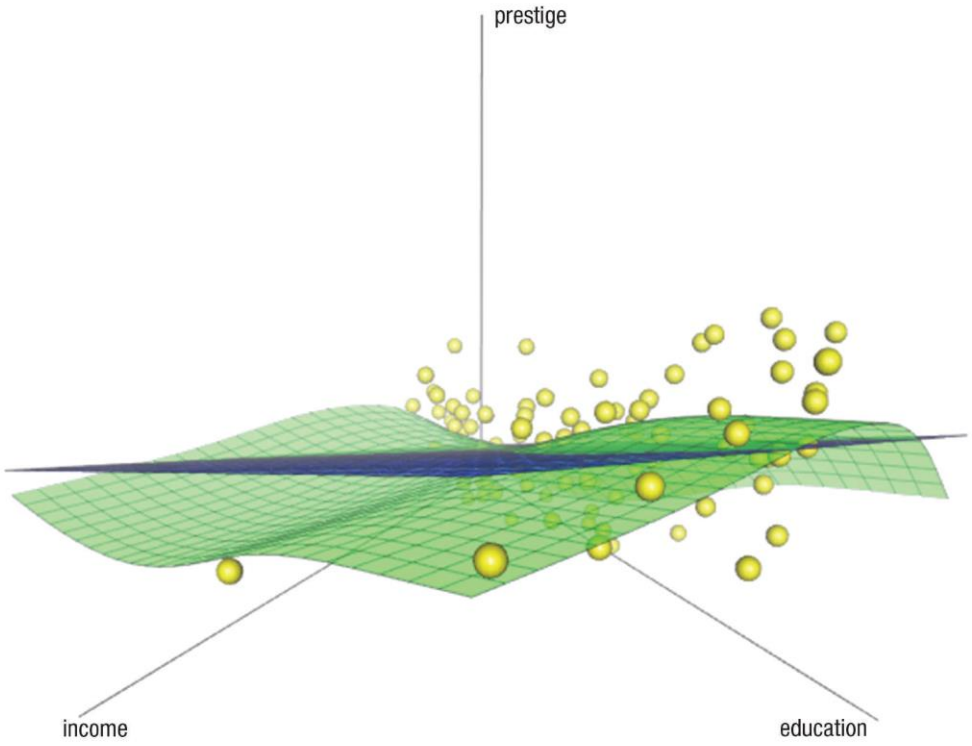
```
> tail(Duncan)
           type income education prestige
cook         bc     14        22       16
soda.clerk   bc     12        30        6
watchman     bc     17        25       11
janitor      bc      7        20        8
policeman    bc     34        47       41
waiter       bc      8        32       10
```

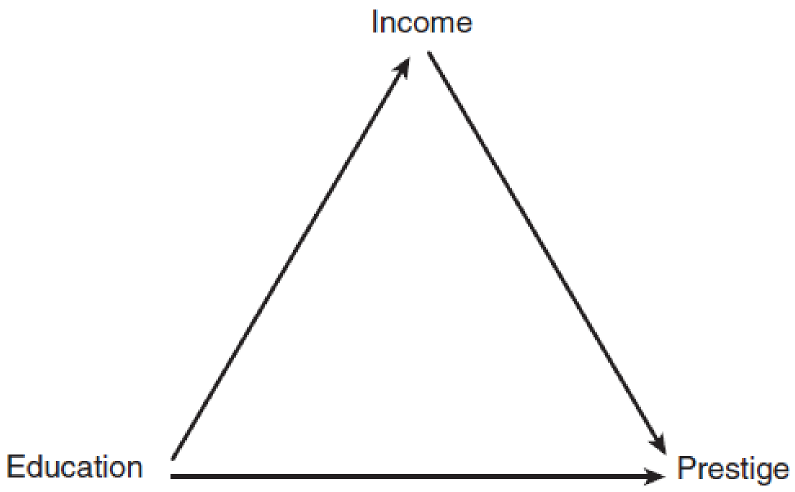
- Data on the prestige, income and education on 45 US occupations (in 1950).

Example: Duncan data



Example: Duncan data





“Causal model” by Fox.

$$Prestige_i = \beta_0 + \beta_1 Income_i + \beta_2 Education_i + \epsilon_i$$

For each $i = 1, \ldots, 45$

Example: Duncan data

```
> head(Duncan)
           type income education prestige
accountant prof     62        86       82
pilot      prof     72        76       83
architect  prof     75        92       90
author     prof     55        90       76
chemist    prof     64        86       90
minister   prof     21        84       87
```

```
> tail(Duncan)
           type income education prestige
cook         bc     14        22       16
soda.clerk   bc     12        30        6
watchman     bc     17        25       11
janitor      bc      7        20        8
policeman    bc     34        47       41
waiter       bc      8        32       10
```

- Linear model in matrix form:

$$\begin{bmatrix} 82 \\ 83 \\ \vdots \\ 41 \\ 10 \end{bmatrix} = \begin{bmatrix} 1 & 62 & 86 \\ 1 & 72 & 76 \\ \vdots & \vdots & \vdots \\ 1 & 34 & 47 \\ 1 & 8 & 32 \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{bmatrix} + \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_{44} \\ \epsilon_{45} \end{bmatrix}$$

$$\mathbf{Prestige}_{45\times 1} = [\mathbf{1}; \mathbf{Income};\ \mathbf{Education}]_{45\times(2+1)}\ \ \boldsymbol{\beta}_{(2+1)\times 1}\ + \boldsymbol{\epsilon}_{45\times 1}$$

Example: Duncan data

- Dependent variables
- Response
- Estimated coefficients

```
> X <- cbind(Duncan$education, Duncan$income)
> Y <- Duncan$prestige
> # Calculations through R
> pres.model <- lm(Y~X)
> # Another way
> # lm(prestige ~ education + income, data = Duncan)
> summary(pres.model)

Call:
lm(formula = Y ~ X)

Residuals:
    Min      1Q  Median      3Q     Max 
-29.538  -6.417   0.655   6.605  34.641 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) -6.06466    4.27194  -1.420    0.163    
X1           0.54583    0.09825   5.555 1.73e-06 ***
X2           0.59873    0.11967   5.003 1.05e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 13.37 on 42 degrees of freedom
Multiple R-squared:  0.8282,	Adjusted R-squared:   0.82 
F-statistic: 101.2 on 2 and 42 DF,  p-value: < 2.2e-16
```

Assumptions linear models in matrix form

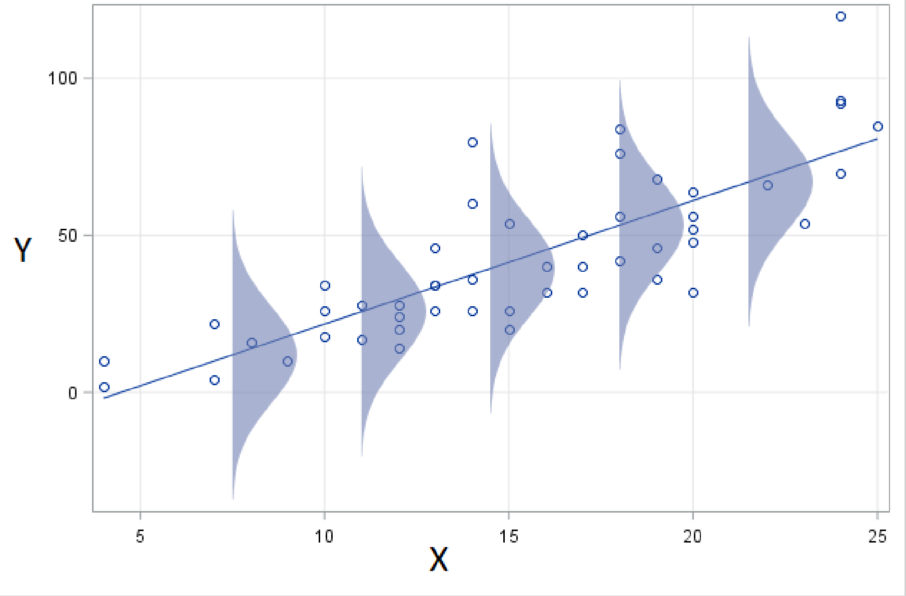
Recall the linear model

$$\mathbf{y}_{n\times 1} = \mathbf{X}_{n\times(k+1)}\boldsymbol{\beta}_{(k+1)\times 1} + \boldsymbol{\epsilon}_{n\times 1}$$

- For the error term $\boldsymbol{\epsilon}$

$$\epsilon \sim N_n(\mathbf{0}, \sigma_\epsilon^2 \mathbf{I}_n)$$

- Expectation: $E(\epsilon) = \mathbf{0}$
- Variance- covariance: $V(\epsilon) = E(\epsilon\epsilon') = \sigma_\epsilon^2 \mathbf{I}_n$.



- For the response $\mathbf{y}$

$$\mathbf{y} \sim N(\mathbf{X}\beta, \sigma_\epsilon^2 \mathbf{I}_n)$$

- Expectation: $\mu = E(\mathbf{y}) = \mathbf{X}\beta$ (Why?)
- Variance- covariance: $\Sigma = V(\mathbf{y}) = \sigma_\epsilon^2 \mathbf{I}_n$. (Why?)

Dummy regression and analysis of variance

- One-way ANOVA model:

$$Y_{ij} = \mu + \alpha_j + \epsilon_{ij} \text{ for groups } j = 1, \ldots, m.$$

$$\begin{bmatrix} Y_{11} \\ \vdots \\ Y_{n_1,1} \\ \hline Y_{12} \\ \vdots \\ Y_{n_2,2} \\ \hline \vdots \\ \hline Y_{1,m-1} \\ \vdots \\ Y_{n_{m-1},m-1} \\ \hline Y_{1m} \\ \vdots \\ Y_{n_m,m} \end{bmatrix} = \begin{bmatrix} 1 & 1 & 0 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & & \vdots & \vdots \\ 1 & 1 & 0 & \cdots & 0 & 0 \\ \hline 1 & 0 & 1 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & & \vdots & \vdots \\ 1 & 0 & 1 & \cdots & 0 & 0 \\ \hline \vdots & \vdots & \vdots & & \vdots & \vdots \\ \hline 1 & 0 & 0 & \cdots & 1 & 0 \\ \vdots & \vdots & \vdots & & \vdots & \vdots \\ 1 & 0 & 0 & \cdots & 1 & 0 \\ \hline 1 & 0 & 0 & \cdots & 0 & 1 \\ \vdots & \vdots & \vdots & & \vdots & \vdots \\ 1 & 0 & 0 & \cdots & 0 & 1 \end{bmatrix} \begin{bmatrix} \mu \\ \alpha_1 \\ \alpha_2 \\ \vdots \\ \alpha_{m-1} \\ \alpha_m \end{bmatrix} + \begin{bmatrix} \epsilon_{11} \\ \vdots \\ \epsilon_{n_1,1} \\ \hline \epsilon_{12} \\ \vdots \\ \epsilon_{n_2,2} \\ \hline \vdots \\ \hline \epsilon_{1,m-1} \\ \vdots \\ \epsilon_{n_{m-1},m-1} \\ \hline \epsilon_{1m} \\ \vdots \\ \epsilon_{n_m,m} \end{bmatrix}$$

Design matrix $\mathbf{X}$ become highly structured

We delete one column, implicitly setting corresponding parameter to 0 (e.g., **R** sets $\alpha_1 = 0$, by default).

Recall the cornerstone representation.

In practice, either α_m or α_1 is set to zero.

Dummy regression and analysis of variance

- Alternative model representations

sum-to-zero condition (sigma constraint or deviation coding):

$$Y_{ij} = \mu + \alpha_j + \epsilon_{ij} \text{ for groups } j = 1, \dots, m.$$

$$\textstyle\sum_j^m \alpha_j = 0$$

- μ is the overall mean of the m types (check it!)

$$\mathbf{X}_F\beta = \left[\begin{array}{ccccc} 1 & 1 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & 1 & 0 & \cdots & 0 \\ \hline 1 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & 0 & 1 & \cdots & 0 \\ \hline \vdots & \vdots & \vdots & & \vdots \\ \hline 1 & 0 & 0 & \cdots & 1 \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & 0 & 0 & \cdots & 1 \\ \hline 1 & -1 & -1 & \cdots & -1 \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & -1 & -1 & \cdots & -1 \end{array}\right] \left[\begin{array}{c} \mu \\ \alpha_1 \\ \alpha_2 \\ \vdots \\ \alpha_{m-1} \end{array}\right]$$

$\mathbf{X}_F$ is of full column rank

Dummy regression and analysis of variance

- Group means: $\boldsymbol{\mu} = [\mu_1, \mu_2, \dots, \mu_m]'$

$$\begin{bmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_{m-1} \\ \mu_m \end{bmatrix} = \begin{bmatrix} 1 & 1 & 0 & \cdots & 0 \\ 1 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & 0 & 0 & \cdots & 1 \\ 1 & -1 & -1 & \cdots & -1 \end{bmatrix} \begin{bmatrix} \mu \\ \alpha_1 \\ \alpha_2 \\ \vdots \\ \alpha_{m-1} \end{bmatrix}$$

$$\boldsymbol{\mu} = \mathbf{X}_{\mathrm{B}} \beta_F$$

- Rows of $\mathbf{X}_B$ form row basis of full-rank model matrix $\mathbf{X}_{\mathrm{F}}$

- We can solve uniquely for the constrained parameters (from ***Linear Algebra***):

$$\beta_F = \mathbf{X}_{\mathrm{B}}^{-1} \boldsymbol{\mu}$$

Dummy regression and analysis of variance

- The solution for the constrained parameters using sigma-constraint is

$$\mu = \mu_{.}$$

$$\alpha_1 = \mu_1 - \mu_{.}$$
$$\alpha_2 = \mu_2 - \mu_{.}$$
$$\dots$$
$$\alpha_{m-1} = \mu_{m-1} - \mu_{.}$$

- Check it.

Dummy regression and analysis of variance

- E.g., ANCOVA model

$$Y_i = \alpha + \beta x_i + \gamma d_i + \delta(x_i d_i) + \epsilon_i$$

Y income

x (i.e., covariate) years of education,

dummy regressor d coding for men ($d = \{0,1\}$ dichotomous factor)

- Model in matrix form is:

$$\begin{bmatrix} Y_1 \\ \vdots \\ Y_{n_1} \\ \hline Y_{n_1+1} \\ \vdots \\ Y_n \end{bmatrix} = \begin{bmatrix} 1 & x_1 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 1 & x_{n_1} & 0 & 0 \\ \hline 1 & x_{n_1+1} & 1 & x_{n_1+1} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & x_n & 1 & x_n \end{bmatrix} \begin{bmatrix} \alpha \\ \beta \\ \gamma \\ \delta \end{bmatrix} + \begin{bmatrix} \epsilon_1 \\ \vdots \\ \epsilon_{n_1} \\ \hline \epsilon_{n_1+1} \\ \vdots \\ \epsilon_n \end{bmatrix}$$

Statistical theory linear models

- Linear models in matrix form
- Linear contrasts
- Least squares estimation (quadratic form)
- Variance covariance matrix

One-way ANOVA: Revisited

- Does sweet taste differ between three types of tomato?

	taste	type
1	25.44	r
2	28.10	r
3	46.46	r
4	36.96	r
5	24.83	b
6	28.47	b
7	48.15	b
8	31.78	b
9	53.42	c
10	70.87	c
11	57.07	c
12	38.08	c

- $Y_{ij} = \mu + \alpha_i + \epsilon_{ij}$
- i = 1, 2, 3 for type and j = 1, 2, 3, 4 for tomato
- Test if the three types have the same population mean for sweet taste or not.
- $H_0: \tau_1 = \tau_2 = \tau_3 = 0$

  is the same as

- $H_0: \mu_1 = \mu_2 = \mu_3$

An example of contrast among the groups

Linear contrasts

- A comparison among m population means is done through **Linear contrasts**

$$l = a_1\mu_1 + a_2\mu_2 + \cdots + a_m\mu_m = \sum_{i=1}^{m} a_i\mu_i$$

Where $\sum_{i=1}^{m} a_i = 0$.

- Example: for checking $\mu_1 = \mu_2$, we can write $l = \mu_1 - \mu_2 =$?
- What are the a_i ?

Linear contrasts

- Recall relationship between group means and parameters in the ANOVA model the :

$$\boldsymbol{\mu} = \mathbf{X}_{\mathrm{B}}\beta_F$$

- Or as a linear function of the means:

$$\beta_F = \mathbf{X}_{\mathrm{B}}^{-1}\boldsymbol{\mu}$$

- Full rank parameterizations allow easy testing of the null:

H_0 : no differences among group means

Or

$$H_0 : \mu_1 = \cdots = \mu_m$$

- Sometimes, we want to formulate $\mathbf{X}_{\mathrm{B}}$ so that individual parameters of β_F represent interesting contrasts among group means.

Linear contrasts: Example

- Data From Friendly and Franklin's (1980) "Experiment on the Effects of Presentation on Recall"

- Memory experiment with 3 experimental conditions:
  - SFR (''standard free recall) - *control group*
  - B ("before") - *experimental group*
  - M ("meshed") - *experimental group*

Condition		
SFR	*B*	*M*
39	40	40
25	38	39
37	39	34
25	37	37
29	39	40
39	24	36
21	30	36
39	39	38
24	40	36
25	40	30

Linear contrasts: Example

- Consider linear contrasts for the following two null hypotheses according to Friendly and Franklin:

1. The mean for the control group is no different from the average of the means for the experimental groups

$$H_0: \mu_1 = \frac{\mu_2 + \mu_3}{2}$$

2. The means for the two experimental groups are the same

$$H_0:\ \mu_2 = \mu_3$$

Rewrite these tests according to the linear contrasts

- We can code each hypothesis as parameter of model, employing the matrix $\mathbf{X}_{\mathrm{B}}^{-1}$.

Linear contrasts: Example

- Recall:

$$\beta_F = \mathbf{X}_{\mathrm{B}}^{-1}\boldsymbol{\mu}$$

- We can write:

$$\begin{bmatrix} \mu \\ \zeta_1 \\ \zeta_2 \end{bmatrix} = \begin{bmatrix} \frac{1}{3} & \frac{1}{3} & \frac{1}{3} \\ 1 & -\frac{1}{2} & -\frac{1}{2} \\ 0 & 1 & -1 \end{bmatrix} \begin{bmatrix} \mu_1 \\ \mu_2 \\ \mu_3 \end{bmatrix}$$

The a_i values.
Check the sums.

- Therefore,
  - $\mu = (\mu_1 + \mu_2 + \mu_3)/3$
  - $\zeta_1 = \mu_1 - \mu_2\ /2 - \mu_3/2$ (First hypothesis is H_0: $\zeta_1 = 0$)
  - $\zeta_2 = \mu_2 - \mu_3$ (Second hypothesis H_0: $\zeta_2 = 0$)

Linear contrasts: Example

- The rows of $\mathbf{X}_\mathrm{B}^{-1}$ are orthogonal.

```
> XBinv<- matrix(c(1/3,1/3,1/3,1,-1/2,-1/2,0,1,-1), ncol=3, byrow = TRUE)
> XBinv %*% t(XBinv)
          [,1] [,2] [,3]
[1,] 0.3333333  0.0    0
[2,] 0.0000000  1.5    0
[3,] 0.0000000  0.0    2
> (XB <- solve(XBinv))
     [,1]       [,2] [,3]
[1,]    1  0.6666667  0.0
[2,]    1 -0.3333333  0.5
[3,]    1 -0.3333333 -0.5
```

- Here $\mathbf{X}_\mathbf{B} = \begin{bmatrix} 1 & \frac{2}{3} & 0 \\ 1 & -\frac{1}{3} & \frac{1}{2} \\ 1 & -\frac{1}{3} & -\frac{1}{2} \end{bmatrix}$, but we could rescale it $\mathbf{X}_\mathbf{B} = \begin{bmatrix} 1 & 2 & 0 \\ 1 & -1 & 1 \\ 1 & -1 & -1 \end{bmatrix}$.

- Check $\mathbf{X}_\mathrm{B}^{-1}\boldsymbol{\mu}$ with the scaled $\mathbf{X}_\mathbf{B}$.

Statistical theory linear models

- Linear models in matrix form
- Linear contrasts
- Least squares estimation (quadratic form)
- Variance covariance matrix

Least squares estimation

- Recall: $y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \epsilon_i$
- Find estimating values $\mathrm{B}_0 = \hat{\beta}_0$, $\mathrm{B}_1 = \hat{\beta}_1$ and $B_2 = \hat{\beta}_2$ estimates for β_0, β_1 and β_2 that minimize the sum of squared errors:

$$\sum_{i=1}^{n}\left(y_i - (B_0 + B_1 x_{1i} + B_2 x_{2i})\right)^2$$

- Same terminology as before:

$$\hat{y}_i = B_0 + B_1 x_{1i} + B_2 x_{2i} \quad \text{is a fitted value}$$

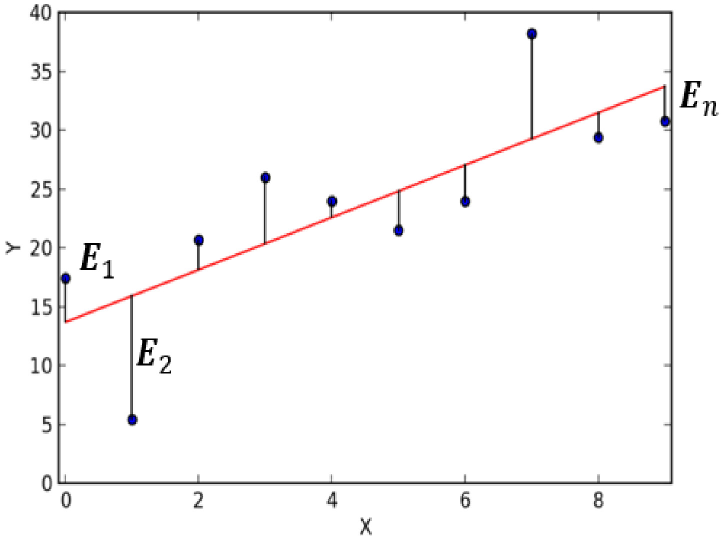
$$e_i = y_i - (B_0 + B_1 x_{1i} + B_2 x_{2i}) = (y_i - \hat{y}_i) \quad \text{is a residual}$$

- See the difference between ϵ_i and e_i.

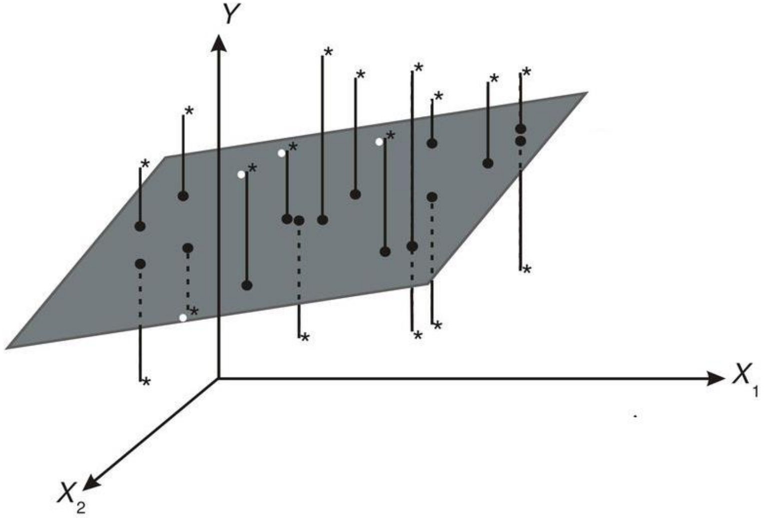
Note: Fox uses E_i instead of e_i.

Least-squares fit: Minimize the sum of squares (SS) of distances

Carl Friedrich Gauss



Simple linear regression



Multiple linear regression

Least-squares fit

- Recall the linear model:

Unknown parameters.

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}, \quad \text{with } \boldsymbol{\epsilon} \sim N_n(\mathbf{0}, \sigma_\epsilon^2 \mathbf{I}_n)$$

Unknown Errors.

- Fitting model to data gives vectors of fitted values and residuals:

$$\mathbf{y}_{n\times 1} = \mathbf{X}\begin{bmatrix} \boldsymbol{B}_0 \\ \boldsymbol{B}_1 \\ \boldsymbol{B}_2 \\ \vdots \\ \boldsymbol{B}_k \end{bmatrix} + \begin{bmatrix} \boldsymbol{e}_1 \\ \boldsymbol{e}_2 \\ \vdots \\ \boldsymbol{e}_n \end{bmatrix} = \mathbf{X}_{n\times(k+1)}\mathbf{b}_{(k+1)\times 1} + \boldsymbol{e}_{n\times 1}$$

- $\mathbf{b} = [\mathbf{B_0}, \mathbf{B_1}, \dots, \mathbf{B_k}]'$ is the vector of estimated coefficients

- $\mathbf{e} = [\boldsymbol{e}_1, \dots, \boldsymbol{e}_n]'$ is the vector of residuals (***distance of the observation from the line/plane***).

Least-squares fit

- Question: *how is* $\boldsymbol{b}$ *obtained*?
- Answer: *Find* $\boldsymbol{b}$ *that minimizes residual sum of squares*:

$$S(\mathbf{b}) = \sum_{i=1}^{n} \boldsymbol{e}_i^2 = [\boldsymbol{e}_1 \quad \boldsymbol{e}_2 \quad \cdots \quad \boldsymbol{e}_n] \begin{bmatrix} \boldsymbol{e}_1 \\ \boldsymbol{e}_2 \\ \vdots \\ \boldsymbol{e}_n \end{bmatrix} = \mathbf{e}'\mathbf{e} = (\mathbf{y} - \mathbf{X}\mathbf{b})'(\mathbf{y} - \mathbf{X}\mathbf{b}) =$$

$$= \mathbf{y}'\mathbf{y} - \mathbf{y}'\mathbf{X}\mathbf{b} - \mathbf{b}'\mathbf{X}'\mathbf{y} + \mathbf{b}'\mathbf{X}'\mathbf{X}\mathbf{b} = \mathbf{y}'\mathbf{y} - 2\mathbf{y}'\mathbf{X}\mathbf{b} + \mathbf{b}'\mathbf{X}'\mathbf{X}\mathbf{b}$$

Note that $\mathbf{y}'\mathbf{X}\mathbf{b} = \mathbf{b}'\mathbf{X}'\mathbf{y}$ (why?)

- The term $(\mathbf{y} - \mathbf{X}\mathbf{b})'(\mathbf{y} - \mathbf{X}\mathbf{b})$ is called a quadratic form.

Least-squares fit

- Obtain LS estimators $\mathbf{b}$, by minimizing $S(\mathbf{b})$.

- Set the vector of partial derivatives w.r.t. $\mathbf{b}$ to zero:

$$\frac{\partial S(\mathbf{b})}{\partial \mathbf{b}} = \mathbf{0} - 2\mathbf{X}'\mathbf{y} + 2\mathbf{X}'\mathbf{X}\mathbf{b} = \mathbf{0}$$

From Linear Algebra Course

- Normal equations:

$$(\mathbf{X}'\mathbf{X})\mathbf{b} = \mathbf{X}'\mathbf{y}$$

- If $\mathbf{X}'\mathbf{X}$ is nonsingular, then the unique least-squares solution is:

$$\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$$

Least-squares fit

- Second partial derivatives of sum of squared residuals:

$$\frac{\partial^2 S(\mathbf{b})}{\partial \mathbf{b}^2} = 2\mathbf{X}'\mathbf{X}.$$

Why?

- Note that $\mathbf{X}'\mathbf{X} = \begin{bmatrix} 1 & 1 & \dots & 1 \\ x_{11} & x_{21} & \dots & x_{n1} \\ \vdots & \vdots & \dots & \vdots \\ x_{1k} & x_{2k} & \dots & x_{nk} \end{bmatrix}'_{(k+1)\times n} \begin{bmatrix} 1 & x_{11} & \dots & x_{1k} \\ 1 & x_{21} & \dots & x_{2k} \\ \vdots & \vdots & \dots & \vdots \\ 1 & x_{n1} & \dots & x_{nk} \end{bmatrix}_{n\times(k+1)} \succcurlyeq 0$

- The solution $\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$ represents the minimum of $S(\mathbf{b})$.

Example: Duncan data, R outcome

```
> X <- cbind(Duncan$education, Duncan$income)
> modX <- cbind(1, X)
> Y <- Duncan$prestige
> t(modX) %*% modX
     [,1]   [,2]   [,3]
[1,]   45   2365   1884
[2,] 2365 163265 122197
[3,] 1884 122197 105148
> t(modX) %*% Y
       [,1]
[1,]   2146
[2,] 147936
[3,] 118229
> (b <- solve(t(modX) %*% modX)%*%(t(modX) %*% Y))
           [,1]
[1,] -6.0646629
[2,]  0.5458339
[3,]  0.5987328
```

β_0

β_1

β_2

```
> # Calculations through R
> pres.model <- lm(Y~X)
> # Another way
> # lm(prestige ~ education + income, data = Duncan)
>
> coef(pres.model)
(Intercept)          X1          X2
 -6.0646629   0.5458339   0.5987328
>
> # Design (model) matrix
> head(model.matrix(pres.model))
  (Intercept) X1 X2
1           1 86 62
2           1 76 72
3           1 92 75
4           1 90 55
5           1 86 64
6           1 84 21
```

Statistical theory linear models

- Linear models in matrix form
- Linear contrasts
- Least squares estimation (quadratic form)
- Variance covariance matrix

Properties of least-squares estimator

- Least-squares estimator: $\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$

1. $\mathbf{b}$ is a linear estimator:

$$\mathbf{b} = \mathbf{M}\mathbf{y}, \quad \text{where } \mathbf{M} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$$

2. $\mathbf{b}$ is an unbiased estimator:

$$E(\mathbf{b}) = \beta$$

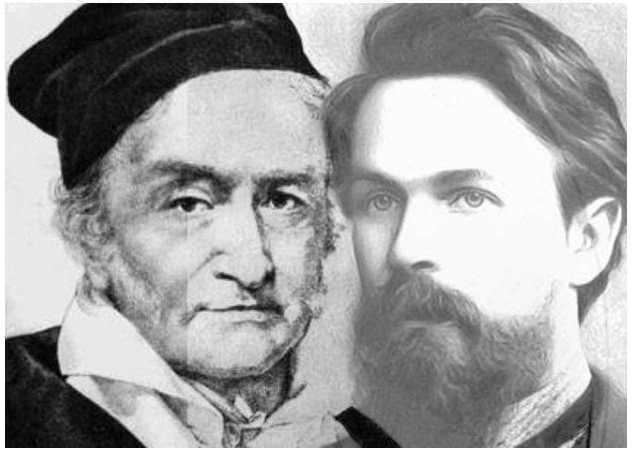
3. $\mathbf{b}$ has a variance-covariance matrix :

$$V(\mathbf{b}) = \sigma_\epsilon^2(\mathbf{X}'\mathbf{X})^{-1}$$

4. $\mathbf{b}$ has a normal distribution, *if* $\mathbf{y}$ is normally distributed:

$$\mathbf{b} \sim N(\beta, \sigma_\epsilon^2(\mathbf{X}'\mathbf{X})^{-1})$$

Gauss-Markov theorem



- Recall $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$
- If errors ϵ_i are
  - independent
  - with zero expectation
  - constant variance

then the least-squares estimator $\mathbf{b}$ is the most efficient (i.e., has the smallest sampling variance) estimator within the class of linear unbiased estimators.

- Least-squares estimator is BLUE (*B*est *L*inear *U*nbiased *E*stimator).

- Under normality, the least-squares estimator is the most efficient of all unbiased estimators.

OLS and Maximum-likelihood estimation

- Under assumptions of linear model, LS estimator $\mathbf{b}$ is also maximum-likelihood estimator of β.
- Linear model with assumption for i-th observation Y_i:

$$Y_i \sim N(x_i'\beta, \sigma_\epsilon^2 \mathbf{I}_n) \quad \text{or} \quad \epsilon_i \sim N(0, \sigma_\epsilon^2 \mathbf{I}_n)$$

- Probability function for observation i:

$$p(Y_i) = \frac{1}{\sigma_\epsilon \sqrt{2\pi}} \exp\left(-\frac{\left(Y_i - x_i'\beta\right)^2}{2\sigma_\epsilon^2}\right).$$

- Joint probability density (i.e., likelihood function):

$$L(\beta, \sigma^2) = p(\mathbf{y}) = \prod p(Y_i) = \frac{1}{\left(\sigma_\epsilon \sqrt{2\pi}\right)^n} \exp\left(-\frac{\sum\left(Y_i - x_i'\beta\right)^2}{2\sigma_\epsilon^2}\right) = \frac{1}{\left(\sigma_\epsilon \sqrt{2\pi}\right)^n} \exp\left(-\frac{(\mathbf{y}-\mathbf{X}\beta)'(\mathbf{y}-\mathbf{X}\beta)}{2\sigma_\epsilon^2}\right).$$

OLS and Maximum-likelihood estimation

- Maximum likelihood estimators (practical exercise)

$$\hat{\beta}_{MLE} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y},$$

$$\hat{\sigma}^2_{MLE} = \frac{\mathbf{e}'\mathbf{e}}{n}.$$

- The least square estimate $\mathbf{b}$ is the same as the $\hat{\beta}_{MLE}$.

- Minimizing sum of squared residuals maximizes the likelihood.
- The estimator $\hat{\sigma}^2_{MLE}$ is biased (although asymptotically unbiased).
- The unbiased estimator is preferred:

$$S_E^2 = \frac{\mathbf{e}'\mathbf{e}}{n-(k+1)}.$$

Estimation of error variance σ_ϵ^2: Revisited

$$\hat{\sigma}_\epsilon^2 = (e_1^2 + \ldots + e_n^2) \,/\, (n - (k+1))$$

Minimized sum of squares
= sum of squares for error
= residual sum of squares
= *SSE*.

degrees of freedom =
number of observations *n*
minus number of β
parameters (*k* slopes + 1
intercept)

$$S_E^2 = \hat{\sigma}_\epsilon^2 = \frac{\mathbf{e}'\mathbf{e}}{n-(k+1)} = SSE/(n-(k+1))$$

is the mean square for error (or residual): *MSE*