1 Linear Models: Preliminaries

1.1 All models are wrong.
1.2 Some models are more useful than others.
1.3 The "correct" or "true" model never can be known.
1.4 In general, simpler models are better than more complex models.

2 Variables in Statistical Models

2.1 Response variable (categorical, numerical; binary, proportion, count, continuous)
2.2 Explanatory variables (categorical, numerical).

3 Model Content

3.1 Content range: null model, minimal adequate model, maximal model, saturated model

3.2 Model properties

Model	# parameters	Model fit	Degrees of freedom	Explanatory power
Null	1: mean y ($\bar{y}$)	none: $SSE = SSY$	n - 1	none
Minimal adequate	$0 \le p' \le p$	$\le$ maximal model	$n - p' - 1$	$r^2 = 1 - SSR / SSY$
Maximal	p+1		$n - p - 1$	
Saturated	n	perfect	none	none

4 Linear Model Structure

4.1 Linear model: General form, numerical explanatory variable(s)

$$y_i = \alpha + \beta_1 x_{i,1} + \beta_2 x_{i,2} + \beta_3 x_{i,3} + \cdots + \beta_p x_{i,p} + \varepsilon_i$$

$$y_i = \alpha + \sum_j \beta_j x_{ij} + \varepsilon_i$$

y_i = response variable

x_i = explanatory variable(s)

α, β = parameters

ε_i = unexplained deviation

4.2 Assumptions

4.2.1 ε_i are independently and identically distributed ("iid").
4.2.2 mean(ε_i) = 0.
4.2.3 $\varepsilon_i \sim N(0,\sigma)$ [ε_i are normally distributed].

5 Model Fitting: Least Squares

Minimize residual sum of squares (SSR)

$$\sum_{i=1}^{n}\left[y_i - \left(\alpha + \sum_{j=1}^{p} \beta_j x_{ij} \right) \right]^2$$

6 Statistical Methods in Model Context

Method	Model	Response var.	Explanatory var(s)
t-test(s)	$y_i = \mu_j + \varepsilon_i$	continuous	categorical
ANOVA, 1-factor	$y_i = \mu + \beta_j + \varepsilon_i$	continuous	categorical
ANOVA, 2-factor	$y_{ijk} = \mu + \beta_i + \beta_j + (\beta_i \beta_j) + \varepsilon_{ijk}$	continuous	categorical (2)
Regression, simple	$y_i = \alpha + \beta x_i + \varepsilon_i$	continuous	continuous
Regression, multiple	$y_i = \alpha + \beta_1 x_{i,1} + \cdots + \beta_p x_{i,p} + \varepsilon_i$	continuous	continuous (>1)
Regression, logistic	$y_i = \dfrac{e^{\alpha + \beta x_i + \varepsilon_i}}{1 + e^{\alpha + \beta x_i}}$	binary (0,1) (categorical)	continuous
Regression, Poisson	$y_i = e^{\alpha + \beta x_i + \varepsilon_i}$	count data	continuous
Goodness of fit	$y_{ij} = np_j + \varepsilon_i$	categorical	categorical
Contingency	$y_{ijk} = np_j p_k + \varepsilon_i$	categorical	categorical

7 Strategies to Improve Models

7.1 Transform response variable (y).

7.2 Transform one or more explanatory variables (x).

7.3 Include different explanatory variable(s), if available.

7.4 Use different error structure.