EVERYTHING QUAL

(just like everything bagel)

Anqi Zhao

08/20/2025

Reviews (from previous students who used this note and took the Qual)

You're invited to leave a review as well if you find this note helpful or if you hate this ¯_(ツ)_/¯
I will also add an acknowledgment section if you find any typos/ errors (there will be a lot, but still first come first serve ofc)
Please contact Anqi Zhao (the one who wrote the note and uploaded) at azhao6@ncsu.edu or find me on discord or at LAU

I may add a table of contents later here

PhD Qualifying EXAM Syllabus
(Updates to this document must be updated on the Student Copy in the PhD Qualifying Exam ***Student Resources*** shared drive. The Graduate Services Coordinator can edit that shared drive.)

ST 703-704: Statistical Methods

Representative Texts

- Rao, P.V. Statistical Research Methods in the Life Sciences, Brooks/Cole.
- Ott, R. Lyman and Longnecker, Michael T. Introduction to Statistical Methods and Data Analysis.
- Damon. Jr., Richard A. and Harvey, Walter R. Experimental Design, ANOVA, and Regression, Harper and Row, Publishers, 9187. 508 pp.
- Neeter, John, Wasserman, William, and Kutner, Michael H. Applied Linear Statistical Models 3rd Ed., Richard D. Irwin, Inc., 1990, 1182 pp.
- Ostle, Bernard and Mensing, Richard W. Statistics in Research, 3rd Ed., Iowa State University Press.
- Snedecor, George W. and Cochran, William G. Statistical Methods, 7th Ed., Iowa State University.
- Steel, Robert G.D. and Torrie, James H. Principles and Procedures of Statistics: A Biometrical Approach, 2nd Ed., McGraw – Hill.
- SAS Institute Inc., SAS/STAT User's Guide, Release 6.03 ED., Cary, NC: SAS Institute Inc., 1988. 1028pp.
- SAS Institute Inc., SAS Language: Reference, Version 6, First Ed., Cary, NC: SAS Institute Inc., 1990. 1042 pp.

Topics to Review

- Definition and computation of elementary descriptive statistics
- Populations and samples
- Sampling distributions and the Central Limit Theorem
- Use of the Z, t, Chi Square, and F tables
- Logical basis of confidence and tests of hypothesis
- Inference on mean, variance and proportion of one population
- Inference on means and proportions from two populations – independent and paired samples
- Inference on variances from two populations
- Power and sample size calculations
- Basic concepts of experimental design including the concept of experimental unit, experimental error, replication, relative efficiency, blocking, covariance, and randomization.
- For each of the following experimental designs:
    - Completely Randomized Design (CRD)

- Randomized Complete Block Design (RCBD)
- Split Plot and Repeated Measures designs

Students Should:

- Know the models (including alternative parameterizations) and related distributional assumptions.
- Be able to recognize which design is appropriate for a given experiment and understand the advantages and disadvantages of each design.
- Know how to construct the ANOVA table with degrees of freedom, sums of squares and mean squares.
- Know what hypotheses can be tested and how to test them, including use of expected mean squares to find appropriate denominators.
- Know how to estimate and place confidence intervals on meaningful linear combinations of the fixed effects such as treatment contrasts, treatment means and other linear combinations.
- Know how to estimate the variance components in the model (using method of moments) and how to use them to obtain variances for linear combinations of (estimated) fixed effects (and understand the correlation structure as a function of these variance components.)
- Be able to recognize replication and subsampling, and account for them in the model, ANOVA table and analysis.
- Know how to make multiple comparisons using a number of procedures that adjust for multiple testing, including the Tukey-Kramer, Scheffe, Bonferroni, and Benjamini-Hochberg procedures.
- Know how to account for one or more covariates.

- Basic concepts of treatment designs including:
  - Treatments and treatment combinations
  - Control versus experimental treatments
  - Factorial treatment designs
  - Factors and their levels
  - Main effects and interactions (1st order, 2nd order, etc.)
  - Nested designs
  - Nested-Factorial designs

- For balanced data, partitioning of the treatment sum of squares in the ANOVA table for each of the treatment designs for fixed, random and mixed models and interpret expected mean squares.
- Multiple regression using matrix notation, including
  - Model and assumptions
  - Normal equations and parameter estimators
  - Properties of the estimators
  - Inference in multiple regression, comparing subsetted models
  - Lack of fit and pure error
  - Residual diagnostics, including

SS df

SSE $\sum_i (Y_i - \hat{Y}_i)^2$ $n-p$

Lack of fit $\sum_i (\hat{Y}_i - \bar{Y}_{i\cdot})^2$ n^*-p

Pure error $\sum_i (Y_i - \bar{Y}_{i\cdot})^2$ $n-n^*$

        - Misspecified mean model
        - Misspecified covariance model
        - Outliers and influential points
        - Multicollinearity
    - Model assessment and exploration, including a variety of criteria, k-fold cross validation, and stepwise or subset exploration
- Correlation
- Analysis of Covariance
- Biased Regression
    - Penalized regression (e.g., ridge, lasso, etc.)
    - Reduced dimensions (e.g., principal components, partial least squares, etc.)
- Generalized Linear Models
    - Model and assumptions
    - Inference
    - Logistic regression
    - Poisson modeling
- Linear Mixed Models
    - Models, assumptions, implied covariance structure
    - Subject-specific versus marginal model
    - ANOVA-type mixed models
    - Inference
    - Clustered/repeated measures/longitudinal data
- Sampling Strategies: simple random, stratified random, cluster, and systematic sampling
- Be able to read and interpret code and output from SAS and R for conducting the above analyses

ST 701-702: Statistical Theory

Representative Texts

- Casella, G. and Berger, R.L. Statistical Inference, 2nd Ed., Wadsworth/Brooks Cole, Pacific Grove, CA, 2001.
- Hogg, R.V., and Craig, A.T. Introduction to Mathematical Statistics, 4th Ed., MacMillan.
- Rohatgi, V.K. An Introduction to Probability Theory and Mathematical Statistics, John Wiley & Sons, New York, 1976.

Topics

- Basic probability calculus
- Random variables, probability distributions, density functions and distribution functions
- Discrete probability models: e.g. binomial, Poisson, geometric, negative binomial, hypergeometric, etc.
- Continuous probability models: e.g. uniform, exponential, beta, gamma, normal Weibull, Cauchy, extreme value, log-normal, etc.
- Multivariate probability models: multinomial, bivariate normal

● Expected value, variance, covariance, correlation, moments (about zero and about the mean) and moment-generating functions
● Moments of functions of random variables
● Joint distributions, conditional distributions, marginal distributions and expectations
● Distributions of functions of random variables, order statistics
● Chebyshev's, Markov's and Jensen's inequalities
● Normal theory: joint distribution of the sample mean and variance, central and noncentral distributions for Student t, Chi square and F
● Convergence in probability and the weak law of large numbers
● Convergence in distribution, the central limit theorem, asymptotic normality, Slutsky's theorem and the "delta method"

● Sufficient and minimal sufficient statistics
● Ancillary statistics
● Complete statistics
● Basu's theorem
● Method of moments, maximum likelihood estimation
Jan 2015 Part 1 Q2
● Bayesian inference: prior and posterior probability distributions, conjugate priors,
Aug 2017 Part 2 Q3
Bayes estimators based on squared error loss; hierarchical models
Aug 2019 Part 2 Q3
● Properties of estimators -- unbiasedness, mean squared error, Cramer-Rao lower bound, Rao-Blackwell and Lehmann-Scheffe theorems, UMVUE, consistency, asymptotic efficiency
● Logical basis for and properties of hypothesis tests
● Type I and II error, level of significance and power
● Simple and composite hypotheses
● Unbiased tests
● Likelihood ratio tests
● Neyman-Pearson lemma for MP tests; Karlin-Rubin theorem for UMP tests, UMPU tests
● Asymptotic tests: chi-square, Wald, score
● Confidence interval construction by inversion of hypothesis tests
● Confidence interval construction using pivots
● Properties of confidence intervals: shortest length, UMA and UMA unbiased

ST 705: Linear Models

Representative Texts

● Graybill, F.A. Theory and Applications of the Linear Model, Duxbury, N. Scituate, Mass, 1976.
● Monahan, J.F. A Primer on Linear Models, CRC Press, 2008.
● Searle, S.R. Linear Models, John Wiley, New York, 1971.

- Seber, G.A.F. Linear Regression Analysis, Wiley, 2003.

Topics

- Review of linear systems of equations, generalized inverses and projection matrices, vector spaces and subspaces
- Linear statistical models and reparameterization
- Least squares theory and computation, including normal equations and partition of sums
of squares
- Estimability and estimable linear functions, restricted models
- Gauss-Markov theorem, BLUE, and generalized least squares
- Theory and application of multivariate normal distribution and related distributions of quadratic forms: central and noncentral chi-squared, central and noncentral F. Cochran's Theorem
- Testing the general linear hypothesis and Likelihood Ratio Tests
- Joint distribution of several BLUEs under normality
- Confidence intervals and sets for parameters and predictions
- Random effects, mixed models, and variance component estimation

confidence interval

General Rules:

① Interval estimator (random):

$$P(\theta \in (\hat{\theta}_L, \hat{\theta}_U)) \geq 1-\alpha$$

② Interval estimates (fixed):

$(\tilde{\theta}_L, \tilde{\theta}_U)$ has actual numbers

Confidence interval for μ:

(one sample - known σ^2)

Assumptions: ① $Y_1 \cdots Y_n$ iid

② $E(Y_i) = \mu$; $Var(Y_i) = \sigma^2$

$Y_i \overset{iid}{\sim} (\mu, \sigma^2)$ ← may not normally distributed

③ $E(\bar{Y}) = \mu$

$SE(\bar{Y}) = \sqrt{\frac{\sigma^2}{n}} = \frac{\sigma}{\sqrt{n}}$

	Y_i normal	Y_i not normal
n large	$\bar{Y} \pm z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$	$\bar{Y} \pm z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$
n small	$\bar{Y} \pm z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$	?

(one sample - unknown σ^2)

	Y_i normal	Y_i not normal
n large	$\bar{Y} \pm z_{\alpha/2}\frac{s}{\sqrt{n}}$	$\bar{Y} \pm z_{\alpha/2}\frac{s}{\sqrt{n}}$
n small	$\bar{Y} \pm t_{\alpha/2, n-1}\frac{s}{\sqrt{n}}$	?

(two sample (μ_1, μ_2) – independent)

Assumptions: ① Samples from 2 different populations

Group 1: $Y_{11} \cdots Y_{1n_1}$ } mutually independent

Group 2: $Y_{21} \cdots Y_{2n_2}$ }

② $Y_{1j} \overset{iid}{\sim} N(\mu_1, \sigma_1^2)$ $\quad Y_{2j} \overset{iid}{\sim} N(\mu_2, \sigma_2^2)$

③ $E(\bar{Y}_{1\cdot} - \bar{Y}_{2\cdot}) = \mu_1 - \mu_2$

$\hat{SE}(\bar{Y}_{1\cdot} - \bar{Y}_{2\cdot}) = \sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}$

$\hat{SE}(\bar{Y}_{1\cdot} - \bar{Y}_{2\cdot}) = S_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$ ← for this case, only consider unknown σ^2

If $\sigma_1^2 \neq \sigma_2^2$:

$$(\bar{Y}_{1\cdot} - \bar{Y}_{2\cdot}) \pm t_{\alpha/2, v}\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}$$

$$v = \frac{\left(\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}\right)^2}{\frac{(S_1^2/n_1)^2}{n_1 - 1} + \frac{(S_2^2/n_2)^2}{n_2 - 1}}$$ Satterthwaite approximate (same type)

If $\sigma_1^2 = \sigma_2^2$:

$$(\bar{Y}_{1\cdot} - \bar{Y}_{2\cdot}) \pm t_{\alpha/2, n_1+n_2-2} \cdot S_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

$$S_p = \sqrt{\frac{(n_1-1)S_1^2 + (n_2-1)S_2^2}{(n_1+n_2-2)}}$$ pooled SD

(two-sample (μ_1, μ_2) – paired data)

Assumptions: ① Samples from 2 measurements of same unit

Group 1: $Y_{11} \cdots Y_{1n}$ } NOT independent

Group 2: $Y_{21} \cdots Y_{2n}$ }

② $Y_{1j} \overset{iid}{\sim} N(\mu_1, \sigma_1^2)$ $\quad Y_{2j} \overset{iid}{\sim} N(\mu_2, \sigma_2^2)$

③ $E(\bar{Y}_{1\cdot} - \bar{Y}_{2\cdot}) = \mu_1 - \mu_2$

$Var(\bar{Y}_{1\cdot} - \bar{Y}_{2\cdot}) = \frac{\sigma_1^2}{n} + \frac{\sigma_2^2}{n} - 2Cov(\bar{Y}_{1\cdot}, \bar{Y}_{2\cdot})$

Let $W_j = Y_{1j} - Y_{2j}$, thus $W_j \overset{iid}{\sim} N(\mu_1 - \mu_2, \sigma_W^2)$ → $\sigma_1^2 + \sigma_2^2$

$$\bar{w} \pm t_{\alpha/2, n-1} \cdot SE(\bar{Y}_{1\cdot} - \bar{Y}_{2\cdot})$$

confidence interval for π :

(one sample)

Assumptions: ① $Y_i = 0$ (failure) ; $Y_i = 1$ (success)

② $Y_1 \cdots Y_n$ are iid

③ $E(Y_i) = \pi$; $Var(Y_i) = \pi(1-\pi)$

④ $E(\bar{Y}) = \pi$; $SE(\bar{Y}) = \sqrt{\frac{\pi(1-\pi)}{n}}$

Wald-typed CI (from CLT) ← know how to implement this and know its drawbacks

When n is large, $\frac{p(1-p)}{n} \approx \frac{\pi(1-\pi)}{n}$, and $p = \bar{y}$ is normal :

$$p \pm z_{\alpha/2} * \sqrt{\frac{p(1-p)}{n}}$$

Issues: ① can include negative values or > 1 values

② doesn't work at $p = 0$ or $p = 1$

③ no coverage guarantee, especially when n is smaller.

solution

Wilson - Score Interval ← at least remember the name and know this is better

$$\frac{p + z^2_{\alpha/2}/2n}{1 + z^2_{\alpha/2}/n} \pm z_{\alpha/2} \frac{\sqrt{p(1-p)/n + z^2_{\alpha/2}/4n^2}}{1 + z^2_{\alpha/2}/n}$$

(two sample - independent)

Assumptions: ① Samples from 2 different populations

Group 1: $Y_{11} \cdots Y_{1n_1}$ } mutually independent

Group 2: $Y_{21} \cdots Y_{2n_2}$ }

② $E(Y_{1j}) = \pi_1$ $E(Y_{2j}) = \pi_2$

③ $E(p_1 - p_2) = \pi_1 - \pi_2$

$$SE(p_1 - p_2) = \sqrt{\frac{\pi_1(1-\pi_1)}{n_1} + \frac{\pi_2(1-\pi_2)}{n_2}}$$

$$(p_1 - p_2) \pm z_{\alpha/2} \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}$$

Confidence Interval v.s. Prediction Interval

* The confidence interval provides information about the mean of Y at a given $X = x_0$.
  - Expect $100(1-\alpha)\%$ of intervals to capture $\beta_0 + \beta_1 x_0$
  - $P(\beta_0 + \beta_1 x_0 \in CI) = 1-\alpha$
* The prediction interval captures Y at $X = x_0$.
  - $Y = \beta_0 + \beta_1 x_0 + E$ is random and independent of the collected data
  - $P(Y \in PI) = 1-\alpha$
  - Both Y and PI are random quantities.

$$\text{Var}(Y-\hat{Y}) = \text{var}(Y) + \text{var}(\hat{Y}) = \sigma^2\left(1 + \left(\frac{1}{n} + \frac{(\bar{x}-x_0)^2}{\Sigma(x_i-\bar{x})^2}\right)\right)$$

↑ assume Y and $\hat{Y}$ independent, because Y_{n+1} is the future data, $\hat{Y}$ from observed

$$Y-\hat{Y} \sim N(0, \text{var}(Y-\hat{Y})) \Rightarrow \frac{Y-\hat{Y}}{SE(Y-\hat{Y})} \sim t_{n-2}$$

df of MSE: df total : $n-1$, df Reg : 1

$$\Rightarrow P\left(-t_{n-2,\alpha/2} \le \frac{Y-\hat{Y}}{\sqrt{MSE\left(1+\frac{1}{n}+\frac{(\bar{x}-x_0)^2}{\Sigma(x_i-\bar{x})^2}\right)}} \le t_{n-2,\alpha/2}\right) = 1-\alpha$$

$$\Rightarrow \hat{Y} \pm t_{n-2,\alpha/2} \cdot \sqrt{MSE\left(1+\frac{1}{n}+\frac{(\bar{x}-x_0)^2}{\Sigma(x_i-\bar{x})^2}\right)}$$

↳ if already have $\text{var}(\hat{Y})$, simply "$+\hat{\sigma}^2$"

This example is in SLR case only, but the idea remains the same, that is:
with $\hat{\text{var}}(\hat{Y})$ known, $\hat{\text{var}}(Y_{n+1}) = MSE + \hat{\text{var}}(\hat{Y})$.

Shortest CI: Minimum width for fixed coverage
UMA CI: Highest accurate coverage (dual to UMP test)
UMAU CI: Best unbiased CI (dual to UMPU test)

Example:	CI-Type	Notes
Normal mean with known variance	Equal-tailed interval	UMA; Shortest
Normal mean with unknown variance	t- interval	UMA unbiased
Exponential mean	Shortest interval not symmetric	UMA, not equal tail
Binomial proportion	Clopper-Pearson interval	unbiased, not shortest

Hypothesis Testing (1)

General Rules

Size ↑

	H_0 true	H_0 false
Reject H_0	Type I error (α)	Power ($1-\beta$)
Accept H_0	Correct	Type II error (β)

P-value

Probability $T(\hat{\theta})$ takes on a value as or more extreme than $T(\hat{\theta})$ assuming H_0.

Pivotal quantity

A pivotal quantity is a function of the data and parameters that has a probability distribution that does not depend on any unknown parameters.

Suppose $X \sim f(x;\theta)$, then

- $Q(X,\theta)$ is a pivotal quantity $\Leftrightarrow$ the distribution of Q does not depend on θ
- then can construct CI using the distribution:

$$P(Dist_{0.975} \le Q(X,\theta) \le Dist_{0.05}) = 0.95$$

$$\Rightarrow \quad P(\ (\cdot) \ \le \theta \le \ (\cdot)\ ) = 0.95$$

Critical value

The critical value is a quantile (cutoff point) of the pivotal quantity's known distribution.

e.g. $z_{0.025}$, $t_{df,0.025}$, $\chi^2_{df,0.05}$, $F_{df_1,df_2,0.05}$

Definition 8.3.5 For $0 \le \alpha \le 1$, a test with power function $\beta(\theta)$ is a *size* α *test* if $\sup_{\theta\in\Theta_0} \beta(\theta) = \alpha$.

Definition 8.3.6 For $0 \le \alpha \le 1$, a test with power function $\beta(\theta)$ is a *level* α *test* if $\sup_{\theta\in\Theta_0} \beta(\theta) \le \alpha$.

Risk function of a decision rule: $\hat{d(x)}$: $\bar{x}$, x_{median}, S_n^2

$R_d(\theta) = E_\theta[L(d(x),\theta)]$ the expected loss of decision rule $d(x)$ when θ is true.

Estimation: $R_d(\theta) = E_\theta[(d(x)-\theta)^2]$; $R_d(\theta) = E_\theta[|d(x)-\theta|]$

Testing:

$$R_\phi(\theta) = \begin{cases} \text{loss(false reject)}\, P_\theta(\phi \text{ rejects } H_0) & \text{if } \theta \in \Omega_0 \\ \text{loss(false accept)}\, P_\theta(\phi \text{ accepts } H_0) & \text{if } \theta \in \Omega_1 \end{cases}$$

$$= \begin{cases} \text{loss(false reject)}\, \pi_\phi(\theta) & \text{if } \theta \in \Omega_0 \\ \text{loss(false accept)}(1-\pi_\phi(\theta)) & \text{if } \theta \in \Omega_1 \end{cases}$$

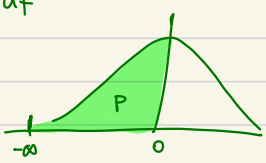
$$\pi_\phi(\theta) = P_\theta(\phi \text{ rejects } H_0) = E_\theta(\phi(X))$$

χ^2 Goodness of fit test

H_0: Observation follows * distribution

$\chi^2 = \sum_i \frac{(O_i - E_i)^2}{E_i} \sim \chi^2_{\#\text{group} - \#\text{parameters} - 1}$

pdf

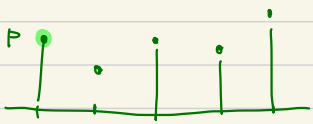


$(-\infty, 0)$ | $(0, 1$ ...

$O_i =$

$E_i = n \cdot p$

pmf



χ^2 independent test

e.g. testing proportion the same across periods:

$H_0: p_1 = p_2 = p_3$

$\chi^2 = \sum_i \sum_j \frac{(E_{ij} - O_{ij})^2}{E_{ij}} \overset{H_0}{\sim} \chi^2_{(I-1)(J-1)}$ with $E_{ij} = \frac{n_{i\cdot} n_{\cdot j}}{n_{\cdot\cdot}}$

T-test / F-test

$Y_{ij} = \mu + \beta_i + \varepsilon_{ij}$ $i = 1, \cdots, 4$. $\sum \beta_i = 0$

$H_0: \beta_1 = \beta_4$ $(c'\beta = 0)$

① Find $\hat{\beta}_i$ (using restricted normal model OR Lagrangian method).

② Must know $\hat{\beta}_1 - \hat{\beta}_4 \sim N(\cdot)$ under null

$$T = \frac{c'\hat{\beta}}{\sqrt{MSE}\sqrt{h'h}} \sim t_{n - \text{rank}X}$$ → df of MSE

$$F = T^2 = \frac{(c'\hat{\beta})^2}{MSE} \sim F_{1, n-\text{rank}(X)}$$

since only one c

$$X = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \end{pmatrix} \quad \beta = \begin{pmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \end{pmatrix}$$ under $H_0: \beta_1 = \beta_4$

$SSE = \sum (Y_i - \hat{Y}_i)^2$	$N-P$
$SSR = \sum (\hat{Y}_i - \bar{Y}_i)^2$	$P-1$
$SST = \sum (Y_i - \bar{Y}_i)^2$	$N-1$

Inference on variance of two populations

$H_0: \sigma_1^2 = \sigma_2^2$ $\quad H_1: \sigma_1^2 \neq \sigma_2^2$

Recall $S_1^2 (n_1 - 1) / \sigma_1^2 \sim \chi^2_{n_1 - 1}$

$S_2^2 (n_2 - 1) / \sigma_2^2 \sim \chi^2_{n_2 - 1}$

Under H_0,

$$\frac{S_1^2 (n_1 - 1) / (n_1 - 1)}{S_2^2 (n_2 - 1) / (n_1 - 1)} = \frac{S_1^2}{S_2^2} \sim F_{n_1 - 1, n_2 - 2}$$

S_1^2: larger variance

S_2^2: smaller variance

Lack of fit test

n^* : unique x value

p : number of parameters

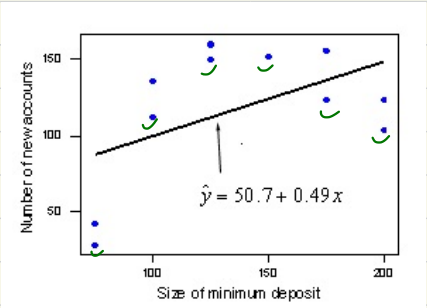
n : total # of observations

H_0: There is no lack of fit $\mu_i = \beta_0 + \beta_1 x_i$

$n^* = 6$

$p = 2$

$r_i = 2, 2, 2, 1, 2, 2$



Source	DF	Adj SS	Adj MS	F-value
Regression	$p-1$	$\sum_{ij}(\bar{Y}_{..} - \hat{Y}_{ij})^2$	SSR / (p-1)	MSR/MSE
Residual Error	$n-p$	$\sum_{ij}(Y_{ij} - \hat{Y}_{ij})^2$	SSE / (n-p)	
LoF Error	n^*-p	$\sum_{ij}(\bar{Y}_{i.} - \hat{Y}_{ij})^2$	SSLF / (n^*-p)	MSLF/MSPE
Pure Error	$n-n^*$	$\sum_{ij}(Y_{ij} - \bar{Y}_{i.})^2$	SSPE / (n-n^*)	
Total	$n-1$	$\sum_{ij}(Y_{ij} - \bar{Y}_{..})^2$		

← if any point is unique cannot measure pure error. Then MSLOF = MSE

related to fit

doesn't relate to fit

$\sum_{i=1}^{n^*} \sum_{j=1}^{r_i} (Y_{ij} - \bar{Y}_{..})^2$

↑ unique x ↑ replicates

Try it!

The lack of fit test

Fill in the missing numbers (??) in the following analysis of variance table resulting from a simple linear regression analysis.

Click on the light bulb in each cell to reveal the correct answer.

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Regression	??	12.597	??	??	0.000
Residual Error	??	??	??		
Lack of Fit	3	??	??	??	??
Pure Error	??	0.157	??		
Total	14	15.522			

Table A.2 Coefficients c_i for orthogonal polynomial trend contrasts

$v = 3$

Trend	c_1	c_2	c_3
Linear	−1	0	1
Quadratic	1	−2	1

$v = 4$

Trend	c_1	c_2	c_3	c_4
Linear	−3	−1	1	3
Quadratic	1	−1	−1	1
Cubic	−1	3	−3	1

$v = 5$

Trend	c_1	c_2	c_3	c_4	c_5
Linear	−2	−1	0	1	2
Quadratic	2	−1	−2	−1	2
Cubic	−1	2	0	−2	1
Quartic	1	−4	6	−4	1

$v = 6$

Trend	c_1	c_2	c_3	c_4	c_5	c_6
Linear	−5	−3	−1	1	3	5
Quadratic	5	−1	−4	−4	−1	5
Cubic	−5	7	4	−4	−7	5
Quartic	1	−3	2	2	−3	1
Quintic	−1	5	−10	10	−5	1

$v = 7$

Trend	c_1	c_2	c_3	c_4	c_5	c_6	c_7
Linear	−3	−2	−1	0	1	2	3
Quadratic	5	0	−3	−4	−3	0	5
Cubic	−1	1	1	0	−1	−1	1
Quartic	3	−7	1	6	1	−7	3
Quintic	−1	4	−5	0	5	−4	1
Sextic	1	−6	15	−20	15	−6	1

Hypothesis testing in ST704

Lack of fit test using deviance ← this is particularly used in GLM.

If Y_i has a distribution "close to normal" with link "close to identity", then $\frac{D_M}{\phi} \approx \chi^2_{n-p}$.

(Approximation will not improve as n increases)

e.g. consider n is the # of groups, where n_i is each group's size. wanna $n_i \uparrow$ instead of n.

$Bin(n_i, p_i)$, $i = 1, \cdots, n$

H_0: model M fits the data H_1: H_0 is not true

When n_i is reasonably large, can use this → Reject if $\frac{D_M}{\phi} > \chi^2_{n-p,\alpha}$

However, this should only be applied when having normal / identity. ① If used this test under overdispersion (Bin/p_i) Type I error ↑. ② may perform poorly when expected counts are small.

Solve 1: Fit a quasi-binomial / quasi-poisson model to estimate dispersion

Solve 2: Use a bootstrap or simulation-based test

H_0: Model M_0 with q reg. parameters H_1: Model M (with $M_0 \subseteq M$) with p reg. parameters

$$T = \frac{D_{M_0} - D_M}{\phi} \xrightarrow{d} \chi^2_{p-q} \quad \text{under } H_0$$

need to distinguish t distribution or χ^2 distribution, and distinguish related df.

$E(\frac{D_M}{\phi}) = n-p \Rightarrow \hat{\phi}_{Mom} = \frac{D_M}{n-p}$, when $\hat{\phi}$ large, we should concern:

Reason 1: Inadequate linear predictor (underspecification)

Reason 2: Over-dispersion $Var(Y_i) > \phi w_i h(\mu_i)$ where ϕ is given by family ← think this

- correlated $Y_1 \cdots Y_n$ may lead to over-dispersion

also see next page

Asymptote Wald Test

$Y \sim N(X\beta, \sigma^2 I)$

H_0: $A\beta = m$ H_1: $A\beta \neq m$ ← $Var(\hat{\beta}) = \sigma^2(X'X)^{-1} \Rightarrow \hat{Var}(\hat{\beta}) = \hat{\sigma}^2(X'X)^{-1}$

$$T_W = (A\hat{\beta} - m)'(A(\hat{F}'\hat{V}^{-1}\hat{F})^{-1}A')^{-1}(A\hat{\beta} - m) \sim \chi^2_{rank(A)}$$

H_0: $h(\theta) = 0$ H_1: $h(\theta) \neq 0$

$$T_W = h(\hat{\theta})'(H(\hat{\theta})(I(\hat{\theta})^{-1})H(\hat{\theta})')^{-1}h(\hat{\theta}) \sim \chi^2_{rank(A)}$$

Test for correlation ρ_{xy}

$$\rho_{xy} = \frac{E((x-\mu_x)(y-\mu_y))}{\sigma_x \sigma_y} \qquad r_{xy} = \frac{1}{n-1}\sum_{i=1}^{n} \frac{(x_i-\bar{x})}{s_x}\frac{(y_i-\bar{y})}{s_y}$$

$$\begin{pmatrix} x_i \\ y_i \end{pmatrix} \overset{iid}{\sim} N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_x^2 & \rho_{xy}\cdot\sigma_y\cdot\sigma_x \\ \rho_{xy}\cdot\sigma_x\cdot\sigma_y & \sigma_y^2 \end{pmatrix}\right)$$

$H_0: \rho_{xy} = 0$

$H_1: \rho_{xy} \neq 0$

$$z = \frac{1}{2}\log\left(\frac{1+r_{xy}}{1-r_{xy}}\right) \sim N\left(\frac{1}{2}\log\left(\frac{1+\rho_{xy}}{1-\rho_{xy}}\right), \frac{1}{n-3}\right)$$

$$\overset{H_0}{\sim} N\left(0, \frac{1}{n-3}\right)$$

RR: $|z\sqrt{(n-3)}| > z_{\alpha/2}$

Test on $h(\theta)$ with $\hat{\theta}_{MLE}$

This part was discussed under LRM setting, but may apply on other occasions

$H_0: h(\theta) = 0 \qquad H_1: h(\theta) \neq 0$ e.g. $h(\theta)_{r\times 1} = A\beta - m$

- $H(\theta) = \frac{\partial h(\theta)}{\partial \theta^T}_{r\times d}$ first partial derivatives of $h(\theta)$
- $\hat{\theta}$ is MLE, $\hat{\theta}_0$ is MLE under H_0
- When H_0 is true, all of $T_W, T_{LR}, T_S \xrightarrow{d} \chi^2_r$ ← See here, unlike the t-statistics, the df is not dfError, but rank(A).

- **Wald statistics:**

$$T_W = h(\hat{\theta})^T \{ H(\hat{\theta}) \underbrace{I(\hat{\theta})^{-1}}_{\text{var}(\hat{\theta})} H(\hat{\theta})^T \}^{-1} h(\hat{\theta})$$

$\{\cdots\}^{-1}$ = $\text{Var}(h(\hat{\theta}))$; only involves $\hat{\theta}$

- **Likelihood ratio statistics:**

$$T_{LR} = -2[\ell(\hat{\theta}_0; y) - \ell(\hat{\theta}; y)]$$

- **Score statistics:**

$$T_S = S(\hat{\theta}_0)^T I(\hat{\theta}_0)^{-1} S(\hat{\theta}_0)$$ ← only involves $\hat{\theta}_0$

HT terminologies

1. A test ϕ is consistent if $\text{power}(\phi) = \pi_1(\phi) \to 1$ ← as data sample size ↑
2. A test ϕ is unbiased if $\sup_{\theta\in\Omega_0} \pi_\phi(\theta) \leq \inf_{\theta\in\Omega_1} \pi_\phi(\theta)$ ← $P(\text{reject}\mid H_0) \leq P(\text{reject}\mid H_1)$
   ≡ Power ≥ α always
3. When $H_0: \theta = a$; $H_1: \theta = b$, try using Neyman-Pearson Lemma show MP (LR)
   When $H_0: \theta = a$; $H_1: \theta > a$ or $\theta < a$, using Karlin-Rubin Lemma show UMP(LR)
   Both show LR monotonic with test statistics $T(x)$.
   When $H_0: \theta = a$; $H_1: \theta \neq a$ cannot use Neyman Pearson / Karlin-Rubin Lemma.
   Cannot be two-sided H_1

One-sample HT

(HT for μ – known σ^2)

$H_0: \mu = \mu_0$

$\bar{Y} \sim N(\mu, \frac{\sigma^2}{n}) \Rightarrow \frac{\bar{Y}-\mu}{\sigma/\sqrt{n}} \sim N(0,1)$

$Z = \frac{\bar{Y}-\mu_0}{\sigma/\sqrt{n}} \sim N(\frac{\mu-\mu_0}{\sigma/\sqrt{n}}, 1) \overset{H_0}{\sim} N(0,1)$

$H_1: \mu > \mu_0$	$RR = Z > C = Z_\alpha$
$H_1: \mu < \mu_0$	$RR = Z < C = -Z_\alpha$
$H_1: \mu \neq \mu_0$	$RR = Z < -Z_{\alpha/2}$ or $Z > Z_{\alpha/2}$

remember when in favor of H_1, reject H_0.
Thus rejection region depends on H_1.

(HT for μ – unknown σ^2)

$H_0: \mu = \mu_0$

$\bar{Y} \sim N(\mu, \frac{\sigma^2}{n}) \Rightarrow \frac{\bar{Y}-\mu}{S/\sqrt{n}} \sim t_{n-1}$

$T = \frac{\bar{Y}-\mu_0}{S/\sqrt{n}} \sim t_{n-1,\lambda} \overset{H_0}{\sim} t_{n-1}$

non-central parameter $\lambda = \frac{\mu-\mu_0}{\sigma/\sqrt{n}}$

$H_1: \mu > \mu_0$	$RR = Z > C = t_{n-1,\alpha}$
$H_1: \mu < \mu_0$	$RR = Z < C = -t_{n-1,\alpha}$
$H_1: \mu \neq \mu_0$	$RR = Z < -t_{n-1,\alpha/2}$ or $Z > t_{n-1,\alpha/2}$

Fact: $T(\hat{\theta}) \in RR$ iff p-value $< \alpha$.

P-value

H_1	known σ^2	unknown σ^2
$\mu > \mu_0$	$1-\Phi(z) = \Phi(-z)$	$1-F_{t_{n-1,0}}(t) = F_{t_{n-1,0}}(-t)$
$\mu < \mu_0$	$\Phi(z)$	$F_{t_{n-1,0}}(t)$
$\mu \neq \mu_0$	$2(1-\Phi(\lvert z\rvert))$	$2(1-F_{t_{n-1,0}}(\lvert t\rvert))$

→ CDF for t dist.

when z small, $\Phi(z)$ is small, then reject

when z large, $\Phi(-z)$ is small, thus reject H_0.

when $|z|$ large, $1-\Phi(|z|)$ small, thus reject

(HT for π)

Exact test: ← uses binomial dist.

$$Y_i \overset{iid}{\sim} Ber(\pi)$$

$H_0: \pi = \pi_0$

Test statistics: $Y = \sum_i Y_i \sim Bn(n, \pi)$

$$P(Y=y) = \binom{n}{y} \pi^y (1-\pi)^{n-y} \qquad y = 0, 1, \cdots, n$$

$H_1: \pi > \pi_0$ $\quad p = P(Y \geq y \mid \pi_0) = \sum_{j=y}^{n} \binom{n}{j} \pi_0^j (1-\pi_0)^{n-j}$

$H_1: \pi < \pi_0$ $\quad p = P(Y \leq y \mid \pi_0) = \sum_{j=0}^{y} \binom{n}{j} \pi_0^j (1-\pi_0)^{n-j}$

P-values: maybe no closed form for RR.

$H_1: \pi \neq \pi_0$ $\quad K = \{k : P(Y=k \mid \pi_0) \leq P(Y=y \mid \pi_0)\}$ (observed y) → means more extreme

$$p = P(Y \in K \mid \pi_0) = \sum_{k \in K} P(Y=k \mid \pi_0)$$

Approximation:

$$p = 2 \cdot \min\{P(Y \geq y \mid \pi_0), P(Y \leq y \mid \pi_0)\}$$

Rao-score test: ← use CLT

$H_0: \pi = \pi_0$

$$Z = \frac{p - \pi_0}{\sqrt{\frac{\pi_0(1-\pi_0)}{n}}} \overset{H_0}{\sim} N(0,1) \qquad p = \bar{Y}$$

← be careful for this part

$H_1: \pi > \pi_0$ $\quad RR: Z > z_\alpha$

$H_1: \pi < \pi_0$ $\quad RR: Z < -z_\alpha$

$H_1: \pi \neq \pi_0$ $\quad RR: Z < -z_{\alpha/2}$ or $Z > z_{\alpha/2}$

← same as one sample procedure with known σ^2

Two-sample HT

(μ_1, μ_2 independent)

$Y_{1j} \overset{iid}{\sim} N(\mu_1, \sigma_1^2)$ $\quad Y_{2j} \overset{iid}{\sim} N(\mu_2, \sigma_2^2)$ indep.

$\bar{Y}_{1\cdot} - \bar{Y}_{2\cdot} \sim N\left(\mu_1 - \mu_2, \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}\right)$

$H_0 : \mu_1 - \mu_2 = \Delta_0$

	Test Statistics	Null Distribution	Alternative Distribution
$\sigma_1^2 = \sigma_2^2$	$\frac{(\bar{Y}_1 - \bar{Y}_2 - \Delta_0)}{sp\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$	$t_{n_1+n_2-2}$	$t_{n_1+n_2-2, \lambda}$
$\sigma_1^2 \neq \sigma_2^2$	$\frac{(\bar{Y}_1 - \bar{Y}_2 - \Delta_0)}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}}$	t_ν	$t_{\nu, \tilde{\lambda}}$

$: sp = \sqrt{\frac{(n_1-1)S_1^2 + (n_2-1)S_2^2}{(n_1+n_2-2)}}$

$: \nu = \frac{(S_1^2/n_1 + S_2^2/n_1)^2}{\frac{(S_1^2/n_1)^2}{n_1-1} + \frac{(S_2^2/n_2)^2}{n_2-1}}$

$\lambda = \frac{(\mu_1 - \mu_2 - \Delta_0)}{\sigma^2\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$ $\qquad \tilde{\lambda} = \frac{(\mu_1 - \mu_2 - \Delta_0)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$

☆

When Y_{1j}, Y_{2j} paired, $W_j = Y_{1j} - Y_{2j}$, $H_0 : \mu_1 - \mu_2 = \Delta_0$

Test Statistics	Null Distribution	Alternative Distribution
$\frac{(\bar{Y}_{1\cdot} - \bar{Y}_{2\cdot} - \Delta_0)}{\sqrt{\frac{\hat{Var}(W_j)}{n}}}$	t_{n-1}	$t_{n-1, \lambda}$

$\lambda = \frac{(\mu_1 - \mu_2 - \Delta_0)}{\sigma_W^2/\sqrt{n}}$

(π_1, π_2 independent)

$Y_{1j} \overset{iid}{\sim} Ber(\pi_1)$ $\quad Y_{2j} \overset{iid}{\sim} Ber(\pi_2)$ indep.

$H_0 : \pi_1 - \pi_2 = \Delta_0$

H_0	Test statistics	Null Distribution
$\pi_1 - \pi_2 = 0$	$Z = \frac{p_1 - p_2}{p(1-p)\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$	$N(0,1)$
$\pi_1 - \pi_2 = \Delta_0$	$Z = \frac{p_1 - p_2 - \Delta_0}{\sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}}$	$N(0,1)$

$p = \frac{n_1 p_1 + n_2 p_2}{n_1 + n_2}$ estimates π (if $H_0 : \pi_1 = \pi_2 = \pi$)

Power & Sample Size

Power = $P(\text{Reject } H_0 \mid H_1)$

(Power for μ, known σ^2)

$\bar{x} \sim N(\mu, \frac{\sigma^2}{n})$

$\Rightarrow \bar{x} - \mu_0 \sim N(\mu - \mu_0, \frac{\sigma^2}{n})$

$\Rightarrow \frac{\bar{x}-\mu_0}{\sigma/\sqrt{n}} \sim N(\frac{\mu-\mu_0}{\sigma/\sqrt{n}}, 1)$

H_1	RR	Power
$\mu > \mu_0$	$z > z_\alpha$	$1 - \Phi(z_\alpha - \frac{\mu-\mu_0}{\sigma/\sqrt{n}}) = \Phi(\frac{\mu-\mu_0}{\sigma/\sqrt{n}} - z_\alpha)$
$\mu < \mu_0$	$z < -z_\alpha$	$\Phi(-z_\alpha - \frac{\mu-\mu_0}{\sigma/\sqrt{n}})$
$\mu \neq \mu_0$	$\lvert z\rvert > z_{\alpha/2}$	$\Phi(\frac{\mu-\mu_0}{\sigma/\sqrt{n}} - z_{\alpha/2}) + \Phi(-z_{\alpha/2} - \frac{\mu-\mu_0}{\sigma/\sqrt{n}})$

$P(z > z_\alpha \mid \mu > \mu_0)$

$= P(\frac{\bar{y}-\mu_0}{\sigma/\sqrt{n}} > z_\alpha \mid \mu > \mu_0)$

$= P(\frac{\bar{y}-\mu_0}{\sigma/\sqrt{n}} - \frac{\mu-\mu_0}{\sigma/\sqrt{n}} > z_\alpha - \frac{\mu-\mu_0}{\sigma/\sqrt{n}} \mid \mu > \mu_0)$

$= 1 - \Phi(z_\alpha - \frac{\mu-\mu_0}{\sigma/\sqrt{n}})$

effect size $\mu - \mu_0$

power depends on the alternative distribution

Q: what is the smallest sample size to achieve $1-\beta$ sample size.

e.g. $1 - \Phi(z_\alpha - \frac{\mu-\mu_0}{\sigma}\sqrt{n}) \geq 1-\beta$

$\Phi(z_\alpha - \frac{\mu-\mu_0}{\sigma}\sqrt{n}) \leq \beta$

$z_\alpha - \frac{\mu-\mu_0}{\sigma}\sqrt{n} \leq z_\beta$

$(\frac{z_\alpha - z_\beta}{\mu-\mu_0/\sigma})^2 \leq n$

$\Phi(-z_\alpha - \frac{\mu-\mu_0}{\sigma}\sqrt{n}) \geq 1-\beta$

$\Phi(z_\alpha + \frac{\mu-\mu_0}{\sigma}\sqrt{n}) \leq \beta$

$z_\alpha + \frac{\mu-\mu_0}{\sigma}\sqrt{n} \leq z_\beta$

$(\frac{z_\alpha - z_\beta}{\mu-\mu_0/\sigma})^2 \leq n$

$\mu \neq \mu_0 \Rightarrow (\frac{z_{\alpha/2} - z_\beta}{\mu-\mu_0/\sigma})^2 \leq n$

(Power for μ, unknown σ^2)

Recall when σ^2 unknown $\frac{\bar{y}-\mu_0}{s/\sqrt{n}} \sim t_{n-1,\lambda} = \frac{\mu-\mu_0}{\sigma/\sqrt{n}}$

H_1	RR	Power
$\mu > \mu_0$	$t > t_{n-1,\alpha}$	$1 - F_{t_{n-1,\lambda}}(t_{n-1,\alpha})$
$\mu < \mu_0$	$t < -t_{n-1,\alpha}$	$F_{t_{n-1,\lambda}}(-t_{n-1,\alpha})$
$\mu \neq \mu_0$	$\lvert t\rvert > t_{n-1,\alpha/2}$	$1 - F_{t_{n-1,\lambda}}(t_{n-1,\alpha/2}) + F_{t_{n-1,\lambda}}(-t_{n-1,\alpha/2})$

$P(t > t_{n-1,\alpha} \mid \mu > \mu_0)$

$= P(\frac{\bar{y}-\mu_0}{s/\sqrt{n}} > t_{n-1,\alpha} \mid \mu > \mu_0)$

$= 1 - F_{t_{n-1,\lambda}}(t_{n-1,\alpha})$

(Power for $\mu_1 - \mu_2$, independent and paired)

indep: $\lambda = \frac{\mu_1 - \mu_2 - \Delta_0}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$

pairs: $\lambda = \frac{\mu_1 - \mu_2 - \Delta_0}{\sigma_w/\sqrt{n}}$

(power for π)

① Power for exact test hard to derive (no alternative)

② Alternative for Rao Score test is Normal.

Recall $Z = \frac{p-\pi_0}{\sqrt{\pi_0(1-\pi_0)/n}} \sim N\left(\frac{\pi-\pi_0}{\sqrt{\pi_0(1-\pi_0)/n}}, \frac{\pi(1-\pi)}{\pi_0(1-\pi_0)}\right)$

Thus,

$$Z \cdot \sqrt{\frac{\pi_0(1-\pi_0)}{\pi(1-\pi)}} - \frac{\pi-\pi_0}{\sqrt{\pi(1-\pi)}}\sqrt{n} \sim N(0,1)$$

H_1	RR	Power
$\pi > \pi_0$	$Z > Z_\alpha$	$\Phi\left(\frac{\pi-\pi_0}{\sqrt{\pi(1-\pi)}}\sqrt{n} - Z_\alpha\sqrt{\frac{\pi_0(1-\pi_0)}{\pi(1-\pi)}}\right)$
$\pi < \pi_0$	$Z < -Z_\alpha$	$\Phi\left(-Z_\alpha\sqrt{\frac{\pi_0(1-\pi_0)}{\pi(1-\pi)}} - \frac{\pi-\pi_0}{\sqrt{\pi(1-\pi)}}\sqrt{n}\right)$
$\pi \neq \pi_0$	$\lvert Z\rvert > Z_\alpha$	$\Phi\left(\frac{\pi-\pi_0}{\sqrt{\pi(1-\pi)}}\sqrt{n} - Z_{\alpha/2}\sqrt{\frac{\pi_0(1-\pi_0)}{\pi(1-\pi)}}\right) + \Phi\left(-Z_{\alpha/2}\sqrt{\frac{\pi_0(1-\pi_0)}{\pi(1-\pi)}} - \frac{\pi-\pi_0}{\sqrt{\pi(1-\pi)}}\sqrt{n}\right)$

$$\begin{aligned}
&P(Z > Z_\alpha \mid \pi > \pi_0) \\
&= P\left(\frac{p-\pi_0}{\sqrt{\pi_0(1-\pi_0)}}\sqrt{n} > Z_\alpha \mid \pi > \pi_0\right) \\
&= P\left(\frac{p-\pi_0}{\sqrt{\pi_0(1-\pi_0)}}\sqrt{n}\cdot\frac{\sqrt{\pi_0(1-\pi_0)}}{\sqrt{\pi(1-\pi)}} - \frac{\pi-\pi_0}{\sqrt{\pi(1-\pi)}}\sqrt{n} > Z_\alpha\cdot\frac{\sqrt{\pi_0(1-\pi_0)}}{\sqrt{\pi(1-\pi)}} - \frac{\pi-\pi_0}{\sqrt{\pi(1-\pi)}}\sqrt{n} \mid \pi > \pi_0\right) \\
&= 1 - \Phi\left(Z_\alpha\cdot\sqrt{\frac{\pi_0(1-\pi_0)}{\pi(1-\pi)}} - \frac{\pi-\pi_0}{\sqrt{\pi(1-\pi)}}\sqrt{n}\right) \\
&= \Phi\left(\frac{\pi-\pi_0}{\sqrt{\pi(1-\pi)}}\sqrt{n} - Z_\alpha\sqrt{\frac{\pi_0(1-\pi_0)}{\pi(1-\pi)}}\right)
\end{aligned}$$

H_1	RR	Sample Size
$\pi > \pi_0$	$Z > Z_\alpha$	$\left(\frac{Z_\alpha\sqrt{\pi_0(1-\pi_0)} - Z_\beta\sqrt{\pi(1-\pi)}}{\pi-\pi_0}\right)^2$
$\pi < \pi_0$	$Z < -Z_\alpha$	$\left(\frac{Z_\alpha\sqrt{\pi_0(1-\pi_0)} - Z_\beta\sqrt{\pi(1-\pi)}}{\pi-\pi_0}\right)^2$
$\pi \neq \pi_0$	$\lvert Z\rvert > Z_\alpha$	$\left(\frac{Z_{\alpha/2}\sqrt{\pi_0(1-\pi_0)} - Z_\beta\sqrt{\pi(1-\pi)}}{\pi-\pi_0}\right)^2$

$$\begin{aligned}
&\Phi\left(\frac{\pi-\pi_0}{\sqrt{\pi(1-\pi)}}\sqrt{n} - Z_\alpha\sqrt{\frac{\pi_0(1-\pi_0)}{\pi(1-\pi)}}\right) \geq 1-\beta \\
&\Phi\left(Z_\alpha\sqrt{\frac{\pi_0(1-\pi_0)}{\pi(1-\pi)}} - \frac{\pi-\pi_0}{\sqrt{\pi(1-\pi)}}\sqrt{n}\right) \leq \beta \\
&Z_\alpha\sqrt{\frac{\pi_0(1-\pi_0)}{\pi(1-\pi)}} - \frac{\pi-\pi_0}{\sqrt{\pi(1-\pi)}}\sqrt{n} \leq Z_\beta \\
&Z_\alpha\sqrt{\frac{\pi_0(1-\pi_0)}{\pi(1-\pi)}} - Z_\beta \leq \frac{\pi-\pi_0}{\sqrt{\pi(1-\pi)}}\sqrt{n} \\
&Z_\alpha\frac{\sqrt{\pi_0(1-\pi_0)}}{\pi-\pi_0} - Z_\beta\frac{\sqrt{\pi(1-\pi)}}{\pi-\pi_0} \leq \sqrt{n} \\
&\left(\frac{Z_\alpha\sqrt{\pi_0(1-\pi_0)} - Z_\beta\sqrt{\pi(1-\pi)}}{\pi-\pi_0}\right)^2 \leq n
\end{aligned}$$

Multiple Testing

Family-wise Type I error rate :

1. $P(\text{Reject } H_0^{ij} \mid \mu_i = \mu_j) = \alpha_i$ ← individual test error rate
2. $H_0: \mu_1 = \cdots = \mu_K$ is then

$$\alpha = P\left(\bigcup_{i<j} \text{Reject } H_0^{ij} \mid \mu_1 = \cdots = \mu_K\right)$$ ← family-wise error rate (FWER)

$$\leq \sum_{i<j} P(\text{Reject } H_0^{ij} \mid \mu_1 = \cdots = \mu_K)$$

$$= \binom{K}{2} \alpha_i$$

I'm not sure if this $\binom{K}{2}$ is fine, because $(K-1)$ tests are enough, but that may considered as Dunnett method (?)

Global F-test

$H_0: L\theta = 0$

① $L\theta$ is estimable

② $\text{rank}(L) = \text{rank}(X) - 1$

③ H_0 true $\Rightarrow \theta$ in the null space of L

$$F = \frac{SSR / \text{rank}(L)}{SSE / (n - 1 - \text{rank}(L))}$$ ← whatever L is used, the test statistic is the same (for global F)

$$\overset{H_0}{\sim} F_{\text{rank}(X)-1,\, n-\text{rank}(X)}$$

Global F-test : Full v.s. Reduced model

$$F = \frac{\frac{SSE(H_0) - SSE(H_1)}{dfE(H_0) - dfE(H_1)}}{SSE(H_1)/dfE(H_1)} = \frac{\frac{SSE(H_0) - SSE(H_1)}{\text{rank}(X) - 1}}{SSE(H_1)/(n - \text{rank}(X))}$$ ← since global / ← full model

General Rules:

1. If all pairwise, use Tukey
2. If pairwise with control, use Dunnett
3. If general $L\theta$ with small rank(L), use Bonferroni ← when # of test is small
4. When in doubt, use scheffe
5. When improve Bonferroni, avoid massive rejection, use Benjamini
6. Fisher does nothing

Fisher

$c'\hat{\theta} \pm t_{df_E, \alpha/2} * SE(c'\hat{\theta})$ ← obviously incorrect

Bonferroni

$c'\hat{\theta} \pm t_{df_E, \alpha/2p} * SE(c'\hat{\theta})$ $\quad p = \binom{k}{2}$

k: # of parameters involved in H_0.

Benjamini

Benjamini-Hochberg Step-Up Procedure (For large k to avoid mass rejection)

1. Compute p-values $p_j = 2 \cdot \Pr(t_{df_{error}} > |\hat{\theta}_j / SE(\hat{\theta}_j)|)$. Goal: Control FDR $< \frac{k_0}{k}\alpha$ → two-sided p-value
2. Reject H_j^0 if $p_j < T_{BH}$ where $T_{BH} := \max\{p_{(j)} : p_{(j)} < \alpha \frac{j}{k}; 1 \le j \le k\}$ → rule

Note: FDR is the expected ratio of the number of ***falsely*** rejected null hypotheses.

Tukey

$c'\hat{\theta} \pm q t_{k, df_E, \alpha} \cdot SE(c'\hat{\theta})$

Tukey HSD : $q = \dfrac{|c'\hat{\theta}|}{SE(c'\hat{\theta})} \Rightarrow$ e.g. $\dfrac{|\bar{Y}_{i\cdot} - \bar{Y}_{j\cdot}|}{\sqrt{MSE(\frac{1}{n_i} + \frac{1}{n_j})}}$

This Tukey HSD follows a studentized range dist

Scheffe:

No interception : ← e.g. one-way ANOVA

$$c'\hat{\theta} \pm \sqrt{(rank(X)-1) F_{rank(X)-1,\, n-rank(X),\, \alpha}}\, SE(c'\hat{\theta})$$

$(rank(X)-1)$: df regression : # of estimable functions (independent) ← max · rank(L)

$n - rank(X)$: df Error

Intercept : ← e.g. full rank linear model or maybe ANOVA with μ ?

$$c'\hat{\theta} \pm \sqrt{rank(X) F_{rank(X),\, n-rank(X),\, \alpha}}\, SE(c'\hat{\theta})$$

$rank(X)$: for full rank linear model : $(1+p)$

Sum of square in testing

$$R(\beta_1 | \beta_0) = SSR(\beta_0, \beta_1) - SSR(\beta_0)$$

$$R(\beta_2, \beta_3 | \beta_0, \beta_1) = SSR(\beta_0, \beta_1, \beta_2, \beta_3) - SSR(\beta_0, \beta_1)$$

$$= R(\beta_1, \beta_2, \beta_3 | \beta_0) + SSR(\beta_0) - (R(\beta_1 | \beta_0) + SSR(\beta_0))$$

$$= R(\beta_1, \beta_2, \beta_3 | \beta_0) - R(\beta_1 | \beta_0)$$

- Special case: model A sets $p - q$ parameters to 0
- Arrange model B parameters: $(\beta_0, \ldots, \beta_q, \beta_{q+1}, \ldots, \beta_p)$
- H_0: Model A "true" $\Leftrightarrow \beta_{q+1} = \cdots = \beta_p = 0$
- H_1: Model B "true" $\Leftrightarrow$ Not all $\beta_{q+1}, \cdots, \beta_p$ equal 0
- $R(\beta_{q+1}, \ldots, \beta_p | \beta_0, \ldots, \beta_q) = SSR(H_1) - SSR(H_0)$
  $= R(\beta_1, \ldots, \beta_p | \beta_0) - R(\beta_1, \ldots, \beta_q | \beta_0)$

Hypothesis Testing (2)

F-test

$Y \sim N_n(X\beta, \sigma^2 I_n)$

$\Rightarrow \lambda'\hat{\beta} \sim N(\lambda'\beta, \sigma^2\lambda'(X'X)^{-}\lambda)$ independent of $\hat{\sigma}^2 = \frac{1}{n-r} Y'(I-P_X)Y$ where

SSE: $\frac{Y'(I-P_X)Y}{\sigma^2} \sim \chi^2_{n-r}$

$(n-r)\frac{\hat{\sigma}^2}{\sigma^2} \sim \chi^2_{n-r}$

Theorem: In the model $Y \sim N_n(X\beta, \sigma^2 I_n)$ with unknown β, σ^2.

$\hat{\beta}_{MLE} = (X'X)^{-}X'y$ and $\hat{\sigma}_{MLE} = \frac{1}{n} Y'(I-P_X)Y$

Definition:

$H_0: K'\beta = m$ $\quad$ $H_1: K'\beta \neq m$ ← estimable if $K_i \in Col(X')$

where $K \in \mathbb{R}^{p \times s}$ with $rank(K) = s$ and every column of K is in $Col(X')$.

The general linear hypothesis is said to be testable iff $K'\beta$ is estimable and has full rank.

Lemma: If $K'\beta$ is estimable, then $H := K'(X'X)^{-}K \in \mathbb{R}^{s \times s}$ is non-singular.

$H = K'(X'X)^{-}K = K'(X'X)^{-}(X'X)(X'X)^{-}K = K'(X'X)^{-}X'W$

$rank(W) \leq \min\{n, s\} \leq s = rank(K) = rank(X'X(X'X)^{-}K) = rank(X'W) \leq rank(W)$

$s = rank(W) = rank(W'W) = rank(H)$

Example: $Y \sim N_n(X\beta, \sigma^2 I_n)$ $\quad K'\hat{\beta} \sim N(K'\beta, \sigma^2 K'(X'X)^{-}K) \equiv N(K'\beta, \sigma^2 H)$

$E(K'\hat{\beta}) \overset{H_0}{=} M \Rightarrow K'\hat{\beta} - M \overset{H_0}{\sim} N(0, \sigma^2 H)$

$(K'\hat{\beta}-M)' H^{-1}(K'\hat{\beta}-M)/\sigma^2 \sim \chi^2_s$

Lemma: Let $Y_n \sim N_n(\mu, V)$ and A, B symmetric matrices. Then if $BVA = 0$, $Y'AY$ and $Y'BY$ are independent.

$\sigma^{-2} Y'(I-P_X)Y \sim \chi^2_{n-r}$

Consider $V = \sigma^2 I$, $A = (I-P_X)$, $H = K'(X'X)^{-}K = LL'$, $B = L^{-1}K'(X'X)^{-}X'\sigma^{-1}$

Because $BVA = 0$, BY is independent of $Y'AY$.

$$F = \frac{(K'\hat{\beta}-M)'H^{-1}(K'\hat{\beta}-M)/\sigma^2 s}{Y'(I-P_X)Y/\sigma^2(n-r)} \overset{H_0}{\sim} F_{s,n-r}$$

$\overset{H_1}{\sim} F_{s,n-r}\left(\frac{1}{2\sigma^2}(K'\beta-M)'(K'\beta-M)\right)$

Likelihood Ratio Test

$\Omega_0 := \{(\beta, \sigma^2) : K'\beta = m, \sigma > 0\}$

$\Omega := \{(\beta, \sigma^2) : \beta \in \mathbb{R}^p, \sigma > 0\}$

$$LR := \frac{\max_{\Omega_0}\{L(\beta, \sigma^2)\}}{\max_{\Omega}\{L(\beta, \sigma^2)\}}$$

$P(LR < c \mid H_0) = \alpha$

Example:

$$LR = \left(\frac{\|y - X\hat{\beta}\|^2}{\|y - X\hat{\beta}_H\|^2}\right)^{\frac{n}{2}} = \left(\frac{Q(\hat{\beta})}{Q(\hat{\beta}_H)}\right)^{\frac{n}{2}}$$

$\hat{\beta}_H$ is from RNE

$$\begin{pmatrix} X'X & P \\ P' & 0 \end{pmatrix}\begin{pmatrix} \beta_H \\ \theta_H \end{pmatrix} = \begin{pmatrix} X'y \\ \delta \end{pmatrix}$$

with $p = k$, $\delta = m$

Reject H_0 if $\left(\frac{Q(\hat{\beta})}{Q(\hat{\beta}_H)}\right)^{\frac{n}{2}} < c$ or $\frac{[Q(\hat{\beta}_H) - Q(\hat{\beta})]/s}{Q(\hat{\beta})/(n-r)} > \frac{n-r}{s}(c^{-\frac{2}{n}} - 1)$

This is actually an F-test.

Theorem: If $K'\beta$ is estimable and $\hat{\beta}_H$ is part of a solution to the RNEs with constrain $K'\beta = m$, then

$$Q(\hat{\beta}_H) - Q(\hat{\beta}) \quad \textcircled{1} = (\hat{\beta}_H - \hat{\beta})'X'X(\hat{\beta}_H - \hat{\beta})$$

$$\textcircled{2} = (K'\hat{\beta} - m)'(K'(X'X)^g K)^{-1}(K'\hat{\beta} - m)$$

$$\textcircled{3} = (K'\hat{\beta} - m)'H^{-1}(K'\hat{\beta} - m)$$

Proof: $Q(\hat{\beta}_H) - Q(\hat{\beta}) = \|y - X\hat{\beta}_H\|^2 - \|y - X\hat{\beta}\|^2$

$= \|X(\hat{\beta} - \hat{\beta}_H)\|^2$

$\textcircled{1} = (\hat{\beta} - \hat{\beta}_H)'X'X(\hat{\beta} - \hat{\beta}_H) \longrightarrow = (\hat{\beta} - \hat{\beta}_H)'K\hat{\theta}_H$

$= (\hat{\beta} - \hat{\beta}_H)'X'X(X'X)^g X'X(\hat{\beta} - \hat{\beta}_H) \quad = (\hat{\beta} - \hat{\beta}_H)'KH^{-1}K(\hat{\beta} - \hat{\beta}_H)$

$= \hat{\theta}_H'K'(X'X)^g K\hat{\theta}_H \quad \textcircled{3} = (K'\hat{\beta} - m)'H^{-1}(K'\hat{\beta} - m)$

$= \hat{\theta}_H H \hat{\theta}_H$

$X'X\hat{\beta}_H + K\hat{\theta}_H = X'y$

$X'X\hat{\beta}_H = X'y - K\hat{\theta}_H$

$X'X(\hat{\beta} - \hat{\beta}_H) = K\hat{\theta}_H$

$K'(X'X)^g X'X(\hat{\beta} - \hat{\beta}_H) = K'(X'X)^g K\hat{\theta}_H$

$\hat{\theta}_H = H^{-1}K'(X'X)^g X'X(\hat{\beta} - \hat{\beta}_H)$

$= H^{-1}(A'X)(X'X)^g X'X(\hat{\beta} - \hat{\beta}_H)$

$= H^{-1}K'(\hat{\beta} - \hat{\beta}_H)$

since $K = X'A$

Corollary: If $K'\beta$ is estimable and $\hat{\beta}$ solves the normal equations, then the $\hat{\beta}_H$ component of a solution to the RNE solver,

$$X'X\beta = X'y - K(K'(X'X)^{g}K)^{-1}(K'\hat{\beta} - m)$$

Proof: $\hat{\beta}_H$ solves $X'X\beta + K\hat{\theta}_H = X'y$

$$\begin{aligned} X'X\beta &= X'y - K\hat{\theta}_H \\ &= X'y - KH^{-1}K'(\hat{\beta} - \hat{\beta}_H) \\ &= X'y - KH^{-1}(K'\hat{\beta} - m) \end{aligned}$$

Theorem: If $P'\beta$ is a system of linear independent, non-estimable functions and $\hat{\beta}_H$ is part of a solution to the RNEs with constraint $P'\beta = \delta$, then $Q(\hat{\beta}) = Q(\hat{\beta}_H)$ and $\hat{\theta}_H = 0$

Proof: $X'X\hat{\beta}_H + P\hat{\theta}_H = X'y$, and so $P\hat{\theta}_H = X'(y - X\hat{\beta}_H) \in col(X')$

Since $P'\beta$ is non-estimable, $col(P) \cap col(X') = \{0\}$, which implies $P\hat{\theta}_H = 0$.

P has full column rank, so it must be the case that $\hat{\theta}_H = 0$. Thus,

$$\hat{\beta}_H \in \{X'X\beta = X'y\}$$

So that $Q(\hat{\beta}) = Q(\hat{\beta}_H)$

Confidence Interval & Multiple Comparison

Single estimable function $\lambda'\beta$

$Y \sim N_n(X\beta, \sigma^2 I_n)$, recall if $\lambda'\beta$ is estimable, then for $\hat{\beta} = (X'X)^{g}X'y$, $\lambda'\hat{\beta} \sim N_1(\lambda'\beta, \sigma^2\lambda'(X'X)^{g}\lambda)$. If we estimate σ^2 with $\hat{\sigma}^2 = \frac{1}{n-r}y'(I-P_X)y$, then

$$t = \frac{\lambda'\hat{\beta} - \lambda'\beta}{\sqrt{\hat{\sigma}^2\lambda'(X'X)^{g}\lambda}} \sim T_{n-r}$$

thus $$P(|t| \le T_{n-r,\alpha/2}) = 1 - \alpha \quad \leftarrow \text{CI coverage}$$

CI: $$\lambda'\beta \in \left[\lambda'\hat{\beta} - T_{n-r,\alpha/2}\cdot\sqrt{\hat{\sigma}^2\lambda'(X'X)^{g}\lambda}\ ,\ \lambda'\hat{\beta} + T_{n-r,\alpha/2}\cdot\sqrt{\hat{\sigma}^2\lambda'(X'X)^{g}\lambda}\right]$$

Multiple estimable function $\Lambda'\beta$

Consider s estimable functions $\lambda_1'\beta, \lambda_2'\beta, \cdots, \lambda_s'\beta$, then let $\Lambda = (\lambda_1 \cdots \lambda_s)$ with linearly independent columns, $\Lambda' = \begin{pmatrix} \lambda_1' \\ \vdots \\ \lambda_s' \end{pmatrix}$

$$\Lambda'\hat{\beta} \sim N_s(\Lambda'\beta, \sigma^2\Lambda'(X'X)^g\Lambda)$$

In this case,

$$P(a_j \le \Lambda_j'\beta \le b_j) = 1-\alpha \quad , \forall j \in \{1, \cdots, s\}, \text{ where}$$
$$a_j := \lambda'\hat{\beta} - t_{n-r,\alpha/2} \cdot \sqrt{\hat{\sigma}^2\lambda'(X'X)^g\lambda}$$
$$b_j := \lambda'\hat{\beta} + t_{n-r,\alpha/2} \cdot \sqrt{\hat{\sigma}^2\lambda'(X'X)^g\lambda}$$

However,

$$P(\bigcap_{j=1}^{s}\{a_j \le \Lambda_j'\beta \le b_j\}) \le P(a_1 \le \Lambda_1'\beta \le b_1) = 1-\alpha$$

↳ Thus, need to adjust family-wise Type-I-error

Theorem: Let $\{E_j\}$ be a collection of measurable events, then:

(i). $P(\bigcup_j E_j) \le \sum_j P(E_j)$

(ii). $P(\bigcap_j E_j) \ge 1 - \sum_j P(E_j^c)$

$$P(\bigcap_j E_j) = 1 - P((\bigcap_j E_j)^c)$$
$$= 1 - P(\bigcup_j E_j^c) \quad \text{by DeMorgan's rule}$$
$$\ge 1 - \Sigma_{j=1} P(E_j^c) \quad \text{by (i)}$$

① **Bonferroni method:** adjust level of each interval as α/s.

Bonferroni method interval:

$$P(a_j \le \Lambda_j'\beta \le b_j) = 1-\alpha \quad , \forall j \in \{1, \cdots, s\}, \text{ where}$$
$$a_j := \lambda'\hat{\beta} - t_{n-r,\alpha/(2s)} \cdot \sqrt{\hat{\sigma}^2\lambda'(X'X)^g\lambda}$$
$$b_j := \lambda'\hat{\beta} + t_{n-r,\alpha/(2s)} \cdot \sqrt{\hat{\sigma}^2\lambda'(X'X)^g\lambda}$$

$$P(\bigcap_{j=1}^{s}\{a_j \le \Lambda_j'\beta \le b_j\}) \ge 1 - \sum_{i=1}^{s} P(\{a_j \le \Lambda_j'\beta \le b_j\}^c)$$
$$= 1 - \sum_{i=1}^{s} \frac{\alpha}{s}$$
$$= 1 - \alpha \quad \text{CI coverage} \ge 1-\alpha$$

② Scheffé method Construct a confidence interval $\forall$ linear combination $u'\Lambda'\beta$ of the form

$$\Gamma(u,c) = u'\Lambda'\hat{\beta} \pm c \cdot \sqrt{\hat{\sigma}^2 \cdot u'\Lambda'(x'x)^{g}\Lambda u},$$

where c is such that $P(u'\Lambda\beta \in \Gamma(u,c)) = 1-\alpha$

CI: $$u'\Lambda'\hat{\beta} \pm \sqrt{s \cdot F_{s,n-r,\alpha}} \cdot \sqrt{\hat{\sigma}^2\, u'\Lambda'(x'x)^{g}\Lambda u}$$

(s: rank of test; MSE df)

③ Tukey method

Replace $t_{n-r,\alpha/2}$ by $q^{*}_{K,n-r,\alpha}$ (studentized range distribution quantile; K: # of parameters involved; MSE df)

$$P(a_j \le \Lambda_j'\beta \le b_j) = 1-\alpha \quad , \forall j \in \{1, \cdots, s\}, \text{ where}$$

$$a_j := \lambda'\hat{\beta} - q^{*}_{K,n-r,\alpha} \cdot \sqrt{\hat{\sigma}^2\lambda'(x'x)^{g}\lambda}$$

$$b_j := \lambda'\hat{\beta} + q^{*}_{K,n-r,\alpha} \cdot \sqrt{\hat{\sigma}^2\lambda'(x'x)^{g}\lambda}$$

Hypothesis Testing (3)

- decision rules d_1 & d_2.
- d_1 **dominates** d_2 if $R_{d_1}(\theta) \le R_{d_2}(\theta)\ \forall \theta \in \Omega$, with "$<$" for at least one θ.
- If d is dominated by some rule, then it's **inadmissible**, otherwise it's admissible.
- d is **minimax** if $\max_{\theta\in\Omega} R_d(\theta) = \min_{d'} \max_{\theta\in\Omega} R_{d'}(\theta)$

Bayes Decision Rule

prior distribution of Θ with pdf (pmf) $\varphi(\theta)$.

- Bayes risk is the expected loss:

$$\begin{aligned} r_d(\varphi) &= E[L(d(x),\theta)] \\ &= E_\varphi[E[L(d(x),\theta)|\theta]] \\ &= E_\varphi[R_d(\theta)] \end{aligned}$$

$$\begin{aligned} r_d(\varphi) &= E[L(d(x),\theta)] \\ &= E[E[L(d(x),\theta)|X]] \\ &= E[\text{expected posterior loss for } d(x)] \end{aligned}$$

- Bayes decision rule:

$$d:\ r_{d_\varphi}(\varphi) = \min_d r_d(\varphi)$$

* may not exist, may not be unique

- Bayes estimator:

$$L(a,\theta) = (a-\theta)^2 \quad \Rightarrow \quad d_\varphi(x) = E[\theta|X=x]$$

$$L(a,\theta) = |a-\theta| \quad \Rightarrow \quad d_\varphi(x) = \text{median}[\theta|X=x]$$

- Bayes testing:

Expected posterior loss:

$$E[L(a,\Theta)|X=x] = \begin{cases} \text{Loss(false accept)}\ P(\theta=\Omega_1|X=x) & \text{if accept } H_0 \\ \text{Loss(false reject)}\ P(\theta=\Omega_0|X=x) & \text{if reject } H_0 \end{cases}$$

① prob of H_1 true; ② prob of H_0 true

Bayes test:

$$\phi_\varphi(x) = \begin{cases} H_0 & \text{if } ① < ② \\ H_1 & \text{if } ① > ② \end{cases}$$

A **unique** Bayes rule w.r.t. a general loss function L is admissible w.r.t. L.

ϕ admissible $\Rightarrow$ ϕ bayes rule $\Rightarrow$ LRT

Parameter Estimation

Simple Linear Regression

Quantities:

Covariance:

$Cov(X_i, Y_i) = E((X_i - \mu_X)(Y_i - \mu_Y)) \Rightarrow \hat{Cov}(X_i, Y_i) = \frac{1}{n-1}\Sigma_i (X_i - \bar{x})(Y_i - \bar{y})$

Correlation:

$\rho_{xy} = \frac{Cov(X_i, Y_i)}{\sigma_X \sigma_Y} = \frac{E((X_i - \mu_X)(Y_i - \mu_Y))}{\sigma_X \sigma_Y}$

$= E\left(\left(\frac{X_i - \mu_X}{\sigma_X}\right)\left(\frac{Y_i - \mu_Y}{\sigma_Y}\right)\right) \Rightarrow r_{xy} = \frac{1}{n-1}\Sigma_i \frac{(X_i - \bar{x})}{S_X}\frac{(Y_i - \bar{y})}{S_Y}$

↑ test for this see previous pages

SST & SSR:

$\underbrace{\Sigma(Y_i - \bar{y})^2}_{SST} = \underbrace{\Sigma(Y_i - \hat{y})^2}_{SSE} + \underbrace{\Sigma(\hat{y} - \bar{y})^2}_{SSR}$ ← sum of square regression: compares fitted value under SLR to fitted value assuming $E(Y|X=x) = \beta_0$

$\Downarrow$

$\frac{\Sigma(Y_i - \hat{y})^2}{(n-2)}$: MSE

$0 \le R^2 \le 1$: is the % of total variation explained by the simple linear regression model.

$R^2 = \frac{SSR}{SST} \overset{\text{in SLR}}{=} (r_{xy})^2$

Residuals & assumption check: $R_i = Y_i - \hat{Y}_i$

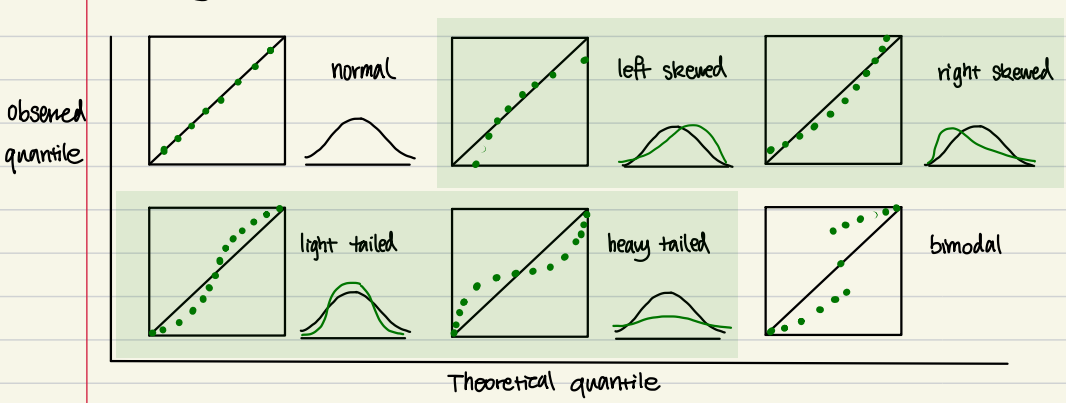
Assumption: $E_i \overset{iid}{\sim} N(0, \sigma^2)$

① independent assumption is impossible to check.

② Residual constant variance.

③ Normality assumption

④ Lack of fit: when see a trend in residual ★ see previous table for this



Estimators

$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i \qquad \varepsilon_i \overset{iid}{\sim} N(0, \sigma^2)$

> $\hat{\beta_0} = \bar{Y} - \hat{\beta_1}\bar{X}$

$SE(\hat{\beta_0}) = \sqrt{MSE\left(\frac{1}{n} + \frac{\bar{X}^2}{\Sigma(X_i - \bar{X})^2}\right)}$

> $\hat{\beta_1} = \frac{r_{xy}}{s_x} s_y = \frac{s_{xy}}{s_x s_y}\frac{s_y}{s_x} = \frac{s_{xy}}{s_{xx}} = \frac{\Sigma_i (X_i - \bar{X})(Y_i - \bar{Y})}{\Sigma_i (X_i - \bar{X})^2}$

$SE(\hat{\beta_1}) = \sqrt{MSE\left(\frac{1}{\Sigma(X_i - \bar{X})^2}\right)}$

think these as $\sigma^2 (X'X)^{-1} = \sigma^2 \begin{pmatrix} \frac{1}{n} + \frac{\bar{X}^2}{\Sigma(X_i - \bar{X})^2} & \cdots \\ \vdots & \frac{1}{\Sigma(X_i - \bar{X})^2} \end{pmatrix}$

> $\hat{Y} = \hat{\beta_0} + \hat{\beta_1} X_0$

$SE(\hat{Y}) = \sqrt{MSE\left(\frac{1}{n} + \frac{(\bar{X} - X_0)^2}{\Sigma(X_i - \bar{X})^2}\right)}$ ← I think how we do NOT assume $\hat{\beta_0} \perp \hat{\beta_1}$

> $\tilde{Y} = \hat{\beta_0} + \hat{\beta_1} X_0 + \varepsilon_{n+1}$

$SE(\tilde{Y}) = \sqrt{MSE\left(1 + \frac{1}{n} + \frac{(\bar{X} - X_0)^2}{\Sigma(X_i - \bar{X})^2}\right)}$ ← see PI in previous pages

Linear Regression v.s. Penalization

(Ordinary) Least Square Regression:

$$\hat{\beta} = \underset{\beta}{\operatorname{argmin}} \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 X_{1i} - \cdots - \beta_p X_{pi})^2$$

> $\hat{\beta}$ may or may not be uniquely determined

> convention: center & scale all $X_{ji} \to (X_{ji} - \bar{X_j})/s_j$

then $\hat{\beta_0} = \bar{y} = \sum_{i=1}^{n} Y_i / n$ really like centering y as well

↳ $Y_i \mapsto Y_i - \bar{Y}$ with no intercept column in X

> $\hat{\beta}^{OLS} = (X'X + \lambda I)^{-1} X'Y$

linear in Y

Ridge Regression

$$\hat{\beta} = \underset{\beta}{\operatorname{argmin}} \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 X_{1i} - \cdots - \beta_p X_{pi})^2 + \lambda \sum_{j=1}^{n} \beta_j^2$$

> Balance: minimize SSE v.s. make length of slope vector close to 0. $\beta_1 \cdots \beta_p$

> Highly correlated predictor variables:

- OLS slope is not well determined, have large variance
- shrinking $\beta_1 \cdots \beta_j$ to close to 0 helps determination

> convention: center & scale all $X_{ji} \to (X_{ji} - \bar{X_j})/s_j$

$Y_i \mapsto Y_i - \bar{Y}$ with no intercept column in X.

> $\hat{\beta}^{ridge} = (X'X + \lambda I)^{-1} X'Y$ ← shrinkage estimator

> Bias $= E(\hat{\beta}^{ridge} - \beta) = -\lambda (X'X + \lambda I)^{-1} \beta$ ~ use Sherman-Morrison Woodbury formula

> $var(\hat{\beta}^{ridge}) = \sigma^2 (X'X + \lambda I)^{-1} (X'X) (X'X + \lambda I)^{-1}$ ↖ biased regression

Lasso Regression

$$\hat{\beta} = \underset{\beta}{\operatorname{argmin}} \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_{1i} - \cdots - \beta_p x_{pi})^2 + \lambda \sum_{j=1}^{p} |\beta_j|$$

> Convention: center & scale all $x_{ji} \to (x_{ji} - \bar{x}_{j\cdot})/s_j$

$y_i \mapsto y_i - \bar{y}$ with no intercept column in X.

> $\hat{\beta}^{lasso}$ has no closed-form as is non-linear.

> Closed form expression s exist in a special case:

→ ① centered & ② scaled predictors, when X has ③ orthonormal columns : $X'X = I$

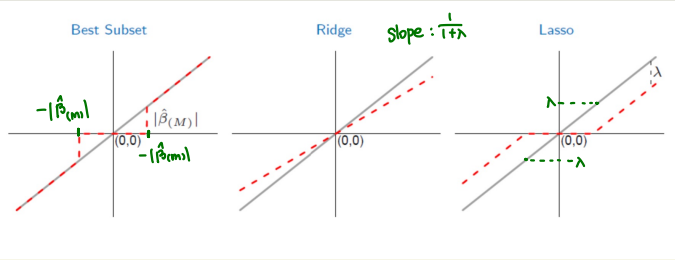
LS(p): $\hat{\beta} = (X'X)^{-1}X'y \Rightarrow \hat{\beta}_j = x_j'y$

LS(k): keep k predictors have the largest impact ← Best Subset

- $\text{Var}(\hat{\beta}_j)$ equals
- order $|\hat{\beta}_{(1)}| \geq \cdots \geq |\hat{\beta}_{(k)}| \geq \cdots \geq |\hat{\beta}_{(p)}|$
- if $|\hat{\beta}_{(j)}| \geq |\hat{\beta}_{(k)}|$ report $\hat{\beta}_{(j)}$; otherwise, report 0.
- $\hat{\beta}_j^{LS(k)} = \hat{\beta}_j \cdot I(|\hat{\beta}_j| \geq |\hat{\beta}_k|)$ ← hard threshold

Ridge(λ): $\hat{\beta}_j^{ridge} = \frac{x_j'y}{1+\lambda} = \frac{\hat{\beta}_j}{1+\lambda}$ ← proportional shrinkage

Lasso(λ): $\hat{\beta}_j^{lasso} = \text{sign}(\hat{\beta}_j)(|\hat{\beta}_j| - \lambda)_+ = \text{sign}(\hat{\beta}_j)\max(|\hat{\beta}_j| - \lambda, 0)$ ← soft thresholding



> Properties

Setting: Multicolinearity or p relatively to n

Must center and scale predictors

- shrinkage applied to partial slopes, not intercept
- scaling matters for predictions $\hat{y}$ (?)

As $\lambda \uparrow$, $\hat{\beta} \to 0$: bias ↑ variance ↓

Ridge v.s. Lasso

Ridge : 1. $\hat{\beta} \to 0$ not exactly 0

(use when most predictors are important)

2. Shrinks to similar coefficient estimates when correlated

Lasso : 1. $\hat{\beta} \to 0$ equals to 0

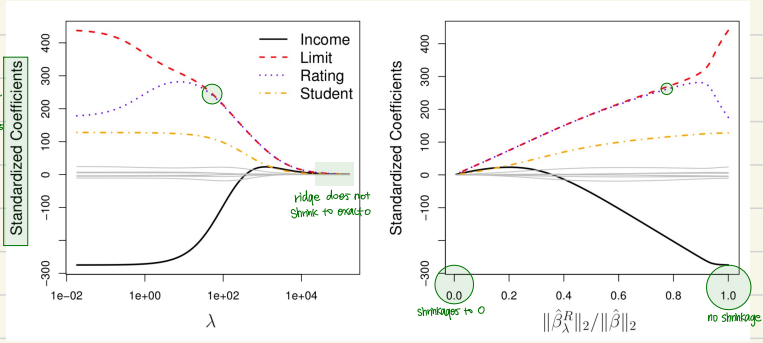
(implicit variable selection)

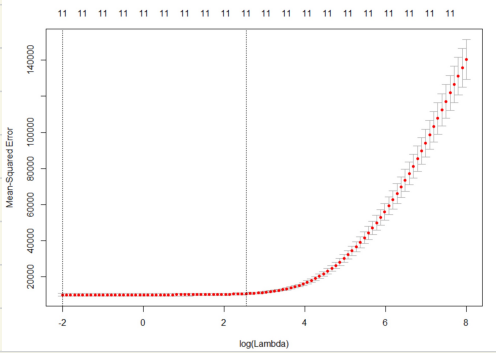
can be misleading and arbitrary + various adaptations exist

2. Picks one and discard others when correlated

* Both useful for prediction
* Both require care when focus is estimation of β.

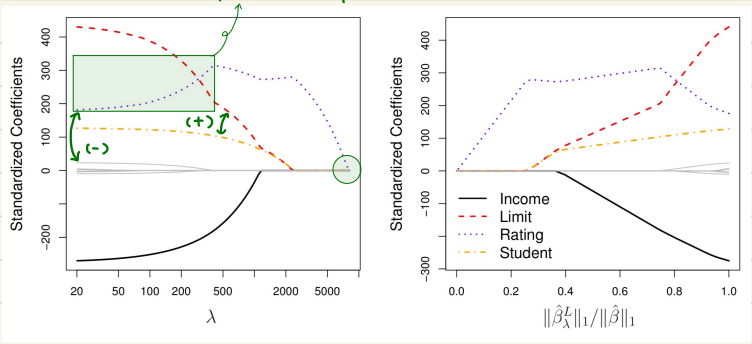
We look at the standardized coefficients because may be hard to see similar but significant coefficients

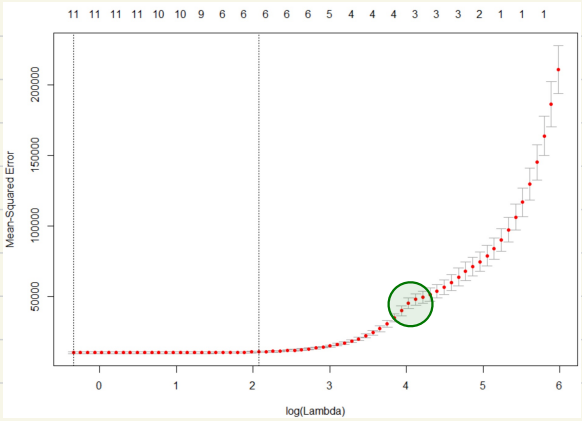




Smooth

Reallocation of correlated features





has kink

Elastic net Regression

normally λ is selected — α is specified

$$\hat{\beta} = \underset{\beta}{\operatorname{argmin}} \; SSE + \lambda\left(\frac{(1-\alpha)}{2}\|\beta\|_2^2 + \alpha\|\beta\|_1\right)$$

Dimension Reduction

Convention: center and scale predictors

1. X-space ⟶ W space (can always be done)
2. Columns of W (scores) = linear combinations of columns of X
3. Columns of W is ordered by relevance
4. Drop irrelavent columns to get W_{reg}-space
5. Perform back conversion $W_{reg} \to X$ to do inference.

Pricipal Component Regression (PCR)

- chooses "scores" without info Y

PCR: "$\boldsymbol{W}$-space" from eigenvalue-eigenvector decomposition of $\boldsymbol{X}'\boldsymbol{X}$

- Eigenvalues $\lambda_1 \geq \cdots \geq \lambda_p$. Corresponding orthogonal eigenvectors $\boldsymbol{v}_1, \ldots, \boldsymbol{v}_p$ as columns of matrix $\boldsymbol{V}$
- $\boldsymbol{X}\boldsymbol{v}_1$ explains the most variation in the $\boldsymbol{X}$-space, namely $\left(\frac{\lambda_1}{\lambda_1+\cdots+\lambda_p}\right)$ 100% of the variation in the $\boldsymbol{X}$-space
- $\boldsymbol{X}\boldsymbol{v}_1$ and $\boldsymbol{X}\boldsymbol{v}_2$ together explain $\left(\frac{\lambda_1+\lambda_2}{\lambda_1+\cdots+\lambda_p}\right)$ 100% of the variation in the $\boldsymbol{X}$-space
- $\boldsymbol{W} = \boldsymbol{X}\boldsymbol{V}$ has p columns, also known as p "principal components"
- The kth principal component is "irrelevant" if its
  - eigenvalue is "small" compared to the largest eigenvalue, i.e., if kth condition index $\delta_k = \sqrt{\frac{\lambda_1}{\lambda_k}}$ is large (say > 10)
  - estimated regression coefficient is statistically zero

Understanding principal components when $p = 2$:

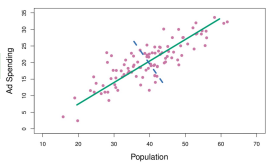


Fig 6.14 from ISLR

$pc_1 = .839 * popn + .544 * ad$
$pc_2 = .544 * popn - .839 * ad$

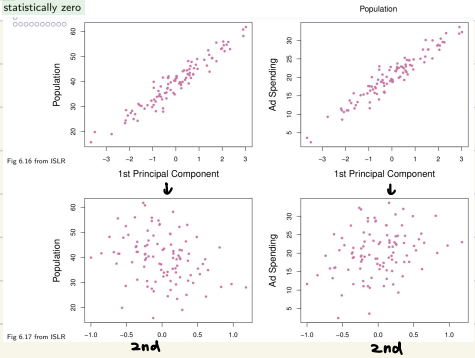


Fig 6.16 from ISLR

Fig 6.17 from ISLR

Partial Least Squares Regression (PLSR)

- choose "scores" with info Y

PLSR: "$\boldsymbol{W}$-space" simultaneously seeks high levels of variation in the $\boldsymbol{X}$-space **and** strong correlation with $\boldsymbol{Y}$

- Does not naturally arise from elegant matrix algebra
- More algorithmic, less theoretical
- Can be quite effective

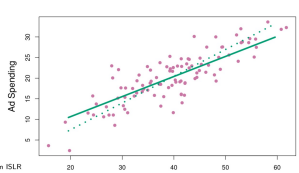


Fig 6.21 from ISLR

Multiple Linear Regression ← weird I don't see much interpretations, but typically

Fitted value: $\hat{y}_i = \hat{\beta}_0 + \sum_j \hat{\beta}_j x_{ij} = x_i^T \hat{\beta}$ are normal x^2 coefficient means curvature.

$$E(\hat{y}_i \mid x_i) = x_i^T \beta$$

$$\text{Var}(\hat{y}_i \mid x_i) = \sigma^2 x_i'(x'x)^{-1} x_i \quad \begin{matrix} < \sigma^2 \\ > 0 \end{matrix}$$

↳ $h_{ii} = x_i^T (x'x)^{-1} x_i$ leverage (≤ 1 for models with intercept)

Fitted vector: $\hat{y} = X\hat{\beta} = (x'x)^{-1}x'y = Hy$

$$E(\hat{y} \mid x) = X\beta$$

$$\text{Var}(\hat{y} \mid x) = \sigma^2 H$$

Residue vector: $r = y - \hat{y} = y - Hy = (I-H)y$

Fact 1: $\sum_i r_i = 0$ ∀ models with intercept

Fact 2: Residuals are always correlated

- $\text{var}(r) = \sigma^2(I-H)$
- complicates abilities to check independence assumption

$$SSE = (y-\hat{y})'(y-\hat{y}) = y'(I-H)y$$

Example: Let $x_0^T = (1 \; x_1 \; x_2 \cdots x_p)$ be values of prediction

Prediction value: $\hat{y}_0 = \hat{\beta}_0 + \sum_j \hat{\beta}_j x_j = x_0' \hat{\beta}$ is normal

$$E(\hat{y}_0 \mid x_0) = x_0^T \beta$$

$$\text{Var}(\hat{y}_0 \mid x_0) = \sigma^2 x_0'(x'x)^{-1} x_0$$

CI $x_0'\beta$: $\hat{y} \pm t_{\alpha/2, n-p-1} \sqrt{MSE} \sqrt{x_0'(x'x)^{-1}x_0}$

PI $x_0'\beta$: $\hat{y} \pm t_{\alpha/2, n-p-1} \sqrt{MSE} \sqrt{1 + x_0'(x'x)^{-1}x_0}$

Coefficient of Determination: $R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST}$

* However, R^2 always increases as adding predictors.

$$\text{adjusted } R^2 = 1 - \frac{SSE/(n-p-1)}{SST/(n-1)} \quad \downarrow \uparrow \downarrow$$

$$= 1 - (1-R^2)\frac{n-1}{n-p-1}$$

$$< R^2 \text{ always}$$

* Adding predictor that does not increases R^2 causing adj R^2 to decrease.

* However, adj R^2 is NOT great for small dataset.

Model Specification

Over-specification

True Model: $E(Y_i | X = x_i) = \sum_{j=0}^{P_1} \theta_j x_{ij} + \sum_{j=P_1+1}^{P-P_1-1} 0 \cdot x_{ij}$ OR $E(y | X_1) = X_1\theta_1$

Oracle estimator: $\hat{\theta}_1 = (X_1'X_1)^{-1}X_1'y$

$E(\hat{\theta}_1) = \theta_1$

$\text{var}(\hat{\theta}_1) = \sigma^2(X_1'X_1)^{-1}$

$E(MSE) = \sigma^2 \Rightarrow n - (1+P_1)$ DF

Assumed Model:

$E(Y_i | X = x_i) = \sum_{j=0}^{P_1} \theta_j x_{ij} + \sum_{j=P_1+1}^{P-P_1-1} \theta_j x_{ij}$ OR $E(y | X_1, X_2) = X_1\theta_1 + X_2\theta_2 = X\theta$

Estimator: $\hat{\theta} = (X'X)^{-1}X'y$

$E(\hat{\theta}) = (X'X)^{-1}X'X_1\theta_1 = \begin{pmatrix} \theta_1 \\ 0 \end{pmatrix}$

$\text{Var}(\hat{\theta}) = \sigma^2(X'X)^{-1} > \sigma^2(X_1'X_1)^{-1}$

$E(MSE) = \sigma^2 \Rightarrow n-(1+p)$ DF $\Leftarrow$ dfError ↓ F statistics ↑ Power ↓

Summary: Both Type I Error & Type II Error are correct.

Under over-specification, correct variables have $\hat{SE}$ ↑ p-value ↑. $\Leftarrow$ CI ↑

Over-specification causes problems as well:

Inflating Estimator Variance:

For assumed model (with p+1 parameters):

$\text{var}(\hat{\beta}_j) = \frac{\sigma^2}{\sum_i (x_{ij} - \bar{x}_j)^2} \times \frac{1}{1 - R_j^2}$ → R-squared value treating x_j as response and other predictors as covariates.

↓ variance for SLR with x_j

$R_j^2 \to 1$, $\text{var}(\hat{\beta}_j) \to \infty$

Variance inflation factor (VIF): $\frac{1}{1 - R_j^2}$

VIF > 10 means predictors are highly correlated, complicating ability to identify which variables are important.

① model is overly complicated

② the predictors are naturally correlated

1. Global F-test p-value ↓ but individual p-value ↑
2. Normally when colinearity happens, VIF ↑ p-value ↑

Under-specification

True Model: $E(Y_i | X = X_i) = \sum_{j=0}^{P_1} \theta_j X_{ij} + \sum_{j=P_1+1}^{P-P_1-1} \theta_j X_{ij}$ OR $E(y | X_1, X_2) = X_1\theta_1 + X_2\theta_2 = X\theta$

Oracle estimator: $\hat{\theta} = (X'X)^{-1}X'y$

$E(\hat{\theta}) = \theta$

$Var(\hat{\theta}) = \sigma^2(X'X)^{-1}$

$E(MSE) = \sigma^2 \Rightarrow n - (1+P)$ DF

Assumed Model:

$E(Y_i | X = x_i) = \sum_{j=0}^{P_1} \theta_j X_{ij} + \sum_{j=P_1+1}^{P-P_1-1} 0 \cdot X_{ij}$ OR $E(y | X_1) = X_1\theta_1$

Estimator: $\hat{\theta}_1 = (X_1'X_1)^{-1}X_1'y$

$E(\hat{\theta}_1) = (X_1'X_1)^{-1}X_1'(X_1\theta_1 + X_2\theta_2) = \theta_1 + (X_1'X_1)^{-1}X_1'X_2\theta_2$

unbiased when X_1, X_2 are orthonormal

$Var(\hat{\theta}_1) = \sigma^2(X_1'X_1)^{-1}$

$E(MSE) = \sigma^2 + Q(X_2, \theta_2) > \sigma^2 \Rightarrow n - (1+P_1)$ DF $<$ DF

Summary: Hypothesis Test Type I & Type II errors unpredictable. $\Leftarrow$ CI x

Replication and lack-of-fit can detect underspecified model.

pure error may be used to instead of σ^2. since MSpureE always unbiased σ^2.

Linear Mixed Model

1. In linear models, we consider r.v. ε_i account for the variability in the response not explained by the predictors.
2. $\hat{\varepsilon}_i = y_i - \hat{\mu}_i$

   $\hat{\varepsilon} = (I - X(X'X)^{-}X')y \sim N(0, (I - X(X'X)^{-}X')\sigma^2)$
3. Model formula in "R": `y ~ a*b + x:z + I(v^2) - 1`

   `a*b` : `a + b + a:b`

   `x:z` : interaction of x and z

   `I(v^2)` : linear predictor depends on v^2.

   `-1` : no intercept

   Model : $y = a + b + ab + xz + v^2$
4. Kronecker product notation:

$$\begin{pmatrix} 1_3 & & \\ & 1_3 & \\ & & 1_3 \end{pmatrix} = I_3 \otimes 1_3 \qquad \begin{pmatrix} 1_3 & 0 \\ 0 & 1_3 \\ 1_3 & 0 \\ 0 & 1_3 \end{pmatrix} = 1_2 \otimes (I_2 \otimes 1_3)$$

5. Linear Mixed Model:

   When estimating σ^2_A, Type 3: used to see the tests, but $\hat{\sigma}^2_A < 0$ with MSR < MSE

   ML: Provides $\hat{\sigma}^2_A \geq 0$, biased

   $\hat{\sigma}^2_A(ML) = \frac{1}{n}(y - X\hat{\beta})'(y - X\hat{\beta})$ negatively biased

   REML: Provides $\hat{\sigma}^2_A \geq 0$, $E(\hat{\sigma}^2_A) = \sigma^2_A$

   $\hat{\sigma}^2_A(REML) = \frac{1}{n-p}(y - X\hat{\beta})'(y - X\hat{\beta})$ unbiased

   In SAS: proc mixed

   In R: `lme()` for nested

   `lmer()` for nested & crossed

   `(1|a)` : random effect for a

   `(1|a:b)` : random effect for ab interaction (slope)

   } see following pages

Prediction

Two varieties ...

- Marginal, aka population-averaged:
  $\hat{Y} = X\hat{\beta}$
- Conditional, aka cluster-specific:
  $\hat{Y} = X\hat{\beta} + Z\hat{\alpha}$ where $\hat{\alpha}$ is best linear unbiased predictor
  i.e., $\hat{\alpha}$ is conditional mean of α given observed y

$$\hat{\alpha} = \hat{G}Z^T\hat{V}^{-1}(y - X\hat{\beta})$$

BLUP property relies on known G and R

$\alpha \sim N(0, G)$ notation in this case

$\varepsilon \sim N(0, R)$

$Cov(\alpha, \varepsilon) = 0$

lmer(·) usage and output is crucial

Model specification

The following formula extensions for specifying random-effects structures in R are used by

- `lme4`
- `nlme` (nested effects only, although crossed effects can be specified with more work)
- `glmmADMB` and `glmmTMB`

`MCMCglmm` uses a different specification, inherited from AS-REML.

(Modified from Robin Jeffries, UCLA:)

formula	meaning
`(1|group)`	random group intercept
`(x|group) = (1+x|group)`	random slope of x within group with correlated intercept
`(0+x|group) = (-1+x|group)`	random slope of x within group: no variation in intercept
`(1|group) + (0+x|group)`	uncorrelated random intercept and random slope within group
`(1|site/block) = (1|site)+(1|site:block)`	intercept varying among sites and among blocks within sites (nested random effects)
`site+(1|site:block)`	*fixed* effect of sites plus random variation in intercept among blocks within sites
`(x|site/block) = (x|site)+(x|site:block) = (1 + x|site)+(1+x|site:block)`	slope and intercept varying among sites and among blocks within sites
`(x1|site)+(x2|block)`	two different effects, varying at different levels
`x*site+(x|site:block)`	fixed effect variation of slope and intercept varying among sites and random variation of slope and intercept among blocks within sites
`(1|group1)+(1|group2)`	intercept varying among crossed random effects (e.g. site, year)

Covered in lecture note

+ "random group"

+ "random intercept /subject = group"

+ "random block(site)"

Consider under Aitken model, want to test on $\hat{\beta}$ but notice $\varepsilon \sim N(0, V)$

1. If V is known, $H_0: A\beta = m$ inference is exact.
2. If $V = \sigma^2 D$ with D known, use REML method
3. If $V = \sigma^2 D$ with both σ^2, D unknown, use Type III + Satterthwaite / Kenward Roger find df

Situation	Estimation	Inference Method
V known	GLS / Aitken	Exact F-test / Type III
$V = \sigma^2 D$, D known	REML (for σ^2)	Type III with known D (?)
$V = \sigma^2 D$, both unknown	REML / ML	Type III + Satterthwaite / Kenward Roger

↑ REML for $\hat{\sigma}^2$

↑ Type III used for testing

Example: Fabric

Two types of yarn are used to weave fabric on looms. In our shop, we have three operators and we randomly select three of our many looms. We will have each operator run each yarn type on each loom twice. The order of these runs will be at random.

The operators are the only three we have who are qualified for this type of weaving. We are interested in comparing them. The yarn types are the only two available for this particular application and are different in the type of fiber used. The looms, on the other hand, are of no particular interest and are thought to be representative of the collection of looms that might be used for this type of application.

The data are average puncture resistance measurements made on five randomly selected spots on the cloth. Here are the data:

Loom	1	1	1	2	2	2	3	3	3
Oper	1	2	3	1	2	3	1	2	3
Yarn 1	59, 52	43, 42	48, 54	88, 86	68, 71	83, 82	50, 43	35, 30	42, 41
Yarn 2	81, 73	54, 56	59, 68	99, 98	82, 88	98, 94	63, 62	49, 40	60, 56

```
proc mixed data=fabric; class loom operator yarn;
  model resistance=operator|yarn / solution;
  random loom loom*operator loom*yarn loom*operator*yarn;
```

same model, diffrent method

```
proc mixed data=fabric method=ml; class loom operator yarn;
  model resistance=operator|yarn / solution;
  random loom loom*operator loom*yarn loom*operator*yarn;

proc mixed data=fabric method=type3; class loom operator yarn;
  model resistance=operator|yarn / solution;
  random loom loom*operator loom*yarn loom*operator*yarn;

proc mixed data=fabric method=type3; class loom operator yarn;
  model resistance=operator|yarn / solution ddfm=satterth;
  random loom loom*operator loom*yarn loom*operator*yarn;

proc mixed data=fabric method=type3; class loom operator yarn;
  model resistance=operator|yarn / solution ddfm=kenwardroger;
  random loom loom*operator loom*yarn loom*operator*yarn;
```

need this adjustment for $\hat{\sigma}^2$ unknown test

REML: $\widehat{\sigma}^2_L = 406.43$, $\widehat{\sigma}^2_{LO} = 0$, $\widehat{\sigma}^2_{LY} = 0$, $\widehat{\sigma}^2_{LOY} = 0$, $\widehat{\sigma}^2 = 12.06$

ML: $\widehat{\sigma}^2_L = 270.84$, $\widehat{\sigma}^2_{LO} = 0$, $\widehat{\sigma}^2_{LY} = 0$, $\widehat{\sigma}^2_{LOY} = 0$, $\widehat{\sigma}^2 = 10.23$

type3: $\widehat{\sigma}^2_L = 407.90$, $\widehat{\sigma}^2_{LO} = -1.60$, $\widehat{\sigma}^2_{LY} = -1.97$, $\widehat{\sigma}^2_{LOY} = 0.875$, $\widehat{\sigma}^2 = 13.19$

type3 w/ Satterthwaite or Kenward-Roger:

$\widehat{\sigma}^2_L = 407.90$, $\widehat{\sigma}^2_{LO} = -1.60$, $\widehat{\sigma}^2_{LY} = -1.97$, $\widehat{\sigma}^2_{LOY} = 0.875$, $\widehat{\sigma}^2 = 13.19$

Effect	ndf	ddf	F value	Pr>F	Param	Estim	SE	df	Pr> $\lvert t \rvert$
Oper	2	4	68.31	.0008	$O_1 - O_3$	6.83	2.00	4	.0271
			80.51	.0006			1.85	4	.0208
			96.59	.0004			1.98	4	.0259
			96.59	.0004			1.98	7.44	.0096
Yarn	1	2	159.34	.0062	$O_2 - O_3$	−11.0	2.00	4	.0054
			187.80	.0053			1.85	4	.0040
			617.58	.0016			1.98	4	.0051
			617.58	.0016			1.98	7.44	.0007
O*Y	2	4	0.60	.5932					
			0.70	.5474					
			0.48	.6496					
			0.48	.6496					

effects only has the same F value as type III (as $\hat{\beta}$ the same with df the same)

Computing . . . SAS: proc glimmix

Example **[Corn]**: Consider an experiment to study the effects of 4 planting methods and 3 pesticides on yields of corn. Each pesticide is applied by aerial spraying to 2 large fields. Each large field is subdivided into 4 subplots, and the 4 planting methods are randomly assigned to the 4 subplots. $3 \times 2 = 6$ fields total

- Completely randomized split-plot design
- Y_{ijk} = yield of jth planting method within kth field receiving ith pesticide

$$Y_{ijk} = \mu + \alpha_i + \gamma_{k(i)} + \beta_j + (\alpha\beta)_{ij} + \epsilon_{ijk}, \quad \gamma_{k(i)} \sim^{iid} N(0, \sigma^2_\gamma),$$
$$\epsilon_{ijk} \sim^{iid} N(0, \sigma^2), \quad Cov(\gamma_{k(i)}, \epsilon_{ijk}) = 0,$$
$$\rightsquigarrow Var(Y_{ijk}) = \sigma^2_\gamma + \sigma^2, \quad Cov(Y_{ijk}, Y_{i\ell k}) = \sigma^2_\gamma$$

(field, pesticide, method, nested)

```
proc glimmix data=corn;
  class pesticide field method;
  model yield=pesticide|method / solution;
  random field(pesticide);

proc glimmix data=fabric;
  class loom operator yarn;
  model resistance=operator|yarn / solution;
  random loom loom*operator loom*yarn loom*operator*yarn;
```

** Results for FABRIC differ (a little) between `mixed` and `glimmix` **

Computing . . . R

- Function `lme()` in package `nlme` is very popular, but is limited to nested random effects.
- Function `lmer()` in package `lme4` can handle both nested and crossed random effects. It is also newer and better able to deal with large datasets.

But more structure and flexibility in specifying matrices $\boldsymbol{G}$ and $\boldsymbol{R}$ is offered by `lme()` compared to `lmer()`.

This is nested, I think this is because: there is no outside random factor: field.

```
fit.corn = lmer(yield ~ pesticide * method + (1|field:pesticide), data=corn)
```

The interaction cannot stand by itself

```
fit.fabric = lmer(resistance ~ operator * yarn + (1|loom) +
    (1|loom:operator) + (1|loom:yarn)+ (1|loom:operator:yarn), data=fabric)
```

Sleep Study

Interested in **average reaction time** for sleep-deprived subjects.

- Before day 0, subjects had their normal amount of sleep.
- On following nights, subjects were limited to 3 hours of sleep.
- Subjects were given a series of tests on each day.
- This dataset is balanced with no missing observations.

`Reaction` is average reaction time (ms) on a series of tests given each `Days` (values 0,1,...,9) to each `Subject` (there are 18 of them).

Y_{ij}: reaction time for subject i on day j

X_{ij}: day j for subject i on day j

$$E(Y_{ij}) = \beta_0 + \beta_1 X_{ij} \qquad \text{expect } \beta_1 > 0$$

But different people may respond differently ...

$$E(Y_{ij}) = \beta_{0i} + \beta_{1i} X_{ij}$$

But different people may respond differently . . .

$$E(Y_{ij}) = \beta_{0i} + \beta_{1i} X_{ij}$$

- $\beta_{0i},\ \beta_{1i}$ as fixed effects limits us to inference on these 18 subjects
- $\beta_{0i},\ \beta_{1i}$ as random effects allows inference on a broader scale

$$\begin{aligned} \beta_{0i} &\equiv \beta_0 + \alpha_{0i}, & \alpha_{0i} &\sim^{ind} N(0, \sigma_0^2) \\ \beta_{1i} &\equiv \beta_1 + \alpha_{1i}, & \alpha_{1i} &\sim^{ind} N(0, \sigma_1^2) \end{aligned}$$

$$Y_{ij} = \beta_0 + \beta_1 X_{ij} + \alpha_{0i} + \alpha_{1i} X_{ij} + \epsilon_{ij}$$

for different person, effect different

Called the "random intercepts, random slopes" model.

Independently over i, for $i = 1, \ldots, N = 18$,

$$\boldsymbol{Y}_i = \boldsymbol{X}_i \boldsymbol{\beta} + \boldsymbol{Z}_i \boldsymbol{\alpha}_i + \boldsymbol{\epsilon}_i, \qquad \boldsymbol{\alpha}_i \sim N(\boldsymbol{0}, \boldsymbol{D}), \qquad \boldsymbol{\epsilon}_i \sim N(\boldsymbol{0}, \boldsymbol{R}_i), \qquad \boldsymbol{\alpha}_i, \boldsymbol{\epsilon}_i \text{ indep}$$

Clustered Data / Repeated Measures / Longitudinal Data

"Single-level LMM"

Independently over i, for $i = 1, \ldots, N$,

$$\boldsymbol{Y}_i = \boldsymbol{X}_i\beta + \boldsymbol{Z}_i\boldsymbol{\alpha}_i + \boldsymbol{\epsilon}_i, \qquad \boldsymbol{\alpha}_i \sim N(\boldsymbol{0}, \boldsymbol{D}), \qquad \boldsymbol{\epsilon}_i \sim N(\boldsymbol{0}, \boldsymbol{R}_i), \qquad \boldsymbol{\alpha}_i, \boldsymbol{\epsilon}_i \text{ indep}$$

- $\boldsymbol{Y}_i$ is $n_i \times 1$ observed response vector for ith cluster
- $\boldsymbol{X}_i$ is $n_i \times p$ known covariate matrix for ith cluster
- β is $p \times 1$ unknown fixed effects
- $\boldsymbol{Z}_i$ is $n_i \times q$ known matrix for ith cluster
- $\boldsymbol{\alpha}_i$ is $q \times 1$ random, unobserved, for ith cluster

may not involve all the features in the random effect.

$$\boldsymbol{Y} = \begin{bmatrix} \boldsymbol{Y}_1 \\ \vdots \\ \boldsymbol{Y}_N \end{bmatrix}, \boldsymbol{X} = \begin{bmatrix} \boldsymbol{X}_1 \\ \vdots \\ \boldsymbol{X}_N \end{bmatrix}, \boldsymbol{Z} = \begin{bmatrix} \boldsymbol{Z}_1 & & \boldsymbol{0} \\ & \ddots & \\ \boldsymbol{0} & & \boldsymbol{Z}_N \end{bmatrix}, \boldsymbol{\alpha} = \begin{bmatrix} \boldsymbol{\alpha}_1 \\ \vdots \\ \boldsymbol{\alpha}_N \end{bmatrix}, \boldsymbol{\epsilon} = \begin{bmatrix} \boldsymbol{\epsilon}_1 \\ \vdots \\ \boldsymbol{\epsilon}_N \end{bmatrix}, \boldsymbol{R} = \begin{bmatrix} \boldsymbol{R}_1 & & \boldsymbol{0} \\ & \ddots & \\ \boldsymbol{0} & & \boldsymbol{R}_N \end{bmatrix}$$

$$\boldsymbol{G} = \boldsymbol{I}_N \otimes \boldsymbol{D}, \qquad n = \sum_{i=1}^{N} n_i, \qquad \boldsymbol{V} = \text{diag}\{\boldsymbol{V}_1, \ldots, \boldsymbol{V}_N\}, \qquad \boldsymbol{V}_i = \boldsymbol{Z}_i\boldsymbol{D}\boldsymbol{Z}_i^T + \boldsymbol{R}_i$$

Single-level LMM: N groups (indexed $i = 1, \ldots, N$) each containing n_i observations

$$\boldsymbol{Y}_i = \boldsymbol{X}_i\boldsymbol{\beta} + \boldsymbol{Z}_i\boldsymbol{\alpha}_i + \boldsymbol{\epsilon}_i, \qquad \boldsymbol{\alpha}_i \sim N(\boldsymbol{0}, \boldsymbol{D}), \qquad \boldsymbol{\epsilon}_i \sim N(\boldsymbol{0}, \boldsymbol{R}_i), \qquad \boldsymbol{\alpha}_i, \boldsymbol{\epsilon}_i \text{ indep}$$

Two-level LMM: N first-level groups (indexed $i = 1, \ldots, N$) each with n_i second-level groups (indexed $j = 1, \ldots, n_i$) containing n_{ij} observations

$$\boldsymbol{Y}_{ij} = \boldsymbol{X}_{ij}\boldsymbol{\beta} + \boldsymbol{Z}_{1,ij}\boldsymbol{\alpha}_i + \boldsymbol{Z}_{2,ij}\boldsymbol{\alpha}_{ij} + \boldsymbol{\epsilon}_{ij}, \qquad \boldsymbol{\alpha}_i \sim N_{q_1}(\boldsymbol{0}, \boldsymbol{D}_1), \quad \boldsymbol{\alpha}_{ij} \sim N_{q_2}(\boldsymbol{0}, \boldsymbol{D}_2),$$

$$\boldsymbol{\epsilon}_{ij} \sim N_{n_{ij}}(\boldsymbol{0}, \boldsymbol{R}_{ij}), \qquad \boldsymbol{\alpha}_i, \boldsymbol{\alpha}_{ij}, \boldsymbol{\epsilon}_{ij} \text{ indep}$$

- Single-level and two-level LMMs are great for **nested** groupings. Not for **crossed** random effects.
- Not happy with assuming vectors of random effects are independent? Use "extended LMM" or "hierarchical specification":

$$\text{Replace} \qquad \boldsymbol{\alpha}_i, \boldsymbol{\epsilon}_i \text{ indep} \qquad \text{with} \qquad \boldsymbol{\epsilon}_i | \boldsymbol{\alpha}_i \sim N(\boldsymbol{0}, \boldsymbol{R}_i)$$

Sleep Study: Software – R

```
> str(sleepstudy)
'data.frame': 180 obs. of  3 variables:
$ Reaction: num  250 259 251 321 357 ...
$ Days    : num  0 1 2 3 4 5 6 7 8 9 ...
$ Subject : Factor w/ 18 levels "308","309","310",..: 1 1 1 1 1 1 1 1 1 1 ...
> plot(sleepstudy)
> xyplot(Reaction~Days|Subject,sleepstudy)
```

$Y_i = \beta_0 + \beta_1 Day_i + a_{0i} + a_{1i} Day_i + \varepsilon_i \quad ; \; a_{0i} \perp \varepsilon_i \; ; \; a_{1i} \perp \varepsilon_i$

```
fit.sleep = lmer(Reaction ~ Days + (Days|Subject) , sleepstudy)
                                  #includes random intercept; allows correlated REs
summary(fit.sleep)
```

$Y_i = \beta_0 + \beta_1 Day_i + a_{0i} + a_{1i} Day_i + \varepsilon_i \quad a_{0i} \perp a_{1i} \perp \varepsilon_i$

```
fit.sleep = lmer(Reaction ~ Days + (Days||Subject) , sleepstudy)
                                  #includes random intercept; REs uncorrelated
summary(fit.sleep)
```

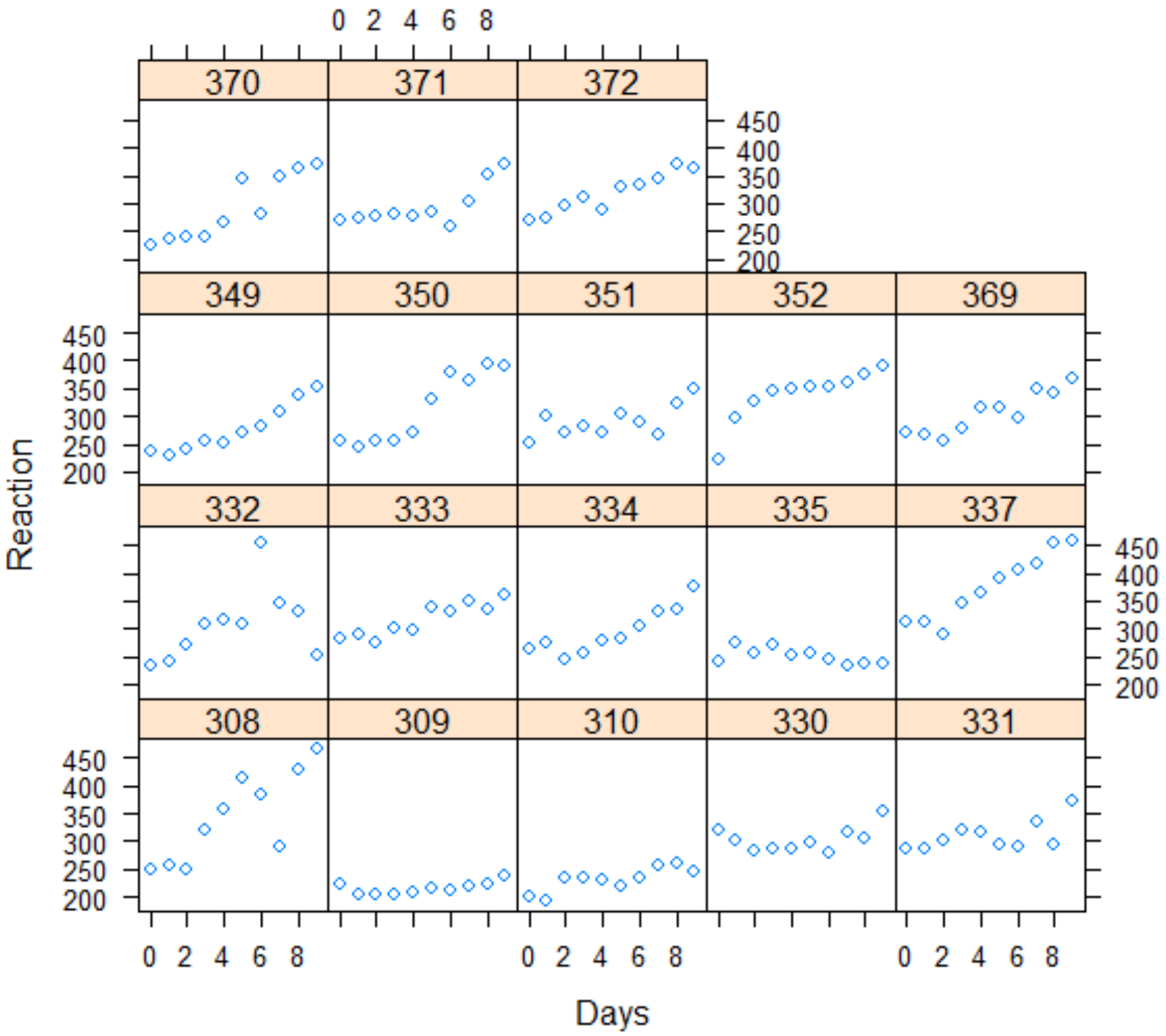
$Y_i = \beta_0 + \beta_1 Day_i + a_{0i} + a_{1i} Day_i + \varepsilon_i \quad a_{0i} \perp a_{1i} \perp \varepsilon_i$

```
fit.sleep = lmer(Reaction ~ Days + (1|Subject) + (0 + Days|Subject), sleepstudy)
                                  #REs uncorrelated
summary(fit.sleep)
```

$Y_i = \beta_0 + \beta_1 Day_i + a_{0i} + a_{1i} Day_i + \varepsilon_i \quad a_{0i} \perp a_{1i} \perp \varepsilon_i$

```
fit.sleep = lmer(Reaction ~ Days + (1|Subject) + (-1 + Days|Subject), sleepstudy)
                                  #REs uncorrelated
summary(fit.sleep)
```

those three were equivalent.


370
371
372
349
350
351
352
369
332
333
334
335
337
308
309
310
330
331
Reaction
Days
0 2 4 6 8
200 250 300 350 400 450

```
> fit.sleep = lmer(Reaction ~ Days + (Days|Subject) , sleepstudy)
> summary(fit.sleep)
Linear mixed model fit by REML ['lmerMod']
Formula: Reaction ~ Days + (Days | Subject)
   Data: sleepstudy

REML criterion at convergence: 1743.6

Scaled residuals:
    Min      1Q  Median      3Q     Max
-3.9536 -0.4634  0.0231  0.4634  5.1793

Random effects:
 Groups   Name        Variance Std.Dev. Corr
 Subject  (Intercept) 612.09   24.740
          Days         35.07    5.922   0.07
 Residual             654.94   25.592
Number of obs: 180, groups:  Subject, 18

Fixed effects:
            Estimate Std. Error t value
(Intercept)  251.405      6.825  36.838
Days          10.467      1.546   6.771

Correlation of Fixed Effects:
     (Intr)
Days -0.138
```

comparison between models fitted by REML, lower is better.

standardized residuals: showing the distribution of model errors (should ideally be symmetric & centered at 0)

allow both intercept and slope varies by group: $+ A_{0i} + A_{1i}X_{ij}$

subject differs in their baseline reaction time.

subject differ in how much Days affect their Reaction

between intercept and slope

expected reaction time on Day 0 (baseline)

Average Y change by each additional day.

variance not explained by either fixed or random effect.

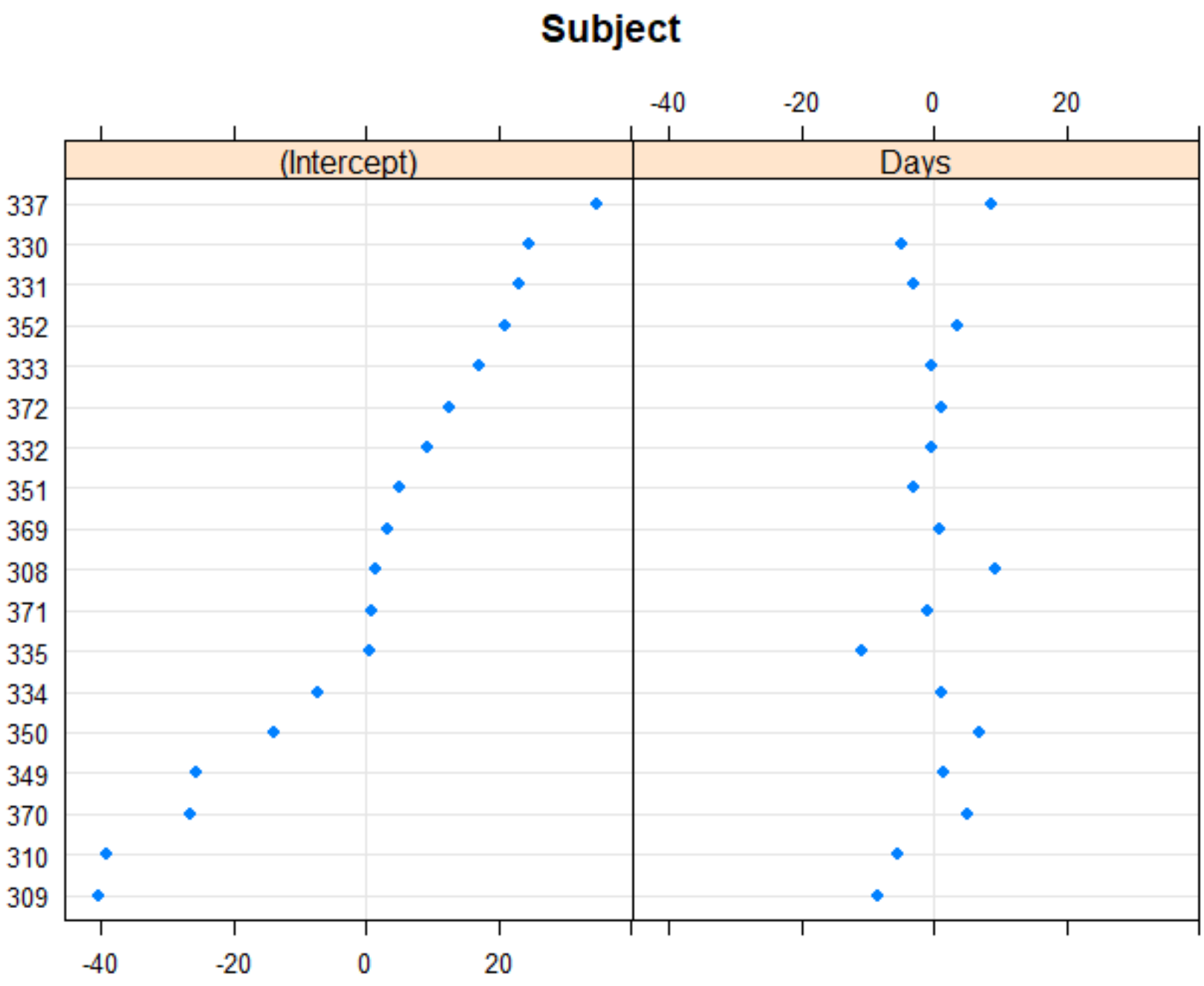
for model diagnostic

$$Y_i = \beta_0 + \beta_1 X_i + a_{0i} + a_{1i} X_i$$

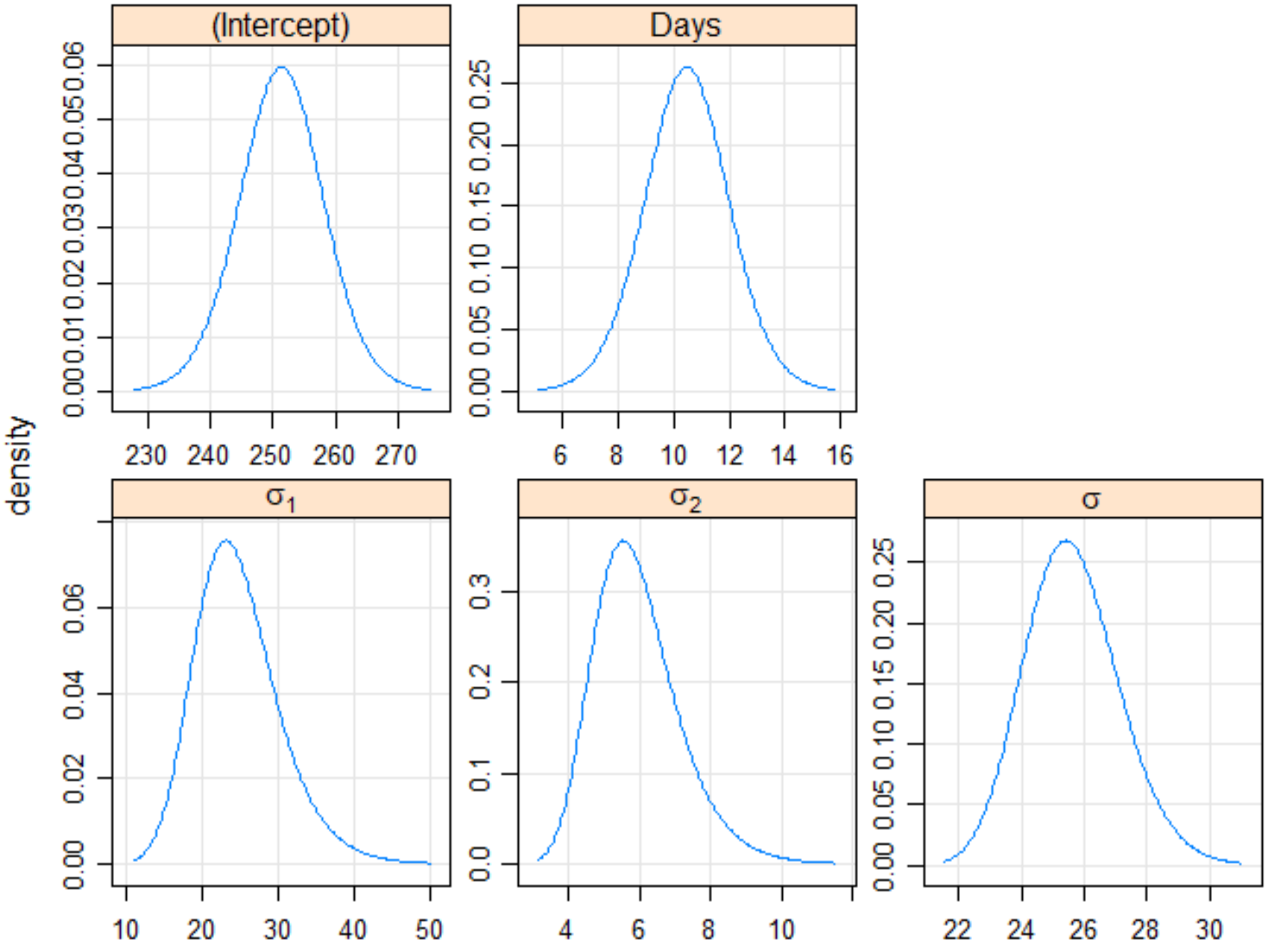
```
> fit.sleep = lmer(Reaction ~ Days + (Days||Subject) , sleepstudy)
> summary(fit.sleep)
Linear mixed model fit by REML ['lmerMod']
Formula: Reaction ~ Days + ((1 | Subject) + (0 + Days | Subject))
   Data: sleepstudy

REML criterion at convergence: 1743.7

Scaled residuals:
    Min      1Q  Median      3Q     Max
-3.9626 -0.4625  0.0204  0.4653  5.1860

Random effects:
 Groups    Name        Variance Std.Dev.
 Subject   (Intercept) 627.57   25.051
 Subject.1 Days         35.86    5.988
 Residual              653.58   25.565
Number of obs: 180, groups:  Subject, 18

Fixed effects:
            Estimate Std. Error t value
(Intercept)  251.405      6.885  36.513
Days          10.467      1.560   6.712

Correlation of Fixed Effects:
(Intr)
Days -0.184
```

Correlation now not estimated

naming convention only: now two separate random effects




```
> dotplot(ranef(fit.sleep))
```



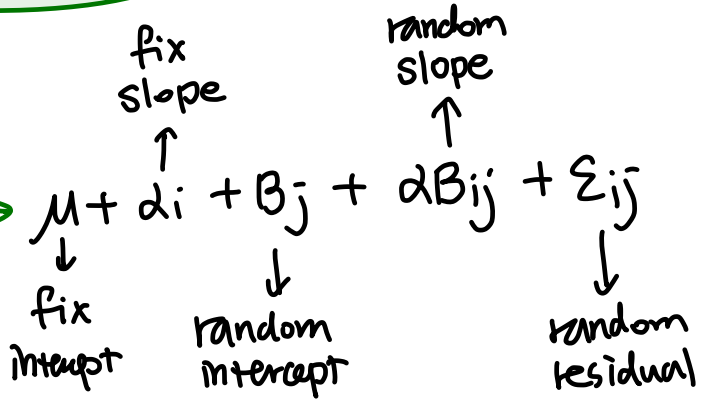

```
> fitsleep.profile = profile(fit.sleep); densityplot(fitsleep.profile)
```

Software – SAS

Recall R command

```
> fit.sleep = lmer(Reaction ~ Days + (Days||Subject) , sleepstudy)
```

The following yield equivalent results . . .




```
proc mixed data=sleep covtest cl;
  class subject;
  model reaction=days / solution cl;
  random subject days*subject / solution;
proc mixed data=sleep covtest cl;
  class subject;
  model reaction=days / solution cl;
  random intercept days / subject=subject solution;
proc glimmix data=sleep;
  class subject;
  model reaction=days / solution cl;
  random subject days*subject / solution;
proc glimmix data=sleep;
  class subject;
  model reaction=days / solution cl;
  random intercept days / subject=subject solution;
```

Correlated $\boldsymbol{R}$ & $\boldsymbol{G}$ Matrices

`lme4::lmer`: $\boldsymbol{R} = \sigma^2\boldsymbol{I}$ only; $\boldsymbol{G}$ diagonal (use || instead of |) or completely unstructured (use | instead of ||)

`proc mixed`: $\boldsymbol{R}$ controlled by `repeated` stmt; $\boldsymbol{G}$ controlled by `random` stmt

`proc glimmix`: $\boldsymbol{R}$ controlled by `random` stmt, with `_residual_` before "/"; $\boldsymbol{G}$ controlled by `random` stmt

---

- Autoregressive of order one, i.e., AR(1) `type=ar(1)`

$$\boldsymbol{R} = \sigma^2 \begin{bmatrix} 1 & \rho & \rho^2 & \rho^3 \\ \rho & 1 & \rho & \rho^2 \\ \rho^2 & \rho & 1 & \rho \\ \rho^3 & \rho^2 & \rho & 1 \end{bmatrix}$$

- Compound symmetry `type=cs`

$$\boldsymbol{R} = \sigma^2 \begin{bmatrix} 1 & \rho & \rho & \rho \\ \rho & 1 & \rho & \rho \\ \rho & \rho & 1 & \rho \\ \rho & \rho & \rho & 1 \end{bmatrix}$$

- Toeplitz, two bands `type=toep(2)`

$$\boldsymbol{R} = \sigma^2 \begin{bmatrix} 1 & \rho_1 & 0 & 0 \\ \rho_1 & 1 & \rho_1 & 0 \\ 0 & \rho_1 & 1 & \rho_1 \\ 0 & 0 & \rho_1 & 1 \end{bmatrix}$$

- Toeplitz, three bands `type=toep(3)`

$$\boldsymbol{R} = \sigma^2 \begin{bmatrix} 1 & \rho_1 & \rho_2 & 0 \\ \rho_1 & 1 & \rho_1 & \rho_2 \\ \rho_2 & \rho_1 & 1 & \rho_1 \\ 0 & \rho_2 & \rho_1 & 1 \end{bmatrix}$$

- Unstructured `type=un`

$$\boldsymbol{R} = \begin{bmatrix} \sigma_1^2 & \sigma_{12} & \sigma_{13} & \sigma_{14} \\ \sigma_{12} & \sigma_2^2 & \sigma_{23} & \sigma_{24} \\ \sigma_{13} & \sigma_{23} & \sigma_3^2 & \sigma_{34} \\ \sigma_{14} & \sigma_{24} & \sigma_{34} & \sigma_4^2 \end{bmatrix}$$

- Unstructured, one band `type=un(1)`

$$\boldsymbol{R} = \begin{bmatrix} \sigma_1^2 & 0 & 0 & 0 \\ 0 & \sigma_2^2 & 0 & 0 \\ 0 & 0 & \sigma_3^2 & 0 \\ 0 & 0 & 0 & \sigma_4^2 \end{bmatrix}$$

- Unstructured, two bands `type=un(2)`

$$\boldsymbol{R} = \begin{bmatrix} \sigma_1^2 & \sigma_{12} & 0 & 0 \\ \sigma_{12} & \sigma_2^2 & \sigma_{23} & 0 \\ 0 & \sigma_{23} & \sigma_3^2 & \sigma_{34} \\ 0 & 0 & \sigma_{34} & \sigma_4^2 \end{bmatrix}$$

Software – SAS

Recall `R` command

```
> fit.sleep = lmer(Reaction ~ Days + (Days|Subject) , sleepstudy)
```

The following yield equivalent results . . .

```
proc mixed data=sleep covtest cl;
  class subject;
  model reaction=days / solution cl;
  random intercept days / subject=subject solution type=un;
proc glimmix data=sleep;
  class subject;
  model reaction=days / solution cl;
  random intercept days / subject=subject solution type=un;
  covtest diagg / wald cl;
```

Longitudinal modeling that allows correlation across time for each subject . . .

```
proc mixed data=sleep covtest cl;
  class subject;
  model reaction=days / solution cl;
  random intercept days / subject=subject solution;
  repeated / subject=subject type=ar(1);
proc glimmix data=sleep;
  class subject;
  model reaction=days / solution cl;
  random intercept days / subject=subject solution;
  random _residual_ / subject=subject type=ar(1);
  covtest diagr / wald cl;
```

ST704, Sujit K. Ghosh

Generalized Linear Models, Part I

Sample questions :

Jan 2024 Part 2 Q2
Aug 2024 Part 2 Q3
Aug 2022 Part 1 Q4
Aug 2021 Part 1 Q5

Classical Linear Model (LM):

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}, \qquad \boldsymbol{\epsilon} \sim N(\boldsymbol{0}, \sigma^2 \boldsymbol{I})$$

- $\boldsymbol{Y}$ is $n \times 1$, observed
- $\boldsymbol{X}$ is $n \times p$, known covariate matrix
- $\boldsymbol{\beta}$ is $p \times 1$, unknown fixed effects
- Parameters to be estimated: $\boldsymbol{\beta}$, σ^2

In other words . . .

- Linear Predictor, aka Systematic Component:

  $\boldsymbol{X}\boldsymbol{\beta}$ . . . $\boldsymbol{x}_i^T\boldsymbol{\beta}$ is ith entry, often η_i

- Identity Link:

  $E(Y_i) = \boldsymbol{x}_i^T\boldsymbol{\beta}$ . . . "link" transforms $E(Y_i)$ to linear predictor scale

- Random Component:

  $Y_1, \ldots, Y_n$ indep normal with $var(Y_i) = \sigma^2$

Generalized Linear Model (GLM)

- Linear Predictor, aka Systematic Component:

  $\boldsymbol{X}\boldsymbol{\beta}$ ... $\boldsymbol{x}_i^T\boldsymbol{\beta}$ is ith entry, often η_i

- Link:

  $g\left(E(Y_i)\right) = \boldsymbol{x}_i^T\boldsymbol{\beta}$ ... $E(Y_i) = g^{-1}\left(\boldsymbol{x}_i^T\boldsymbol{\beta}\right)$

  Note that $g(\cdot)$ is link, $g^{-1}(\cdot)$ is inverse link

- Random Component:

  $Y_1, \ldots, Y_n$ indep from the **exponential family** with *dispersion parameter* ψ

Parameters to be estimated: $\boldsymbol{\beta}$ and ψ

Properties of link function $g(\cdot)$: • monotonic and invertible

- maps mean response to a scale where covariate effects are additive
- ensures range restriction for mean response
- distns in exponential family have "canonical" or "natural" link functions
- any suitable link function may be paired with any distribution in the exponential family

Goals for GLM

- Fit the model, i.e., estimate $\boldsymbol{\beta}$ and ψ
- CI and HT for $\boldsymbol{\beta}$: need distribution, standard error
- Inference on functions of $\boldsymbol{\beta}$:
  - $H_0 : \boldsymbol{A\beta} = \boldsymbol{m}$?
  - Estimate mean response $g^{-1}\left(\boldsymbol{x}_i^T \boldsymbol{\beta}\right)$ . . . CI & HT?
- How well does the model fit?

Exponential family with natural parameter $\theta = \theta(\mu)$, dispersion (scale) parameter ψ, and known weight w: pdf has form ...

$$f(y_i) = \exp\left\{\frac{y_i\theta_i - b(\theta_i)}{\psi w_i} + c(y_i, \psi w_i)\right\}$$

for some functions $b(\cdot)$ and $c(\cdot)$.

$N(\mu_i, \sigma^2)$

$$\begin{aligned}
f(y_i) &= \frac{1}{\sqrt{2\pi\sigma^2}}\exp\left\{-\frac{(y_i-\mu_i)^2}{2\sigma^2}\right\} = \exp\left\{-\frac{(y_i-\mu_i)^2}{2\sigma^2} - \frac{1}{2}\ln(2\pi) - \frac{1}{2}\ln(\sigma^2)\right\} \\
&= \exp\left\{\frac{-y_i^2 - \mu_i^2 + 2y_i\mu_i}{2\sigma^2} - \frac{1}{2}\ln(2\pi) - \frac{1}{2}\ln(\sigma^2)\right\} \\
&= \exp\left\{\underbrace{\frac{y_i\mu_i - \frac{1}{2}\mu_i^2}{\sigma^2}}_{\substack{\theta_i = \mu_i, \psi = \sigma^2 \\ b(\theta_i) = \frac{1}{2}\theta_i^2, w_i = 1}} \quad \underbrace{-\frac{y_i^2}{2\sigma^2} - \frac{1}{2}\ln(2\pi) - \frac{1}{2}\ln(\sigma^2)}_{c(y_i, \psi w_i)}\right\}
\end{aligned}$$

Exponential family with natural parameter $\theta = \theta(\mu)$, dispersion (scale) parameter ψ, and known weight w: pdf has form ...

$$f(y_i) = \exp\left\{\frac{y_i\theta_i - b(\theta_i)}{\psi w_i} + c(y_i, \psi w_i)\right\}$$

for some functions $b(\cdot)$ and $c(\cdot)$.

Y_i is proportion of successes: $\frac{1}{n_i} Bin(n_i, p_i)$

$$\begin{aligned}
f(y_i) &= \binom{n_i}{n_i y_i} p_i^{n_i y_i}(1-p_i)^{n_i - n_i y_i} = \binom{n_i}{n_i y_i}\left(\frac{p_i}{1-p_i}\right)^{n_i y_i}(1-p_i)^{n_i} \\
&= \exp\left\{n_i y_i \ln\left(\frac{p_i}{1-p_i}\right) + n_i \ln(1-p_i) + \ln\binom{n_i}{n_i y_i}\right\} \\
&= \exp\left\{\underbrace{\frac{y_i \ln\left(\frac{p_i}{1-p_i}\right) - \ln\left(\frac{1}{1-p_i}\right)}{1/n_i}}_{\theta_i = \ln\left(\frac{p_i}{1-p_i}\right),\ \psi = 1,\ w_i = 1/n_i,\ b(\theta_i) = \ln\left(\frac{1}{1-p_i}\right) = \ln(1+e^{\theta_i})} + \underbrace{\ln\binom{n_i}{n_i y_i}}_{c(y_i, \psi w_i)}\right\}
\end{aligned}$$

Exponential family with natural parameter $\theta = \theta(\mu)$, dispersion (scale) parameter ψ, and known weight w: pdf has form ...

$$f(y_i) = \exp\left\{\frac{y_i\theta_i - b(\theta_i)}{\psi w_i} + c(y_i, \psi w_i)\right\}$$

for some functions $b(\cdot)$ and $c(\cdot)$.

Poisson(λ_i)

$$f(y_i) = \frac{e^{-\lambda_i}\lambda_i^{y_i}}{y_i!} = \exp\left\{\begin{array}{cc} \underbrace{y_i \ln \lambda_i - \lambda_i}_{\theta_i = \ln\lambda_i,\ \psi = 1,\ w_i = 1} & \underbrace{-\ln(y_i!)}_{c(y_i,\psi w_i)} \\ b(\theta_i) = \lambda_i = e^{\theta_i} & \end{array}\right\}$$

Properties:

- $\mu_i = b'(\theta_i)$ is mean function
- $b''(\theta_i)$ is 'variance' function, with $var(Y_i) = \psi w_i b''(\theta_i)$

Distribution for Y_i	ψ	w_i	$\boxed{\theta_i(\mu_i)}$ Canonical link $g(\mu_i) = \boldsymbol{x}_i^T\boldsymbol{\beta}$	Inverse link $g^{-1}(\boldsymbol{x}_i^T\boldsymbol{\beta}) = \mu_i$	$\boxed{b'(\theta_i)}$ Mean function $E(Y_i)$	$\boxed{b''(\theta_i)}$ 'Variance' fnc $\frac{1}{\psi w_i} var(Y_i)$
$N(\mu_i, \sigma^2)$	σ^2	1	μ_i; identity	$\boldsymbol{x}_i^T\boldsymbol{\beta}$; identity	μ_i	1
$\frac{1}{n_i} Bin(n_i, p_i)$	1	$\frac{1}{n_i}$	$\ln\left(\frac{p_i}{1-p_i}\right)$; logit	$\frac{e^{\boldsymbol{x}_i^T\boldsymbol{\beta}}}{1+e^{\boldsymbol{x}_i^T\boldsymbol{\beta}}}$; expit	p_i	$p_i(1-p_i)$
$Poisson(\lambda_i)$	1	1	$\ln\lambda_i$; log	$e^{\boldsymbol{x}_i^T\boldsymbol{\beta}}$; exp	λ_i	λ_i

Note: The 'variance' function is sometimes denoted as $h(\mu_i)$

Inference on $\boldsymbol{\beta}$

- Estimate $\boldsymbol{\beta}$ by maximum likelihood
  - $\sqrt{n}\left(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right) \xrightarrow{d} N\left(\mathbf{0}, [\mathrm{I}_1(\boldsymbol{\beta})]^{-1}\right)$ where $n\mathrm{I}_1(\boldsymbol{\beta}) = \boldsymbol{F}^T \boldsymbol{V}^{-1} \boldsymbol{F}$

$$\boldsymbol{V} = \begin{bmatrix} var(Y_1) & & 0 \\ & \ddots & \\ 0 & & var(Y_n) \end{bmatrix}, \qquad \boldsymbol{F} = \frac{\partial \boldsymbol{\mu}}{\partial \boldsymbol{\beta}^T}$$

In other words, $\widehat{\boldsymbol{\beta}} \approx N\left(\boldsymbol{\beta}, \left[\boldsymbol{F}^T \boldsymbol{V}^{-1} \boldsymbol{F}\right]^{-1}\right)$ . . . use this for CIs

- Hypothesis testing for $\boldsymbol{\beta}$
  - Wald test: $T_W \overset{H_0}{\approx} \chi^2_\nu$
  - Score test: $T_S \overset{H_0}{\approx} \chi^2_\nu$
  - Asymptotic likelihood ratio test: $T_{LR} \overset{H_0}{\approx} \chi^2_\nu$ $\quad T_{LR} = 2\{\ell_{H_a} - \ell_{H_0}\}$
  - Exact likelihood ratio test

Deviance

$\ell(\boldsymbol{\mu}, \psi; \boldsymbol{y})$ denotes loglikelihood

- A *saturated* model has $\widehat{\boldsymbol{\mu}} = \boldsymbol{y}$, and "fits the data perfectly"
- Consider a model M, less complex than *saturated*, with estimated mean response $\widehat{\boldsymbol{\mu}}$ based on p regression parameters.

How well does Model M fit, relative to the saturated model?

- **Deviance for model M is**
$$D_M = \psi \cdot T_{LR:\text{ saturated vs. model M}} = \psi \cdot 2\{\ell(\boldsymbol{y}, \psi; \boldsymbol{y}) - \ell(\widehat{\boldsymbol{\mu}}, \psi; \boldsymbol{y})\}$$
- **Scaled deviance for model M is** simply D_M/ψ

$N(\mu_i, \sigma^2)$

$$\ell(\boldsymbol{\mu}, \psi; \boldsymbol{y}) = \sum_{i=1}^{n} \left\{ -\frac{(y_i - \mu_i)^2}{2\sigma^2} - \frac{1}{2}\ln(2\pi) - \frac{1}{2}\ln(\sigma^2) \right\}$$

for normal

$$\Rightarrow \ell(\boldsymbol{y}, \psi; \boldsymbol{y}) = \sum_{i=1}^{n} \left\{ -\frac{1}{2}\ln(2\pi) - \frac{1}{2}\ln(\sigma^2) \right\}$$

$$\ell(\widehat{\boldsymbol{\mu}}, \psi; \boldsymbol{y}) = \sum_{i=1}^{n} \left\{ -\frac{(y_i - \widehat{\mu}_i)^2}{2\sigma^2} - \frac{1}{2}\ln(2\pi) - \frac{1}{2}\ln(\sigma^2) \right\}$$

★ exact !

$$D_M = \psi \cdot 2 \sum_{i=1}^{n} \left\{ \frac{(y_i - \widehat{\mu}_i)^2}{2\sigma^2} \right\} = \sum_{i=1}^{n} (y_i - \widehat{\mu}_i)^2 = SSE \ldots\ldots D_M/\psi \sim \chi^2_{n-p} \text{ is exact}$$

- If Y_i has a distribution "close to normal" with link "close to identity," then $D_M/\psi \approx \chi^2_{n-p}$
  - Approximation will often NOT improve as n increases!
  - Suppose data are grouped, n is the # of groups and is fixed. We want the size of each group to be large

$$Bin(n_i, p_i),\ i = 1, \ldots, n \quad : \quad \text{want } n_i \text{ large}$$

  - Lack of fit testing ... H_0: model M fits the data vs. H_a: not H_0

Global test

$$\text{Reject if} \qquad D_M/\psi > \chi^2_{n-p,\alpha}$$

- Consider testing H_0: model M_0 vs. H_a: model M, where $M_0 \subset M$ is a submodel (i.e., nested) with $q < p$ regression parameters. Note that

$$T_{LR} = \frac{D_{M_0} - D_M}{\psi} \quad \xrightarrow{\text{d}} \quad \chi^2_{p-q} \quad \text{when } H_0 \text{ is true}$$

Distribution	D_M
$N(\mu_i, \sigma^2)$	$\sum_{i=1}^{n}(y_i - \widehat{\mu}_i)^2$
$\frac{1}{n_i} Bin(n_i, p_i)$	$2\sum_{i=1}^{n}\left\{n_i y_i \ln\left(\frac{y_i}{\widehat{\mu}_i}\right) + (n_i - n_i y_i)\ln\left(\frac{1-y_i}{1-\widehat{\mu}_i}\right)\right\}$
$Bin(n_i, p_i)$	$2\sum_{i=1}^{n}\left\{y_i \ln\left(\frac{y_i}{\widehat{\mu}_i}\right) + (n_i - y_i)\ln\left(\frac{n_i-y_i}{n_i-\widehat{\mu}_i}\right)\right\}$
$Poisson(\lambda_i)$	$2\sum_{i=1}^{n}\left\{y_i \ln\left(\frac{y_i}{\widehat{\mu}_i}\right) - (y_i - \widehat{\mu}_i)\right\}$

Estimating ψ

- Some families set a value for ψ, but it is good practice to estimate it
- Assuming $D_M/\psi \approx \chi^2_{n-p}$, a method of moments estimator of ψ is

$$D_M/\psi \overset{\text{set}}{=} E\left(\chi^2_{n-p}\right) = n-p \quad \Rightarrow \quad \underset{\text{MoM}}{\widehat{\psi}} = \frac{D_M}{n-p}$$

- $\widehat{\psi}$ "large" may be due to
  - inadequate linear predictor, i.e., missing predictors
  - overdispersion, i.e., $var(Y_i) > \psi w_i h(\mu_i)$ where ψ is given by family
    - Correlation between $Y_1, \ldots, Y_n$ can lead to overdispersion
- Recall

$$var(\widehat{\boldsymbol{\beta}}) \doteq \psi \left(\boldsymbol{F}^T [\text{diag}(w_1 h(\mu_1), \ldots, w_n h(\mu_n))]^{-1} \boldsymbol{F}\right)^{-1}$$

so over(under)-reporting the value of ψ will lead to SEs that are too large(small).

A Basic Checklist

- Choose family (distribution + link)
- Select covariates and estimate $\boldsymbol{\beta}$
  - Lack of fit testing using D_M/ψ if appropriate
  - Compare scaled deviances of nested models
- Estimate ψ
- Report $SE(\widehat{\beta}_j)$, $SE(\widehat{\mu}_i)$
- Inference on nonlinear function of $\boldsymbol{\beta}$ will need Taylor's approx for SE

- Model interpretation – odds, odds ratio, log odds, log odds ratio
- Problems with convergence – complete and quasi-complete separation

Inference, treating ψ as known

Maximum likelihood estimation for $\boldsymbol{\beta}$: ..... need $\ell(\boldsymbol{\beta}, \psi; \boldsymbol{y})$

- $\mu_i = g^{-1}(\boldsymbol{x}_i^T \boldsymbol{\beta})$ is a function of $\boldsymbol{\beta}$
- θ_i is a function of μ_i, and hence a function of $\boldsymbol{\beta}$ → natural parameter
- $\ell(\boldsymbol{\mu}, \psi; \boldsymbol{y})$ is loglikelihood function

$$\ell(\boldsymbol{\mu}, \psi; \boldsymbol{y}) = \sum_{i=1}^{n} \left\{ \frac{y_i \theta_i - b(\theta_i)}{\psi w_i} + c(y_i, \psi w_i) \right\}$$

- Score wrt $\boldsymbol{\theta}$:

$$\frac{\partial \ell(\boldsymbol{\mu}, \psi; \boldsymbol{y})}{\partial \theta_i} = \frac{y_i - b'(\theta_i)}{\psi w_i} = \frac{y_i - \mu_i}{\psi w_i}$$

- Score wrt $\boldsymbol{\beta}$:

$$\frac{\partial \ell(\boldsymbol{\mu}, \psi; \boldsymbol{y})}{\partial \boldsymbol{\beta}} = \boldsymbol{F}^T \boldsymbol{V}^{-1} (\boldsymbol{y} - \boldsymbol{\mu})$$

is $b''(\theta_i) = h(\mu_i)$? ?

where $\boldsymbol{V} = \text{diag}(\psi w_1 h(\mu_1), \ldots, \psi w_n h(\mu_n))$ has $var(Y_i)$ on its diagonal, and $\boldsymbol{F} = \frac{\partial \boldsymbol{\mu}}{\partial \boldsymbol{\beta}^T}$ has n rows and same number of columns as $\boldsymbol{X}$

- Example: $N(\mu_i, \sigma^2)$:
  $\mu_i = \boldsymbol{x}_i^T\boldsymbol{\beta}, \quad \boldsymbol{\mu} = \boldsymbol{X}\boldsymbol{\beta}, \quad h(\mu_i) = 1, \quad \psi = \sigma^2, \quad w_i = 1$
  $\hookrightarrow \boldsymbol{V} = \sigma^2\boldsymbol{I}, \quad \boldsymbol{F} = \frac{\partial \boldsymbol{X}\boldsymbol{\beta}}{\partial \boldsymbol{\beta}^T} = \boldsymbol{X}$
  $\hookrightarrow \boldsymbol{X}^T(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})/\sigma^2 = \boldsymbol{0} \quad \Rightarrow \quad \boldsymbol{X}^T\boldsymbol{y} = \boldsymbol{X}^T\boldsymbol{X}\boldsymbol{\beta}$, usual normal eqns

- An estimate of the large-sample variance of $\widehat{\boldsymbol{\beta}}$ is

$$\widehat{var(\widehat{\boldsymbol{\beta}})} = \left(\widehat{\boldsymbol{F}}^T\widehat{\boldsymbol{V}}^{-1}\widehat{\boldsymbol{F}}\right)^{-1} \overset{\text{normal, exact}}{=} \sigma^2(\boldsymbol{X}^T\boldsymbol{X})^{-1}$$

- Estimate mean response ...
  - $\widehat{\mu} = g^{-1}(\boldsymbol{x}^T\widehat{\boldsymbol{\beta}})$ usu. biased, even if $\widehat{\boldsymbol{\beta}}$ is unbiased
  - $SE(\widehat{\mu})$: Use Taylor series of $g^{-1}(\boldsymbol{x}^T\widehat{\boldsymbol{\beta}})$ around $\boldsymbol{\beta}^*$ to get linear approximation, then use variance of linearization:

delta theorem

$$\widehat{var(\widehat{\mu})} \doteq \left[\left.\frac{\partial g^{-1}(\eta)}{\partial \eta}\right|_{\eta=\widehat{\eta}}\right]^2 \boldsymbol{x}^T\left(\widehat{\boldsymbol{F}}^T\widehat{\boldsymbol{V}}^{-1}\widehat{\boldsymbol{F}}\right)^{-1}\boldsymbol{x}$$

  - If the canonical link (i.e., $\eta = \theta(\mu)$) is used: $\frac{\partial g^{-1}(\eta)}{\partial \eta} = \frac{\partial \mu}{\partial \theta(\mu)} = h(\mu)$

$$\widehat{var(\widehat{\mu})} \doteq h(\widehat{\mu})^2\boldsymbol{x}^T\left(\widehat{\boldsymbol{F}}^T\widehat{\boldsymbol{V}}^{-1}\widehat{\boldsymbol{F}}\right)^{-1}\boldsymbol{x}$$

- Tests for functions of $\boldsymbol{\beta}$ may be conducted using asymptotic Wald, likelihood ratio, and score tests, comparing each of T_W, T_{LR}, T_S to a chi-squared distribution with degrees of freedom matching the number of constraints placed on $\boldsymbol{\beta}$.

E.g., for testing $H_0 : \boldsymbol{A\beta} = \boldsymbol{m}$, then

$$T_W = (\boldsymbol{A}\widehat{\boldsymbol{\beta}} - \boldsymbol{m})^T \left(\boldsymbol{A} \left(\widehat{\boldsymbol{F}}^T \widehat{\boldsymbol{V}}^{-1} \widehat{\boldsymbol{F}} \right)^{-1} \boldsymbol{A}^T \right)^{-1} (\boldsymbol{A}\widehat{\boldsymbol{\beta}} - \boldsymbol{m})$$

compares to a $\chi^2_{rank(\boldsymbol{A})}$ distribution

ST704, Sujit K. Ghosh

GLM, Part II

Logistic Regression

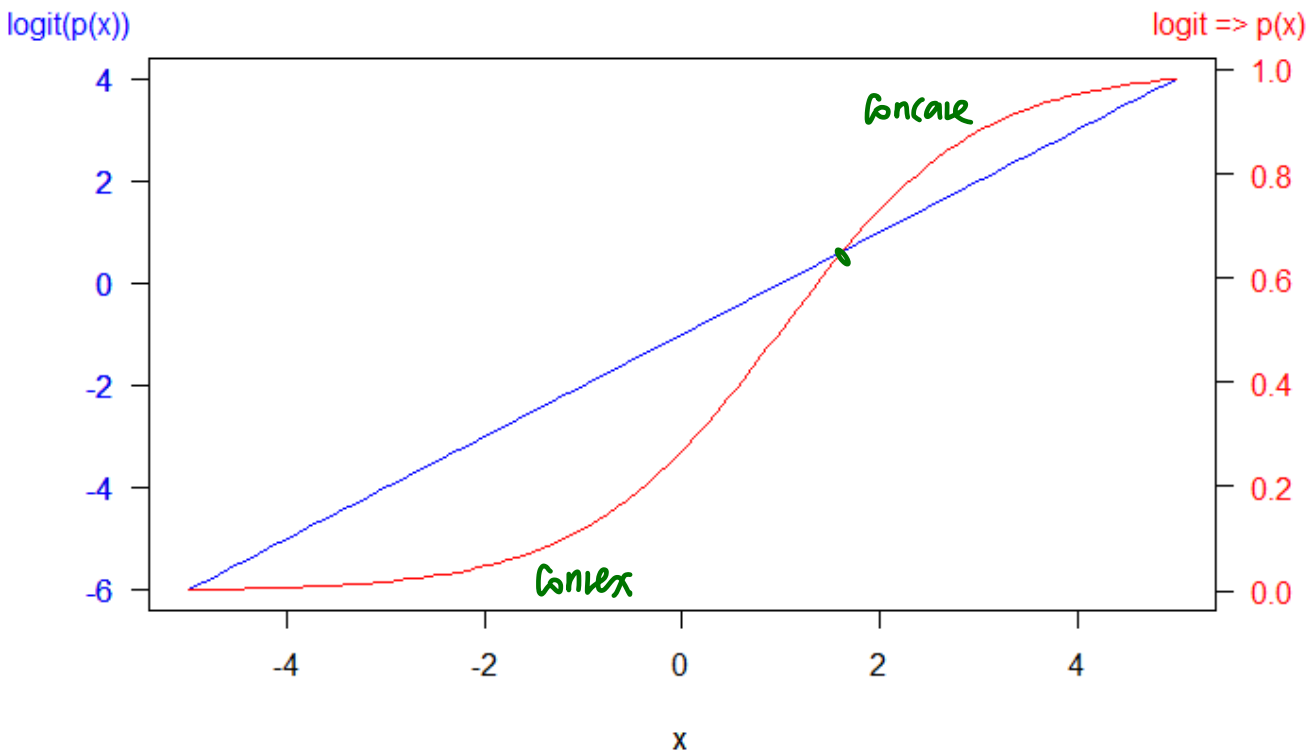
Linear predictor: $\boldsymbol{x}_i^T \boldsymbol{\beta}$

Link function: logit, i.e., $\ln\left(\frac{p_i}{1-p_i}\right)$ also $\ln\left(\frac{p(x)}{1-p(x)}\right)$

Random component: Y_i independent $\frac{1}{n_i} Bin(n_i, p_i)$, $i = 1, 2, \ldots, n$

$$\text{logit}(p(x)) = \ln\left(\frac{p(x)}{1-p(x)}\right) = \beta_0 + \beta_1 x = -1 + x$$

$g^{-1}(\cdot)$



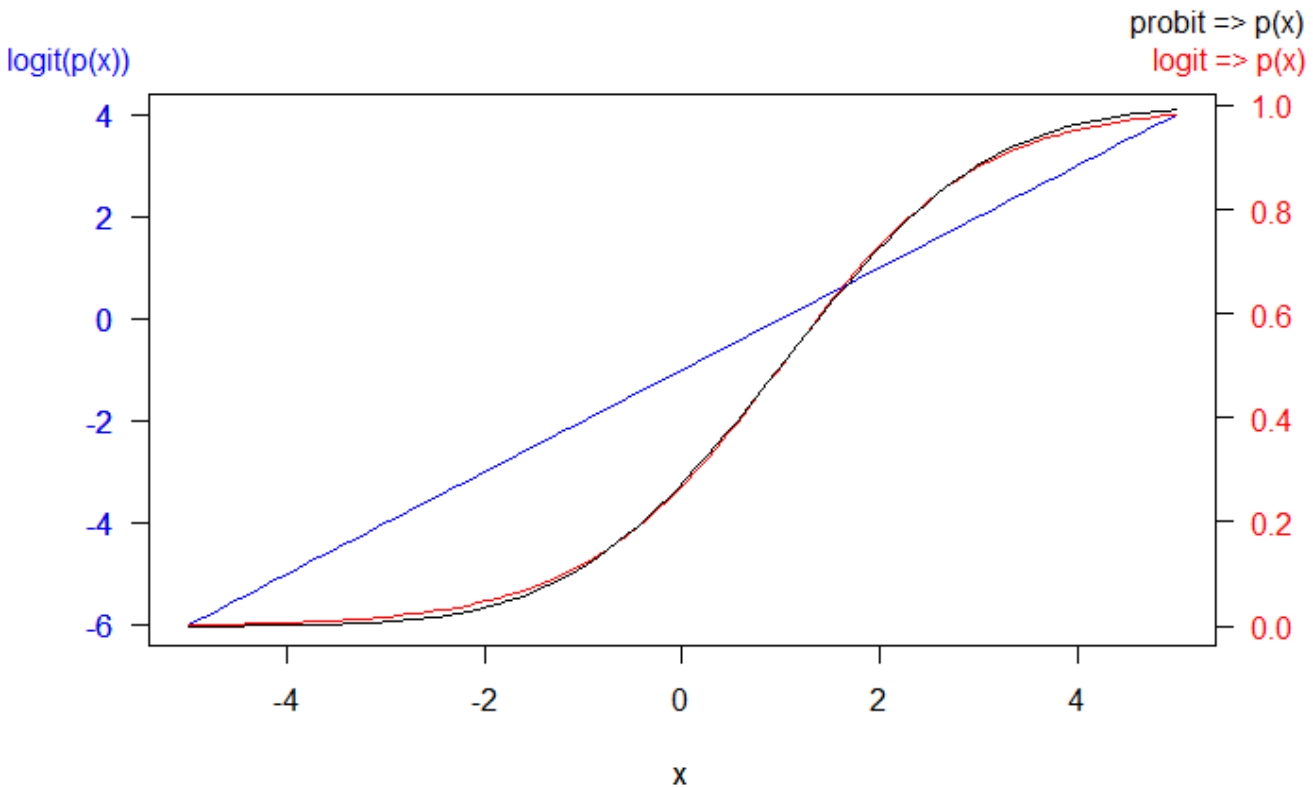
Probit Regression

Linear predictor: $\boldsymbol{x}_i^T \boldsymbol{\beta}$

Link function: probit, i.e., $\Phi^{-1}(p_i)$, where $\Phi(\cdot)$ is N(0,1) cdf also $\Phi^{-1}(p(\boldsymbol{x}))$

Random component: Y_i independent $\frac{1}{n_i} Bin(n_i, p_i)$, $i = 1, 2, \ldots, n$

$$\Phi^{-1}(p(x)) = \beta_0 + \beta_1 x = -0.6 + 0.6x$$



Interpreting the Logistic Model

$$\ln\left(\frac{p(x)}{1-p(x)}\right) = \beta_0 + \beta_1 x$$

$\text{odds}(x) = \frac{p(x)}{1-p(x)}$

- $= 1 \rightsquigarrow$ S & F equally likely, aka 'odds are even'
- $< 1 \rightsquigarrow$ S less likely
- $> 1 \rightsquigarrow$ S more likely

$\ln\left(\frac{p(x)}{1-p(x)}\right) \overset{\beta_0=\beta_1=0}{=} 0 \rightsquigarrow$ odds $(= e^0 = 1)$ are even and don't change with x

$\ln\left(\frac{p(x)}{1-p(x)}\right) \overset{\beta_1=0}{=} \beta_0 \rightsquigarrow$ odds $(= e^{\beta_0})$ are not even but don't change with x

- $\star$ $\beta_0 > 0$: S is more likely odd > 1
- $\star$ $\beta_0 < 0$: S is less likely odd < 1

$$\left[\frac{p(x+1)}{1-p(x+1)}\right] \Big/ \left[\frac{p(x)}{1-p(x)}\right] = \frac{e^{\beta_0+\beta_1(x+1)}}{e^{\beta_0+\beta_1 x}} = e^{\beta_1}$$

is 'odds ratio for 1 unit increase in x'

$$\ln\left(\frac{p(x)}{1-p(x)}\right) = \beta_0 + \beta_1 x \quad \Rightarrow \quad \frac{p(x)}{1-p(x)} = e^{\beta_0}\left(e^{\beta_1}\right)^x$$

- The odds at $x = 0$ is e^{β_0}
- The odds increase multiplicatively by e^{β_1} for 1 unit increase in x
  - $\beta_1 > 0$: $p(x) \uparrow$ as $x \uparrow$
  - $\beta_1 < 0$: $p(x) \downarrow$ as $x \uparrow$

Dose Response Modeling

Dose: % concentration of insecticide

Response: proportion of larvae killed, $\widehat{p}_i$: $y_i = k_i/n_i, \quad n_i = 20, \quad i = 1, \ldots, 7$

Assume:

(1) larvae react independently

(2) larvae exposed to same conc have equal probability of survival

concentration (%)	.375	.75	1.5	3	6	12	24
# of larvae killed	0	1	8	11	16	18	20

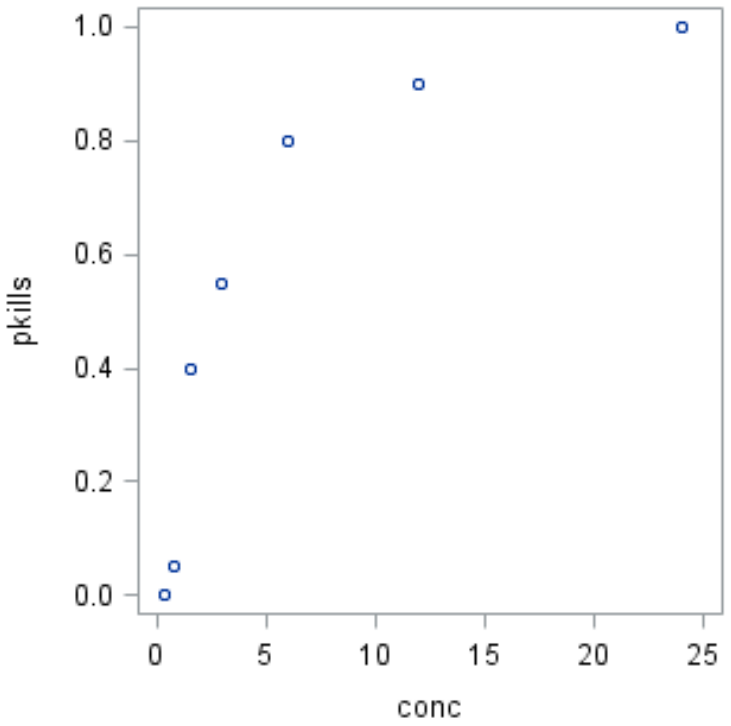
Grouped data ... $n = 7$ groups NOT $n = 7 \times 20 = 140$

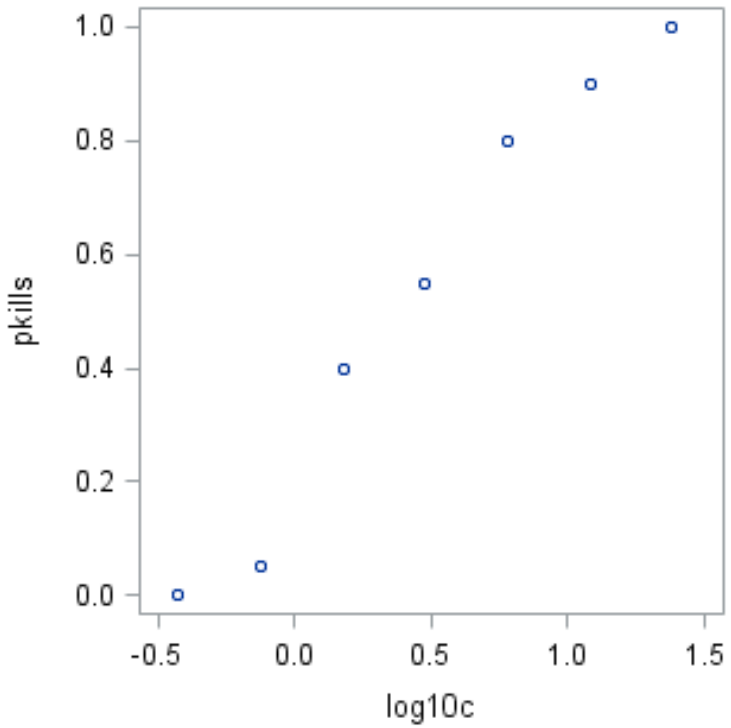
- Choose family : $Bin(n_i, p_i)$ + logit link
- Select covariates ... look at plots ... but $\widehat{p}_i = 0$, $\widehat{p}_i = 1$ cause problems

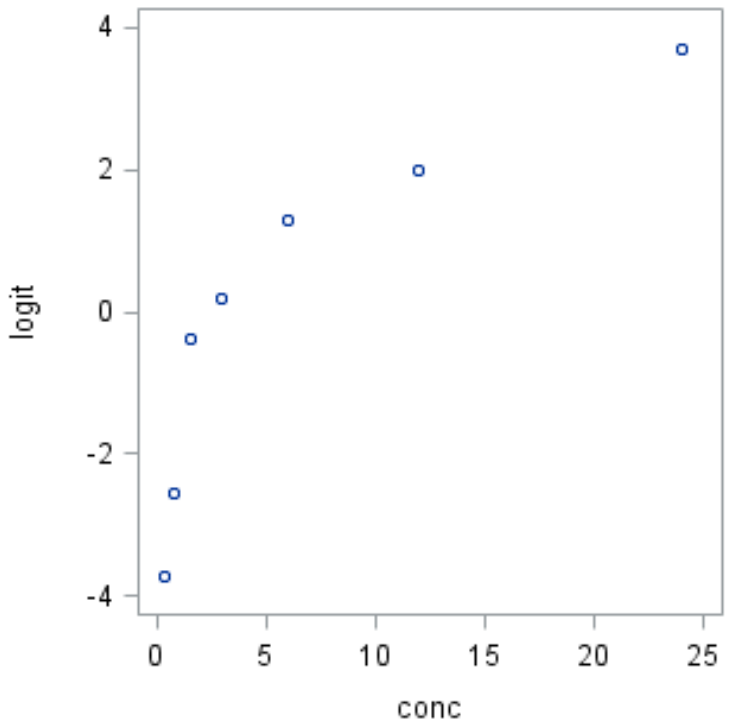
empirical logit: $$\ln\left(\frac{k_i + 0.5}{n_i - k_i + 0.5}\right)$$

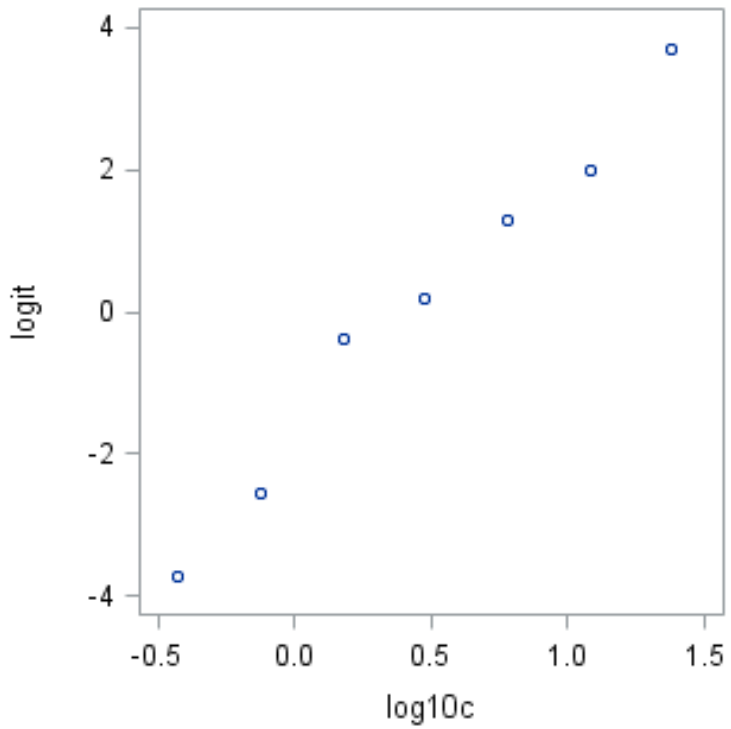
$\ln\left(\frac{p}{1-p}\right) = \ln\left(\frac{k_i/n_i}{1 - k_i/n_i}\right) = \ln\left(\frac{k_i}{n_i - k_i}\right)$

what is this 0.5


pkills
1.0
0.8
0.6
0.4
0.2
0.0
0
5
10
15
20
25
conc


pkills
1.0
0.8
0.6
0.4
0.2
0.0
-0.5
0.0
0.5
1.0
1.5
log10c


logit
4
2
0
-2
-4
0
5
10
15
20
25
conc


logit
4
2
0
-2
-4
-0.5
0.0
0.5
1.0
1.5
log10c

Logistic Model

Linear predictor: $\beta_0 + \beta_1 x$, $\quad x = \log_{10}(\text{conc})$

Link function: logit, i.e., $\ln\left(\frac{p_i}{1-p_i}\right)$

Random component: Y_i independent $Bin(n_i = 20, p_i)$, $\quad i = 1, 2, \ldots, 7$

$$\ln\left(\frac{p(x)}{1-p(x)}\right) = \beta_0 + \beta_1 x$$

$$p(x) = \frac{1}{1+e^{-\beta_0-\beta_1 x}}$$

Fit to data:

$$\widehat{\ln\left(\frac{p(x)}{1-p(x)}\right)} = -1.7305 + 4.1651x = \widehat{\beta}_0 + \widehat{\beta}_1 x$$

$$\widehat{\ln\left(\frac{p(x)}{1-p(x)}\right)} \quad = \quad -1.7305 + 4.1651x \quad = \quad \widehat{\beta}_0 + \widehat{\beta}_1 x$$

- $e^{\widehat{\beta}_0} = 0.177$ estimates "odds at $x = 0$" (i.e., odds at $c = 10^0 = 1$)
  ... survival is more likely (more than $5\times$) than death at % conc of 1

  CI: LR $\rightarrow \left(e^{-2.5351}, e^{-1.0557}\right) = (0.079, 0.348)$
  CI: Wald $\rightarrow \left(e^{-2.4637}, e^{-0.9973}\right) = (0.085, 0.369)$

- $e^{\widehat{\beta}_1} = 64.399$ estimates "odds ratio for 1 unit increase in x"
  ... odds of death increase by a factor of 64.399 for each increase of 1% conc

  CI: LR $\rightarrow \left(e^{3.0174}, e^{5.6003}\right) = (20.438, 270.508)$
  CI: Wald $\rightarrow \left(e^{2.8872}, e^{5.4430}\right) = (17.943, 231.135)$

Lack of fit?

- $D_M = 4.6206$ — Is this "large," suggesting a bad model?
- Ok to do a formal test because $n_i = 20$ is reasonable

H_0: model fits data well vs. H_a: not H_0

$$p - \text{value} = \Pr\left(\chi^2_{n-p} > D_M/\psi\right) = \Pr\left(\chi^2_{7-2} > 4.6206\right) = 0.4639$$

Estimate ψ: $\widehat{\psi} = \frac{D_M}{n-p} = \frac{4.6206}{7-2} = 0.9241$ — "close to" theoretical value 1

Inverse regn: What concentration kills 50% of larvae? ...80%? ...100δ%?

$$\ln\left(\frac{\delta}{1-\delta}\right) = \beta_0 + \beta_1 x_\delta \quad \Rightarrow \quad x_\delta = \frac{1}{\beta_1}\left\{\ln\left(\frac{\delta}{1-\delta}\right) - \beta_0\right\}$$

x_δ is a nonlinear function of $\boldsymbol{\beta}$, so need Taylor expansion for standard error!

	Estimate	SE
LD50	2.6030	0.3647
LD80	5.6016	1.0246

SAS code

```
data kills;
  input conc kills;
  trials=20;
  pkills=kills/trials;
  logit=log((kills+.5)/(trials - kills+.5));
  log10c=log10(conc);
datalines;
0.375 0
0.75  1
1.5   8
3.0  11
6.0  16
12.0  18
24.0  20
;
proc sgscatter data=kills;
  plot (pkills logit)*(conc log10c) / columns=2 rows=2;
proc genmod data=kills;
  model kills/trials=log10c / dist=binomial link=logit type1 type3 lrci;
                                                    *scale=deviance;
proc glimmix data=kills;
  model kills/trials=log10c / dist=binomial link=logit solution cl;
```

SAS code

```
proc logistic data=kills plots=all;
  model kills/trials=log10c / link=logit clparm=both clodds=both;
```

Options offered by `proc logistic` include:

- lots of diagnostics plots, checks
- forward, backward, stepwise, best subset selection
- produce a receiver operating characteristic (ROC) curve for fitted model
- BUT only for independent binomial or multinomial data

SAS code

```
proc nlmixed data=kills df=5;
  parameters b0=-1.7 b1=4.0;
  p = 1/(1+exp(- b0 - b1*log10c));
  model kills ~ binomial(trials,p);
  estimate 'LD50' -b0/b1;
  estimate 'LD50 original' 10**(-b0/b1);
  estimate 'LD80' ( log(0.8/0.2) - b0 ) /b1;
  estimate 'LD80 original' 10**(( log(0.8/0.2) - b0 )/b1);
  estimate 'OR' exp(b1);
```

R code

```
library(MASS)
library(car)

killsDF = data.frame(conc=c(.375,.75,1.5,3,6,12,24), dead=c(0,1,8,11,16,18,20))
killsDF$log10c = log(killsDF$conc,10)
killsDF$alive = 20 - killsDF$dead
fit = glm( cbind(dead,alive) ~ log10c, family=binomial(link=logit), data=killsDF )
summary( fit, correlation=T )

deviance(fit)
anova(fit)

confint(fit)  #profile likelihood CI by default
confint.default(fit)  #Wald CI, using z

( dose.p( fit, cf=1:2, p=c(.5,.8)) )  #Does inverse regression

predict(fit, type="response")  #estimated mu
predict(fit, type="link")      #estimated linear predictor (default)
predict(fit, type="response", se.fit=TRUE)
fitted(fit)                    #estimated mu
```

```
resid(fit, type="deviance")   #default
resid(fit, type="pearson")

residualPlots(fit)  #loess smooth replaces quadratic
outlierTest(fit)
influencePlot(fit)
crPlots(fit)

library(visreg)
visreg(fit,scale="linear",ylab="log odds (death)",points=list(cex=1))
visreg(fit,scale="response",ylab="Pr(death)",partial=TRUE,points=list(cex=1))

library(glmnet)
fit2 = glmnet( cbind(killsDF$log10c, killsDF$conc),
cbind(killsDF$alive, killsDF$dead),
family="binomial" )
```

Residuals and Diagnostics

- Several types of residuals are commonly used:
  - Raw or response: $y_i - \widehat{\mu}_i$ — not very useful
  - Pearson: $r_i^P \equiv \frac{y_i - \widehat{\mu}_i}{\sqrt{w_i h(\widehat{\mu}_i)}} \rightsquigarrow \mathbb{X}^2 = \sum_{i=1}^n (r_i^P)^2$ — very useful for diagnostics

    also $\frac{r_i^P}{\sqrt{\psi}\sqrt{1-h_i}}$ where h_i is leverage
  - Deviance: $r_i^D \equiv \text{sign}(y_i - \widehat{\mu}_i)\sqrt{d_i} \rightsquigarrow D_M = \sum_{i=1}^n d_i = \sum_{i=1}^n (r_i^D)^2$

    also $\frac{r_i^D}{\sqrt{\psi}\sqrt{1-h_i}}$ — very useful for diagnostics

  Note: The chi-squared statistic $\mathbb{X}^2$ and the scaled chi-squared statistic $\mathbb{X}^2/\psi$ are often used interchangeably with the deviance D_M and scaled deviance D_M/ψ.
- New diagnostic measure: likelihood displacement:
$$LD_i \equiv 2\left\{\ell_M(\widehat{\boldsymbol{\theta}}; \boldsymbol{y}) - \ell_M(\widehat{\boldsymbol{\theta}}_{(-i)}; \boldsymbol{y})\right\}$$
  where $\widehat{\boldsymbol{\theta}}_{(-i)}$ is MLE from excluding ith observation. Likelihood evaluated with *all* observations.
- Predicted values: $\widehat{\mu}_i$, response or $\boldsymbol{x}_i^T\widehat{\boldsymbol{\beta}}$, linear predictor

Complete and Quasi-complete Separation[1]

What is it? A linear predictor (almost) completely separates response values

Complete:

y	x1	x2
0	1	3
0	2	2
0	3	-1
0	3	-1
1	5	2
1	6	4
1	10	1
1	11	0

$\Pr(Y = 1 | x_1 > 3)$ best estimated as 1
$\Pr(Y = 1 | x_1 \leq 3)$ best estimated as 0

$$p(x_1) = \frac{e^{\beta x_1}}{1 + e^{\beta x_1}} \quad \xrightarrow{x_1 > 3:\ \beta \to \infty} \quad 1$$

$$p(x_1) = \frac{e^{\beta x_1}}{1 + e^{\beta x_1}} \quad \xrightarrow{0 < x_1 \leq 3:\ \beta \to -\infty} \quad 0$$

Quasi-complete:

y	x1	x2
0	1	3
0	2	2
0	3	-1
1	3	-1
1	5	2
1	6	4
1	10	1
1	11	0

Consequence: Difficulty getting convergence

Fix: Biased regression . . . Firth's penalized procedure. Other penalties

Common with: Rare events; very large predictor space; many binary predictors; small sample size

[1] https://stats.idre.ucla.edu/other/mult-pkg/faq/general/faqwhat-is-complete-or-quasi-complete-separation-in-logisticprobit-regression-and-how-do-we-deal-with-them/

SAS Code

```
data one;
input y x1 x2;
datalines;
0 1 3
0 2 2
0 3 -1
0 3 -1
1 5 2
1 6 4
1 10 1
1 11 0
;
proc genmod data=one descending;
  model y = x1 x2 / lrci lrcl ;
proc logistic data=one;
  model y(event='1') = x1 x2 / cl plcl;
proc logistic data=one;
  model y(event='1') = x1 x2 / cl plcl firth;
```

① ② ③

to model the prob. of 1

```
proc genmod data=one descending;
  model y = x1 x2 / dist=binomial lrci lrcl ; run;
```

①

```
NOTE: PROC GENMOD is modeling the probability that y='1'.

NOTE: The Pearson chi-square and deviance are not computed since the AGGREGATE
option is not specified.

NOTE: Algorithm converged.

WARNING: Convergence not attained for at least one side of profile likelihood
confidence interval for Prm1. Number of iterations = 50.
WARNING: Convergence not attained for at least one side of profile likelihood
confidence interval for Prm2. Number of iterations = 50.
WARNING: Convergence not attained for at least one side of profile likelihood
confidence interval for Prm3. Number of iterations = 50.

NOTE: The scale parameter was held fixed.
NOTE: PROCEDURE GENMOD used (Total process time):
real time            0.05 seconds
cpu time             0.04 seconds
```

appropriate for binomial models.

The GENMOD Procedure

①

Criteria For Assessing Goodness Of Fit

Criterion	DF	Value	Value/DF
Log Likelihood		0.0000	
Full Log Likelihood		0.0000	
AIC (smaller is better)		6.0000	
AICC (smaller is better)		12.0000	
BIC (smaller is better)		6.2383	

model perfectly separates the data, can find parameters that perfectly classify $y=0$ vs. $y=1$

Algorithm converged.

Analysis Of Maximum Likelihood Parameter Estimates

Parameter	DF	Estimate	Standard Error	Likelihood Ratio 95% Confidence Limits		Wald Chi-Square	Pr > ChiSq
Intercept	1	-107.266	4.912E8	-107.266	-107.266	0.00	1.0000
x1	1	25.1805	75202009	25.1805	25.1805	0.00	1.0000
x2	1	9.5189	2.1684E8	9.5189	9.5189	0.00	1.0000
Scale	0	1.0000	0.0000	1.0000	1.0000		

NOTE: The scale parameter was held fixed.

As cannot find proper CI

huge : non-meaningful / unstable

```
12   proc logistic data=one;* descending;
13     model y(event='1') = x1 x2 / cl plcl;
```

(2) wald ↗ Cprofile-likelihood ↘

```
NOTE: PROC LOGISTIC is modeling the probability that y=1.

WARNING: There is a complete separation of data points. The maximum likelihood
estimate does not exist.

WARNING: The LOGISTIC procedure continues in spite of the above warning. Results
shown are based on the last maximum likelihood iteration. Validity of the model
fit is questionable.

NOTE: There were 8 observations read from the data set WORK.ONE.
NOTE: PROCEDURE LOGISTIC used (Total process time):
real time           0.08 seconds
cpu time            0.03 seconds
```

```
14   proc logistic data=one;* descending;
15     model y(event='1') = x1 x2 / cl plcl firth;
16   run;
```

(3) Firth penalized procedure

```
NOTE: PROC LOGISTIC is modeling the probability that y=1.

NOTE: Convergence criterion (GCONV=1E-8) satisfied for the intercept-only model.
NOTE: Convergence criterion (GCONV=1E-8) satisfied.

NOTE: There were 8 observations read from the data set WORK.ONE.
NOTE: PROCEDURE LOGISTIC used (Total process time):
real time           0.08 seconds
cpu time            0.04 seconds
```

The LOGISTIC Procedure (2)

Complete separation of data points detected.

WARNING: The maximum likelihood estimate does not exist.

WARNING: The LOGISTIC procedure continues in spite of the above warning. Results shown are based on the last maximum likelihood iteration. Validity of the model fit is questionable.

Model Fit Statistics

Criterion	Intercept Only	Intercept and Covariates
AIC	13.090	6.005
SC	13.170	6.244
-2 Log L	11.090	0.005

⇒ dramatic improvement, but the "perfect fit" is misleading

Testing Global Null Hypothesis: BETA=0

Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	11.0850	2	0.0039
Score	6.8932	2	0.0319
Wald	0.1302	2	0.9370

wald directly depend on SE, only wald insignificant indicates large SE.

Analysis of Maximum Likelihood Estimates

Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	-20.7083	73.7757	0.0788	0.7789
x1	1	4.4921	12.7425	0.1243	0.7244
x2	1	2.3960	27.9875	0.0073	0.9318

very large

unstable

Odds Ratio Estimates ← $\exp(\beta_1)$ →

Effect	Point Estimate	95% Wald Confidence Limits
x1	89.311	<0.001 >999.999
x2	10.980	<0.001 >999.999

massive

meaningless

Parameter Estimates and Profile-Likelihood Confidence Intervals

Parameter	Estimate	95% Confidence Limits	
Intercept	-20.7083	.	-2.2738
x1	4.4921	0.4161	.
x2	2.3960	.	.

→ Does not compute limits, because likelihood is flat or infinite.

Parameter Estimates and Wald Confidence Intervals

Parameter	Estimate	95% Confidence Limits	
Intercept	-20.7083	-165.3	123.9
x1	4.4921	-20.4827	29.4669
x2	2.3960	-52.4584	57.2505

- includes 0, insignificant
- unreliable due to inflated standard errors

The LOGISTIC Procedure (3)

Intercept-Only Model Convergence Status
Convergence criterion (GCONV=1E-8) satisfied.

Model Convergence Status
Convergence criterion (GCONV=1E-8) satisfied.

Model Fit Statistics

Criterion	Intercept Only	Intercept and Covariates
AIC	7.478	5.995
SC	7.557	6.233
-2 Log L	5.478	-0.005

rare but okay with penalized likelihood

Testing Global Null Hypothesis: BETA=0

Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	5.4830	2	0.0645
Score	6.8932	2	0.0319
Wald	2.6789	2	0.2620

wald tends to be less reliable in small or noisy data

Analysis of Penalized Maximum Likelihood Estimates

Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	-3.2433	2.1696	2.2348	0.1349
x1	1	0.5163	0.3318	2.4212	0.1197
x2	1	0.4777	0.6322	0.5711	0.4498

β_1 β_2

Odds Ratio Estimates $\longleftrightarrow$ e^{β}

Effect	Point Estimate	95% Wald Confidence Limits	
x1	1.676	0.875	3.211
x2	1.612	0.467	5.567

CI includes 1
So neither significant

Parameter Estimates and Profile-Likelihood Confidence Intervals

Parameter	Estimate	95% Confidence Limits	
Intercept	-3.2433	-15.6408	-0.1134
x1	0.5163	0.0516	2.1353
x2	0.4777	-0.6628	4.5243

Generally more accurate than the wald test.
(when small sample size)

Parameter Estimates and Wald Confidence Intervals

Parameter	Estimate	95% Confidence Limits	
Intercept	-3.2433	-7.4956	1.0089
x1	0.5163	-0.1340	1.1667
x2	0.4777	-0.7613	1.7168

less reliable

R Code

```
> one = data.frame( y=c(0,0,0,0,1,1,1,1),x1=c(1,2,3,3,5,6,10,11),
+                   x2=c(3,2,-1,-1,2,4,1,0) )
> fit1 = glm(y ~ x1+x2, data=one, family=binomial)
Warning message: glm.fit: fitted probabilities numerically 0 or 1 occurred
> summary(fit1)
Call: glm(formula = y ~ x1 + x2, family = binomial, data = one)

Deviance Residuals:
         1           2           3           4           5           6           7           8
-2.110e-08  -1.404e-05  -2.522e-06  -2.522e-06   1.564e-05   2.110e-08   2.110e-08   2.110e-08

Coefficients:
             Estimate  Std. Error  z value  Pr(>|z|)
(Intercept)   -66.098  183471.722    0.000         1
x1             15.288   27362.843    0.001         1
x2              6.241   81543.720    0.000         1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 1.1090e+01  on 7  degrees of freedom
Residual deviance: 4.5454e-10  on 5  degrees of freedom
AIC: 6

Number of Fisher Scoring iterations: 24
```

① separation detected

huge, very large, meaningless

fake perfect fit : due to overfitting

$T = \frac{D_0 - D_M}{\varphi} \sim \chi^2_{df_0 - df_M}$

Under completely separation thus deviance / LR tests are not reliable. (Wald also highly affected)

firth

```
> library(logistf)
> fit2 = logistf(y~., data=one)
> summary(fit2)
logistf(formula = y ~ ., data = one)

Model fitted by Penalized ML
Confidence intervals and p-values by Profile Likelihood Profile Likelihood Profile Likelihood

                  coef  se(coef)   lower 0.95 upper 0.95     Chisq          p
(Intercept) -2.9748886 2.0332566 -15.47721061 -0.1208941 4.2179522 0.03999841
x1           0.4908484 0.3241088   0.05268297  2.1275832 5.0225056 0.02501994
x2           0.4313730 0.5941957  -0.65793072  4.4758930 0.7807099 0.37692411

Likelihood ratio test=5.505687 on 2 df, p=0.06374636, n=8
Wald test = 2.569861 on 2 df, p = 0.2766698

Covariance-Matrix:
           [,1]       [,2]       [,3]
[1,]  4.1341324 -0.4970381 -0.6764776
[2,] -0.4970381  0.1050465  0.0260937
[3,] -0.6764776  0.0260937  0.3530685
```

penalized estimates : shrink extreme values

significant

under small sample and separation, use LRT.

A Designed Binomial Study

- An experiment involves a 4×4 factorial design with factors

  $$\text{temperature: } T_1, T_2, T_3, T_4 \qquad \text{and} \qquad \text{concentration: } 0, 0.1, 1.0, 10$$

  Completely randomized design
- What is the effect on germination probability of seeds?
- For each treatment combination, there are 4 dishes each with 50 seeds. Count number that germinate, Y_{ijk}.
- Assume
  - seeds germinate independently
  - seeds treated similarly have the same probability of germinating
  - Then $Y_{ijk} \sim Bin(50, \pi_{ij})$
- Some questions:
  - $p_{1j} = p_{2j} = p_{3j} = p_{4j}$
  - $p_{i1} = p_{i2} = p_{i3} = p_{i4}$
  - $p_{1.} = p_{2.} = p_{3.} = p_{4.}$
  - $p_{.1} = p_{.2} = p_{.3} = p_{.4}$

Code

```
proc genmod data=germrate;
  class temp conc;
  model germ/trials = temp conc temp*conc /
                  link=logit dist=binomial type1 type3;
```

```
fit = glm( cbind(germ,trials-germ) ~ temp*conc,
                  family=binomial(link=logit), data=germrate )
summary( fit, correlation=F )
deviance( fit )
anova( fit )
```

Code

```
fit = glm( cbind(germ,trials-germ) ~ temp*conc,
                  family=binomial(link=logit), data=germrate )
summary( fit, correlation=F )
deviance( fit )
anova( fit )
```

---

how much deviance each term reduces when added sequentially to the model.

```
> anova(fit)
Analysis of Deviance Table
Model: binomial, link: logit
Response: cbind(germ, trials - germ)
Terms added sequentially (first to last)

          Df Deviance Resid. Df Resid. Dev
NULL                        63    1193.80
temp       3   763.69        60     430.11
conc       3   282.01        57     148.11
temp:conc  9    92.46        48      55.64
```

very significant

Good model : most deviance explained

From this, can test null v.s. temp ; temp v.s. temp + conc ; temp + conc v.s. temp + conc + temp*conc

Poisson Modeling

GLMs with log link are often called log-linear models.

AT&T 1988 soldering experiment : R data `solder` in package `rpart`, $n = 720$

- `skips`: number of defects (solder skips) on a circuit board [response]
- `Opening`: amount of clearance around the mounting pad (3 levels)
- `Solder`: amount of solder (Thick or Thin)
- `Mask`: type and thickness of the material used for the solder mask (A1.5, A3, A6, B3, B6)
- `PadType`: geometry and size of the mounting pad (10 levels)
- `Panel`: each board was divided into 3 panels

Code

```
library(rpart)
plot(skips~.,solder)
solder$Panel = as.factor(solder$Panel)
summary(solder)

fit = glm(skips~., family=poisson, data=solder); summary(fit)
anova(fit, test="Chisq")
```

```
proc genmod data=solder;
  class opening solder mask padtype panel;
  model skips = opening solder mask padtype panel /
                link=log dist=poisson type1 type3;
```

Poisson Modeling: $n_{ij} \sim \text{Poisson}(\lambda_{ij})$

Assuming independence:
$$n_{i\cdot} \sim \text{Poisson}(\lambda_{i\cdot}) \quad \text{where } \lambda_{i\cdot} = \sum_j \lambda_{ij}$$
$$n_{\cdot j} \sim \text{Poisson}(\lambda_{\cdot j}) \quad \text{where } \lambda_{\cdot j} = \sum_i \lambda_{ij}$$
$$n_{\cdot\cdot} \sim \text{Poisson}(\lambda_{\cdot\cdot}) \quad \text{where } \lambda_{\cdot\cdot} = \sum_{i,j} \lambda_{ij}$$
$$\lambda_{ij} = \frac{\lambda_{i\cdot}\lambda_{\cdot j}}{\lambda_{\cdot\cdot}}$$

$$\rightsquigarrow \ln(\lambda_{ij}) = -\ln(\lambda_{\cdot\cdot}) + \ln(\lambda_{i\cdot}) + \ln(\lambda_{\cdot j}) \equiv \mu + \alpha_i + \beta_j$$

Looks like a two-way ANOVA without interaction!

Multinomial Modeling: Assumes $n_{\cdot\cdot}$ is fixed. $\left(\{n_{ij}\}_{\forall i,j}\right) \sim \text{Multinomial}\left(n_{\cdot\cdot}, \{p_{ij}\}_{\forall i,j}\right)$

Assuming independence: $p_{ij} = p_{i\cdot}p_{\cdot j}$ $\rightsquigarrow$ $\lambda_{ij} = E(n_{ij}) = n_{\cdot\cdot}p_{i\cdot}p_{\cdot j}$

$$\rightsquigarrow \ln(\lambda_{ij}) = \ln(n_{\cdot\cdot}) + \ln(p_{i\cdot}) + \ln(p_{\cdot j}) \equiv \mu^* + \alpha_i^* + \beta_j^*$$

Looks like a two-way ANOVA without interaction!

Poisson Modeling: $n_{ij} \sim \text{Poisson}(\lambda_{ij})$

Independence: $\ln(\lambda_{ij}) = \mu + \alpha_i + \beta_j$

Saturated model: $\ln(\lambda_{ij}) = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij}$

Is there an in-between model?

Linear-by-linear association: $\ln(\lambda_{ij}) = \mu + \alpha_i + \beta_j + \gamma u_i v_j$

u_i, v_j represent "scores" e.g., $u_i = i - 2$, $v_j = j - 2.5$

Agreement: $\ln(\lambda_{ij}) = \mu + \alpha_i + \beta_j + \gamma \mathbb{I}(i = j)$

Example: Results of rating the same 236 units by two different raters on an ordinal scale from 1 to 5.

	Rater 1					
Rater 2	1	2	3	4	5	Total
1	10	6	4	2	2	24
2	12	20	16	7	2	57
3	1	12	30	20	6	69
4	4	5	10	25	12	56
5	1	3	3	8	15	30
Total	28	46	63	62	37	236

ST704, Sujit K. Ghosh

Generalized Linear Mixed Models

Generalized Linear Mixed Model (GLMM)

LM: $E(\boldsymbol{Y}) = \boldsymbol{X}\beta$ LMM: $E(\boldsymbol{Y}|\boldsymbol{\alpha}) = \boldsymbol{X}\beta + \boldsymbol{Z}\boldsymbol{\alpha}$

GLM: $E(\boldsymbol{Y}) = g^{-1}(\boldsymbol{X}\beta)$ GLMM: $E(\boldsymbol{Y}|\boldsymbol{\alpha}) = g^{-1}(\boldsymbol{X}\beta + \boldsymbol{Z}\boldsymbol{\alpha})$

- $\boldsymbol{Y}$ is $n \times 1$, observed
- $\boldsymbol{X}$ is $n \times p$, known covariate matrix
- β is $p \times 1$, unknown fixed effects
- $\boldsymbol{Z}$ is $n \times q$, known covariate matrix
- $\boldsymbol{\alpha}$ is $q \times 1$, random effects, where $\boldsymbol{\alpha} \sim N(\boldsymbol{0}, \boldsymbol{G})$
- $g(\cdot)$ is link function and $g^{-1}(\cdot)$ is inverse link function
- GLMM: $\text{Var}(\boldsymbol{Y}|\boldsymbol{\alpha}) = \boldsymbol{A}^{1/2}\boldsymbol{R}\boldsymbol{A}^{1/2}$
  - $\boldsymbol{R} = \psi\boldsymbol{I}$ as default
  - $\boldsymbol{A} = \text{diag}\left(w_1 h(\mu_1), \ldots, w_n h(\mu_n)\right)$

Slight abuse of notation in that functions $g(\cdot)$ and $g^{-1}(\cdot)$ apply to scalars, not vectors.

Standardized Mortality Ratio

Standardized Mortality Ratio (SMR) is a ratio between the observed number of deaths in an study population and the number of deaths would be expected, based on the age- and sex-specific rates in a standard population and the population size of the study population by the same age/sex groups. If the ratio of observed:expected deaths is greater than 1.0, there is said to be "excess deaths" in the study population.

. . . The SMR is used to compare the mortality risk of an study population to that of a standard population. It is especially applicable where the two populations have dissimilar age distributions, and in cases where direct age standardization may not be appropriate because the study population is small, or when lack of age-specific mortality rates precludes calculation of directly-age-standardized mortality rates. [1]

Y_i: # deaths in region i

E_i: expected # deaths in region i, according to age & sex death rates

SMR_i: $\frac{Y_i}{E_i}$

[1] https://ibis.doh.nm.gov/resource/SMR_ISR.html

Source: SAS glimmix "Example 44.3 Smoothing Disease Rates"

Lip cancer in 56 counties of Scotland, 1975-1980

X_i is % of employees in agriculture, fishing, forestry.

May be a surrogate for exposure to sunlight.

Does X_i help explain variability in SMR_i across counties?

Belief: $Y_i \sim Poisson(\lambda_i)$, so $SMR_i \equiv \frac{Y_i}{E_i}$ has mean $\frac{\lambda_i}{E_i}$

Model 1: $\ln\left(\frac{\lambda_i}{E_i}\right) = \beta_0 + \beta_1 x_i \quad \rightsquigarrow \quad \ln(\lambda_i) = \ln(E_i) + \beta_0 + \beta_1 x_i$

Model 2: $\ln\left(\frac{\lambda_i}{E_i}\right) = \beta_0 + \beta_1 x_i + \alpha_i \quad \rightsquigarrow \quad \ln(\lambda_i) = \ln(E_i) + \beta_0 + \beta_1 x_i + \alpha_i$

random

```
proc glimmix data=lipcancer plots=(studentpanel residualpanel);
  class county;
  loge = log(expected);
  model observed = employment / dist=poisson offset=loge
                   solution cl ddfm=none;
  random county;
  covtest zerog / cl(type=profile);
  covtest indep;
  SMR = observed/expected;
  SMR_pred = exp(_zgamma_ + _xbeta_);
  id county employment SMR SMR_pred;
  output out=glimmixout;
proc sgplot data=glimmixout;
  reg x=smr y=smr_pred / datalabel=county clm curvelabel;
run;
```

R function to use: glmer in package lme4

Description

Fit a generalized linear mixed-effects model (GLMM). Both fixed effects and random effects are specified via the model formula.

Usage

```
glmer(formula, data = NULL, family = gaussian, control = glmerControl(),
      start = NULL, verbose = 0L, nAGQ = 1L, subset, weights, na.action,
      offset, contrasts = NULL, mustart, etastart,
      devFunOnly = FALSE, ...)
```

Arguments

formula: a two-sided linear formula object describing both the fixed-effect and random effects part of the model, with the response on the left of a ~ operator and the terms, separated by + operators, on the right. Random-effects terms are distinguished by vertical bars ("|") separating expressions for design matrices from grouping factors.

Model Design

Simple Linear Model

Model: $Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \quad i = 1, \cdots, n \quad \varepsilon_i \overset{iid}{\sim} N(0, \sigma^2)$

Estimates: $\hat{\beta}_0 = \bar{y} - \bar{x}\hat{\beta}$

$\hat{\beta}_1 = r_{xy}\frac{s_y}{s_x} = \frac{\Sigma(x-\bar{x})(y-\bar{y})}{S_x^2}$

$r_{xy} = \frac{1}{n-1}\sum_{i=1}^{n}\left(\frac{x_i-\bar{x}}{s_x}\right)\left(\frac{y_i-\bar{y}}{s_y}\right)$

$S_x = \sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2}$

Source	df	SS	MS	E(MS)	F
Regression	1	$\Sigma_i(\hat{\bar{y}}_\cdot - \hat{y}_i)^2$	SSR	$\sigma^2 + \beta_1^2\Sigma_{i=1}^{n}(x_i-\bar{x})^2$	$MSR/MSE \sim F_{1,n-2}\left(\frac{\beta_1^2\Sigma_{i=1}^{n}(x_i-\bar{x})^2}{2\sigma^2}\right)$
Error	n-2	$\Sigma_i(y_i - \hat{y}_i)^2$	$SSE/(n-2)$	σ^2	
Total	n-1	$\Sigma_i(y_i - \bar{y}_\cdot)^2$			

$H_0: \beta_1 = 0 \qquad T = \frac{\hat{\beta}_1}{SE(\hat{\beta}_1)} \overset{H_0}{\sim} t_{n-2}$

Multiple Linear Regression

Model: $Y_{ij} = \beta_0 + \beta_i x_{ij} + \varepsilon_{ij} \quad i = 1, \cdots, p \quad j = 1, \cdots, n \quad \varepsilon_{ij} \overset{iid}{\sim} N(0, \sigma^2)$

Estimates: $\hat{\beta} = (x'x)^{-1}x'y$ (same if centered)

Source	df	SS	E(MS)	β_0 not included	F
Regression	p	$\Sigma_i\Sigma_j(\hat{\bar{y}}_\cdot - \hat{y}_{ij})^2$	$\sigma^2 + \frac{1}{p}\beta^T X^T X\beta$		$MSR/MSE \sim F_{p,n-p-1}\left(\frac{\beta^T X^T X\beta}{2\sigma^2}\right)$
Error	n-(p+1)	$\Sigma_i\Sigma_j(y_{ij} - \hat{y}_{ij})^2$	σ^2		
Total	n-1	$\Sigma_i\Sigma_j(y_{ij} - \bar{y}_\cdot)^2$	or $\frac{1}{p}\beta^T X^T(I-P_1)X\beta$ if non-centered		

Cell-Means

Models: $Y_{ij} = \mu_i + \varepsilon_{ij} \quad i = 1, \cdots, I \quad j = 1, \cdots, n_i \quad \varepsilon_{ij} \overset{iid}{\sim} N(0, \sigma^2)$

Estimates: $\hat{\mu}_i = \bar{y}_{i\cdot}$

Source	df	SS	E(MS)	F
Trt	I-1	$\Sigma_i\Sigma_j(\bar{y}_\cdot - \bar{y}_{i\cdot})^2 = y^T(P_X - P_1)y$	$\sigma^2 + \frac{1}{I-1}\sum_{i=1}^{I} n_i(\mu_i - \bar{\mu})^2$	$MSTrt/MSE \sim F_{I-1,I(J-1)}\left(\frac{\Sigma_{i=1}^{I} n_i(\mu_i-\bar{\mu})^2}{2\sigma^2}\right)$
Error	I(J-1)	$\Sigma_i\Sigma_j(y_{ij} - \bar{y}_{i\cdot})^2 = y^T(I-P_X)y$	σ^2	
Total	IJ-1	$\Sigma_i\Sigma_j(y_{ij} - \bar{y}_\cdot)^2 = y^T(I-P_1)y$	$= \frac{1}{I-1}M^T X^T(I-P_1)XM$	

$H_0: \mu_1 = \cdots = \mu_I \qquad F = \frac{MSTrt}{MSE}$

$H_0: \mu_i = 0 \qquad T = \frac{\hat{\mu}_i}{SE(\hat{\mu}_i)} = \frac{\bar{y}_{i\cdot}}{\sqrt{MSE/n_i}}$

Factorial (effect) Model (ANOVA)

① Fixed effect one-way (unbalanced)

Models: $Y_{ij} = \mu + \alpha_i + \varepsilon_{ij}$ $\quad i = 1, \cdots, I \quad j = 1, \cdots, n_i \quad \varepsilon_{ij} \overset{iid}{\sim} N(0, \sigma^2)$

Models: $Y_{ij} = \mu + \alpha_i + \varepsilon_{ij}$ $\quad i = 1, \cdots, I \quad j = 1, \cdots, \frac{N}{I} \quad \varepsilon_{ij} \overset{iid}{\sim} N(0, \sigma^2)$ completely randomized design (CRD)

notice only pure error in this case since $\hat{Y}_{ij} = \bar{Y}_{i\cdot}$

Estimates: $\hat{\mu + \alpha_i} = \bar{Y}_{i\cdot}$ $\quad \sum c_i \hat{\alpha}_i = \sum c_i \bar{Y}_{i\cdot}$ s.t. $\sum c_i = 0$

$\hat{\mu + \bar{\alpha}_\cdot} = \bar{Y}_{\cdot\cdot}$

Source	df	SS		E(MS)	F
Mean	1	$N\bar{Y}_{\cdot\cdot}^2$	$= Y^T P_1 Y$	$\sigma^2 + \sum_{i=1}^{I}\sum_{j=1}^{n_i} (\mu + \bar{\alpha})^2$	$\frac{MSM}{MSE} \sim F_{1, N-I}(\frac{N(\mu+\bar{\alpha})^2}{2\sigma^2})$
Trt	$I-1$	$\sum_{ij}(\bar{Y}_{i\cdot} - \bar{Y}_{\cdot\cdot})^2$	$= Y^T(P_X - P_1)Y$	$\sigma^2 + \frac{1}{I-1}\sum_{i=1}^{I} n_i(\alpha_i - \bar{\alpha}_\cdot)^2$	$\frac{MST}{MSE} \sim F_{I-1, N-I}(\frac{1}{2\sigma^2}\alpha^T X^T (I - P_1) X\alpha)$
Error	$N-I$	$\sum_{ij}(Y_{ij} - \bar{Y}_{i\cdot})^2$	$= Y^T(I - P_X)Y$	σ^2	
Total	$N-1$	$\sum_{ij}(Y_{ij} - \bar{Y}_{\cdot\cdot})^2$			

$\updownarrow$

$\frac{1}{I-1}\alpha^T X^T (I - P_1) X\alpha$

can disgard this

Ho: $\alpha_1 = \cdots = \alpha_I$

$F = \frac{MSTrt}{MSE}$

Ho: $\alpha_i - \alpha_j = 0$

$T = \frac{c'\hat{\theta}}{SE(c'\hat{\theta})} = \frac{c'\hat{\theta}}{\sqrt{MSE}\sqrt{h'h}} \sim t_{n-rank(X)}$ $(N-I)$ $\quad$ with $c'\hat{\theta} = h'Y$

$F = \frac{c'\hat{\theta}^2 / h'h}{MSE} \sim F_{1, n-rank(X)}$

Ho: $\sum_{i=1}^{I} c_i \alpha_i = 0$

$T = \frac{c'\hat{\theta}}{\sqrt{MSE}\sqrt{h'h}} = \frac{\sum c_i \bar{Y}_{i\cdot}}{\sqrt{MSE}\sqrt{\sum \frac{c_i^2}{n_i}}} \sim t_{n-rank(X)} = df\ error$

② Random effect one-way (balanced)

Models: $Y_{ij} = \mu + A_i + \varepsilon_{ij}$ $\quad i = 1, \cdots, a \quad j = 1, \cdots, \tilde{n}$ $\quad \varepsilon_{ij} \overset{iid}{\sim} N(0, \sigma^2)$, $A_i \overset{iid}{\sim} N(0, \sigma_A^2)$ } mutually independent

Estimates: $\hat{\mu} = \bar{Y}_{\cdot\cdot}$

Source	df	SS		E(MS)	F
Mean	1	$N\bar{Y}_{\cdot\cdot}^2$	$= Y' P_1 Y$		
A	$a-1$	$\sum_i\sum_j(\bar{Y}_{\cdot\cdot} - \bar{Y}_{i\cdot})^2$	$= Y'(P_X - P_1)Y$	$\sigma^2 + \tilde{n}\sigma_A^2$	$\frac{MSA}{MSE}(\frac{\sigma^2}{\sigma^2 + n\sigma_A^2}) \sim F_{a-1, a(n-1)}$
Error	$a(\tilde{n}-1)$	$\sum_i\sum_j(Y_{ij} - \bar{Y}_{i\cdot})^2$	$= Y'(I - P_X)Y$	σ^2	
Total	$n-1$	$\sum_i\sum_j(Y_{ij} - \bar{Y}_{\cdot\cdot})^2$	$= Y'(I - P_1)Y$		

Ho: $\sigma_A^2 = 0$

$F = \frac{MSA}{MSE} \overset{Ho}{\sim} F_{a-1, a(n-1)}$ (0)

③ Fixed effect (two way full factorial : crossed) ← simple contrasts are NOT estimable

Models: $Y_{ijk} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \varepsilon_{ijk}$

$i = 1, \cdots, a \quad j = 1, \cdots, b \quad k = 1, \cdots, \tilde{n}_{ij} \quad \varepsilon_{ijk} \overset{iid}{\sim} N(0, \sigma^2)$

Estimates:

equivalent

$\hat{\mu} + \hat{\bar{\alpha}}_{\cdot} + \hat{\bar{\beta}}_{\cdot} + \widehat{(\alpha\beta)}_{\cdot\cdot} = \bar{Y}_{\cdots}$

$\hat{\mu} + \hat{\alpha}_i + \hat{\beta}_j + \widehat{(\alpha\beta)}_{ij} = \bar{Y}_{ij\cdot}$

$\hat{\mu} + \hat{\alpha}_i + \hat{\bar{\beta}}_{\cdot} + \widehat{(\alpha\beta)}_{i\cdot} = \bar{Y}_{i\cdot\cdot}$

$\hat{\mu} + \hat{\bar{\alpha}}_{\cdot} + \hat{\beta}_j + \widehat{(\alpha\beta)}_{\cdot j} = \bar{Y}_{\cdot j\cdot}$

for unbalanced design, marginal mean biased

when balanced LSMean = Mean

unbiased and BLUE

$\Sigma_i c_i (\hat{\alpha}_i + \widehat{(\alpha\beta)}_{ij}) = \Sigma_i c_i \bar{Y}_{ij\cdot}$ simple effect of α at j

$\Sigma_i c_i (\hat{\alpha}_i + \widehat{(\bar{\alpha}\beta)}_{i\cdot}) = \Sigma_i c_i \bar{Y}_{i\cdot\cdot}$ main effect of α

$\Sigma_j d_j (\hat{\beta}_j + \widehat{(\alpha\beta)}_{ij}) = \Sigma_j d_j \bar{Y}_{ij\cdot}$ simple effect of β at i

$\Sigma_j d_j (\hat{\beta}_j + \widehat{(\bar{\alpha}\beta)}_{\cdot j}) = \Sigma_j d_j \bar{Y}_{\cdot j\cdot}$ main effect of β

$\Sigma_i \alpha_i \Sigma_j d_j (\alpha\beta)_{ij} = \Sigma_i \alpha_i \Sigma_j d_j \bar{Y}_{ij\cdot}$

interaction effect of $\alpha\beta$ (consistency of simple effect)

Models: $Y_{ijk} = \mu^* + \alpha_i^* + \beta_j^* + (\alpha\beta)_{ij}^* + \varepsilon_{ijk}$

$i = 1, \cdots, a \quad j = 1, \cdots, b \quad k = 1, \cdots, n \quad \varepsilon_{ijk} \overset{iid}{\sim} N(0, \sigma^2)$

$\mu^* = \mu - \bar{\alpha\beta}_{\cdot\cdot}$

$\alpha_i^* = \alpha_i + \bar{\alpha\beta}_{i\cdot}$

$\beta_j^* = \beta_j + \bar{\alpha\beta}_{\cdot j}$

$(\alpha\beta)_{ij}^* = (\alpha\beta)_{ij} - \bar{\alpha\beta}_{i\cdot} - \bar{\alpha\beta}_{\cdot j} + \bar{\alpha\beta}_{\cdot\cdot}$

Source	df	SS	E(MS)	F
can discard → Mean	1	$N\bar{Y}_{\cdots}^2$		
α	$a-1$	$\sum_{ijk}(\bar{Y}_{i\cdot\cdot} - \bar{Y}_{\cdots})^2$	$\sigma^2 + b\tilde{n}Q(\alpha^*)$	$\frac{MS_\alpha}{MSE} \sim F\left((a-1), \frac{b\tilde{n}Q(\alpha^*)}{z\sigma^2}\right)$
β	$b-1$	$\sum_{ijk}(\bar{Y}_{\cdot j\cdot} - \bar{Y}_{\cdots})^2$	$\sigma^2 + a\tilde{n}Q(\beta^*)$	$\frac{MS_\beta}{MSE} \sim F\left((b-1), \frac{a\tilde{n}Q(\beta^*)}{z\sigma^2}\right)$
$\alpha\beta$	$(a-1)(b-1)$	$\sum_{ijk}(\bar{Y}_{ij\cdot} - \bar{Y}_{i\cdot\cdot} - \bar{Y}_{\cdot j\cdot} + \bar{Y}_{\cdots})^2$	$\sigma^2 + \tilde{n}Q((\alpha\beta)^*)$	$\frac{MS_{\alpha\beta}}{MSE} \sim F\left((a-1)(b-1), \frac{\tilde{n}Q((\alpha\beta)^*)}{z\sigma^2}\right)$
Error	$n-ab$	$\sum_{ijk}(\bar{Y}_{ij\cdot} - \bar{Y}_{\cdots})^2$	σ^2	
Total	$n-1$	$\sum_{ijk}(Y_{ijk} - \bar{Y}_{\cdots})^2$		

$Q(\alpha^*) = \frac{1}{a-1}\sum_{i=1}^{a}(\alpha_i^* - \bar{\alpha}_{\cdot}^*)^2$

$= \frac{1}{a-1}\sum_{i=1}^{a}(\alpha_i + \bar{\alpha\beta}_{i\cdot} - \bar{\alpha}_{\cdot} - \bar{\alpha\beta}_{\cdot\cdot})^2$

$H_0: \alpha_1^* = \cdots = \alpha_a^*$

$H_0: \beta_1^* = \cdots = \beta_b^*$

$H_0: (\alpha\beta)_{ij}^* = 0 \ \forall i, j$

$\alpha_i^* = \alpha_i + \bar{\alpha\beta}_{i\cdot}$ ← $\bar{Y}_{i\cdot\cdot} - \bar{Y}_{\cdots} \Rightarrow \alpha_i^* - \bar{\alpha}_{\cdot}$

$\beta_j^* = \beta_j + \bar{\alpha\beta}_{\cdot j}$ ← $\bar{Y}_{\cdot j\cdot} - \bar{Y}_{\cdots} \Rightarrow \beta_j^* - \bar{\beta}_{\cdot}$

$(\alpha\beta)_{ij}^* = (\alpha\beta)_{ij} - \bar{\alpha\beta}_{i\cdot} - \bar{\alpha\beta}_{\cdot j} + \bar{\alpha\beta}_{\cdot\cdot}$ ← $\bar{Y}_{ij\cdot} - \bar{Y}_{i\cdot\cdot} - \bar{Y}_{\cdot j\cdot} + \bar{Y}_{\cdots} \Rightarrow (\alpha\beta)_{ij}^*$

$(\alpha_i + \beta_j + (\alpha\beta)_{ij}) - (\alpha_i + \bar{\beta}_{\cdot} + \bar{\alpha\beta}_{i\cdot}) - (\bar{\alpha}_{\cdot} + \beta_j + \bar{\alpha\beta}_{\cdot j}) + (\bar{\alpha}_{\cdot} + \bar{\beta}_{\cdot} + \bar{\alpha\beta}_{\cdot\cdot})$

$= (\alpha\beta)_{ij} - \bar{\alpha\beta}_{i\cdot} - \bar{\alpha\beta}_{\cdot j} + \bar{\alpha\beta}_{\cdot\cdot}$

$= (\alpha\beta)_{ij}^*$

④ Mixed two-way effect model (balanced, crossed)

Models: $Y_{ijk} = \mu + \alpha_i + B_j + \alpha B_{ij} + \varepsilon_{ijk}$

$i=1,\cdots,a \quad j=1,\cdots,b \quad k=1,\cdots,n$

$B_j \overset{iid}{\sim} N(0, \sigma_B^2)$, $\alpha B_{ij} \overset{iid}{\sim} N(0, \sigma_{\alpha B}^2)$, $\varepsilon_{ijk} \overset{iid}{\sim} N(0, \sigma^2)$ } mutually independent

Estimates: $\hat{\mu} + \hat{\alpha}_\cdot = \bar{Y}_{\cdots}$

$\hat{\mu} + \hat{\alpha}_i = \bar{Y}_{i\cdots}$

Source	df	SS	E(MS)	F
α	$a-1$	$\sum_{ijk}(\bar{Y}_{\cdots} - \bar{Y}_{i\cdots})^2$	$\sigma^2 + \tilde{n}\sigma_{\alpha B}^2 + b\tilde{n}Q(\alpha)$	$\frac{MSA}{MSAB} \sim F_{a-1,(a-1)(b-1)}(\frac{(a-1)b\tilde{n}Q(\alpha)}{2(\sigma^2+\tilde{n}\sigma_{\alpha B}^2)})$
B	$b-1$	$\sum_{ijk}(\bar{Y}_{\cdots} - \bar{Y}_{\cdot j\cdot})^2$	$\sigma^2 + \tilde{n}\sigma_{\alpha B}^2 + a\tilde{n}\sigma_B^2$	$\frac{MSB}{MSAB}(\frac{\sigma^2+\tilde{n}\sigma_{\alpha B}^2}{\sigma^2+\tilde{n}\sigma_{\alpha B}^2+a\tilde{n}\sigma_B^2}) \sim F_{b-1,(a-1)(b-1)}$
αB	$(a-1)(b-1)$	$\sum_{ijk}(\bar{Y}_{ij\cdot} - \bar{Y}_{i\cdots} - \bar{Y}_{\cdot j\cdot} + \bar{Y}_{\cdots})^2$	$\sigma^2 + \tilde{n}\sigma_{\alpha B}^2$	$\frac{MSAB}{MSE}(\frac{\sigma^2}{\sigma^2+\tilde{n}\sigma_{\alpha B}^2}) \sim F_{(a-1)(b-1),n-ab}$
Error	$n-ab$	$\sum_{ijk}(Y_{ijk} - \bar{Y}_{ij\cdot})^2$	σ^2	
Total	$n-1$	$\sum_{ijk}(Y_{ijk} - \bar{Y}_{\cdots})^2$		

$H_0: \alpha_1 = \cdots = \alpha_a$: $MSA/MSAB \overset{H_0}{\sim} F_{(a-1),(a-1)(b-1)}$

$H_0: \sigma_B^2 = 0$: $MSB/MSAB \overset{H_0}{\sim} F_{(b-1),(a-1)(b-1)}$

$H_0: \sigma_{\alpha B}^2 = 0$: $MSAB/MSE \overset{H_0}{\sim} F_{(a-1)(b-1),n-ab}$

⑤ Random two-way effect model (balanced, crossed)

Models: $Y_{ijk} = \mu + A_i + B_j + AB_{ij} + \varepsilon_{ijk}$

$i=1,\cdots,a \quad j=1,\cdots,b \quad k=1,\cdots,\tilde{n}$

$A_i \overset{iid}{\sim} N(0, \sigma_A^2)$, $B_j \overset{iid}{\sim} N(0, \sigma_B^2)$, $AB_{ij} \overset{iid}{\sim} N(0, \sigma_{AB}^2)$, $\varepsilon_{ijk} \overset{iid}{\sim} N(0, \sigma^2)$ } mutually independent

Source	df	SS	E(MS)	F
α	$a-1$	$\sum_{ijk}(\bar{Y}_{\cdots} - \bar{Y}_{i\cdots})^2$	$\sigma^2 + \tilde{n}\sigma_{\alpha B}^2 + b\tilde{n}\sigma_A^2$	$\frac{MSA}{MSAB}(\frac{\sigma^2+\tilde{n}\sigma_{\alpha B}^2}{\sigma^2+\tilde{n}\sigma_{\alpha B}^2+b\tilde{n}\sigma_A^2}) \sim F_{a-1,(a-1)(b-1)}$
B	$b-1$	$\sum_{ijk}(\bar{Y}_{\cdots} - \bar{Y}_{\cdot j\cdot})^2$	$\sigma^2 + \tilde{n}\sigma_{\alpha B}^2 + a\tilde{n}\sigma_B^2$	$\frac{MSB}{MSAB}(\frac{\sigma^2+\tilde{n}\sigma_{\alpha B}^2}{\sigma^2+\tilde{n}\sigma_{\alpha B}^2+a\tilde{n}\sigma_B^2}) \sim F_{b-1,(a-1)(b-1)}$
αB	$(a-1)(b-1)$	$\sum_{ijk}(\bar{Y}_{ij\cdot} - \bar{Y}_{i\cdots} - \bar{Y}_{\cdot j\cdot} + \bar{Y}_{\cdots})^2$	$\sigma^2 + \tilde{n}\sigma_{\alpha B}^2$	$\frac{MSAB}{MSE}(\frac{\sigma^2}{\sigma^2+\tilde{n}\sigma_{\alpha B}^2}) \sim F_{(a-1)(b-1),n-ab}$
Error	$n-ab$	$\sum_{ijk}(Y_{ijk} - \bar{Y}_{ij\cdot})^2$	σ^2	
Total	$n-1$	$\sum_{ijk}(Y_{ijk} - \bar{Y}_{\cdots})^2$		

$H_0: \sigma_A^2 = 0$: $MSA/MSAB \overset{H_0}{\sim} F_{(a-1),(a-1)(b-1)}$

$H_0: \sigma_B^2 = 0$: $MSB/MSAB \overset{H_0}{\sim} F_{(b-1),(a-1)(b-1)}$

$H_0: \sigma_{AB}^2 = 0$: $MSAB/MSE \overset{H_0}{\sim} F_{(a-1)(b-1),n-ab}$

⑥ Fixed effect (two way full factorial : nested)

Models: $Y_{ijk} = \mu + \alpha_i + \beta_{j(i)} + \varepsilon_{ijk}$ $\quad \varepsilon_{ijk} \overset{iid}{\sim} N(0,\sigma^2)$

$i = 1, \cdots, a \quad j = 1, \cdots, b_i \quad k = 1, \cdots, n_i$

Estimates: A main : $\frac{1}{b_i}\sum_{j=1}^{b_i}\sum_{i=1}^{a} c_i(\alpha_i + \beta_{j(i)}) = \sum_{i=1}^{a} c_i(\alpha_i + \bar{\beta}_{\cdot(i)})$

B simple for given i : $\sum_{j=1}^{b_i} d_j(\alpha_i + \beta_{j(i)}) = \sum_{j=1}^{b_i} \beta_{j(i)}$

→ There are $b_i - 1$ linearly independent B simple effects at i (from d_j independent contrasts)

$B_i', \cdots, B_i^{b_i-1}$

Source	df	SS	E(MS)	F
α	$a-1$	$\sum_{ijk}(\bar{Y}_{\cdots} - \bar{Y}_{i\cdot\cdot})^2$	$\sigma^2 + \frac{1}{a-1}\sum_i\sum_j(\alpha_i^* - \bar{\alpha}^*)^2$	$\frac{MSA}{MSE} \sim F\left(\frac{\sum_i\sum_j(\alpha_i^* - \bar{\alpha}^*)^2}{2\sigma^2}\right)$
$\beta(\alpha)$	$\sum_{i=1}^{a}(b_i-1)$	$\sum_{ijk}(\bar{Y}_{i\cdot\cdot} - \bar{Y}_{ij\cdot})^2$	$\sigma^2 + \frac{1}{\sum(b_i-1)}\sum_i\sum_j(\beta_{j(i)} - \bar{\beta}_{\cdot(i)})^2$	$\frac{MSB}{MSE} \sim F\left(\frac{\sum_i\sum_j(\beta_{j(i)} - \bar{\beta}_{\cdot(i)})^2}{2\sigma^2}\right)$
Error	$n - \sum_i b_i$	$\sum_{ijk}(Y_{ijk} - \bar{Y}_{ij\cdot})^2$	σ^2	
Total	$n-1$	$\sum_{ijk}(Y_{ijk} - \bar{Y}_{\cdots})^2$		

$H_0: A' = \cdots = A^{a-1} = 0$ — main effect A

$H_0: B_1' = \cdots = B_1^{b_1-1} = 0$ — overall test for factor B.

$B_2' = \cdots = B_1^{b_2-1} = 0$

$B_a' = \cdots = B_a^{b_a-1} = 0$

⑦ Mixed effect (nested, balanced)

Models: $Y_{ijk} = \mu + \alpha_i + \beta_{j(i)} + \varepsilon_{ijk}$

$\varepsilon_{ijk} \overset{iid}{\sim} N(0,\sigma^2)$, $\beta_{j(i)} \overset{iid}{\sim} N(0,\sigma^2_{B(A)})$ } mutually independent

$i = 1, \cdots, a \quad j = 1, \cdots, b \quad k = 1, \cdots, \tilde{n}$

Estimates: $\hat{\mu + \alpha_i} = \bar{Y}_{i\cdot\cdot}$

$\hat{\mu + \alpha_i} = \bar{Y}_{i\cdot\cdot}$

Source	df	SS	E(MS)	F
α	$a-1$	$\sum_{ijk}(\bar{Y}_{\cdots} - \bar{Y}_{i\cdot\cdot})^2$	$\sigma^2 + \tilde{n}\sigma^2_{B(A)} + b\tilde{n}Q(\alpha)$	$\frac{MSA}{MSB(A)} \sim F_{a-1,a(b-1)}\left(\frac{(a-1)b\tilde{n}Q(\alpha)}{2(\sigma^2+\tilde{n}\sigma^2_{B(A)})}\right)$
$\beta(\alpha)$	$a(b-1)$	$\sum_{ijk}(\bar{Y}_{i\cdot\cdot} - \bar{Y}_{ij\cdot})^2$	$\sigma^2 + \tilde{n}\sigma^2_{B(A)}$	$\frac{MSB(A)}{MSE}\left(\frac{\sigma^2}{\sigma^2+\tilde{n}\sigma^2_{B(A)}}\right) \sim F_{a(b-1),n-ab}$
Error	$n-ab$	$\sum_{ijk}(Y_{ijk} - \bar{Y}_{ij\cdot})^2$	σ^2	
Total	$n-1$	$\sum_{ijk}(Y_{ijk} - \bar{Y}_{\cdots})^2$		

$H_0: \sum_i c_i\alpha_i = 0$ — main effect of A with contrast c → $H_0: \alpha_1 = \cdots = \alpha_a$

$H_0: \sigma^2_{B(A)} = 0$

⑧ Random effect (nested)

Models: $Y_{ijk} = \mu + A_i + B_{j(i)} + \varepsilon_{ijk}$ $\quad i=1,\cdots,a \quad j=1,\cdots,b \quad k=1,\cdots,\tilde{n}$

$\varepsilon_{ijk} \overset{iid}{\sim} N(0,\sigma^2)$, $B_{j(i)} \overset{iid}{\sim} N(0,\sigma^2_{B(A)})$, $A_i \overset{iid}{\sim} N(0,\sigma^2_A)$ } mutually independent

Source	df	SS	E(MS)	F
A	$a-1$	$\sum_{ijk}(\bar{Y}_{...}-\bar{Y}_{i..})^2$	$\sigma^2+\tilde{n}\sigma^2_{B(A)}+b\tilde{n}\sigma^2_A$	$\frac{MSA}{MSB(A)}\left(\frac{\sigma^2+\tilde{n}\sigma^2_{B(A)}}{\sigma^2+\tilde{n}\sigma^2_{B(A)}+b\tilde{n}\sigma^2_A}\right)\sim F_{a-1,\,a(b-1)}$
B(A)	$a(b-1)$	$\sum_{ijk}(\bar{Y}_{i..}-\bar{Y}_{ij.})^2$	$\sigma^2+\tilde{n}\sigma^2_{B(A)}$	$\frac{MSB(A)}{MSE}\left(\frac{\sigma^2}{\sigma^2+\tilde{n}\sigma^2_{B(A)}}\right)\sim F_{a(b-1),\,n-ab}$
Error	$n-ab$	$\sum_{ijk}(Y_{ijk}-\bar{Y}_{ij.})^2$	σ^2	
Total	$n-1$	$\sum_{ijk}(Y_{ijk}-\bar{Y}_{...})^2$		

$H_0: \sigma^2_A = 0$

$H_0: \sigma^2_{B(A)} = 0$

ANCOVA

Models: $Y_{ij} = \mu + \alpha_i + \beta X_{ij} + \varepsilon_{ij}$ $\quad i=1,\cdots,k \quad j=1,\cdots,n_i$ $\quad \varepsilon_{ij} \overset{iid}{\sim} N(0,\sigma^2)$

Estimates: $\hat{\beta}$ by SLR

$\hat{\mu+\alpha_i} = \bar{Y}_{i.} - \hat{\beta}\bar{X}_{i.}$

$\sum_i c_i\hat{\alpha_i} = \sum_i c_i\bar{Y}_{i.} - \hat{\beta}\sum_i c_i\bar{X}_{i.}$

$\hat{\mu+\alpha_i} + \hat{\beta}X_{ij} = Y_{ij}$

$\hat{\mu+\alpha_i} + \hat{\beta}\bar{X}_{..} = \bar{Y}_{i.} - \hat{\beta}(\bar{X}_{i.} - \bar{X}_{..})$

Source	df	SS	E(MS)	F
α	$k-1$	$\sum_{ij}(\bar{Y}_{..}-\hat{Y}_{i.})^2$	$\sigma^2 + \frac{1}{k-1}\sum_i\sum_j(\alpha_i-\bar{\alpha}_.)^2$	
β	1	$\sum_{ij}(\bar{Y}_{..}-\hat{Y}_{.j})^2$	?	
Error	$N-K-1$	$\sum_{ij}(Y_{ij}-\hat{Y}_{ij})^2$	σ^2	
Total	$N-1$	$\sum_{ij}(Y_{ij}-\bar{Y}_{..})^2$		

$H_0: \beta = 0$ and $\alpha_1 = \cdots = \alpha_k$ Global F test

$H_0: \alpha_1 = \cdots = \alpha_k$ treatment test

$H_0: \beta = 0$ covariate test

Two-way effect model

① Two-way effect model (crossed)

Models: $Y_{ijk} = \mu + \alpha_{ij} + \varepsilon_{ijk}$

$i = 1, \cdots, a \quad j = 1, \cdots, b, \quad k = 1, \cdots, n_{ij} \qquad \varepsilon_{ijk} \overset{iid}{\sim} N(0, \sigma^2)$

if $n_{ij} = \tilde{n}$, balanced $\Rightarrow$ complete

Estimates:

$\hat{\mu + \alpha_{ij}} = \bar{Y}_{ij\cdot}$ $\qquad A_j = \sum_{i=1}^{a} c_i \alpha_{ij} \Rightarrow \hat{A_j} = \sum_{i=1}^{a} c_i \bar{Y}_{ij\cdot}$ A simple effect at j

$\hat{\mu + \bar{\alpha}_{\cdot\cdot}} = \bar{Y}_{\cdots}$ $\qquad A = \frac{1}{b}\sum_{j=1}^{b} A_j \Rightarrow \hat{A} = \frac{1}{b}\sum_{j=1}^{b}\sum_{i=1}^{a} c_i \bar{Y}_{ij\cdot}$ A main effect

$\hat{\mu + \bar{\alpha}_{i\cdot}} = \bar{Y}_{i\cdot\cdot}$ $\qquad AB = \sum_{j=1}^{b} d_j A_j \Rightarrow \widehat{AB} = \sum_j \sum_i c_i d_j \bar{Y}_{ij\cdot}$ AB interaction effect

$\hat{\mu + \bar{\alpha}_{\cdot j}} = \bar{Y}_{\cdot j\cdot}$ $\qquad$ When $a > 2$ or $b > 2$, more than one contrasts available.

Contrast table: (2 x 3 factorial)

	A	
	1	2
c_1	1	-1

	B		
	1	2	3
d_1	1	-1	0
d_2	1	0	-1

	α_{11}	α_{12}	α_{13}	α_{21}	α_{22}	α_{23}	fraction
A'	1	1	1	-1	-1	-1	1/3
B' (main effect)	1	-1	0	1	-1	0	1/2
B^2 (main effect)	1	0	-1	1	0	-1	1/2
A'B' (interaction effect)	1	-1	0	-1	1	0	
$A'B^2$ (interaction effect)	1	0	-1	-1	0	1	

Source	df	SS	E(MS)	F
Model	ab-1	$\sum_{ijk}(\bar{Y}_{\cdots} - \bar{Y}_{ij\cdot})^2$	$\sigma^2 + \tilde{n}Q(\alpha)$	$\frac{MSM}{MSE} \sim F_{ab-1, n-ab}\left((ab-1)\frac{\tilde{n}Q(\alpha)}{2\sigma^2}\right)$
Error	n-ab	$\sum_{ijk}(\bar{Y}_{\cdots} - Y_{ijk})^2$	σ^2	
Total	n-1	$\sum_{ijk}(Y_{ijk} - \bar{Y}_{\cdots})^2$		

$H_0: A' = 0$ No A main effect $\quad df = a - 1 = 1$

$H_0: B' = B^2 = 0$ No B main effect $\quad df = b - 1 = 2$

$H_0: A'B' = A'B^2 = 0$ NO AB interaction effect $\quad df = (a-1)(b-1) = 2$

② Two-way effect model (nested)

Models: $Y_{ijk} = \mu + \tau_{j(i)} + \varepsilon_{ijk} \qquad \varepsilon_{ijk} \overset{iid}{\sim} N(0, \sigma^2)$

$i = 1, \cdots, a \quad j(i) = 1, \cdots, b_i \quad k = 1, \cdots, n_{ij}$

Estimates: $\sum_i c_i \hat{\bar{\tau}}_{\cdot(i)} = \sum_i c_i \bar{Y}_{i\cdot\cdot}$ A main effect $\quad$ since nested, B main effect

$\sum_j d_j \hat{\tau}_{j(i)} = \sum_j d_j \bar{Y}_{ij\cdot}$ B simple effect ← doesn't make sense.

Block Design :

Models: $Y_{ij} = \mu + \alpha_i + \varepsilon_{ij}$ $i = 1, \cdots, I$ $j = 1, \cdots, \frac{N}{I}$ $\varepsilon_{ij} \overset{iid}{\sim} N(0, \sigma^2)$ Completely randomized design (CRD)

Models: $Y_{hi} = \mu + \beta_h + \tau_i + \varepsilon_{hi}$ $\varepsilon_{hi} \overset{iid}{\sim} N(0, \sigma^2)$ Randomized complete block design (RCBD)

$h = 1, \cdots, b$ ↑ block effect $\quad i = 1, \cdots, t$ ← treatment effect ← no replicas per combination thus no interaction → treatment applied on all blocks

Estimates: $\sum_i \hat{c_i \tau_i} = \sum_i c_i \bar{Y}_{\cdot i}$ (BLUE)

Source	df	SS	E(MS)	F
Block	$b-1$	$\sum_{hi} (\bar{Y}_{\cdot\cdot} - \bar{Y}_{h\cdot})^2$	$\sigma^2 + tQ(\beta)$	Typically not interested in this
Treatment	$t-1$	$\sum_{hi} (\bar{Y}_{\cdot\cdot} - \bar{Y}_{\cdot i})^2$	$\sigma^2 + bQ(\tau)$	$\frac{MSI}{MSE} \sim F_{t-1,(t-1)(b-1)}\left((t-1)\frac{bQ(c)}{2\sigma^2}\right)$
Error	$(t-1)(b-1)$	$\sum_{hi} (\bar{Y}_{\cdot\cdot} - \bar{Y}_{h\cdot} - \bar{Y}_{\cdot i} + Y_{hi})^2$	σ^2	
Total	$tb-1$	$\sum_{hi} (Y_{hi} - \bar{Y}_{\cdot\cdot})^2$		

→ only in the blocking case, SSE isn't in easy form.

Models: $Y_{hi} = \mu + \beta_h + \tau_i + \varepsilon_{hi}$ $\varepsilon_{hi} \overset{iid}{\sim} N(0, \sigma^2)$ Incomplete block design (IBD)

$h = 1, \cdots, b$ $\quad i = 1, \cdots, t$ when $n_{hi} = 1$

Estimates: Relies on softwares to construct BLUE for $\sum_i c_i \tau_i$

Problems: 1. May not be able to estimate all treatment contrast.

2. May have pairwise contrasts with different variances.

Models: Balanced incomplete block design (BIBD)

Rules: b: block number

k: experimental units in one block

t: treatment number

Each treatment occurs in exact $r = \frac{bk}{t}$ blocks

Each pair of treatment appears in $\lambda = \frac{r(k-1)}{t-1}$ blocks

→ Thus all pairwise contrasts have equal variance

df same as RCBD

Models: $Y_{ijk} = \mu + \beta_i + \gamma_j + \tau_k + \varepsilon_{ijk}$ $\varepsilon_{ijk} \overset{iid}{\sim} N(0, \sigma^2)$ Latin Square Designs (LSD)

$i = 1, \cdots, k$ $\quad j = 1, \cdots, k$ $\quad k = 1, \cdots, t$ where $n_{ijk} = 1$ $\quad \sum_i n_{ijk} = \sum_j n_{ijk} = 1$; $\sum_{ij} n_{ijk} = k$

Estimates:

$$\hat{\mu + \bar{\beta}_{\cdot} + \bar{\gamma}_{\cdot}} + \tau_k = \bar{Y}_{\cdot\cdot k}$$

$$\sum_k c_k \tau_k = \sum_k c_k \bar{Y}_{\cdot\cdot k}$$

Split-plot design (CRD-RCBD)

Models:

Whole plot: $Y_{ijk} = \mu + \alpha_i + W_{j(i)} + \varepsilon_{ijk}$ ($W_{j(i)}$: random blocking factor)

$i = 1, \cdots, t \quad j = 1, \cdots, r \quad k = 1, \cdots, s$ (j: WP main effect)

$W_{j(i)} \overset{iid}{\sim} N(0, \sigma_w^2)$, $\varepsilon_{ijk} \overset{iid}{\sim} N(0, \sigma_s^2)$ } mutually independent

Sub-plot: $Y_{ijk} = \mu + W_{j(i)} + \beta_k + \varepsilon_{ijk}$

$i = 1, \cdots, t \quad j = 1, \cdots, r \quad k = 1, \cdots, s$ (k: SP main effect)

$W_{j(i)} \overset{iid}{\sim} N(0, \sigma_w^2)$, $\varepsilon_{ijk} \overset{iid}{\sim} N(0, \sigma_s^2)$ } mutually independent

Full model: $Y_{ijk} = \mu + \alpha_i + W_{j(i)} + \beta_k + \alpha\beta_{ik} + \varepsilon_{ijk}$

$i = 1, \cdots, t \quad j = 1, \cdots, r \quad k = 1, \cdots, s$

$W_{j(i)} \overset{iid}{\sim} N(0, \sigma_w^2)$, $\varepsilon_{ijk} \overset{iid}{\sim} N(0, \sigma_s^2)$ } mutually independent

Source	df	SS	E(MS)	F
α	$t-1$	$\sum_{ijk}(\bar{Y}_{...} - \bar{Y}_{i..})^2$	$\sigma_s^2 + s\sigma_w^2 + rsQ(\alpha^*)$	MSA/MSW
Wonor	$t(r-1)$	$\sum_{ijk}(\bar{Y}_{i..} - \bar{Y}_{ij.})^2$	$\sigma_s^2 + s\sigma_w^2$	MSW/MSE
β	$s-1$	$\sum_{ijk}(\bar{Y}_{...} - \bar{Y}_{..k})^2$	$\sigma_s^2 + tr Q(\beta^*)$	MSB/MSE
$\alpha\beta$	$(s-1)(t-1)$	$\sum_{ijk}(\bar{Y}_{i.k} - \bar{Y}_{i..} - \bar{Y}_{..k} + \bar{Y}_{...})^2$	$\sigma_s^2 + r Q(\alpha\beta^*)$	MSAB/MSE
Error	$t(r-1)(s-1)$	$\sum_{ijk}(Y_{ijk} - \bar{Y}_{ij.} - \bar{Y}_{i.k} + \bar{Y}_{i..})^2$	σ_s^2	
Total	$trs-1$	$\sum_{ijk}(Y_{ijk} - \bar{Y}_{...})^2$		

Source	A	B	AB	W(AB)
df	$a-1$	$b-1$	$(a-1)(b-1)$	$ab(r-1)$

Source	C	AC	BC	ABC	Error	Total
df	$c-1$	$(a-1)(c-1)$	$(b-1)(c-1)$	$(a-1)(b-1)(c-1)$	$ab(r-1)(c-1)$	$abrc-1$

$H_0: \sigma_w^2 = 0$

$H_0: \alpha_i^* = 0 \;/\; \beta_j^* = 0 \;/\; \alpha\beta_{ij}^* = 0$ ← be careful with this

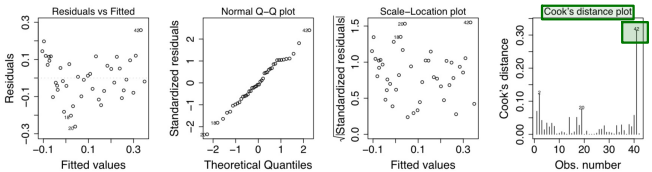
Effect Principles

1. Effect hierarchy: main effect and lower-order interactions more likely to be significant
2. Effect heredity: If interaction significant, keep parent terms
3. Effect sparsity: Few effects are significant
4. Factor sparsity: Few factors are significant

Model Diagnostics

Example Model Diagnostics

- Some default residual plots are produced by `plot(gw.mod1)`.



cook's distance

- There is a trend in the mean of the residuals, violating **independence**.
- The QQ plot is close to a straight line, so **normality** is OK.
- The residual magnitudes seem consistent with **constant variance**.
- The 42nd observation has a very high influence on the results.

unusual data point

Definition:

$\hat{Y} = X\hat{\beta} = \underbrace{X(X'X)X'}_{P:\text{ hat matrix}} Y = PY$ ← projection of Y onto the column space of "X", $\hat{Y} \in Col(X)$

↳ $\hat{Y}_i = P_{i1}Y_1 + \cdots + P_{i,i-1}Y_{i-1} + P_{ii}Y_i + P_{i,i+1}Y_{i+1} + \cdots + P_{in}Y_n$

By examine the elements of P, get to know the influence of individual response values on fitted values.

P_{ii}: leverage of ith component: leverage ↑ influence on fit ↑

centered design matrix with intercept in model →

1. $\frac{1}{n} \le P_{ii} \le 1$, $P_{ii} = \sum_{j=1}^{n} P_{ij}^2$ ← in SLR, $P_{ii} = \frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum_i (x_i - \bar{x})^2}$
2. If $P_{ii} = 1$, then rest of row(& column) equals 0: $\hat{Y}_i = Y_i$
3. If $P_{ii} = 0$, then rest of row(& column) equals 0: $\hat{Y}_i = 0$
4. $-1 \le P_{ij} \le 1$
5. P is determined entirely by the design matrix X, and not at all by Y.

P properties:

1. P is symmetric even if $(X'X)^-$ is not. ← orthogonal projection
2. P is invariant of $(X'X)^-$ choice since P unique orthogonal projection, $(X'X)^-$ varies
3. $X = PX \Rightarrow X(X'X)^-X'X = X \Rightarrow (X'X)^-X'X = X^-$
   $X'P = X' \Rightarrow X'X(X'X)^-X' = X' \Rightarrow X(X'X)^- = (X')^-$
4. $\text{rank}(P) = \text{trace}(P) = \text{rank}(X)$, where $\text{trace}(P) = \sum_i P_{ii}$
5. If $\underline{1 \in Col(X)}$, then each row (column) of P sums to 1. $\sum_i P_{ij} = \sum_j P_{ij} = 1$

↑ with intercept

Potential issues in regression

1. Assumption violation

Assumptions: 1.1: X is known, without error

1.2: $E(\varepsilon)=0$, i.e. each model is correctly specified

1.3: $Var(\varepsilon)=\sigma I$, i.e. equal variance, uncorrelated

1.4: $\varepsilon \sim N(\cdot)$

2. Unusual Data

3. Computational instability

1.1. X observed with error (measurement error, errors-in-variables models)

- Estimators biased: usually towards 0
- Affects everything
- instrumental variable may help

1.2. Mean model is misspecified

- underspecification
- Model not additive
- Nonlinear relationship
- Diagnose: Plot e_i v.s. $\hat{Y}_i$ should show random scatter.

(a). $\Sigma_i e_i = 0$ if model includes intercept

(b). $\Sigma_i e_i \hat{Y}_i = 0$ (a) & (b) $\Rightarrow \hat{Corr}(e, \hat{Y}) = 0$ $\Rightarrow$ only implies no linear relationship

(c). $\Sigma_i e_i X_{ij} = 0$ (a) & (c) $\Rightarrow \hat{Corr}(e, X_j) = 0$

Plot e_i v.s. X_j should show random scatter.

Partial Residual (leverage) plots should show straight lines.

has more information than usual residual plot (when no much multicollinearity)

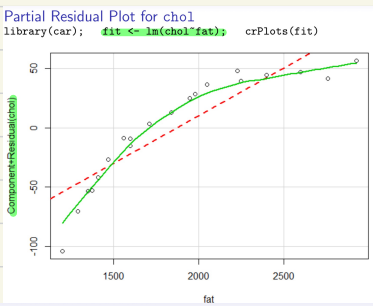
e.g. straight line says add X_j to model

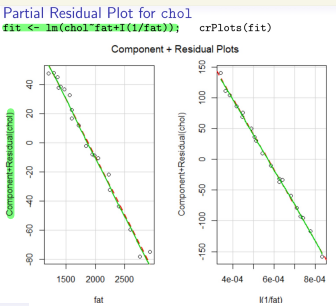
Nonlinear relationship says to transform X_j before adding to model.

Partial Residual for X: $e^* = Y - (\hat{\beta_0} + \hat{\beta_1} X_1 + \cdots + \hat{\beta_{j-1}} X_{j-1} + \cancel{\hat{\beta_j} X_j} + \hat{\beta_{j+1}} X_{j+1} + \cdots + \hat{\beta_p} X_p)$

$= Y - \hat{Y} + \hat{\beta_j} X_j$

$= e + \hat{\beta_j} X_j$





Simulation Study: Effect of Misspecifying the Mean Model

Truth: $\boldsymbol{Y} = \boldsymbol{X}\beta + \boldsymbol{Z}\gamma + \epsilon \quad \Rightarrow \quad \boldsymbol{Y} \sim \boldsymbol{N}(\boldsymbol{X}\beta + \boldsymbol{Z}\gamma, \sigma^2\boldsymbol{I})$

Assume: $\boldsymbol{Y} = \boldsymbol{X}\beta + \epsilon$ to get $\hat{\beta} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y}$

Then

	$\gamma = 0$	$Z = XB$	$X^TZ = 0$	other
$E(\hat{\beta}) = \beta$	√		√	
$E(\hat{Y}) = X\beta + Z\gamma$	√	√		
$E(\hat{Y}) = X\beta$	√		√	

Simulation Setup:

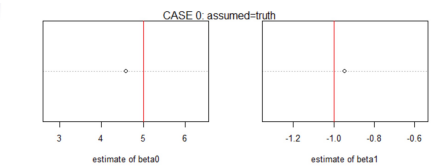
$n = 25, \quad \boldsymbol{X}$ is $n \times 2, \quad \beta = \begin{bmatrix} 5 \\ -1 \end{bmatrix}, \quad \boldsymbol{Z}$ is $n \times 1, \quad \gamma = 5$

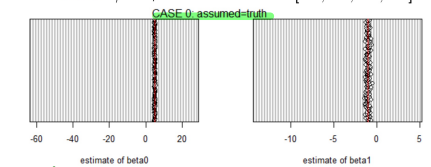
Case 0: $\gamma = 0$ . . . demonstrates usual properties

Case 1: $\gamma = 5$, $\boldsymbol{Z}$ in column space of $\boldsymbol{X}$ . . . $\boldsymbol{Z} = \boldsymbol{X}\begin{bmatrix} 4 \\ 1 \end{bmatrix}$

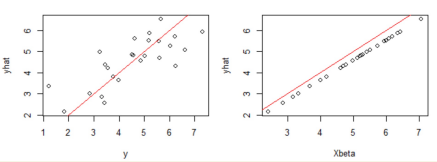
Case 2: $\gamma = 5$, columns of $\boldsymbol{Z}$ orthogonal to columns of $\boldsymbol{X}$ . . . $\boldsymbol{Z} = (\boldsymbol{I} - \boldsymbol{P_X})(4 + N(0,1))$

Case 3: $\gamma = 5$, other . . . $\boldsymbol{Z} = [-1, -2, \ldots, -n]^T$

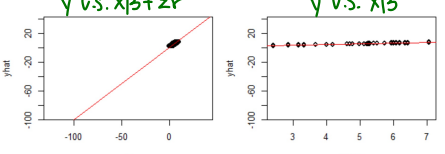




unbiased $\hat{\beta}$

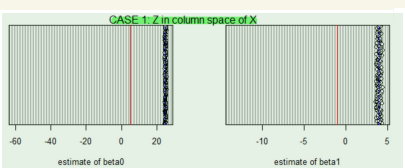


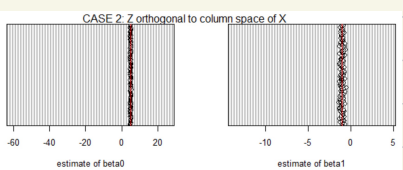
$\hat{y}$ v.s. $X\beta + Z\gamma$ | $\hat{y}$ v.s. $X\beta$



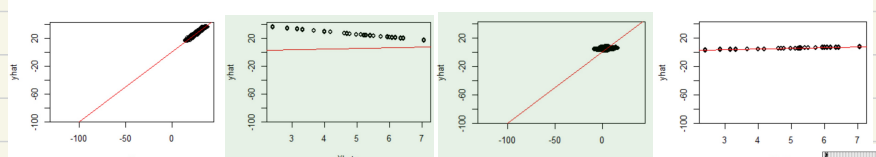
$\hat{y}$ v.s. $E(Y)$

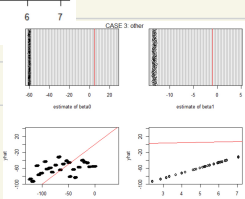
$\hat{\beta}$ biased





unbiased $\hat{\beta}$





Note: 1. If want to have $E(\hat{\beta}) = \beta$, design $X'Z = 0$

- $E(\hat{\beta}) = \beta$
- $E(\hat{Y}) = X\beta$
- e can be used to estimate γ as $e \sim N(Z\gamma, \sigma^2(I-P))$

2. If want to have $E(\hat{Y}) = X\beta + Z\gamma$, let $Z \in C(X)$

Thus "estimate β" and "estimate $E(Y)$" have different requirements

1.3 (i) Heterogeneity i.e. unequal variance

- estimator still unbiased, but not best
- ANOVA sum of squares still ok
- standard errors are wrong, tests, CI wrong
- Transformation may fix ★

Truth: $Y = X\beta + \varepsilon \quad \varepsilon \sim N(0, \sigma^2 V)$

$Y = X\beta + \varepsilon \quad \varepsilon \sim N(0, \sigma^2 I)$

(ii). Correlated errors

- estimator still unbiased, but not best
- ANOVA sum of squares still ok
- standard errors are wrong, tests, CI wrong
- e.g: time series, spatial, split plot, subsampling
- Other estimation procedure may help

1.4 Checking for normality

Why not use $e = \hat{y} - y$ to check for normality?

1. Residuals are not independent

$Cov(e_i, e_j) = -\sigma^2 P_{ij}$

2. Residuals are not identically distributed

$Var(e_i) = \sigma^2(1 - P_{ii})$ not constant

3. Residuals are not a random sample from error distribution ε

Standardized residual: $r_i = e_i / \hat{\sigma}\sqrt{1 - P_{ii}}$

Studentized residual: $r_i = e_i / \hat{\sigma}_{(i)}\sqrt{1 - P_{ii}}$

Deleted/Jackniff/LOOCV results:

1. Assumptions violated:

1.3 (i) Heterogeneity, i.e., unequal variances

1.3 (ii) Correlated errors

Truth: $\boldsymbol{Y} = \boldsymbol{X}\beta + \epsilon, \quad \epsilon \sim \boldsymbol{N}(\boldsymbol{0}, \sigma^2\boldsymbol{V}) \quad \Rightarrow \quad \boldsymbol{Y} \sim \boldsymbol{N}(\boldsymbol{X}\beta, \sigma^2\boldsymbol{V})$

Assume: $\boldsymbol{Y} = \boldsymbol{X}\beta + \epsilon, \quad \epsilon \sim \boldsymbol{N}(\boldsymbol{0}, \sigma^2\boldsymbol{I})$ to get $\widehat{\beta} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y}$

Then

- $\mathrm{E}(\widehat{\beta}) = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{X}\beta = \beta$ unbiased
- $\mathrm{E}(\widehat{\boldsymbol{Y}}) = \boldsymbol{X}\mathrm{E}(\widehat{\beta}) = \boldsymbol{X}\beta$ unbiased
- $\mathrm{E}(\boldsymbol{e}) = (\boldsymbol{I} - \boldsymbol{P})\mathrm{E}(\boldsymbol{Y}) = \boldsymbol{0}$ unbiased for 0
- $\mathrm{var}(\widehat{\beta}) = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\sigma^2\boldsymbol{V}\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}$
  is not as "small" as possible among unbiased estimators

$\mathrm{cov}(\widehat{\boldsymbol{Y}}, \boldsymbol{e}) = \boldsymbol{P}\sigma^2\boldsymbol{V}(\boldsymbol{I} - \boldsymbol{P}) \neq \boldsymbol{0}$ correlated

even though $\widehat{\boldsymbol{Y}}^T\boldsymbol{e} = \boldsymbol{Y}^T\boldsymbol{P} \cdot (\boldsymbol{I} - \boldsymbol{P})\boldsymbol{Y} = \boldsymbol{0}$ orthogonal

Possible remedial actions . . .

- Use Weighted Least Squares (WLS)
  if know $\text{var}(Y) = a\sigma^2$, know a, know independent
- Use Interatively Reweighted Least Squares (IRLS)
  if know $\text{var}(Y) = f[\text{E}(Y)]$, know independent
- Use Estimated Generalized Least Squares (EGLS)
- Use Generalized Linear Model (GLM)
- Transform!

There is the Box-Cox Family of Transformations:

$$Y_i^{(\lambda)} = \begin{cases} \frac{Y_i^{\lambda}-1}{\lambda \dot{Y}^{\lambda-1}} & \lambda \neq 0 \\ \dot{Y} \ln(Y_i) & \lambda = 0 \end{cases} \qquad \text{where} \quad \dot{Y} = (Y_1 Y_2 \cdots Y_n)^{1/n}$$

Denominator and -1 in numerator are just for scaling.

Converts scale back to original units, and thus allows direct comparison of SSE across models with different powers.

- Fit ANOVA for several values of λ, e.g., $\lambda = -1, -.9, \ldots, 1$. Record $SSE^{(\lambda)}$ for each value of λ.
- Plot $\left\{\lambda, SSE^{(\lambda)}\right\}$ and determine $SSE_{\text{min}}^{(\lambda)}$
- In the end, choose any value of λ that causes

$$SSE^{(\lambda)} \leq SSE_{\text{min}}^{(\lambda)} \left[1 + \frac{t^2_{df_e,\alpha/2}}{df_e}\right],$$

where df_e is the degrees of freedom associated with any $SSE^{(\lambda)}$

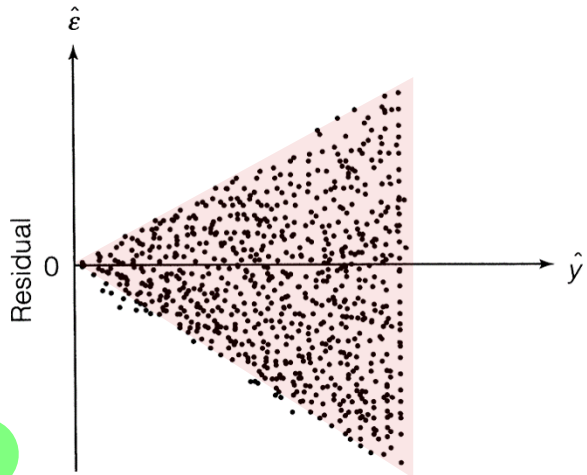
Note: $\widehat{\boldsymbol{\beta}}^{(\lambda)}$ can be very different as λ changes! Also sensitive to $\boldsymbol{X}\boldsymbol{\beta}$.

Other Guidance on Choosing a Transformation

If $V(Y) \propto [E(Y)]^{2k}$ then transform to Y^{1-k}, where "$Y^0 = \ln(Y)$".

$k = 1 : V(Y) \propto \{E(Y)\}^2$

- Use transformation $Y^{1-k} = Y^0 = \ln(Y)$
- Great if $Y \sim \text{Gamma}(\alpha, \beta)$, with $V(Y) = [E(Y)]^2/\alpha$
  - Example: $Y =$ survival time of mice subjected to a treatment

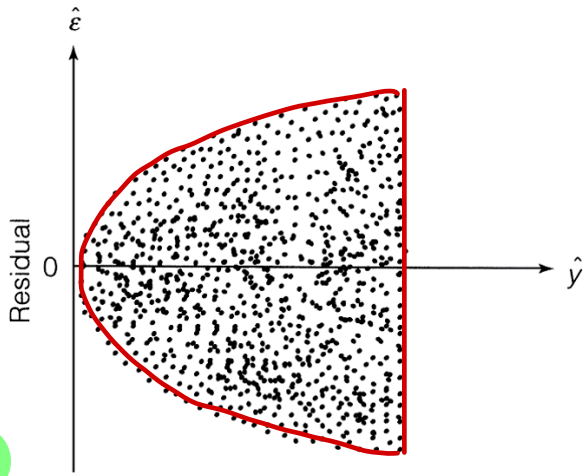


- Plot e_i versus $\widehat{Y}_i$ is fan-shaped
- *Possible alternative approach:* *generalized linear model*

If $V(Y) \propto [E(Y)]^{2k}$ then transform to Y^{1-k}, where "$Y^0 = \ln(Y)$".

$k = 0.5 : \; V(Y) \propto E(Y)$

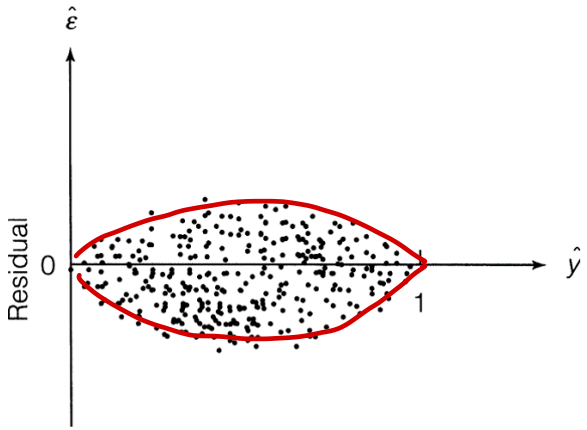
- Use transformation $Y^{1-k} = \sqrt{Y}$
- Great if $Y \sim$ Poisson, with $V(Y) = E(Y)$
  - Example: $Y = \#$ trees in 1000 acres of a forested area
    This is a count having a very large (possible "infinite") upper bound



- Plot e_i versus $\widehat{Y}_i$ is fan-shaped
- *Possible alternative approach:* *loglinear regression model*

$V(Y) \propto \{E(Y)\}\{1 - E(Y)\}$

- Use transformation $\arcsin(\sqrt{Y}) = \sin^{-1}(\sqrt{Y})$
- Great if $nY \sim$ Binomial, with $V(Y) = [E(Y)][1 - E(Y)]/n$
  - Example: $Y =$ proportion of 30 trees that are afflicted by a fungus
  - (arcsin good when $E(Y) < 0.3$ or $E(Y) > 0.7$)
- Plot e_i versus $\widehat{Y}_i$ has bulge for $\widehat{Y}$ close to 0.5



- *Possible alternative approach:* *logistic regression*

Unequal variances often coexist with nonnormality!

1. Assumptions violated:

1.3 (ii) Correlated errors:

- Estimators are still unbiased, but not best.
- ANOVA sum-of-squares are still ok
- Standard errors are wrong, and hence tests and CIs are wrong
- Creates the problem: Time series, spatial, split plot, subsampling
- Tests
- Possible fix: Estimation procedures other than OLS

How to diagnose?

- Be guided by the type of data
- Look for patterns among residuals over time/space/etc.

Detecting Correlation

Positively(Negatively) correlated data can lead to *standard errors that are seriously under-(over-)estimated*, thus drastically affecting hypothesis tests and confidence intervals.

Durbin-Watson Test of Autocorrelation

Durbin-Watson test statistic:

$$d = \frac{\sum_{i=2}^{n}(e_i - e_{i-1})^2}{\sum_{i=1}^{n} e_i^2} \approx 2(1-\widehat{\rho}) \approx \begin{cases} 0 & \text{strong positive autocorrelation} \\ 2 & \text{no autocorrelation} \\ 4 & \text{strong negative autocorrelation} \end{cases},$$

where $\widehat{\rho}$ is the sample correlation between e_i and e_{i-1}

- Ordering of the data matters. Looking for correlation with immediate neighbors, following sequence.
- Null distribution complicated

Example: 35-year sales history of a company

Y : annual sales, in thousands of dollars [SALES]

X : year [T]

OLS Regression:

$\widehat{\beta}_0 = 0.402$, with $se(\widehat{\beta}_0) = 2.206$

$\widehat{\beta}_1 = 4.296$, with $se(\widehat{\beta}_1) = 0.107$

AR(1) Regression:

$\widehat{\beta}_0 = 0.422$, with $se(\widehat{\beta}_0) = 3.670$

$\widehat{\beta}_1 = 4.295$, with $se(\widehat{\beta}_1) = 0.179$

```
proc reg data=sales35;
  model sales=t / dwprob; **d=0.821, rho=0.590, pval[Ha:+ve corr] is <.0001;
proc arima data=sales35;
  identify var=sales crosscorr=t;
  estimate p=1 input=t; run;
```

```
fit = lm(SALES ~ T, data=SALES35); summary(fit)
library(car); durbinWatsonTest(fit)

# AR(1) regression:
fit = arima(SALES35$SALES, order=c(1, 0, 0), xreg = SALES35$T); fit
```

Check for Normality

- **Checking for normality ranks low relative to other checks**
  - Expectation & variance of estimators and sums of squares are unaffected by nonnormality
  - ANOVA F-test is reasonably robust to nonnormality
  - HTs and CIs are more affected by nonnormality, but robust in large samples
- Testing for normality can be overkill. Instead,
  - use histogram, with normal curve overlaid
  - use normal quantile-quantile (Q-Q) plot and look for pattern[1]
    - straight line $\rightsquigarrow$ normal distribution (intercept is mean, slope is std. dev.)
    - "S" $\rightsquigarrow$ symmetric, light-tailed distribution
    - "tangent" $\rightsquigarrow$ symmetric, heavy-tailed dist'n. **Unequal variances? Outliers?**
    - "J" $\rightsquigarrow$ positively skewed dist'n. **Log transformation?**
    - "r" $\rightsquigarrow$ negatively skewed dist'n
    - line not through origin $\rightsquigarrow$ missing important predictor variable

[1] heavily dependent on sample size!

- Is $\boldsymbol{e}$ a good choice for testing normality?
  - Is it true that $e_1, \ldots, e_n$ forms a random sample?

    No because $\text{var}(e_i) = \sigma^2(1 - P_{ii})$ & $\text{Cov}(e_i, e_j) = -\sigma^2 P_{ij}$
- internally studentized residual

  (R=standardized residual, SAS=studentized residual):

  $$r_i = \frac{e_i}{s\sqrt{1 - P_{ii}}}, \quad s = \sqrt{MSE}, \quad MSE \text{ from regn with all } n \text{ obs}$$

  * $\text{var}(r_i) \approx 1$.
  * r_i, r_j are likely dependent
  * $r_i \approx t_{df_e}$ (dependent numerator & denominator)
- externally studentized residual or studentized deleted residual

  (R & SAS=rstudent residual):

  $$r_i^* = \frac{Y_i - \widehat{Y}_{i(i)}}{s_{(i)}\sqrt{1 + \boldsymbol{x}_i(\boldsymbol{X}_{(i)}^T \boldsymbol{X}_{(i)})^{-1}\boldsymbol{x}_i^T}} = \frac{e_i}{s_{(i)}\sqrt{1 - P_{ii}}}$$

  * $\text{var}(r_i^*) \approx 1$.
  * r_i^*, r_j^* are likely dependent
  * $r_i^* \approx t_{df_e - 1}$ (better approx than r_i) (df_e is from regression with all n observations)
  * r_i^* reflects large values more dramatically than r_i (Atkinson 1983)

- deleted/Jackknife/LOOCV results: run regression n times, with the ith observation excluded during the ith run: (Belsley, Kuh, Welsch 1980)
  - $\boldsymbol{X}_{(i)}$ is new design matrix from omitting the ith observation
  - $s^2_{(i)} = MSE_{(i)}$ is the MSE from omitting the ith observation
    No need to rerun the regn: $(n-p-1)s^2_{(i)} = (n-p)s^2 - r_i^2/(1-P_{ii})$
  - $\widehat{\boldsymbol{\beta}}_{(i)}$ is estimate of full p-dimensional vector $\boldsymbol{\beta}$ from omitting the ith obs.
  - $\widehat{\beta}_{j(i)}$ is estimate of parameter β_j from omitting the ith obs.
  - $\widehat{\boldsymbol{Y}}_{(i)}$ is prediction of full n-dimensional vector $\boldsymbol{Y}$ from omitting the ith obs.
  - $\widehat{Y}_{i(i)}$ is prediction of ith observation Y_i from omitting the ith obs.
  - $Y_i - \widehat{Y}_{i(i)}$ is called the ith deleted residual
  - r_i^* is the t-statistic for agreement between Y_i and $\widehat{Y}_{i(i)}$:

$$r_i^* = \frac{Y_i - \widehat{Y}_{i(i)}}{\text{s.e.}(Y_i - \widehat{Y}_{i(i)})} = \frac{Y_i - \widehat{Y}_{i(i)}}{s_{(i)}\sqrt{1 + \boldsymbol{x}_i(\boldsymbol{X}_{(i)}^T\boldsymbol{X}_{(i)})^{-1}\boldsymbol{x}_i^T}} = \frac{Y_i - \widehat{Y}_i}{s_{(i)}\sqrt{1-P_{ii}}}$$

Creating and interpreting a normal Q-Q plot:

- y-axis: residual, e.g. r_i^*, in increasing order
  x-axis: quantiles from a normal distribution
  For jth ordered residual, plot $\left\{\Phi^{-1}\left[\frac{j-3/8}{n+1/4}\right], j\text{th ordered residual}\right\}$
- Good when $n \geq 30$, better when $n \geq 50$. (theory says ok when $n \geq 5$)
- Look for pattern

How to correct nonnormality?

TRANSFORMATION, e.g. Box-Cox

Make variances equal

Address outliers

Illustration

Salary as a function of years of experience

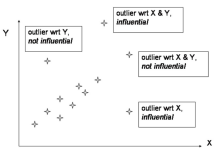
3. Unusual Data

(i) Outlier

(ii) Influential point

Outliers and Influential data points . . .*

- **Outlier**: data point that is "unusual" when compared to other points in the dataset. May not be influential. Can be outlier wrt X (aka high leverage) or wrt Y.
- **Influential data point:** has undue influence in that inference can change drastically depending on whether that single point is included in the analysis.



What to do about outlier wrt Y?
- improved model is needed
- claim that point has other features driving response that model doesn't account for

What to do about outlier wrt X?
- try to fill in gap in X space, then refit

What to do about influential points? Fit model with and without points. Change in inference?
- Yes : report both results
- No : report results from full set

Outlier wrt Y:

Is Y_i well predicted?

- A "no" answer could result from outlier wrt Y or an inadequate model!
- Recall that r_i^* is the t-statistic for agreement between Y_i and $\widehat{Y}_{i(i)}$:

$$r_i^* = \frac{Y_i - \widehat{Y}_{i(i)}}{\text{s.e.}(Y_i - \widehat{Y}_{i(i)})} = \frac{Y_i - \widehat{Y}_{i(i)}}{s_{(i)}\sqrt{1 + \boldsymbol{x}_i(\boldsymbol{X}_{(i)}^T\boldsymbol{X}_{(i)})^{-1}\boldsymbol{x}_i^T}} = \boxed{\frac{Y_i - \widehat{Y}_i}{s_{(i)}\sqrt{1 - P_{ii}}}}$$

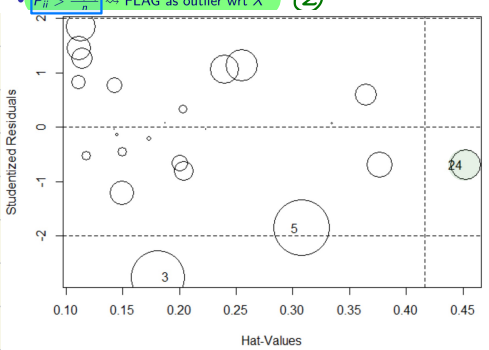
Studentized residual

- $|r_i^*| > 3$ ⇝ FLAG as outlier wrt Y (1)
  - `proc reg` in SAS applies threshold of 2 by default – too low
- Bonferroni: $|r_i^*| > t_{df_e - 1, \frac{\alpha}{2n}}$ ⇝ FLAG as outlier wrt Y
  We're actually making inferences about n residuals!
  Must make adjustment for multiple testing.
  - R has a function for this, namely `car::outlierTest`

Outlier wrt X (aka High Leverage Point):

Which data points are far from the centroid of the X-space?
Does a data point have "high leverage," i.e., large value of P_{ii}?

- P_{ii} is weight of ith observation in determining $\widehat{Y}_i$, since $\widehat{\boldsymbol{Y}} = \boldsymbol{PY}$
- $\frac{1}{n} \le P_{ii} \le \frac{1}{c}$, c is number of rows same as ith row
  - $P_{ii} = \frac{1}{n}$ only when $x_{ij} = \overline{\boldsymbol{X}}_{j\cdot}, j = 1, \ldots, p$, i.e., the ith point is at the centroid
    - model should capture this point well, if it is a good model
    - relatively speaking, this point deserves little scrutiny
    - omitting this point will cause little change in the analysis
  - $P_{ii} = 1$ only when x_i is far from the centroid
    - regression line fits the point exactly, i.e., $\widehat{Y}_i = Y_i$. Indicative of good model? Not really
    - this point may or may not be in line with the other points
    - relatively speaking, this point deserves heavy scrutiny
    - omitting this point may cause great change in the analysis
- $\sum_{i=1}^n P_{ii} = r(\boldsymbol{X})$ since $\text{tr}(\boldsymbol{P}) = r(\boldsymbol{X})$
- $P_{ii} > \frac{2r(\boldsymbol{X})}{n}$ ⇝ FLAG as outlier wrt X (2)



Influential Data Points:

Examine effect on $\widehat{\beta}$, $\widehat{\boldsymbol{Y}}$ (also possible: $\widehat{Y}_i$, $\widehat{\beta}_j$, $\widehat{\text{Var}}(\widehat{\beta})$)

- Effect on $\widehat{\beta}$, $\widehat{\boldsymbol{Y}}$: Cook's D(istance)

$$D_i = \frac{(\widehat{\beta}_{(i)} - \widehat{\beta})'\,(\boldsymbol{X}^T\boldsymbol{X})\,(\widehat{\beta}_{(i)} - \widehat{\beta})}{r(\boldsymbol{X})s^2}$$

$$= \frac{(\widehat{\boldsymbol{Y}}_{(i)} - \widehat{\boldsymbol{Y}})^T(\widehat{\boldsymbol{Y}}_{(i)} - \widehat{\boldsymbol{Y}})}{r(\boldsymbol{X})s^2} \overset{\text{computing}}{=} \boxed{\frac{r_i^2}{r(\boldsymbol{X})}\left(\frac{P_{ii}}{1 - P_{ii}}\right)}$$

$D_i > 1$ (some use > 0.8) ⇝ FLAG as influential (3)

$D_i > F_{r(\boldsymbol{X}), n - r(\boldsymbol{X}), 0.5}$ ⇝ FLAG as influential

- `proc reg` in SAS applies a different threshold by default – too low

Cook's distance

3. Computational Instability, aka Multicollinearity:
   - Do different predictor variables provide redundant information?
   - Affected by choice of values for predictors $X_1, X_2, \ldots, X_p$

Definition:

Multicollinearity exists when two or more of the predictor variables used in regression are moderately or highly correlated.

$$\boldsymbol{X} = \begin{bmatrix} 1 & 4 & 4.01 \\ 1 & 6 & 5.98 \\ 1 & 7 & 7.02 \\ 1 & 8 & 7.99 \end{bmatrix}, \qquad r(\boldsymbol{X}) = 3$$

But solving $\boldsymbol{X}^T\boldsymbol{X}\widehat{\beta} = \boldsymbol{X}^T\boldsymbol{Y}$ will be highly unstable!

NEEDS:

- Recognize when multicollinearity is a problem
- Take corrective action

Diagnosis ... ☆ this is very important

Indirect:

- Look for symptoms, although they can happen for other reasons:
  - Large changes in $\widehat{\beta}$ when predictor variable is added/omitted
  - $\widehat{\beta}_j$ has a sign opposite what is expected
  - Model has significant F but many non-significant t tests
  - Sequential SS and partial SS are very different
  - Variances of $\widehat{\beta}_j$ are very large
  - High correlation between $\widehat{\beta}_j$ and $\widehat{\beta}_k$. This may violate simple interpretation of regression coefficients as measuring change in E(Y) when a given predictor variable is increased by 1 while all others are held constant.

Direct:

- Large (simple) correlations between predictor variables
- Large variance inflation factors
- Condition number and indices, i.e., scrutinize structure of $\boldsymbol{X}^T\boldsymbol{X}$ (won't cover)

Variance Inflation Factors (VIF) . . .

- One for each predictor variable in the model, except intercept
- $VIF_j = \frac{1}{1-R_j^2}$, R_j^2 is coefficient of determination from regressing X_j on all other predictor variables (including intercept)
  - $0 \leq R_j^2 \leq 1$ implies $1 \leq VIF_j$
  - $R_j^2 \approx 0$ implies $VIF_j \approx 1$ and $R_j^2 \approx 1$ implies VIF_j very large

  X_j involved in linear dependency with other non-intercept predictor variables
  $\Rightarrow R_j^2 \approx 1 \Rightarrow VIF_j$ very large
- $\text{var}(\widehat{\beta}_j) = \sigma^2 \cdot VIF_j$† : jth variable $\perp$ all others $\Rightarrow VIF_j = 1$
- FLAG: $VIF_j > 10$ indicates problem, $VIF_j > 30$ severe problem
- *Unaffected by centering predictor variables*
- *Detects overall collinearity problems with more than just the intercept*
- *No direct indication of:*
  - *number of linear dependencies*
  - *which other variables are involved in linear dependency with X_j*

†assumes centered & scaled predictors

Possible Corrective Actions . . .

- Drop one or more of the correlated predictor variables
- If you decide to keep all predictor variables in the model:
  - Avoid making inferences about the individual β values, and don't try to determine "relative importance" of the predictor variables
- Sacrifice unbiasedness to get smaller variance $\rightsquigarrow$ **Biased Regression**:
  - Principal Components Regression (PCR)
  - Partial Least Squares Regression (PLSR)
  - Ridge Regression
  - LASSO, ElasticNet, etc.

Example: Cigarettes [dataset FTCCIGAR.txt]

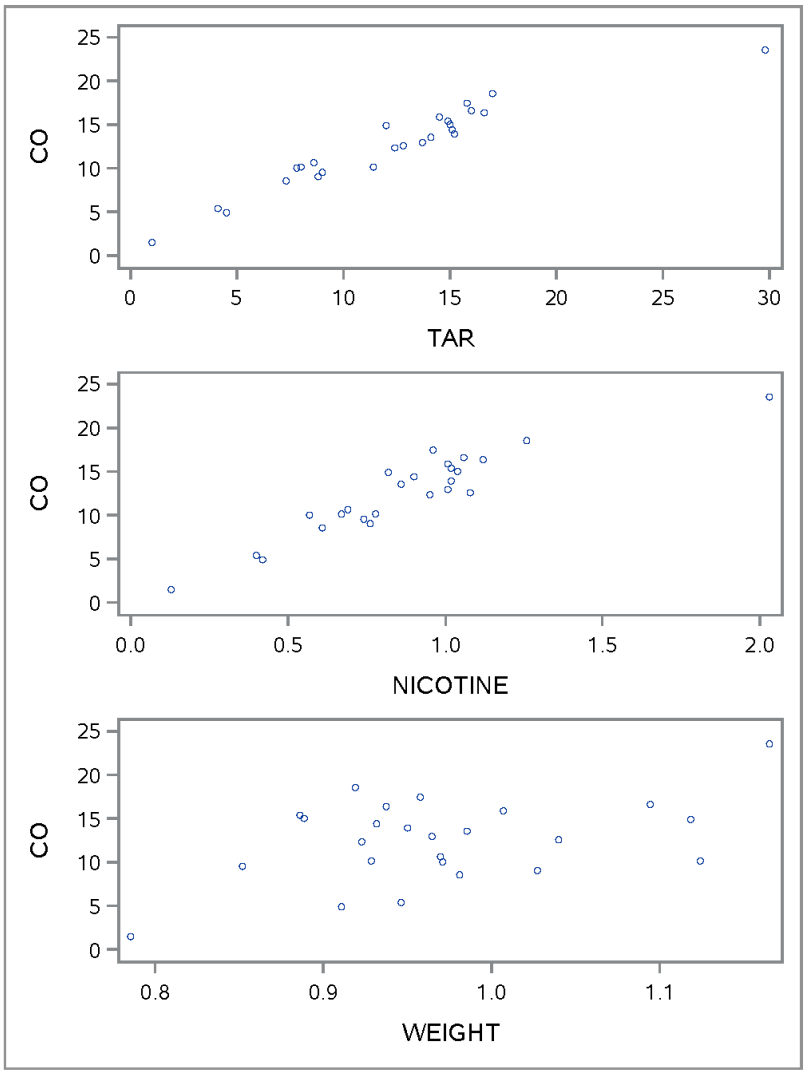
Can we model the carbon monoxide content of cigarettes as a function of their tar content, their nicotine content, and their weight?

Y : carbon monoxide content from cigarette smoke [CO]

X_1 : tar content of cigarette [TAR]

X_2 : nicotine content of cigarette [NICOTINE]

X_3 : weight of cigarette [WEIGHT]



```
proc reg data=cigar plots=none;
model co=tar nicotine weight /vif;
```

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	495.25781	165.08594	78.98	<.0001
Error	21	43.89259	2.09012		
Corrected Total	24	539.15040			

Root MSE	1.44573	R-Square	0.9186
Dependent Mean	12.52800	Adj R-Sq	0.9070
Coeff Var	11.53996		

Parameter Estimates						
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variance Inflation
Intercept	1	3.20219	3.46175	0.93	0.3655	0
TAR	1	0.96257	0.24224	3.97	0.0007	21.63071
NICOTINE	1	-2.63166	3.90056	-0.67	0.5072	21.89992
WEIGHT	1	-0.13048	3.88534	-0.03	0.9735	1.33386

```
proc reg data=cigar plots=none;
model co=nicotine weight /vif;
```

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	462.25639	231.12820	66.13	<.0001
Error	22	76.89401	3.49518		
Corrected Total	24	539.15040			

Root MSE	1.86954	R-Square	0.8574
Dependent Mean	12.52800	Adj R-Sq	0.8444
Coeff Var	14.92290		

Parameter Estimates						
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variance Inflation
Intercept	1	1.61398	4.44663	0.36	0.7201	0
NICOTINE	1	12.38812	1.24473	9.95	<.0001	1.33366
WEIGHT	1	0.05883	5.02395	0.01	0.9908	1.33366

```
proc reg data=cigar plots=none;
model co=tar weight /vif;
```

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	494.30638	247.15319	121.25	<.0001
Error	22	44.84402	2.03836		
Corrected Total	24	539.15040			

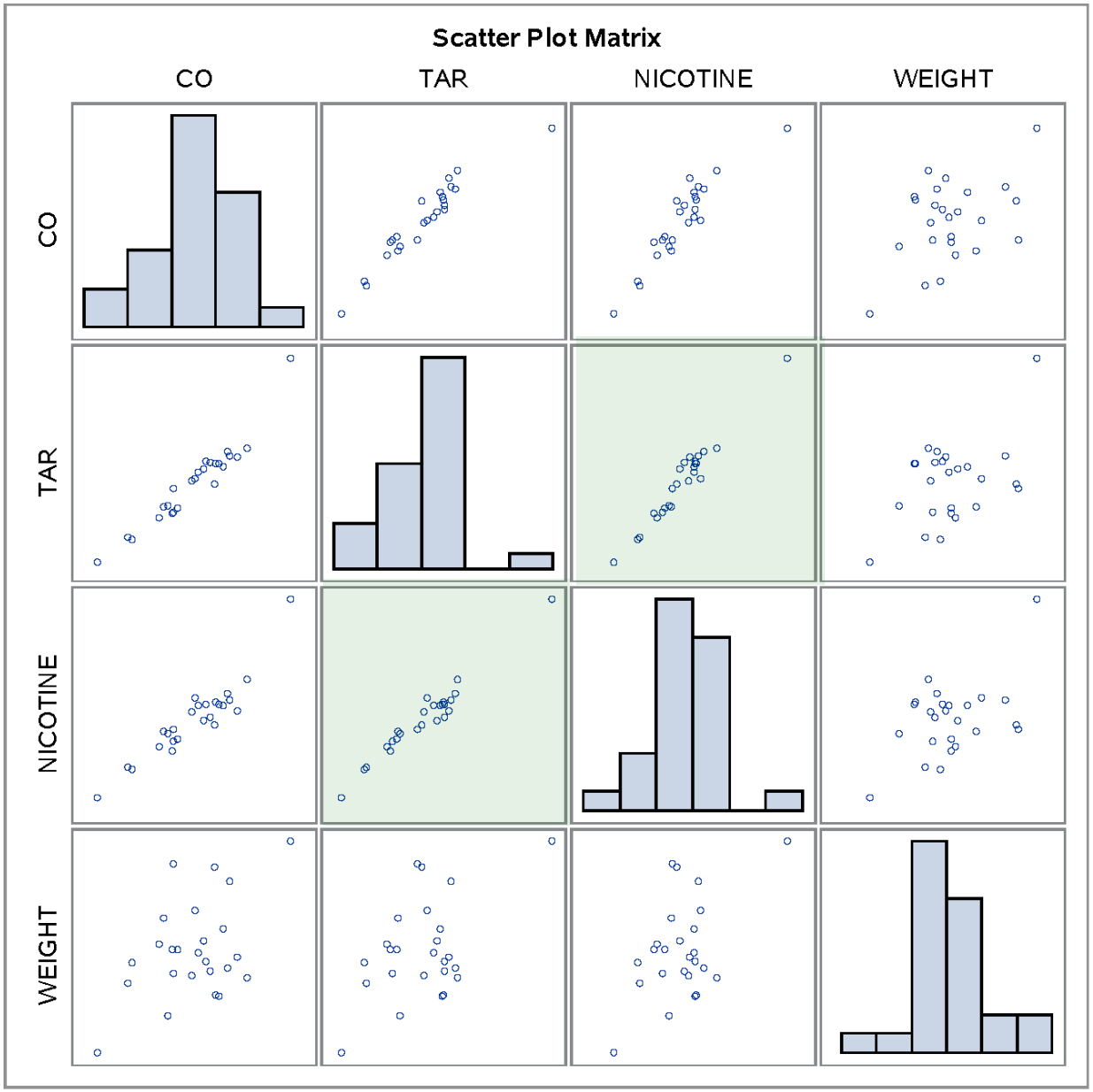
Root MSE	1.42771	R-Square	0.9168
Dependent Mean	12.52800	Adj R-Sq	0.9093
Coeff Var	11.39618		

Parameter Estimates						
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variance Inflation
Intercept	1	3.11433	3.41620	0.91	0.3718	0
TAR	1	0.80419	0.05904	13.62	<.0001	1.31726
WEIGHT	1	-0.42287	3.81299	-0.11	0.9127	1.31726

```
proc corr data=cigar
plots=matrix(histogram);
var co tar nicotine weight;
```

Pearson Correlation Coefficients, N = 25 Prob > |r| under H0: Rho=0				
	CO	TAR	NICOTINE	WEIGHT
CO	1.00000	0.95749 <.0001	0.92595 <.0001	0.46396 0.0195
TAR	0.95749 <.0001	1.00000	0.97661 <.0001	0.49077 0.0127
NICOTINE	0.92595 <.0001	0.97661 <.0001	1.00000	0.50018 0.0109
WEIGHT	0.46396 0.0195	0.49077 0.0127	0.50018 0.0109	1.00000

The CORR Procedure



Using R ...

```
> fit = lm(CO ~ TAR + NICOTINE + WEIGHT, data=FTCCIGAR)

> install.packages("car")
> library(car)
> vif(fit)
      TAR  NICOTINE    WEIGHT
21.630706 21.899917  1.333859

> cor(FTCCIGAR)
                TAR  NICOTINE    WEIGHT        CO
TAR       1.0000000 0.9766076 0.4907654 0.9574853
NICOTINE  0.9766076 1.0000000 0.5001827 0.9259473
WEIGHT    0.4907654 0.5001827 1.0000000 0.4639592
CO        0.9574853 0.9259473 0.4639592 1.0000000
>
```

Recommendation: Drop `nicotine` from the model.

from R^2 & R^2 adjust

Diagnostic Measures for GLM

Recall: Some residuals

- Is $\boldsymbol{e}$ a good choice for testing normality?
  - Is it true that $e_1, \ldots, e_n$ forms a random sample?
    No because $\text{var}(e_i) = \sigma^2(1 - P_{ii})$ & $\text{Cov}(e_i, e_j) = -\sigma^2 P_{ij}$
- internally studentized residual (2)
  (R=standardized residual, SAS=studentized residual):
  $$r_i = \frac{e_i}{s\sqrt{1 - P_{ii}}}, \quad s = \sqrt{MSE}, \quad MSE \text{ from regn with all } n \text{ obs}$$
  - $\text{var}(r_i) \approx 1$.
  - r_i, r_j are likely dependent
  - $r_i \approx t_{df_e}$ (dependent numerator & denominator)
- externally studentized residual or studentized deleted residual (3)
  (R & SAS=rstudent residual):
  $$r_i^* = \frac{Y_i - \hat{Y}_{i(i)}}{s_{(i)}\sqrt{1 + \boldsymbol{x}_i(\boldsymbol{X}_{(i)}^T\boldsymbol{X}_{(i)})^{-1}\boldsymbol{x}_i^T}} = \frac{e_i}{s_{(i)}\sqrt{1 - P_{ii}}}$$
  - $\text{var}(r_i^*) \approx 1$.
  - r_i^*, r_j^* are likely dependent
  - $r_i^* \approx t_{df_e - 1}$ (better approx than r_i) (df_e is from regression with all n observations)
  - r_i^* reflects large values more dramatically than r_i (Atkinson 1983)
- deleted/Jackknife/LOOCV results: run regression n times, with the ith observation excluded during the ith run: (Belsley, Kuh, Welsch 1980)
  - $\boldsymbol{X}_{(i)}$ is new design matrix from omitting the ith observation
  - $s_{(i)}^2 = MSE_{(i)}$ is the MSE from omitting the ith observation
    No need to rerun the regn: $(n - p - 1)s_{(i)}^2 = (n - p)s^2 - r_i^2/(1 - P_{ii})$
  - $\hat{\boldsymbol{\beta}}_{(i)}$ is estimate of full p-dimensional vector $\boldsymbol{\beta}$ from omitting the ith obs.
  - $\hat{\beta}_{j(i)}$ is estimate of parameter β_j from omitting the ith obs.
  - $\hat{\boldsymbol{Y}}_{(i)}$ is prediction of full n-dimensional vector $\boldsymbol{Y}$ from omitting the ith obs.
  - $\hat{Y}_{i(i)}$ is prediction of ith observation Y_i from omitting the ith obs.
  - $Y_i - \hat{Y}_{i(i)}$ is called the ith deleted residual
  - r_i^* is the t-statistic for agreement between Y_i and $\hat{Y}_{i(i)}$:
    $$r_i^* = \frac{Y_i - \hat{Y}_{i(i)}}{\text{s.e.}(Y_i - \hat{Y}_{i(i)})} = \frac{Y_i - \hat{Y}_{i(i)}}{s_{(i)}\sqrt{1 + \boldsymbol{x}_i(\boldsymbol{X}_{(i)}^T\boldsymbol{X}_{(i)})^{-1}\boldsymbol{x}_i^T}} = \frac{Y_i - \hat{Y}_i}{s_{(i)}\sqrt{1 - P_{ii}}}$$

GLM Diagnostic

1. + Deviance lack of fit test
2. Residual
   - Several types of residuals are commonly used:
     - (1) Raw or response: $y_i - \widehat{\mu}_i$ — not very useful
     - (2) Pearson: $r_i^P \equiv \frac{y_i - \widehat{\mu}_i}{\sqrt{w_i \widehat{h}(\widehat{\mu}_i)}} \rightsquigarrow \mathbb{X}^2 = \sum_{i=1}^n (r_i^P)^2$ — very useful for diagnostics
       also $\frac{r_i^P}{\sqrt{\widehat{\psi}}\sqrt{1 - h_i}}$ where h_i is leverage
     - (4) Deviance: $r_i^D \equiv \text{sign}(y_i - \widehat{\mu}_i)\sqrt{d_i} \rightsquigarrow D_M = \sum_{i=1}^n d_i = \sum_{i=1}^n (r_i^D)^2$
       also $\frac{r_i^D}{\sqrt{\widehat{\psi}}\sqrt{1 - h_i}}$ — very useful for diagnostics

     Note: The chi-squared statistic $\mathbb{X}^2$ and the scaled chi-squared statistic $\mathbb{X}^2/\psi$ are often used interchangeably with the deviance D_M and scaled deviance D_M/ψ.
3. Influence (analog Cook's distance)
   - New diagnostic measure: likelihood displacement:
     $$LD_i \equiv 2\left\{\ell_M(\widehat{\boldsymbol{\theta}}; \mathbf{y}) - \ell_M(\widehat{\boldsymbol{\theta}}_{(-i)}; \mathbf{y})\right\}$$
     where $\widehat{\boldsymbol{\theta}}_{(-i)}$ is MLE from excluding ith observation. Likelihood evaluated with *all* observations.
   - Predicted values: $\widehat{\mu}_i$, response or $\boldsymbol{x}_i^T\widehat{\boldsymbol{\beta}}$, linear predictor
4. Complete / quasi complete separation
   - non-convergence, high SE, LR ≈ 0.
   - happens when rare events, large prediction space many binary prediction, small sample size.
   - biased regression (Firth's procedure) (to fix)
   - quick check: deviance / df Error
     ≈ 1 model fits well
     $<< 1$ possibly overfitting
     $>> 1$ underfitting, overdispersion

Exponential Family

General Forms & Rules

Definition 1:

$$\ell(\mu, \phi; y) = \sum_{i=1}^{n} \left\{ \frac{y_i\theta_i - b(\theta_i)}{\phi w_i} + c(y_i, \phi; w_i) \right\}$$ ← this is in a very weird form, it doesn't put T(x) altogether.

- $\mu_i = g^{-1}(x_i'\beta)$ is some function of β
- θ is some function of μ (thus β)

Definition 2:

$$Y \sim EF(\theta, \phi)$$

$$f(y;\theta,\phi) = \exp\left\{ \frac{T(y)g(\theta) - b(\theta)}{a(\phi)} \right\} h(y,\phi)$$

- ϕ dispersion parameter
- $E(T(y)) = b'(\theta)$
- $Var(T(y)) = a(\phi)b''(\theta)$

Definition 3: $f(y;\theta) = \exp\{T(y)g(\theta) - b(\theta)\}h(y)$

$g(\theta)$: natural (canonical) parameter

$T(y)$: sufficient statistics

* Assume $\theta = g(\theta)$ is the canonical parameter:

$$\int \exp\{T(y)g(\theta) - b(\theta)\}h(y)\,dy = 1$$

$$\int e^{T(y)g(\theta)} e^{-b(\theta)} h(y)\,dy = 1$$

$$\int e^{T(y)g(\theta)} h(y)\,dy = e^{b(\theta)}$$

$$\Rightarrow \int e^{T(y)\theta} h(y)\,dy = e^{b(\theta)} \quad * g(\theta) = \theta$$

$$\Rightarrow \int T(y) e^{T(y)\theta} h(y)\,dy = b'(\theta) e^{b(\theta)} \quad * \text{derivate w.r.t. } \theta \text{ on both sides}$$

$$\int T(y) e^{T(y)\theta - b(\theta)} h(y)\,dy = b'(\theta)$$

$$\Rightarrow E(T(y)) = b'(\theta)$$

$Var(T(y)) = b''(\theta)$ similarly

Univariate Normal

$$f(x) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x_i-\mu)^2}{2\sigma^2}}$$

$$= \exp\left(-\frac{n}{2}\log(2\pi\sigma^2) - \frac{\Sigma(x_i-\mu)^2}{2\sigma^2}\right)$$

$$= \exp\left(-\frac{n}{2}\log(2\pi\sigma^2) - \frac{\Sigma x_i^2 - 2\mu\Sigma x_i + n\mu^2}{2\sigma^2}\right) \longrightarrow \text{Complete \& sufficient statistics } (\Sigma x_i, \Sigma x_i^2)$$

$$= \exp\left(-\frac{n}{2}\log(2\pi\sigma^2) + \frac{\mu\Sigma x_i - n\mu^2/2}{\sigma^2} - \frac{\Sigma x_i^2}{2\sigma^2}\right)$$

$$= \exp\left(\frac{\mu\Sigma x_i - n\mu^2/2}{\sigma^2} - \frac{n}{2}\log(\sigma^2) - \frac{n}{2}\log(2\pi) - \frac{\Sigma x_i^2}{2\sigma^2}\right)$$

$\phi = \sigma^2 \quad \theta = \mu \quad b(\theta) = \mu^2/2$

$w_i = 1 \quad c(y, \phi, w_i) = -\frac{n}{2}\log(\sigma^2) - \frac{n}{2}\log(2\pi) - \frac{\Sigma x_i^2}{2\sigma^2}$

* MGF derivation

$$M_X(t) = E(e^{tx}) \qquad EX = \mu \qquad \mathrm{Var}\,X = \sigma^2$$

$$= \int_{-\infty}^{\infty} e^{tx} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} dx$$

$$= \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{x^2-2x\mu+\mu^2}{2\sigma^2} + \frac{2\sigma^2 tx}{2\sigma^2}} dx$$

$$= \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{x^2-2x(\mu+\sigma^2 t)+\mu^2}{2\sigma^2}} dx$$

$$= \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-(\mu+\sigma^2 t))^2}{2\sigma^2}} \cdot e^{\frac{t^2\sigma^2}{2} + \mu\sigma^2 t^2} dx$$

$$= e^{t^2\sigma^2/2 + \mu t}$$

Multivariate Normal

$$f_X(x) = \frac{1}{\det|2\pi\Sigma|} \exp\left(-\frac{1}{2}(x-\mu)'\Sigma^{-1}(x-\mu)\right)$$

$$= \exp\left(-\log\det|2\pi\Sigma| - \frac{1}{2}(x-\mu)'\Sigma^{-1}(x-\mu)\right)$$

$$= \exp\left(-\log\det|2\pi\Sigma| - \frac{1}{2}(x'\Sigma^{-1}x - 2\mu'\Sigma^{-1}x + \mu'\Sigma^{-1}\mu)\right)$$

$$\text{MGF} = e^{\frac{1}{2}t'\Sigma t + t'\mu}$$

* Conditional Normal

$$\begin{pmatrix} X \\ Y \end{pmatrix} \sim N\left(\begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}\right)$$

$$X|Y \sim N(\mu_1 + \Sigma_{12}'\Sigma_{22}^{-1}(Y-\mu_2),\ \Sigma_{11} - \Sigma_{12}'\Sigma_{22}^{-1}\Sigma_{21})$$

Gamma

1. Shape α, Scale β ← this is the most common case

$$f_X(x) = \frac{1}{\Gamma(\alpha)\beta^\alpha} x^{\alpha-1} e^{-x/\beta} \qquad 0<x \qquad EX = \alpha\beta \quad VarX = \alpha\beta^2$$

$$f(x) = \prod_{i=1}^n \frac{1}{\Gamma(\alpha)\beta^\alpha} x_i^{\alpha-1} e^{-x_i/\beta} \qquad 0<x_i$$

$$= \exp(-n\log(\Gamma(\alpha)) - n\alpha\log(\beta) + (\alpha-1)\textstyle\sum_{i=1}^n \log(x_i) - \frac{1}{\beta}\sum_{i=1}^n x_i)$$

$$= \exp(-\tfrac{1}{\beta}\textstyle\sum_{i=1}^n x_i - n\log(\Gamma(\alpha)) - n\alpha\log(\beta) + (\alpha-1)\sum_{i=1}^n \log x_i)$$

$\phi = 1 \quad \theta = -\frac{1}{\beta} \quad b(\theta) = -n\log(\Gamma(\alpha)) - n\alpha\log(\beta)$

$w_i = 1 \quad a(y,\phi,w_i) = (\alpha-1)\sum_{i=1}^n \log x_i$

$$Ex = \int_0^\infty x \frac{1}{\Gamma(\alpha)\beta^\alpha} x^{\alpha-1} e^{-x/\beta} dx$$

$$= \int_0^\infty \frac{1}{\Gamma(\alpha)\beta^\alpha} x^{(\alpha+1)-1} e^{-x/\beta} dx$$

$$= \int_0^\infty \frac{1}{\Gamma(\alpha+1)\beta^{\alpha+1}} \cdot \alpha\beta \, x^{(\alpha+1)-1} e^{-x/\beta} dx$$

$$= \alpha\beta$$

Sufficient & complete statistics

$T(x) = (\sum_{i=1}^n x_i, \sum_{i=1}^n \log x_i)$

2. Shape α & rate θ * this is not being tested too much.

As long as always remember $\theta = \frac{1}{\beta}$.

Beta

$$f_X(x) = \frac{x^{\alpha-1}(1-x)^{\beta-1}}{B(\alpha,\beta)} \quad \text{with } B(\alpha,\beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)} \qquad 0 \le x \le 1$$

$$= \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1}(1-x)^{\beta-1}$$

$$= \exp(\log\left(\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\right) + (\alpha-1)\log(x) + (\beta-1)\log(1-x))$$

$$= \exp((\alpha-1)\log(x) + (\beta-1)\log(1-x) + \log\left(\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\right))$$

$\phi = 1 \quad w = 1 \quad \theta = 0 \quad b(\theta) = \log\left(\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\right)$

$a(y,\phi,w_i) = (\alpha-1)\log x + (\beta-1)\log(1-x)$

$EX = \frac{\alpha}{\alpha+\beta}$

$VarX = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$

Sufficient & complete statistics

$T(x) = (\Sigma \log x_i, \Sigma \log(1-x_i))$

$Beta(1,1) = Unif(0,1)$

Exponential * time between poisson events

1. Scale = mean = β

$f_X(x) = \frac{1}{\beta} e^{-x/\beta} \sim \text{Gamma}(\text{shape} = 1, \text{scale} = \beta) \quad 0 < x$

$= \exp(-\log\beta - \frac{1}{\beta}x)$

Sufficient & complete statistics

$T(x) = \sum_{i=1}^{n} x_i$

$\phi = 1 \quad w = 1 \quad b(\theta) = -\log\beta \quad \theta = \frac{-1}{\beta}$

$a(y, \phi, w) = 0$

$E(X) = \int_0^\infty x \frac{1}{\beta} e^{-x/\beta} dx$

$= -xe^{-x/\beta} + \int_0^\infty e^{-x/\beta} dx \qquad u = x \quad v' = \frac{1}{\beta} e^{-x/\beta}$

$= -xe^{-x/\beta} - \beta e^{-x/\beta} \Big|_0^\infty \qquad u' = 1 \quad v = -e^{-x/\beta}$

$= \beta$

$E(X) = \beta$

$Var(X) = \beta^2$

$M_X(t) = E(e^{tx})$

$= \int_0^\infty e^{tx} \frac{1}{\beta} e^{-x/\beta} dx$

$= \frac{1}{\beta} \int_0^\infty e^{-x(\frac{1}{\beta} - t)} dx$

$= \frac{1}{\beta} \int_0^\infty e^{-x(\frac{1-t\beta}{\beta})} dx \qquad \text{let's think } \frac{1}{\beta^*} = \frac{1-t\beta}{\beta}$

$= \frac{1}{\beta} \cdot \frac{\beta}{1-t\beta} \qquad \text{then } \frac{1}{\beta^*} \cdot \int \cdot dx = 1$

$= \frac{1}{1-t\beta} \qquad \Rightarrow \int \cdot dx = \beta^*$

2. rate = λ ($= 1/\beta$)

$f_X(x) = \lambda e^{-x\lambda}$

$= \exp(\log(\lambda) - \lambda \Sigma x_i)$

$\theta = -\lambda \quad \phi = 1 \quad w = 1$

$b(\theta) = \log(\lambda)$

Sufficient & complete statistics is Σx_i

$E(X) = 1/\lambda$

$Var(X) = 1/\lambda^2$

$Ex = \int_0^\infty x\lambda e^{-x\lambda}$

$= -xe^{-x\lambda} + \int_0^\infty e^{-x\lambda} dx \qquad u = x \quad v' = \lambda e^{-x\lambda}$

$= -xe^{-x\lambda} - \frac{1}{\lambda} e^{-x\lambda} \Big|_0^\infty \qquad u' = 1 \quad v = -e^{-x\lambda}$

$= \frac{1}{\lambda}$

$$M_X(t) = E(e^{tx})$$
$$= \int_0^\infty e^{tx} \lambda e^{-x\lambda} \, dx$$
$$= \int_0^\infty \lambda e^{-x(\lambda - t)} \, dx$$
$$= \frac{\lambda}{\lambda - t}$$

Weibull * not exp family

* Just know some relationship.

$\exp(\beta) = \text{weibull}(1, \beta) = \text{Gamma}(1, \beta)$

$X \sim \text{weibull}(a, b) \Leftrightarrow X^a \sim \exp(b^a)$ both scale-liked.

Cauchy * not exp family

* Just know some relationship.
* Undefined mean & variance

X indep $Y \sim N(0,1)$; $\frac{X}{Y} \sim \text{Cauchy}(0,1)$

Log-normal * not exp family (?)

$$f(x) = \frac{1}{x\sigma\sqrt{2\pi}} \exp\left(-\frac{\log x - \mu}{2\sigma^2}\right)$$
$$= \exp\left(-\log(x) - \tfrac{1}{2}\log(2\pi\sigma^2) - \frac{\log x - \mu}{2\sigma^2}\right)$$

sufficient and complete statistics $\sum_i \log(x_i)$ (?)

Uniform * not exponential family

$X \sim \text{Unif}(0,1)$

$f(x) = 1 \quad I(0 \le x \le 1)$

$E(X) = \frac{1}{2}$

$\text{Var}(X) = \frac{1}{12}$

$$f(\underline{x}) = \prod_{i=1} 1 \; I(0 \le x_i \le 1)$$
$$= I(0 < x_{(1)}, x_{(n)} < 1)$$

Sufficient statistics $X_{(1)}, X_{(n)}$

Not complete.

$(X_{(1)} + X_{(n)}) - (a+b)$ has mean 0 every where, but is non-trivial.

$$P(X_{(1)} \le x) = 1 - P(X_{(1)} \ge x)$$
$$= 1 - P(X_1 \ge x)^n$$
$$= 1 - (1-x)^n$$
$$f_{X_{(1)}}(x) = \frac{\partial}{\partial x} 1 - (1-x)^n$$
$$= n \cdot (1-x)^{n-1}$$
$$= n(1-x)^{n-1} \sim \text{Beta}(1, n)$$

$$P(X_{(n)} \le x) = P(X_1 \le x)^n$$
$$= x^n$$
$$f_{X_{(n)}}(x) = \frac{\partial}{\partial x} P(X_{(n)} \le x)$$
$$= n x^{n-1} \sim \text{Beta}(n, 1)$$

Poisson * counts in time / space

$$\prod_{i=1}^{n} P(X_i = x_i) = \prod_{i=1}^{n} \frac{\lambda^x e^{-\lambda}}{x!} \qquad \Leftarrow \sum_{k=0}^{\infty} \frac{\lambda^k}{k!} = e^{\lambda}$$

$$= \exp\left(\sum_{i=1}^{n} \log(x_i!) + \sum_{i=1}^{n} x_i \log\lambda - n\lambda\right)$$ sufficient & complete statistics $\sum X_i$

$\phi = 1 \quad w = 1$

$\theta = \log\lambda \quad b(\theta) = \lambda = e^{\theta}$

$$EX = \sum_{x=0}^{\infty} x \cdot \lambda^x e^{-\lambda}/x! \qquad EX = \lambda \qquad \mathrm{Var}X = \lambda$$

$$= \sum_{x=1}^{\infty} \frac{\lambda^x e^{-\lambda}}{(x-1)!}$$

$$= \lambda \sum_{x=1}^{\infty} \frac{\lambda^{x-1} e^{-\lambda}}{(x-1)!}$$

$$= \lambda$$

$$M_X(t) = E(e^{tx})$$

$$= \sum_{x=0}^{\infty} e^{tx} \frac{\lambda^x e^{-\lambda}}{x!}$$

$$= \sum_{x=0}^{\infty} \frac{(e^t\lambda)^x e^{-\lambda}}{x!}$$

$$= e^{-\lambda} \sum_{x=0}^{\infty} \frac{(e^t\lambda)^x}{x!}$$

$$= e^{-\lambda} e^{e^t\lambda}$$

$$= e^{\lambda(e^t - 1)}$$

Bernoulli * occurance of event in one trail

$$\prod_{i=1}^{n} P(X_i = x_i) = \prod_{i=1}^{n} p^{x_i} (1-p)^{1-x_i}$$

$$= \exp\left(\sum x_i \log\left(\frac{p}{1-p}\right) + n\log(1-p)\right)$$ sufficient & complete statistics $\sum X_i$

$$\theta = \log\left(\frac{p}{1-p}\right)$$

$$EX = \sum_{x=0}^{1} x\, p^x (1-p)^{1-x} \qquad EX = p \qquad \mathrm{Var}X = p(1-p)$$

$$= p$$

$$M_X(t) = E(e^{tx})$$

$$= \sum_{x=0}^{1} e^{tx} p^x (1-p)^{1-x}$$

$$= pe^t + (1-p)$$

Binomial *

Occurance of event in n trails

$$\prod_{i=1}^{n} P(X_i=x_i) = \prod_{i=1}^{n} \binom{n}{x_i} p^{x_i} (1-p)^{n-x_i}$$

$$= \exp\left\{ \sum_{i=1}^{n} \log\binom{n}{x_i} + \sum x_i \log\left(\frac{p}{1-p}\right) + n\log(1-p) \right\}$$

Sufficient & complete statistics is $\sum X_i$

$\theta = \log\left(\frac{p}{1-p}\right)$ $\quad b(\theta) = \log(1-p) = \ln(1+e^{\theta})$

$\phi = 1 \quad w = n$

$$EX = \sum_{x=0}^{n} \binom{n}{x} \cdot x \; p^x (1-p)^{n-x}$$

$$= \sum_{x=1}^{n} \frac{n!}{(x-1)!(n-x)!} p^{x-1} p \, (1-p)^{n-x}$$

$$= n \cdot p \sum_{x=1}^{n} \frac{(n-1)!}{(x-1)!(n-x)!} p^{x-1} (1-p)^{n-x}$$

$$= np$$

$EX = np$

$Var X = np(1-p)$

$$M_X(t) = E(e^{tx})$$

$$= E(e^{t\sum Y_i}) \quad \text{with } Y_i \overset{iid}{\sim} Bor(p)$$

$$= E(e^{tY_1}) E(e^{tY_2}) \cdots$$

$$= (pe^t + (1-p))^n \quad \leftarrow \text{using bernoulli MGF}$$

Multinomial

$$\prod_{i=1}^{n} P(X_i=x_i) = \frac{n!}{x_1! \cdots x_n!} p_1^{x_1} \cdots p_n^{x_n}$$

$$= \exp\left\{ \log\left(\frac{n}{x_1! \cdots x_n!}\right) + x_1 \log\left(\frac{p_1}{p_n}\right) + \cdots + n\log(p_n) \right\}$$

$$E(x_i) = \sum_{x_i=0}^{n} x_i \frac{n!}{x_1! \cdots x_n!} p_1^{x_1} \cdots p_n^{x_n}$$

$$= n \cdot p_i$$

$EX_i = np_i$

$Var X_i = np_i(1-p_i)$

$$M_X(t) = \left(\sum_{i=1}^{n} p_i e^{t_i} \right)^n \quad \leftarrow \text{inspired by binomial dist.}$$

Geometric

1. # of failure until first success at prob of success p (include success)

$$P(X=x) = (1-p)^{x-1}p \qquad x = 1, \cdots$$

$$\prod_{i=1}^{n} P(X_i = x_i) = (1-p)^{\sum x_i - n} p^n$$

$$= \exp\left\{ \sum x_i \log(1-p) + n \log\left(\frac{p}{1-p}\right) \right\}$$

Sufficient & complete statistics is $\sum X_i$

$$\theta = \log(1-p) \quad b(\theta) = -\log\left(\frac{p}{1-p}\right)$$

$$\phi = 1 \quad w = \frac{1}{n}$$

$$EX = \sum_{x=1}^{\infty} x(1-p)^{x-1}p$$

$$= p\sum_{x=1}^{\infty} x(1-p)^{x-1}$$

$$= p\left(-\frac{\partial}{\partial p}\sum_{x=1}^{\infty}(1-p)^x\right)$$

$$= p\left(-\frac{\partial}{\partial p}\sum_{x=1}^{\infty}(1-p)^{x-1} - 1\right)$$

$$= p\left(-\frac{\partial}{\partial p}\frac{1}{p} - 1\right)$$

$$= p\frac{1}{p^2}$$

$$= \frac{1}{p}$$

Recall Geometric Summation:

$$\sum_{x=0}^{\infty} a r^x = \frac{a}{1-r} \qquad 0 < r < 1$$

$$EX = \frac{1}{p}$$

$$VarX = \frac{1-p}{p^2}$$

$$M_X(t) = \sum_{x=1}^{\infty} e^{tx}(1-p)^{x-1}p$$

$$= \sum_{x=1}^{\infty}(e^t(1-p))^x \frac{p}{1-p}$$

$$= \frac{p}{1-p}\sum_{x=1}^{\infty}(e^t(1-p))^x$$

$$= \frac{p}{1-p}\left(\sum_{x=1}^{\infty}(e^t(1-p))^{x-1} - 1\right)$$ converges when $0 < e^t(1-p) < 1$

$$= \frac{p}{1-p}\frac{e^t(1-p)}{1-e^t(1-p)}$$

$$= \frac{pe^t}{1-(1-p)e^t}$$ when $e^t(1-p) < 1$

$$e^t < \frac{1}{1-p}$$

$$t < \ln\left(\frac{1}{1-p}\right)$$

2. Number of failures before success (does not include success) ☆ may use more of this case

waiting time (discrete) to hit the first success (only count for failures)

$$P(X=x) = (1-p)^x p \qquad x = 0, \cdots$$

$$\begin{aligned} EX &= \sum_{x=0}^{\infty} x(1-p)^x p \\ &= p(1-p)\sum_{x=0}^{\infty} x(1-p)^{x-1} \\ &= p(1-p)\left(-\frac{\partial}{\partial p}\sum_{x=0}^{\infty}(1-p)^x\right) \\ &= p(1-p)\left(-\frac{\partial}{\partial p}\left(\frac{1}{p}\right)\right) \\ &= p(1-p)\frac{1}{p^2} \\ &= \frac{1-p}{p} \end{aligned}$$

$EX = \frac{1-p}{p}$

$Var X = \frac{1-p}{p^2}$

$$\begin{aligned} M_X(t) &= E(e^{tx}) \\ &= \sum_{x=0}^{\infty} e^{tx} p(1-p)^x \\ &= p\sum_{x=0}^{\infty}(e^t(1-p))^x \\ &= p \cdot \frac{1}{1-e^t(1-p)} \\ &= \frac{p}{1-e^t(1-p)} \end{aligned}$$

with $0 < e^t(1-p) < 1$

$t < \log\left(\frac{1}{1-p}\right)$

Negative-Binomial

Number of failures to hit the r^{th} success (does not include r)

waiting time (discrete) to hit the rth success (only count for failures)

$$P(X=x) = \binom{x+r-1}{r-1} p^r (1-p)^x \qquad x = 0, 1, \cdots$$

$$\begin{aligned} E(X) &= \sum_{x=0}^{\infty} x\binom{x+r-1}{r-1} p^r(1-p)^x \\ &= \sum_{x=1}^{\infty} \frac{(x+r-1)!}{(r-1)!(x-1)!} p^r(1-p)^x \\ &= \sum_{x=1}^{\infty} \frac{(x+r-1)!}{r!(x-1)!} \cdot r\, p^{r+1} \cdot \frac{1}{p}(1-p)^{x-1}(1-p) \\ &= r\frac{1-p}{p} \end{aligned}$$

$EX = r\frac{(1-p)}{p}$

$Var X = r\frac{(1-p)}{p^2}$

$$M_X(t) = \left(\frac{p}{1-(1-p)e^t}\right)^r$$ ← inspired from geometric MGF

Since Geometric(p) ≡ Neg-binomial(1, p)

Hypergeometric * Like binomial but without replacement

$$P(X=k) = \frac{\binom{K}{k}\binom{N-K}{n-k}}{\binom{N}{n}}$$

$$= \frac{k \text{ success}, n-k \text{ fail}}{\text{among all possible results}}$$

N : population size

K : number of success

n : number of draw

k : number of observed success

$$EX = \sum_{x=0}^{K} x \frac{\binom{K}{x}\binom{N-K}{n-x}}{\binom{N}{n}}$$

$$= \sum_{x=1}^{K} \frac{1}{\binom{N}{n}} \frac{K!}{(x-1)!(K-x)!} \frac{(N-K)!}{(n-x)!(N-n)!}$$

$$= \sum_{x=1}^{K} \frac{1}{\binom{N}{n}} \frac{(K-1)!}{(x-1)!(K-x)!} K \frac{(N-K)!}{(n-x)!(N-n)!}$$

$$= \sum_{x=1}^{K} \frac{(n-1)!(N-n)!}{(N-1)!} \cdot \frac{n}{N} \frac{(K-1)!}{(x-1)!(K-x)!} K \frac{(N-K)!}{(n-x)!(N-n)!}$$

$$= \sum_{x=1}^{K} \frac{\binom{K-1}{x-1}\binom{N-K}{n-x}}{\binom{N-1}{n-1}} \cdot \frac{nK}{N}$$

$$= \frac{nK}{N}$$

Distribution Theory

Multivariate normal distribution

$Y \in \mathbb{R}^p \sim N_p(\mu, \Sigma) \Rightarrow v'Y \sim N(v'\mu, v'\Sigma v)$ with $v'\Sigma v > 0$

$f_Y(y) = \det(2\pi\Sigma)^{-\frac{1}{2}} \exp(-\frac{1}{2}(y-\mu)'\Sigma(y-\mu)) \Rightarrow M_Y(t) = e^{t'\mu + \frac{1}{2}t'\Sigma t}$

$E(Y) = \mu$

$\mathrm{Var}(Y) = \Sigma$

t distribution

$X \sim N(\mu, 1)$ indep. $U \sim \chi^2_p$

$X/\sqrt{U/p} \sim T_p(\mu)$

← Consider $\frac{X}{\sqrt{\sigma^2}} \sim N(0,1)$

$X \sim N(0, \sigma^2)$

$\frac{S^2(p)}{\sigma^2} \sim \chi^2_p \Rightarrow \frac{S^2}{\sigma^2} \sim \frac{1}{p}\chi^2_p$

$\frac{N(0,\sigma^2)}{\sqrt{\chi^2_p/p}} \sim T_p$

F distribution

① $U_1 \sim \chi^2_{p_1}$ indep. $U_2 \sim \chi^2_{p_2}$

$F = \frac{U_1/p_1}{U_2/p_2} \sim F_{p_1, p_2}$

② $U_1 \sim \chi^2_{p_1}(\phi)$ indep. $U_2 \sim \chi^2_{p_2}$

$F = \frac{U_1/p_1}{U_2/p_2} \sim F_{p_1, p_2}(\phi)$

Chi-squared distribution

① $Z_i \overset{iid}{\sim} N(0,1) \Rightarrow Z \sim N_p(0, I)$

$V: \sum_{i=1}^p Z_i^2 \sim \chi^2_p \Rightarrow M_V(t) = (1-2t)^{-p/2}$

$E(V) = p$

$\mathrm{Var}(V) = 2p$

② $Z_i \overset{iid}{\sim} N(\mu_i, 1) \Rightarrow Z \sim N_p(\mu, I)$

$V: \sum_{i=1}^p Z_i^2 \sim \chi^2_p(\frac{1}{2}\sum_{i=1}^p \mu_i^2) \Rightarrow M_V(t) = (1-2t)^{-p/2} e^{\frac{2\phi t}{1-2t}}$ with $t < \frac{1}{2}$

$E(V) = p + 2\phi$

$\mathrm{Var}(V) = 2p + 4\phi$

③ $U_i \overset{iid}{\sim} \chi^2_{p_i}(\phi_i)$

$U := \sum_{i=1}^m U_i \sim \chi^2_{\sum p_i}(\sum \phi_i)$

④ $X \sim N_p(u, V) \Rightarrow X'V^{-1}X \sim \chi^2_p(\frac{1}{2}\mu'V^{-1}\mu)$

Lemma: $A \in \mathbb{R}^{n\times n}$ is symmetric and idempotent matrix with $\text{rank}(A) = s$ iff $A = GG'$ for some matrix $G \in \mathbb{R}^{n\times s}$ with $G'G = I_s$ and $\text{rank}(G) = s$.

★ **Theorem:** Let $X \sim N_n(\mu, I)$. If A is a symmetric and idempotent matrix with $\text{rank}(A) = s$, then $X'AX \sim \chi^2_s(\frac{1}{2}\mu'A\mu)$

Proof: $A = GG' \Rightarrow G'X \sim N_s(G'\mu, G'IG) \equiv N_s(G'\mu, I_s)$

$\Rightarrow XGG'X \equiv XAX \sim \chi^2_s(\frac{1}{2}\mu'GG'\mu) \equiv \chi^2_s(\frac{1}{2}\mu'A\mu)$

Usage: $Y \sim N(X\beta, \sigma^2 I)$

$\frac{Y}{\sigma} \sim N(\frac{X\beta}{\sigma}, I)$

$Y'(P_X - P_{X_1})Y \sim \chi^2_{\text{rank}(P_X - P_{X_1})}\left(\frac{\beta'X(P_X - P_{X_1})X\beta}{2\sigma^2}\right)$

Theorem: Let $X \sim N_n(\mu, V)$. If A is a symmetric and AV idempotent with $\text{rank}(AV) = s$, then $X'AX \sim \chi^2_s(\frac{1}{2}\mu'A\mu)$

Proof: $V = LL'$ for V symmetric & P.D. $\Rightarrow Y := L^{-1}X \sim N_p(L^{-1}\mu, I_p)$

$X'AX = X'(L')^{-1}L'AL L^{-1}X$

$= Y'L'ALY$

$= Y'BY \sim \chi^2_s(\frac{1}{2}(L^{-1}\mu)'B(L^{-1}\mu))$

$\equiv \chi^2_s(\frac{1}{2}\mu'A\mu)$

$\Rightarrow$ since $B' = (L'AL)' = L'AL$ is symmetric

$BB = L'ALL'AL = L'AVAL = L'AVAL\cdot L'(L')^{-1}$

$= L'AVAV(L')^{-1} = L'AV(L')^{-1} = L'AL$

is idempotent

$\text{rank}(B) = \text{trace}(B) = \text{trace}(L'AL) = \text{trace}(LL'A) = \text{trace}(VA) = \text{rank}(AV) = s$

want to know how to proof this

Theorem: Let $X \sim N_p(\mu, V)$ and A symmetric. $\text{rank}(A) = s$. If $BVA = 0$ for a given B, then BX and $X'AX$ are independent.

Corollary: Let $X \sim N_p(\mu, V)$, Let A be a symmetric matrix with $\text{rank}(A) = r$. let B be a symmetric matrix with $\text{rank}(B) = s$. If $BVA = 0$, then $X'BX$ and $X'AX$ are independent.

Cochran's theorem: Let $Y \sim N_n(\mu, \sigma^2 I_n)$ and let A_i be symmetric and idempotent with $\text{rank}(A_i) = s_i$ $\forall i = (1, \cdots, k)$. If $\sum_{i=1}^k A_i = I_n$, then $\frac{1}{\sigma^2}Y'A_iY \sim \chi^2_{s_i}(\frac{1}{2\sigma^2}\mu'A_i\mu)$, $\sum_{i=1}^k s_i = n$, and $\frac{1}{\sigma^2}Y'A_1Y, \cdots, \frac{1}{\sigma^2}Y'A_kY$ are independent.

Distribution relationship

$N(0,1)^2 \sim \chi^2_1$; $N(\mu,1)^2 \sim \chi^2_1(\frac{1}{2}\mu^2)$; $N(\mu,\sigma^2)^2 \sim \sigma^2\chi^2_1(\frac{\mu^2}{2\sigma^2})$

$\frac{N(0,1)}{\sqrt{\frac{1}{(n-p)}\chi^2_{n-p}}} \sim t_{n-p}$; $t_p^2 \sim F_{1,p}$; χ^2_a indep. χ^2_b, $\frac{\chi^2_a/a}{\chi^2_b/b} \sim F_{a,b}$

$\chi^2_k \equiv \text{Gamma}(\frac{k}{2}, 2)$; $\overset{iid}{\Sigma_i} \chi^2_{ai(bi)} \equiv \chi^2_{\Sigma ai(\Sigma bi)}$ (2: scale)

$\equiv \text{Gamma}(\frac{k}{2}, \frac{1}{2})$ (rate)

$\overset{iid}{\Sigma_i} \exp(\lambda) \equiv \text{Gamma}(n, \lambda)$; $\frac{1}{n}\overset{iid}{\Sigma_i} \exp(\lambda) \equiv \text{Gamma}(n, \frac{\lambda}{n})$ (scale (mean))

$\equiv \text{Gamma}(n, \frac{n}{\lambda})$ (rate)

scale = 1 / rate

$f_x = \frac{1}{\lambda} e^{-x/\lambda}$ λ: scale $1/\lambda$: rate

$\overset{iid}{\Sigma_i} \text{Gamma}(a_i, b) \equiv \text{Gamma}(\Sigma_i a_i, b)$; $\frac{1}{b}\text{Gamma}(a, b) \equiv \text{Gamma}(a, 1)$

$\text{Poi}(\mu)$ indep. $\text{Poi}(\lambda)$, $\text{pois}(\mu) + \text{poi}(\lambda) = \text{poi}(\mu+\lambda)$

For poisson, rate = scale = mean = λ

$\exp(\beta) = \text{weibull}(1, \beta) = \text{Gamma}(1, \beta)$

$X \sim \text{weibull}(a,b) \Leftrightarrow X^a \sim \exp(b^a)$ both scale-liked.

X indep $Y \sim N(0,1)$; $\frac{X}{Y} \sim \text{Cauchy}(0,1)$

$\text{Beta}(1,1) = \text{Unif}(0,1)$

neg-binomial$(1,p)$ = geometric(p)

Inequalities

Boole's Inequality

$$P\left(\bigcup_{i=1}^{\infty} E_i\right) \leq \sum_{i=1}^{\infty} P(E_i)$$

Triangular Inequality

$$\|X+Y\| \leq \|X\| + \|Y\|$$

Product Inequality

matrix norm

$$\|X \times Y\| \leq \|X\| \times \|Y\|$$

Markov's Inequality

$$P(X > \varepsilon) \leq \frac{E(X)}{\varepsilon}$$

Chebyshev's Inequality

$$P(|X - E(X)| > \varepsilon) \leq \frac{Var(X)}{\varepsilon^2}$$

Jonson's inequality

$E(g(X)) \geq g(E(X))$ if $g(\cdot)$ is convex $\quad \Sigma \theta_i f(x_i) \geq f(\Sigma \theta_i x_i)$

$E(g(X)) \leq g(E(X))$ if $g(\cdot)$ is concave

* Equality meet when :

1. X is constant almost surely :

$$P(X=c) = 1 \Rightarrow E(\phi(X)) = \phi(E(X))$$

2. ϕ is linear (affine) over the range of X :

$$\phi(x) = ax+b$$

a.k.a. if ϕ is differentiable and convex, then

$$\phi(X) \geq \phi(EX) + \phi'(EX)(X-EX)$$

{ by Taylor expansion
{ first-order condition in convexity

with equality iff $\forall$ tangent line of $g(\cdot)$ at EX ,

$$P(\phi(X) = \phi(EX) + \phi'(EX)(X-EX)) = 1$$

Cauchy-Schwartz Inequality

$$Cov(U,V) \leq \sqrt{Var(U)\,Var(V)} \Leftrightarrow (E(XY))^2 \leq E(X^2)E(Y^2)$$ equal iff $X = aY$

Set Operations

Properties of set operations

Commutativity:

$$A \cup B = B \cup A \; ; \; A \cap B = B \cap A \; ; \; A \triangle B = B \triangle A$$

$$A \triangle B = (A \setminus B) \cup (B \setminus A) = (A \cup B) \setminus (A \cap B) = (A \cap B^c) \cup (B \cap A^c)$$

Associativity:

$$(A \cup B) \cup C = A \cup (B \cup C) \; ; \; (A \cap B) \cap C = A \cap (B \cap C) \; ;$$

$$(A \triangle B) \triangle C = A \triangle (B \triangle C)$$

Distributivity:

$$A \cap (B \cup C) = (A \cap B) \cup (A \cap C)$$

$$A \cup (B \cap C) = (A \cup B) \cap (A \cup C)$$

De Morgan's law:

$$(A \cup B)^c = A^c \cap B^c \; ; \; (A \cap B)^c = A^c \cup B^c$$

Variable transformation

General Methods:

1. By CDF , MGF
2. By pdf / pmf :
   1. Law of total probability (when in $\mathbb{R}$)
   2. Jacobian (when in $\mathbb{R}^k$)
      (when sign changes)
      ↑ be aware the range of new parameters ← from the range of olds.
3. By distribution relationship recognition

Example

Problem

Find the value of the constant c which make the following function a valid probability density function of (X, Y). And then calculate $P(X \geq 2, Y \geq 3)$

$$f(x,y) = ce^{-x-y} I_{(0 \leq x \leq y)}$$

$$\iint_{y} f(x,y)\,dx\,dy \overset{\text{set}}{=} 1$$

$$= \int_0^\infty \int_0^y ce^{-x-y}\,dx\,dy$$

$$= \int_0^\infty ce^{-y}(-e^{-x})\Big|_0^y\,dy$$

$$= \int_0^\infty ce^{-y}(1-e^{-y})\,dy$$

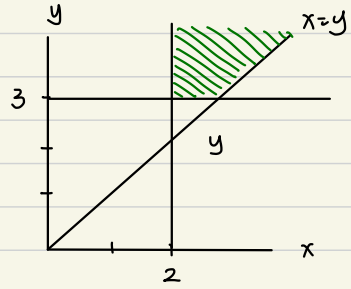
$$= c\int_0^\infty e^{-y} - e^{-2y}\,dy$$

$$= c(-e^{-y} + \tfrac{1}{2}e^{-2y})\Big|_0^\infty$$

$$= c(-0+0+1-\tfrac{1}{2})$$

$$= \frac{c}{2} \overset{\text{set}}{=} 1$$

$$\Rightarrow c = 2$$



$$P(X \geq 2, Y \geq 3) = \int_3^\infty \int_2^y 2e^{-x-y}\,dx\,dy$$

$$= \int_3^\infty 2e^{-y}(-e^{-x})\Big|_2^y\,dy$$

$$= \int_3^\infty 2e^{-y}(-e^{-y}+e^{-2})\,dy$$

$$= -2\int e^{-2y}\,dy + 2e^{-2}\int e^{-y}\,dy$$

$$= -2\cdot(-\tfrac{1}{2})e^{-2y} - 2e^{-2-y}$$

$$= e^{-2y} - 2e^{-y-2}\Big|_3^\infty$$

$$= -e^{-6} + 2e^{-5}$$

$$= 2e^{-5} - e^{-6}$$

$2 \leq x \leq y$

$3 \leq y \leq \infty$

Problem

When a random vector (X, Y) has the following joint pdf

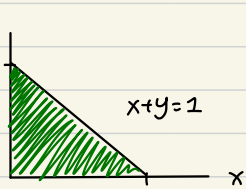
$$f(x,y) = 120xy(1-x-y)I_{(x\geq 0,\ y\geq 0,\ x+y\leq 1)}$$

, find the value of $\mathbb{E}(XY)$

$$E(XY) = \iint_{x,y} xy\, f(x,y)\, dy dx$$

$$= \int_0^1 \int_0^{1-y} xy \cdot 120xy(1-x-y)\, dx dy$$

$$= \int_0^1 \int_0^{1-y} 120x^2y^2(1-x-y)\, dx dy$$



$0 \leq x \leq 1-y$

$0 \leq y \leq 1$

$$= \int_y \int_x 120x^2y^2\, dx dy - \iint_x 120x^3y^2\, dx dy - \int_y \int_x 120x^2y^3\, dx dy$$

$$= \int_y 40x^3y^2 \Big|_0^{1-y} dy - \int_y 30x^4y^2 \Big|_0^{1-y} dy - \int_y 40x^3y^3 \Big|_0^{1-y} dy$$

$$= \int_y 40(1-y)^3y^2 - 30(1-y)^4y^2 - 40(1-y)^3y^3\, dy$$

$$= \int_y (1-y)^3y^2(40 - 30(1-y) - 40y)\, dy$$

$$= \int_y (1-y)^3y^2(10 - 10y)\, dy$$

$$= 10\int_y (1-y)^4y^2\, dy$$

$t = 1-y \quad dt = -dy$

$y = 1-t$

$0 \leq y \leq 1$

$t=1 \quad t=0$

$$= -10\int_t t^4(1-t)^2\, dt$$

$$= -10\int_t t^4(1-2t+t^2)\, dt$$

$$= -10\int_t t^4 - 2t^5 + t^6\, dt$$

$$= -10\left(\frac{1}{5}t^5 - \frac{1}{3}t^6 + \frac{1}{7}t^7\right)\Big|_1^0$$

$$= 10\left(\frac{1}{5} - \frac{1}{3} + \frac{1}{7}\right)$$

$$= 10\left(\frac{21}{105} - \frac{35}{105} + \frac{25}{105}\right)$$

$$= 10\left(\frac{11}{105}\right)$$

$$= \frac{110}{105}$$

$$= \frac{22}{21}$$

Question

If a random vector $(X_1, X_2, X_3)^\intercal$ has the following joint pdf,

$$f(x_1, x_2, x_3) = 120x_1(1 - x_1 - x_2 - x_3)I_{(x_1 \geq 0,\ x_2 \geq 0,\ x_3 \geq 0,\ x_1+x_2+x_3 \leq 1)}$$

calculate $\mathbb{E}(X_1X_2)$

easy to start with z
last to integrate x

$$\begin{aligned}
E(XY) &= \int_0^1 \int_0^{1-x} \int_0^{1-x-y} xy\ 120x(1-x-y-z)\,dz\,dy\,dx \\
&= \int_0^1 120x^2 \int_0^{1-x} y \int_0^{1-x-y} (1-x-y-z)\,dz\,dy\,dx \\
&= \int_0^1 120x^2 \int_0^{1-x} y\left(z - xz - yz - \tfrac{1}{2}z^2\right)\Big|_0^{1-x-y} dy\,dx \\
&= \int_0^1 120x^2 \int_0^{1-x} y(1-x-y - x(1-x-y) - y(1-x-y) - \tfrac{1}{2}(1-x-y)^2)\,dy\,dx \\
&= \int_0^1 120x^2 \int_0^{1-x} y(1-x-y-x+x^2+xy-y+xy+y^2-\tfrac{1}{2}(1-2xy-2x-2y+x^2+y^2)\,dy\,dx \\
&= \int_0^1 120x^2 \int_0^{1-x} -2xy + x^2y + 2xy^2 - y^2 + y^3 - \tfrac{1}{2}y + xy^2 + xy + y^2 - \tfrac{1}{2}x^2y - \tfrac{1}{2}y^3\,dy\,dx \\
&= -xy + \tfrac{1}{2}x^2y + 3xy^2 + \tfrac{1}{2}y^3 - \tfrac{1}{2}y \\
&\vdots
\end{aligned}$$

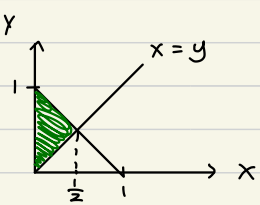
Problem

When a random vector (X, Y) has the following joint pdf

$$f_{1,2}(x, y) = 10xy^2 I_{(0<x<y<1)}$$

calculate $P(X + Y \leq 1)$

$$\begin{aligned}
P(X+Y \leq 1) &= \int_0^{\frac{1}{2}} \int_x^{1-x} 10xy^2\,dy\,dx \\
&= \int_0^{\frac{1}{2}} \tfrac{10}{3}xy^3\Big|_x^{1-x} dx \\
&= \int_0^{\frac{1}{2}} \tfrac{10}{3}(x(1-x)^3 - x^4)\,dx \\
&= \int_0^{\frac{1}{2}} \tfrac{10}{3}(x(1-2x+x^2)(1-x) - x^4)\,dx \\
&= \int_0^{\frac{1}{2}} \tfrac{10}{3}(x - 3x^2 + 3x^3 - 2x^4)\,dx \\
&= \tfrac{5}{3}x^2 - \tfrac{10}{3}x^3 + \tfrac{5}{2}x^4 - \tfrac{4}{3}x^5\Big|_0^{\frac{1}{2}}
\end{aligned}$$



$$\begin{cases} x < y < 1-x \\ 0 < x < \frac{1}{2} \end{cases}$$

$$+ \begin{cases} 0 < x < y \\ 0 < y < \frac{1}{2} \end{cases} + \begin{cases} 0 < x < 1-y \\ \frac{1}{2} < y < 1 \end{cases}$$

Problem

When a random vector (X, Y) has the following joint pdf

$$f_{1,2}(x, y) = cx^2 y I_{(0<x<y<1)}$$

find the constant c, and calculate $P(Y \le 2X)$

$$\int_{x,y} f(x,y)\,dx\,dy \overset{\text{set}}{=} 1$$

$$= c \int_0^1 \int_0^y x^2 y \,dx\,dy$$

$$= c \int_0^1 y \int_0^y x^2 dx\,dy$$

$$= c \int_0^1 y \tfrac{1}{3} x^3 \Big|_0^y dy$$

$$= \frac{c}{3} \int_0^1 y^4 dy$$

$$= \frac{c}{15} y^5 \Big|_0^1$$

$$= \frac{c}{15} \overset{\text{set}}{=} 1$$

$$c = 15$$

$$P(Y \le 2X)$$

$$= \int_0^1 \int_{\frac{y}{2}}^y 15x^2 y \,dx\,dy$$

$$= \int_0^1 5x^3 y \Big|_{\frac{y}{2}}^y dy$$

$$= \int_0^1 5y^4 - \frac{5}{8} y^4 dy$$

$$= y^5 - \frac{1}{8} y^5 \Big|_0^1$$

$$= \frac{7}{8}$$

should always find the clear cut-off

Problem

When a random vector (X, Y) has the following joint pdf

$$f_{1,2}(x, y) = e^{-y} I_{(0<x<y<\infty)}$$

calculate $P(X + Y \le 1)$

$$P(X+Y \le 1) = \int_0^{1/2} \int_x^{1-x} e^{-y} dy\,dx$$

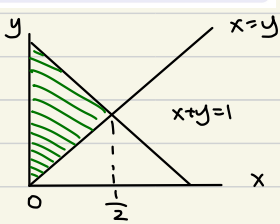
$$= \int_0^{\frac{1}{2}} -e^{-y} \Big|_x^{1-x} dx$$

$$= \int_0^{\frac{1}{2}} -e^{-(1-x)} + e^{-x} dx$$

$$= -e^{-(1-x)} - e^{-x} \Big|_0^{\frac{1}{2}}$$

$$= -e^{-\frac{1}{2}} - e^{-\frac{1}{2}} + e^{-1} + 1$$

$$= 1 + e^{-1} - 2e^{-\frac{1}{2}}$$



Reparameterization & BLUE

Reparameterization

Definition (Reparameterization): Let $X \in \mathbb{R}^{n \times p}$ and $W \in \mathbb{R}^{n \times q}$
the linear models $Y = X\beta + U$ and $Y = Wr + U$ are reparameterizations of each other, if
$Col(X) = col(W)$

Theorem: Suppose $col(X) = col(W)$, then $P_X = P_W$.

Proposition: If $col(X) = col(W)$, then $\exists$ a matrix S s.t. $X = WS$
Similarly, $\exists\ T$ s.t. $W = XT$

Theorem: Let $col(X) = col(W)$.
Suppose $W = XT$ and $\hat{r}$ solves the normal equations in W.
1). The fitted data & residuals are the same:
$\hat{y} = P_X y = P_W y \Leftarrow X(X'X)^g X'y = W(W'W)^g W'y$ (P_X, P_W)
$\hat{e} = (I - P_X)y = (I - P_W)y$
2). $\hat{\beta} = T\hat{r}$ solves the normal equations in X

$W = XT$ is when $col(X) \subseteq col(W)$

$Y = X\beta = Wr$
$= XTr \Rightarrow \hat{\beta} = T\hat{r}$

Theorem: If $\lambda'\beta$ is estimatable in the model with design X, and $\hat{r}$ solves the normal equation in design W, then $\lambda' T \hat{r}$ is the least square estimator of $\lambda'\beta$.

Theorem: If $\delta' r$ is estimatable in design W; then $\delta' S\beta$ is estimatable in design X, and its least square estimator is $\delta'\hat{r}$, where $\hat{r} \in \{w'wr = w'y\}$

$Y = Wr = X\beta$
$= WS\beta$
$\Rightarrow \hat{r} = S\hat{\beta}$

Constraint model

$col(X') \perp col(C')$

Lemma: Suppose $C \in \mathbb{R}^{(p-r)\times p}$, $r := \text{rank}(X)$, and $col(X') \cap col(C') = \{0\}$. Then the following systems are equivalent:

(1) $\begin{pmatrix} X'X \\ C'C \end{pmatrix}\beta = \begin{pmatrix} X'y \\ 0 \end{pmatrix}$

(2) $\begin{pmatrix} X'X \\ C \end{pmatrix}\beta = \begin{pmatrix} X'y \\ 0 \end{pmatrix}$

(3) $(X'X + C'C)\beta = X'y$

Theorem: Let $C \in \mathbb{R}^{(p-r)\times p}$ with $\text{rank}(C) = p-r$ and $col(X') \cap col(C') = \{0\}$. Then:

(1) The matrix $X'X + C'C$ is non-singular

(2) $(X'X + C'C)^{-1}X'y$ uniquely solves $X'X\beta = X'y$, $C'C\beta = 0$

(3) $(X'X + C'C)^{-1}$ is a generalized inverse of $X'X$

(4) $C(X'X + C'C)^{-1}X' = 0$

(5) $C(X'X + C'C)^{-1}C' = I$

Definition: The function $\lambda'\beta$ is estimatable in the restricted model if and only if there exists a scalar c and vector a s.t. $E(c + a'y) = \lambda'\beta$, $\forall \beta \in \{P'\beta = \delta\}$

$$\begin{pmatrix} X'X & P \\ P' & 0 \end{pmatrix}\begin{pmatrix} \beta \\ \theta \end{pmatrix} = \begin{pmatrix} X'y \\ \delta \end{pmatrix}$$

Restricted Model: $Y = X\beta + U$, $E(U) = 0$, $\beta \in \{\beta : \underbrace{P'\beta = \delta}_{\text{full-col rank}}\}$

Remember this

Theorem: In the restricted model, $c + a'y$ is unbiased for $\lambda'\beta$ if and only if $\exists d$ s.t. $\lambda = X'a + Pd$ and $c = d'\delta$.

($\Rightarrow$) $E(c + a'y)$
$= d'\delta + a'X\beta$
$= d'\delta + a'X\beta + d'P'\beta - d'P'\beta$
$= d'\delta + \lambda'\beta - d'\delta$
$= \lambda'\beta$

($\Leftarrow$) $\lambda'\beta = d'P'\beta + a'X\beta$
$= d'\delta + a'E(y)$

Theorem: If $\delta \in Col(P')$, then there exists a solution to the RNEs.

Restriction: $P'\beta = \delta \Leftarrow \delta \in Col(P')$

Theorem: If $\hat{\beta}_H$ denotes the first component of a solution to true RNE's:

$$\begin{pmatrix} \hat{\beta}_H \\ \hat{\theta}_H \end{pmatrix} \in \left\{ \begin{pmatrix} X'X & P \\ P' & 0 \end{pmatrix} \begin{pmatrix} \beta \\ \theta \end{pmatrix} = \begin{pmatrix} X'y \\ \delta \end{pmatrix} \right\}$$

then $\hat{\beta}_H$ minimizes $Q(\beta) = \|y - X\beta\|^2$ over $\{P'\beta = \delta\}$

Theorem: If $\hat{\beta}_H$ denotes the first component of a solution to true RNE's:

$$\begin{pmatrix} \hat{\beta}_H \\ \hat{\theta}_H \end{pmatrix} \in \left\{ \begin{pmatrix} X'X & P \\ P' & 0 \end{pmatrix} \begin{pmatrix} \beta \\ \theta \end{pmatrix} = \begin{pmatrix} X'y \\ \delta \end{pmatrix} \right\}$$

if $\beta \in \{P'\beta = \delta\}$, then $Q(\beta) = Q(\hat{\beta}_H)$ iff β also solves RNE.

Not only RNE gives a L.S. solution, but a L.S solution MUST solves RNE.

Best Linear Unbiased Estimator

* a.k.a unbiased estimator that achieves CRLB.

Gauss-Markov Model

$$\left.\begin{aligned} &Y = X\beta + U \\ &E(U) = 0 \\ &Cov(U) = \sigma^2 I \quad \text{for some } \sigma > 0 \end{aligned}\right\} \text{ Gauss-Markov Assumption}$$

Theorem (Gauss-Markov): Under the Gauss-Markov assumptions/model, if $\lambda'\beta$ is estimable, then $\lambda'\hat{\beta}$ is the best (minimum variance) linear unbiased estimator of $\lambda'\beta$, $\forall \hat{\beta} \in \{X'X\beta = X'y\}$.

(For any unbiased linear estimator, the one with least variance is the LSE).

Theorem: An unbiased estimator of σ^2 is $Y'(I-P_X)Y/(n-p)$

Aitken Model

$Y = X\beta + u$

$E(u) = 0$

$\text{Var}(u) = \sigma^2 V$

$V > 0$ known

$w'Vw > 0 \quad \forall w \in \mathbb{R}^n$

Theorem (AITKEN)

So this is a Gauss-Markov Model. Consequently, the BLUE for any estimable $\lambda'\beta$ is $\lambda'\hat{\beta}_{GLS}$ where $\hat{\beta}_{GLS} \in \{x'v^{-1}x\beta = x'v^{-1}y\}$

Remember this ⇒

Theorem: The estimator $t'y$ is BLUE for $E(t'y)$ if and only if $t'y$ is uncorrelated with all unbiased estimators of zero.

Corollary: Under the Aitken Model, $t'y$ is BLUE for $E(t'y)$ iff $Vt \in \text{col}(X)$

Theorem: Under the aitken model, $\lambda'\hat{\beta}_{OLS}$ is the BLUE for estimable $\lambda'\beta$ iff $\exists Q$ s.t. $VX = XQ$

Bayes prior & posterior

Conjugate Priors

Prior	Data	Posterior	
Beta	Bernoulli	Beta	
Beta	Binomial	Beta	think parameters in these are p.
Beta	Geometric	Beta	with Beta $0<x<1$ continuous fits well.
Beta	Neg-Binomial	Beta	
Gamma	Poisson	Gamma	
Gamma	exponential	Gamma	may think mirroring the Beta
Gamma	Gamma	Gamma	
Dirichlet	Multinomial	Dirichlet	remember these two separately
Dirichlet	Hypergeometric	Dirichlet	
Normal	Normal	Normal	

Convergence Orders

Notation	Deterministic or Stochastic?	Formal Definition	Intuition	Example
$O(b_n)$	Deterministic	$\exists C > 0$ s.t. $begin:math:text$	a_n	\le C
$o(b_n)$	Deterministic	$\frac{a_n}{b_n} \to 0$ as $n \to \infty$	Negligible compared to b_n	$n = o(n^2)$
$O_p(a_n)$	Stochastic	$\forall \varepsilon > 0, \exists M > 0$ s.t. $begin:math:text$\Pr(	X_n	> M a_n) < \varepsilon$end:math:text$ for large n
$o_p(a_n)$	Stochastic	$X_n/a_n \xrightarrow{p} 0$	Small compared to a_n with high probability	If $X_n \sim N(0, 1/n^2)$, then $X_n = o_p(n^{-1/2})$

Ancillary & Completeness

Show ancillary by family

Location family ①	Scale family ②	location & Scale family ③
$f(x;\mu) = f_0(x-\mu)$	$f(x;\sigma) = \frac{1}{\sigma} f_0(\frac{x}{\sigma})$: σ scale	$f(x;\mu,\sigma) = \frac{1}{\sigma} f_0(\frac{x-\mu}{\sigma})$
$N(\mu,\sigma^2)$ with known σ^2	exponential	$N(\mu,\sigma^2)$
Laplace with fixed scale	chi-squared with ϕ	Cauchy
Cauchy with fixed scale	Gamma with fixed shape	Laplace
Uniform $(\mu, \mu+c)$	Weibull with fixed shape	Logistic
Logistic with fixed scale		Uniform $(\mu-c\sigma, \mu+c\sigma)$

Location family ancillary statistics ①

- Sample spacing : $(X_{(2)}-X_{(1)}, X_{(3)}-X_{(2)}, \cdots, X_{(n)}-X_{(n-1)})$
- Sample range : $X_{(n)}-X_{(1)}$
- Sample variance : $S_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X_n})^2$

Example 1:

Location family, want to show $X_{(n)}-X_{(1)}$ ancillary.

Show $X-\theta \sim \exp(1)$ invariant of θ, thus $(X_{(n)}-\theta)-(X_{(1)}-\theta)$ is also invariant of θ

Thus $X_{(n)}-X_{(1)}$ is ancillary.

Scale family ancillary statistics ②

- Sample ratio : $(\frac{X_{(1)}}{X_{(n)}}, \cdots, \frac{X_{(n-1)}}{X_{(n)}})$
- The t-statistics : $\frac{\bar{X_n}}{S_n}$ OR robust $\tilde{t} = \frac{\tilde{X_n}}{X_{(\frac{3n}{4})}-X_{(\frac{n}{4})}}$ ← median

location - scale family ancillary statistics ③

- normalized sample spacing : $(\frac{X_{(2)}-X_{(1)}}{X_{(n)}-X_{(1)}}, \cdots, \frac{X_{(n)}-X_{(n-1)}}{X_{(n)}-X_{(1)}})$
- sample range / sample s.d. ratio : $\frac{X_{(n)}-X_{(1)}}{S_n}$

Uniqueness of power series

★ use this to proof completeness.

Example 1

$X_i \overset{iid}{\sim} \text{Poisson}(\lambda)$; $\sum X_i \overset{iid}{\sim} \text{Poisson}(n\lambda)$

$$E(g(t)) = \sum_{t=0}^{\infty} g(t)\, e^{-n\lambda} \frac{(n\lambda)^t}{t!} \overset{\text{set}}{=} 0$$

$$\Rightarrow \quad \sum_{t=0}^{\infty} g(t) \frac{(n\lambda)^t}{t!} = 0$$

$$\Rightarrow \quad \sum_{t=0}^{\infty} \frac{g(t)}{t!} (n\lambda)^t = 0$$

$$\Rightarrow \quad \frac{g(t)}{t!} = 0 \ \forall t \quad \text{by uniqueness of power series}$$

$$\Rightarrow \quad g(t) = 0 \ \forall t$$

Example 2

$X_i \overset{iid}{\sim} \exp(\lambda)$; $\sum X_i \overset{iid}{\sim} \text{Gamma}(n, \lambda)$

$$E(g(x)) = \int_0^{\infty} g(x) \frac{\lambda^a}{\Gamma(n)} x^{a-1} e^{-\lambda x}\, dx \overset{\text{set}}{=} 0$$

$$\Rightarrow \quad \int_0^{\infty} g(x)\, x^{a-1} e^{-\lambda x} dx = 0$$

$$\frac{\partial}{\partial \lambda} \Rightarrow \quad \int_0^{\infty} g(x)\, x^{a-1} (-x) e^{-\lambda x} dx = 0 \quad \Rightarrow \int_0^{\infty} g(x)(-x^a) e^{-\lambda x} dx = 0$$

$$\frac{\partial^2}{\partial \lambda^2} \Rightarrow \quad \int_0^{\infty} g(x)\, x^{a+1} e^{-\lambda x} dx = 0$$

$$\vdots$$

Thus $g(x) e^{-\lambda x} = 0$ everywhere $\forall x$

Thus $g(x) = 0$

Asymptotic Theory

Cramér-Rao Lower bound (RCLB) ← called Cramér condition

Under regularity conditions, the inverse of the fisher information is attained:

1. Model is correctly specified
2. The log-likelihood is sufficiently smooth (e.g. differentable up to second order)
3. The fisher information is positive definite and finite
4. The MLE exists and consistent
5. The score function has mean zero and finite variance

and this estimator reaches the CRLB

* Violation: Score has to be a linear function of an unbiased estimator.
But for MLE estimators, the CRLB always met asymptotically.

Wilk's Theorem assumptions

$H_0: \theta \in \Theta_0$ $H_1: \theta \in \Theta \setminus \Theta_0$

$$\Lambda = \frac{\sup_{\theta \in \Theta_0} L(\theta)}{\sup_{\theta \in \Theta} L(\theta)} \qquad -2\log\Lambda \xrightarrow{d} \chi^2_K \text{ (by wilk's theorem)} \; ; K = \dim(\Theta) - \dim(\Theta_0)$$

Assumptions:

1. $\theta \in \Theta$ uniquely determine the distribution of data
   $f(x|\theta_1) \neq f(x|\theta_2)$ if $\theta_1 \neq \theta_2$
2. True parameter θ_0 must lie in the interior of parameter space Θ.
   If on the boundary, the limiting distribution might not be chi-square.
3. Regularity of Likelihood
   Twice differentiable in θ
   Well-defined, non-singular fisher information
   Expectation of derivatives exists and can be interchanged with integration
4. Large Sample $n \to \infty$ to use Wilk
5. MLE behavior Consistent & asymptotic normal
   $\hat{\theta}_{MLE} \xrightarrow{p} \theta_0 \; ; \sqrt{n}(\hat{\theta} - \theta_0) \xrightarrow{d} N(0, I^{-1}(\theta_0))$
6. Nested Model
   $\Theta_0 \subseteq \Theta$ ← ensures the likelihood ratio is well-defined

Violations: Zero-inflation Poisson & mixture distribution

Cannot test for $H_0: \pi = 0, 1$; $H_0: p =$ any value $\in [0, 1]$

Can test for $H_0: \pi =$ some value $\in (0, 1)$

Law of large number (weak)

Suppose $X \sim D(\cdot)$, $E(X) = \mu$, $Var(X) = \sigma^2 < \infty$

Thus $\bar{X} \xrightarrow{P} \mu$

Convergence in Probability

$P(|X_n - X| > \varepsilon) \longrightarrow 0$ — can be some scalar

1. By definition
2. Use WLLN (But must know at least mean & variance)
3. If $\lim Var(X_n) = 0$, and $\lim E(X_n) = a$, then $X_n \xrightarrow{P} a$

$$P(|X_n - a| > \varepsilon) \leq \frac{E(X_n - a)^2}{\varepsilon^2}$$

$$= \frac{E(X_n^2 - 2X_n a + a^2)}{\varepsilon^2}$$

$$= \frac{E(X_n^2) - 2E(X_n)a + a^2}{\varepsilon^2} \longrightarrow Var(X_n)$$

$$\rightarrow \frac{Var(X_n)}{\varepsilon^2}$$

$$\rightarrow 0$$

When can't get rid of $|\cdot|$, use $E(|\cdot|^2)$ or $Var(|\cdot|)$, otherwise, use probability method.

Convergence in distribution

$X_n \xrightarrow{d} X$ if $F_{X_n} \longrightarrow F_X \ \forall x$

1. Central Limit Theorem (CLT)
2. Delta theorem
3. Slutsky theorem (usually CLT + WLLN)
4. By definition

Central Limit Theorem

1. For sample mean

$X_i \overset{iid}{\sim} D$ $\quad E(X_i) = \mu$ $\quad Var(X_i) = \sigma^2 < \infty$

$\sqrt{n}(\bar{X} - \mu) \xrightarrow{d} N(0, \sigma^2)$

2. For some estimator

$E(\hat{\theta_n}) = \theta_0$ or bias $= o(\frac{1}{\sqrt{n}})$ $\quad \hat{\theta_n} \xrightarrow{P} \theta_0$ $\quad \sigma^2 < \infty$

$\sqrt{n}(\hat{\theta_n} - \theta_0) \xrightarrow{d} N(0, \sigma^2)$

Delta Theorem

$\sqrt{n}(\bar{X}-\mu) \xrightarrow{d} N(0, \sigma^2)$

Consider $g(x)$ differentiable and $g'(\mu) \neq 0$.

$\sqrt{n}(g(\bar{x})-g(\mu)) \xrightarrow{d} N(0, g'(\mu)^2\sigma^2)$

Slutsky Theorem

Suppose $X_n \xrightarrow{d} X$, $Y_n \xrightarrow{p} c$, then

- $aX_n + bY_n \xrightarrow{d} aX + bc$
- $X_n Y_n \xrightarrow{d} cX$

Continuous Mapping Theorem

$\{X_n\}$ be a k-dimensional random vectors. Let $g: \mathbb{R}^k \to \mathbb{R}^L$ be a continuous function. Then

$$X_n \xrightarrow{p} X \Rightarrow g(X_n) \xrightarrow{p} g(X)$$

$$X_n \xrightarrow{d} X \Rightarrow g(X_n) \xrightarrow{d} g(X)$$

$$X_n \xrightarrow{a.s.} X \Rightarrow g(X_n) \xrightarrow{a.s.} g(X)$$

Examples:

Question

Let's define $Y_n \triangleq \min_{1\leq i\leq n} X_i$, where $X_1, \cdots, X_n \overset{i.i.d}{\sim} f(x;\theta) = e^{-(x-\theta)} I_{(\theta,\infty)}(x)$ then,

$$Y_n \xrightarrow[n\to\infty]{p} \theta$$

$$\begin{aligned}
&P(|Y_n-\theta|>\varepsilon) \\
&= P(|\min X_i-\theta|>\varepsilon) \\
&= P(\min X_i > \varepsilon+\theta) \quad \text{as } \theta<X_i \\
&= P(X_i > \varepsilon+\theta)^n \\
&= \left(\int_{\varepsilon+\theta}^{\infty} e^{-(x-\theta)}dx\right)^n \\
&= \left(e^{\theta}(-e^{-x})\Big|_{\varepsilon+\theta}^{\infty}\right)^n \\
&= \left(e^{\theta}e^{-(\varepsilon+\theta)}\right)^n \\
&= e^{-n\varepsilon} \longrightarrow 0 \quad \text{as } n\to\infty
\end{aligned}$$

Question

If $X_1 \cdots, X_n \overset{i.i.d}{\sim} \underline{Beta(1,\alpha)}$, what is the pdf of the limit distribution of $n^{1/\alpha}(1 - \max_{1\le i\le n} X_i)$?

$P(n^{\frac{1}{\alpha}}(1-\max X_i) \le x)$ ← CDF

$= P(1-\max X_i \le x \cdot n^{-\frac{1}{\alpha}})$

$= P(\max X_i \ge 1 - x n^{-\frac{1}{\alpha}})$

$= 1 - P(\min X_i \le 1 - x\cdot n^{-\frac{1}{\alpha}})$

$= 1 - P(X_i \le 1 - x\cdot n^{-\frac{1}{\alpha}})^n$

$= 1 - (1 - \underbrace{(1 - \frac{x}{n^{1/\alpha}})^{\alpha}}_{\longrightarrow \frac{x^\alpha}{n}})^n$

$\longrightarrow 1 - e^{-x^\alpha}$

$X \sim Beta(1,\alpha)$

$f_X(x) = \frac{\Gamma(\alpha+1)}{\Gamma(1)\Gamma(\alpha)} (x)^0 (1-x)^{\alpha-1}$

$= \alpha \cdot (1-x)^{\alpha-1}$

$F_X(x) = \int_0^x \alpha(1-x)^{\alpha-1} dx$

$= -(1-x)^\alpha \Big|_0^x$

$= 1 - (1-x)^\alpha \qquad 0 < x < 1$

$f_X(x) = \frac{\partial}{\partial x}(1 - e^{-x^\alpha})$

$= \frac{\partial g}{\partial h}\frac{\partial h}{\partial x}$

$= (-1)(-e^{-x^\alpha})\,\alpha x^{\alpha-1}$

$= \alpha x^{\alpha-1} e^{-x^\alpha} \quad I(0<x<\infty)$

cheat sheet

Law of total expectation (and related definitions)

$$E(Y) = E_X(E(Y|X))$$
$$= P(X>x)E(Y|X>x) + P(X\le x)E(Y|X\le x)$$
$$= \int_0^\infty P(Y>t)\,dt$$

Law of total variance

$$Var(Y) = E_X(Var(Y|X)) + Var_X(E(Y|X))$$
$$= E_Y(Y-EY)(Y-EY)$$
$$Cov(X,Y) = Cov(X, E(Y|X))$$
$$= E_x(XE(Y|X)) - E(X)E(Y)$$

Taylor Expansion:

$$g(x) = \sum_{k=0}^{\infty} \frac{g^{(k)}(a)(x-a)^k}{k!}$$

$\Rightarrow g(x) \ge g(a) + \nabla g(a)^T(x-a)$ when $g(\cdot)$ convex and differentiable.

Geometric summation:

$$\ln(1+x) = \sum_{n=0}^{\infty} \frac{(-1)^n}{n+1} x^{n+1} \quad \text{with } x\in(-1,1]$$

two instances

$$\frac{1}{1-x} = \sum_{n=0}^{\infty} x^n \quad \text{with } x\in(-1,1)$$
$$\frac{1}{1+x} = \sum_{n=0}^{\infty} (-x)^n \quad \text{with } x\in(-1,1)$$
$$e^x = \lim_{n\to\infty}\left(1+\frac{x}{n}\right)^n = \sum_{k=0}^{\infty}\frac{x^k}{k!}$$
$$\sum_{i=1}^{n} ar^{i-1} = \frac{a(1-r^n)}{1-r} \Rightarrow \sum_{i=1}^{\infty} ar^{i-1} = \frac{a}{1-r}$$
$$\sum_{i=1}^{n}(2i-1) = n^2$$

Usefull Property Given $z \sim E(z)=\mu,\ Var(z)=V$; $E(z'Az) = tr(AV) + \mu'A\mu$

use to find the expectation of SSR or SSE in forms that $y'Ay$

$$E(z'Az) = E((z-\mu+\mu)'A(z-\mu+\mu))$$
$$= E((z-\mu)'A(z-\mu) + \mu'A(z-\mu) + (z-\mu)'A\mu + \mu'A\mu)$$
$$= E((z-\mu)'A(z-\mu)) + 0 + 0 + \mu'A\mu$$
$$= E(trace((z-\mu)'A(z-\mu))) + \mu'A\mu$$
$$= E(trace(A(z-\mu)(z-\mu)')) + \mu'A\mu$$
$$= tr(E(A(z-\mu)(z-\mu)')) + \mu'A\mu$$
$$= tr(AV) + \mu'A\mu$$

Gamma function

$$\Gamma(x) = (x-1)\Gamma(x-1)$$
$$= (x-1)!$$
$$\Gamma(\tfrac{1}{2}) = \sqrt{\pi}$$
$$\Gamma(\tfrac{3}{2}) = \tfrac{1}{2}\cdot\Gamma(\tfrac{1}{2}) = \tfrac{\sqrt{\pi}}{2}$$

MGF for Normal

$$M_X(t) = e^{\frac{1}{2}t'\Sigma t + t'\mu}$$

MGF for χ^2

$$M_X(t) = (1-2t)^{-\frac{p}{2}} e^{\frac{2\phi t}{1-2t}}$$
$$EX = p + 2\phi$$
$$\mathrm{Var}X = 2p + 4\phi$$

Theorem (Gradient)

$a, b \in \mathbb{R}^n \quad A \in \mathbb{R}^{n\times n}$

(a). $\nabla_b a^T b = a$

(b). $\nabla_b b^T A b = Ab + A^T b = (A + A^T)b$

Sequential Sum of Square

If intercept : P_1

TABLE 7.1 ANOVA Table for Sequential SS

Source	df	Projection	SS	noncentrality
$\mathbf{b}_0$	$r(\mathbf{X}_0)$	$\mathbf{P}_{\mathbf{X}_0}$	$R(\mathbf{b}_0)$	$(2\sigma^2)^{-1}(\mathbf{Xb})^T\mathbf{P}_{\mathbf{X}_0}(\mathbf{Xb})$
$\mathbf{b}_1$ after $\mathbf{b}_0$	$r(\mathbf{X}_1^*) - r(\mathbf{X}_0)$	$\mathbf{P}_{\mathbf{X}_1^*} - \mathbf{P}_{\mathbf{X}_0}$	$R(\mathbf{b}_0, \mathbf{b}_1) - R(\mathbf{b}_0)$	$(2\sigma^2)^{-1}(\mathbf{Xb})^T(\mathbf{P}_{\mathbf{X}_1^*} - \mathbf{P}_{\mathbf{X}_0})(\mathbf{Xb})$
...	...			
$\mathbf{b}_j$ after $\mathbf{b}_0, \ldots, \mathbf{b}_{j-1}$	$r(\mathbf{X}_j^*) - r(\mathbf{X}_{j-1}^*)$	$\mathbf{P}_{\mathbf{X}_j^*} - \mathbf{P}_{\mathbf{X}_{j-1}^*}$	$R(\mathbf{b}_0, \ldots, \mathbf{b}_j) - R(\mathbf{b}_0, \ldots, \mathbf{b}_{j-1})$	$(2\sigma^2)^{-1}(\mathbf{Xb})^T(\mathbf{P}_{\mathbf{X}_j^*} - \mathbf{P}_{\mathbf{X}_{j-1}^*})(\mathbf{Xb})$
...	...			
$\mathbf{b}_k$ after $\mathbf{b}_0, \ldots, \mathbf{b}_{k-1}$	$r(\mathbf{X}_k^*) - r(\mathbf{X}_{k-1}^*)$	$\mathbf{P}_{X_k^*} - \mathbf{P}_{\mathbf{X}_{k-1}^*}$	$R(\mathbf{b}_0, \ldots, \mathbf{b}_k) - R(\mathbf{b}_0, \ldots, \mathbf{b}_{k-1})$	$(2\sigma^2)^{-1}(\mathbf{Xb})^T(\mathbf{P}_{\mathbf{X}} - \mathbf{P}_{\mathbf{X}_{k-1}^*})(\mathbf{Xb})$
Error	$N - r(X)$	$\mathbf{I} - \mathbf{P}_{\mathbf{X}}$	$\mathbf{y}^T\mathbf{y} - R(\mathbf{b})$	0
Total	N	$\mathbf{I}$	$\mathbf{y}^T\mathbf{y}$	$(2\sigma^2)^{-1}(\mathbf{Xb})^T(\mathbf{Xb})$