BIO753: Advanced Methods in Biostatistics, III *TERM 3, 2011*
T. A. Louis

March 17, 2011

Exercise 1 (Incorrect regressors: Due Feb. 01): You have data $(Y_i, X_i, T_i), i = 1, \ldots, n$ and use the linear model,

$$Y_i = \alpha_x + \beta_x X_i + \epsilon_i, \quad \epsilon_i \; iid \; (0, \sigma^2),$$

producing the LSE, $\hat{\beta}_x$. However, the *true linear model* is,

$$Y_i = \alpha_t + \beta_t T_i + e_i, \quad e_i \; iid \; (0, \sigma_*^2).$$

If you estimate using this model, you obtain the LSE, $\hat{\beta}_t$.

(a) Find $\beta_x(\mathbf{X}, \mathbf{T}, \beta_t) = E[\hat{\beta}_x \mid \mathbf{X}, \mathbf{T}, \beta_t]$ and $\beta_t(\mathbf{X}, \mathbf{T}, \beta_t) = E[\hat{\beta}_t \mid \mathbf{X}, \mathbf{T}, \beta_t]$.

(b) Represent the relation between $\beta_x(\mathbf{X}, \mathbf{T}, \beta_t)$ and $\beta_t(\mathbf{X}, \mathbf{T}, \beta_t)$ in terms of the LSE slope of T on X and $\rho(\mathbf{X}, \mathbf{T})$, the correlation between the Xs and the Ts.

(c) Briefly discuss your results.

§

Exercise 2 (Conservative LR Test: Due Feb. 04): In Bio752 we proved that the likelihood ratio test was optimal for testing a simple hypothesis versus a simple alternative and that one could guarantee control of the type I error at α by using the rule “Reject H_0 if LR $> c = 1/\alpha$.” Because this value of c provides control for all situations, it will be conservative (possibly very conservative) for a specific situation.

To study this issue, consider testing $H_0 : Y \sim N(0, 1)$ versus $H_1 : Y \sim N(\mu, 1)$ with a fixed $\mu > 0$. The optimal test for this Gaussian testing situation with type I error equal to α is “Reject if $Y > Z_{1-\alpha}$ (recall that Z_a leaves a area to the left).

1. For comparison, derive the rule that results from, "Reject H_0 if LR $> c = 1/\alpha$" and compare its actual type I error to α (the value for the optimal, Gaussian rule).
2. Note that this type I error will depend on the specific value of μ and so also find the μ value that produces the largest Type I error
3. Briefly discuss your results.

§

Exercise 3 (Concordance Correlation: Due Feb. 08): The concordance correlation, unlike the standard Pearson, product-moment correlation (ρ_{XY}^{pm}), is designed to penalize for differences in location and scale. One representation of it is,

$$\rho_{XY}^{con} = 1 - \frac{E[(X-Y)^2]}{E_{indep}[(X-Y)^2]},$$

where the denominator is computed assuming that X and Y are independent (or at least uncorrelated).

1. Show that ρ_{XY}^{con} is invariant to common changes in location and scale (both X and Y are transformed by the same location and scale values), but that it will change if the location and scale changes are not the same for X and Y.
2. Represent ρ_{XY}^{con} in terms of $(\mu_X, \sigma_X, \mu_Y, \sigma_Y, \text{cov}_{XY})$.
3. Represent ρ_{XY}^{con} by the Pearson correlation multiplied by two penalty factors, one that penalizes for unequal variances and one that penalizes for unequal means.
4. Briefly discuss situations when you would use ρ_{XY}^{pm} and ρ_{XY}^{con}.

§

Exercise 4 (CI for a variance: Due Feb. 11): Work out the CI for the variance using either the pivot approach or inverting a hypothesis test. §

Exercise 5 (Boundary situations: Due Feb. 15): Identify two interesting and relevant "on the boundary" situations and simulations you could do to study the large-sample theory. You don't have to do the simulations. §

Exercise 6 (Bonferroni vs Multivariate: Due Feb 18): Do the following.

1. Make a table for $\alpha = (0.001, 0.010, 0.025, 0.050)$ crosstabulated with $d = (2, 3, 5, 10, 100)$ that displays $(Z_{1-\alpha/2}, Z_{1-\alpha_{bf}(d,\alpha)/2}, \chi_{\{d,\alpha\}})$. The last component is the square-root of the α cutoff for a chi-square on d df.

2. For each d and α, find the $d^*(d, \alpha) : \chi_{\{d^*(d,\alpha),\alpha\}} = Z_{1-\alpha_{bf}(d,\alpha)/2}$. Note that $d^*(d, \alpha) \leq d$ is the effective number of dimensions for a multivariate approach that would match the limits produced by the Bonferroni approach. In general d^* will not be an integer and you need a recursive approach to find it.

3. Discuss the relations you find in parts 1 and 2.

§

Exercise 7 (Cluster Effects: Due Feb. 23): Consider a study with K clinics (clusters). We want to estimate the population mean. Within-cluster sample sizes are n_k and the vector of sample sizes is $\mathbf{n}$. The variance of the estimated population mean, when computed under the assumption of independence is V_{ind}. The variance V_{ind} needs to be adjusted by the "design effect." To this end we need the *intra-class correlation*,

$$\rho = \frac{\tau^2}{\sigma^2 + \tau^2}.$$

Here σ^2 is the within-cluster variability of a single observation (in ANOVA parlance, the within-column variability) and τ^2 is the between-cluster variability (the between-column variance component). Note that V_{ind} is computed using σ^2.

The required variance is $F(\mathbf{n}, \rho)V_{ind}$. With $n_+ = \sum_k n_k$ etc., we have,

$$F(\mathbf{n}, \rho) = \left[1 + \rho\left(\frac{\sum_k n_k^2}{n_+} - 1\right)\right] \tag{1}$$

If $n_k \equiv n$

$$= [1 + \rho\,(n - 1)]\,. \tag{2}$$

Here's what I want you to do:

1. See if you can derive formulas (**??**) and (**??**).
2. Let $\bar{n} = n_+/K$ and show that equation **??** is greater than or equal to equation **??** with $n = \bar{n}$. Indeed, represent $\frac{\sum_k n_k^2}{n_+}$ in terms of $\bar{n}$ and the coefficient of variation of the n_k.

§

Exercise 8 ("Center" of a Category: Due Mar. 01): You have a measured covariate but have partitioned it into categories (e.g, quartiles). You want to compute a summary score for each category and then use that summary as a numeric regressor in a linear regression. Propose what you would use for each category-specific "center" and (informally) justify your proposal. Work out the details for partitioning into two categories (e.g., above/below the median). §

End of Exercises

Criteria for estimates, Continued

Estimating the variance: Unbiased, MLE, Min MSE

- $(Y_1, \ldots, Y_n)$ iid. $S^2 = \frac{1}{n-1}\sum(Y_i - \bar{Y}_n)^2$
- $E(S^2) = \sigma^2$, so it is unbiased
- The MLE is: $\frac{n-1}{n}S^2$

- The minimum MSE estimate is: $\frac{n-1}{n+1}S^2$
- More generally, with df the degrees of freedom for a sum of squares, divide by df for the UBE, by n for the MLE and by $(df + 2)$ for minimum MSE

Variance component with known σ^2

- $\theta \sim N(\zeta, \tau^2)$, τ^2 unknown
- $Y \mid \theta \sim N(\theta, \sigma^2)$, σ^2 known
- $V(Y) = \sigma^2 + \tau^2$
- $(S^2 - \sigma^2)^+$ is biased, but better than $(S^2 - \sigma^2)$ which is unbiased

Other MLE Examples

- Gaussian $[X_i \mid \mu, \gamma_i\sigma] \sim N(\mu, \gamma_i^2\sigma^2)$, with known γ_i. This model produces a weighted mean and a weighted sum of squares.
- Poisson, with offsets (exposure times):
  $[Y_i \mid t_i, \lambda] \sim Poi(t_i\lambda)$, $\hat{\lambda}_{mle} = Y_+/\sum t_i$
- Bilateral exponential (Laplace) → the sample median
  $[X_1, \ldots, X_n \mid \mu]\ iid\ \frac{1}{2}e^{-|x_i - \mu|}, -\infty < x_i < \infty$
  $\hat{\mu}_{mle} =$ median$(X_1, \ldots, X_n)$
- Cauchy: Needs iteration
- A location/scale family with $-\log$(density) $\propto |x - \mu|^p$: p = 1 is Laplace, 2 is Gaussian, p = 1.5 produces good efficiency for the Gaussian, but more robustness than assuming Gaussian (i.e., using $\bar{X}_n$)

MLEs depend on having a likelihood: The MLE depends on the likelihood. It is desirable to have a likelihood, but sometimes they are challenging to produce, sometimes there is controversy on what is "the" or "a" likelihood, sometimes they don't exist:

- Let $[Y_i \mid p_i]$ $indep$ Bernoulli$(p_i), i = 1, \ldots, n$. Let $S_n = \sum Y_i$. Formulas for $E(S_n)$ and $V(S_n)$ are straightforward. The full likelihood for the vector $(Y_1, \ldots, Y_n)$ is also straightforward (product of Bernoulli RVs). But, the likelihood for S_n is complicated (write it out) and unless we restrict the p_i in some manner, we can't make useful progress. If we parameterize via $p_i = p(\theta, X_i)$ for covariates X and a low-dimension θ (e.g., logistic regression) we can produce the MLE and "combine evidence."

- Spatial correlations using the Conditional Auto-regressive Model (CAR). This produces the Gibbs distribution which only exists as the ergodic state of a Markov chain (the Gibbs sampler). Geman and Geman (Hopkins) did the fundamental work.

- Sometimes there isn't a likelihood: "Hey, I'm just looking at the data, don't bother me about likelihoods!"

- Sometimes data-analytic algorithms evolve into likelihood based or approximately so: Mann-Whitney/Wilcoxon, M-estimates, . . .
  - MW/W: $X \sim F, Y \sim F(Y - \theta), H_0 : \theta = 0$. Combine data, rank, sum the ranks for the Ys. Null distribution does not depend on F and can get test and CI. (More generally $H_0 : \theta = \theta_0$). CI and point estimate.
  - This procedure was not likelihood-based. Subsequently, it was shown that the test is locally most powerful Rank test (LMPRT) for location changes when F is the logistic distribution. It's Asymptotic Relative Efficiency if F is Gaussian is 95% and the ARE is never worse than 86% for unimodal, translation families.
  - Similarly, if F is assumed to be Gaussian, the LMPRT converts the ranks to Gaussian scores $\{$score $= \phi^{-1}(R/(n+1)\}$ and then computes the sum, etc. ARE if F is Gaussian is 100%
  - Can produce the optimal scores for any differentiable, location family F.

- Though likelihood ratio tests are very effective and attractive, sometimes the alternative is hard to specify:
  - $H_0 : X_1, \ldots, X_n \; iid$ (a general statement) H_A : not so
  - Come up with a rejection region that is unlikely under H_0. For example, if the Xs are binary, compute the run lengths.
  - Better, if possible, to embed in a family with a specified alternative.

Procedure-generation and evaluation
The MLE and associated inferences are procedures, algorithms, based on distributional assumptions. We can evaluate them under those assumptions and also under other assumptions to see how it performs. For example, the sample mean is the MLE for the Gaussian likelihood, but we can evaluate it's properties for other truths.

- Ditto for the median, trimmed mean, M-estimates, Lasso, ...
- Ditto for Bayesian Procedures: we can evaluate their Bayesian and frequentist properties
- Ditto for stepwise regression, . . .

Summary

- Generate procedures and study their properties under those assumptions. Then, mount aggressive evaluations under other distributional scenarios and relevant comparisons with reasonable competitors.
- The FDA insists on these evaluations; *JASA/ACS, Biometrics, Biostatistics, . . .* insist on them
- You should insist on them!

Generalized Likelihood ratio test

Assume you are interested in θ, but there are nuisance parameters(η). For example, you are interested in the Gaussian mean (μ), but don't know the variance (σ^2). What to do? Assume complete specification depends on (θ, η), dim$(\theta) = p$, dim$(\eta) = q < \infty$. For the focused $H_0\colon \theta = \theta_0$ (η unrestricted) and general alternative, $H_1 : \theta \neq \theta_0$ (η unrestricted), proceed as follows.

$$\begin{aligned} L(\theta_0) = GLR &= \frac{\sup_\eta L(\theta_0, \eta)}{\sup_{\theta,\eta} L(\theta, \eta)} = \frac{L(\theta_0, \hat{\eta}(\theta_0))}{L(\hat{\theta}, \hat{\eta})} \\ \hat{\eta}(\theta_0) &= \arg\sup_\eta L(\theta_0, \eta) \end{aligned}$$

Test: Reject if $\ell(\theta_0) < c$.

This is a test statistic, an algorithm, so either work out the exact distribution or use the boostrap to find the rejection region or use large sample theory.

Large Sample Theory: Let Θ be the parameter space for θ. If θ_0 is **in the interior** of Θ, then under H_0 and a few other conditions:

$$\lim_{\text{information}\rightarrow\infty} -2\log(GLR) \rightarrow \chi^2_p(0),\ p = \dim(\theta).$$

The proof of this depends on large sample theory for the MLE; see below.

Note: The GLRT justifies creating confidenced regions by including all θs such that $\ell(\theta) = \log(L(\theta)) \geq c$. The large sample theory shows that c can be selected using the chi-square distribution.

Gaussian example: $\mu = 0$ vs $\mu \neq 0$, σ^2 unknown. Let,

$$\hat{\sigma}^2(\mu) = \frac{1}{n}\sum_i (X_i - \mu)^2$$

Substitute $\mu = 0, \hat{\sigma}^2(0)$ in the numerator of the GLR and $\hat{\mu} = \bar{X}_n, \hat{\sigma}^2(\bar{X}_n)$ in the denominator. Note that for each the $(2\pi)^{-n/2}$ cancel and also the exponential part for each is: $e^{-n/2}$ (and cancel). You are left with rejecting if,

$$\left(\frac{\sum_i X_i^2}{\sum_i (X_i - \bar{X}_n)^2}\right)^{n/2} = \left(1 + \frac{\bar{X}_n^2}{\frac{1}{n}\sum(X_i - \bar{X}_n)^2}\right)^{n/2} > \text{ const}$$

or

$$n \log \left(1 + \frac{\bar{X}_n^2}{\frac{1}{n}\sum(X_i - \bar{X}_n)^2} \right) > c^* = 2\text{log(c)} \tag{3}$$

This is an Analysis of Variance (ANOVA) comparing sums of squares. Interestingly, we compare sums of squares to investigate a mean.

Distribution under $H_0 : \mu = 0$: In this case, $\bar{X}_n^2 \to 0$ at order $1/n$ (it is $O_p(1/n)$). Expanding the log we get:

$$n \log \left(1 + \frac{\bar{X}_n^2}{\frac{1}{n}\sum(X_i - \bar{X}_n)^2} \right) \approx \left(\frac{n\bar{X}_n^2}{\frac{1}{n}\sum(X_i - \bar{X}_n)^2} \right) = \left(\frac{n}{n-1} \right) \mathcal{F}_{1,n-1} \to \chi_1^2(0)$$

The $\mathcal{F}$ part is only strictly true for the Gaussian case, but the $\chi_1^2(0)$ holds for a broad range of one-dimensional θ_0.

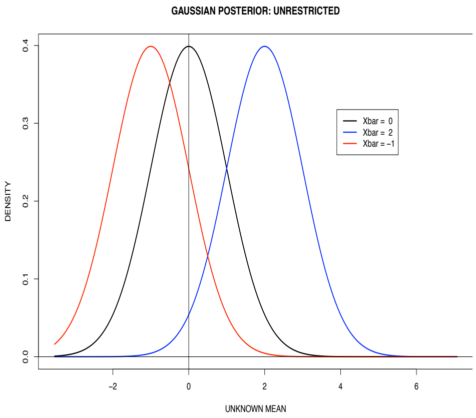
Distribution for a general $\mu \neq 0$: It's easy to see that the test statistic goes to infinity at rate $O_p(n)$ because the argument of the log in (**??**) converges to $(1 + \mu^2/\sigma^2)$.

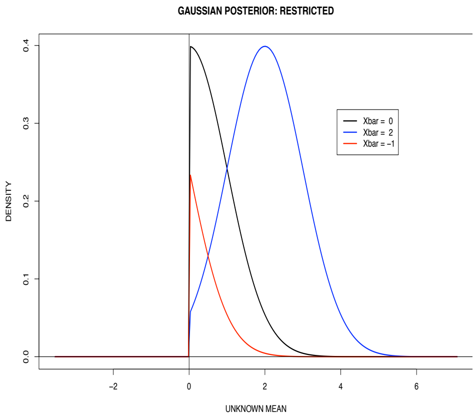
Local asymptotics: The middle ground case is when we set $\mu_n = \mu_0/\sqrt{n}$. Then, similar to when $\mu_n \equiv 0$,

$$n \log \left(1 + \frac{\bar{X}_n^2}{\frac{1}{n}\sum_{(} X_i - \bar{X}_n)^2} \right) \approx \left(\frac{n\bar{X}_n^2}{\frac{1}{n}\sum_{(} X_i - \bar{X}_n)^2} \right) \to \chi_1^2(\mu_0^2/2),$$

a non-central chi-square.

If H_0 is on the boundary: If H_0: $\mu = 0,\ H_1$: $\mu > 0$ [parameter space is $[0, \infty)$] under H_0, equation (**??**) has $(\bar{X}_n^+)^2$ in place of $\bar{X}_n^2$ and the limiting distribution has probability $\frac{1}{2}$ at 0, and $\frac{1}{2}$ on a central chi-square, i.e.: $\frac{1}{2}$dirac$(0) + \frac{1}{2}\chi_1^2(0)$. More generally, the asymptotic distribution will depend on p and on the shape of the boundary. (Ciprian knows a lot about the asymptotic theory for "on the boundary" models).





Large sample theory for the MLE

Here's a general representation and then the proof for the iid case. The result holds more generally.

Let Θ be the parameter space and $\boldsymbol{\theta}$ a p dimensional vector parameter. Assume, $[Y_i \mid \theta] \sim f(y_i \mid \theta)$, where f is twice differentiable in θ in the interior of Θ. Assume that θ_0, the true θ, is in the interior of Θ. Let,

$$\begin{aligned} \text{Score:} \qquad \mathbf{S}_n(\mathbf{Y};\boldsymbol{\theta}) &= \frac{\partial}{\partial \boldsymbol{\theta}} \log(f(\mathbf{Y} \mid \boldsymbol{\theta})) \\ \text{Observed Information:} \qquad \mathbf{B}_n(\mathbf{Y};\boldsymbol{\theta}) &= -\frac{\partial}{\partial \boldsymbol{\theta}} \mathbf{S}_n(\mathbf{Y} \mid \boldsymbol{\theta}) = -\frac{\partial^2}{\partial \boldsymbol{\theta}^2} \log(f(\mathbf{Y} \mid \boldsymbol{\theta})) \end{aligned}$$

S is a $p \times 1$ vector and $\mathbf{B}$ is a $p \times p$ matrix.
Let $\hat{\boldsymbol{\theta}}(\mathbf{Y})$ be the MLE and write,

$$0 = \mathbf{S}_n(\mathbf{Y};\hat{\boldsymbol{\theta}}(\mathbf{Y})) \;=\; \mathbf{S}_n(\mathbf{Y};\boldsymbol{\theta}_0) - \mathbf{B}_n(\mathbf{Y};\boldsymbol{\theta})(\hat{\boldsymbol{\theta}}(\mathbf{Y}) - \boldsymbol{\theta}_0) + \text{remainder}$$

Solving this, we obtain,

$$\{\hat{\boldsymbol{\theta}}(\mathbf{Y}) - \boldsymbol{\theta}_0\} \;\doteq\; \mathbf{B}_n(\mathbf{Y};\boldsymbol{\theta}_0)^{-1}\mathbf{S}_n(\mathbf{Y};\boldsymbol{\theta}_0)$$

or,

$$\sqrt{n}\{\hat{\boldsymbol{\theta}}(\mathbf{Y}) - \boldsymbol{\theta}_0\} \;\doteq\; \left(\frac{1}{n}\mathbf{B}_n(\mathbf{Y};\boldsymbol{\theta}_0)\right)^{-1} \frac{\mathbf{S}_n(\mathbf{Y};\boldsymbol{\theta}_0)}{\sqrt{n}} \qquad (4)$$

We have that,

$$\begin{aligned} E(\mathbf{S}_n(\mathbf{Y};\boldsymbol{\theta}_0) \mid \boldsymbol{\theta}_0) &= 0 \\ V(\mathbf{S}_n(\mathbf{Y};\boldsymbol{\theta}_0) \mid \boldsymbol{\theta}_0) &= \mathbf{B}_n(\boldsymbol{\theta}_0) = E(\mathbf{B}_n(\mathbf{Y};\boldsymbol{\theta}_0) \mid \boldsymbol{\theta}_0) \end{aligned}$$

Can continue from here to show, subject to $\frac{\mathbf{B}_n(\mathbf{Y};\boldsymbol{\theta})}{\mathbf{B}_n(\boldsymbol{\theta})} \to 1$, $\mathbf{B}_n(\boldsymbol{\theta}) \to \infty$ and conditions like Lindeberg, that

$$\mathbf{B}_n(\boldsymbol{\theta}_0)^{\frac{1}{2}}\{\hat{\boldsymbol{\theta}}(\mathbf{Y}) - \boldsymbol{\theta}_0\} \to N(0, \mathbf{I}).$$

For data analysis, we use the "observed information" $\mathbf{B}_n(\mathbf{Y};\boldsymbol{\theta})$,

$$\mathbf{B}_n(\mathbf{Y}, \boldsymbol{\theta}_0)^{\frac{1}{2}}\{\hat{\boldsymbol{\theta}}(\mathbf{Y}) - \boldsymbol{\theta}_0\} \approx N(0, \mathbf{I}).$$

The iid case: $\mathbf{S}_n$ and $\mathbf{B}_n$ are each the sum of iid random variables,

$$\begin{aligned} \mathbf{S}_n(\mathbf{Y};\boldsymbol{\theta}_0) &= \sum_i \mathbf{S}(Y_i;\boldsymbol{\theta}_0) \\ \mathbf{B}_n(\mathbf{Y};\boldsymbol{\theta}_0) &= \sum_i \mathbf{B}(Y_i;\boldsymbol{\theta}_0) \\ E(\mathbf{S}(Y_i;\boldsymbol{\theta}_0) \mid \boldsymbol{\theta}_0) &= 0 \\ V(\mathbf{S}(Y_i;\boldsymbol{\theta}_0) \mid \boldsymbol{\theta}_0) &= \mathbf{B}(\boldsymbol{\theta}_0) = \mathbf{B}_1(\boldsymbol{\theta}_0) \end{aligned}$$

To verify these, reverse the order of expectation and differentiation of f. Substituting these into (**??**) gives,

$$\sqrt{n}\{\hat{\boldsymbol{\theta}}(\mathbf{Y}) - \boldsymbol{\theta}_0\} \doteq \left(\frac{1}{n}\sum_i \mathbf{B}(Y_i;\boldsymbol{\theta}_0)\right)^{-1} \frac{\sum_i \mathbf{S}(Y_i;\boldsymbol{\theta}_0)}{\sqrt{n}}$$

But, the first term on the RHS converges to $\mathbf{B}(\boldsymbol{\theta}_0)$ and a central limit theorem applies to the second term, producing for large n,

$$\sqrt{n}\{\hat{\boldsymbol{\theta}}(\mathbf{Y}) - \boldsymbol{\theta}_0\} \sim N_p\left(\mathbf{0}, \mathbf{B}(\boldsymbol{\theta}_0)^{-1}\right) \approx N_p\left(\mathbf{0}, \mathbf{B}(\mathbf{Y};\hat{\boldsymbol{\theta}})^{-1}\right).$$

Information

Observed Information:

$$I_n(\mathbf{Y};\hat{\boldsymbol{\theta}}) = \mathbf{B}_n(\mathbf{Y};\hat{\boldsymbol{\theta}}) \approx \mathbf{S}_n(\mathbf{Y};\hat{\boldsymbol{\theta}})\mathbf{S}_n(\mathbf{Y};\hat{\boldsymbol{\theta}})'$$

This is the curvature of the log-likelihood at the mle.

Fisher Information: Under the regularity conditions,

$$I_n(\boldsymbol{\theta}) \;=\; \mathbf{B}_n(\boldsymbol{\theta}) = E[\mathbf{S}_n(\mathbf{Y};\boldsymbol{\theta})\mathbf{S}_n(\mathbf{Y};\boldsymbol{\theta})']$$

This is the expected curvature of the log-likelihood at the mle.

In regression, we always use the observed information: $\frac{1}{\sigma^2}(\mathbf{X}'\mathbf{X})$ in data analysis, but we can use the Fisher Information (the expectation of this) if we are designing a study and don't get to pick the Xs.

Notes

- This result justifies the asymptotic chi-square distribution for the GLRT under H_0 (see Cox & Hinkley).
- A large variance for the score, produces a small variance for the MLE (the score is informative).
- In general, it is not sufficient for $n \rightarrow \infty$. For consistency you must have $\mathbf{B}_n \rightarrow \infty$. In addition, the CLT needs the Lindeberg or Lyapunov condition. For example, take simple regression with a centered regressor. Then then,

  $$\mathbf{B}_n = \frac{1}{\hat{\sigma}^2}\begin{pmatrix} n & 0 \\ 0 & \sum_i X_i^2 \end{pmatrix}$$

  We must have $\sum_i X_i^2 \rightarrow \infty$ to get the CLT result. For example, if $X_{2k} = k^{-1} = -X_{2k-1}$, the sum of squares will be finite and there is no CLT (in fact the slope estimate does not converge to β).
- These results show how unknown nuisance parameters inflate the variance of the MLE. To wit, we have the general matrix result:

  $$\begin{aligned} \mathbf{A} \;&=\; \begin{pmatrix} \mathbf{A}_{11} & \mathbf{A}_{12} \\ \mathbf{A}_{21} & \mathbf{A}_{22} \end{pmatrix} \\ (\mathbf{A}^{-1})_{11} \;&=\; \left(\mathbf{A}_{11} - \mathbf{A}_{12}\mathbf{A}_{22}^{-1}\mathbf{A}_{21}\right)^{-1} \end{aligned}$$

  So, with A the information matrix for $\boldsymbol{\theta} = (\boldsymbol{\theta}_1 : \boldsymbol{\theta}_2)'$, the covariance matrix for the mle of $\boldsymbol{\theta}_1$ is $(\mathbf{A}^{-1})_{11} \geq \mathbf{A}_{11}^{-1}$.

- Consider transforms to help convergence, but for the vector case this can get difficult.
- As the trinomial example (below) shows, it is better to use the observed information in computing tests and CIs. See Efron and Hinkley on Observed and Expected Information.
  - ○ Numerical maximization routines can find the observed information.
- Importantly, it is better to use the outer product of $\mathbf{S}$ representation of the observed information because it is more robust. It relies less on the details of the assumed distribution. Indeed, it is the "robust variance" in GenMod, etc. See Louis (1982) *JRSSB* on speeding up the convergence of the EM.

- *GLRT for Regression:* as for the mean only case, depends on the ratio of two variance estimates:

$$\frac{\hat{\sigma}^2_{int}}{\hat{\sigma}^2_{int,slope}} = 1 + \frac{SSregr}{SSE}$$

which produces the F-test, with p the number of regressors (not including the intercept):

$$\begin{aligned} F &= \frac{\frac{SSRegr}{p}}{\frac{SSE}{n-p}} \\ \mathcal{F}_{\nu_1,\nu_2}(0) &\sim \frac{\chi^2_{\nu_1}(0)/\nu_1}{\chi^2_{\nu_2}(0)/v_2} \quad \text{independent rvs} \end{aligned}$$

Under H_0, as $\nu_2 \rightarrow \infty$, $\mathcal{F}_{\nu_1,\nu_2} \rightarrow \chi^2_{\nu_1}(0)/\nu_1$,

- Note that in the simple case, under $H_0 : \beta = 0$,

$$\begin{aligned} E[SSregr \mid \mathbf{X}] &= E[\hat{\beta}^2 \mid \mathbf{X}] \sum (X_i - \bar{X})^2 \\ &= V[\hat{\beta} \mid \mathbf{X}] \sum (X_i - \bar{X})^2 = \sigma^2 \end{aligned}$$

and that SSregr is asymptotically $\sigma^2\chi^2_1(0)$

- Can use the Wald test (the t-test) or the LR test (the F-test). For simple linear regression, the 2-sided Wald and the GLRT test are equivalent.

CI for the line (when $\bar{X} = 0$)

Variance for a given X_i expands like X_i^2.

$$\begin{aligned} \hat{V}(\hat{Y}_i \mid X_i) &= V(\hat{\alpha} + \hat{\beta} X_i) = \hat{\sigma}^2 \left(\frac{1}{n} + \frac{X_i^2}{\sum_\ell X_\ell^2} \right) \\ \sum_i \hat{V}(\hat{Y}_i \mid X_i) &= \hat{\sigma}^2 (1+1) = 2\hat{\sigma}^2 \\ & \qquad\qquad\qquad \text{2 df} \end{aligned}$$

Prediction interval for a Y

The variance for predicting a new Y at an X is $\hat{\sigma}^2 + \hat{V}(\hat{Y}_i \mid X_i)$.

(FYI: Not to be covered in class)

Delta theorem for large-sample distn of order statistics:

Probability Integral Transform: F continuous with density f.

$U \sim F(X) \sim U(0,1), \; X \sim F^{-1}(U)$

For $0 < p < 1$ (note the strict inequalities)

$\lim_{n\to\infty} \sqrt{n}(U_{([pn])} - p) \sim N(0, p(1-p))$

Can prove this by noting that $U_{(k)} \sim$ Beta$(k, n-k+1)$ or using the Bernoulli variables representation for the order statistics.

Now, the delta theorem shows: $\lim_{n\to\infty} \sqrt{n}(X_{[(pn])} - \xi_p) \sim N\left(0, \frac{p(1-p)}{f^2(\xi_p)}\right)$ with $\xi_p = F^{-1}(p)$. Need derivative of $F^{-1}(u)$.

Generalized Least squares

Ingo's notes, ch09 Generalized Least Squares

Ingo's notes, ch14 Residuals and Influence (review this section)

$\mathbf{Y}_{n\times 1}$ = stacked dependent variable
$\mathbf{X}_{n\times p}$ = stacked full rank
p columns, $\boldsymbol{\beta}_{p\times 1}$.

$$\begin{aligned}\mathbf{Y} &= \mathbf{X}\boldsymbol{\beta}+\boldsymbol{\epsilon}\\ E(\mathbf{Y}\mid\mathbf{X}) &= \mathbf{X}\boldsymbol{\beta}\\ E(\boldsymbol{\epsilon}) &= \mathbf{0}\\ V(\boldsymbol{\epsilon}) &= \boldsymbol{\Sigma}_{n\times n}\end{aligned}$$

$$\begin{aligned}\hat{\boldsymbol{\beta}}_{\boldsymbol{\Sigma}} &= \left(\mathbf{X}'\boldsymbol{\Sigma}^{-1}\mathbf{X}\right)^{-1}\mathbf{X}'\boldsymbol{\Sigma}^{-1}\mathbf{Y}\\ E\left(\hat{\boldsymbol{\beta}}_{\boldsymbol{\Sigma}}\right) &= \boldsymbol{\beta}\\ V\left(\hat{\boldsymbol{\beta}}_{\boldsymbol{\Sigma}}\right) &= \left(\mathbf{X}'\boldsymbol{\Sigma}^{-1}\mathbf{X}\right)^{-1}\end{aligned}$$

Proof, representation, transformation and insight: Let,

$$\mathbf{Y}_* = \boldsymbol{\Sigma}^{-\frac{1}{2}}\mathbf{Y}$$

and so

$$\begin{aligned}\mathbf{X}_* &= \boldsymbol{\Sigma}^{-\frac{1}{2}}\mathbf{X}\\ \boldsymbol{\epsilon}_* &= \boldsymbol{\Sigma}^{-\frac{1}{2}}\boldsymbol{\epsilon}\\ \hat{\boldsymbol{\beta}} &= \left(\mathbf{X}_*'\mathbf{X}_*\right)^{-1}\mathbf{X}_*'\mathbf{Y}_*\\ \mathbf{H}_* &= \mathbf{X}_*\left(\mathbf{X}_*'\mathbf{X}_*\right)^{-1}\mathbf{X}_*'\end{aligned}$$

Use these with the standard LM and you have GLS.

Notes:

- This transform pre-multiplies $\mathbf{X}$ (a l.c. of the rows) and preserves the parameterization. Previously, we post-multiplied $\mathbf{X}$ (a l.c. of the columns) which gave us a different parameterization but the same $\hat{\mathbf{Y}}$.
- Be careful using diagnostics; leverage, influence, residual plots, added variable plots, ... For example, in standard LSE the residuals, $\hat{\boldsymbol{\epsilon}} = (\mathbf{Y}-\hat{\mathbf{Y}})$, have mean 0 and are uncorrelated with the

$\hat{\mathbf{Y}}$ $\{(\mathbf{I}-\mathbf{H})'\mathbf{H}=\mathbf{0}\}$. In GLS all of these relations apply to the $*$ versions and these relations will generally not hold for the unstarred versions. For example, there is no requirement that the residuals in the original scale have mean 0 or are uncorrelated with the $\hat{\mathbf{Y}}$.

- You don't have to use $\boldsymbol{\Sigma}^{-\frac{1}{2}}$ for the transform. Any $\mathbf{A}$ such that $\mathbf{A}\boldsymbol{\Sigma}\mathbf{A}' \propto \mathbf{I}$ will work and it can be revealing to cater to the structure of the data. For example, if each unit has a pair of observations and the second follows the first in time sequence, if

$$\boldsymbol{\Sigma} = \sigma^2 \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}, \text{ then } \mathbf{A} = \begin{pmatrix} 1 & 0 \\ -\frac{\rho}{\sqrt{1-\rho^2}} & \frac{1}{\sqrt{1-\rho^2}} \end{pmatrix}$$

respects the longitudinal structure. This is the lower Cholesky decomposition.

$\boxed{\text{Sketch residuals}}$

- If you don't know $\boldsymbol{\Sigma}$, then you can do iteratively re-weighted least squares, by producing estimates with $\boldsymbol{\Sigma}^{(\nu)}$, re-estimating the covariance using the residuals to get $\boldsymbol{\Sigma}^{(\nu+1)}$, etc.

- In general, you need to parameterize $\boldsymbol{\Sigma}$, using $\boldsymbol{\Sigma}(\boldsymbol{\psi})$ and recursively estimate $\hat{\boldsymbol{\psi}}^{(\nu)}$. You need a program that can produce the MLE for $\boldsymbol{\psi}$.

- Since, $\hat{\boldsymbol{\beta}}_{\boldsymbol{\Sigma}}$ is unbiased for $\boldsymbol{\beta}$, and starting with the independence model and then doing 1 iteration is asymptotically, fully efficient.

- Sometimes, you can estimate an "unstructured" $\boldsymbol{\Sigma}$, for example when it is block diagonal (repeated measures on independent individuals with common follow-up assessments).

$$\boldsymbol{\Sigma}_{jj} \;=\; \tilde{\boldsymbol{\Sigma}},\ \boldsymbol{\Sigma}_{jj'} = \mathbf{0}$$

With no missing data, the MLE is,

$$\tilde{\boldsymbol{\Sigma}}^{(\nu+1)} = \frac{1}{J-p}\sum_{j=1}^{J}\left(\mathbf{Y}_j - \mathbf{X}_j\boldsymbol{\beta}^{(\nu)}\right)\left(\mathbf{Y}_j - \mathbf{X}_j\boldsymbol{\beta}^{(\nu)}\right)'$$

- Note that when some aspects of $\mathbf{\Sigma}$ are unknown (other than a leading multiplier), the regression is "non-linear" in that some unknown parameters enter non-linearly.

Efficiency of GLS relative to OLS

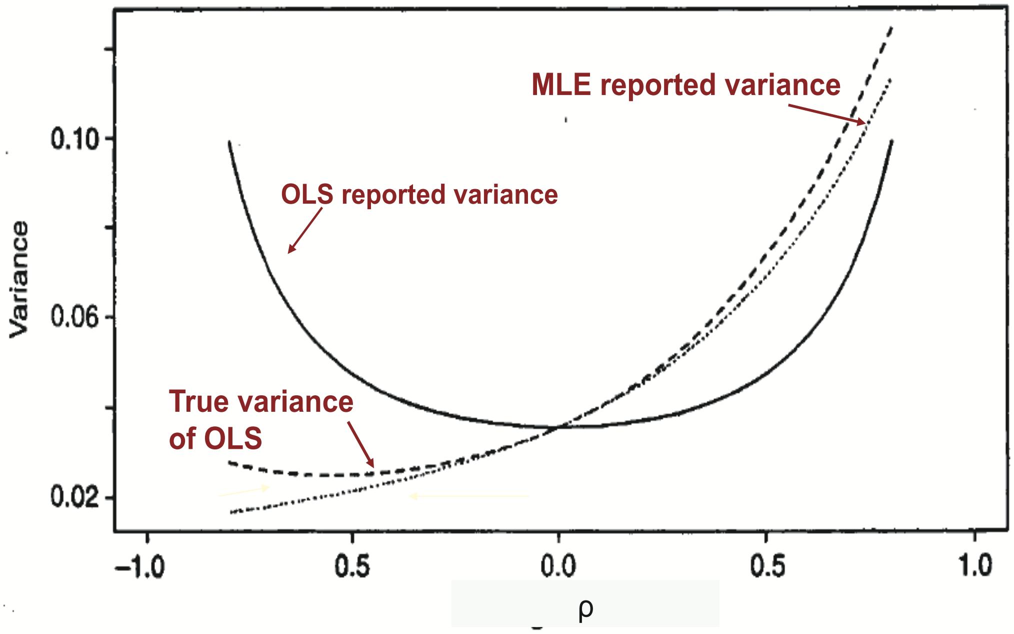
Using the correct $\mathbf{\Sigma}$ is efficient and protects against bias from certain missing data processes (MAR, see later). However, it is somewhat "fragile" (also see later) and so there is some attraction to estimating via OLS and then using a "robust variance" to get the correct SEs for the OLS estimate. We will consider robust variance estimation shortly, but here's a look at the efficiency gain.

- Generate 5 observations per unit (individual)
- Response is a linear function of time

$$\begin{aligned} Y_{it} &= \alpha + \beta t + e_{it} \\ \text{cov}(e_{i,t+s}, e_{it}) &= \sigma^2 \rho^s \quad \text{AR(1)} \end{aligned}$$

- Estimate β by OLS and GLS (take account of ρ)
- Compare $V_{ols}(\hat{\beta}_{ols})$, $V_{ar1}(\hat{\beta}_{ols})$, $V_{ar1}(\hat{\beta}_{ar1})$

Variance of OLS & MLE Estimates of β versus ρ, the lag-1 correlation



$\boxed{\text{Ingo's notes, ch10}}$ Gauss-Markov

$\boxed{\text{Ingo's notes, ch11}}$ Hypothesis Testing

Wishart Distribution

$\mathbf{Y}_i, i = 1, \ldots, n > p$ are iid $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$
$\mathbf{V} = \sum_i \mathbf{Y}_i \mathbf{Y}_i'$
$\mathbf{V}$ is a p-dimensional Wishart matrix with n degrees of freedom, covariance matrix $\boldsymbol{\Sigma}$ and non-centrality $\lambda = \frac{1}{2}\mathbf{u}'\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}$. That is $\mathbf{V} \sim \mathbf{W}_p(n, \boldsymbol{\Sigma}, \lambda)$. The density exists, etc.

If $n > p + 1$, the $\hat{\mathbf{V}} = \sum_i (\mathbf{Y}_i - \bar{\mathbf{Y}})(\mathbf{Y} - \bar{\mathbf{Y}})_i'$ is $\mathbf{W}_p(n-1, \boldsymbol{\Sigma}, 0)$.

There are lots of theorems associated with the Wishart, see Rao, *Linear Models* for all of them. Here are a few results.

Let $\mathbf{V} \sim \mathbf{W}_p(n, \boldsymbol{\Sigma}, 0)$. Then,

$$\begin{aligned}
\frac{\sigma^{pp}}{\mathbf{V}^{pp}} &\sim \chi^2_{n-(p-1)}(0) \text{ and is independent of } \mathbf{V}_{ij}, (i,j) \leq (p-1) \\
\frac{L'\boldsymbol{\Sigma}^{-1}L}{L'\mathbf{V}^{-1}L} &\sim \chi^2_{n-(p-1)}(0) \\
\frac{\mid \mathbf{V} \mid}{\mid \boldsymbol{\Sigma} \mid} &\sim \prod_{j=1}^{p} \chi^2_{n-j+1}(0), \text{ independent chi-square RVs}
\end{aligned}$$

$\boxed{\textbf{Comparing nested or non-nested models}}$

R^2_{adj}, PRESS . . . : Defined earlier.

AIC, BIC and DIC

In all cases, models with small values are candidates.

AIC: Akaike Information Criterion

AIC $=-2\times$(maximum log-likelihood) $+ 2df$:

Bayesian Information Criterion
BIC $= -2\times$(maximum log-likelihood) $+ \log(n)df$:

Note that BIC is more stringent than AIC and that for models with the same df, both compare log likelihoods.

What is df and what is n?: In a complicated, hierarchical models, we don't know n and we don't know df (df is data-dependent). The Deviance Information Criterion (DIC) partially solves this problem. See, Spiegelhalter DJ, Best NG, Carlin BP and van der Linde A (2002). Bayesian measures of model complexity and fit (with discussion). *J. Roy. Statist. Soc. B.*, **64:** 583-640.

Departures from assumptions

Incorrect covariance matrix $\mathbf{W}$ = the "working covariance"

$\mathbf{T}$ = the "true covariance"

$$\begin{aligned}
\hat{\boldsymbol\beta}_{\mathbf{W}} &= \left(\mathbf{X}'\mathbf{W}^{-1}\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{W}^{-1}\mathbf{Y} \\
E(\hat{\boldsymbol\beta}_{\mathbf{W}}) &= \boldsymbol\beta \\
V(\hat{\boldsymbol\beta}_{\mathbf{W}}) &= \left(\mathbf{X}'\mathbf{W}^{-1}\mathbf{X}\right)^{-1}\left(\mathbf{X}'\mathbf{W}^{-1}\mathbf{T}\mathbf{W}^{-1}\mathbf{X}\right)\left(\mathbf{X}'\mathbf{W}^{-\mathbf{1}}\mathbf{X}\right)^{-1}
\end{aligned}$$

Let cov(assumed | true) = the true covariance with estimates generated from the assumed model.

$$\begin{aligned}
\text{cov}(\mathbf{T} \mid \mathbf{T}) &= \left(\mathbf{X}\mathbf{T}^{-1}\mathbf{X}\right)^{-1} \\
\text{cov}(\mathbf{W} \mid \mathbf{W}) &= \left(\mathbf{X}\mathbf{W}^{-1}\mathbf{X}\right)^{-1}\text{wrong!} \\
\text{cov}(\mathbf{W} \mid \mathbf{T}) &= \left(\mathbf{X}'\mathbf{W}^{-1}\mathbf{X}\right)^{-1}\left(\mathbf{X}'\mathbf{W}^{-1}\mathbf{T}\mathbf{W}^{-1}\mathbf{X}\right)\left(\mathbf{X}'\mathbf{W}^{-\mathbf{1}}\mathbf{X}\right)^{-1}
\end{aligned}$$

For W = I, OLS

$$\text{cov}(\mathbf{I} \mid \mathbf{T}) = \left(\mathbf{X}'\mathbf{X}\right)^{-1}\left(\mathbf{X}'\mathbf{T}\mathbf{X}\right)\left(\mathbf{X}'\mathbf{X}\right)^{-1}$$

In general,

$$\mathsf{cov}(\mathbf{W} \mid \mathbf{T}) \;\geq\; \mathsf{cov}(\mathbf{T} \mid \mathbf{T});$$

using the correct covariance is most efficient.

Example

$$\begin{aligned}
E(Y_i \mid X_i) &= \beta X_i \\
V(Y_i \mid X_i) &= \sigma^2 w_i^{-1} \\
\beta^* &= \frac{\sum w_i X_i Y_i}{\sum w_i X_i^2} \\
\hat{\beta} &= \frac{\sum X_i Y_i}{\sum X_i^2} \\
V(\beta^*) &= \frac{\sigma^2}{\sum w_i X_i^2} \\
V(\hat{\beta}) &= \sigma^2 \frac{\sum w_i^{-1} X_i^2}{\{\sum X_i^2\}^2} \geq V(\beta^*)
\end{aligned}$$

The sandwich, robust variance estimate:

$$\begin{aligned}
\widehat{\mathsf{cov}}(\mathbf{W} \mid \mathbf{T}) &= \left(\mathbf{X}'\mathbf{W}^{-1}\mathbf{X}\right)^{-1} \left(\mathbf{X}'\mathbf{W}^{-1}\hat{\mathbf{\Sigma}}\mathbf{W}^{-1}\mathbf{X}\right) \left(\mathbf{X}'\mathbf{W}^{-\mathbf{1}}\mathbf{X}\right)^{-1} \\
\hat{\mathbf{\Sigma}} &= \text{an empirical estimate}
\end{aligned}$$

This computation can be represented as coming from a bootstrap re-sampling.

If $\mathbf{Y}_{(nJ)\times p} = (\mathbf{Y}_1 : \ldots : \mathbf{Y}_n)'$ and $\mathbf{\Sigma}$ is block diagonal with blocks denoted $\mathbf{A}_{J\times J}$, and similarly for $\mathbf{X}$, then

$$\hat{\mathbf{A}} = \frac{1}{nJ - p} \sum_{i=1}^{n} (\mathbf{Y}_i - \mathbf{X}_i\hat{\boldsymbol{\beta}})(\mathbf{Y}_i - \mathbf{X}_i\hat{\boldsymbol{\beta}})'$$

- The foregoing cn be generalized for non-balanced situations and for structured $\mathbf{A}$ matrices (compound symmetry, Toeplitz, AR(1), ...).
- SAS, R, Stata and other programs compute this for Gaussian models and for Generalized Linear Models.

- When a convenient structure is not available for a sandwich estimate, a robust adjustment can be computed using the jackknife or the bootstrap. Discuss the MUST example
- It is good practice (very good practice) to polish up a covariance estimate with a robust adjustment.

Incorrect design matrix; correct covariance

Basic example: Standard confounding: Y depends on X_1 and X_2, but you use only X_1. Residuals contain (slope)×$E(X_2 \mid X_1)$.

$$\begin{aligned}[Y \mid X_1, X_2] &= \alpha_0 + \alpha_1 X_1 + \alpha_2 X_2 + \epsilon \\ [Y \mid X_1] &= \alpha_0 + \alpha_1 X_1 + \alpha_2 [X_2 \mid X_1] + \epsilon \\ E(Y \mid X_1) &= \alpha_0 + \alpha_1 X_1 + \alpha_2 E(X_2 \mid X_1)\end{aligned}$$

Note that $E(X_2 \mid X_1)$ does not have to be linear in X_1. For non-zero X_1 we can write:

$$E(X_2 \mid X_1) = \left[\frac{E(X_2 \mid X_1)}{X_1}\right] X_1 = \gamma(X_1) X_1$$

So,

$$E(Y \mid X_1) = \alpha_0 + \{\alpha_1 + \alpha_2 \gamma(X_1)\} X_1$$

Ingo's notes, ch13 Departures from Assumptions

Longitudinal misspecification: Based on:

- Pan W, Louis TA, Connett JE (2000). A Note On Marginal Linear Regression With Correlated Response Data. *The American Statistician,* **54:** 191-195.

See also,

- Emond, Ritz and Oakes (1997). Bias in GEE estimates from misspecified models for longitudinal data. *Comm. Statist.*, **26:** 15-32.

- Pepe & Anderson (1994). A cautionary note on inference for marginal regression models with longitudinal data and general correlated response data. *Comm. Statist. Simuln Computn,* **23:** 939-951.
- Lai & Small (2007). Marginal regression analysis of longitudinal data with time-dependent covariates: a generalized method-of-moments approach, *JRSS(B),* **69:** 79-99.

Accounting for correlation in General Linear Models (or, in fact in Generalized Linear Models) potentially confers the benefits of increased efficiency relative to "working independence" and protection from some missing data processes (some violations of missing completely at random).

But, as the foregoing articles show "working independence" is more robust to design mis-specification. For the non-diagonal matrix, large-sample consistency is more fragile and even the "sandwich estimate" of the estimated parameter covariance is more fragile

Consider the linear model: $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}; \quad \boldsymbol{\epsilon} \sim N(\mathbf{0}, \boldsymbol{\Sigma})$ and let $\mathbf{W}$ be any full-rank, working covariance matrix. Then, the GLS estimate of $\boldsymbol{\beta}$ is:
$\hat{\boldsymbol{\beta}}_{\mathbf{W}} = \left(\mathbf{X}'\mathbf{W}^{-1}\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{W}^{-1}\mathbf{Y}$

With expectation: $E[\hat{\boldsymbol{\beta}}_{\mathbf{W}} \mid \mathbf{X}] = \left(\mathbf{X}'\mathbf{W}^{-1}\mathbf{X}\right)^{-1}\left(\mathbf{X}'\mathbf{W}^{-1}\mathbf{X}\right)\boldsymbol{\beta} = \boldsymbol{\beta}$

Notes:

- $\hat{\boldsymbol{\beta}}_{\mathbf{W}}$ is unbiased (and C.A.N.), for all, non-singular $\mathbf{W}$ and $\boldsymbol{\Sigma}$
- A robust (e.g., sandwich) estimate of cov($\hat{\beta}$) produces valid inferences
- Use of a $\mathbf{W}$ that is close to the true covariance enhances efficiency and missing-data protection relative to working independence
- But, these desirable properties depend on the (asymptotic) unbiasedness of the estimating equation $[S(\boldsymbol{\beta} \mid \mathbf{X}, \mathbf{Y}) = \mathbf{0}]$ and if $\mathbf{W}$ is not diagonal (is not working independence) this unbiasedness depends on validity of:

$[\mathbf{Y} \mid \mathbf{X}] = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$, with conditioning on $\boxed{\text{all of } \mathbf{X}}$, not just row-specific conditioning.

Example

For $i = 1, ..., n$ and $t = 1, ..., m$:

$$\begin{aligned} Y_{it} &= \text{the response for participant "i" at time "t"} \\ X_{it} &= \text{the corresponding covariate vector} \\ \mathbf{X}_i &= \text{the full covariate vector for participant "i"} \end{aligned}$$

The *fully conditioned regression* is: $E(Y_{it}|X_{i1}, X_{i2}, ..., X_{im})$

The *marginal regression* (conditioning only on covariates at time t), is:
$E[Y_{it} \mid X_{it}] = X_{it}\beta$

For non-diagonal $\mathbf{W}$, if $E(Y_{it}|X_{it}) \neq E(Y_{it}|X_{i1}, X_{i2}, ..., X_{im})$, then, $\hat{\boldsymbol{\beta}}_{\mathbf{W}}$ estimated assuming that $E(Y_{it}|X_{it}) = X_{it}\beta$ can be biased.

Consider the AR(1) model on the uncentered Ys,

$$\begin{aligned} Y_{it}|(Y_{i,t-1}, X_{it}) &= Y_{i,t-1} + X_{it}\beta + \epsilon_{it} \\ X_{it} \;\; iid & \;\; N(0, \sigma^2) \\ e_{it} \;\; iid & \;\; N(0, \tau^2) \end{aligned}$$

- X_{it} and e_{it} are independent and independent of $Y_{i,t-1}$
- Y$_{i0} \equiv 0$

$\boxed{\textbf{Note}}$ This is not a very good specification. For example, it can't be produced by the MVN.

Then,

$$E(Y_{it}|\, X_{i1}, ..., X_{im}) = E(Y_{it}|\, X_{i1}, ..., X_{it}) = \left(\sum_{j=1}^{t} X_{ij} \right) \beta$$

However, the t-specific mean is:

$$E(Y_{it}|X_{it}) = X_{it}\beta$$

So it might be ok to fit the model,

$$[Y_i|X_i] = X_{it}\beta + \epsilon_i$$

With $\mathbf{W} = \mathbf{I}$, the OLS estimate is:

$$\hat{\beta}_I = \left(\sum_{i=1}^{n} X_i'X_i\right)^{-1} \sum_{i=1}^{n} X_i'Y_i = \frac{\sum_{i=1}^{n}\sum_{t=1}^{m} X_{it}Y_{it}}{\sum_{i=1}^{n}\sum_{t=1}^{m} X_{it}^2}$$

And,

$$\begin{aligned} E(\hat{\beta}_I|X) &= \left[1 + \frac{\sum_{i=1}^{n}\sum_{t=2}^{m} X_{it}\sum_{j=1}^{t-1} X_{ij})}{\sum_{i=1}^{n}\sum_{t=1}^{m} X_{it}^2}\right]\beta \\ &= \left[1 + \frac{\sum_{t=2}^{m}\sum_{j=1}^{t-1} \text{cov}(X_{it}, X_{ij})}{m\sigma^2}\right]\beta \\ &\approx \beta \end{aligned}$$

$\hat{\beta}_I$ is conditionally biased, but asymptotically unbiased.

The GLS Estimate: Assume the same $\mathbf{W}$ for all i and let $\mathbf{C} = \mathbf{W}^{-1} = (c_{ij})$. Then, the GLS estimate is:

$$\hat{\beta}_W = (\sum_{i=1}^{n} X_i'\mathbf{C}X_i)^{-1} \sum_{i=1}^{n} X_i'\mathbf{C}Y_i = \frac{\sum_{i=1}^{n}\sum_{j=1}^{m}(\sum_{t=1}^{m} c_{tj}X_{it})Y_{ij}}{\sum_{i=1}^{n} X_i'\mathbf{C}X_i}$$

So,

$$E(\hat{\beta}_W|X) = \left[1 + \frac{\sum_{i=1}^{n}\sum_{j=2}^{m}[(\sum_{t=1}^{m} c_{tj}X_{it})\sum_{k=1}^{j-1} X_{ik}]}{\sum_{i=1}^{n} X_i'\mathbf{C}X_i}\right]\beta$$

In general, $\hat{\beta}_W$ is conditionally biased and its unconditional bias can be substantial, even for large n.

Asymptotics

For large n, $\frac{1}{n}\sum_{i=1}^{n} X_i'\mathbf{C}X_i \approx E(X_i\mathbf{C}X_i)$

$X_i\mathbf{C}X_i$ is distributed as a mixture of chi-squares, so $E(X_i\mathbf{C}X_i) = \sigma^2\sum_{j=1}^{m}\lambda_j$, where the λ_j are eigenvalues of $\mathbf{C}$. So,

$$E(\hat{\beta}_W) = E[E(\hat{\beta}_W|X)] \;\approx\; \left[1 + \frac{\sum_{t=1}^{m-1}\sum_{j=t+1}^{m} c_{tj}}{\sum_{j=1}^{m}\lambda_j}\right]\beta$$

If W is diagonal, so is $\mathbf{C}$, and: $\sum_{t=1}^{m-1}\sum_{j=t+1}^{m} c_{tj} = 0$

So, $\hat{\beta}_{DIAG}$ is unconditionally unbiased.

When W = cov($\boldsymbol{Y_i}$): In the AR1 example, for $0 < j \leq k \leq m$:

$$V(Y_{ij} \mid X_{ij}) = j\tau^2, \quad Cov(Y_{ij}, Y_{ik}|X_i) = j\tau^2$$

With $W = Cov(Y_i|X_i) = \mathbf{C}^{-1}$, $\sum_{t=1}^{m-1}\sum_{j=t+1}^{m} c_{tj} = -m+1$

And, $\sum_{j=1}^{m}\lambda_j = (2m-1)/\tau^2$

$\boxed{\textbf{The large-sample bias of } \hat{\beta}_W \textbf{ is: } -\left[\frac{m-1}{2m-1}\right]\boldsymbol{\tau^2\beta}}$

Sandwich variance estimate: For an incorrectly specified marginal model and a non-diagonal working covariance matrix, in general the sandwich variance estimate is biased. However, the sandwich estimate based on a diagonal working covariance matrix is asymptotically unbiased (for the variance of the estimate based on the independence working model).

Lung Health Study Example: The LHS was a RCT of smoking cessation and broncho-dilator use with $n = 4491$ participants and $m = 5$ anticipated visits.

Y_{it} = FEV$_1$, the forced expiratory volume in one second

$$X_{it} = \begin{cases} 0 & = \text{ smoking} \\ 1 & = \text{ not smoking} \end{cases}$$

Analysis is based on data from participants who completed all 5 annual visits. Thus, comparisons are not influenced by missing data. BUT, since the missing data process violates *missing completely at random*, validity is compromised.

Comparisons: Fit the marginal linear model: $E(Y_{it}|X_{it}) = \mu_t + \beta X_{it}$. Here, β is the overall, year-specific *non-smoking* effect on FEV$_1$ using only current year smoking. (But, smoking in previous years obviously influences current year FEV$_1$).

Use the working models:

IND:	Independence
CS:	Compound symmetry
AR1:	First-order autoregressive
UN:	Unstructured
YS:	Year-specific, cross-sectional: $(\beta_1, \ldots, \beta_5)$

Results

LONGITUDINAL

	IND	CS	AR1	UN
$1000\hat{\beta}$	168	59	55	56
1000SE	9	4	4	4

CROSS-SECTIONAL

t	1	2	3	4	5
$1000\hat{\beta}_t$	124	164	170	184	192
1000SE	21	21	21	21	21

- $\hat{\beta}_{ind} = \frac{1}{5}\sum \hat{\beta}_t$
- $\hat{\beta}_{ind}$ is very different from the $\hat{\beta}_{dependent}$
- The Ys are partial surrogates for Xs in “other years” and so $\hat{\beta}_{dependent}$ is closer to 0 than is $\hat{\beta}_{ind}$
- All the $\hat{\beta}_{dependent}$ are similar and are biased estimates of β. They do have smaller reported standard errors than $\hat{\beta}_{ind}$, but in the face of the

substantial bias, having a small SE is not a virtue.

Informative Sample Size (similar to informative censoring)

Louis GB, Dukic VM, Heagerty PJ, Louis TA, Lynch CD, Ryan LM, Schisterman EF, Trumble A and the pregnancy modeling working group (2006). Analysis of Repeated Pregnancy Outcomes. *Statistical Methods in Medical Research,* **15:** 103-126.

- Y = birth weight
- X = time varying regressor
- Model the marginal mean: $E(Y \mid X, \text{live birth})$
- 300 potential mothers with up to 5 completed pregnancies
- Data generated by $Y = \beta_0 + \beta_1 X + \epsilon$
    - The ϵ correlation structure is generated by a random intercept, random slopes and auto-correlated errors
    - True slopes are $\beta_0 = \beta_1 = 1$

Estimation methods (Robust SEs for all approaches)

1. Use the first birth only
2. Use all births and GEE with working independence
3. Use all births and a random intercept working covariance. This is GEE with working compound symmetry, which leaves out the random slope and autocorrelation and so is not the correct likelihood.

Babies at random: Probability of a baby in the next time period does not depend on the past.

	β_0	β_1
OLS, first-born only	$0.98_{(.37)}$	$0.99_{(.17)}$
GEE, working independence	$1.00_{(.04)}$	$1.00_{(.04)}$
GEE, random intercept	$1.00_{(.04)}$	$1.00_{(.04)}$

Simulation mean$_{(se)}$

- All approaches are unbiased
- Using only the first birth is inefficient

History-dependent babies: The probability of having another pregnancy depends on the previous birth outcome.

	β_0	β_1
OLS, first-born only	$1.05_{(.35)}$	$1.02_{(.18)}$
GEE, working independence	$1.01_{(.04)}$	$1.00_{(.04)}$
GEE, random intercept	$0.74_{(.06)}$	$0.91_{(.04)}$

Simulation mean$_{(se)}$

- First-born and OLS are unbiased
- GEE (random intercept) is negatively biased (an attenuated covariate effect) and in this example is no more efficient than OLS.

Lessons Learned

- Means confound covariances and vice-versa.
- Dropout-induced bias ("non-ignorable" missingness) further complicates model selection.

- It is worth questioning whether the potential efficiency and dropout protection of nearly-correct working models are worth the risk.
- A proposed analytic strategy
  - ○ Start with a working independence model.
  - ○ Vigorously investigate models with full covariate conditioning.
    - ◇ Use robust SEs.
  - ○ Then, refit with non-independence working assumptions and robust SEs.
  - ○ Re-run, evaluate, re-run, ...
- A complicated business!

Keeping the Analysis Targeted on Your Goals

Class/Indicator Variables: Class variables, how you represent matters if you are going to interpret them; with an intercept, requiring that a weighted sum of the "effects" = 0 removes the extra degree of freedom, but you don't get "direct" estimates of the expectation for each group. Fitting without an intercept eliminates the need to remove a degree of freedom and you can estimate group-specific expectations and combine them. For the intercept-only model, you get an estimate of the overall mean, but you lose control of the weights.

J categories, $j = 1, \ldots, J$. Y_j a column mean. Model the means,

$$\begin{aligned} E(Y_j) &= \mu_j = \alpha + \delta_j \\ V(Y_j) &= \sigma_j^2 \ (= \sigma^2/n_j) \\ \sum_j a_j \delta_j &= 0, \ a_+ = 1 \end{aligned}$$

SAS uses $a_1 = a_2 = \ldots = a_{J-1} = 0, \ a_J = 1.$

Some programs use: $a_1 = a_2 = \ldots = a_J = 1/J.$

Coupled with the relative sizes of the σ_j^2, each produces a different set of $\hat{\alpha}$s.

Using the "noint" option, we get direct estimates $\hat{\mu}_j = Y_j$ and can then control how we weight them to produce $\hat{\mu}(w) = \sum_j w_j \hat{\mu}_j,\ w_+ = 1.$ This estimate has variance:

$$V(\hat{\mu}\{\mathbf{w}\}) \;=\; \sum_j w_j^2 \sigma_j^2 \geq \left(\frac{1}{\sum_j \sigma_j^{-2}} \right) = V(\hat{\mu}\{w_j = \sigma_j^{-2}\}) = \text{“optimal” variance.}$$

If you fit the sub-model with $\delta_j \equiv 0, \forall j$, you will get,

$$\begin{aligned} \hat{\alpha} &= \hat{\mu}\{w_j = \sigma_j^{-2}\} \\ &= \sum_j n_j Y_j, \text{ if } \sigma_j^2 = \sigma^2/n_j \end{aligned}$$

So, for complete control of the weights, fit the "noint" model, get the $\hat{\mu}_j$ and then use your desired weights to combine them.

Of course, all of this is straightforward if you have no covariates; just do whatever you want with group means. However these issues also apply when you want "adjusted means," that is in the context of a covariate adjustment.

Individual means: What if you want to estimate the individual μ_j? Should you just use $\hat{\mu}_j = Y_j$ or should you "combine evidence?" The answer depends on whether column "j" provides information on column "j'". It will for a Random Effects ANOVA and won't for a Fixed Effects ANOVA.

Simultaneous CIs and tests; multiple comparisons

(S and T methods, k-ratio method)

Ingo's notes, ch12 Simultaneous Confidence Intervals

S-method: Control simultaneous coverage for all l.c. $c'\hat{\boldsymbol{\beta}}$
Bayesian version is more transparent:
$pr(\boldsymbol{\beta} \in R) = 1 - \alpha$ and R has smallest volume for optimal.

Then, for the CI of $c'\boldsymbol{\beta}$ use,

$$\begin{aligned} \text{lower} &= \inf_{\boldsymbol{\beta} \in R} (c'\boldsymbol{\beta}) \\ \text{upper} &= \sup_{\boldsymbol{\beta} \in R} (c'\boldsymbol{\beta}) \end{aligned}$$

The idea is to move the plane of constancy until it just touches R on the two extremes.

This is conservative compared to the single contrast CI, but in general not as conservative as Bonferroni.

Bonferroni: Staying with the Bayesian formalism, consider the $d = 2$ case with the goal of producing a confidence interval for θ_1 when:
$\boldsymbol{\theta} = (\theta_1,\ \theta_2) \sim N_2(\mathbf{0}, I_2)$. Using polar coordinates, the $(1 - \alpha)$ HPD region is

$$\|\ \boldsymbol{\theta}\ \|^2 \leq -2\log(\alpha)\ .$$

Projecting this region to the first coordinate gives

$$-\sqrt{-2\log(\alpha)} \leq \theta_1 \leq \sqrt{-2\log(\alpha)}\ .$$

Table **??** shows that for $d = 2$ the Bayesian HPD interval is longer than either the unadjusted or the Bonferroni adjusted intervals (α_{bf} is the Bonferroni non-coverage probability). That is, the Bayesian interval is actually more conservative, as is the S-method interval (indeed, they are essentially the same). However, both of these approaches control coverage for $\boxed{\text{all}}$ linear combinations; Bonferroni does not.

α	$Z_{\alpha/2}$	$Z_{\alpha_{bf}/2}$	$\sqrt{-2\log(\alpha)}$
.01	2.57	2.81	3.03
.02	2.33	2.57	2.80
.05	1.96	2.24	2.45
.10	1.65	1.95	2.15

Table 1: Comparison of upper univariate confindence limits, multiple comparisons setting with $K = 2$.

The d-dimensional region is,

$$||\boldsymbol{\theta}||^2 < \chi^2_d(\alpha)$$

where $\chi^2_d(\alpha)$ is the cutpoint that leaves α probability to the right in a central chi-square distribution.

Projecting onto the first coordinate gives,

$$-\chi_d(\alpha) < \theta_1 < \chi_d(\alpha)$$

For the Bonferroni approach, use,

$$-Z_{\alpha_{bf}(d,\alpha)/2} < \theta_1 < Z_{\alpha_{bf}(d,\alpha)/2}$$

where

$$\alpha_{bf}(d,\alpha) = 1 - (1-\alpha)^{1/d}$$

And, of course, for the no-adjustment approach use

$$-Z_{\alpha/2} < \theta_1 < Z_{\alpha/2}$$

As exercise 6 shows, as d increases, the Bayesian HPD interval and the S-method interval for a single linear combination remain longer than the Bonferroni interval that controls d comparisons. However, the penalty is small and the benefit is the ability to control coverage for all linear combinations.

More on Multiplicity

Candidate Comparisons for Three Treatment Groups

What follows are some considerations when choosing comparisons when there are three, statistically independent, treatment groups along with some of the consequences. I label the groups "C" for control, "M" for middle level and "H" for high. The groups are hierarchical in that,

$$\begin{aligned} \mathsf{M} &= \ \mathsf{C} + \mathsf{A} \\ \mathsf{H} &= \ \mathsf{M} + \mathsf{B} = \mathsf{C} + \mathsf{A} + \mathsf{B} \end{aligned}$$

with "A" and "B" being a component interventions.

Pairs of Statistically Independent Comparisons

With three treatment groups there are a several candidate pairs of statistically independent comparisons. Independence requires orthogonality of the contrasts. Here are a two pairs that pay attention to the hierarchy:

$\{$(M&H vs C) and (H vs M)$\}$ or $\{$(M vs C) and (H vs M&C)$\}$

Unnormalized Contrast	C	M	H
(M&H vs C)	-2	+1	+1
(H vs M)	0	-1	+1

Unnormalized Contrast	C	M	H
(M vs C)	-1	+1	0
(H vs M&C)	-1	-1	+2

Linear and quadratic contrasts also attend to the hierarchical structure,

Unnormalized Contrast	C	M	H
Linear	-1	0	+1
Quadratic	+1	-2	+1

Triples of Comparisons

Three comparisons can be constructed, but they won't be mutually independent. Here are two sets of contrasts:

{(M vs C) (H vs C) (H vs M)} or {(M vs C) (H&M vs C) (H vs M)}

Unnormalized Contrast	C	M	H
(M vs C)	-1	+1	0
(H vs C)	-1	0	+1
(H vs M)	0	-1	1

Unnormalized Contrast	C	M	H
(M vs C)	-1	+1	0
(H&M vs C)	-2	+1	+1
(H vs M)	0	-1	+1

An Incomplete Factorial Representation

In addition, the design can be represented as an incomplete factorial by indicating whether the group has or does not have component A and whether or not it has component B. This representation produces,

	b	B
a	✓	
A	✓	✓

The (aB) combination is missing. This construct encourages estimation and testing of the main effects for A and for B, with these main effects addressing comparisons somewhat different from any of the foregoing.

Bonferroni and Beyond

There are many choices of comparisons with at most two being statistically independent. Of course, lack of statistical independence should not impede powering for three comparisons. If 2 independent comparisons are targeted, the exact Bonferroni adjustment will be neither liberal nor conservative (exact meaning using $1-(1-\alpha)^{\frac{1}{2}}$ as the nominal level) and the $\alpha/2$ approach is virtually exact (this is not the case for a large number of independent comparisons).

When there is dependency among the comparisons, (either 2 dependent comparisons or ≥ 3 comparisons), the Bonferroni approach can be conservative with the degree of conservatism depending on the correlation structure. It's more efficient to take dependence into account, though the efficiency gain can be small. To see this, consider the comparisons,

$$\{(\text{M vs C}) \ (\text{H vs C}) \ (\text{H vs M})\} \text{ or } \{(\text{M vs C}) \ (\text{H\&M vs C}) \ (\text{H vs M})\}$$

With equal sample sizes per group, the Z-scores for these comparisons have correlation matrces,

$$\text{cor}(Z_1, Z_2, Z_3) = \begin{pmatrix} 1.0 & 0.5 & -0.5 \\ 0.5 & 1.0 & 0.5 \\ -0.5 & 0.5 & 1.0 \end{pmatrix}$$

and

$$\text{cor}(Z_1, Z_2, Z_3) = \begin{pmatrix} 1.00 & 0.87 & -0.50 \\ 0.87 & 1.00 & 0 \\ 0.50 & 0 & 1.00 \end{pmatrix}$$

respectively. An R-program was used to simulate, under H_0, the probability distribution of $\min(Z_1, Z_2, Z_3)$ and the appropriate Z-value cutpoint for various nominal Type I error levels. Tables 1 and 2 report results.

As can be seen, for three comparisons taking account of the correlation produces only a modest efficiency gain and that the gain depends on the set of contrasts. In Table 4, for the two-sided 5% test (the 0.0250 column in the table) a p-value of 0.0092 rather than 0.0083 can be used. Over all Type I errors, the nominal Type I error is multiplied by about 0.36 rather than by 0.33. In Table 5, for the two-sided 5% test a p-value of 0.0100 rather than 0.0083 can be used. Over all Type I errors, the nominal Type I error is multiplied by about 0.40 rather than by 0.33. Related comparisons hold for the Z-score cutpoints.

```
alpha             0.0500 0.0250 0.01250 0.0200 0.0100 0.0050
alphadiv2         0.0250 0.0125 0.00625 0.0100 0.0050 0.0025
truepvalue        0.0188 0.0092 0.00450 0.0073 0.0036 0.0017
alphadiv3         0.0167 0.0083 0.00420 0.0067 0.0033 0.0017
ratio_truetoalpha 0.3770 0.3670 0.35800 0.3640 0.3580 0.3480

Z_alphadiv2       1.9600 2.2414 2.49770 2.3263 2.5758 2.8070
Z_truep           2.0783 2.3587 2.61350 2.4436 2.6889 2.9215
Z_alphadiv3       2.1280 2.3940 2.63830 2.4747 2.7131 2.9352
```

Table 2: Results for {(M vs C) (H vs C) (H vs M)}:
alpha is the nominal one-sided type I error; alphadiv2 = alpha/2; truepvlaue is the p-value that should be used to produce the nominal type I error; alphadiv3 = alpha/3; ratio_truetoalpha = truepvalue/alpha; the Z-values are those associated with the foregoing probabilities.

```
alpha             0.0500 0.0250 0.01250 0.0200 0.0100 0.0050
alphadiv2         0.0250 0.0125 0.00625 0.0100 0.0050 0.0025
truepvalue        0.0204 0.0100 0.00490 0.0079 0.0039 0.0019
alphadiv3         0.0167 0.0083 0.00420 0.0067 0.0033 0.0017
ratio_truetoalpha 0.4080 0.3990 0.39100 0.3970 0.3890 0.3850

Z_alphadiv2       1.9600 2.2414 2.49770 2.3263 2.5758 2.8070
Z_truep           2.0460 2.3273 2.58410 2.4115 2.6616 2.8900
Z_alphadiv3       2.1280 2.3940 2.63830 2.4747 2.7131 2.9352
```

Table 3: Results for {(M vs C) (H&M vs C) (H vs M)}:
alpha is the nominal one-sided type I error; alphadiv2 = alpha/2; truepvlaue is the p-value that should be used to produce the nominal type I error; alphadiv3 = alpha/3; ratio_truetoalpha = truepvalue/alpha; the Z-values are those associated with the foregoing probabilities.

Design

We are all Bayesians in the design phase

General Issues

- Accommodate uncertainty in inputs
  - Expected power, expected CI length
  - pr(power $>$ 0.80) $>$ 0.90
  - pr(CI length $>$ L) $<$ 0.07
- Allow for model checking
  - 2-point design is “optimal” for linear regression, but has no information to check linearity

- Robustness
- Ease of implementation

Present, the "sample size" file

List of some designs

- A optimal; D optimal; targeted parameter optimal
- Sequential: e.g., Go until SE is sufficiently small
  - In general, for sequential monitoring the analysis needs to be adjusted (e.g., monitor the $\hat{\beta}$. But,
  - If monitoring the design matrix, it is ancillary (we condition on $\mathbf{X}$) and so standard analysis is valid
  - If monitoring $\hat{\sigma}$ and residuals are Gaussian, this is ancillary to $\hat{\boldsymbol{\beta}}$ and standard analysis is valid.
- Orthogonal Structure: ANOVA
  - One-way: options orthog and not
  - Two-way: ditto
- Balanced incomplete block
- Factorial and Fractional factorial
- Rotatable
- Crossover
- Nested
- Latin Square, Graeco-Latin Square (display it), Latin Hypercube, ...

$\boxed{\text{Ingo's notes, ch15}}$ Introduction to Design

$\boxed{\text{Ingo's notes, ch16}}$ One-Way ANOVA

$\boxed{\text{Present: Fixed-effects vs Random Effects}}$

$\boxed{\text{Present: best paper}}$

Bayes/RE & Multiplicity

- In a random effects ANOVA, exploiting the RE structure provides "automatic" multiplicity control for comparing columns via use of posterior means (aka, BLUP)
- With additive, component-specific losses each comparison is optimized separately with no accounting for the number of comparisons
- However, use of a hyper-prior (or EB) links the components since the posterior "borrows information"
  - ○ Shrinkage as a multiplicity control
- If collective penalties are needed, use a multiplicity-explicit loss function

Shrinkage as a multiplicity control: The K-ratio approach, (Dixon&Duncan)

RE ANOVA

- $\theta_1, \ldots, \theta_K \quad iid \quad N(\mu, \tau^2)$
- $[Y_{ik} \mid \theta_k] \quad ind \quad N(\theta_k, \sigma^2)$
- $[\theta_k \mid Y_{.k}] \quad \sim \quad N\left(\mu + (1-B)(Y_{.k} - \mu), (1-B)\frac{\sigma^2}{n}\right)$

$$\begin{aligned} F &= 1/\hat{B} = (\hat{\sigma}^2 + n\hat{\tau}^2)/\hat{\sigma}^2 \\ (1-\hat{B}) &= \frac{n\hat{\tau}^2}{\hat{\sigma}^2 + n\hat{\tau}^2} = \frac{F-1}{F} \end{aligned}$$

Compare columns 1 and 2:

$$Z_{12}^{Bayes} = Z_{12}^{freq} \left\{ \frac{(\mathbf{F} - \mathbf{1})^{+}}{\mathbf{F}} \right\}^{\frac{1}{2}} = \left(\frac{\sqrt{n}(Y_{.1} - Y_{.2})}{\hat{\sigma}\sqrt{2}} \right) \left\{ \frac{(F-1)^{+}}{F} \right\}^{\frac{1}{2}}$$

Comments

- The magnitude of F adjusts the test statistic
- For large K, under the global null hypothesis ($\tau^2 = 0$), pr[all $Z_{ij} = 0] \geq 0.5$
- The FW rejection rate is much smaller than 0.5
- "Scoping" is important because the number of candidate comparisons influences the value of $\hat{\mu}$ and $\hat{B}$ and performance more generally
- Non-additive loss functions can be used: e.g., 1 + 1 = 2.5

Ingo's notes, ch17	Nested and higher-order designs
Ingo's notes, ch18	Miscellaneous ANOVA

The Flexibility of Likelihood-Based Methods

Nested, crossed, factorial, fractional factorial, . . . are very important in designing a study. They enable economical use of research "units" to achieve primary goals.

With the advent of likelihood-based methods implemented by recursive algorithms, we no longer have to deal with "pure forms," forms for which we can compute closed-form sums of squares. This is true even for Gaussian, linear models, and of course supremely so for non-Gaussian models.

Still, care is needed (more care) on estimabilty, etc.

Respondents and interviewers, different ones at different visits

Rodents in cages, change cages

Old concept: "Unit of analysis" (the measurement, the person, the household, . . .)

New concept: Variance components and RHS models,

Random Effects should replace "unit of analysis"

- What is your unit of analysis?
- Well, sometimes it's the serum level, sometimes it's the patient, sometimes it's the clinician, sometimes the clinic, . . .
- There are many "units" and so in effect no single set of units
- Models contain Fixed-effects, Random effects (via Variance Components) and other correlation-inducers
- Random Effects induce unexplained (co)variance
- Some of the unexplained may be explicable by including additional covariates

Value added modeling of teacher effects (not presented in class)

Initial Goal: The initial goal is to isolate the effect of the teacher on student achievement growth during a single academic year in a single subject. However, students go from teacher to teacher, class to class, . . . and aren't "nested." The following notation allows for this progression.

Assumptions
Linear trend and additive error structure.
Exam scores have been adjusted/calibrated so that the "Ys" are calibrated over grades and over time.

Notation

t = Time period, calendar quarter or other time period sequence that lines up with school instruction cycles and allows determining "summer," etc. For example, if the school year is in semesters with a summer break, we can use 1995.1, 1995.2, 1995.3, where 1995.1 is the spring 1995 semester, 1995.2 is summer; 1995.3 is fall. Or we can use another coding.

s = School, s = 1, . . . , S

C = Course indicators (e.g., third-grade math, section 1); unique labels within a school, We use index k = 1, . . . , K in basic models.

j = Teacher in school, j = 1, . . . , J$_s$.

g = Grade (e.g., 1st ....)

i = Student, indexed by “i = 1, . . . I$_s$.

$\mathbf{C}_{stu}(i, s, t)$ = Vector of course indicators for student “i” in school “s” in time period “t.”

$\mathbf{C}_{tea}(j, s, t)$ = Vector of course indicators for teacher “j” in school “s” in time period “t.” In most grade schools a student has one teacher (in the tested courses). In this case the $\mathbf{C}_{stu}(i, s, t)$ vector of course indicators will be a subset of the $\mathbf{C}_{tea}(j, s, t)$ course indicators.

$\mathbf{X}_{ist}$ = Covariate (race/ethnicity, gender, special ed status, age) vector for student “i” in school “s” in time period “t.”

$\mathbf{Y}_{ist}$ = Test score vector for student “i” in school “s” in time period “t.”

$\mathbf{Y}_{is}$ = Test score vector for student “i” in school “s” over all time periods considered.

$\mathbf{Y}_{s}$ = Test score vector for school “s” over all time periods considered.

A first-difference model: First, consider a model for the student-specific *test score increments*. It provides a good basis for understanding the longitudinal model that comes next. The increment occurs during one teacher/class contact. Student-specific change is comprised of a general trend, a course effect, a school effect, a teacher effect, a student effect and covariate effects. Not all of these will be estimable in all data sets, but the model is interesting nevertheless.

For ease of notation, assume that the tests are taken at unit intervals and write include "*l*" as part of a term name to indicate that it is a "Longitudinal" term, "*c*" to indicate a cross-sectional term, "*f*" to indicate that it is a *fixed effect* and "*r*" to indicate that it is a *random effect*. Since we are dealing with only one subject (math), Y_{ist} is a scalar. For the i^th^ student with the j^th^ teacher in the s^th^ school at time t, we have:

$$\begin{aligned}(Y_{ijst} - Y_{ijs(t-1)}) &= [INT^{lf} + CRS^{lf}_{math}] + SCH^{l?}_s + TEA^{lr}_j + STU^{lr}_i \\ &\quad + (X_{ijst} - X_{ijs(t-1)})\beta + \delta_{ijst}.\end{aligned}$$

The random effects are assumed drawn from mean 0 distributions with variances that produce the specific heterogeneities. The distributions don't have to be Gaussian. Note that the student/teacher-specific covariate matrix is also first-differenced, that we can decide whether SCH should be fixed or random, that I'm not including all of the "*t*" indices on the RHS terms. Since we have only one class (k $\equiv$ 1), CRS^{lf}_k is confounded with INT^{lf}. There are a wide range of generalizations. I have not included the teacher by student interaction, but it can be included, if the data support estimating it.

The associated cross-sectional model: Now, put this in a "cross-sectional" form by undoing the first differences in (**??**):

$$\begin{aligned}Y_{ijst} &= [INT^{cf} + CRS^{cf}_{math}] + SCH^{c?}_s + TEA^{cr}_j + STU^{cr}_i + \\ &\quad \left\{[INT^{lf} + CRS^{lf}_{math}] + SCH^{l?}_s + TEA^{lr}_j + STU^{lr}_i\right\} \cdot t + X_{ist}\beta + \epsilon_{ijst},\end{aligned} \qquad (5)$$

where the ϵ_{ist} are independent for different schools $\left(\delta_{ijst} = (\epsilon_{ijst} - \epsilon_{ijs(t-1)}\right)$. The ϵ s can be dependent or independent for different students within a school/teacher, and are definitely dependent within student. A good starting point is to assume that they follow a model which is the sum of compound symmetry (random intercept)and AR(1).

Multi-course, multi-year models

Use the notation in the notation section to develop a mega-model. Have fun!

Missing data

Daniels MJ, Hogan JW (2008). *Missing Data in Longitudinal Studies, Strategies for Bayesian Modeling and Sensitivity Analysis*, Chapman & Hall/CRC, Boca Raton, FL

Missing data processes: Let $\mathbf{T}$ represent all the potential data (dependent variables, regressors, . . .). For a data item i (a full unit or a subset of the data), let $R_i = 0$ or 1 according as the item is unobserved or observed. The general missing data process is $pr(R_i = 0 \mid \text{all observeds}, \mathbf{T}_i, \theta)$ and the various types of missing data (really, missing data processes!) are determined by crossing things out of this conditional probability.

Missing Completely at Random (MCAR): $pr(R_i = 0)$

Missing at Random (MAR): $pr(R_i = 0 \mid \text{all observeds})$

Not missing at random (NMAR): $pr(R_i = 0 \mid \text{all observeds}, \mathbf{T}_i, \theta)$
There are many flavors of NMAR.

Note that MCAR $\subset$ MAR $\subset$ NMAR.

Ignorability: If the missing process is MAR and you have the correct likelihood for the observed data, then you can ignore (don't have to model) the missing data process and still have the valid, full likelihood inference. However, if you model for the observed data is not correct, you can have a biased analysis. So, ignorability is not an inherent property of a missing data process. MAR sets up the potential for ignorability, but you need to correct likelihood for the observeds to achieve it.

Example: For $i = 1, \dots, n$ let

$$(Y_{i1}, Y_{i2}) \sim N_2\left[\underline{0}, \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}\right]$$

be iid. In the study Y_{i1} is observed for everyone, but Y_{i2} is observed only if

$Y_{i1} > 0.7$. This is MAR and so the missing data process is potentially ignorable. To be ignorable, you need to use the correct likelihood. For example, the average of all of the observed Y_{i2} is a biased estimate of $\underline{0}$. However, if you pose a model with an unknown mean vector $\boldsymbol{\mu} = (\mu_1, \mu_2)'$ and unknown covariance matrix Σ and maximize the correlated data likelihood, you get consistent estimates that are nearly unbiased. To maximize the likelihood, use that $[Y_{i2} \mid Y_{i1}] \sim N[\mu_2 + \rho(Y_{i1} - \mu_1), (1 - \rho^2)]$. However, as the following numerical example shows, if you restrict Σ, you can get non-consistent, biased estimates.

Numerical Example

- Generate 10,000 pairs of data values with $\rho = 0.60$
  - The sample variances are 1 and the sample correlation is 0.596
- Analyze for baseline intercept and (visit2 - visit1) change (via a model, not by taking differences)
- Analyze with working independence (WI) and with unstructured (UN) covariance.
- Three datasets
  1. The 10,000 pairs
  2. The 10,000 baseline values, but delete a random 2,500 of the follow-up values (missing completely at random)
  3. The 10,000 baseline values, but delete the second observation for cases with the first observation > 0.70 (produces about 25% missing second visits

Working Model	Full Data	Random Deletion	Delete if $Y_1 > 0.70$
Independence	$108_{(90)}$	$125_{(107)}$	$-232_{(120)}$
Unstructured	$108_{(90)}$	$90_{(101)}$	$116_{(100)}$

Table 4: $10^4 \hat{E}(Y_2 - Y_1)_{(se)}$

Example of the formalism

Two potential observations; the first is always observed; the second possibly observed. Write out the general likelihood with R_i indicating that the second observation is observed ($R_i = 1$ rather than 0):

$$L(Y_{i1}, Y_{i2}, R_i; \theta) = L(Y_{i1}; \theta) L(R_i \mid Y_{i1}; \theta) \{L(Y_{i2} \mid Y_{i1}, R_i; \theta)\}^{R_i}$$

If MAR, for $R_i = 1$ we have (no θ in the second term),

$$= L(Y_{i1}; \theta) L(R_i = 1 \mid Y_{i1}) L(Y_{i2} \mid Y_{i1}, R_i = 1; \theta).$$

By the conditional independence of Y_{i2} and R_i given Y_{i1}

$$= L(Y_{i1}; \theta) L(Y_{i2} \mid Y_{i1}; \theta) L(R_i = 1 \mid Y_{i1}). \tag{6}$$

For $R_i = 0$, we have

$$= L(Y_{i1}; \theta) L(R_i = 0 \mid Y_{i1}) \tag{7}$$

Note that θ-dependence in (**??**) and (**??**) is only via the likelihood for the observed data and so the missing data process is "ignorable."

This derivation depends on the conditional independence of Y_{i2} and R_i given Y_{i1} using the correct probability model for this conditioning. If you don't use the correct probability model, the likelihood (the score) can be biased. For example, assuming that (Y_{i1}, Y_{i2}) are independent when they are not will produce a biased result.

Present: missing data power point

MEASUREMENT ERROR

Effects of Measurement Error

- Attenuation of the true, structural exposure effect
  - ○ Under-appreciation of regulatory and public health potentials

- Variance inflation/deflation
- Change in the dose-response shape
  - Linear to non-linear
  - Elimination of a threshold
- Challenges research synthesis
  - Findings depend on "methodologic variation"
  - Contributes to between-study heterogeneity

Dose-response Modeling

- Determinants of Slope and Shape
  - Basic form of the model
  - Exposure metameter
  - Measurement error
  - Type of background
  - Population heterogeneity

Notation

Y = response variable

$X_t =$ true exposure

$X_o =$ observed exposure

$g(X_t,\ X_o) =$ joint $(X_t,\ X_o)$ distribution

Types of Measurement Error

Berkson

$$X_t = X_o + \delta$$

$$E[X_t \mid X_o = x_o] = x_o$$

"Standard"

$X_o = X_t + \delta$

$E[X_t \mid X_o = x_o] = \rho x_o$

$\rho = \sigma_t^2/(\sigma_t^2 + \sigma_\delta^2) = \text{corr}(X_t, X_o)^2$

Unified Approach

$$\begin{aligned} R(x_t \mid \beta) &= \text{the dose response relation} \\ Y &\sim \text{statistical model } \{R(x_t \mid \beta),\, \epsilon\} \\ R(x_t \mid \beta) &= E[Y \mid X_t = x_t] \\ \bar{R}(x_o \mid \beta) &= \int R(x_t \mid \beta) g(x_t \mid X_o = x_o) dx_t \end{aligned}$$

Linear Model

$$\begin{aligned} Y &= R(x_t \mid \beta) + \epsilon \\ R(x_t \mid \beta) &= \text{int} + \beta x_t \\ \bar{R}(x_o \mid \beta) &= \text{int} + \beta E[X_t \mid X_o = x_o] \end{aligned}$$

$\boxed{\text{Berkson}}$

$$\begin{aligned} X_t &= X_o + \delta \\ \bar{R}(x_o \mid \beta) &= \text{int} + \beta x_o \\ E[X_t \mid X_o = x_o] &= x_o \\ \text{residual variance} &= \beta^2 \sigma_\delta^2 + \sigma_\epsilon^2 \end{aligned}$$

$\boxed{\text{“Standard”}}$

$$\begin{aligned} X_o &= X_t + \delta \\ \bar{R}(x_o \mid \beta) &= = \mathsf{int} + (\rho\beta)x_o \\ &= \mathsf{int} + \beta^* x_o \\ E[X_t \mid X_o = x_o] &= \rho x_o \\ \rho &= \sigma_t^2/(\sigma_t^2 + \sigma_\delta^2) \end{aligned}$$

$\boxed{\text{“Hybrid”}}$

Neither of the above

Deattenuation

Best done by the unified approach, automatically de-attenuation and variance adjustment.

- Traditionally, done by using a known or estimated ρ and computing $\hat{\beta}_1 = \hat{\beta}^*/\rho$
- $V(\hat{\beta}_1) = V(\hat{\beta}^*)/\rho^2$
- MSE($\hat{\beta}_1$) $=$ you do it.

This adjustment can induce a high variance, so instead,

- Partially deattenuate (minimize Mean Squared Error) using,

$$\begin{aligned} \hat{\beta}_c &= c(\hat{\beta}^*/\rho) \\ c_{opt} &= \frac{(\rho\beta)^2}{V(\hat{\beta}^*) + (\rho\beta)^2} \\ \hat{c}_{opt} &= \frac{(\hat{\beta}^*)^2}{V(\hat{\beta}^*) + (\hat{\beta}^*)^2} \end{aligned}$$

- Shrinkage because $(0 \leq c \leq 1)$, producing an effective variance/bias trade-off

Linear can become non-linear

- Standard measurement error, but exposures are log-normal

$$\begin{aligned} E[X_t \mid X_o = x_o] &= \text{slope} \times x_o^{\rho} \\ \log(\text{slope}) &= (1-\rho)\mu + \frac{1}{2}\rho\sigma_\delta^2 \end{aligned}$$

Exponential Model

$$\begin{aligned} 1 - R(x_t \mid \beta) &= e^{-\beta x_t} \\ 1 - \bar{R}(x_o \mid \beta) &= \int e^{-\beta x_t} g(x_t \mid x_o) dx_t \\ &= MGF_{X_t \mid X_o = x_o}(\beta) \end{aligned}$$

- If G is Gamma with: Mean $= x_o$ and Variance $= x_o^2/\alpha$, then

$$1 - \bar{R}(x_o) \quad = \quad \left(\beta\frac{x_o}{\alpha} + 1\right)^{-\alpha}$$

$$\rightarrow e^{-\beta x_o} \quad \text{if } \alpha \rightarrow \infty$$

$$\rightarrow 1 \quad \text{if } \alpha \rightarrow 0$$

Threshold with Measurement Error

- x^* is the threshold
- $X_t \mid X_o \sim$ exponential: $E[X_t \mid X_o] = X_o$
- Because of the lack of memory property for the exponential:

$$\begin{aligned} 1 - R(x_t \mid \beta) &= e^{-\beta(x_t - x^*)^+} \\ 1 - \bar{R}(x_o \mid \beta) &= \frac{e^{-\frac{x^*}{x_o}}}{\beta x_o + 1} \end{aligned}$$

- There is no threshold

- $\bar{R}(x_o \mid \beta)$ is "sub-linear" at $x_o = 0$

Score Equations

$$\begin{aligned} f(Y \mid x_t, \beta) &= \text{the structural likelihood} \\ \bar{f}(Y \mid x_0, \beta) &= \int f(Y \mid x_t, \beta) g(x_t \mid X_o = x_o) dx_t \end{aligned}$$

Take logs and differentiate:

$$\begin{aligned} \frac{\partial}{\partial\beta} \log(\bar{f}(Y \mid x_o, \beta)) &= \frac{\int \left\{ \frac{f'(Y|x_t,\beta)}{f(Y|x_t,\beta)} \right\} f(Y \mid x_t, \beta) g(x_t \mid X_o = x_o) dx_t}{\int f(Y \mid x_t, \beta) g(x_t \mid X_o = x_o) dx_t} \\ &= \frac{\int S(x_t \mid \beta) f(Y \mid x_t, \beta) g(x_t \mid X_o = x_o) dx_t}{\int f(Y \mid x_t, \beta) g(x_t \mid X_o = x_o) dx_t} \\ &= E(S(X_t \mid \beta) \mid x_o, Y, \beta) \end{aligned}$$

Notes

- This expectation conditions on everything and in general **depends on** Y and β even though g does not.
- Can be implemented by the EM algorithm
- Can be implemented by multiple imputations to do the integral
- Uses the full analysis model to do them
- So, different from MI that doesn't use the Ys or doesn't use the analysis model

Multivariate Measurement Error

- $\mathbf{X}_t$ and $\mathbf{X}_o$ are vectors
- For example, one coordinate is the exposure of interest and the other is a potential confounder

- Measurement error, especially correlated error, can confound confounding adjustments
- Formal modeling appropriately accounts for the measurement error process
- Commonly producing non-intuitive adjustments
- Of course, information on the joint distribution is necessary

	Coefficients ($\times 10^4$)			
Regressor	Unadj.	Univ adj.	Mult. adj.	Actual
sodium	7	19	23	21
potassium	7	14	-20	-15
calcium	3	7	11	11
caffeine	-19	-30	-31	-30
alcohol	903	1474	1528	1528
bmi	1348	1443	1645	1657

Measurement Error: High, Moderate, Low

Present: Uncertain Genome Calls

Summary

- Measurement error influences all studies
- Prevention is better than (partial) cure
- Prevention depends on personalized (or "unitized") exposure information
- Partial cures depend on side studies that provide calibration information, information on g

Penalized LSE: Ridge and Lasso

Penalties: Minimize

$$\sum_{i=1}^{n}(Y_i - \hat{Y}(\boldsymbol{\beta}))^2 + \lambda\mathcal{P}(\boldsymbol{\beta}) \tag{8}$$

Examples:

$$\begin{aligned}\mathcal{P}(\boldsymbol{\beta}) &= \sum_{\ell=1}^{p} \beta_\ell^2 \quad \text{ridge regression}\\ \mathcal{P}(\boldsymbol{\beta}) &= \sum_{\ell=1}^{p} \mid \beta_\ell \mid \quad \text{Lasso regression}\end{aligned}$$

Application to the intercept-only model: $Y_i = \beta + \epsilon_i$

Ridge:

$$\hat{\beta} = \left(\frac{n}{n+\lambda}\right)\bar{Y}$$

Shrinks towards 0, just like a Bayesian estimate.

Lasso: Take the case where $\bar{Y} \geq 0$. Then,

$$\hat{\beta} = \left(\bar{Y} - \frac{\lambda}{2n}\right)^{+},$$

where $(\ldots)^{+}$ is the positive part.

- So, if $\frac{\lambda}{2n} < \bar{Y}$, $\bar{Y}$ is “slid” towards 0.
  When $\frac{\lambda}{2n} \geq \bar{Y}, \hat{\beta} = 0$ (thresholding). Similarly for $\bar{Y} < 0$.
- In both situations λ can be selected by cross validation. And, in the $p > 1$ situation it is interesting steadily to increase λ and watch estimates get thresholded.
- Since in this basic case there is only one λ, for $p > 1$ scaling is important. Usually, regressors are standardized so the units are comparable.

Robust Methods: Sample Reuse

Present: the Jackknife Powerpoints

- Influence Curve (IC(x)): A functional derivative of $T(F)$, with F a distribution function.
- Idea is

$$IC_F(G) \;=\; \lim_{\epsilon \to 0} \frac{T((1-\epsilon)F + \epsilon G) - T(F)}{\epsilon}$$

Let,

$$G \;=\; I_{\{u \geq x\}} \text{ to obtain } IC(x)$$

- $n \times \hat{V}(\hat{\theta}) = \frac{1}{n} \sum_i IC^2(x_i)$
- You can uncover the implicit or explicit score function via the IC

$$\text{Plot } \widehat{IC}(x_i) = (n-1)\{\hat{\theta}_{-\bullet} - \hat{\theta}_{-i}\} \text{ versus } x_i$$

- M-estimates and weighted averages Solve,

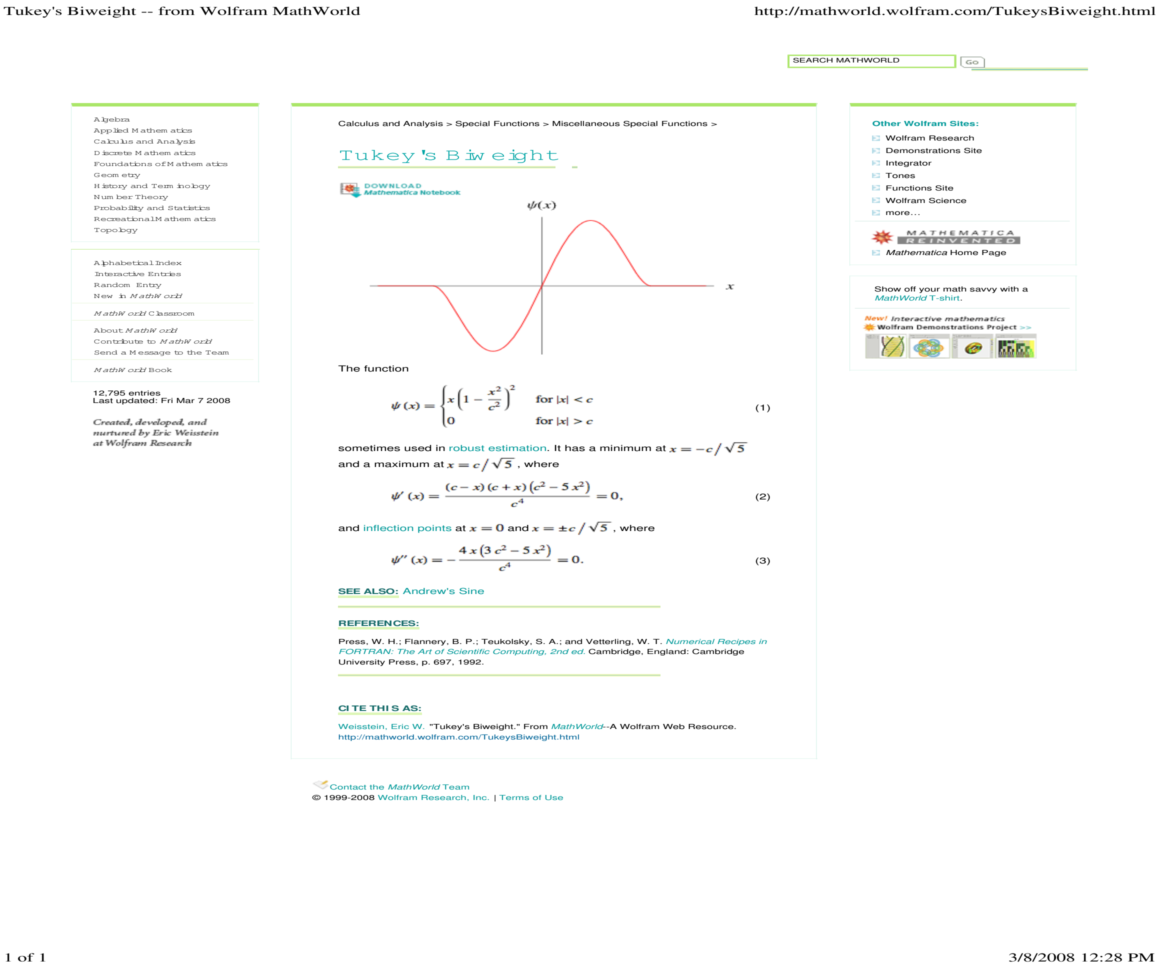
$$\sum_i \psi(u_i) \;=\; 0,\; u_i = \frac{(x_i - T)}{c \cdot \text{scale}}$$

Equivalent to,

$$\sum_i w_i u_i \;=\; 0,\; w_i = \frac{\psi(u_i)}{u_i}$$

Examples, including Tukey's bi-weight with:

$$\begin{aligned} \psi(u) &= u(1-u^2)^2 I_{\{|u| \leq 1\}} \\ w(u) &= (1-u^2)^2 I_{\{|u| \leq 1\}} \end{aligned}$$



Algebra
Applied Mathematics
Calculus and Analysis
Discrete Mathematics
Foundations of Mathematics
Geometry
History and Terminology
Number Theory
Probability and Statistics
Recreational Mathematics
Topology

Alphabetical Index
Interactive Entries
Random Entry
New in *MathWorld*

MathWorld Classroom

About *MathWorld*
Contribute to *MathWorld*
Send a Message to the Team

MathWorld Book

12,795 entries
Last updated: Fri Mar 7 2008

Created, developed, and
nurtured by Eric Weisstein
at Wolfram Research

Calculus and Analysis > Special Functions > Miscellaneous Special Functions >

Tukey's Biweight

The function

$$\psi(x) = \begin{cases} x\left(1-\dfrac{x^2}{c^2}\right)^2 & \text{for } |x| < c \\ 0 & \text{for } |x| > c \end{cases} \qquad (1)$$

sometimes used in robust estimation. It has a minimum at $x = -c/\sqrt{5}$ and a maximum at $x = c/\sqrt{5}$, where

$$\psi'(x) = \frac{(c-x)(c+x)(c^2-5x^2)}{c^4} = 0, \qquad (2)$$

and inflection points at $x = 0$ and $x = \pm c/\sqrt{5}$, where

$$\psi''(x) = -\frac{4x(3c^2-5x^2)}{c^4} = 0. \qquad (3)$$

SEE ALSO: Andrew's Sine

REFERENCES:

Press, W. H.; Flannery, B. P.; Teukolsky, S. A.; and Vetterling, W. T. *Numerical Recipes in FORTRAN: The Art of Scientific Computing, 2nd ed.* Cambridge, England: Cambridge University Press, p. 697, 1992.

CITE THIS AS:

Weisstein, Eric W. "Tukey's Biweight." From *MathWorld*--A Wolfram Web Resource. http://mathworld.wolfram.com/TukeysBiweight.html

Other Wolfram Sites:
Wolfram Research
Demonstrations Site
Integrator
Tones
Functions Site
Wolfram Science
more…

MATHEMATICA REINVENTED
Mathematica Home Page

Show off your math savvy with a *MathWorld* T-shirt.

New! Interactive mathematics
Wolfram Demonstrations Project >>

Contact the *MathWorld* Team

© 1999-2008 Wolfram Research, Inc. | Terms of Use

Figure 1: Tukey's Bi-weight

Iteratively reweighted least squares

- x is the residual (Y_i- model(β))

$$\begin{aligned}\psi(x) &= x\left(1-\frac{x^2}{c^2}\right)^2, \quad \mid x \mid\leq c\\ w(x) &= \left(1-\frac{x^2}{c^2}\right)^2, \quad \mid x \mid\leq c\end{aligned}$$

- Do iteratively reweighted least squares with w.
- If the "tuning constant" $c=\infty$ (i.e., big relative to the scale of x), you get equally weighted LSE
- For finite c, is approximately equal weighting whe
- Bounded influence regression
  Influence = (leverage)×(residual)
  Need to control the impact of both
- Regression Diagnostics and/versus Robust Methods

$\boxed{\text{Look at: M-estimates}}$

FINIS