1 Introduction

What is measurement error?

Occurs when we cannot observe exactly some of the variables which define our model of interest. Usually this is instrument error or sampling error.

A few examples of error-prone variables:

True Value	Observed
presence or absence of disease	result of diagnostic test
dietary intake or physical activity	measure from questionnaire, diary, replication, etc.
air pollution exposure	indoor measures or local monitors
Food expenditures	self report
Employment status	self report
Education level	self report
Land area classification	value from satellite image
animal abundance	estimate from sampling
Delivered “dose” (e.g.,concentration, speed on treadmill)	target dose

The main ingredients in a M.E. problem.

- MODEL FOR THE TRUE VALUES. e.g.,

  1. Estimation of a single proportion, mean, variance, etc.
  2. Contingency tables
  3. Linear regression
  4. Nonlinear Regression (including Binary regression)

- MODEL FOR MEASUREMENT ERROR

  Specification of relationship between error-prone measurements and true values (more later)

- EXTRA INFORMATION/DATA (allowing us to correct for ME)

  1. Knowledge about some of ME parameters or functions of them.
  2. Replicate values
  3. Estimated standard errors attached to error prone variable
  4. Validation data (internal or external)
  5. Instrumental variables.

General Questions

- What are the consequences of analyses which ignore the measurement error (so-called **naive analyses**) on estimation, testing, confidence intervals, etc.?

- How do we correct for measurement error?

  This usually requires information or extra data.

  - Direct bias corrections.

  - Moment approaches.

  - Likelihood based techniques.

  - Regression calibration.

  - Simex.

  - Modified estimating equations.

Differentiality, surrogacy and conditional independence.

X = true value, W = error-prone version of X

Y = another variable (usually a response)

- Differential and non-differential measurement error.

  Nondifferential error: Measurement error in X is non-differential with respect to y if the distribution of $W|x, y$ = distribution of $W|x$

  Differential error: distribution of $W|x, y$ depends on y.

- **Surrogacy:** Distribution of $Y|x, w$ = that of $Y|x$

  w contains no information about y beyond that contained in x.

- **Conditional independence:**

  Y and W are independent given $X = x$

$$\text{Surrogacy} \Leftrightarrow \text{conditional independence} \Leftrightarrow \text{nondifferential error}$$

2 Misclassification in Estimating a Proportion

2.1 Examples

Example 1 : Prevalence of HIV Virus.

Mali et al. (1995): Estimate 4.9% of 4404 women attending a family planning clinic in Nairobi have HIV.

What if assessment is subject to misclassification?

		ELISA(w)		
		0	1	
Truth(x)	0	293 (275)	4 (22)	297
	1	2	86	88

External validation data for the ELISA test for presence of HIV. x is the true value, with indicating being HIV positive, and w is the error prone measure from the assay. Based on Weiss et al. (1985).

If what Weiss et al. (1985) considered borderline are classified as negative then there are only 4 false positives, while if they are considered positive there are 22 false positives.

Example 2 . From Fleiss (1981, p. 202). Goal is to estimate the proportion of heavy smokers in a population.

- $x =$ "Truth" from blood test. (1 = heavy smoker)
- $w =$ self report

Internal validation sample. Overall random sample of 200. 50 subsampled for validation.

		w		
		0	1	
x =	0	24	2	26
	1	6	18	24
		30	20	50
	?	82	68	150
		112	88	200

Example 3. Lara et al. (2004) use as the randomized response technique (RRT) to estimate the proportion of induced abortions in Mexico. With probability 1/2 the woman answered the question "Did you ever interrupt a pregnancy?" and with probability 1/2 answered the question "Were you born in April?". Only the respondent knew which question they answered. Of 370 women interviewed, 56 answered yes.

The RRT intentionally introduces misclassification but with known misclassification probabilities (sensitivity $= 13/24 = .5 + (1/2)(1/12)$; specificity $= 23/24 = 1/2 + (1/2)11/12)$).

Example 4. Use satellite imagery to classify habitat types. Validation data taken from Canada Centre for Remote Sensing for five classes of land-use/land-cover maps produced from a Landsat Thematic Mapper image of a rural area in southern Canada. Unit is a pixel.

	LANDSTAT				
	Water	BareGround	Deciduous	Coniferous	Urban
TRUTH					
Water	367	2	4	3	6
BareGround	2	418	8	9	17
Deciduous	3	14	329	24	25
Coniferous	12	5	26	294	23
Urban	16	26	29	43	422

Other Examples.

- Theau et al. (2005): 10 categories involved in mapping lichen in Caribou habitat
- Chandramohan et al. (2001). Misclassification based on “verbal autopsy” in determination of the cause of death.
- Ekelund et al. (2006). Measure if someone is meeting PA quidelines using 7-day International Physical Activity Questionnaire (IPAQ). This is w; 1 = meets standard. x = assessment by accelerometer (treated as truth).

		w		
		0	1	
x =	0	25	30	55
	1	30	100	130

2.2 Models

$$X = \text{true value} \qquad = 1 \text{ if "success"}, \qquad = 0 \text{ if "failure"}$$

$X_1, \ldots, X_n$ a random sample(i.i.d.)

OBJECTIVE: Inferences for $\pi = P(X_i = 1)$.

With no measurement error: $T =$ number of successes is Binomial(n, π).

$p = T/n =$ proportion of successes in sample, $E(p) = \pi$, $V(p) = \pi(1-\pi)/n$.

- Exact techniques are based on the Binomial.

- Large sample confidence interval: $p \pm z_{\alpha/2}(p(1-p)/n)^{1/2}$.

- Large sample tests based on normal approximation.

With misclassification

Observe W, fallible/error-prone measurement, instead of X.

$P(W = 1|X = 1) = \theta_{1|1}$ (**sensitivity**), $\quad P(W = 0|X = 1) = 1 - \theta_{1|1}$

$P(W = 0|X = 0) = \theta_{0|0}$ (**specificity**), $\quad P(W = 1|X = 0) = 1 - \theta_{0|0}$

Since X is random, we can also derive **reclassification/predictive probabilities**

$$\lambda_{x|w} = P(X = x|W = w).$$

Observe $W_1, \ldots, W_n$: p_W = sample proportion with $W = 1$. **Naive analysis** uses p_W.

$$\pi_W = P(W_i = 1) = \pi(\theta_{1|1} + \theta_{0|0} - 1) + 1 - \theta_{0|0}. \tag{1}$$

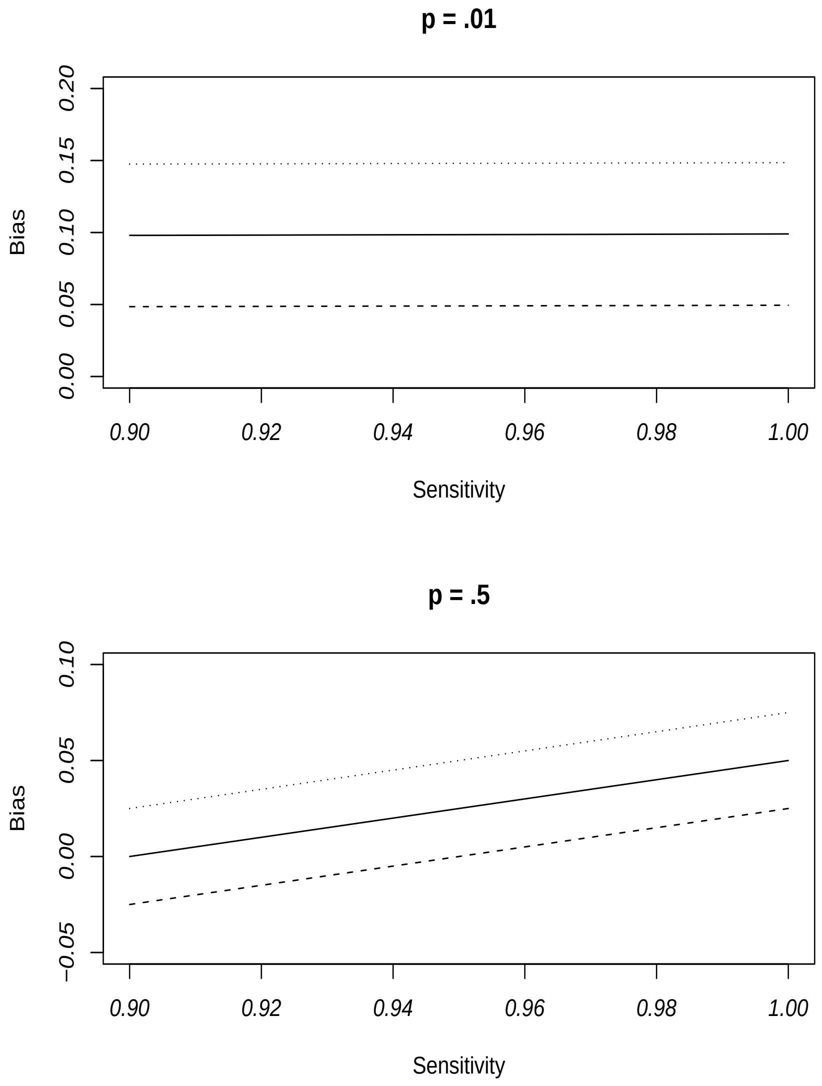
We can also reverse the roles of X and W, leading to

$$\pi = \pi_W(\lambda_{1|1} + \lambda_{0|0} - 1) + 1 - \lambda_{0|0}. \tag{2}$$

Since $E(p_W) = \pi_W$ rather than π, the naive estimator has a bias of

$$BIAS(p_w) = \pi_W - \pi = \pi(\theta_{1|1} + \theta_{0|0} - 2) + (1 - \theta_{0|0}).$$

For a simple problem the nature of the bias can be surprisingly complex. The absolute bias can actually increase as the sensitivity or specificity increases with the other held fixed.



Plot of bias in naive proportion plotted versus sensitivity for different values of specificity (dotted line = .85; solid line = .90; dashed line = .95).

2.3 Correcting for misclassification

Using known or estimated misclassification rates,

$$\hat{\pi} = \frac{p_W - (1 - \hat{\theta}_{0|0})}{\hat{\theta}_{0|0} + \hat{\theta}_{1|1} - 1}. \tag{3}$$

Ignoring uncertainty in the misclassification rates

$$SE(\hat{\pi}) = \left[\frac{p_W(1 - p_W)}{n(\theta_{1|1} + \theta_{0|0} - 1)^2}\right]^{1/2}$$

and an approximate large sample **Wald** confidence interval for π is

$$\hat{\pi} \pm z_{\alpha/2} SE(\hat{\pi}). \tag{4}$$

Exact CI. Get an "exact" CI (L_W, U_W) for π_W (based on the Binomial) and transform it.

Assuming $\hat{\theta}_{1|1} + \hat{\theta}_{0|0} - 1 > 0$ (it usually is)

$$[L, U] = \left[\frac{(L_W - (1 - \hat{\theta}_{0|0}))}{(\hat{\theta}_{1|1} + \hat{\theta}_{0|0} - 1)}, \frac{(U_W - (1 - \hat{\theta}_{0|0}))}{(\hat{\theta}_{1|1} + \hat{\theta}_{0|0} - 1)}\right]. \tag{5}$$

Example 3 *Abortion example (n = 370).*

Sensitivity $\theta_{1|1} = 13/24$ and specificity $\theta_{0|0} = 23/24$.

Method	Estimate	SE	Wald Interval	Exact Interval
Naive	.1514	.0186	(.1148,.1878)	(.1161,.1920)
Corrected	.2194	.0373	(.1463, .2924)	(.1488, .3007)

2.3.1 Correction using external validation data.

- n_V independent observations (*not involving units from the main study*), on which W and X are both observed.

- *An important assumption is that the misclassification model is the same for this data as the main data; that is, the measurement error model is* **exportable.**

Representation of external validation data

		W 0	W 1	
X	0	n_{V00}	n_{V01}	$n_{V0.}$
	1	n_{V10}	n_{V11}	$n_{V1.}$
		$n_{V.0}$	$n_{V.1}$	n_V

Estimated specificity: $\hat{\theta}_{0|0} = n_{V00}/n_{V0.}$

Estimated sensitivity: $\hat{\theta}_{1|1} = n_{V11}/n_{V1.}$

$$\hat{\pi} = \frac{p_W - (1 - \hat{\theta}_{00})}{\hat{\theta}_{00} + \hat{\theta}_{11} - 1}.$$

This is the Maximum Likelihood Estimate (MLE) of π if $0 \leq \hat{\pi} \leq 1$.

An "exact" approach: Get confidence intervals for each π_W, $\theta_{0|0}$ and $\theta_{1|1}$ and get a confidence set for π by using the minimum and maximum value of π computed over π_W, $\theta_{0|0}$ and $\theta_{1|1}$ ranging over their intervals. If each of the intervals has confidence coefficient $1-\alpha^*$, the confidence coefficient for the interval for π is $\geq (1-\alpha^*)^3$.

Delta Method/Approximate normal based interval:

$$\hat{\pi} = Z_1/Z_2, \text{ where } Z_1 = p_W - (1-\hat{\theta}_{00}),\ Z_2 = \hat{\theta}_{00} + \hat{\theta}_{11} - 1$$

Account for uncertainty in estimates of sensitivity and specificity to get the approximate variance

$$V(\hat{\pi}) \approx \frac{1}{(\theta_{0|0}+\theta_{1|1}-1)^2}\left[V(Z_1) - 2\pi cov(Z_1, Z_2) + \pi^2 V(Z_2)\right]$$

$$V(Z_1) = V(p_W) + V(\hat{\theta}_{00}) = \frac{\pi_W(1-\pi_W)}{n} + \frac{\theta_{0|0}(1-\theta_{0|0})}{n_{V0.}}$$

$$V(Z_2) = V(\hat{\theta}_{00}+\hat{\theta}_{11}) = \frac{\theta_{0|0}(1-\theta_{0|0})}{n_{V0.}} + \frac{\theta_{1|1}(1-\theta_{1|1})}{n_{V1.}}$$

$$Cov(Z_1, Z_2) = V(\hat{\theta}_{00}) = \frac{\theta_{0|0}(1-\theta_{0|0})}{n_{V0.}}$$

Get estimated variance $\hat{\sigma}^2_{\hat{\pi}}$, by replacing parameters in $\sigma^2_{\hat{\pi}}$ by estimates and use approximate confidence interval $\hat{\pi} \pm z_{\alpha/2}\hat{\sigma}_{\hat{\pi}}$.

Fieller Method: Use Fieller's approach for getting confidence set (usually an interval) for a ratio.

Bootstrapping.

For the bth bootstrap sample, $b = 1, \dots, B$, where B is large:

1. Generate $p_{wb} = T_b/n$, $\hat{\theta}_{0|0b} = n_{V00b}/n_{V0.}$ and $\hat{\theta}_{1|1b} = n_{V11b}/n_{V1.}$, where T_b, n_{V00b} and n_{V11b} are generated independently as $\text{Bin}(n, p_w)$, $\text{Bin}(n_{V0.}, \hat{\theta}_{0|0})$ and Bin $(n_{V1.}, \hat{\theta}_{1|1})$, respectively.

2. Use the generated quantities to obtain $\widehat{\pi}_b = (p_{wb} - (1 - \hat{\theta}_{0|0b}))/(\hat{\theta}_{0|0b} + \hat{\theta}_{1|1b} - 1)$ (truncated to 0 or 1 if necessary).

3. Use the B bootstrap values $\widehat{\pi}_1, \dots, \widehat{\pi}_B$ to obtain bootstrap estimates of bias and standard error for $\widehat{\pi}$, and to compute a bootstrap confidence interval. Here, simple bootstrap percentile intervals will be used.

Estimation of prevalence in HIV example. Cor-K refers to corrected estimates and intervals treating the misclassification rates as known, while Cor-U accounts for uncertainty in the estimated misclassification rates. Boot indicates the bootstrap analysis. Bootstrap means are .0369 and -.0274 for 4 and 22 false positives, respectively. Confidence intervals are 95% Wald intervals except for the bootstrap where they are percentile intervals.

Method	Est.	SE	CI	Fieller	Exact
Naive	.049	.0033	(.043, .055)		(.043, .056)
4 false positives: $\widehat{\theta}_{1\|1} = .977$, $\widehat{\theta}_{0\|0} = .986$					
Cor-K	.0369	.0033	(.030, .043)		(.030, .044)
Cor-U	.0369	.0075	(.022,.052)	(.022, .052)	(.002, .047)
Boot		.0075	(.021, .050)		
22 false positives: $\widehat{\theta}_{1\|1} = .977$, $\widehat{\theta}_{0\|0} = .926$					
Cor-K	-.0278	.0036	(-.035, -.021)		(-.035, -.020)
Cor-U	-.0278	.0177	(-.062, .007)	(-.064, .006)	(-.098, .004)
Boot		.0179	(-.065, .006)		

2.3.2 Correcting using internal validation data

Observe W_i on all n units. Also observe X_i on random sample of size n_V.

Representation of main study and internal validation data.

		W		
		0	1	
X =	0	n_{V00}	n_{V01}	$n_{V0.}$
	1	n_{V10}	n_{V11}	$n_{V1.}$
		$n_{V.0}$	$n_{V.1}$	n_V
	?	n_{I0}	n_{I1}	n_I
		$n_{.0}$	$n_{.1}$	n

Could mimic what was done with external validation data, using estimated misclassification rates, but this is inefficient. More efficient approach is to use the estimated reclassification rates

$$\hat{\lambda}_{1|1} = n_{V11}/n_{V.1} \text{ and } \hat{\lambda}_{0|0} = n_{V00}/n_{V.0}$$

and

$$\widehat{\pi}_r = p_W(\hat{\lambda}_{1|1} + \hat{\lambda}_{0|0} - 1) + 1 - \hat{\lambda}_{0|0}.$$

This is the maximum likelihood estimator (MLE) as long as $0 < \widehat{\pi}_r < 1$.

$$SE(\widehat{\pi}_r) = \left[p_W^2 \widehat{V}(\hat{\lambda}_{1|1}) + (p_W - 1)^2 \widehat{V}(\hat{\lambda}_{0|0}) + (\hat{\lambda}_{0|0} + \hat{\lambda}_{1|1} - 1)^2 \widehat{V}(p_w)\right]^{1/2} \tag{6}$$

where $\widehat{V}(\hat{\lambda}_{1|1}) = \hat{\lambda}_{1|1}(1-\hat{\lambda}_{1|1})/n_{V.1}$, $\widehat{V}(\hat{\lambda}_{0|0}) = \hat{\lambda}_{0|0}(1-\hat{\lambda}_{0|0})/n_{V.0}$, and $\widehat{V}(p_W) = p_W(1 - p_W)/n$. A Wald confidence interval for π is given by $\widehat{\pi}_r \pm z_{\alpha/2} SE(\widehat{\pi}_r)$.

Bootstrapping. With the validation sample chosen as a random sample from the main study units, the bootstrap can be carried out as follows. For the bth bootstrap sample, $b = 1, \ldots, B$:

1. For $i = 1$ to n generate W_{bi} distributed Bernoulli (0 or 1) with $P(W_{bi} = 1) = p_w$.
2. For $i = 1$ to n_V generate X_{bi} distributed Bernoulli, with $P(X_{bi} = 1)$ equal to $1 - \hat{\lambda}_{0|0} = \hat{\lambda}_{1|0}$ if $W_{bi} = 0$, and equal to $\hat{\lambda}_{1|1}$ if $W_{bi} = 1$.
3. Calculate $\widehat{\pi}_{rb} = p_{wb}(\hat{\lambda}_{1|1b} + \hat{\lambda}_{0|0b} - 1) + 1 - \hat{\lambda}_{0|0b}$ which is the estimate of π based on the reclassification rates for the bth bootstrap sample.
4. Use the B bootstrap values to obtain bootstrap inferences.

Exact confidence interval. Using Bonferonni's inequality, form $100(1 - \alpha/3)$ % confidence intervals for π_w, $\lambda_{1|1}$ and $\lambda_{0|0}$ based on the binomial (or modifications of it). Confidence set for π is formed by value of $\pi_W(\lambda_{1|1} + \lambda_{0|0} - 1) + 1 - \lambda_{0|0}$ as the three parameters range over their respective intervals.

Smoking Example: Estimation of proportion of heavy smokers.

The estimated reclassification rates are

$$\hat{\lambda}_{0|0} = 24/30 = .8 \text{ and } \hat{\lambda}_{1|1} = 18/20 = .9$$

Method	Estimate	SE	95% Wald Interval	95% Exact Interval
Naive	.44	.0351	(0.3712, .5088)	(.3680, 0.5118)
Corrected	.508	.0561	(.3980,.6180)	(.0982, .9502)
Bootstrap	.509(mean)	.0571	95% Bootstrap CI (.3973,.6213)	

- The bootstrap estimate of bias (.509 - .508) is small.

- The "exact" corrected interval, which is conservative, is large, and is not recommended given the relatively large sample sizes involved.

Extensions

- Handling two-phase/designed double sampling with internal validation (methods above still good except bootstrap needs modification.)
- Finite population adjustments.
- Multiple measures (rather than validation).

More than two categories.

$M \geq 2$ categories with $\pi_x = P(X \text{ is in category } x)$.

$$\boldsymbol{\pi}' = (\pi_1, \ldots \pi_M)$$

$\pi_M = 1 - \pi_1 - \ldots - \pi_{M-1}$.

Observed W could have a different number of categories but here assume it has M categories also with

$$P(W = w|X = x) = \theta_{w|x}.$$

$\gamma_w = P(W_i = w) = \Sigma_{x=1}^{M} \theta_{w|x}\pi_x$.

With $\boldsymbol{\gamma}' = (\gamma_1, \ldots, \gamma_M)$ then

$$\boldsymbol{\gamma} = \boldsymbol{\Theta}\boldsymbol{\pi} \text{ and } \boldsymbol{\pi} = \boldsymbol{\Lambda}\boldsymbol{\gamma},$$

where $\boldsymbol{\Theta}$ has $\theta_{w|1}, \ldots \theta_{w|M}$ in the wth row.

Similarly, $\boldsymbol{\Lambda}$ is a matrix of reclassification probabilities, using $\lambda_{x|w} = P(X = x|W = w)$.

Vector of naive proportions, $\mathbf{p}_W$, estimate $\boldsymbol{\Theta}\boldsymbol{\pi}$ rather than $\boldsymbol{\pi}$.

Bias = $(\boldsymbol{\Theta} - \mathbf{I})\boldsymbol{\pi}$ where $\mathbf{I}$ is an identity matrix.

If we have estimated misclassification rates from external validation data $\hat{\boldsymbol{\pi}} = \hat{\boldsymbol{\Theta}}^{-1}\mathbf{p}_W$.

With internal validation data the maximum likelihood estimator (assuming all components of $\widehat{\boldsymbol{\pi}}_r$ are between 0 and 1) is $\hat{\boldsymbol{\pi}}_r = \hat{\boldsymbol{\Lambda}}\mathbf{p}$, where $\mathbf{p}$ is the vector of proportions formed using all of the observations.

Can get analytical expressions for the covariance of either $\widehat{\boldsymbol{\pi}}$ or $\widehat{\boldsymbol{\pi}}_r$ using theory at end of Ch. 3. See also Greenland (1988) and Kuha et al. (1998).

Mapping example. To illustrate consider the mapping validation data given in Example 3 and suppose there is separate random sample of an area with 1000 pixels yielding proportions $\mathbf{p}'_W = (.25, .10, .15, .20, .30)$, where .25 is the proportion of pixels that are sensed as water, etc. The combined proportions for W over the main and validation data are $\mathbf{p}' = (.208, .181, .175, .183, .245)$.

From the validation data, the estimated misclassification and reclassification rates are

$$\hat{\boldsymbol{\Theta}} = \begin{bmatrix} 0.961 & 0.005 & 0.010 & 0.008 & 0.016 \\ 0.004 & 0.921 & 0.018 & 0.020 & 0.037 \\ 0.008 & 0.035 & 0.833 & 0.061 & 0.063 \\ 0.033 & 0.014 & 0.072 & 0.817 & 0.064 \\ 0.030 & 0.049 & 0.054 & 0.080 & 0.787 \end{bmatrix}$$

and

$$\hat{\Lambda} = \begin{bmatrix} 0.918 & 0.005 & 0.008 & 0.030 & 0.040 \\ 0.004 & 0.899 & 0.030 & 0.011 & 0.056 \\ 0.010 & 0.020 & 0.831 & 0.066 & 0.073 \\ 0.008 & 0.024 & 0.064 & 0.788 & 0.115 \\ 0.012 & 0.034 & 0.051 & 0.047 & 0.856 \end{bmatrix}.$$

Treating the validation data as external and internal respectively, leads to

$$\widehat{\pi} = \begin{bmatrix} 0.251 \\ 0.087 \\ 0.134 \\ 0.195 \\ 0.337 \end{bmatrix} \quad \text{and} \quad \widehat{\pi}_r = \begin{bmatrix} 0.209 \\ 0.185 \\ 0.181 \\ 0.191 \\ 0.243 \end{bmatrix}.$$

Treating validation data as external. 1000 bootstrap samples.

	Estimate		Bootstrap	
TYPE	Naive	Corrected	SE	90% CI
Water	0.25	0.251	0.015	(0.227, 0.275
BareGround	0.1	0.087	0.011	(0.069, 0.105)
Deciduous	0.15	0.134	0.015	(0.111,0.158)
Coniferous	0.2	0.195	0.017	(0.167, 0.224)
Urban	0.3	0.337	0.021	(0.302, 0.371)

3 Misclassification in two-way tables

Example 1: Antibiotics/SIDS. From Greenland (1988). Case-control study of the association of antibiotic use by the mother during pregnancy, X, and the occurrence of sudden infant death syndrome (SIDS), Y. W = antibiotic use based on a self report from the mother.

Validation study of 428 women. X = from medical records.

Antibiotic use and SIDS example.

MAIN STUDY

		Y	
		Controls(0)	Cases(1)
W	No Use (0)	479	442
	Use (1)	101	122
		580	564

VALIDATION DATA

		Control (Y=0)		Cases (Y=1)	
		X		X	
		0 (no use)	1 (use)	0 (no use)	1 (use)
W =	0 (no use)	168	16	143	17
	1 (use)	12	21	22	29
		180	37	165	46

For controls ($Y = 0$) estimated specificity and sensitivity are .933 and .568

For cases ($Y = 1$), .867 and .630.

Suggests misclassification may be differential.

Accident Example. Partial data from Hochberg (1977). Look at seat belt use and injury, accidents in North Carolina in 1974-1975 with high damage to the car and involving male drivers.

Error prone measures (from police report):

D (1 = injured, 0 = not) and W (0 = no seat belt use, 1 = use).

True values (based on more intensive inquiry):

Y (injury) and X (seat-belt use).

MAIN STUDY (error-prone values)

		D			
		No injury(0)	Injured(1)		Prop. inj.
W	No Seat Belt (0)	17476	6746	2422	.278
	Seat Belt (1)	2155	583	2738	.213

VALIDATION DATA

		D = 0		D = 1	
Y	X	W= 0	W= 1	W = 0	W = 1
0	0	299	4	11	1
0	1	20	30	2	2
1	0	59	1	118	0
1	1	9	6	5	9

Common ecological settings.

X = habitat type or category (may be different types or categories formed from a categorizing a quantitative measure; e.g., low, medium or high level of vegetation cover) or different geographic regions.

Y = outcome; presence or absence of one or many "species", nesting success, categorized population abundance, etc.

3.1 Models for True values

There are three formulations:

- joint model for X and Y

$$\gamma_{xy} = P(X = x, Y = y).$$

		Y		
		0	1	
X	0	γ_{00}	γ_{01}	$\gamma_{0.}$
	1	γ_{10}	γ_{11}	$\gamma_{1.}$
		$\gamma_{.0}$	$\gamma_{.1}$	1

- The conditional model for Y given X, which is specified by

$$\pi_x = P(Y = 1|X = x),$$

		Y	
		0	1
x	0	$1 - \pi_0$	π_0
	1	$1 - \pi_1$	π_1

- The conditional model for X given Y, specified by

$$\alpha_y = P(X = 1|Y = y)$$

		y	
		0	1
X	0	$1 - \alpha_0$	$1 - \alpha_1$
	1	α_0	α_1

Three types of study.

- Random Sample.
- “Prospective” or ”cohort” study, pre-stratified on X.

  Can only estimate π’s and functions of them
- “Case-Control” study, pre-stratified on Y

  Can only estimate α’s and functions of them

$$\text{Relative Risk} = \frac{\pi_1}{\pi_0}$$

$$\text{Odds Ratio} = \Psi = \frac{\pi_1/(1-\pi_1)}{\pi_0/(1-\pi_0)} = \frac{\pi_1(1-\pi_0)}{\pi_0(1-\pi_1)}.$$

OR $\approx$ R.R. if π_0 and π_1 are small. (These can be defined in terms of γ’s if have a random sample.

$$\frac{\pi_1(1-\pi_0)}{\pi_0(1-\pi_1)} = \frac{\alpha_1(1-\alpha_0)}{\alpha_0(1-\alpha_1)} = \frac{\gamma_{11}\gamma_{00}}{\gamma_{01}\gamma_{10}}.$$

Can estimate the odds ratio from a case-control study.

$$X \text{ is independent of } Y \Leftrightarrow \pi_0 = \pi_1 \Leftrightarrow \alpha_0 = \alpha_1 \Leftrightarrow OR = 1.$$

Tested via Pearson chi-square, likelihood ratio or Fisher’s exact test.

Observed main study data

		D		
		0	1	
W	0	n_{00}	n_{01}	$n_{0.}$
	1	n_{10}	n_{11}	$n_{1.}$
		$n_{.0}$	$n_{.1}$	n

Observed Proportions from naive 2 × 2 table.

Proportion	Quantity	Naive Estimate of:
Overall	$p_{wd} = n_{wd}/n$	γ_{wd}
Row	$p_w = n_{w1}/n_{w.}$	π_w
Column	$a_d = n_{1d}/n_{.d}$	α_d

The naive estimator of the odds ratio

$$\hat{\Psi}_{naive} = \frac{a_1(1-a_0)}{a_0(1-a_1)} = \frac{n_{11}n_{00}}{n_{10}n_{01}} = \frac{p_1(1-p_0)}{p_0(1-p_1)}$$

Other naive estimators are similarly defined.

There are a huge number of cases based on combining

- Sampling design

- Whether one or both of the variables are misclassified

- Whether validation data is internal or external

- If internal, is a random sample or a two-phase design used.

3.2 Models and biases

Error in one variable (X) with other perfectly measured. With change in notation this handles error in Y only (e.g., replaces α_j with π_j). This will illustrate the main points.

Observe W instead of X. General misclassification probabilities are given by

$$\theta_{w|xy} = P(W = w|X = x, Y = y).$$

Differential M. Error: $\theta_{w|xy}$ depends on y.

Nondifferential M. Error: $\theta_{w|xy} = \theta_{w|x}$ (doesn't depend on y.)

$$\theta_{1|1y} = \textbf{sensitivity} \text{ at } y. \qquad \theta_{0|0y} = \textbf{specificity} \text{ at } y.$$

Biases of naive estimators: RS or stratified on y.

a_y (proportion of data with $Y = y$ at $W = 1$): naive estimator of $\alpha_y = P(X = 1|Y = y)$. $E(a_y) = \mu_y$, where

$$\mu_0 = \theta_{1|00} + \alpha_0(\theta_{1|10} + \theta_{0|00} - 1) \text{ and } \mu_1 = \theta_{1|01} + \alpha_1(\theta_{1|11} + \theta_{0|01} - 1).$$

With nondifferential misclassification

$$\mu_0 = \theta_{1|0} + \alpha_0(\theta_{1|1} + \theta_{0|0} - 1) \text{ and } \mu_1 = \theta_{1|0} + \alpha_1(\theta_{1|1} + \theta_{0|0} - 1)$$

$\mu_1 - \mu_0 = (\alpha_1 - \alpha_0)(\theta_{1|1} + \theta_{0|0} - 1)$.

$\mu_1 = \mu_0$ is equivalent to $\alpha_0 = \alpha_1$ (independence) under non-differential misclassification.

- Naive estimator of $\alpha_1 - \alpha_0$ estimates $\mu_1 - \mu_0$.

- *With nondifferential misclassification in one variable and no misclassification in the other, naive tests of independence are valid in that they have the correct size (approximately). With differential misclassification these tests are not valid.*

Bias in naive estimators of odd-ratio.

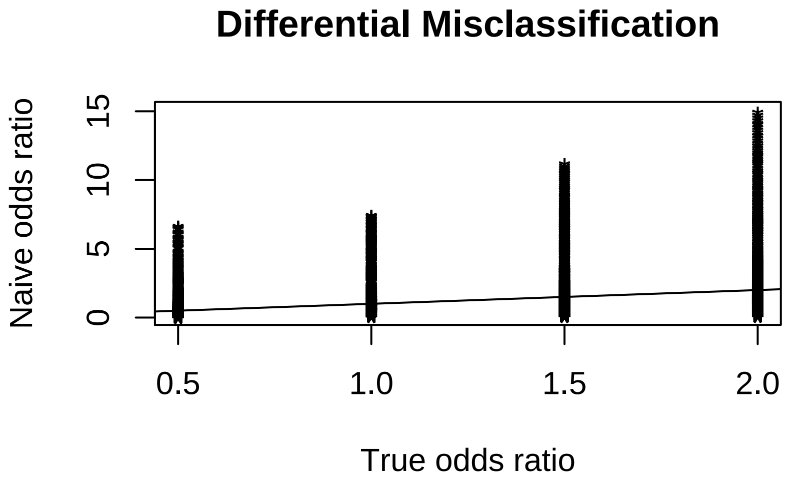
The naive estimator of the odd ratio, $\hat{\Psi}_{naive} = a_1(1-a_0)/a_0(1-a_1)$, is a consistent estimator of

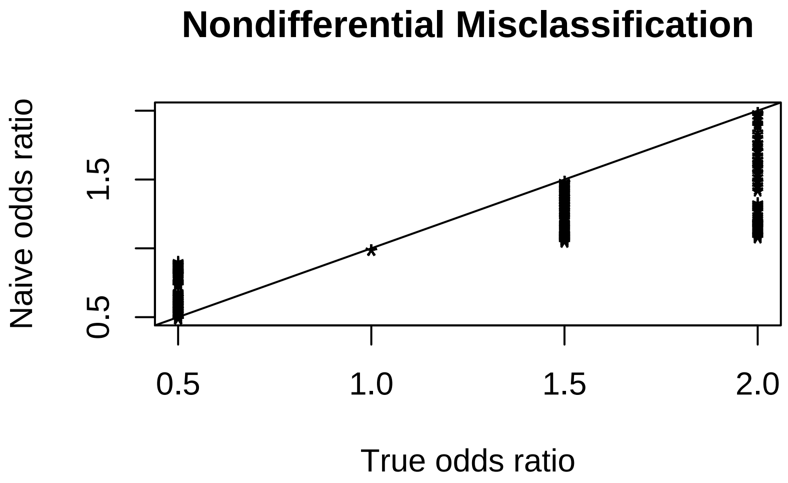
$$\Psi_{naive} = \frac{\mu_1(1-\mu_0)}{\mu_0(1-\mu_1)}.$$

- The bias is difficult to characterize. It may go in either direction, either over or under estimating the odds-ratio (on average), when the misclassification is differential.

- With nondifferential misclassification as long as $\theta_{1|1} + \theta_{0|0} - 1 > 0$ (which will almost always hold) the naive estimator is biased towards the value of 1 in that either $1 \leq \Psi_{naive} \leq \Psi$ or $\Psi \leq \Psi_{naive} \leq 1$.(Gustafson (2004, Section 3.3))

- Bias results are asymptotic/approximate. May not work for small samples. No guarantee of direction in a specific sample (see Jurek et al. (2005)).





Plot of limiting value of naive estimator.

- odds ratio $\Psi = .5, 1, 1.5$ or 2.
- $\alpha_1 = .05$ to .95 by .3 (α_0 to yield the desired Ψ).
- Sensitivity and specificity ranged from .8 to 1 by .5. Taken equal for non-differential cases.

Error in both variables.

The general misclassification model is

$$\theta_{wd|xy} = P(W = w, D = d|X = x, Y = y).$$

If misclassification is both nondifferential and independent, $P(W = w, D = d|x, y) = \theta_{wd|xy} = \theta_{w|x}\theta^*_{d|y}$, with
$\theta_{w|x} = P(W = w|X = x)$ and $\theta^*_{d|y} = P(D = d|Y = y)$.

For a **random sample**, it is relatively easy to address bias in the naive estimators. The naive estimate of γ_{wv} is p_{wd} with

$$E(p_{wd}) = \mu_{wd} = \sum_x \sum_y \theta_{wd|xy}\gamma_{xy}. \tag{7}$$

This yields the bias in p_{wd} as an estimator of γ_{wd} and can be used to determine biases (sometimes exact, often approximate) for other quantities in the same manner as in the previous sections.

For example, the approximate bias in the naive estimate of $\pi_1 - \pi_0$ is

$$\frac{\mu_{11}}{\mu_{1.}} - \frac{\mu_{01}}{\mu_{0.}} - (\pi_1 - \pi_0)$$

and for the odd-ratio is

$$\frac{\mu_{11}\mu_{00}}{\mu_{01}\mu_{10}} - \frac{\gamma_{11}\gamma_{00}}{\gamma_{01}\gamma_{10}}.$$

Bias expressions are complicated. No general characterizations of the bias.

More complicated expressions and derivations if stratify on a one of the misclassified variables.

3.3 Correcting using external validation data.

The general approach using external data.

- $\mathbf{p}$ = vector of proportions from main study (either cell proportions or conditional proportions.)
- $\boldsymbol{\phi}$ = parameters of interest (what $\mathbf{p}$ naively estimates.)

  E.g., $\mathbf{p}' = (p_{00}, p_{01}, p_{10}, p_{11})$ and $\boldsymbol{\phi}' = (\gamma_{00}, \gamma_{01}, \gamma_{10}, \gamma_{11})$.
- Find $E(\mathbf{p}) = \mathbf{b} + \mathbf{B}\boldsymbol{\phi}$, where $\mathbf{b}$ and $\mathbf{B}$ are functions of the misclassification rates. In some cases $\mathbf{b}$ is $\mathbf{0}$.
- Estimate misclassification rates, and hence, $\mathbf{b}$ and $\mathbf{B}$, using external validation data.
- Get corrected estimate

  $$\hat{\boldsymbol{\phi}} = \hat{\mathbf{B}}^{-1}(\mathbf{p} - \hat{\mathbf{b}}).$$

  This feeds into other estimates (e.g., odds ratios.)
- Getting an expressions for $Cov(\hat{\boldsymbol{\phi}})$ is "straightforward" in principle, but can be tedious. There are exact expression are available for some special cases.
- Bootstrapping: Resample main study data and external validation (similar to with one proportion).

Back to misclassification in X only.

External Validation data at $Y = y$.

		W 0	1	
X =	0	$n_{V00(y)}$	$n_{V01(y)}$	$n_{V0.(y)}$
	1	$n_{V10(y)}$	$n_{V11(y)}$	$n_{V1.(y)}$
		$n_{V.0(y)}$	$n_{V.1(y)}$	$n_{V(y)}$

$$\widehat{\theta}_{1|1y} = n_{V11(y)}/n_{V1.(y)} \text{ (estimated sensitivity at} Y = y)$$

$$\widehat{\theta}_{0|0y} = n_{V00(y)}/n_{V0.(y)} \text{ (estimated specificity at} Y = y),$$

$\widehat{\theta}_{0|1y} = 1 - \widehat{\theta}_{1|1y}$ and $\widehat{\theta}_{1|0y} = 1 - \widehat{\theta}_{0|0y}$

With non-differential misclassification, have $\widehat{\theta}_{1|1y} = \widehat{\theta}_{1|1}$, etc.

Suppose the design is either a random sample or case-control with α's or the odds-ratio of primary interest.

Recall: With a change of notation this covers the case of misclassification of a response over perfectly measured categories (for RS or prospective study).

$$\hat{\alpha}_0 = (a_0 - \hat{\theta}_{1|00})/(\hat{\theta}_{1|10} - \hat{\theta}_{1|00}) \quad \text{and} \quad \hat{\alpha}_1 = (a_1 - \hat{\theta}_{1|01})/(\hat{\theta}_{1|11} - \hat{\theta}_{1|01})$$

For nondifferential misclassification, replace $\widehat{\theta}_{1|10}$ with $\widehat{\theta}_{1|1}$, etc.

- Differential misclassification:

  $cov(\widehat{\alpha}_0, \widehat{\alpha}_1) = 0$

$$V(\hat{\alpha}_0) \approx \frac{1}{\Delta_0^2}\left[V(a_0) + (1-\alpha_0)^2 V(\hat{\theta}_{1|00}) + \alpha_0^2 V(\hat{\theta}_{1|10})\right]$$

$$V(\hat{\alpha}_1) \approx \frac{1}{\Delta_1^2}\left[V(a_1) + (1-\alpha_1)^2 V(\hat{\theta}_{1|01}) + \alpha_1^2 V(\hat{\theta}_{1|11})\right]$$

  where $V(a_y) = \mu_y(1-\mu_y)/n_{.y}$, $V(\hat{\theta}_{1|xy}) = \theta_{1|xy}(1-\theta_{1|xy})/n_{Vx.(y)}$,

  $\Delta_0 = \theta_{1|10} - \theta_{1|00}$, and $\Delta_1 = \theta_{1|11} - \theta_{1|01}$.

- Non-differential

  $\Delta_0 = \Delta_1 = \Delta = \theta_{1|1} - \theta_{1|0}$, while $n_{Vx.(y)}$ is replaced by $n_{Vx.}$.

$$cov(\hat{\alpha}_0, \hat{\alpha}_1)) \approx \frac{1}{\Delta^2}\left[(1-\alpha_0)(1-\alpha_1)V(\hat{\theta}_{1|0}) + \alpha_0\alpha_1 V(\hat{\theta}_{1|1})\right].$$

$$V(\widehat{\alpha}_0 - \widehat{\alpha}_1)) = V(\hat{\alpha}_0) + V(\hat{\alpha}_1) - 2cov(\hat{\alpha}_0, \hat{\alpha}_1).$$

The estimate of $L = log(OR)$ is

$$\hat{L} = log(\hat{\alpha}_1) + log(1-\hat{\alpha}_0) - log(\hat{\alpha}_0) - log(1-\hat{\alpha}_1).$$

(Common to first get a confidence interval for L and then exponentiate to get CI for odds ratio.)

$$V(\hat{L}) \approx \frac{V(\hat{\alpha}_0)}{\alpha_0^2(1-\alpha_0)^2} + \frac{V(\hat{\alpha}_1)}{\alpha_1^2(1-\alpha_1)^2} - 2\frac{cov(\hat{\alpha}_0, \hat{\alpha}_1)}{\alpha_0(1-\alpha_0)\alpha_1(1-\alpha_1)}. \qquad (8)$$

SIDS/Antibiotic Example. Case/control study.

$$\hat{\theta}_{0|00} = .933 = \text{estimated specificity at } y = 0,$$
$$\hat{\theta}_{1|10} = .568 = \text{estimated sensitivity at } y = 0,$$
$$\hat{\theta}_{0|01} = .867 = \text{estimated specificity at } y = 1, \text{ and}$$
$$\hat{\theta}_{1|11} = .630 = \text{estimated sensitivity at } y = 1.$$

Differential Misclassification

Quantity	Naive	Corrected	SE	Wald CI	Boot SE	Boot CI
α_0	.174	.215				
α_1	.216	.167				
Odds ratio	1.309	.7335	.403	(.25,2.15)	.532	(.167,2.17)

nondifferential Misclassification

Quantity	Naive	Corrected	SE	Wald CI	Boot SE	Boot CI
α_0	.174	.150				
α_1	.216	.234				
Odds ratio	1.309	1.73	.555	(.92,3.24)	.222	(.963, 4.00)

- The assumption of nondifferential misclassification is questionable.
- As Greenland (1988) illustrated and discussed, allowing differential misclassification produces very different results.
- Under differential misclassification the bootstrap mean is .8260 leading to a bootstrap estimate of bias in the corrected estimator of .8260 - .7335 = .0925.
- The calculations for the example were calculated using SAS-IML. Certain aspects of the analysis can also be carried out using the GLLAMM procedure in STATA; see Skrondal and Rabe-Hesketh (2004).

External data with both Misclassified.

3.4 Error in X and Y both

$$E\begin{bmatrix} p_{00} \\ p_{01} \\ p_{10} \\ p_{11} \end{bmatrix} = \begin{bmatrix} \theta_{00|00} & \theta_{00|01} & \theta_{00|10} & \theta_{00|11} \\ \theta_{01|00} & \theta_{01|01} & \theta_{01|10} & \theta_{01|11} \\ \theta_{10|00} & \theta_{10|01} & \theta_{10|10} & \theta_{10|11} \\ \theta_{11|00} & \theta_{11|01} & \theta_{11|10} & \theta_{11|11} \end{bmatrix} \begin{bmatrix} \gamma_{00} \\ \gamma_{01} \\ \gamma_{10} \\ \gamma_{11} \end{bmatrix}.$$

In the most general case there are 12 distinct misclassification rates involved, since each column of $\mathbf{B}$ sums to 1.

$$\hat{\gamma} = \hat{\mathbf{B}}^{-1}\mathbf{p}. \tag{9}$$

$Cov(\widehat{\boldsymbol{\theta}})$ and $Cov(\mathbf{p})$ can be obtained using multinomial results. These can be used along with multivariate delta method to obtain an approximate expression for $Cov(\widehat{\boldsymbol{\gamma}})$ and to obtain approximate variances for the estimated π's, their difference, the odds ratio, etc. Example below only utilizes the bootstrap for obtaining standard errors and confidence intervals.

Accident Example. Treat validation data as external, with both variables measured with error.

For bootstrapping, first a vector of main sample counts is generated using a multinomial with sample size $n_I = 26960$ and a vector of probabilities $\mathbf{p} = (0.6482, .2502, .0799, .0216)$ based on the observe cell proportions. The validation data is generated by generating a set of counts for each (x, y) combination using a multinomial with sample size equal to n_{Vxy}, the number of observations in validation data with $X = x$ and $Y = y$, and probabilities given by the estimated reclassification rates associated with that (x, y) combination.

π_0 and π_1 are probability of injury without and with seat belts, respectively.

	Estimates		Bootstrap Analysis			
Parameter	Naive	Cor.	Mean	Median	SE	90% CI
γ_{00}	.648	.508	0.505	0.507	0.0278	(0.460,0.547)
γ_{01}	.250	.331	0.329	0.329	0.027	(0.289,0.369)
γ_{10}	.080	.110	0.108	0.108	0.025	(0.071, 0.148)
γ_{11}	.022	.051	0.060	0.053	0.063	(0.026,0.104)
π_0	.279	.395	0.394	0.392	0.031	(0.350, 0.445)
π_1	.213	.319	0.344	0.333	0.135	(0.160, 0.571)
$\pi_1 - \pi_0$	-.066	-.076	-0.050	-0.070	0.147	(-0.252, 0.198)

NAIVE: $\widehat{\pi}_0 = .2785$, $\widehat{\pi}_1 = .2129$, estimated difference = -.0656, SE = .0083, 90% confidence interval of $(-0.0793, -0.0518)$.

3.5 Correcting using internal validation data.

General Strategies

- Use the validation data to estimate the misclassification rates then correct the naive estimators as above for external data. Sometimes referred to as the **matrix method.**

- Weighted average approach (assuming a random subsample). Use true values from validation data to get estimates in usual fashion and to estimate misclassification rates. Then correct estimates from part of main study that is not validated (as above using external validation) and use a weighted average.

- Correct using estimated reclassification rates. This usually produces the MLEs and is more efficient. Can be used with designed two-phase studies. Sometimes called the **inverse matrix method** (ironically).

Illustrate last approach here with overall random sample. (Same developments work for other cases with redefinition of $\boldsymbol{\gamma}$, $\boldsymbol{\Lambda}$, $\boldsymbol{\mu}$ and $\mathbf{p}$)

Error in X only: $P(X = x|W = w, Y = y) = \lambda_{x|wy}$

$$\gamma_{xy} = P(X = x, Y = y) = \sum_w \lambda_{x|wy}\mu_{wy}, \ \mu_{wy} = P(W = w, Y = y).$$

$$\boldsymbol{\gamma} = \begin{bmatrix} \gamma_{00} \\ \gamma_{10} \\ \gamma_{01} \\ \gamma_{11} \end{bmatrix} = \begin{bmatrix} \lambda_{0|00} & \lambda_{0|10} & 0 & 0 \\ \lambda_{1|00} & \lambda_{1|10} & 0 & 0 \\ 0 & 0 & \lambda_{0|01} & \lambda_{0|11} \\ 0 & 0 & \lambda_{1|01} & \lambda_{1|11} \end{bmatrix} \begin{bmatrix} \mu_{00} \\ \mu_{10} \\ \mu_{01} \\ \mu_{11} \end{bmatrix} = \boldsymbol{\Lambda\mu}.$$

Error in X and Y both: $P(X = x, Y = y|W = w, D = d) = \lambda_{xy|wd}$

$$\boldsymbol{\gamma} = \begin{bmatrix} \gamma_{00} \\ \gamma_{01} \\ \gamma_{10} \\ \gamma_{11} \end{bmatrix} = \begin{bmatrix} \lambda_{00|00} & \lambda_{00|01} & \lambda_{00|10} & \lambda_{00|11} \\ \lambda_{01|00} & \lambda_{01|01} & \lambda_{01|10} & \lambda_{01|11} \\ \lambda_{10|00} & \lambda_{10|01} & \lambda_{10|10} & \lambda_{10|11} \\ \lambda_{11|00} & \lambda_{11|01} & \lambda_{11|10} & \lambda_{11|11} \end{bmatrix} \begin{bmatrix} \mu_{00} \\ \mu_{01} \\ \mu_{10} \\ \mu_{11} \end{bmatrix} = \boldsymbol{\Lambda\mu}.$$

$\mu_{wd} = P(W = w, D = d)$ and $\lambda_{wd|xy} = P(X = x, Y = y|W = w, D = d)$.

$E(\mathbf{p}) = \boldsymbol{\mu}$ ($\mathbf{p}$ contains cell proportions.)

$$\hat{\gamma} = \hat{\boldsymbol{\Lambda}}\mathbf{p}.$$

- $Cov(\widehat{\gamma})$ obtained using the delta method in combination with $Cov(\mathbf{p})$ and the variance-covariance structure of the $\widehat{\gamma}$'s (from binomial and multinomial results.)

$$Cov(\hat{\gamma}) = \Sigma_{\widehat{\gamma}} \approx \mathbf{A} + \Lambda Cov(\mathbf{p})\Lambda'$$

$\mathbf{A}$ is a 4×4 matrix with $(j,k)th$ element equal to $\boldsymbol{\mu}' C_{jk} \boldsymbol{\mu}$, with $C_{jk} = cov(\mathbf{Z}_j, \mathbf{Z}_k)$, $\mathbf{Z}_j'$ =e jth row of $\widehat{\Lambda}$.

Accident Example with internal validation data. Treated as overall random sample with both variables misclassified.

$n_I = 26969$ cases with W and D only.

$n_V = 576$ validated cases.

$$\hat{\Lambda} = \begin{bmatrix} 0.7726098 & 0.0808824 & 0.097561 & 0.0833333 \\ 0.1524548 & 0.8676471 & 0.0243902 & 0 \\ 0.0516796 & 0.0147059 & 0.7317073 & 0.1666667 \\ 0.0232558 & 0.0367647 & 0.1463415 & 0.75 \end{bmatrix}$$

Naive analysis: $\mathbf{p} = (0.6487, .249, .0796, .0216)$, $\widehat{\pi}_0 = .2781$, $\widehat{\pi}_1 = .2132$, estimated difference of -.0649 (SE = .0083)

Corrected estimate and bootstrap analysis of accident data with internal validation and both variables misclassified.

		Bootstrap		
Variable	Estimate	Mean	SE	90% CI
γ_{00}	.5310	.5310	0.5819	(0.5049,0.5573)
γ_{01}	.3177	.3175	0.3682	(0.2949, 0.3411)
γ_{10}	.0994	.0994	0.1303	(0.0836, 0.1148)
γ_{11}	.0520	.0520	0.0799	(0.0395, 0.0651)
π_0	.3743	.3742	0.4267	(0.3481, 0.4001)
π_1	.3447	.3433	0.4883	(0.2712, 0.4181)
$\pi_1 - \pi_0$	-.0297	-.0309	0.1121	(-0.1070, 0.0489)

4 Simple Linear Regression with additive error

Look at simple linear case first to get main ideas across in simple fashion. Then move to multiple regression.

$$Y_i|x_i = \beta_0 + \beta_1 x_i + \epsilon_i$$

$E(\epsilon_i) = 0$, $V(\epsilon_i) = \sigma^2$.

The ϵ_i (referred to as **error in the equation**) are uncorrelated,

With random X_i there may be interest in the correlation

$$\rho = \sigma_{XY}/\sigma_X\sigma_Y = \beta_1(\sigma_X/\sigma_Y).$$

$D =$ observed value of Y (this is equal to Y if no error in response)

$W =$ measured value of X.

Examples

- Soil nitrogen/corn yield example from Fuller (1987).

  Nitrogen content (X) estimated through sampling. Measurement error variance treated as known. Y treated as measured without error.

- Gypsy moth egg mass concentrations and defoliation.

Model is defined in terms of values on sixty hectare units. Very costly to get exact values. Subsample m subplots that are .01 hectare in size.

w = estimated gypsy moth egg mass (w)

d = estimated defoliation (percent)

FOREST	m_i	d_i	$\hat{\sigma}_{qi}$	w_i	$\hat{\sigma}_{ui}$	$\hat{\sigma}_{uqi}$
GW	20	45.50	6.13	69.50	12.47	13.20
GW	20	100.00	0.00	116.30	28.28	0.00
GW	20	25.00	2.67	13.85	3.64	1.64
GW	20	97.50	1.23	104.95	19.54	0.76
GW	20	48.00	2.77	39.45	7.95	-2.74
GW	20	83.50	4.66	29.60	6.47	12.23
MD	15	42.00	8.12	29.47	7.16	-5.07
MD	15	80.67	7.27	41.07	9.57	18.66
MD	15	35.33	7.92	18.20	3.88	18.07
MD	15	14.67	1.65	15.80	4.32	-2.07
MD	15	16.00	1.90	54.99	26.28	25.67
MD	15	18.67	1.92	54.20	12.98	-1.31
MD	15	13.33	2.32	21.13	5.40	-3.17
MM	10	30.50	5.65	72.00	26.53	18.22
MM	10	6.00	0.67	24.00	8.84	1.78
MM	10	7.00	0.82	56.00	14.85	-5.78
MM	10	14.00	4.46	12.00	6.11	12.44
MM	10	11.50	2.48	44.00	28.25	8.22

- **Functional case:** x_i's treated as fixed values
- **Structural case:** X_i's are random and x_i = realized value.

  Usually assume $X_1, \ldots, X_n$ independent with $E(X_i) = \mu_X$, $V(X_i) = \sigma_X^2$ but can be relaxed.

Naive estimators

$$\hat{\beta}_{1naive} = S_{WD}/S_{WW}, \quad \hat{\beta}_{0naive} = \bar{D} - \hat{\beta}_{1naive}\bar{W},$$

$$\hat{\sigma}^2_{naive} = \sum_i (D_i - (\hat{\beta}_{0naive} + \hat{\beta}_{1naive}W_i))^2/(n-2),$$

$\overline{W} = \Sigma_i W_i/n$, $\overline{D} = \Sigma_i D_i/n$,

$$S_{WD} = \frac{\Sigma_i(W_i - \bar{W})(D_i - \overline{D})}{n-1} \text{ and } S_{WW} = \frac{\Sigma_i(W_i - \bar{W})^2}{n-1}.$$

4.1 The additive Berkson model and consequences

The "error-prone value", w_i, is fixed value while the true value, X_i, is random. This occurs when w_i is a target "dose" (dose, temperature, speed). There are numerous nitrogen intake/nitrogen balance studies used to determine nutritional requirements. Target intake is w.

$$X_i = w_i + e_i, \quad E(e_i) = 0, \text{ and } V(e_i) = \sigma_e^2.$$

If error in y then $D_i = y_i + q_i$ with $E(q_i) = 0$ and $V(q_i) = \sigma_q^2$.

Reasonable to assumed q_i is independent of e_i.

$$D_i = \beta_0 + \beta_1 w_i + \eta_i,$$

where $\eta_i = \beta_1 e_i + \epsilon_i + q_i$ with

$$E(\eta_i) = 0, \quad V(\eta_i) = \beta_1^2 \sigma_e^2 + \sigma^2 + \sigma_q^2.$$

Key point: with w_i fixed, η_i is uncorrelated with w_i (since any random variable is uncorrelated with a constant)

For the additive Berkson model, naive inferences for the coefficients, based on regressing D on w, are correct. If there is no error in Y, naive predictions of Y from w, including prediction intervals, are correct. If there is measurement error in Y then the naive prediction intervals for Y from w are not correct.

- This result depends on both the additivity of the error and the linear model.

If $X_i = \lambda_0 + \lambda_1 w_i + \delta_i$, the naive estimate of the slope estimates $\beta_1 \lambda_1$ rather than β_1.

4.1.1 Additive Error and consequences

Given x_i *and* y_i,

$$D_i = y_i + q_i \qquad \text{and} \qquad W_i = x_i + u_i$$

$E(u_i|x_i) = 0$, $E(q_i|y_i) = 0$,

$$V(u_i|x_i) = \sigma^2_{ui}, \;\; V(q_i|y_i) = \sigma^2_{qi} \;\; \text{and} \;\; Cov(u_i, q_i|x_i, y_i) = \sigma_{uqi}.$$

$q_i =$ measurement error in D_i as an estimate of y_i

$u_i =$ measurement error in W_i as an estimate of x_i.

(u_i, q_i) independent over i, uncorrelated with the ϵ_i.

- Either of the variables may be observed exactly in which case the appropriate measurement error is set to 0, as is the corresponding variance and the covariance.

- The most general form here allows a heteroscedastic measurement error model in which the (conditional) measurement error variances, or covariance if applicable, may change with i. This can arise for a number of reasons, including unequal sampling effort on a unit or the fact that the variance may be related to the true value.

  If X is random then it may be necessary to carefully distinguish the conditional behavior of u given $X = x$, from the behavior of the "unconditional measurement error" $W - X$. This is a subtle point that we'll put aside here.

- Uncorrelatedness of (u_i, q_i) with ϵ_i is weaker than independence. It is implied just by the assumption that the conditional means are 0! The variances could depend on the true values (a type of dependence.)

Constant ME variances/covariance:

$\sigma^2_{ui} = \sigma^2_u, \quad \sigma^2_{qi} = \sigma^2_q, \quad \sigma^2_{uqi} = \sigma_{uq}$.

The behavior of naive analyses under additive measurement error.

Why should there be any trouble?

With x_i fixed,

$$D_i = \beta_0 + \beta_1 W_i + \epsilon^*_i, \qquad \epsilon^*_i = -\beta_1 u_i + \epsilon_i + q_i, \tag{10}$$

ϵ^*_i is correlated with W_i since $Cov(W_i, \epsilon^*_i) = \sigma_{uqi} - \beta_1 \sigma^2_{ui}$. This violates one of the standard regression assumptions. Can also show $E(\epsilon^*_i | W_i = w_i)$ is *not* 0 (unless $\beta_1 = 0$).

The naive estimates of the regression coefficients or the error variance are typically biased. The direction and magnitude of bias depends on a variety of things. If there is error in both variables, the correlation in the measurement errors plays a role.

Approximate/Asymptotic Biases:

$$E(\hat{\beta}_{0naive}) \approx \gamma_0,\ E(\hat{\beta}_{1naive}) \approx \gamma_1 \text{ and } E(\hat{\sigma}^2_{naive}) \approx \sigma^2_\delta.$$

$$\gamma_0 = \beta_0 + \frac{\mu_X}{\sigma^2_X + \sigma^2_u}(\beta_1\sigma^2_u - \sigma_{uq})$$

$$\gamma_1 = \left[\frac{\sigma^2_X}{\sigma^2_X + \sigma^2_u}\right]\beta_1 + \frac{\sigma_{uq}}{\sigma^2_X + \sigma^2_u} = \beta_1 - \left[\frac{\sigma^2_u}{\sigma^2_X + \sigma^2_u}\right]\beta_1 + \frac{\sigma_{uq}}{\sigma^2_X + \sigma^2_u}$$

$$\sigma^2_\delta = \sigma^2 + \sigma^2_q + \beta_1^2\sigma^2_x - (\beta_1\sigma^2_x + \sigma_{uq})^2/(\sigma^2_x + \sigma^2_u).$$

$$\mu_X = \sum_i E(X_i)/n, \quad \sigma^2_X = E(S_{XX}),$$

$S_{XX} = \Sigma_i(X_i - \overline{X})^2/(n-1)$.

$$\sigma^2_u = \sum_{i=1}^n \sigma^2_{ui}/n, \quad \sigma^2_q = \sum_{i=1}^n \sigma^2_{qi}/n \quad \text{and } \sigma_{uq} = \sum_{i=1}^n \sigma_{uqi}/n.$$

This handles either the structural case (without the X_i necessarily being i.i.d.) or the functional cases. In the functional case, $X_i = x_i$ fixed, so $\sigma^2_X = \Sigma_{i=1}^n (x_i - \overline{x})^2/n$ and $\mu_X = \Sigma_{i=1}^n x_i/n$.

- If the X's are i.i.d. normal and the measurement errors are normal with constant variance and covariances then the result above is exact in the sense that $D_i|w_i = \gamma_0 + \gamma_1 w_i + \delta_i$ where δ_i has mean 0 and variance σ^2_δ.

Special case: No error in the response or uncorrelated measurement errors.

Important and widely studied special case, with $\sigma_{uq} = 0$.

- The naive estimate of the slope estimates

$$\gamma_1 = \left[\frac{\sigma_X^2}{\sigma_X^2 + \sigma_u^2}\right]\beta_1 = \kappa\beta_1,$$

  where

$$\kappa = \left[\frac{\sigma_X^2}{\sigma_X^2 + \sigma_u^2}\right]$$

  is the **reliability ratio**. When $\beta_1 \neq 0$ and $\sigma_u^2 > 0$ then $|\gamma_1| < |\beta_1|$ leading to **attenuation** (slope is biased towards 0.)

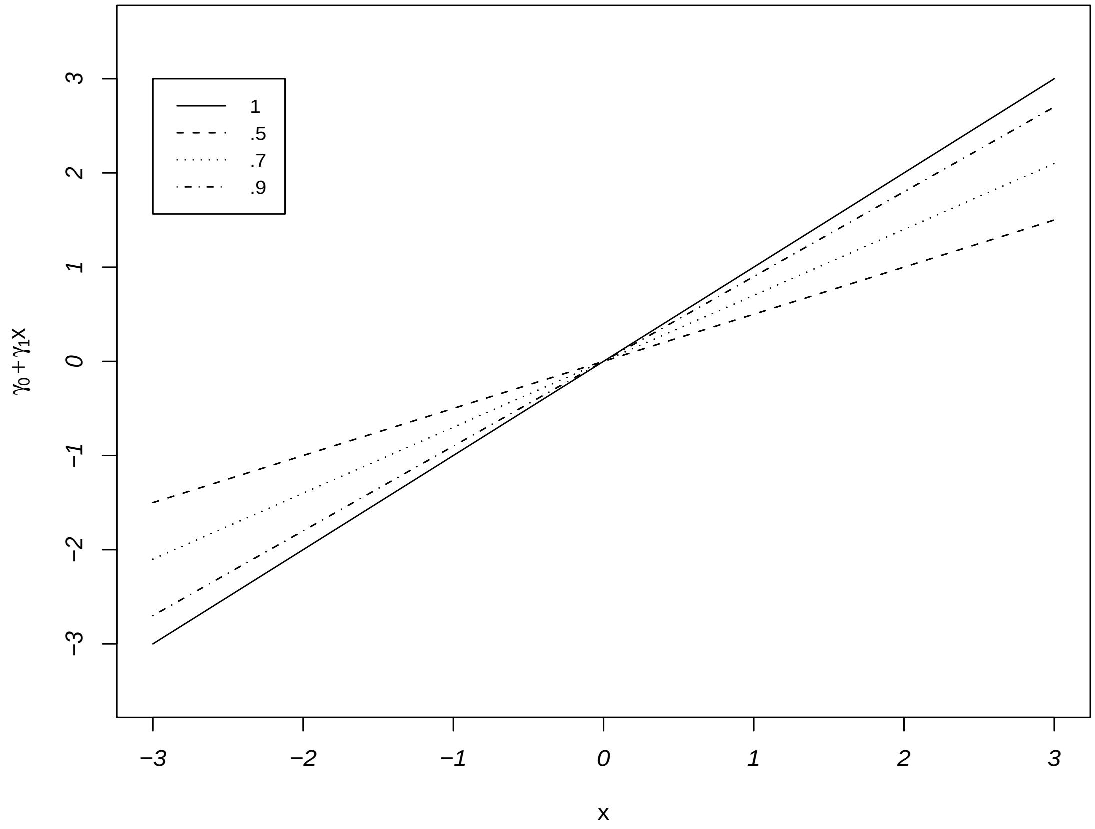
- $\gamma_1 = 0$ if and only if $\beta_1 = 0$, so naive test of $H_0 : \beta_1 = 0$ is essentially correct.

The general statements that "the effect of measurement error is to underestimate the slope but the test for 0 slope are still okay" are not always true. They do hold when there is additive error and either there is error in X only or, if there is measurement error in both variables, the measurement errors are uncorrelated.

Numerical illustrations of bias with error in X only.

$\beta_0 = 0$, $\beta_1 = 1$, $\mu_X = 1$,$\sigma^2_X = 1$ and $\sigma = 1$.

Truth: $\kappa = 1$ (solid line)



Shows attenuation in the slope and that a point on the regression line, $\beta_0 + \beta_1 x_0$ is underestimated if $x_0 > \mu_X = 0$ and overestimated if $x_0 < \mu_X = 0$.

4.2 Correcting For Measurement Error

Long history and a plethora of techniques available depending on what is assumed known; Fuller (1987) and Cheng and Van Ness (1996).

Focus on a simple moment corrected estimator.

Let $\hat{\sigma}^2_{ui}$, $\hat{\sigma}^2_{qi}$ and $\hat{\sigma}_{uqi}$ denote estimates of measurement error variances and covariance.

- If the measurement error variances are constant across i, drop the i subscript.
- In some cases (and in much of the traditional literature) it is assumed that these are known in which case the "hats" would be dropped.
- there are a variety of ways these can be estimated depending on the context.

Replication.

With additive error the measurement error variances and covariances are often estimated through replicate measurements.

$$W_{i1}, \ldots W_{imi} \text{ (replicate values of error-prone measure of } x\text{)}$$

$$W_{ij} = x_i + u_{ij}$$

$E(u_{ij}) = 0$, $V(u_{ij}) = \sigma^2_{ui1}$ (per-replicate variance).

$$W_i = \sum_{j=1}^{m_i} W_{ij}/m_i$$

$$\sigma^2_{ui} = \sigma^2_{ui1}/m_i, \quad \hat{\sigma}^2_{ui} = S^2_{Wi}/m_i$$

S^2_{Wi} = sample variance of $W_{i1}, \ldots, W_{im_i}$.

Similarly if error in Y,

$$\hat{\sigma}^2_{qi} = s^2_{Di}/k_i \qquad \text{and (if paired)} \qquad \hat{\sigma}_{uqi} = S_{WDi}/m_i.$$

k_i = number of replicates of D on unit i ($= m_i$ when paired).

s_{Wdi} = sample covariance of replicate values.

REMARK: If the per-replicate measurement error variance is constant, can use pooling across units without replicates on each individual.

More generally the manner in which estimated variances and covariance are estimated depends on the manner in which W_i (and D_i) are generated on a unit.

Moment based correction:

$$\hat{\beta}_1 = \frac{S_{WD} - \Sigma_i \hat{\sigma}_{uqi}/n}{S_{WW} - \Sigma_i \hat{\sigma}^2_{ui}/n} \qquad \hat{\beta}_0 = \bar{D} - \hat{\beta}_1 \bar{W} \tag{11}$$

$$\hat{\rho} = \frac{S_{WD} - \Sigma_i \hat{\sigma}_{uqi}/n}{(S_{WW} - \Sigma_i \hat{\sigma}^2_{ui}/n)^{1/2}(S_{DD} - \Sigma_i \hat{\sigma}^2_{di}/n)^{1/2}}.$$

$$\hat{\sigma}^2 = \sum_i (r_i^2/(n-2)) - \sum_i (\hat{\sigma}^2_{qi} + \hat{\beta}^2_1 \hat{\sigma}^2_{ui} - 2\hat{\beta}_1 \hat{\sigma}_{uqi})/n,$$

where $r_i = D_i - (\hat{\beta}_0 + \hat{\beta}_1 W_i)$.

NOTE: If either $S_{WW} - \Sigma_i \hat{\sigma}^2_{ui}/n$, which estimates σ^2_X, or $\widehat{\sigma}^2$ are negative, some adjustments need to be made.

Motivation: Conditional on the $x's$ and $y's$,

$E(S_{WD} - \Sigma_i \hat{\sigma}_{uqi}/n) = S_{xy}$ and $E(S_{WW} - \Sigma_i \hat{\sigma}^2_{ui}/n) = S_{xx}$.

Or can view as bias correction to naive estimators.

- The corrected estimators are not unbiased, but are they are consistent under general conditions.

- Under the normal structural model with normal measurement errors and constant known ME variances/covariances, these are close to the maximum Likelihood estimators (as long as $\hat{\sigma}^2_X = S_{WW} - \sigma^2_u$ and $\hat{\sigma}^2$ are nonnegative.) Differ by divisors being used.

- The sampling properties and how to get standard errors, etc. are discussing later under multiple linear regression.

- SPECIAL CASE: Uncorrelated measurement errors and constant variance.

$$\hat{\beta}_1 = \hat{\beta}_{1naive}\hat{\kappa}^{-1}, \qquad \hat{\kappa} = \frac{\hat{\sigma}^2_X}{\hat{\sigma}^2_X + \hat{\sigma}^2_u} \tag{12}$$

$\hat{\kappa}$ = estimated reliability ratio and $\hat{\sigma}^2_X = S_{WW} - \hat{\sigma}^2_u$.

- Regression Calibration Approach: No error in y.

  Motivation for RC will come later. With random X and some assumptions the best linear predictor of X given W is $\mu_X + \kappa(w - \mu_X)$. (Under normality and constant measurement error variance this is exactly $E(X|W = w)$.) If we regress Y_i on $\hat{X}_i = \bar{W} + \hat{\kappa}(w_i - \bar{W})$, the estimated coefficients are identical to the moment corrected estimators.

Bootstrapping.

Bootstrapping needs to be approached carefully (more general discussion later.)

Data on the *ith* unit is $(D_i, W_i, \widehat{\boldsymbol{\theta}}_i)$ where $\widehat{\boldsymbol{\theta}}_i$ contains quantities associated with estimating the measurement error parameters. With error in X only then $\widehat{\boldsymbol{\theta}}_i = \widehat{\sigma}^2_{ui}$, while with correlated errors in both variables, $\widehat{\boldsymbol{\theta}}'_i = (\widehat{\sigma}^2_{ui}, \widehat{\sigma}^2_{qi}, \widehat{\sigma}_{uqi})$.

One-stage bootstrap. Resample units; i.e., resample from $\{(D_i, W_i, \widehat{\boldsymbol{\theta}}_i)\}$.

Only justified if overall random sample of units and common manner of estimating the measurement error parameters on each unit (e.g., equal sampling effort). Will not work for non random-samples or if sampling effort changes and is attached to the position in the sample (rather than being attached to the unit occurring in that position.)

Two-stage bootstrap. Mimic the whole process (e.g., occurrence of measurement error and of error in the equation.).

Consider error in x only.

$\hat{x}_i$ = estimate of x_i (could be W_i or predicted value from $E(X|w)$)

On bth bootstrap sample the ith observation is $(Y_{bi}, W_{bi}, \widehat{\theta}_{bi})$ where

$$Y_{bi} = \widehat{\beta}_0 + \widehat{\beta}_1 \hat{x}_i + e_{bi}$$

where e_{bi} is a generated value having mean 0 and variance $\widehat{\sigma}^2$.

W_{bi} and $\widehat{\theta}_{bi}$ are created by mimicking how W_i and $\widehat{\theta}_i$ arose. For example with replication we would generate replicates $W_{bij} = \hat{x}_i + u_{bij}$ where u_{bij} has mean 0 and per-replicate variance $m_i \widehat{\sigma}^2_{ui}$ (so the mean W_{bi} has variance $\widehat{\sigma}^2_{ui}$.)

How to generate e_{bi}? Easy if assume normality, but difficult to do nonparametrically. With measurement error can't just use residuals from the corrected fit since even with known coefficients

$$r_i = D_i - (\hat{\beta}_0 + \hat{\beta}_1 W_i) = \epsilon_i + q_i - \beta_1 u_i.$$

Need to unmix/deconvolve to remove the influence of the measurement error.

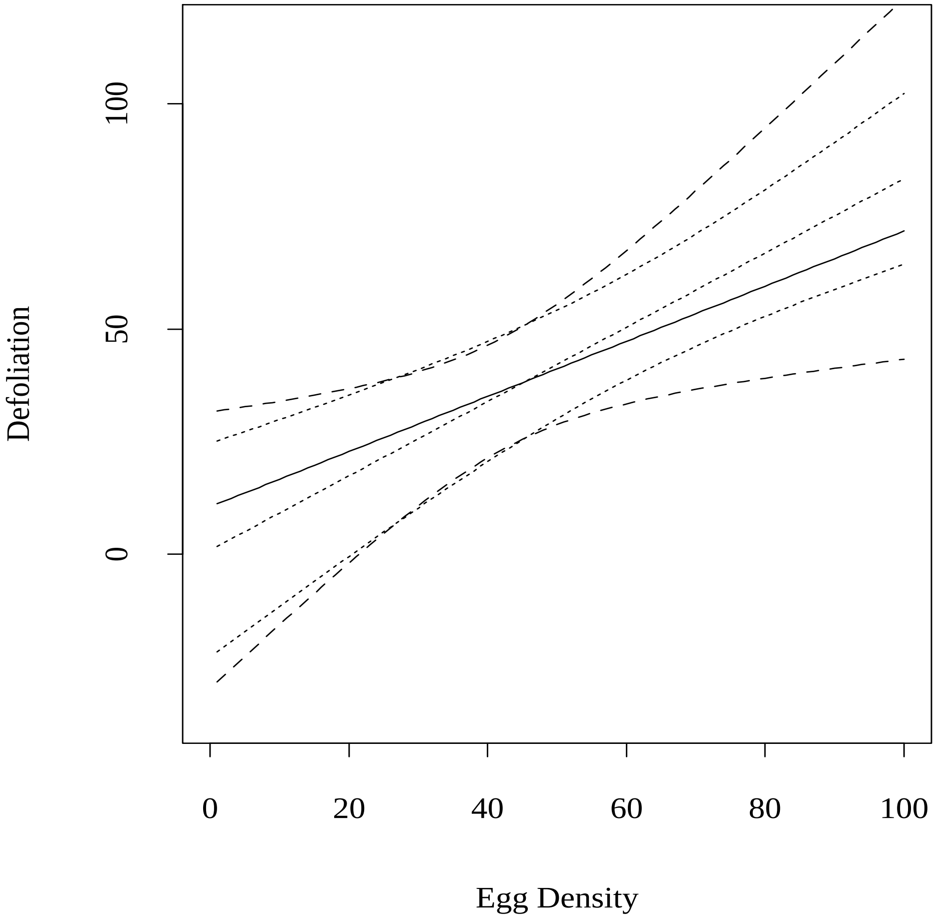
4.3 Examples

Defoliation example with error in both variables

Assume a common linear regression model holds over all three forests (more on this later).

No bootstrap analysis not provided here. Since the 18 stands were not an overall random sample and the sample effort is unequal an analysis based on resampling observations could be misleading.

Method	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\sigma}^2$	$SE(\hat{\beta}_0)$	$SE(\hat{\beta}_1)$	v_{01}	CI for β_1
NAIVE	10.504	.6125	696.63	11.459	.2122	-2.042	(.163,1.063)
COR-R	.8545	.8252	567.25	12.133	.182	-1.918	(.469,1.181)
COR-N	.8545	.8252	567.25	15.680	.338	-4.848	(.163,1.488)



Naive: solid; corrected: = dots; dots = Robust CIs; dashed = normal CIs

Nitrogen-yield example. Y = corn yield (no error).

X = nitrogen level. Measured with error with $\sigma_u^2 = 57$.

v_{01} = estimated covariance of $\widehat{\beta}_0$ and $\widehat{\beta}_1$.

Method	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\sigma}^2$	$SE(\hat{\beta}_0)$	$SE(\hat{\beta}_1)$	v_{01}	CI (β_1)
NAIVE	73.153	.344	57.32	9.951	.137	-1.328	(.034, .654)
COR-N	67.56	.423	49.235	12.56	.176	-2.157	(.081,.766)
COR-R	67.56	.423	49.234	10.787	.146	-1.547	(.137,.710)
COR-RC	67.56	.423	57.32	12.130	.169		

Using the rcal function in STATA (non-bootstrap SE's for intercept are incorrect).

```
. mat input suu=(57)        . rcal (yield=) (w:nitrogen), suuinit(suu)
----------------------------------------------------------------
 yield |    Coef.  Std. Err.  t     P>|t|    [95% Conf. Interval]
-------+--------------------------------------------------------
     w |   .4232   .1712    2.47   0.035    .0358   .8105
 _cons |   67.56   9.951    6.79   0.000    45.053   90.075
----------------------------------------------------------------
  rcal (yield=) (w: nitrogen), suuinit(suu) robust
----------------------------------------------------------------
       |        Semi-Robust
 yield |  Coef.  Std. Err.  t      P>|t|    [95% Conf. Interval]
-------+--------------------------------------------------------
     w | .4232   .1491      2.84   0.019   .0860   .7604
 _cons |  67.56   8.559     7.89   0.000    48.202  86.926
----------------------------------------------------------------
. rcal (yield = ) (w: nitrogen), suuinit(suu) bstrap brep(5000)
----------------------------------------------------------------
       |        Bootstrap
 yield |  Coef.  Std. Err.  t     P>|t|    [95% Conf. Interval]
-------+--------------------------------------------------------
     w |.4232   .1927     2.20   0.056    -.01276    .8591
 _cons | 67.56   14.16     4.77   0.001    35.530     99.599
```

5 Multiple Linear Regression With Additive Error

5.1 Model and the performance of naive estimators.

$$Y_i|\mathbf{x}_i = \beta_0 + \sum_{j=1}^{p-1} \beta_j x_{ij} + \epsilon_i = \boldsymbol{\beta}'\mathbf{x}_{i*} + \epsilon_i = \beta_0 + \boldsymbol{\beta}_1'\mathbf{x}_i + \epsilon_i$$

$\mathbf{x}_i = (x_1, \ldots, x_{p-1})'$ and $\mathbf{x}_{i*}' = (1, x_1, \ldots, x_{p-1})$,

$$\boldsymbol{\beta}' = (\beta_0, \beta_1, \ldots, \beta_{p-1}) = (\beta_0, \boldsymbol{\beta}_1'), \quad \boldsymbol{\beta}_1' = (\beta_1, \ldots, \beta_{p-1}).$$

ϵ_i are assumed uncorrelated with mean 0 and variance σ^2.

Without an intercept, $\mathbf{x}_i = \mathbf{x}_{i*}$ and $\boldsymbol{\beta} = \boldsymbol{\beta}_1$.

In matrix form: $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$,

$$\mathbf{Y} \begin{bmatrix} Y_1 \\ . \\ Y_n \end{bmatrix} \qquad \mathbf{X} = \begin{bmatrix} \mathbf{x}_{1*}' \\ . \\ \mathbf{x}_{n*}' \end{bmatrix}, \qquad \boldsymbol{\epsilon} = \begin{bmatrix} \epsilon_1 \\ . \\ \epsilon_n) \end{bmatrix}$$

$E(\boldsymbol{\epsilon}) = \mathbf{0}$ and $Cov(\boldsymbol{\epsilon}) = \sigma^2 I$.

Define

$$\bar{\mathbf{X}} = \sum_i \mathbf{X}_i/n, \bar{Y} = \sum_i Y_i/n, S_Y^2 = \sum_i (Y_i - \bar{Y})^2/(n-1),$$

$$\mathbf{S}_{XX} = \frac{\Sigma_i(\mathbf{X}_i - \bar{\mathbf{X}})(\mathbf{X}_i - \bar{\mathbf{X}})'}{n-1} \text{ and } \mathbf{S}_{XY} = \frac{\Sigma_i(\mathbf{X}_i - \bar{\mathbf{X}}_)(Y_i - \bar{Y})}{n-1}.$$

$\mathbf{S}_{XX}$ is a $(p-1) \times (p-1)$ matrix and $\mathbf{S}_{XY}$ is a $(p-1) \times 1$ vector.

Mimicking what was done in the simple linear case

$$\boldsymbol{\mu}_X = \sum_{i=1}^{n} E(\mathbf{X}_i) \text{ and } \Sigma_{XX} = E(\mathbf{S}_{XX}). \tag{13}$$

Accommodates any mix of random and fixed predictors.

- In the structural case with $\mathbf{X}_i$ i.i.d., $\boldsymbol{\mu}_X = E(\mathbf{X}_i)$ and $\Sigma_{XX} = Cov(\mathbf{X}_i)$.

- In functional case with all fixed predictors, $\boldsymbol{\mu}_X = \Sigma_i \mathbf{x}_i/n$ and $\Sigma_{XX} = \mathbf{S}_{xx} = \Sigma_i(\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})'/(n-1)$.

- With a combination of fixed and random predictors the expected value and covariance of $\mathbf{X}_i$ are conditional on any fixed components.

Additive Measurement Error model:

Given y_i and $\mathbf{x}_i$,

$$D_i = y_i + q_i, \qquad \mathbf{W}_i = \mathbf{x}_i + \mathbf{u}_i,$$

with $E(q_i|y_i, \mathbf{x}_i) = 0$, $E(\mathbf{u}_i|y_i, \mathbf{x}_i) = \mathbf{0}$,

$$V(q_i|y_i, \mathbf{x}_i) = \sigma^2_{qi}, Cov(\mathbf{u}_i|y_i, \mathbf{x}_i) = \Sigma_{ui}, \text{ and } Cov(\mathbf{u}_i, q_i|y_i, \mathbf{x}_i) = \Sigma_{uqi}.$$

- Allows error in the response as well as the predictors. If there is no error in y then $\sigma^2_{qi} = 0$ and $\Sigma_{uqi} = \mathbf{0}$.

- If parts of $\mathbf{x}_i$ are measured without error, the appropriate components of $\mathbf{u}_i$ equal 0 and all of the components in Σ_{ui} and Σ_{uqi} involving that variable also equal 0.

- The measurement error variances and covariances are allowed to change with i with

$$\Sigma_u = \sum_{i=1}^{n} \Sigma_{ui}/n, \quad \Sigma_{uq} = \sum_{i=1}^{n} \Sigma_{uqi}/n \quad \text{and} \quad \sigma_q^2 = \sum_{i=1}^{n} \sigma_{qi}^2/n.$$

Naive estimators.

$$\hat{\boldsymbol{\beta}}_{naive} = (\mathbf{W}'\mathbf{W})^{-1}\mathbf{W}'\mathbf{D},$$

$\mathbf{W}$ is the same as $\mathbf{X}$ but $\mathbf{W}_i'$ in place of $\mathbf{x}_i'$.

$$\hat{\boldsymbol{\beta}}_{1naive} = \mathbf{S}_{WW}^{-1}\mathbf{S}_{WD}, \; \textit{and} \; \hat{\beta}_{0naive} = \bar{D} - \hat{\boldsymbol{\beta}}_{1naive}'\bar{\mathbf{W}},$$

$$\widehat{\sigma}_{naive}^2 = \frac{\Sigma_i (D_i - \widehat{D}_i)^2}{n-p} = \frac{n-1}{n-p}\left[S_D^2 - \hat{\boldsymbol{\beta}}_{1naive}'\mathbf{S}_{WW}\hat{\boldsymbol{\beta}}_{1naive}\right],$$

where $\widehat{D}_i$ is the fitted value using the naive estimates.

Approximate/asymptotic properties of naive estimators

$E(\hat{\boldsymbol{\beta}}_{1naive}) \approx \boldsymbol{\gamma}_1 = (\Sigma_{XX} + \Sigma_u)^{-1}\Sigma_{XX}\boldsymbol{\beta}_1 + (\Sigma_{XX} + \Sigma_u)^{-1}\Sigma_{uq}$

$E(\hat{\boldsymbol{\beta}}_{0naive}) \approx \gamma_0 = \beta_0 + (\boldsymbol{\beta}_1 - \boldsymbol{\gamma}_1)'\boldsymbol{\mu}_X$

$E(\widehat{\sigma}_{naive}^2) \approx \sigma^2 + \sigma_q^2 + \boldsymbol{\beta}_1'\Sigma_{XX}\boldsymbol{\beta}_1 - \boldsymbol{\gamma}_1'(\Sigma_{XX} + \Sigma_u)\boldsymbol{\gamma}_1.$

Alternate expression:

$$E(\hat{\boldsymbol{\beta}}_{naive}) \approx \boldsymbol{\gamma} = (\mathbf{M}_{XX} + \Sigma_{u*})^{-1}(\mathbf{M}_{XX}\boldsymbol{\beta} + \Sigma_{uq*}),$$

$$\mathbf{M}_{XX} = \sum_i E(\mathbf{X}_{i*}\mathbf{X}'_{i*})/n = \Sigma_{XX*} + \boldsymbol{\mu}_{X*}\boldsymbol{\mu}'_{X*},$$

$$\boldsymbol{\mu}_{X*} = \begin{bmatrix} 1 \\ \boldsymbol{\mu}'_X \end{bmatrix}, \Sigma_{XX*} = \begin{bmatrix} 0 & \mathbf{0}' \\ \mathbf{0} & \Sigma_{XX} \end{bmatrix}, \Sigma_{u*} = \begin{bmatrix} 0 & \mathbf{0}' \\ \mathbf{0} & \Sigma_u \end{bmatrix} \text{ and } \Sigma_{uq*} = \begin{bmatrix} 0 \\ \Sigma_{uq} \end{bmatrix}.$$

In functional case $\mathbf{M}_{XX} = \mathbf{X}'\mathbf{X}/n$.

- Similar to simple linear regression the expectations are exact under the normal structural model with normal measurement error and constant measurement error covariance matrix. Otherwise they are only approximate/asymptotic.

- If $\Sigma_{uq} = \mathbf{0}$ (i.e., no error in the response or the error in the response is uncorrelated with any errors in the predictors) then

  $$E(\hat{\boldsymbol{\beta}}_{1naive}) = \boldsymbol{\gamma}_1 = (\Sigma_{XX} + \Sigma_u)^{-1}\Sigma_{XX}\boldsymbol{\beta}_1 = \boldsymbol{\kappa}\boldsymbol{\beta}_1 \qquad (14)$$

  where

  $$\boldsymbol{\kappa} = (\Sigma_{XX} + \Sigma_u)^{-1}\Sigma_{XX} \qquad (15)$$

  is referred to as the **reliability matrix.**

- These same approximations work for a number of generalized linear models (including logistic regression) with additive error in the predictors.

- *Measurement error in one of the variables often induces bias in the estimates of all of the coefficients including those that are not measured with error.*

Ex. Two predictors x_1 and x_2 with regression function $\beta_0 + \beta_1 x_1 + \beta_2 x_2$. x_1 subject to measurement error with variance σ_u^2. No measurement error in either x_2 or y. Then

$$\mathbf{X}_i = \begin{bmatrix} X_{i1} \\ X_{i2} \end{bmatrix} \text{ and } \Sigma_{XX} = \begin{bmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2 \end{bmatrix},$$

with the components of Σ_{XX} interpreted accordingly.

$$\Sigma_{ui} = \begin{bmatrix} \sigma_u^2 & 0 \\ 0 & 0 \end{bmatrix}$$

$$E(\hat{\beta}_{1naive}) \approx \left[\frac{\sigma_2^2 \sigma_1^2 - \sigma_{12}^2}{\sigma_2^2(\sigma_1^2 + \sigma_u^2) - \sigma_{12}^2} \right] \beta_1$$

$$\text{Bias in } \hat{\beta}_{1naive} \approx \left[\frac{-\sigma_2^2 \sigma_u^2}{\sigma_2^2(\sigma_1^2 + \sigma_u^2) - \sigma_{12}^2} \right] \beta_1$$

$$\text{Bias in } \hat{\beta}_{2naive} \approx \left[\frac{\sigma_{12} \sigma_u^2}{\sigma_2^2(\sigma_1^2 + \sigma_u^2) - \sigma_{12}^2} \right] \beta_1.$$

REMARKS:

- If $\sigma_{12} = 0$, then there is no bias in the naive estimator of β_2. If the two predictors are "correlated" ($\sigma_{12} \neq 0$) then the measurement error in x_1 induces bias in the estimated coefficient for x_2.

- The bias in the naive estimator of β_2 (the coefficient associated with the variable measured without error) can be either positive or negative. It does not depend on β_2, but does depend on β_1.

- If $\sigma_{12} = 0$, then $\hat{\beta}_{1naive}$ estimates $\kappa_1\beta_1$, where $\kappa_1 = \sigma_1^2/(\sigma_1^2 + \sigma_u^2)$. This means the attenuation is the same as in simple linear regression. *It is critical to note that this requires X_1 and X_2 to be uncorrelated.*

Illustration based on values in Armstrong et al.(1989)

x_2 = caloric intake (in 100's of grams), x_1 = fiber intake (in gms).

$V(X_2) = 33$, $V(X_1) = 72.0$, $cov(X_1, X_2) = 16.4$ (correlation of .336).

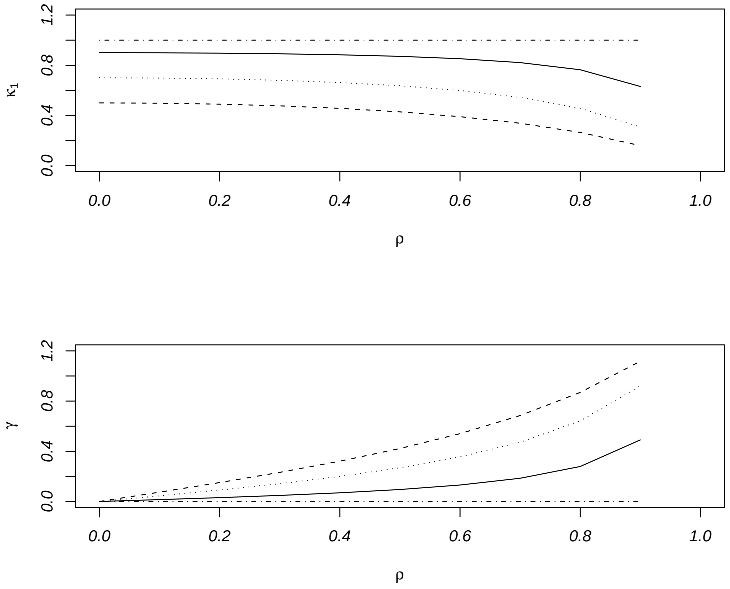
Measurement error in fiber with measurement variance $\sigma_u^2 = 70.1$.

- In a simple linear regression model with only x_1 in it the naive estimator of the coefficient for x_1 is estimating $\kappa\times$(true coefficient for x_1), where $\kappa = 72/(72 + 70.1) = .507$.

  In a model with both x_1 and x_2, then naive estimate of coefficient of x_1 is estimating $\kappa_1\beta_1$ where $\kappa_1 = ((33)(72) - 16.4^2)/(33(72 + 70.1) - 16.4^2) = .4767$. In this case, if we interpret the bias (or correct the naive estimator) by assuming the attenuation factor of .507 applies in the multiple setting, then we don't go too wrong. This is because the correlation between X_1 and X_2 is relatively low.

- If we change ρ to .8, then $\kappa = .507$, but $\kappa_1 = .270$, while with $\rho = .8$ and σ_u^2 reduced to 10, then $\kappa = .878$, but $\kappa_1 = .722$. This illustrates how it can be misleading to try and characterize the bias in β_1 using κ.

Plot of attenuation factor (κ_1) in the naive estimator of the coefficient of the mismeasured predictor in a model with two variables versus ρ = correlation between the mismeasured variable and the second, perfectly measured, predictor. The reliability on the predictor by itself is $\kappa = \sigma_1^2/(\sigma_1^2 + \sigma_u^2) =$ .9 (solid line), .7 (dotted line) and .5 (dashed line).



The above can be generalized easily to multivariate $\mathbf{X}_1$ and $\mathbf{X}_2$ with error in $\mathbf{X}_1$ but not $\mathbf{X}_2$.

Analysis of Covariance: Interested in comparing group means where individuals (plants, people, etc.) are randomized to groups (e.g., treatments). The analysis of covariance adjusts for the effect of an additional variable x, (the covariate), which is associated with the unit.

Example 1: Randomize varieties of tomato plants over an area observe $Y = yield$. Use $x =$ measure of soil nitrogen at the location as a covariate.

Example 2: Randomize individuals with high cholesterol to different treatment regimes. Look at change in cholesterol after a certain period of time. $x =$ person's level of physical activity (e.g., METS/day.)

For kth individual assigned to group j, model is $Y_{jk} = \tau_j + \beta_x x_{jk} + \epsilon_{jk}$, $x_{jk} =$ covariate.

$\mu_j =$ population mean for group $j =$ the expected average response if all units in the population get treatment j.

With randomization $\mu_j = \tau_j + \beta_x \mu_X$, where μ_X is the population mean of the covariate.

The τ_j are the adjusted group means; i.e., $\tau_j = \mu_j - \beta_x \mu_X$.

Can show that the dummy variables indicating group membership and the covariate are uncorrelated over the randomization. This is because the expected value of the covariate given the group assignment is always μ_X.

In the standard analysis of covariance model, which involves randomization to treatment groups and a constant coefficient for the covariate in each group, measurement error in the covariate does not influence inferences for the τ's and functions of them. In particular inferences for contrasts in the group means (e.g., testing equality of group means and estimating difference in group means) are correct.

It is important to notice however that the naive estimators of the group means are not unbiased since $\hat{\tau}_{j,naive} + \widehat{\beta}_{xnaive}\bar{x}$ is a biased estimator of μ_j.

5.2 Correcting for Measurement Error

Moment correction: Generalizes the result for simple linear regression case:

$$\hat{\boldsymbol{\beta}}_1 = \widehat{\Sigma}_{XX}^{-1}\widehat{\Sigma}_{XY} \text{ and } \widehat{\beta}_0 = \bar{D} - \widehat{\boldsymbol{\beta}}_1'\bar{\mathbf{W}},$$

$\widehat{\Sigma}_{XX} = \mathbf{S}_{WW} - \hat{\Sigma}_u$ and $\widehat{\Sigma}_{XY} = \mathbf{S}_{WD} - \hat{\Sigma}_{uq}$ (with no error in the response $\widehat{\Sigma}_{XY} = \mathbf{S}_{WD}$).

Motivated by $E(\mathbf{S}_{WW}) = \Sigma_{XX} + \Sigma_u$ and $E(\mathbf{S}_{WD}) = \Sigma_{XY} + \Sigma_{uq}$.

Alternate form

$$\hat{\boldsymbol{\beta}} = \widehat{\mathbf{M}}_{XX}^{-1}\widehat{\mathbf{M}}_{XY}$$

where

$$\widehat{\mathbf{M}}_{XX} = \frac{\mathbf{W}'\mathbf{W}}{n} - c\widehat{\Sigma}_{u*} \quad \text{and} \quad \widehat{\mathbf{M}}_{XY} = \frac{\mathbf{W}'\mathbf{D}}{n} - c\widehat{\Sigma}_{uq*},$$

$$\widehat{\Sigma}_{u*} = \begin{bmatrix} 0 & \mathbf{0}' \\ \mathbf{0} & \widehat{\Sigma}_u \end{bmatrix}$$

The constant c can be taken to be either $c = 1$ or $c = (n-1)/n$. $c = (n-1)/n$ yields estimators above.

$$\hat{\sigma}^2 = \frac{\Sigma_i(D_i - (\widehat{\beta}_0 + \widehat{\boldsymbol{\beta}}_1'\mathbf{W}_i))^2}{n-p} - (\widehat{\sigma}_q^2 - 2\hat{\boldsymbol{\beta}}_1'\hat{\Sigma}_{uq} + \hat{\boldsymbol{\beta}}_1'\hat{\Sigma}_u\hat{\boldsymbol{\beta}}_1).$$

If a divisor of $n-1$ is used in the first piece above then

$$\hat{\sigma}^2 = S_d^2 - \widehat{\sigma}_q^2 - \hat{\boldsymbol{\beta}}_1'\widehat{\Sigma}_{XX}\hat{\boldsymbol{\beta}}_1.$$

- These are essentially the maximum likelihood estimates under normality.
- With no error in Y this is equivalent to a regression calibration estimator obtained by regressing Y_i on $\hat{\mathbf{x}}_i = \bar{\mathbf{w}} + \hat{\boldsymbol{\kappa}}'(\mathbf{w}_i - \bar{\mathbf{w}})$.
- Similar to simple linear regression, we need a modification if certain matrices (e.g., $\hat{M}_{xx}$) are negative or if $\widehat{\sigma}^2 < 0$.

5.3 Sampling Properties and Approximate Inferences

Small sample properties of $\widehat{\boldsymbol{\beta}}$ difficult to ascertain. Asymptotic properties treated in detail in Fuller (1987) and summarized in Buonaccorsi (2010).

Under some general conditions

$$\widehat{\boldsymbol{\beta}} \sim AN(\boldsymbol{\beta}, \Sigma_\beta)$$

$$\Sigma_\beta = \mathbf{M}_{XX}^{-1}\mathbf{H}\mathbf{M}_{XX}^{-1}$$

with

$$\mathbf{H} = Cov(\sum_{i=1}^{n} \Delta_i)/n^2,$$

$$\Delta_i = \mathbf{W}_{i*}(D_i - \mathbf{W}'_{i*}\boldsymbol{\beta}) - (\hat{\Sigma}_{uqi*} - \hat{\Sigma}_{ui*}\boldsymbol{\beta}) = \mathbf{T}_i - \mathbf{Z}_i,$$

$$e_i = \epsilon_i + q_i - \mathbf{u}_i'\boldsymbol{\beta}_1, \quad T_i = \mathbf{W}_{i*}e_i \quad \text{and} \quad \mathbf{Z}_i = (\hat{\Sigma}_{uqi*} - \hat{\Sigma}_{ui*}\boldsymbol{\beta}).$$

$E(\Delta_i) = \mathbf{0}$.

With an estimate $\widehat{\Sigma}_\beta$, can use large sample normal based Wald methods

Considerations in the form of Σ_β and how to estimate it.

- Do we have separately estimated measurement error parameters for each i?
- Are the measurement error parameters constant over i?
- Can the estimated measurement error parameters be treated as known?
- Is ϵ_i assumed normal?
- Are the measurement errors assumed normal?
- If the measurement error variances and covariances are estimated what can we say about their sampling properties?
- If replication is involved are the replicates assumed normal?
- **Robust estimate**

$$\hat{\Sigma}_{\beta,Rob} = \widehat{\mathbf{M}}_{XX}^{-1}\widehat{\mathbf{H}}_R\widehat{\mathbf{M}}_{XX}^{-1},$$

- **Normal based** with known M.E. parameters ((3.1.23) in Fuller.)

$$\hat{\boldsymbol{\Sigma}}_{\beta,N} = \widehat{\mathbf{M}}_{XX}^{-1}\widehat{\mathbf{H}}_N\widehat{\mathbf{M}}_{XX}^{-1},$$

$\hat{\mathbf{H}}_R = \Sigma_i \hat{\Delta}_i\hat{\Delta}_i^{'}/n(n-p)$ and $\hat{\mathbf{H}}_N = \Sigma_{i=1}^n(\mathbf{W}_{i*}\mathbf{W}'_{i*}\widehat{\sigma}^2_{ei} + \widehat{\mathbf{Z}}_i\widehat{\mathbf{Z}}'_i)/n^2$

$$\hat{\Delta}_i = \mathbf{W}_{i*}r_i - (\hat{\Sigma}_{uqi*} - \hat{\Sigma}_{ui*}\widehat{\boldsymbol{\beta}}) \text{ and } \widehat{\mathbf{Z}}_i = \hat{\Sigma}_{uqi*} - \hat{\Sigma}_{ui*}\widehat{\boldsymbol{\beta}}$$

$$\widehat{\sigma}^2_{ei} = \widehat{\sigma}^2 + \widehat{\sigma}^2_{qi} - 2\widehat{\boldsymbol{\beta}}'_1\widehat{\Sigma}_{uqi} + \widehat{\boldsymbol{\beta}}'_1\widehat{\Sigma}_{ui}\widehat{\boldsymbol{\beta}}_1.$$

- Without measurement error $\widehat{\Sigma}_{\beta,Rob}$ is White's robust estimate and $\widehat{\Sigma}_{\beta,N} = \widehat{\sigma}^2(\mathbf{X}'\mathbf{X})^{-1}$ (the usual covariance estimate for linear regression with constant variances.)

- There are other possibilities under various assumptions.

Bootstrapping. Same comments made for simple linear regression apply here.

- Be careful just resampling the units. Only applicable under certain assumptions.

- Two stage resampling (mimicking the randomness in original model) is difficult to do non-parametrically but parametric bootstrap is pretty straightforward.

There are a variety of modifications and alternate estimators that can be considered. These include (but are not limited to)

- Maximum likelihood estimators under normality and particular restrictions on some parameters (includes orthogonal least squares).
- Small sample modifications.
- Weighted corrected estimators.

- Estimation using instrumental variables.

  Instead of replication or estimated M.E. variances and covariances have an $r \times 1$ vector of instruments $\mathbf{R}_i$ ($r \geq p =$ number of elements of $\mathbf{x}_{i*}$).

  Variables measured without error are included in $\mathbf{R}_i$, so they are instruments for themselves. For $\mathbf{R}_i$ to be instrumental for $\mathbf{x}_{i*}$ requires:

  1. $\mathbf{R}_i$ is uncorrelated with $\mathbf{u}_i$, q_i and ϵ_i
  2. $\mathbf{R}_i$ is correlated with $\mathbf{x}_{i*}$.

  SAS- SYSLIN, STATA-IVREG and other programs will run this.

Some other issues.

- Residual analysis for model assessment
- Prediction

5.4 Examples

Defoliation example revisited

Allow a different intercept for each forest.

Forest j model: $\tau_j + \beta_x x$.

$Y_i = \boldsymbol{\beta}'\mathbf{x}_i + \epsilon_i$, $\mathbf{x}_i' = (x_{i1}, x_{i2}, x_{i3}, x_i)$, $\boldsymbol{\beta}' = (\tau_1, \tau_3, \tau_3, \beta_x)$

x_{i1}, x_{i2} and x_{i3} are dummy variables indicating forest membership.

$$\Sigma_{ui} = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & \sigma^2_{ui} \end{bmatrix} \quad \text{and} \quad \Sigma_{uqi} = \begin{bmatrix} 0 \\ 0 \\ 0 \\ \sigma_{uqi} \end{bmatrix}.$$

	Naive (SE)	Cor	SE-R	CI-R	SE-N	CI-N
τ_1	39.21 (14.65)	26.86	21.26	(-14.82,68.54)	20.66	(-13.63, 67.35)
τ_2	16.78 (10.26)	10.12	11.31	(-12.05, 32.30)	13.26	(-15.86, 36.11)
τ_3	-4.49 (12.37)	-12.73	12.75	(-37.72, 12.25)	16.31	(-44.70, 19.24)
β_x	0.44 (0.19)	0.638	0.239	(0.17, 1.11)	0.321	(0.01, 1.27)
σ^2	455.24	381.87				

R: robust standard error. N: “normal” based.

- Substantial correction for attenuation in the estimate of β_x from .44 to a corrected estimate of .64.

- Some relatively large changes in the estimated intercepts. Recall that naive estimates of coefficients of perfectly measured predictors can be biased. Here due to different estimated egg mass densities(62.3, 33.6 and 41.6) over the three different forests, indicating ”correlation” between the dummy variables indicating forest stand and the egg mass density.

Hypothesis testing:

Null Hypothesis	Naive (Pvalue)	Robust (Pvalue)	Normal (Pvalue)
$\beta_x = 0$	2.32 (.036)	2.67 (.008)	1.98 (.047)
$\tau_1 = \tau_2 = \tau_2$	5.24 (.020)	12.58 (.002)	7.29 (.026)

$$\hat{\Sigma}_{\beta,naive} = \begin{bmatrix} 214.75 & 74.82 & 92.77 & -2.23 \\ 74.82 & 105.34 & 49.98 & -1.20 \\ 92.77 & 49.98 & 153.02 & -1.49 \\ -2.23 & -1.20 & -1.49 & 0.04 \end{bmatrix}$$

$$\hat{\Sigma}_{\beta,rob} = \begin{bmatrix} 452.17 & 122.02 & 220.91 & -4.77 \\ 122.02 & 128.03 & 66.47 & -1.31 \\ 220.91 & 66.47 & 162.52 & -2.74 \\ -4.77 & -1.31 & -2.74 & 0.06 \end{bmatrix}$$

$$\hat{\Sigma}_{\beta,norm} = \begin{bmatrix} 426.75 & 194.84 & 234.53 & -5.99 \\ 194.84 & 175.77 & 132.19 & -3.37 \\ 234.53 & 132.19 & 266.06 & -4.06 \\ -5.99 & -3.37 & -4.06 & 0.10 \end{bmatrix}.$$

House price example

Data from DASL (Data and Story Library) in StatLib (*"a random sample of records of resales of homes from Feb 15 to Apr 30, 1993 from the files maintained by the Albuquerque Board of Realtors. This type of data is collected by multiple listing agencies in many cities and is used by realtors as an information base."*)

Variable	Mean	SD	Min	Max
PRICE	1077	383.99	540.00000	2150
SQFT	1667	531.34	837.00000	3750
TAX	793.49	308.18	223.00000	1765

House price is in hundreds of dollars so the mean price is $10,770 (the good old days). The correlation between recorded square footage and tax level is .8586.

Regress sale price on square footage and the yearly tax using the 107 (out of 117) cases that have all three variables assuming square footage is measured with error having a constant, known measurement error variance, σ_u^2.

- The bootstrap is one-stage (resampling cases). 52 out of 5000 bootstrap samples required modification for some negative matrices.

- Fairly big difference between robust and the "normal" based estimates of Σ_β. Probably due to heterocedasticity in the error in the equation. Use robust or bootstrap methods.

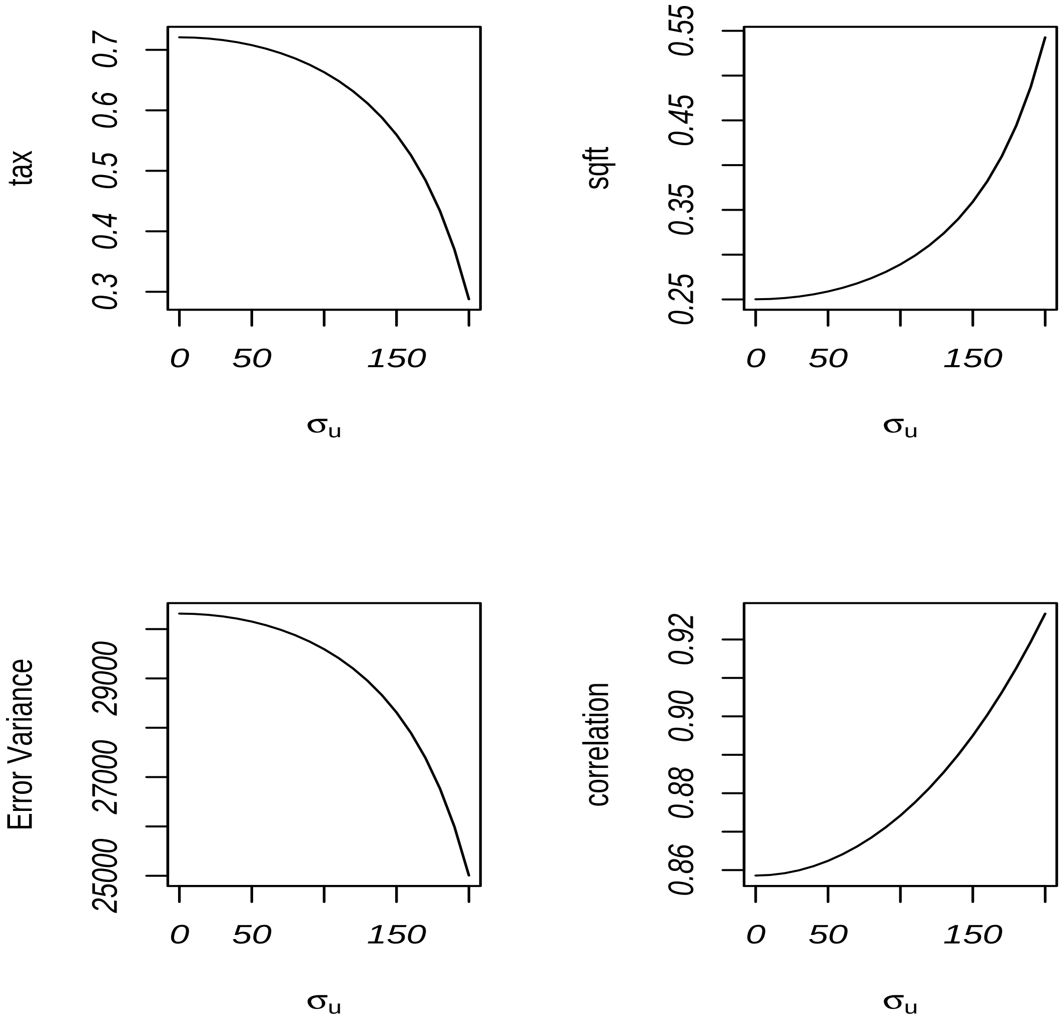
$\sigma_u = 175$; Cor-R uses robust; Cor-N normal based. Standard errors in parentheses.

Method	Intercept	Tax	SqFt	σ^2
Naive:	88.34 (55.81)	.721 (.107)	.250 (.062)	30314.0
Naive: RobSE	88.34 (87.70)	.721 (.178)	.250 (.131)	
Cor-R:	1.71 (171.05)	.46 (.41)	.43 (.30)	27094.1
Cor-N:	1.71 (74)	.46 (.18)	.43 (.11)	27094.1
Bootstrap				
Mean(SE)	-104.15(241.84)	.21 (.59)	.61 (.43)	
Confidence intervals				
Naive: CI	(-22.34, 199.01)	(.508, .933)	(.127,.373)	
Cor-R: CI	(-333.55,336.98)	(-0.34,1.26)	(-0.16, 1.01)	
Cor-N: CI	(-143.3,146.75)	(0.11, 0.81)	(0.20, 0.65)	
Boot: CI	(-628.66, 195.33)	(-1.08, 0.96)	(0.09,1.56)	

STATA analyses using rcal function.

```
----------------------------------------------------------------
 price |  Coef.    Std. Err  t    P>|t|    [95% Conf. Interval]
-------+--------------------------------------------------------
   tax | .46    .3218453     1.43    0.156    -.1777726     1.09869
     w | .43    .2236257     1.90    0.060    -.0174522    .8694642
 _cons | 1.71   55.80885     0.03    0.976     -108.9563    112.3858
-----------------------------------------------------------------
       |         Semi-Robust
 price |  Coef.    Std. Err   t    P>|t|    [95% Conf. Interval]
-------+---------------------------------------------------------
   tax | .46    .4149749     1.11    0.270     -.362452    1.283369
     w | .43    .2979442     1.43    0.156    -.1648285    1.016841
 _cons | 1.71  87.70255     0.02    0.984    -172.2027    175.6322
-----------------------------------------------------------------
       |         Bootstrap
 price |  Coef.    Std. Err.  t    P>|t|     [95% Conf. Interval]
-------+---------------------------------------------------------
   tax | .46    .5837385     0.79    0.432    -.6971165    1.618034
     w | .43    .4241468     1.00    0.318    -.415093    1.267105
 _cons | 1.71   239.2451     0.01    0.994    -472.7173    476.1468
-----------------------------------------------------------------
```

Illustration of impact of measurement error on estimated coefficients for tax and square footage, their correlation and variance in error in the equation as σ_u ranged from 0 to 200 in steps of 10 (reliability ratio, for square footage alone goes from 1 down to .876.)


tax
σ_u
sqft
σ_u
Error Variance
σ_u
correlation
σ_u