Variable Selection "Insurance"
aka
Valid Statistical Inference After Model Selection

L. Brown
Wharton School, Univ. of Pennsylvania

with

Richard Berk, Andreas Buja, Ed George, Michael Traskin, Linda Zhao, Kai Zhang, Emil Pitkin

Purdue: June 23, 2012

Subtext:

What we think we know (or think we have learned)

may not be so.

Plagiarized from?

Maybe from John Ruskin?:

What John Ruskin actually said was:

"What we think, or what we know, or what we believe is, in the end, of little consequence. The only consequence is what we *do.*"

BUT Forbes magazine did write (4/12/2012)

"What we know, and what we think we know"
(about the new IPhone)

Probably from Will Rogers:

"It isn't what we don't know that gives us trouble,

it's what we know that ain't so."

Classical Perspective
On Statistical Model Building and Inference

1. Conceptualize Problem
2. Build MODEL for inferential analysis
3. Gather and process Data
4. Produce Inference

Standard inference (estimates, tests and CIs) is based on this schematic plan.

Contemporary Pragmatism

1. Conceptualize Problem
2. Gather and process Data
3. Build MODEL for inferential analysis
4. Produce Inference

Note reversal of steps #2 and #3.

- There is no currently accepted theory that applies when the steps are carried out this way.
- But the risks of doing so have long been discussed in the statistical literature.

References later…

Layout of Talk

1. Introduction: The dangers of variable selection
   Berk, et al. (2010), Berk et al. (2012b)

2. Insurance Plan #1: "POSI"
   Berk, et al (2012a), in revision

3. Insurance Plan #2: Split sample bootstrap
   Berk, et al (2012c), in preparation

Part 1: Examples

Is Race a factor in criminal sentencing,
after controlling for concomitant factors?

Data: Random sample of 500 criminal sentences (in years) with Race of individual, Sex, Marital status, etc + severity of crime (on a standard numeric scale, 3 = most severe, 60 = least severe) + #of prior criminal charges + prior record of drug crime, or psychological problems, or alcohol use

Outcome variable: $Y = \text{Log}_e(\text{sentence} + 0.5)$

Regression of Y on Race

Parameter Estimates

Term	**Estimate**	**Std Error**	**t Ratio**	**Prob>|t|**
Race [B vs W = ±1]	0.142	0.0940	1.51	**0.1321**

By itself, race does not have a significant (linear) effect.

But this is not really the question of interest!

Better procedure – a regression of Y on all variables:

Regression of Y on All Variables

RSquare 0.1832

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	4.229	0.3489	12.12	<.0001*
Race	0.118	0.0917	1.28	**0.1999**
intage	-0.031	0.0057	-5.37	<.0001*
sex (1=Male)	-0.124	0.145	-0.85	0.3933
Married?	-0.072	0.1212	-0.59	0.5521
Employed?	-0.591	0.1557	-3.80	0.0002*
priors	0.0138	0.0032	4.28	<.0001*
seriousness	-0.0165	0.0042	-3.89	0.0001*
HS educ	0.0841	0.1355	0.62	0.5350
psych	-0.574	0.4128	-1.39	0.1648
drugs	0.535	0.1642	3.26	0.0012*
alco	-0.409	0.1673	-2.44	0.0149*

Conclusion: After controlling on all other var's in the study, **Race** is (still) **not stat sig**.

Best Procedure (??) = :

Regression of Y on Race plus

other covariates that strongly influence the race effect

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	3.99	0.2499	15.97	<.0001*
Race	0.219	0.0920	2.38	**0.0179***
intage	-0.0307	0.00569	-5.40	<.0001*
seriousness	-0.0135	0.00441	-3.05	0.0024*
drugs	0.371	0.1334	2.78	0.0056*

Classical **Conclusion**: Race **is** a significant factor in sentencing, after controlling for relevant covariates in the study.

(P-Value = 0.018)

Other Objectives

Such a Y on X model might be analyzed for different reasons.

For example --

Objective: Find the X-factors that collectively significantly influence Y and/or Best model for predicting Y

- Model Selection Procedure: All subsets BIC

Result: **4** variables in model (intage, Employed?, seriousness and priors). **All P-values <0.001**.

- Model Selection Procedure: All subsets AIC.

Result: **7** factor model (4 factors of BIC model & 3 others.
Not "Race" in either model.
But "psych" has **P-value <0.001** in 7 factor model.)

Question: Are the conventional P-values in these models "legitimate"?

Answer: **NO!!**
We need variable selection insurance.

How badly mis-aligned can conventional analysis be when applied after variable selection?

A simulation example

- Model: Balanced two-way random effects model. p = 10 blocks. Error df $\triangleq d \approx \infty$.
- Model selection over Treatments $\times$ Blocks interactions.
- "Best" model should control for all Blocks that interact with treatment/control contrast
- Special assumptions:

  Error variance known (= 1) [since $d \approx \infty$] &
  First block (a dummy) has no interaction term.
- Model selection, $\hat{M}$:

  Include treatment effect & all interaction terms that are significant in the full analysis at level $\alpha = 0.05$.

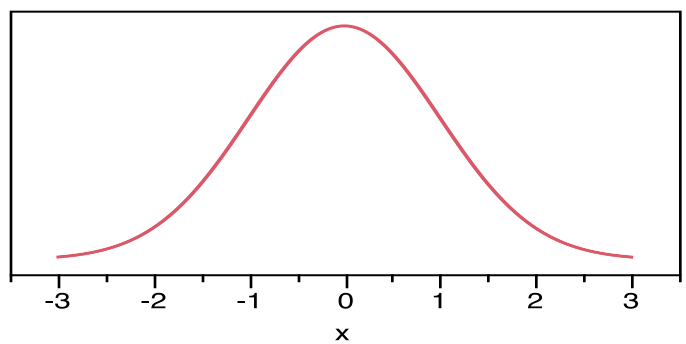
"Conventional" Analysis

The conventional t-statistic for model $\hat{M}$ is

$$t = \frac{\hat{\beta}_{t\bullet\hat{M}} - \beta_{t\bullet\hat{M}}}{\text{StdError}\left(\hat{\beta}_{t\bullet\hat{M}}\right)}.$$

<u>IF</u> $\hat{M}$ had been chosen without reference to the data, then t would have the t-distribution with d deg of freedom.

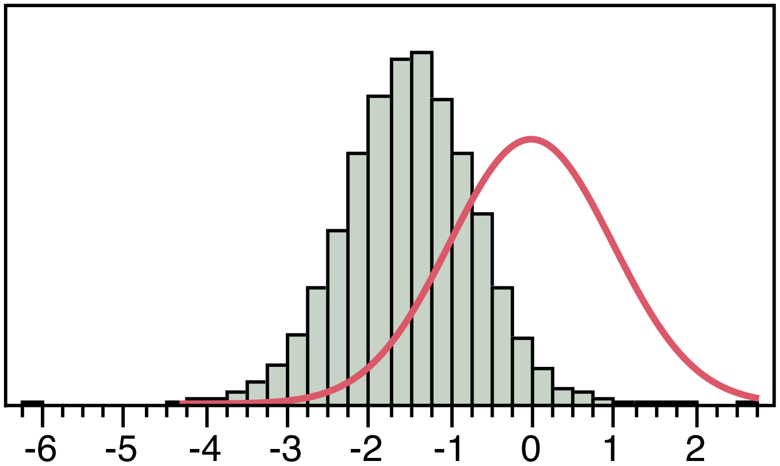
By simplifying assumption, $d \approx \infty$. So histogram of t would then be



And, $|t| < 1.96$ (approx.) means true $\beta_{t\bullet\hat{M}} \in 95\%$ conf. interval.

Actually --

The data IS used to help select the model. The histogram of $t = \dfrac{\hat{\beta}_{t\bullet\hat{M}} - \beta_{t\bullet\hat{M}}}{\text{StdError}\left(\hat{\beta}_{t\bullet\hat{M}}\right)}$ (via a large simulation) is



Prob (95% CI covers true value) ≈ 0.725

For situations with larger *p* the disparity can be greater. For other sampling designs and objectives it can be less or somewhat more. [For p = 30 coverage can be low as 39%.]

Part 2

- Described in the following is our "PoSI" algorithm, and the associated computational method for computing the PoSI constant.

- Formulation conditional on Design matrix

$$\mathbf{X} = \left\{ x_{ij} : i = 1,..,n,\ j = 1,..,p \right\}$$

i.e., Results are for "Fixed **X** " formulation;

and also hold, via conditioning, even when **X** is random

- Pre model-selection observations are assumed as

$$Y \sim N_n\left(\mathbf{X}\beta, \sigma^2 I\right), \ \mathbf{X} \text{ is } n \times p \text{ and full rank}.$$

Convention: If “intercept” not of model selection interest, assume the columns of **X** have been centered.

[Non-normal models are also of interest, but not treated in this part.]

- **Also assume**, $\boldsymbol{n > p}$.
- Then $\hat{\sigma}^2 = MSE_{\text{fullmodel}}$

is a valid estimate of σ^2, free of any model selection effects.

- A (sub)Model, *M*, is a subset of $\{1,..,p\}$, and leads to

$$\mathbf{X}_M = \left[X_{(j)} : j \in M\right]$$

$X_{(j)}$ denotes the *j*-th column of $\mathbf{X}$. $\mathbf{X}_M$ is an $n \times (\# M)$ matrix.

Prologue:
Parameters for a given (sub) model, *M*

- Denote the corresponding, usual LS estimate by

$$\hat{\beta}_{\bullet M} = \left(\mathbf{X}'_M \mathbf{X}_M\right)^{-1} \mathbf{X}'_M Y.$$

Coordinates of $\hat{\beta}_{\bullet M}$ are $\hat{\beta}_{k\bullet M}$, $k \in M$.

- Then

$$\hat{\beta}_{k\bullet M} \triangleq \ell'_{k\bullet M} Y$$

Note $\ell_{k\bullet M} = x_{k\bullet M} \Big/ \left\| x_{k\bullet M} \right\|^2$, where $x_{k\bullet M}$ is the residual vector of x_k from $\mathrm{ColSp}\left(X_{M-\{k\}}\right)$.

Tests and CIs

- Conventional test of $H_0 : \beta_{k\bullet M} = 0$ is based on

$$t_{k\bullet M} = \frac{\ell'_{k\bullet M} Y}{\left\|\ell_{k\bullet M}\right\| \hat{\sigma}}$$

with Student's-t null distribution & $n - p$ df.

- Conventional CI is

$$\mathrm{CI}_{k\bullet M} \triangleq \hat{\beta}_{k\bullet M} \pm \mathrm{t}_{n-p;1-\alpha/2}\, \hat{\sigma} \Big/ \left\|x_{k\bullet M}\right\| .$$

Note: $n - p$ df, not $n - \#M$. If σ^2 is known, replace $\hat{\sigma}$ by σ and **t** by Z .

Meaning of Correlation Coefficients within *M*

Description #1

- For given *M* define $\beta_{k\bullet M}$ by

$$\beta_{k\bullet M} = E\left(\hat{\beta}_{k\bullet M}\right) = \ell'_{k\bullet M}\mathbf{X}\beta.$$

The interpretation of each coefficient in *M* depends on which other coefficients are in *M*.

[Exception: If the col's of **X** are orthogonal then Sub-model parameters = Full-model parameters --- $\beta_{k\cdot M} = \beta_k \;\; \forall k \in M$]

Alternate Descriptions:

- **Description #2:** $\beta_{\bullet M}$ provides the linear "slopes" within the model with predictors X_M -- ie,

$$\beta_{\bullet M} = \arg\min_{\vec{b}} E\left(Y - X_M \vec{b}\right)^2$$

So (within the linear model) -- $\beta_{k\bullet M}$ *is the expected change in E(Y) for a unit change in* x_k *when the other factors in M are held fixed.*

- **Description #3:** When the chosen model is proposed for future use, the $\beta_{k\bullet M}$ are the coefficients for prediction of *Y*.

i.e., The $\beta_{k\bullet M}$ *are the coefficients for <u>prediction</u> when the chosen model is being used.*

Model Selection

The data is examined; a "model" $\hat{M} = M\left(Y\right)$ is chosen.

[Model selection here is (only) about choice of predictor variables, not about – *e.g.* – transformation of *Y*.]

This yields a post-selection design –

$$\mathbf{X}_{\hat{M}} = \text{the columns of } \mathbf{X} \text{ with indices in } \hat{M}.$$

Conventional inference after Model Selection is Invalid:

Typically

$$P\left(\beta_{k \bullet \hat{M}} \in \mathrm{CI}_{k \bullet \hat{M}}\right) < 1 - \alpha$$

[instead of desired $\geq 1 - \alpha$].

We propose to construct **Po**st **S**election **I**nference with valid tests and multiple confidence statements (Family Wise Error Rate).

Some Digressions

1. Bock, Judge, et al and also Sclove, et al (1970s) looked at “pre-test estimators”. Although a sub-model is involved in these estimators, they are actually estimators of the full parameter vector, β, and not of the parameters of the sub-model $\hat{M}_{\neq 0} = \left\{ k : \hat{\beta}_k \neq 0 \right\}$.

2. Similarly, Lasso and other penalization algorithms should be viewed as giving estimators of the full parameter vector, rather than as estimators within the sub-model of parameters estimated to be $\neq 0$.

3. The problem has been known for decades:

Koopmans (1949); Buehler and Fedderson (1963); Brown (1967); and Olshen (1973); Mosteller and Tukey (1977); Sen (1979); Sen and Saleh (1987); Dijkstra and Veldkamp (1988); Arabatzis et al. (1989); Hurvich and Tsai (1990); Regal and Hook (1991); Pötscher (1991); Chiou and Han (1995a,b); Giles (1992); Giles and Srivastava (1993); Kabaila (1998); Brockwell and Gordeon (2001); **Leeb and Pötscher** (2003; 2005; 2006a; 2006b; 2008a; 2008b); Kabaila (2005); Kabaila and Leeb (2006): Berk, Brown and Zhao (2009); Kabaila (2009).

Ahrens, C., Altman, N., Casella, G., Eaton, M., Hwang, J.T.G, Staudenmeyer, J. and Stefanescu, C. (2001). Leukemia Clusters and TCE Wastesites in Upstate New York; How Adding Covariates Changes the Story. *Environmetrics* ***12*** 659-672.

---

ie: **The interpretation of each coefficient depends on which other coefficients are in *M*.**

PoSI Algorithm

Define constant **K** so that $\forall\ \hat{M}$ the CI of the form

(CI*) $$\mathrm{CI}^*_{k\bullet M} \triangleq \hat{\beta}_{k\bullet\hat{M}} \pm \mathbf{K}\hat{\sigma} \big/ \left\| x_{k\bullet\hat{M}} \right\|$$

satisfies

(*) $$P_\beta\left(\beta_{k\bullet\hat{M}} \in \mathrm{CI}^*_{k\bullet\hat{M}} \text{ for all } k \in \hat{M}\right) \geq 1-\alpha\,.$$

K allowed to depend on α, p, $n-p$ and $\mathbf{X}$.

But K does not depend on the rule leading to $\hat{M}$, or on $\hat{M}$ itself. So (*) true for all β, all rules $\hat{M}(Y)$, all $k \in \hat{M}$.

- A restricted version of (*) – call it (*k) – is also of interest. This is (*) but only for a previously fixed k, under the restriction $\hat{M} \supset k$. [ie, not family-wise]

Key Equation (**), below

Recall –requirement is that $\forall$ β and $\forall\ \hat{M}$

(*) $$P_\beta\left(\beta_{k\bullet\hat{M}} \in \mathrm{CI}^*_{k\bullet\hat{M}} \text{ for all } k \in \hat{M}\right) \geq 1-\alpha .$$

Linearity and normality of $\hat{\beta}_{k\bullet M}$ and centered-ness yield --

$\mathrm{CI}^*_{k\bullet\hat{M}}$ (*) is implied by

(**) $$1-\alpha \leq P_{\mathbf{0}}\left[\max_{M\,;\,k\in M}\left|t_{k\bullet M}\right| \leq \mathrm{K}\right].$$

As a reminder, $t_{k\bullet M} = \dfrac{\ell'_{k\bullet M}Y}{\left\|\ell_{k\bullet M}\right\|\hat{\sigma}}$.

Canonical Form

Rotation of Y reduces problem to canonical form with new $X, Y, \hat{\sigma}^2$ without affecting meaning of β or M (*eg*, *TSH*, Chapter 7), where now

$$\textbf{(CF)} \quad \mathbf{X} \sim p \times p, Y \sim N_p\left(\mathbf{X}\beta, \sigma^2 I_p\right), \left(n-p\right)\hat{\sigma}^2 \sim \sigma^2 \chi^2_{n-p}.$$

PoSI is Possible

Review: PoSI needs $\mathbf{K}$ such that

(**) $$1-\alpha \le P_0\left[\max_{M;\,k\in M}\left|t_{k\bullet M}\right| \le \mathbf{K}\right].$$

Theorem: Scheffe's $\mathrm{K}_{\mathrm{S}} = \sqrt{pF_{p,n-p;1-\alpha}}$ satisfies (**).

Proof: By its construction

$$1-\alpha = P_0\left[\max_{\mathbf{c}} \frac{\mathbf{c}'Y}{\|\mathbf{c}\|\hat{\sigma}} \le \mathrm{K}_{\mathrm{S}}\right] < P_0\left[\max_{M;\,k\in M} \frac{\ell'_{k\bullet M}Y}{\left\|\ell_{k\bullet M}\right\|\hat{\sigma}} \le \mathrm{K}_{\mathrm{S}}\right]$$

- K_{S} may give very conservative CIs. (Inequality can be Big.)
- *It's possible to do better*. Here's our plan:

Our Proposal for PoSI

- For fixed α, *p*, *n* – *p* and full-design *X* ***computationally*** find **K** = K(*X*) such that (**) holds – ie,

$$1-\alpha = P_0\left[\max_{M;\,k\in M}\left|t_{k\bullet M}\right| \le \mathbf{K}\right].$$

- For modest *p* we can always do so by simulating the null distribution of *Y* under $\beta = 0$, and using Monte-Carlo.

[Computational limitation: the max step involves looking at $p2^{p-1}$ possibilities. Clever, "naive" version of this requires (approx) $p \le 20$.]

- **Alternatives to "naïve" computation are possible!**

“SPAR” (and “SPAR(*k)”)

- A model-selection routine for which (**) is sharp --

$$P_0\left[\max_{M;\,k\in M}\left|t_{k\bullet M}\right|\le \mathrm{K}\right]=P_0\left[\left|t_{k\bullet \hat{M}}\right|\le \mathrm{K}\right]$$

This is “**S**ingle **P**arameter **A**djusted **R**egression”; formally

$$\hat{M}_{\text{spar}}=\left\{\hat{M}=\hat{M}\left(Y\right):\exists\hat{k}\in\hat{M}\ni\left|t_{\hat{k}\bullet\hat{M}}\right|=\max_{M,k\in M}\left|t_{k\bullet M}\right|\right\}.$$

- Though artificial, this approximates what a naive scientist might do - one who combs a large data set looking for the most “publishable” result.
- A modification of this is somewhat more plausible in the setting (*k). Here one first settles on a co-variate of principle interest, and looks for the set of control variates that make this have the largest apparent effect after including those controls.

Bounds on PoSI **K** (for known $\sigma^2 = 1$)

- Lower Bound:

$$\mathrm{K}\left(\mathbf{X}\right) \geq \mathrm{K}\left(\mathbf{I}\right) = \Phi^{-1}\left(\left(1+\left(1-\alpha\right)^{1/p}\right)\Big/2\right).$$

$$\sim \sqrt{2\log p} \text{ as } p \to \infty,\ n-p \to \infty,\ \alpha \text{ fixed}$$

- Upper Bound:

$$.6363\sqrt{p} \leq \sup_{\mathbf{X}} \mathrm{K}\left(\mathbf{X}\right) \leq \sqrt{p}\sqrt{2/\pi} + o\left(\sqrt{p}\right) < \mathrm{K}_S \sim \sqrt{p}.$$

- Moral from comparison of these bounds:
  Calculation of $\mathrm{K}\left(\mathbf{X}\right)$ matters since the value can turn

out anywhere from about $\sqrt{2\log p}$ to about $0.6363\sqrt{p}$.
(Or, maybe upper bound is as large as $\sqrt{2/\pi}\sqrt{p}$.)

Part 3:

Random Design;

Linear Estimation
– but not necessarily a linear model

Inference involves a special Split-Sample bootstrap

Observational Data

Sample:

$$\left(X_i, Y_i\right) \ iid \ \ i = 1,..,n$$

with $X_i \in \Re^m$ being absolutely cont. (for simplicity) and

$$\mathrm{E}\left(Y \middle| X = x\right) = \mu\left(x\right), \ \mathrm{var}\left(Y \middle| X = x\right) = \sigma^2\left(x\right) < \infty .$$

Objective: (*observe the data and*)

- Create a linear model to "explain" the data and/or
- To predict future observations
- Make inference for the slope coefficients in that model.

Our focus:

- Produce valid inference for these coefficients. *ie,*

Slope Coefficients for Linear Analytical Models:

- Define the "BestLinearApproximation" coefficients $\beta = \left(\beta_1, .., \beta_m\right)'$:

$$\beta = \beta^M = \arg\min_{\vec{b}} \left\{ \mathrm{E}\left(\left(Y - X'\vec{b} \right)^2 \right) \right\}.$$

- Formula:

$$\beta = \left[\mathrm{E}\left(XX' \right) \right]^{-1} \mathrm{E}\left(X'Y \right)$$

The Split-Sample Bootstrap

- Begin with a sample $S = \{(X_i, Y_i)\}$.
- Split it at random into two disjoint parts – S_{Mod}, S_{Inf} .
- Use S_{Mod} to construct linear model covariates, $X(W)$.
- Then apply a bootstrap to S_{Inf} with these covariates.
  (details on next overhead)
- This gives asymptotically valid confidence intervals (and estimates) of the corresponding BLA coefficients.
- Asymptotics are valid (as $n \to \infty$ for fixed p)[1]
  - under mild moment conditions on S, and
  - uniformly over all distributions satisfying these conditions. (*This uniformity is important.*)
  - *Allowing p to grow slowly with n may also be OK*.

1. Key theory comes from Mammen (1993).

Details of the Bootstrap

- Calculate the LSE $\hat{\beta}$ within S_{Inf}
- Generate a resample (with replacement) of size n_{Inf} from.
- For this resample calculate LSE $\hat{\beta}^*$.
- Repeat K times (large #) to get $\left\{\hat{\beta}^{*k} : k = 1,..,K\right\}$.
- Create the histogram of $\left\{\hat{\beta}^{*k} - \hat{\beta} : k = 1,..,K\right\}$.
- The quantiles of this histogram estimate the quantiles of the true CDF of $\hat{\beta} - \beta$.
- Invert this estimated CDF to get confidence intervals.
- Use the median (or the mean) of the estimated CDF as an estimate.
- Simultaneous intervals (FWER) can be estimated from the p-dimensional CDF of the bootstrap sample.

Remarks

- This split sample procedure sacrifices a portion of the data to the sole purpose of model selection.
- Consequently the inference is based on fewer than n observations. (And also the model selection.)
- How to balance sample sizes between S_{Mod} and S_{Inf} is under investigation.
- Scientific embarrassment is quite possible. – e.g.
  - The statistician may choose a model from S_{Mod} claimed to contain only significant linear factors.
  - Then find in the bootstrap on S_{Inf} that some (or all) of these factors are not significantly different from 0!
- InsuranceAgainstEmbarrassment *(when the setting allows)* would be provided by selection using POSI on S_{Mod}, followed by the bootstrap on S_{Inf}.

Results with the two insurance plans in Criminology Data

- For Plan #2 (Bootstrap) we took the first portion of the data as 250 of the original 500 observations
- Then we performed an "all subsets BIC" variable selection, and ended up with 7 variables in the model
- The second portion of the data was used to construct "bootstrap" confidence intervals
- These were compared with the POSI intervals (=Plan #1) for the same model. These can be, and were, constructed from the full 500 observations
- Here are the resulting 95% confidence intervals

- All intervals #1 and #2 considerably overlap (as they should)
- Surprisingly (to us) CI#1 are generally shorter than #2
- "BIC" claims to yield only significant variables, **but**
- Only 4 variables with #1 are stat'ly significant at 0.05; and only 1 variable with #2. (This 'price of insurance' realistically reflects the fact that "All subsets BIC" is actually too greedy, and may choose non-significant variables.)

Chosen Variable	**CI#1**	**CI #2**
Seriousness	[-0.024, -0.001]*	[-0.021, 0.005]
Prior Drug use	[-0.024, 0.859]	[-0.341, 0,818]
Prior Alcohol use	[-0.912, -0.014]*	[-0.772, 0.379]
# of Prior records	[0.004, 0.021]*	[0.001, 0.023]
Age at incidence	[-0.037, -0.008]*	[-0.040, -0.009]*
Gender	[-0.189, 0.584]	[-0.467, 0.505]
Employment (Y or N)	[-0.805, 0.016]	[-0.770, 0.312]

Summary

Main themes:

1. Examples to remind that analysis involving both model selection and inference is problematic

- Main example involved question of effect on sentencing of Race after controlling for other relevant factors.

2. POSI methodology.

- Creates simultaneously valid CIs for all factors in the model. These do not depend on the model selection methodology or the selected model. Except in least favorable case they will be conservative. Can be modified to treat case of a single variable of interest (as in the criminology race example).

3. Bootstrap methodology.

- Splits the sample in an unconventional manner, into parts (1)model selection and (2)inference portions. Applies the bootstrap on the inference portion, using the selected model.