Homoskedastic adjustment inflation factors in model selection

With examples from the Avon Longitudinal Study of Parents and Children (ALSPAC) cohort study at Bristol University, UK
http://www.bristol.ac.uk/alspac/

Roger B. Newson
r.newson@imperial.ac.uk
http://www.imperial.ac.uk/nhli/r.newson/

National Heart and Lung Institute
Imperial College London

15th UK Stata Users' Group Meeting, 10–11 September, 2009
Downloadable from the conference website at
http://ideas.repec.org/s/boc/usug09.html

Concomitant variables in observational studies

- In an observational study, the **outcome** variable is the variable that we would like to change. (Such as asthma or lung capacity in children.)
- An **exposure** variable is a variable that we might propose to change (or fantasize about changing), in order to cause a change in the outcome. (Such as smoking or paracetamol use during pregnancy.)
- Other variables included in the model are known as **concomitant** variables. (Such as housing tenure, income, and education level.)
- They should have the feature that we do not expect them to be changed by our proposed (or fantasized) intervention to change the exposure.
- We aim to estimate the effect of changing the exposure by comparing the outcome in subjects with different exposure levels and the same concomitant values.

Concomitant variables in observational studies

- In an observational study, the **outcome** variable is the variable that we would like to change. (Such as asthma or lung capacity in children.)
- An **exposure** variable is a variable that we might propose to change (or fantasize about changing), in order to cause a change in the outcome. (Such as smoking or paracetamol use during pregnancy.)
- Other variables included in the model are known as **concomitant** variables. (Such as housing tenure, income, and education level.)
- They should have the feature that we do not expect them to be changed by our proposed (or fantasized) intervention to change the exposure.
- We aim to estimate the effect of changing the exposure by comparing the outcome in subjects with different exposure levels and the same concomitant values.

Concomitant variables in observational studies

- In an observational study, the **outcome** variable is the variable that we would like to change. (Such as asthma or lung capacity in children.)
- An **exposure** variable is a variable that we might propose to change (or fantasize about changing), in order to cause a change in the outcome. (Such as smoking or paracetamol use during pregnancy.)
- Other variables included in the model are known as **concomitant** variables. (Such as housing tenure, income, and education level.)
- They should have the feature that we do not expect them to be changed by our proposed (or fantasized) intervention to change the exposure.
- We aim to estimate the effect of changing the exposure by comparing the outcome in subjects with different exposure levels and the same concomitant values.

Concomitant variables in observational studies

- In an observational study, the **outcome** variable is the variable that we would like to change. (Such as asthma or lung capacity in children.)
- An **exposure** variable is a variable that we might propose to change (or fantasize about changing), in order to cause a change in the outcome. (Such as smoking or paracetamol use during pregnancy.)
- Other variables included in the model are known as **concomitant** variables. (Such as housing tenure, income, and education level.)
- They should have the feature that we do not expect them to be changed by our proposed (or fantasized) intervention to change the exposure.
- We aim to estimate the effect of changing the exposure by comparing the outcome in subjects with different exposure levels and the same concomitant values.

Concomitant variables in observational studies

- In an observational study, the **outcome** variable is the variable that we would like to change. (Such as asthma or lung capacity in children.)
- An **exposure** variable is a variable that we might propose to change (or fantasize about changing), in order to cause a change in the outcome. (Such as smoking or paracetamol use during pregnancy.)
- Other variables included in the model are known as **concomitant** variables. (Such as housing tenure, income, and education level.)
- They should have the feature that we do not expect them to be changed by our proposed (or fantasized) intervention to change the exposure.
- We aim to estimate the effect of changing the exposure by comparing the outcome in subjects with different exposure levels and the same concomitant values.

Concomitant variables in observational studies

- In an observational study, the **outcome** variable is the variable that we would like to change. (Such as asthma or lung capacity in children.)
- An **exposure** variable is a variable that we might propose to change (or fantasize about changing), in order to cause a change in the outcome. (Such as smoking or paracetamol use during pregnancy.)
- Other variables included in the model are known as **concomitant** variables. (Such as housing tenure, income, and education level.)
- They should have the feature that we do not expect them to be changed by our proposed (or fantasized) intervention to change the exposure.
- We aim to estimate the effect of changing the exposure by comparing the outcome in subjects with different exposure levels and the same concomitant values.

What concomitant variables should we include?

- In the epidemiology sector, most people nowadays think that there is a historic culture of under–adjustment. (See, for instance, Davey Smith and Ebrahim (2002)[1].)
- *Therefore*, if we have a large cohort study with many concomitant variables, then we may instinctively use large confounder sets in order to be safe.
- *However*, we are then likely to be accused by journal referees of "over–adjusting".
- It is not clear what these referees mean, except that we should have adjusted for fewer concomitants.

What concomitant variables should we include?

- In the epidemiology sector, most people nowadays think that there is a historic culture of under–adjustment. (See, for instance, Davey Smith and Ebrahim (2002)[1].)
- *Therefore*, if we have a large cohort study with many concomitant variables, then we may instinctively use large confounder sets in order to be safe.
- *However*, we are then likely to be accused by journal referees of "over–adjusting".
- It is not clear what these referees mean, except that we should have adjusted for fewer concomitants.

What concomitant variables should we include?

- In the epidemiology sector, most people nowadays think that there is a historic culture of under–adjustment. (See, for instance, Davey Smith and Ebrahim (2002)[1].)
- *Therefore*, if we have a large cohort study with many concomitant variables, then we may instinctively use large confounder sets in order to be safe.
- *However*, we are then likely to be accused by journal referees of "over–adjusting".
- It is not clear what these referees mean, except that we should have adjusted for fewer concomitants.

What concomitant variables should we include?

- In the epidemiology sector, most people nowadays think that there is a historic culture of under–adjustment. (See, for instance, Davey Smith and Ebrahim (2002)[1].)
- *Therefore*, if we have a large cohort study with many concomitant variables, then we may instinctively use large confounder sets in order to be safe.
- *However*, we are then likely to be accused by journal referees of "over–adjusting".
- It is not clear what these referees mean, except that we should have adjusted for fewer concomitants.

What concomitant variables should we include?

- In the epidemiology sector, most people nowadays think that there is a historic culture of under–adjustment. (See, for instance, Davey Smith and Ebrahim (2002)[1].)
- *Therefore*, if we have a large cohort study with many concomitant variables, then we may instinctively use large confounder sets in order to be safe.
- *However*, we are then likely to be accused by journal referees of "over–adjusting".
- It is not clear what these referees mean, except that we should have adjusted for fewer concomitants.

So what do we mean by "over–adjusting"?

I would argue that "over–adjusting" can mean one of two completely different things:

- **Some of the concomitants are causally downstream from the exposure.** For instance, if we think that a proposed intervention to reduce smoking exposure during pregnancy will have the side effect of increasing birthweight, then we should not include birthweight as a concomitant, when estimating the effect of this intervention on child lung capacity at 7 years of age.
- **The concomitants predict the exposure "too well".** This may cause loss of power to detect exposure effects, especially if the number of concomitants becomes too close to the number of study subjects.

In this presentation, we focus on the second problem.

So what do we mean by "over–adjusting"?

I would argue that "over–adjusting" can mean one of two completely different things:

- **Some of the concomitants are causally downstream from the exposure.** For instance, if we think that a proposed intervention to reduce smoking exposure during pregnancy will have the side effect of increasing birthweight, then we should not include birthweight as a concomitant, when estimating the effect of this intervention on child lung capacity at 7 years of age.
- **The concomitants predict the exposure "too well".** This may cause loss of power to detect exposure effects, especially if the number of concomitants becomes too close to the number of study subjects.

In this presentation, we focus on the second problem.

So what do we mean by "over–adjusting"?

I would argue that "over–adjusting" can mean one of two completely different things:

- **Some of the concomitants are causally downstream from the exposure.** For instance, if we think that a proposed intervention to reduce smoking exposure during pregnancy will have the side effect of increasing birthweight, then we should not include birthweight as a concomitant, when estimating the effect of this intervention on child lung capacity at 7 years of age.
- **The concomitants predict the exposure "too well".** This may cause loss of power to detect exposure effects, especially if the number of concomitants becomes too close to the number of study subjects.

In this presentation, we focus on the second problem.

So what do we mean by "over–adjusting"?

I would argue that "over–adjusting" can mean one of two completely different things:

- **Some of the concomitants are causally downstream from the exposure.** For instance, if we think that a proposed intervention to reduce smoking exposure during pregnancy will have the side effect of increasing birthweight, then we should not include birthweight as a concomitant, when estimating the effect of this intervention on child lung capacity at 7 years of age.
- **The concomitants predict the exposure "too well".** This may cause loss of power to detect exposure effects, especially if the number of concomitants becomes too close to the number of study subjects.

In this presentation, we focus on the second problem.

So what do we mean by "over–adjusting"?

I would argue that "over–adjusting" can mean one of two completely different things:

- **Some of the concomitants are causally downstream from the exposure.** For instance, if we think that a proposed intervention to reduce smoking exposure during pregnancy will have the side effect of increasing birthweight, then we should not include birthweight as a concomitant, when estimating the effect of this intervention on child lung capacity at 7 years of age.
- **The concomitants predict the exposure "too well".** This may cause loss of power to detect exposure effects, especially if the number of concomitants becomes too close to the number of study subjects.

In this presentation, we focus on the second problem.

Problems with "stepwise" variable selection

- These are discussed extensively (with references) at *http://www.stata.com/support/faqs/stat/stepwise.html*.
- The main problem is that confidence interval formulas do *not* cover us for finding a model in the data from which the parameters of that model will later be estimated.
- *However*, they do cover us for finding a model in the sample distribution of the exposure and the concomitants ("Step 1").
- We can then estimate the parameters of that model from the *conditional* distribution of the outcome, given the exposure and the concomitants ("Step 2").
- In this presentation, we focus on Step 1 of this strategy.

Problems with "stepwise" variable selection

- These are discussed extensively (with references) at *http://www.stata.com/support/faqs/stat/stepwise.html*.
- The main problem is that confidence interval formulas do *not* cover us for finding a model in the data from which the parameters of that model will later be estimated.
- *However*, they do cover us for finding a model in the sample distribution of the exposure and the concomitants ("Step 1").
- We can then estimate the parameters of that model from the *conditional* distribution of the outcome, given the exposure and the concomitants ("Step 2").
- In this presentation, we focus on Step 1 of this strategy.

Problems with "stepwise" variable selection

- These are discussed extensively (with references) at *http://www.stata.com/support/faqs/stat/stepwise.html*.
- The main problem is that confidence interval formulas do *not* cover us for finding a model in the data from which the parameters of that model will later be estimated.
- *However*, they do cover us for finding a model in the sample distribution of the exposure and the concomitants ("Step 1").
- We can then estimate the parameters of that model from the *conditional* distribution of the outcome, given the exposure and the concomitants ("Step 2").
- In this presentation, we focus on Step 1 of this strategy.

Problems with "stepwise" variable selection

- These are discussed extensively (with references) at *http://www.stata.com/support/faqs/stat/stepwise.html*.
- The main problem is that confidence interval formulas do *not* cover us for finding a model in the data from which the parameters of that model will later be estimated.
- *However*, they do cover us for finding a model in the sample distribution of the exposure and the concomitants ("Step 1").
- We can then estimate the parameters of that model from the *conditional* distribution of the outcome, given the exposure and the concomitants ("Step 2").
- In this presentation, we focus on Step 1 of this strategy.

Problems with "stepwise" variable selection

- These are discussed extensively (with references) at *http://www.stata.com/support/faqs/stat/stepwise.html*.
- The main problem is that confidence interval formulas do *not* cover us for finding a model in the data from which the parameters of that model will later be estimated.
- *However*, they do cover us for finding a model in the sample distribution of the exposure and the concomitants ("Step 1").
- We can then estimate the parameters of that model from the *conditional* distribution of the outcome, given the exposure and the concomitants ("Step 2").
- In this presentation, we focus on Step 1 of this strategy.

Problems with "stepwise" variable selection

- These are discussed extensively (with references) at *http://www.stata.com/support/faqs/stat/stepwise.html*.
- The main problem is that confidence interval formulas do *not* cover us for finding a model in the data from which the parameters of that model will later be estimated.
- *However*, they do cover us for finding a model in the sample distribution of the exposure and the concomitants ("Step 1").
- We can then estimate the parameters of that model from the *conditional* distribution of the outcome, given the exposure and the concomitants ("Step 2").
- In this presentation, we focus on Step 1 of this strategy.

Homoskedastic adjustment inflation factors: the `haif` package

- The `haif` package, downloadable from SSC, is a more comprehensive version of `estat vif`.
- It inputs a core variable list, defining a core design matrix X, and an additional variable list, defining an additional submatrix A.
- It outputs the **homoskedastic adjustment inflation factors (HAIFs)**, by which the variances and standard errors of the coefficients for the X–variables are scaled (or inflated) by adjusting additionally for the A–variables, assuming X to be the true design matrix.
- Note that these factors are calculated assuming (a) that the A–variables have no independent effect on the mean of the outcome, and (b) that the variance of the outcome is not affected either by the X-variables or by the A-variables (or, in other words, that the outcome is homoskedastic).
- *Therefore*, the HAIFs represent a "worst case" scenario, assuming that the A–variables are not really necessary.

Homoskedastic adjustment inflation factors: the `haif` package

- The `haif` package, downloadable from SSC, is a more comprehensive version of `estat vif`.
- It inputs a core variable list, defining a core design matrix X, and an additional variable list, defining an additional submatrix A.
- It outputs the **homoskedastic adjustment inflation factors (HAIFs)**, by which the variances and standard errors of the coefficients for the X–variables are scaled (or inflated) by adjusting additionally for the A–variables, assuming X to be the true design matrix.
- Note that these factors are calculated assuming (a) that the A–variables have no independent effect on the mean of the outcome, and (b) that the variance of the outcome is not affected either by the X-variables or by the A-variables (or, in other words, that the outcome is homoskedastic).
- *Therefore*, the HAIFs represent a "worst case" scenario, assuming that the A–variables are not really necessary.

Homoskedastic adjustment inflation factors: the `haif` package

- The `haif` package, downloadable from SSC, is a more comprehensive version of `estat vif`.
- It inputs a core variable list, defining a core design matrix X, and an additional variable list, defining an additional submatrix A.
- It outputs the **homoskedastic adjustment inflation factors (HAIFs)**, by which the variances and standard errors of the coefficients for the X–variables are scaled (or inflated) by adjusting additionally for the A–variables, assuming X to be the true design matrix.
- Note that these factors are calculated assuming (a) that the A–variables have no independent effect on the mean of the outcome, and (b) that the variance of the outcome is not affected either by the X-variables or by the A-variables (or, in other words, that the outcome is homoskedastic).
- *Therefore*, the HAIFs represent a "worst case" scenario, assuming that the A–variables are not really necessary.

Homoskedastic adjustment inflation factors: the `haif` package

- The `haif` package, downloadable from SSC, is a more comprehensive version of `estat vif`.
- It inputs a core variable list, defining a core design matrix X, and an additional variable list, defining an additional submatrix A.
- It outputs the **homoskedastic adjustment inflation factors (HAIFs)**, by which the variances and standard errors of the coefficients for the X–variables are scaled (or inflated) by adjusting additionally for the A–variables, assuming X to be the true design matrix.
- Note that these factors are calculated assuming (a) that the A–variables have no independent effect on the mean of the outcome, and (b) that the variance of the outcome is not affected either by the X-variables or by the A-variables (or, in other words, that the outcome is homoskedastic).
- *Therefore*, the HAIFs represent a "worst case" scenario, assuming that the A–variables are not really necessary.

Homoskedastic adjustment inflation factors: the `haif` package

- The `haif` package, downloadable from SSC, is a more comprehensive version of `estat vif`.
- It inputs a core variable list, defining a core design matrix X, and an additional variable list, defining an additional submatrix A.
- It outputs the **homoskedastic adjustment inflation factors (HAIFs)**, by which the variances and standard errors of the coefficients for the X–variables are scaled (or inflated) by adjusting additionally for the A–variables, assuming X to be the true design matrix.
- Note that these factors are calculated assuming (a) that the A–variables have no independent effect on the mean of the outcome, and (b) that the variance of the outcome is not affected either by the X-variables or by the A-variables (or, in other words, that the outcome is homoskedastic).
- *Therefore*, the HAIFs represent a "worst case" scenario, assuming that the A–variables are not really necessary.

Homoskedastic adjustment inflation factors: the `haif` package

- The `haif` package, downloadable from SSC, is a more comprehensive version of `estat vif`.
- It inputs a core variable list, defining a core design matrix X, and an additional variable list, defining an additional submatrix A.
- It outputs the **homoskedastic adjustment inflation factors (HAIFs)**, by which the variances and standard errors of the coefficients for the X–variables are scaled (or inflated) by adjusting additionally for the A–variables, assuming X to be the true design matrix.
- Note that these factors are calculated assuming (a) that the A–variables have no independent effect on the mean of the outcome, and (b) that the variance of the outcome is not affected either by the X-variables or by the A-variables (or, in other words, that the outcome is homoskedastic).
- *Therefore*, the HAIFs represent a "worst case" scenario, assuming that the A–variables are not really necessary.

Example: Adjusting weight effects by car origin in the `auto` data

In the `auto` data, we might want to estimate effect of weight (per pound) on fuel consumption. And we might consider adjusting this effect for origin (US or non–US), which might or might not have independent predictive value. How much power might this lose?

```
. sysuse auto, clear;
(1978 Automobile Data)

. haif weight, addvars(foreign);
Number of observations: 74
Homoskedastic adjustment inflation factors
for variances and standard errors:
                Variance          SE
      weight   1.5418947   1.2417305
       _cons   1.8340183   1.3542593
```

We see that, *if* the variable `foreign` has no independent predictive power, *then* adjusting for it will inflate the confidence interval for the weight effect (per pound) by a factor of 1.24. This could be cancelled out by increasing the sample size by a factor of 1.54, assuming the sample composition to stay the same. (And homoskedasticity.)

Example: Adjusting weight effects by car origin in the `auto` data

In the `auto` data, we might want to estimate effect of weight (per pound) on fuel consumption. And we might consider adjusting this effect for origin (US or non–US), which might or might not have independent predictive value. How much power might this lose?

```
. sysuse auto, clear;
(1978 Automobile Data)

. haif weight, addvars(foreign);
Number of observations: 74
Homoskedastic adjustment inflation factors
for variances and standard errors:
                Variance          SE
      weight   1.5418947   1.2417305
       _cons   1.8340183   1.3542593
```

We see that, *if* the variable `foreign` has no independent predictive power, *then* adjusting for it will inflate the confidence interval for the weight effect (per pound) by a factor of 1.24. This could be cancelled out by increasing the sample size by a factor of 1.54, assuming the sample composition to stay the same. (And homoskedasticity.)

Example: Adjusting weight effects by car origin in the `auto` data

In the `auto` data, we might want to estimate effect of weight (per pound) on fuel consumption. And we might consider adjusting this effect for origin (US or non–US), which might or might not have independent predictive value. How much power might this lose?

```
. sysuse auto, clear;
(1978 Automobile Data)
```

```
. haif weight, addvars(foreign);
Number of observations: 74
Homoskedastic adjustment inflation factors
for variances and standard errors:
                Variance          SE
      weight   1.5418947   1.2417305
       _cons   1.8340183   1.3542593
```

We see that, *if* the variable `foreign` has no independent predictive power, *then* adjusting for it will inflate the confidence interval for the weight effect (per pound) by a factor of 1.24. This could be cancelled out by increasing the sample size by a factor of 1.54, assuming the sample composition to stay the same. (And homoskedasticity.)

Example: Adjusting weight effects by car origin in the `auto` data

In the `auto` data, we might want to estimate effect of weight (per pound) on fuel consumption. And we might consider adjusting this effect for origin (US or non–US), which might or might not have independent predictive value. How much power might this lose?

```
. sysuse auto, clear;
(1978 Automobile Data)

. haif weight, addvars(foreign);
Number of observations: 74
Homoskedastic adjustment inflation factors
for variances and standard errors:
                Variance          SE
      weight   1.5418947   1.2417305
       _cons   1.8340183   1.3542593
```

We see that, *if* the variable `foreign` has no independent predictive power, *then* adjusting for it will inflate the confidence interval for the weight effect (per pound) by a factor of 1.24. This could be cancelled out by increasing the sample size by a factor of 1.54, assuming the sample composition to stay the same. (And homoskedasticity.)

Example: Adjusting weight effects by car origin in the `auto` data

In the `auto` data, we might want to estimate effect of weight (per pound) on fuel consumption. And we might consider adjusting this effect for origin (US or non–US), which might or might not have independent predictive value. How much power might this lose?

```
. sysuse auto, clear;
(1978 Automobile Data)

. haif weight, addvars(foreign);
Number of observations: 74
Homoskedastic adjustment inflation factors
for variances and standard errors:
                Variance           SE
      weight   1.5418947   1.2417305
       _cons   1.8340183   1.3542593
```

We see that, *if* the variable `foreign` has no independent predictive power, *then* adjusting for it will inflate the confidence interval for the weight effect (per pound) by a factor of 1.24. This could be cancelled out by increasing the sample size by a factor of 1.54, assuming the sample composition to stay the same. (And homoskedasticity.)

Formulas for homoskedastic adjustment inflation factors (HAIFs)

- Suppose that X is the core design matrix, A is the additional–variables matrix, and D is the diagonal matrix of weights.
- Then the variance HAIF for the kth column of X is a ratio, whose denominator is the kth diagonal element of
$$(X'DX)^{-1}$$
and whose numerator is the kth diagonal element of
$$[(X,A)'D(X,A)]^{-1}$$
where X,A is the horizontal concatenation matrix of X and A.
- The kth standard error HAIF is the square root of the kth variance HAIF.
- The weight matrix D is either the default identity matrix, or a diagonal matrix containing inverse variance weights.
- In the second case, the HAIF is a *heteroskedastic* adjustment inflation factor, assuming that we guessed the form of the heteroskedasticity correctly in advance.

Formulas for homoskedastic adjustment inflation factors (HAIFs)

- Suppose that X is the core design matrix, A is the additional–variables matrix, and D is the diagonal matrix of weights.
- Then the variance HAIF for the kth column of X is a ratio, whose denominator is the kth diagonal element of

$$(X'DX)^{-1}$$

  and whose numerator is the kth diagonal element of

$$[(X,A)'D(X,A)]^{-1}$$

  where X,A is the horizontal concatenation matrix of X and A.
- The kth standard error HAIF is the square root of the kth variance HAIF.
- The weight matrix D is either the default identity matrix, or a diagonal matrix containing inverse variance weights.
- In the second case, the HAIF is a *heteroskedastic* adjustment inflation factor, assuming that we guessed the form of the heteroskedasticity correctly in advance.

Formulas for homoskedastic adjustment inflation factors (HAIFs)

- Suppose that X is the core design matrix, A is the additional–variables matrix, and D is the diagonal matrix of weights.
- Then the variance HAIF for the kth column of X is a ratio, whose denominator is the kth diagonal element of

$$(X'DX)^{-1}$$

and whose numerator is the kth diagonal element of

$$[(X,A)'D(X,A)]^{-1}$$

where X,A is the horizontal concatenation matrix of X and A.
- The kth standard error HAIF is the square root of the kth variance HAIF.
- The weight matrix D is either the default identity matrix, or a diagonal matrix containing inverse variance weights.
- In the second case, the HAIF is a *heteroskedastic* adjustment inflation factor, assuming that we guessed the form of the heteroskedasticity correctly in advance.

Formulas for homoskedastic adjustment inflation factors (HAIFs)

- Suppose that X is the core design matrix, A is the additional–variables matrix, and D is the diagonal matrix of weights.
- Then the variance HAIF for the kth column of X is a ratio, whose denominator is the kth diagonal element of

$$(X'DX)^{-1}$$

and whose numerator is the kth diagonal element of

$$[(X,A)'D(X,A)]^{-1}$$

where X,A is the horizontal concatenation matrix of X and A.
- The kth standard error HAIF is the square root of the kth variance HAIF.
- The weight matrix D is either the default identity matrix, or a diagonal matrix containing inverse variance weights.
- In the second case, the HAIF is a *heteroskedastic* adjustment inflation factor, assuming that we guessed the form of the heteroskedasticity correctly in advance.

Formulas for homoskedastic adjustment inflation factors (HAIFs)

- Suppose that X is the core design matrix, A is the additional–variables matrix, and D is the diagonal matrix of weights.
- Then the variance HAIF for the kth column of X is a ratio, whose denominator is the kth diagonal element of

$$(X'DX)^{-1}$$

and whose numerator is the kth diagonal element of

$$[(X,A)'D(X,A)]^{-1}$$

where X,A is the horizontal concatenation matrix of X and A.
- The kth standard error HAIF is the square root of the kth variance HAIF.
- The weight matrix D is either the default identity matrix, or a diagonal matrix containing inverse variance weights.
- In the second case, the HAIF is a *heteroskedastic* adjustment inflation factor, assuming that we guessed the form of the heteroskedasticity correctly in advance.

Formulas for homoskedastic adjustment inflation factors (HAIFs)

- Suppose that X is the core design matrix, A is the additional–variables matrix, and D is the diagonal matrix of weights.
- Then the variance HAIF for the kth column of X is a ratio, whose denominator is the kth diagonal element of

$$(X'DX)^{-1}$$

  and whose numerator is the kth diagonal element of

$$[(X,A)'D(X,A)]^{-1}$$

  where X,A is the horizontal concatenation matrix of X and A.
- The kth standard error HAIF is the square root of the kth variance HAIF.
- The weight matrix D is either the default identity matrix, or a diagonal matrix containing inverse variance weights.
- In the second case, the HAIF is a *heteroskedastic* adjustment inflation factor, assuming that we guessed the form of the heteroskedasticity correctly in advance.

Note that the HAIF represents a "worst–case" scenario

- The HAIFs are calculated assuming that the additional variables A do not predict the outcome, independently of the core variables X.
- This is why they can be calculated from A and X, without looking at the outcome.
- If the variables A predict *only* the distribution of the outcome, *conditional* on the X–values, then including them may actually *decrease* the sampling variance of the X–effects.
- If the variables in A are in fact confounders properly so called, predicting the exposure *and* the outcome, then the effect of including them will probably be intermediate between these two extremes.
- See Seber (1977)[4] for a rigorous discussion of these issues.

Note that the HAIF represents a "worst–case" scenario

- The HAIFs are calculated assuming that the additional variables A do not predict the outcome, independently of the core variables X.
- This is why they can be calculated from A and X, without looking at the outcome.
- If the variables A predict *only* the distribution of the outcome, *conditional* on the X–values, then including them may actually *decrease* the sampling variance of the X–effects.
- If the variables in A are in fact confounders properly so called, predicting the exposure *and* the outcome, then the effect of including them will probably be intermediate between these two extremes.
- See Seber (1977)[4] for a rigorous discussion of these issues.

Note that the HAIF represents a "worst–case" scenario

- The HAIFs are calculated assuming that the additional variables A do not predict the outcome, independently of the core variables X.
- This is why they can be calculated from A and X, without looking at the outcome.
- If the variables A predict *only* the distribution of the outcome, *conditional* on the X–values, then including them may actually *decrease* the sampling variance of the X–effects.
- If the variables in A are in fact confounders properly so called, predicting the exposure *and* the outcome, then the effect of including them will probably be intermediate between these two extremes.
- See Seber (1977)[4] for a rigorous discussion of these issues.

Note that the HAIF represents a "worst–case" scenario

- The HAIFs are calculated assuming that the additional variables A do not predict the outcome, independently of the core variables X.
- This is why they can be calculated from A and X, without looking at the outcome.
- If the variables A predict *only* the distribution of the outcome, *conditional* on the X–values, then including them may actually *decrease* the sampling variance of the X–effects.
- If the variables in A are in fact confounders properly so called, predicting the exposure *and* the outcome, then the effect of including them will probably be intermediate between these two extremes.
- See Seber (1977)[4] for a rigorous discussion of these issues.

Note that the HAIF represents a "worst–case" scenario

- The HAIFs are calculated assuming that the additional variables A do not predict the outcome, independently of the core variables X.
- This is why they can be calculated from A and X, without looking at the outcome.
- If the variables A predict *only* the distribution of the outcome, *conditional* on the X–values, then including them may actually *decrease* the sampling variance of the X–effects.
- If the variables in A are in fact confounders properly so called, predicting the exposure *and* the outcome, then the effect of including them will probably be intermediate between these two extremes.
- See Seber (1977)[4] for a rigorous discussion of these issues.

Note that the HAIF represents a "worst–case" scenario

- The HAIFs are calculated assuming that the additional variables A do not predict the outcome, independently of the core variables X.
- This is why they can be calculated from A and X, without looking at the outcome.
- If the variables A predict *only* the distribution of the outcome, *conditional* on the X–values, then including them may actually *decrease* the sampling variance of the X–effects.
- If the variables in A are in fact confounders properly so called, predicting the exposure *and* the outcome, then the effect of including them will probably be intermediate between these two extremes.
- See Seber (1977)[4] for a rigorous discussion of these issues.

HAIF ratios: the `haifcomp` module

- The `haif` package has two modules, `haif` and `haifcomp`.
- The `haifcomp` module inputs a core variable list, defining a core design matrix X, and two alternative additional variable lists, defining two alternative submatrices B and C.
- It outputs, for each core variable in X, the ratio between its variance HAIF for adding the submatrix C and its variance HAIF for adding the submatrix B. (And the corresponding standard error HAIFs.)
- `haifcomp` is useful if the columns of the denominator submatrix B are linearly dependent on the columns of the numerator submatrix C, without being a subset of the columns of C.
- In that case, the HAIF ratios are the inflation factors caused by unnecessarily using the design matrix (X, C), when in fact (X, B) is the true design matrix.

HAIF ratios: the `haifcomp` module

- The `haif` package has two modules, `haif` and `haifcomp`.
- The `haifcomp` module inputs a core variable list, defining a core design matrix X, and two alternative additional variable lists, defining two alternative submatrices B and C.
- It outputs, for each core variable in X, the ratio between its variance HAIF for adding the submatrix C and its variance HAIF for adding the submatrix B. (And the corresponding standard error HAIFs.)
- `haifcomp` is useful if the columns of the denominator submatrix B are linearly dependent on the columns of the numerator submatrix C, without being a subset of the columns of C.
- In that case, the HAIF ratios are the inflation factors caused by unnecessarily using the design matrix (X, C), when in fact (X, B) is the true design matrix.

HAIF ratios: the `haifcomp` module

- The `haif` package has two modules, `haif` and `haifcomp`.
- The `haifcomp` module inputs a core variable list, defining a core design matrix X, and two alternative additional variable lists, defining two alternative submatrices B and C.
- It outputs, for each core variable in X, the ratio between its variance HAIF for adding the submatrix C and its variance HAIF for adding the submatrix B. (And the corresponding standard error HAIFs.)
- `haifcomp` is useful if the columns of the denominator submatrix B are linearly dependent on the columns of the numerator submatrix C, without being a subset of the columns of C.
- In that case, the HAIF ratios are the inflation factors caused by unnecessarily using the design matrix (X, C), when in fact (X, B) is the true design matrix.

HAIF ratios: the `haifcomp` module

- The `haif` package has two modules, `haif` and `haifcomp`.
- The `haifcomp` module inputs a core variable list, defining a core design matrix X, and two alternative additional variable lists, defining two alternative submatrices B and C.
- It outputs, for each core variable in X, the ratio between its variance HAIF for adding the submatrix C and its variance HAIF for adding the submatrix B. (And the corresponding standard error HAIFs.)
- `haifcomp` is useful if the columns of the denominator submatrix B are linearly dependent on the columns of the numerator submatrix C, without being a subset of the columns of C.
- In that case, the HAIF ratios are the inflation factors caused by unnecessarily using the design matrix (X, C), when in fact (X, B) is the true design matrix.

HAIF ratios: the **haifcomp** module

- The `haif` package has two modules, `haif` and `haifcomp`.
- The `haifcomp` module inputs a core variable list, defining a core design matrix X, and two alternative additional variable lists, defining two alternative submatrices B and C.
- It outputs, for each core variable in X, the ratio between its variance HAIF for adding the submatrix C and its variance HAIF for adding the submatrix B. (And the corresponding standard error HAIFs.)
- `haifcomp` is useful if the columns of the denominator submatrix B are linearly dependent on the columns of the numerator submatrix C, without being a subset of the columns of C.
- In that case, the HAIF ratios are the inflation factors caused by unnecessarily using the design matrix (X, C), when in fact (X, B) is the true design matrix.

HAIF ratios: the `haifcomp` module

- The `haif` package has two modules, `haif` and `haifcomp`.
- The `haifcomp` module inputs a core variable list, defining a core design matrix X, and two alternative additional variable lists, defining two alternative submatrices B and C.
- It outputs, for each core variable in X, the ratio between its variance HAIF for adding the submatrix C and its variance HAIF for adding the submatrix B. (And the corresponding standard error HAIFs.)
- `haifcomp` is useful if the columns of the denominator submatrix B are linearly dependent on the columns of the numerator submatrix C, without being a subset of the columns of C.
- In that case, the HAIF ratios are the inflation factors caused by unnecessarily using the design matrix (X, C), when in fact (X, B) is the true design matrix.

Formulas for HAIF ratios

- Suppose that X is the core design matrix, B and C are the two alternative additional–variables matrices, and D is the diagonal matrix of weights.
- Then the variance HAIF ratio for the kth column of X is a ratio, whose denominator is the kth diagonal element of

$$[(X,B)'D(X,B)]^{-1}$$

and whose numerator is the kth diagonal element of

$$[(X,C)'D(X,C)]^{-1}$$

- A typical application is choosing between two linear regression models, each with a common slope and an array of intercepts corresponding to a grouping of the data.
- In that case, X may contain the variable defining the slope, B may contain identifier variables for the groups in a coarser grouping, and C may contain identifiers for the groups in a finer grouping.

Formulas for HAIF ratios

- Suppose that X is the core design matrix, B and C are the two alternative additional–variables matrices, and D is the diagonal matrix of weights.
- Then the variance HAIF ratio for the kth column of X is a ratio, whose denominator is the kth diagonal element of

$$[(X,B)'D(X,B)]^{-1}$$

and whose numerator is the kth diagonal element of

$$[(X,C)'D(X,C)]^{-1}$$

- A typical application is choosing between two linear regression models, each with a common slope and an array of intercepts corresponding to a grouping of the data.
- In that case, X may contain the variable defining the slope, B may contain identifier variables for the groups in a coarser grouping, and C may contain identifiers for the groups in a finer grouping.

Formulas for HAIF ratios

- Suppose that X is the core design matrix, B and C are the two alternative additional–variables matrices, and D is the diagonal matrix of weights.
- Then the variance HAIF ratio for the kth column of X is a ratio, whose denominator is the kth diagonal element of

$$[(X, B)'D(X, B)]^{-1}$$

and whose numerator is the kth diagonal element of

$$[(X, C)'D(X, C)]^{-1}$$

- A typical application is choosing between two linear regression models, each with a common slope and an array of intercepts corresponding to a grouping of the data.
- In that case, X may contain the variable defining the slope, B may contain identifier variables for the groups in a coarser grouping, and C may contain identifiers for the groups in a finer grouping.

Formulas for HAIF ratios

- Suppose that X is the core design matrix, B and C are the two alternative additional–variables matrices, and D is the diagonal matrix of weights.
- Then the variance HAIF ratio for the kth column of X is a ratio, whose denominator is the kth diagonal element of

$$[(X,B)'D(X,B)]^{-1}$$

and whose numerator is the kth diagonal element of

$$[(X,C)'D(X,C)]^{-1}$$

- A typical application is choosing between two linear regression models, each with a common slope and an array of intercepts corresponding to a grouping of the data.
- In that case, X may contain the variable defining the slope, B may contain identifier variables for the groups in a coarser grouping, and C may contain identifiers for the groups in a finer grouping.

Formulas for HAIF ratios

- Suppose that X is the core design matrix, B and C are the two alternative additional–variables matrices, and D is the diagonal matrix of weights.
- Then the variance HAIF ratio for the kth column of X is a ratio, whose denominator is the kth diagonal element of

$$[(X, B)'D(X, B)]^{-1}$$

and whose numerator is the kth diagonal element of

$$[(X, C)'D(X, C)]^{-1}$$

- A typical application is choosing between two linear regression models, each with a common slope and an array of intercepts corresponding to a grouping of the data.
- In that case, X may contain the variable defining the slope, B may contain identifier variables for the groups in a coarser grouping, and C may contain identifiers for the groups in a finer grouping.

Example: Adjusting weight effects for length in the `auto` data

- When measuring weight effects (per pound) on fuel consumption (in gallons/mile), we might want to adjust for length (in inches).
- We might decide to fit a multi–intercept model, with one intercept for each of a number of length categories, and a common slope (per pound).
- And we might be wondering whether to group length into 4 quartiles (the submodel), or to group length into 8 octiles (the supermodel).
- *So* we might use `haifcomp` to assess the loss of power caused by fitting the supermodel, assuming that the submodel is true.

Example: Adjusting weight effects for length in the `auto` data

- When measuring weight effects (per pound) on fuel consumption (in gallons/mile), we might want to adjust for length (in inches).
- We might decide to fit a multi–intercept model, with one intercept for each of a number of length categories, and a common slope (per pound).
- And we might be wondering whether to group length into 4 quartiles (the submodel), or to group length into 8 octiles (the supermodel).
- *So* we might use `haifcomp` to assess the loss of power caused by fitting the supermodel, assuming that the submodel is true.

Example: Adjusting weight effects for length in the `auto` data

- When measuring weight effects (per pound) on fuel consumption (in gallons/mile), we might want to adjust for length (in inches).
- We might decide to fit a multi–intercept model, with one intercept for each of a number of length categories, and a common slope (per pound).
- And we might be wondering whether to group length into 4 quartiles (the submodel), or to group length into 8 octiles (the supermodel).
- *So* we might use `haifcomp` to assess the loss of power caused by fitting the supermodel, assuming that the submodel is true.

Example: Adjusting weight effects for length in the `auto` data

- When measuring weight effects (per pound) on fuel consumption (in gallons/mile), we might want to adjust for length (in inches).
- We might decide to fit a multi–intercept model, with one intercept for each of a number of length categories, and a common slope (per pound).
- And we might be wondering whether to group length into 4 quartiles (the submodel), or to group length into 8 octiles (the supermodel).
- *So* we might use `haifcomp` to assess the loss of power caused by fitting the supermodel, assuming that the submodel is true.

Example: Adjusting weight effects for length in the `auto` data

- When measuring weight effects (per pound) on fuel consumption (in gallons/mile), we might want to adjust for length (in inches).
- We might decide to fit a multi–intercept model, with one intercept for each of a number of length categories, and a common slope (per pound).
- And we might be wondering whether to group length into 4 quartiles (the submodel), or to group length into 8 octiles (the supermodel).
- *So* we might use `haifcomp` to assess the loss of power caused by fitting the supermodel, assuming that the submodel is true.

Setting up the grouping variables for length groups

In the `auto` data, the following Stata 11 code sets up two new grouping variables, containing length quartiles and octiles, using `xtile`:

```
. xtile lengp4=length, nquantiles(4);

. xtile lengp8=length, nquantiles(8);
```

(Note that, in Stata 10, we would have needed two `xi` commands at this point, to create indicator variables for these two grouping variables. These `xi` commands are no longer needed in Stata 11, which has factor variables instead.)

Setting up the grouping variables for length groups

In the `auto` data, the following Stata 11 code sets up two new grouping variables, containing length quartiles and octiles, using `xtile`:

```
. xtile lengp4=length, nquantiles(4);

. xtile lengp8=length, nquantiles(8);
```

(Note that, in Stata 10, we would have needed two `xi` commands at this point, to create indicator variables for these two grouping variables. These `xi` commands are no longer needed in Stata 11, which has factor variables instead.)

Setting up the grouping variables for length groups

In the `auto` data, the following Stata 11 code sets up two new grouping variables, containing length quartiles and octiles, using `xtile`:

```
. xtile lengp4=length, nquantiles(4);

. xtile lengp8=length, nquantiles(8);
```

(Note that, in Stata 10, we would have needed two `xi` commands at this point, to create indicator variables for these two grouping variables. These `xi` commands are no longer needed in Stata 11, which has factor variables instead.)

Setting up the grouping variables for length groups

In the `auto` data, the following Stata 11 code sets up two new grouping variables, containing length quartiles and octiles, using `xtile`:

```
. xtile lengp4=length, nquantiles(4);

. xtile lengp8=length, nquantiles(8);
```

(Note that, in Stata 10, we would have needed two `xi` commands at this point, to create indicator variables for these two grouping variables. These `xi` commands are no longer needed in Stata 11, which has factor variables instead.)

Computing HAIF ratios for weight between the length octile and quartile models

We now use `haifcomp`, with the `noconst` option, to calculate the variance and SE HAIF ratios between the length–octile–adjusted model and the length–quartile–adjusted model for the per-pound weight effect:

```
. haifcomp weight, noconst daddvars(ibn.lengp4) naddvars(ibn.lengp8);
Number of observations: 74
Homoskedastic adjustment inflation factor ratios
for variances and standard errors:
                Variance          SE
      weight   1.3477577   1.1609297
```

We see that, *if* the length–quartile–adjusted model is true, *then* using the length–octile–adjusted model will inflate the confidence interval for the weight effect (per pound) by a factor of 1.16. This could be cancelled out by increasing the sample size by a factor of 1.35, assuming the sample composition to stay the same.

Computing HAIF ratios for weight between the length octile and quartile models

We now use `haifcomp`, with the `noconst` option, to calculate the variance and SE HAIF ratios between the length–octile–adjusted model and the length–quartile–adjusted model for the per-pound weight effect:

```
. haifcomp weight, noconst daddvars(ibn.lengp4) naddvars(ibn.lengp8);
Number of observations: 74
Homoskedastic adjustment inflation factor ratios
for variances and standard errors:
                Variance           SE
      weight   1.3477577   1.1609297
```

We see that, *if* the length–quartile–adjusted model is true, *then* using the length–octile–adjusted model will inflate the confidence interval for the weight effect (per pound) by a factor of 1.16. This could be cancelled out by increasing the sample size by a factor of 1.35, assuming the sample composition to stay the same.

Computing HAIF ratios for weight between the length octile and quartile models

We now use `haifcomp`, with the `noconst` option, to calculate the variance and SE HAIF ratios between the length–octile–adjusted model and the length–quartile–adjusted model for the per-pound weight effect:

```
. haifcomp weight, noconst daddvars(ibn.lengp4) naddvars(ibn.lengp8);
Number of observations: 74
Homoskedastic adjustment inflation factor ratios
for variances and standard errors:
               Variance          SE
      weight   1.3477577   1.1609297
```

We see that, *if* the length–quartile–adjusted model is true, *then* using the length–octile–adjusted model will inflate the confidence interval for the weight effect (per pound) by a factor of 1.16. This could be cancelled out by increasing the sample size by a factor of 1.35, assuming the sample composition to stay the same.

Computing HAIF ratios for weight between the length octile and quartile models

We now use `haifcomp`, with the `noconst` option, to calculate the variance and SE HAIF ratios between the length–octile–adjusted model and the length–quartile–adjusted model for the per-pound weight effect:

```
. haifcomp weight, noconst daddvars(ibn.lengp4) naddvars(ibn.lengp8);
Number of observations: 74
Homoskedastic adjustment inflation factor ratios
for variances and standard errors:
                Variance          SE
      weight   1.3477577   1.1609297
```

We see that, *if* the length–quartile–adjusted model is true, *then* using the length–octile–adjusted model will inflate the confidence interval for the weight effect (per pound) by a factor of 1.16. This could be cancelled out by increasing the sample size by a factor of 1.35, assuming the sample composition to stay the same.

Real–world example: Prenatal smoking and lung capacity in the ALSPAC cohort

- In the ALSPAC birth cohort study in Bristol, mothers of 13383 children gave information on smoking habits over pregnancy.
- Outcomes were lung capacity measures of the children at 7 years of age, converted to standardized residuals (in SD units) with respect to gender and height.
- Prenatal tobacco exposure was defined as a 5–level ordinal variable ("Not exposed", "Passive only", "Mother 1–9/day", "Mother 10–19/day", or "Mother 20+/day").
- 32 concomitant variables (suspected as confounders) were also measured. (These mostly were "socio–economic", or referred to previous maternal disease history.)
- These were used to define a propensity score (Lu *et al.*, 2001)[2], using `ologit` to fit an ordinal logistic regression model, with prenatal tobacco exposure as the outcome.

Real–world example: Prenatal smoking and lung capacity in the ALSPAC cohort

- In the ALSPAC birth cohort study in Bristol, mothers of 13383 children gave information on smoking habits over pregnancy.
- Outcomes were lung capacity measures of the children at 7 years of age, converted to standardized residuals (in SD units) with respect to gender and height.
- Prenatal tobacco exposure was defined as a 5–level ordinal variable ("Not exposed", "Passive only", "Mother 1–9/day", "Mother 10–19/day", or "Mother 20+/day").
- 32 concomitant variables (suspected as confounders) were also measured. (These mostly were "socio–economic", or referred to previous maternal disease history.)
- These were used to define a propensity score (Lu *et al.*, 2001)[2], using `ologit` to fit an ordinal logistic regression model, with prenatal tobacco exposure as the outcome.

Real–world example: Prenatal smoking and lung capacity in the ALSPAC cohort

- In the ALSPAC birth cohort study in Bristol, mothers of 13383 children gave information on smoking habits over pregnancy.
- Outcomes were lung capacity measures of the children at 7 years of age, converted to standardized residuals (in SD units) with respect to gender and height.
- Prenatal tobacco exposure was defined as a 5–level ordinal variable ("Not exposed", "Passive only", "Mother 1–9/day", "Mother 10–19/day", or "Mother 20+/day").
- 32 concomitant variables (suspected as confounders) were also measured. (These mostly were "socio–economic", or referred to previous maternal disease history.)
- These were used to define a propensity score (Lu *et al.*, 2001)[2], using `ologit` to fit an ordinal logistic regression model, with prenatal tobacco exposure as the outcome.

Real–world example: Prenatal smoking and lung capacity in the ALSPAC cohort

- In the ALSPAC birth cohort study in Bristol, mothers of 13383 children gave information on smoking habits over pregnancy.
- Outcomes were lung capacity measures of the children at 7 years of age, converted to standardized residuals (in SD units) with respect to gender and height.
- Prenatal tobacco exposure was defined as a 5–level ordinal variable ("Not exposed", "Passive only", "Mother 1–9/day", "Mother 10–19/day", or "Mother 20+/day").
- 32 concomitant variables (suspected as confounders) were also measured. (These mostly were "socio–economic", or referred to previous maternal disease history.)
- These were used to define a propensity score (Lu *et al.*, 2001)[2], using `ologit` to fit an ordinal logistic regression model, with prenatal tobacco exposure as the outcome.

Real–world example: Prenatal smoking and lung capacity in the ALSPAC cohort

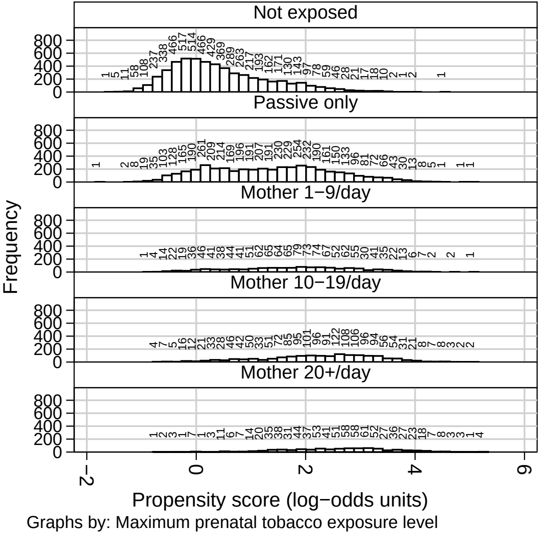
- In the ALSPAC birth cohort study in Bristol, mothers of 13383 children gave information on smoking habits over pregnancy.
- Outcomes were lung capacity measures of the children at 7 years of age, converted to standardized residuals (in SD units) with respect to gender and height.
- Prenatal tobacco exposure was defined as a 5–level ordinal variable ("Not exposed", "Passive only", "Mother 1–9/day", "Mother 10–19/day", or "Mother 20+/day").
- 32 concomitant variables (suspected as confounders) were also measured. (These mostly were "socio–economic", or referred to previous maternal disease history.)
- These were used to define a propensity score (Lu *et al.*, 2001)[2], using `ologit` to fit an ordinal logistic regression model, with prenatal tobacco exposure as the outcome.

Real–world example: Prenatal smoking and lung capacity in the ALSPAC cohort

- In the ALSPAC birth cohort study in Bristol, mothers of 13383 children gave information on smoking habits over pregnancy.
- Outcomes were lung capacity measures of the children at 7 years of age, converted to standardized residuals (in SD units) with respect to gender and height.
- Prenatal tobacco exposure was defined as a 5–level ordinal variable ("Not exposed", "Passive only", "Mother 1–9/day", "Mother 10–19/day", or "Mother 20+/day").
- 32 concomitant variables (suspected as confounders) were also measured. (These mostly were "socio–economic", or referred to previous maternal disease history.)
- These were used to define a propensity score (Lu *et al.*, 2001)[2], using `ologit` to fit an ordinal logistic regression model, with prenatal tobacco exposure as the outcome.

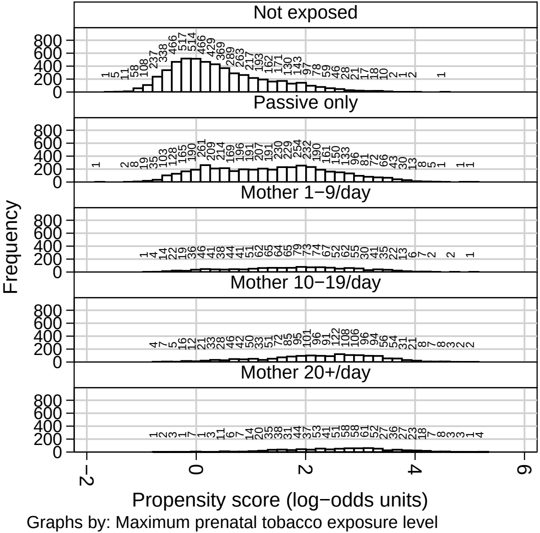
Histograms of the propensity score at each exposure level

- The panels correspond to the 5 levels of prenatal tobacco exposure.
- The histograms give the distribution of the propensity score in children at each exposure level.
- "Tobacco–proneness" seems to predict tobacco exposure, but not *too* well, as there is a lot of overlap between groups.



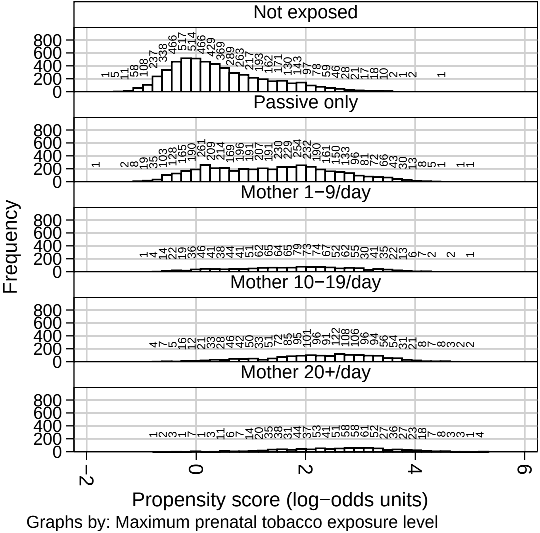
Histograms of the propensity score at each exposure level

- The panels correspond to the 5 levels of prenatal tobacco exposure.
- The histograms give the distribution of the propensity score in children at each exposure level.
- "Tobacco–proneness" seems to predict tobacco exposure, but not *too* well, as there is a lot of overlap between groups.



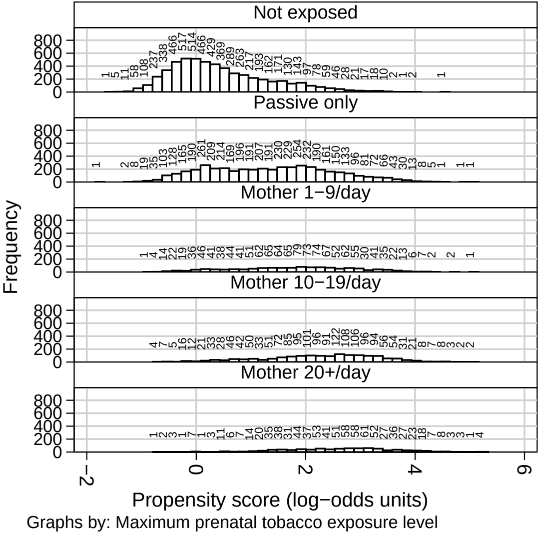
Histograms of the propensity score at each exposure level

- The panels correspond to the 5 levels of prenatal tobacco exposure.
- The histograms give the distribution of the propensity score in children at each exposure level.
- "Tobacco–proneness" seems to predict tobacco exposure, but not *too* well, as there is a lot of overlap between groups.



Histograms of the propensity score at each exposure level

- The panels correspond to the 5 levels of prenatal tobacco exposure.
- The histograms give the distribution of the propensity score in children at each exposure level.
- “Tobacco–proneness” seems to predict tobacco exposure, but not *too* well, as there is a lot of overlap between groups.



Choosing a model to measure tobacco exposure effects on lung capacity

- We planned to measure overall lung capacity trend using linear regression models, containing a single per–category slope with respect to tobacco exposure.
- We planned to fit an unadjusted model, with a single intercept, representing the unexposed mean outcome.
- And we also planned to fit a propensity–adjusted model, with one intercept for each of a number of propensity groups, generated from the propensity score using `xtile`.
- This, of course, poses the question. . .

Choosing a model to measure tobacco exposure effects on lung capacity

- We planned to measure overall lung capacity trend using linear regression models, containing a single per–category slope with respect to tobacco exposure.
- We planned to fit an unadjusted model, with a single intercept, representing the unexposed mean outcome.
- And we also planned to fit a propensity–adjusted model, with one intercept for each of a number of propensity groups, generated from the propensity score using `xtile`.
- This, of course, poses the question. . .

Choosing a model to measure tobacco exposure effects on lung capacity

- We planned to measure overall lung capacity trend using linear regression models, containing a single per–category slope with respect to tobacco exposure.
- We planned to fit an unadjusted model, with a single intercept, representing the unexposed mean outcome.
- And we also planned to fit a propensity–adjusted model, with one intercept for each of a number of propensity groups, generated from the propensity score using `xtile`.
- This, of course, poses the question. . .

Choosing a model to measure tobacco exposure effects on lung capacity

- We planned to measure overall lung capacity trend using linear regression models, containing a single per–category slope with respect to tobacco exposure.
- We planned to fit an unadjusted model, with a single intercept, representing the unexposed mean outcome.
- And we also planned to fit a propensity–adjusted model, with one intercept for each of a number of propensity groups, generated from the propensity score using `xtile`.
- This, of course, poses the question. . .

Choosing a model to measure tobacco exposure effects on lung capacity

- We planned to measure overall lung capacity trend using linear regression models, containing a single per–category slope with respect to tobacco exposure.
- We planned to fit an unadjusted model, with a single intercept, representing the unexposed mean outcome.
- And we also planned to fit a propensity–adjusted model, with one intercept for each of a number of propensity groups, generated from the propensity score using `xtile`.
- This, of course, poses the question. . .

How many propensity groups to use?

- We considered groupings with 1, 2, 4, 8, 16, 32, 64 and 128 nearly–equal groups, generated using `xtile`.
- Note that each successive grouping is defined by splitting each group of the previous grouping into two nearly–equal subgroups.
- *Therefore*, the multi–intercept models for these groupings are a sequence of nested models, in which the earlier models in the sequence are submodels of the later models in the sequence (the supermodels).
- We used `haifcomp` to compute the HAIF ratios for the linear per–category tobacco exposure effect, with the lists of multiple group factor variables as the numerator lists, and the single constant term for the single–intercept model as the denominator list.

How many propensity groups to use?

- We considered groupings with 1, 2, 4, 8, 16, 32, 64 and 128 nearly–equal groups, generated using `xtile`.
- Note that each successive grouping is defined by splitting each group of the previous grouping into two nearly–equal subgroups.
- *Therefore*, the multi–intercept models for these groupings are a sequence of nested models, in which the earlier models in the sequence are submodels of the later models in the sequence (the supermodels).
- We used `haifcomp` to compute the HAIF ratios for the linear per–category tobacco exposure effect, with the lists of multiple group factor variables as the numerator lists, and the single constant term for the single–intercept model as the denominator list.

How many propensity groups to use?

- We considered groupings with 1, 2, 4, 8, 16, 32, 64 and 128 nearly–equal groups, generated using `xtile`.
- Note that each successive grouping is defined by splitting each group of the previous grouping into two nearly–equal subgroups.
- *Therefore*, the multi–intercept models for these groupings are a sequence of nested models, in which the earlier models in the sequence are submodels of the later models in the sequence (the supermodels).
- We used `haifcomp` to compute the HAIF ratios for the linear per–category tobacco exposure effect, with the lists of multiple group factor variables as the numerator lists, and the single constant term for the single–intercept model as the denominator list.

How many propensity groups to use?

- We considered groupings with 1, 2, 4, 8, 16, 32, 64 and 128 nearly–equal groups, generated using `xtile`.
- Note that each successive grouping is defined by splitting each group of the previous grouping into two nearly–equal subgroups.
- *Therefore*, the multi–intercept models for these groupings are a sequence of nested models, in which the earlier models in the sequence are submodels of the later models in the sequence (the supermodels).
- We used `haifcomp` to compute the HAIF ratios for the linear per–category tobacco exposure effect, with the lists of multiple group factor variables as the numerator lists, and the single constant term for the single–intercept model as the denominator list.

How many propensity groups to use?

- We considered groupings with 1, 2, 4, 8, 16, 32, 64 and 128 nearly–equal groups, generated using `xtile`.
- Note that each successive grouping is defined by splitting each group of the previous grouping into two nearly–equal subgroups.
- *Therefore*, the multi–intercept models for these groupings are a sequence of nested models, in which the earlier models in the sequence are submodels of the later models in the sequence (the supermodels).
- We used `haifcomp` to compute the HAIF ratios for the linear per–category tobacco exposure effect, with the lists of multiple group factor variables as the numerator lists, and the single constant term for the single–intercept model as the denominator list.

Costs and benefits of nested propensity groupings: Methods

- The **cost** of each number of groups was measured using its variance and SE HAIF ratios for the linear per–category tobacco exposure effect, compared to the single–intercept model.
- Note that, for any 2 unequal numbers of groups, the ratio between their HAIF ratios measures the cost of using the model with more groups, assuming that the model with fewer groups is true.
- The **benefit** of each number of groups was measured using the within–strata Somers' D of propensity *score* with respect to exposure, restricted to comparisons within propensity *groups* (Newson, 2006)[3].
- Somers' D is 1 for a perfect positive ordinal predictor, -1 for a perfect negative ordinal predictor, and 0 for an ordinal non–predictor.
- In this case, it is interpreted as the highest level of within–group exposure–outcome association which might be caused by residual confounding by the propensity score within groups.

Costs and benefits of nested propensity groupings: Methods

- The **cost** of each number of groups was measured using its variance and SE HAIF ratios for the linear per–category tobacco exposure effect, compared to the single–intercept model.
- Note that, for any 2 unequal numbers of groups, the ratio between their HAIF ratios measures the cost of using the model with more groups, assuming that the model with fewer groups is true.
- The **benefit** of each number of groups was measured using the within–strata Somers' D of propensity *score* with respect to exposure, restricted to comparisons within propensity *groups* (Newson, 2006)[3].
- Somers' D is 1 for a perfect positive ordinal predictor, -1 for a perfect negative ordinal predictor, and 0 for an ordinal non–predictor.
- In this case, it is interpreted as the highest level of within–group exposure–outcome association which might be caused by residual confounding by the propensity score within groups.

Costs and benefits of nested propensity groupings: Methods

- The **cost** of each number of groups was measured using its variance and SE HAIF ratios for the linear per–category tobacco exposure effect, compared to the single–intercept model.
- Note that, for any 2 unequal numbers of groups, the ratio between their HAIF ratios measures the cost of using the model with more groups, assuming that the model with fewer groups is true.
- The **benefit** of each number of groups was measured using the within–strata Somers' D of propensity *score* with respect to exposure, restricted to comparisons within propensity *groups* (Newson, 2006)[3].
- Somers' D is 1 for a perfect positive ordinal predictor, -1 for a perfect negative ordinal predictor, and 0 for an ordinal non–predictor.
- In this case, it is interpreted as the highest level of within–group exposure–outcome association which might be caused by residual confounding by the propensity score within groups.

Costs and benefits of nested propensity groupings: Methods

- The **cost** of each number of groups was measured using its variance and SE HAIF ratios for the linear per–category tobacco exposure effect, compared to the single–intercept model.
- Note that, for any 2 unequal numbers of groups, the ratio between their HAIF ratios measures the cost of using the model with more groups, assuming that the model with fewer groups is true.
- The **benefit** of each number of groups was measured using the within–strata Somers' *D* of propensity *score* with respect to exposure, restricted to comparisons within propensity *groups* (Newson, 2006)[3].
- Somers' *D* is 1 for a perfect positive ordinal predictor, -1 for a perfect negative ordinal predictor, and 0 for an ordinal non–predictor.
- In this case, it is interpreted as the highest level of within–group exposure–outcome association which might be caused by residual confounding by the propensity score within groups.

Costs and benefits of nested propensity groupings: Methods

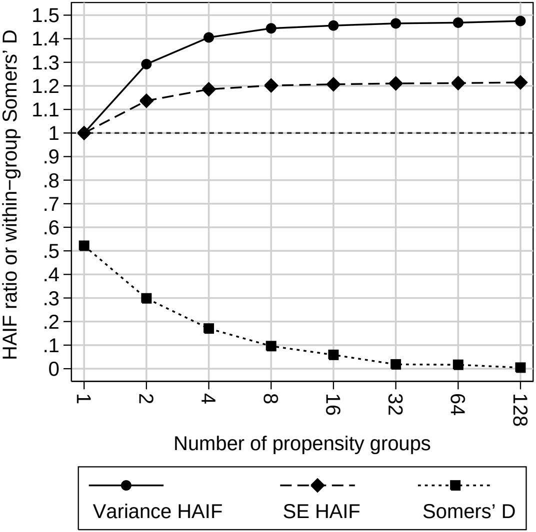
- The **cost** of each number of groups was measured using its variance and SE HAIF ratios for the linear per–category tobacco exposure effect, compared to the single–intercept model.
- Note that, for any 2 unequal numbers of groups, the ratio between their HAIF ratios measures the cost of using the model with more groups, assuming that the model with fewer groups is true.
- The **benefit** of each number of groups was measured using the within–strata Somers' *D* of propensity *score* with respect to exposure, restricted to comparisons within propensity *groups* (Newson, 2006)[3].
- Somers' *D* is 1 for a perfect positive ordinal predictor, -1 for a perfect negative ordinal predictor, and 0 for an ordinal non–predictor.
- In this case, it is interpreted as the highest level of within–group exposure–outcome association which might be caused by residual confounding by the propensity score within groups.

Costs and benefits of nested propensity groupings: Methods

- The **cost** of each number of groups was measured using its variance and SE HAIF ratios for the linear per–category tobacco exposure effect, compared to the single–intercept model.
- Note that, for any 2 unequal numbers of groups, the ratio between their HAIF ratios measures the cost of using the model with more groups, assuming that the model with fewer groups is true.
- The **benefit** of each number of groups was measured using the within–strata Somers' D of propensity *score* with respect to exposure, restricted to comparisons within propensity *groups* (Newson, 2006)[3].
- Somers' D is 1 for a perfect positive ordinal predictor, -1 for a perfect negative ordinal predictor, and 0 for an ordinal non–predictor.
- In this case, it is interpreted as the highest level of within–group exposure–outcome association which might be caused by residual confounding by the propensity score within groups.

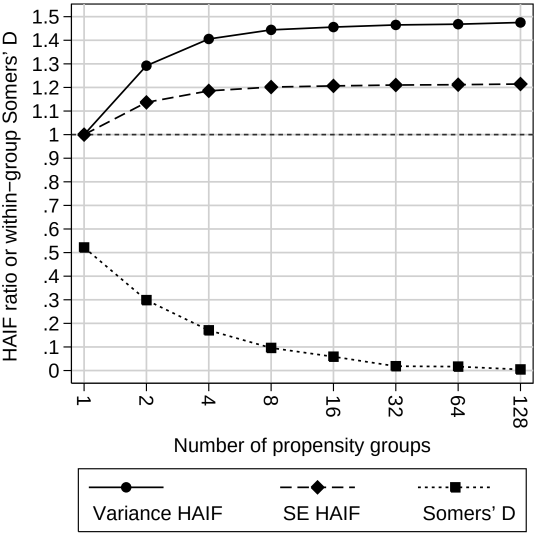
Costs and benefits of nested propensity groupings: Results

- The variance and SE HAIF ratios measure costs, in inflated sample sizes and confidence intervals.
- The within–group Somers' *D* measures benefits, in reducing residual confounding.
- We decided that 32 propensity groups would be enough for our analyses.



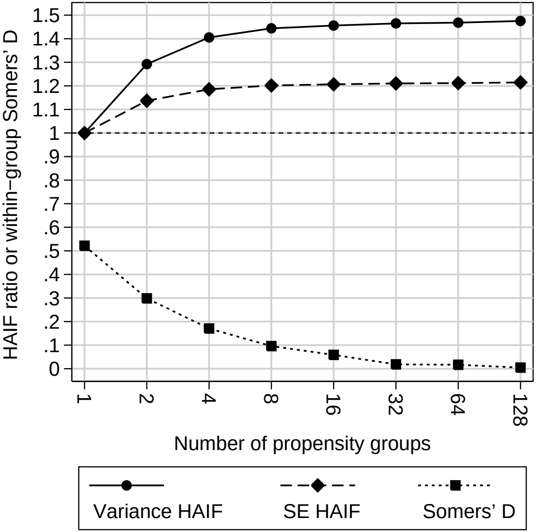
Costs and benefits of nested propensity groupings: Results

- The variance and SE HAIF ratios measure costs, in inflated sample sizes and confidence intervals.
- The within–group Somers' *D* measures benefits, in reducing residual confounding.
- We decided that 32 propensity groups would be enough for our analyses.



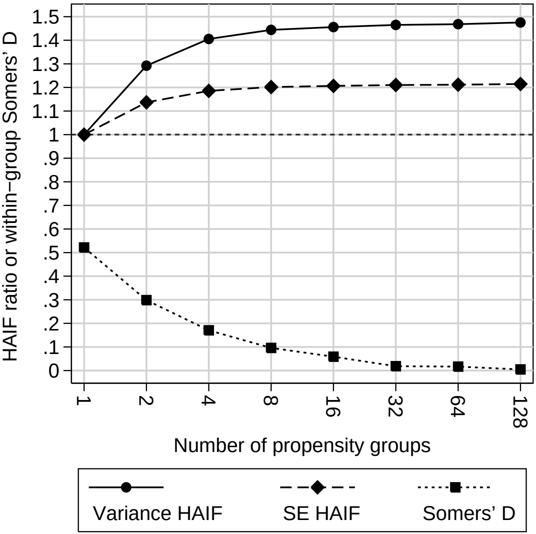
Costs and benefits of nested propensity groupings: Results

- The variance and SE HAIF ratios measure costs, in inflated sample sizes and confidence intervals.
- The within–group Somers' *D* measures benefits, in reducing residual confounding.
- We decided that 32 propensity groups would be enough for our analyses.



Costs and benefits of nested propensity groupings: Results

- The variance and SE HAIF ratios measure costs, in inflated sample sizes and confidence intervals.
- The within–group Somers' *D* measures benefits, in reducing residual confounding.
- We decided that 32 propensity groups would be enough for our analyses.



References

[1] Davey Smith G. and Ebrahim S. 2002. Data dredging, bias or confounding. They can all get you into the BMJ and the Friday papers. *British Medical Journal* **325**: 1437–1438.

[2] Lu B., Zanutto E., Hornik R. and Rosenbaum P. R. 2002. Matching with doses in an observational study in a media campaign against drug use. *Journal of the American Statistical Association* **96(456)**: 1245–1253.

[3] Newson R. 2006. Confidence intervals for rank statistics: Somers' *D* and extensions. *The Stata Journal* **6(3)**: 309–334.

[4] Seber G. A. F. 1977. *Linear Regression Analysis.* New York: John Wiley & Sons.

This presentation can be downloaded from the conference website at *http://ideas.repec.org/s/boc/usug09.html*

The `haif`, `xsvmat`, `parmest` and `dsconcat` packages, used in producing this presentation, can be downloaded from SSC, using the `ssc` command.