Imperial College London

The role of Somers' *D* in propensity modelling

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

Department of Primary Care and Public Health, Imperial College London

22nd UK Stata Users' Group Meeting, 8–9 September, 2016
Downloadable from the conference website at
http://ideas.repec.org/s/boc/usug16.html

What is Somers' D?

- We assume that pairs of (X, Y)–pairs (X_i, Y_i) and (X_j, Y_j) are sampled from a specified population of (X, Y)–pairs, under a specified sampling scheme.
- **Kendall's** τ_a is defined as the expectation

$$\tau_{XY} = \mathrm{E}[\mathrm{sign}(X_i - X_j)\mathrm{sign}(Y_i - Y_j)],$$

  or as the difference between the probabilities of **concordance** and **discordance** between the two (X, Y)–pairs.
- **Somers'** D is defined as the ratio

$$D(Y|X) = \tau_{XY}/\tau_{XX},$$

  or as the difference between the two corresponding *conditional* probabilities, *given* that one X–value is known to be larger than the other X–value.
- Somers' D has the useful property that a higher–magnitude $D(Y|X)$ cannot be secondary to a lower–magnitude $D(W|X)$.

What is Somers' D?

- We assume that pairs of (X,Y)–pairs (X_i,Y_i) and (X_j,Y_j) are sampled from a specified population of (X,Y)–pairs, under a specified sampling scheme.
- **Kendall's** τ_a is defined as the expectation

$$\tau_{XY} = \mathrm{E}[\mathrm{sign}(X_i - X_j)\mathrm{sign}(Y_i - Y_j)],$$

  or as the difference between the probabilities of **concordance** and **discordance** between the two (X,Y)–pairs.
- **Somers'** D is defined as the ratio

$$D(Y|X) = \tau_{XY}/\tau_{XX},$$

  or as the difference between the two corresponding *conditional* probabilities, *given* that one X–value is known to be larger than the other X–value.
- Somers' D has the useful property that a higher–magnitude $D(Y|X)$ cannot be secondary to a lower–magnitude $D(W|X)$.

What is Somers' D?

- We assume that pairs of (X, Y)–pairs (X_i, Y_i) and (X_j, Y_j) are sampled from a specified population of (X, Y)–pairs, under a specified sampling scheme.
- **Kendall's** τ_a is defined as the expectation

$$\tau_{XY} = \mathrm{E}[\mathrm{sign}(X_i - X_j)\mathrm{sign}(Y_i - Y_j)],$$

  or as the difference between the probabilities of **concordance** and **discordance** between the two (X, Y)–pairs.
- **Somers'** D is defined as the ratio

$$D(Y|X) = \tau_{XY}/\tau_{XX},$$

  or as the difference between the two corresponding *conditional* probabilities, *given* that one X–value is known to be larger than the other X–value.
- Somers' D has the useful property that a higher–magnitude $D(Y|X)$ cannot be secondary to a lower–magnitude $D(W|X)$.

What is Somers' D?

- We assume that pairs of (X, Y)–pairs (X_i, Y_i) and (X_j, Y_j) are sampled from a specified population of (X, Y)–pairs, under a specified sampling scheme.
- **Kendall's** τ_a is defined as the expectation

$$\tau_{XY} = \mathrm{E}[\mathrm{sign}(X_i - X_j)\mathrm{sign}(Y_i - Y_j)],$$

  or as the difference between the probabilities of **concordance** and **discordance** between the two (X, Y)–pairs.
- **Somers'** D is defined as the ratio

$$D(Y|X) = \tau_{XY}/\tau_{XX},$$

  or as the difference between the two corresponding *conditional* probabilities, *given* that one X–value is known to be larger than the other X–value.
- Somers' D has the useful property that a higher–magnitude $D(Y|X)$ cannot be secondary to a lower–magnitude $D(W|X)$.

What is Somers' D?

- We assume that pairs of (X, Y)–pairs (X_i, Y_i) and (X_j, Y_j) are sampled from a specified population of (X, Y)–pairs, under a specified sampling scheme.
- **Kendall's** τ_a is defined as the expectation

$$\tau_{XY} = \mathrm{E}[\mathrm{sign}(X_i - X_j)\mathrm{sign}(Y_i - Y_j)],$$

  or as the difference between the probabilities of **concordance** and **discordance** between the two (X, Y)–pairs.
- **Somers'** D is defined as the ratio

$$D(Y|X) = \tau_{XY}/\tau_{XX},$$

  or as the difference between the two corresponding *conditional* probabilities, *given* that one X–value is known to be larger than the other X–value.
- Somers' D has the useful property that a higher–magnitude $D(Y|X)$ cannot be secondary to a lower–magnitude $D(W|X)$.

What is the 21st–century Rubin method of confounder adjustment?

- The Rubin method of confounder adjustment, in its 21st–century version[6], is a 2–phase method for estimating the **causal effect** of a proposed **intervention**, using observational data.
- In Phase 1 ("design"), we find a model in the sample data, predicting the **exposure** (which we propose to intervene to change) from **confounders** (expected to be unaffected).
- This model is used to define a **propensity score**, predicting "exposure–proneness" as a function of the confounders.
- In Phase 2 ("analysis"), we add in the outcome data, and use the propensity score in a regression model to estimate a propensity–adjusted exposure effect on the outcome.
- This adjusted effect is interpreted as a difference between mean outcomes in two **scenario populations**, with the same propensity distribution, but different exposure levels.
- This is usually done using propensity matching, propensity weighting, or propensity stratification[2].

What is the 21st–century Rubin method of confounder adjustment?

- The Rubin method of confounder adjustment, in its 21st–century version[6], is a 2–phase method for estimating the **causal effect** of a proposed **intervention**, using observational data.
- In Phase 1 ("design"), we find a model in the sample data, predicting the **exposure** (which we propose to intervene to change) from **confounders** (expected to be unaffected).
- This model is used to define a **propensity score**, predicting "exposure–proneness" as a function of the confounders.
- In Phase 2 ("analysis"), we add in the outcome data, and use the propensity score in a regression model to estimate a propensity–adjusted exposure effect on the outcome.
- This adjusted effect is interpreted as a difference between mean outcomes in two **scenario populations**, with the same propensity distribution, but different exposure levels.
- This is usually done using propensity matching, propensity weighting, or propensity stratification[2].

What is the 21st–century Rubin method of confounder adjustment?

- The Rubin method of confounder adjustment, in its 21st–century version[6], is a 2–phase method for estimating the **causal effect** of a proposed **intervention**, using observational data.
- In Phase 1 ("design"), we find a model in the sample data, predicting the **exposure** (which we propose to intervene to change) from **confounders** (expected to be unaffected).
- This model is used to define a **propensity score**, predicting "exposure–proneness" as a function of the confounders.
- In Phase 2 ("analysis"), we add in the outcome data, and use the propensity score in a regression model to estimate a propensity–adjusted exposure effect on the outcome.
- This adjusted effect is interpreted as a difference between mean outcomes in two **scenario populations**, with the same propensity distribution, but different exposure levels.
- This is usually done using propensity matching, propensity weighting, or propensity stratification[2].

What is the 21st–century Rubin method of confounder adjustment?

- The Rubin method of confounder adjustment, in its 21st–century version[6], is a 2–phase method for estimating the **causal effect** of a proposed **intervention**, using observational data.
- In Phase 1 ("design"), we find a model in the sample data, predicting the **exposure** (which we propose to intervene to change) from **confounders** (expected to be unaffected).
- This model is used to define a **propensity score**, predicting "exposure–proneness" as a function of the confounders.
- In Phase 2 ("analysis"), we add in the outcome data, and use the propensity score in a regression model to estimate a propensity–adjusted exposure effect on the outcome.
- This adjusted effect is interpreted as a difference between mean outcomes in two **scenario populations**, with the same propensity distribution, but different exposure levels.
- This is usually done using propensity matching, propensity weighting, or propensity stratification[2].

What is the 21st–century Rubin method of confounder adjustment?

- The Rubin method of confounder adjustment, in its 21st–century version[6], is a 2–phase method for estimating the **causal effect** of a proposed **intervention**, using observational data.
- In Phase 1 ("design"), we find a model in the sample data, predicting the **exposure** (which we propose to intervene to change) from **confounders** (expected to be unaffected).
- This model is used to define a **propensity score**, predicting "exposure–proneness" as a function of the confounders.
- In Phase 2 ("analysis"), we add in the outcome data, and use the propensity score in a regression model to estimate a propensity–adjusted exposure effect on the outcome.
- This adjusted effect is interpreted as a difference between mean outcomes in two **scenario populations**, with the same propensity distribution, but different exposure levels.
- This is usually done using propensity matching, propensity weighting, or propensity stratification[2].

What is the 21st–century Rubin method of confounder adjustment?

- The Rubin method of confounder adjustment, in its 21st–century version[6], is a 2–phase method for estimating the **causal effect** of a proposed **intervention**, using observational data.
- In Phase 1 ("design"), we find a model in the sample data, predicting the **exposure** (which we propose to intervene to change) from **confounders** (expected to be unaffected).
- This model is used to define a **propensity score**, predicting "exposure–proneness" as a function of the confounders.
- In Phase 2 ("analysis"), we add in the outcome data, and use the propensity score in a regression model to estimate a propensity–adjusted exposure effect on the outcome.
- This adjusted effect is interpreted as a difference between mean outcomes in two **scenario populations**, with the same propensity distribution, but different exposure levels.
- This is usually done using propensity matching, propensity weighting, or propensity stratification[2].

What is the 21st–century Rubin method of confounder adjustment?

- The Rubin method of confounder adjustment, in its 21st–century version[6], is a 2–phase method for estimating the **causal effect** of a proposed **intervention**, using observational data.
- In Phase 1 (“design”), we find a model in the sample data, predicting the **exposure** (which we propose to intervene to change) from **confounders** (expected to be unaffected).
- This model is used to define a **propensity score**, predicting “exposure–proneness” as a function of the confounders.
- In Phase 2 (“analysis”), we add in the outcome data, and use the propensity score in a regression model to estimate a propensity–adjusted exposure effect on the outcome.
- This adjusted effect is interpreted as a difference between mean outcomes in two **scenario populations**, with the same propensity distribution, but different exposure levels.
- This is usually done using propensity matching, propensity weighting, or propensity stratification[2].

So what is the role of Somers' D in propensity modelling?

- The package `somersd`[5] can be downloaded from SSC, and estimates many versions of Somers' D.
- These may be weighted or matched (using `pweights`), or within–strata (using the `wstrata()` option).
- In propensity modelling, we want to limit the level of spurious treatment effect that may remain, after propensity matching and/or weighting and/or stratification.
- A good measure of this limit is Somers' $D(W|X)$, where X is an exposure, W is a confounder or a propensity score, and Somers' D is matched and/or weighted and/or stratified.
- If Y is an outcome, then a higher–magnitude $D(Y|X)$ cannot be secondary to a lower–magnitude $D(W|X)$, defined using the same matching and/or weighting and/or stratification.

So what is the role of Somers' D in propensity modelling?

- The package `somersd`[5] can be downloaded from SSC, and estimates many versions of Somers' D.
- These may be weighted or matched (using `pweights`), or within–strata (using the `wstrata()` option).
- In propensity modelling, we want to limit the level of spurious treatment effect that may remain, after propensity matching and/or weighting and/or stratification.
- A good measure of this limit is Somers' $D(W|X)$, where X is an exposure, W is a confounder or a propensity score, and Somers' D is matched and/or weighted and/or stratified.
- If Y is an outcome, then a higher–magnitude $D(Y|X)$ cannot be secondary to a lower–magnitude $D(W|X)$, defined using the same matching and/or weighting and/or stratification.

So what is the role of Somers' D in propensity modelling?

- The package `somersd`[5] can be downloaded from SSC, and estimates many versions of Somers' D.
- These may be weighted or matched (using `pweights`), or within–strata (using the `wstrata()` option).
- In propensity modelling, we want to limit the level of spurious treatment effect that may remain, after propensity matching and/or weighting and/or stratification.
- A good measure of this limit is Somers' $D(W|X)$, where X is an exposure, W is a confounder or a propensity score, and Somers' D is matched and/or weighted and/or stratified.
- If Y is an outcome, then a higher–magnitude $D(Y|X)$ cannot be secondary to a lower–magnitude $D(W|X)$, defined using the same matching and/or weighting and/or stratification.

So what is the role of Somers' D in propensity modelling?

- The package `somersd`[5] can be downloaded from SSC, and estimates many versions of Somers' D.
- These may be weighted or matched (using `pweights`), or within–strata (using the `wstrata()` option).
- In propensity modelling, we want to limit the level of spurious treatment effect that may remain, after propensity matching and/or weighting and/or stratification.
- A good measure of this limit is Somers' $D(W|X)$, where X is an exposure, W is a confounder or a propensity score, and Somers' D is matched and/or weighted and/or stratified.
- If Y is an outcome, then a higher–magnitude $D(Y|X)$ cannot be secondary to a lower–magnitude $D(W|X)$, defined using the same matching and/or weighting and/or stratification.

So what is the role of Somers' D in propensity modelling?

- The package `somersd`[5] can be downloaded from SSC, and estimates many versions of Somers' D.
- These may be weighted or matched (using `pweights`), or within–strata (using the `wstrata()` option).
- In propensity modelling, we want to limit the level of spurious treatment effect that may remain, after propensity matching and/or weighting and/or stratification.
- A good measure of this limit is Somers' $D(W|X)$, where X is an exposure, W is a confounder or a propensity score, and Somers' D is matched and/or weighted and/or stratified.
- If Y is an outcome, then a higher–magnitude $D(Y|X)$ cannot be secondary to a lower–magnitude $D(W|X)$, defined using the same matching and/or weighting and/or stratification.

So what is the role of Somers' *D* in propensity modelling?

- The package `somersd`[5] can be downloaded from SSC, and estimates many versions of Somers' *D*.
- These may be weighted or matched (using `pweights`), or within–strata (using the `wstrata()` option).
- In propensity modelling, we want to limit the level of spurious treatment effect that may remain, after propensity matching and/or weighting and/or stratification.
- A good measure of this limit is Somers' $D(W|X)$, where X is an exposure, W is a confounder or a propensity score, and Somers' *D* is matched and/or weighted and/or stratified.
- If Y is an outcome, then a higher–magnitude $D(Y|X)$ cannot be secondary to a lower–magnitude $D(W|X)$, defined using the same matching and/or weighting and/or stratification.

But what is the meaning of Somers' $D(Y|X)$?

- Under a wide variety of regression models, $D(Y|X)$ can be transformed to give a treatment effect of X on Y[3].
- *For instance*, if X and Y are both binary, then $D(Y|X)$ is *exactly* the difference $\Pr(Y=1|X=1) - \Pr(Y=1|X=0)$.
- *Similarly*, if X is binary, and Y is Normally distributed with standard deviation σ in both sub–populations defined by X, and $-0.5 < D(Y|X) < 0.5$, then $2D(Y|X)$ is *approximately* the standardized mean difference $(\mu_1 - \mu_0)/\sigma$.
- *So*, either way, a small Somers' $D(W|X)$ (matched and/or weighted and/or stratified) can be used to give an upper bound to the spurious treatment effect attributable to the confounder (or propensity score) W.
- *And*, a large $D(W|X)$ (matched and/or weighted and/or stratified) indicates a problem of **non–overlap**, which our matching and/or weighting and/or stratification has not balanced.

But what is the meaning of Somers' $D(Y|X)$?

- Under a wide variety of regression models, $D(Y|X)$ can be transformed to give a treatment effect of X on Y[3].
- *For instance*, if X and Y are both binary, then $D(Y|X)$ is *exactly* the difference $\Pr(Y = 1|X = 1) - \Pr(Y = 1|X = 0)$.
- *Similarly*, if X is binary, and Y is Normally distributed with standard deviation σ in both sub–populations defined by X, and $-0.5 < D(Y|X) < 0.5$, then $2D(Y|X)$ is *approximately* the standardized mean difference $(\mu_1 - \mu_0)/\sigma$.
- *So*, either way, a small Somers' $D(W|X)$ (matched and/or weighted and/or stratified) can be used to give an upper bound to the spurious treatment effect attributable to the confounder (or propensity score) W.
- *And*, a large $D(W|X)$ (matched and/or weighted and/or stratified) indicates a problem of **non–overlap**, which our matching and/or weighting and/or stratification has not balanced.

But what is the meaning of Somers' $D(Y|X)$?

- Under a wide variety of regression models, $D(Y|X)$ can be transformed to give a treatment effect of X on Y[3].
- *For instance*, if X and Y are both binary, then $D(Y|X)$ is *exactly* the difference $\Pr(Y = 1|X = 1) - \Pr(Y = 1|X = 0)$.
- *Similarly*, if X is binary, and Y is Normally distributed with standard deviation σ in both sub–populations defined by X, and $-0.5 < D(Y|X) < 0.5$, then $2D(Y|X)$ is *approximately* the standardized mean difference $(\mu_1 - \mu_0)/\sigma$.
- *So*, either way, a small Somers' $D(W|X)$ (matched and/or weighted and/or stratified) can be used to give an upper bound to the spurious treatment effect attributable to the confounder (or propensity score) W.
- *And*, a large $D(W|X)$ (matched and/or weighted and/or stratified) indicates a problem of **non–overlap**, which our matching and/or weighting and/or stratification has not balanced.

But what is the meaning of Somers' $D(Y|X)$?

- Under a wide variety of regression models, $D(Y|X)$ can be transformed to give a treatment effect of X on Y[3].
- *For instance*, if X and Y are both binary, then $D(Y|X)$ is *exactly* the difference $\Pr(Y = 1|X = 1) - \Pr(Y = 1|X = 0)$.
- *Similarly*, if X is binary, and Y is Normally distributed with standard deviation σ in both sub–populations defined by X, and $-0.5 < D(Y|X) < 0.5$, then $2D(Y|X)$ is *approximately* the standardized mean difference $(\mu_1 - \mu_0)/\sigma$.
- *So*, either way, a small Somers' $D(W|X)$ (matched and/or weighted and/or stratified) can be used to give an upper bound to the spurious treatment effect attributable to the confounder (or propensity score) W.
- *And*, a large $D(W|X)$ (matched and/or weighted and/or stratified) indicates a problem of **non–overlap**, which our matching and/or weighting and/or stratification has not balanced.

But what is the meaning of Somers' $D(Y|X)$?

- Under a wide variety of regression models, $D(Y|X)$ can be transformed to give a treatment effect of X on Y[3].
- *For instance*, if X and Y are both binary, then $D(Y|X)$ is *exactly* the difference $\Pr(Y = 1|X = 1) - \Pr(Y = 1|X = 0)$.
- *Similarly*, if X is binary, and Y is Normally distributed with standard deviation σ in both sub–populations defined by X, and $-0.5 < D(Y|X) < 0.5$, then $2D(Y|X)$ is *approximately* the standardized mean difference $(\mu_1 - \mu_0)/\sigma$.
- *So*, either way, a small Somers' $D(W|X)$ (matched and/or weighted and/or stratified) can be used to give an upper bound to the spurious treatment effect attributable to the confounder (or propensity score) W.
- *And*, a large $D(W|X)$ (matched and/or weighted and/or stratified) indicates a problem of **non–overlap**, which our matching and/or weighting and/or stratification has not balanced.

But what is the meaning of Somers' $D(Y|X)$?

- Under a wide variety of regression models, $D(Y|X)$ can be transformed to give a treatment effect of X on Y[3].
- *For instance*, if X and Y are both binary, then $D(Y|X)$ is *exactly* the difference $\Pr(Y = 1|X = 1) - \Pr(Y = 1|X = 0)$.
- *Similarly*, if X is binary, and Y is Normally distributed with standard deviation σ in both sub–populations defined by X, and $-0.5 < D(Y|X) < 0.5$, then $2D(Y|X)$ is *approximately* the standardized mean difference $(\mu_1 - \mu_0)/\sigma$.
- *So*, either way, a small Somers' $D(W|X)$ (matched and/or weighted and/or stratified) can be used to give an upper bound to the spurious treatment effect attributable to the confounder (or propensity score) W.
- *And*, a large $D(W|X)$ (matched and/or weighted and/or stratified) indicates a problem of **non–overlap**, which our matching and/or weighting and/or stratification has not balanced.

Example: the `ldw_exper` dataset of Abadie *et al.*, 2004

- We will demonstrate our methods using a dataset distributed by *The Stata Journal* as online supplemental material for an article on propensity matching[1].
- The dataset has 1 observation per subject in a 1970s observational study, in which 185 subjects participated in a job training program and 260 did not.
- We aim to measure the effect of the training program on 1978 earnings (in 1000s of 1978 dollars), adjusted for a list of 10 confounding covariates, using a logit propensity score computed by the SSC package `psmatch2`.
- We demonstrate propensity adjustment, using matching, weighting, and stratification.
- In Phase 1 of the Rubin method, we check for balance and variance inflation, using the SSC packages `somersd` and `haif`[4], respectively.
- And, in Phase 2, we measure the **average treatment effect on the treated (ATET)**, using the SSC package `scenttest`.

Example: the `ldw_exper` dataset of Abadie *et al.*, 2004

- We will demonstrate our methods using a dataset distributed by *The Stata Journal* as online supplemental material for an article on propensity matching[1].
- The dataset has 1 observation per subject in a 1970s observational study, in which 185 subjects participated in a job training program and 260 did not.
- We aim to measure the effect of the training program on 1978 earnings (in 1000s of 1978 dollars), adjusted for a list of 10 confounding covariates, using a logit propensity score computed by the SSC package `psmatch2`.
- We demonstrate propensity adjustment, using matching, weighting, and stratification.
- In Phase 1 of the Rubin method, we check for balance and variance inflation, using the SSC packages `somersd` and `haif`[4], respectively.
- And, in Phase 2, we measure the **average treatment effect on the treated (ATET)**, using the SSC package `scenttest`.

Example: the `ldw_exper` dataset of Abadie *et al.*, 2004

- We will demonstrate our methods using a dataset distributed by *The Stata Journal* as online supplemental material for an article on propensity matching[1].
- The dataset has 1 observation per subject in a 1970s observational study, in which 185 subjects participated in a job training program and 260 did not.
- We aim to measure the effect of the training program on 1978 earnings (in 1000s of 1978 dollars), adjusted for a list of 10 confounding covariates, using a logit propensity score computed by the SSC package `psmatch2`.
- We demonstrate propensity adjustment, using matching, weighting, and stratification.
- In Phase 1 of the Rubin method, we check for balance and variance inflation, using the SSC packages `somersd` and `haif`[4], respectively.
- And, in Phase 2, we measure the **average treatment effect on the treated (ATET)**, using the SSC package `scenttest`.

Example: the `ldw_exper` dataset of Abadie *et al.*, 2004

- We will demonstrate our methods using a dataset distributed by *The Stata Journal* as online supplemental material for an article on propensity matching[1].
- The dataset has 1 observation per subject in a 1970s observational study, in which 185 subjects participated in a job training program and 260 did not.
- We aim to measure the effect of the training program on 1978 earnings (in 1000s of 1978 dollars), adjusted for a list of 10 confounding covariates, using a logit propensity score computed by the SSC package `psmatch2`.
- We demonstrate propensity adjustment, using matching, weighting, and stratification.
- In Phase 1 of the Rubin method, we check for balance and variance inflation, using the SSC packages `somersd` and `haif`[4], respectively.
- And, in Phase 2, we measure the **average treatment effect on the treated (ATET)**, using the SSC package `scenttest`.

Example: the `ldw_exper` dataset of Abadie *et al.*, 2004

- We will demonstrate our methods using a dataset distributed by *The Stata Journal* as online supplemental material for an article on propensity matching[1].
- The dataset has 1 observation per subject in a 1970s observational study, in which 185 subjects participated in a job training program and 260 did not.
- We aim to measure the effect of the training program on 1978 earnings (in 1000s of 1978 dollars), adjusted for a list of 10 confounding covariates, using a logit propensity score computed by the SSC package `psmatch2`.
- We demonstrate propensity adjustment, using matching, weighting, and stratification.
- In Phase 1 of the Rubin method, we check for balance and variance inflation, using the SSC packages `somersd` and `haif`[4], respectively.
- And, in Phase 2, we measure the **average treatment effect on the treated (ATET)**, using the SSC package `scenttest`.

Example: the `ldw_exper` dataset of Abadie *et al.*, 2004

- We will demonstrate our methods using a dataset distributed by *The Stata Journal* as online supplemental material for an article on propensity matching[1].
- The dataset has 1 observation per subject in a 1970s observational study, in which 185 subjects participated in a job training program and 260 did not.
- We aim to measure the effect of the training program on 1978 earnings (in 1000s of 1978 dollars), adjusted for a list of 10 confounding covariates, using a logit propensity score computed by the SSC package `psmatch2`.
- We demonstrate propensity adjustment, using matching, weighting, and stratification.
- In Phase 1 of the Rubin method, we check for balance and variance inflation, using the SSC packages `somersd` and `haif`[4], respectively.
- And, in Phase 2, we measure the **average treatment effect on the treated (ATET)**, using the SSC package `scenttest`.

Example: the `ldw_exper` dataset of Abadie *et al.*, 2004

- We will demonstrate our methods using a dataset distributed by *The Stata Journal* as online supplemental material for an article on propensity matching[1].
- The dataset has 1 observation per subject in a 1970s observational study, in which 185 subjects participated in a job training program and 260 did not.
- We aim to measure the effect of the training program on 1978 earnings (in 1000s of 1978 dollars), adjusted for a list of 10 confounding covariates, using a logit propensity score computed by the SSC package `psmatch2`.
- We demonstrate propensity adjustment, using matching, weighting, and stratification.
- In Phase 1 of the Rubin method, we check for balance and variance inflation, using the SSC packages `somersd` and `haif`[4], respectively.
- And, in Phase 2, we measure the **average treatment effect on the treated (ATET)**, using the SSC package `scenttest`.

The `ldw_exper` data

And here are the variables, after `compress`ing and adding variable labels:

```
. desc, fu;

Contains data from ldw_exper.dta
  obs:           445
 vars:            12                          7 Apr 2004 21:48
 size:         9,345
------------------------------------------------------------------------------------
              storage  display    value
variable name   type   format     label      variable label
------------------------------------------------------------------------------------
t               byte   %16.0g     t          Participation in the job training program
age             byte   %8.0g                 Age
educ            byte   %8.0g                 Years of education
black           byte   %8.0g                 Indicator for African-American
hisp            byte   %8.0g                 Indicator for Hispanic
married         byte   %8.0g                 Indicator for married
nodegree        byte   %8.0g                 Indicator for > grade school but < high-school e
re74            float  %9.0g                 Earnings in 1974 (1000s of 1978 $)
re75            float  %9.0g                 Earnings in 1975 (1000s of 1978 $)
re78            float  %9.0g                 Earnings in 1978 (1000s of 1978 $)
u74             byte   %9.0g                 Indicator for unemployed in 1974
u75             byte   %9.0g                 Indicator for unemployed in 1975
------------------------------------------------------------------------------------
Sorted by:
```

The outcome is `re78`, the exposure is `t`, and the other 10 are confounders.

Adding propensity scores and weights using `psmatch2`

We use the SSC package `psmatch2`, with the `logit` option (output omitted):

```
. psmatch2 t age educ black hisp married nodegree re74 re75 u74 u75, logit;
```

This adds some new underscored variables, of which the most important are a propensity score and a weight:

```
. desc _pscore _weight, fu;

              storage   display    value
variable name   type    format     label      variable label
-------------------------------------------------------------------------------------------
_pscore         double  %10.0g                psmatch2: Propensity Score
_weight         double  %10.0g                psmatch2: weight of matched controls
```

The weight is 1 for trainees, missing for unmatched controls, and equal to number of matched trainees for matched controls.

Balance checks for propensity matching

To compute the *unadjusted* Somers' *D* of the propensity score and confounding covariates with repect to exposure to training, we use the `somersd` command:

```
somersd t _pscore age educ black hisp married nodegree re74 re75 u74 u75, tdist;
```

To compute sensible matching weights for balance checks, we recall that matching is a special case of weighting, with zero weights for unmatched controls:

```
gene matchwei=cond(missing(_weight),0,_weight);
lab var matchwei "Propensity-matching weight";
```

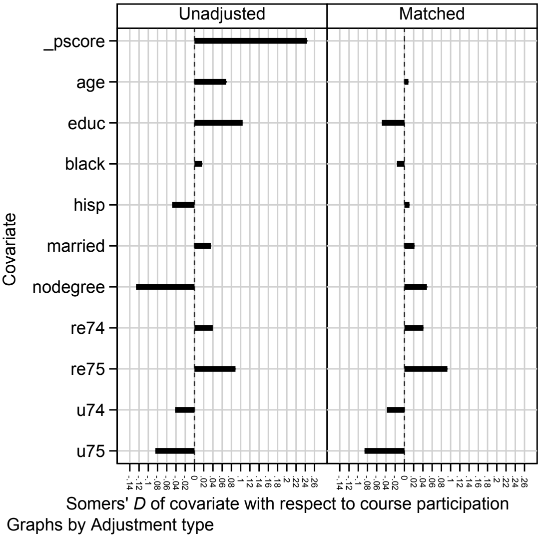
We can now do balance checks for matching by computing *adjusted* Somers' *D* statistics, weighted by the matching weight:

```
somersd t _pscore age educ black hisp married nodegree re74 re75 u74 u75
  [pwei=matchwei], tdist;
```

Both `somersd` commands generate output for 11 parameter estimates, which we will omit. *However*. . .

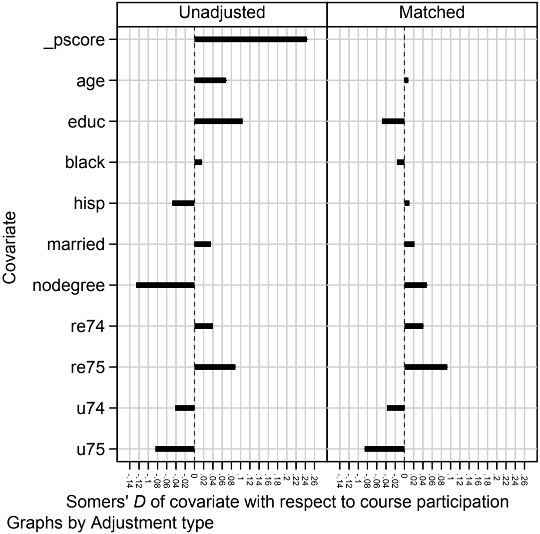
Unadjusted and matched Somers' *D* of covariates with respect to training

- . . . we can plot the two types of Somers' *D* and see instantly how well matching has balanced the covariates.
- Matching has balanced the propensity score well, but *not* all the component covariates.
- Note that confidence limits and *P*–values are not really interesting here.



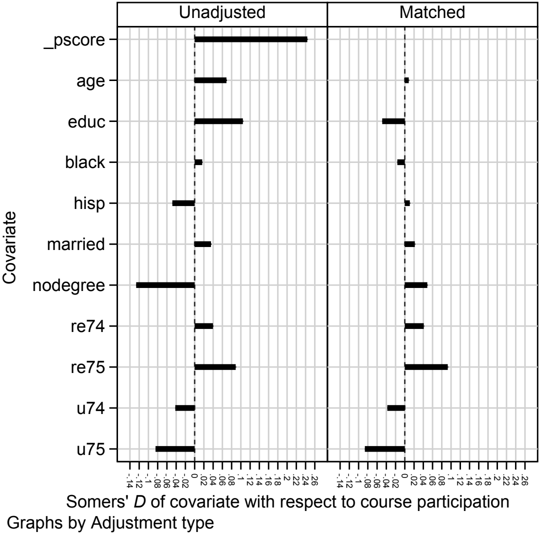
Unadjusted and matched Somers' *D* of covariates with respect to training

- . . . we can plot the two types of Somers' *D* and see instantly how well matching has balanced the covariates.
- Matching has balanced the propensity score well, but *not* all the component covariates.
- Note that confidence limits and *P*–values are not really interesting here.



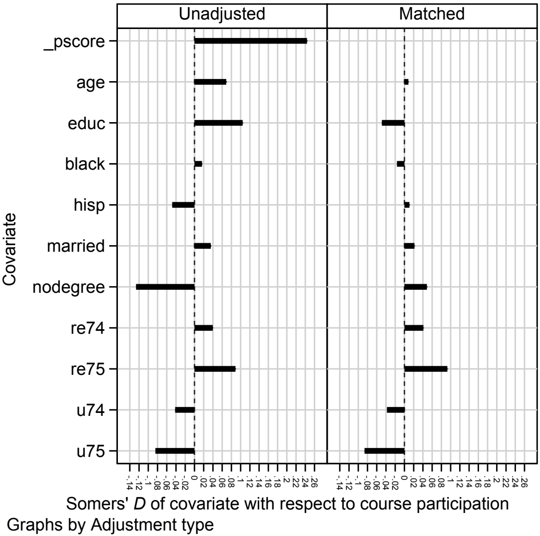
Unadjusted and matched Somers' *D* of covariates with respect to training

- . . . we can plot the two types of Somers' *D* and see instantly how well matching has balanced the covariates.
- Matching has balanced the propensity score well, but *not* all the component covariates.
- Note that confidence limits and *P*–values are not really interesting here.



Unadjusted and matched Somers' *D* of covariates with respect to training

- . . . we can plot the two types of Somers' *D* and see instantly how well matching has balanced the covariates.
- Matching has balanced the propensity score well, but *not* all the component covariates.
- Note that confidence limits and *P*–values are not really interesting here.



Variance inflation for propensity matching

The *costs* of matching are summarized using the `haif` package[4], which measures how much propensity–matching would inflate the required sample number and the confidence interval widths for an equal–variance regression, *assuming* that propensity–matching was not really necessary:

```
. haif t, pweight(matchwei);
Number of observations: 445
Homoskedastic adjustment inflation factors
for variances and standard errors:
              Variance         SE
          t   1.989675   1.410558
      _cons    3.38057   1.838633
```

We see that the variance and standard error of the treatment effect `t` may be greatly inflated. This is not surprising, as matching discards a lot of controls, and weights the others unequally. *However*...

Proceeding to Phase 2 after propensity matching

. . . if we decide to proceed to Phase 2 after all, and add in the outcome (earnings in 1978 Kdollars), then we use a regression command:

```
regress re78 t [pweight=matchwei], vce(robust);
```

This produces some alien–looking output (omitted), but we then use the `scenttest` command to do a scenario *t*–test, comparing treated and untreated scenarios in trainees and their matched controls:

```
. scenttest, at(t=0) atzero(t=1);
Scenario 0: t=1
Scenario 1: t=0
Confidence intervals for the arithmetic means under Scenario 0 and Scenario 1
and for their comparison (arithmetic mean difference)
Total number of observations used: 295
------------------------------------------------------------------------------
             |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
  Scenario_0 |   6.349145   .5788231    10.97   0.000     5.209967    7.488323
  Scenario_1 |   4.207207   .4140692    10.16   0.000      3.39228    5.022134
  Comparison |   2.141939   .7116808     3.01   0.003     .7412844    3.542593
------------------------------------------------------------------------------
```

We see that these subjects are expected to earn 6.349K dollars if trained, or 4.207K dollars if untrained. The difference is 2.142K dollars (95% CI, 0.741K to 3.543K dollars).

Balance checks for propensity weighting

On the other hand, we might decide *not* to proceed to Phase 2, and to ask ourselves whether we should use weighting instead of matching, in order to use *all* the controls. To compute sensible ATET weights for balance checks, we compute weights to be equal to 1 for treated subjects, and to the exposure odds for control subjects:

```
gene atetwei=cond(t==1,1,_pscore/(1-_pscore));
lab var atetwei "Propensity weight for ATET";
```

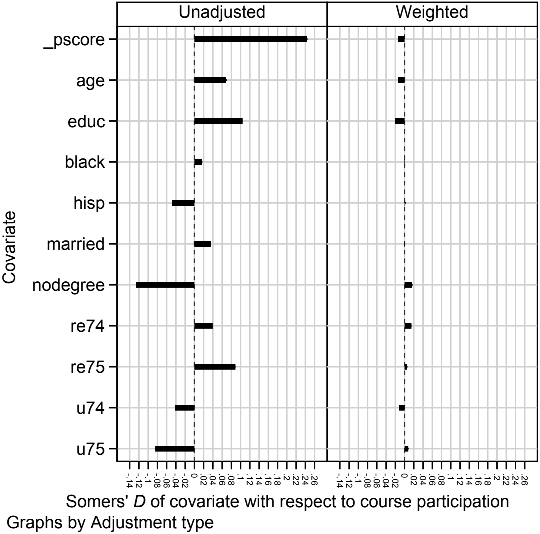
We can now do balance checks for weighting by computing Somers' *D* statistics, weighted by the ATET propensity weights:

```
somersd t _pscore age educ black hisp married nodegree re74 re75 u74 u75
  [pwei=atetwei], tdist;
```

Again, we will omit the command output.

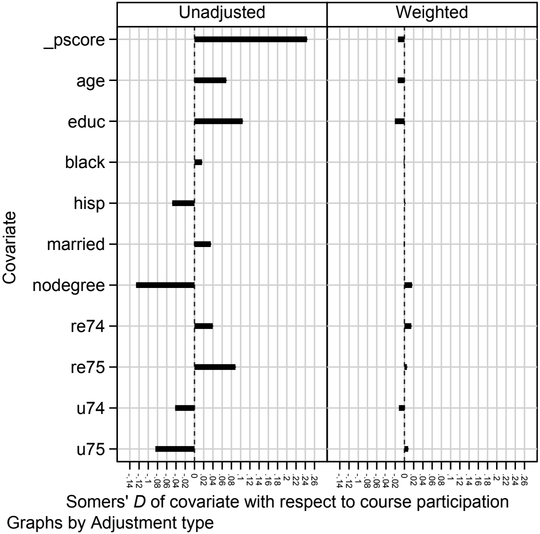
Unadjusted and weighted Somers' *D* of covariates with respect to training

- This time, the weighted Somers' *D* values are *much* closer to zero than the unadjusted ones.
- This is the case for the propensity score *and* for the component covariates.
- *So*, the possibilities for spurious exposure–outcome associations are limited, if we use weighting to compute ATETs.



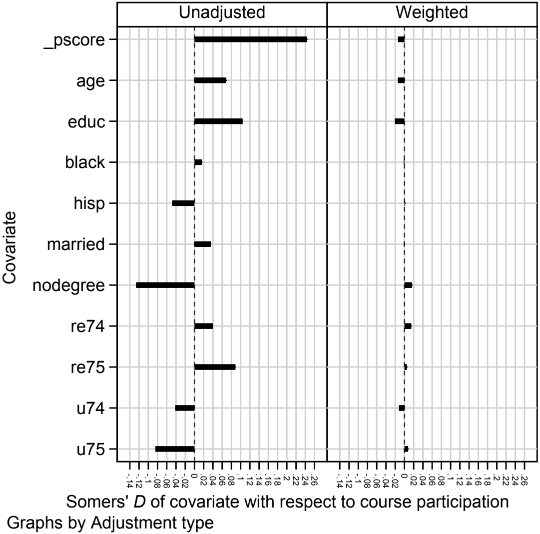
Unadjusted and weighted Somers' *D* of covariates with respect to training

- This time, the weighted Somers' *D* values are *much* closer to zero than the unadjusted ones.
- This is the case for the propensity score *and* for the component covariates.
- *So*, the possibilities for spurious exposure–outcome associations are limited, if we use weighting to compute ATETs.



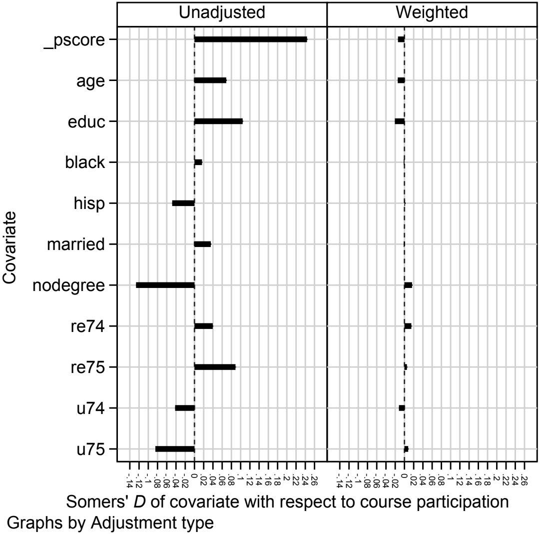
Unadjusted and weighted Somers' *D* of covariates with respect to training

- This time, the weighted Somers' *D* values are *much* closer to zero than the unadjusted ones.
- This is the case for the propensity score *and* for the component covariates.
- *So*, the possibilities for spurious exposure–outcome associations are limited, if we use weighting to compute ATETs.



Unadjusted and weighted Somers' *D* of covariates with respect to training

- This time, the weighted Somers' *D* values are *much* closer to zero than the unadjusted ones.
- This is the case for the propensity score *and* for the component covariates.
- *So*, the possibilities for spurious exposure–outcome associations are limited, if we use weighting to compute ATETs.



Variance inflation for propensity weighting

When we measure the *costs* of weighting using `haif`, the results are again encouraging:

```
. haif t, pweight(atetwei);
Number of observations: 445
Homoskedastic adjustment inflation factors
for variances and standard errors:
               Variance         SE
           t   1.098882   1.048276
       _cons   1.237852   1.112588
```

We see that the variance and standard error of the treatment effect `t` will only be 10 percent and 5 percent larger, respectively, even if the propensity weighting is not really necessary. This is a benefit of using *all* the controls.

Proceeding to Phase 2 after propensity weighting

This time, we might have better reason to proceed to Phase 2, and add in the outcome (earnings in 1978 Kdollars), again using a weighted regression command:

```
regress re78 t [pweight=atetwei], vce(robust);
```

Again, we omit the regression output, and use `scenttest` to do a scenario *t*–test on the ATET:

```
. scenttest, at(t=0) atzero(t=1);
Scenario 0: t=1
Scenario 1: t=0
Confidence intervals for the arithmetic means under Scenario 0 and Scenario 1
and for their comparison (arithmetic mean difference)
Total number of observations used: 445
------------------------------------------------------------------------------
             |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
  Scenario_0 |   6.349145   .5781584    10.98   0.000     5.212871    7.485419
  Scenario_1 |   4.594593   .3984515    11.53   0.000     3.811503    5.377683
  Comparison |   1.754553   .7021615     2.50   0.013     .3745712    3.134534
------------------------------------------------------------------------------
```

This time, subjects like the trained ones are expected to earn 6.349K dollars if trained, or 4.595K dollars if untrained. The difference is 1.755K dollars (95% CI, 0.375K to 3.135K dollars).

Balance checks for propensity stratification

Alternatively, we might use propensity stratification. The strata will be quintiles, which are thought by some to be a fine enough stratification most of the time. For this, we use `xtile`:

```
xtile propgp=_pscore, nq(5);
lab var propgp "Propensity group";
```

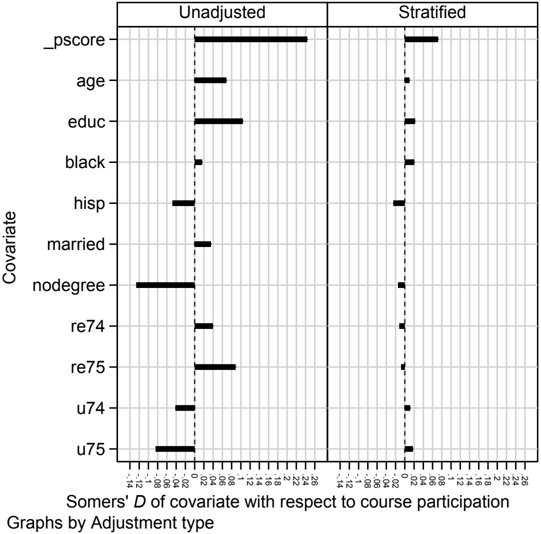
This time, we do balance checks for stratification by computing Somers' *D* statistics, limited to within–strata comparisons by the `wstrata()` option:

```
somersd t _pscore age educ black hisp married nodegree re74 re75 u74 u75,
  tdist wstrata(propgp);
```

Again, we will omit the command output.

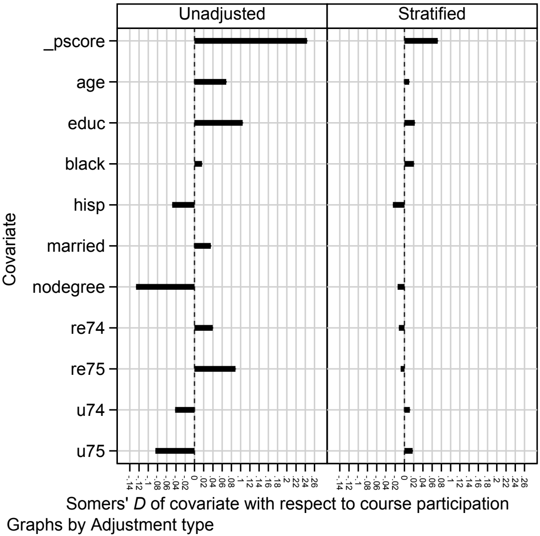
Unadjusted and stratified Somers' *D* of covariates with respect to training

- The stratified Somers' *D* values are close to zero for the component covariates.
- *However*, the Somers' *D* for the propensity score is suspiciously positive.
- This suggests that there is residual exposure–propensity association *within* the quintiles, implying that 5 equal groups are not enough after all.



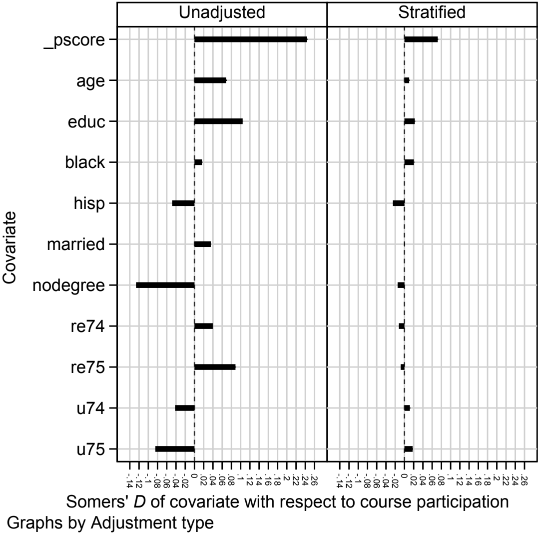
Unadjusted and stratified Somers' D of covariates with respect to training

- The stratified Somers' D values are close to zero for the component covariates.
- *However*, the Somers' D for the propensity score is suspiciously positive.
- This suggests that there is residual exposure–propensity association *within* the quintiles, implying that 5 equal groups are not enough after all.



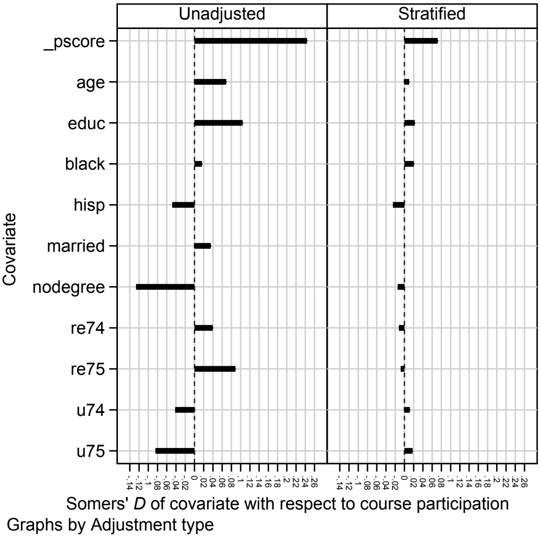
Unadjusted and stratified Somers' *D* of covariates with respect to training

- The stratified Somers' *D* values are close to zero for the component covariates.
- *However*, the Somers' *D* for the propensity score is suspiciously positive.
- This suggests that there is residual exposure–propensity association *within* the quintiles, implying that 5 equal groups are not enough after all.



Unadjusted and stratified Somers' *D* of covariates with respect to training

- The stratified Somers' *D* values are close to zero for the component covariates.
- *However*, the Somers' *D* for the propensity score is suspiciously positive.
- This suggests that there is residual exposure–propensity association *within* the quintiles, implying that 5 equal groups are not enough after all.



Variance inflation for propensity stratification

This time, we measure the *costs* of stratification using the `haifcomp` module of `haif`, and a generated unit variable `const`:

```
. haifcomp t, nadd(ibn.propgp) dadd(const) noconst;
Number of observations: 445
Homoskedastic adjustment inflation factor ratios
for variances and standard errors:
                Variance          SE
           t   1.0549393   1.0271024
```

We see that the variance and standard error of the treatment effect `t` will only be 6 percent and 3 percent larger, respectively, if the propensity stratification is not really necessary. Note that we are assuming a non–interactive regression model. *However*. . .

Proceeding to Phase 2 after propensity stratification

... if we then decide to proceed to Phase 2, and add in the outcome (earnings in 1978 Kdollars), then we should use an *interactive* model:

```
regress re78 ibn.propgp ibn.propgp#c.t, noconst vce(robust);
```

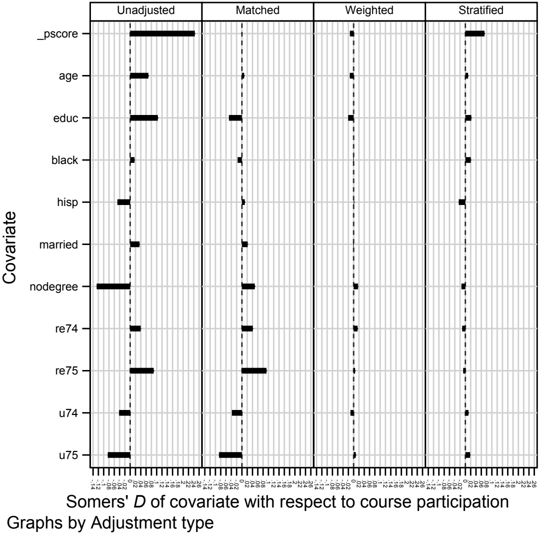
This time, there is even more regression output (omitted), as we have a 10–parameter model, with 1 parameter per treatment level per propensity quintile. `scenttest` summarizes the ATET:

```
. scenttest, at(t=0) atzero(t=1) subpop(if t==1);
Scenario 0: t=1
Scenario 1: t=0
Confidence intervals for the arithmetic means under Scenario 0 and Scenario 1
and for their comparison (arithmetic mean difference)
Total number of observations used: 445
------------------------------------------------------------------------------
             |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
  Scenario_0 |   6.349145   .5811033    10.93   0.000     5.207026    7.491265
  Scenario_1 |   4.498153   .3630063    12.39   0.000     3.784689    5.211617
  Comparison |   1.850993   .6851676     2.70   0.007     .5043419    3.197643
------------------------------------------------------------------------------
```

This time, the trained subjects are expected to earn 6.349K dollars if trained, or 4.498K dollars if untrained. The difference is 1.851K dollars (95% CI, 0.504K to 3.198K dollars).

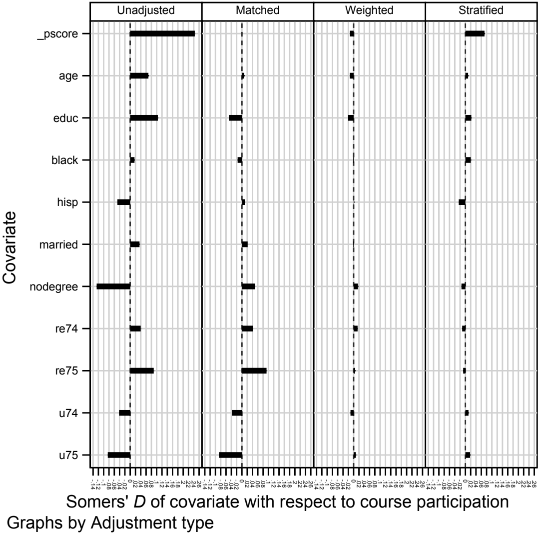
Summary: Balance checks using Somers' *D*

- Here are the unadjusted, matched, weighted and stratified Somers' *D* parameters, for the propensity score and component covariates.
- Of the 3 adjustment methods, weighting seems best at balancing the propensity score *and* the component covariates.
- Propensity weighting therefore seems to be the "best buy".



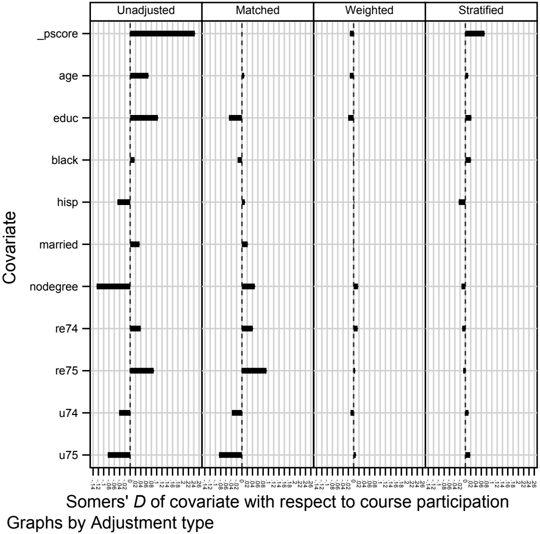
Summary: Balance checks using Somers' *D*

- Here are the unadjusted, matched, weighted and stratified Somers' *D* parameters, for the propensity score and component covariates.
- Of the 3 adjustment methods, weighting seems best at balancing the propensity score *and* the component covariates.
- Propensity weighting therefore seems to be the "best buy".



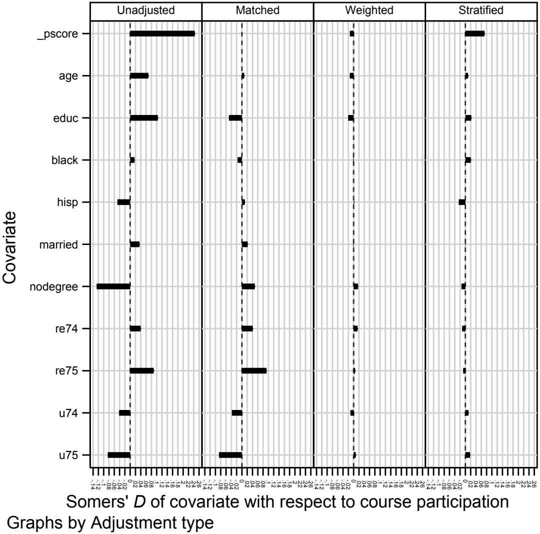
Summary: Balance checks using Somers' *D*

- Here are the unadjusted, matched, weighted and stratified Somers' *D* parameters, for the propensity score and component covariates.
- Of the 3 adjustment methods, weighting seems best at balancing the propensity score *and* the component covariates.
- Propensity weighting therefore seems to be the "best buy".



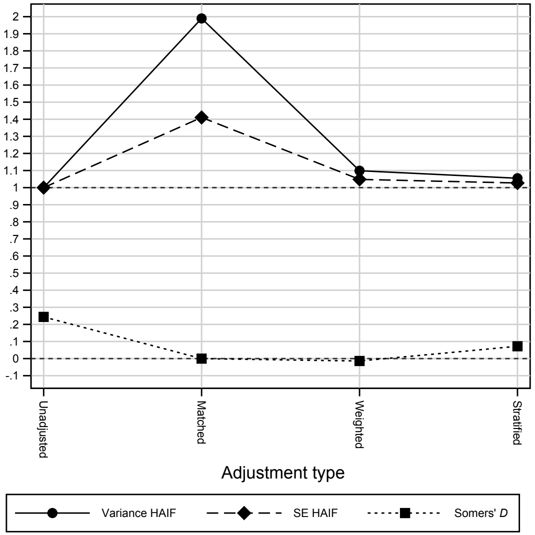
Summary: Balance checks using Somers' *D*

- Here are the unadjusted, matched, weighted and stratified Somers' *D* parameters, for the propensity score and component covariates.
- Of the 3 adjustment methods, weighting seems best at balancing the propensity score *and* the component covariates.
- Propensity weighting therefore seems to be the "best buy".



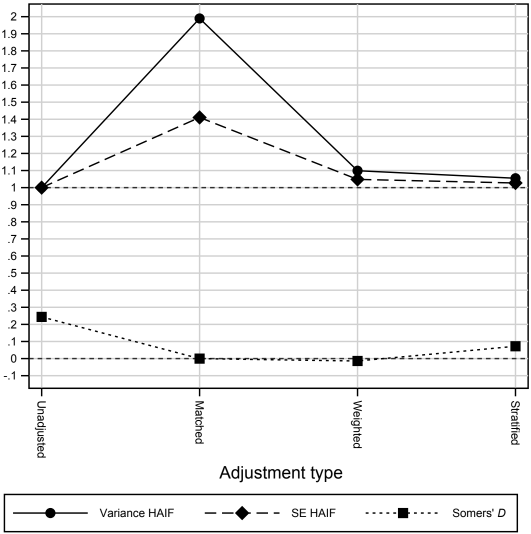
Summary: Costs and benefits of adjustment methods

- The *costs* of adjustment are measured using the variance and SE inflation factors.
- The *benefits* of adjustment are measured using reduction in Somers' *D* of propensity score with respect to exposure.
- Again, propensity weighting seems to be the "best buy".



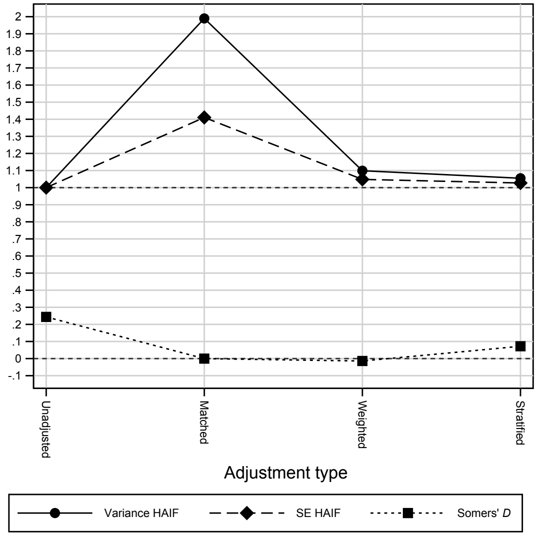
Summary: Costs and benefits of adjustment methods

- The *costs* of adjustment are measured using the variance and SE inflation factors.
- The *benefits* of adjustment are measured using reduction in Somers' *D* of propensity score with respect to exposure.
- Again, propensity weighting seems to be the "best buy".



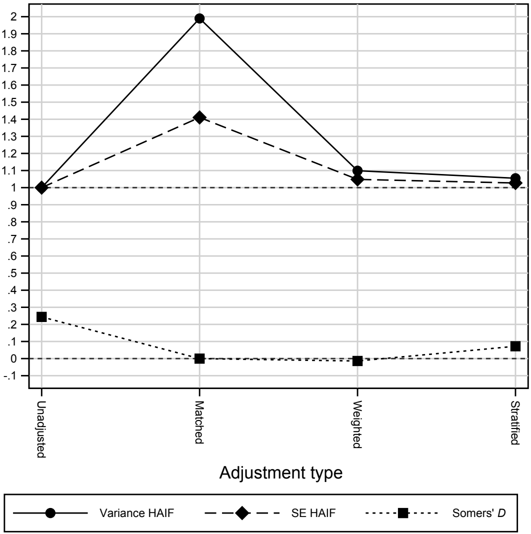
Summary: Costs and benefits of adjustment methods

- The *costs* of adjustment are measured using the variance and SE inflation factors.
- The *benefits* of adjustment are measured using reduction in Somers' *D* of propensity score with respect to exposure.
- Again, propensity weighting seems to be the "best buy".



Summary: Costs and benefits of adjustment methods

- The *costs* of adjustment are measured using the variance and SE inflation factors.
- The *benefits* of adjustment are measured using reduction in Somers' *D* of propensity score with respect to exposure.
- Again, propensity weighting seems to be the "best buy".



References

[1] Abadie, A., Drukker, D., Leber Herr, J. and Imbens, G. W. 2004. Implementing matching estimators for average treatment effects in Stata. *The Stata Journal* **4(3)**: 290–311.

[2] Guo, S. and Fraser, M. W. 2014. *Propensity score analysis. Second edition.* Los Angeles, CA: Sage.

[3] Newson, R. B. 2015. Somers' *D*: A common currency for associations. Presented at the *21st UK Stata User Meeting*, 10–11 September, 2015. Downloadable from the conference website at *http://ideas.repec.org/p/boc/usug15/01.html*

[4] Newson, R. B. 2009. Homoskedastic adjustment inflation factors in model selection. Presented at the *15th UK Stata User Meeting*, 10-11 September, 2009. Downloadable from the conference website at *http://ideas.repec.org/p/boc/usug09/15.html*

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

[6] Rubin, D. B. 2008. For objective causal inference, design trumps analysis. *The Annals of Applied Statistics* **2(3)**: 808–840.

This presentation, and the do–file producing the examples, can be downloaded from the conference website at *http://ideas.repec.org/s/boc/usug16.html*

The packages described and used in this presentation can be downloaded from SSC, using the `ssc` command.