Empirical Bayes Moderation of Asymptotically Linear Parameters

Nima Hejazi

Division of Biostatistics
University of California, Berkeley
stat.berkeley.edu/~nhejazi

nimahejazi.org
twitter/@nshejazi
github/nhejazi

slides: goo.gl/6ou8YR

Preview

1. Linear models are the standard approach for analyzing microarray and next-generation sequencing data (e.g., R package “limma”).

Preview

1. Linear models are the standard approach for analyzing microarray and next-generation sequencing data (e.g., R package “limma”).
2. Moderated statistics help reduce false positives by using an empirical Bayes method to perform standard deviation shrinkage for test statistics.

Preview

1. Linear models are the standard approach for analyzing microarray and next-generation sequencing data (e.g., R package “limma”).
2. Moderated statistics help reduce false positives by using an empirical Bayes method to perform standard deviation shrinkage for test statistics.
3. *Beyond linear models:* we can assess evidence using parameters that are more scientifically interesting (e.g., ATE) by way of TMLE.

Preview

1. Linear models are the standard approach for analyzing microarray and next-generation sequencing data (e.g., R package “limma”).

2. Moderated statistics help reduce false positives by using an empirical Bayes method to perform standard deviation shrinkage for test statistics.

3. *Beyond linear models:* we can assess evidence using parameters that are more scientifically interesting (e.g., ATE) by way of TMLE.

4. The approach of moderated statistics easily extends to the case of asymptotically linear parameters.

Motivation: Let’s meet the data

- Observational study of the impact of occupational exposure (to benzene), with data collected on 125 subjects and roughly $22,000$ biomarkers.

Motivation: Let’s meet the data

- Observational study of the impact of occupational exposure (to benzene), with data collected on 125 subjects and roughly $22,000$ biomarkers.

- Biomarkers of interest are in the form of **miRNA**, assessed using the *Illumina Human Ref-8 BeadChips* platform.

Motivation: Let's meet the data

- Observational study of the impact of occupational exposure (to benzene), with data collected on 125 subjects and roughly $22,000$ biomarkers.
- Biomarkers of interest are in the form of **miRNA**, assessed using the *Illumina Human Ref-8 BeadChips* platform.
- Occupational exposure to benzene reported as discrete values of interest (to epidemiologists): none, $<$ 1ppm, $>$ 5ppm.

Motivation: Let's meet the data

- Observational study of the impact of occupational exposure (to benzene), with data collected on 125 subjects and roughly $22,000$ biomarkers.
- Biomarkers of interest are in the form of **miRNA**, assessed using the *Illumina Human Ref-8 BeadChips* platform.
- Occupational exposure to benzene reported as discrete values of interest (to epidemiologists): none, $<$ 1ppm, $>$ 5ppm.
- Background (phenotype-level) information available on each subject, including age, sex, smoking status.

Data analysis? Linear models!

- For each biomarker ($b = 1, \ldots, B$), fit a linear model:

$$\mathbb{E}[y_b] = X\beta_b$$

Data analysis? Linear models!

- For each biomarker ($b = 1, \ldots, B$), fit a linear model:

$$\mathbb{E}[y_b] = X\beta_b$$

- Generally, we have a particular model coefficent in which we are interested (e.g., effect of benzene on biomarker expression).

Data analysis? Linear models!

- For each biomarker ($b = 1, \ldots, B$), fit a linear model:

$$\mathbb{E}[y_b] = X\beta_b$$

- Generally, we have a particular model coefficent in which we are interested (e.g., effect of benzene on biomarker expression).

- Controlling for baseline covariates, batch effects, and potential confounders happens by adding terms to the linear model.

Data analysis? Linear models!

- For each biomarker ($b = 1, \ldots, B$), fit a linear model:

$$\mathbb{E}[y_b] = X\beta_b$$

- Generally, we have a particular model coefficent in which we are interested (e.g., effect of benzene on biomarker expression).
- Controlling for baseline covariates, batch effects, and potential confounders happens by adding terms to the linear model.
- Test the coefficent of interest using a standard t-test:

$$t_b = \frac{\hat{\beta}_b - \beta_{b,H_0}}{s_b}$$

LIMMA: Linear Models for Microarray Data

- When the sample size is small, s_b^2 may be so small that small differences $(\hat{\beta}_b - \beta_{b,H_0})$ lead to large t_b.

LIMMA: Linear Models for Microarray Data

- When the sample size is small, s_b^2 may be so small that small differences $(\hat{\beta}_b - \beta_{b,H_0})$ lead to large t_b.
- Uncertainty in the variance is an acute problem when the sample size is small.

LIMMA: Linear Models for Microarray Data

- When the sample size is small, s_b^2 may be so small that small differences $(\hat{\beta}_b - \beta_{b,H_0})$ lead to large t_b.
- Uncertainty in the variance is an acute problem when the sample size is small.
- This results in false positives. Smyth proposes we get around this by an empirical Bayes shrinkage of the s_b^2.

LIMMA: Linear Models for Microarray Data

- When the sample size is small, s_b^2 may be so small that small differences $(\hat{\beta}_b - \beta_{b,H_0})$ lead to large t_b.
- Uncertainty in the variance is an acute problem when the sample size is small.
- This results in false positives. Smyth proposes we get around this by an empirical Bayes shrinkage of the s_b^2.
- Test the coefficent of interest with a **moderated** t-test:

$$\tilde{t}_b = \frac{\hat{\beta}_b - \beta_{b,H_0}}{\tilde{s}_b}, \tilde{s}_b^2 = \frac{s_b^2 d_b + s_0^2 d_0}{d_b + d_0}$$

LIMMA: Linear Models for Microarray Data

- When the sample size is small, s_b^2 may be so small that small differences $(\hat{\beta}_b - \beta_{b,H_0})$ lead to large t_b.
- Uncertainty in the variance is an acute problem when the sample size is small.
- This results in false positives. Smyth proposes we get around this by an empirical Bayes shrinkage of the s_b^2.
- Test the coefficent of interest with a **moderated** t-test:

$$\tilde{t}_b = \frac{\hat{\beta}_b - \beta_{b,H_0}}{\tilde{s}_b}, \; \tilde{s}_b^2 = \frac{s_b^2 d_b + s_0^2 d_0}{d_b + d_0}$$

- Eliminates large t-statistics merely from very small s_b.

Beyond linear models

- It's not always desirable to specify a functional form: perhaps we can do better than linear models?

Beyond linear models

- It’s not always desirable to specify a functional form: perhaps we can do better than linear models?
- Such models are a matter of convenience and not honest scientific practice: does $\hat{\beta}_b$ really answer our questions?

Beyond linear models

- It's not always desirable to specify a functional form: perhaps we can do better than linear models?
- Such models are a matter of convenience and not honest scientific practice: does $\hat{\beta}_b$ really answer our questions?
- We can do better by using parameters motivated by causal models (n.b., these will reduce to "variable importance measures" in our case).

Beyond linear models

- It's not always desirable to specify a functional form: perhaps we can do better than linear models?
- Such models are a matter of convenience and not honest scientific practice: does $\hat{\beta}_b$ really answer our questions?
- We can do better by using parameters motivated by causal models (n.b., these will reduce to "variable importance measures" in our case).
- As long as the parameters we seek to estimate have asymptotically linear estimators, we can readily apply the approach of moderated statistics.

Target parameters for complex questions

- Rather than being satisfied with $\hat{\beta}_b$ as an answer to our questions, let's consider a simple target parameter: the average treatment effect (ATE):

$$\Psi_b(P_0) = \mathbb{E}_{W,0}[\mathbb{E}_0[Y_b \mid A = a_{high}, W] - \mathbb{E}_0[Y_b \mid A = a_{low}, W]]$$

Target parameters for complex questions

- Rather than being satisfied with $\hat{\beta}_b$ as an answer to our questions, let's consider a simple target parameter: the average treatment effect (ATE):

$$\Psi_b(P_0) = \mathbb{E}_{W,0}[\mathbb{E}_0[Y_b \mid A = a_{high}, W] - \mathbb{E}_0[Y_b \mid A = a_{low}, W]]$$

- No need to specify a functional form or assume that we know the true data-generating distribution P_0.

Target parameters for complex questions

- Rather than being satisfied with $\hat{\beta}_b$ as an answer to our questions, let's consider a simple target parameter: the average treatment effect (ATE):

  $$\Psi_b(P_0) = \mathbb{E}_{W,0}[\mathbb{E}_0[Y_b \mid A = a_{high}, W] - \mathbb{E}_0[Y_b \mid A = a_{low}, W]]$$

- No need to specify a functional form or assume that we know the true data-generating distribution P_0.

- Parameters like this can be estimated using *targeted minimum loss-based estimation* (TMLE).

Target parameters for complex questions

- Rather than being satisfied with $\hat{\beta}_b$ as an answer to our questions, let's consider a simple target parameter: the average treatment effect (ATE):

$$\Psi_b(P_0) = \mathbb{E}_{W,0}[\mathbb{E}_0[Y_b \mid A = a_{high}, W] - \mathbb{E}_0[Y_b \mid A = a_{low}, W]]$$

- No need to specify a functional form or assume that we know the true data-generating distribution P_0.
- Parameters like this can be estimated using *targeted minimum loss-based estimation* (TMLE).
- **Asymptotic linearity:**

$$\Psi_b(P_n^*) - \Psi_b(P_0) = \frac{1}{n}\sum_{i=1}^{n} IC(O_i) + o_P(\frac{1}{\sqrt{n}})$$

Targeted Minimum Loss-Based Estimation

- TMLE produces a well-defined, unbiased, efficient substitution estimator of target parameters of a data-generating distribution.

Targeted Minimum Loss-Based Estimation

- TMLE produces a well-defined, unbiased, efficient substitution estimator of target parameters of a data-generating distribution.

- Iterative procedure (though there is a one-step now) that updates an initial estimate of the relevant part (Q_0) of the data generating distribution (P_0).

Targeted Minimum Loss-Based Estimation

- TMLE produces a well-defined, unbiased, efficient substitution estimator of target parameters of a data-generating distribution.
- Iterative procedure (though there is a one-step now) that updates an initial estimate of the relevant part (Q_0) of the data generating distribution (P_0).
- Like corresponding A-IPTW estimators, removes asymptotic residual bias of initial estimator for the target parameter. If it uses a consistent estimator of g_0 (nuisance parameter), it is *doubly robust*.

Targeted Minimum Loss-Based Estimation

- TMLE produces a well-defined, unbiased, efficient substitution estimator of target parameters of a data-generating distribution.
- Iterative procedure (though there is a one-step now) that updates an initial estimate of the relevant part (Q_0) of the data generating distribution (P_0).
- Like corresponding A-IPTW estimators, removes asymptotic residual bias of initial estimator for the target parameter. If it uses a consistent estimator of g_0 (nuisance parameter), it is *doubly robust*.
- We can estimate the target parameter:

$$\Psi_b(P_n^*) = \frac{1}{n}\sum_{i=1}^{n}[Q_n^{(b,1)}(A_i = a_h, W_i) - Q_n^{(b,1)}(A_i = a_l, W_i)]$$

Inference with influence curves

- The influence curve for the estimator is:

$$
\begin{aligned}
IC_{b,n}(O_i) = &\left(\frac{\mathbb{1}(A_i = a_h)}{g_n(a_h \mid W_i)} - \frac{\mathbb{1}(A_i = a_l)}{g_n(a_l \mid W_i)} \right) \\
&\cdot (Y_{b,i} - \bar{Q}_n^{(b,1)}(A_i, W_i)) + \bar{Q}_n^{(b,1)}(a_h, W_i) \\
&- \bar{Q}_n^{(b,1)}(a_l, W_i) - \Psi_b(P_n^*)
\end{aligned} \tag{1}
$$

Inference with influence curves

- The influence curve for the estimator is:

$$
\begin{aligned}
IC_{b,n}(O_i) = & \left(\frac{\mathbb{1}(A_i = a_h)}{g_n(a_h \mid W_i)} - \frac{\mathbb{1}(A_i = a_l)}{g_n(a_l \mid W_i)} \right) \\
& \cdot (Y_{b,i} - \bar{Q}_n^{(b,1)}(A_i, W_i)) + \bar{Q}_n^{(b,1)}(a_h, W_i) \\
& - \bar{Q}_n^{(b,1)}(a_l, W_i) - \Psi_b(P_n^*)
\end{aligned}
\qquad (1)
$$

- Sample variance of the influence curve:

$$s^2(IC_n) = \frac{1}{n} \sum_{i=1}^{n} (IC_n(O_i))^2$$

Inference with influence curves

- The influence curve for the estimator is:

$$
\begin{aligned}
IC_{b,n}(O_i) = &\left(\frac{\mathbb{1}(A_i = a_h)}{g_n(a_h \mid W_i)} - \frac{\mathbb{1}(A_i = a_l)}{g_n(a_l \mid W_i)} \right) \\
&\cdot \left(Y_{b,i} - \bar{Q}_n^{(b,1)}(A_i, W_i) \right) + \bar{Q}_n^{(b,1)}(a_h, W_i) \\
&- \bar{Q}_n^{(b,1)}(a_l, W_i) - \Psi_b(P_n^*)
\end{aligned}
\tag{1}
$$

- Sample variance of the influence curve:

$$s^2(IC_n) = \tfrac{1}{n} \sum_{i=1}^{n} \left(IC_n(O_i) \right)^2$$

- Use sample variance to estimate the standard error:

$$se_n = \sqrt{\frac{s^2(IC_n)}{n}}$$

Inference with influence curves

- The influence curve for the estimator is:

$$
\begin{aligned}
IC_{b,n}(O_i) = {} & \left(\frac{\mathbb{1}(A_i = a_h)}{g_n(a_h \mid W_i)} - \frac{\mathbb{1}(A_i = a_l)}{g_n(a_l \mid W_i)} \right) \\
& \cdot \left(Y_{b,i} - \bar{Q}_n^{(b,1)}(A_i, W_i)\right) + \bar{Q}_n^{(b,1)}(a_h, W_i) \\
& - \bar{Q}_n^{(b,1)}(a_l, W_i) - \Psi_b(P_n^*)
\end{aligned}
\tag{1}
$$

- Sample variance of the influence curve:

$$s^2(IC_n) = \tfrac{1}{n} \textstyle\sum_{i=1}^{n} \left(IC_n(O_i)\right)^2$$

- Use sample variance to estimate the standard error:

$$se_n = \sqrt{\frac{s^2(IC_n)}{n}}$$

- Use this for inference — that is, to derive uncertainty measures (i.e., p-values, confidence intervals).

Moderated statistics for target parameters

- One can define a standard t-test statistic for an estimator of an asymptotically linear parameter (over $b = 1, \ldots, B$) as:

$$t_b = \frac{\sqrt{n}(\Psi_b(P_n^*) - \Psi_0)}{s_b(IC_{b,n})}$$

Moderated statistics for target parameters

- One can define a standard t-test statistic for an estimator of an asymptotically linear parameter (over $b = 1, \ldots, B$) as:

$$t_b = \frac{\sqrt{n}(\Psi_b(P_n^*) - \Psi_0)}{s_b(IC_{b,n})}$$

- This naturally extends to the moderated t-statistic of Smyth:

$$\tilde{t}_b = \frac{\sqrt{n}(\Psi_b(P_n^*) - \Psi_0)}{\tilde{s}_b}$$

  where the posterior estimate of the variance of the influence curve is

$$\tilde{s}_b^2 = \frac{s_b^2(IC_{b,n})d_b + s_0^2 d_0}{d_b + d_0}$$

An influence curve transform

- Need the estimate for each biomarker (b) and the IC for every observation for that biomarker, repeating for all $b = 1, \ldots, B$.

An influence curve transform

- Need the estimate for each biomarker (*b*) and the IC for every observation for that biomarker, repeating for all $b = 1, \ldots, B$.

- Essentially, transform original data matrix such that new entries are:

$$Y_{b,i}^{*} = IC_{b,n}(O_i; P_n) + \Psi_b(P_n^{*})$$

An influence curve transform

- Need the estimate for each biomarker (b) and the IC for every observation for that biomarker, repeating for all $b = 1, \ldots, B$.

- Essentially, transform original data matrix such that new entries are:

$$Y^*_{b,i} = IC_{b,n}(O_i; P_n) + \Psi_b(P^*_n)$$

- Since $\mathbb{E}[IC_{b,n}] = 0$ across the columns (units) for each b, the average will be the original estimate $\Psi_b(P^*_n)$.

An influence curve transform

- Need the estimate for each biomarker (b) and the IC for every observation for that biomarker, repeating for all $b = 1, \ldots, B$.

- Essentially, transform original data matrix such that new entries are:

$$Y^*_{b,i} = IC_{b,n}(O_i; P_n) + \Psi_b(P^*_n)$$

- Since $\mathbb{E}[IC_{b,n}] = 0$ across the columns (units) for each b, the average will be the original estimate $\Psi_b(P^*_n)$.

- For simplicity, let's assume the null value is $\Psi_0 = 0$ for all b. Then, applying the moderated t-test to $Y^*_{b,i}$ will generate corrected, conservative test statistics $\tilde{t}_b$.

Why moderated statistics in this context?

- Often times, such data analyses are based on relatively small samples.

Why moderated statistics in this context?

- Often times, such data analyses are based on relatively small samples.
- To get a data-adaptive estimate, with standard implementation of these estimates, standard errors can be non-robust.

Why moderated statistics in this context?

- Often times, such data analyses are based on relatively small samples.
- To get a data-adaptive estimate, with standard implementation of these estimates, standard errors can be non-robust.
- Practically, “significant” estimates of variable importance measures may be driven by poorly and underestimated $s_b^2(IC_{b,n})$.

Why moderated statistics in this context?

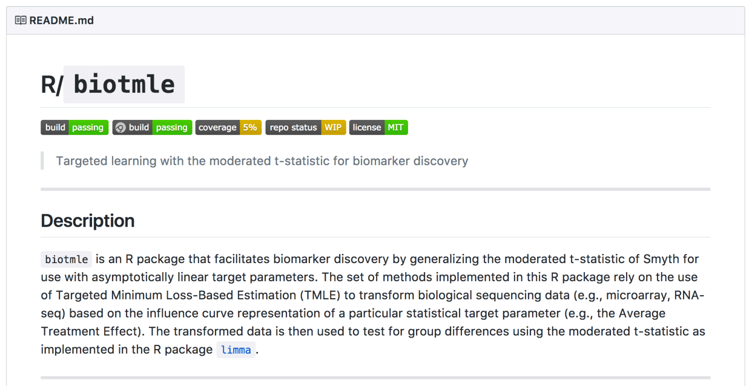
- Often times, such data analyses are based on relatively small samples.
- To get a data-adaptive estimate, with standard implementation of these estimates, standard errors can be non-robust.
- Practically, "significant" estimates of variable importance measures may be driven by poorly and underestimated $s_b^2(IC_{b,n})$.
- Moderated statistics shrink these $s_b^2(IC_{b,n})$ (making them bigger), thus taking biomarkers with small parameter estimates but very small $s_b^2(IC_{b,n})$ out of statistical significance.

Software implementation: “R/biotmle”

- An R package that “facilitates biomarker discovery by generalizing the moderated t-statistic of Smyth for use with asymptotically linear parameters.”

Software implementation: “R/biotmle”

- An R package that “facilitates biomarker discovery by generalizing the moderated t-statistic of Smyth for use with asymptotically linear parameters.”
- Check it out on GitHub: nhejazi/biotmle



README.md

R/ `biotmle`

> Targeted learning with the moderated t-statistic for biomarker discovery

Description

`biotmle` is an R package that facilitates biomarker discovery by generalizing the moderated t-statistic of Smyth for use with asymptotically linear target parameters. The set of methods implemented in this R package rely on the use of Targeted Minimum Loss-Based Estimation (TMLE) to transform biological sequencing data (e.g., microarray, RNA-seq) based on the influence curve representation of a particular statistical target parameter (e.g., the Average Treatment Effect). The transformed data is then used to test for group differences using the moderated t-statistic as implemented in the R package `limma`.

Data analysis with “R/biotmle”

- Observational study of the impact of occupational exposure (to benzene), with data collected on 125 subjects and roughly $22,000$ biomarkers.

Data analysis with “R/biotmle”

- Observational study of the impact of occupational exposure (to benzene), with data collected on 125 subjects and roughly $22,000$ biomarkers.
- Baseline covariates W: age, sex, smoking status; all were discretized.

Data analysis with "R/biotmle"

- Observational study of the impact of occupational exposure (to benzene), with data collected on 125 subjects and roughly $22,000$ biomarkers.
- Baseline covariates W: age, sex, smoking status; all were discretized.
- Treatment A is degree of Benzene exposure: none, $<$ 1ppm, and $>$ 5ppm.

Data analysis with “R/biotmle”

- Observational study of the impact of occupational exposure (to benzene), with data collected on 125 subjects and roughly $22,000$ biomarkers.
- Baseline covariates W: age, sex, smoking status; all were discretized.
- Treatment A is degree of Benzene exposure: none, $<$ 1ppm, and $>$ 5ppm.
- Outcome Y is miRNA expression, median normalized.

Data analysis with “R/biotmle”

- Observational study of the impact of occupational exposure (to benzene), with data collected on 125 subjects and roughly $22,000$ biomarkers.
- Baseline covariates W: age, sex, smoking status; all were discretized.
- Treatment A is degree of Benzene exposure: none, < 1ppm, and > 5ppm.
- Outcome Y is miRNA expression, median normalized.
- Estimate the parameter:

$$\Psi_b(P_n^*) = \mathbb{E}[\mathbb{E}[Y_b \mid A = \max(A), W] - \mathbb{E}[Y_b \mid A = \min(A), W]]$$

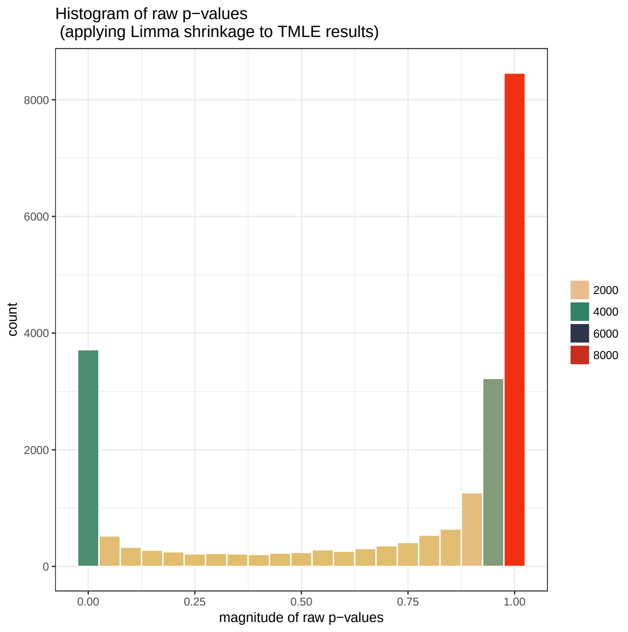
Data analysis with “R/biotmle”

- Observational study of the impact of occupational exposure (to benzene), with data collected on 125 subjects and roughly $22,000$ biomarkers.
- Baseline covariates W: age, sex, smoking status; all were discretized.
- Treatment A is degree of Benzene exposure: none, $<$ 1ppm, and $>$ 5ppm.
- Outcome Y is miRNA expression, median normalized.
- Estimate the parameter:

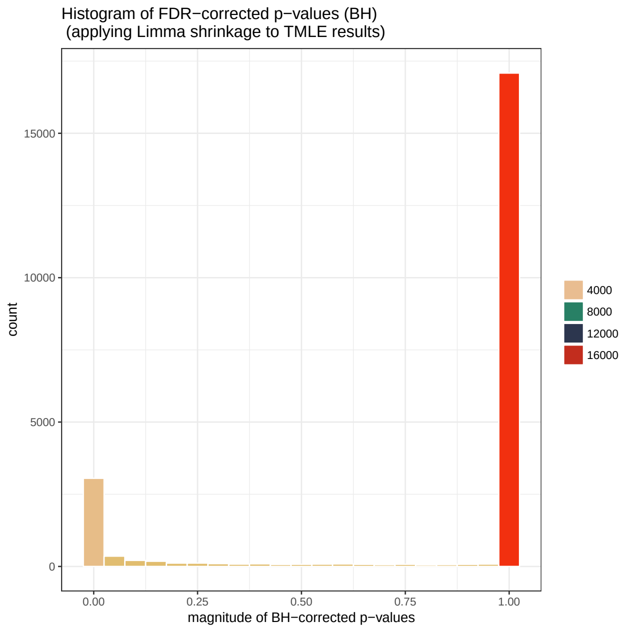
$$\Psi_b(P_n^*) = \mathbb{E}[\mathbb{E}[Y_b \mid A = \max(A), W] - \mathbb{E}[Y_b \mid A = \min(A), W]]$$

- Apply moderated t-test as previously discussed.

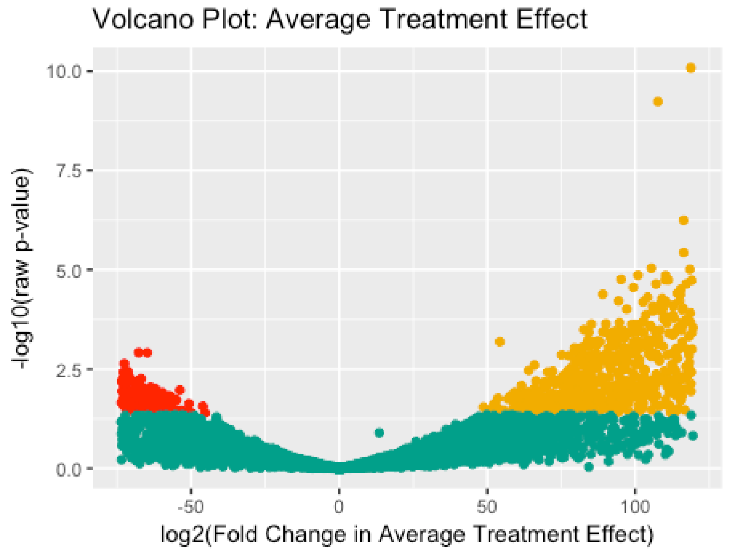
Analysis results I: Uncorrected tests



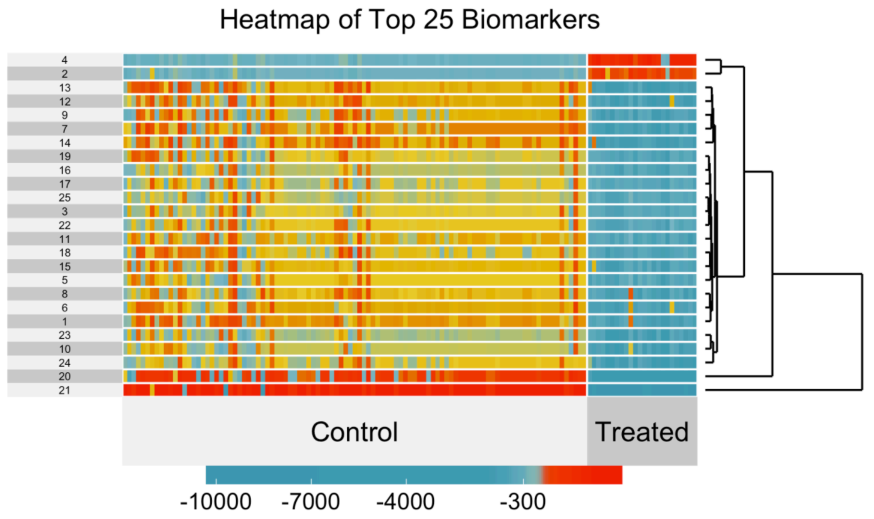
Analysis results II: Corrected tests



Analysis results III: Volcano plot



Analysis results IV: Heatmap of IC estimates



Review

1. Linear models are the standard approach for analyzing microarray and next-generation sequencing data (e.g., R package “limma”).

Review

1. Linear models are the standard approach for analyzing microarray and next-generation sequencing data (e.g., R package “limma”).

2. Moderated statistics help reduce false positives by using an empirical Bayes method to perform standard deviation shrinkage for test statistics.

Review

1. Linear models are the standard approach for analyzing microarray and next-generation sequencing data (e.g., R package “limma”).

2. Moderated statistics help reduce false positives by using an empirical Bayes method to perform standard deviation shrinkage for test statistics.

3. *Beyond linear models:* we can assess evidence using parameters that are more scientifically interesting (e.g., ATE) by way of TMLE.

Review

1. Linear models are the standard approach for analyzing microarray and next-generation sequencing data (e.g., R package "limma").
2. Moderated statistics help reduce false positives by using an empirical Bayes method to perform standard deviation shrinkage for test statistics.
3. *Beyond linear models:* we can assess evidence using parameters that are more scientifically interesting (e.g., ATE) by way of TMLE.
4. The approach of moderated statistics easily extends to the case of asymptotically linear parameters.

References I

Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the royal statistical society. Series B (Methodological)*, pages 289–300.

References II

Smyth, G. K. (2005). Limma: linear models for microarray data. In *Bioinformatics and computational biology solutions using R and Bioconductor*, pages 397–420. Springer.

Tuglus, C. and van der Laan, M. J. (2011). Targeted methods for biomarker discovery. In *Targeted Learning*, pages 367–382. Springer.

van der Laan, M. J. and Rose, S. (2011). *Targeted learning: causal inference for observational and experimental data*. Springer Science & Business Media.

Acknowledgments

Alan Hubbard	University of California, Berkeley
Mark van der Laan	University of California, Berkeley

Slides: goo.gl/6ou8YR

stat.berkeley.edu/~nhejazi

nimahejazi.org

twitter/@nshejazi

github/nhejazi