A Structured Approach to Evaluating Life-Course Hypotheses: Moving Beyond Analyses of Exposed Versus Unexposed in the -Omics Context

Yiwen Zhu, Andrew J. Simpkin, Matthew J. Suderman, Alexandre A. Lussier, Esther Walton, Erin C. Dunn§, and Andrew D.A.C. Smith[§]

[§]Both senior authors contributed equally to this work. Their names appear alphabetically.

Contents

1 Web Appendix 1

This section provides details on the statistical methods examined in the current study. We introduce the regression setup formally, followed by an overview of two variable selection procedures and five methods for making statistical inference in the structured life course modeling approach (SLCMA), which were assessed in the current study. Details on a new confidence interval calculation for the max-$|t|$-test are also provided. A summary of the technical details is provided in **Web Table 3** for quick reference.

For a sample of size n, let $\boldsymbol{y}$ be the vector of responses and $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_p$ be vectors of p predictors. These are assumed to be centered and standardized such that $\sum_{i=1}^n y_i = 0$, and $\sum_{i=1}^n x_{ij} = 0$ and $\sum_{i=1}^n x_{ij}^2 = 1$ for all $j = 1, \ldots, p$. The response $\boldsymbol{y}$ is assumed to be a realization of the random vector generated by the model

$$\boldsymbol{Y} = X\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \tag{1}$$

where $X = (\boldsymbol{x}_1 \cdots \boldsymbol{x}_p)$ and $\boldsymbol{\varepsilon} \sim \mathrm{N}(0, \sigma^2 I)$. The predictor that explains the most variation in the response is the predictor with the largest correlation with the response, i.e. the predictor that maximizes $|\boldsymbol{x}_j^T \boldsymbol{y}|$.

1.1 Variable selection procedures

Two variable selection procedures that find the single predictor that explains the most variation in the outcome (in their first step) are forward stepwise regression and least angle regression (LARS). We therefore considered post-selection inference methods that were developed for these two procedures. This section contains a short overview of these two procedures.

Forward stepwise regression fits a sequence of models with an increasing number of predictors. At each step, the procedure selects the predictor not already in the model that has the largest correlation with the residuals obtained from the current model. At the first step, there are no predictors in the model, the residuals are therefore simply $\boldsymbol{y}$, and the correlations with the residuals are contained in $\boldsymbol{r}$ Hence the first-selected predictor maximizes $|\boldsymbol{x}_j^T \boldsymbol{y}|$.

LARS [7] is related to the lasso [19]. The lasso estimate $\hat{\boldsymbol{\beta}}$ minimizes $\frac{1}{2}||\boldsymbol{y} - X\boldsymbol{\beta}||_2^2 + \lambda||\boldsymbol{\beta}||_1$ for some fixed positive value of the smoothing parameter λ. For sufficiently large λ we have $\hat{\boldsymbol{\beta}} = \boldsymbol{0}$, i.e. the lasso has selected a model with no predictors. As λ decreases the model selected by the lasso increases in complexity. LARS is a procedure for identifying the sequence of lasso models and estimates produced as λ decreases. The first step of LARS identifies the value λ_1 below which the lasso selects its first predictor, the second step identifies the value λ_2 below which a second predictor is selected. The predictor selected at λ_1 is that which maximizes $|\boldsymbol{x}_j^T \boldsymbol{y}|$, as in forward stepwise regression. This second predictor is not necessarily the same as that selected in the second step of forward stepwise regression.

For simplicity of notation, we will assume that $\boldsymbol{x}_1$ is the predictor selected in the first step of both forward stepwise regression and LARS. The model containing only the first-selected predictor simplifies to

$$\boldsymbol{Y} = \boldsymbol{x}_1 \beta_1 + \boldsymbol{\varepsilon}. \tag{2}$$

Note that r_1 is the ordinary least squares estimate for the regression coefficient β_1 in this model.

1.2 Post-selection inference methods

This section gives an overview of methods for calculating P values and confidence intervals for the regression coefficient of the first-selected predictor.

1.2.1 Naïve calculations

A typical implementation of forward stepwise regression ignores the selective nature of the procedure. Inference is essentially based on the test statistic for the first selected predictor, ignoring the fact that this predictor was selected due to its having the largest correlation with the residuals, which would artificially give it a test statistic larger than that expected under the null hypothesis.

In the context of the first predictor selected by forward stepwise regression, this naïve method of inference would involve fitting the simple linear regression model in (2) and testing the hypothesis $H_0 : \beta_1 = 0$ against a two-sided alternative. If H_0 is rejected at the α level of significance, then the probability of making a type I error will be potentially much larger than α. This is because the usual hypothesis test assumes that $\boldsymbol{x}_1$ has been selected from the predictors at random, rather than selected because it has the largest correlation with the response (and hence largest standardized regression coefficient). For $p = 10$ and $\alpha = 5\%$, the probability of a type I error would be approximately 39% using this method [12].

1.2.2 Bonferroni correction

It is possible to control the type I error in the naïve approach by means of a Bonferroni correction, dividing the significance level α by the number of predictors, p (or equivalently, multiplying the P value by the number of predictors and capping at 1). The resulting probability of type I error would be less than α. However, this would be a very conservative approach, as the Bonferroni correction assumes that the regression coefficients are uncorrelated. In practice, this will only occur if the predictors are orthogonal. In the case of general predictors, Bonferroni correction will result in a loss of statistical power. The methods in the following sections attempt to control the probability of type I error without loss of statistical power, by further making use of the correlation between predictors.

1.2.3 Covariance test

Lockhart et al. [12] developed the covariance test as “a significance test for the lasso”. It provides a P value for the selected variable that takes into account the selective nature of the sequence of lasso models.

For the first predictor selected by LARS, the null hypothesis considered by the covariance test is $H_0 : \boldsymbol{\beta} = \mathbf{0}$, and the test statistic is

$$\lambda_1(\lambda_1 - \lambda_2)/\sigma^2. \tag{3}$$

If σ^2 is known this test statistic will have, asymptotically, an Exp(1) distribution under the null hypothesis. Since λ_1 does not depend on the correlation between a particular predictor and the response, but on the maximum correlation between predictor and response, this test statistic takes into account the fact that LARS has not selected a predictor at random, but selected the predictor with the largest correlation with the response.

Lockhart et al. [12] demonstrated the distribution of T_1 with an example with $n = 100$ and $p = 10$ orthogonal predictors. The authors claim that, in their example and under the null hypothesis, the quantiles of T_1 were "decently matched" to those of an Exp(1) distribution.

If the variance σ^2 is unknown it can be replaced by an estimate. Provided that $n > p$, the variance can be estimated by fitting the full linear model (1). When an estimated variance is used, the covariance test statistic will have, asymptotically, an $\mathrm{F}(2, n-p)$ distribution under the null hypothesis. Further details regarding variance estimation for post-selection inference is discussed in Reid et al. [15].

The covariance test does not directly yield a confidence interval for the regression coefficient β_1. Smith et al. [18] proposed a method for modifying the usual confidence interval to account for the selective nature of the model being presented, based on the covariance test P value. They showed using simulations that 95% confidence intervals calculated this way had 95% coverage in a typical structured approach application. However, the usual confidence interval and the covariance test P value are not based on a common set of statistics, and as a result the confidence intervals of Smith et al. [18] and the covariance test P values are not consistent. That is, the 95% confidence interval may contain 0 even if the P value is less than 5%, and vice versa. This can cause confusion if, as is typical in many applications, confidence intervals and P values are displayed side by side in results.

1.2.4 Selective inference

Tibshirani et al. [21] proposed a new set of inference tools applicable to forward stepwise regression and LARS, which are available in the `selectiveInference` R package [20]. The authors identified variable selection procedures that made estimates under polyhedral constraints. As a result, P values and confidence intervals are calculated based on a truncated Gaussian distribution.

For the first step of the variable selection procedure, the standard implementation of the `selectiveInference` package calculates a P value pertaining to the null hypothesis $H_0 : \beta_1 = 0$. The P value is the probability that the estimated regression coefficient would be more extreme than r_1, under H_0 and conditional on the fact that $\boldsymbol{x}_1$ is the first-selected predictor and that the estimated regression coefficient has the same sign as r_1. This P value can be shown to be

$$\frac{1-\Phi(\lambda_1/\sigma)}{1-\Phi(\lambda_2/\sigma)} \tag{4}$$

where Φ is the cumulative distribution function for a standard normal distribution.

In its standard implementation the P values and confidence intervals calculated by the `selectiveInference` package are not consistent. The 95% confidence intervals will contain 0 if and only if the corresponding P value is greater than 2.5%, not 5%. The reason for this is the P value in (4) is effectively that of a one-sided test, due to conditioning on the regression coefficient having the same sign as r_1. Thus a confidence interval would have to be one-sided to be consistent with the P value in (4).

1.2.5 max-$|t|$-test

Buja and Brown [4] proposed a hypothesis test for forward selection, based on the largest t-value of all predictors not yet included in the model at a certain step. We present basic details of this hypothesis test at the first step of the procedure, and a novel method for calculating consistent confidence intervals for the regression coefficient in the first-selected model.

As the predictors $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_p$ share a common scale, the t-values in the first step of the procedure will be proportional to the correlations between the predictors and the response. Therefore we can use the largest correlation as a test statistic. Let $\boldsymbol{r} = X^T\boldsymbol{y}$ be the vector of observed correlations, and let $\boldsymbol{R} = X^T\boldsymbol{Y}$, so that $\boldsymbol{R} \sim \mathrm{N}(X^T\boldsymbol{\beta}, \sigma^2 X^T X)$ under model (1). As we have assumed that $\boldsymbol{x}_1$ is the first predictor selected by LARS and forward selection, as it has the largest correlation with $\boldsymbol{y}$, then r_1 is the observed value of the test statistic, and $|r_1| = \max_j |r_j|$.

We consider the hypothesis $H_0 : \boldsymbol{\beta} = \mathbf{0}$ versus $H_A : \boldsymbol{\beta} \neq \mathbf{0}$. The P value for the max-$|t|$-test equals

$$\begin{aligned} & P_0\Big(|R_1| > |r_1| \Big| |R_1| = \max_j |R_j|\Big) \\ = \quad & 1 - P_0\big(-|r_1| \leq R_1 \leq |r_1| \cap \cdots \cap -|r_1| \leq R_p \leq |r_1|\big). \end{aligned} \tag{5}$$

where P_0 refers to the probability under H_0.

Hence the P value is the probability that $\boldsymbol{R}$ lies outside a cube of radius $|r_1|$. Under H_0, the distribution of $\boldsymbol{R}$ is $\mathrm{N}(\mathbf{0}, \sigma^2 X^T X)$. The probability in (5) can be calculated using existing software for the multivariate normal distribution if σ^2 is known. If σ^2 is unknown, we can estimate it from the residuals of the full linear model as for the covariance test or selective Inference package and calculate the P value from a multivariate t-distribution with $n - p$ degrees of freedom.

Having selected the model in (2) we can construct a 95% confidence interval for β_1 that is consistent with the P value in (5). This confidence interval is the set of all β values that would give a P value not less than 0.05 when testing $H_0 : \beta_1 = \beta$ against a two-sided alternative. Under the model in (2), we have $E(R_1) = \beta_1$, so a suitable test statistic would be $R_1 - \beta$. A 95% confidence interval is given by

$$\begin{aligned} & \Big\{\beta : P_\beta\Big(|R_1 - \beta| > |r_1 - \beta| \Big| |R_1| = \max_j |R_j|\Big) \geq 0.05\Big\} \\ = \quad & \Big\{\beta : P_\beta\Big(|R_1 - \beta| \leq |r_1 - \beta| \Big| |R_1| = \max_j |R_j|\Big)) \leq 0.95\Big\}, \end{aligned}$$

where P_β refers to the probability under $H_0 : \beta_1 = \beta$. The limits of this interval must be found using numerical methods. To calculate whether a certain value of β belongs inside the interval requires calculation of the form

$$\begin{aligned} & P_\beta\Big(R_1 \leq r \Big| |R_1| = \max_j R_j\Big) \\ = \quad & \frac{P_\beta\big(R_1 \leq r \cap |R_1| = \max_j |R_j|\big)}{P_\beta\big(R_1 \leq \infty \cap |R_1| = \max_j |R_j|\big)} \end{aligned} \tag{6}$$

for $r = \beta \pm |r_1 - \beta|$.

Under H_0 we have $\boldsymbol{R} \sim \mathrm{N}(X^T\boldsymbol{x}_1\beta, \sigma^2 X^T X)$. If σ^2 is estimated then a multivariate t-distribution with $n - p$ degrees of freedom should be used for calculation instead of a multivariate normal distribution. Existing software for multivariate normal and t distributions can calculate probabilities over (potentially infinite) cuboid regions. However, probabilities of the form encountered in (6) require calculation over non-cuboid regions. A simple linear transformation allows these probabilities to be calculated using existing software.

The probabilities in (6) can be written as follows

$$\begin{aligned} & P_\beta\big(R_1 \leq r \cap |R_1| = \max_j |R_j|\big) \\ = & \left\{ \begin{array}{ll} 1 - P_\beta\big(R_1 \geq r \cap |R_1| = \max_j |R_j|\big) & r \geq 0 \\ P_{-\beta}\big(R_1 \geq -r \cap |R_1| = \max_j |R_j|\big) & r < 0. \end{array} \right. \end{aligned}$$

We will discuss how to calculate a general probability

$$P\Big(R_1 \geq r \cap |R_1| = \max_j |R_j| \,\Big|\, \boldsymbol{R} \sim \mathrm{N}(\boldsymbol{\mu}, \Sigma)\Big) \tag{7}$$

for $r \geq 0$. Note that

$$\begin{aligned} & R_1 \geq r \cap |R_1| = \max_j |R_j| \\ = \quad & R_1 \geq r \cap -R_1 \leq R_2 \leq R_1 \cap \cdots \cap -R_1 \leq R_p \leq R_1 \end{aligned}$$

and this set of inequalities is equivalent to the intersection of the following set of inequalities:

$$\begin{aligned} R_1 &\geq r \\ R_1 - R_2 &\geq 0 \\ R_1 + R_2 &\geq 0 \\ &\vdots \\ R_1 - R_p &\geq 0 \\ R_1 + R_p &\geq 0. \end{aligned} \tag{8}$$

Let C be a $2p - 1 \times p$ matrix with

$$C_{i,j} = \left\{ \begin{array}{rl} 1 & j = 1 \\ -1 & i = j \text{ even} \\ 1 & i = j \text{ odd} \\ 0 & \text{otherwise} \end{array} \right.$$

Then the inequalities in (8) are satisfied by $\boldsymbol{CR} \geq \boldsymbol{r'}$ where $\boldsymbol{r'} = (r, 0, \ldots, 0)^T$. As C is a linear transformation and $\boldsymbol{R}$ has either a multivariate normal or multivariate t distribution, then $\boldsymbol{R'} = \boldsymbol{CR}$ will have a multivariate normal or t distribution. Hence the probability in (7) is equal to

$$P\Big(\boldsymbol{R'} \geq \boldsymbol{r'} \,\Big|\, \boldsymbol{R'} \sim \mathrm{N}(C\boldsymbol{\mu}, C\Sigma C^T)\Big)$$

and this can be calculated using existing software.

1.3 Simulations setup and data generating process

We included a brief description of the simulations setup in the main text. Here we present full details on the data generating process. We leveraged observed data from the empirical example such that the simulation analyses closely resembled a real-world example of SLCMA application in omics.

1.3.1 Scenario 1: normal outcomes

In the first scenario, the seven exposure variables were resampled from observed data, consisting of five sensitive periods (binary variables taking the value of 0 or 1), accumulation (sum of all five sensitive periods, ranging from 0 to 5), and recency (a weighted sum, with weights defined as age at assessment). The j^{th} outcome (i.e., DNA methylation, $j = 1, \cdots, 485000$) was simulated from a standard normal distribution, $y_j \sim \mathcal{N}(0, 1)$. Because DNA methylation values ('beta' values) may not following a normal distribution, we considered the consequence of having a non-normally distributed outcome in the second scenario. Considering a normally distributed outcome was still useful, as it would help illustrate the performance of the methods when the assumption held.

To assess FWER under the null hypothesis, we ran a single simulation of 485,000 tests and examined the distributions of observed P values against their expected distribution. To assess statistical power and CI coverage, we ran simulations in which the outcome variable was correlated with one of the predictors and then varied the correlation between outcome and predictor such that the variance explained by the predictor r^2 varied from 0.01 to 0.1. When only the first sensitive period hypothesis, denoted by X_1, was simulated to be associated with the outcome, we generated the j^{th} outcome as follows:

$$y_j = X_{1,i}\beta_j + \epsilon_{ij}, \text{where } \beta_j = \sqrt{\frac{r^2}{1 - r^2}}, \epsilon_{ij} \sim \mathcal{N}(0, 1)$$

1.3.2 Scenario 2: empirical outcomes

We aimed to consider non-normally distributed outcomes that closely resemble observed DNA methylation values. The simulation of the exposures was identical to the process described in scenario 1. To assess the FWER under the null hypothesis, we resampled the real predictors and DNAm values from ALSPAC separately. The resampling breaks the predictor-outcome link and hence removes any observed association between the two, while maintaining the empirical distributions of DNAm. In the assessment of statistical power and confidence interval coverage, outcomes were simulated to follow beta distributions and effect sizes were parameterized as the difference in mean levels of DNAm between the exposed and unexposed at the first sensitive period (ΔDNAm), ranging from 0.05 to 0.5.

The number of tests and P value threshold were the same as Scenario 1. We additionally considered a transformation of the DNAm values from beta values (y) to M values equivalent to $M = \log_2 \frac{y}{1-y}$, which are sometimes used to stabilize variance [5].

1.4 Discussion

In addition to the discussion provided in the main text, we would like to highlight a few technical details that may also influence one's preference for one post-selection inference method over another. First of all, the consistency of the confidence interval (CI) and the corresponding P value may make the max-$|t|$-test more favorable. While the selective inference method provided desired confidence interval coverage, the confidence intervals and P values calculated are not consistent: the confidence intervals are two-sided but the P value effectively tests a one-sided hypothesis. Tibshirani et al. [21] argued in favor of this inconsistency, giving the reasons that the one-sided P value would be expected to have more statistical power, while practitioners would prefer to report two-sided confidence intervals. We would further argue that practitioners would prefer to report confidence intervals consistent with observed P values, given that both are frequently reported side by side. Consistent confidence intervals can be provided by the max-$|t|$-test introduced in this paper.

Second, when considering a compound hypothesis in the simulations, we noticed that the statistical power of the selective inference was reduced. As has been noted by Fan and Ke [8] and Bühlmann et al. [3], if there is a second predictor with a non-zero regression coefficient, then λ_2 will be closer to λ_1 and the covariance test statistic will be smaller than if there were no such second predictor. We further note that the selective inference P value (4) will be larger under this scenario. Hence the statistical power of the covariance test and selective inference method may be severely reduced if another predictor also has a large contribution to the outcome variation, even when only considering inference regarding the first-selected predictor. There is no theoretical basis for such a reduction in power when using the max-$|t|$-test, which was consistent with our observation in the simulations. We recommend that practitioners conduct their own simulations to determine statistical power if there is any uncertainty on this point.

Third, post-selection inference methods are available for generalized linear models. Although implementations of the covariance test were available in the `covTest` R package [11], these are no longer recommended by the package authors. The `selectiveInference` package can be used for binary or Cox regression [20], but further simulation is required to confirm its suitability in high-throughput applications. The fact that further Bonferroni correction did not result in a significant loss of statistical power indicates that this conservative method could potentially be used if post-selection software is not available for certain nonlinear regression models.

1.5 Estimating family-wise error rate (FWER)

To further investigate the FWER, we performed repeated simulation experiments under a theoretical scenario. Specifically, we based the setup on the simulations described by Lockhart et al. [12], who used a simulated example to investigate the distribution of the covariance test statistic. We ran 2 000 simulation experiments for each set of parameters to allow the confidence interval of the FWER to have a radius of 1%, setting α to 5%. In each of the 2 000 simulations, we simulated a sample size of $n = 100$ and $p = 10$ uncorrelated predictors with a Gaussian distribution. The response was also generated from a Gaussian distribution. We set $m = 1$, as in Lockhart et al. [12], but also investigated values of m =10, 100, or 1 000 to assess how calculations were affected by the number of tests. The residual variance σ^2 was considered fixed and known, hence the covariance test statistic was considered against an Exp(1) distribution, and a multivariate normal distribution was used for the max-$|t|$-test.

The estimates of FWER for varying numbers of tests performed are presented in **Web Table 4**. As predicted by Lockhart et al. [12], the FWER for the naïve method was not significantly different from 39%, no matter how many tests were performed. In contrast, the conservative Bonferroni correction (using an individual test significance level of $5/pm\%$) gave a FWER that was not significantly different from 5% for all considered values of m. In this scenario, Bonferroni correction is not overly conservative as the predictors are uncorrelated. Hence the p tests of regression coefficients that are implicitly considered during variable selection are independent.

The selective inference method and the max-$|t|$-test gave FWER that were not significantly different from 5% for all considered values of m. The covariance test gave a FWER of approximately 5% for $m = 1$, but for increasing m the FWER increased. For m=1 000 the covariance test FWER was similar to that of the naive method. We took this to indicate that, below 0.05, the P values generated by the covariance test under the null hypothesis are smaller than expected.

The conclusions drawn from this set of simulations are consistent with the conclusions drawn from **Figure 1**.

2 Web Appendix 2

The follow R code shows how the P values and the confidence intervals of the post-selection inference methods compared in this study can be computed in R, where X_hypos is the design matrix, y the outcome, and npred the number of predictors. The code is also available on GitHub: https://github.com/thedunnlab/simulations

```
library(lars)
# archived version of the covTest package can be retrieved here:
# https://cran.r-project.org/src/contrib/Archive/covTest/
library(covTest)
library(selectiveInference)
library(mvtnorm)


## X_hypos: a matrix of the predictors
## y: outcome
## npred: number of predictors
## n: sample size


#### functions for confidence interval for the max-|t|-test ---

# Calculates the probability in (5)
Psi <- function(z, p, mu, df, s2, Corr) {
  C <- rbind(diag(p),-diag(p))
  C <- C[-(p+1),]
  C[,1] <- 1
  pmvt(lower=c(z, rep(0,2*p-2)), upper=rep(Inf,2*p-1),
       delta=as.vector(C %*% mu), df=df, sigma=s2*C %*% Corr %*% t(C), type="shifted")
}


# Calculates the probability in (4)
Pconditional <- function(r, largest, mu, df, s2, Corr) {
  # Reorder so that variable in position 1 is the first one selected
  p <- length(mu)
  mu <- mu[c(largest, (1:p)[-largest])]
  Corr <- Corr[c(largest, (1:p)[-largest]),]
  Corr <- Corr[,c(largest, (1:p)[-largest])]
  # Calculate denominator in (4)
  lower.denom <- Psi(0, p, -mu, df, s2, Corr)
  upper.denom <- Psi(0, p,  mu, df, s2, Corr)
  # Calculate numerator in (4), according to page x
  if(r >= 0) {
    numer <- Psi( r, p,  mu, df, s2, Corr)
    prob <- 1 - numer / (lower.denom + upper.denom)
  } else {
    numer <- Psi(-r, p, -mu, df, s2, Corr)
    prob <- numer / (lower.denom + upper.denom)
  }
  prob
}

Paccept <- function(beta) {
  Pconditional(beta+abs(Xty[selection]-beta), selection,
               XtX[,selection] * beta, n-7,  s^2, XtX) -
    Pconditional(beta-abs(Xty[selection]-beta), selection,
                 XtX[,selection] * beta, n-7,  s^2, XtX) - 0.95
}

#### Run SLCMA ----

# Normalize the design matrix
col_mean <- apply(X_hypos, 2, mean)
X_centered <- X_hypos - rep(col_mean, rep(n, npred)) #subtract mean
col_sss <- apply(X_centered, 2, function(x) sqrt(sum(x^2)))
X_normed <- X_centered / rep(col_sss, rep(n, npred)) #divide by sqrt sum squares

Xt <- t(X_normed)
XtX <- Xt %*% X_normed
Xty <- Xt %*% y

y_centered <- y - mean(y)

## select the predictor with the highest correlation
selection <- which.max(abs(Xty))

## fit OLS
coeftable <- summary(lm(y ~ X_normed[,selection]))$coef


## Naive calculations ----
p.naive <- coeftable[2,4]
lower.naive <- coeftable[2,1] + qt(0.025, n-n_hypo)*coeftable[2,2]
upper.naive <- coeftable[2,1] + qt(0.975, n-n_hypo)*coeftable[2,2]

## Naive calculations + Bonferroni correction ----
p.bonf <- ifelse(p.naive*npred <= 1, p.naive*npred, 1)


## Covariance test ----
lasso <- lars(X_hypos, y)
tt <- covTest(lasso,X_hypos,sigma.est=1,y,maxp=2)$results[1,2]
p.covTest <- 1 - pexp(tt, 1)
# Code from Smith et al. (2015)
thep <- p.covTest/2
lower.covTest <- -1
upper.covTest <- 1
if(thep < 0.05) {
  lower.covTest <- coeftable[2,1]+qnorm((0.025-thep/2)/(1-thep))*coeftable[2,2]
  upper.covTest <- coeftable[2,1]+qnorm((0.975-thep/2)/(1-thep))*coeftable[2,2]
}
if(lower.covTest <= 0 & upper.covTest >= 0 & thep < 0.975) {
  lower.covTest <- coeftable[2,1]+qnorm(0.025/(1-thep))*coeftable[2,2]
  upper.covTest <- coeftable[2,1]+qnorm((0.975-thep)/(1-thep))*coeftable[2,2]
}
if(thep >= 0.975) {
  lower.covTest <- 0
  upper.covTest <- 0
}

## Selective inference ----
larfit <- lar(X_normed, y, maxsteps=3)
inference <- larInf(larfit, type="active", alpha=0.05)
p.sI <- inference$pv[1]
lower.sI <- inference$ci[1,1]
upper.sI <- inference$ci[1,2]

## Max-|t| test ----
absbeta <- abs(Xty[selection])
s <- summary(lm(y_centered ~ X_normed))$sigma
p.maxt <- 1 -
  pmvt(lower=-rep(absbeta,npred),
       upper= rep(absbeta,npred),
       delta= rep(0,npred),
       df= n-npred, sigma= s^2 * XtX)

search_middle <- Xty[selection]
search_radius <- 3*s*XtX[selection,selection]
# lower limit
lower.maxt <- uniroot(Paccept,
        lower=search_middle-search_radius, upper=search_middle)$root
# upper limit
upper.maxt <- uniroot(Paccept,
        lower=search_middle, upper=search_middle+search_radius)$root
```

3 Web Appendix 3

3.1 Sample and procedure

The empirical exposure and outcome data used in our simulations came from the Avon Longitudinal Study of Parents and Children (ALSPAC), a prospective, longitudinal birth cohort of children born to mothers living in the county of Avon, England (120 miles west of London) with estimated delivery dates between April 1991 and December 1992 [9, 2]. Approximately 85 percent of eligible pregnant women agreed to participate (N=14,541), and 99% of eligible live births (n=14,062) who were alive at one year of age (n=13,988 children) were enrolled. Response rates to data collection have been good (75% have completed at least one follow-up). Informed consent for the use of data collected via questionnaires and clinics was obtained from participants following the recommendations of the ALSPAC Ethics and Law Committee at the time. Ethical approval for the study was obtained from the ALSPAC Ethics and Law Committee and the Local Research Ethics Committees. Consent for biological samples has been collected in accordance with the Human Tissue Act 2004 [1]. More details are available on the ALSPAC website, including a fully searchable data dictionary: `http://www.bristol.ac.uk/alspac/researchers/our-data/`. The ALSPAC generated blood-based DNAm profiles at 7 years of age as part of the Accessible Resource for Integrated Epigenomic Studies (ARIES), a subsample of 1018 mother–child pairs from the ALSPAC. The ARIES mother-child pairs were randomly selected out of those with complete data across at least five waves of data collection [16].

3.2 Measures

We used data capturing the exposure to sexual or physical abuse (by anyone) and constructed the following hypotheses: five sensitive periods (at ages 1.5 years, 2.5 years, 3.5 years, 4.75 years, 5.75 years, and 6.75 years), accumulation, and recency. The first sensitive period hypothesis was set to be the true underlying hypothesis in the power and confidence interval coverage simulations. The prevalence of the exposure at the six time points ranged from 2.61% to 3.96%. Exposure to sexual or physical abuse was determined through an item asking the mother to indicate whether or not the child had been exposed to either sexual or physical abuse from anyone at each of the six time points listed above. Reports of sexual or physical abuse were not reported to child welfare agencies. Other available types of exposure to childhood adversity in ALSPAC are described by Dunn et al. [6].

DNAm was measured at 485,000 CpG dinucleotide sites across the genome using the Illumina Infinium Human Methylation 450K BeadChip microarray. DNA for this assay was extracted from peripheral blood leukocytes at age 7. DNAm levels are expressed as a 'beta' value, representing the proportion of cells methylated at each interrogated CpG site. Detailed descriptions of the preprocessing and quality control procedures are provided elsewhere [6, 16].

The covariates included in the empirical analyses were consistent with the adjustment by Dunn et al. [6]. They were: child sex, child race and/or ethnicity; child birth weight; maternal age; number of previous pregnancies; sustained maternal smoking during pregnancy; and parent social class.

3.3 Adjusting for covariates

While we focused on assessing relationships between exposures capturing life course theories and omics outcomes in the simulations, the associations are usually confounded by other factors in practice, such as socioeconomic status or maternal smoking status during pregnancy [6]. In previous studies, adjustment had been formerly done by regressing exposures on the covariates and using the residuals of the exposures in the SLCMA [18].

An alternative method that can be used to adjust for covariates in a linear regression setting is to apply the Frisch-Waugh-Lovell (FWL) theorem, or partitioned regression [10, 13]. The method has been proven to yield the same regression coefficients and residual variance as a fully adjusted model [14]. The FWL theorem was first proposed by two econometricians, Frisch and Waugh [10], to highlight a useful property of ordinary least squares such that a two-step approach to detrend the independent and dependent variables yields the same regression coefficients as a fully adjusted regression model with the trend variables included as covariates. Lovell showed that the adjustment remains true for any nonempty subset of explanatory variables (i.e., it does not just apply to trend variables) [13]. The proof of this theorem can be found in several previous publications [10, 13, 14]. It has since been proven in the context of penalized regression as well, such as the lasso or ridge regression [22].

However, it remained unclear whether this theorem would be applicable to post-selection inference methods (such as selective inference or the max-$|t|$-test) and whether additionally regressing the outcomes on the covariates would result in smaller residual variances and larger test statistics. Therefore, we assessed the FWER in a similar manner as in scenario 2 presented in the main results, by running one simulation experiment with resampled empirical outcomes (n=700). As seen in **Web Figure 5**, the P value distributions were similar to what we observed without applying the FWL theorem. There was no inflation in the observed P value distributions.

To evaluate the potential improvement in statistical power, we repeated the empirical analyses included in the current study using the selective inference method and max-$|t|$-test, additionally regressing DNAm values on the confounders. Comparing the P values of the five top CpG sites obtained from the selective inference method and max-$|t|$-test before and after applying the FWL theorem, we found that the P values decreased at all five loci (**Web Figure 4**). The P value at *cg*06430102 exceeded the estimated 450K array-wide threshold after the additional adjustment [17], suggesting that the approach improved statistical power.

Given the evidence observed here, we recommend applying the FWL theorem and regress both the exposure variables and the outcome on confounders before subsequent SLCMA analyses. This approach may effectively increase statistical power and overcome bias due to confounding.

4 Web Tables 1 to 4

Web Table 1: Comparison of effect estimates and confidence intervals of the top CpG sites in the empirical analyses, calculated using the covariance test, selective inference, and max-$|t|$-test.

CpG	**First hypothesis chosen**	**DNAm in unexposed group**	**DNAm in exposed group**	**Increase in** R^2	**Effect estimate**	**Post-selection inference method**	***P* value**	**Lower 95% CI**	**Upper 95% CI**
cg01370449	Very early childhood (2.5 years of age)	0.2439	0.3341	0.0297	0.084	Max-$|t|$-test	1.23E-05	0.0532	0.118
						Covariance test	8.87E-08	0.0501	0.1179
						Selective inference	8.09E-06	0.0493	0.1183
cg06430102	Very early childhood (2.5 years of age)	0.9257	0.8619	0.0368	-0.058	Max-$|t|$-test	5.58E-07	-0.0789	-0.0384
						Covariance test	1.69E-09	-0.0789	-0.037
						Selective inference	5.32E-07	-0.0791	-0.0367
cg19170021	Early childhood (4.75 years of age)	0.7342	0.8275	0.0275	0.0958	Max-$|t|$-test	5.79E-05	0.0578	0.1374
						Covariance test	6.41E-08	0.0542	0.1373
						Selective inference	1.47E-05	0.0536	0.1378
cg05072819	Early childhood (5.75 years of age)	0.0401	0.0534	0.0305	0.0141	Max-$|t|$-test	8.87E-06	0.0089	0.0198
						Covariance test	3.49E-08	0.0084	0.0198
						Selective inference	5.70E-06	0.0083	0.0199
cg05936516	Middle childhood (6.75 years of age)	0.1279	0.1532	0.0311	0.0255	Max-$|t|$-test	3.26E-06	0.0164	0.0354
						Covariance test	7.47E-08	0.0156	0.0354
						selective inference	5.43E-06	0.0153	0.0355

Web Table 2: Overlap in most strongly associated loci based on results obtained from the covariance test and the two recommended methods (max-$|t|$-test and selective inference) in the empirical analyses.

Number of top loci	**Selective inference**	**Max-$\|t\|$-test**
10	100%	50%
50	84 %	54%
100	89%	56%
1000	91%	55%
2000	93%	56%
5000	94%	58%

For example, the first line indicates that for the first 10 loci identified by the covariance test, all of them were also among the top 10 based on the selective inference results. However, only half of them appeared among the top 10 identified using the max-t-test.

Web Table 3: Summary of the most popular statistical inference methods used in the SLCMA to identify the best fitting theoretical model.

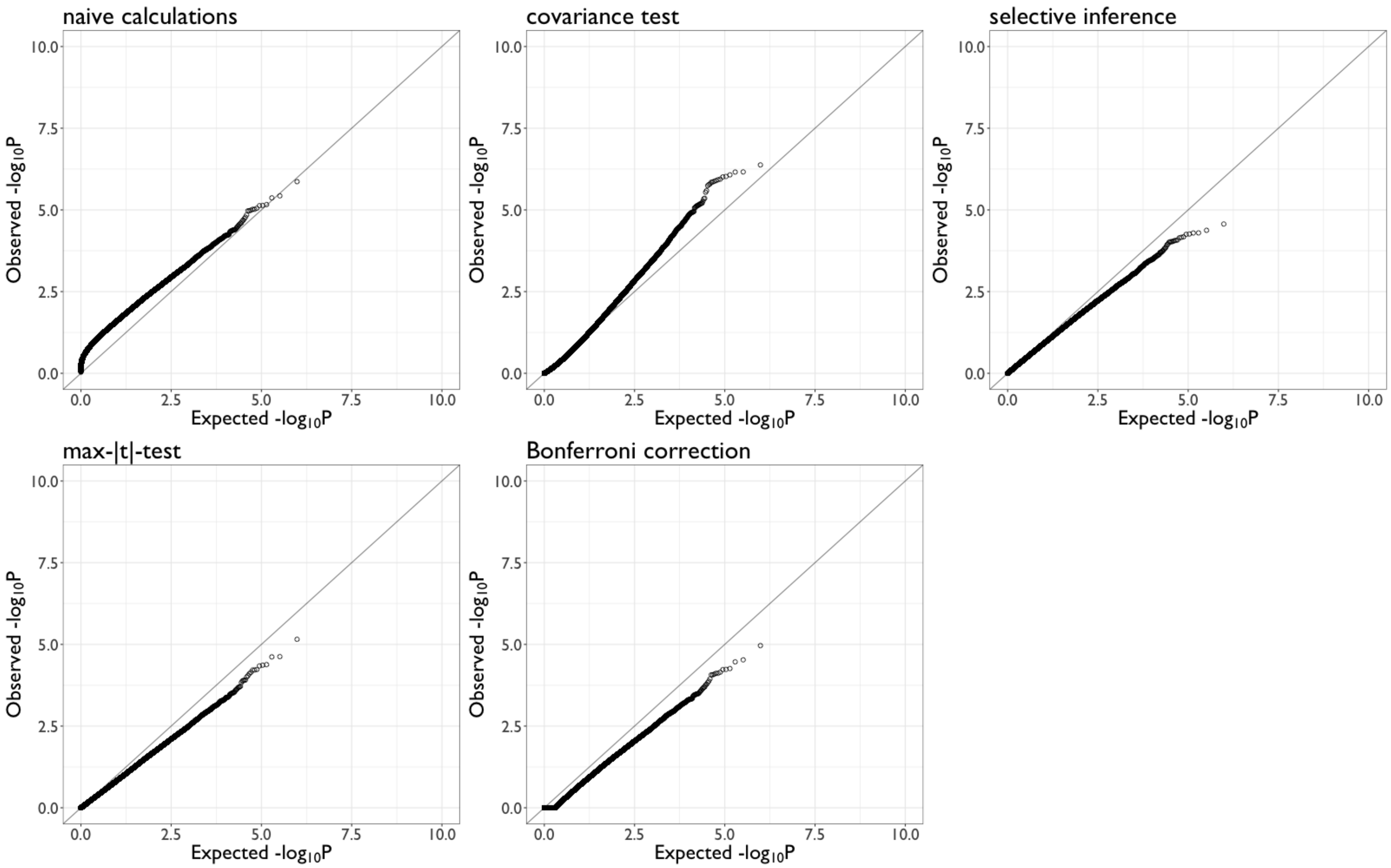
Method	**Model selection procedure**	**Post Selection Inference**		
		Test statistic	**Strategy to address multiple testing and selection**	**Procedure to derive confidence intervals**
Naïve calculations	Forward stepwise regression and least angle regression are equivalent when considering just the predictor with the largest correlation with the outcome	$\hat{\beta}_{OLS} = x_1^T y$	NA	Ordinary least squares (OLS)
Bonferroni correction		$\hat{\beta}_{OLS} = x_1^T y$	Bonferroni correction	NA
Max-$\|t\|$-test		$r_1 = x_1^T y$ where x_1 is the predictor that has the largest correlation with the outcome	Condition the test statistic distribution on it having the maximal correlation with the outcome	Linear transformation of non-cuboid space; can be calculated using existing software
Covariance test		$\lambda_1(\lambda_1 - \lambda_2)/\sigma^2$, where λ_1 and λ_2 are the values of the smoothing parameters at the first and second step of LARS	Condition the test statistic distribution on it having the maximal correlation with the outcome	A modification of the OLS confidence intervals using the corresponding covariance test pvalues
Selective inference		p-value can be shown to be: $$\frac{1-\Phi(\lambda_1/\sigma)}{1-\Phi(\lambda_2/\sigma)}$$ where Φ is the cumulative distribution function for a standard normal distribution	Conceptualized the selection as responses being in a polyhedral set	Inverting the test statistic

Web Table 4: Estimated family-wise error rate and corresponding 95% CI in a theoretical scenario, after 2 000 simulation experiments

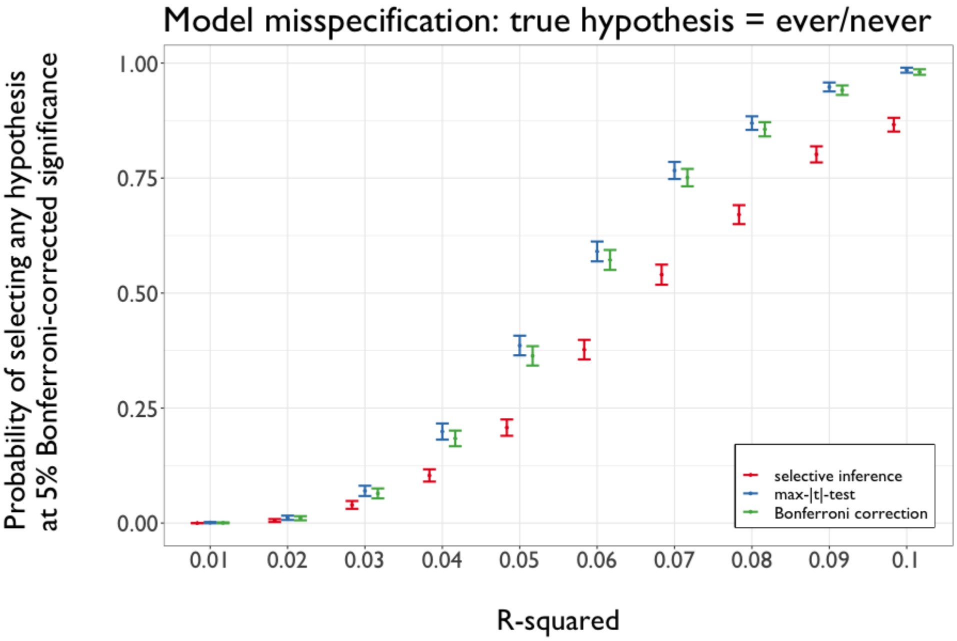
Number of tests	**1**	**10**	**100**	**1000**
Naïve calculation	38.6% (36.5-40.8)	38.8% (36.7-40.9)	38.8% (36.7-40.9)	38.1% (36.0-40.3)
Bonferroni correction	4.9%(3.9-5.8)	4.2% (3.4-5.1)	5.3% (4.4-6.3)	5.1% (4.1-6.0)
Covariance test	6.0% (5.0-7.0)	12.1% (10.7-13.5)	22.5% (20.7-24.3)	42.7% (40.5-44.9)
Selective inference	5.1% (4.1-6.0)	5.2% (4.3-6.2)	5.1% (4.1-6.1)	5.1% (4.1-6.0)
Max-$\|t\|$-test	5.0% (4.0-5.9)	4.2% (3.4-5.1)	5.3% (4.4-6.3)	5.1% (4.1-6.0)

5 Web Figures 1 to 5

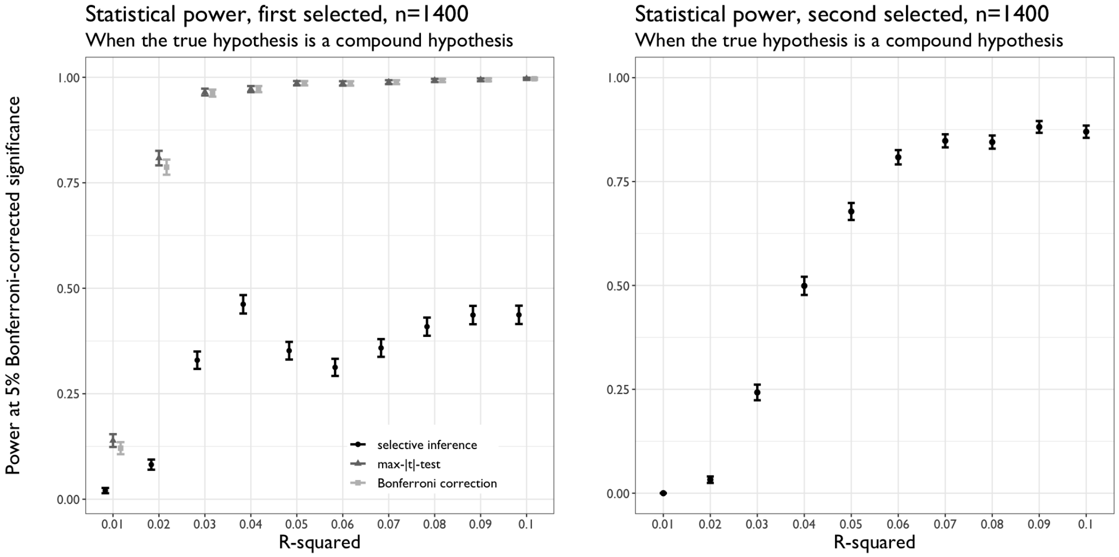
Web Figure 1: Q-Q plots comparing the expected and observed P values simulated under the null for all five methods with empirical outcomes (N=700), where the outcome variables were resampled from observed DNAm values and transformed to M-values.



Web Figure 2: Estimated probability of selecting any hypothesis with a 5% Bonferroni corrected P value threshold under model misspecification.



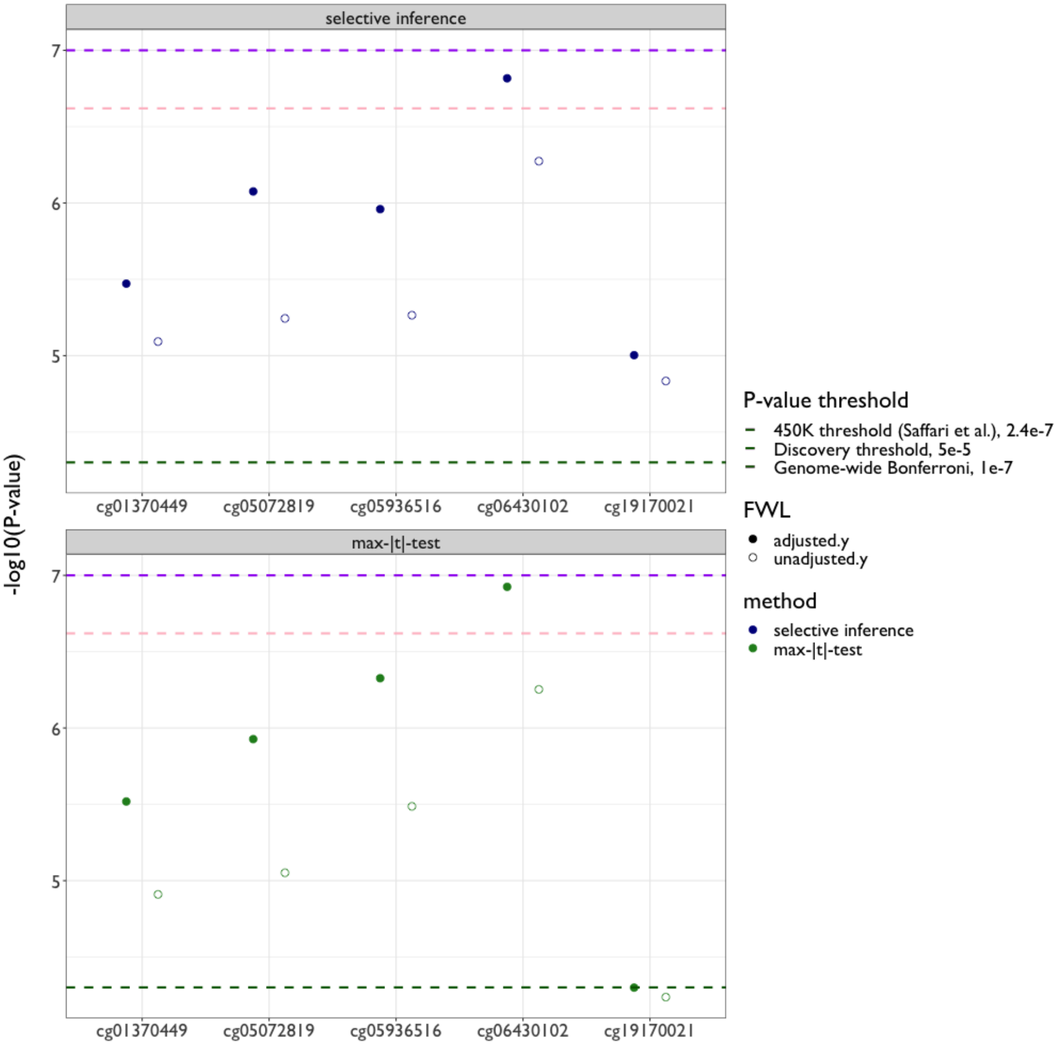
Web Figure 3: Estimated statistical power and corresponding 95% CI in simulated epigenome-wide analyses with increased sample size (n=1400), with varying effect sizes, when the true causal relationship was represented by two hypotheses working in combination.



Statistical power, first selected, n=1400
When the true hypothesis is a compound hypothesis

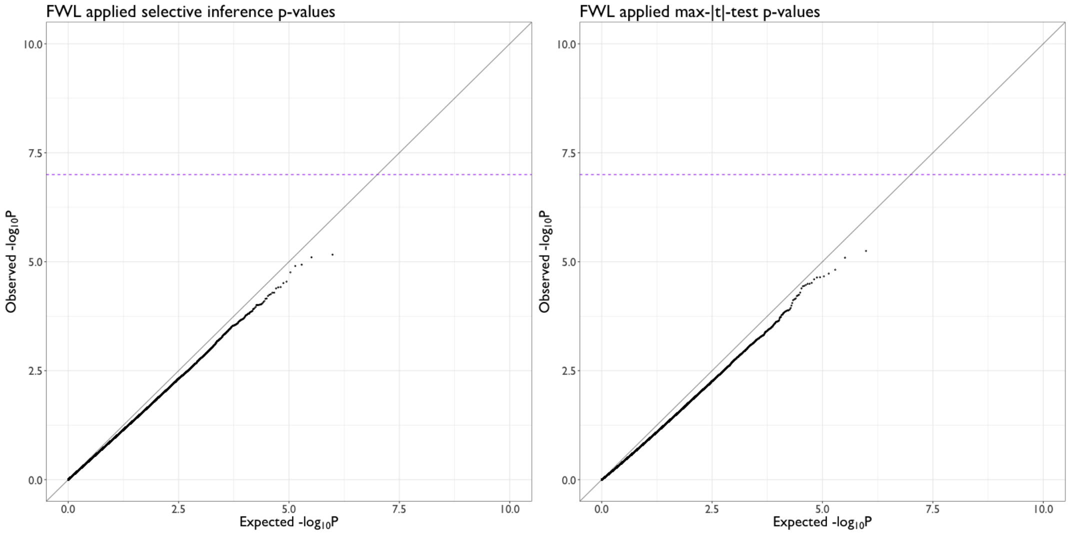
Statistical power, second selected, n=1400
When the true hypothesis is a compound hypothesis

Web Figure 4: Differences in P values of the top CpG sites, before and after applying the FWL theorem, obtained from the selective inference method and max-$|t|$-test



The plot shows the change in $-\log_{10}(p)$ obtained using the two recommended methods: selective inference and the max-$|t|$-test. Open dots represent P values before additionally adjusting for the covariates following the FWL theorem; solid dots represent P values after the FWL adjustment (i.e., regressing the outcome on the covariates and analyzing the residuals). The three dashed lines in different colors denote three commonly used threshold considered in genome-wide DNA methylation studies.

Web Figure 5: Q-Q plots comparing the expected versus observed P values simulated under the null for selective inference and max-—t—-test with empirical outcomes (N=700), after applying the Frisch-Waugh-Lovell theorem to adjust for covariates.



FWL applied selective inference p-values

FWL applied max-|t|-test p-values

References

[1] Human Tissue Act 2004. http://www.legislation.gov.uk/ukpga/2004/30/pdfs/ukpga_20040030_en.pdf, 2004.

[2] Boyd, A., Golding, J., Macleod, J., Lawlor, D. A., Fraser, A., Henderson, J., Molloy, L., Ness, A., Ring, S., and Davey Smith, G. Cohort Profile: The 'Children of the 90s'—the index offspring of the Avon Longitudinal Study of Parents and Children. *International Journal of Epidemiology 42*, 1 (Feb. 2013), 111–127.

[3] Bühlmann, P., Meier, L., and van de Geer, S. Discussion: "A significance test for the lasso". *Annals of Statistics 42*, 2, 469–477.

[4] Buja, A., and Brown, L. Discussion: "A significance test for the lasso". *Annals of Statistics 42*, 2 (Apr. 2014), 509–517.

[5] Du, P., Zhang, X., Huang, C.-C., Jafari, N., Kibbe, W. A., Hou, L., and Lin, S. M. Comparison of Beta-value and M-value methods for quantifying methylation levels by microarray analysis. *BMC Bioinformatics 11* (Nov. 2010), 587.

[6] Dunn, E. C., Soare, T. W., Zhu, Y., Simpkin, A. J., Suderman, M. J., Klengel, T., Smith, A. D. A. C., Ressler, K. J., and Relton, C. L. Sensitive Periods for the Effect of Childhood Adversity on DNA Methylation: Results From a Prospective, Longitudinal Study. *Biological Psychiatry 85*, 10 (May 2019), 838–849.

[7] Efron, B., Hastie, T., Johnstone, I., and Tibshirani, R. Least angle regression. *The Annals of statistics 32*, 2 (2004), 407–499.

[8] Fan, J., and Ke, Z. T. Discussion: "A significance test for the lasso". *Annals of Statistics 42*, 2, 483–492.

[9] Fraser, A., Macdonald-Wallis, C., Tilling, K., Boyd, A., Golding, J., Davey Smith, G., Henderson, J., Macleod, J., Molloy, L., Ness, A., Ring, S., Nelson, S. M., and Lawlor, D. A. Cohort Profile: The Avon Longitudinal Study of Parents and Children: ALSPAC mothers cohort. *International Journal of Epidemiology 42*, 1 (Feb. 2013), 97–110.

[10] Frisch, R., and Waugh, F. V. Partial Time Regressions as Compared with Individual Trends. *Econometrica 1*, 4 (Oct. 1933), 387.

[11] Lockhart, R., Taylor, J., Tibshirani, R., and Tibshirani, R. covTest: Computes covariance test for adaptive linear modelling, R package version 1.02.

[12] Lockhart, R., Taylor, J., Tibshirani, R. J., and Tibshirani, R. A significance test for the lasso. *Annals of Statistics 42*, 2 (Apr. 2014), 413–468.

[13] Lovell, M. C. Seasonal Adjustment of Economic Time Series and Multiple Regression Analysis. *Journal of the American Statistical Association 58*, 304 (Dec. 1963), 993.

[14] Lovell, M. C. A Simple Proof of the FWL Theorem. *The Journal of Economic Education 39*, 1 (Jan. 2008), 88–91.

[15] Reid, S., Tibshirani, R., and Friedman, J. A study of error variance estimation in Lasso regression. *Statistica Sinica* (2016).

[16] Relton, C. L., Gaunt, T., McArdle, W., Ho, K., Duggirala, A., Shihab, H., Woodward, G., Lyttleton, O., Evans, D. M., Reik, W., Paul, Y.-L., Ficz, G., Ozanne, S. E., Wipat, A., Flanagan, K., Lister, A., Heijmans, B. T., Ring, S. M., and Davey Smith, G. Data Resource Profile: Accessible Resource for Integrated Epigenomic Studies (ARIES). *International journal of epidemiology 44*, 4 (Aug. 2015), 1181–1190.

[17] Saffari, A., Silver, M. J., Zavattari, P., Moi, L., Columbano, A., Meaburn, E. L., and Dudbridge, F. Estimation of a significance threshold for epigenome-wide association studies. *Genetic Epidemiology 42*, 1 (Feb. 2018), 20–33.

[18] Smith, A. D. A. C., Heron, J., Mishra, G., Gilthorpe, M. S., Ben-Shlomo, Y., and Tilling, K. Model Selection of the Effect of Binary Exposures over the Life Course. *Epidemiology (Cambridge, Mass.) 26*, 5 (Sept. 2015), 719–726.

[19] Tibshirani, R. Regression Shrinkage and Selection via the Lasso. *Journal of the Royal Statistical Society. Series B (Methodological) 58*, 1 (1996), 267–288.

[20] Tibshirani, R., Tibshirani, R., Taylor, J., Loftus, J., and Reid, S. selective-Inference: Tools for post-selection inference, R package version 1.2.0.

[21] Tibshirani, R. J., Taylor, J., Lockhart, R., and Tibshirani, R. Exact Post-Selection Inference for Sequential Regression Procedures. *Journal of the American Statistical Association 111*, 514 (Apr. 2016), 600–620.

[22] Yamada, H. The Frisch–Waugh–Lovell theorem for the lasso and the ridge regression. *Communications in Statistics - Theory and Methods 46*, 21 (Nov. 2017), 10897–10902.