# "A Significance Test for the Lasso"

Lockhart R, Taylor J, Tibshirani R, and Tibshirani R

Ashley Petersen

May 14, 2013

# Last time

- Problem: Many clinical covariates – which are important to a certain medical outcome?
- Want to choose the important variables and say how important these variables are
- Bad solution: Forward stepwise regression $\rightarrow$ very anti-conservative p-values
- Better solution: Lasso with p-values from newly proposed covariance test statistic

# Framework

Consider regression setup with outcome vector $y \in \mathbb{R}^n$ with covariate matrix $X \in \mathbb{R}^{n \times p}$ and

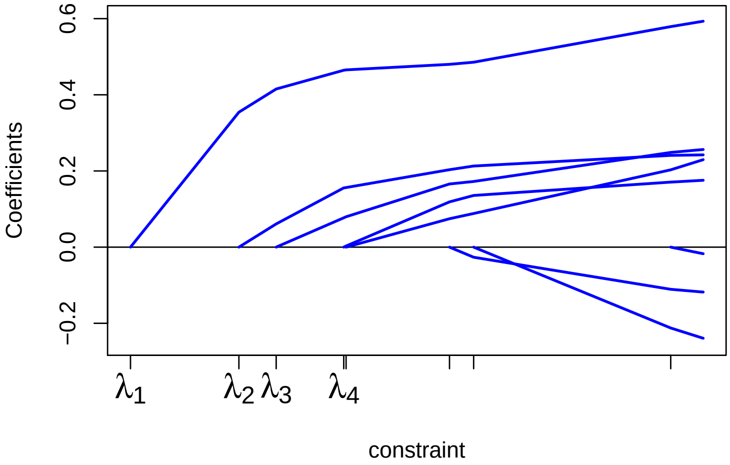
$$y = \boldsymbol{\beta} X + \boldsymbol{\epsilon} \text{ with } \boldsymbol{\epsilon} \sim N(0, \sigma^2 I).$$

The lasso estimator is obtained by finding $\boldsymbol{\beta}$ that minimizes

$$\frac{1}{2}\|y - X\boldsymbol{\beta}\|^2 + \lambda \sum_{i=1}^{p} |\beta_i|,$$

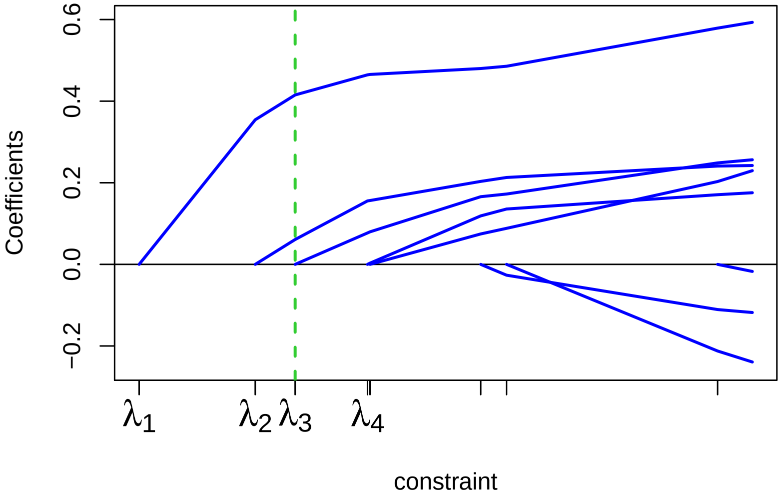
where $\lambda$ is the lasso penalty.

# Lasso solution path ($\lambda_1 > \lambda_2 > \lambda_3 > \lambda_4 > \ldots$)




$$\hat{\boldsymbol{\beta}}_{lasso} = \arg\min_{\boldsymbol{\beta}} \frac{1}{2}\|y - X\boldsymbol{\beta}\|^2 + \lambda \sum_{i=1}^{p} |\beta_i|$$

# Obtain p-value for covariate entering the model



# Form of test statistic [1]

Forward stepwise regression:

$$\begin{aligned}\frac{RSS_{null} - RSS}{\sigma^2} &= \frac{\|y - \hat{y}_{null}\|^2 - \|y - \hat{y}\|^2}{\sigma^2} \\ &= 2\left[\frac{y^T\hat{y} - y^T\hat{y}_{null}}{\sigma^2}\right] + \frac{\|\hat{y}_{null}\|^2 - \|\hat{y}\|^2}{\sigma^2}\end{aligned}$$

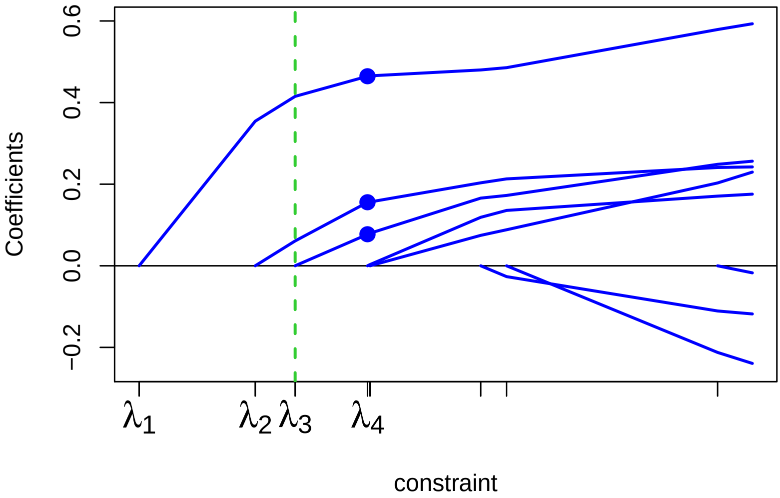
Lasso:

$$T_k = \frac{y^T\hat{y} - y^T\hat{y}_{null}}{\sigma^2}$$

[1] Taking $\sigma^2$ as known (for now)

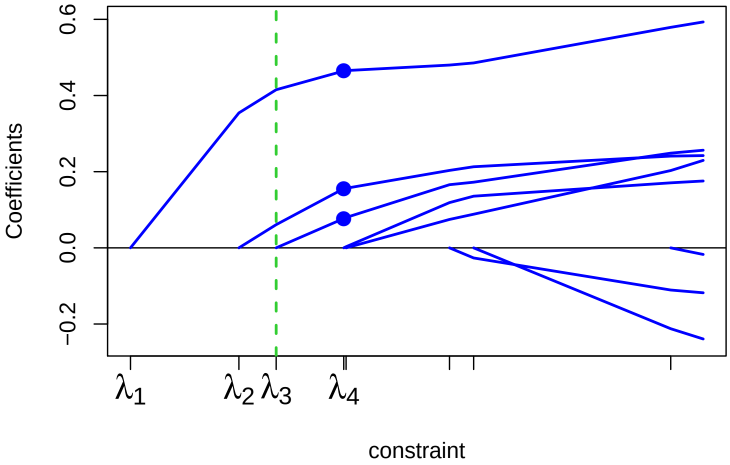
# What is $\hat{y}$?

- testing that variable that enters at $\lambda_3$ has $\beta = 0$
- $\hat{y} = X\hat{\beta}(\lambda_4)$



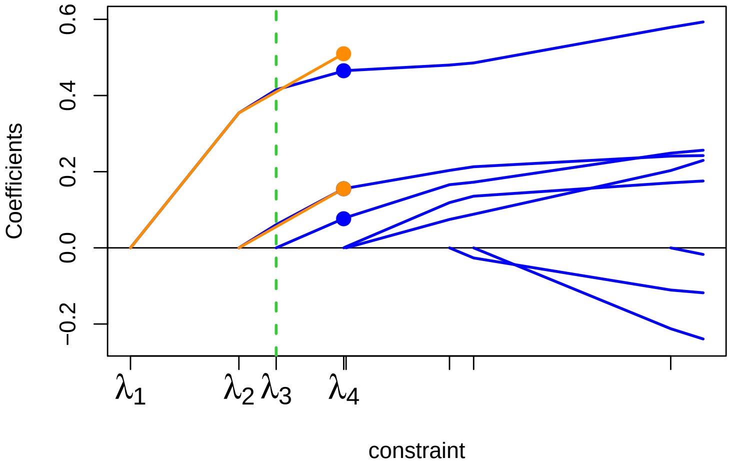
# What about $\hat{y}_{null}$?

- testing that variable that enters at $\lambda_3$ has $\beta = 0$
- $\hat{y} = X\hat{\beta}(\lambda_4)$
- $\hat{y}_{null} = X_{null}\hat{\beta}_{null}(\lambda_4)$



# What about $\hat{y}_{null}$?

- testing that variable that enters at $\lambda_3$ has $\beta = 0$
- $\hat{y} = X\hat{\beta}(\lambda_4)$
- $\hat{y}_{null} = X_{null}\hat{\beta}_{null}(\lambda_4)$



# Putting this together

The covariance test statistic for testing the predictor that enters at the $k$th step is

$$
\begin{aligned}
T_k &= \frac{y^T \hat{y} - y^T \hat{y}_{null}}{\sigma^2} \\
&= \frac{y^T X \hat{\beta}(\lambda_{k+1}) - y^T X_{null} \hat{\beta}_{null}(\lambda_{k+1})}{\sigma^2}.
\end{aligned}
$$

# What exactly is the null?

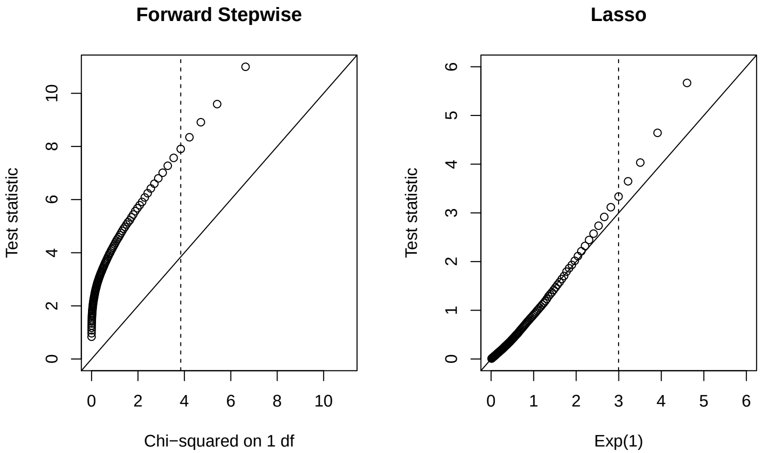
Under the global null ($\beta = 0$), then

$$\begin{aligned} T_1 &\to_d Exp(1) \\ T_2 &\to_d Exp(1/2) \\ T_3 &\to_d Exp(1/3) \\ &\vdots \end{aligned}$$

for orthogonal predictor matrix $X$. Asymptotic distributions are stochastically smaller for general $X$.

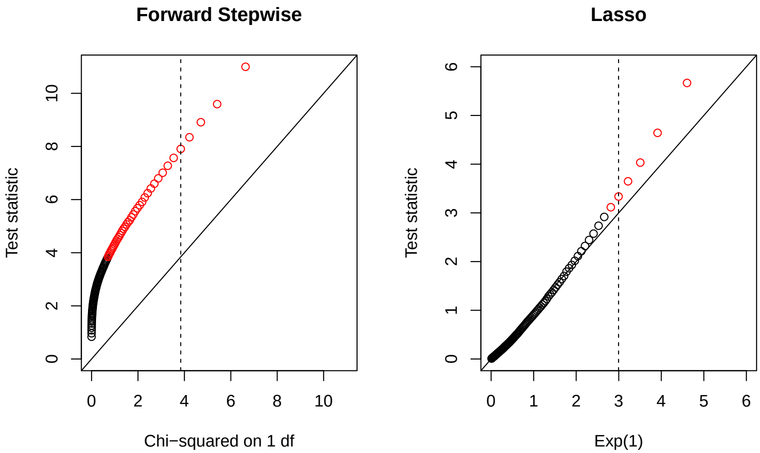
# Does it work for finite samples?

- Simulation of distribution of test statistics for first covariate to enter model under global null ($\beta = 0$)
- $n = 100$, $p = 10$



# Does it work for finite samples?

- Simulation of distribution of test statistics for first covariate to enter model under global null ($\beta = 0$)
- $n = 100$, $p = 10$



# What exactly is the null?

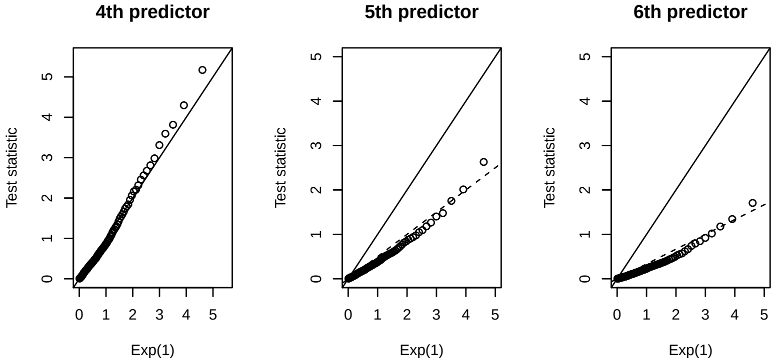
Under the weaker null where there are $k_0$ truly active covariates (and they have entered the model), then

$$\begin{aligned} T_{k_0+1} &\to_d Exp(1) \\ T_{k_0+2} &\to_d Exp(1/2) \\ T_{k_0+3} &\to_d Exp(1/3) \\ &\vdots \end{aligned}$$

for orthogonal predictor matrix $X$. Asymptotic distributions are stochastically smaller for general $X$.

# See, it works...

- Simulation of distribution of test statistics when true $\beta$ has three non-zero components
- $n = 100$, $p = 10$
- $F^{-1}(p) = -\theta \log(1 - p)$ for $Exp(\theta)$



# Simulation setup

- Distribution of $T_1$ under global null ($\boldsymbol{\beta} = 0$)
- $n = 100$ and $p \in (10, 50, 200)$
- Varying correlation structure of predictors with $\rho \in (0, 0.2, 0.4, 0.6, 0.8)$
  - Exchangeable
  - AR(1)
  - Block diagonal
- Mean, variance, and tail probability of distribution

# The authors' results

| $n = 100,\ p = 10$ | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| $\rho$ | Equal pop'n corr | | | $AR(1)$ | | | Block diagonal | | |
| | Mean | Var | Tail pr | Mean | Var | Tail pr | Mean | Var | Tail pr |
| 0 | 1.120 | 1.951 | 0.090 | 1.017 | 1.484 | 0.070 | 1.058 | 1.548 | 0.060 |
| 0.2 | 1.119 | 1.844 | 0.086 | 1.034 | 1.497 | 0.074 | 1.069 | 1.614 | 0.078 |
| 0.4 | 1.115 | 1.724 | 0.092 | 1.045 | 1.469 | 0.060 | 1.077 | 1.701 | 0.076 |
| 0.6 | 1.095 | 1.648 | 0.086 | 1.048 | 1.485 | 0.066 | 1.074 | 1.719 | 0.086 |
| 0.8 | 1.062 | 1.624 | 0.092 | 1.034 | 1.471 | 0.062 | 1.062 | 1.687 | 0.072 |
| se | 0.010 | 0.049 | 0.001 | 0.013 | 0.043 | 0.001 | 0.010 | 0.047 | 0.001 |
| $n = 100,\ p = 50$ | | | | | | | | | |
| 0 | 1.078 | 1.721 | 0.074 | 1.039 | 1.415 | 0.070 | 0.999 | 1.578 | 0.048 |
| 0.2 | 1.090 | 1.476 | 0.074 | 0.998 | 1.391 | 0.054 | 1.064 | 2.062 | 0.052 |
| 0.4 | 1.079 | 1.382 | 0.068 | 0.985 | 1.373 | 0.060 | 1.076 | 2.168 | 0.062 |
| 0.6 | 1.057 | 1.312 | 0.060 | 0.978 | 1.425 | 0.054 | 1.060 | 2.138 | 0.060 |
| 0.8 | 1.035 | 1.346 | 0.056 | 0.973 | 1.439 | 0.060 | 1.046 | 2.066 | 0.068 |
| se | 0.011 | 0.037 | 0.001 | 0.009 | 0.041 | 0.001 | 0.011 | 0.103 | 0.001 |
| $n = 100,\ p = 200$ | | | | | | | | | |
| 0 | 1.004 | 1.017 | 0.054 | 1.029 | 1.240 | 0.062 | 0.930 | 1.166 | 0.042 |
| 0.2 | 0.996 | 1.164 | 0.052 | 1.000 | 1.182 | 0.062 | 0.927 | 1.185 | 0.046 |
| 0.4 | 1.003 | 1.262 | 0.058 | 0.984 | 1.016 | 0.058 | 0.935 | 1.193 | 0.048 |
| 0.6 | 1.007 | 1.327 | 0.062 | 0.954 | 1.000 | 0.050 | 0.915 | 1.231 | 0.044 |
| 0.8 | 0.989 | 1.264 | 0.066 | 0.961 | 1.135 | 0.060 | 0.914 | 1.258 | 0.056 |
| se | 0.008 | 0.039 | 0.001 | 0.009 | 0.028 | 0.001 | 0.007 | 0.032 | 0.001 |

# Some commentary...

"I don't have any applied or technical comments on the paper at hand (except for feeling strongly that Tables 2 and 3 should really really really be made into a graph . . . do we really care that a certain number is 315.216?)"

–Andrew Gelman[2]

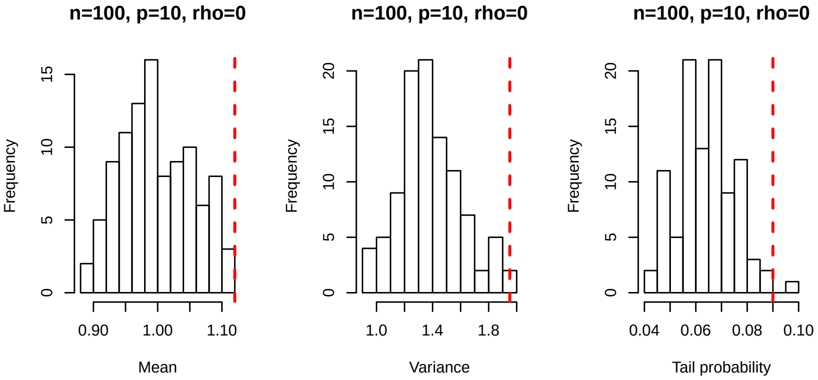
[2]via his blog

# The authors' results

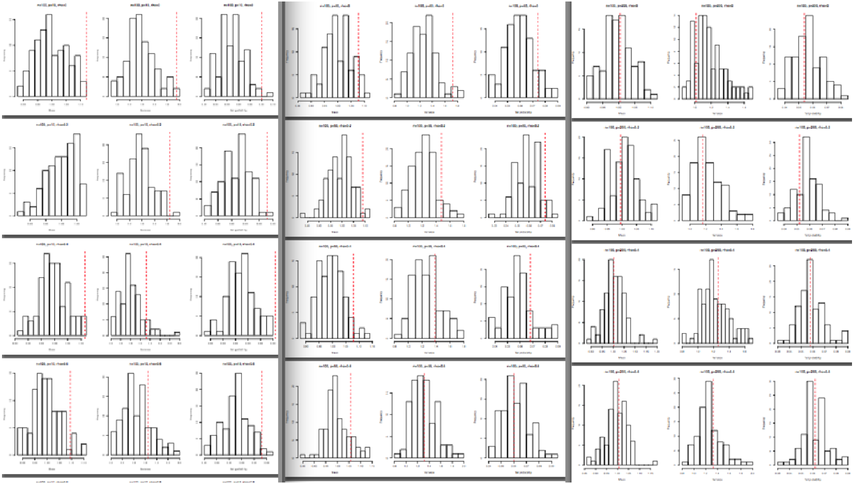
| $\rho$ | Equal pop'n corr | | | $AR(1)$ | | | Block diagonal | | |
|---|---|---|---|---|---|---|---|---|---|
| | Mean | Var | Tail pr | Mean | Var | Tail pr | Mean | Var | Tail pr |
| | $n = 100,\ p = 10$ | | | | | | | | |
| 0 | 1.120 | 1.951 | 0.090 | 1.017 | 1.484 | 0.070 | 1.058 | 1.548 | 0.060 |
| 0.2 | 1.119 | 1.844 | 0.086 | 1.034 | 1.497 | 0.074 | 1.069 | 1.614 | 0.078 |
| 0.4 | 1.115 | 1.724 | 0.092 | 1.045 | 1.469 | 0.060 | 1.077 | 1.701 | 0.076 |
| 0.6 | 1.095 | 1.648 | 0.086 | 1.048 | 1.485 | 0.066 | 1.074 | 1.719 | 0.086 |
| 0.8 | 1.062 | 1.624 | 0.092 | 1.034 | 1.471 | 0.062 | 1.062 | 1.687 | 0.072 |
| se | 0.010 | 0.049 | 0.001 | 0.013 | 0.043 | 0.001 | 0.010 | 0.047 | 0.001 |
| | $n = 100,\ p = 50$ | | | | | | | | |
| 0 | 1.078 | 1.721 | 0.074 | 1.039 | 1.415 | 0.070 | 0.999 | 1.578 | 0.048 |
| 0.2 | 1.090 | 1.476 | 0.074 | 0.998 | 1.391 | 0.054 | 1.064 | 2.062 | 0.052 |
| 0.4 | 1.079 | 1.382 | 0.068 | 0.985 | 1.373 | 0.060 | 1.076 | 2.168 | 0.062 |
| 0.6 | 1.057 | 1.312 | 0.060 | 0.978 | 1.425 | 0.054 | 1.060 | 2.138 | 0.060 |
| 0.8 | 1.035 | 1.346 | 0.056 | 0.973 | 1.439 | 0.060 | 1.046 | 2.066 | 0.068 |
| se | 0.011 | 0.037 | 0.001 | 0.009 | 0.041 | 0.001 | 0.011 | 0.103 | 0.001 |
| | $n = 100,\ p = 200$ | | | | | | | | |
| 0 | 1.004 | 1.017 | 0.054 | 1.029 | 1.240 | 0.062 | 0.930 | 1.166 | 0.042 |
| 0.2 | 0.996 | 1.164 | 0.052 | 1.000 | 1.182 | 0.062 | 0.927 | 1.185 | 0.046 |
| 0.4 | 1.003 | 1.262 | 0.058 | 0.984 | 1.016 | 0.058 | 0.935 | 1.193 | 0.048 |
| 0.6 | 1.007 | 1.327 | 0.062 | 0.954 | 1.000 | 0.050 | 0.915 | 1.231 | 0.044 |
| 0.8 | 0.989 | 1.264 | 0.066 | 0.961 | 1.135 | 0.060 | 0.914 | 1.258 | 0.056 |
| se | 0.008 | 0.039 | 0.001 | 0.009 | 0.028 | 0.001 | 0.007 | 0.032 | 0.001 |

# ‘Sampling distribution’ of simulation results

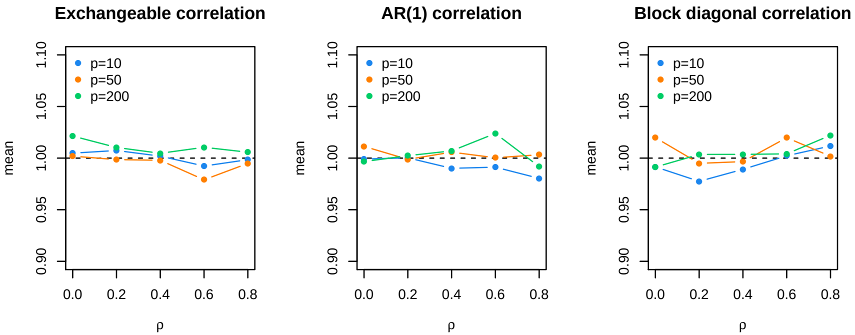
- 100 replications of the simulation for given parameters
- Note large variance of each distribution
- Larger number of replications needed for accurate estimate



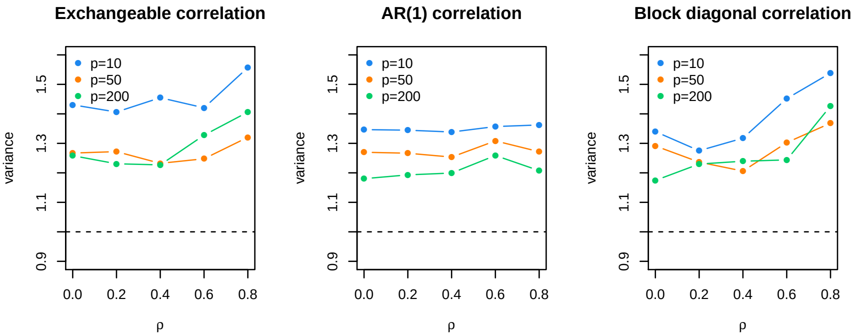
# Lots of sampling distributions



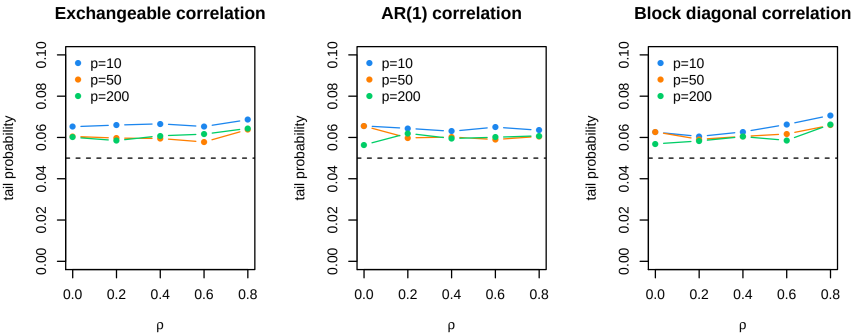
# My results – mean



# My results – variance



# My results – tail probability



# What to do when $\sigma^2$ is unknown? $(n > p)$

- Estimate in the usual way:

$$\hat{\sigma}^2 = \frac{RSS}{n-p} = \frac{\|y - X\hat{\beta}_{LS}\|^2}{n-p}$$

- Asymptotic distribution is now $F_{2,n-p}$
  - Numerator is $Exp(1) = \chi^2_2/2$
  - Denominator is $\chi^2_{n-p}/(n-p)$
  - Numerator and denominator independent

# What to do when $\sigma^2$ is unknown? ($n \leq p$)

- Estimate from least squares fit from model selected by cross-validation
- No rigorous theory here (fingers crossed!)

# What's the big idea?

- Use covariance test statistic to obtain p-value for covariates as they enter the lasso model
- Compare to asymptotic distribution – $Exp(1)$ – to obtain p-values
- Reasonable performance in finite samples
- Possibly extend this to obtaining inference for all coefficients from a model for a specific lasso penalty

# What's next!

To do:

- Obtain data for $p > n$ case (HIV data)
- Finish simulations

Next time:

- 'Real' data examples
- More on assumptions and theory