Degrees of Freedom for Multiple Regression with Latent Variables

Patrick Perry, NYU
(joint with, Natesh Pillai)

Finding Age-Related Genes

AGEMAP Dataset:

- Activation level measured for 18,000 genes in 39 mice
- Mouse-specific covariates: Age (Months), Sex (Female/Male)

Goal Find **all genes** related to age.

Easier Problem Test if **a particular gene** is related to age.

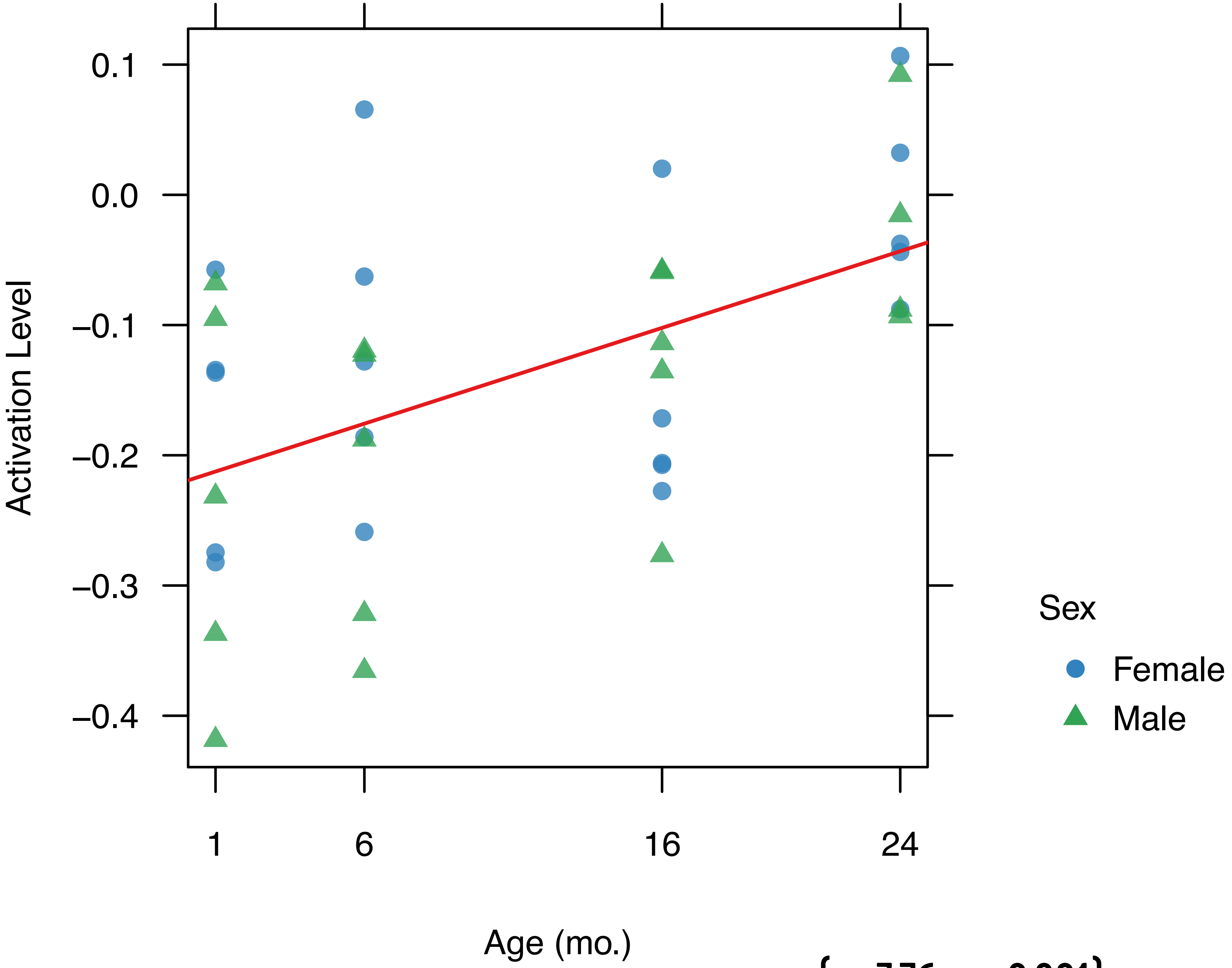
Regression: estimate associations between covariates and response

Factor analysis: estimate latent covariates

Degrees of freedom: adjust the regression estimates of significance

Regression

Main Tool: Linear Regression



(t = 3.76; p < 0.001)

Regression Model

$$\begin{array}{cccccc}
Y & = & X & \beta & + & \varepsilon \\
\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} & = & \begin{bmatrix} 1 & \text{age}_1 \\ 1 & \text{age}_2 \\ \vdots & \\ 1 & \text{age}_n \end{bmatrix} & \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} & + & \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix} \\
 & & & & & \Big\uparrow \\
 & & & & & \text{Independent Normal } (0, \sigma^2)
\end{array}$$

Goal: Test whether $\beta_1 \neq 0$

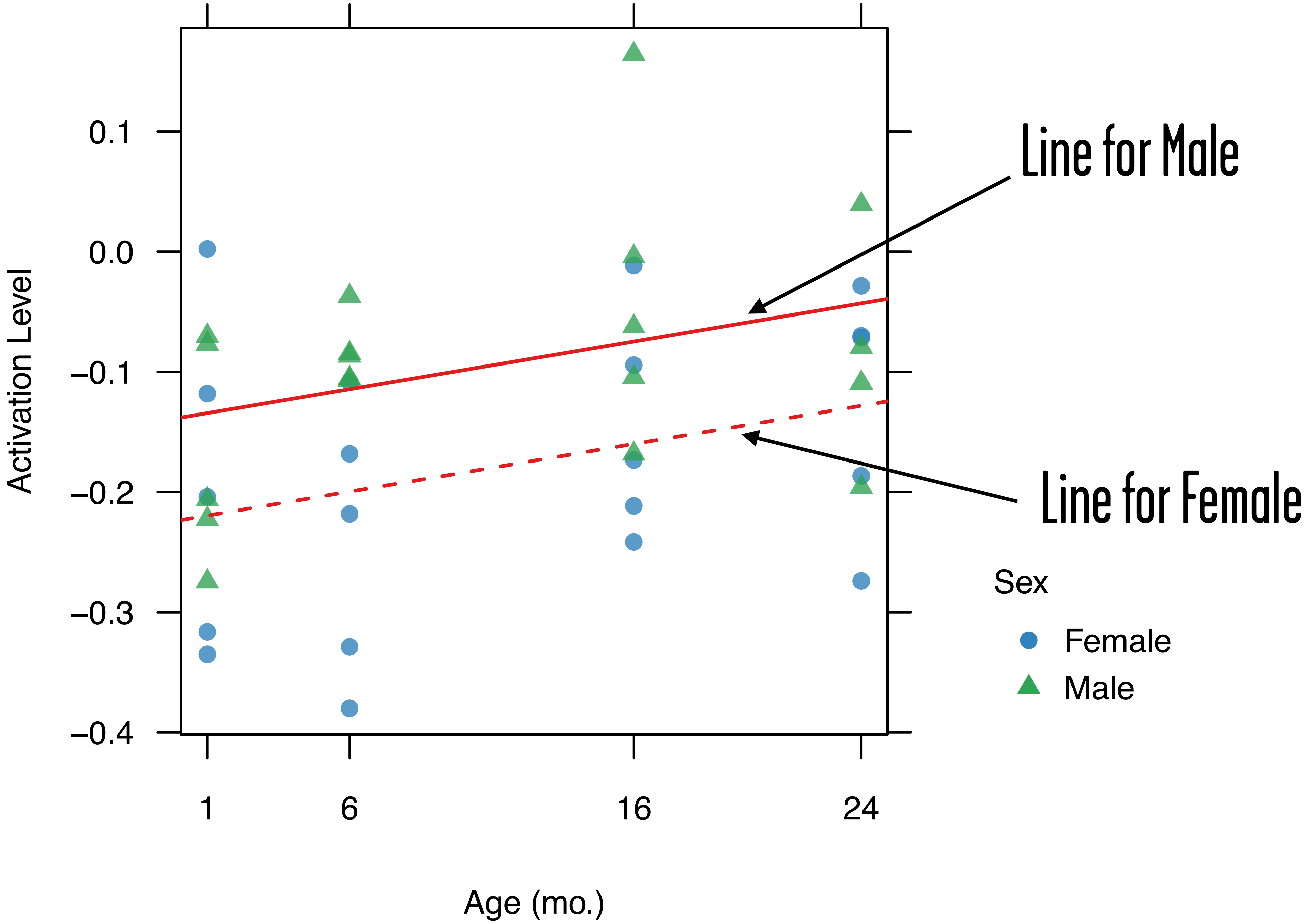
T Test

1. Compute least squares fit $\hat{\beta} = (X^T X)^{-1} X^T y, \quad \hat{y} = X\hat{\beta}$

2. Compute noise variance estimate $\hat{\sigma}^2 = \frac{1}{n-2} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$

3. Reject null hypothesis if t-statistic is large $t_1 = \frac{\hat{\beta}_1}{\{\hat{\sigma}^2 (X^T X)^{-1}_{11}\}^{1/2}}$

Adjusting for other Covariates



Line for Male

Line for Female

(Without adjustment: t = 1.90, p = 0.07. With adjustment: t = 2.14, p = 0.04)

Degrees of Freedom

$$y = \underset{\uparrow}{X}\beta + \varepsilon$$

p columns

$$\hat{\sigma}^2 = \frac{1}{n-p}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

“n - p degrees of freedom in the residual errors”

Recap

1. Estimate coefficient of age after adjusting for other covariates (sex)

2. Estimate noise variance, using n - p degrees of freedom in the residual errors

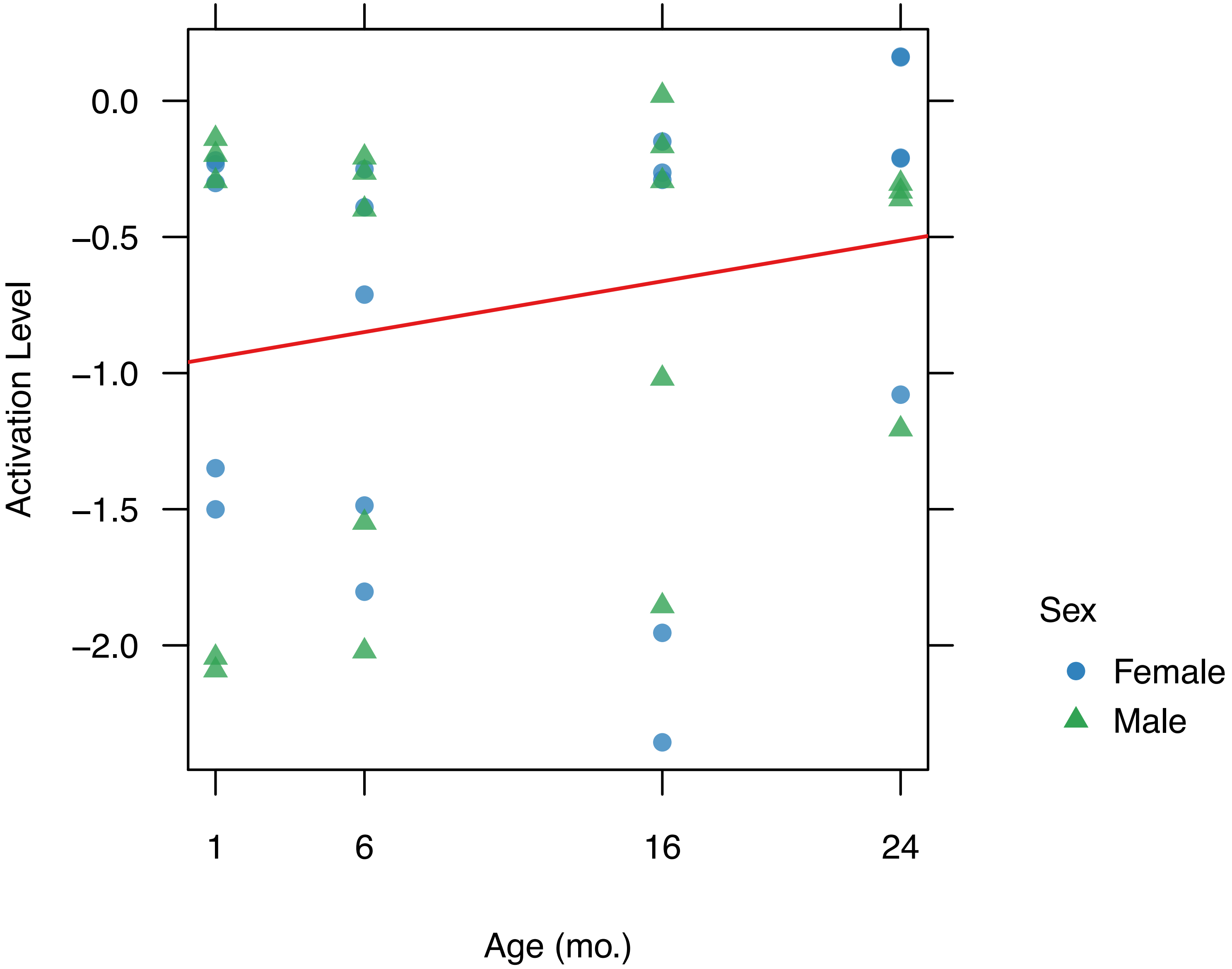
$$\hat{\sigma}^2 = \frac{1}{n-p} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

3. Declare the gene "age related" if t-statistic is large

$$t_1 = \frac{\hat{\beta}_1}{\{\hat{\sigma}^2 (X^T X)^{-1}_{11}\}^{1/2}}$$

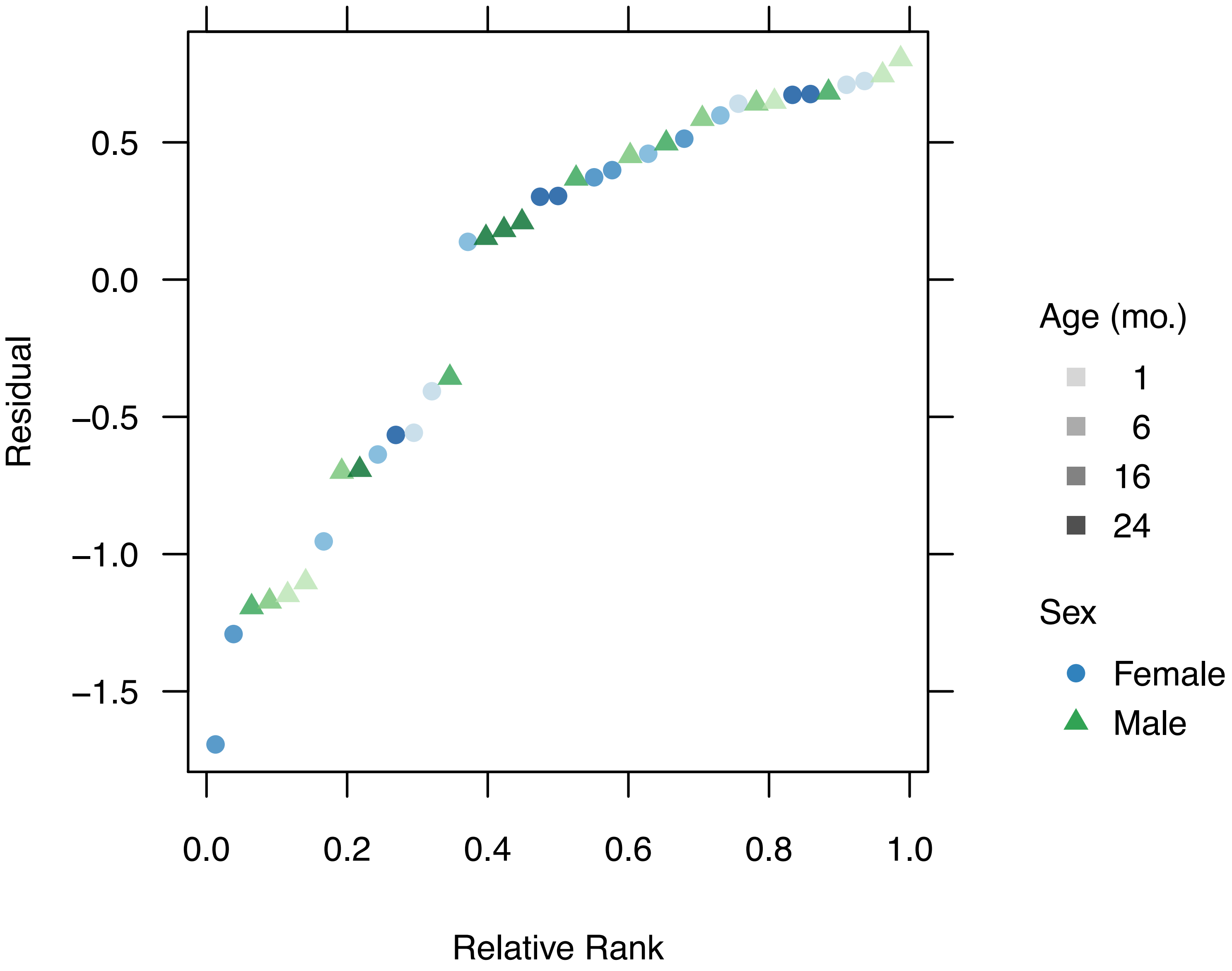
Factor Analysis

Is Cerebellum Mm.71015 Related to Age?

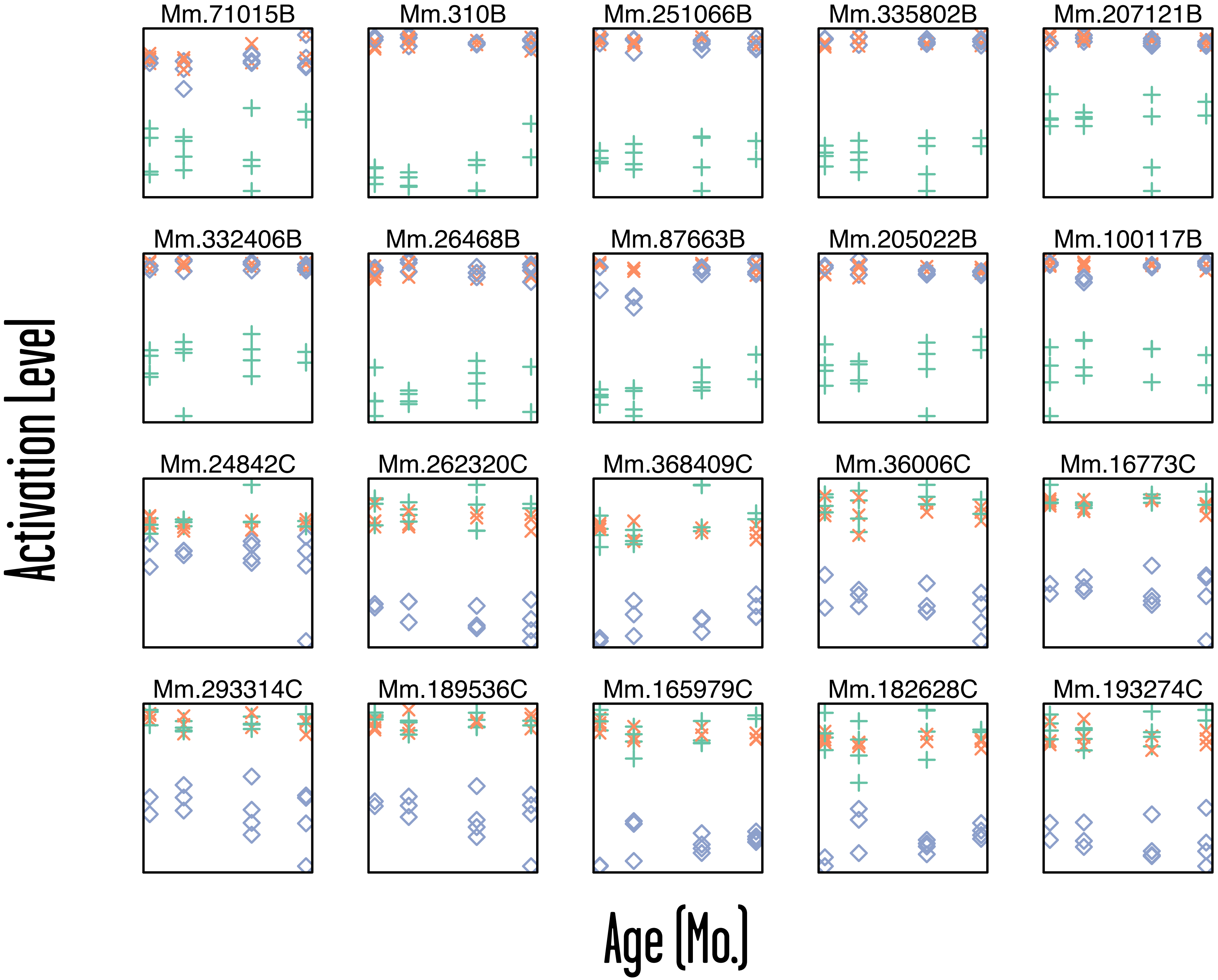


(Test statistic for Age coefficient: t = 1.35, df=36; p = 0.18)

Residuals Exhibit Dependence



Three Mouse Types?



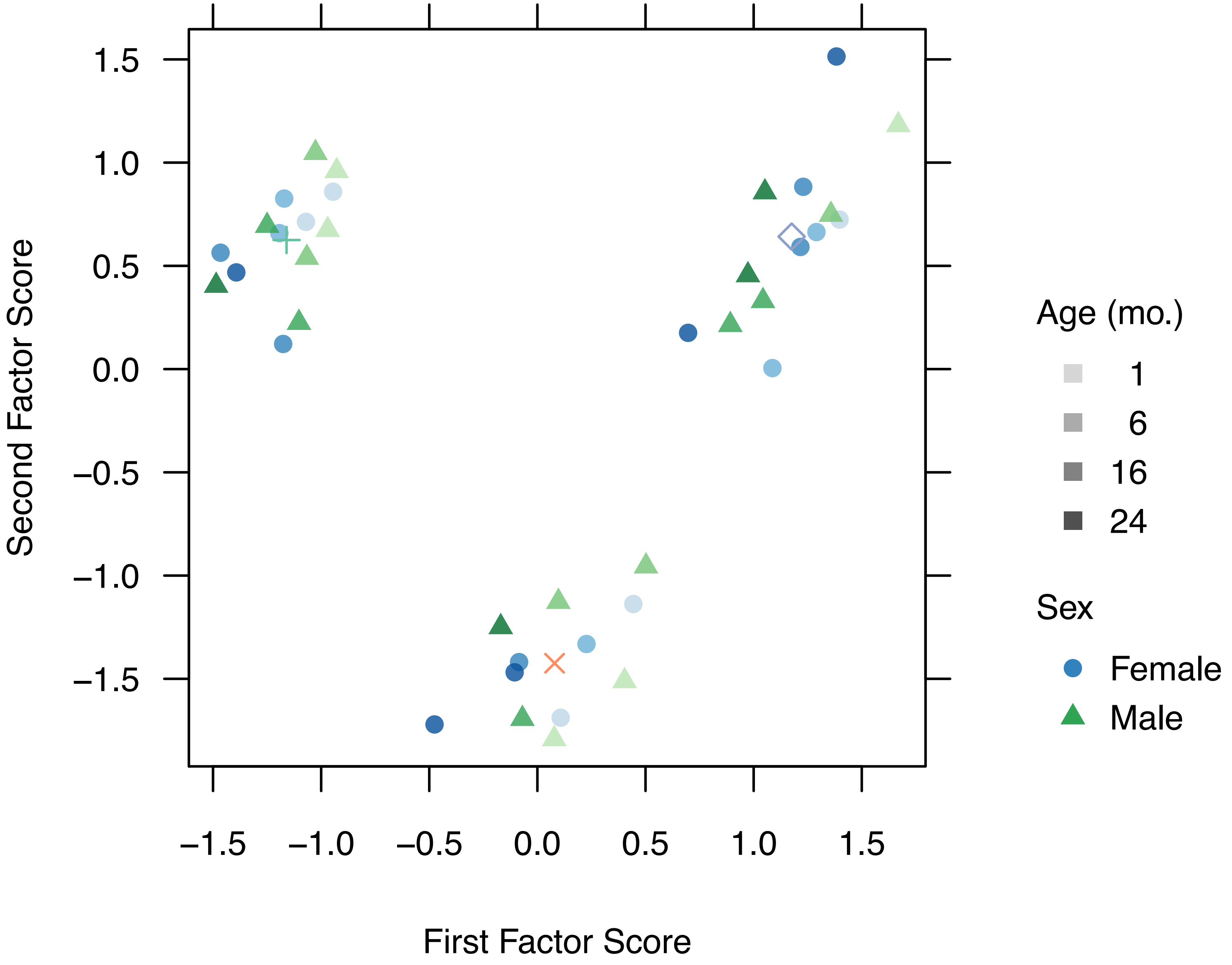
Factor Analysis on Residuals

Model: $$\underset{\underset{n \times m}{\nearrow}}{Y} = XB^T + \underset{\underset{\text{low rank}}{\uparrow}}{UV^T} + \underset{\underset{\text{Gaussian noise}}{\nwarrow}}{E}$$

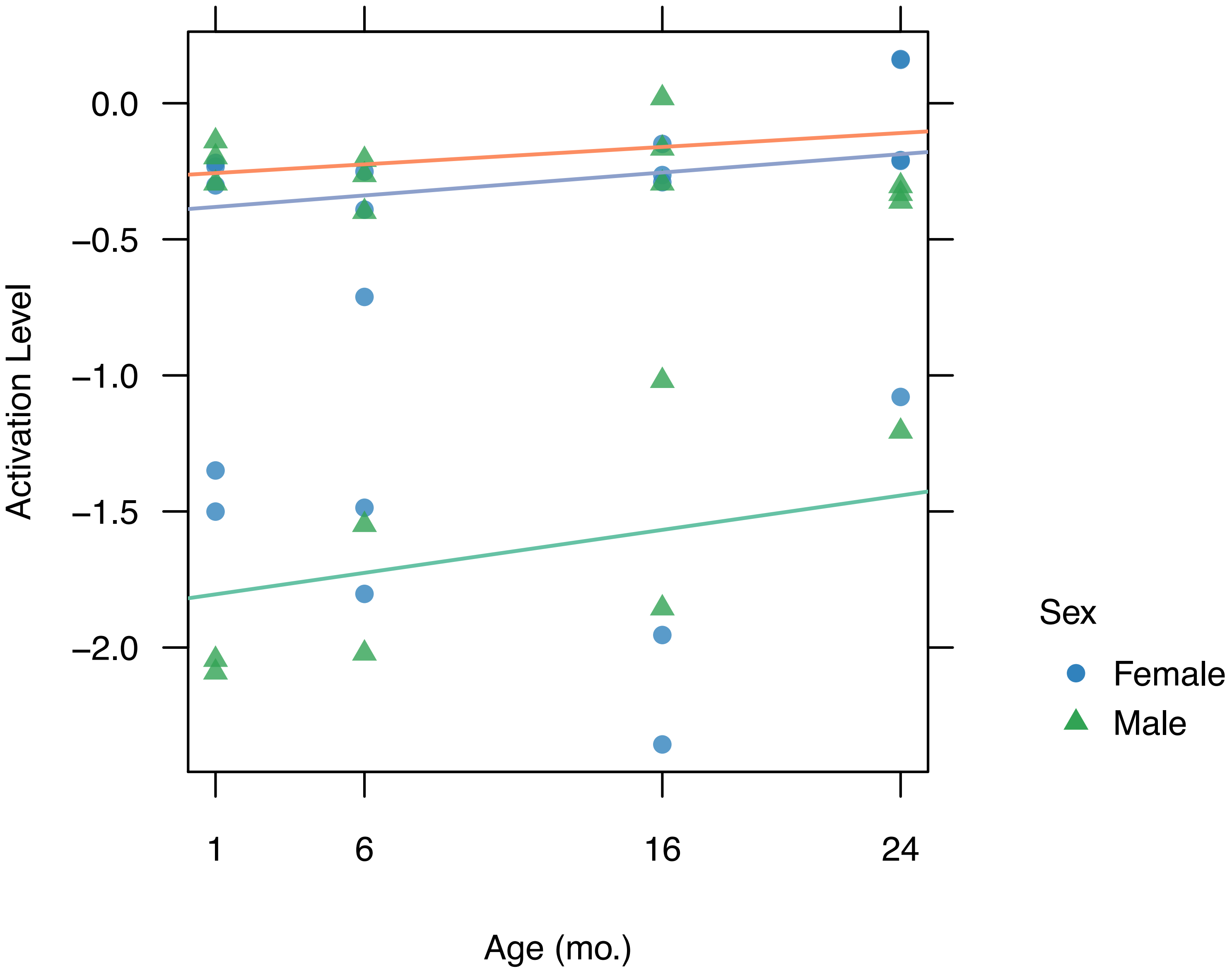
Estimate latent factors via singular value decomposition of $Y - \hat{Y}$

Evidence of Mouse Types in AGEMAP Data

- Measure log activation for 18,000 genes in 40 mice
- Regress on mouse age and sex, gene tissue type
- Plot first two estimated factor scores from residuals



How Many Degrees of Freedom?



(Test statistic for Age coefficient: $t = 3.86$, **df=36-2?**; $p < 0.001$)

Adjusting for Latent Factors

Even if we are not interested in the latent sources of variability, ignoring it affects are conclusions about the measured sources of variability!

Problem: how can we adjust for latent sources of variability?

Degrees of Freedom

Defining “Degrees of Freedom” for Estimated Factor Terms

Model

$$Y = UV^T + E$$

Independent rows, Normal $(0, \Sigma)$

Residual Matrix

$$\hat{E} = Y - \hat{U}\hat{V}^T$$

From SVD of Y

Degrees of Freedom

$$\mathbb{E}[s^T \hat{E}^T \hat{E} s] = \{n - \mathrm{df}(s)\} \cdot s^T \Sigma s$$

$$\hat{\sigma}^2(s) = \frac{s^T \hat{E}^T \hat{E} s}{n - \mathrm{df}(s)}$$

Degrees of Freedom Estimates

Parameter Counting:	$\{(n-1)+(m-1)+1\}/m$	(Gollob, 1968)
Parametric Bootstrap:	Monte Carlo estimate	(Mandel, 1971)
Maximum Likelihood:	1 or 0	

Theory for Noise Case

Theorem: Under IID noise, if there are no true latent factors, then the appropriate degrees of freedom adjustment for each estimated latent term is approximately

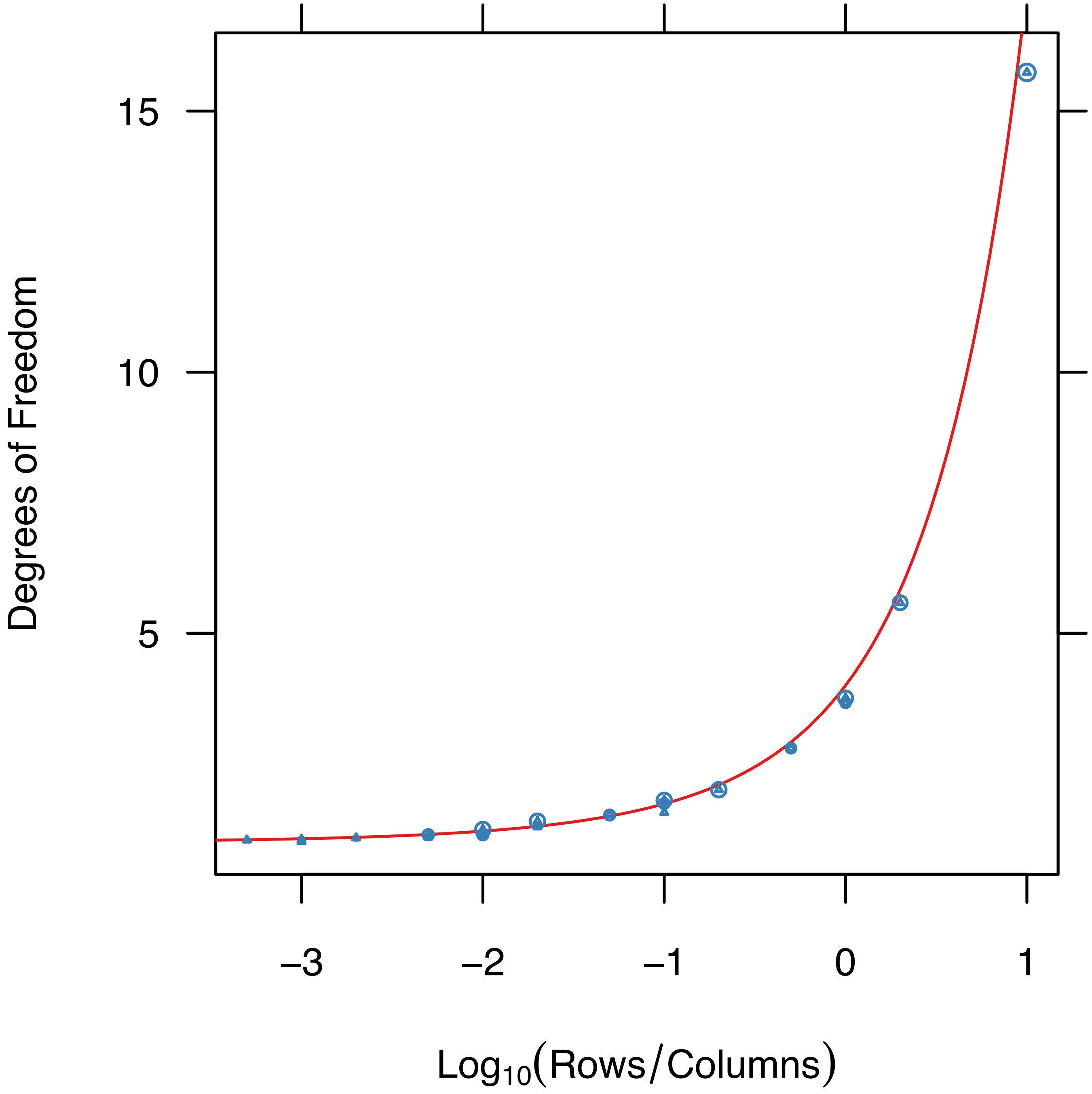
$$1 + n/m + 2\sqrt{n/m}$$

Proof: follows from Yin, Bai, Krishnaiah (1988) result on maximum Wishart eigenvalue

Noise Case Simulation

1. Generate from model $Y = E$
2. Fit $\hat{Y} = \hat{U}\hat{V}^T$
3. Compute residuals $\hat{E} = Y - \hat{Y}$
4. Compute residual sum of squares $\mathrm{RSS}(s) = s^T \hat{E}^T \hat{E} s$
5. Take empirical df $\widehat{\mathrm{df}}(s) = n - \mathrm{RSS}(s)/\sigma^2$
6. Compare to theoretical df $\mathrm{df}(s) = 1 + n/m + 2\sqrt{n/m}$

Noise Case Simulations



Theory for Signal Case

Theorem: Under IID noise, if the kth signal is detectable and uncorrelated with the test direction, then the appropriate degrees of freedom adjustment for the estimated latent term is approximately

$$1 + n/m + \sigma^2/\mu_k$$

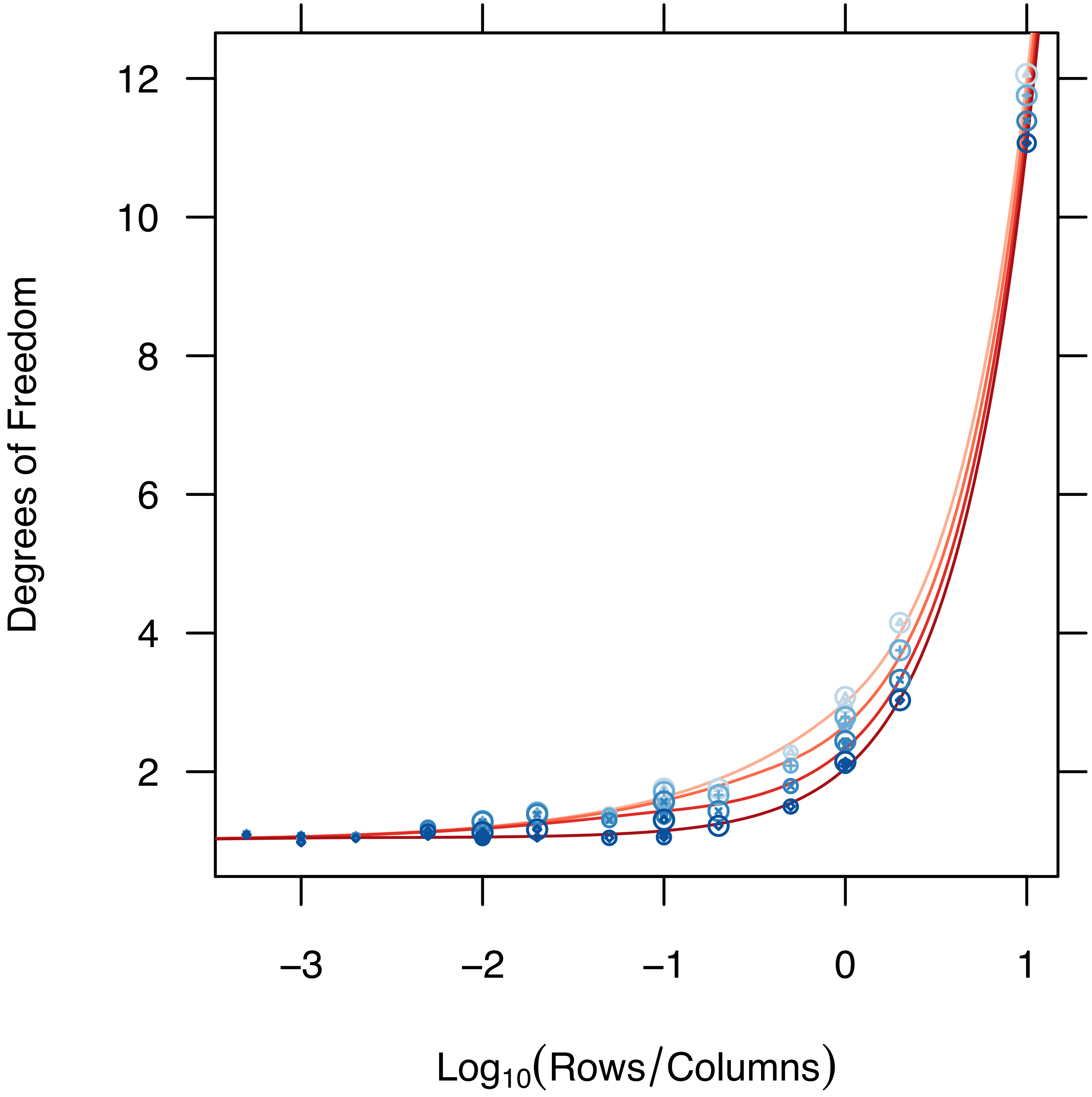
Proof: follows from Onatski (2007), Benaych-Georges and Rao Nadakuditi (2012) SVD asymptotics.

Signal Case Simulations

1. Generate from model $Y = E + \sqrt{n\mu} \cdot UV^T$
2. Fit $\hat{Y} = \hat{U}\hat{V}^T$
3. Compute residuals $\hat{E} = Y - \hat{Y}$
4. Compute residual sum of squares $\mathrm{RSS}(s) = s^T \hat{E}^T \hat{E} s$
5. Take empirical df $\widehat{\mathrm{df}}(s) = n - \mathrm{RSS}(s)/\sigma^2$
6. Compare to theoretical df $\mathrm{df}(s) = 1 + n/m + \sigma^2/\mu$

Simulations for Signal Case

(Random test vector)



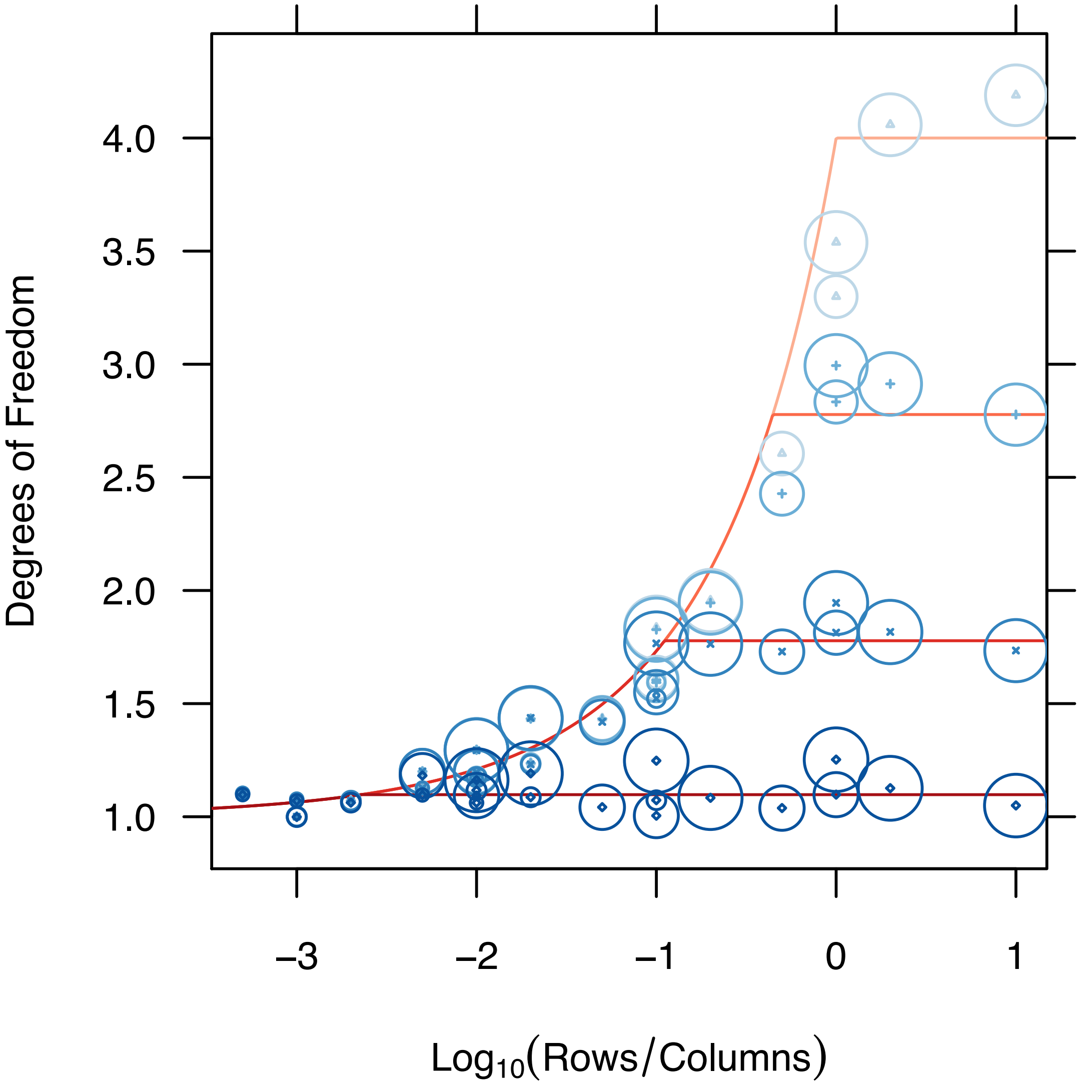
General Case

Theorem: Under IID noise, the appropriate degrees of freedom adjustment for the estimated latent term is approximately

$$\begin{cases} n\left(1 - \frac{m\sigma^2}{n\mu_k} - \frac{m\sigma^4}{n\mu_k^2}\right)\frac{(v_k^T s)^2}{s^T s} + \left(1 + \frac{\sigma^2}{\mu_k}\right)^2\left(1 - \frac{(v_k^T s)^2}{s^T s}\right) & \text{if } \mu_k > \sigma^2\sqrt{m/n}, \\ (1 + \sqrt{n/m})^2 - n\frac{\mu_k}{\sigma^2}\frac{(v_k^T s)^2}{s^T s} & \text{otherwise.} \end{cases}$$

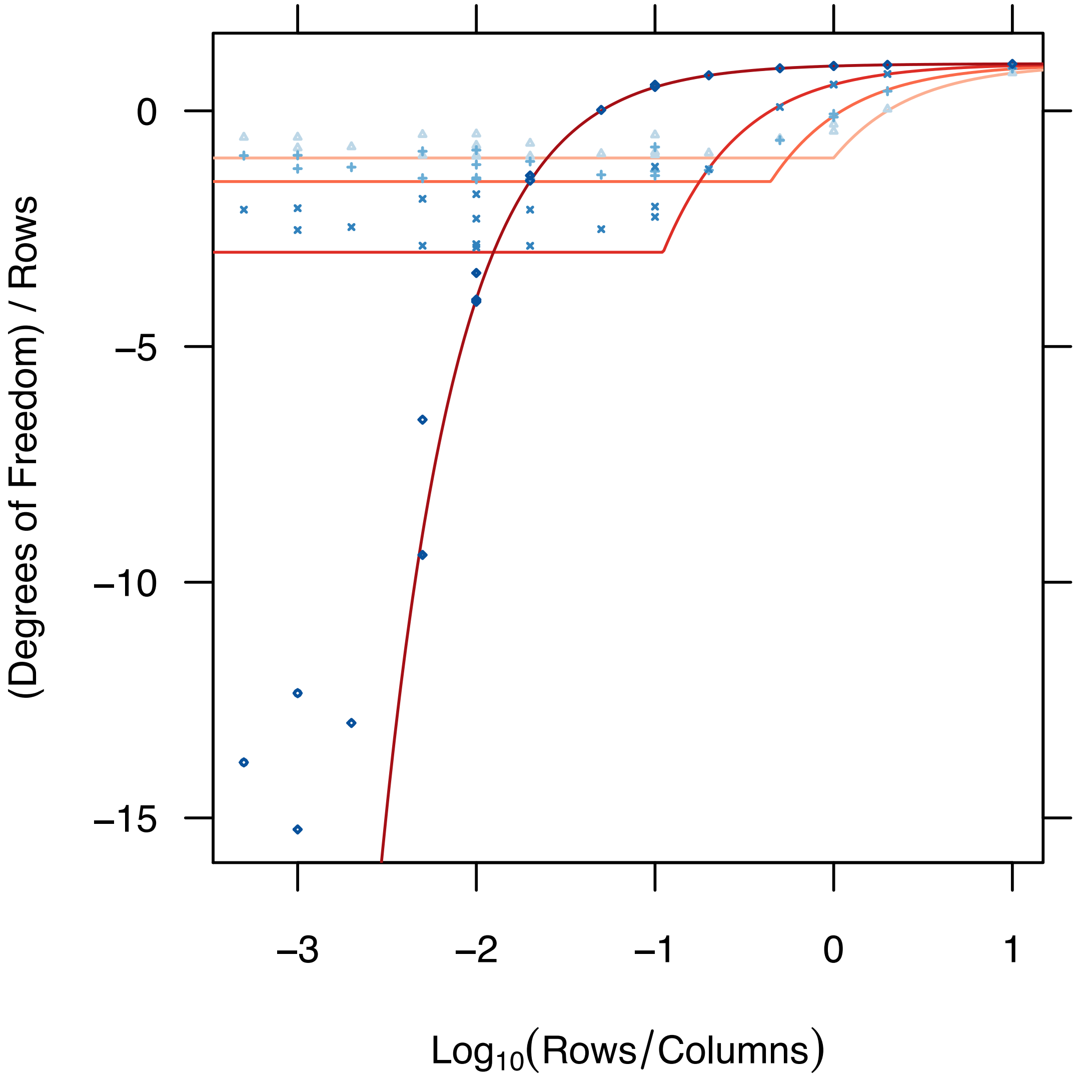
Orthogonal Test Direction

$$\begin{cases} 1 + (\sigma^2/\mu)^2 & \text{if } \mu > \sigma^2\sqrt{m/n}, \\ (1 + \sqrt{n/m})^2 & \text{otherwise.} \end{cases}$$



Parallel Test Direction

$$\begin{cases} 1 - m\sigma^2/(n\mu) - m\sigma^4/(n\mu^2) & \text{if } \mu > \sigma^2\sqrt{m/n}, \\ -(\mu/\sigma^2) & \text{otherwise.} \end{cases}$$



Application to AGEMAP Data

Caveat

Disconnect between theory and application: independent noise $(\Sigma = I)$ versus correlated gene responses (general Σ)

Conjecture: results hold under mild correlation (some theoretical support for this, but still incomplete)

Estimating Degrees of Freedom

True Degrees of Freedom:

$$1 + n/m + 2\sqrt{n/m}$$ (noise)

$$1 + n/m + \sigma^2/\mu_k$$ (signal)

Conservative Estimate for Signal (uncorrelated test direction):

$$1 + n/m + 2\sqrt{n/m}$$

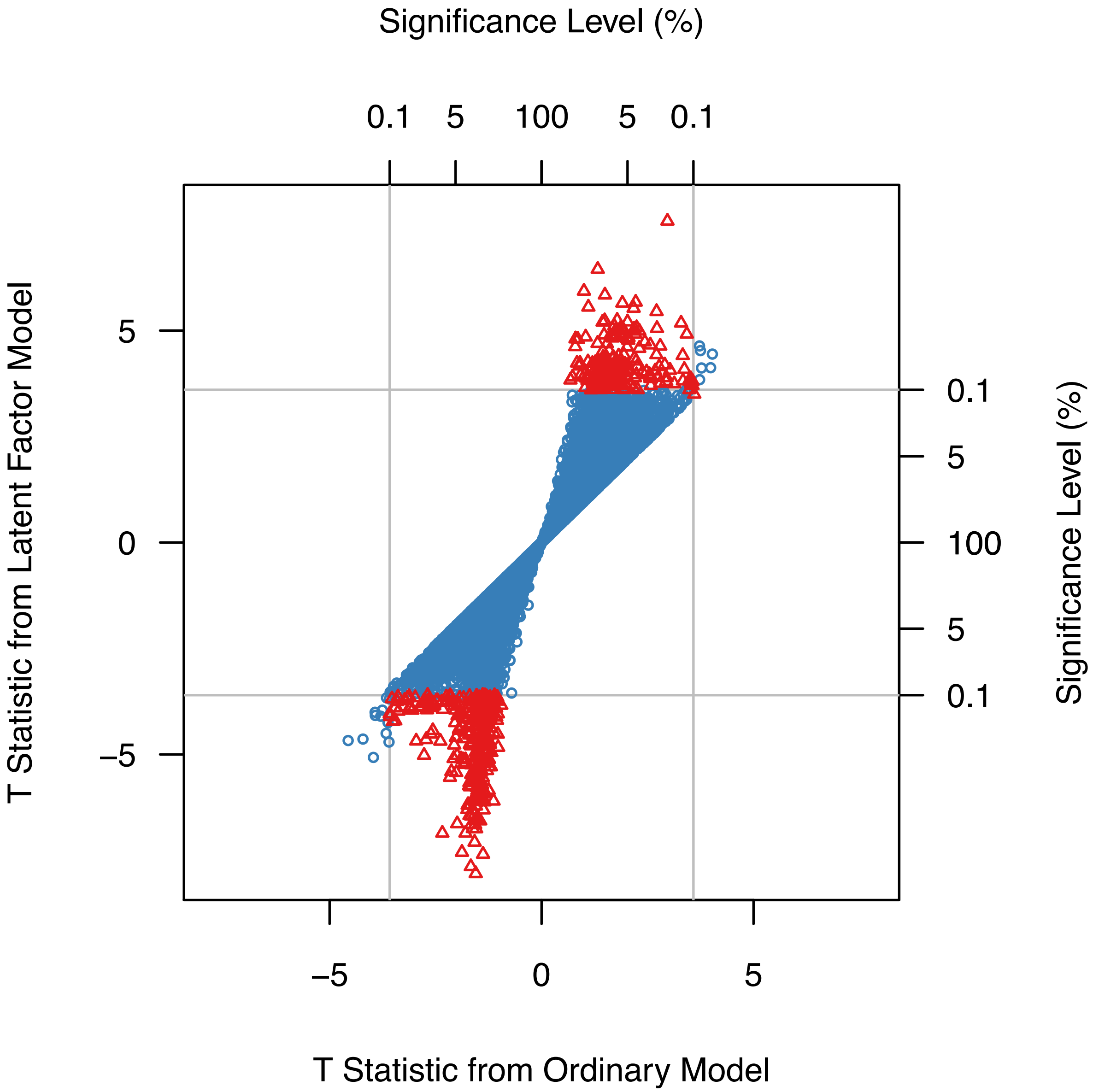
Results for Cerebrum Mm.71015

Without latent factor adjustment: $T = 1.356$, $df = 36$; $p = .18$

With latent factor adjustment: $T = 3.73$, $df = 33.8$; $p < .001$

Each latent factor uses 1.1 degrees of freedom

Results for All Genes



Summary

With matrix-variate response, we can **use regression to model associations between covariates and response.**

Latent sources of variability affect our inferences. We can **adjust for latent sources of variability using a latent factor model.**

We can use **conservative degrees of freedom estimates** for the factor terms.

Thank You

Extra Innings

Linking Covariates to Response

Row covariates only: $Y = X B^T + E$

Column covariates only: $Y = A Z^T + E$

Both sets of covariates: $Y = A Z^T + X B^T + E$

Identifying Parameters

Can't identify A, B:

$$AZ^T + XB^T = (A + XC^T)Z^T + X(B - ZC)^T$$

Can only identify residuals from regressing A on X, B on Z!

If s is any vector such that $Z^T s = 0$, then we can identify $B^T s$.

Example: $Z = 1_{M,1}$ $s = e_j - \frac{1}{M} 1_M$

Interpreting Parameters and Estimates

Parameters: "After adjusting for gene-specific affects, holding sex constant, increasing age by 1 unit is associated with increasing expected log activation of gene j by $\beta_{j,age}$"

Estimates (incorrect): "The estimated value of $\beta_{j,age}$ is …"

Estimates (correct): "The estimated difference between $\beta_{j,age}$ and the average value of $\beta_{j',age}$ for all genes j' of the same tissue type is …"

Regression: explain data variability with row and column effects

Factor Analysis: explain data variability with latent factors

Bilinear Model: combine regression with factor analysis

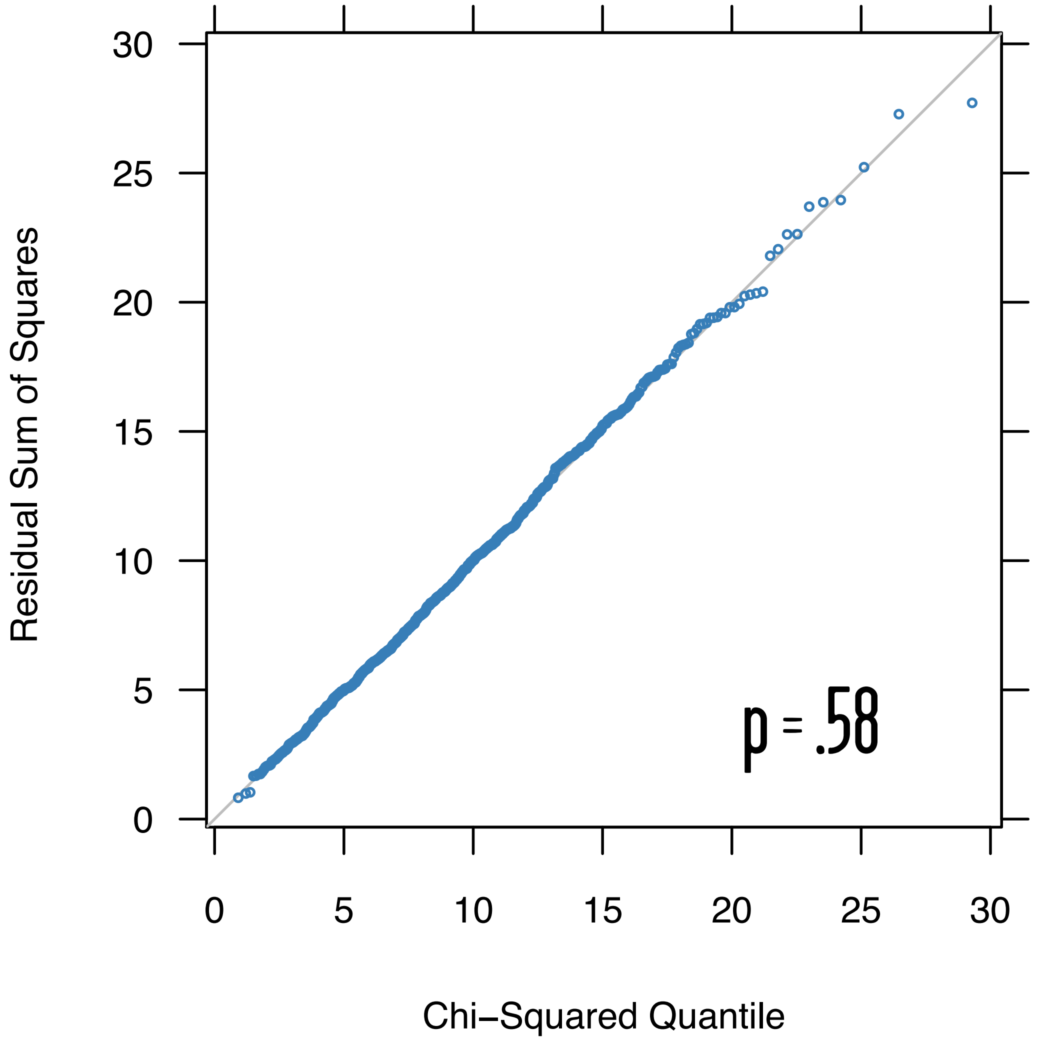
History

Early approaches: Fisher and Mackenzie (1923), Cochran (1943), Williams (1952)

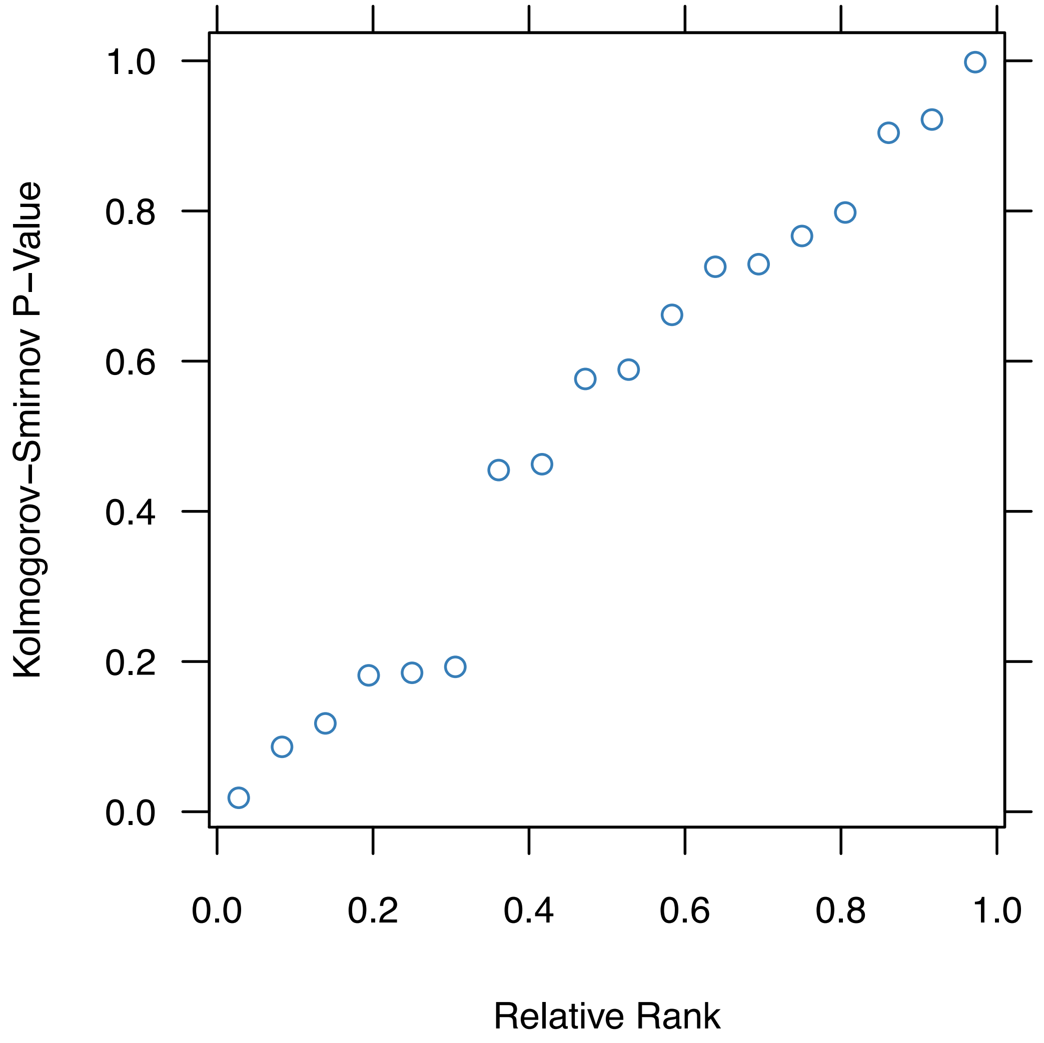
Further development: Tukey (1962), Gollob (1968), Mandel (1969, 1972), Gabriel (1978)

Recent work: van Euwijk (1995), Cornelius and Seyedsadr (1997), Gabriel (1998), Crossa et al. (2002), dos S. Dias and Krzanowski (2003), West (2003), Hoff (2007), Carvalho et al. (2008), Leek and Storey (2008), Friguet et al. (2009), Sun et al. (2012)

Comparing with Chi-Squared (Noise Case)



(n = 10, m = 1000)



(All Simulations)

Why Not Use a Random Effects Model?

Random effects models make even more independence assumptions!!

Evaluating Estimator Bias

153 configurations for m, n, mu

Test if E[dfhat] > df using 1000 replicates

T statistics summary:

```
     Min.  1st Qu.   Median     Mean  3rd Qu.     Max. 
  -22.140  -13.540   -6.569   -8.183   -3.212    3.887 
```

1/153 test is significant at level 5%