OVERVIEW OF LEAST-SQUARES, MAXIMUM LIKELIHOOD AND BLUP PART 1

"ALL MODELS ARE WRONG,
BUT SOME ARE USEFUL"

(Box, 1976)

"ALL MODELS ARE USEFUL,

SOME ARE WRONG"

(Aike Kagada, unpublished)

MODEL BUILDING AS TAUGHT IN STATISTICS COURSES

- Postulate model (abstraction)
- Identify parameters and sources of variation
- Make distributional assumptions (e.g., normality)
- Iterate as follows:

- Examine model residuals ("goodness of fit")

random behavior → take model as "true"

non-random → residuals may contain information about parameters: revise model

MODELS CAN BE USED FOR

- **Description** (e.g., fraction of observed variability accounted for)
- **Inference** (e.g., expected difference between varieties is 100 kg/ha)
- **Prediction or forecast** (e.g., disease free progression of a patient treated in some manner is 3.4 years)

TYPICALLY: A model that is good or sensible for inference may not be good for prediction

IN PREDICTION

- Measure prediction errors and their distribution.
- Use cross-validation and analytical treatment, when possible or reasonable.
- Model complexity matters.
- Generalization ability: “do not extrapolate beyond the experimental region”.
- Statistical significance and predictive outcomes are distinct matters.
- Predictive outcomes do not necessarily guide about state of nature.

LEARNING UNOBSERVABLES FROM OBSERVABLES

MARGINAL $\beta \sim (0, \sigma_\beta^2)$

JOINT $p(\beta, y_1, y_2, \ldots, y_n)$

CONDITIONAL $p(y_1, y_2, \ldots, y_n|\beta)$

CONDITIONAL $p(\beta|y_1, y_2, \ldots, y_n)$

MARGINAL $p(y_1, y_2, \ldots, y_n)$

$$p(\beta|y_1, y_2, \ldots, y_n) = \frac{p(\beta, y_1, y_2, \ldots, y_n)}{p(y_1, y_2, \ldots, y_n)}$$

Uncertainty about unobservable, given observables

LEARNING OBSERVABLES (POTENTIALLY) FROM OBSERVABLES: PREDICTIVE DISTRIBUTIONS

H: set of hyper-parameters ("tuning knobs")

β=set of random variables in a predictive model

$$p(y_{n+1},...,y_{n+n_f}|y_1,y_2,...,y_n,H) = \int \frac{p(y_{n+1},...,y_{n+n_f},y_1,y_2,...,y_n,\beta,H)}{p\left(y_1,y_2,...,y_n|H\right)}d\beta$$

$$= \int \frac{p(y_{n+1},...,y_{n+n_f}|y_1,y_2,...,y_n,\beta,H)}{p\left(y_1,y_2,...,y_n|H\right)}p(y_1,y_2,...,y_n,\beta,H)d\beta$$

$$= \int p(y_{n+1},...,y_{n+n_f}|y_1,y_2,...,y_n,\beta,H)p(\beta|y_1,y_2,...,y_n,H)d\beta$$

$$= \int p(y_{n+1},...,y_{n+n_f}|\beta,H)p(\beta|y_1,y_2,...,y_n,H)d\beta$$

$$\text{IF } y_{n+1},...,y_{n+n_f} \perp y_1,y_2,...,y_n \text{ GIVEN } \beta,H$$

OFTEN, PREDICTIVE DISTRIBUTION CAN BE ESTIMATED USING SAMPLING METHODS. MODEL MAY BE VERY WRONG IN WHICH CASE OUTCOMES WILL NOT APPEAR WITH APPRECIABLE DENSITY

ESTIMATION 1: ORDINARY LEAST-SQUARES (FIXED MODELS)

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e}$$

minimize w.r. β $(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})$

$$\frac{d}{d\boldsymbol{\beta}}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})$$

$$= -2\mathbf{X}'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})$$

$$\mathbf{X}'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) = \mathbf{0}$$

$$\mathbf{X}'\mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{X}'\mathbf{y} \Rightarrow \begin{cases} \hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} \\ \beta^0 = (\mathbf{X}'\mathbf{X})^{-}\mathbf{X}'\mathbf{y} \end{cases}$$

Unique solution
If **X** has full-column rank

One of an infinite Number of solutions for a given g-inverse

A generalized inverse satisfies $\Rightarrow$ $\mathbf{X}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-}\mathbf{X}'\mathbf{X} = \mathbf{X}'\mathbf{X}$

Point and interval estimation

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e}; \quad \mathbf{e} \sim (0, \mathbf{I}\sigma_e^2)$$

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$$

$$E\left(\hat{\boldsymbol{\beta}}|\boldsymbol{\beta}, \mathbf{X}\right) = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'(\mathbf{X}\boldsymbol{\beta} + \mathbf{e})$$

$$= \boldsymbol{\beta} + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'E(\mathbf{e}) = \boldsymbol{\beta}$$

Ols unbiased

$$Var\left(\hat{\boldsymbol{\beta}}|\boldsymbol{\beta}, \mathbf{X}\right) = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'Var(\mathbf{y})\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}$$

$$= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{I}\sigma_e^2\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}$$

$$= (\mathbf{X}'\mathbf{X})^{-1}\sigma_e^2$$

Cov.matrix

Tests of hypothesis (assume normality)

$$H_0 : \boldsymbol{\beta} = \mathbf{0} \text{ vs } H_0 : \boldsymbol{\beta} \neq \mathbf{0} \text{ test: } \frac{\frac{\hat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{X}\hat{\boldsymbol{\beta}}}{rank(\mathbf{X})}}{\hat{\sigma}_e^2} \sim F_{(rank(\mathbf{X}), n-rank(\mathbf{X}))}$$

Entire model

$$\hat{\sigma}_e^2 = \frac{\left(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}\right)'\left(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}\right)}{n - rank(\mathbf{X})}$$

$$\widehat{Var}\left(\hat{\boldsymbol{\beta}}\right) = (\mathbf{X}'\mathbf{X})^{-1}\hat{\sigma}_e^2 = \{c^{ii}\}\hat{\sigma}_e^2$$

Sole regression (given other regs in model)

$$t_i = \frac{\hat{\beta}_i}{\sqrt{\{c^{ii}\}\hat{\sigma}_e^2}} \sim t(0, n - rank(\mathbf{X}), \sigma_e^2) \text{ under } H_0$$

Model complexity (effective no. parameters)

$$tr\left[\mathbf{X}\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\right] = tr\left[\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{X}\right] = p = rank\left(\mathbf{X}\right),$$

Example of ordinary least-squares: polynomial regression model

Suppose **third** order model $y = \beta_0 + \beta_1 x_1 + \beta_{11} x_1^2 + \beta_{111} x_1^3 + e$

$e \sim (0, \sigma^2)$

$$\mathbf{y} = \begin{bmatrix} 4 \\ 2 \\ 1 \\ 7 \\ 8 \\ 10 \\ 9 \\ 10 \end{bmatrix} \quad X = \begin{bmatrix} 1 & 2 & 4 & 8 \\ 1 & 1 & 1 & 1 \\ 1 & 2 & 4 & 8 \\ 1 & 1 & 1 & 1 \\ 1 & 2 & 4 & 8 \\ 1 & 4 & 16 & 64 \\ 1 & 3 & 9 & 27 \\ 1 & 3 & 9 & 27 \end{bmatrix}$$

$$X^T X = \begin{bmatrix} 8.0 & 18.0 & 48.0 & 144.0 \\ 18.0 & 48.0 & 144.0 & 468.0 \\ 48.0 & 144.0 & 468.0 & 1608.0 \\ 144.0 & 468.0 & 1608.0 & 5748.0 \end{bmatrix}$$

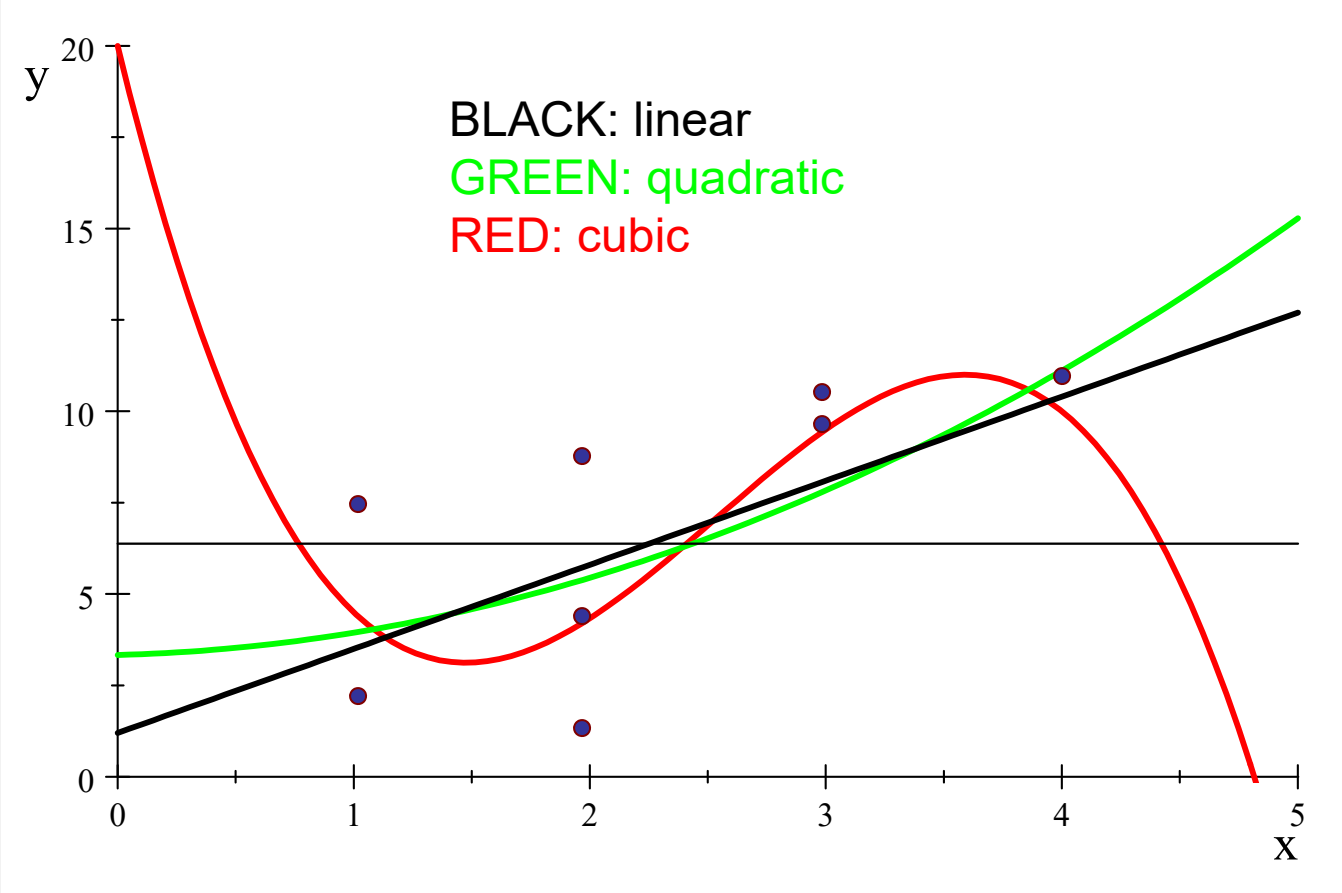
$$X^T y = \begin{bmatrix} 51.0 \\ 132.0 \\ 392.0 \\ 1266.0 \end{bmatrix}$$

rank: 4

Comparison of four fits (see how coefficients change, e.g., **linear**

Mean	6.375
Linear	$2.299\,999\,996x + 1.2$
Quadratic	$0.444\,444\,446\,1x^2 + 0.166\,666\,609\,6x + 3.333\,333\,346$
Cubic	$-1.666\,666\,536x^3 + 12.666\,665\,78x^2 - 26.500\,000\,44x + 20.0$

$$y = \begin{bmatrix} 4 \\ 2 \\ 1 \\ 7 \\ 8 \\ 10 \\ 9 \\ 10 \end{bmatrix} \quad X = \begin{bmatrix} 1 & 2 & 4 & 8 \\ 1 & 1 & 1 & 1 \\ 1 & 2 & 4 & 8 \\ 1 & 1 & 1 & 1 \\ 1 & 2 & 4 & 8 \\ 1 & 4 & 16 & 64 \\ 1 & 3 & 9 & 27 \\ 1 & 3 & 9 & 27 \end{bmatrix}$$



EXAMPLE OF ORDINARY LEAST-SQUARES: GWAS VIA SINGLE MARKER REGRESSION

GWAS: *search for association between some marker or genomic region and a phenotype*

Marker allele substitution effect

Marker genotype (0,1,2)

$$y_{ij} = \beta_0 + \beta_j x_{ij} + e_{ij}$$

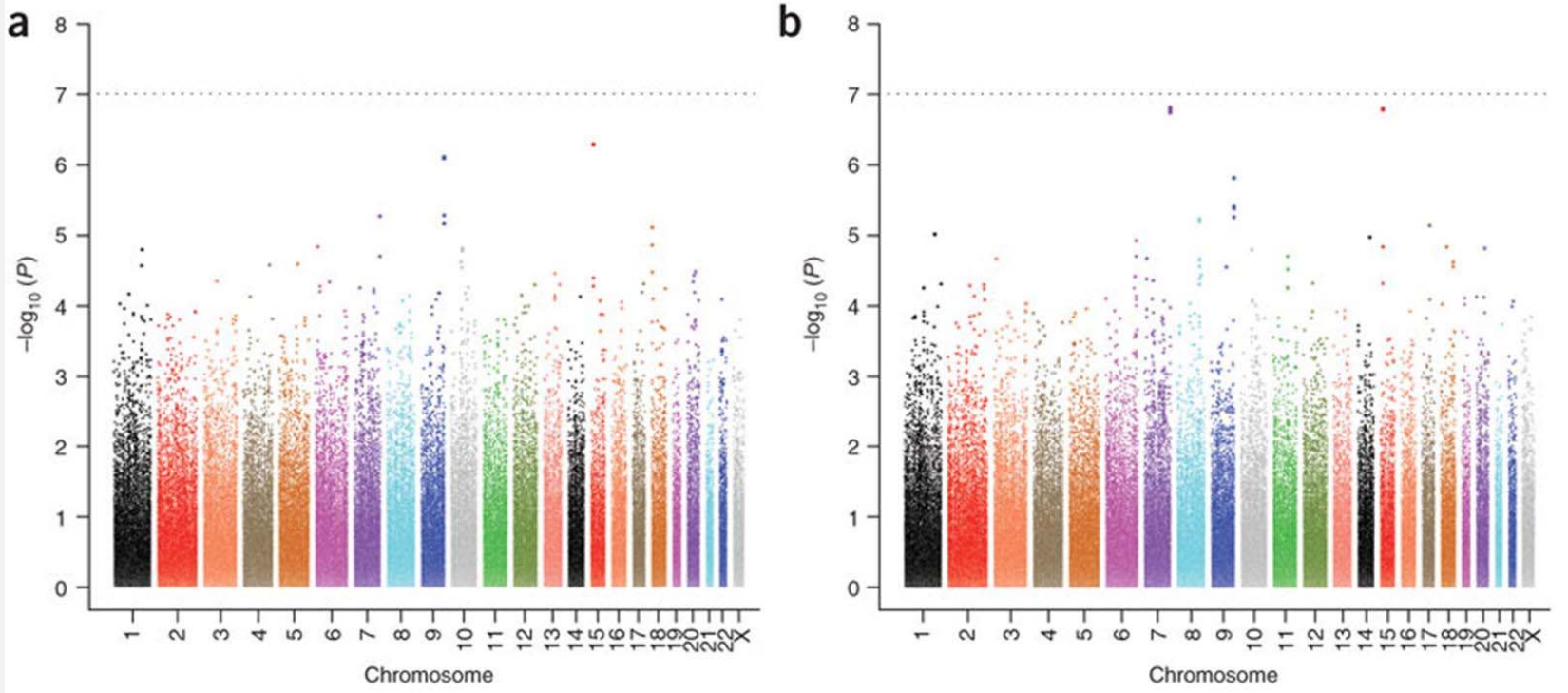
$$i = 1,2,\ldots,n$$

$$j = 1,2,\ldots,p$$

$$e_{ij} \sim N\left(0, \sigma^2_{e_j}\right)$$

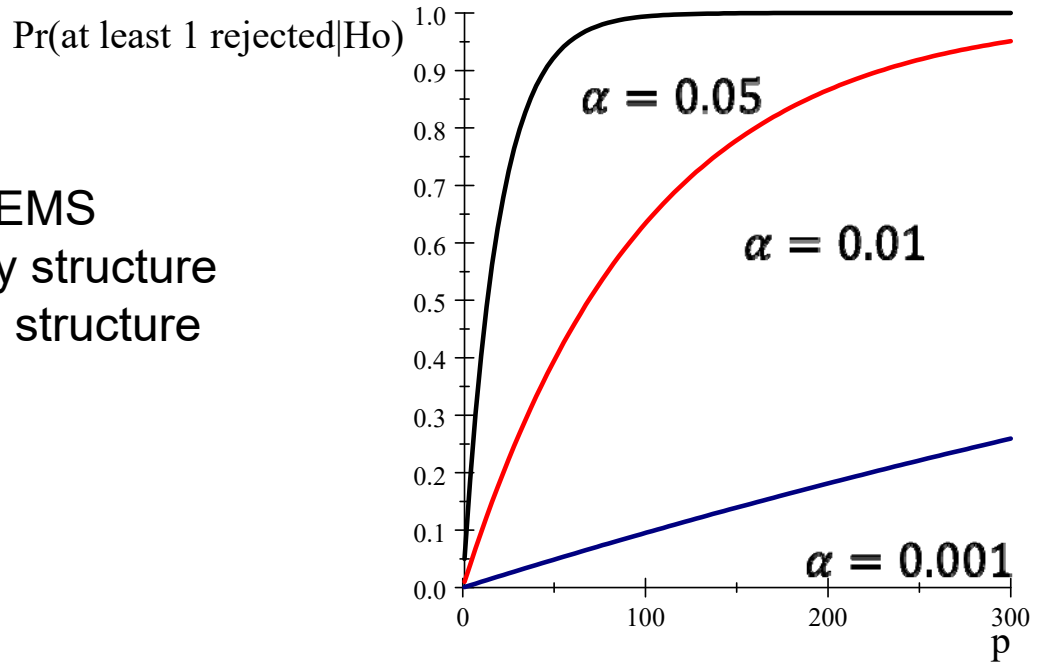
Residual variance typically differs over markers

“MANHATTAN” PLOTS OF P-VALUES
(points are -log(p-values, base 10) by marker)



GWAS FOR PANCREATIC CANCER…
(Nature Genetics)

AT LEAST THREE PROBLEMS
-Naïve model in explanatory structure
-Naïve model in covariance structure
-Multiple-testing issue



H_{0j} : $\beta_j = 0$ vs. $\beta_j \neq 0$

$j = 1, 2, \ldots, p$ $\alpha =$ significance level

$$\Pr(\text{at least one significance}|H_0) = 1 - \Pr(\text{all NS})$$
$$= 1 - (1-\alpha)^p$$

Bonferroni method: given p-values $p_1, p_2 \ldots, p_p$ reject hypothesis if $p_j \leq \frac{\alpha}{p}$

Genetics: Early Online, published on July 9, 2014 as 10.1534/genetics.114.164442

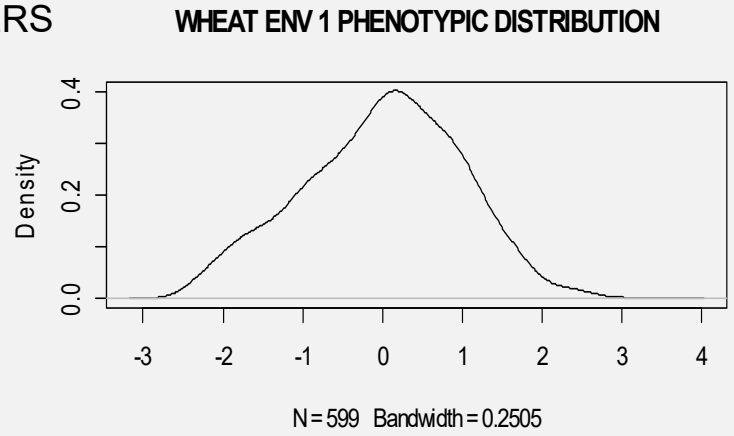
Genome-wide Regression & Prediction with the BGLR statistical package

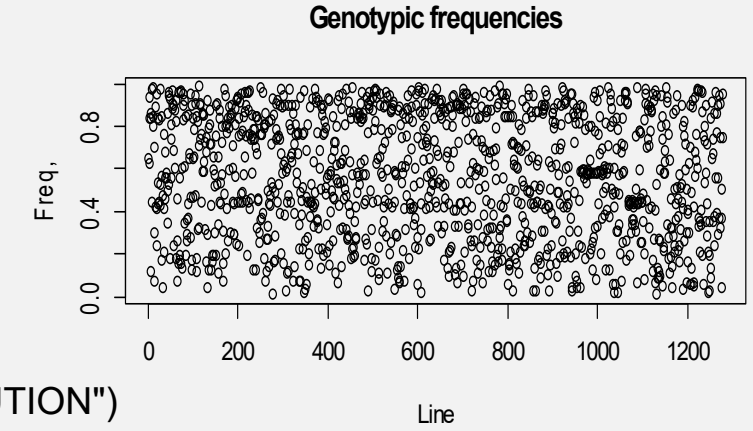
Paulino Pérez
Socio Economía Estadística e Informática,
Colegio de Postgraduados, México
perpdgo@colpos.mx

Gustavo de los Campos
Department of Biostatistics, Section on Statistical Genetics
University of Alabama at Birmingham, USA
gcampos@uab.edu

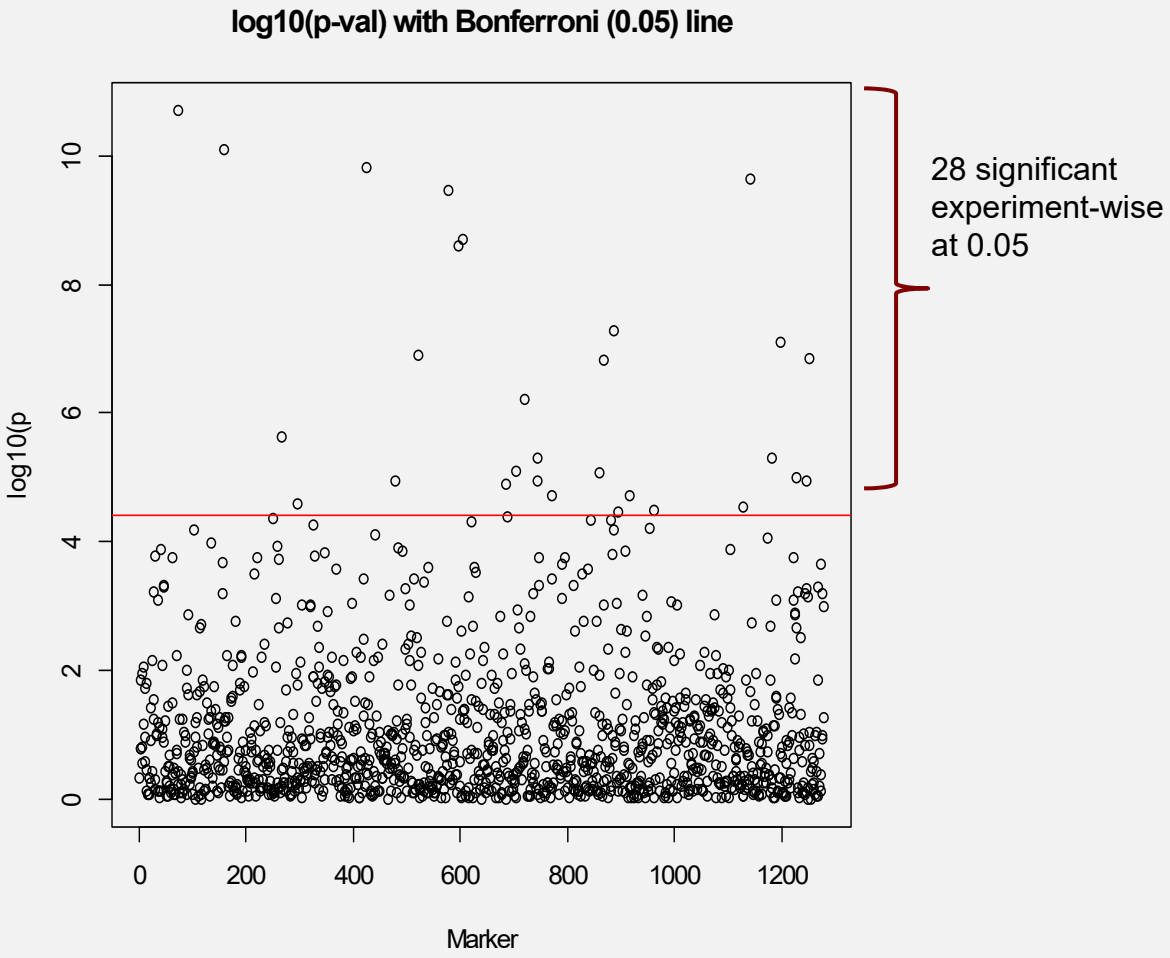
PSEUDO-GWAS OF WHEAT DATA SET n=599 p=1279

```
library(BGLR)
set.seed(1234567)

###LOAD DATA ###599 INBRED LINES, 1279 BINARY MARKERS
data(wheat)
Y<-wheat.Y
X<-wheat.X
y<-Y[,1]
n<-nrow(X)
p<-ncol(X)

###CHECK PHENOTYPES
mean(y)
var(y)

####CHECK FREQUENCIES
freq<-numeric(p)
for (i in 1:p){
freq[i]<-mean(X[,i])
}
####PLOT

par(mfrow=c(1,2))
plot(density(y),main="WHEAT ENV 1 PHENOTYPIC DISTRIBUTION")
plot(freq,main="Genotypic frequencies",xlab="Line",ylab="Freq,")
Par(mfrow=c(1,1))
```





```
########ALLOCATE SPACE
bols<-numeric(p)
pvaluesols<-numeric(p)
Vresols<-numeric(p)
Rsquaredols<-numeric(p)
logpols<-numeric(p)

#####FIT MODEL WITH INTERCEPT
for(i in 1:p){

GWASols<-lm(y~X[,i])
bols[i]<-GWASols$coefficients[2]
opresstdev<-summary(GWASols)$sigma
Vresols[i]<-opresstdev**2
Rsquaredols[i]<-
summary(GWASols)$r.squared
pvaluesols[i]<-
summary(GWASols)$coef[2,4]

}

logpols<--log(pvaluesols,base=10)

plot(logpols,main="log10(p-val) with
Bonferroni (0.05) line",
ylab="log10(p",xlab="Marker")
abline(a=-
log(0.05/p,base=10),b=0,col="red")
```



GWAS DONE!

- Found many “significant” genomic regions.
- Issue: grain yield is probably multi-factorial.
- Implication: many genomic regions should be at play jointly
- How do we interpret results?

CAUSAL VERSUS INSTRUMENTAL MODEL

QUANTITATIVE TRAIT LOCI "DETERMINE" PHENOTYPES. ASSUME:

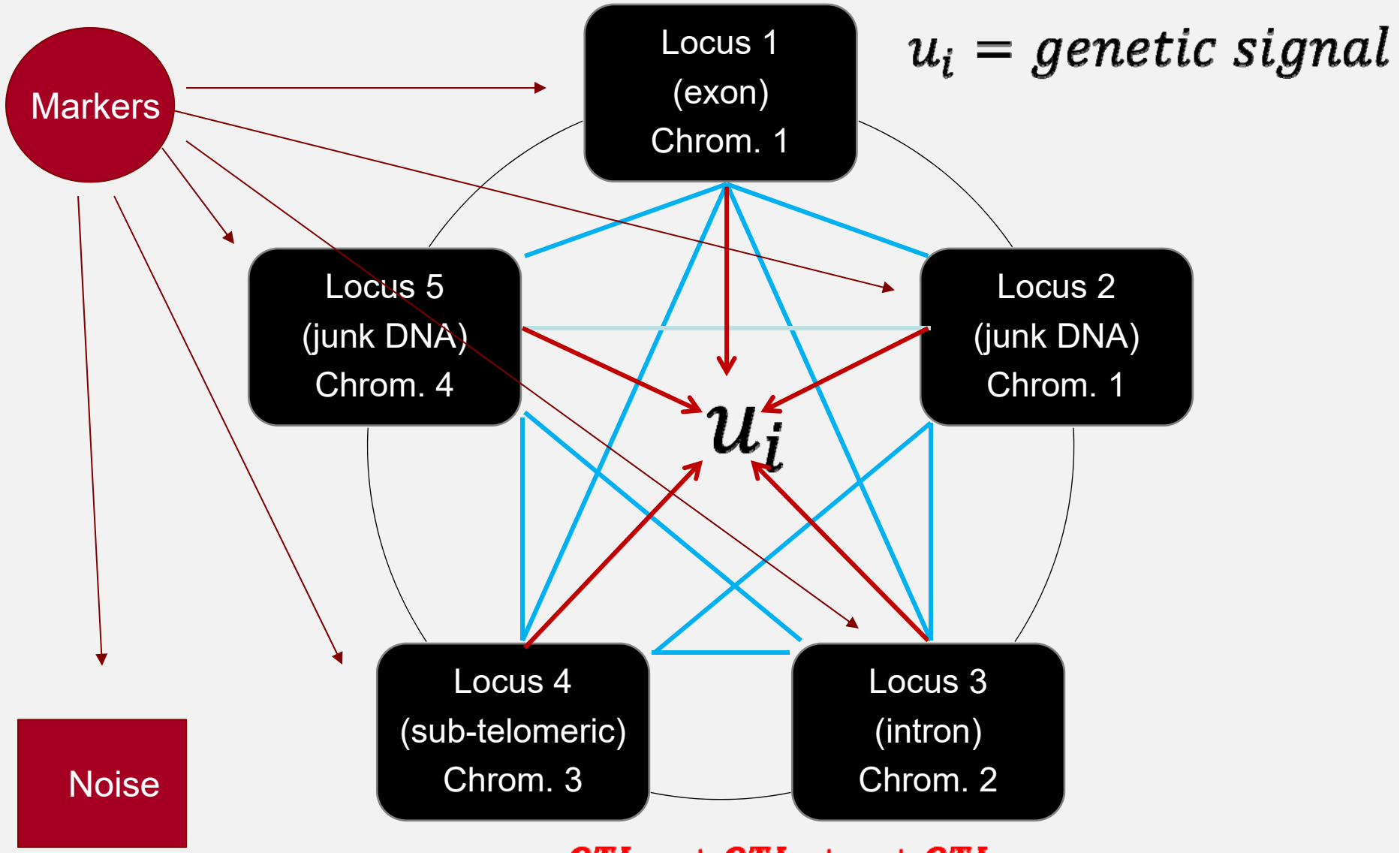
$$E(\mathbf{y}|\mathbf{Q}) = \mathbf{Q}\mathbf{q} = \boldsymbol{\mu}$$

MARKERS ARE NOT QTL

$$\begin{aligned} E_{\mathbf{y}|\mathbf{X},\mathbf{Q}}\left(\widehat{\boldsymbol{\beta}}|\mathbf{X},\mathbf{Q}\right) &= E_{\mathbf{y}|\mathbf{X},\mathbf{Q}}[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}] \\ &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'E_{\mathbf{y}|\mathbf{Q}}[(\mathbf{y})] \\ &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Q}\mathbf{q} \\ &= \mathbf{T}\mathbf{q}, \end{aligned}$$

where $\mathbf{T}_{n\times l}$ projects true genetic effects into marker effects. A given column of $\mathbf{T}$ can be interpreted as the multiple regression of the corresponding column of $\mathbf{Q}$ on the p markers, and $\mathbf{T}$ is a realization of a stochastic process called linkage disequilibrium (LD). If any of the columns of $\mathbf{T}$ is null, then markers do not inform about genotypes for a specific genetic effect. All, some or none of the markers may be in LD with one of more QTL. Although $\widehat{\boldsymbol{\beta}}$ informs in some manner about $\mathbf{q}$, the fact that $\mathbf{X}'\mathbf{Q}$ is typically unknown ($\mathbf{Q}$ is not an observable matrix, contrary to $\mathbf{X}$) indicates that genetic or causal interpretation of estimates of marker effects must be done with caution. The term LD actually refers to the population as a whole; it may be that $E(\mathbf{T})$ is a null matrix, but a particular realization of $\mathbf{T}$ may suggest linkage disequilibrium when none exists.

GWAS can be ambiguous: Five locus system in linkage disequilibrium. Arrows represent direct effects on additive genetic value (*u*) ; undirected lines and arcs represent correlations between genotypes stemming from LD.



$$u = QTL_1 + QTL_2 + \cdots + QTL_5$$

How many QTLs? “Honey I shrunk epistasis!”

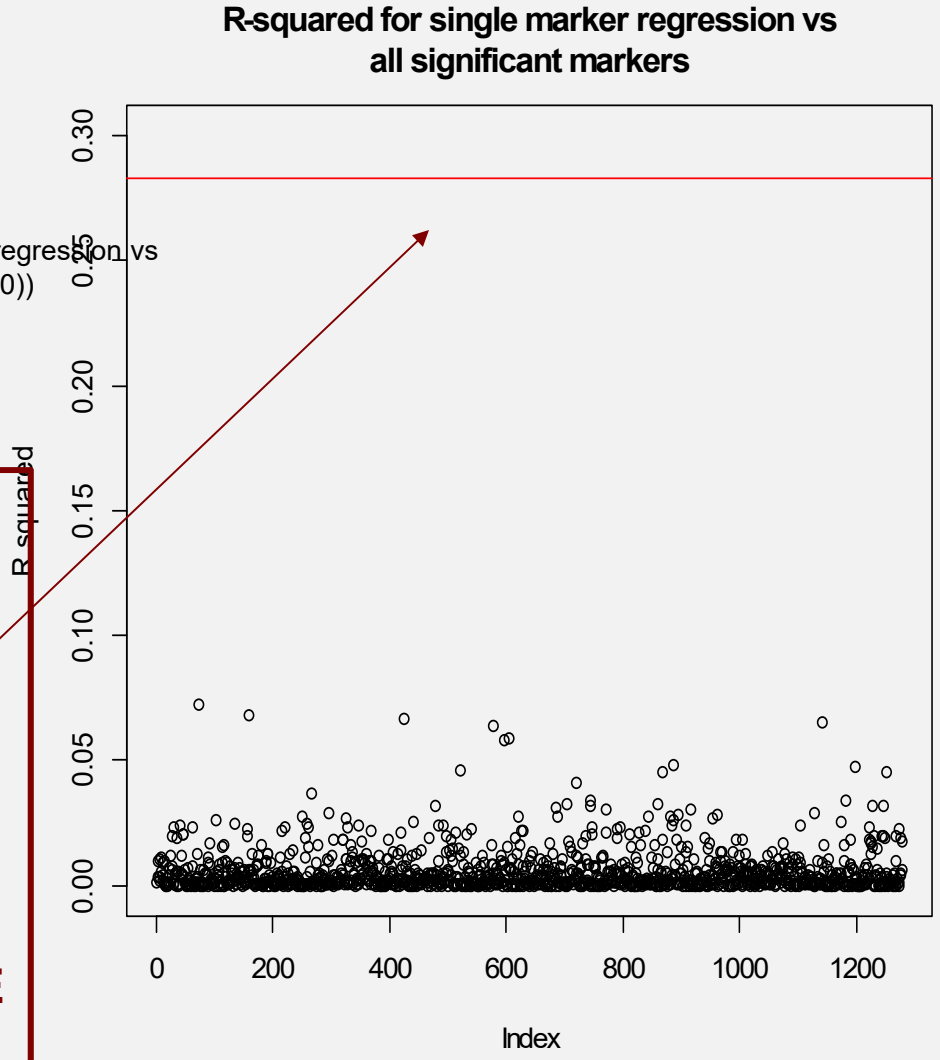
MODEL MODIFICATION

####Fitting significant markers together In a multiple regression

SOME MARKERS THAT WERE SIGNIFICANT ARE NO LONGER SO, GIVEN OTHER MARKERS IN MODEL

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  -2.599943   0.487964  -5.328 1.43e-07 ***
XsigwPt.2185  0.701304   0.398665   1.759 0.079093 .
XsigwPt.3697  0.472445   0.214281   2.205 0.027868 *
XsigwPt.1325 -0.035093   0.196546  -0.179 0.858355
XsigwPt.0921  0.418388   0.275447   1.519 0.129333
XsigwPt.2087 -0.010930   0.456455  -0.024 0.980904
XsigwPt.3132  0.025580   0.208431   0.123 0.902366
XsigwPt.9256  0.285042   0.077435   3.681 0.000254 ***
XsigwPt.2448  0.653652   0.997529   0.655 0.512558
XsigwPt.4533  0.028555   0.277646   0.103 0.918121
XsigwPt.9422  0.007167   0.209006   0.034 0.972657
Xsigc.304224 -0.163335   0.364688  -0.448 0.654414
Xsigc.304430 -0.190215   0.522496  -0.364 0.715956
Xsigc.304701 -0.444809   0.324604  -1.370 0.171130
Xsigc.305115 -0.010829   0.194624  -0.056 0.955646
Xsigc.305166 -0.257800   0.170849  -1.509 0.131870
Xsigc.305742  0.496554   0.276734   1.794 0.073290 .
Xsigc.344673  0.026619   0.370227   0.072 0.942708
Xsigc.344809 -0.751904   0.263346  -2.855 0.004458 **
Xsigc.345107  0.149420   0.235180   0.635 0.525459
Xsigc.345237 -0.151406   0.216330  -0.700 0.484286
Xsigc.345583  0.018636   0.109260   0.171 0.864623
Xsigc.346394 -0.169585   0.206248  -0.822 0.411285
Xsigc.375520  0.697958   0.206274   3.384 0.000765 ***
Xsigc.376463 -0.536227   1.082800  -0.495 0.620634
Xsigc.378288 -0.252120   0.110455  -2.283 0.022825 *
Xsigc.378625  0.054522   0.195429   0.279 0.780356
Xsigc.379670  0.109837   0.522986   0.210 0.833728
Xsigc.380286 -0.423306   0.308241  -1.373 0.170201
Xsigc.381104  0.204391   0.106374   1.921 0.055176 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.8681 on 569 degrees of freedom
Multiple R-squared:  0.2829,    Adjusted R-squared:  0.2464
F-statistic: 7.742 on 29 and 569 DF,  p-value: < 2.2e-16
```

```
###VARIATION EXPLAINED BY SINGLE MARKERS
### AND BY ALL SIGNIFICANT MARKERS
###FITTED TOGETHER

Rsquaredsig<-summary(GWASsig)$r.squared

plot(Rsquaredols,main="R-squared for single marker regression vs
all significant markers", ylab="R.squared",ylim=c(0,0.30))
abline(a=Rsquaredsig,b=0,col="red")
```



NOTE:

sumRsq<-sum(Rsquaredols)
sumRsq
[1] 7.081406

Rsquaredsig=**0.28**

CANNOT ADD UP VARIANCE DUE TO LOCI IN LD

SIMPLE STATISTICAL EXPLANATION FOR PART OF "MISSING HERITABILITY"!

ESTIMATION 2: GENERALIZED LEAST-SQUARES ESTIMATION OF FIXED EFFECTS (GAUSS-MARKOV: V matrix is known)

More general covariance structure, e.g., due to spatial or genetic similarities

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}; \quad \mathbf{e} \sim (\mathbf{0}, \mathbf{V})$$

$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \left[(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})' \mathbf{V}^{-1} (\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) \right]$$

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{-1}\mathbf{y}$$

$$E(\hat{\boldsymbol{\beta}}) = \boldsymbol{\beta}$$

unbiased

$$Var(\hat{\boldsymbol{\beta}}) = (\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}$$

Test statistic for model fit:

$$\mathbf{y} \sim N(\mathbf{X}\boldsymbol{\beta}, \mathbf{V})$$

$$H_0 : \boldsymbol{\beta} = \mathbf{0} \rightarrow \hat{\boldsymbol{\beta}}\mathbf{X}'\mathbf{V}^{-1}\mathbf{X}\hat{\boldsymbol{\beta}} \sim \chi^2_{rank(\mathbf{X})} \text{ if } \mathbf{V} \text{ known}$$

EXAMPLE OF POSSIBLE FORM OF **V**

$$\mathbf{y} = \mathbf{X\beta} + \mathbf{Zg} + \mathbf{W\gamma}_{ge} + \mathbf{e}$$

$$\boldsymbol{\epsilon} = \mathbf{Zg} + \mathbf{W\gamma}_{ge} + \mathbf{e}$$

$$\mathbf{V} = Var\left(\mathbf{Zg} + \mathbf{W\gamma}_{ge} + \mathbf{e}\right)$$

IF $\mathbf{g}, \boldsymbol{\gamma}_{ge}, \boldsymbol{\gamma}_{ge}$ UNCORRELATED

$$= \mathbf{Z}Var(\mathbf{g})\mathbf{Z}' + \mathbf{W}Var\left(\boldsymbol{\gamma}_{ge}\right)\mathbf{W}' + Var(\mathbf{e})$$

IF SINGLE TRAIT AND

VARIANCE COMPONENT MODEL

$$= \mathbf{ZGZ}'\sigma_g^2 + \mathbf{WTW}'\sigma_{ge}^2 + \mathbf{I}\sigma_e^2$$

EXAMPLE OF GWAS MODEL WITH A RANDOM FACTOR IN MODEL

Set of fixed effects to be estimated or tested (e.g., GWAS)

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{g} + \mathbf{e}$$

$$\mathbf{g} \sim N(0, \mathbf{G}\sigma_g^2)$$

Vector of random "line genotypic effects"

Genomic variance

$$\mathbf{e} \sim N(0, \mathbf{I}\sigma_e^2)$$

IF $\mathbf{G}$ IS marker based kinship matrix

$$h_g^2 = \frac{\sigma_g^2}{\sigma_g^2 + \sigma_e^2}$$ "genomic heritability"

Can write

$$\begin{aligned}\mathbf{V} &= \mathbf{G}\sigma_g^2 + \mathbf{I}\sigma_e^2 \\ &= \left(\mathbf{G}\frac{h_g^2}{1-h_g^2} + \mathbf{I}\right)\sigma_e^2 \\ &= \mathbf{V}^*_{h_g^2}\sigma_e^2\end{aligned}$$

$$\widehat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{-1}\mathbf{y}$$

$$= \left(\mathbf{X}'\left(\mathbf{V}^{*}_{h_g^2}\sigma_e^2\right)^{-1}\mathbf{X}\right)^{-1}\mathbf{X}'\left(\mathbf{V}^{*}_{h_g^2}\sigma_e^2\right)^{-1}\mathbf{y}$$

$$= \left(\mathbf{X}'\mathbf{V}^{*-1}_{h_g^2}\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{V}^{*-1}_{h_g^2}\mathbf{y}$$

GLS DEPENDS ON h_g^2

(and not on the two variances)

Computing the GLS estimator using an OLS program

In single trait model can write

$$\mathbf{V} = \mathbf{V}^{*}\sigma_e^2$$

Do Cholesky decomposition

$$\mathbf{V}^{*} = \mathbf{C}'\mathbf{C} \rightarrow \mathbf{V}^{*-1} = (\mathbf{C})^{-1}(\mathbf{C}')^{-1}$$

Transform data and model

$$\mathbf{y}^{*} = (\mathbf{C}')^{-1}\mathbf{y} = (\mathbf{C}')^{-1}\mathbf{X}\boldsymbol{\beta} + (\mathbf{C}')^{-1}\boldsymbol{\varepsilon} = \mathbf{X}^{*}\boldsymbol{\beta} + \boldsymbol{\varepsilon}^{*}$$

$$Var(\boldsymbol{\varepsilon}^{*}) = (\mathbf{C}')^{-1}\mathbf{C}'\mathbf{C}\sigma_e^2(\mathbf{C})^{-1} = \mathbf{I}\sigma_e^2$$

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^{*\prime}\mathbf{X}^{*})^{-1}\mathbf{X}^{*\prime}\mathbf{y}^{*} = \left(\mathbf{X}'(\mathbf{C})^{-1}(\mathbf{C}')^{-1}\mathbf{X}\right)^{-1}\mathbf{X}'(\mathbf{C})^{-1}(\mathbf{C}')^{-1}\mathbf{y}$$

$$= (\mathbf{X}'\mathbf{V}^{*-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{*-1}\mathbf{y}$$

$$Var(\hat{\boldsymbol{\beta}}) = (\mathbf{X}'\mathbf{V}^{*-1}\mathbf{X})^{-1}\sigma_e^2$$

GLS estimator and its variance retrieved

Means that we can do GWAS with GLS using a grid of genomic heritability values and an ordinary least-squares program

EXAMPLE OF GENERALIZED LEAST-SQUARES GWAS VIA SINGLE MARKER REGRESSION

FIXED RANDOM

$$y_{ij} = \beta_0 + \beta_j x_{ij} + g_i + e_{ij}$$

$$i = 1, 2, \ldots, n$$

$$j = 1, 2, \ldots, p$$

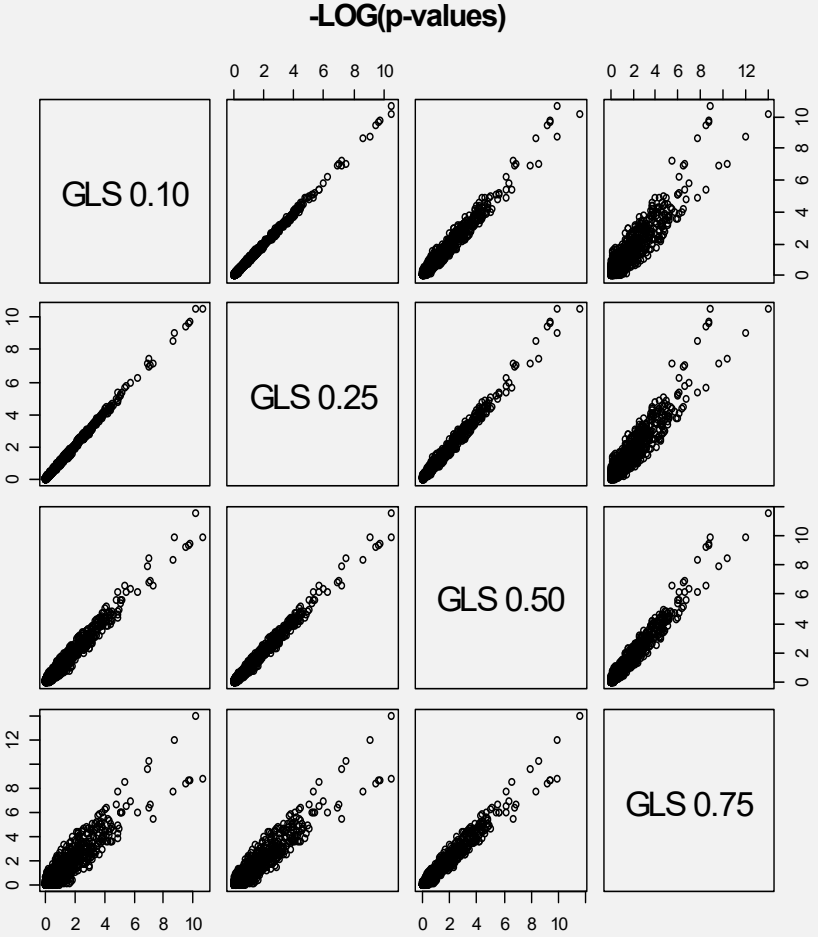
$\mathbf{g} = \{g_i\} \sim N(0, \mathbf{G}\sigma_g^2)$ for some similarity matrix $\mathbf{G}$

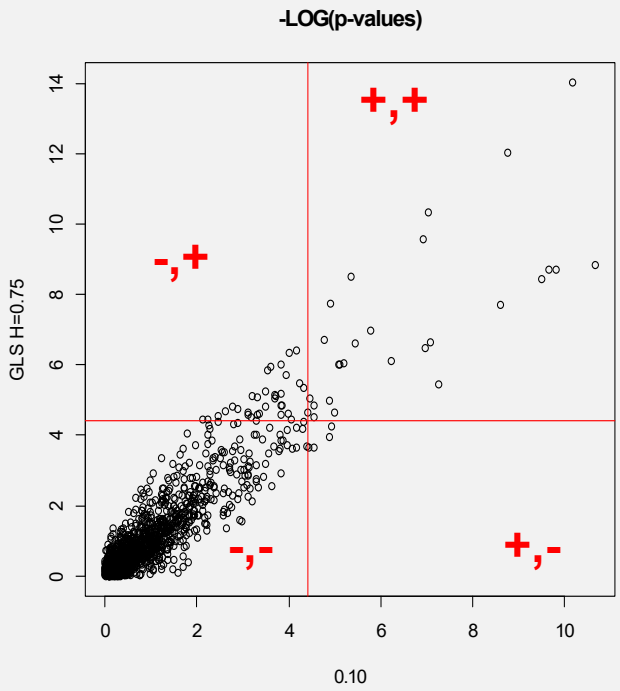
here we will use pedigree matrix $\mathbf{A}$

$$e_{ij} \sim N\left(0, \sigma_{e_j}^2\right)$$

$$H^2 = \frac{\sigma_e^2}{\sigma_g^2 + \sigma_e^2} \Rightarrow \lambda = \frac{\sigma_e^2}{\sigma_g^2} = \frac{1 - H^2}{H^2}$$

$\sigma^2{}_g$





ACCOUNTING FOR COVARIANCE STRUCTURE CAN MAKE A DIFFERENCE

THE LEAST-SQUARES PREDICTOR

Consider the problem of using least-squares for out-of-sample prediction. The setting presented here is one of a single training and a single testing set. Suppose that the distribution of residuals in left-out data (testing set) of size n_{test} is as in a training set of size n_{train}, with the training phenotypes represented as $\mathbf{y}_{train}$. Residuals in training and testing sets are assumed to be mutually independent and normally distributed. The predictor of future data $\mathbf{y}_{test} \sim N\left(\boldsymbol{\mu}_{test}, \mathbf{I}\sigma_e^2\right)$ having the smallest averaged squared prediction error (Henderson 1973; Searle 1974) is called the best predictor (BP). Under normality the BP is the linear function of training data

$$E\left(\mathbf{y}_{test}|\mathbf{y}_{train}\right) = \boldsymbol{\mu}_{test} + Cov\left(\mathbf{y}_{test}, \mathbf{y}'_{train}\right) Var\left(\mathbf{y}_{train}\right)\left[\mathbf{y}_{train} - \boldsymbol{\mu}_{train}\right] \tag{7}$$

$$\begin{aligned} BLUP\left(\mathbf{y}_{test}\right) &= \widehat{\mathbf{y}}_{test} \\ &= \mathbf{X}_{test}\widehat{\boldsymbol{\beta}} \\ &= \mathbf{X}_{test}\left(\mathbf{X}'_{train}\mathbf{X}_{train}\right)^{-1}\mathbf{X}'_{train}\mathbf{y}_{train} \\ &= \mathbf{H}_{test,train}\mathbf{y}_{train}, \end{aligned}$$

$$\begin{aligned} \frac{\partial \widehat{\mathbf{y}}_{train}}{\partial \mathbf{y}'_{train}} &= \frac{\partial \mathbf{X}_{train}\left(\mathbf{X}'_{train}\mathbf{X}_{train}\right)^{-1}\mathbf{X}'_{train}\mathbf{y}_{train}}{\partial \mathbf{y}'_{train}} \\ &= \mathbf{H}_{train,train}, \end{aligned}$$

$$\frac{\partial \widehat{\mathbf{y}}_{test}}{\partial \mathbf{y}'_{train}} = \mathbf{H}_{test,train}.$$

Statistical uncertainty of predictions

$$Var\left(\mathbf{y}_{test} - \mathbf{X}_{test}\widehat{\boldsymbol{\beta}}\right) = Var\left(\mathbf{y}_{test}\right) + Var\left(\mathbf{X}_{test}\widehat{\boldsymbol{\beta}}\right)$$

$$\begin{aligned} Var\left(\mathbf{y}_{test} - \mathbf{X}_{test}\widehat{\boldsymbol{\beta}}\right) &= \mathbf{I}\sigma_e^2 + \mathbf{X}_{test} Var\left(\widehat{\boldsymbol{\beta}}\right) \mathbf{X}'_{test} \\ &= \left[\mathbf{I} + \mathbf{X}_{test}\left(\mathbf{X}'_{train}\mathbf{X}_{train}\right)^{-1}\mathbf{X}'_{test}\right]\sigma_e^2. \end{aligned} \tag{13}$$

The first term in the expression above is the variability of "new errors", stemming from the fact that residuals in the testing set have not been "seen" during the training process; the second terms is the covariance matrix of the predictor. Note that even if when the predictor is very precise, there will be uncertainty due to the fact that there is no (or imperfect) information on testing set residuals.

PREDICTION PERFORMANCE

(least-squares formulae but concepts carry to other methods)

Measuring agreement between predictions and outcomes: correlation or some distance metric?

Perturb distributions by adding biases
(covariance structure is insensitive)
x is a predictor of y

$$\begin{aligned} E(\mathbf{x}) &= \boldsymbol{\mu}, \quad E(\mathbf{y}) = \boldsymbol{\mu}' \\ Cov(\mathbf{x}, \mathbf{y}) &= \mathbf{V} \\ Cov(\mathbf{x} + \boldsymbol{\delta}, \mathbf{y} + \boldsymbol{\delta}') &= \mathbf{V} \end{aligned}$$

Euclidean distance between
predictor (x) and predictand (y)

$$\begin{aligned} d^2(\mathbf{x}, \mathbf{y}) &= (\mathbf{x} - \mathbf{y})'(\mathbf{x} - \mathbf{y}) \\ E\left[d^2(\mathbf{x}, \mathbf{y})\right] &= (\boldsymbol{\mu} - \boldsymbol{\mu}')'(\boldsymbol{\mu} - \boldsymbol{\mu}') + tr\left[Var(\mathbf{x} - \mathbf{y})\right] \end{aligned}$$

Covariance matrix of prediction errors
(trace gives sum of prediction error variances)

Euclidean distance between non-randomly disturbed predictor (x) and non-randomly disturbed predictand (y)

$$
\begin{aligned}
E\left[d^2(\mathbf{x}+\boldsymbol{\delta},\mathbf{y}+\boldsymbol{\delta}')\right] &= E\left[(\mathbf{x}-\mathbf{y}+\boldsymbol{\delta}-\boldsymbol{\delta}')'(\mathbf{x}-\mathbf{y}+\boldsymbol{\delta}-\boldsymbol{\delta}')\right] \\
&= E\left[(\mathbf{x}-\mathbf{y})'(\mathbf{x}-\mathbf{y})\right] + E(\mathbf{x}-\mathbf{y})'(\boldsymbol{\delta}-\boldsymbol{\delta}') \\
&\quad + (\boldsymbol{\delta}-\boldsymbol{\delta}')' E(\mathbf{x}-\mathbf{y}) + (\boldsymbol{\delta}-\boldsymbol{\delta}')'(\boldsymbol{\delta}-\boldsymbol{\delta}') \\
&= (\boldsymbol{\mu}-\boldsymbol{\mu}')'(\boldsymbol{\mu}-\boldsymbol{\mu}') + tr\left[Var(\mathbf{x}-\mathbf{y})\right] + \\
&\quad 2(\boldsymbol{\mu}-\boldsymbol{\mu}')'(\boldsymbol{\delta}-\boldsymbol{\delta}') + (\boldsymbol{\delta}-\boldsymbol{\delta}')'(\boldsymbol{\delta}-\boldsymbol{\delta}') \\
\text{IF } \boldsymbol{\mu} &= \boldsymbol{\mu}' \\
E\left[d^2(\mathbf{x}+\boldsymbol{\delta},\mathbf{y}+\boldsymbol{\delta}')\right] &= tr\left[Var(\mathbf{x}-\mathbf{y})\right] + (\boldsymbol{\delta}-\boldsymbol{\delta}')'(\boldsymbol{\delta}-\boldsymbol{\delta}') \\
&\quad \text{COVARIANCE (CORRELATION)} \\
&\quad \text{DOES NOT PICK UP BIAS}
\end{aligned}
$$

CORRELATION BETWEEN PREDICTOR AND PREDICTAND (“accuracy”) DOES NOT TELL THE WHOLE STORY.

MISLEADING HUNG-UP BASED ON FALSE BELIEF OF MODEL CORRECTNESS!

simplystats ABOUT CONFERENCE COURSES INTERVIEWS

12 AUG

Correlation is not a measure of reproducibility

POSTED BY RAFAEL IRIZARRY / UNCATEGORIZED

Professor of Biostatistics,
T.H. Chan School of Public Health
Harvard University

Suppose you have collected data from an experiment

$$x_1, x_2, \ldots, x_n$$

and want to determine if a second experiment replicates these findings

$$y_1=x_1+d_1,\ y_2=x_2+d_2,\ \ldots\ y_n=x_n+d_n.$$

For us to claim reproducibility we want the differences

to be as small as possible $d_1=y_1-x_1,\ d_2=y_2-x_2,\ldots\ ,d_n=y_n-x_n$

La-la Land

But aren't correlations and distances directly related? Sort of, and this actually brings up another problem. If the *x* and *y* are standardized to have average 0 and standard deviation 1 then, yes, correlation and distance are directly related:

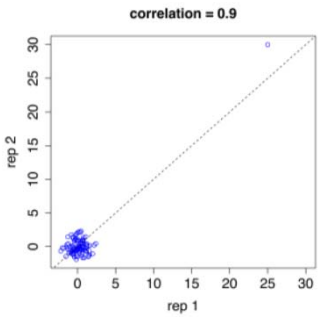
$$\frac{1}{2n}\mathrm{dist}(x,y)^2 = 1 - \mathrm{cor}(x,y)$$

Hacksaw ridge

However, if instead *x* and *y* have different average values, which would put into question reproducibility, then distance is sensitive to this problem while correlation is not. If the standard devtiation is 1, the formula is:

$$\frac{1}{2n}\mathrm{dist}(x,y)^2 = 1 + \frac{1}{2}\{\mathrm{avg}(y) - \mathrm{avg}(x)\}^2 - \mathrm{cor}(x,y)$$

Add one point to uncorrelated data: 0.9→



1) Sampling over infinite number of test sets, conditionally on training set and genotypes

$$
\begin{aligned}
E\left(PMSE|\mathbf{y}_{train},\mathbf{X}_{train},\mathbf{X}_{test}\right) &= \frac{1}{n_{test}}\left\{\left(\boldsymbol{\mu}_{test}-\mathbf{X}_{test}\widehat{\boldsymbol{\beta}}\right)'\left(\boldsymbol{\mu}_{test}-\mathbf{X}_{test}\widehat{\boldsymbol{\beta}}\right)+tr\left[Var\left(\mathbf{y}_{test}\right)\right]\right\} \\
&= \frac{1}{n_{test}}\left[\left(\boldsymbol{\mu}_{test}-\mathbf{X}_{test}\widehat{\boldsymbol{\beta}}\right)'\left(\boldsymbol{\mu}_{test}-\mathbf{X}_{test}\widehat{\boldsymbol{\beta}}\right)+n_{test}\sigma_e^2\right],
\end{aligned}
$$

2) Sampling over infinite number of test and train sets, conditionally on genotypes

$$
E\left(PMSE|\mathbf{X}_{train},\mathbf{X}_{test}\right)=\frac{1}{n_{test}}\left\{\boldsymbol{\delta}'\boldsymbol{\delta}+\sigma_e^2 tr\left[\mathbf{H}_{test,train}\right]+n_{test}\sigma_e^2\right\}.
$$

Pred. bias

$$
\begin{aligned}
\boldsymbol{\delta} &= \boldsymbol{\mu}_{test}-\mathbf{X}_{test}E\left(\widehat{\boldsymbol{\beta}}\right)=\boldsymbol{\mu}_{test}-\mathbf{X}_{test}\left(\mathbf{X}'_{train}\mathbf{X}_{train}\right)^{-1}\mathbf{X}'_{train}\boldsymbol{\mu}_{train} \\
&= \boldsymbol{\mu}_{test}-\mathbf{H}_{test,train}\boldsymbol{\mu}_{train}
\end{aligned}
$$

Var-bias trade off

It is unreasonable to dismiss prediction bias in more general settings because $\boldsymbol{\mu}_{test}\neq\mathbf{X}_{test}\boldsymbol{\beta}$ and $\boldsymbol{\mu}_{train}\neq\mathbf{X}_{train}\boldsymbol{\beta}$. Suppose now that the model is "wrong", that $n_{train}=n_{test}$, and that $\mathbf{X}_{test}=\mathbf{X}_{train}=\mathbf{X}$. In such a situation $\mathbf{H}_{test,train}=\mathbf{H}_{train,train}$, and (20) can be written as

$$
E\left(PMSE|\boldsymbol{X}\right)=\frac{1}{n_{train}}\sum_{i=1}^{n_{train}}\delta_i^2+\left(1+\frac{p}{n_{train}}\right)\sigma_e^2, \qquad (22)
$$

2) Sampling over infinite number of test and train sets, AND genotypes

Observe that (22) gives the expected mean-squared error of prediction, conditionally on the realized values of $\mathbf{X}$. However, in genome-enabled prediction matrix $\mathbf{X}$ has some distribution F that reflects linkage or linkage disequilibrium relationships (creating correlations among columns) as well as how genotypes are distributed in the target population, for example, a Hardy-Weinberg distribution. If the prediction model is to be applied repeatedly to a population, random variation of $\mathbf{X}$ must be accommodated. The fully unconditional predictive mean-squared error is then

$$\begin{aligned} E(PMSE) &= \frac{1}{n_{train}} E\left[\sum_{i=1}^{n_{train}} \delta_i^2\right] + \left(1 + \frac{p}{n_{train}}\right)\sigma_e^2 \\ &= \frac{1}{n_{train}} E\left[(\boldsymbol{\mu}_{test} - \mathbf{H}\boldsymbol{\mu}_{train})'(\boldsymbol{\mu}_{test} - \mathbf{H}\boldsymbol{\mu}_{train})\right] \\ &\quad + \left(1 + \frac{p}{n_{train}}\right)\sigma_e^2. \end{aligned} \tag{23}$$

is $\mathbf{H} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$. Letting $E(\mathbf{H}) = \overline{\mathbf{H}}$

$$E(PMSE) = \left\{\frac{1}{n_{train}}(\boldsymbol{\mu}_{test} - \overline{\mathbf{H}}\boldsymbol{\mu}_{train})'(\boldsymbol{\mu}_{test} - \overline{\mathbf{H}}\boldsymbol{\mu}_{train}) + tr\left[Var(\mathbf{H}\boldsymbol{\mu}_{train})\right]\right\} + \left(1 + \frac{p}{n_{train}}\right)\sigma_e^2. \tag{24}$$

The preceding implies that the contribution of bias (first part of the expression above) is a function of the unknown population means and of the distribution of genotypes in the population. Perhaps an elaborate model can paliate the adverse impact of bias on predictive performance, but the second part of the expression indicates that a highly parameterized model will produce predictions with larger variance than a "smaller" model. The upper limit of p is n_{train} (otherwise, the OLS estimator would not be unique), so the prediction error variance can almost double the residual variance in a model with many parameters. Unfortunately, the impact of model complexity on prediction bias is impossible to quantify in the absence of mechanistic knowledge.

LEAST-SQUARES PREDICTION: AN EXAMPLE

A population of $n = 100,000$ individuals was simulated employing the `R-script Prediction 1` given at the end of the chapter (computations are slow for larger populations). The main steps in the program are as follows.

- Genotypes for 40 marker loci and 4 QTL that were linear combinations of marker genotypes were created. Markers intervening in QTL are the equivalent of "SNP in genes".
- Expected frequencies in the population were $\Pr(aa) = 0.15$, $\Pr(Aa) = 0.25$ and $\Pr(AA) = 0.60$ at any of the 40 marker loci. Markers were assumed to be placed contiguously over a single chromosome. Decay in linkage disequilibrium was mimicked by combined use of two "tricks": a) an exponential decay function (called "decay" in the script) gradually reduced frequencies of Aa and AA genotypes as markers became more distant from the first marker. b) For every individual, a random deviate from a $Dirichlet\,(0.15, 0.25, 0.60)$ distribution was drawn. The decay function was applied to each marker locus, probabilities were normalized (summing to 1) and multinomial sampling with the new probabilities was applied to generate genotypes. Since the same Dirichlet distribution was used for each column of the marker matrix $\mathbf{X}$, multinomial sampling over $0:2$ produced correlations among its 40 columns, creating an LD structure (correlations among columns of matrix $\mathbf{X}$) that resembled the decay often observed as members of a pair of markers are further apart.
- The four QTL genotypes were created by taking averages of 4 pairs of columns of $\mathbf{X}$: (3,37), (6,34), (9,31), (12,28) and then rounding such that QTL genotypes were in $0:2$ as well. Since marker genotypes were in LD, QTL genotypes were in LD as well.
- Phenotypes were simulated for each of the 100,000 individuals (i denotes a generic individual and $Q_1 - Q_4$ are QTL genotypes) as

$$\begin{aligned} y_i &= add_i + epi_i + e_i, \\ add_i &= \frac{1}{2}Q_{i1} - Q_{i2} + \frac{1}{8}Q_{i3} + \frac{1}{2}Q_{i4} \\ epi_i &= -Q_{i1}Q_{i2}, \\ e_i &\sim N(0,1)\,. \end{aligned} \tag{25}$$

Hence, gene action involved both additive and inter-locus interaction (epistasis) effects, denoted in their aggregate as add and epi, respectively. The partial regressions of phenotypes on genotypes at QTL 1-4 $(Q_{i1}, Q_{i2}, Q_{i3}, Q_{i4})$ were $\frac{1}{2}, -1, \frac{1}{8}$,and $\frac{1}{2}$, respectively; the partial regression (epistasis) on co-genotype $Q_{i1}Q_{i2}$ was -1. Enviromental effects (e_i) were created by randomly and independently sampling from a $N(0,1)$ distribution.

```
#FIT TRUE MODEL
regtrue<-lm(y~Qqtl[,1]+Qqtl[,2]+Qqtl[,3]+Qqtl[,4]+Qqtl[,1]*Qqtl[,2])
summary(regtrue)
Coefficients:
 Estimate Std.  Error t value Pr(>|t|)
(Intercept) 9.222e-05 4.668e-03 0.02 0.984
Qqtl[, 1] 5.058e-01 8.104e-03 62.41 <2e-16 ***
Qqtl[, 2] -9.973e-01 1.345e-02 -74.17 <2e-16 ***
Qqtl[, 3] 1.031e-01 1.063e-02 9.70 <2e-16 ***
Qqtl[, 4] 5.155e-01 1.075e-02 47.95 <2e-16 ***
Qqtl[, 1]:Qqtl[, 2] -9.995e-01 1.451e-02 -68.86 <2e-16 ***
---
Signif.  codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.'  0.1 ' ' 1
Residual standard error:  0.9974 on 99994 degrees of freedom
Multiple R-squared:  0.3476, Adjusted R-squared:  0.3476
F-statistic:  1.066e+04 on 5 and 99994 DF, p-value:  < 2.2e-16
```

```
#FIT ADDITIVE MODEL
regadd<-lm(y~Qqtl[,1]+Qqtl[,2]+Qqtl[,3]+Qqtl[,4])
summary(regadd)

lm(formula = y ~Qqtl[, 1] + Qqtl[, 2] + Qqtl[, 3] + Qqtl[, 4])
Coefficients:
 Estimate Std.  Error t value Pr(>|t|)
(Intercept) 0.091849 0.004579 20.059 < 2e-16 ***
Qqtl[, 1] 0.262856 0.007467 35.204 < 2e-16 ***
Qqtl[, 2] -1.701595 0.008934 -190.466 < 2e-16 ***
Qqtl[, 3] 0.040395 0.010841 3.726 0.000195 ***
Qqtl[, 4] 0.438468 0.010942 40.071 < 2e-16 ***
---
Signif.  codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.'  0.1 ' ' 1
Residual standard error:  1.021 on 99995 degrees of freedom
Multiple R-squared:  0.3167, Adjusted R-squared:  0.3167
F-statistic:  1.159e+04 on 4 and 99995 DF, p-value:  < 2.2e-16
```

Using a simulation similar to the one described above (see `R-script Prediction 2` at the end of the chapter) we drew random samples of sizes $n = 100$ through $n = 800$ (8 sizes in increments of 100), to examine the effect of n on accuracy and variability of predictions resulting from models with varying number of markers ($p = 1, 6, 12, 24, 36$), to study impacts of model complexity.

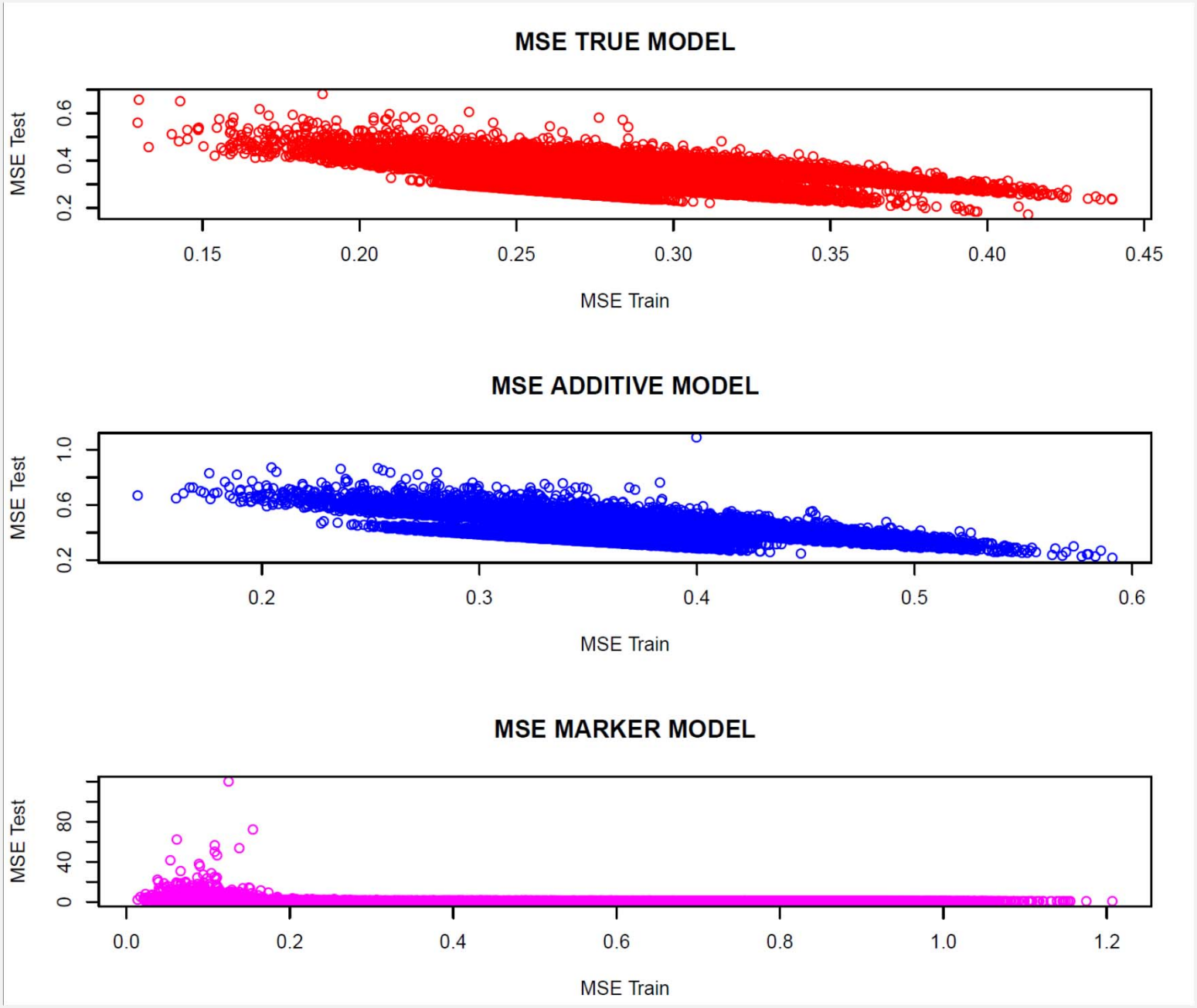
Samples were divided at random into training and testing sets of equal size; e.g., when $n = 800$, the training and testing sets had 400 individuals each. For each of $8 \times 5 = 40$ size-complexity scenarios, 1000 replications were made by randomly allocating individuals into disjoint training and testing sets. Gene action was epistatic as per the simulation described above; decay in LD was not simulated because it often led to singular $\mathbf{X}$ matrices, especialy for the more parameterized models and smaller sample sizes. "Baseline" genotypic frequencies used were $0.25, 0.50$ and 0.25 for aa, Aa and AA individuals at each locus.

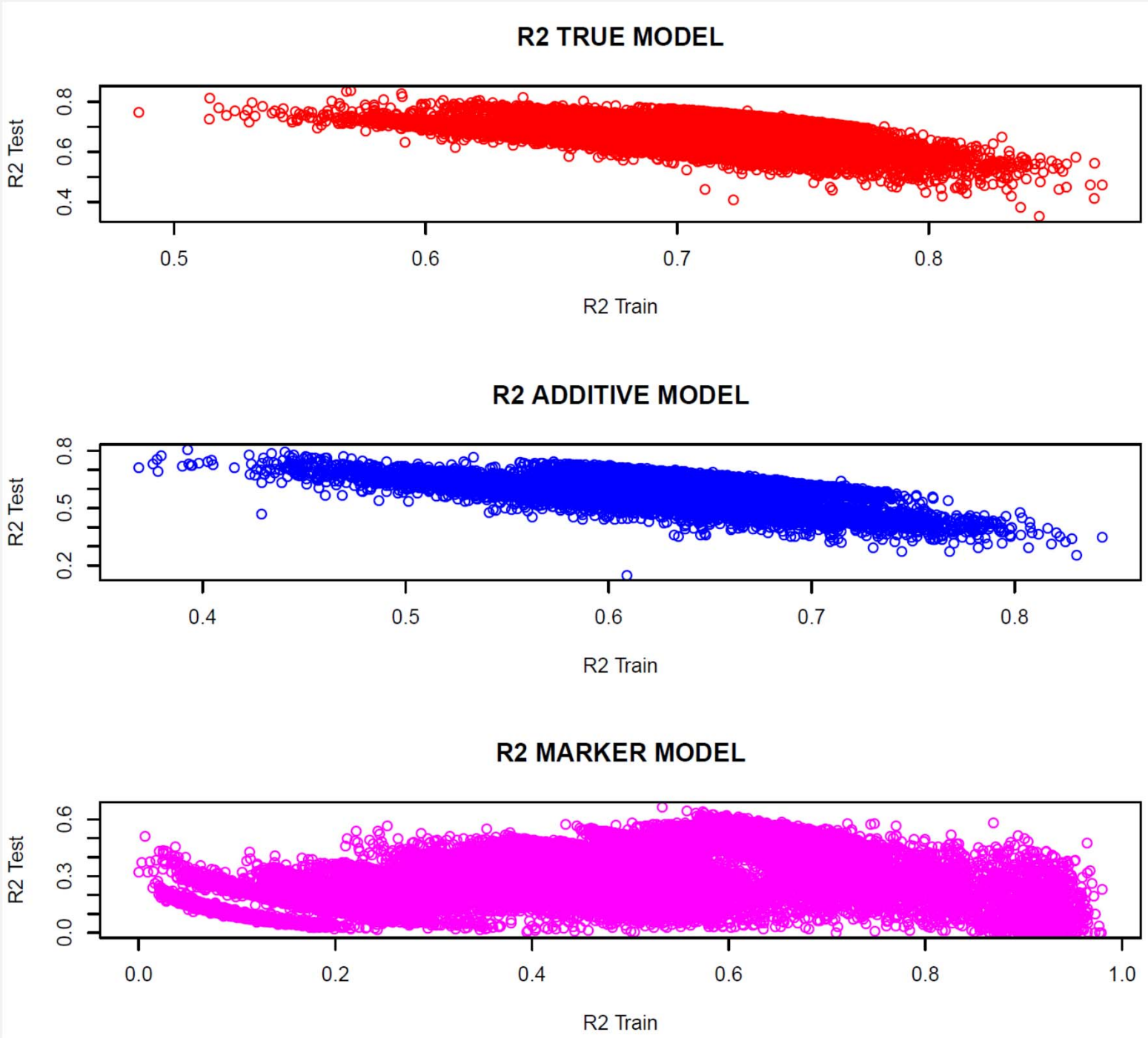
In each of the 40,000 runs the models fitted were: 1) true model using the 4 known QTL genotypes and including the epistatic term; 2) model with additive effects of the QTL; 3) marker based models with varying degrees of complexity. For example, for $p = 36$, markers were sampled by drawing at random (without replacement) 36 columns from $\mathbf{X}$. Metrics calculated were R^2 and MSE (mean-squared error) in training and testing sets using the three models fitted.

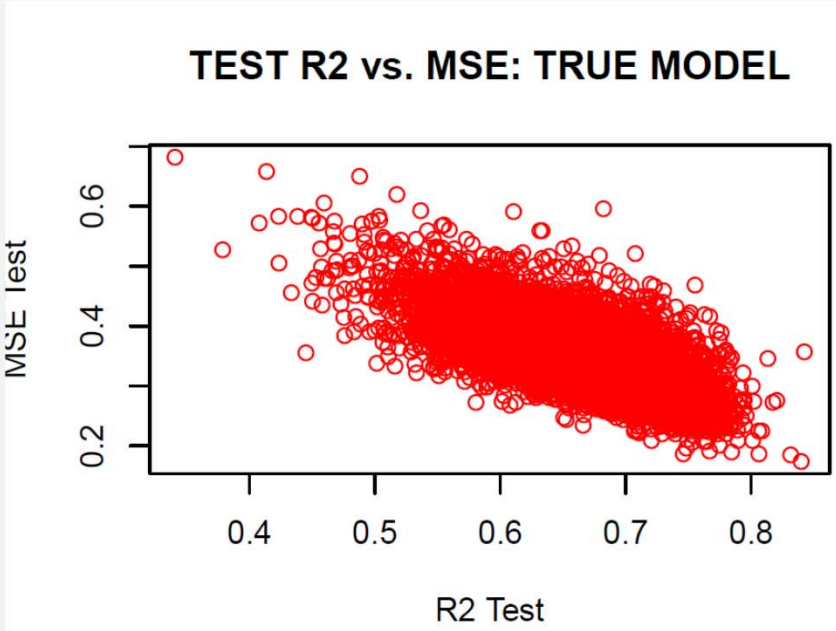
Overview of predictive performance The ranges of the R^2 metric in training and testing sets over the 40,000 runs are below.

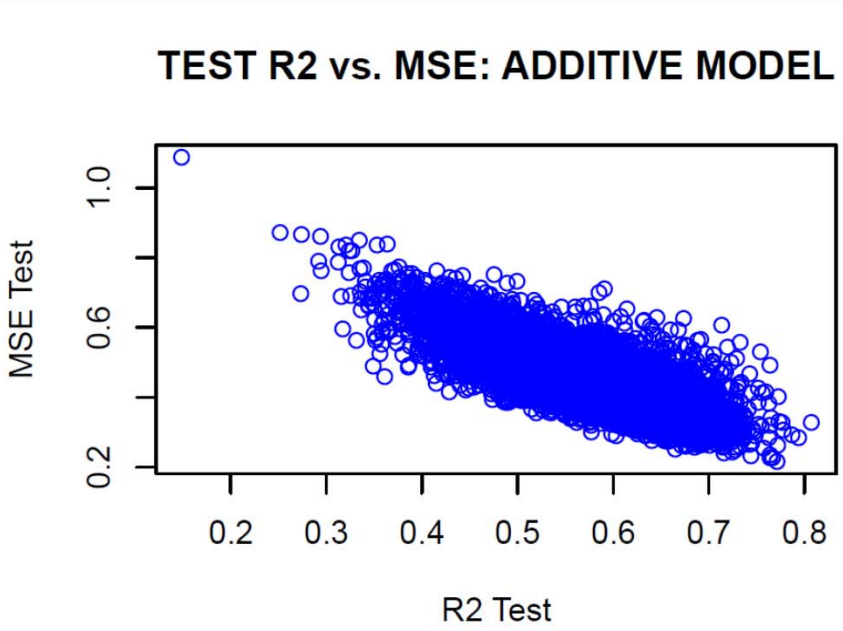
```
range(R2traintrue)
#[1] 0.4859632 0.8690036
range(R2trainadd)
#[1] 0.3685519 0.8431821
range(R2trainmark)
#[1] 0.0000425017 0.9799664439
##RANGE TEST
range(R2testtrue)
#[1] 0.3405665 0.8426000
range(R2testadd)
#[1] 0.1484958 0.8070288
range(R2testmark)
#[1] 3.074384e-05 6.643009e-01
```

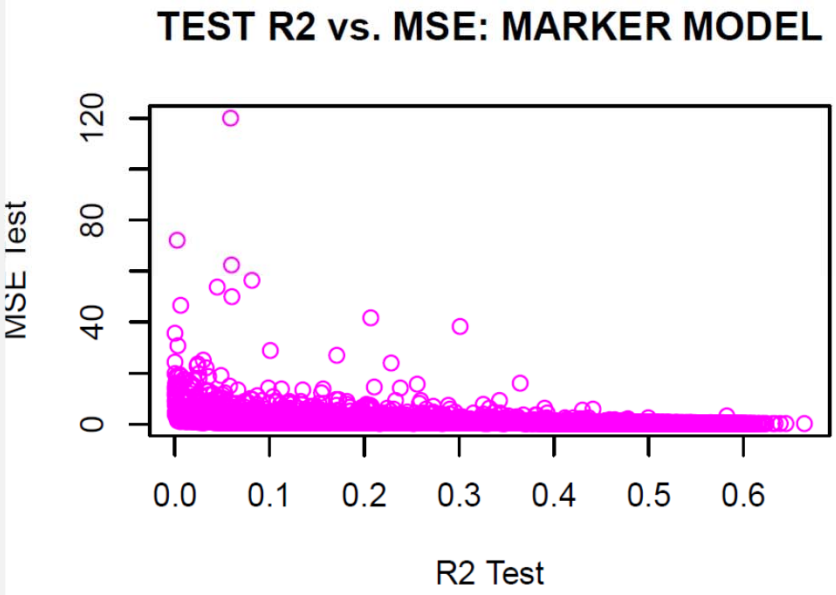
OVERALL PATTERNS

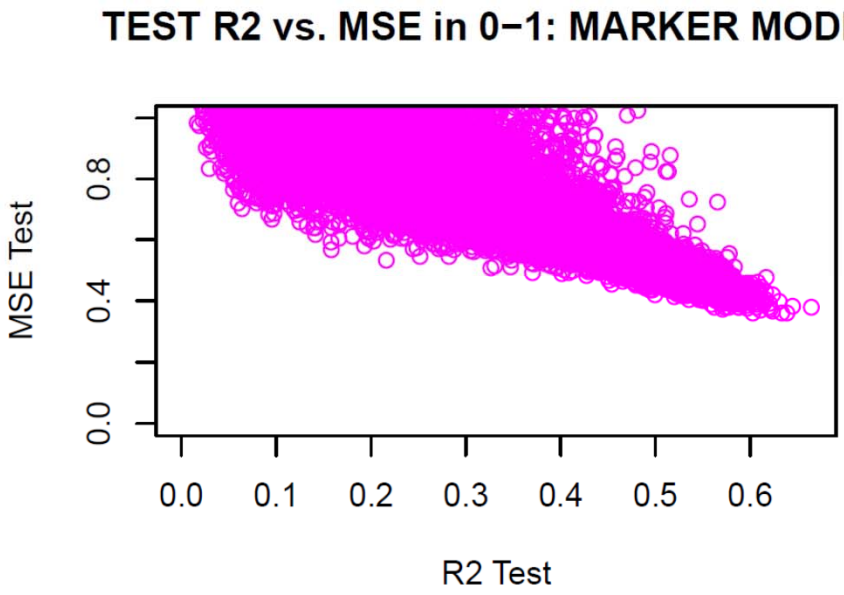



R2 TRUE MODEL
R2 Test
0.4
0.6
0.8
0.5
0.6
0.7
0.8
R2 Train
R2 ADDITIVE MODEL
R2 Test
0.2
0.5
0.8
0.4
0.5
0.6
0.7
0.8
R2 Train
R2 MARKER MODEL
R2 Test
0.0
0.3
0.6
0.0
0.2
0.4
0.6
0.8
1.0
R2 Train


TEST R2 vs. MSE: TRUE MODEL
MSE Test
0.2
0.4
0.6
0.4
0.5
0.6
0.7
0.8
R2 Test

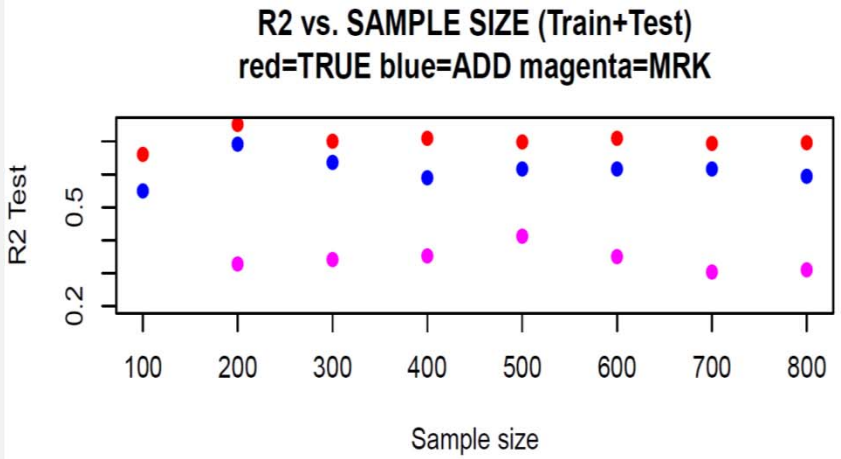

TEST R2 vs. MSE: ADDITIVE MODEL
MSE Test
0.2
0.6
1.0
0.2
0.3
0.4
0.5
0.6
0.7
0.8
R2 Test

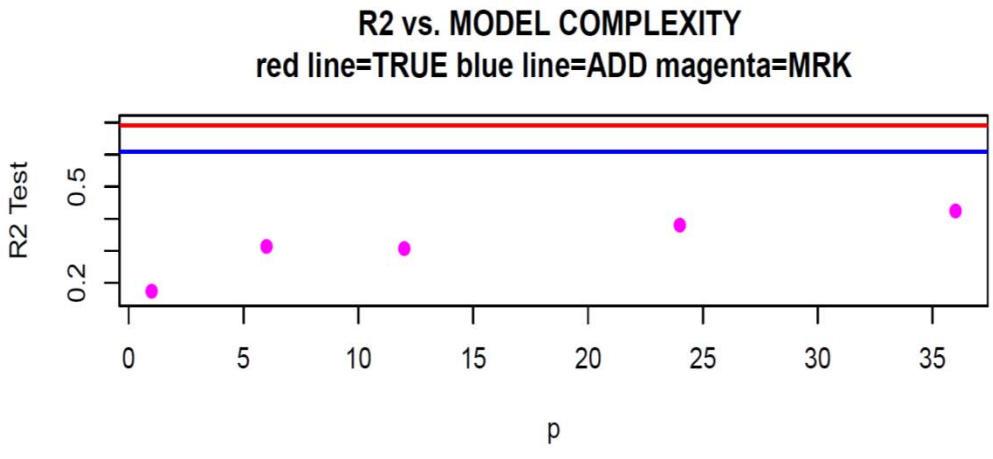

TEST R2 vs. MSE: MARKER MODEL
MSE Test
0
40
80
120
0.0
0.1
0.2
0.3
0.4
0.5
0.6
R2 Test

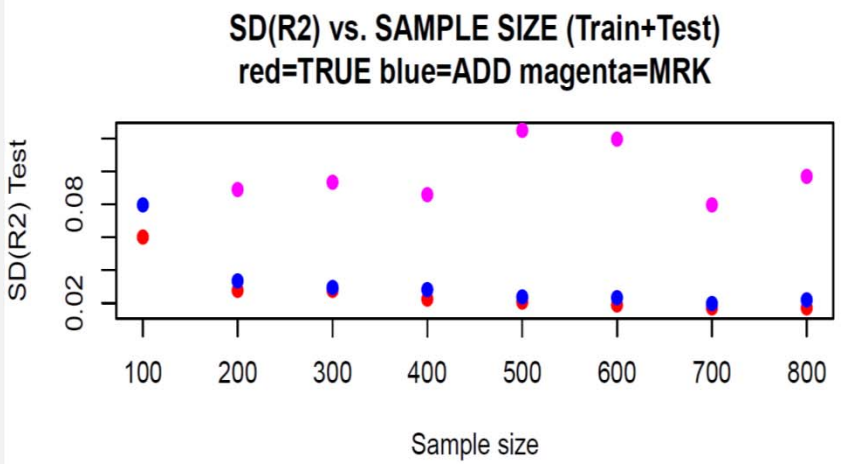

TEST R2 vs. MSE in 0−1: MARKER MODEL
MSE Test
0.0
0.4
0.8
0.0
0.1
0.2
0.3
0.4
0.5
0.6
R2 Test

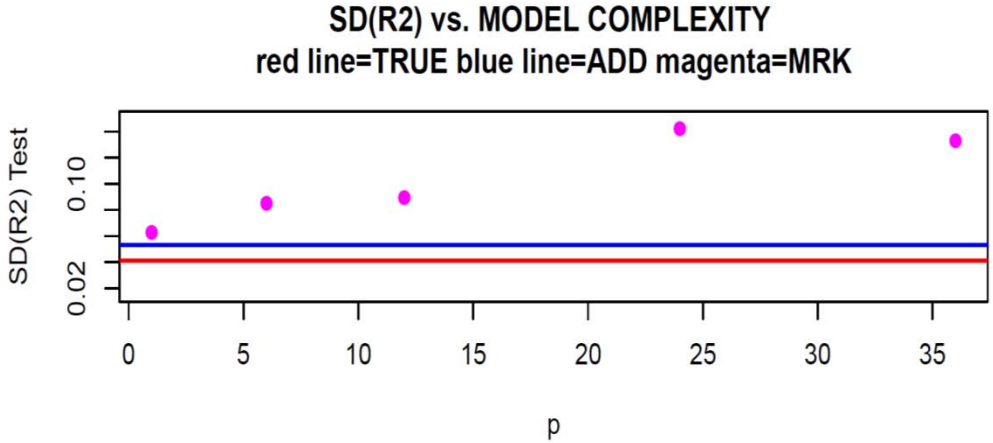
EFFECT OF SAMPLE SIZE

EFFECT OF MODEL COMPLEXITY

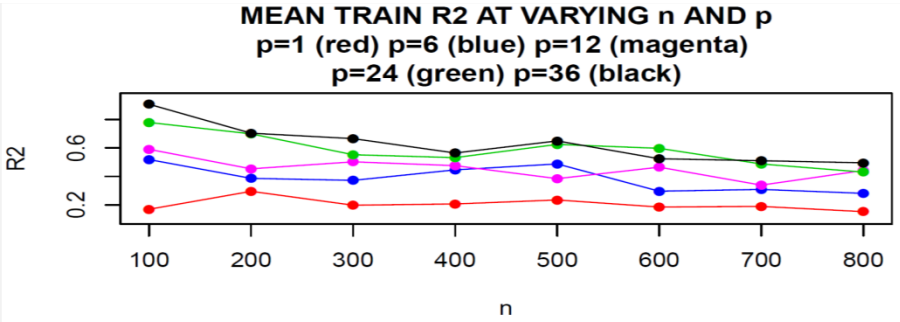


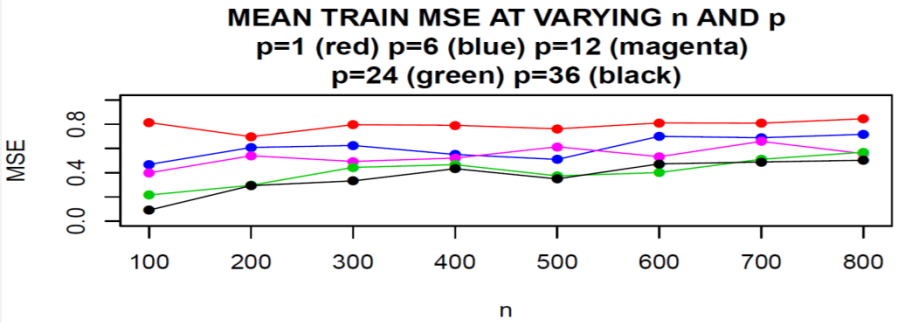


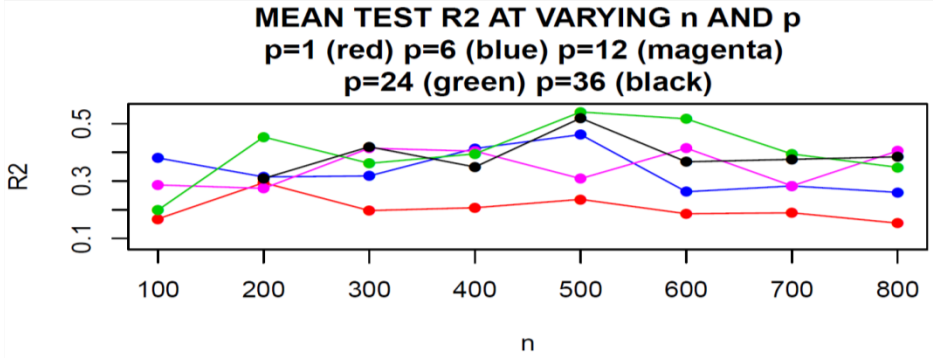


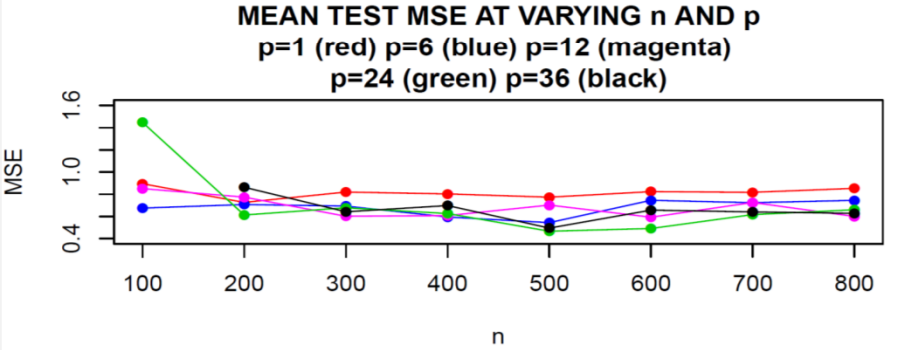


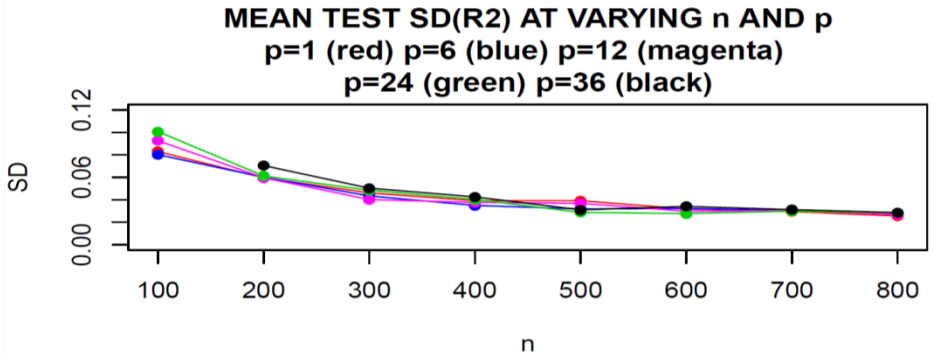
INTERPLAY BETWEEN SAMPLE SIZE AND MODEL COMPLEXITY

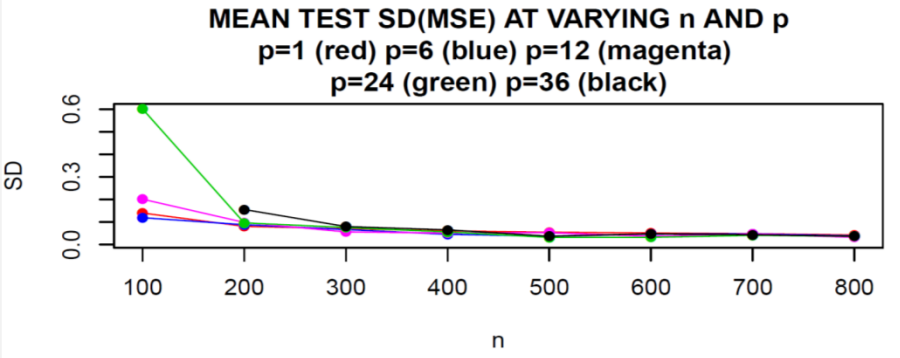












CROSS-VALIDATION WITH LEAST-SQUARES

LEAVE-ONE OUT CV

the training folds. Let $\mathbf{X}_{[-i]}$ be $\mathbf{X}$ with its i^{th} row $(\mathbf{x}_i')$ removed, so that its order is $(n-1)\times p$. Since $\mathbf{X}'\mathbf{X}=\sum_{i=1}^{n}\mathbf{x}_i\mathbf{x}_i'$,

$$\mathbf{X}_{[-i]}'\mathbf{X}_{[-i]}=\mathbf{X}'\mathbf{X}-\mathbf{x}_i\mathbf{x}_i';\ i=1,2,...,n, \tag{29}$$

$$\mathbf{X}_{[-i]}'\mathbf{y}_{[-i]}=\sum_{i=1}^{n}\mathbf{x}_i\mathbf{y}-\mathbf{x}_iy_i=\mathbf{X}'\mathbf{y}-\mathbf{x}_iy_i$$

$$\widehat{\boldsymbol{\beta}}_{[-i]}=\widehat{\boldsymbol{\beta}}-\frac{(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_i\widehat{e}_i}{1-h_{ii}},$$

$$h_{ii}=\mathbf{x}_i'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_i \text{ and } \widehat{e}_i=y_i-\mathbf{x}_i'\widehat{\boldsymbol{\beta}}$$

$$r_i=y_i-\mathbf{x}_i'\widehat{\boldsymbol{\beta}}_{[-i]}=\frac{y_i-\mathbf{x}_i'\widehat{\boldsymbol{\beta}}}{1-h_{ii}}.$$

$$MSE(1)=\frac{1}{n}\sum_{i=1}^{n}\left(\frac{y_i-\mathbf{x}_i'\widehat{\boldsymbol{\beta}}}{1-h_{ii}}\right)^2$$

A simple way of obtaining an estimate of the uncertainty of $MSE(1)$ is by using the bootstrap (Efron and Tibshirani 1993). Let the LOO prediction errors be denoted as $\mathbf{r} = (r_1, r_2, ..., r_n)'$ where r_i is as in (33). A bootstrap sample b $(b = 1, 2, ..., B)$ is obtained by sampling (with replacement) n elements of $\mathbf{r}$, so that some may appear more than once and some may not appear at all. The prediction errors in sample b are $\mathbf{r}_b = (r^*_{b1}, r^*_{b2}, ..., r^*_{bn})$ and the corresponding prediction mean squared error is $MSE_b(1) = \frac{1}{n}\sum_{i=1}^{n} r^{*2}_{bi}$. Then $MSE_b(1), b = 1, 2, ..., B$ can be used to estimate the distribution of LOO mean-squared error of prediction, or of LOO R^2. For the latter metric, one can bootstrap over pairs represented as $\mathbf{p}_i = \left(y_i, \mathbf{x}_i'\widehat{\boldsymbol{\beta}}_{[-i]}\right), i = 1, 2, ..., n$, with a bootstrap sample being $\mathbf{p}_b$. The B samples can be used to estimated the distribution of the LOO R^2.

K-FOLD CV

Here, the data are divided into K groups or "folds" of about the same size. The model is trained using data in $K - 1$ partitions with the remaining fold used as testing set. The procedure is repeated for all K folds, producing, e.g., $MSE_1, MSE_2, ..., MSE_K$, and an overall estimate of predictive ability is

$$MSE_K = \frac{1}{\sum_{k=1}^{K} w_k} \sum_{k=1}^{K} w_k MSE_k, \tag{38}$$

where w_k is a weight reflecting fold size; if all folds have equal size, $MSE_K = \sum_{k=1}^{K} MSE_k/K$. If $K = n$, the procedure becomes LOOCV, so that $n - 1$ observations are used in each of n training instances. The $K - fold$ CV may be repeated several times by reconstructing folds at random, to obtaine estimates of the distributions of MSE_K or R^2_K.

EXAMPLES OF K-FOLD CV

We simulated data as before and carried out LOO CV and $K-fold$ cross-validation for samples sizes $n = 400, 600, 800$ and 1000 and K from 2 through 20 in increments of 2 folds. The following models were fitted by OLS: a) true model on known epistatic QTL; b) additive model on QTL; c) markers known to reside in QTL (the markers whose linear combinations of genotypes were used to create QTL), and d) a regression on 40 markers simultaneously. Computations were done with the R package "boot" versions 1.3.18. Details are in `R-script Prediction 4` (end of chapter).

RESULTS OF LOO CV

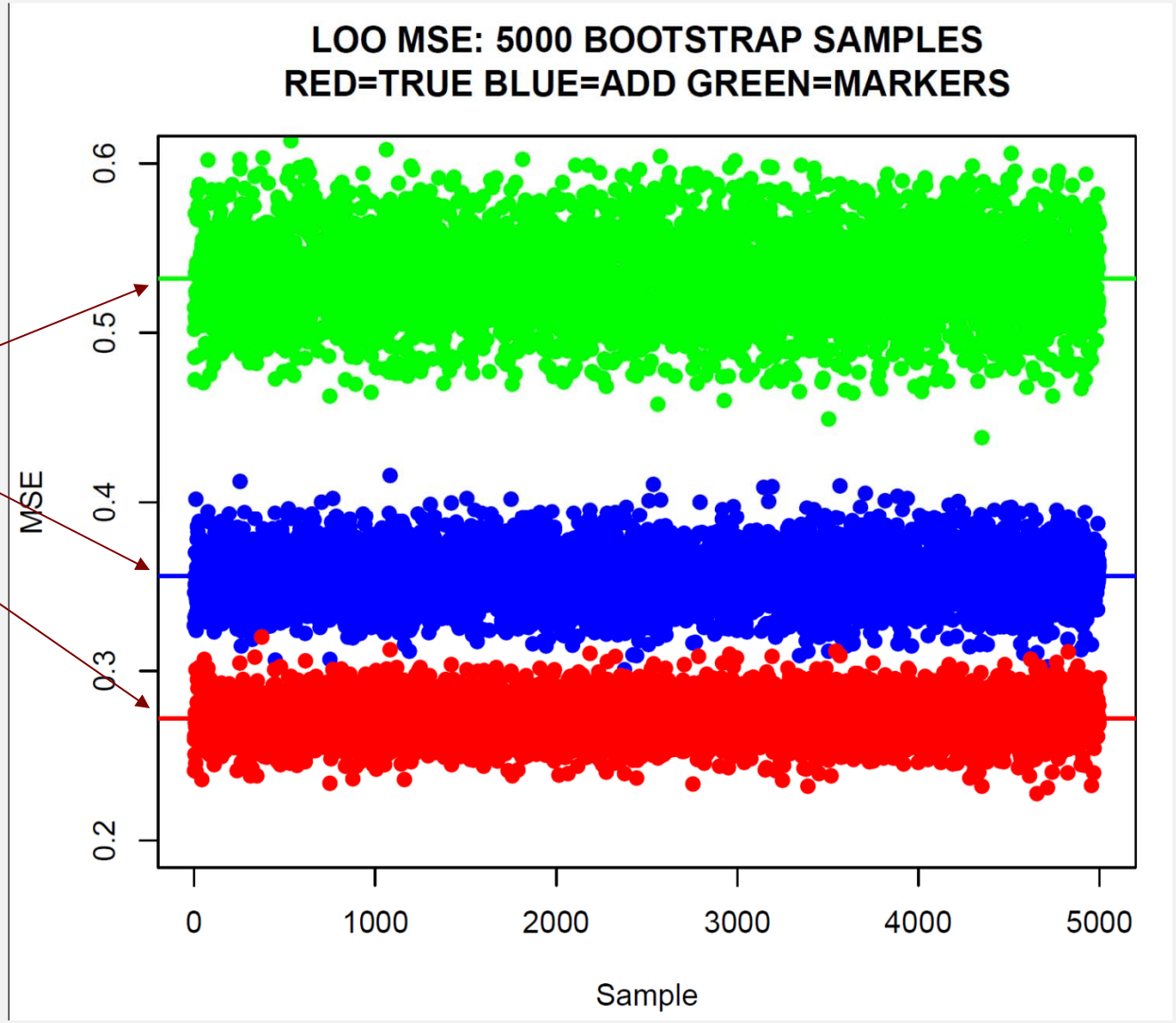
n= 400 600 800 1000

```
> mseloomarkqtl<-round(mseloomarkqtl,3)
> mseloomark<-round(mseloomark,3)
> mselootrue
[1] 0.295 0.303 0.283 0.304
> mselooadd
[1] 0.403 0.397 0.368 0.401
> mseloomarkqtl
[1] 0.466 0.470 0.425 0.461
> mseloomark
[1] 0.505 0.500 0.441 0.470
```

RESULTS OF K-FOLD CV

```
                K= 2      4      6      8     10     12     14     16     18     20
> msecvtrue
n
400  [1,] 0.294 0.291 0.296 0.296 0.295 0.298 0.297 0.295 0.294 0.296
600  [2,] 0.319 0.302 0.307 0.301 0.303 0.303 0.304 0.302 0.303 0.303
800  [3,] 0.283 0.281 0.286 0.284 0.282 0.283 0.284 0.282 0.283 0.283
1000 [4,] 0.307 0.302 0.304 0.305 0.305 0.303 0.304 0.304 0.303 0.303
> msecvadd
n
400  [1,] 0.402 0.409 0.403 0.401 0.403 0.402 0.401 0.402 0.403 0.403
600  [2,] 0.403 0.393 0.399 0.398 0.395 0.398 0.401 0.396 0.398 0.397
800  [3,] 0.367 0.369 0.368 0.367 0.368 0.369 0.368 0.368 0.368 0.367
1000 [4,] 0.402 0.405 0.402 0.402 0.398 0.400 0.401 0.400 0.401 0.400
> msecvmarkqtl
n
400  [1,] 0.459 0.474 0.466 0.467 0.464 0.467 0.472 0.467 0.467 0.462
600  [2,] 0.463 0.466 0.480 0.471 0.470 0.468 0.470 0.471 0.471 0.472
800  [3,] 0.431 0.425 0.431 0.425 0.426 0.424 0.424 0.426 0.424 0.426
1000 [4,] 0.455 0.469 0.463 0.462 0.464 0.462 0.463 0.460 0.461 0.462
> msecvmark
n
400  [1,] 0.559 0.502 0.523 0.513 0.526 0.505 0.518 0.501 0.499 0.509
600  [2,] 0.531 0.507 0.511 0.506 0.502 0.501 0.498 0.507 0.500 0.503
800  [3,] 0.464 0.452 0.455 0.442 0.447 0.447 0.442 0.445 0.443 0.450
1000 [4,] 0.492 0.477 0.472 0.478 0.469 0.470 0.477 0.469 0.469 0.466
```

p=200 markers fitted, n=1000



Analytical MSE

Variability of K-fold (K= 5, 10) CV: n=500, 1000, p=40
1000 replications of the K-fold CV