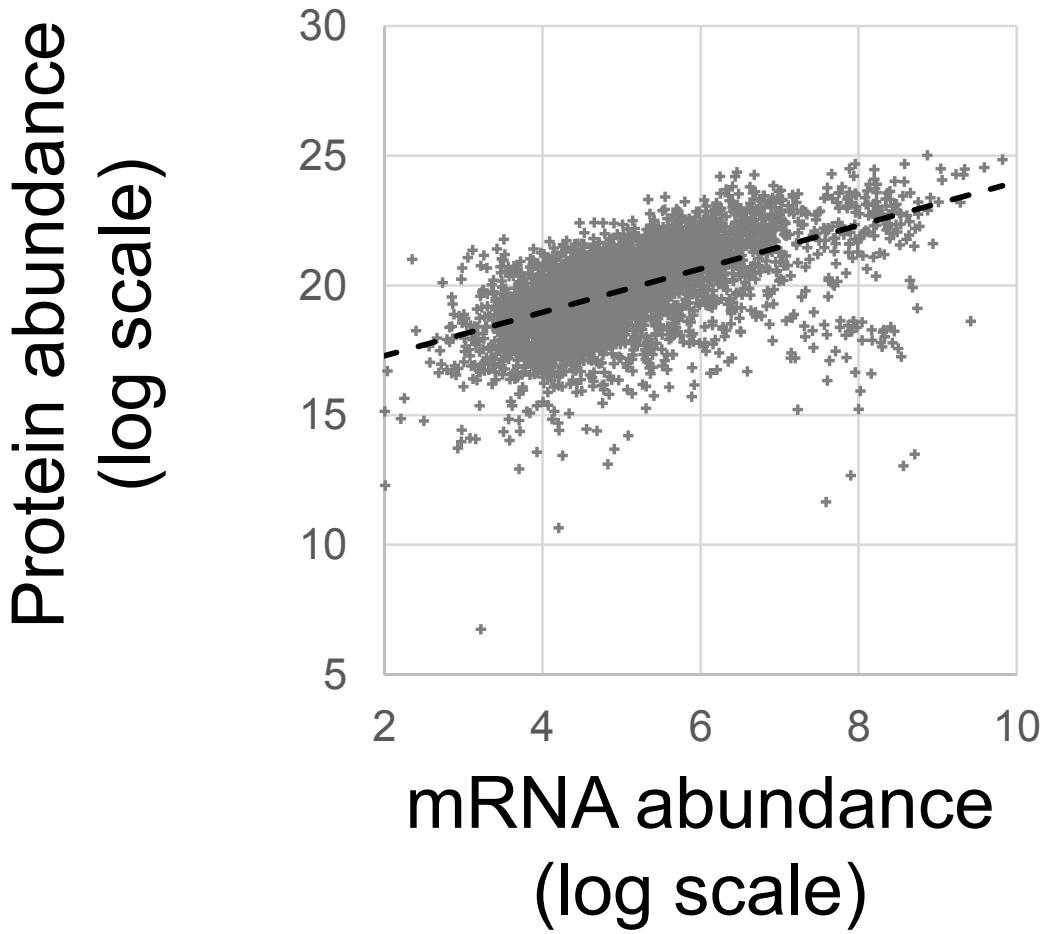
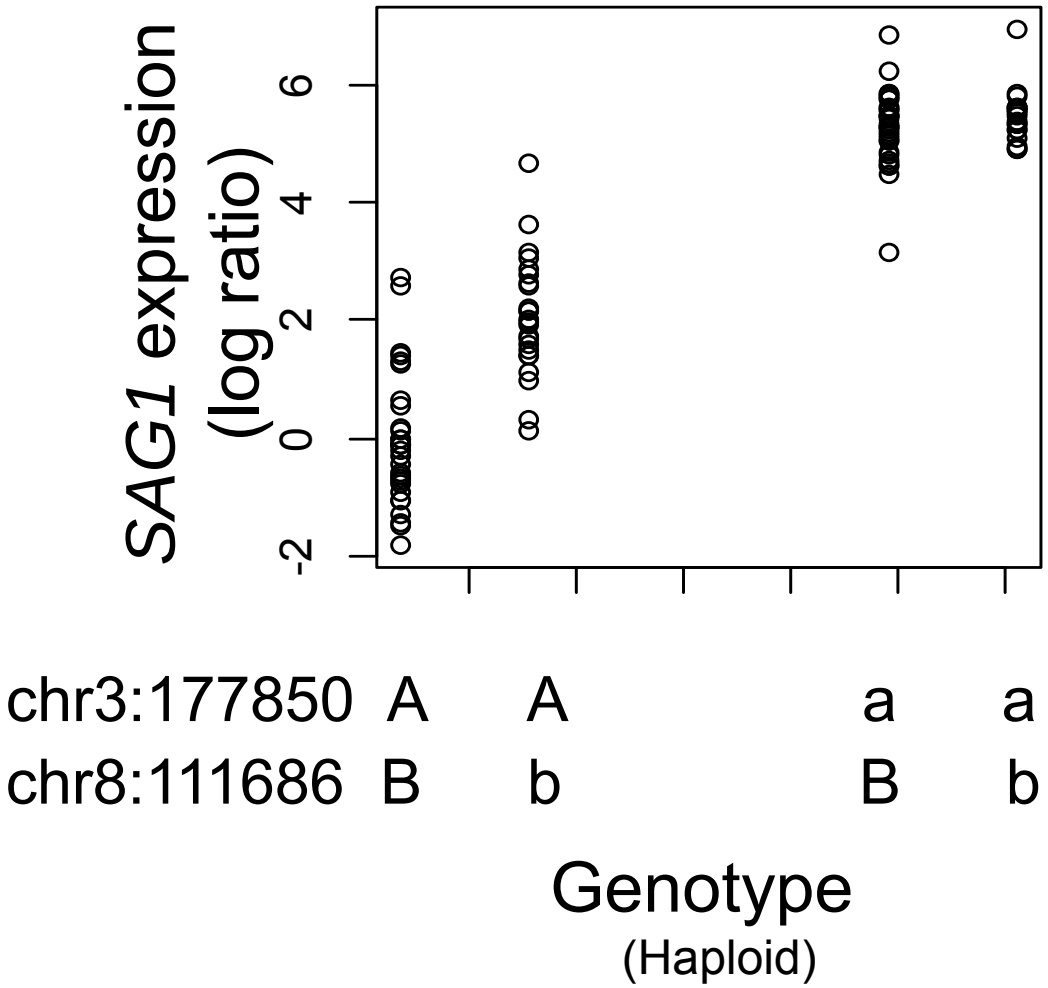
Regression

Alan Moses
ML4bio



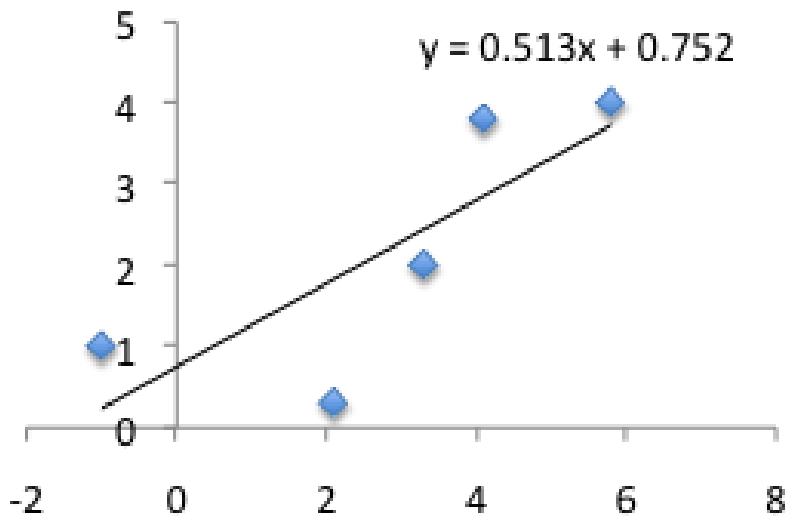
data collated in Csardi et al. *PLoS Genetics* 2015



eQTL data from Brem et al. *Science* 2002

Topics for today

- Univariate simple linear regression
- Local regression
- Multiple regression (with regularization!)
- Generalized linear models


y = 0.513x + 0.752
5
4
3
2
1
0
-2
0
2
4
6
8

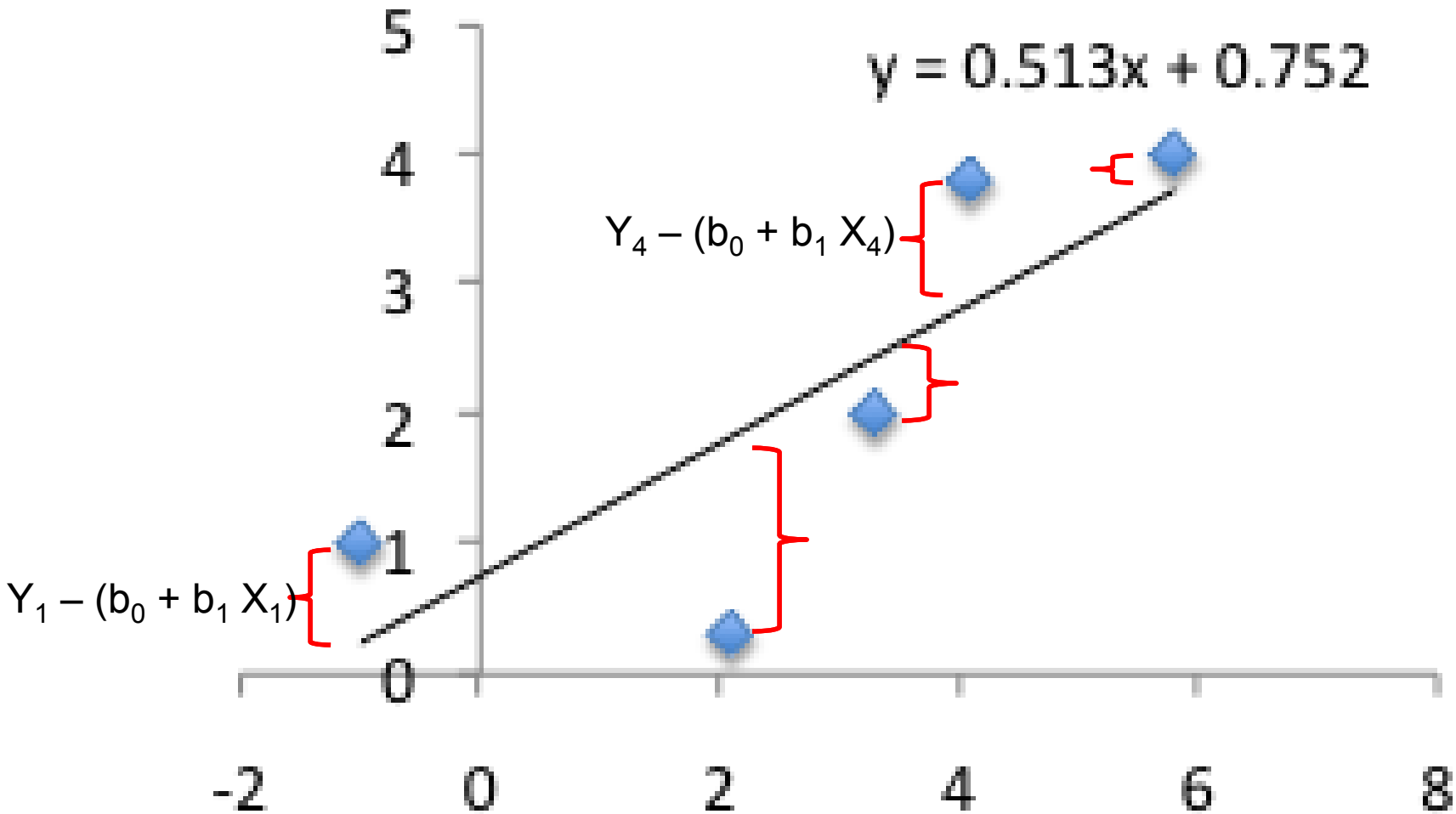
How to do this estimation ?

- The usual way to fit simple linear regressions is using the “ordinary least squares” (or OLS) method
- Predicted $\hat{Y} = b_0 + b_1 X$
- choose parameters to minimize sum of squared residuals cost function

$$SSR = \sum_i (Y_i - \hat{Y}_i)^2 = \sum_i [Y_i - (b_0 + b_1 X_i)]^2$$

- In the simple case, this can be done ***analytically***

What is the SSR ?



What's so great about simple linear regression?

- $Y = b_0 + b_1 X$
- Simple formulas for parameter estimates

$$b_1 = r \frac{s_Y}{s_X}$$

Where s are the standard deviations, and r is "Pearson's correlation coefficient"

$$r = \frac{E[\,(X - E[X])\,(Y - E[Y])\,]}{s_X s_Y}$$

What's so great about simple linear regression?

- $Y = b_0 + b_1 X$
- Simple formulas for parameter estimates

- Rigorous interpretation of model in terms of variance explained

$$r^2 = 1 - \frac{\sum_i [Y_i - (b_0 + b_1 X_i)]^2}{\sum_i (Y_i - E[Y])^2} = 1 - \frac{\sum_i [Y_i - \hat{Y}]^2}{\sum_i (Y_i - E[Y])^2}$$

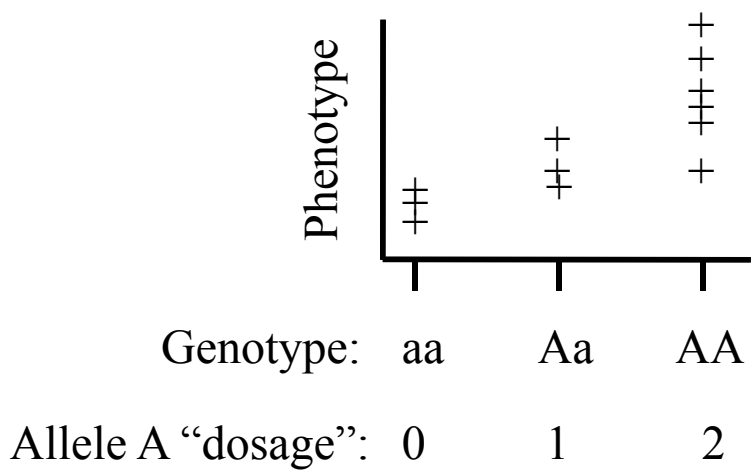
What's so great about simple linear regression?

- $Y = b_0 + b_1 X$
- Simple formulas for parameter estimates

- Rigorous interpretation of model in terms of variance explained
- Ideas like "variance explained" and "least squares" can be applied more generally

Using the correlation to test for association between two variables

- Pearson correlation is a powerful test statistic for association.
- E.g., used in statistical genetics

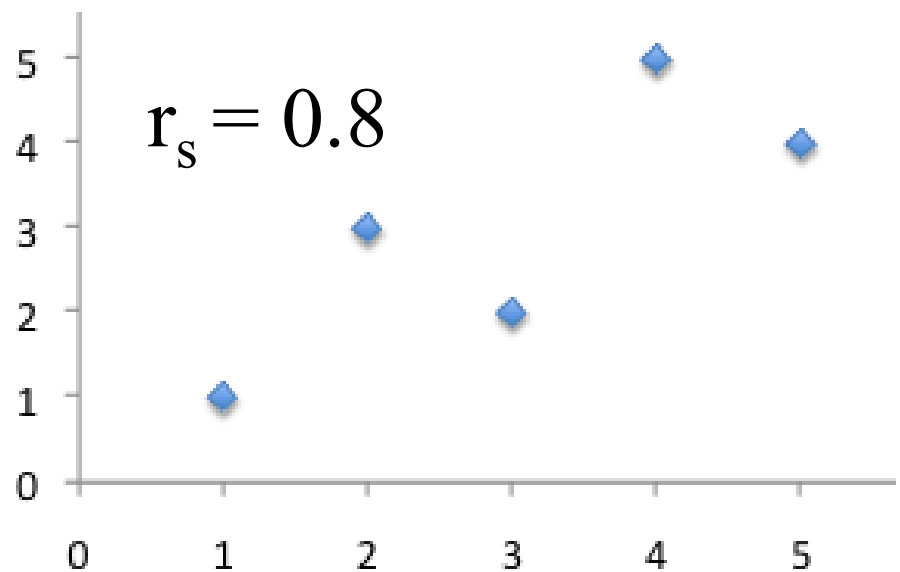


- What is the null hypothesis?

Using the correlation to test for association between two variables

- Non-parametric test for association is the correlation of the ranks, a.k.a Spearman's rank correlation

data	rank	data	rank
X_1	3	Y_1	2
X_2	4	Y_2	5
X_3	1	Y_3	1
X_4	2	Y_4	3
X_5	5	Y_5	4



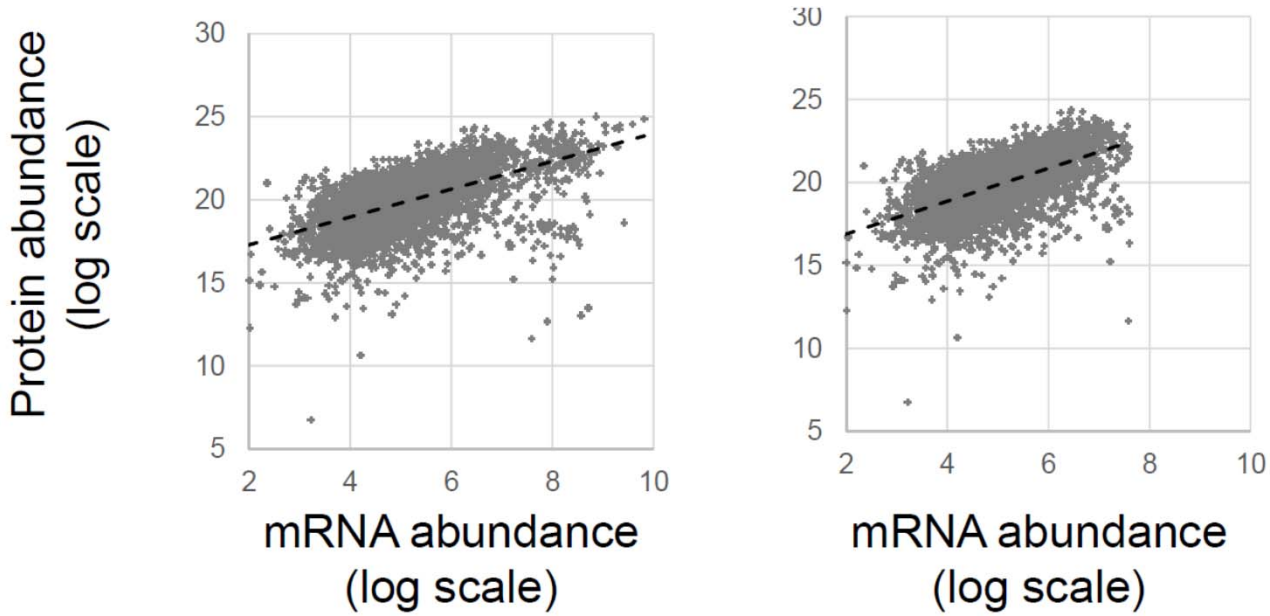
$$t = r\sqrt{\frac{n-2}{1-r^2}} = 2.3 \quad (P = 0.05)$$

df = n-2

Topics for today

- Univariate simple linear regression
- Local regression
- Multiple regression (with regularization!)
- Generalized linear models

What if the data deviate from a line?



data collated in Csardi et al. *PLoS Genetics* 2015

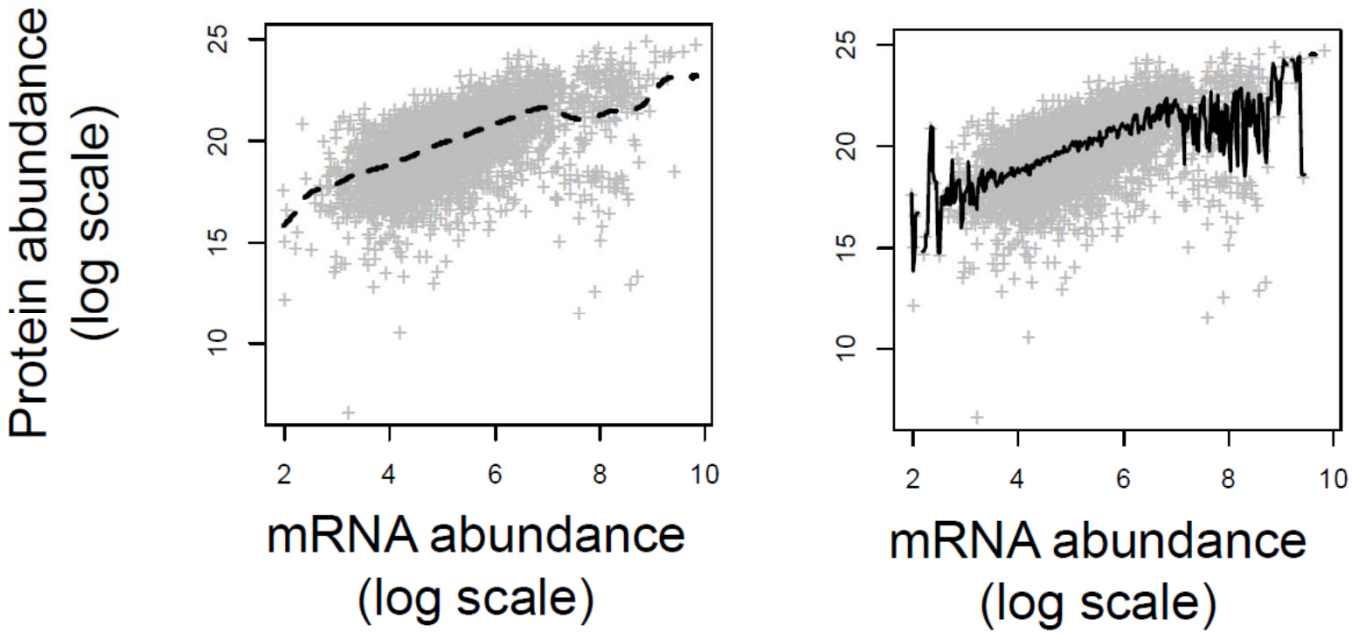
Kernel regression

- Predict Y to be a weighted average of nearby datapoints.

Kernel

$$\hat{Y}_{\text{at } X_0} = \frac{\sum_{i=1}^{n} K(|X_i - X_0|) Y_i}{\sum_{i=1}^{n} K(|X_i - X_0|)}$$

- Weights are a function of distance between nearby datapoint and point where you want the prediction
- This function is called the “Kernel”
- Kernel (usually) depends on a “bandwidth” (a hyperparameter) that determines the distance scale for the averaging



data collated in Csardi et al. *PLoS Genetics* 2015

Local regression and smoothing

- LOESS is a type of local regression that fits a polynomial (instead of a simple weighted average)
- These methods are often used to “smooth” data by fitting a curve to the points
- Popular because they don’t assume any particular shape or form for the curve they fit
- Drawback is that you have to remember the whole training set to make predictions

Topics for today

- Univariate simple linear regression
- Local regression
- Multiple regression (with regularization!)
- Generalized linear models

Multiple regression

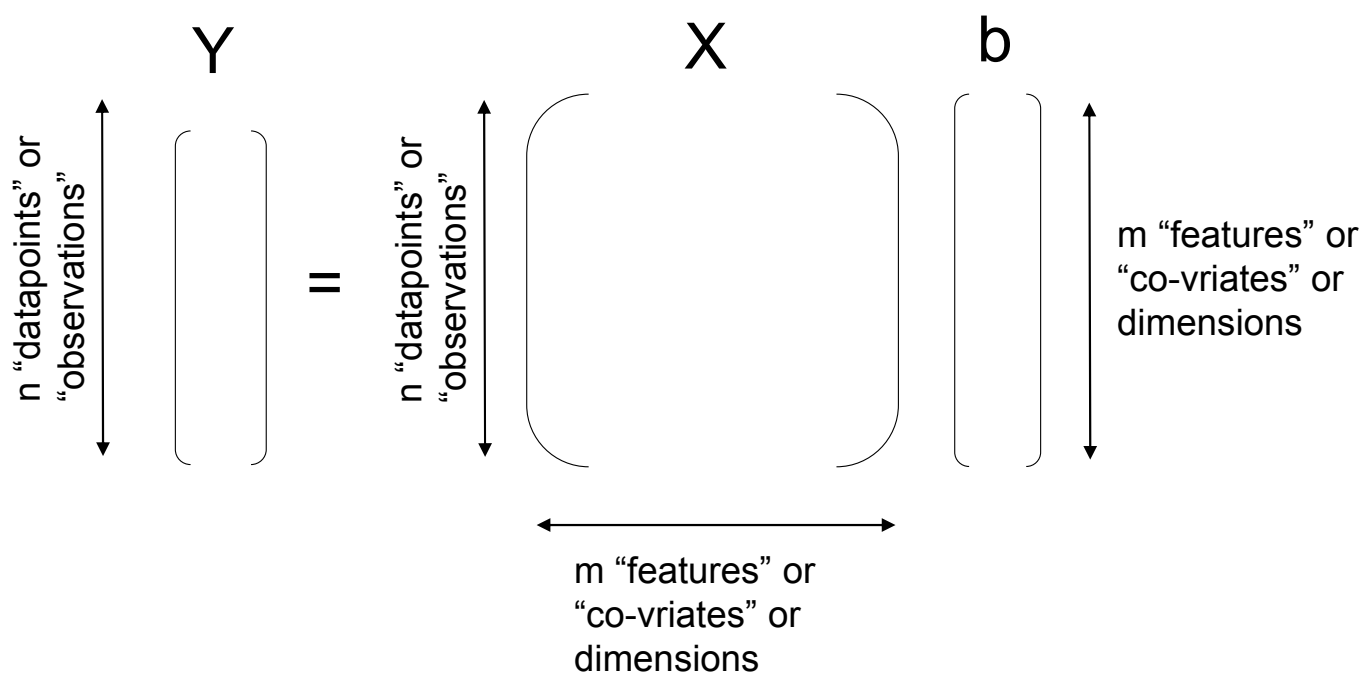
- Predict Y given arbitrarily many dimensions of X

$$\hat{Y}_i = b_0 + b_1 X_{1i} + b_2 X_{2i} + b_3 X_{3i} + \ldots$$

$$\hat{Y}_i = \sum_j b_j X_{ji}$$

$$\hat{Y}_i = X_i\, b$$

$$\hat{Y} = X\, b$$



- How do we fit polynomials?

Overfitting in multiple regression

- Multiple regression will find covariates that match the noise in the data
- Under assumptions, we can write a logL objective function for regression and use the AIC to choose which covariates to include.

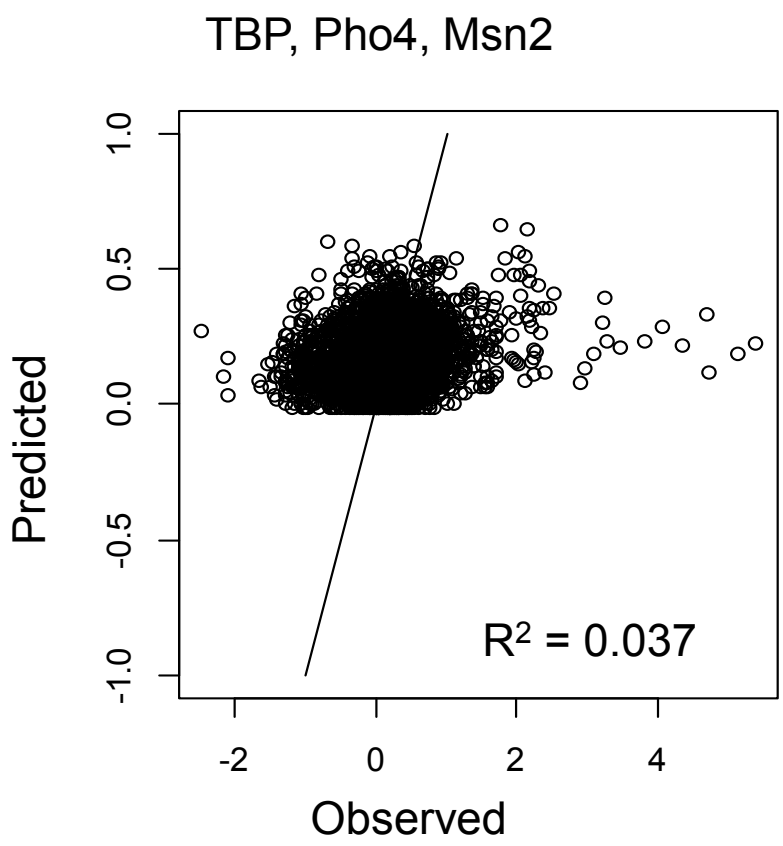
Predicting gene expression based on TFBS motifs

- YETFASCO database has motifs for more than 200 yeast transcription factors

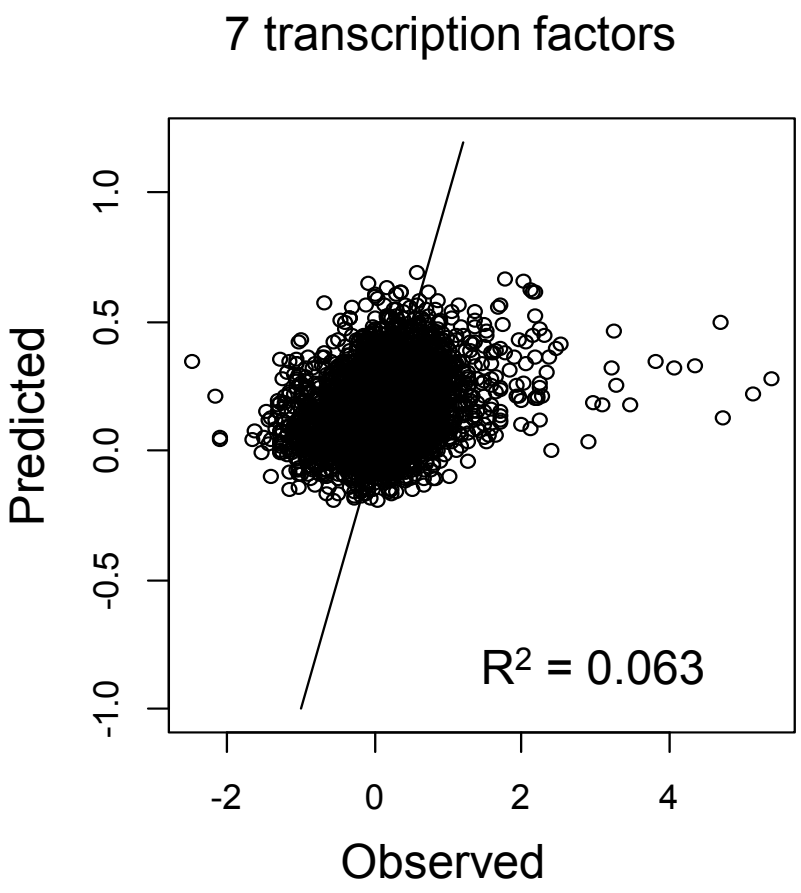
de Boer et al. *NAR* 2011

- Predict matches for each of these to promoters of all yeast genes.

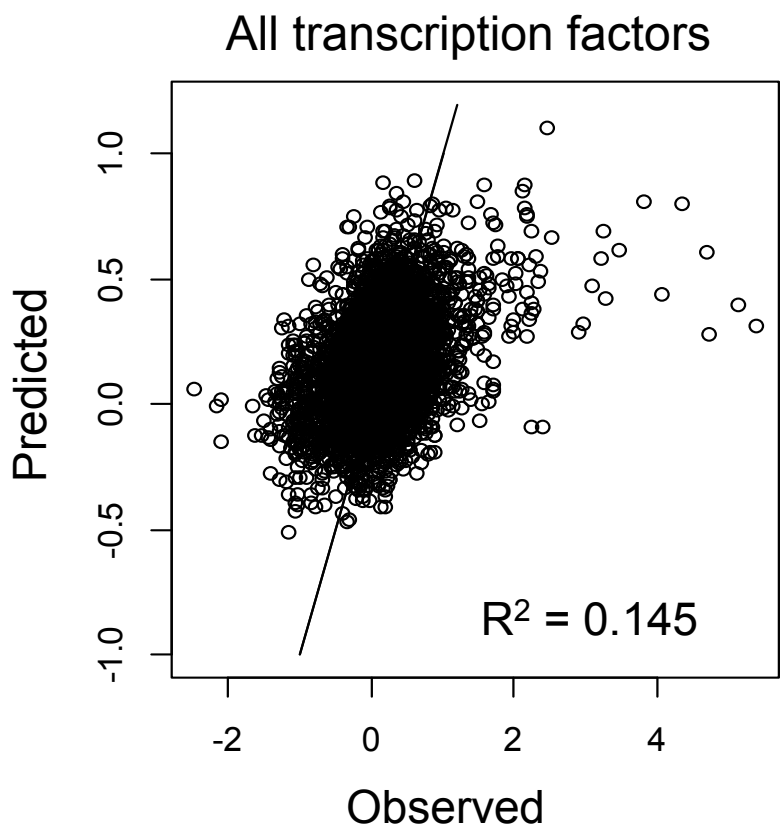
- Predict gene expression (Y) based on these matches (X)



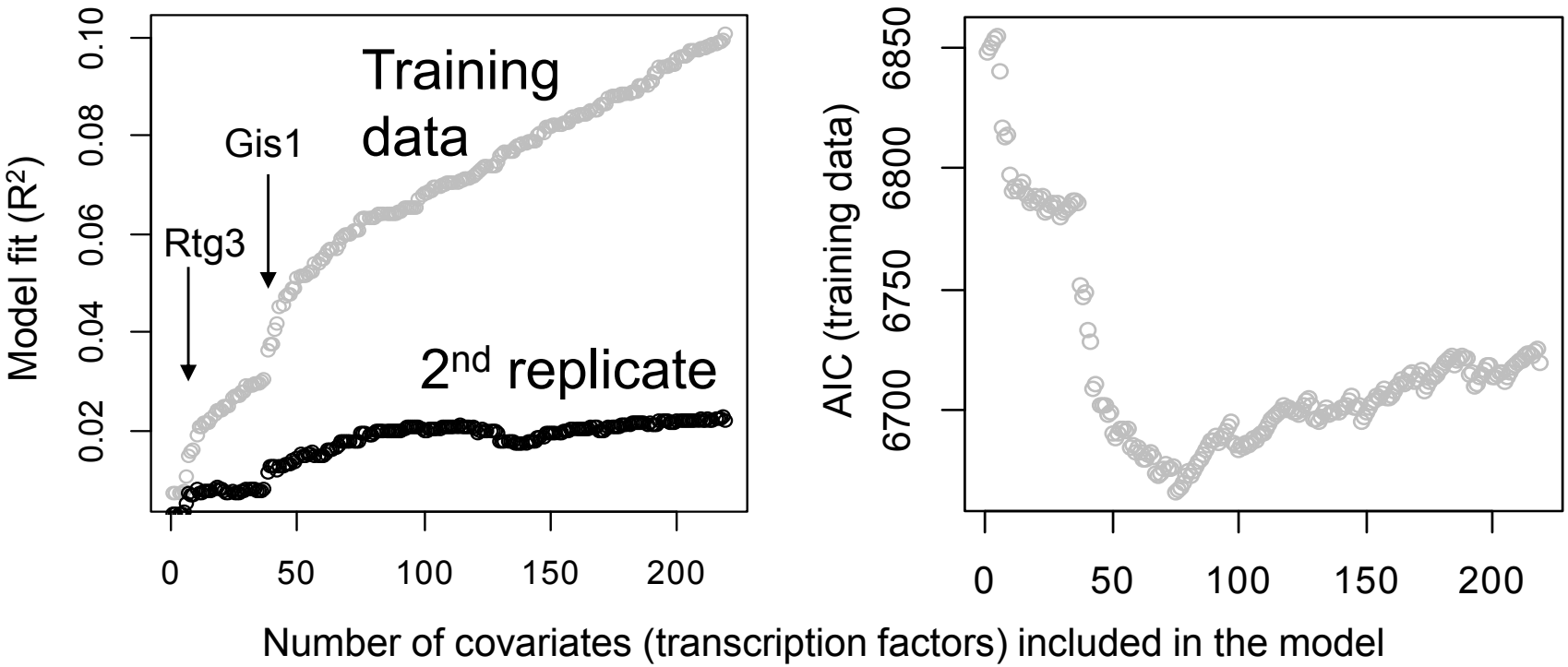
Expression data from Ogawa et al. *MCB* 2001



Expression data from Ogawa et al. *MCB* 2001


All transcription factors
Predicted
Observed
1.0
0.5
0.0
-0.5
-1.0
-2
0
2
4
R2 = 0.145

Add transcription factors sequentially
(in alphabetical order)



Expression data from Ogawa et al. *MCB* 2001

Overfitting in multiple regression

- Multiple regression will find covariates that match the noise in the data
- With large numbers of covariates, there are simply too many models to try
- If the number of covariates approaches or surpasses the number of observations, multiple regression breaks down

Regularization

- We often expect most of the 'b's to be 0 (Y is independent of most of the Xs)
- "Regularization": Modify the cost function to penalize the number of non-zero 'b's

$$SSR = \sum_i [Y_i - b\,X_i]^2$$ ← OLS — Gauss ~1794, Legendre 1805

$$SSR_{L1} = \sum_i [Y_i - b\,X_i]^2 + \alpha \sum_j |b_j|$$ ← LASSO — Tibshirani 1996

$$SSR_{L2} = \sum_i [Y_i - b\,X_i]^2 + \alpha \sum_j b_j^2$$ ← Ridge — Hoerl 1962, Tychonoff 1943

$$SSR_{EN} = \sum_i [Y_i - b\,X_i]^2 + \lambda_1 \sum_j |b_j| + \lambda_2 \sum_j b_j^2$$ ← Elastic Net — Zhou & Hastie 2005

- controls the trade-off between a "good fit" and "over fit"

  Regularization will "choose" the X's that are most useful for explaining Y

$$SSR_{L1} = \sum_i [Y_i - b\,X_i]^2 + \alpha \sum_j |b_j| \longleftarrow$$ LASSO

Tibshirani 1996

- α controls the importance of the regularization relative to the "data fit"
- We can always minimize the cost function by setting α to 0 …

But then we have no regularization.

- α is a "hyperparameter" – it can't be chosen to optimize the cost function.
- How do we choose it?

Trade off between L1 and L2 regularization

- L1 does great at removing co-variates that don’t predict, but if there are two correlated variables it will choose one of them
- L2 “shares” the fit among the correlated variables, but doesn’t work very well for removing the uncorrelated variables

More general types of regularization

- So far, regularization has been used to encourage sparsity
- Regularization can be used to influence model structure in other ways
- E.g., population structure in GWAS or cell-lineage structure in gene expression modeling

Puniyani et al. *Bioinformatics* 2010

Jojic et al. *Nat Immunol.* 2013

Topics for today

- Univariate simple linear regression
- Local regression
- Multiple regression (with regularization!)
- Generalized linear models

Regression is not just "linear"

- Linear regression means linear in the b's, not in X.
- So any function of bX is allowed.
- E.g., Predicted $\hat{Y} = \sqrt{b_0 + b_1 X}$
- Or more generally, $\hat{Y} = f_L(bX)$

Vector of b's

"link" function

- We've seen logistic regression, where

$$f_L(t) = \frac{1}{1 + e^{-t}}$$

Generalized linear models

- Choose the “link function” to match the type of data in Y.
- Binary classification: positive or negative
- Multi-way classification: e.g., 5 cell types
- Natural numbers: 0,1,2,3 …

What if the data in X are not real numbers?
E.g., genotypes or number of matches to TFBS motifs?