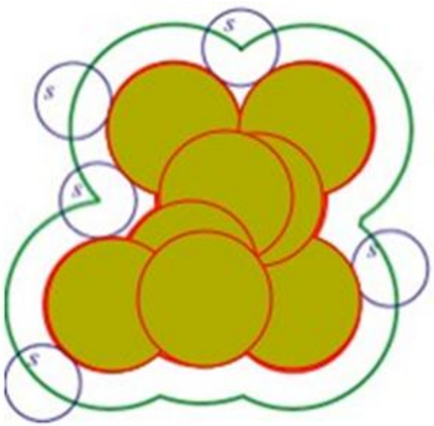
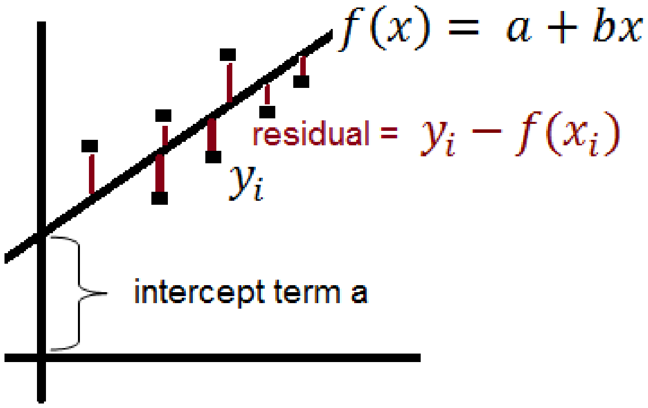
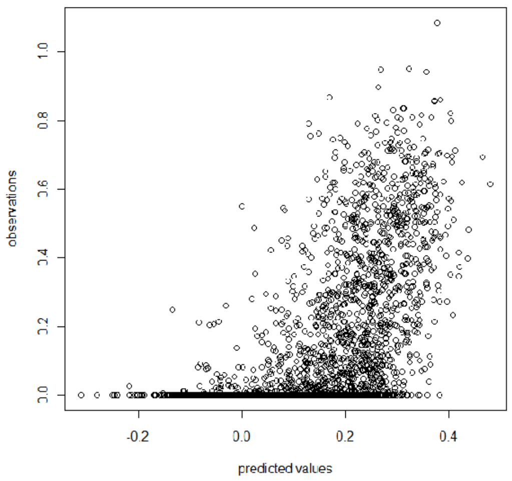
Prediction of the burial status of transmembrane residues of helical membrane proteins (with Linear Regression)

Jing Cui

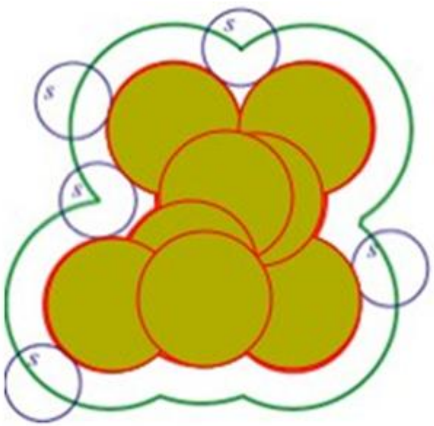
Outline

- burial status 
- linear regression 

- practical 

Outline

- burial status



HMPs

- Helical Membrane Proteins (HMPs)
  - play a crucial role in diverse cellular processes
- **Why** we need to predict their structures ?
  - hard by experimental techniques
    - <1% of proteins with known structure are HMPs
- **How** to predict their structures ?
  - solvent accessibility (burial status)

Park, Y. (2007) BMC Bioinformatics

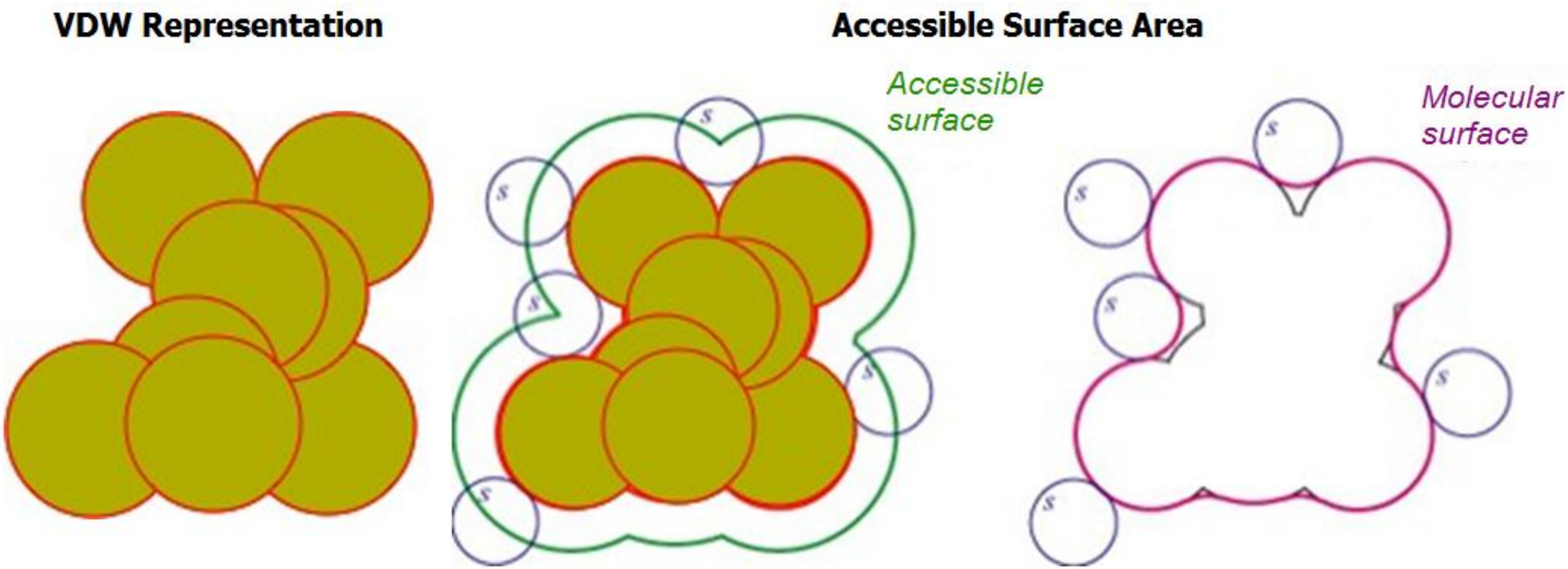
Burial Status

- **Prediction Burial Status** : transmembrane(TM) residues of HMPs buried in the protein structure vs. exposed to the membrane
- rSASA value ➡ buried vs. exposed
- rSASA value = $\frac{SASA}{reference\ value}$

rSASA	Burial Status
> 0.00	exposed residue
0.00	buried residue

Park, Y. (2007) BMC Bioinformatics

Solvent Accessible Surface Area (SASA)

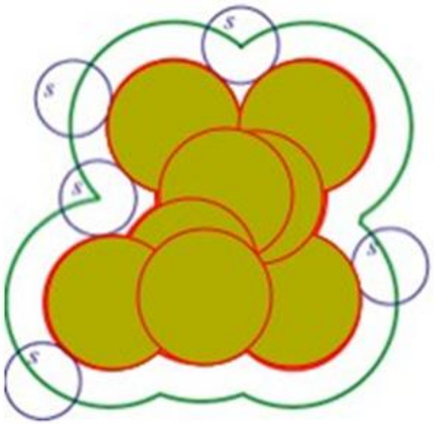


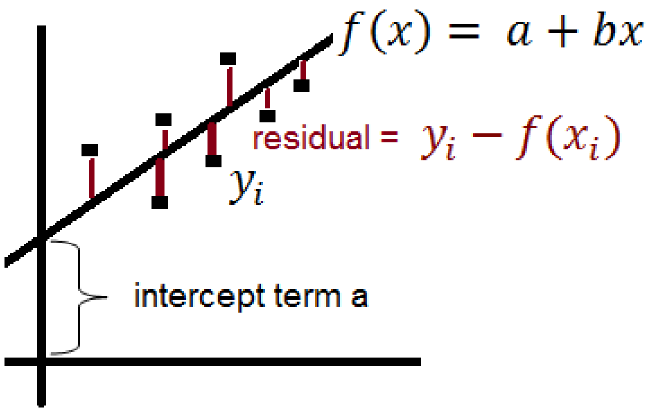
Solvent Accessible Surface is the surface defined by rolling a sphere the size of solvent over the molecule.

Richards, F.M. (1977) Annu. Rev. Biophys. Bioeng
Kavraki, L.E. (2007) The Connexions Project

Outline

- burial status
- linear regression





Multiple Linear Regression Models

- Simple
- Interpretable description of how the inputs affect the output

$$\textit{input vector}: \; X^T = (X_1, X_2, \dots, X_p) \rightarrow Y$$

$$\textit{linear regression model}: \; f(x) = \beta_0 + \sum_{j=1}^{p} X_j \beta_j$$

$\beta_0, \beta_1, \dots, \beta_p$: parameters or coefficients

$X_1, X_2, \dots, X_p$: variables or features

Hastie, T. (2009) Springer

Multiple Linear Regression Models

The model is linear in the parameters.

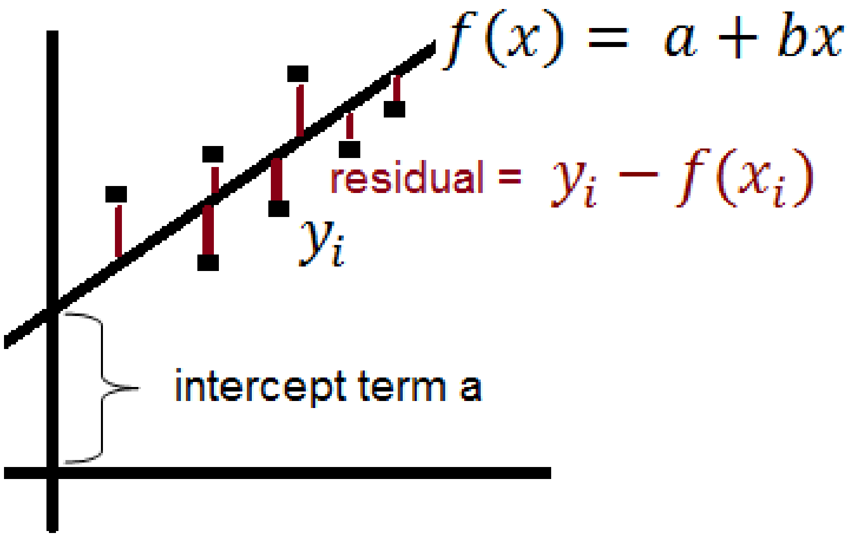
$$X_j \rightarrow X_j^{\,2} \;,\; X_j^{\,3}$$
$$X_j \rightarrow \sqrt{X_j} \;,\; logX_j$$
$$X_j \rightarrow X_j \cdot X_i$$

$$\begin{array}{ccccc}
y_1 & x_{11} & x_{12} & \cdots & x_{1p} \\
y_2 & x_{21} & x_{22} & & x_{2p} \\
\cdots & & \cdots & & \\
y_n & x_{n1} & x_{n2} & \cdots & x_{np}
\end{array}$$

one case
n cases in total

one feature
p features in total

Multiple Linear Regression Models



(x_1, y_1)

.

.

(x_N, y_N)

estimation method : *least squares*

$$\beta = (\beta_0, \beta_1, \ldots, \beta_p)^T$$

minimize

$$RSS(\beta): residual\ sum\ of\ squares$$

$$= \sum_{i=1}^{N} (y_i - f(x_i))^2$$

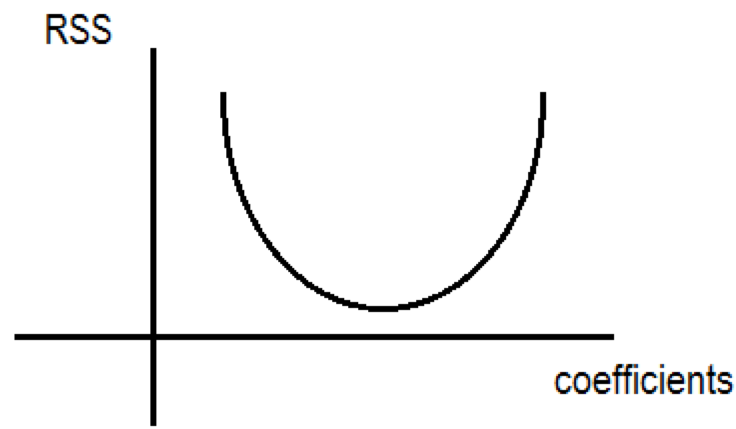
Hastie, T. (2009) Springer

Multiple Linear Regression Models

$$f(x) = \beta_0 + \sum_{j=1}^{p} X_j \beta_j \xrightarrow{X = \begin{pmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1p} \\ 1 & x_{21} & x_{22} & & x_{2p} \\ \cdots & & \cdots & & \cdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{np} \end{pmatrix}} f(x) = X\beta$$

$RSS(\beta)$: *residual sum of squares*

$$= \sum_{i=1}^{N} (y_i - f(x_i))^2 = (y - X\beta)^T (y - X\beta)$$



$$\frac{\partial RSS}{\partial \beta} = -2X^T(y - X\beta)$$

$$\frac{\partial RSS}{\partial \beta \partial \beta^T} = 2X^T X$$

$$\Rightarrow \quad \hat{\beta} = (X^T X)^{-1} X^T y$$

Hastie, T. (2009) Springer

Assumptions

- The Gauss-Markov Theorem

$$Y = \beta_0 + \sum_{j=1}^{p} X_j \beta_j + \varepsilon$$

With $E[\varepsilon | X = x] = 0,$ errors have expectation zero

$Var[\varepsilon | X = x] = \sigma^2,$ constant variance

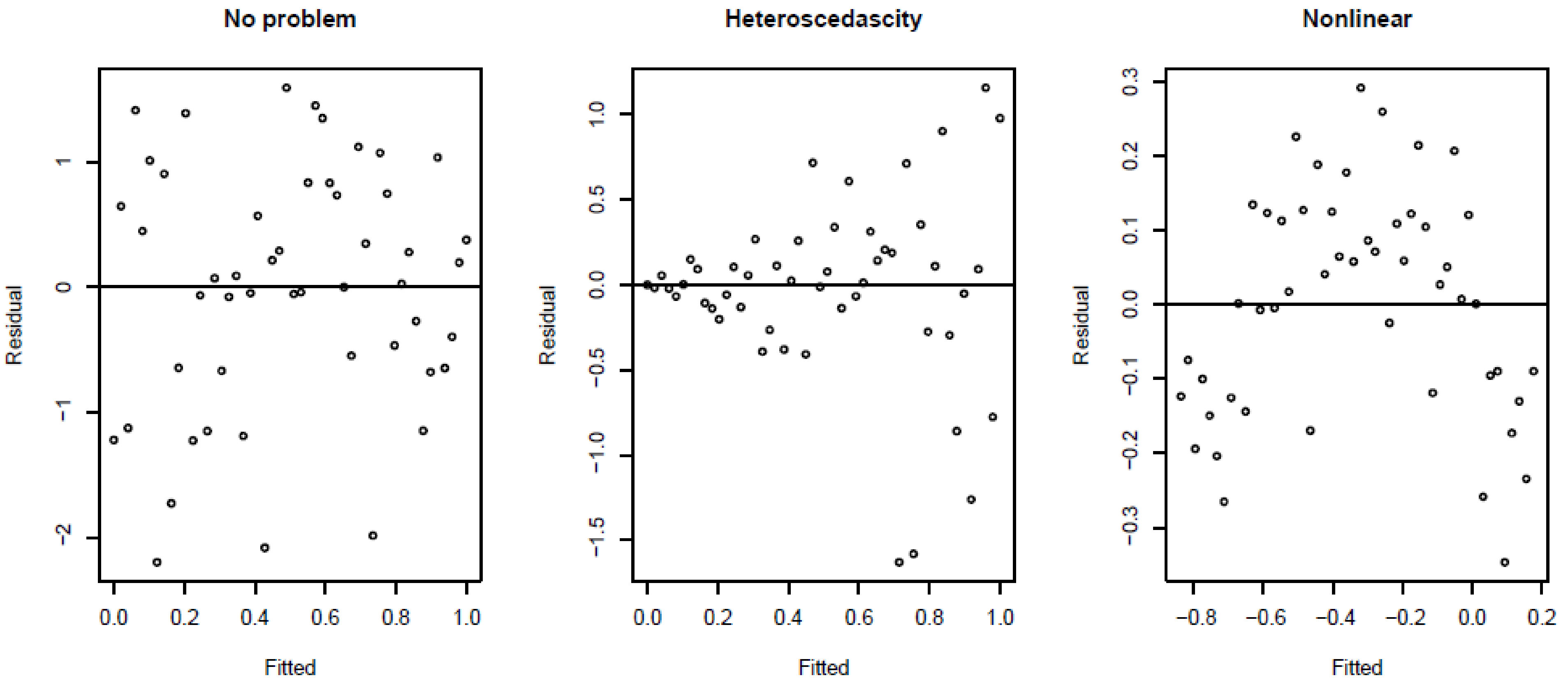
$E[\varepsilon_x \varepsilon_z] = 0$ the errors are uncorrelated

Under these assumptions, the maximum likelihood is given by the method of least squares.

Crawley, M.J. (2007) John Wiley

Model Assessment and Selection

- Model checking



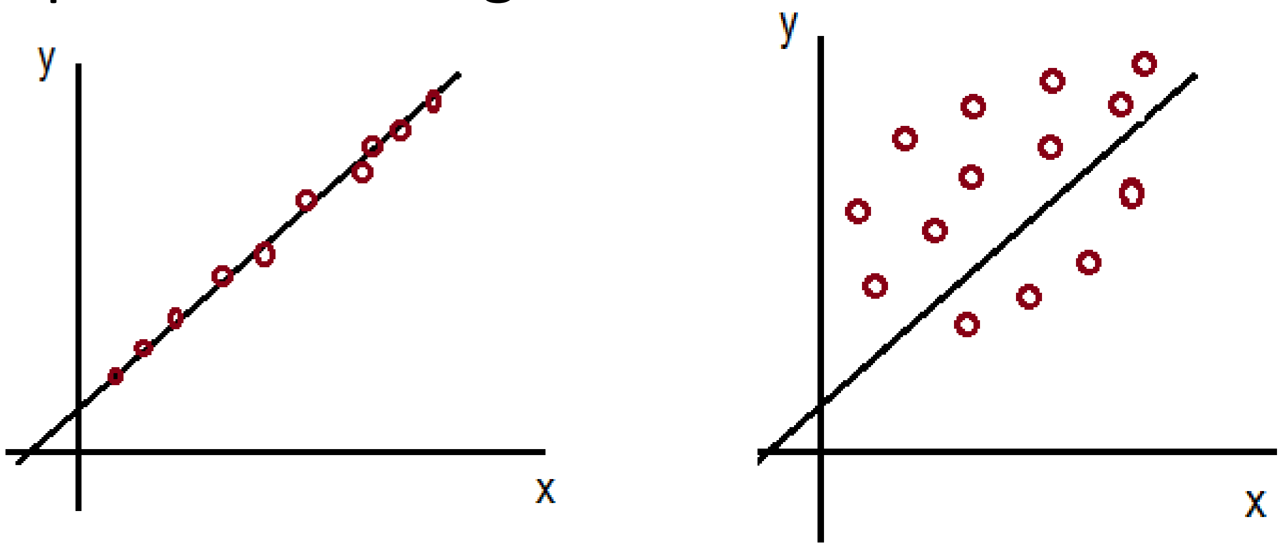
Faraway, J.J. (2002) CRAN

Model Assessment and Selection

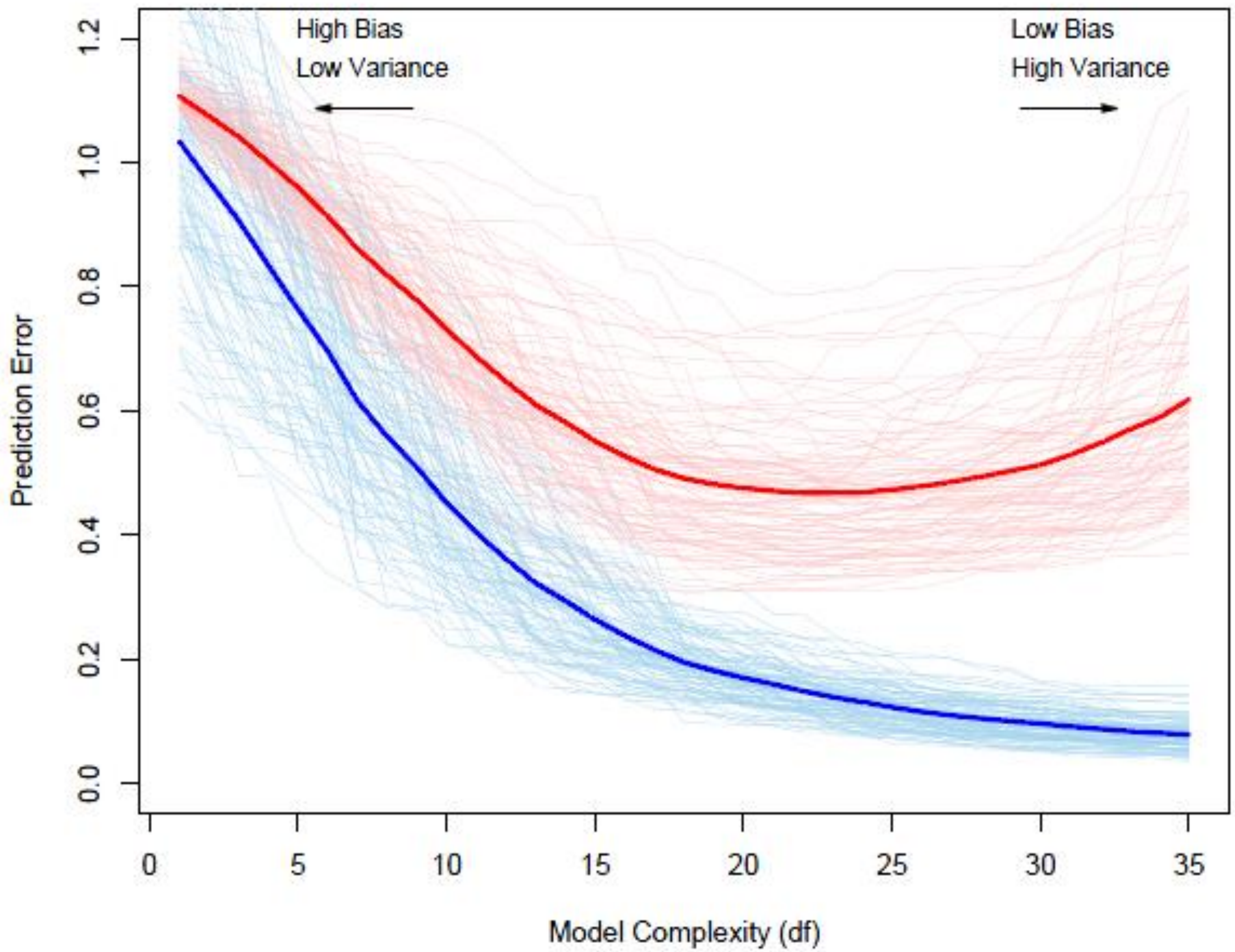
- Goodness of Fit
  - Coefficient of determination R^2

$$R^2 = 1 - \frac{RSS}{Total\ SS} = 1 - \frac{\sum(\hat{y}_i - y_i)^2}{\sum(y_i - \bar{y})^2}$$

perfect fit will get 1



Test Error and Training Error



Hastie, T. (2009) Springer

Model Assessment and Selection

- AIC (<u>A</u>kaike <u>i</u>nformation <u>c</u>riterion)

$$AIC = -\frac{2}{N} * loglik + 2 * \frac{d}{N}$$

d : the number of parameters in the model

N: sample size

log-likelihood loss function is used

Hastie, T. (2009) Springer

Cross Validation

1	2	3	4	5
Train	Train	Validation	Train	Train

K-Fold Cross-Validation

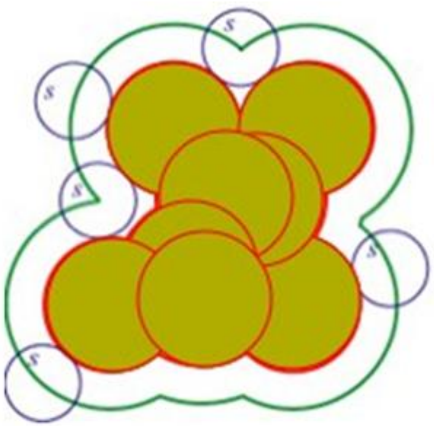
Leave-One-Out cross-validation   K = N

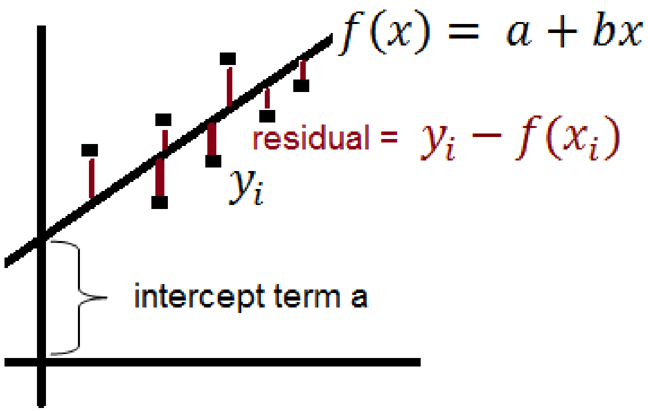
$$\mathrm{CV}(\hat{f}) = \frac{1}{N}\sum_{i=1}^{N} L(y_i, \hat{f}^{-\kappa(i)}(x_i))$$

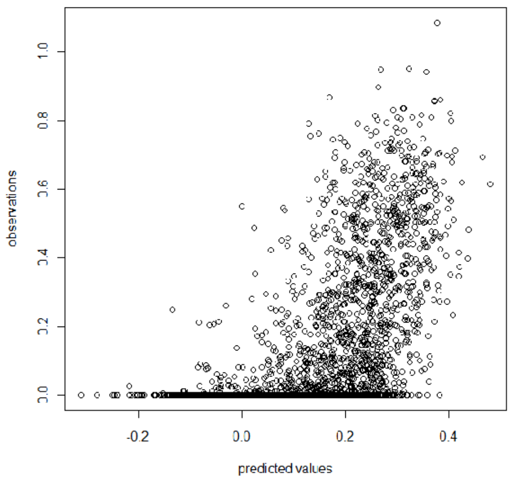
Hastie, T. (2009) Springer

Outline

- burial status
- linear regression
- practical







Dataset

```
pdbid chain number type rsasa freq1 freq2 freq3 freq4 freq5 freq6 freq7 freq8 freq9 freq10
3ddl A 18 F 0.642 0.0 0.234043 0.0 0.0 0.531915 0.085106 0.148936 0.0 0.0 0.0 0.0 0.0 0.0 (
3ddl A 19 T 0.0 0.0 0.0 0.0 0.0 0.0 0.06383 0.425532 0.191489 0.0 0.234043 0.0 0.06383 0.02
3ddl A 20 V 0.168 0.0 0.021277 0.021277 0.0 0.021277 0.170213 0.042553 0.021277 0.042553 0.
3ddl A 21 A 0.792 0.0 0.021277 0.0 0.148936 0.0 0.0 0.021277 0.787234 0.0 0.021277 0.0 0.0
```

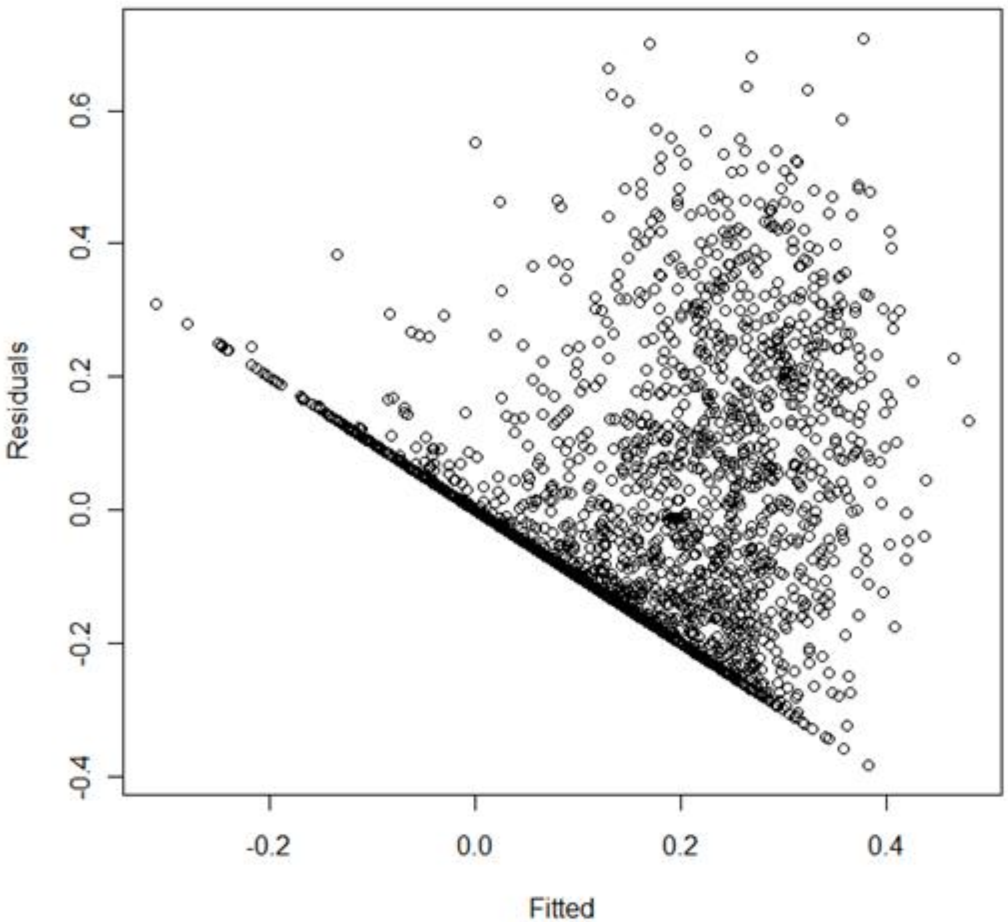
41 features

rSASA	frequencies per aa (20)	PSSM (PSI-Blast results) (20)	conservation index
2595	2595	2595	2595

2595 cases without missing values

1^{st} Linear Regression Model

$$f(x) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_{41} X_{41}$$



correlation coefficient = 0.5949052

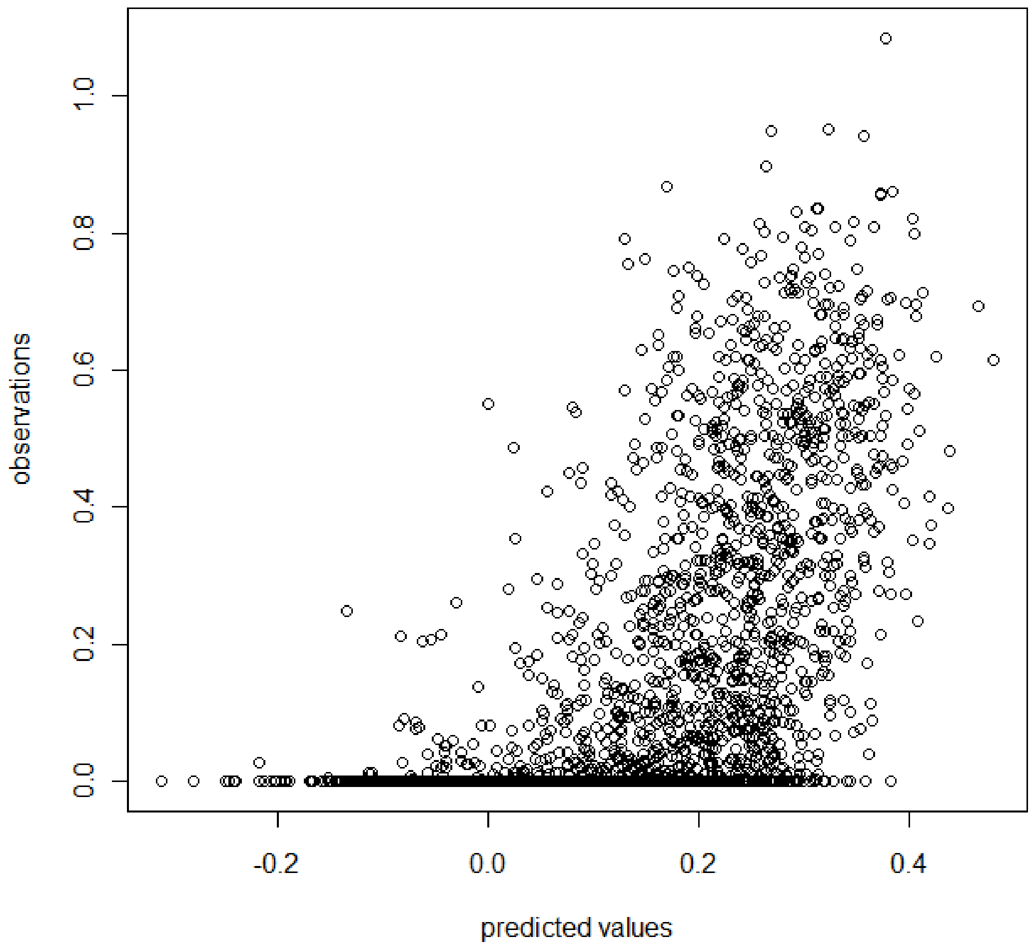
$\mathbf{R^2}$ = 0.353912197

prediction error = 0.02994921 (cross-validation estimate)

AIC = -1653.624

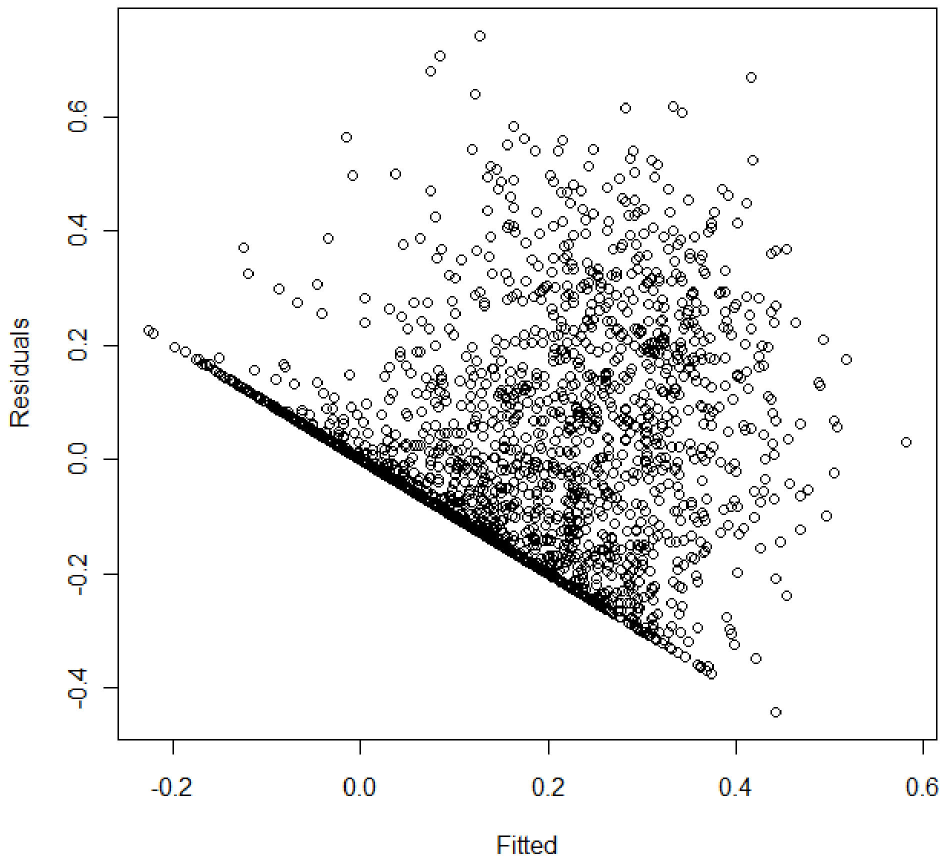
1st Linear Regression Model

$$f(x) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_{41} X_{41}$$



2^{nd} Linear Regression Model

features : *conservation, score11, freq2, score13, score20, score17, freq4* with quadratic term



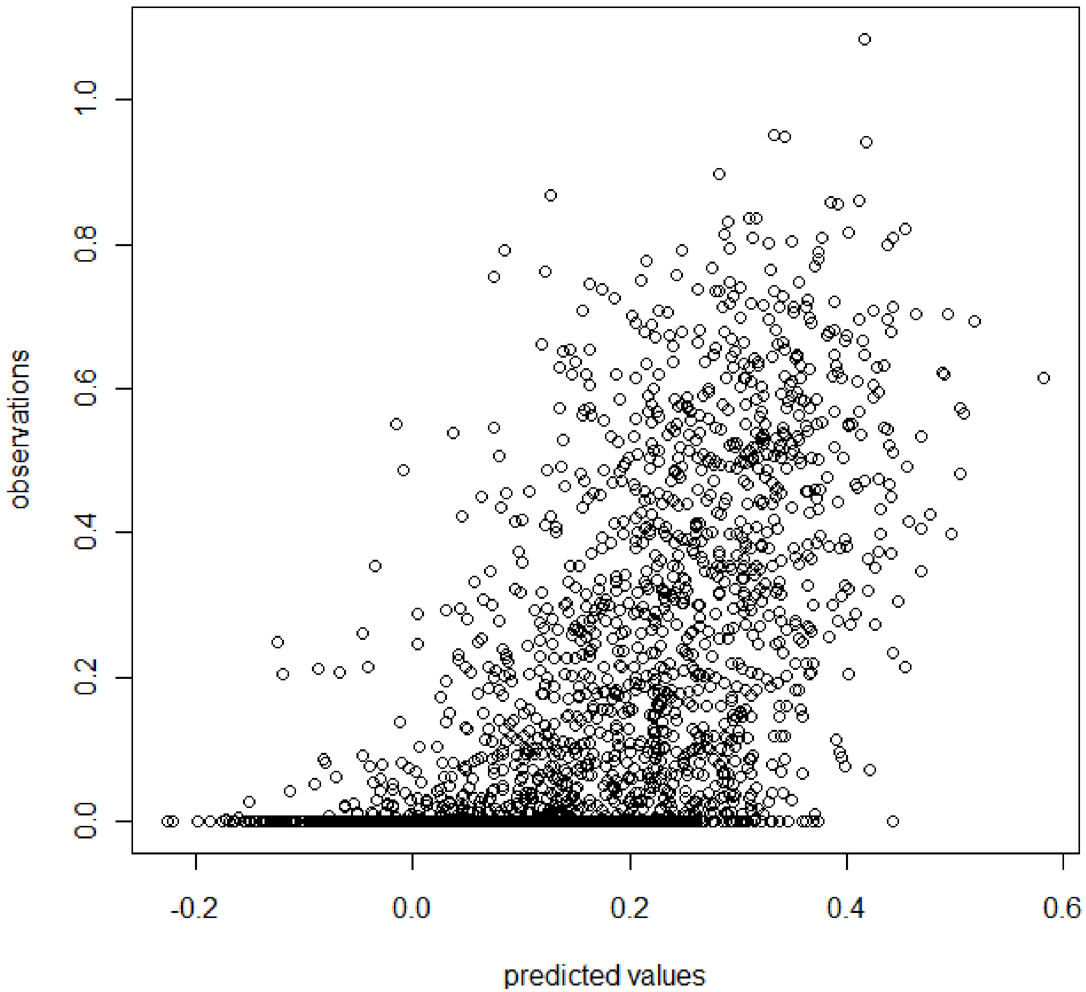
correlation coefficient =0.6233053

$\mathbf{R^2}$ = 0.3885095

prediction error = 0.02834546 (cross-validation estimate)

AIC = -1794.442

2^{nd} Linear Regression Model



References

- *Park, Y.* **Prediction of the burial status of transmembrane residues of helical membrane proteins.**(*2007)BMC Bioinformatics*
- *Richards, F. M.* **Areas, volumes, packing and protein structure.**(1977) *Annu Rev Biophys Bioeng*
- *Kavraki, L.E.* **Molecular Shapes and Surfaces.**(2007*) The Connexions Project*
- *Faraway, J.J.* **Practical Regression and Anova using R.**(2002)*CRAN*
- *Hastie, T.* **The Elements of Statistical Learning**.(2009)*Springer*
- *Crawley, M.J.* **The R book.** (2007)*Wiley*

Thank you for your attention