What You See May Not be What You Get – A Brief Introduction to Overfitting

Fei Ye, PhD
Alex (Zhiguo) Zhao, MS

April 16th, 2010

Cancer Biostatistics Workshop

The problem of overfitting

- Overfitting: an old problem since 19th century

- Generally recognized to be a violation of Ockham's razor (*All other things being equal, the simplest solution is the best – 14th century logician William of Ockham*).

- Definition of overfitting: fitting a statistical model with too many degrees of freedom in the modelling process.

- Conclusion: overfitting makes you too optimistic about the performance of your model. Overfitting costs money.

Causes of overfitting

- Model is too complex (too many predictors).
- Training data too noisy.
- Model being refined over time with ever increasing data inputs.
- Training set too small.
- A very rich hypothesis space.

Phenomena of overfitting

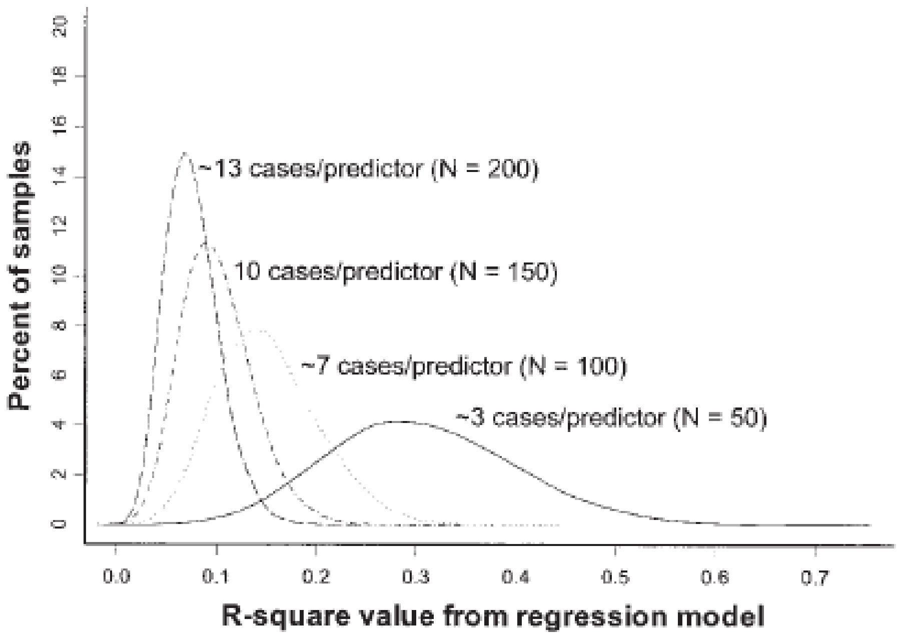
- Exaggerate minor fluctuations in the data
- A complex model may fit the noise, not just the signal (real underlying relationship)

simulation study:

one outcome,

15 predictors

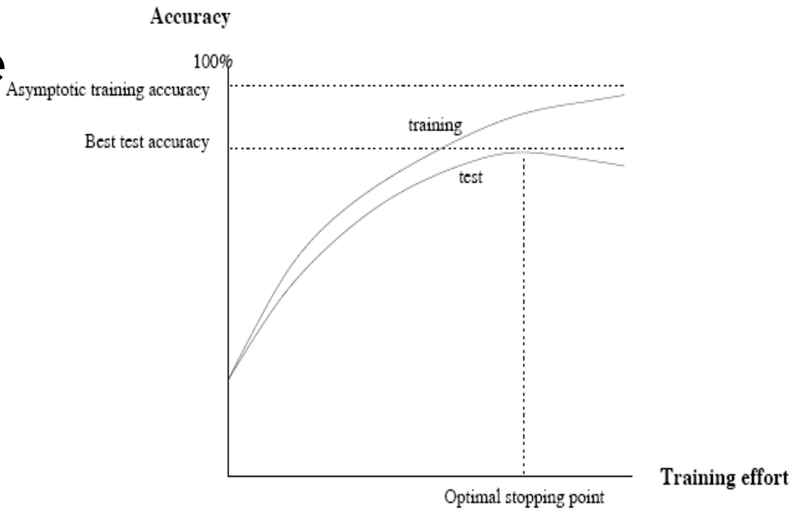
multiple regression model



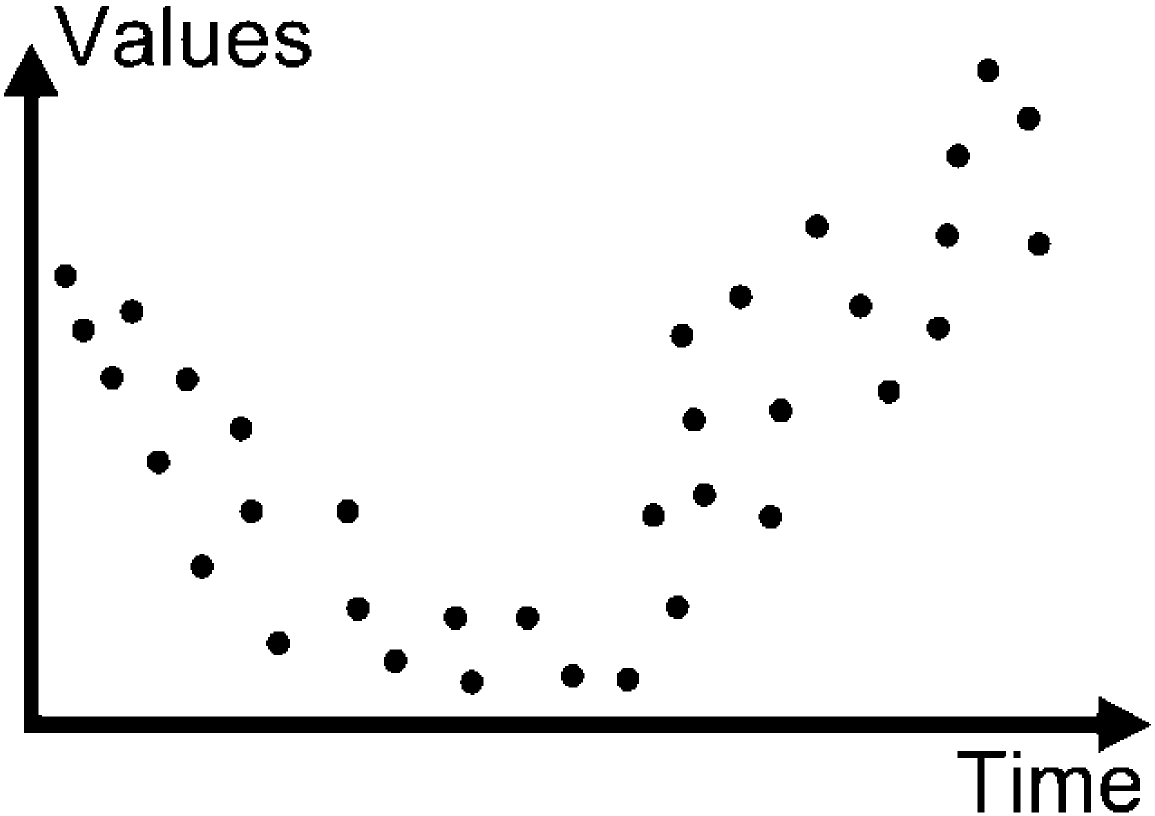
Phenomena of overfitting – *cont.*

- Data are well described, but the predictions do not generalize to new data outside the study sample.

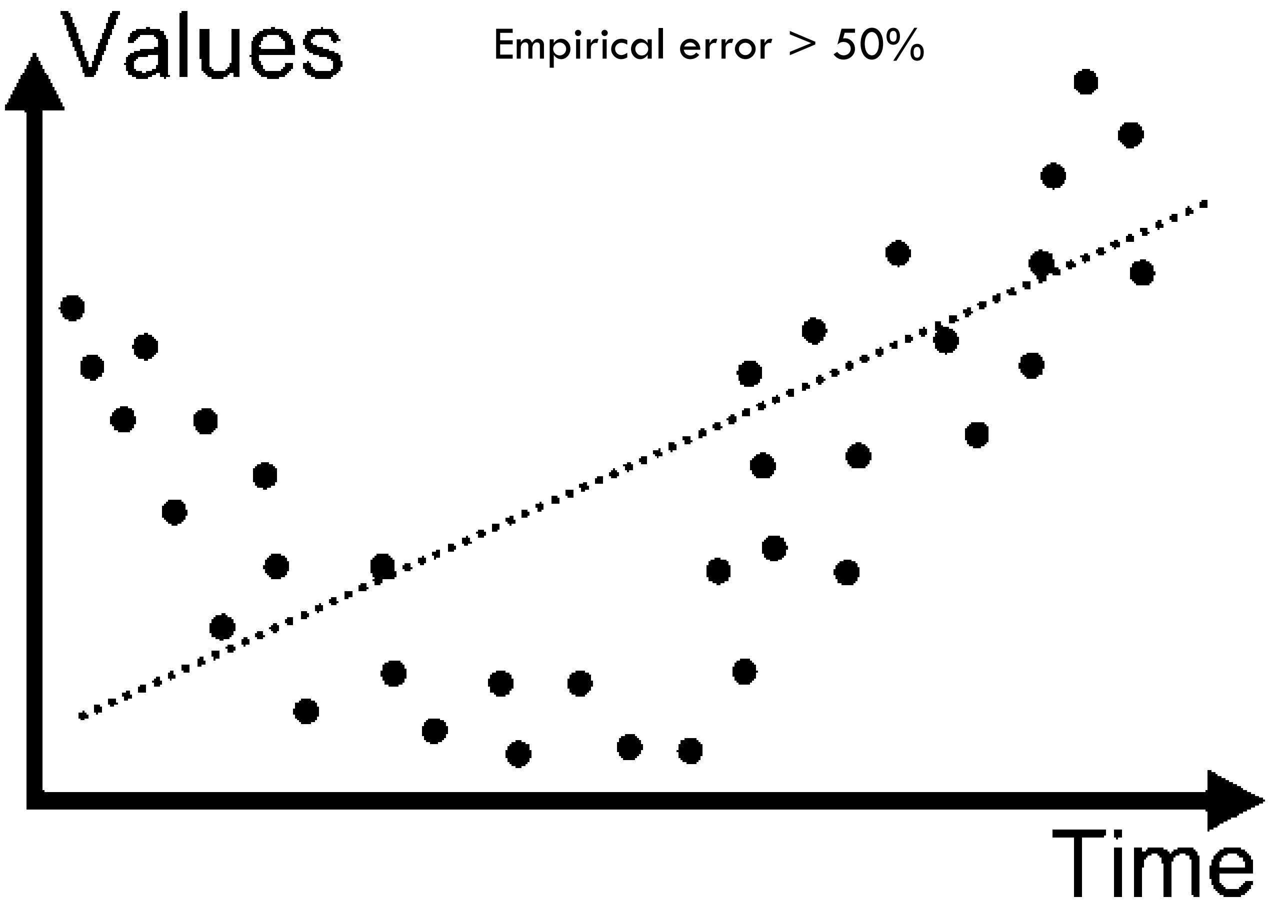
- The training set performance continues to improve, while the test set performance no longer does.



A simple example: data

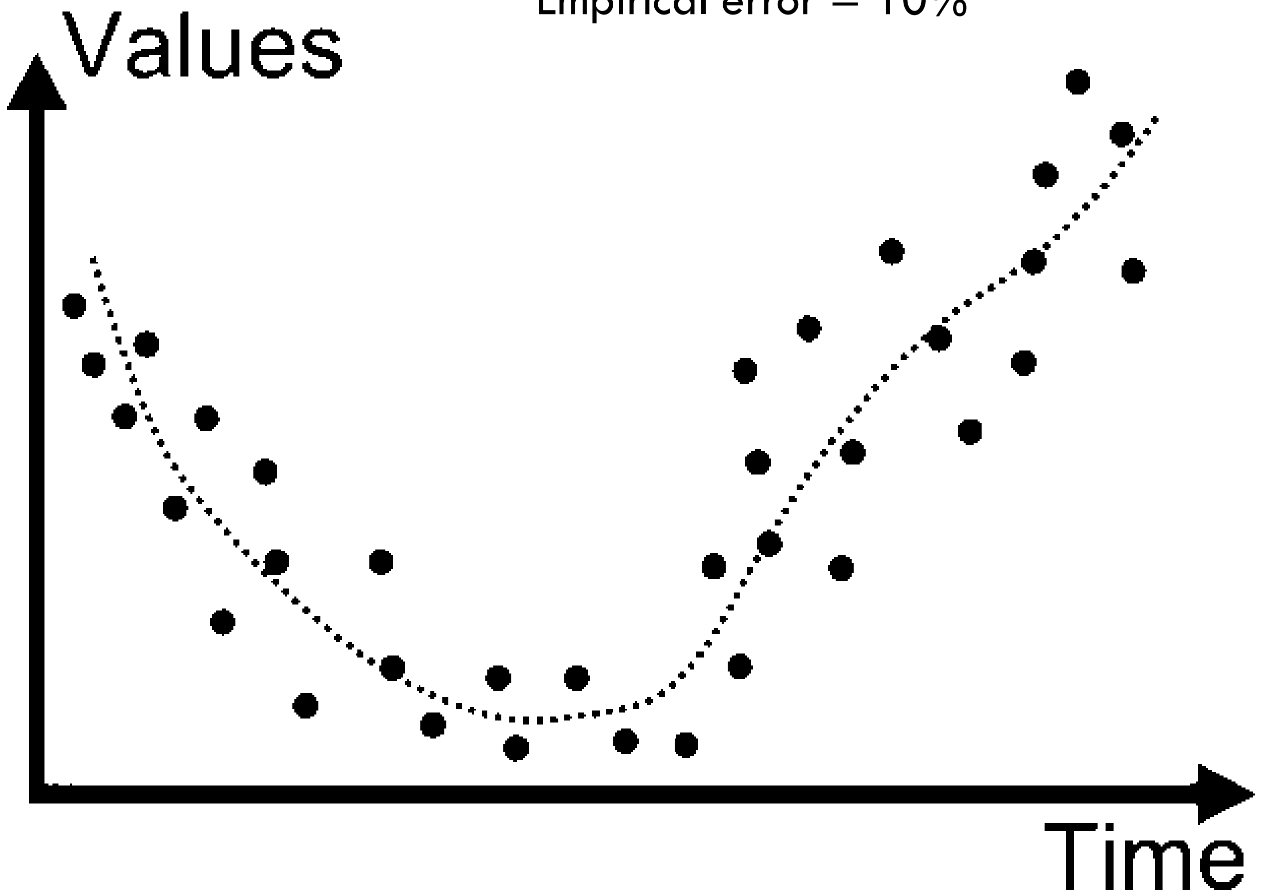


Linear model



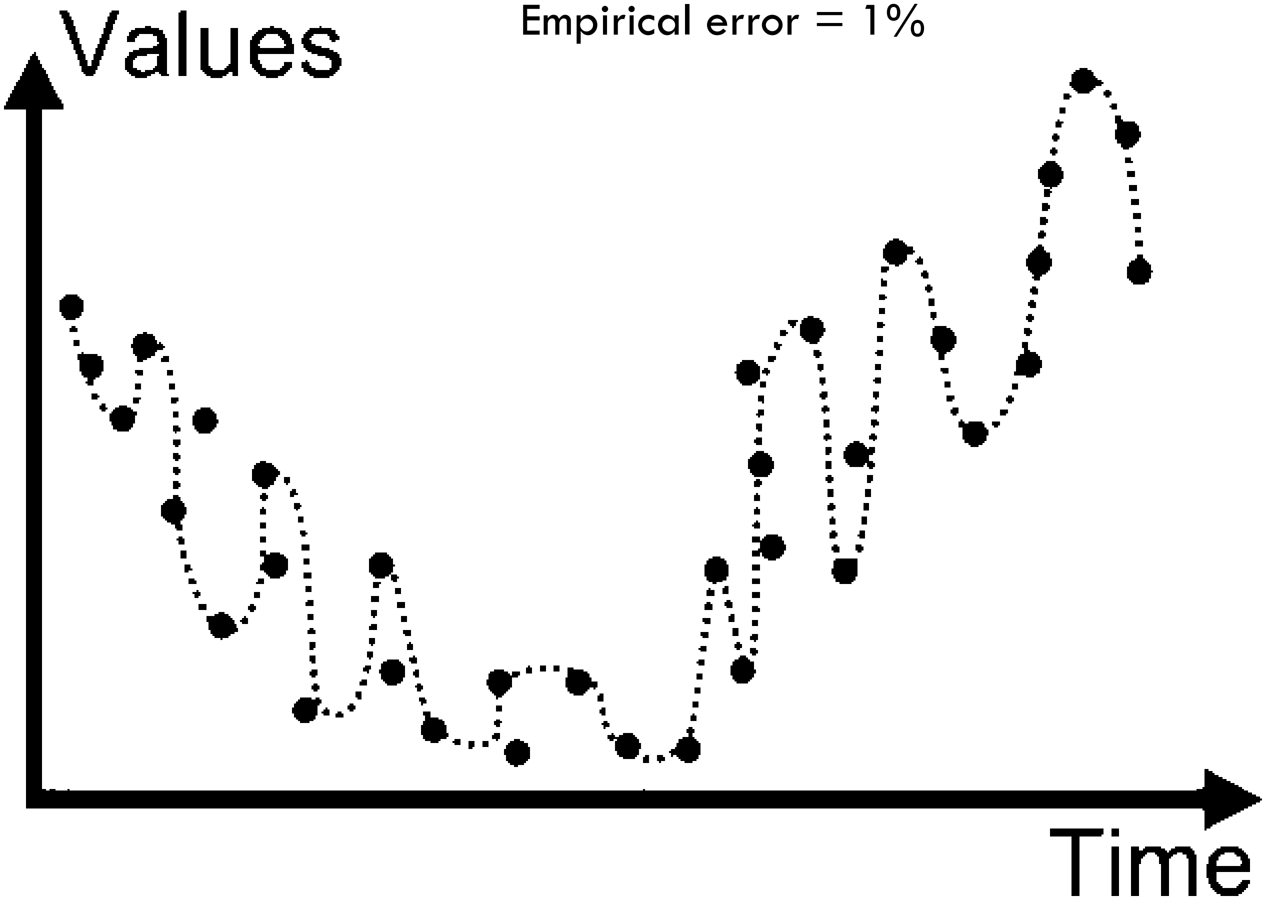
More complex model

Empirical error = 10%



Even more complexity

Empirical error = 1%



Which model is better?

Bias versus Variance

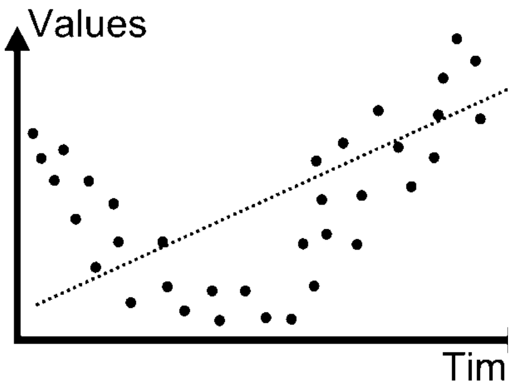
- **Bias**: average difference between our estimate and the true mean.

  Model too "simple" $\Longrightarrow$ does not fit the data well $\Longrightarrow$ a biased model

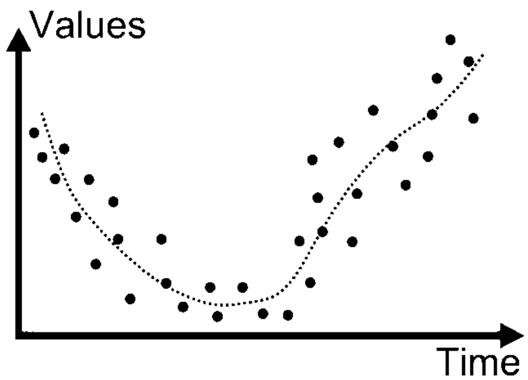
- **Variance**: squared deviation of the predictor around its mean (variation explained by the model).

  Model too complex $\Longrightarrow$ small changes to the data changes the model a lot $\Longrightarrow$ a high-variance model

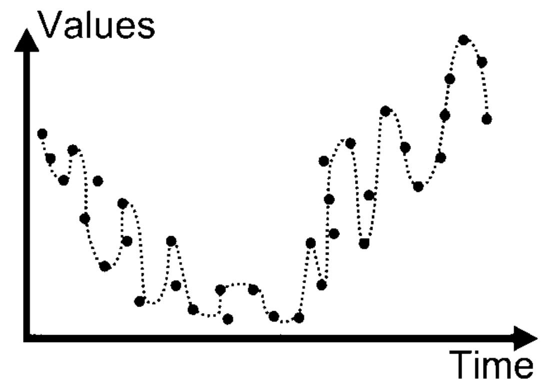
Bias-Variance trade-off



Large bias
Small variance



Moderate bias
Moderate variance



Small bias
Large variance

Prediction accuracy

- The easiest way to measure the prediction accuracy: make a prediction and wait for the event to happen.
- Drawbacks of the above method:
  - Can't wait
  - Method of prediction is changing
  - It only tells you about the accuracy of **past** predictions.
- It matters little to know the past prediction accuracy.
- Primary aspect of statistical models: not to provide good or bad predictions, but to provide repeatable predictions (**generalization**).

Decomposition of prediction error

- The expected squared error (MSE) of a predictor/model $\widehat{r}(x)$ from Y can (also referred as the *loss function*) be decomposed into:

Let $\widehat{r}(x)$ be any predictor. Then

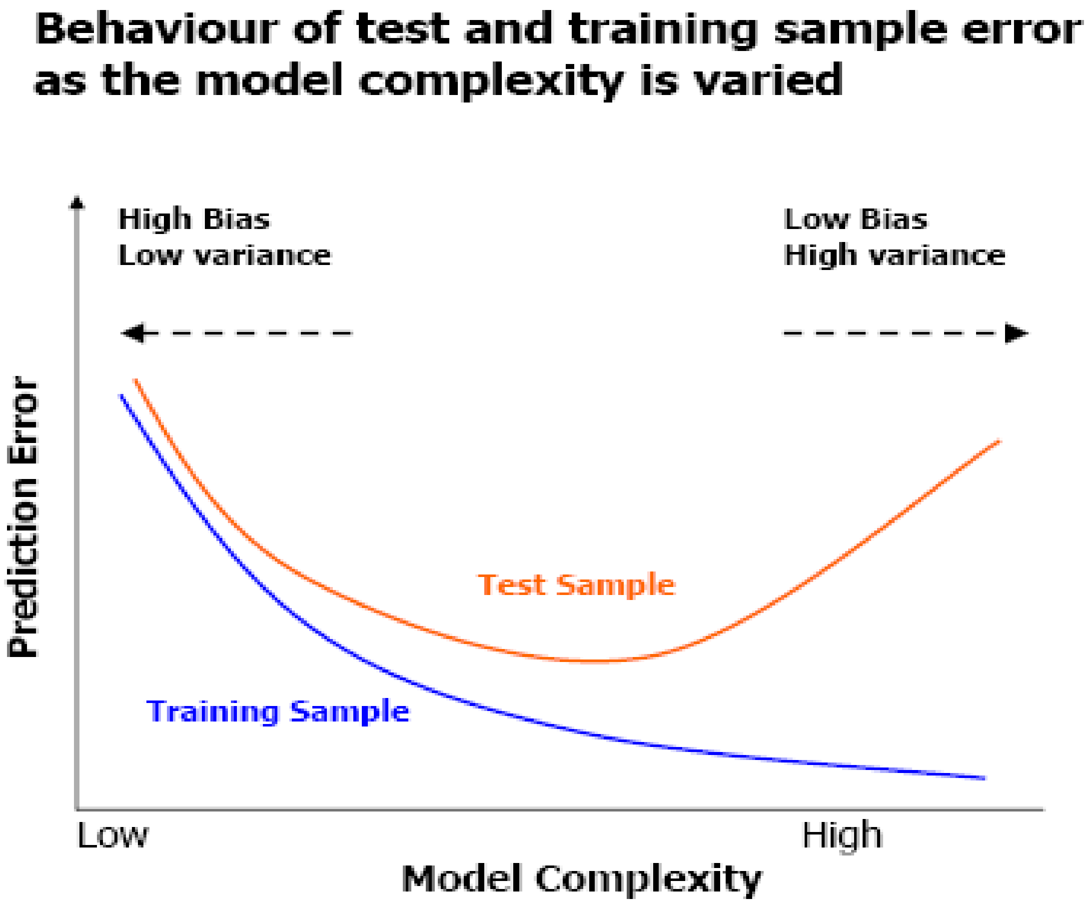
$$R = \mathbb{E}(Y - \widehat{r}(X))^2 = \int R(x) f(x) dx$$

where $R(x) = \mathbb{E}((Y - \widehat{r}(X))^2 | X = x)$. Let $\overline{r}(x) = \mathbb{E}(\widehat{r}(x))$, $V(x) = \mathbb{V}(\widehat{r}(x))$ and $\sigma^2(x) = \mathbb{V}(Y | X = x)$.

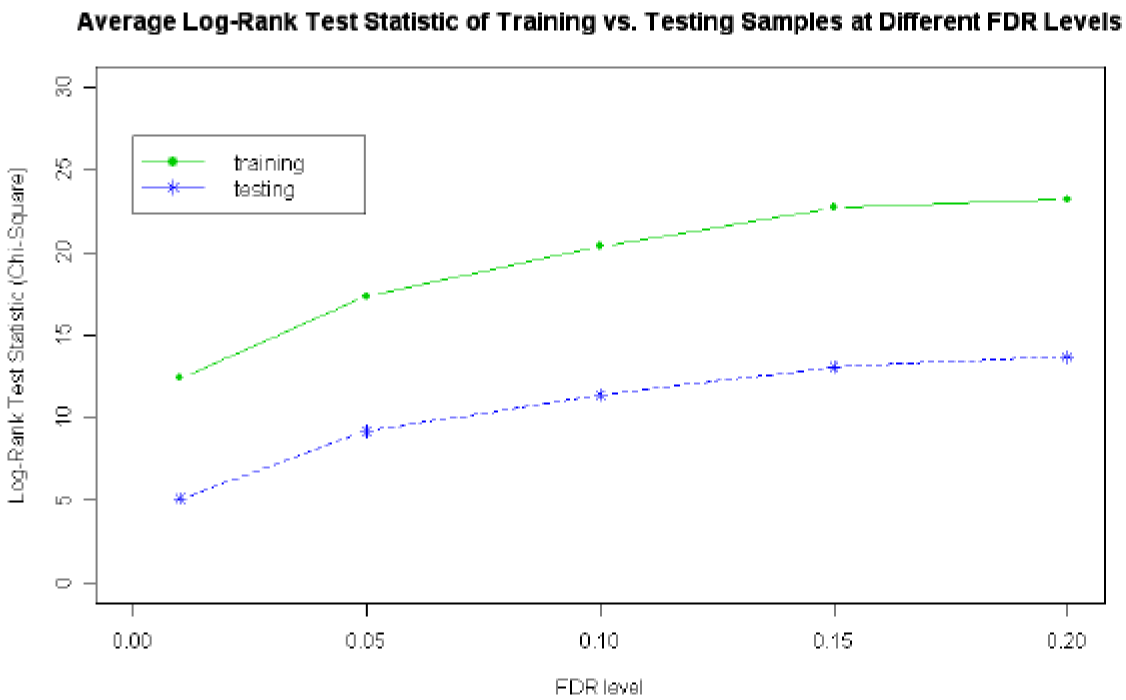
$$\begin{aligned} R(x) &= \mathbb{E}((Y - \widehat{r}(X))^2 | X = x) \\ &= \mathbb{E}\left(((Y - r(x)) + (r(x) - \overline{r}(x)) + (\overline{r}(x) - \widehat{r}(x)))^2 \,\middle|\, X = x \right) \\ &= \underbrace{\sigma^2(x)}_{\text{irreducible error}} + \underbrace{(r(x) - \overline{r}(x))^2}_{\text{bias squared}} + \underbrace{V(x)}_{\text{variance}} . \end{aligned}$$

- The first term: random error/noise; beyond our control.
- The remaining two terms, the bias and the variance are functions of our predictor/model and therefore can potentially be reduced.

Model complexity and prediction error



An example of overfitting in high-dimensional data



Survival data: 75 predictors, 66 patients

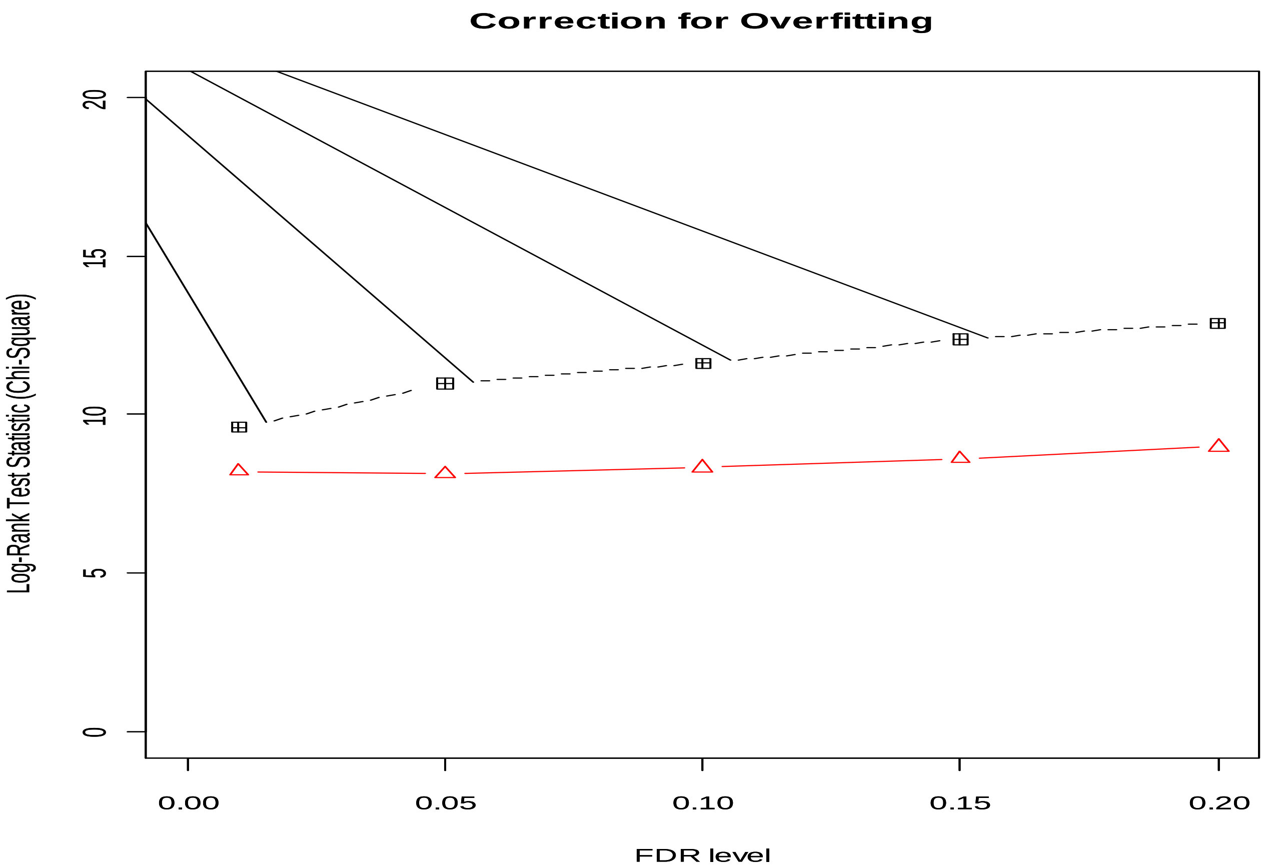
Correction for overfitting – a simple approach

- How well will a model perform on new data drawn from the same population?
- Adjustments applied to the statistic of the original data to correct inflated signals due to overfitting.
- A simple approach to correct the inflated Chi-square test statistic

$$\chi^2_{corrected} = \frac{1}{M}\sum_m \left(\frac{\chi^2_{test_m}}{\chi^2_{training_m}}\right)\cdot\chi^2_{orig}$$

M: number of bootstrap samples (M=1000)

Corrected statistic



Bootstrap vs. Cross-Validation in overfitting correction

- Bootstrap overfitting-corrected estimates of model performance can also be biased in favor of the model. Although cross-validation methods are less biased than the bootstrap, Efron showed that it in fact has much higher variance in estimating overfitting-corrected predictive accuracy than bootstrapping. In other words, cross-validation methods can yield significantly different estimates when the entire validation process is repeated.

Does your model overfit your data?

Rule of thumb to shrinkage estimate

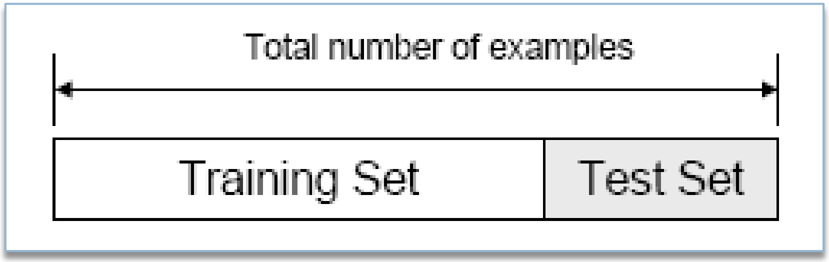
- Rough rule of thumb: 1:10 or 1:15
  - Number of candidate predictors
  - Limiting sample size
- Use shrinkage estimate (quantification of the amount of overfitting present)
  - Heuristic formula

$$\hat{\gamma} = (\chi^2 - p) / \chi^2$$

  - Bootstrap
  - If $\hat{\gamma}$ falls below 0.9, then lack of calibration on new data is very likely

Model validation – data splitting

- Data splitting

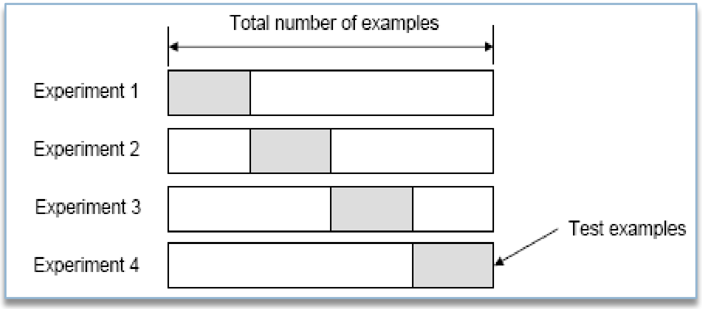


  - Simple: modeling steps are only done once.
  - Problem: holding back data from model fitting results in lower precision and power!!

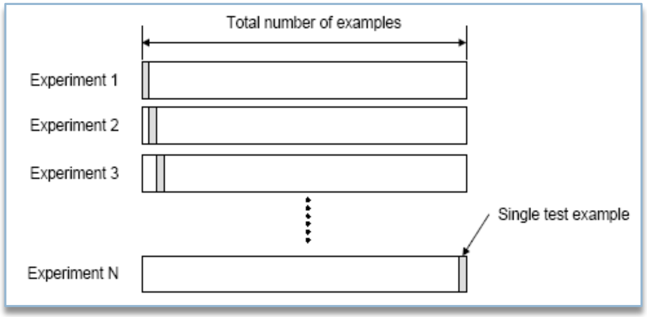
Model validation – cross validation

- Cross validation (repeated data splitting)

K-fold cross-validation:



Leave-one-out cross-validation:



  - Less data are discarded from the estimation process. Reduces variability by not relying on a single sample split.
  - Cross-validation is relatively inefficient. Data splitting is even worse

Model validation – bootstrap

- Bootstrap

The ordinary bootstrap:

.632 and .632+ bootstrap:

- Nearly unbiased estimates of predictive accuracy that are of relatively low variance.
- Entire dataset is used for model development

Model validation – bootstrap cont.

- Original data: c-statistic
- Bootstrap samples 1 to 200,
  - we will have $model_i$, c_i
  - Apply $model_i$ on original data, we get c_i'
  - $optimism_i = c_i - c_i'$
- c_{honest}=c-mean(optimism)
- Penalized for overfitting

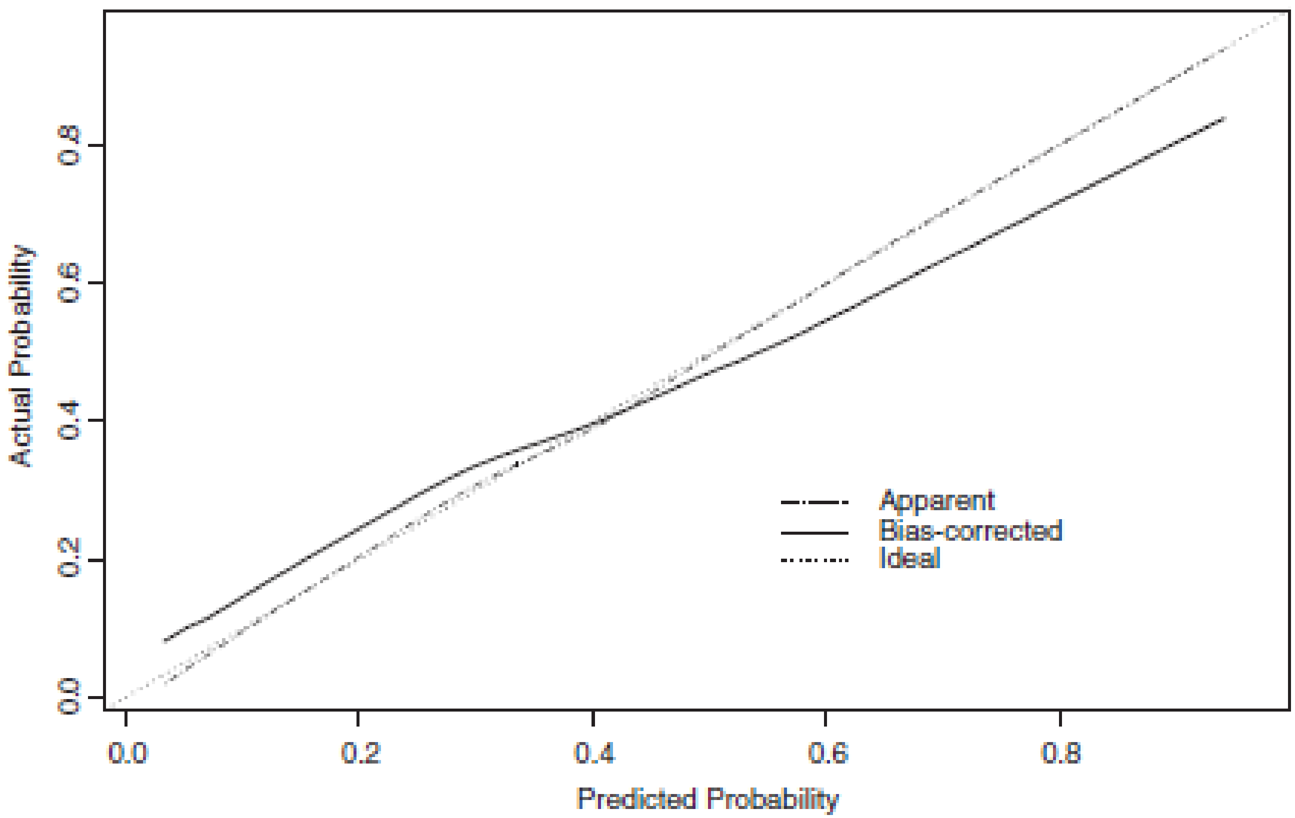
Model validation – example

- Cross validation and bootstrap validation can be easily done using ***validate( )*** function in ***Design*** package.
- Example: a fitted logistic regression with df 12, $\min(n_1,n_2)=105$

```
> validate(f, B=200)
           index.orig     training       test   optimism index.corrected   n
Dxy       0.400776014  0.474314595 0.35445926  0.11985534      0.280920679 200
R2        0.168742420  0.229191997 0.12082715  0.10836484      0.060377578 200
Intercept 0.000000000  0.000000000 0.09772866 -0.09772866      0.097728664 200
Slope     1.000000000  1.000000000 0.64964887  0.35035113      0.649648870 200
Emax      0.000000000  0.000000000 0.11327079  0.11327079      0.113270787 200
D         0.130383226  0.184272382 0.09045234  0.09382004      0.036563186 200
U        -0.008333333 -0.008333333 0.02917206 -0.03750539      0.029172059 200
Q         0.138716560  0.192605715 0.06128028  0.13132543      0.007391127 200
B         0.213888041  0.201883214 0.22594768 -0.02406446      0.237952506 200
```

Calibration

- Agreement between predictions and actual values.
- Calibration plot for binary outcome



Borrowed from K.G.M. Moons, etc. 2004

Prevent overfitting – increase n

- Increase sample size
  - Collect more data
  - Impute missing values
- Use a simpler model with fewer predictors.
  - Rough guide: at least 10 or 15 observations per variable

Prevent overfitting – data reduction

- Use a simpler model with fewer predictors.
  - Data reduction
    - Subject matter knowledge
      - Cost
      - Measure reliability
      - Other studies
    - No data "peeking"
      - Variable screening, stepwise variable selections are not viable
    - Principle Components
    - Variable Clustering
    - Well developed scores or index
- Training with noise (Bishop C M, 1991)

Prevent overfitting – uniform shrinkage

- Uniform shrinkage
  - heuristic formula

$$\beta^* = \hat{\gamma} \times \beta$$

  - Afterwards, uniform shrink

Prevent overfitting – PMLE

- Penalized maximum likelihood estimation (PMLE)
  - Maximizes the penalized log-likelihood

$$PML = \log L - 0.5\lambda \sum (s_i \beta_i)^2$$

- Shrinkage is done during the fitting of the model
- Example: 27 candidate predictors, 170 events

Selected predictors	1. No shrinkage	2. With shrinkage	3. PMLE	Shrinkage per predictor obtained by PMLE[a]
Age (per 10 year)	0.19	0.14	0.17	0.89
Quetelet index (per kg/m^2)	0.17	0.12	0.06	0.35
Days of immobilization (per day)	0.036	0.026	0.026	0.72
Days of symptoms (per day)	−0.020	−0.015	−0.016	0.80
Pain in legs	0.74	0.54	0.60	0.81
Coughing	0.70	0.51	0.52	0.74
Wheezing	−0.93	−0.68	−0.57	0.61
Collapse	1.50	1.10	0.82	0.55
Breathing frequency (breaths/minute)	0.063	0.046	0.058	0.92
Abnormal leg ultrasound	0.96	0.70	0.74	0.77
Abnormal chest X-ray	0.69	0.50	0.57	0.83
Surgery within past 3 months	—[b]	—[b]	0.38	
Crepitations	—[b]	—[b]	−0.31	
ROC area	0.78	0.72	0.75	0.96

Borrowed from K.G.M. Moons, etc. 2004

Prevent overfitting – PMLE

- Can be easily done using ***pentrace()*** and ***update()*** in ***Design*** package

```
> f <- lrm(y ~ blood.pressure + sex * (age + rcs(cholesterol,4)), x=TRUE, y=TRUE)
> p <- pentrace(f, seq(0,2,by=.05))
> f.pen <- update(f, penalty=p$penalty)
```

Prevent overfitting – PMLE

```
> f

Logistic Regression Model

lrm(formula = y ~ blood.pressure + sex * (age + rcs(cholesterol,
    4)), x = TRUE, y = TRUE)


Frequencies of Responses
  0   1
105 135

       Obs  Max Deriv Model L.R.       d.f.          P          C        Dxy      Gamma
       240      2e-07      30.88         10      6e-04      0.697      0.393      0.394
     Tau-a         R2      Brier
     0.194      0.162      0.215

                         Coef      S.E.     Wald Z P
Intercept                 0.462773 6.292137  0.07  0.9414
blood.pressure           -0.019496 0.009375 -2.08  0.0376
sex=male                 -9.517820 8.068298 -1.18  0.2381
age                       0.057228 0.022133  2.59  0.0097
cholesterol              -0.003360 0.035373 -0.09  0.9243
cholesterol'             -0.058262 0.093678 -0.62  0.5340
cholesterol''             0.323125 0.352994  0.92  0.3600
sex=male * age           -0.009549 0.029674 -0.32  0.7476
sex=male * cholesterol    0.056666 0.045251  1.25  0.2105
sex=male * cholesterol'  -0.025193 0.123759 -0.20  0.8387
sex=male * cholesterol'' -0.090383 0.471480 -0.19  0.8480
```

Prevent overfitting – PMLE

```
> f.pen

Logistic Regression Model

lrm(formula = y ~ blood.pressure + sex * (age + rcs(cholesterol,
    4)), x = TRUE, y = TRUE, penalty = p$penalty)


Frequencies of Responses
  0   1
105 135

Penalty factors:

 simple nonlinear interaction nonlinear.interaction
   0.15      0.15        0.15                  0.15

Final penalty on -2 log L: 1.75

       Obs  Max Deriv Model L.R.       d.f.          P          C        Dxy      Gamma
       240      2e-08      29.26       8.39      4e-04      0.694      0.389       0.39
     Tau-a         R2      Brier
     0.192      0.145      0.216

                         Coef     S.E.     Wald Z P      Penalty Scale
Intercept                -2.51449 3.605054 -0.70  0.4855  0.0000
blood.pressure           -0.01997 0.009189 -2.17  0.0297  6.0068
sex=male                 -1.91368 2.615585 -0.73  0.4644  0.2739
age                       0.05967 0.020995  2.84  0.0045  3.8931
cholesterol               0.01244 0.019051  0.65  0.5138  9.8336
cholesterol'             -0.07056 0.050985 -1.38  0.1664  9.2858
cholesterol''             0.31612 0.209110  1.51  0.1306  1.7557
sex=male * age           -0.01573 0.026834 -0.59  0.5578 10.0495
sex=male * cholesterol    0.01540 0.014533  1.06  0.2891 39.0175
sex=male * cholesterol'   0.04548 0.053696  0.85  0.3970  7.8277
sex=male * cholesterol'' -0.27057 0.250903 -1.08  0.2809  1.3836
```

Prevent overfitting – Lasso

- Lasso
  - Least absolute shrinkage and selection operator
  - PMLE with a restriction on the sum of the absolute coefficients.
  - Can be done in ***glmpath*** and some other packages in R or PROC GLMSELECT in SAS

Is what you see what you get?

Questions and comments?

Thanks!