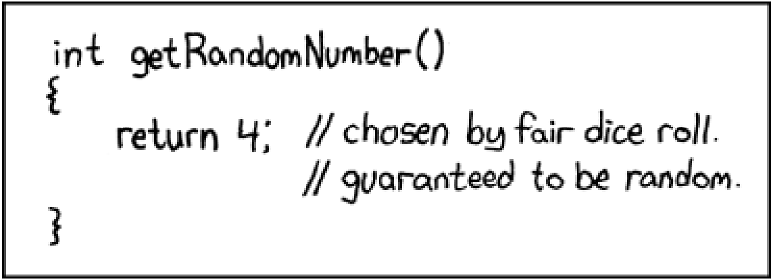
STAT406- Methods of Statistical Learning Lecture 3

Matias Salibian-Barrera

UBC - Sep / Dec 2017



http://xkcd.com/221/

Cross validation

- We can show that

$$E_{\text{data},Y|\mathbf{X}_0}\left[\left(Y-\hat{f}(\mathbf{X}_0)\right)^2\right] \\ = V(\hat{f}(\mathbf{X}_0)) + B^2\left(\hat{f}(\mathbf{X}_0)\right) + V(\epsilon)\,,$$

where V denotes variance and

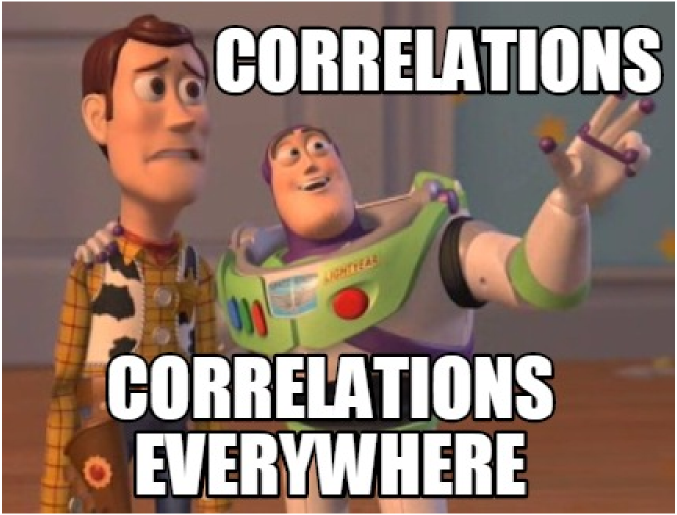
$$B^2\left(\hat{f}(\mathbf{X}_0)\right) = \left[E_{\text{data}}\left(\hat{f}(\mathbf{X}_0) - f(\mathbf{X}_0)\right)\right]^2 .$$

is the squared bias:

Activity!

Cross validation

- Conservative lower bound for MSPE
- Discussion
- Proper way of using CV


CORRELATIONS
CORRELATIONS EVERYWHERE

Model / feature selection

- Simple example:

```
set.seed(123)
x1 <- rnorm(506)
x2 <- rnorm(506, mean=2, sd=1)
x3 <- rexp(506, rate=1)
x4 <- x2 + rnorm(506, sd=.1)
x5 <- x1 + rnorm(506, sd=.1)
x6 <- x1 - x2 + rnorm(506, sd=.1)
x7 <- x1 + x3 + rnorm(506, sd=.1)
y <- x1*3 + x2/3 + rnorm(506, sd=2.2)
```

- Variables X_1 and X_2 are clearly important. But they are also highly correlated to X_4, X_5, X_6 and X_7.

Model / feature selection

- However, nothing is significant?

```
> summary(lm(y~., data=x))

Call:
lm(formula = y ~ ., data = x)
Residuals:
   Min     1Q Median     3Q    Max
-6.882 -1.474 -0.033  1.415  5.823

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.03457    0.23018   0.150   0.8807
x1           3.22612    1.68088   1.919   0.0555 .
x2           0.23867    1.39355   0.171   0.8641
x3          -0.35926    0.98680  -0.364   0.7160
x4          -0.69359    0.99025  -0.700   0.4840
x5           0.09271    0.91162   0.102   0.9190
x6          -0.73887    1.01114  -0.731   0.4653
x7           0.31651    0.98610   0.321   0.7484
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.148 on 498 degrees of freedom
Multiple R-squared: 0.6353,Adjusted R-squared: 0.6302
F-statistic: 123.9 on 7 and 498 DF,  p-value: < 2.2e-16
```

Model / feature selection

- But...

```
> summary(lm(y~x1+x2, data=x))

Call:
lm(formula = y ~ x1 + x2, data = x)

Residuals:
    Min      1Q  Median      3Q     Max
-6.9303 -1.5736 -0.0068  1.3840  5.9567

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.00733    0.20900   0.035  0.97204
x1           2.89168    0.09806  29.490  < 2e-16 ***
x2           0.27903    0.09249   3.017  0.00268 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.141 on 503 degrees of freedom
Multiple R-squared: 0.6343,Adjusted R-squared: 0.6328
F-statistic: 436.2 on 2 and 503 DF,  p-value: < 2.2e-16
```

Model / feature selection

- Even worse...

```
> summary(lm(y~x1+x2+x4, data=x))

Call:
lm(formula = y ~ x1 + x2 + x4, data = x)

Residuals:
    Min      1Q  Median      3Q     Max
-6.8064 -1.5229 -0.0308  1.4226  5.8861

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.0001127  0.2093588   0.001    1.000
x1           2.8964461  0.0983390  29.454   <2e-16 ***
x2           0.9740807  0.9917783   0.982    0.326
x4          -0.6934442  0.9851714  -0.704    0.482
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.142 on 502 degrees of freedom
Multiple R-squared: 0.6347,Adjusted R-squared: 0.6325
F-statistic: 290.7 on 3 and 502 DF,  p-value: < 2.2e-16
```

Discussion points

- Correlated covariates are now prevalent
- Researchers can (and do) collect data "blindly"
- Sometimes, data are collected without a specific question in mind

Discussion points

- Correlated covariates:
- Mask each other when included simultaneously in a model
- May reduce prediction accuracy

Model / feature selection

One strategy:

- (1): Select models to be considered
- (2): Select a quantitative criterion to compare them (e.g. AIC, C_p, CV-based $\widehat{\text{MSPE}}$)
- (3): Choose a strategy to explore the models under consideration

Model / feature selection

For example:

- (1): Consider all possible models
- (2): Use AIC to compare them
- (3): Best subset search (2^p fits!)
- (3’): Stepwise search

- Is this strategy prediction-based?

AIC?

Why not compare models using residual sum of squares, or R^2?