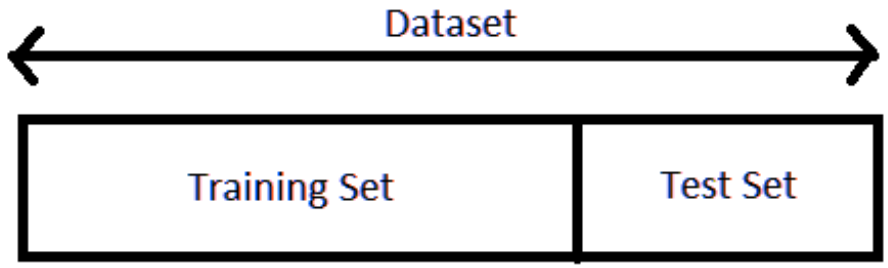
Building the "best" model

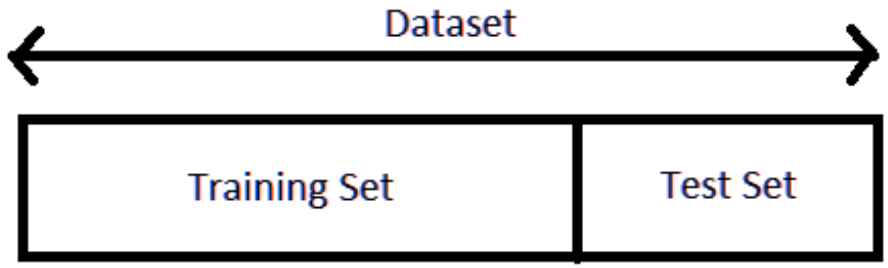
Roberta De Vito

BROWN
Public Health

Division of the data set



Division of the data set



Evaluate performance on an independent validation set (e.g. using cross-validation): Q1.

Evaluation measures

The training set performance gets better as model complexity increases (e.g. likelihood, sum of squares). Two common solutions:

- penalization term to the goodness of fit
- adjusted R^2 involving the term $n - p - 1$

$$R^2_{Adjusted} = 1 - \frac{\frac{RSS}{n-p-1}}{\frac{TSS}{n-1}}.$$

Alternative fitting procedures

Why?

1. Prediction Accuracy: low bias,
   - $n \gg p$: low variance
   - $n > p$ (not so larger): poor prediction
   - $p > n$: no unique estimates
2. Model Interpretability

Alternative fitting procedures

Why?

1. Prediction Accuracy: low bias,
   - $n \gg p$: low variance
   - $n > p$ (not so larger): poor prediction
   - $p > n$: no unique estimates
2. Model Interpretability

Three important methods

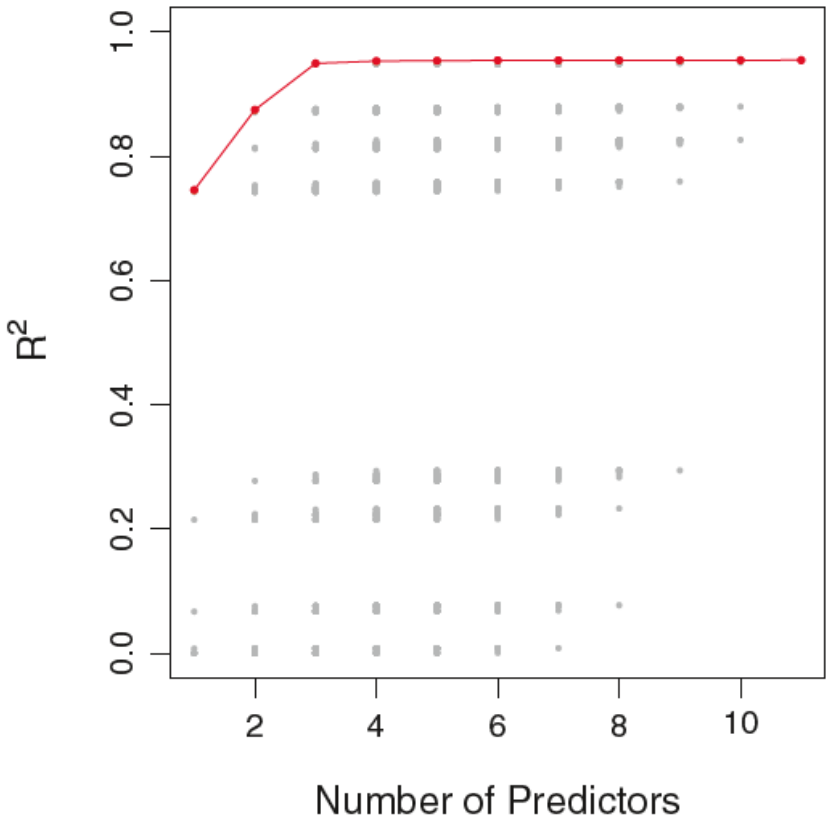
1. Subset Selection
2. Shrinkage
3. Dimension Reduction

Best Subset Selection: Q2

Run a regression for each possible combination of predictors and choose best set

1. Let $\mathcal{M}_0$ denote null model only including the intercept.
2. For $k = 1, 2, \ldots, p$:
   - 2.1 Fit all $\binom{p}{k}$ models that contain exactly k predictors
   - 2.2 Pick best among these $\binom{p}{k}$ models, and call it $\mathcal{M}_k$. Here best is defined as having smallest deviance (-2 times log likelihood).
3. Select single best model from among $\mathcal{M}_0, \ldots, \mathcal{M}_p$ using some evaluation measure e.g. cross-validated prediction error, C_p, AIC, BIC, or adjusted R^2

Best Situation



Deviance based measures

The deviance is defined as

$$Deviance = 2\left(log(Likelihood_f) - log(Likelihood_c)\right)$$

Model selection techniques often minimize

$$\text{Deviance} + \text{penalization term}.$$

For linear regression deviance is equal to RSS/σ^2.

Mallow's C_p and Akaike Information Criterion (AIC)

- For linear regression, Mallow's C_p statistic is

$$C_p = \frac{1}{n}(RSS + 2k\hat{\sigma}^2)$$

- Akaike Information Criterion (AIC)

$$AIC = -2logL + 2k$$

where k is the number of covariates or parameters

Bayesian Information Criterion (BIC)

For linear regression

$$BIC = (RSS + \log(n)k\hat{\sigma}^2)$$

For generalized linear models

$$BIC = (-2\log(L) + \log(n)k)$$

.

AIC, BIC and R-squared in practice: Q3

```
> summary(lm(mort~so2+educ+nonw))

Call:
lm(formula = mort ~ so2 + educ + nonw)

Residuals:
    Min      1Q  Median      3Q     Max
-94.201 -19.410   1.294  16.537  92.986

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 1156.06487   71.68018  16.128  < 2e-16 ***
so2            0.25699    0.08298   3.097 0.003054 **
educ         -24.92413    6.28208  -3.967 0.000209 ***
nonw           3.70485    0.58615   6.321 4.55e-08 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 39.02 on 56 degrees of freedom
Multiple R-squared:  0.6266,  Adjusted R-squared:  0.6066
F-statistic: 31.33 on 3 and 56 DF,  p-value: 5.063e-12

> summary(lm(mort~so2+educ+nonw+poorind))

Call:
lm(formula = mort ~ so2 + educ + nonw + poorind)

Residuals:
     Min       1Q   Median       3Q      Max
-104.330  -21.139   -0.268   17.123  110.441

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 1212.83661   81.76939  14.832  < 2e-16 ***
so2            0.23240    0.08412   2.763 0.007778 **
educ         -29.90378    7.16777  -4.172 0.000108 ***
nonw           4.20543    0.68183   6.168 8.61e-08 ***
poorind      -20.24420   14.42185  -1.404 0.166024
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 38.68 on 55 degrees of freedom
Multiple R-squared:  0.6395,  Adjusted R-squared:  0.6133
F-statistic:  24.4 on 4 and 55 DF,  p-value: 1.209e-11
```

AIC, BIC and R-squared in practice: Q3

```
> AIC(lm(mort~so2+educ+nonw))
[1] 615.8069
> BIC(lm(mort~so2+educ+nonw))
[1] 626.2786
> AIC(lm(mort~so2+educ+nonw+poorind))
[1] 615.695
> BIC(lm(mort~so2+educ+nonw+poorind))
[1] 628.261
```

Best Subset Selection

Run a regression for each possible combination of predictors and choose best set

1. Let $\mathcal{M}_0$ denote null model only including the intercept.
2. For $k = 1, 2, \ldots, p$:
   - 2.1 Fit all $\binom{p}{k}$ models that contain exactly k predictors
   - 2.2 Pick best among these $\binom{p}{k}$ models, and call it $\mathcal{M}_k$. Here best is defined as having smallest deviance (-2 times log likelihood).
3. Select single best model from among $\mathcal{M}_0, \ldots, \mathcal{M}_p$ using some evaluation measure e.g. cross-validated prediction error, C_p, AIC, BIC, or adjusted R^2

Forward Stepwise Selection

Aim: choose $K \leq P$

$$y_i = \beta_0 + \epsilon_i$$

Forward Stepwise Selection

Aim: choose $K \leq P$

$$y_i = \beta_0 + \beta_1 x_{i1} + \epsilon_i$$

Forward Stepwise Selection

Aim: choose $K \leq P$

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \epsilon_i$$

Forward Stepwise Selection

Aim: choose $K \leq P$

$$y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ik} + \epsilon_i$$

Backward Stepwise Selection

Aim: choose $K \leq P$

$$y_i = \beta_0 + \cdots + \beta_p x_{ip} + \epsilon_i$$

Backward Stepwise Selection

Aim: choose $K \leq P$

$$y_i = \beta_0 + \cdots + \beta_{p-1} x_{i(p-1)} \epsilon_i$$

Backward Stepwise Selection

Aim: choose $K \leq P$

$$y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ik} + \epsilon_i$$

Forward Stepwise Selection in practice

```
> fit1 <- lm(mort~., data= data)
> fit2 <- lm(mort ~ 1, data=data)
> mod_forward <- stepAIC(fit2,direction="forward",scope=list(upper=fit1,lower=fit2))
Start:  AIC=496.65
mort ~ 1

        Df Sum of Sq    RSS    AIC
+ nonw   1     94613 133695 466.54
+ educ   1     59612 168696 480.49
+ prec   1     59266 169041 480.61
+ hous   1     41592 186716 486.58
+ so2    1     41411 186896 486.64
+ poor   1     38470 189838 487.57
+ popn   1     29149 199159 490.45
+ wwdrk  1     18518 209789 493.57
+ jult   1     17520 210788 493.86
+ dens   1     16093 212214 494.26
<none>               228308 496.65
+ hc     1      7172 221136 496.73
+ ovr65  1      6960 221347 496.79
+ humid  1      1788 226520 498.17
+ nox    1      1367 226941 498.29
+ jant   1       206 228102 498.59

Step:  AIC=466.54
mort ~ nonw

        Df Sum of Sq    RSS    AIC
+ educ   1     33853  99841 451.02
+ jant   1     29835 103859 453.39
+ so2    1     24492 109203 456.40
+ ovr65  1     21435 112259 458.05
```

Forward Stepwise Selection in practice

```
> summary(mod_forward)

Call:
lm(formula = mort ~ nonw + educ + jant + so2 + prec + jult +
    popn, data = data)

Residuals:
    Min      1Q  Median      3Q     Max
-76.748 -21.036  -3.989  15.555  92.458

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 1429.18663  215.75678   6.624 1.97e-08 ***
nonw           5.21614    0.82709   6.307 6.29e-08 ***
educ         -16.96562    6.81958  -2.488  0.01610 *
jant          -1.89340    0.58900  -3.215  0.00225 **
so2            0.22529    0.08156   2.762  0.00792 **
prec           1.64850    0.60345   2.732  0.00858 **
jult          -2.30061    1.23395  -1.864  0.06791 .
popn         -62.01179   44.78161  -1.385  0.17204
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 33.51 on 52 degrees of freedom
Multiple R-squared:  0.7443,  Adjusted R-squared:  0.7098
F-statistic: 21.62 on 7 and 52 DF,  p-value: 2.409e-13
```

Backward Stepwise Selection in practice

```
> fit1 <- lm(mort~., data= data)
> mod_back <- step(fit1, direction= 'backward')
Start:  AIC=439.79
mort ~ prec + jant + jult + ovr65 + popn + educ + hous + dens +
    nonw + wwdrk + poor + hc + nox + so2 + humid

        Df Sum of Sq   RSS    AIC
- poor   1       3.3 53683 437.79
- humid  1      10.2 53690 437.80
- wwdrk  1      15.5 53695 437.80
- hous   1     165.5 53846 437.97
- so2    1     417.1 54097 438.25
- dens   1     975.2 54655 438.87
- ovr65  1    1392.2 55072 439.32
<none>                53680 439.79
- nox    1    2166.7 55847 440.16
- hc     1    2286.1 55966 440.29
- educ   1    2553.2 56233 440.58
- popn   1    2859.3 56539 440.90
- jult   1    3243.0 56923 441.31
- jant   1    3728.4 57408 441.82
- prec   1    5191.0 58871 443.33
- nonw   1   13774.7 67455 451.49

Step:  AIC=437.79
mort ~ prec + jant + jult + ovr65 + popn + educ + hous + dens +
    nonw + wwdrk + hc + nox + so2 + humid

        Df Sum of Sq   RSS    AIC
- humid  1      12.7 53696 435.80
- wwdrk  1      13.3 53697 435.81
- hous   1     215.2 53898 436.03
- so2    1     425.8 54109 436.26
```

Backward Stepwise Selection in practice

```
> summary(mod_back)

Call:
lm(formula = mort ~ prec + jant + jult + ovr65 + popn + educ +
    nonw + hc + nox, data = data)

Residuals:
    Min      1Q  Median      3Q     Max
-71.084 -22.005   0.678  16.881  79.767

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 1934.0539   333.4957   5.799 4.48e-07 ***
prec           1.8565     0.8373   2.217  0.03118 *
jant          -2.2620     0.6957  -3.252  0.00206 **
jult          -3.3200     1.3971  -2.376  0.02135 *
ovr65        -10.9205     7.1398  -1.530  0.13243
popn        -137.3831    59.7746  -2.298  0.02576 *
educ         -23.4211     7.0619  -3.317  0.00170 **
nonw           4.6623     0.9689   4.812 1.42e-05 ***
hc            -0.9221     0.3192  -2.889  0.00571 **
nox            1.8710     0.6095   3.070  0.00346 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 33.27 on 50 degrees of freedom
Multiple R-squared:  0.7575,  Adjusted R-squared:  0.7139
F-statistic: 17.36 on 9 and 50 DF,  p-value: 1.481e-12
```

Problems with the two procedures

Example: Pollution data set

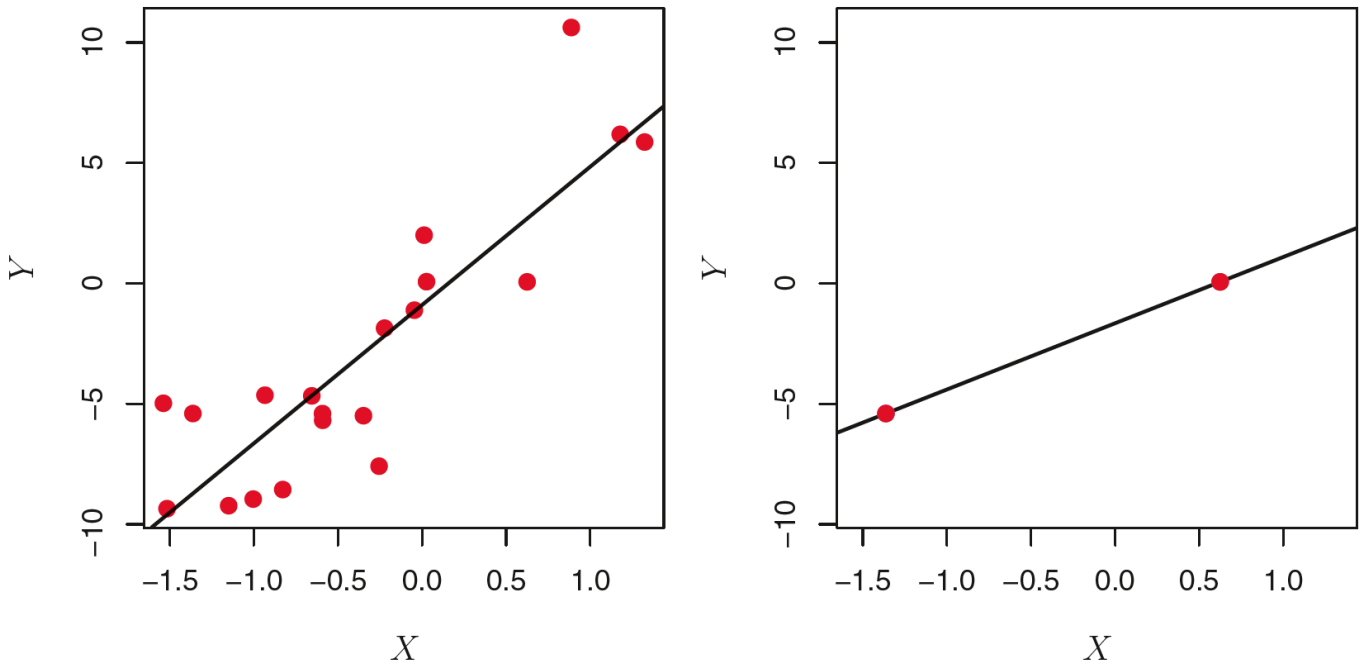
`ID`	code for the identification of the sample
`OVR65`	% of 1960 SMSA population aged 65 or older
`JANT`	Average January temperature in degrees F
`JULT`	Same for July
`EDUC`	Median school years completed by those over 22
`HOUS`	% of housing units with all facilities
`DENS`	Population per sq. mile in urbanized areas, 1960
`NONW`	% non-white population in urbanized areas, 1960
`WWDRK`	% employed in white collar occupations
`POOR`	% of families with income < 3000
`HC`	Relative hydrocarbon pollution potential
`NOX`	Same for nitric oxides
`SO2`	Same for sulphur dioxide
`HUMID`	Annual average % relative humidity at 1pm
`MORT`	Total age-adjusted mortality rate per 100,000
`PREC`	Average annual precipitation in inches

Forward VS Backward Stepwise Selection

```
> summary(mod_forward)

Call:
lm(formula = mort ~ nonw + educ + jant + so2 + prec + jult +
    popn, data = data)

Residuals:
    Min      1Q  Median      3Q     Max
-76.748 -21.036  -3.989  15.555  92.458

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 1429.18663  215.75678   6.624 1.97e-08 ***
nonw           5.21614    0.82709   6.307 6.29e-08 ***
educ         -16.96562    6.81958  -2.488  0.01610 *
jant          -1.89340    0.58900  -3.215  0.00225 **
so2            0.22529    0.08156   2.762  0.00792 **
prec           1.64850    0.60345   2.732  0.00858 **
jult          -2.30061    1.23395  -1.864  0.06791 .
popn         -62.01179   44.78161  -1.385  0.17204
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 33.51 on 52 degrees of freedom
Multiple R-squared:  0.7443,  Adjusted R-squared:  0.7098
F-statistic: 21.62 on 7 and 52 DF,  p-value: 2.409e-13
```

```
> summary(mod_back)

Call:
lm(formula = mort ~ prec + jant + jult + ovr65 + popn + ed
    nonw + hc + nox, data = data)

Residuals:
    Min      1Q  Median      3Q     Max
-71.084 -22.005   0.678  16.881  79.767

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 1934.0539   333.4957   5.799 4.48e-07 ***
prec           1.8565     0.8373   2.217  0.03118 *
jant          -2.2620     0.6957  -3.252  0.00206 **
jult          -3.3200     1.3971  -2.376  0.02135 *
ovr65        -10.9205     7.1398  -1.530  0.13243
popn        -137.3831    59.7746  -2.298  0.02576 *
educ         -23.4211     7.0619  -3.317  0.00170 **
nonw           4.6623     0.9689   4.812 1.42e-05 ***
hc            -0.9221     0.3192  -2.889  0.00571 **
nox            1.8710     0.6095   3.070  0.00346 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘

Residual standard error: 33.27 on 50 degrees of freedom
Multiple R-squared:  0.7575,  Adjusted R-squared:  0.7139
F-statistic: 17.36 on 9 and 50 DF,  p-value: 1.481e-12
```

What Goes Wrong in High Dimensions?



What Goes Wrong in High Dimensions?