Model Selection and Inference

Peter Hoff

STAT 423

Applied Regression and Analysis of Variance

University of Washington

Model selection

Diabetes example:

- 342 subjects
- y_i = diabetes progression
- $\boldsymbol{x}_i$ = explanatory variables.

Each $\boldsymbol{x}_i$ includes

- 13 subject specific measurements ($x_{\text{age}}, x_{\text{sex}}, \ldots$);
- $78 = \binom{13}{2}$ interaction terms ($x_{\text{age}} \cdot x_{\text{sex}}, \ldots$) ;
- 9 quadratic terms (x_{sex} and three genetic variables are binary)

100 explanatory variables total!

Model selection

Diabetes example:

- 342 subjects
- y_i = diabetes progression
- $\boldsymbol{x}_i$ = explanatory variables.

Each $\boldsymbol{x}_i$ includes

- 13 subject specific measurements ($x_{\text{age}}, x_{\text{sex}}, \ldots$);
- $78 = \binom{13}{2}$ interaction terms ($x_{\text{age}} \cdot x_{\text{sex}}, \ldots$) ;
- 9 quadratic terms (x_{sex} and three genetic variables are binary)

100 explanatory variables total!

Model selection

Diabetes example:

- 342 subjects
- y_i = diabetes progression
- $\boldsymbol{x}_i$ = explanatory variables.

Each $\boldsymbol{x}_i$ includes

- 13 subject specific measurements ($x_{\text{age}}, x_{\text{sex}}, \ldots$);
- $78 = \binom{13}{2}$ interaction terms ($x_{\text{age}} \cdot x_{\text{sex}}, \ldots$) ;
- 9 quadratic terms (x_{sex} and three genetic variables are binary)

100 explanatory variables total!

Model selection

Diabetes example:

- 342 subjects
- y_i = diabetes progression
- $\boldsymbol{x}_i$ = explanatory variables.

Each $\boldsymbol{x}_i$ includes

- 13 subject specific measurements ($x_{\text{age}}, x_{\text{sex}}, \ldots$);
- $78 = \binom{13}{2}$ interaction terms ($x_{\text{age}} \cdot x_{\text{sex}}, \ldots$) ;
- 9 quadratic terms (x_{sex} and three genetic variables are binary)

100 explanatory variables total!

Model selection

Diabetes example:

- 342 subjects
- y_i = diabetes progression
- $\boldsymbol{x}_i$ = explanatory variables.

Each $\boldsymbol{x}_i$ includes

- 13 subject specific measurements $(x_{\text{age}}, x_{\text{sex}}, \ldots)$;
- $78 = \binom{13}{2}$ interaction terms $(x_{\text{age}} \cdot x_{\text{sex}}, \ldots)$;
- 9 quadratic terms (x_{sex} and three genetic variables are binary)

100 explanatory variables total!

Model selection

Diabetes example:

- 342 subjects
- y_i = diabetes progression
- $\boldsymbol{x}_i$ = explanatory variables.

Each $\boldsymbol{x}_i$ includes

- 13 subject specific measurements $(x_{\text{age}}, x_{\text{sex}}, \ldots)$;
- $78 = \binom{13}{2}$ interaction terms $(x_{\text{age}} \cdot x_{\text{sex}}, \ldots)$;
- 9 quadratic terms (x_{sex} and three genetic variables are binary)

100 explanatory variables total!

Model selection

Diabetes example:

- 342 subjects
- y_i = diabetes progression
- $\boldsymbol{x}_i$ = explanatory variables.

Each $\boldsymbol{x}_i$ includes

- 13 subject specific measurements $(x_{\text{age}}, x_{\text{sex}}, \ldots)$;
- $78 = \binom{13}{2}$ interaction terms $(x_{\text{age}} \cdot x_{\text{sex}}, \ldots)$;
- 9 quadratic terms (x_{sex} and three genetic variables are binary)

100 explanatory variables total!

Model selection

Diabetes example:

- 342 subjects
- y_i = diabetes progression
- $\boldsymbol{x}_i$ = explanatory variables.

Each $\boldsymbol{x}_i$ includes

- 13 subject specific measurements $(x_{\text{age}}, x_{\text{sex}}, \ldots)$;
- $78 = \binom{13}{2}$ interaction terms $(x_{\text{age}} \cdot x_{\text{sex}}, \ldots)$;
- 9 quadratic terms (x_{sex} and three genetic variables are binary)

100 explanatory variables total!

Model selection

Diabetes example:

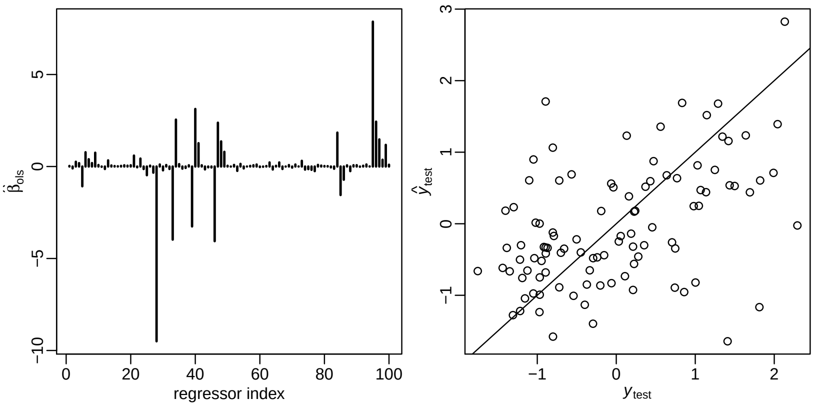
- 342 subjects
- y_i = diabetes progression
- $\boldsymbol{x}_i$ = explanatory variables.

Each $\boldsymbol{x}_i$ includes

- 13 subject specific measurements $(x_{\text{age}}, x_{\text{sex}}, \ldots)$;
- $78 = \binom{13}{2}$ interaction terms $(x_{\text{age}} \cdot x_{\text{sex}}, \ldots)$;
- 9 quadratic terms (x_{sex} and three genetic variables are binary)

100 explanatory variables total!

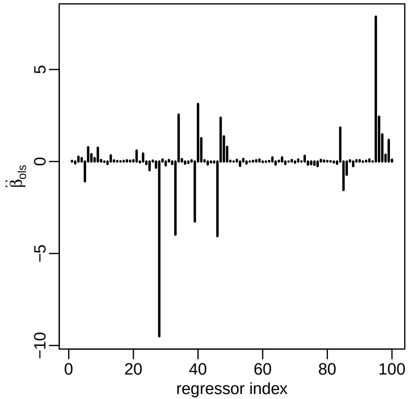
OLS regression

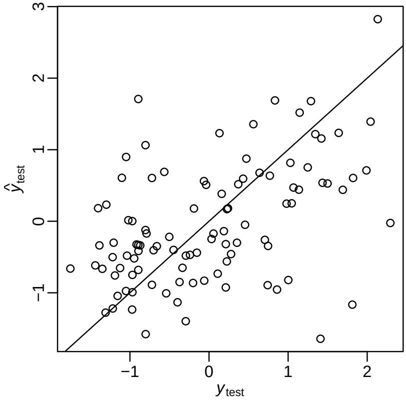


$$\frac{1}{100}\sum(y_{\text{test},i}-\hat{y}_{\text{test},i})^2 = 0.9263277$$

$$\frac{1}{100}\sum(y_{\text{test},i}-0)^2 = \sum y_{\text{test},i}^2 = 1.0094689$$

OLS regression

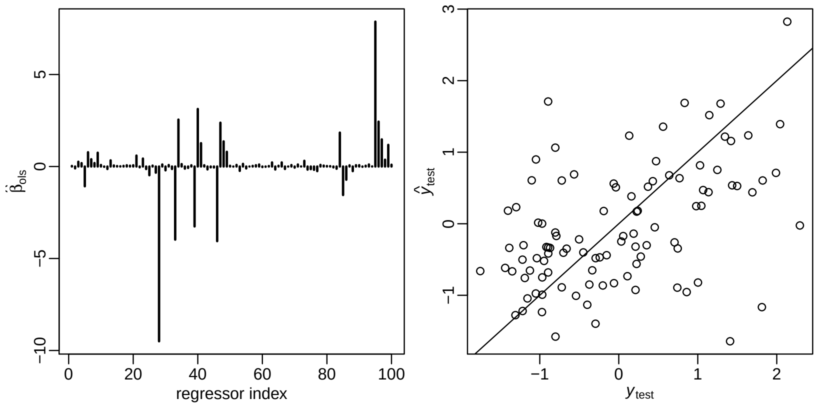




$$\frac{1}{100}\sum(y_{\text{test},i}-\hat{y}_{\text{test},i})^2 = 0.9263277$$

$$\frac{1}{100}\sum(y_{\text{test},i}-0)^2 = \sum y_{\text{test},i}^2 = 1.0094689$$

OLS regression



$$\frac{1}{100}\sum(y_{\text{test},i} - \hat{y}_{\text{test},i})^2 = 0.9263277$$

$$\frac{1}{100}\sum(y_{\text{test},i} - 0)^2 = \sum y_{\text{test},i}^2 = 1.0094689$$

Backwards elimination

1. Obtain the estimator $\hat{\boldsymbol{\beta}}_{\text{ols}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\boldsymbol{y}$ and its t-statistics.
2. If there are any regressors j such that $|t_j| < t_{\text{cutoff}}$,
   2.1 find the regressor $j_{\min}$ having the smallest value of $|t_j|$ and remove column $j_{\min}$ from $\mathbf{X}$.
   2.2 return to step 1.
3. If $|t_j| > t_{\text{cutoff}}$ for all variables j remaining in the model, then stop.

Backwards elimination

1. Obtain the estimator $\hat{\boldsymbol{\beta}}_{\text{ols}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\boldsymbol{y}$ and its t-statistics.
2. If there are any regressors j such that $|t_j| < t_{\text{cutoff}}$,
    2.1 find the regressor $j_{\min}$ having the smallest value of $|t_j|$ and remove column $j_{\min}$ from $\mathbf{X}$.
    2.2 return to step 1.
3. If $|t_j| > t_{\text{cutoff}}$ for all variables j remaining in the model, then stop.

Backwards elimination

1. Obtain the estimator $\hat{\beta}_{\mathrm{ols}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\boldsymbol{y}$ and its t-statistics.
2. If there are any regressors j such that $|t_j| < t_{\mathrm{cutoff}}$,
   2.1 find the regressor j_{min} having the smallest value of $|t_j|$ and remove column j_{min} from $\mathbf{X}$.
   2.2 return to step 1.
3. If $|t_j| > t_{\mathrm{cutoff}}$ for all variables j remaining in the model, then stop.

Backwards elimination

1. Obtain the estimator $\hat{\beta}_{\text{ols}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\boldsymbol{y}$ and its t-statistics.
2. If there are any regressors j such that $|t_j| < t_{\text{cutoff}}$,
   2.1 find the regressor $j_{\min}$ having the smallest value of $|t_j|$ and remove column $j_{\min}$ from $\mathbf{X}$.
   2.2 return to step 1.
3. If $|t_j| > t_{\text{cutoff}}$ for all variables j remaining in the model, then stop.

Backwards elimination

1. Obtain the estimator $\hat{\boldsymbol{\beta}}_{\text{ols}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\boldsymbol{y}$ and its t-statistics.
2. If there are any regressors j such that $|t_j| < t_{\text{cutoff}}$,
   2.1 find the regressor $j_{\min}$ having the smallest value of $|t_j|$ and remove column $j_{\min}$ from $\mathbf{X}$.
   2.2 return to step 1.
3. If $|t_j| > t_{\text{cutoff}}$ for all variables j remaining in the model, then stop.

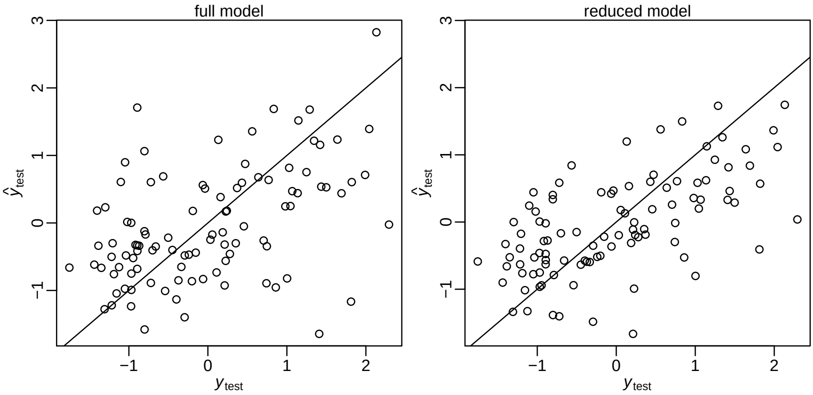
Backwards elimination

1. Obtain the estimator $\hat{\boldsymbol{\beta}}_{\text{ols}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\boldsymbol{y}$ and its t-statistics.
2. If there are any regressors j such that $|t_j| < t_{\text{cutoff}}$,
    2.1 find the regressor $j_{\min}$ having the smallest value of $|t_j|$ and remove column $j_{\min}$ from $\mathbf{X}$.
    2.2 return to step 1.
3. If $|t_j| > t_{\text{cutoff}}$ for all variables j remaining in the model, then stop.

Backwards elimination

1. Obtain the estimator $\hat{\boldsymbol{\beta}}_{\text{ols}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\boldsymbol{y}$ and its t-statistics.
2. If there are any regressors j such that $|t_j| < t_{\text{cutoff}}$,
   2.1 find the regressor j_{min} having the smallest value of $|t_j|$ and remove column j_{min} from $\mathbf{X}$.
   2.2 return to step 1.
3. If $|t_j| > t_{\text{cutoff}}$ for all variables j remaining in the model, then stop.

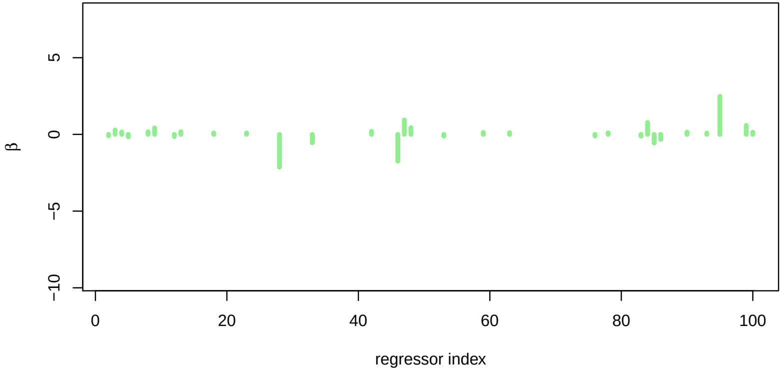
Backwards elimination

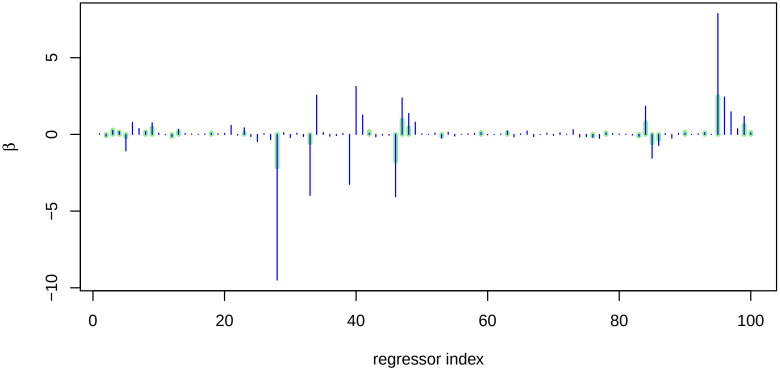


$$\frac{1}{100}\sum(y_{\text{test},i} - \hat{y}_{\text{test}^{bel},i})^2 = 0.6392334$$

```r
summary(bslfit)$coef
```

```
##                                 Estimate Std. Error   t value     Pr(>|t|)
## X[, vars$remain]sex          -0.09328805 0.04025705 -2.317310 2.113382e-02
## X[, vars$remain]bmi           0.28908888 0.05157958  5.604716 4.589992e-08
## X[, vars$remain]map           0.15786992 0.04487884  3.517691 4.999904e-04
## X[, vars$remain]tc           -0.15904952 0.05146590 -3.090386 2.179100e-03
## X[, vars$remain]tch           0.17024181 0.05630380  3.023629 2.704893e-03
## X[, vars$remain]ltg           0.42456399 0.06408671  6.624837 1.523389e-10
## X[, vars$remain]g2           -0.14534596 0.03876275 -3.749630 2.110335e-04
## X[, vars$remain]g3            0.17634891 0.08649395  2.038858 4.230677e-02
## X[, vars$remain]map.sex       0.09829108 0.03752922  2.619055 9.248049e-03
## X[, vars$remain]tc.map        0.09749561 0.04538472  2.148203 3.246722e-02
## X[, vars$remain]ldl.tc       -2.13349833 0.64669555 -3.299077 1.082246e-03
## X[, vars$remain]hdl.tc       -0.54603554 0.18196603 -3.000755 2.910311e-03
## X[, vars$remain]ltg.age       0.19552021 0.04296681  4.550494 7.676157e-06
## X[, vars$remain]ltg.tc       -1.74397252 0.43795688 -3.982064 8.505933e-05
## X[, vars$remain]ltg.ldl       0.93640700 0.22684837  4.127898 4.703360e-05
## X[, vars$remain]ltg.hdl       0.43692066 0.11027766  3.962005 9.215765e-05
## X[, vars$remain]glu.map      -0.11482709 0.05331076 -2.153919 3.201210e-02
## X[, vars$remain]g1.age        0.12558673 0.03895948  3.223522 1.400240e-03
## X[, vars$remain]g1.tc         0.11596046 0.05828432  1.989565 4.751159e-02
## X[, vars$remain]g2.tch       -0.10891318 0.04386530 -2.482900 1.355710e-02
## X[, vars$remain]g2.glu        0.09805674 0.04193307  2.338411 1.999727e-02
## X[, vars$remain]g3.map       -0.11251118 0.05763701 -1.952065 5.182477e-02
## X[, vars$remain]g3.tc         0.77688983 0.25105049  3.094556 2.149609e-03
## X[, vars$remain]g3.ldl       -0.56047786 0.22513369 -2.489534 1.331167e-02
## X[, vars$remain]g3.hdl       -0.32957931 0.09787123 -3.367479 8.535976e-04
## X[, vars$remain]g3.g1         0.14868032 0.05748889  2.586245 1.015561e-02
## X[, vars$remain]bmi2          0.08691802 0.04030775  2.156360 3.181945e-02
## X[, vars$remain]tc2           2.47388908 0.73957486  3.345015 9.231882e-04
## X[, vars$remain]ltg2          0.58268189 0.16696099  3.489928 5.527243e-04
## X[, vars$remain]glu2          0.14192060 0.04795340  2.959553 3.316792e-03
```

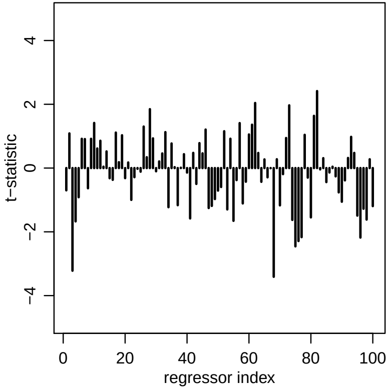

β
5
0
−5
−10
0
20
40
60
80
100
regressor index



β
5
0
−5
−10
0
20
40
60
80
100
regressor index

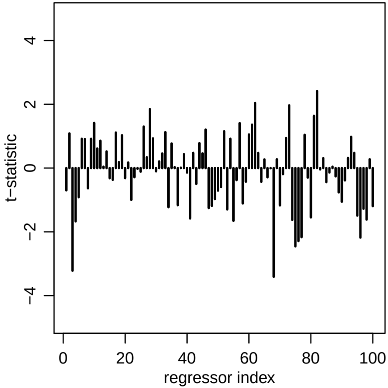
Spurious associations

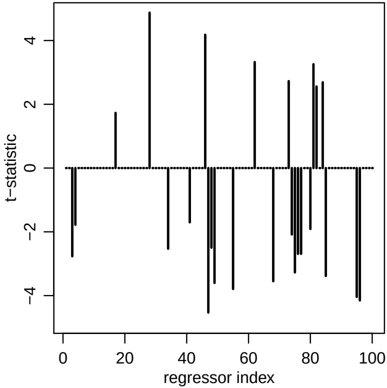
Now try modeling $y_{\pi i} = \boldsymbol{\beta}^T \boldsymbol{x}_i + \epsilon_i$



Spurious associations

Now try modeling $y_{\pi i} = \boldsymbol{\beta}^T \boldsymbol{x}_i + \epsilon_i$





Spurious associations

```r
sum(abs(t.bslperm)>2 )

## [1] 21

sum(abs(t.bslperm)>3 )

## [1] 12

sum(abs(t.bslperm)>4 )

## [1] 5
```

- 21 regressors have t-stats > 2 ($p \approx 0.05$)
- 12 5 regressors have t-stats > 3 ($p \approx 0.003$)
- 5 regressors have t-stats > 4 ($p \approx 0.00006$)

Spurious associations

```
sum(abs(t.bslperm)>2 )

## [1] 21

sum(abs(t.bslperm)>3 )

## [1] 12

sum(abs(t.bslperm)>4 )

## [1] 5
```

- 21 regressors have t-stats > 2 ($p \approx 0.05$)
- 12 5 regressors have t-stats > 3 ($p \approx 0.003$)
- 5 regressors have t-stats > 4 ($p \approx 0.00006$)

Spurious associations

```
sum(abs(t.bslperm)>2 )

## [1] 21

sum(abs(t.bslperm)>3 )

## [1] 12

sum(abs(t.bslperm)>4 )

## [1] 5
```

- 21 regressors have t-stats > 2 ($p \approx 0.05$)
- ~~12~~ 5 regressors have t-stats > 3 ($p \approx 0.003$)
- 5 regressors have t-stats > 4 ($p \approx 0.00006$)

Spurious associations

```
sum(abs(t.bslperm)>2 )

## [1] 21

sum(abs(t.bslperm)>3 )

## [1] 12

sum(abs(t.bslperm)>4 )

## [1] 5
```

- 21 regressors have t-stats > 2 ($p \approx 0.05$)
- 12 5 regressors have t-stats > 3 ($p \approx 0.003$)
- 5 regressors have t-stats > 4 ($p \approx 0.00006$)

Spurious associations

```
sum(abs(t.bslperm)>2 )

## [1] 21

sum(abs(t.bslperm)>3 )

## [1] 12

sum(abs(t.bslperm)>4 )

## [1] 5
```

- 21 regressors have t-stats > 2 ($p \approx 0.05$)
- 12 5 regressors have t-stats > 3 ($p \approx 0.003$)
- 5 regressors have t-stats > 4 ($p \approx 0.00006$)

Spurious associations

```
sum(abs(t.bslperm)>2 )

## [1] 21

sum(abs(t.bslperm)>3 )

## [1] 12

sum(abs(t.bslperm)>4 )

## [1] 5
```

- 21 regressors have t-stats > 2 ($p \approx 0.05$)
- 12 5 regressors have t-stats > 3 ($p \approx 0.003$)
- 5 regressors have t-stats > 4 ($p \approx 0.00006$)