# SF2930 - Regression analysis

KTH Royal Institute of Technology, Stockholm

Lecture 12 – Variable selection and model building (MPV 10, Iz 5.8, HTW 2.1-2.3, 2.4.1-2.4.2, HTF 3.6)

February 18, 2022

# Todays lecture

- Equivalent definitions of ridge regression
- Penalized least squares
- LASSO
- Model selection
  - Bests subsets regression
  - Stepwise forward selection
  - Stepwise backwards selection
  - Criteria for selecting the "best" model
- Consequences of model mis-specification

# Alternative definitions of ridge regression estimates

## Theorem

*The following definitions are equivalent definitions of the RRE for $\boldsymbol{\beta}$.*

1. $\hat{\boldsymbol{\beta}}_T^{(1)}(t) = (X^R X + tI)^{-1} X^T \mathbf{y}$
2. $\hat{\boldsymbol{\beta}}_R^{(2)}(t) = \arg\min_{\boldsymbol{\beta}} \|\mathbf{y} - X\boldsymbol{\beta}\|_2^2 + t\|\boldsymbol{\beta}\|_2^2$
3. $\hat{\boldsymbol{\beta}}_R^{(3)}(t) = \arg\min_{\boldsymbol{\beta}:\ \|\boldsymbol{\beta}\|_2^2 \leq t'(t)} \|\mathbf{y} - X\boldsymbol{\beta}\|_2^2$

Note that in all cases, we do not want to apply the penalty to $\beta_0$, and hence you have to be careful if $X$ and $\mathbf{y}$ are not normalized.

# Ridge regression and penalized least squares

### Ridge regression

$$\hat{\boldsymbol{\beta}}_R = \arg\min_{\boldsymbol{\beta}} \|\mathbf{y} - X\boldsymbol{\beta}\|_2^2 + t\|\boldsymbol{\beta}\|_2^2 = \arg\min_{\beta:\ \|\beta\|_2 \le t'(t)} \|\mathbf{y} - X\boldsymbol{\beta}\|_2^2.$$

### Penalized least squares

More generally, we would replace $\|\boldsymbol{\beta}\|_2^2$ by any penalty function $p(\boldsymbol{\beta})$, to get

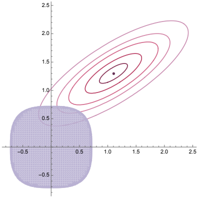
$$\hat{\boldsymbol{\beta}} := \arg\min_{\boldsymbol{\beta}} \|\mathbf{y} - X\boldsymbol{\beta}\|_2^2 + tp(\boldsymbol{\beta}).$$

Setting $p(\boldsymbol{\beta}) = \|\boldsymbol{\beta}\|_q^q$, we obtain

$$\hat{\boldsymbol{\beta}} := \arg\min_{\boldsymbol{\beta}} \|\mathbf{y} - X\boldsymbol{\beta}\|_2^2 + t\|\boldsymbol{\beta}\|_q^q = \arg\min_{\beta:\ \|\beta\|_q \le t'(t)} \|\mathbf{y} - X\boldsymbol{\beta}\|_2^2.$$

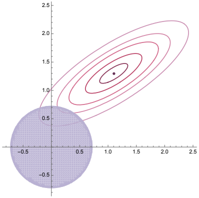
# Ridge regression and penalized least squares

$q = 3$



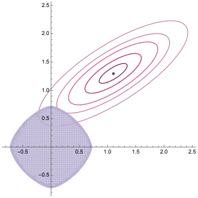

$$\underset{\beta:\ \|\beta\|_3 \le t'(t)}{\arg\min} \|\mathbf{y} - X\boldsymbol{\beta}\|_2^2$$

$q = 2$



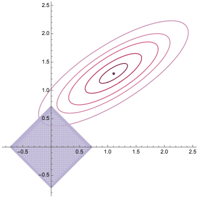

$$\underset{\beta:\ \|\beta\|_2 \le t'(t)}{\arg\min} \|\mathbf{y} - X\boldsymbol{\beta}\|_2^2$$

$q = 1.5$



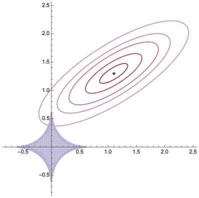

$$\underset{\beta:\ \|\beta\|_{1.5} \le t'(t)}{\arg\min} \|\mathbf{y} - X\boldsymbol{\beta}\|_2^2$$

$q = 1$



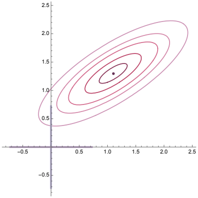

$$\underset{\beta:\ \|\beta\|_1 \le t'(t)}{\arg\min} \|\mathbf{y} - X\boldsymbol{\beta}\|_2^2$$

$q = 0.5$




$$\underset{\beta:\ \|\beta\|_{0.5} \le t'(t)}{\arg\min} \|\mathbf{y} - X\boldsymbol{\beta}\|_2^2$$

$q = 0$




$$\underset{\beta:\ \|\beta\|_0 \le t'(t)}{\arg\min} \|\mathbf{y} - X\boldsymbol{\beta}\|_2^2$$

# Least Absolute Shrinkage and Selection Operator (LASSO)

### LASSO

Choosing $q = 1$, we obtain the *LASSO estimator*

$$\hat{\boldsymbol{\beta}}_L = \arg\min_{\boldsymbol{\beta}} \|\mathbf{y} - X\boldsymbol{\beta}\|_2^2 + t\|\boldsymbol{\beta}\|_1 = \arg\min_{\beta:\ \|\beta\|_1 \le t'(t)} \|\mathbf{y} - X\boldsymbol{\beta}\|_2^2.$$

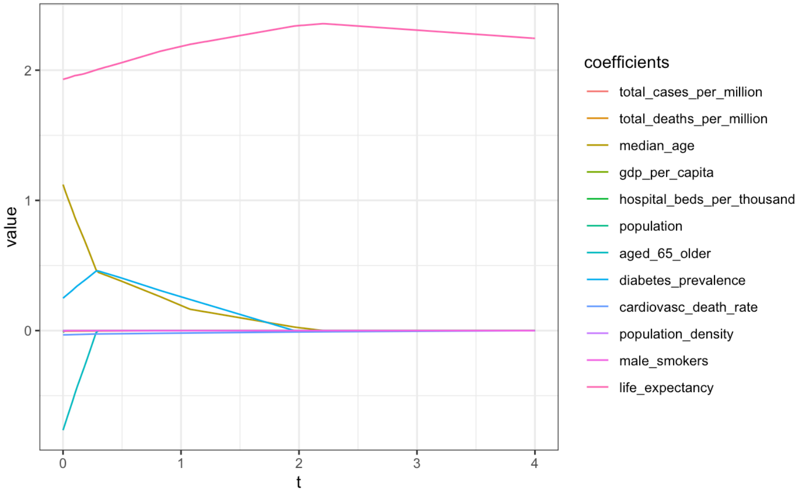
### Properties

- As with ridge regression, one can show that LASSO estimates generally has a smaller variance than LSE, but has a small bias.
- If $X$ is orthonormal, then

  $$\hat{\boldsymbol{\beta}}_R(t) = \hat{\boldsymbol{\beta}}/(1+t) \qquad \text{and} \qquad \hat{\beta}_{L,j} = \operatorname{sgn}\hat{\beta}_j(|\hat{\beta}_j| - t)_+$$

  In particular, LASSO is a *shrinkage operator*.
- There are no formulas for standard errors for LASSO estimates, and hence we have to use bootstrap estimate errors.
- The larger $t$ is, the more coefficients will be set to zero.

# LASSO profiles



# How can we efficiently compute the LASSO profiles?

**Motivation**

Assume first that we only have one regressor, so that $y = \beta x + \varepsilon$, and that $\mathbf{y}$ and $\mathbf{x}$ are both on standard form. Then

$$\hat{\beta}(t) = \arg\min_{\beta} (\mathbf{y} - \mathbf{x}\beta)^T (\mathbf{y} - \mathbf{x}\beta) + t|\beta|$$

$$f(\beta) \coloneqq (\mathbf{y} - \mathbf{x}\beta)^T (\mathbf{y} - \mathbf{x}\beta) + t|\beta|$$
$$f'(\beta) = -2\mathbf{y}^T\mathbf{x} + 2\beta\mathbf{x}^T\mathbf{x} + t \operatorname{sgn} \beta$$
$$\hat{\beta}(t) = \begin{cases} \mathbf{y}^T\mathbf{x} - \frac{t \operatorname{sgn} \mathbf{y}^T\mathbf{x}}{2} & \text{if } \mathbf{y}^T\mathbf{x} > t/2 \\ 0 & \text{else} \end{cases} = \operatorname{sgn}(\mathbf{y}^T\mathbf{x}) \cdot (|\mathbf{y}^T\mathbf{x}| - t/2)_+$$

# How can we efficiently compute the LASSO profiles?

### Cyclic coordinate descent

1. Pick some arbitrary initial value for each regression coefficient, $\beta_1^0,\, \beta_2^0,\, \ldots,\, \beta_k^0$.
2. To update $\beta_i^0$, pick $\hat{\beta}_i$ which minimizes

$$g(\beta_i) \coloneqq \|\mathbf{y} - \sum_{j \neq i} X_{\cdot j}\beta_j^0 - X_{\cdot i}\beta_i\|_2^2 + \overbrace{t\beta_i + t\sum_{j \neq i}\beta_j}^{=t\|\boldsymbol{\beta}\|_1}.$$

   By the above argument, we have $\hat{\beta}_i = (\operatorname{sgn} \mathbf{r}_{-i}^T \mathbf{x}_i) \cdot (|\mathbf{r}_{-i}^T \mathbf{x}_i - t/2|)$, where $\mathbf{r}_{-i} \coloneqq \mathbf{y} - \sum_{j \neq i} X_{\cdot j}\beta_j^0$. Update $\beta_i^0$ by letting $\beta_i^0 \mapsto \hat{\beta}_i$.
3. Repeat this procedure, lopping through all regression coefficients until the coefficients converge.

Since the initial problem is convex and has a unique minimum, $(\beta_1^0,\, \beta_2^0,\, \ldots,\, \beta_k^0)^T$ will converge to $\hat{\boldsymbol{\beta}}_{LASSO}$.

# How can we efficiently compute the LASSO profiles?

**Least angle regression**

1. Start with residual $\mathbf{r} = \mathbf{y}$, $\boldsymbol{\beta}(0) \equiv 0$, and active set $\mathcal{A} = \emptyset$.
2. Find the predictor $\mathbf{x}_j$ which maximizes $\tau_0 \coloneqq |\mathbf{r}_0^T \mathbf{x}_j|$. Define the *active set* $\mathcal{A} \coloneqq \{j\}$ and $X_{\mathcal{A}} = \mathbf{x}_j$.
3. Repeat the following for $i = 1, 2, \ldots$:
   3.1 Define $\delta \coloneqq \tau_{i-1}^{-1}(X_{\mathcal{A}}^T X_{\mathcal{A}})^{-1} X_{\mathcal{A}}^T \mathbf{r}_{i-1}$.
   3.2 Define $\Delta \coloneqq \delta \mathbf{1}_{\mathcal{A}}$.
   3.3 Move $\boldsymbol{\beta}$ in direction $\Delta$ until the time $t$ when another regressor $\ell \notin \mathcal{A}$ has the same correlation with $\mathbf{r}' \coloneqq \mathbf{y} - X\boldsymbol{\beta}(t)$ as the coefficients in $\mathcal{A}$.
   3.4 Set $\mathcal{A} = \mathcal{A} \cup \{\ell\}$ and $\mathbf{r} = \mathbf{r}'$

# How do we choose $t$?

- Small $t$ – better fit and less bias
- Large $t$ – simpler model and smaller variance

**Method (cross-validation)**

1. Partition the data set $S$ up into $m$ samples $V_1, V_2, \ldots, V_m$ of equal size (validation sets).
2. For each $t$ and each $j \in \{1, 2, \ldots, m\}$, use $T_j := S \setminus V_j$ as a training set to find $\hat{\boldsymbol{\beta}}_L^{T_j}(t)$, and estimate the prediction error by

$$\widehat{PE}_j(t) := \frac{1}{|V_j|} \sum_{i \in V_j} \big(y_i - \mathbf{x}_i^T \hat{\boldsymbol{\beta}}_L^{T_j}(t)\big)^2, \qquad \widehat{PE}(t) := \frac{1}{m} \sum_j \widehat{PE}_j(t)$$

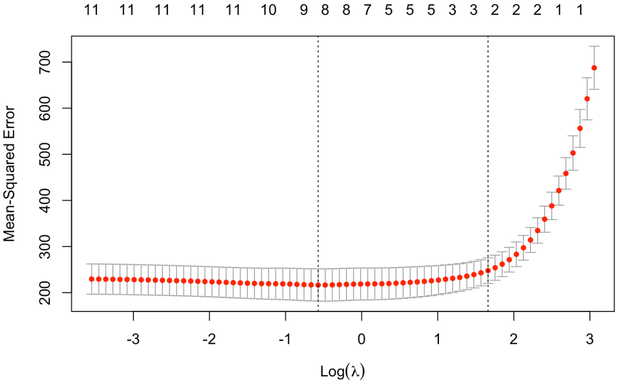
3. Plot $\widehat{PE}(t)$ as a function of $t$. This plot is called a *cross-validation error curve*.
4. Pick $t$ which "almost" minimizes this error.

# How do we choose $t$?

- Small $t$ – better fit and less bias
- Large $t$ – simpler model and smaller variance

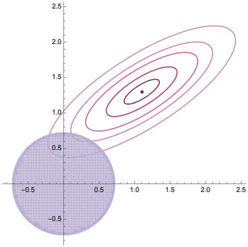
**The cross-validation error curve**

.



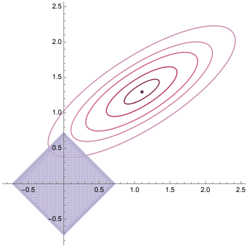
# Ridge vs. LASSO

**Ridge regression**




$$\arg\min_{\beta:\ \|\beta\|_2^2 \le t'(t)} \|\mathbf{y} - X\boldsymbol{\beta}\|_2^2$$

**LASSO**




$$\arg\min_{\beta:\ \|\beta\|_1 \le t'(t)} \|\mathbf{y} - X\boldsymbol{\beta}\|_2^2$$

**Comparison**

- LASSO will in general force some of the coefficients to be equal to zero, which corresponds to deleting the corresponding regressors from the model.
- LASSO estimates tends to be better than RRE when only a few of the "true" coefficients are non-zero, while RRE is generally better than LASSO if $\boldsymbol{\beta}$ is not sparse.

# LASSO and elastic net

**Ridge regression**

$$\underset{\beta:\ \|\beta\|_2^2 \le t'(t)}{\arg\min} \|\mathbf{y} - X\boldsymbol{\beta}\|_2^2 = \underset{\boldsymbol{\beta}}{\arg\min} \|\mathbf{y} - X\boldsymbol{\beta}\|_2^2 + t\|\boldsymbol{\beta}\|_2^2$$

**LASSO**

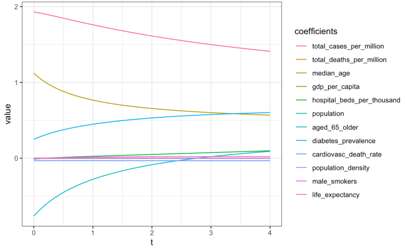
$$\underset{\beta:\ \|\beta\|_1 \le t'(t)}{\arg\min} \|\mathbf{y} - X\boldsymbol{\beta}\|_2^2 = \underset{\boldsymbol{\beta}}{\arg\min} \|\mathbf{y} - X\boldsymbol{\beta}\|_2^2 + t\|\boldsymbol{\beta}\|_1$$

**Elastic net**

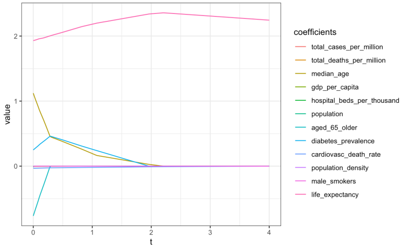
$$\underset{\boldsymbol{\beta}}{\arg\min} \|\mathbf{y} - X\boldsymbol{\beta}\|_2^2 + t\big(\alpha\|\boldsymbol{\beta}\|_1 + \frac{1-\alpha}{2}\|\boldsymbol{\beta}\|_2^2\big)$$

# Ridge vs. LASSO

## Ridge traces




## LASSO traces



# Criteria for selecting the "best" model

**Adjusted $R^2$**

$$R^2 = 1 - \frac{SS_{Res}(S)}{SS_R} \quad \text{and} \quad R^2_{Adj}(S) := 1 - \frac{SS_{Res}(S)/(n - |S| + 1)}{SS_T/(n-1)},$$

We ideally want to choose the model which for which $R^2_{Adj}(S)$ is maximal.

**Residual mean squared**

$$MS_{Res}(S) = \frac{SS_{Res}(S)}{n - |S| - 1}$$

We ideally want to choose the model which for which $MS_{Res}(S)$ is minimal.

**Deviance**

$$\text{deviance} = 2 \log \frac{L_{\text{model}}}{L_{\text{saturated model}}}$$

$$L_{\text{model}} \propto e^{-\|\mathbf{y} - X\hat{\boldsymbol{\beta}}\|_2^2/2\sigma^2}, \qquad L_{\text{saturated model}} \propto \prod e^{(y_i - y_i)^2/2\sigma^2}$$

# Criteria for selecting the "best" model

**The Akaike Information Criterion (AIC)**

$$AIC := -2\log L + 2(|S|+1) \overset{\text{if using LSE}}{=} 2n\log\frac{SS_{Res}(S)}{n} + 2(|S|+1).$$

**Bayesian Information Criterion (BIC)**

$$BIC := -2\log L + (|S|+1)\log n \overset{\text{if using LSE}}{=} n\log\frac{SS_{Res}(S)}{n} + (|S|+1)\log n.$$

# Criteria for selecting the "best" model

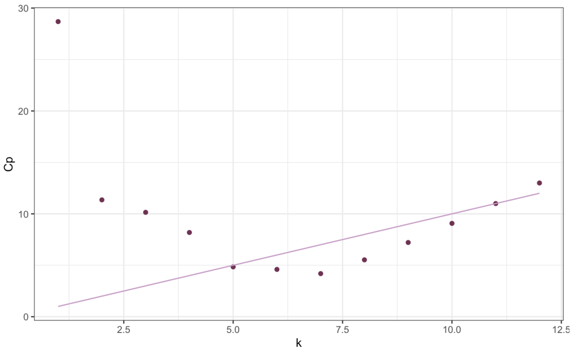
## Mallows' Cp-statistic

$$C_p(S) := \frac{SS_{Res}(S)}{\hat{\sigma}^2} - (n - 2(|S| + 1))$$

which is an estimate of the *standardized total mean square error*

$$\Gamma_S = \frac{\mathbb{E}\left[\left\|\hat{y}(S) - \mathbb{E}[y]\right\|_2^2\right]}{\sigma^2} = \frac{\left\|\mathbb{E}\left[\hat{y}(S)\right] - \mathbb{E}[y]\right\|_2^2 + \operatorname{tr}\operatorname{Var}\left(\hat{y}(S)\right)}{\sigma^2}$$

If the model has little bias, then $C_p(S) \approx |S|$.



# Best subsets regression

**Idea**

Find the best model using some subset of the regressors by calculating the regression coefficients for each possible subset of the regressors, and then choose the "best model".

**Of all models with $j$ regressors, which is best?**

The model which minimizes sum $SS_{Res} = \|\mathbf{e}\|_2^2 = \|\mathbf{y} - \hat{\mathbf{y}}\|_2^2$.

**Comments**

- If we initially have $k$ regressors, excluding the intercept $\beta_0$, then there are $2^k$ different such subsets, and hence if $k$ is large, we will need to compare many models...

# Example

```r
dff <- df00
dff$continent <- NULL
dff$location <- NULL

library("leaps")
model <- regsubsets(people_fully_vaccinated_per_hundred~.,
    data = dff, nvmax = 4)

summary(model)
```

- 1 regressor
  - `life_expectancy`
- 2 regressors
  - `life_expectancy`
  - `gdp_per_capita`
- 3 regressors
  - `life_expectancy`
  - `gdp_per_capita`
  - `cardiovascular_death_rate`
- 4 regressors
  - `life_expectancy`
  - `gdp_per_capita`
  - `cardiovascular_death_rate`
  - `diabetes_prevalence`
- 5 regressors
  - `life_expectancy`
  - `gdp_per_capita`
  - `cardiovascular_death_rate`
  - `aged_65_older`
  - `median_age`

```r
which.max(summary(bsrmodel)$adjr2) # 7
which.min(summary(bsrmodel)$cp) # 7
which.min(summary(bsrmodel)$bic) # 2
```

# Stepwise forward selection

**Algorithm**

Start with a model with no regressors.
Pick a statistic $T$ which can be used to compare models, such as $SS_{Res}$, AIC, etc, and a threshold $t$ for this statistic.

1. For each regressor $x$, calculate
$$T(x \mid \emptyset) = T(x) - T(\emptyset)$$
2. Let $x_1 := \arg\max T(x \mid \emptyset)$. Add $x_1$ to the model if $T(x_1 \mid \emptyset) \geq t$. If $T(x \mid \emptyset) < t$, stop and return $\{\}$.
3. Assume that $x_1$, $x_2$, …, $x_j$ has already been added to the model. For each remaining regressor $x$, calculate
$$T(x|x_1, x_2, \ldots, x_j) = T(x, x_1, x_2, \ldots, x_j) - T(x_1, x_2, \ldots, x_j).$$
4. Let $x_{j+1} := \arg\max T(x|x_1, x_2, \ldots, x_j)$. Include $x_{j+1}$ in the model if $T(x|x_1, x_2, \ldots, x_j) \geq t$. If $T(x|x_1, x_2, \ldots, x_j) < t$, stop and return $\{x_1, x_2, \ldots, x_j\}$.

# Stepwise forward selection

**Comments**

The book suggests using $F$-statistics to define

$$T(x \mid \emptyset) = \frac{SS_R(x)/1}{SS_{Res}(x)/(n-1-1)}$$

and

$$\begin{aligned} T(x|x_1, x_2, \ldots, x_j) &= \frac{SS_R(x \mid x_1, x_2, \ldots, x_j)/((j+1)-1)}{MS_{Res}(x_1, x_2, \ldots, x_j, x)} \\ &= \frac{SS_R(x_1, x_2, \ldots, x_j, x) - SS_R(x_1, x_2, \ldots, x_j)}{SS_{Res}(x_1, x_2, \ldots, x_j, x)/(n-(j+1)-1)}. \end{aligned}$$

With this choice,

- $x_1$ will be the regressor which has the largest simple correlation with $\mathbf{y}$.
- For each $j \geq 2$, $x_j$ will be the regressor not yet included in the model which has the largest simple correlation with the residuals from the model $y = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \ldots + \hat{\beta}_{j-1} x_{j-1}$.

# Example

```r
dff <- df00
dff$continent <- NULL
dff$location <- NULL
```

```r
vaccinations_only <- lm(people_fully_vaccinated_per_hundred
    ~ 1, data=dff)
forward <- step(vaccinations_only, direction='forward',
    scope=formula(all), trace=0)
```

```r
summary(forward)
```

```
Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)     64.5047523 18.0507915   3.574  0.00305 **
cardiovasc_d_r  -0.0597492  0.0298076  -2.004  0.06475 .
gdp_per_capita   0.0005934  0.0003216   1.845  0.08626 .
diabetes_prev   -1.9051997  1.2394378  -1.537  0.14655

Residual standard error: 8.396 on 14 degrees of freedom
Multiple R-squared: 0.8014, Adjusted R-squared:  0.7589
F-statistic: 18.83 on 3 and 14 DF,  p-value: 3.477e-05
```

# Example

```r
dff <- df00
dff$continent <- NULL
dff$location <- NULL
```

```r
vaccinations_only <- lm(people_fully_vaccinated_per_hundred
    ~ 1, data=dff)
forward <- step(vaccinations_only, direction='forward',
    scope=formula(all), trace=0)
```

```r
forward$anova
```

| Step | Df | Deviance | Residual df | Residual deviance | AIC |
|---|---|---|---|---|---|
| | NA | NA | 17 | 4970.4664 | 103.17615 |
| +cardiovasc death rate | -1 | 3576.7428 | 16 | 1393.7236 | 82.28853 |
| +gdp per capita | -1 | 240.1529 | 15 | 1153.5707 | 80.88442 |
| +diabetes prevalence | -1 | 166.5779 | 14 | 986.9928 | 80.07724 |

# Stepwise backward elimination

**Algorithm**

Start with a model that contains all the regressors.

1. Pick some threshold $t$.
2. Calculate a statistic $T(x \mid x_1, \ldots, x_j)$ for each variable as if it were the last to enter the model. Let $x_{j+1}$ be the regressor with the smallest $T$-statistic. Remove $x_1$ from the model if $T(x \mid x_1, \ldots, x_j) < t$.

**Comments**

Backwards selection might be preferred if we want to ensure that we do not miss any information, while forward selection could keep the final model smaller.

# Example

```r
dff <- df00
dff$continent <- NULL
dff$location <- NULL
```

```r
all <- lm(people_fully_vaccinated_per_hundred ~ ., data=dff)
backward <- step(all, direction='backward', scope=formula(all), trace=0)
```

```r
summary(backward)
```

```
Coefficients:
                       Estimate Std. Error t value Pr(>|t|)
(Intercept)           3.089e+02  1.917e+02   1.611   0.1354
gdp_per_capita        7.278e-04  3.638e-04   2.001   0.0707
hospital_beds_per_th -2.325e+00  1.563e+00  -1.487   0.1650
aged_65_older         2.216e+00  1.045e+00   2.121   0.0575
cardiovasc_death_rate -1.318e-01  6.457e-02  -2.041   0.0659
male_smokers          3.879e-01  2.894e-01   1.340   0.2072
life_expectancy      -3.611e+00  2.356e+00  -1.533   0.1536

Residual standard error: 8.068 on 11 degrees of freedom
Multiple R-squared:  0.856, Adjusted R-squared:  0.7774
F-statistic: 10.89 on 6 and 11 DF,  p-value: 0.0004426
```

# Example

```r
dff <- df00
dff$continent <- NULL
dff$location <- NULL
```

```r
all <- lm(people_fully_vaccinated_per_hundred ~ ., data=dff)
backward <- step(all, direction='backward', scope=formula(
    all), trace=0)
```

```r
backward$anova
```

| Step | Df | Deviance | Residual df | Residual deviance | AIC |
|---|---|---|---|---|---|
| | NA | NA | 5 | 582.9232 | 88.59831 |
| - population | 1 | 0.6653029 | 6 | 583.5885 | 86.61884 |
| - diabetes prevalence | 1 | 0.5082647 | 7 | 584.0967 | 84.63451 |
| - median age | 1 | 3.7695471 | 8 | 587.8663 | 82.75030 |
| - population density | 1 | 37.3660036 | 9 | 625.2323 | 81.85953 |
| - total deaths per million | 1 | 35.6410688 | 10 | 660.8734 | 80.85743 |
| - total cases per million | 1 | 55.1191456 | 11 | 715.9925 | 80.29936 |

# The consequences of model misspecification

**True model**

$\mathbf{y} = X\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, $\boldsymbol{\varepsilon} \sim N(0, \sigma^2 I)$.

**Notation**

Let $X = (X_p, X_r)$ and $\beta = (\beta_p^T, \beta_r^T)^T$, so that $\mathbf{y} = X_p\boldsymbol{\beta}_p + X_r\boldsymbol{\beta}_r + \boldsymbol{\varepsilon}$.
Let $\hat{\boldsymbol{\beta}}^* = (\hat{\boldsymbol{\beta}}_p^*, \hat{\boldsymbol{\beta}}_r^*)$, $\hat{\sigma}_*^2$, $\hat{\mathbf{y}}^*$ denote the LS estimates for the full model, and let $\hat{\boldsymbol{\beta}}_p$, $\hat{\sigma}^2$, and $\hat{\mathbf{y}}$ be the corresponding estimates for the *reduced model* $\mathbf{y} = X_p\boldsymbol{\beta}_p + \boldsymbol{\varepsilon}$.

**Properties**

- $\mathbb{E}[\hat{\boldsymbol{\beta}}_p] = \boldsymbol{\beta}_p + \overbrace{(X_p^T X_p)^{-1} X_p^T X_r}^{\Lambda \text{ (the alias matrix)}} \boldsymbol{\beta}_r$. Hence $\hat{\boldsymbol{\beta}}_p$ is a biased estimator of $\boldsymbol{\beta}_p$
- $(y_i - \mathbf{e}_i^T X_p \hat{\boldsymbol{\beta}}_p)^2 \leq (y_i - \mathbf{e}_i^T X \hat{\boldsymbol{\beta}}^*)^2$ In other words, removing regressor never increases the variance of the remaining parameters.
- Since $\hat{\boldsymbol{\beta}}_p$ is biased and the $MSE(\hat{\boldsymbol{\beta}}_p) = \mathrm{Var}(\hat{\boldsymbol{\beta}}_p) + \mathrm{bias}(\hat{\boldsymbol{\beta}}_p)$, we might be able to use it to see interesting differeneces between the models. In fact, one can show that $MSE(\hat{\boldsymbol{\beta}}_p) < MSE(\hat{\boldsymbol{\beta}}_p^*)$ when the deleted variables have regression coefficients which are smaller than the standard errors of their estimates in the full model.
- $\hat{\sigma}_*^2$ is an unbiased estimate of $\sigma^2$, but $\hat{\sigma}^2$ is a biased estimator (generally to large) of $\sigma^2$.

**Conclusion**

Is can often be adventageous to remove variables, even if this means deviating from the true model.