Forward and Backward selection

Forward selection

Vid den s k **forward selection** metoden börjar man enbart ett intercept och lägger sedan till en variabel i taget. Vid varje steg lägger man till en ny variabel tills dess ingen ny variabel har en t-kvot som överstiger tröskeln $t_{obs} = 2$. Valet av just talet 2 som tröskelvärde är rätt godtyckligt, men innebär att man bara lägger till en variabel om den är signifikant på ungefär 5% signifikansnivå (det exakta kritiska värdet är kring 2, men beror ju på antalet frihetsgrader).

Steg 1 - val av den första förklarande variabeln

Vi skattar enkla regressioner för var och en av den fem variablerna separat:

```
lmfit = lm(nRides ~ temp, data = bike)
regsummary(lmfit, anova = F, fit_measures = F)
```

```
##
## Parameter estimates
## -----------------------------------------------
##             Estimate Std. Error t value   Pr(>|t|)
## (Intercept)   1214.6     161.16  7.5367 1.4327e-13
## temp          6640.7     305.19 21.7594 2.8106e-81
```

```
lmfit = lm(nRides ~ hum, data = bike)
regsummary(lmfit, anova = F, fit_measures = F)
```

```
##
## Parameter estimates
## -----------------------------------------------
##             Estimate Std. Error t value   Pr(>|t|)
## (Intercept)   5364.0     322.68 16.6233 7.4356e-53
## hum          -1369.1     501.19 -2.7317 6.4541e-03
```

```
lmfit = lm(nRides ~ windspeed, data = bike)
regsummary(lmfit, anova = F, fit_measures = F)
```

```
##
## Parameter estimates
## -----------------------------------------------
##             Estimate Std. Error t value    Pr(>|t|)
## (Intercept)   5621.2     185.06 30.3744 1.3616e-131
## windspeed    -5862.9     899.99 -6.5144  1.3600e-10
```

```
lmfit = lm(nRides ~ holiday, data = bike)
regsummary(lmfit, anova = F, fit_measures = F)
```

```
##
## Parameter estimates
## -----------------------------------------------
##             Estimate Std. Error t value    Pr(>|t|)
## (Intercept)   4527.1     72.582 62.3723 1.7126e-294
## holiday       -792.1    428.231 -1.8497  6.4759e-02
```

```
lmfit = lm(nRides ~ workingday, data = bike)
regsummary(lmfit, anova = F, fit_measures = F)
```

```
##
## Parameter estimates
## -----------------------------------------------
##             Estimate Std. Error t value    Pr(>|t|)
## (Intercept)  4330.17     127.31 34.0134 1.2970e-152
## workingday    254.65     153.93  1.6543  9.8495e-02
```

Det största t-värdet i absolutbelopp (dvs bortsett från tecknet framför t-värdet) får vi för variabeln `temp` som har $t_{obs} = 21.7594$. Det t-värdet är också större än 2 (i absolutbelopp), så vi inkluderar därför `temp` i modellen. Notera att vi bryr oss inte om t-värdet för interceptet här. Interceptet är alltid med i modellen.

Steg 2 - val av den andra förklarande variabeln

Nu fortsätter vi och lägger till en andra förklarande variabel, givet att `temp` redan är med i modellen. Vi skattar där nya regressioner där var och en har `temp` och ytterligare en variabel som förklarande variabler:

```
lmfit = lm(nRides ~ temp + hum, data = bike)
regsummary(lmfit, anova = F, fit_measures = F)
```

```
##
## Parameter estimates
## ------------------------------------------------
##             Estimate Std. Error t value   Pr(>|t|)
## (Intercept)   2657.9     272.42  9.7565 3.2258e-21
## temp          6887.0     299.38 23.0042 1.9558e-88
## hum          -2492.9     384.76 -6.4789 1.7012e-10
```

```
lmfit = lm(nRides ~ temp + windspeed, data = bike)
regsummary(lmfit, anova = F, fit_measures = F)
```

```
##
## Parameter estimates
## ------------------------------------------------
##             Estimate Std. Error t value   Pr(>|t|)
## (Intercept)   1991.0     225.96  8.8114 8.9857e-18
## temp          6408.5     304.44 21.0500 3.3728e-77
## windspeed    -3472.1     719.10 -4.8284 1.6781e-06
```

```
lmfit = lm(nRides ~ temp + holiday, data = bike)
regsummary(lmfit, anova = F, fit_measures = F)
```

```
##
## Parameter estimates
## ------------------------------------------------
##             Estimate Std. Error t value  Pr(>|t|)
## (Intercept)  1239.00     161.53  7.6702 5.516e-14
## temp         6625.46     304.88 21.7314 4.296e-81
## holiday      -584.92     333.87 -1.7519 8.021e-02
```

```
lmfit = lm(nRides ~ temp + workingday, data = bike)
regsummary(lmfit, anova = F, fit_measures = F)
```

```
##
## Parameter estimates
## ------------------------------------------------
##             Estimate Std. Error  t value   Pr(>|t|)
## (Intercept)  1142.13     177.46  6.43596 2.2248e-10
## temp         6625.00     305.62 21.67711 8.7933e-81
## workingday    117.38     120.25  0.97617 3.2931e-01
```

Givet att `temp` är med i modellen så har `hum` det största t-värdet i absolutbelopp; vi har $t_{obs} = -6.4789$ för `hum` i utskriften ovan. Eftersom detta t-värde är större än 2 (i absolutbelopp) så lägger vi även till variabeln `hum` i modellen. Vi har nu en modell med både `temp` och `hum` som förklarande variabler. Notera att det spelar ingen roll att `temp` har större t-värde än `hum`. Variabeln `temp` är ju redan vald i steg 1 och det är inte längre en fråga om den variabel ska vara med eller inte. Steg 2 här handlar enbart om valet bland de andra variablerna utöver `temp`. Vi fortsätter nu och ser om det är värt att lägga till en tredje variabel.

Steg 3 - val av den tredje förklarande variabeln

Nu skattar vi regressioner med `temp`, `hum` och ytterligare en variabel som förklarande variabler:

```
lmfit = lm(nRides ~ temp + hum + windspeed, data = bike)
regsummary(lmfit, anova = F, fit_measures = F)
```

```
##
## Parameter estimates
## ------------------------------------------------
##             Estimate Std. Error t value   Pr(>|t|)
## (Intercept)   4084.4     337.86 12.0888 8.7098e-31
## temp          6625.5     293.09 22.6062 4.1807e-86
## hum          -3100.1     383.99 -8.0734 2.8330e-15
## windspeed    -4806.9     708.90 -6.7808 2.4754e-11
```

```
lmfit = lm(nRides ~ temp + hum + holiday, data = bike)
regsummary(lmfit, anova = F, fit_measures = F)
```

```
##
## Parameter estimates
## ------------------------------------------------
##             Estimate Std. Error t value   Pr(>|t|)
## (Intercept)  2688.54     272.44  9.8685 1.2144e-21
## temp         6871.92     298.96 22.9857 2.6652e-88
## hum         -2501.83     384.12 -6.5131 1.3739e-10
## holiday      -611.19     324.79 -1.8818 6.0262e-02
```

```
lmfit = lm(nRides ~ temp + hum + workingday, data = bike)
regsummary(lmfit, anova = F, fit_measures = F)
```

```
##
## Parameter estimates
## ------------------------------------------------
##             Estimate Std. Error t value   Pr(>|t|)
## (Intercept)  2581.46     280.81  9.1929 3.9505e-19
## temp         6870.21     299.70 22.9235 6.1121e-88
## hum         -2500.52     384.76 -6.4989 1.5016e-10
## workingday    130.93     117.00  1.1190 2.6350e-01
```

Vi lägger även till `windspeed` som förklarande variabel efter dess t-värde är störst i absolutebelopp i utskrifterna ovan och större än 2.

Steg 4 - fjärde variabeln

Vi fortsätter och frågar oss om det är värt att lägga till någon ytterligare variabel utöver de redan valda `temp`, `hum` och `windspeed`. Vi skattar därför regressionerna:

```
lmfit = lm(nRides ~ temp + hum + windspeed + holiday, data = bike)
regsummary(lmfit, anova = F, fit_measures = F)
```

```
##
## Parameter estimates
## -----------------------------------------------
##             Estimate Std. Error t value   Pr(>|t|)
## (Intercept)  4115.59     337.60 12.1909 3.1149e-31
## temp         6610.34     292.63 22.5895 5.5379e-86
## hum         -3109.33     383.29 -8.1123 2.1175e-15
## windspeed   -4808.48     707.55 -6.7960 2.2444e-11
## holiday      -613.59     315.14 -1.9470 5.1917e-02
```

```
lmfit = lm(nRides ~ temp + hum + windspeed + workingday, data = bike)
regsummary(lmfit, anova = F, fit_measures = F)
```

```
##
## Parameter estimates
## -----------------------------------------------
##             Estimate Std. Error t value   Pr(>|t|)
## (Intercept)   4009.4     344.52 11.6374 8.0395e-29
## temp          6609.7     293.39 22.5289 1.2388e-85
## hum          -3106.8     383.98 -8.0911 2.4838e-15
## windspeed    -4801.7     708.81 -6.7743 2.5843e-11
## workingday     125.8     113.55  1.1079 2.6826e-01
```

Variabeln `holiday` har $t_{obs} = -1.9470$ vilket är större i absolutbelopp än t-värdet för `workingday` (som är $t_{obs} = 1.1079$). `holiday` verkar därför vara en aningens bättre variabel än `workingday`. Men, $t_{obs} = -1.9470$ för `holiday` är mindre än 2 i absolutbelopp vilket innebär att `holiday` inte ska tas med i modellen. Vi stannar där för här och väljer modellen med variablerna `temp`, `hum` och `windspeed`:

```
lmfit = lm(nRides ~ temp + hum + windspeed , data = bike)
regsummary(lmfit, anova = F)
```

```
##
## Measures of model fit
## -----------------------------------------------
##   Root MSE         R2     R2-adj
## 1425.30539    0.46090    0.45867
##
## Parameter estimates
## -----------------------------------------------
##             Estimate Std. Error t value   Pr(>|t|)
## (Intercept)   4084.4     337.86 12.0888 8.7098e-31
## temp          6625.5     293.09 22.6062 4.1807e-86
## hum          -3100.1     383.99 -8.0734 2.8330e-15
## windspeed    -4806.9     708.90 -6.7808 2.4754e-11
```

Backward selection

Backward selection metoden börjar med alla variabler i modellen och börjar sen ta bort icke-signifikanta variabler, en och en i taget. I varje steg tar vi bort den variabel som har lägst t-kvot, om den t-kvoten är mindre än 2 i absolutbelopp. Om den minsta t-kvoten är större än 2 i absolutbelopp drar vi slutsatsen att alla kvarvarande variabler behövs i modellen, och vi stannar då med den modellen som vårt slutgiltiga modell. Here we go:

Steg 1 - kan någon variabel tas bort?

Vi skattar först modellen med alla fem förklarande variabler:

```
lmfit = lm(nRides ~ temp + hum + windspeed + holiday + workingday , data = bike)
regsummary(lmfit)
```

```
##
## Analysis of variance - ANOVA
## -----------------------------------------------
##        df         SS        MS      F    Pr(>F)
## Regr    5 1271139382 254227876 125.52 1.102e-95
## Error 725 1468396010   2025374
## Total 730 2739535392
##
## Measures of model fit
## -----------------------------------------------
##  Root MSE        R2    R2-adj
## 1423.1563    0.4640    0.4603
##
## Parameter estimates
## -----------------------------------------------
##              Estimate Std. Error  t value   Pr(>|t|)
## (Intercept)  4068.222     345.75 11.76632 2.2554e-29
## temp         6602.207     293.02 22.53134 1.2737e-85
## hum         -3112.540     383.47 -8.11668 2.0526e-15
## windspeed   -4805.227     707.85 -6.78844 2.3596e-11
## holiday      -561.092     325.77 -1.72236 8.5430e-02
## workingday     74.983     117.17  0.63994 5.2242e-01
```

Vi ser från utskriften att variabeln `workingday` har lägst t-kvot (i absolutebelopp, dvs $|0.63994| < |-1.72236|$). Eftersom och eftersom t-kvoten för `workingday` är mindre än 2 i absolutbelopp så kastar vi ut denna variabel från modellen, och fortsätter med variablerna `temp`, `hum`, `windspeed` och `holiday` till nästa steg.

Steg 2 - kan ytterligare en variabel tas bort?

```
lmfit = lm(nRides ~ temp + hum + windspeed + holiday, data = bike)
regsummary(lmfit)
```

```
##
## Analysis of variance - ANOVA
## ------------------------------------------------
##        df         SS        MS      F     Pr(>F)
## Regr    4 1270309954 317577488 156.93 1.0118e-96
## Error 726 1469225438   2023726
## Total 730 2739535392
##
## Measures of model fit
## ------------------------------------------------
##   Root MSE         R2     R2-adj
## 1422.57741    0.46370    0.46074
##
## Parameter estimates
## ------------------------------------------------
##             Estimate Std. Error t value   Pr(>|t|)
## (Intercept)  4115.59     337.60 12.1909 3.1149e-31
## temp         6610.34     292.63 22.5895 5.5379e-86
## hum         -3109.33     383.29 -8.1123 2.1175e-15
## windspeed   -4808.48     707.55 -6.7960 2.2444e-11
## holiday      -613.59     315.14 -1.9470 5.1917e-02
```

Vi ser att `holiday` har lägst t-kvot i absolutbelopp, och att denna absoluta t-kvot är mindre än 2. Även `holiday` åker ut ur modellen. Vi fortsätter och ser om vi ska göra oss med ytterligare en variabel.

Steg 3 - kan vi göra oss av med ytterligare en variabel?

```
lmfit = lm(nRides ~ temp + hum + windspeed, data = bike)
regsummary(lmfit)
```

```
##
## Analysis of variance - ANOVA
## ------------------------------------------------
##        df         SS        MS      F     Pr(>F)
## Regr    3 1262638191 420879397 207.18 4.2551e-97
## Error 727 1476897201   2031495
## Total 730 2739535392
##
## Measures of model fit
## ------------------------------------------------
##   Root MSE         R2     R2-adj
## 1425.30539    0.46090    0.45867
##
## Parameter estimates
## ------------------------------------------------
##             Estimate Std. Error t value   Pr(>|t|)
## (Intercept)   4084.4     337.86 12.0888 8.7098e-31
## temp          6625.5     293.09 22.6062 4.1807e-86
## hum          -3100.1     383.99 -8.0734 2.8330e-15
## windspeed    -4806.9     708.90 -6.7808 2.4754e-11
```

Variabeln `windspeed` har lägst t-kvot i absolutbelopp, men $|-6.7808| = 6.7808$ är större än 2, så vi vill inte ta bort `windspeed` från modellen. Vi stannar alltså här och vår slutliga modell blir alltså modellen med `temp`, `hum` och `windspeed`.

I det här exemplet gav forward och backward selection samma slutliga modell. Så är dock inte alltid.