

Stepwise Variable Selection on the Body Fat Data

Columns

```
 3 - 5 Case Number
10 - 13 Percent body fat using Brozek's equation,
      457/Density - 414.2
18 - 21 Percent body fat using Siri's equation,
      495/Density - 450
24 - 29 Density (gm/cm^3)
36 - 37 Age (yrs)
40 - 45 Weight (lbs)
49 - 53 Height (inches)
58 - 61 Adiposity index = Weight/Height^2 (kg/m^2)
65 - 69 Fat Free Weight
      = (1 - fraction of body fat) * Weight,
      using Brozek's formula (lbs)
74 - 77 Neck circumference (cm)
81 - 85 Chest circumference (cm)
89 - 93 Abdomen circumference (cm) "at the umbilicus
      and level with the iliac crest"
97 - 101 Hip circumference (cm)
106 - 109 Thigh circumference (cm)
114 - 117 Knee circumference (cm)
122 - 125 Ankle circumference (cm)
130 - 133 Extended biceps circumference (cm)
138 - 141 Forearm circumference (cm)
146 - 149 Wrist circumference (cm) "distal to the
      styloid processes"
```

```
> bodyfat =
read.table("http://www.utstat.toronto.edu/~brunner/data/illegal/bodyfat.data")
> attach(bodyfat)
> percentfat = (Brozek+Siri)/2
> nothing = lm(percentfat ~ 1) # Just the intercept
> summary(nothing)
```

Call:

```
lm(formula = percentfat ~ 1)
```

Residuals:

Min	1Q	Median	3Q	Max
-19.0446	-6.4071	0.0554	5.9054	27.2554

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	19.0446	0.5077	37.51	<2e-16 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 8.059 on 251 degrees of freedom

```
> everything = lm(percentfat ~ age + weight + height + bmi + neck + chest +
+ belly + hip + thigh + knee + ankle + biceps + forearm + wrist)
> summary(everything)
```

Call:

```
lm(formula = percentfat ~ age + weight + height + bmi + neck +
    chest + belly + hip + thigh + knee + ankle + biceps + forearm +
    wrist)
```

Residuals:

Min	1Q	Median	3Q	Max
-10.709	-2.720	-0.105	3.060	9.611

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	-16.62898	16.74566	-0.993	0.32171
age	0.05953	0.03121	1.907	0.05770 .
weight	-0.08546	0.05186	-1.648	0.10069
height	-0.05454	0.10753	-0.507	0.61250
bmi	0.06645	0.28895	0.230	0.81832
neck	-0.46216	0.22703	-2.036	0.04290 *
chest	-0.03165	0.10166	-0.311	0.75586
belly	0.91307	0.08883	10.279	< 2e-16 ***
hip	-0.20804	0.14249	-1.460	0.14560
thigh	0.22908	0.14092	1.626	0.10536
knee	0.01344	0.23889	0.056	0.95518
ankle	0.16187	0.21577	0.750	0.45388
biceps	0.16253	0.16635	0.977	0.32956
forearm	0.44036	0.19217	2.292	0.02281 *
wrist	-1.55155	0.51630	-3.005	0.00294 **

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.154 on 237 degrees of freedom

Multiple R-squared: 0.7492, Adjusted R-squared: 0.7343

F-statistic: 50.56 on 14 and 237 DF, p-value: < 2.2e-16

```
> # Suggests a model with just neck, belly, forearm and wrist.
```

```
> red1 = lm(percentfat ~ neck + belly + forearm + wrist)
```

```
> anova(red1, everything)
Analysis of Variance Table
```

```
Model 1: percentfat ~ neck + belly + forearm + wrist
```

```
Model 2: percentfat ~ age + weight + height + bmi + neck + chest + belly +
hip + thigh + knee + ankle + biceps + forearm + wrist
```

```
Res.Df  RSS Df Sum of Sq    F    Pr(>F)
1     247 4662.2
2     237 4089.5 10     572.74 3.3192 0.0004701 ***
```

```
---
```

```
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
>
```

Automatic variable selection in R is based on the Akaike information criterion (AIC). The AIC is a measure of how “bad” a model is, based on information theory. Higher SSE is bad, and lots of independent variables is bad.

```
> backwards = step(everything) # Backwards elimination is the default
```

```
Start: AIC=732.26
```

```
percentfat ~ age + weight + height + bmi + neck + chest + belly +
hip + thigh + knee + ankle + biceps + forearm + wrist
```

	Df	Sum of Sq	RSS	AIC
- knee	1	0.05	4089.5	730.26
- bmi	1	0.91	4090.4	730.31
- chest	1	1.67	4091.1	730.36
- height	1	4.44	4093.9	730.53
- ankle	1	9.71	4099.2	730.86
- biceps	1	16.47	4105.9	731.27
<none>			4089.5	732.26
- hip	1	36.78	4126.2	732.51
- thigh	1	45.60	4135.1	733.05
- weight	1	46.86	4136.3	733.13
- age	1	62.77	4152.2	734.10
- neck	1	71.50	4161.0	734.63
- forearm	1	90.61	4180.1	735.78
- wrist	1	155.83	4245.3	739.68
- belly	1	1823.04	5912.5	823.16

Step: AIC=730.26

percentfat ~ age + weight + height + bmi + neck + chest + belly +
hip + thigh + ankle + biceps + forearm + wrist

	Df	Sum of Sq	RSS	AIC
- bmi	1	0.86	4090.4	728.31
- chest	1	1.68	4091.2	728.36
- height	1	4.48	4094.0	728.54
- ankle	1	10.37	4099.9	728.90
- biceps	1	16.42	4105.9	729.27
<none>			4089.5	730.26
- hip	1	36.78	4126.3	730.52
- weight	1	49.55	4139.1	731.30
- thigh	1	51.17	4140.7	731.39
- age	1	67.64	4157.2	732.40
- neck	1	72.86	4162.4	732.71
- forearm	1	91.80	4181.3	733.86
- wrist	1	156.33	4245.8	737.72
- belly	1	1823.42	5912.9	821.18

Step: AIC=728.31

percentfat ~ age + weight + height + neck + chest + belly + hip +
thigh + ankle + biceps + forearm + wrist

	Df	Sum of Sq	RSS	AIC
- chest	1	1.08	4091.4	726.38
- height	1	9.10	4099.5	726.87
- ankle	1	11.11	4101.5	727.00
- biceps	1	17.78	4108.1	727.41
<none>			4090.4	728.31
- hip	1	35.92	4126.3	728.52
- weight	1	49.14	4139.5	729.32
- thigh	1	52.90	4143.3	729.55
- age	1	66.91	4157.3	730.40
- neck	1	72.43	4162.8	730.74
- forearm	1	91.81	4182.2	731.91
- wrist	1	156.15	4246.5	735.76
- belly	1	2109.47	6199.8	831.12

Step: AIC=726.38

percentfat ~ age + weight + height + neck + belly + hip + thigh +
ankle + biceps + forearm + wrist

	Df	Sum of Sq	RSS	AIC
- height	1	8.17	4099.6	724.88
- ankle	1	11.41	4102.9	725.08
- biceps	1	17.15	4108.6	725.43
<none>			4091.4	726.38
- hip	1	34.85	4126.3	726.52
- thigh	1	57.90	4149.3	727.92
- weight	1	65.69	4157.1	728.39
- age	1	66.70	4158.1	728.46
- neck	1	73.40	4164.9	728.86
- forearm	1	90.73	4182.2	729.91
- wrist	1	155.61	4247.1	733.79
- belly	1	2675.31	6766.8	851.17

Step: AIC=724.88

percentfat ~ age + weight + neck + belly + hip + thigh + ankle +
biceps + forearm + wrist

	Df	Sum of Sq	RSS	AIC
- ankle	1	12.22	4111.8	723.63
- biceps	1	19.17	4118.8	724.06
- hip	1	29.28	4128.9	724.68
<none>			4099.6	724.88
- thigh	1	66.20	4165.8	726.92
- neck	1	71.13	4170.7	727.22
- age	1	73.58	4173.2	727.37
- forearm	1	93.04	4192.7	728.54
- weight	1	111.10	4210.7	729.62
- wrist	1	166.05	4265.7	732.89
- belly	1	2953.74	7053.4	859.62

Step: AIC=723.63

percentfat ~ age + weight + neck + belly + hip + thigh + biceps +
forearm + wrist

	Df	Sum of Sq	RSS	AIC
- biceps	1	17.69	4129.5	722.72
- hip	1	30.52	4142.4	723.50
<none>			4111.8	723.63
- thigh	1	69.49	4181.3	725.86
- age	1	70.22	4182.0	725.90
- neck	1	80.61	4192.4	726.53
- forearm	1	92.43	4204.3	727.24
- weight	1	99.65	4211.5	727.67
- wrist	1	153.88	4265.7	730.89
- belly	1	2957.26	7069.1	858.18

Step: AIC=722.72

percentfat ~ age + weight + neck + belly + hip + thigh + forearm +
wrist

	Df	Sum of Sq	RSS	AIC
<none>			4129.5	722.72
- hip	1	34.86	4164.4	722.83
- neck	1	73.34	4202.9	725.15
- age	1	75.61	4205.1	725.29
- weight	1	87.13	4216.7	725.98
- thigh	1	95.44	4225.0	726.47
- forearm	1	131.59	4261.1	728.62
- wrist	1	152.85	4282.4	729.87
- belly	1	2940.68	7070.2	856.22

> # backwards = step(everything,trace=0) would suppress step by step
output.

```
> summary(backwards)
```

```
Call:
```

```
lm(formula = percentfat ~ age + weight + neck + belly + hip +  
    thigh + forearm + wrist)
```

```
Residuals:
```

```
      Min       1Q   Median       3Q      Max  
-10.5166  -2.8607  -0.1727   2.8425   9.8610
```

```
Coefficients:
```

```
              Estimate Std. Error t value Pr(>|t|)  
(Intercept) -21.35925    11.27743  -1.894  0.05942 .  
age           0.06250     0.02963   2.109  0.03594 *  
weight       -0.08699     0.03842  -2.264  0.02444 *  
neck         -0.44923     0.21625  -2.077  0.03882 *  
belly        0.91101     0.06925  13.155 < 2e-16 ***  
hip          -0.19092     0.13331  -1.432  0.15338  
thigh        0.29442     0.12423   2.370  0.01858 *  
forearm      0.49913     0.17937   2.783  0.00581 **  
wrist       -1.47076     0.49041  -2.999  0.00299 **
```

```
---
```

```
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
Residual standard error: 4.122 on 243 degrees of freedom
```

```
Multiple R-squared: 0.7467, Adjusted R-squared: 0.7384
```

```
F-statistic: 89.54 on 8 and 243 DF, p-value: < 2.2e-16
```

R-squared = 0.7467 for a model with 8 variables vs. 0.7492 for model with all 14 variables.

I would be inclined to drop hip.

```

> # Try forward selection
> forwards = step(nothing,
+ scope=list(lower=formula(nothing),upper=formula(everything)),
direction="forward")
Start: AIC=1052.76
percentfat ~ 1

```

	Df	Sum of Sq	RSS	AIC
+ belly	1	10792.1	5510.9	781.43
+ bmi	1	8635.0	7668.0	864.68
+ chest	1	8052.3	8250.6	883.13
+ hip	1	6378.1	9924.8	929.69
+ weight	1	6122.4	10180.6	936.10
+ thigh	1	5120.8	11182.1	959.75
+ knee	1	4211.8	12091.2	979.44
+ biceps	1	3965.4	12337.5	984.53
+ neck	1	3931.2	12371.8	985.22
+ forearm	1	2140.2	14162.8	1019.29
+ wrist	1	1963.9	14339.1	1022.41
+ age	1	1374.7	14928.3	1032.56
+ ankle	1	1156.8	15146.2	1036.21
+ height	1	130.0	16172.9	1052.74
<none>			16303.0	1052.76

```

Step: AIC=781.43
percentfat ~ belly

```

	Df	Sum of Sq	RSS	AIC
+ weight	1	927.38	4583.5	737.00
+ wrist	1	654.46	4856.4	751.57
+ neck	1	566.91	4944.0	756.08
+ hip	1	507.03	5003.8	759.11
+ height	1	424.60	5086.3	763.23
+ knee	1	298.99	5211.9	769.38
+ ankle	1	215.02	5295.9	773.40
+ age	1	182.35	5328.5	774.95
+ chest	1	181.23	5329.6	775.01
+ thigh	1	158.38	5352.5	776.08
+ biceps	1	126.30	5384.6	777.59
+ bmi	1	63.63	5447.2	780.51
+ forearm	1	48.62	5462.3	781.20
<none>			5510.9	781.43

Step: AIC=737

percentfat ~ belly + weight

	Df	Sum of Sq	RSS	AIC
+ wrist	1	144.919	4438.6	730.90
+ neck	1	80.014	4503.5	734.56
+ thigh	1	77.681	4505.8	734.69
+ forearm	1	63.684	4519.8	735.47
+ biceps	1	57.798	4525.7	735.80
+ height	1	37.528	4546.0	736.93
<none>			4583.5	737.00
+ bmi	1	15.229	4568.3	738.16
+ knee	1	7.948	4575.5	738.56
+ age	1	2.171	4581.3	738.88
+ ankle	1	1.441	4582.1	738.92
+ chest	1	0.016	4583.5	739.00
+ hip	1	0.009	4583.5	739.00

Step: AIC=730.9

percentfat ~ belly + weight + wrist

	Df	Sum of Sq	RSS	AIC
+ forearm	1	120.745	4317.8	725.95
+ biceps	1	80.582	4358.0	728.29
+ thigh	1	39.248	4399.3	730.66
<none>			4438.6	730.90
+ neck	1	23.142	4415.4	731.58
+ height	1	21.879	4416.7	731.66
+ age	1	18.181	4420.4	731.87
+ knee	1	17.448	4421.1	731.91
+ ankle	1	13.945	4424.6	732.11
+ bmi	1	13.886	4424.7	732.11
+ hip	1	8.655	4429.9	732.41
+ chest	1	1.094	4437.5	732.84

Step: AIC=725.95

percentfat ~ belly + weight + wrist + forearm

	Df	Sum of Sq	RSS	AIC
+ neck	1	47.303	4270.5	725.18
<none>			4317.8	725.95
+ age	1	33.777	4284.1	725.97
+ biceps	1	29.886	4287.9	726.20
+ thigh	1	26.556	4291.3	726.40
+ ankle	1	16.936	4300.9	726.96
+ height	1	16.840	4301.0	726.97
+ knee	1	16.806	4301.0	726.97
+ bmi	1	4.984	4312.8	727.66
+ hip	1	3.297	4314.5	727.76
+ chest	1	0.523	4317.3	727.92

Step: AIC=725.18

percentfat ~ belly + weight + wrist + forearm + neck

	Df	Sum of Sq	RSS	AIC
+ age	1	42.424	4228.1	724.66
+ biceps	1	40.779	4229.8	724.76
<none>			4270.5	725.18
+ thigh	1	24.545	4246.0	725.72
+ height	1	17.604	4252.9	726.14
+ bmi	1	10.925	4259.6	726.53
+ hip	1	10.230	4260.3	726.57
+ ankle	1	9.969	4260.6	726.59
+ knee	1	8.492	4262.0	726.67
+ chest	1	0.002	4270.5	727.18

Step: AIC=724.66

percentfat ~ belly + weight + wrist + forearm + neck + age

	Df	Sum of Sq	RSS	AIC
+ thigh	1	63.724	4164.4	722.83
+ biceps	1	42.739	4185.4	724.10
<none>			4228.1	724.66
+ height	1	17.708	4210.4	725.60
+ bmi	1	14.981	4213.1	725.77
+ ankle	1	13.703	4214.4	725.84
+ knee	1	5.236	4222.9	726.35
+ hip	1	3.139	4225.0	726.47
+ chest	1	0.834	4227.3	726.61

Step: AIC=722.83

percentfat ~ belly + weight + wrist + forearm + neck + age + thigh

	Df	Sum of Sq	RSS	AIC
+ hip	1	34.857	4129.5	722.72
<none>			4164.4	722.83
+ biceps	1	22.026	4142.4	723.50
+ ankle	1	11.822	4152.6	724.12
+ height	1	3.968	4160.4	724.59
+ bmi	1	3.948	4160.4	724.59
+ chest	1	0.679	4163.7	724.79
+ knee	1	0.068	4164.3	724.83

Step: AIC=722.72

percentfat ~ belly + weight + wrist + forearm + neck + age + thigh + hip

	Df	Sum of Sq	RSS	AIC
<none>			4129.5	722.72
+ biceps	1	17.6920	4111.8	723.63
+ height	1	10.9433	4118.6	724.05
+ ankle	1	10.7437	4118.8	724.06
+ bmi	1	8.8584	4120.7	724.17
+ chest	1	0.0012	4129.5	724.72
+ knee	1	0.0006	4129.5	724.72

```
> summary(forwards)
```

```
Call:
```

```
lm(formula = percentfat ~ belly + weight + wrist + forearm +  
    neck + age + thigh + hip)
```

```
Residuals:
```

```
      Min       1Q   Median       3Q      Max  
-10.5166  -2.8607  -0.1727   2.8425   9.8610
```

```
Coefficients:
```

```
              Estimate Std. Error t value Pr(>|t|)  
(Intercept) -21.35925   11.27743  -1.894  0.05942 .  
belly         0.91101    0.06925  13.155 < 2e-16 ***  
weight       -0.08699    0.03842  -2.264  0.02444 *  
wrist        -1.47076    0.49041  -2.999  0.00299 **  
forearm       0.49913    0.17937   2.783  0.00581 **  
neck         -0.44923    0.21625  -2.077  0.03882 *  
age           0.06250    0.02963   2.109  0.03594 *  
thigh         0.29442    0.12423   2.370  0.01858 *  
hip          -0.19092    0.13331  -1.432  0.15338
```

```
---
```

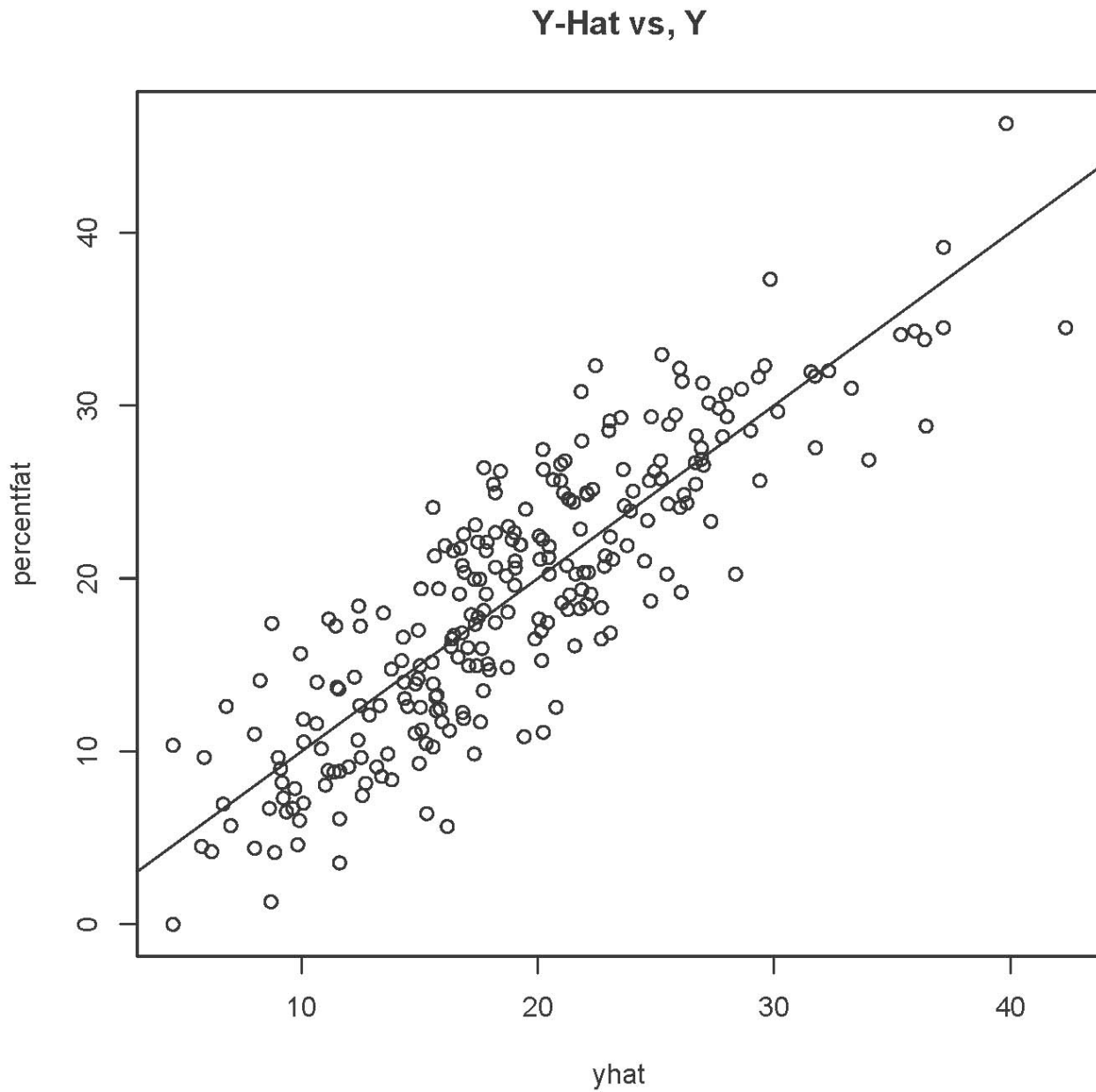
```
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
Residual standard error: 4.122 on 243 degrees of freedom
```

```
Multiple R-squared: 0.7467, Adjusted R-squared: 0.7384
```

```
F-statistic: 89.54 on 8 and 243 DF, p-value: < 2.2e-16
```

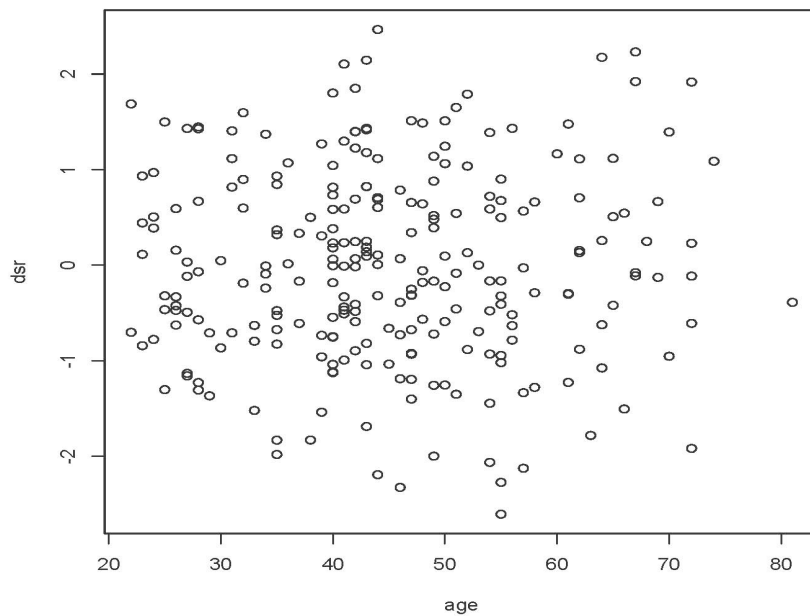
```
>  
> yhat = backwards$fitted.values  
> plot(yhat,percentfat); title("Y-Hat vs, Y")  
> lines(c(0,50),c(0,50))
```



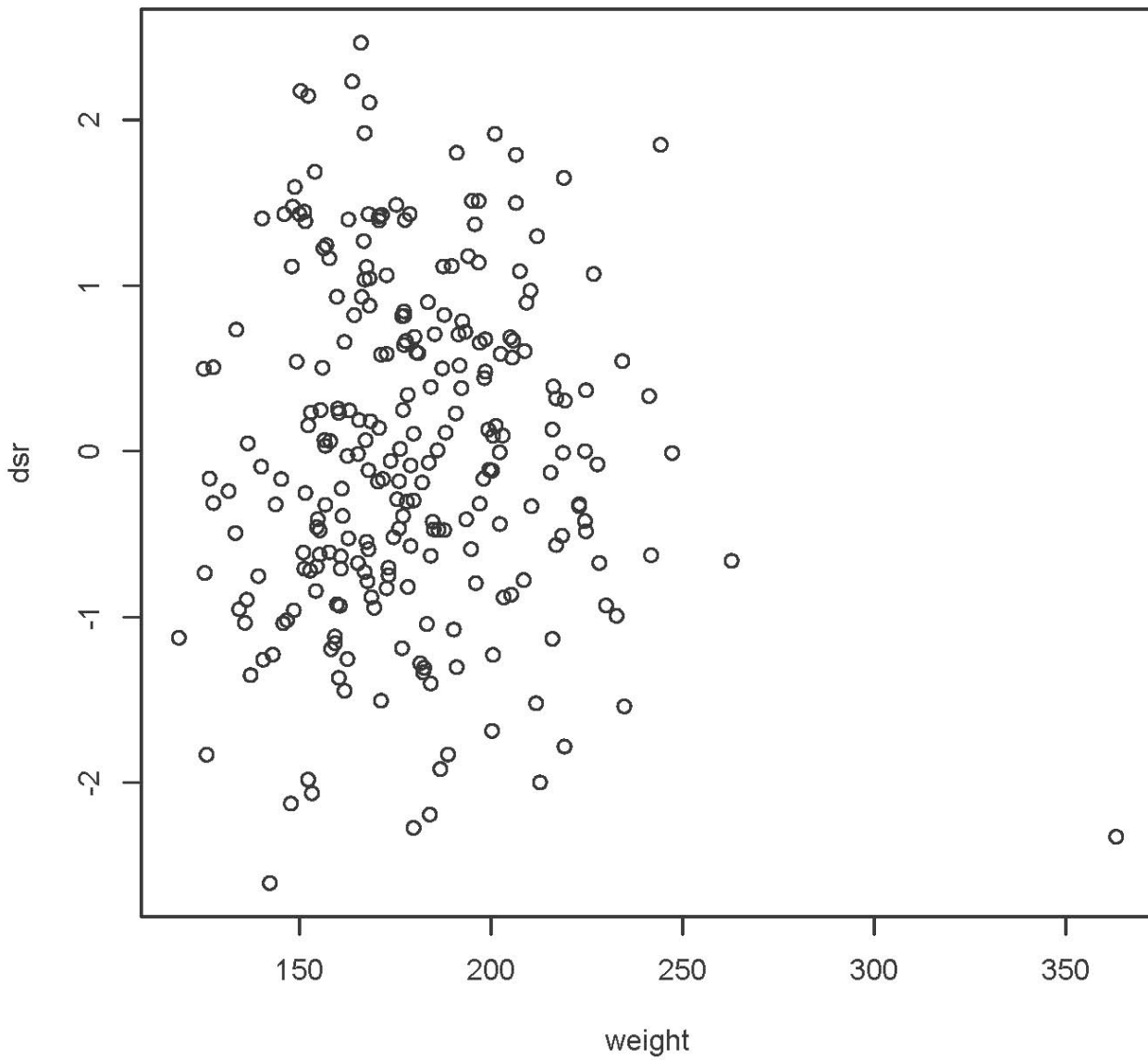
```

>
> # Deleted Studentized residuals
> dsr = rstudent(backwards)
> hist(dsr)
> n = length(age); n
[1] 252
> a = 0.05/n; a
[1] 0.0001984127
> critval = qt(1-a/2,backwards$df.residual-1); critval
[1] 3.778889
> dsr[abs(dsr)>critval]
named numeric(0)
> summary(dsr)
      Min.      1st Qu.      Median      Mean      3rd Qu.      Max.
-2.6044037 -0.7034727 -0.0425321 -0.0006917  0.6950153  2.4655851
>
>
>
> plot(age,dsr)

```



```
> plot(weight,dsr) # Same pattern for all the rest except forearm
> plot(neck,dsr)
> plot(belly,dsr)
> plot(hip,dsr)
> plot(thigh,dsr)
> plot(forearm,dsr)
>
```



```

> bigguy = (1:n)[weight>300]; bigguy
[1] 39
> bodyfat[bigguy,] # Just that row, all columns
  Brozek Siri Density age weight height  bmi fatfree neck chest belly  hip thigh knee ankle biceps
39  33.8 35.2  1.0202  46 363.15  72.25 48.9   240.5 51.2 136.2 148.1 147.7  87.3 49.1  29.6    45
  forearm wrist
39      29  21.4

> kilos = weight[bigguy]/2.2; kilos
[1] 165.0682
>
> percentfat[bigguy]; max(percentfat)
[1] 34.5
[1] 46.3

> dsr[bigguy]
      39
-2.325133

> hii = hatvalues(backwards) # Leverage
> summary(hii)
  Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
0.009032 0.020841 0.028571 0.035714 0.039243 0.317434
> hii[bigguy]
      39
0.3174344
> hii[hii>0.2] # High leverage
      39      175
0.3174344 0.2350573

```

Two possibly influential observations, including the big guy. He is a definite outlier in x .

This document was prepared by [Jerry Brunner](#), University of Toronto. It is licensed under a Creative Commons Attribution - ShareAlike 3.0 Unported License:

http://creativecommons.org/licenses/by-sa/3.0/deed.en_US. Use any part of it as you like and share the result freely. The Open Office document is available from the course website at

<http://www.utstat.toronto.edu/~brunner/oldclass/302f17>