STA221

Neil Montgomery

Last edited: 2017-08-08 19:10

more possibilities - indicators and interactions

indicator, or “dummy” variables

An input variable in a multiple regression model can be just about anything (with minimal technical requirements).

A special and very useful example is a variable with only two possible values: 0 and 1.

indicator, or “dummy” variables

An input variable in a multiple regression model can be just about anything (with minimal technical requirements).

A special and very useful example is a variable with only two possible values: 0 and 1.

This is called an *indicator*, or dummy variable. The 0 and 1 values have no numerical meaning. They only divide the dataset into two groups.

indicator, or "dummy" variables

An input variable in a multiple regression model can be just about anything (with minimal technical requirements).

A special and very useful example is a variable with only two possible values: 0 and 1.

This is called an *indicator*, or dummy variable. The 0 and 1 values have no numerical meaning. They only divide the dataset into two groups.

For example, question 28.2 "Pizza" has results from the assessment of $n = 29$ frozen pizza brands.

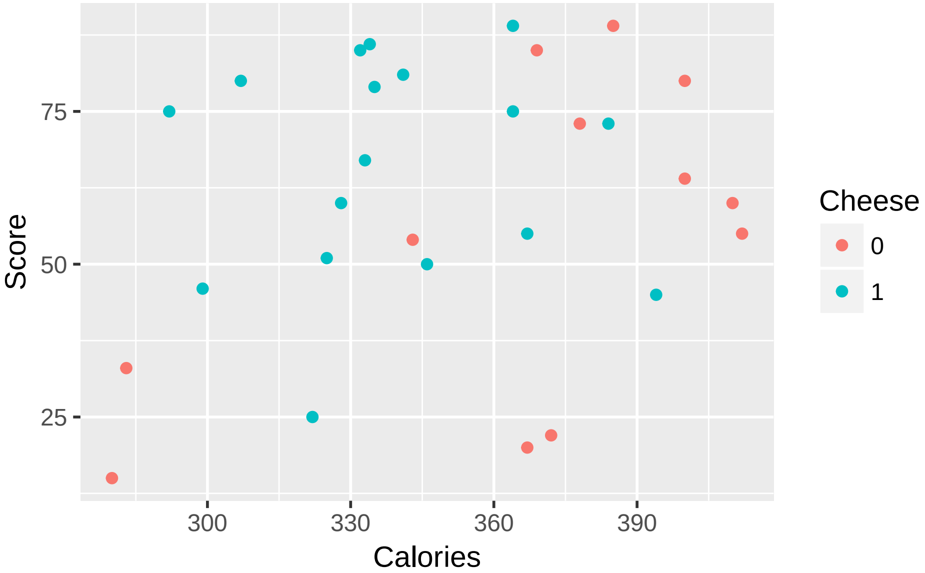
pizza

Here's a glance at the data. The last two columns are redundant.

```
## # A tibble: 29 × 7
##                       Brand Score  Cost Calories   Fat   Type Cheese
##                       <chr> <dbl> <dbl>    <dbl> <dbl> <fctr> <fctr>
## 1         Freshetta 4 Cheese    89  0.98      364    15 cheese      1
## 2 Freschetta stuffed crust    86  1.23      334    11 cheese      1
## 3                  DiGiorno    85  0.94      332    12 cheese      1
## 4              Amy's organic    81  1.92      341    14 cheese      1
## 5                   Safeway    80  0.84      307     9 cheese      1
## # ... with 24 more rows
```

They are there for "software" reasons.

Score versus Calories plotted



model with a dummy variable

What is the meaning of β_2:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$$

when x_2 is a dummy variable?

model with a dummy variable

What is the meaning of β_2:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$$

when x_2 is a dummy variable?

It lets you fit parallel lines with different intercepts.

pizza with Cheese dummy fitted

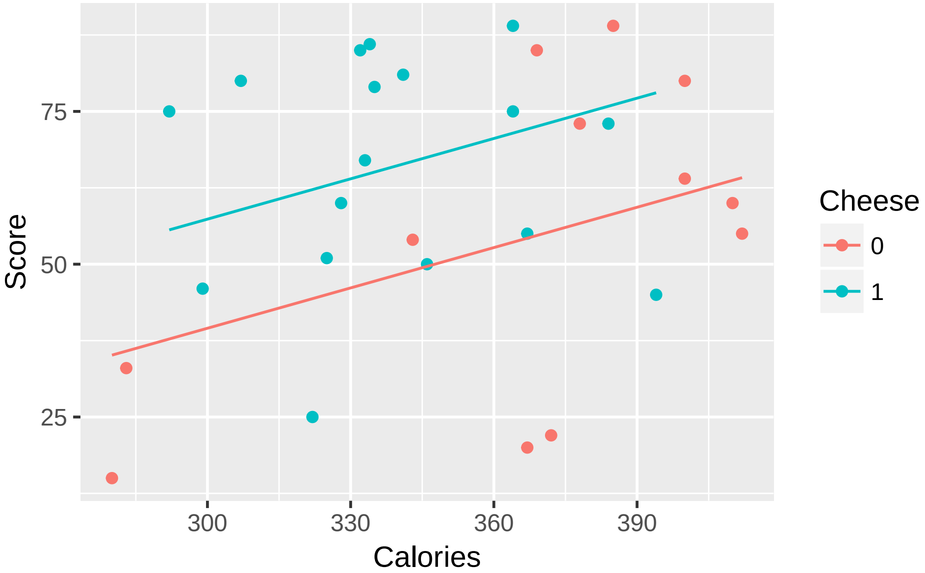
```
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -26.4510    41.2354  -0.641   0.5268
## Calories      0.2199     0.1113   1.976   0.0589
## Cheese1      17.8476     8.3603   2.135   0.0424
##
## Residual standard error: 20.65 on 26 degrees of freedom
## Multiple R-squared:  0.1929, Adjusted R-squared:  0.1308
## F-statistic: 3.107 on 2 and 26 DF,  p-value: 0.06168
```

`Cheese1` is R-speak for *this line is about the impact of 'Cheese' with baseline value '1'.*

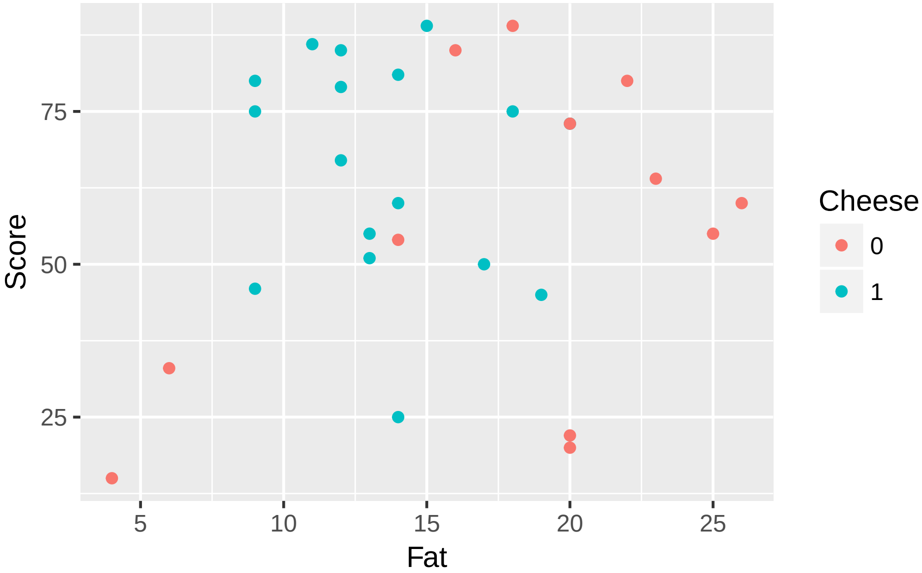
pizza with Cheese dummy fitted

```
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)    -8.6034    38.0890  -0.226   0.8231
## Calories        0.2199     0.1113   1.976   0.0589
## Typepepperoni -17.8476     8.3603  -2.135   0.0424
##
## Residual standard error: 20.65 on 26 degrees of freedom
## Multiple R-squared:  0.1929, Adjusted R-squared:  0.1308
## F-statistic: 3.107 on 2 and 26 DF,  p-value: 0.06168
```

pizza plotted with shifted lines (two intercepts)



Fat and Score by Cheese plotted



interaction with a dummy variable

Another use of dummy variables is to allow for different intercepts *and* slopes.

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 + \varepsilon$$

interaction with a dummy variable

Another use of dummy variables is to allow for different intercepts *and* slopes.

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 + \varepsilon$$

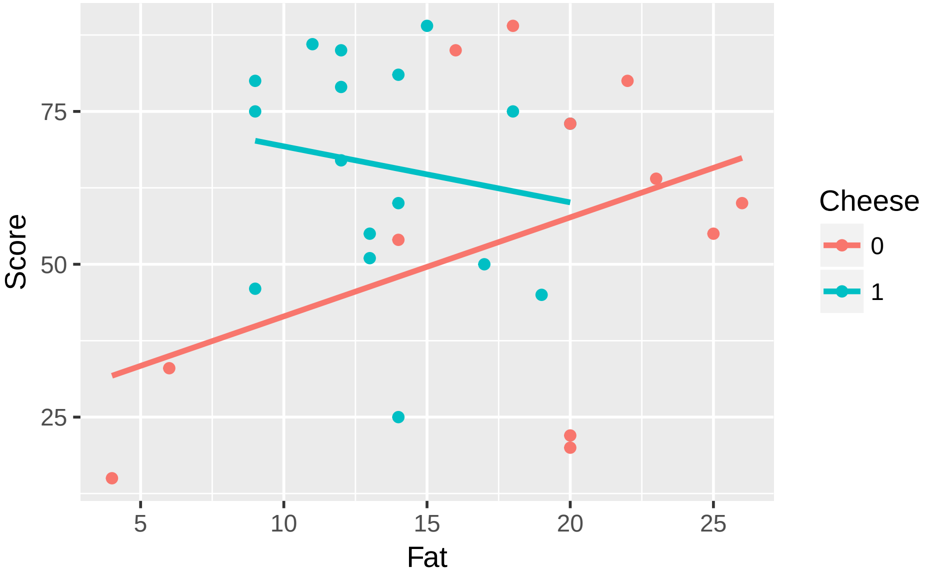
The $x_1 x_2$ term is called an *interaction* term, which allows the impact of x_1 to change as a function of x_2.

Interaction is not limited to the case of one of them being a dummy variable.

pizza with interaction

```
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)   25.2850    17.5776   1.438   0.1627
## Fat            1.6195     0.9241   1.753   0.0919
## Cheese1       53.1752    28.1152   1.891   0.0702
## Fat:Cheese1   -2.5365     1.8217  -1.392   0.1761
##
## Residual standard error: 21.19 on 25 degrees of freedom
## Multiple R-squared:  0.1832, Adjusted R-squared:  0.08518
## F-statistic: 1.869 on 3 and 25 DF,  p-value: 0.1607
```

pizza with two slopes/intercepts



fun fact: t-test versus regression - I

```
##
##  Two Sample t-test
##
## data:  Score by Cheese
## t = -1.4441, df = 27, p-value = 0.1602
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -28.64695   4.98028
## sample estimates:
## mean in group 0 mean in group 1
##        54.16667        66.00000
```

fun fact: t-test versus regression - II

```
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   54.167      6.274   8.634 3.01e-09
## Cheese1       11.833      8.194   1.444     0.16
##
## Residual standard error: 21.73 on 27 degrees of freedom
## Multiple R-squared:  0.0717, Adjusted R-squared:  0.03732
## F-statistic: 2.085 on 1 and 27 DF,  p-value: 0.1602
```

relationships among the inputs

“multicollinearity”

I stated the following fact about the b_i estimates for β_i:

$$\frac{b_i - \beta_i}{\sqrt{MSE}\sqrt{c_i}} \sim t_{n-k-1}$$

where c_i is a number that reflects the relationships between x_i and the other inputs (to be revisited).

“multicollinearity”

I stated the following fact about the b_i estimates for β_i:

$$\frac{b_i - \beta_i}{\sqrt{MSE}\sqrt{c_i}} \sim t_{n-k-1}$$

where c_i is a number that reflects the relationships between x_i and the other inputs (to be revisited).

It turns out that the more accurately x_i can be expressed as a linear combination of the other x_j in the model, the larger c_i gets.

“multicollinearity”

I stated the following fact about the b_i estimates for β_i:

$$\frac{b_i - \beta_i}{\sqrt{MSE}\sqrt{c_i}} \sim t_{n-k-1}$$

where c_i is a number that reflects the relationships between x_i and the other inputs (to be revisited).

It turns out that the more accurately x_i can be expressed as a linear combination of the other x_j in the model, the larger c_i gets.

For example, when x_i and some other x_j are highly “correlated”, it means they are close to linear functions of one another.

“multicollinearity”

I stated the following fact about the b_i estimates for β_i:

$$\frac{b_i - \beta_i}{\sqrt{MSE}\sqrt{c_i}} \sim t_{n-k-1}$$

where c_i is a number that reflects the relationships between x_i and the other inputs (to be revisited).

It turns out that the more accurately x_i can be expressed as a linear combination of the other x_j in the model, the larger c_i gets.

For example, when x_i and some other x_j are highly “correlated”, it means they are close to linear functions of one another.

What happens when c_i is large?

illustration of the problem - two pairs of inputs

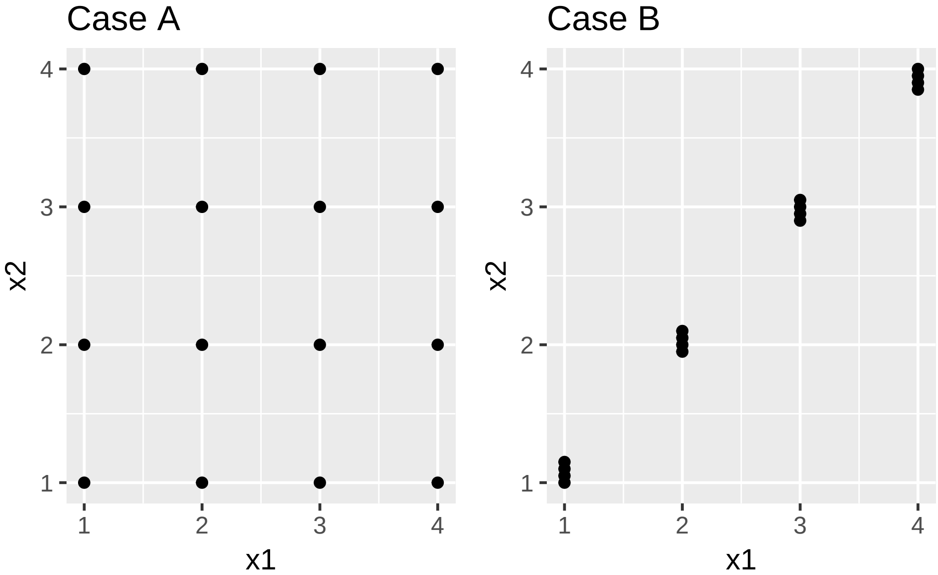


illustration of the problem

I'll generate some data from the same model in each case:

$$y = 1 + 2x_1 + 3x_2 + \varepsilon, \quad \varepsilon \sim N(0, 1)$$

Then fit the two datasets to regression models...

Case A

```
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   1.5331     1.0177   1.506    0.156
## x1            1.9401     0.2744   7.069 8.43e-06
## x2            2.8854     0.2744  10.513 1.00e-07
##
## Residual standard error: 1.227 on 13 degrees of freedom
## Multiple R-squared:  0.9251, Adjusted R-squared:  0.9135
## F-statistic: 80.25 on 2 and 13 DF,  p-value: 4.843e-08
```

Case B

```
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   1.5331     1.0177   1.506    0.156
## x1            4.1181     5.2218   0.789    0.444
## x2            0.7074     5.4890   0.129    0.899
##
## Residual standard error: 1.227 on 13 degrees of freedom
## Multiple R-squared:  0.9591, Adjusted R-squared:  0.9528
## F-statistic: 152.3 on 2 and 13 DF,  p-value: 9.506e-10
```

Note the small p-value for the overall F test.

Note that multicollinearity is merely a *possible* problem

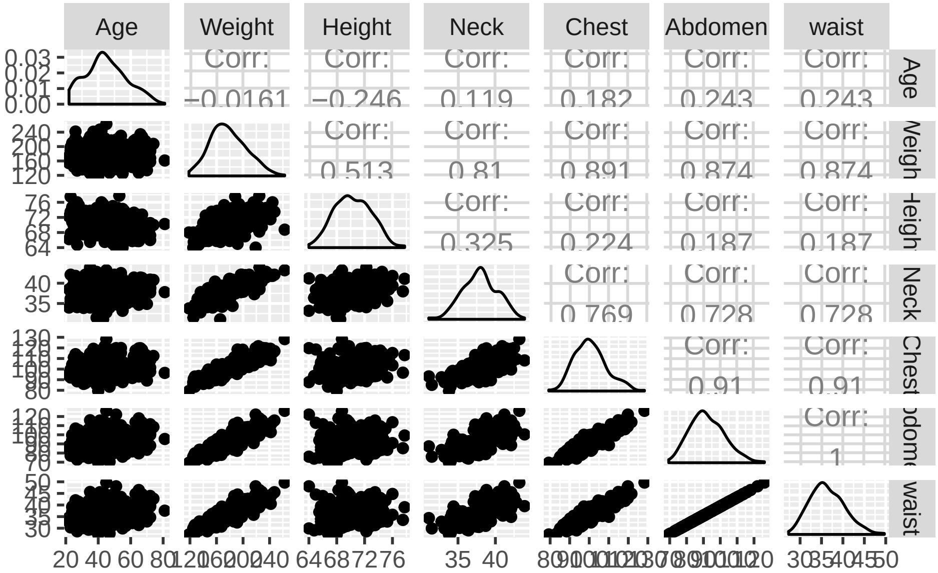
Case C: same model fit to the Case B situation but with $n = 288$

```
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   1.0510     0.1888   5.565 6.03e-08
## x1            2.1419     0.9690   2.210  0.02787
## x2            2.8299     1.0186   2.778  0.00583
##
## Residual standard error: 0.9663 on 285 degrees of freedom
## Multiple R-squared:  0.9693, Adjusted R-squared:  0.9691
## F-statistic:  4502 on 2 and 285 DF,  p-value: < 2.2e-16
```

bodyfat modeling illustration

```
##           Pct BF   Age Weight Height Neck Chest Abdomen waist
## Pct BF      1.00  0.30   0.62  -0.03 0.49  0.70    0.82  0.82
## Age         0.30  1.00  -0.02  -0.25 0.12  0.18    0.24  0.24
## Weight      0.62 -0.02   1.00   0.51 0.81  0.89    0.87  0.87
## Height     -0.03 -0.25   0.51   1.00 0.32  0.22    0.19  0.19
## Neck        0.49  0.12   0.81   0.32 1.00  0.77    0.73  0.73
## Chest       0.70  0.18   0.89   0.22 0.77  1.00    0.91  0.91
## Abdomen     0.82  0.24   0.87   0.19 0.73  0.91    1.00  1.00
## waist       0.82  0.24   0.87   0.19 0.73  0.91    1.00  1.00
```

bodyfat modeling illustration



forwards, backwards, and step-wise model selection

A very simple method is to just fit all possible models and see which one is the best (with small p-values and a nice R^2_{adj} (or any number of other single-number-summaries you might like). But there may be too many models to consider.

forwards, backwards, and step-wise model selection

A very simple method is to just fit all possible models and see which one is the best (with small p-values and a nice R^2_{adj} (or any number of other single-number-summaries you might like). But there may be too many models to consider.

A straightforward, feasible (but nevertheless flawed) model selection strategy involves one or more of:

1. "forward" start with no model terms, and add them one at a time as long as some conditions are met.

forwards, backwards, and step-wise model selection

A very simple method is to just fit all possible models and see which one is the best (with small p-values and a nice R^2_{adj} (or any number of other single-number-summaries you might like). But there may be too many models to consider.

A straightforward, feasible (but nevertheless flawed) model selection strategy involves one or more of:

1. "forward" start with no model terms, and add them one at a time as long as some conditions are met.
2. "backward" start with "all" model terms, and remove them one at a time...

forwards, backwards, and step-wise model selection

A very simple method is to just fit all possible models and see which one is the best (with small p-values and a nice R^2_{adj} (or any number of other single-number-summaries you might like). But there may be too many models to consider.

A straightforward, feasible (but nevertheless flawed) model selection strategy involves one or more of:

1. "forward" start with no model terms, and add them one at a time as long as some conditions are met.
2. "backward" start with "all" model terms, and remove them one at a time...
3. in either strategy, consider adding or removing terms when appropriate.

forwards, backwards, and step-wise model selection

A very simple method is to just fit all possible models and see which one is the best (with small p-values and a nice R^2_{adj} (or any number of other single-number-summaries you might like). But there may be too many models to consider.

A straightforward, feasible (but nevertheless flawed) model selection strategy involves one or more of:

1. "forward" start with no model terms, and add them one at a time as long as some conditions are met.
2. "backward" start with "all" model terms, and remove them one at a time...
3. in either strategy, consider adding or removing terms when appropriate.

forwards, backwards, and step-wise model selection

A very simple method is to just fit all possible models and see which one is the best (with small p-values and a nice R^2_{adj} (or any number of other single-number-summaries you might like). But there may be too many models to consider.

A straightforward, feasible (but nevertheless flawed) model selection strategy involves one or more of:

1. "forward" start with no model terms, and add them one at a time as long as some conditions are met.
2. "backward" start with "all" model terms, and remove them one at a time...
3. in either strategy, consider adding or removing terms when appropriate.

These are accessible strategies for novices, but they are known to have issues, *especially when input variables are highly "correlated".*

forwards, backwards, and step-wise model selection

A very simple method is to just fit all possible models and see which one is the best (with small p-values and a nice R^2_{adj} (or any number of other single-number-summaries you might like). But there may be too many models to consider.

A straightforward, feasible (but nevertheless flawed) model selection strategy involves one or more of:

1. "forward" start with no model terms, and add them one at a time as long as some conditions are met.
2. "backward" start with "all" model terms, and remove them one at a time...
3. in either strategy, consider adding or removing terms when appropriate.

These are accessible strategies for novices, but they are known to have issues, *especially when input variables are highly "correlated".*

There are (significantly) more sophisticated strategies also, which are worth it if you are serious about model selection.

backwards selection

Consider interactions or powers of terms when there is a rational basis for doing so.

Then, start with all input variables and remove the one with the highest p-value.

Repeat until all the p-values are small.

backwards selection

Consider interactions or powers of terms when there is a rational basis for doing so.

Then, start with all input variables and remove the one with the highest p-value.

Repeat until all the p-values are small.

Known problems specific to this procedure:

- sample size may not sensibly suppose "all" input variables

backwards selection

Consider interactions or powers of terms when there is a rational basis for doing so.

Then, start with all input variables and remove the one with the highest p-value.

Repeat until all the p-values are small.

Known problems specific to this procedure:

- sample size may not sensibly suppose “all” input variables
- p-values for variables involved in correlations may be artifically high.

backwards with bodyfat - full model F test

```
##
## Coefficients: (1 not defined because of singularities)
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)    1.68516   23.37412   0.072 0.942587
## Age            0.07189    0.03217   2.234 0.026389
## Weight        -0.01762    0.06714  -0.263 0.793153
## Height        -0.24675    0.19114  -1.291 0.197989
## Neck          -0.38682    0.23486  -1.647 0.100887
## Chest         -0.11919    0.10825  -1.101 0.272004
## Abdomen        0.90452    0.09140   9.897  < 2e-16
## waist               NA         NA      NA       NA
## Hip           -0.15878    0.14586  -1.089 0.277446
## Thigh          0.17299    0.14683   1.178 0.239926
## Knee          -0.04580    0.24560  -0.186 0.852230
## Ankle          0.18502    0.21985   0.842 0.400862
## Bicep          0.17968    0.17039   1.054 0.292732
```

backwards with bodyfat - full model all p-values

```
##
## Coefficients: (1 not defined because of singularities)
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)   1.68516   23.37412   0.072 0.942587
## Age           0.07189    0.03217   2.234 0.026389
## Weight       -0.01762    0.06714  -0.263 0.793153
## Height       -0.24675    0.19114  -1.291 0.197989
## Neck         -0.38682    0.23486  -1.647 0.100887
## Chest        -0.11919    0.10825  -1.101 0.272004
## Abdomen       0.90452    0.09140   9.897  < 2e-16
## waist              NA         NA      NA       NA
## Hip          -0.15878    0.14586  -1.089 0.277446
## Thigh         0.17299    0.14683   1.178 0.239926
## Knee         -0.04580    0.24560  -0.186 0.852230
## Ankle         0.18502    0.21985   0.842 0.400862
## Bicep         0.17968    0.17039   1.054 0.292732
## Forearm       0.27605    0.20692   1.334 0.183454
```

what's up with waist and Abdomen?

```
## # A tibble: 250 × 3
##       waist Abdomen ratio
##       <dbl>   <dbl> <dbl>
## 1 33.54331    85.2  2.54
## 2 32.67717    83.0  2.54
## 3 34.60630    87.9  2.54
## 4 34.01575    86.4  2.54
## 5 39.37008   100.0  2.54
## # ... with 245 more rows
```

backwards with bodyfat - full model all p-values

term	estimate	std.error	statistic	p.value
(Intercept)	1.685	23.374	0.072	0.943
Age	0.072	0.032	2.234	0.026
Weight	-0.018	0.067	-0.263	0.793
Height	-0.247	0.191	-1.291	0.198
Neck	-0.387	0.235	-1.647	0.101
Chest	-0.119	0.108	-1.101	0.272
Abdomen	0.905	0.091	9.897	0.000
Hip	-0.159	0.146	-1.089	0.277
Thigh	0.173	0.147	1.178	0.240
Knee	-0.046	0.246	-0.186	0.852
Ankle	0.185	0.220	0.842	0.401
Bicep	0.180	0.170	1.054	0.293
Forearm	0.276	0.207	1.334	0.183
Wrist	-1.802	0.533	-3.380	0.001

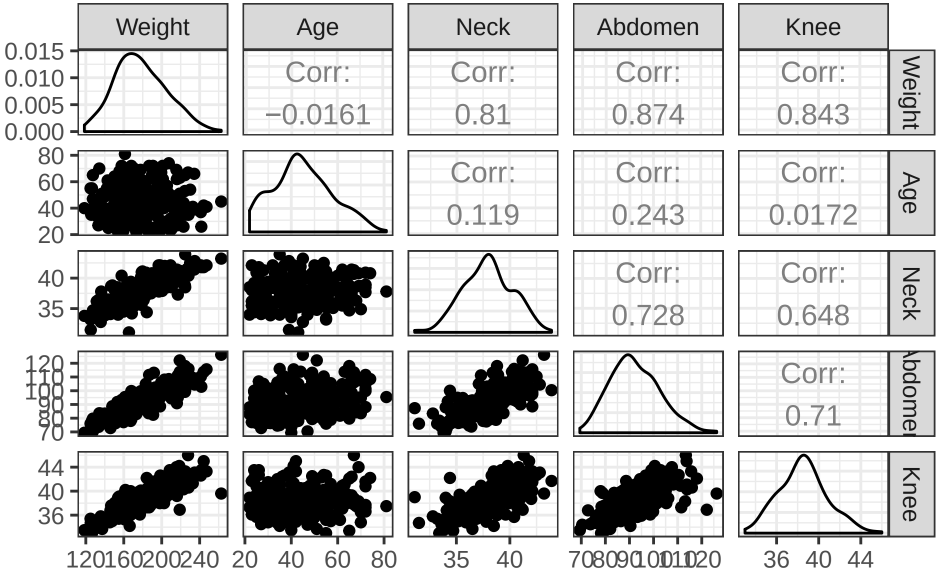
interlude - possibly doesn't mean Knee, Weight, and Ankle are chopped liver!

```
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   3.1215     9.4771   0.329  0.74216
## Knee         -0.1489     0.3366  -0.442  0.65870
## Weight        0.2297     0.0287   8.003  4.8e-14
## Ankle        -0.8348     0.3121  -2.675  0.00798
```

interlude - correlations of Weight with all others

```
##         Pct BF         Age   Height      Neck     Chest   Abdomen
## [1,] 0.6172994 -0.01605487 0.512913 0.8100143 0.8912862 0.8737351
##          waist       Hip    Thigh      Knee     Ankle    Bicep   Forearn
## [1,] 0.8737351 0.9326905 0.852116 0.8427445 0.5809059 0.785214 0.683333
##          Wrist
## [1,] 0.7251042
```

interlude - scatterplots of Weight versus some others



backwards with bodyfat: -Knee

term	estimate	std.error	statistic	p.value
(Intercept)	1.393	23.274	0.060	0.952
Age	0.070	0.031	2.266	0.024
Weight	-0.019	0.066	-0.290	0.772
Height	-0.253	0.188	-1.349	0.179
Neck	-0.383	0.233	-1.640	0.102
Chest	-0.118	0.108	-1.096	0.274
Abdomen	0.905	0.091	9.922	0.000
Hip	-0.161	0.145	-1.107	0.270
Thigh	0.165	0.140	1.176	0.241
Ankle	0.178	0.216	0.823	0.411
Bicep	0.181	0.170	1.067	0.287
Forearm	0.274	0.206	1.329	0.185
Wrist	-1.808	0.531	-3.407	0.001

backwards with bodyfat: -Knee -Weight

term	estimate	std.error	statistic	p.value
(Intercept)	7.665	8.523	0.899	0.369
Age	0.072	0.031	2.359	0.019
Height	-0.293	0.127	-2.299	0.022
Neck	-0.399	0.226	-1.767	0.078
Chest	-0.135	0.090	-1.502	0.134
Abdomen	0.895	0.085	10.575	0.000
Hip	-0.179	0.131	-1.368	0.173
Thigh	0.156	0.136	1.142	0.255
Ankle	0.164	0.210	0.781	0.436
Bicep	0.172	0.166	1.033	0.303
Forearm	0.266	0.204	1.305	0.193
Wrist	-1.837	0.521	-3.527	0.001

backwards with bodyfat: -Knee -Weight -Ankle

term	estimate	std.error	statistic	p.value
Abdomen	0.892	0.085	10.560	0.000
Wrist	-1.713	0.496	-3.456	0.001
Age	0.070	0.030	2.293	0.023
Height	-0.280	0.126	-2.218	0.027
Neck	-0.415	0.225	-1.850	0.066
Chest	-0.130	0.090	-1.447	0.149
Hip	-0.174	0.131	-1.335	0.183
Forearm	0.270	0.204	1.325	0.186
Thigh	0.165	0.136	1.214	0.226
Bicep	0.170	0.166	1.020	0.309
(Intercept)	7.685	8.516	0.902	0.368

backwards with bodyfat: -Knee -Weight -Ankle -Bicep(s)

term	estimate	std.error	statistic	p.value
Abdomen	0.885	0.084	10.511	0.000
Wrist	-1.679	0.495	-3.395	0.001
Age	0.070	0.030	2.324	0.021
Height	-0.279	0.126	-2.207	0.028
Neck	-0.388	0.223	-1.739	0.083
Forearm	0.335	0.194	1.726	0.086
Thigh	0.205	0.130	1.581	0.115
Hip	-0.176	0.131	-1.345	0.180
Chest	-0.114	0.088	-1.287	0.199
(Intercept)	6.251	8.400	0.744	0.458

backwards with bodyfat: -Knee -Weight -Ankle -Bicep(s) -Chest

term	estimate	std.error	statistic	p.value
Abdomen	0.823	0.069	11.958	0.000
Wrist	-1.731	0.494	-3.506	0.001
Age	0.073	0.030	2.396	0.017
Height	-0.268	0.126	-2.125	0.035
Neck	-0.451	0.218	-2.073	0.039
Thigh	0.224	0.129	1.735	0.084
Forearm	0.296	0.192	1.542	0.124
Hip	-0.195	0.130	-1.501	0.135
(Intercept)	5.040	8.359	0.603	0.547

backwards with bodyfat: -Knee -Weight -Ankle -Bicep(s) -Chest -Hip

term	estimate	std.error	statistic	p.value
Abdomen	0.756	0.052	14.408	0.000
Wrist	-1.851	0.488	-3.791	0.000
Age	0.081	0.030	2.718	0.007
Height	-0.322	0.121	-2.657	0.008
Neck	-0.418	0.217	-1.926	0.055
Forearm	0.288	0.192	1.499	0.135
Thigh	0.120	0.109	1.099	0.273
(Intercept)	2.541	8.212	0.309	0.757

backwards with bodyfat: -Knee -Weight -Ankle -Bicep(s) -Chest -Hip -Thigh (could stop here)

term	estimate	std.error	statistic	p.value
Abdomen	0.793	0.040	19.703	0.000
Wrist	-1.789	0.485	-3.686	0.000
Height	-0.315	0.121	-2.601	0.010
Age	0.063	0.025	2.532	0.012
Neck	-0.391	0.216	-1.813	0.071
Forearm	0.315	0.191	1.653	0.100
(Intercept)	3.607	8.159	0.442	0.659

backwards with bodyfat: -Knee -Weight -Ankle -Bicep(s) -Chest -Hip -Thigh -Forearm (could stop here)

term	estimate	std.error	statistic	p.value
Abdomen	0.801	0.040	20.011	0.000
Wrist	-1.587	0.471	-3.367	0.001
Height	-0.314	0.122	-2.582	0.010
Age	0.052	0.024	2.152	0.032
Neck	-0.287	0.207	-1.384	0.168
(Intercept)	4.621	8.164	0.566	0.572

backwards with bodyfat: -Knee -Weight -Ankle -Bicep(s) -Chest -Hip -Thigh -Neck (rather than forearm) (could stop here)

term	estimate	std.error	statistic	p.value
Abdomen	0.758	0.035	21.361	0.000
Wrist	-2.129	0.450	-4.735	0.000
Height	-0.326	0.121	-2.684	0.008
Age	0.065	0.025	2.595	0.010
Forearm	0.214	0.183	1.167	0.244
(Intercept)	1.786	8.134	0.220	0.826

backwards with bodyfat: -Knee -Weight -Ankle -Bicep(s) -Chest -Hip -Thigh -Forearm -Neck (could stop here)

term	estimate	std.error	statistic	p.value
Abdomen	0.771	0.034	22.932	0.000
Wrist	-1.911	0.410	-4.667	0.000
Height	-0.323	0.122	-2.657	0.008
Age	0.056	0.024	2.351	0.020
(Intercept)	2.900	8.084	0.359	0.720

backwards with bodyfat: -Knee -Weight -Ankle -Bicep(s) -Chest -Hip +Thigh -Forearm -Neck -Wrist (trying a few things)

term	estimate	std.error	statistic	p.value
Abdomen	0.693	0.052	13.412	0.000
Height	-0.554	0.117	-4.715	0.000
Age	0.028	0.029	0.960	0.338
(Intercept)	-6.286	8.357	-0.752	0.453
Thigh	-0.017	0.108	-0.157	0.876

backwards with bodyfat: -Knee -Weight -Ankle -Bicep(s) +Chest -Hip -Thigh -Forearm -Neck -Wrist (trying a few things)

term	estimate	std.error	statistic	p.value
Abdomen	0.852	0.067	12.700	0.000
Height	-0.523	0.114	-4.569	0.000
Chest	-0.228	0.083	-2.735	0.007
Age	0.027	0.024	1.115	0.266
(Intercept)	-1.069	8.291	-0.129	0.898

term	estimate	std.error	statistic	p.value
Abdomen	0.771	0.034	22.932	0.000
Wrist	-1.911	0.410	-4.667	0.000
Height	-0.323	0.122	-2.657	0.008
Age	0.056	0.024	2.351	0.020
(Intercept)	2.900	8.084	0.359	0.720

backwards with bodyfat: previous two models compared with R^2_{adj}

```
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1.06890    8.29123  -0.129  0.89753
## Age          0.02664    0.02388   1.115  0.26578
## Height      -0.52316    0.11450  -4.569 7.76e-06
## Chest       -0.22792    0.08333  -2.735  0.00669
## Abdomen      0.85199    0.06709  12.700  < 2e-16
##
## Residual standard error: 4.397 on 245 degrees of freedom
## Multiple R-squared:  0.7235, Adjusted R-squared:  0.719
## F-statistic: 160.3 on 4 and 245 DF,  p-value: < 2.2e-16
```

```
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.90033    8.08402   0.359   0.7201
```

backwards with bodyfat: perspectives

I could try seeing if anything outperforms `Wrist`, for example.

Backwards strategy is a "greedy" method (follows the best path on each short step), which isn't guaranteed to get a "best" model in the end.

The "rankings" of the variables change quite a bit.

Everything is affected by correlations among the inputs.

forwards with bodyfat

This is a little more tedious:

1. Start with the “best” one-term model.

forwards with bodyfat

This is a little more tedious:

1. Start with the “best” one-term model.
2. Look at all two-term models, and choose the best.

forwards with bodyfat

This is a little more tedious:

1. Start with the “best” one-term model.
2. Look at all two-term models, and choose the best.
3. Look at all three-term models. . .

forwards with bodyfat

This is a little more tedious:

1. Start with the “best” one-term model.
2. Look at all two-term models, and choose the best.
3. Look at all three-term models. . .

forwards with bodyfat

This is a little more tedious:

1. Start with the "best" one-term model.
2. Look at all two-term models, and choose the best.
3. Look at all three-term models...

...until you stop, because adding more terms doesn't seem to accomplish anything.

The "best" could be highest R_a^2dj, smallest new p-value, etc.

forwards with bodyfat - step 1

You can easily find the “best” first model just by finding the input most highly correlated with the output.

rowname	r
Height	-0.029
Ankle	0.245
Age	0.295
Wrist	0.339
Forearm	0.365
Bicep	0.482
Neck	0.489
Knee	0.492
Thigh	0.549
Weight	0.617
Hip	0.633
Chest	0.701
Abdomen	0.824
waist	0.824

forwards with bodyfat: +Abdomen

The two-term model "winner" (by R^2_{adj}) is Weight:

```
##   adj.r.squared
## 1     0.7205176
```

Here's for, say Height:

```
##   adj.r.squared
## 1     0.7108945
```

perspectives on forwards

Forwards strategy is also a "greedy" method (follows the best path on each short step), which isn't guaranteed to get a "best" model in the end.

We can immediately see it will result in a different model from the backwards strategy.

The "rankings" of the variables change quite a bit.

Everything is affected by correlations among the inputs.

It is actually the greedy method I tend to use most often.