Statistical Reasoning with Ellipses: The Beta Ellipse

Simple regression, the regression paradox, the anatomy of outliers: fit vs leverage, significance at a glance

Georges Monette

random@yorku.ca

Multiple Regression vs Simple Regression

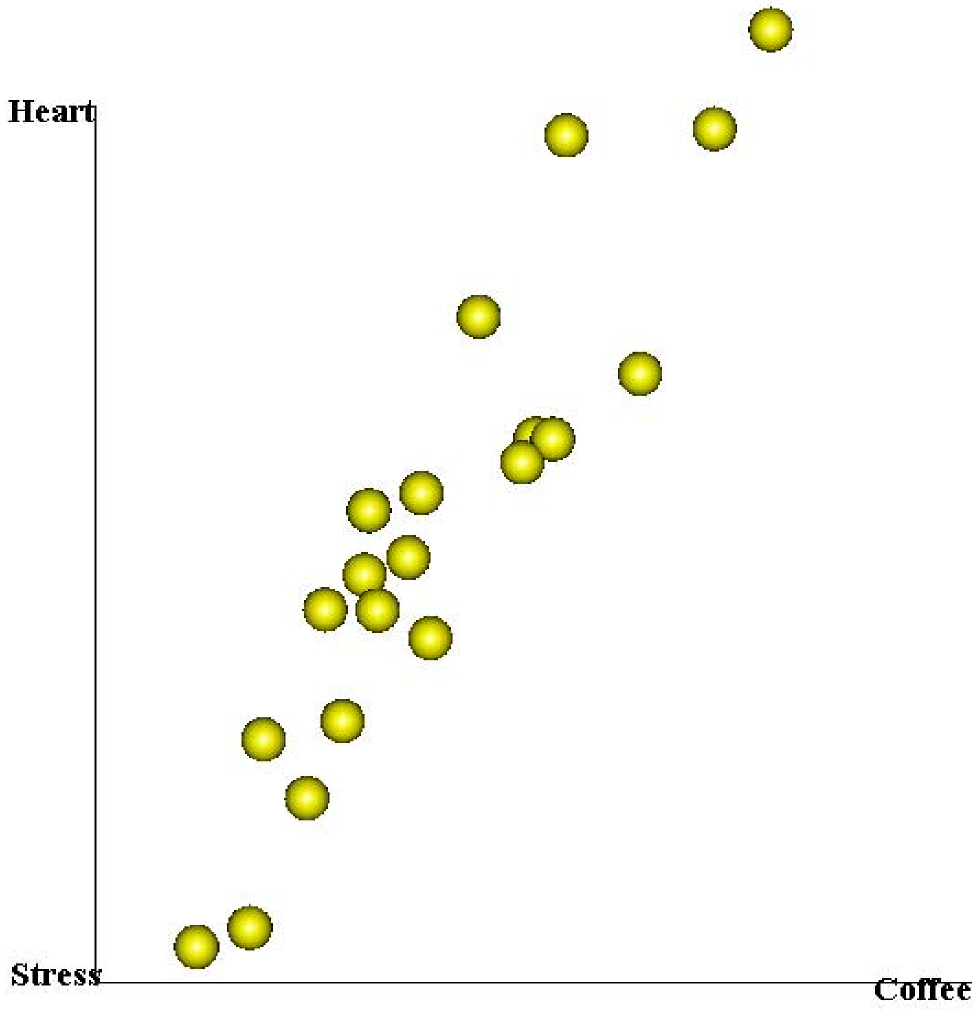
What is the difference between regression on one variable and on two variables?
How can we interpret regression coefficient?

Example:

Small artificial example showing the relationship between Coffee Consumption and Heart Damage

Artificial data: Heart Damage vs Coffee Consumption

Is coffee bad for you?



Model:

$$\text{Heart} = \gamma_0 + \gamma_{Coffee} \times \text{Coffee} + \varepsilon$$

$$Y = \gamma_0 + \gamma_1 X_1 + \varepsilon$$

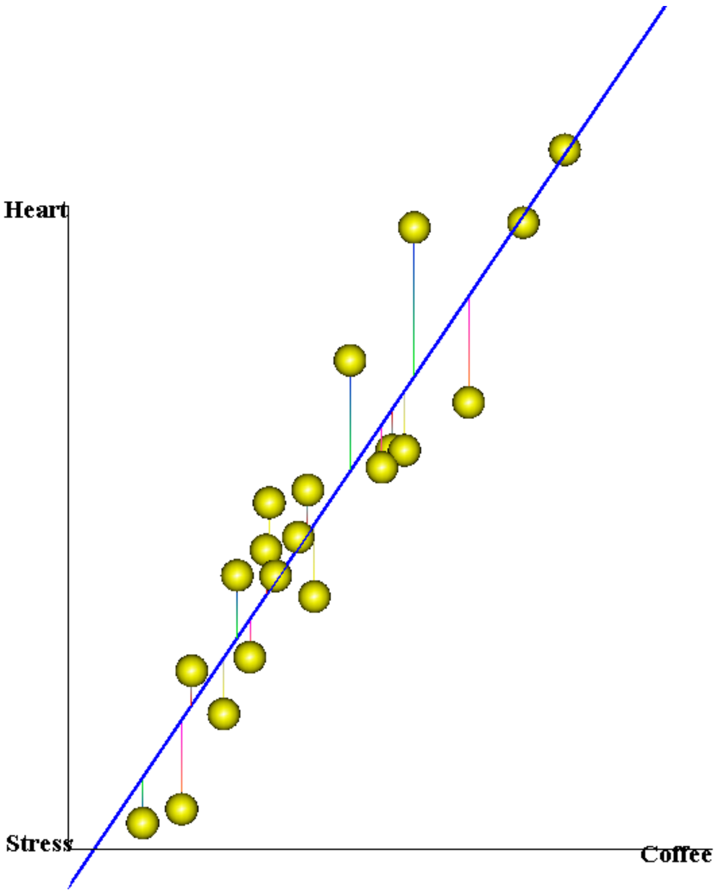
Fit:

$$\widehat{\text{Heart}} = \hat{\gamma}_0 + \hat{\gamma}_{Coffee} \times \text{Coffee}$$

$$\text{Heart} = \widehat{\text{Heart}} + e = \hat{\gamma}_0 + \hat{\gamma}_{Coffee} \times \text{Coffee} + e$$

$$Y = \hat{Y} + e = \hat{\gamma}_0 + \hat{\gamma}_1 X_1 + e$$

[Note: $\gamma =$ 'gamma']



Fitting a simple regression in R:

```
> fit.simple <-lm( Heart ~ Coffee, dd)
> summary(fit.simple)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -9.3138     9.2055  -1.012    0.325
Coffee        1.1082     0.1072  10.339 5.34e-09 ***
---
Residual standard error: 16.48 on 18 degrees of freedom
Multiple R-squared: 0.8559,     Adjusted R-squared: 0.8479
F-statistic: 106.9 on 1 and 18 DF,  p-value: 5.337e-09
```

$$\widehat{\text{Heart}} = -9.3138 + {\color{red}1.1082} \times \text{Coffee}$$

$$\text{Heart} = \widehat{\text{Heart}} + e$$

$$SD(e) = 16.48$$

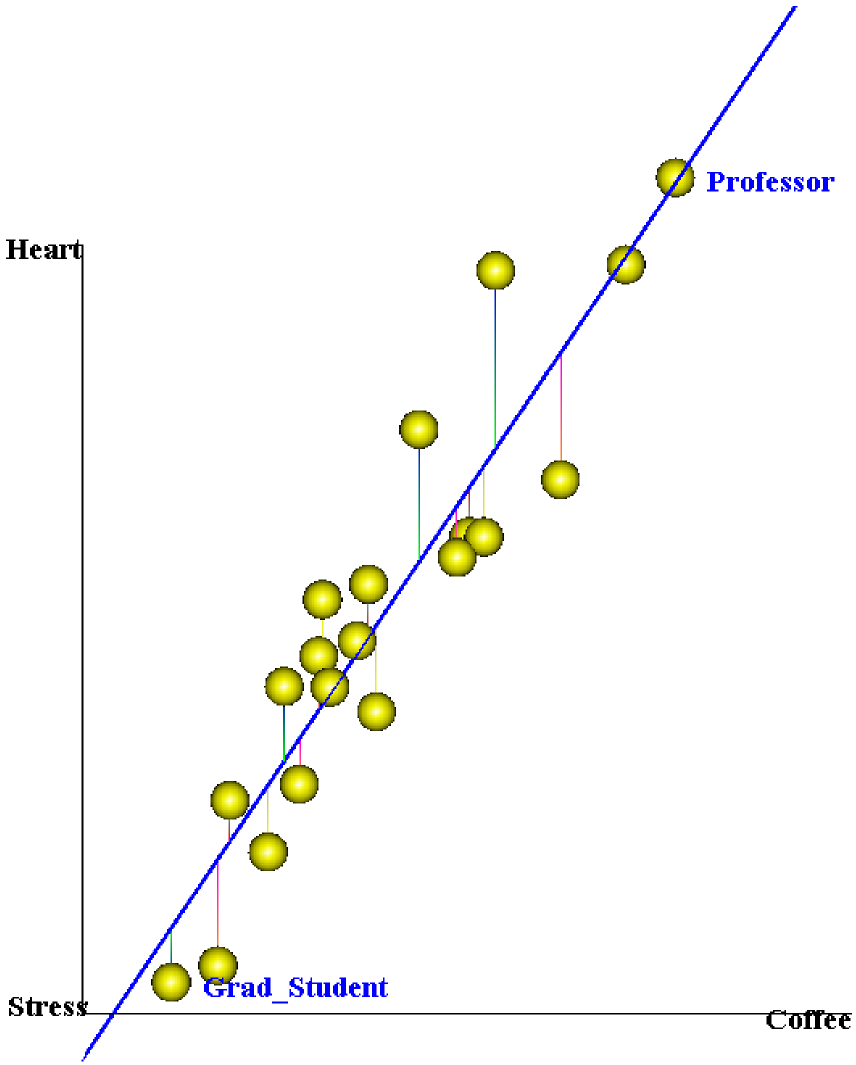
Coffee is terrible!

Is coffee causing Heart Damage?

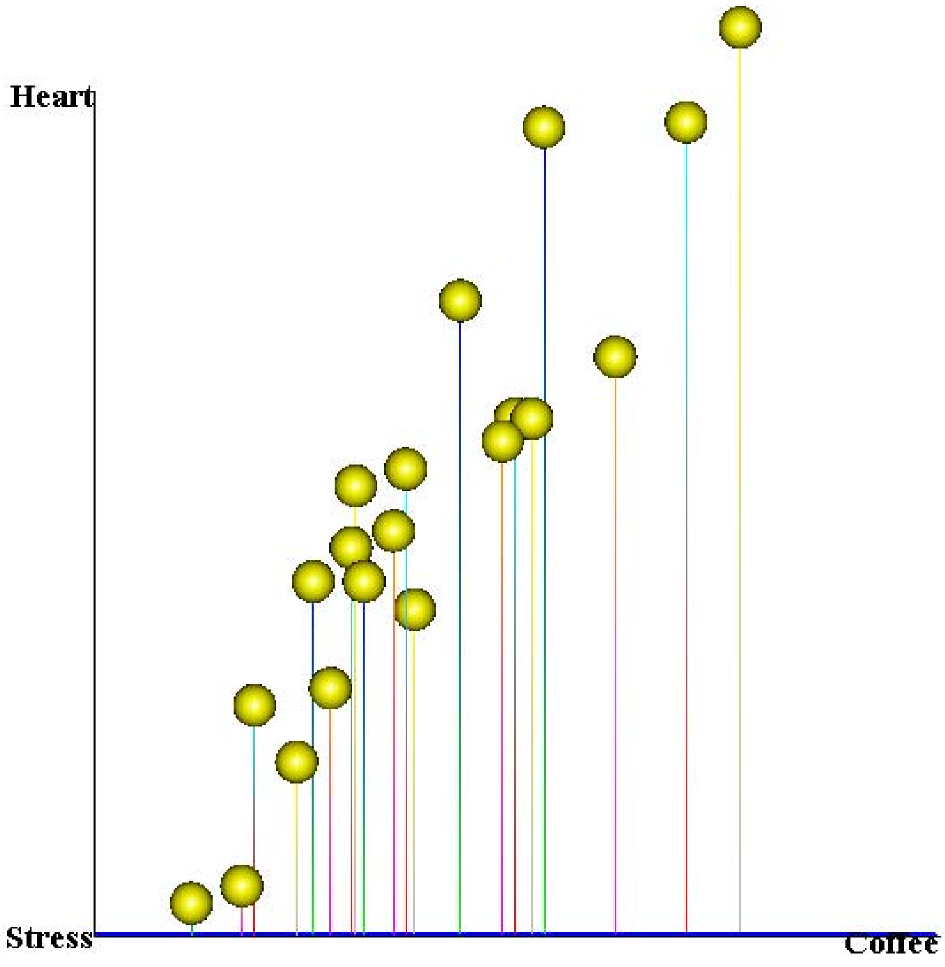
Every unit increase in Coffee Consumption is associated with a 1.1082 increase in Heart Damage.

Occupations suggest the possibility that something else may be responsible!

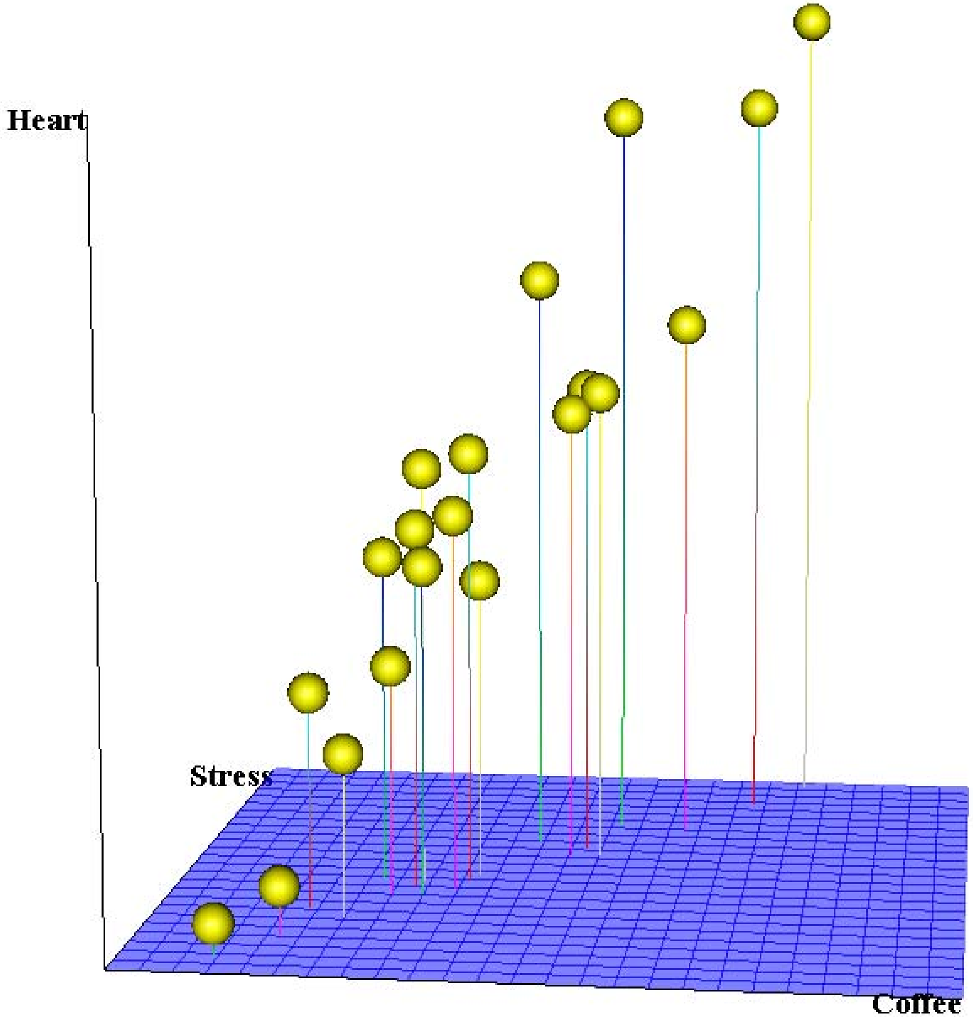
It could be anything but
we've measured Stress and we can see what happens when we include Stress in our analysis.



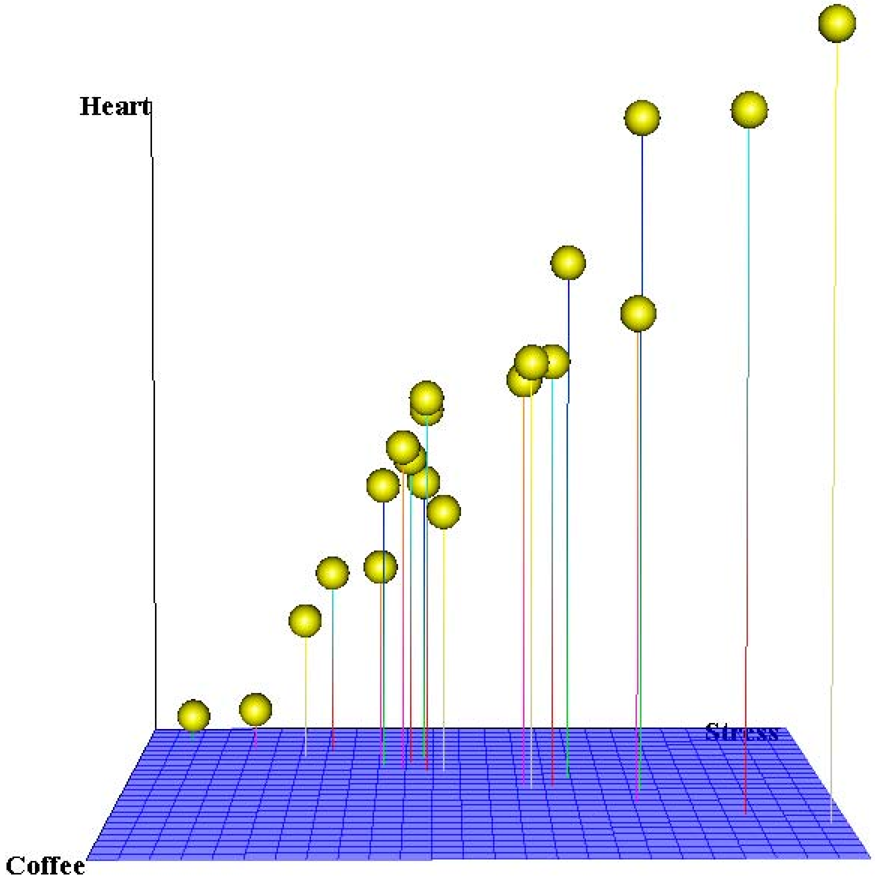
Think of data as balloons floating in space tied down by strings



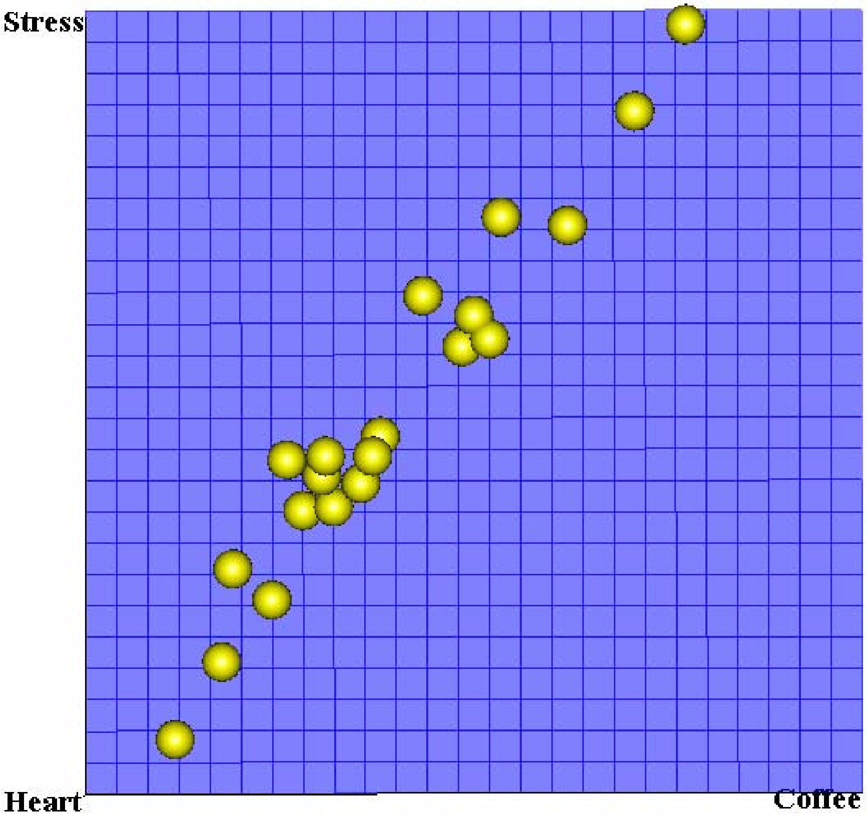
The data show a strong association between any pair of the 3 variables: Heart, Coffee and Stress



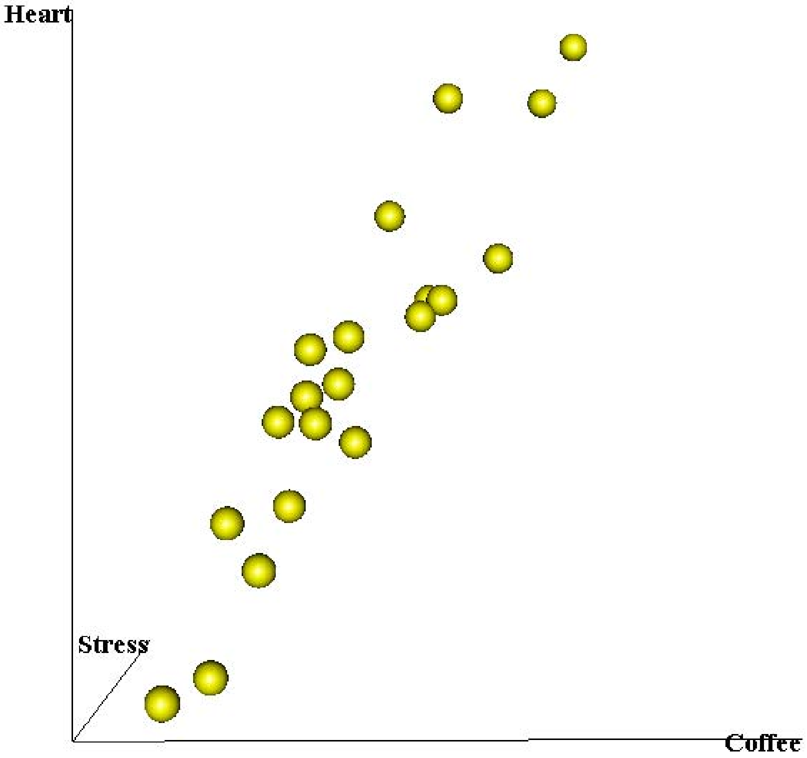
Rotating the data shows the strong association between Heart and Stress



Viewing the data from above shows
a strong association between
Stress and Coffee



How can we consider the relationship between Heart and BOTH Coffee and Stress.



One way: a linear multiple regression model:

Model:

$$\text{Heart} = \beta_0 + \beta_{Coffee} \times \text{Coffee} + \beta_{Stress} \times \text{Stress} + \varepsilon$$

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon \qquad [\beta = \text{'beta'}]$$

Fit:

$$\widehat{\text{Heart}} = \hat{\beta}_0 + \hat{\beta}_{Coffee} \times \text{Coffee} + \hat{\beta}_{Stress} \times \text{Stress}$$

$$\text{Heart} = \widehat{\text{Heart}} + e$$

$$Y = \hat{Y} + e = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2 + e$$

Fitting a multiple linear regression in R:

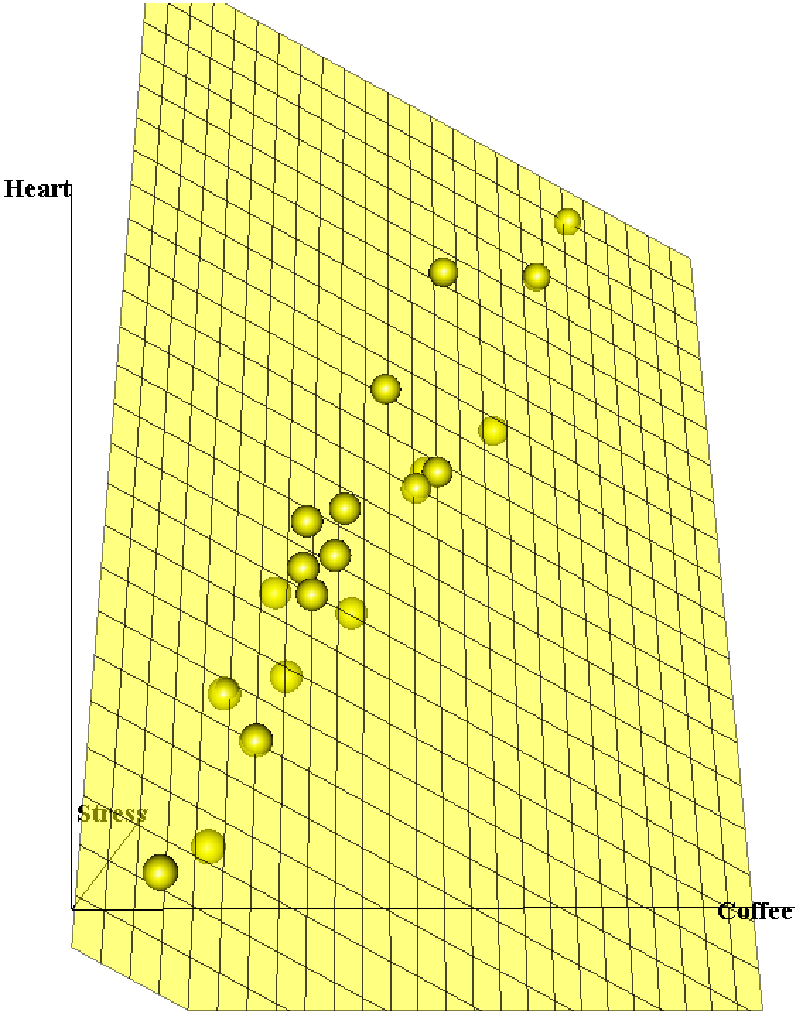
```
> fit.mult <- lm( Heart ~
     Coffee + Stress, dd)
> summary( fit.mult )

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -7.7943     5.7927  -1.346    0.196
Coffee       -0.4091     0.2918  -1.402    0.179
Stress        1.1993     0.2244   5.345 5.36e-05 ***
```

Fit:

$$\widehat{\text{Heart}} = -7.7943 - 0.4091 \times \text{Coffee} + 1.1993 \times \text{Stress}$$

$$\text{Heart} = \widehat{\text{Heart}} + e$$

$$SD(e) = 10.36$$



Coffee is good for you!?!

But using the same data, our previous analysis suggested coffee was bad for you!

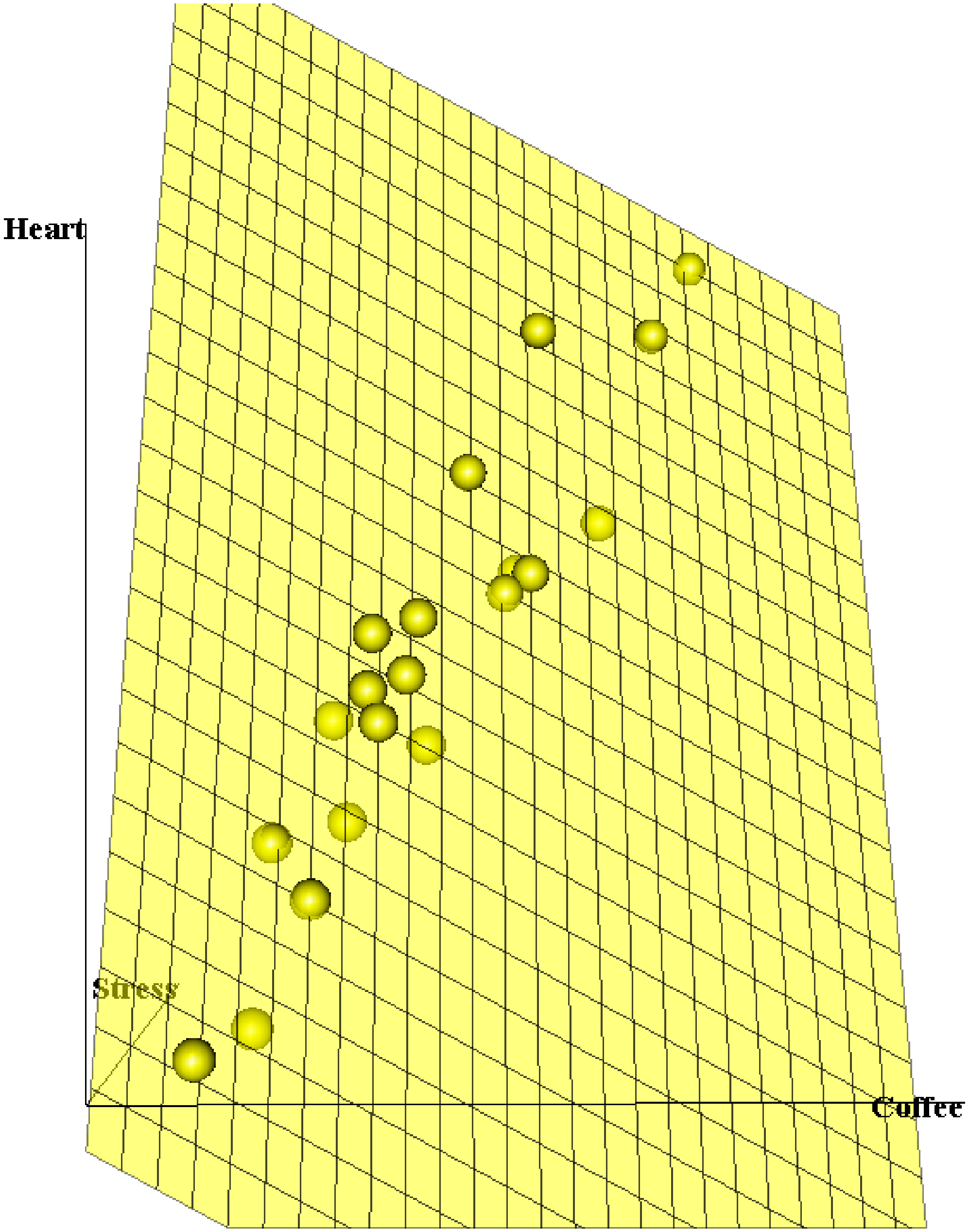
Statistics !#@&!!

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -7.7943     5.7927  -1.346    0.196
Coffee       -0.4091     0.2918  -1.402    0.179
Stress        1.1993     0.2244   5.345 5.36e-05 ***
Residual standard error: 10.36 on 17 degrees of freedom
Multiple R-squared: 0.9462,     Adjusted R-squared: 0.9399
F-statistic: 149.6 on 2 and 17 DF,  p-value: 1.620e-11
```

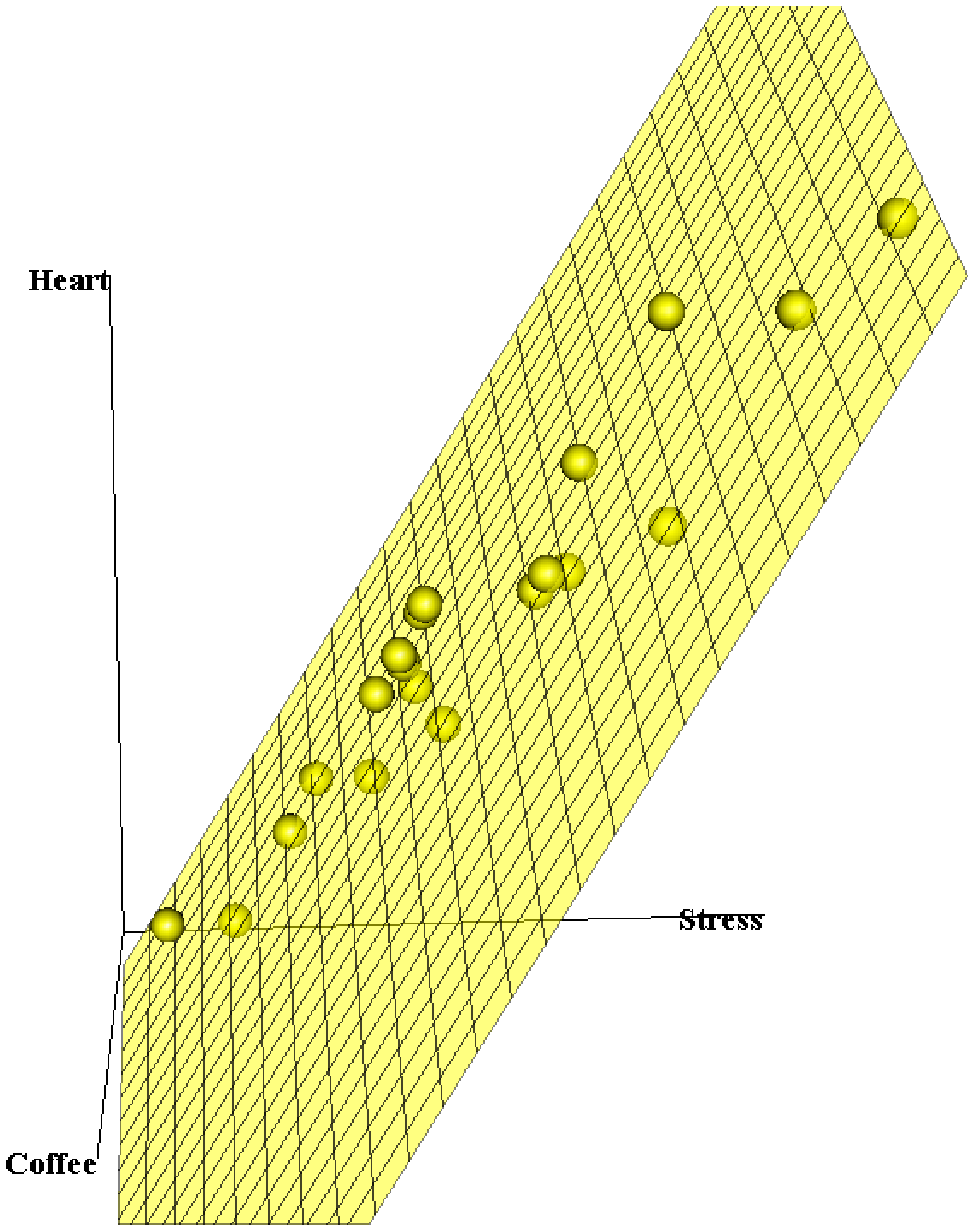
The negative slope with multiple regression measures something different than the positive slope with simple regression.

The slope of **`-0.4091`**
with respect to Coffee tells you the expected difference in Heart Damage when Coffee Consumption is one unit larger keeping the value of **Stress the same.**

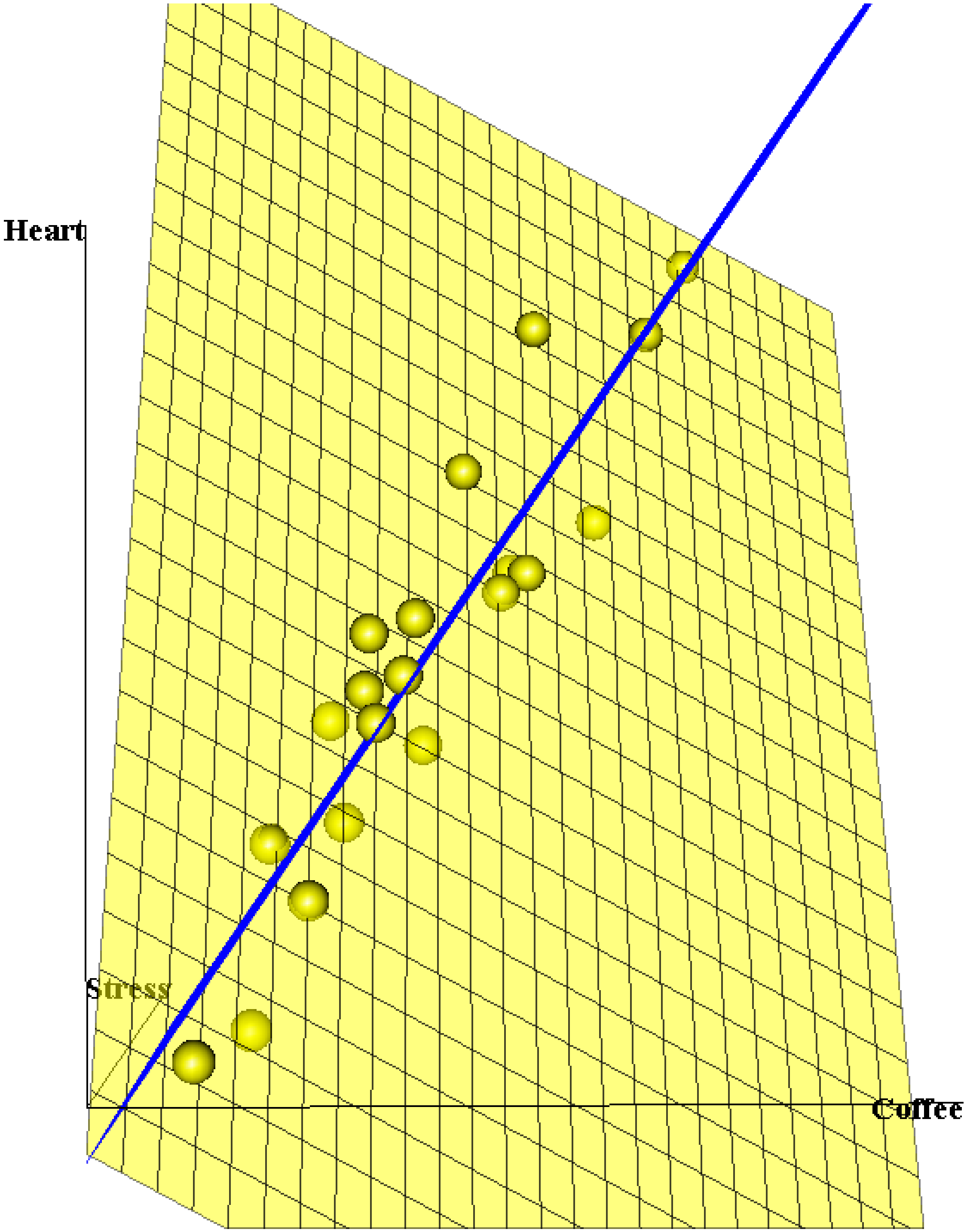
The slope of **`1.1082`** with simple regression tells you the expected difference in Heart Damage when Coffee Consumption is one unit larger but **Stress is allowed to also increase** following the same pattern seen in the data set.



We can take the data for a spin to better see what's happening.



What does the simple regression look like in 3D?



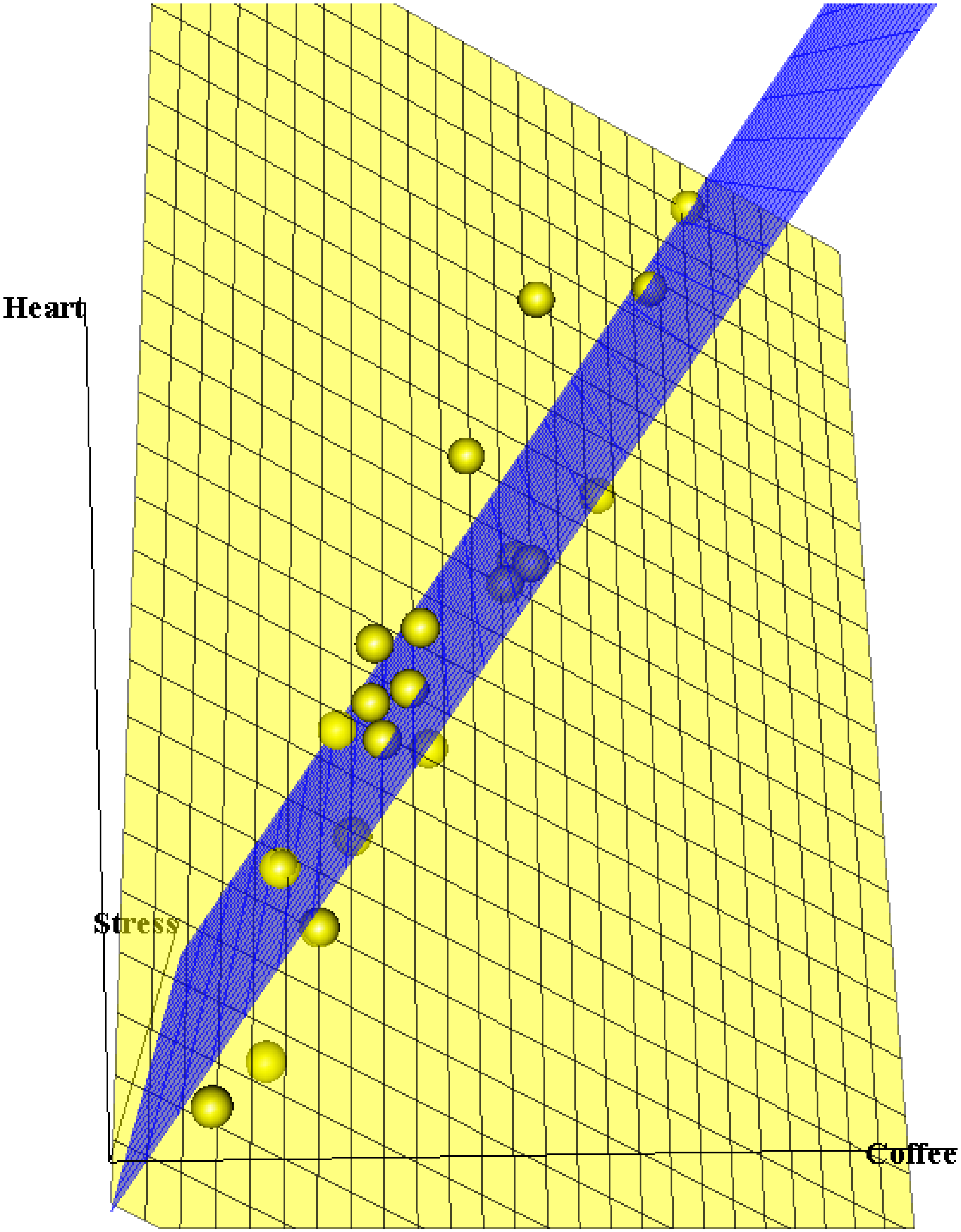
What does the simple regression look like in 3D?

Is it a bird? Is it a plane? Is it Superman?

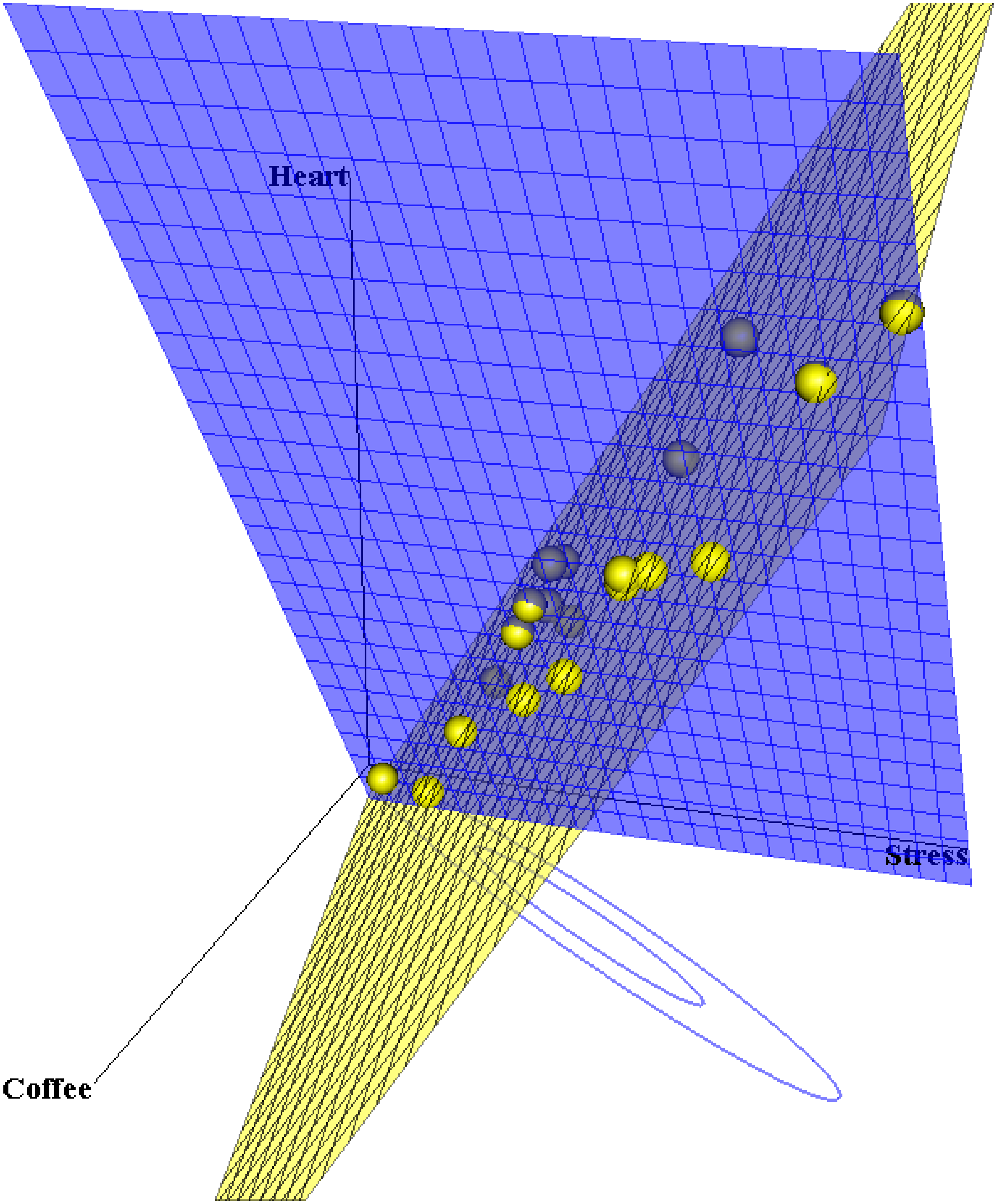
It's just a plane!

It is the best fitting plane (least sum of squared residuals) among all planes that are constrained to have a slope of 0 in the direction of Stress – i.e. it only takes Coffee into account.

The yellow plane is the best fitting plane with no restrictions.



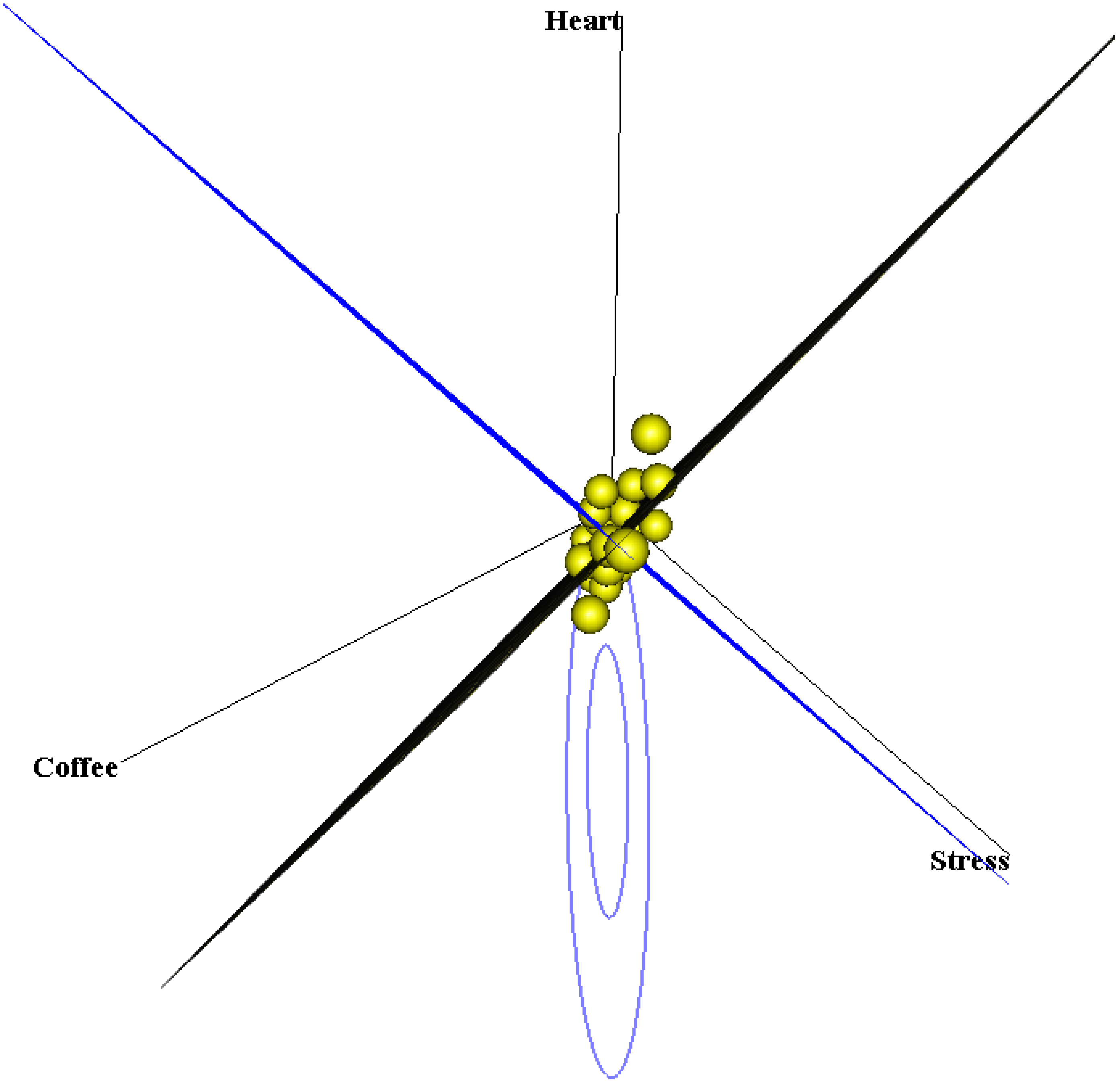
The data ellipses for Coffee and Stress show the strong relationship between these two variables.



Both the blue plane
and the yellow plane
fit the data quite well

but

the yellow plane picks
up a downward tilt in
the data that the blue
planc can't catch because
it is forced to remain
parallel to the Stress
axis.

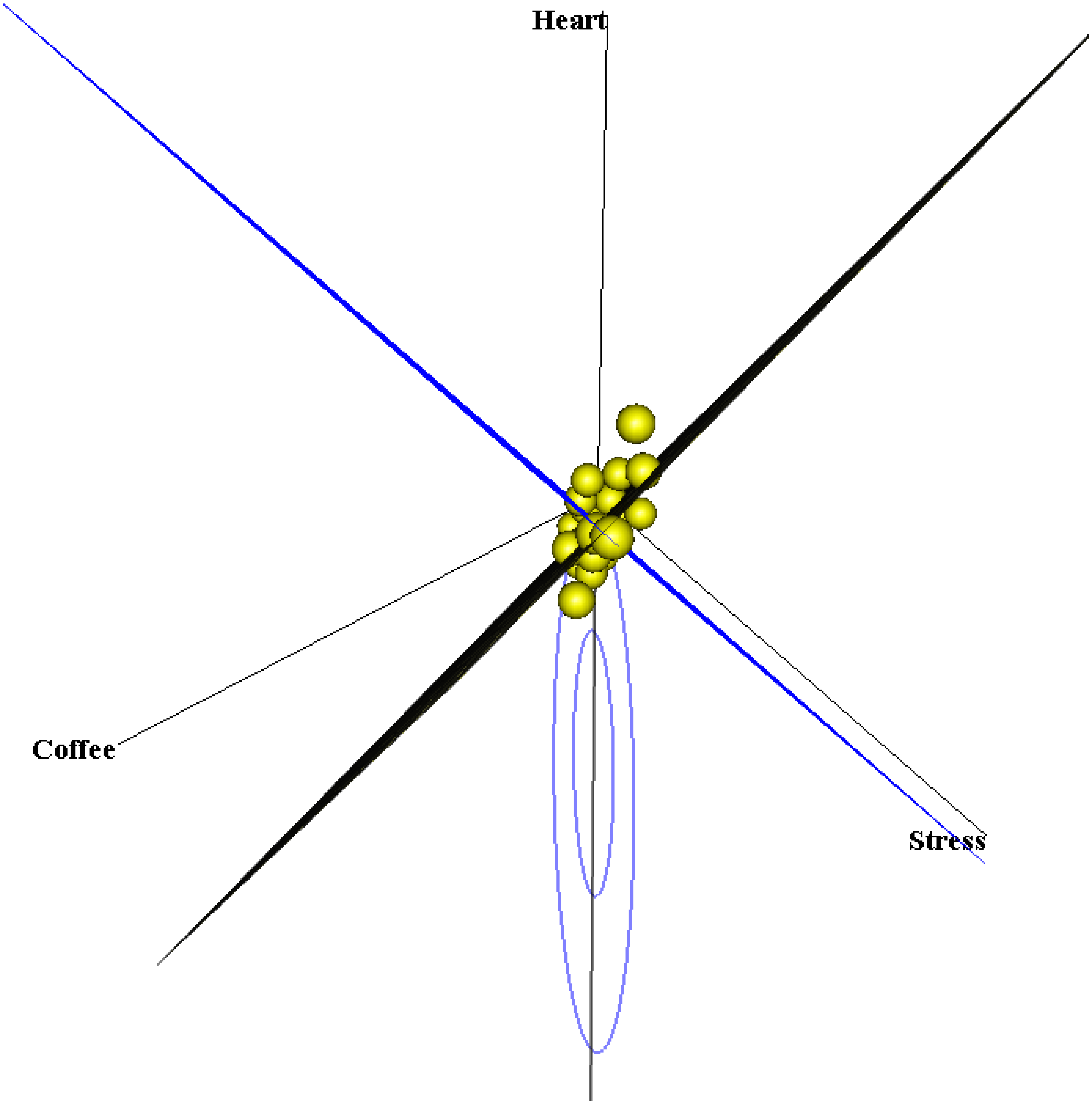


A few points of interest:

If you draw a line for the regression of Stress on Coffee, the line where the two planes intersect will lie vertically above.

The intersection line is the 'multivariate regression' line for the regression of both Heart and Stress on Coffee.

With a shear transformation to make the blue line horizontal, you are now looking at the Added Variable Plot (aka Partial Regression Leverage Plot) for the adding Stress to the model.

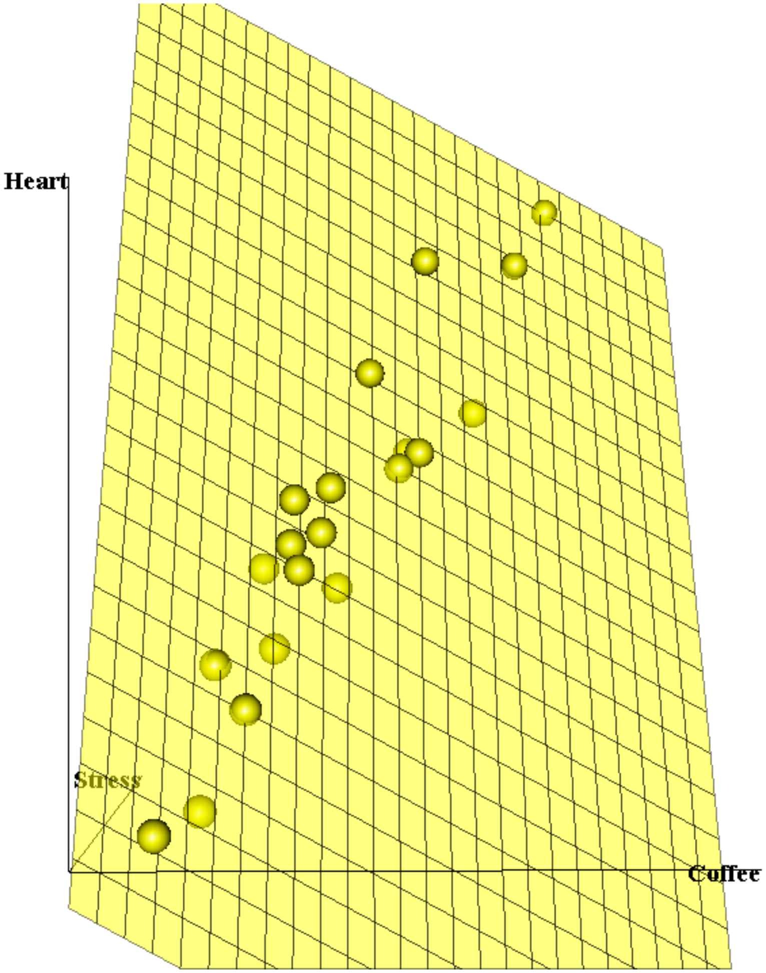


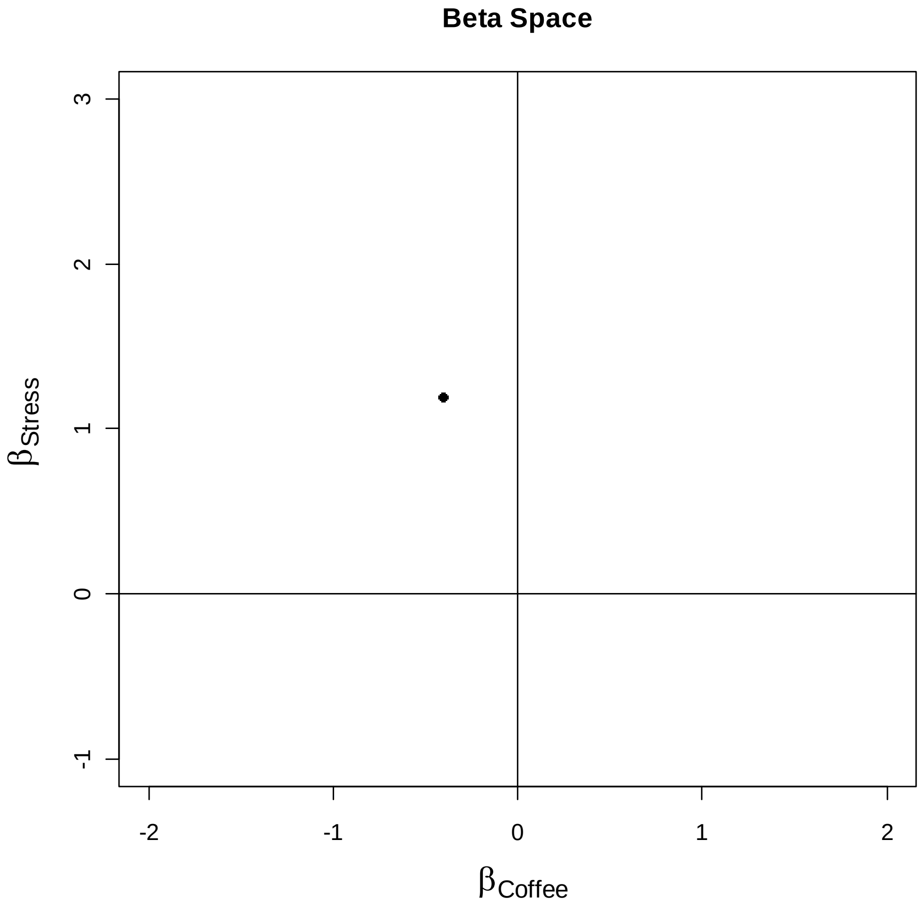
Looking at Models in 'beta-space'

So far we have looked at our data and models in 'data space'. The axes are variables and the points are observations. This is the natural space to look at data.

To understand regression more deeply – which will be particularly useful when we get to hierarchical data – we want to see models in a apace that is more natural for models: 'beta space'. In beta apace, the axes are coefficients, e.g. β_{Coffee} and β_{Stress} and the points are models (true models or fitted models) represented by their coefficients. We can also see confidence regions and confidence intervals in beta space because these are merely sets of models. The simple geometry of beta space elucidates some mysteries of data space.

The multiple regression model represented by a plane in data space is represented by a point showing the fitted slope with respect to Coffee and the fitted slope with respect to Stress in beta space

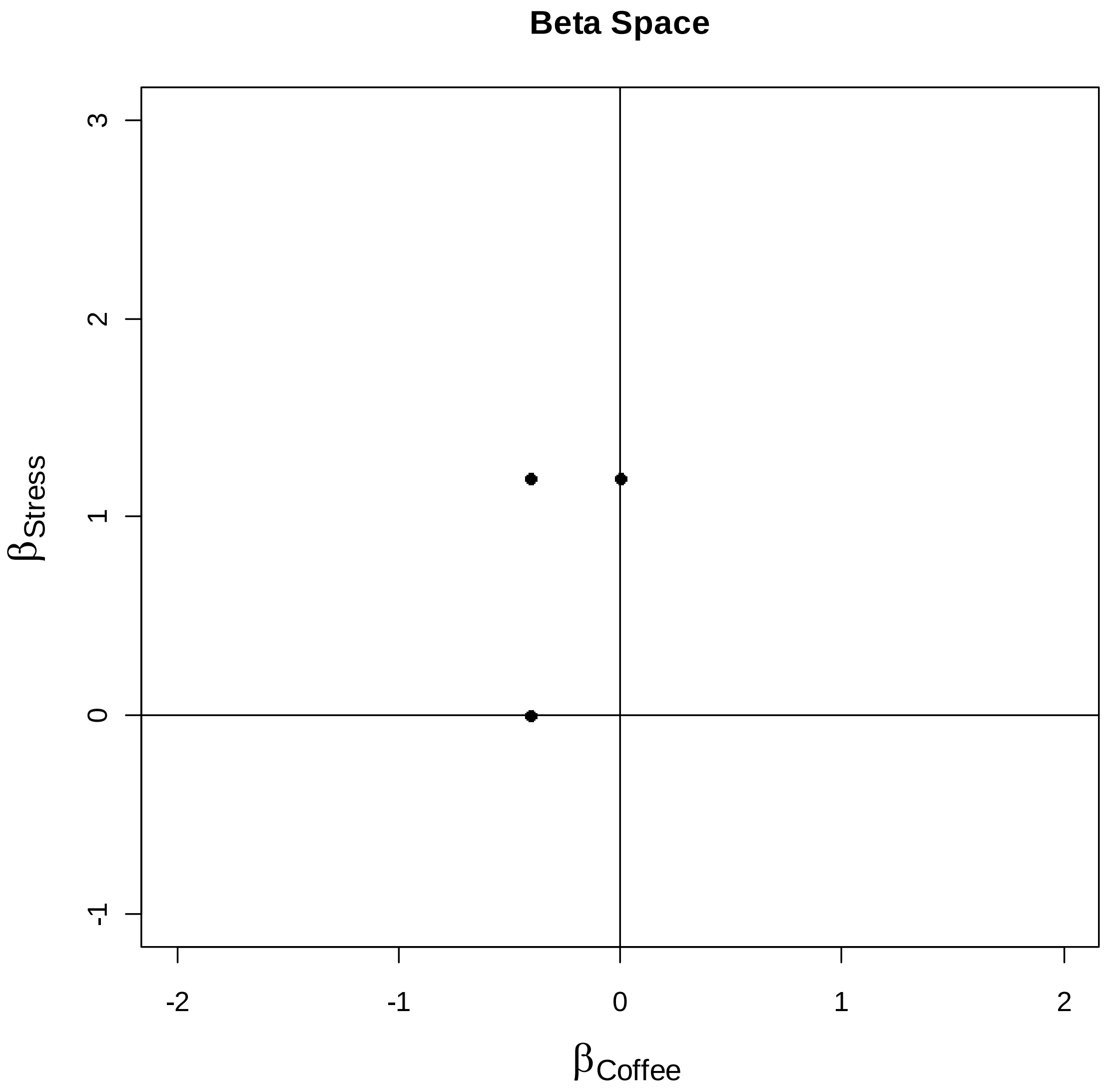




```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -7.7943     5.7927  -1.346    0.196
Coffee       -0.4091     0.2918  -1.402    0.179
Stress        1.1993     0.2244   5.345 5.36e-05 ***
```

Note that to fully represent the model we would need a 3-dimensional beta space to include an axis for the intercept. For our purposes, we only need 2 dimensions, one for each predictor. Note that more complex models, e.g. interaction models, quadratic models, represented by curved surfaces in data space, require more dimensions for beta space.

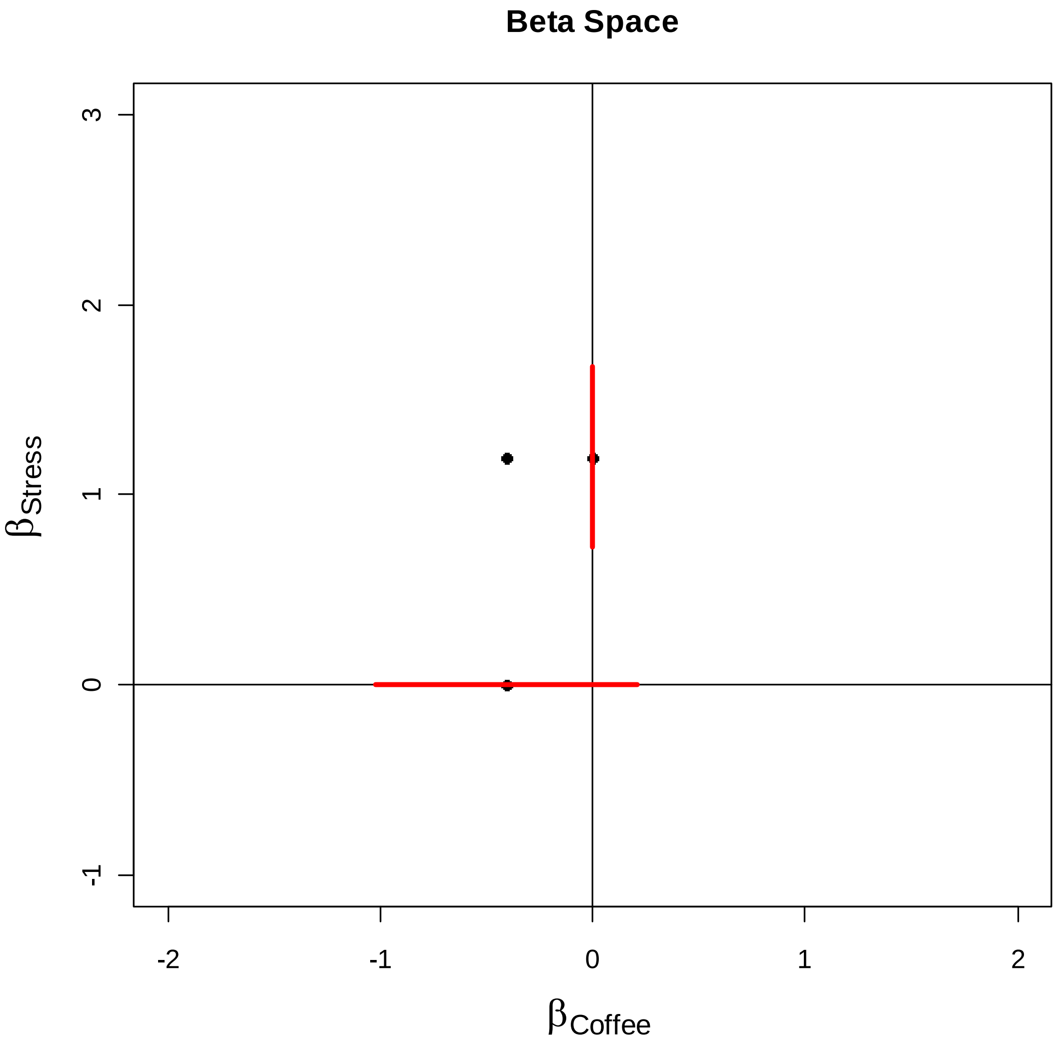
A point represents the two slopes. Projecting the point onto the horizontal and vertical axes gives the estimated slopes for Coffee and Stress respectively.



Confidence intervals for each coefficient (paramenter) can be drawn on the respective axes.

The red lines are 95% confidence intervals. The interval for coffee slope includes the origin (0). This means that we would accept (fail to reject is safer but longer) the null hypothesis that Coffee is not related to Heart Damage when controlling for Stress.

On the other hand, the interval for the slope with respect to Stress excludes the origin and we would reject the null hypothesis that the true slope is 0 when controlling for Coffee Consumption.



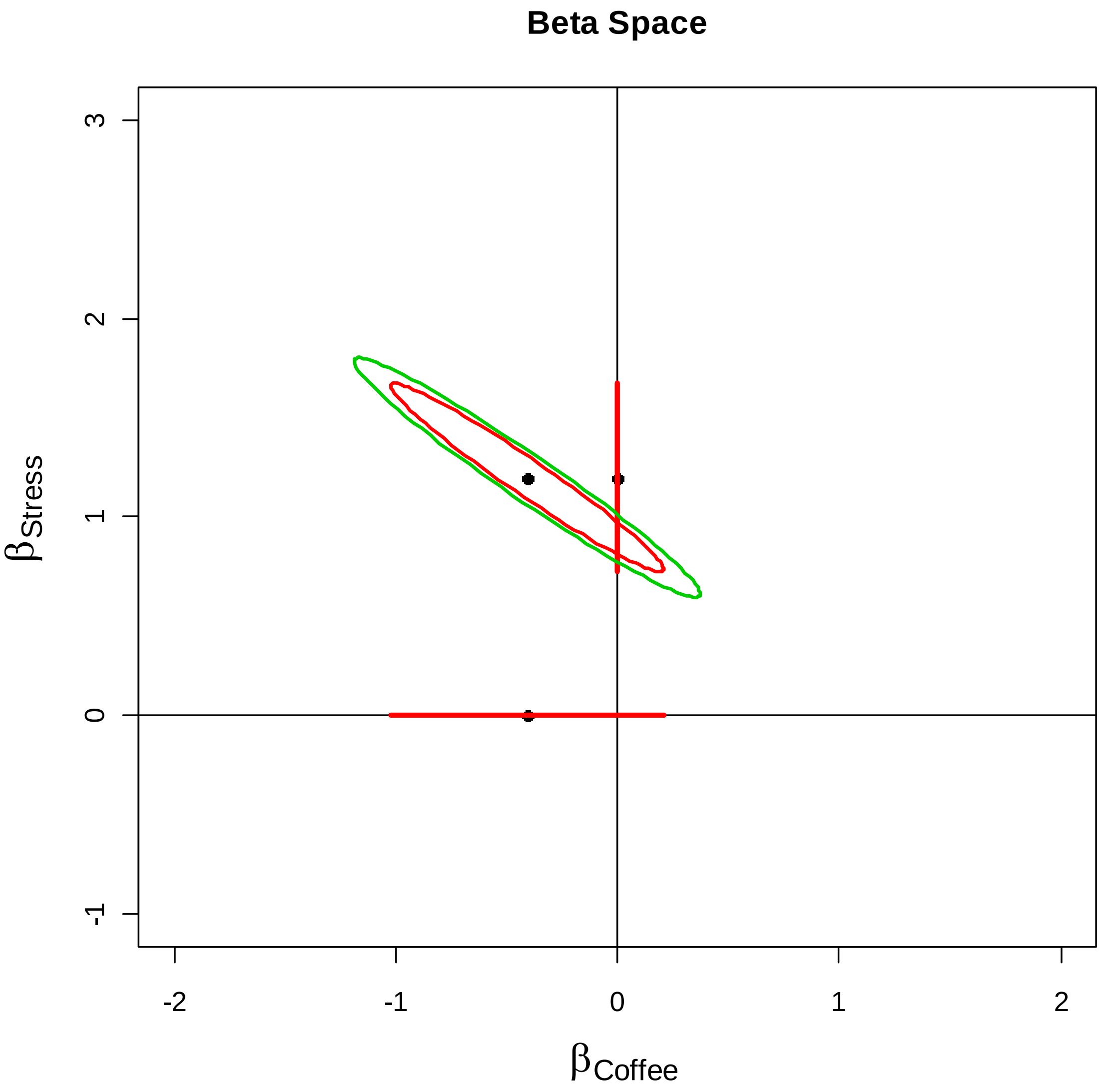
```
> coef( fit.mult)
(Intercept)      Coffee      Stress 
 -7.7942856  -0.4090511   1.1992534 
```

```
> confint(fit.mult)
                  2.5 %    97.5 %
(Intercept) -20.0157291 4.4271578
Coffee       -1.0245942 0.2064921
Stress        0.7258632 1.6726436
```

When we look at computer output we only get to see estimated coefficients and possibly confidence intervals. They only tell part of the story. The estimated coefficients may be far from independent, e.g. you might not be certain of a coefficient but you might be able to know that if it is at the higher end of its confidence interval then the other coefficient must be at the lower end

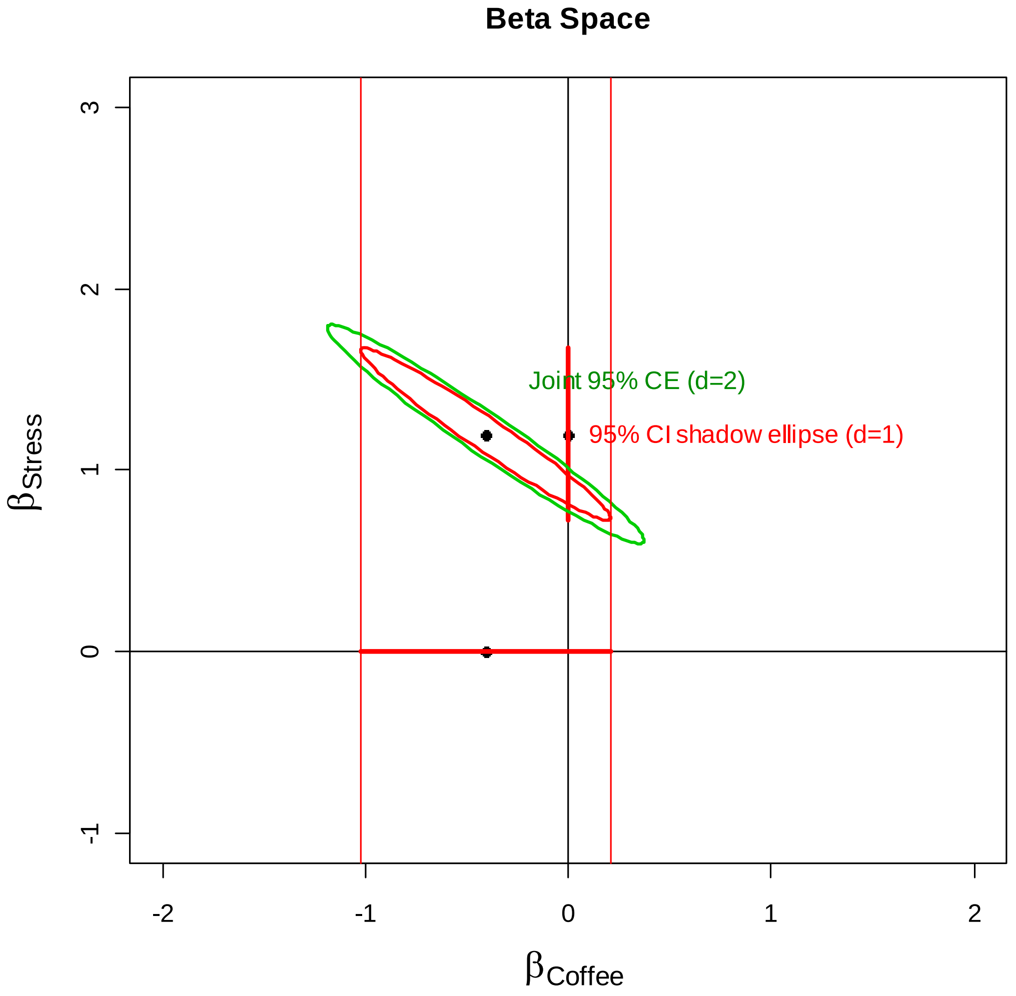
A confidence region shows a set of *combinations* of slopes for Coffee and for Stress that are plausible. We might not know much about the coefficient for Coffee but we do know that if it is high then the coefficient for Stress must be relatively low.

An appendix *will someday* show formulas for the confidence ellipses shown here.



The green ellipse has joint coverage of 95%. This means that the procedure generating the ellipse will cover the true combination of values for β_{Coffee} and β_{Stress} in 95% of samples (assuming that errors are independent with the same normal distribution with mean 0 and that the true model is linear).

The red ellipse has the same shape but is a slightly shrunken version of the green ellipse. It is scaled so that its shadows have 95% coverage as confidence intervals. This means that each interval has 95% coverage. Jointly, they will have less than 95% coverage.



Incidentally, the shadows of the green ellipse are Scheffe joint 95% confidence intervals with protection for a 2-dimensional posterior hypothesis.
A different scaling of the ellipse would produce shadows that are Bonferroni 95% confidence intervals.

In summary, confidence intervals are generated from shadows of ellipses (with traditional normal models) and ellipses generally convey more information than a set of confidence intervals.

What about simple regression? How is it related to multiple regression in beta space?

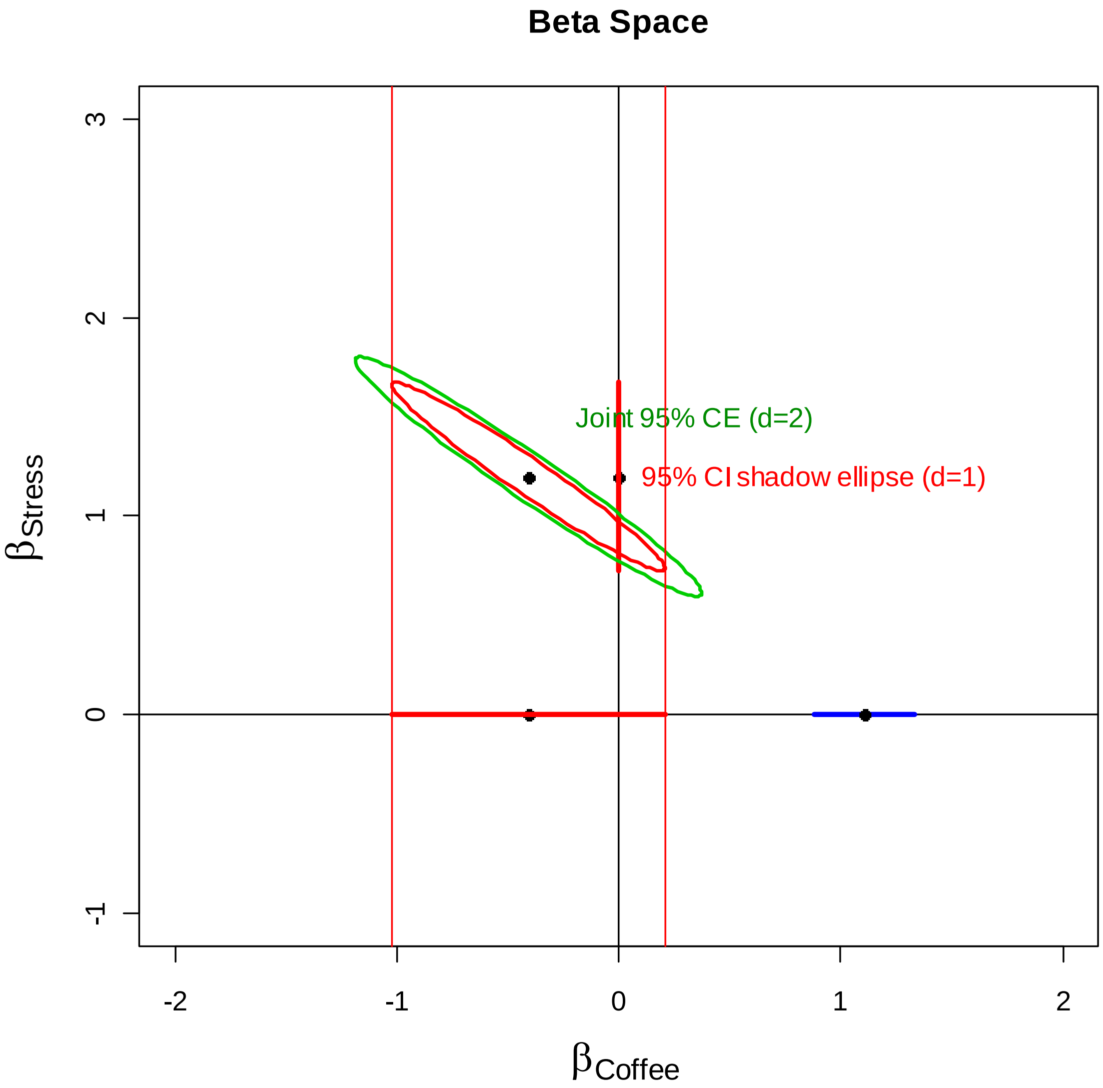
THe blue interval is 95% confidence interval for the slope with respect to Coffee in the simple regression model.

Note that it does not cover 0, thus the null hypothesis is rejected at the 5% level of significance.

What can account for the red interval and the blue interval being so far apart?

The red interval is a vertical shadow of the ellipse.

It turns out that the blue interval is an oblique shadow of the ellipse.



```
> coef(fit.simple)
(Intercept)      Coffee
  -9.313831    1.108181
```

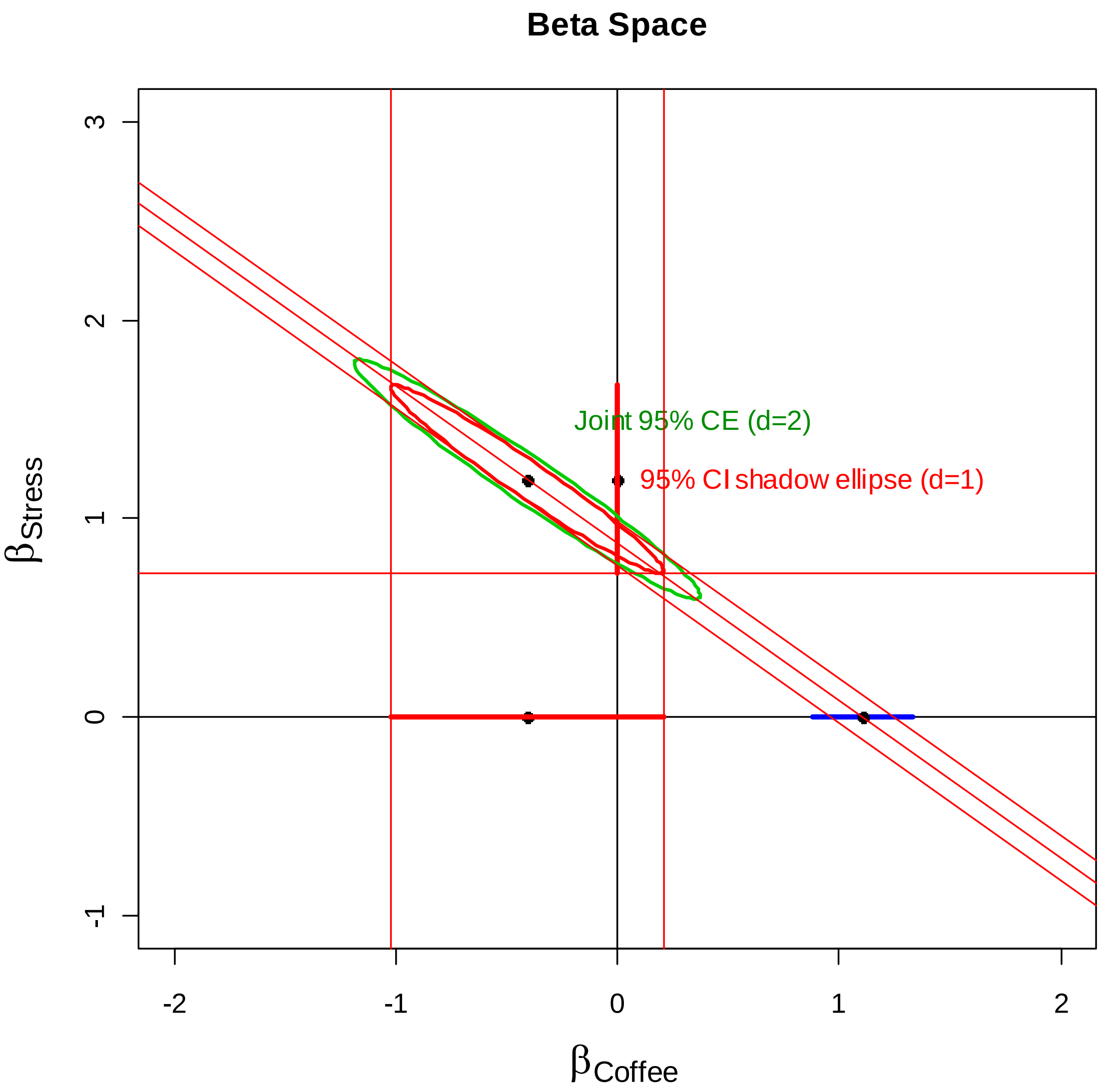
```
> confint( fit.simple )
                    2.5 %    97.5 %
(Intercept) -28.6537939 10.026132
Coffee        0.8829868  1.333375
```

The estimated slope with simple regression is the oblique projection of the centre of the ellipse through the point where the ellipse has a horizontal tangent.

The oblique shadow of the red ellipse lies within the blue interval. How much larger the interval is will be a function of the size of the coefficient for Stress.

A situation like this where the slope for the simple regression has a different sign than the slope for the multiple regression is sometimes known as Simpson's Paradox.

In this case the red interval is not significantly below zero but it is entirely possible to have situations where the two intervals are both significant with opposite signs.



When can this happen?

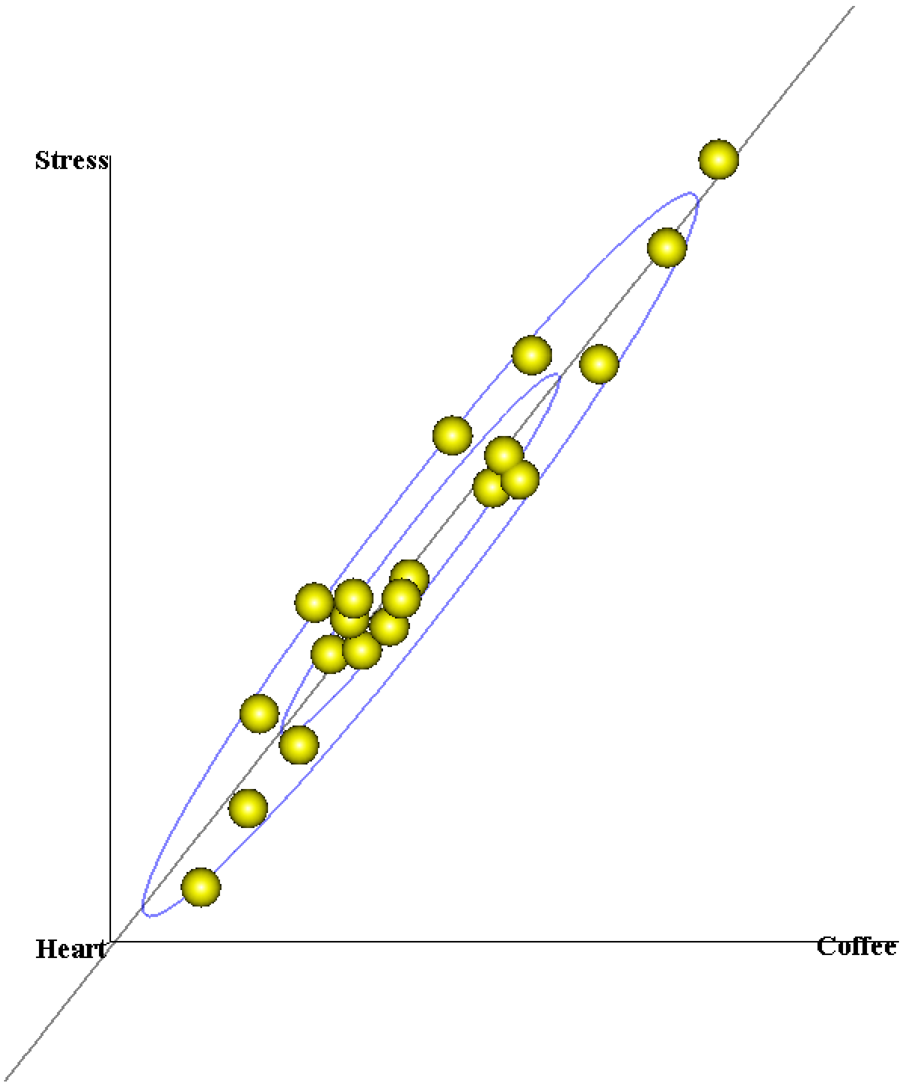
To see this we need to know the relationship between the data ellipse for the predictors, Coffee and Stress, and the confidence interval for their slopes.

Algebraically, the data ellipse has shape given by the 2 x 2 variance covariance matrix of the predictors:

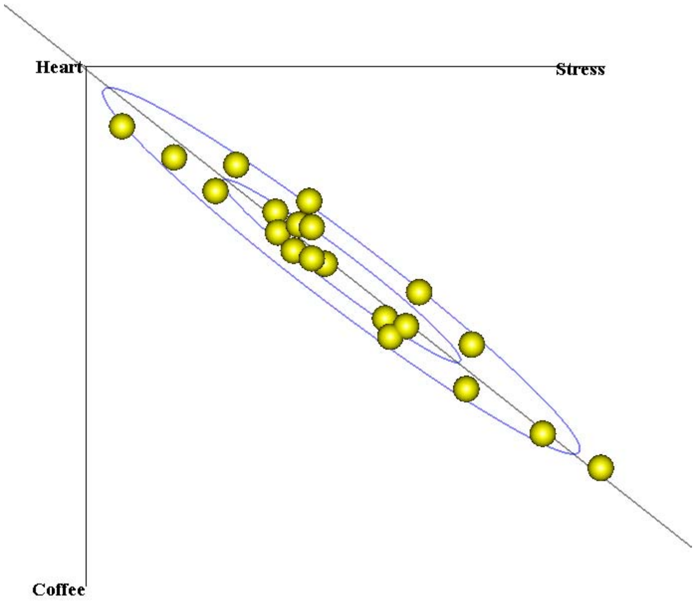
$$\Sigma = \begin{bmatrix} \mathrm{Var}(\mathit{Coffee}) & \mathrm{Cov}(\mathit{Coffee}, \mathit{Stress}) \\ \mathrm{Cov}(\mathit{Coffee}, \mathit{Stress}) & \mathrm{Var}(\mathit{Stress}) \end{bmatrix}$$

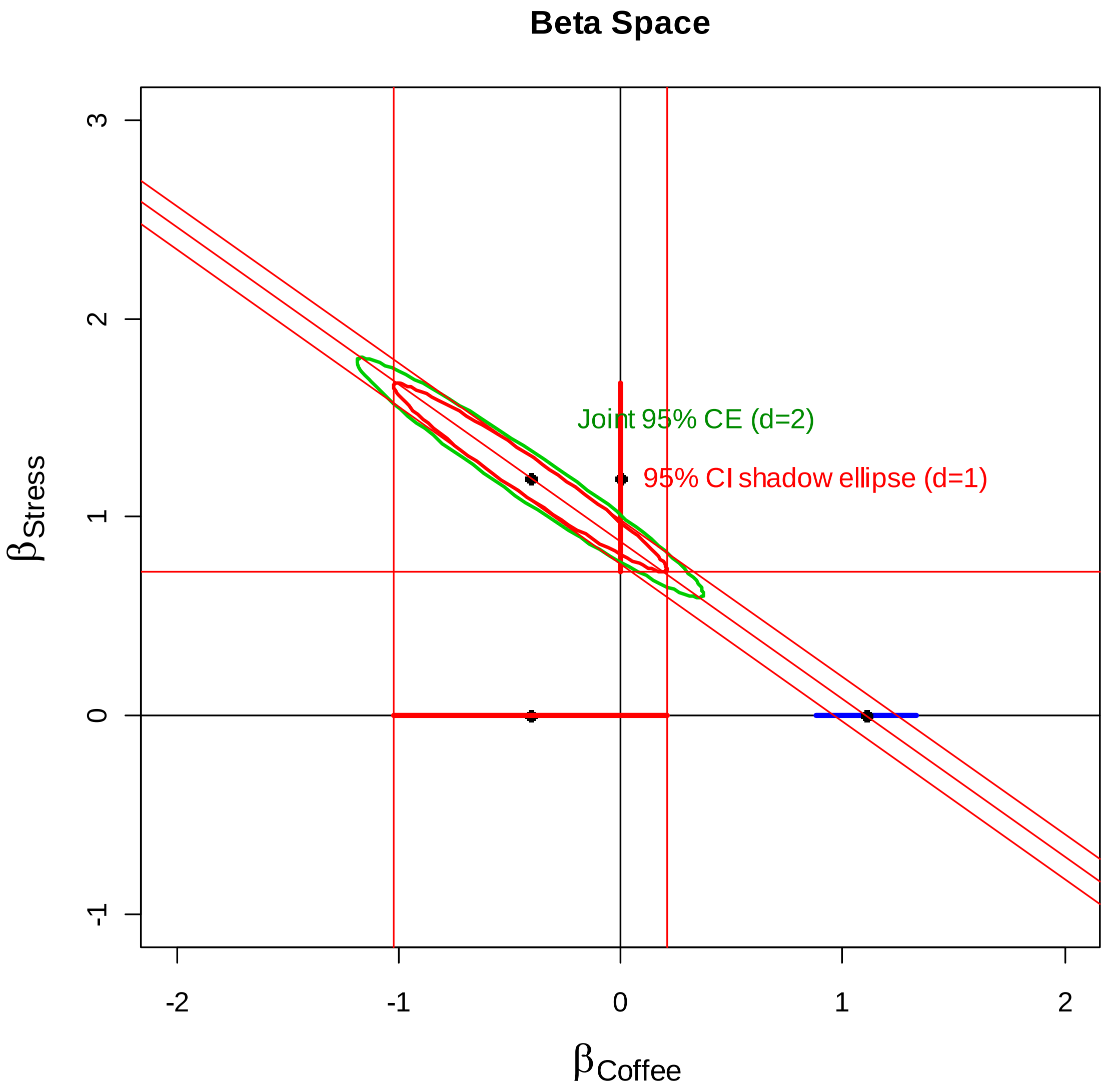
The strong relationship between Coffee and Stress is reflected by a large value for the correlation between Coffee and Stress:

$$\mathrm{Corr}(\mathit{Coffee}, \mathit{Stress}) \\ = \frac{\mathrm{Cov}(\mathit{Coffee}, \mathit{Stress})}{\mathrm{SD}(\mathit{Coffee}) \times \mathrm{SD}(\mathit{Stress})}$$



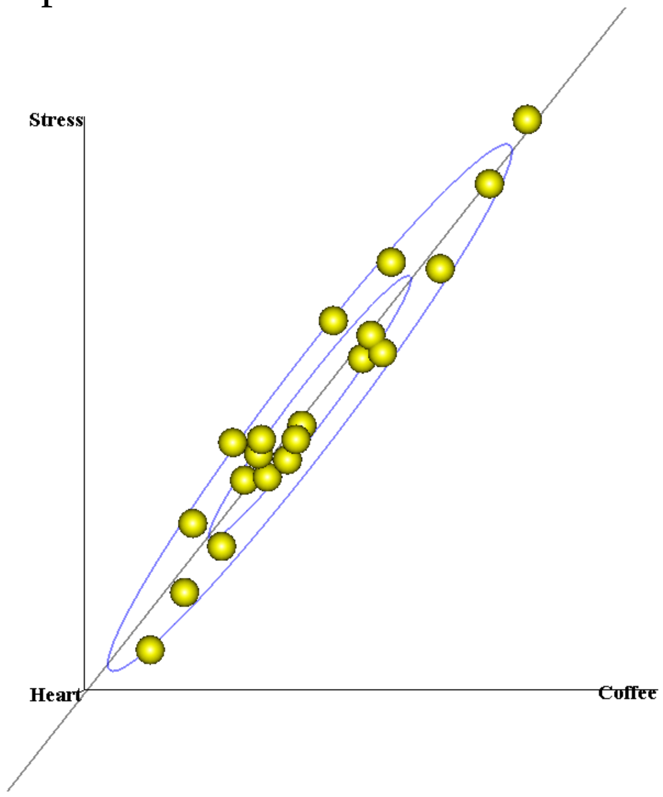
The confidence ellipse has shape Σ^{-1}. This means that the (shadow of) CI is wide in directions where the data ellipse is narrow and vice versa. I.e. we have little information about the slope for change in directions in which we have little date, and vice versa.

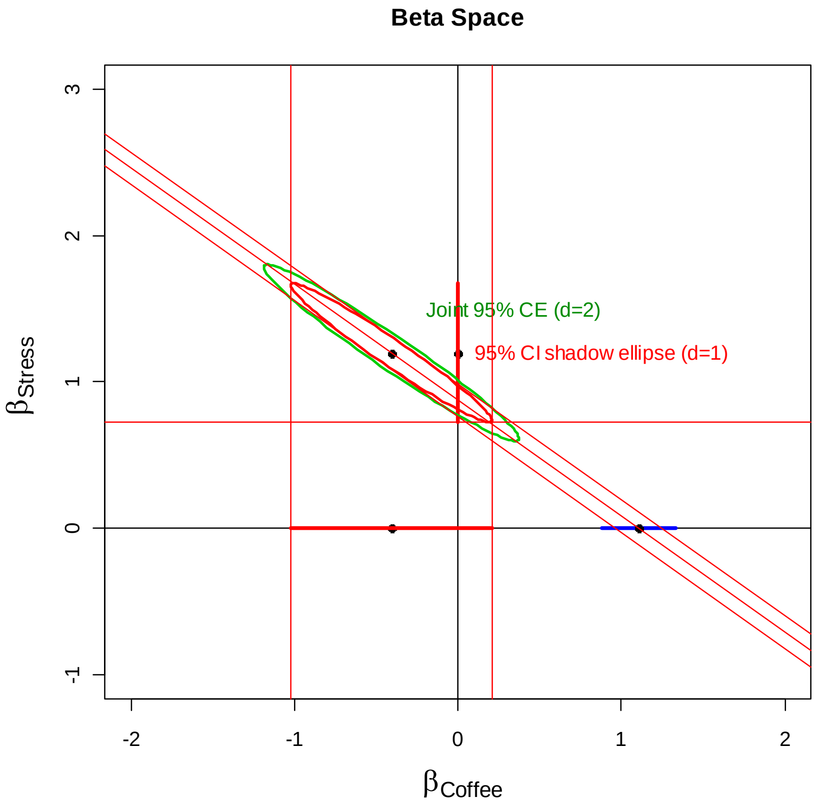




In 2 dimensions, the relationship between Σ and Σ^{-1} is very simple. The Confidence Ellipse has a shape that is the 90 degree rotation of the data ellipse. The projection line for the simple regression is the 90 degree rotation of the line for the regression of Stress on Coffee.

Implications:





1) If X_1 and X_2 are uncorrelated, then the data ellipse is not tilted and neither is the confidenc ellipse. Consequently the downward projection of the center and the oblique projection are identical and $\hat{\gamma}_1 = \hat{\beta}_1$.

2) Conversely, if $\hat{\gamma}_1 \neq \hat{\beta}_1$ then X_1 and X_2 must be correlated. In particular, Simpson's Paradox can only occur if X_1 and X_2 are correlated.

Data ellipse for predictors:

Single predictor	$\bar{x} \pm s_x$
General	$\bar{\mathbf{x}} \oplus \sqrt{\Sigma_X}$ i.e. $\bar{\mathbf{x}} + \Sigma_X^{1/2} U$ where *U* is the unit sphere

Confidence ellipses for slopes:

Let

$$\eta = L\beta$$

have length 2 (could be k) and

$$\hat{\eta} = L\hat{\beta}$$

Then a Scheffé confidence region with 95% coverage in *d* dimensions is

$$\hat{\eta} \oplus \sqrt{dF_{d,\nu}^{0.95}} \sqrt{\widehat{\mathrm{Var}}(\hat{\eta})}$$

where

$$\widehat{\operatorname{Var}}\left(\hat{\eta}\right) = s_e^2 L\left(X'X\right)^{-1} L'$$

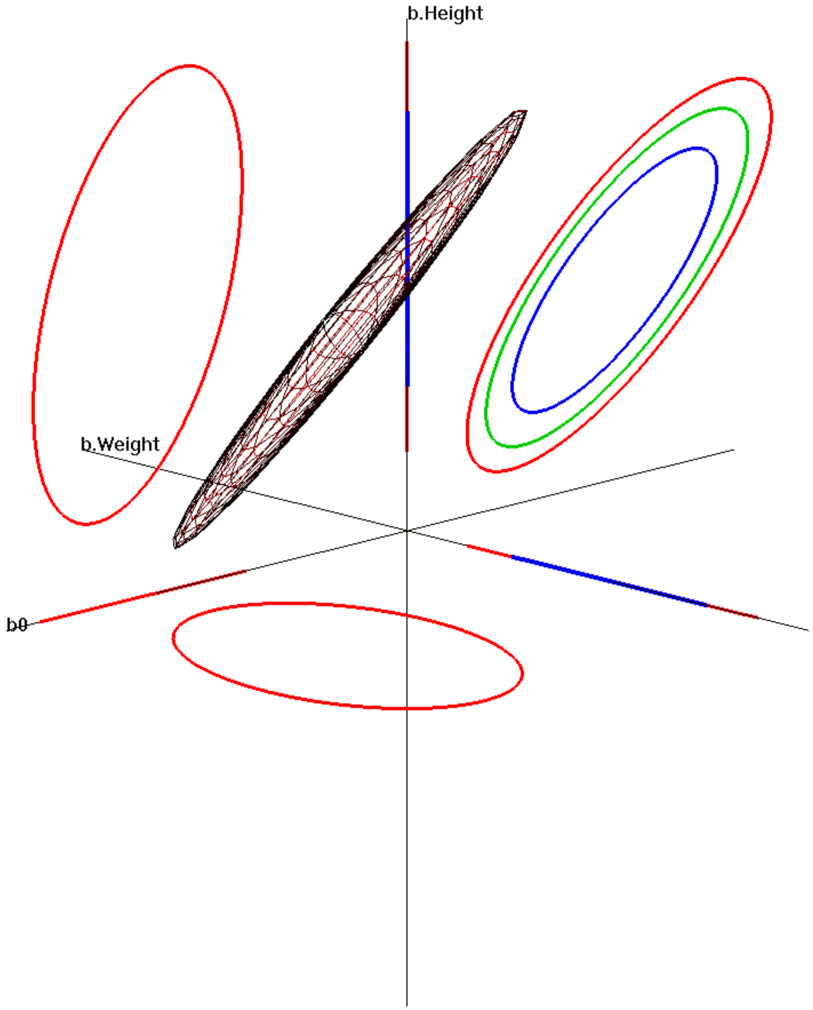
For slopes:

$$\begin{pmatrix} \hat{\beta}_1 \\ \vdots \\ \hat{\beta}_k \end{pmatrix} \oplus \sqrt{dF_{d,\nu}^{0.95}}\,\frac{s_e}{\sqrt{n}}\sqrt{\Sigma_X^{-1}}$$

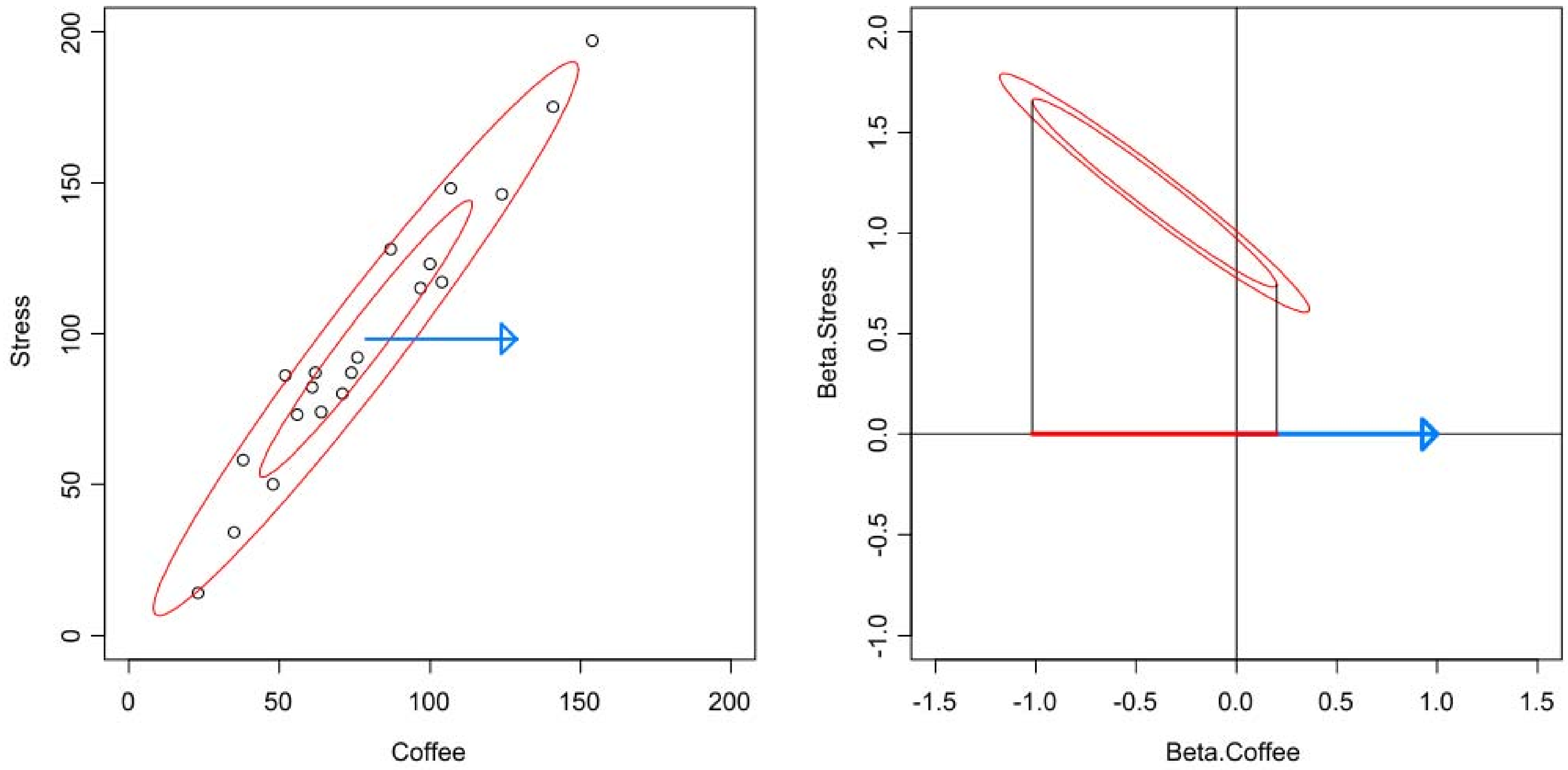
A space of linear images of dimension *d* of a 95% Scheffé confidence region has joint 95% probability of coverage.

Choosing *d* = 1, we get shadows of the ellipse that are ordinary 1-dimensional *t* intervals.

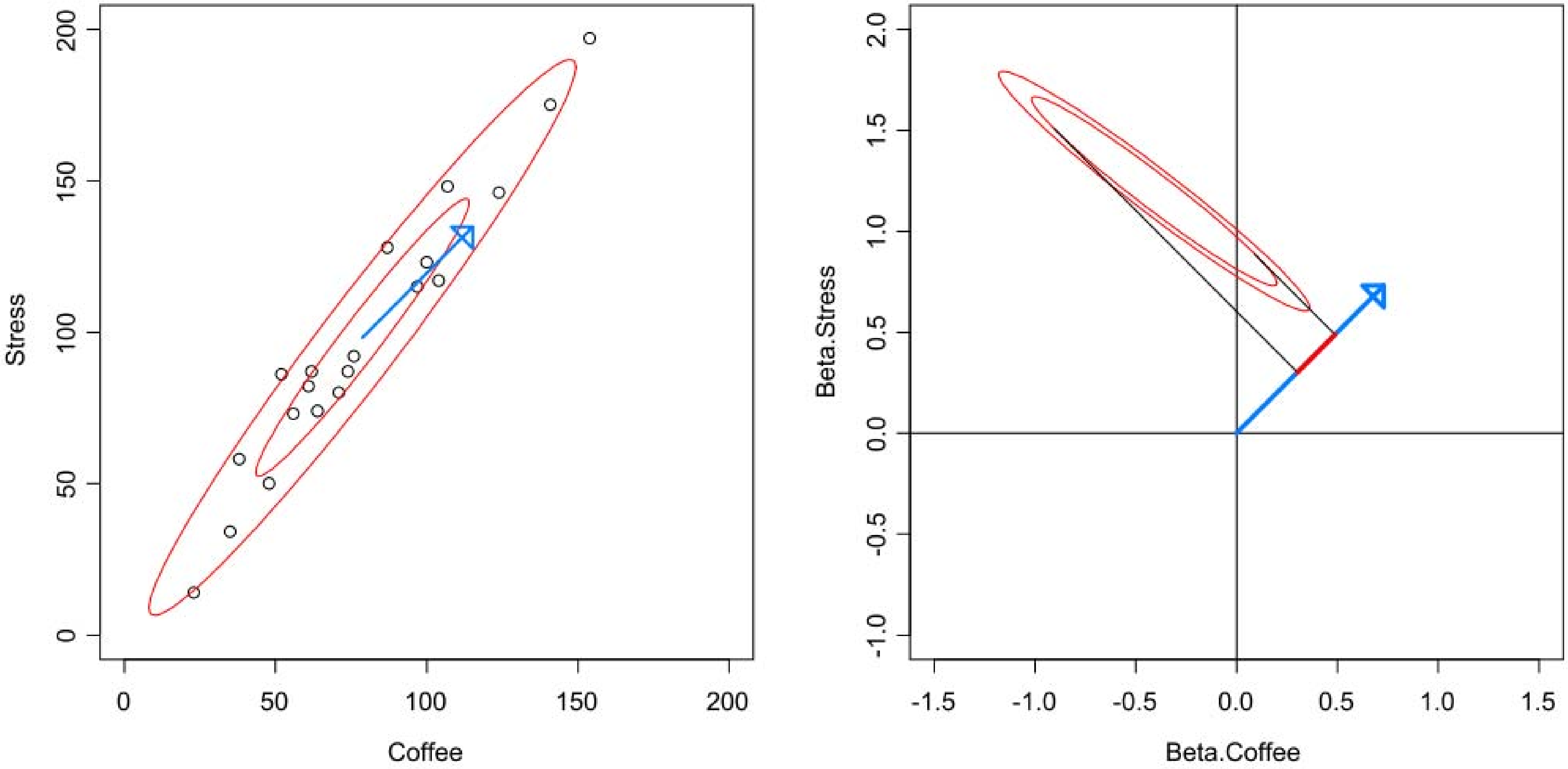
With *d* = 2, and 2 predictors we get the joint 95% confidence region for both slopes simultaneously.



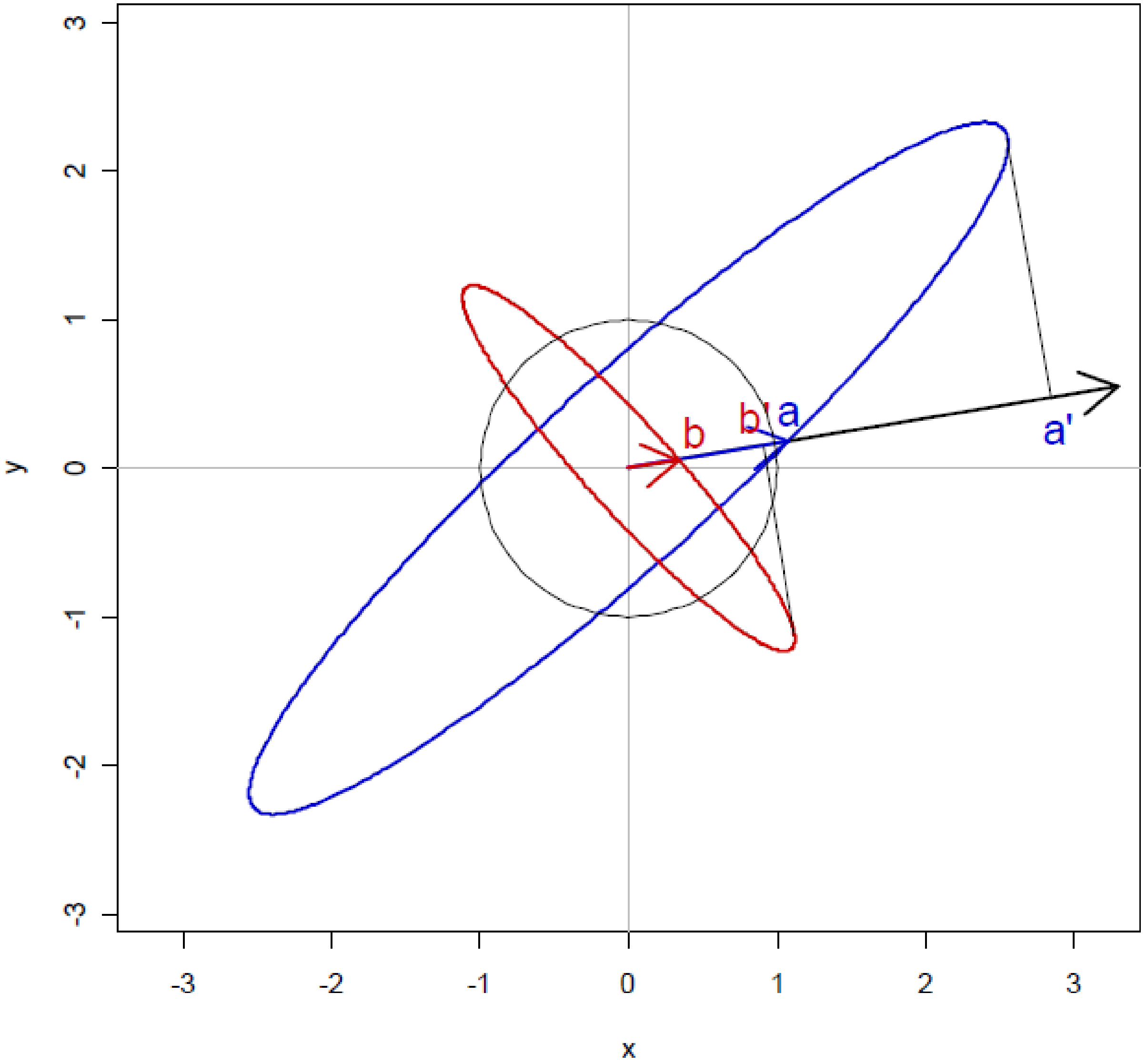
d = 1 (blue), 2 (green) or 3 (red) produces Scheffé confidence regions that have 95% coverage for *a priori* hypotheses of the corresponding dimension.



A general linear combination of β_{Coffee} and β_{Stress} is equivalent to a directional derivative times a constant. If both graphs are 'euclidean' (i.e. a unit of stress is plotted to have the same size as a unit of coffee, the confidence interval is obtained by taking the shadow of the confidence ellipse onto the corresponding axis in beta space.



Since the shape of the confidence ellipse is the ‘inverse’ of the shape of the data ellipse (the length of shadow of one ellipse is inversely proportional to the size of the slice of the other ellipse), the confidence interval is narrower in directions in which the data ellipse is larger.



The relationship between

$\mathcal{E} = \{\mathbf{x} : \mathbf{x}'\Sigma^{-1}\mathbf{x} = 1\}$

and

$\mathcal{E}^* = \{\varphi : \varphi'\Sigma\varphi = 1\}$

slice × shadow = 1
shadow × slice = 1

i.e.

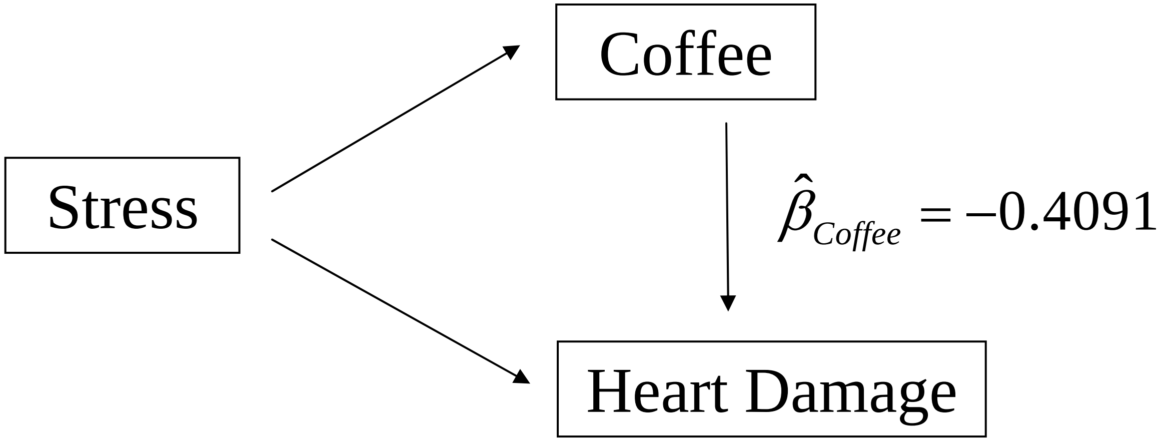
b × a’ = 1
b’ × a = 1

Question: Which estimate $\hat{\gamma}_{Coffee} = 1.1082$ or $\hat{\beta}_{Coffee} = -0.4091$should we use to assess the potential harm of coffee consumption?

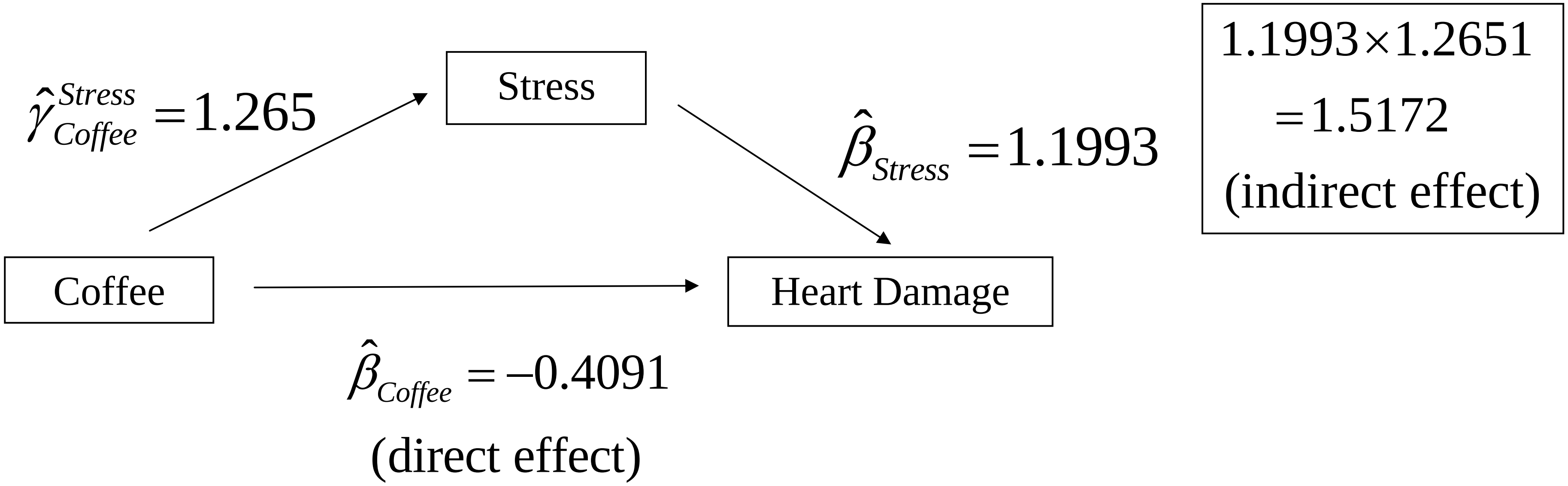
Answer: It depends!

Consider two extreme possibilities:

Extreme I: Stress is a **confounding factor**: Stress causes an increase in coffee consumption and separately causes an increase in heart damage.
In this case an 'exogenous' change in coffee consumption will not change stress and we need to estimate the effect of coffee keeping stress constant.
A plausible example would be if Coffee is used as a palliative or remedy for the effect of Stress. Coffee consumption is highly correlated with Heart Damage because it is taken to mitigate the effect of Stress.



Extreme II: Stress is a **mediating factor**: Coffee Consumption affects Stress which in turn affects Heart Damage in addition to possibly affecting Heart Damage directly:



$\hat{\gamma}_{Coffee}^{Stress} = 1.265$

Stress

$\hat{\beta}_{Stress} = 1.1993$

1.1993×1.2651
$= 1.5172$
(indirect effect)

Coffee

Heart Damage

$\hat{\beta}_{Coffee} = -0.4091$
(direct effect)

Classical path analysis total effect decomposition:

$$\text{total effect} = \text{direct effect} + \text{indirect effect}$$

$$\hat{\gamma}_{Coffee} = \quad \hat{\beta}_{Coffee} \; + \hat{\beta}_{Stress} \times \; \hat{\gamma}_{Coffee}^{Stress}$$

$$1.1081 = -0.4091 + 1.1993 \times 1.2651$$

$$1.1081 = -0.4091 + \qquad 1.5172$$

Note; The interpretation of this decomposition is only useful if Stress is a mediating factor.

It is alway mathematically true that:

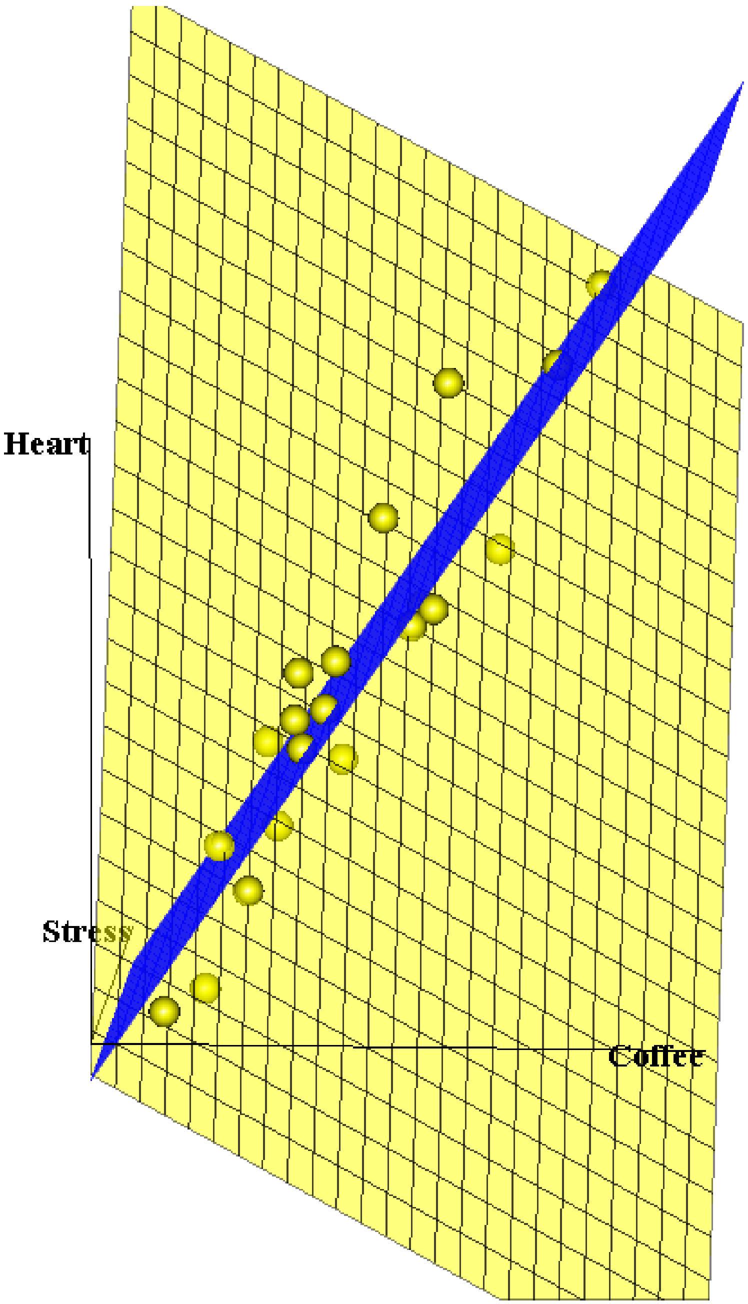
$$\gamma_{Coffee} = \beta_{Coffee} + \beta_{Stress} \times \hat{\gamma}_{Coffee}^{Stress}$$

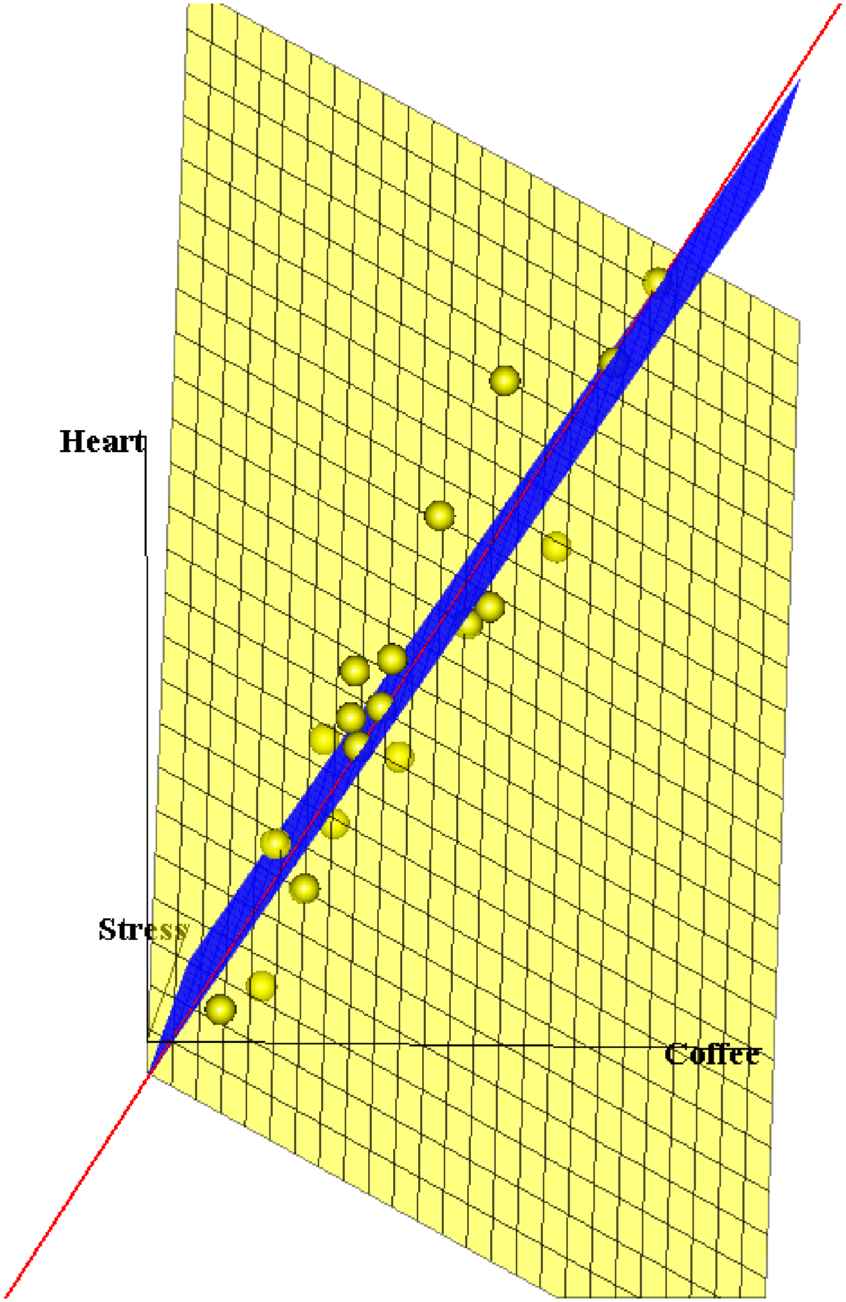
$$\hat{\gamma}_{Coffee} = \hat{\beta}_{Coffee} + \hat{\beta}_{Stress} \times \hat{\gamma}_{Coffee}^{Stress}$$

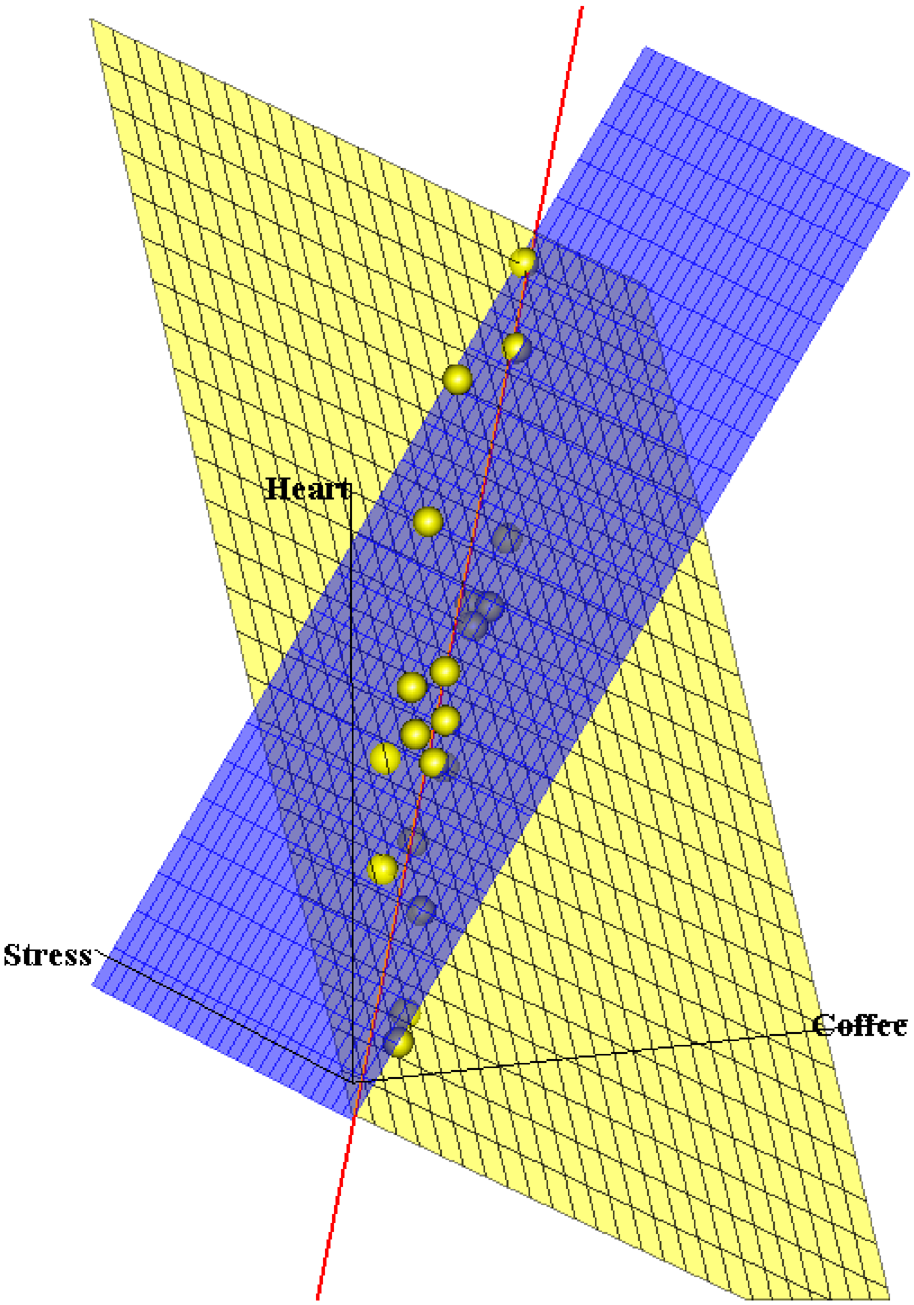
$$\gamma_{Coffee} = \beta_{Stress} + \beta_{Coffee} \times \hat{\gamma}_{Stress}^{Coffee}$$

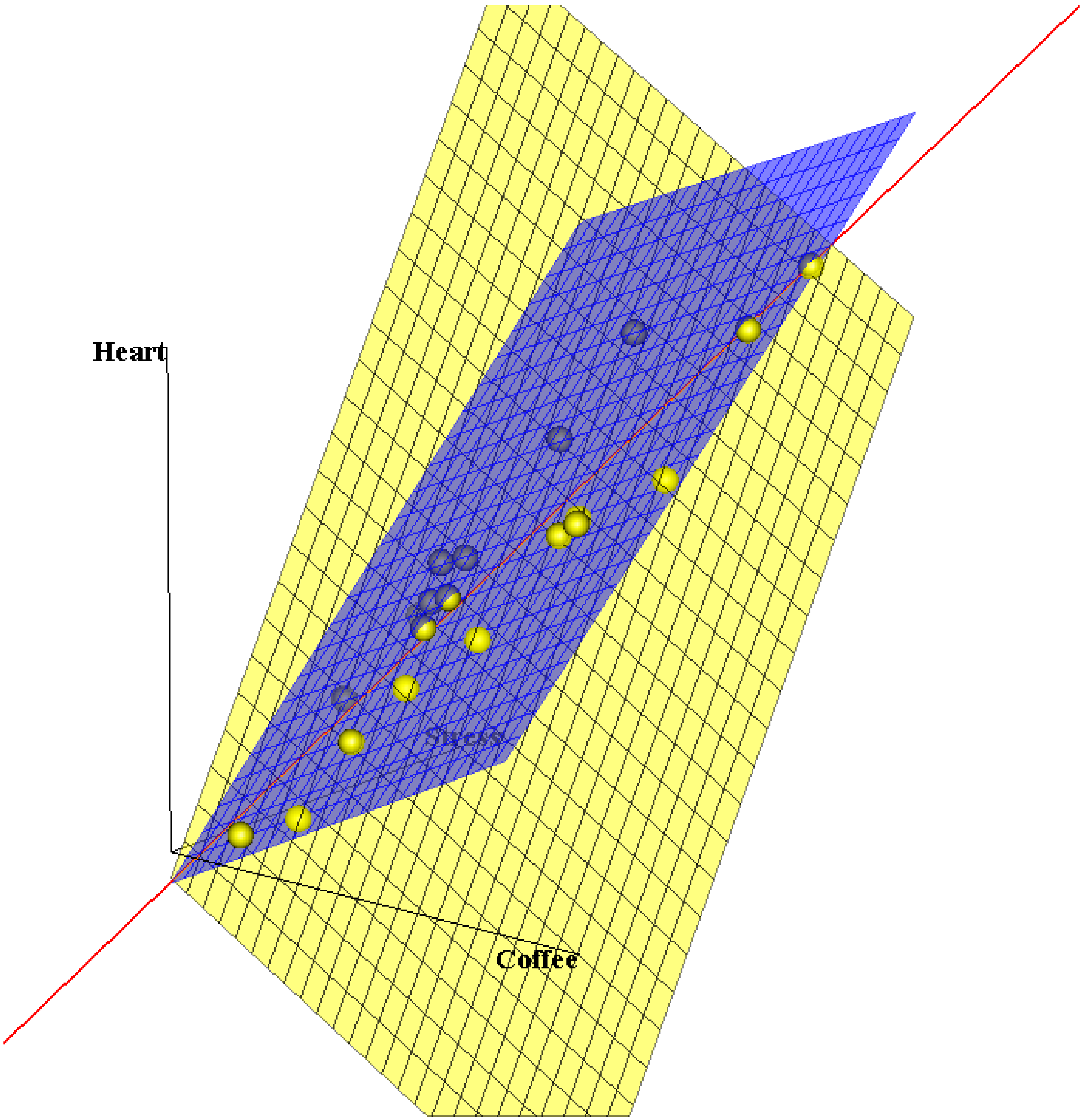
$$\hat{\gamma}_{Coffee} = \hat{\beta}_{Stress} + \hat{\beta}_{Coffee} \times \hat{\gamma}_{Stress}^{Coffee}$$

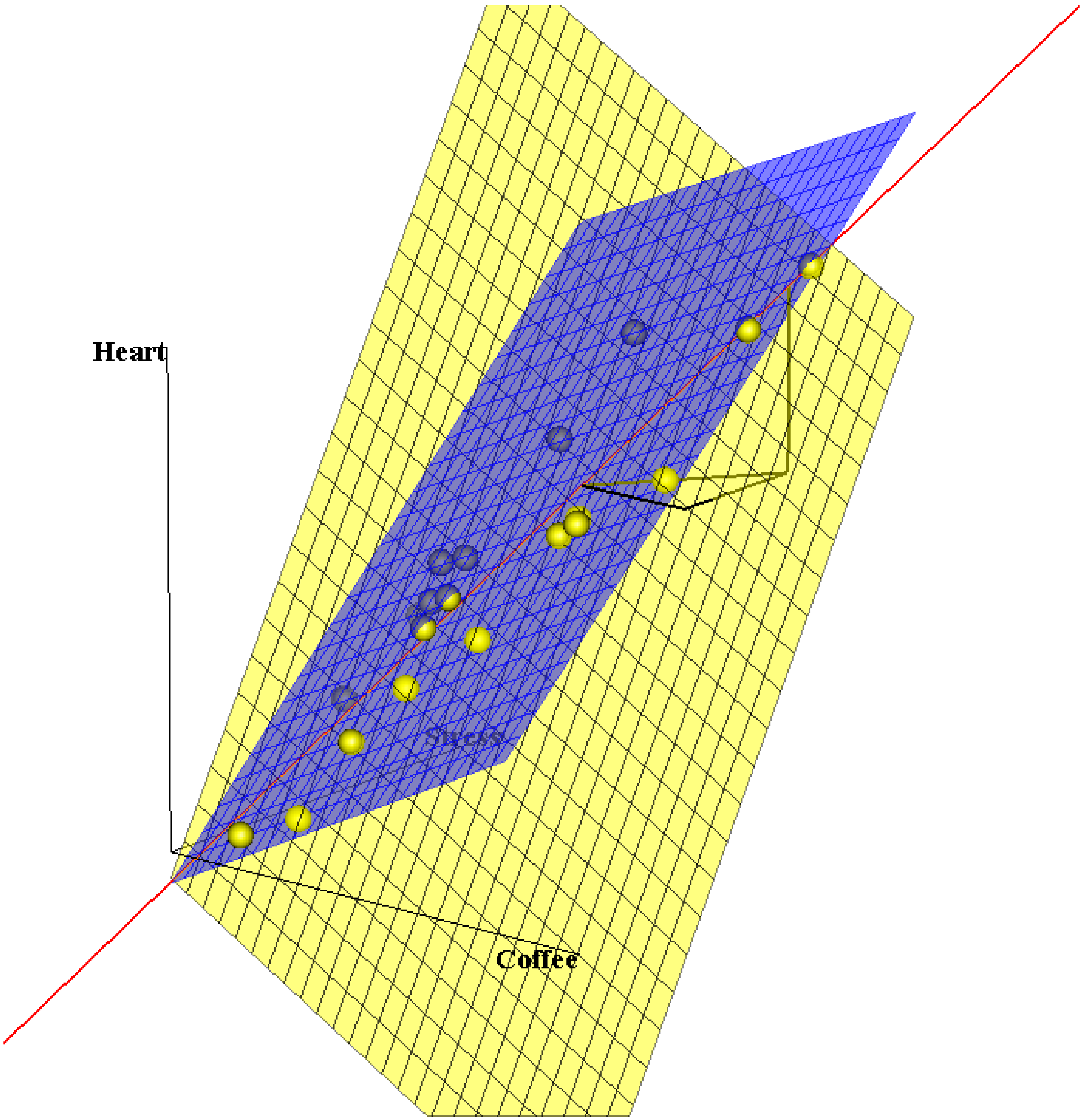
The following pages show the decomposition in data space.

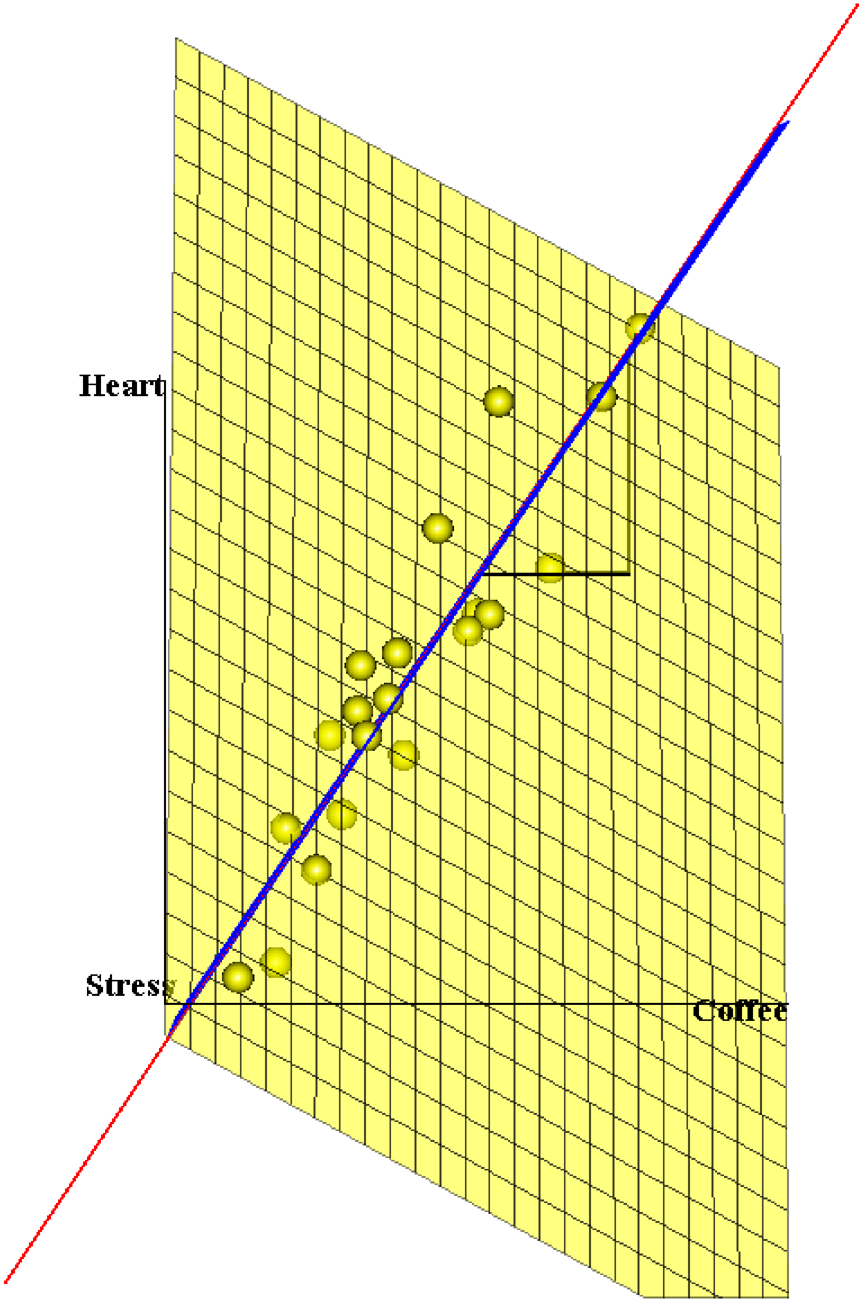

Heart
Stress
Coffee

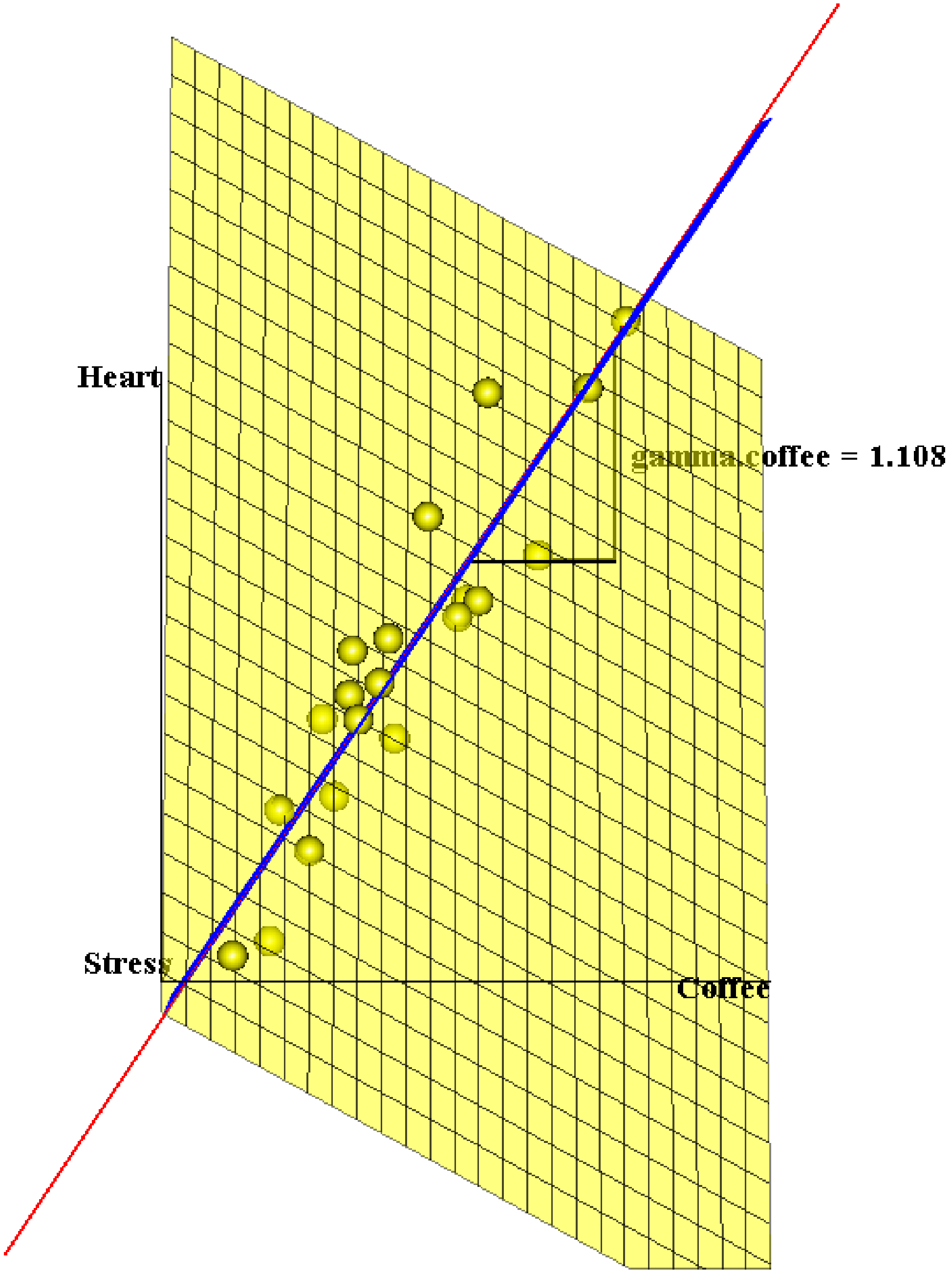

Heart
Stress
Coffee

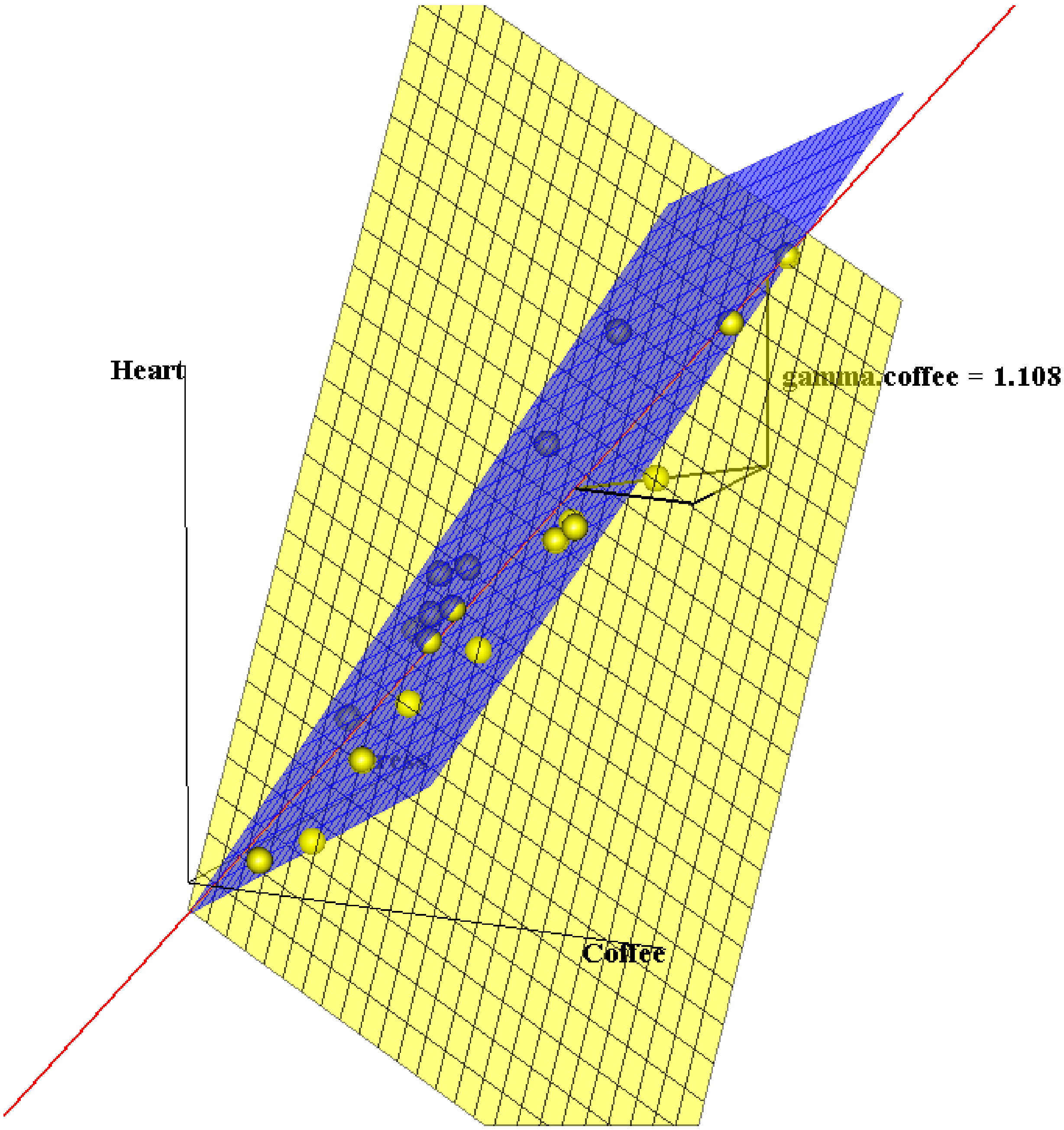

Heart
Stress
Coffee


Heart
Stress
Coffee

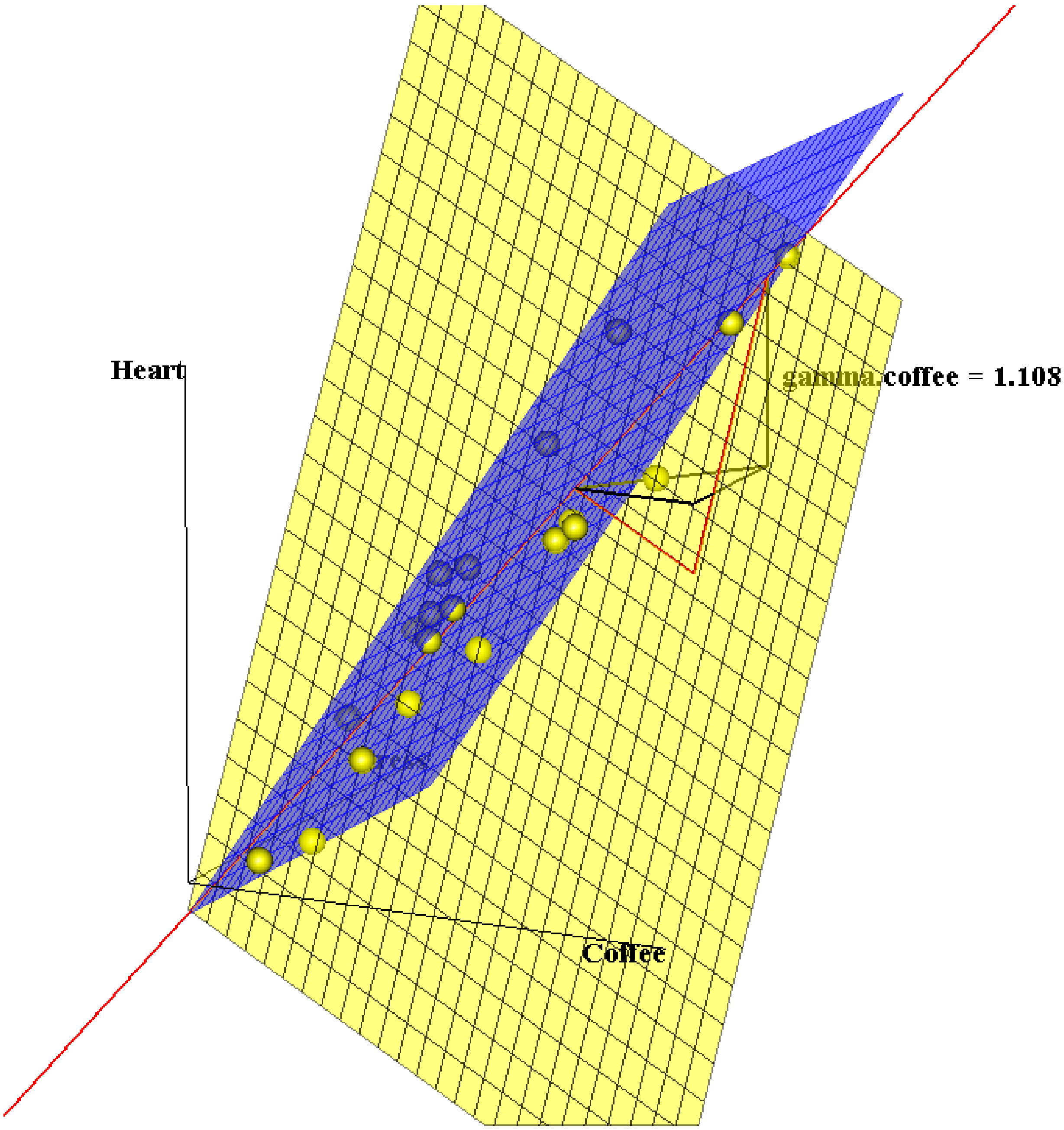

Heart
Stress
Coffee

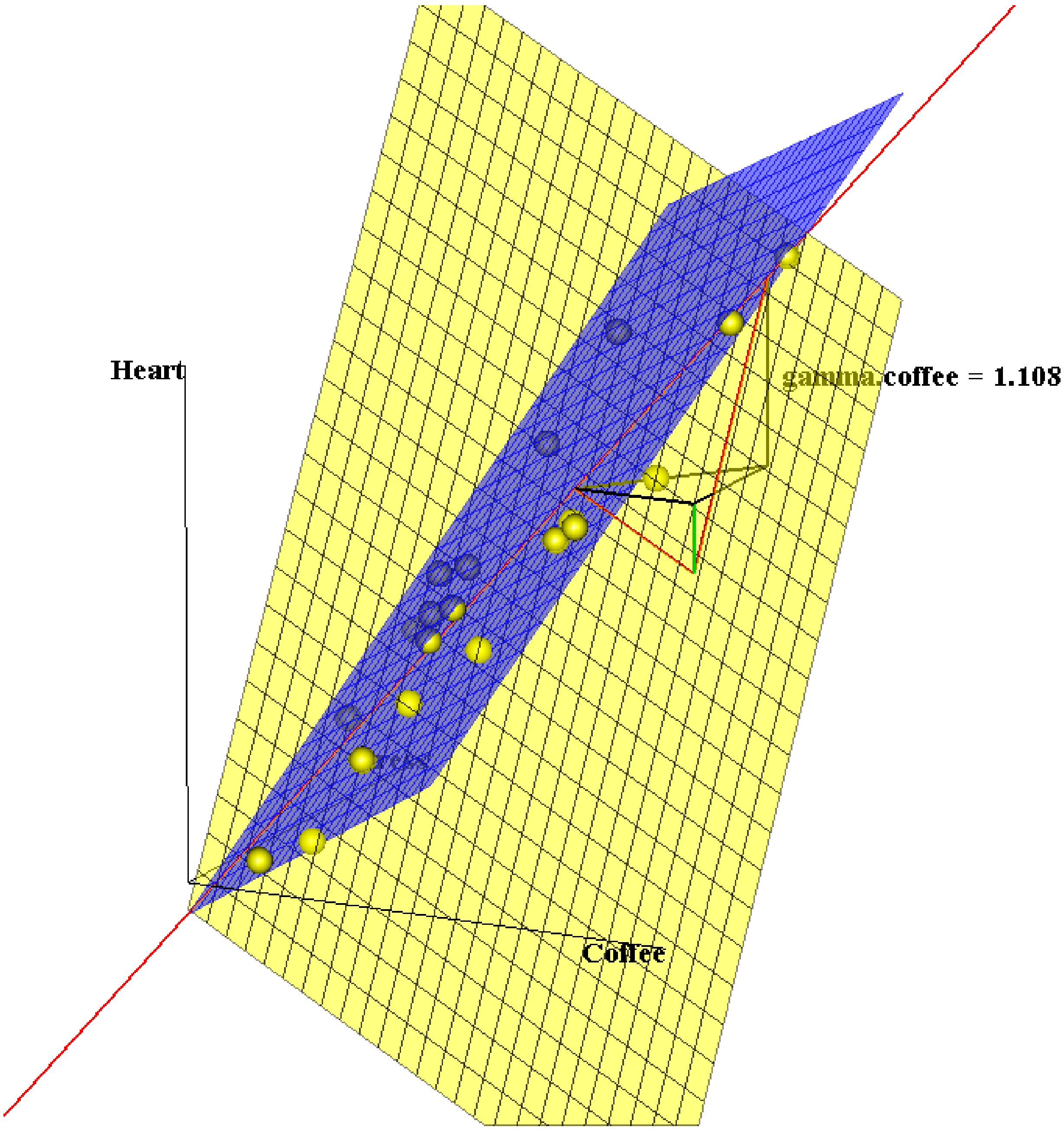

Heart
Stress
Coffee

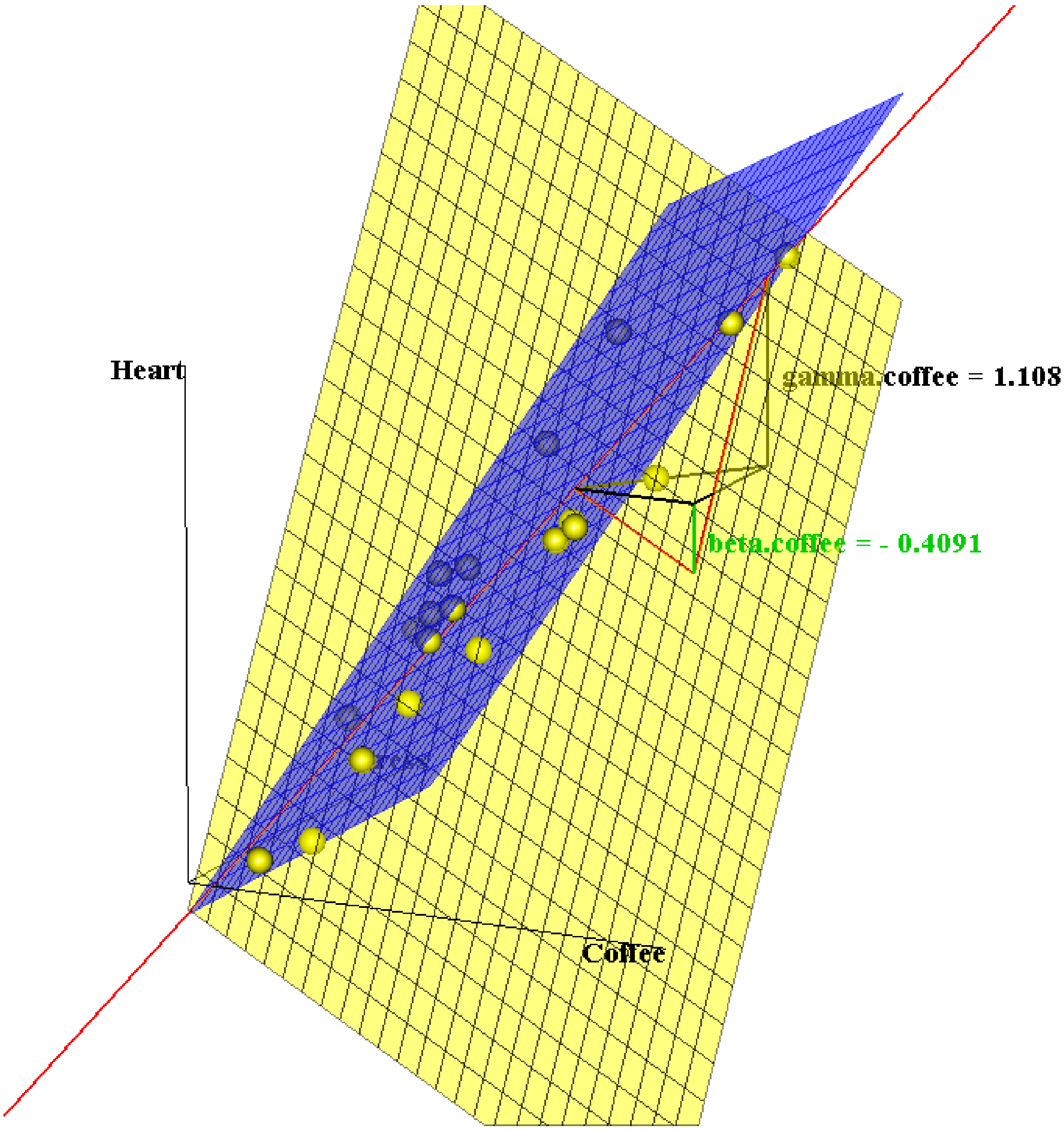

Heart
gamma.coffee = 1.108
Stress
Coffee

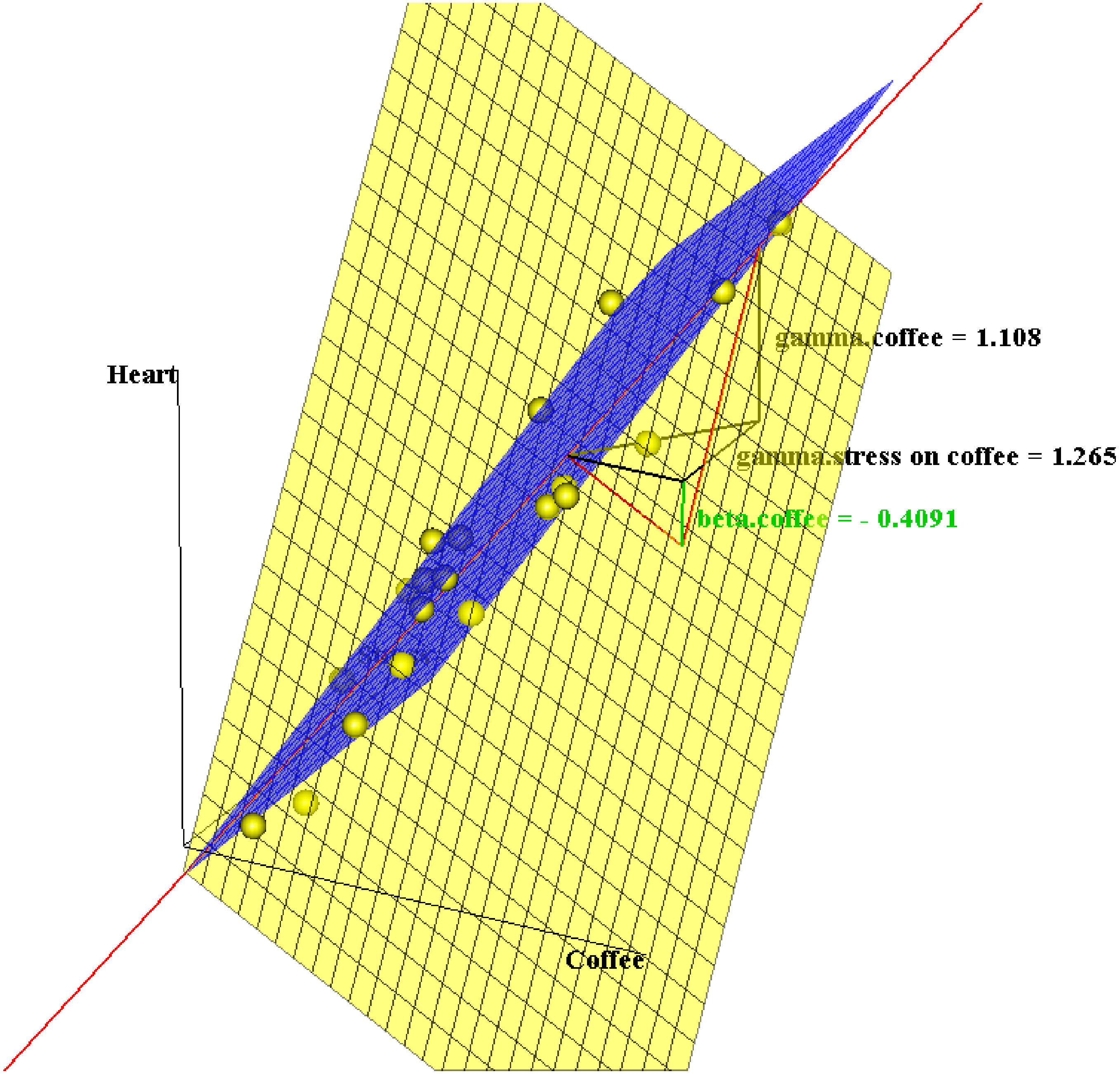


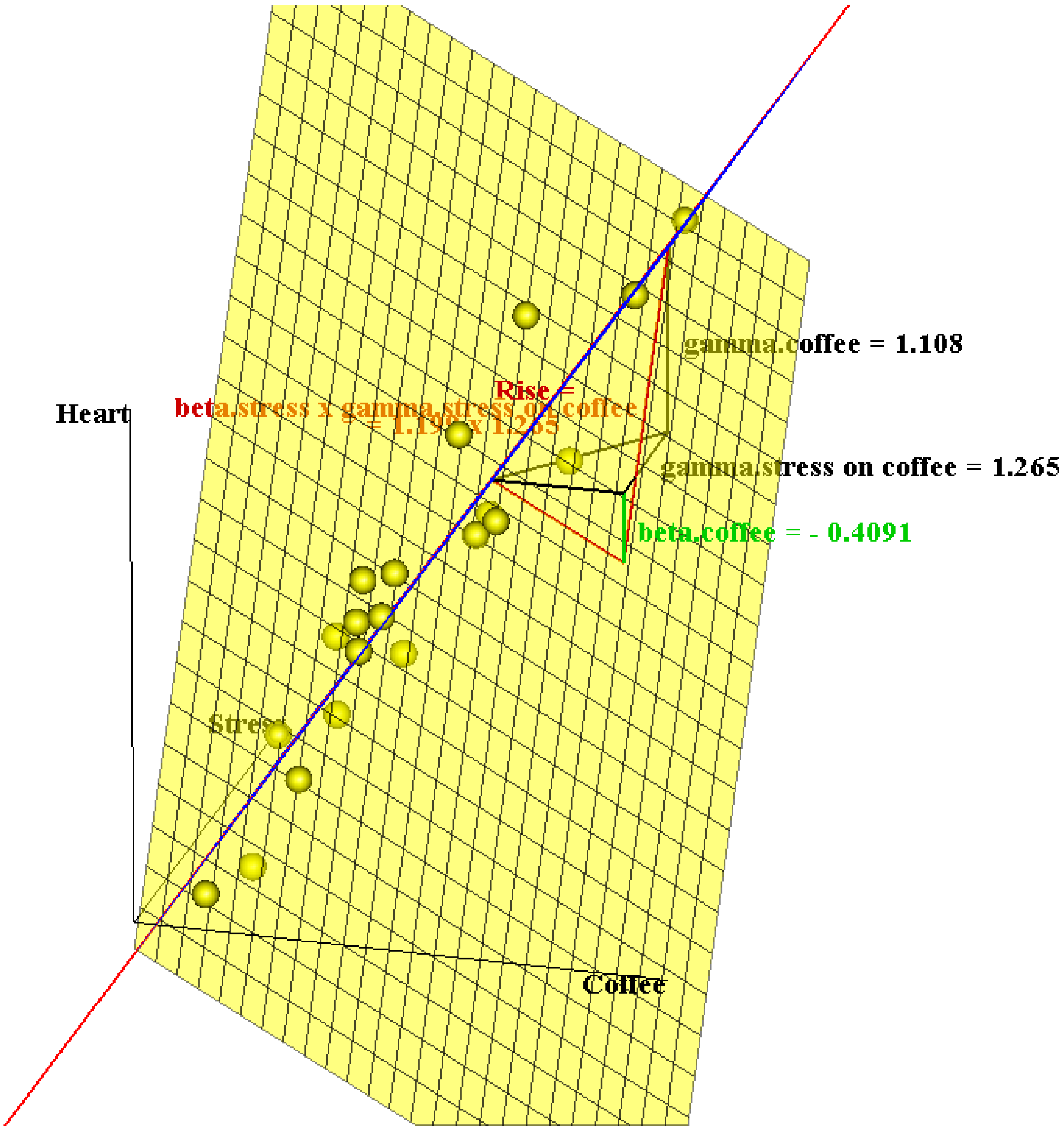
Heart
gamma.coffee = 1.108
Stress
Coffee


Heart
gamma.coffee = 1.108
Stress
Coffee


Heart
gamma.coffee = 1.108
ress
Coffee


Heart
gamma.coffee = 1.108
beta.coffee = - 0.4091
Coffee


Heart
gamma.coffee = 1.108
gamma.stress on coffee = 1.265
beta.coffee = - 0.4091
Coffee


gamma.coffee = 1.108
Rise =
beta.stress x gamma.stress on coffee
= 1.19[illegible] x 1.265
Heart
gamma.stress on coffee = 1.265
beta.coffee = - 0.4091
Stress
Coffee

Discussion

The correct analysis depends on the question we're asking, the data and the *true* model. Since we usually don't know the true model our conclusions are contingent.

The right answer depends on suppositions that cannot be verified from the data. Situation 1 or 2 can produce exactly the same data so they cannot be distinguished on the basis of the data alone without some external information or a willingness to make conclusions that are contigent on suppositions.

The interpretation of regression coefficients depends on suppositions about the relationships among the variables. Generally these suppositions cannot be checked from the data and conclusions are contingent on suppositions.

The interpretation of a regression coefficient depends on what other variables are in the model if those variables are related to the variable of interest.

It might take more than one model to answer a question, e.g. questions regarding mediating factors.

The Fundamental Contingency Table of Statistics

Or: putting what we've just seen in perspective

The Fundamental 2 x 2 Contingency Table of Statistics

		Types of Data	
		Experimental	**Observational**
Types of Inference	**Causal**	Where Fisher wants to be (the gold standard for causality)	**The real challenge**
	Predictive	Very rare but problematic	Okay: This is the topic of Frank Harrell's *Regression Modeling Strategies'*

Why is causal inference with observational data a challenge?

> We can't be sure that an association between X and Y is causal. Some other variable(s), Zs, – confounding factors – might be 'causing' the association.

The magic of experiments: All Zs (measurable or not, known or not) are random with respect to X. So Z can only be the cause by chance which is addressed by the p-value.

Strategy with obervational data:

Control for all the Zs you can, using one of:

Statistical control: use a model that includes Zs, include them in a statistical model and adjust statisitically. Modern thinking: we don't need all the Zs in the model to avoid bias, only the “propensity score” (prediction of X from Zs). This can fail with the wrong model.

Matching: only perform comparisons between observations with similar Zs. Again all that really matters to avoid bias is the propensity score if you have the Zs.

Structural methods: Build a structural causal model.

How do we fail? Don't know all relevant Zs, or can't measure them with enough accuracy

Role of longitudinal data analysis: Controls for Zs (known or not, measure or not) that are time-invariant properties of subjects – provided we use contextual variables … which are really a form of “propensity scores”.

Outliers in Multiple Regression

Three types:

1) Typical values for predictors, Y atypical

Little impact on $\hat{\beta}$

Increases size of confidence intervals

Decreases power

2) Atypical values for predictors but Y consistent with other data

Little impact on $\hat{\beta}$

Shrinks confidence intervals – good if point valid, misleading if not

Creates false sense of power if point not valid

3) Atypical values for predictors and Y not consistent with other data

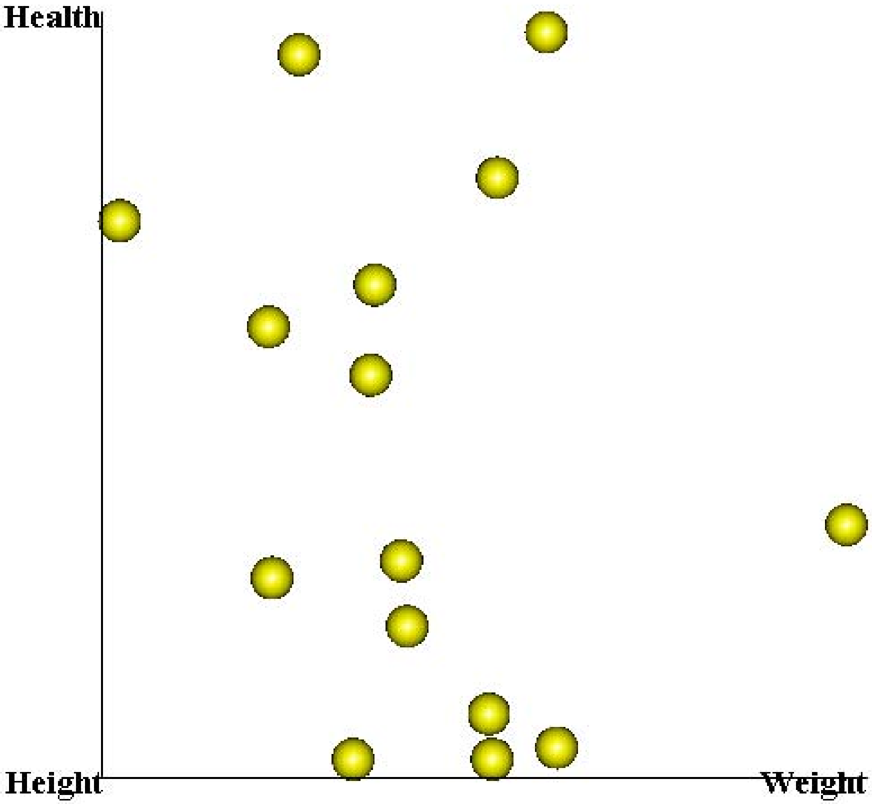
Large impact on $\hat{\beta}$

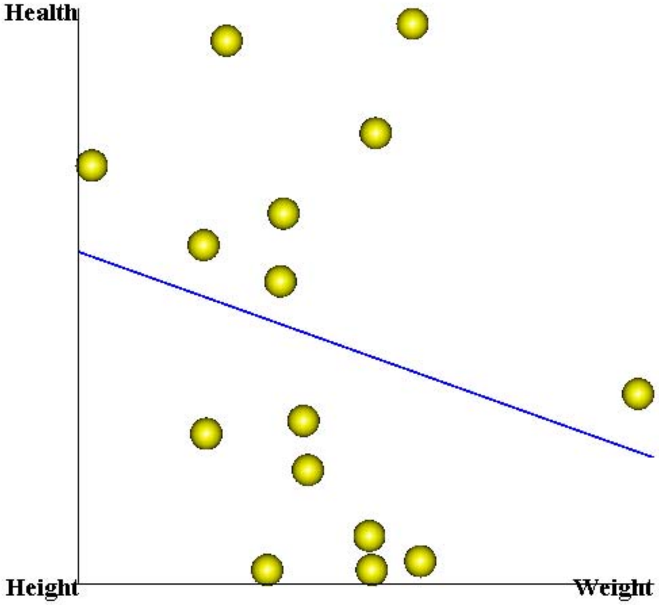
Could shrink or expand CIs

Makes a mess of everything

Another artificial data set:

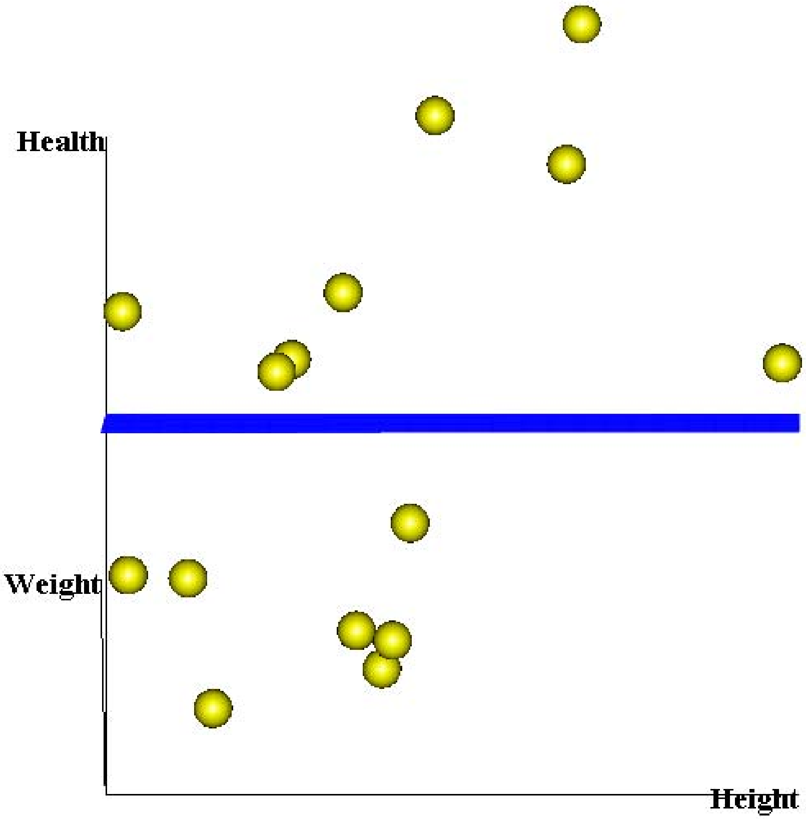
Health as predicted by Weight and Height


Health
Height
Weight



```
> fit.simple <- lm(
Health ~
          Weight, hh,
subset =
          Type == 0)
> summary(fit.simple)
```

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   1.2322     0.1327   9.289 4.21e-07 ***
Weight       -0.1027     0.1185  -0.867    0.402
Residual standard error: 0.1774 on 13 degrees of freedom
Multiple R-squared: 0.05463,    Adjusted R-squared: -0.01809
F-statistic: 0.7512 on 1 and 13 DF,  p-value: 0.4018
```



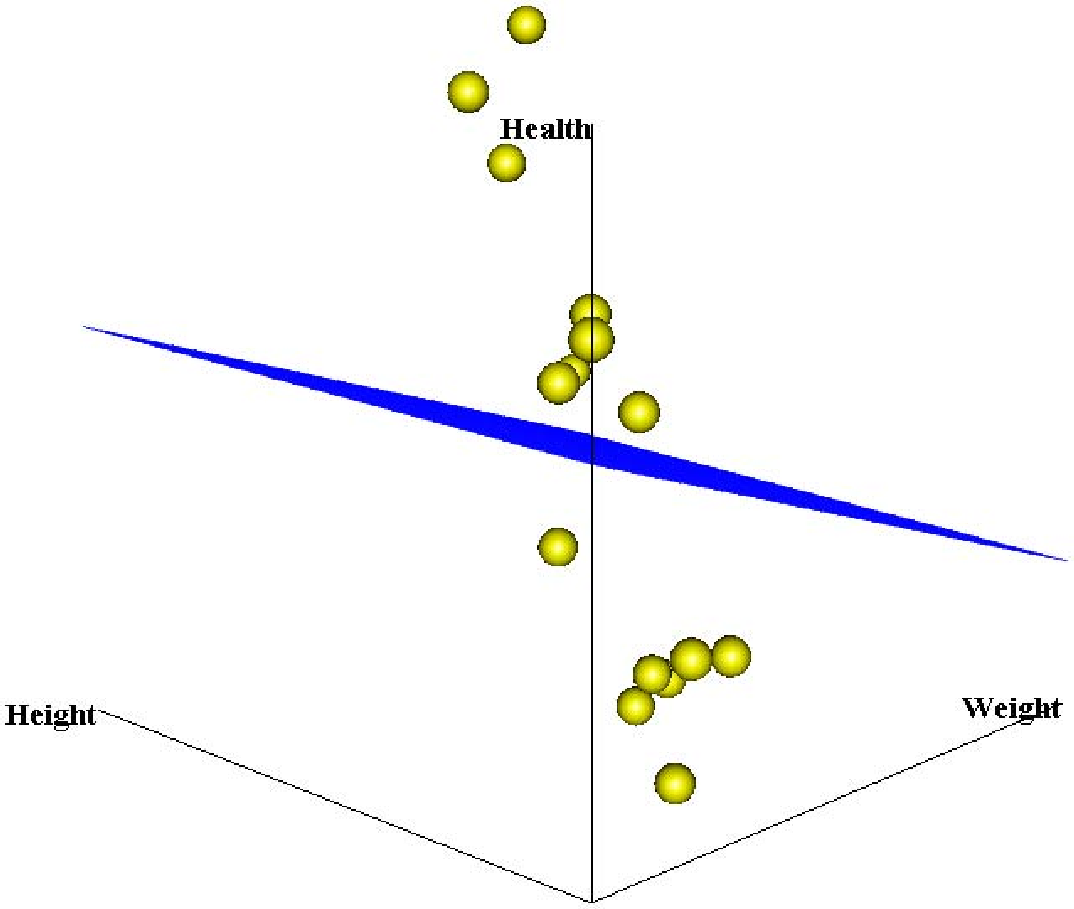
uppose we wonder whether we should
dd Height in the model.

here is no obvious reason to add height.
e don't think that height should have
nuch of an effect on Health.

e can do a traditional diagnostic
esidual plot by plotting the residuals
gainst Height.

his is the same as rotating the data so
eight lies along the horizontal axis and
he plane is on edge.

he traditional residual plot against an
ew variable does not show a strong
attern and we might be tempted to stick
vith our model showing no significant
elationship between Height and Weight.

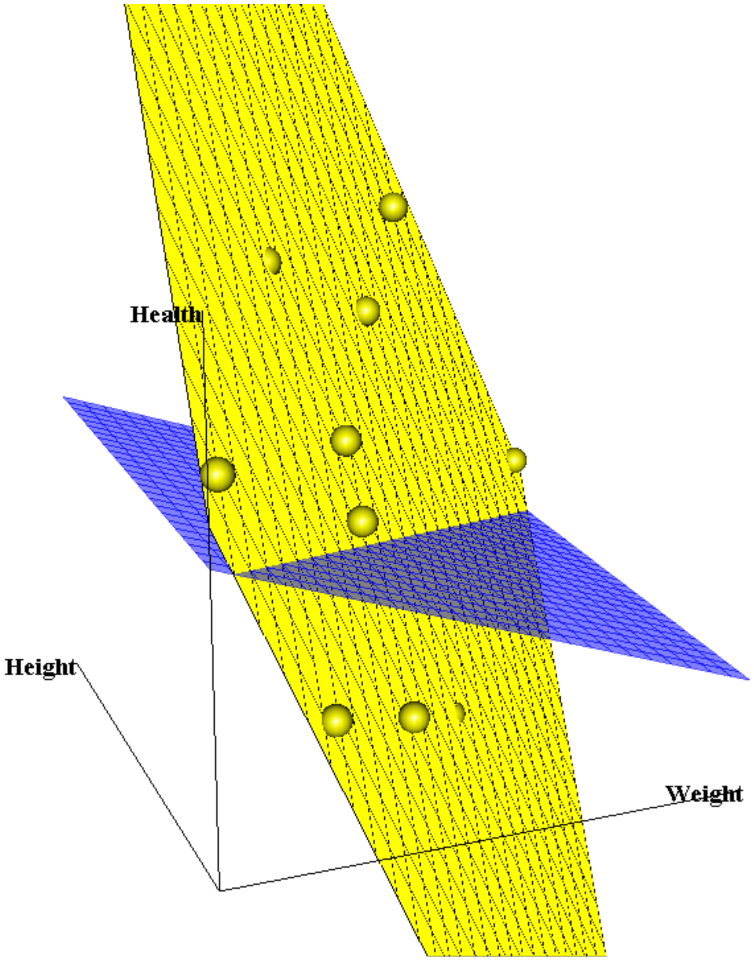


If we rotate the graph around the vertical axis keeping the plane on edge, we get to see all possible plots of residuals against linear combinations of Weight and Height.

There are two points of view that show a very strong relationship.

See
http://www.math.yorku.ca/~georges
and scroll down.

The view that produces maximum correlation is the added variable plot.



Multiple regression of Health on Weight and Height:

```
fit.mult <- lm ( Health
                ~ Height + Weight, hh ,
                          subset = Type ==
0)
```

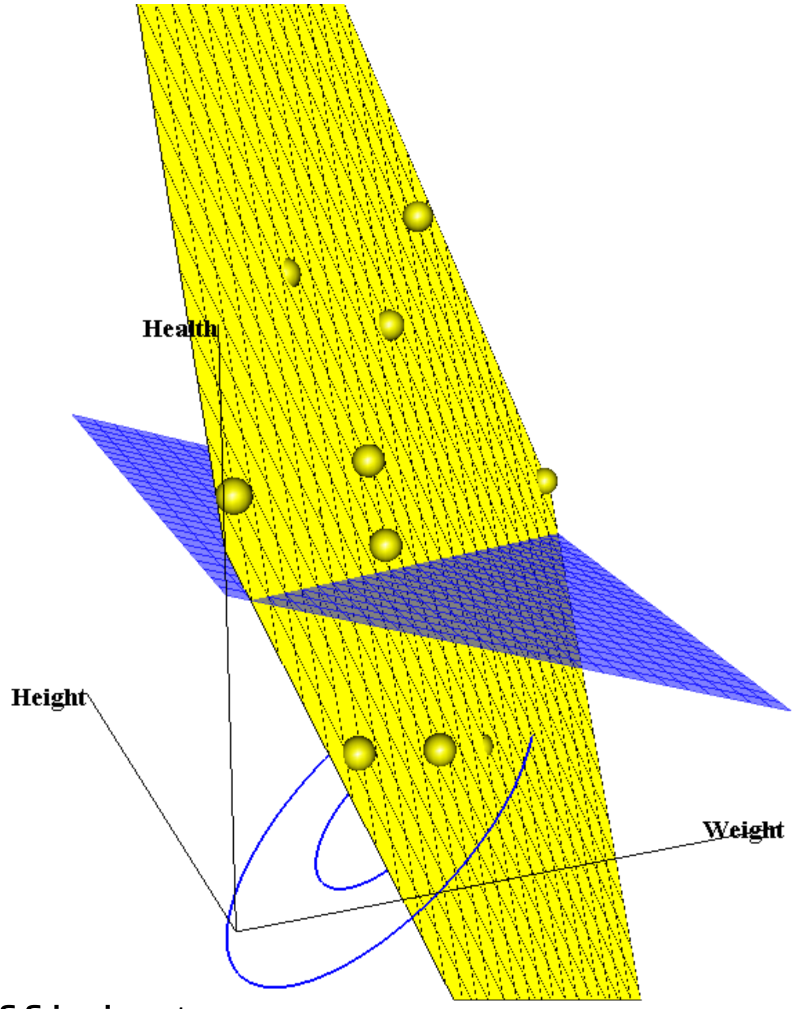
```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.02009    0.08529  11.960 5.02e-08 ***
Height       0.72500    0.13639   5.315 0.000184 ***
Weight      -0.65487    0.12380  -5.290 0.000192 ***

Residual standard error: 0.1008 on 12 degrees of freedom
Multiple R-squared: 0.7182,     Adjusted R-squared: 0.6712
F-statistic: 15.29 on 2 and 12 DF,  p-value: 0.000501
```

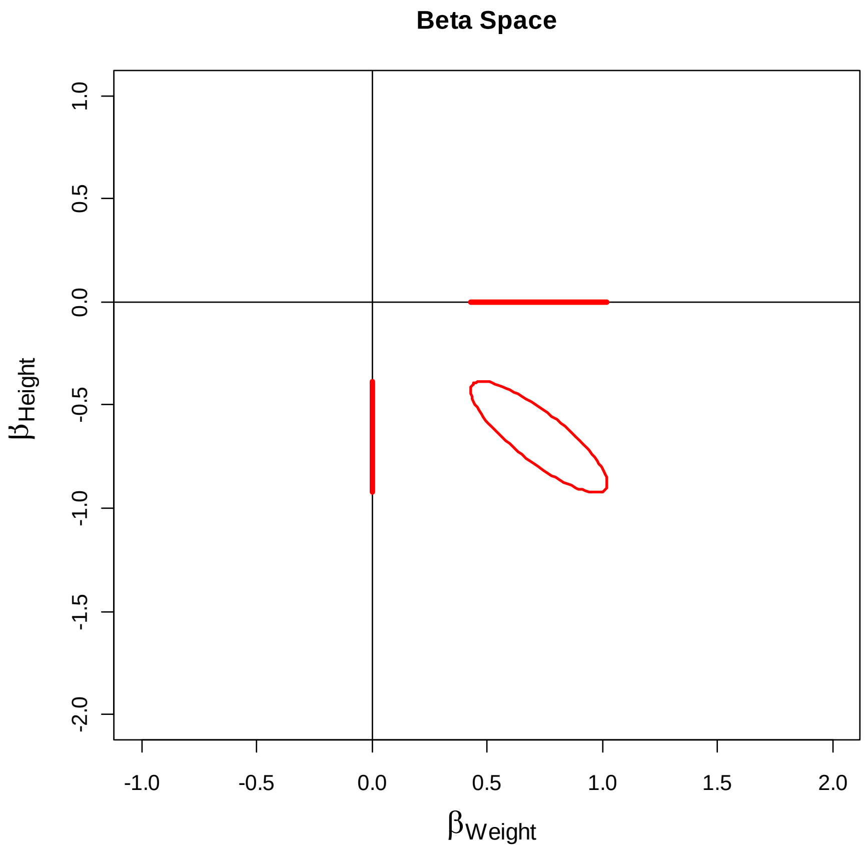
```
> fit.mult <- lm ( Health
        ~ Height + Weight, hh ,
             subset = Type == 0)
> summary( fit.mult )
...
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.02009    0.08529  11.960 5.02e-08 ***
Height       0.72500    0.13639   5.315 0.000184 ***
Weight      -0.65487    0.12380  -5.290 0.000192 ***

Residual standard error: 0.1008 on 12 degrees of freedom
Multiple R-squared: 0.7182,     Adjusted R-squared: 0.6712
F-statistic: 15.29 on 2 and 12 DF,  p-value: 0.000501
```

Does this mean that it's healthy to be tall?
Now weight has a significant negative coefficient!
Why now and not in the simple regression?

The answer is that in a simple regression, when we compare a light person to a heavy person we are not controlling for Height. We are comparing a short light person to a heavy tall person. If the harm in being overweight is in being too heavy relative to height, then we need to control for height to estimate the effect of being overweight.

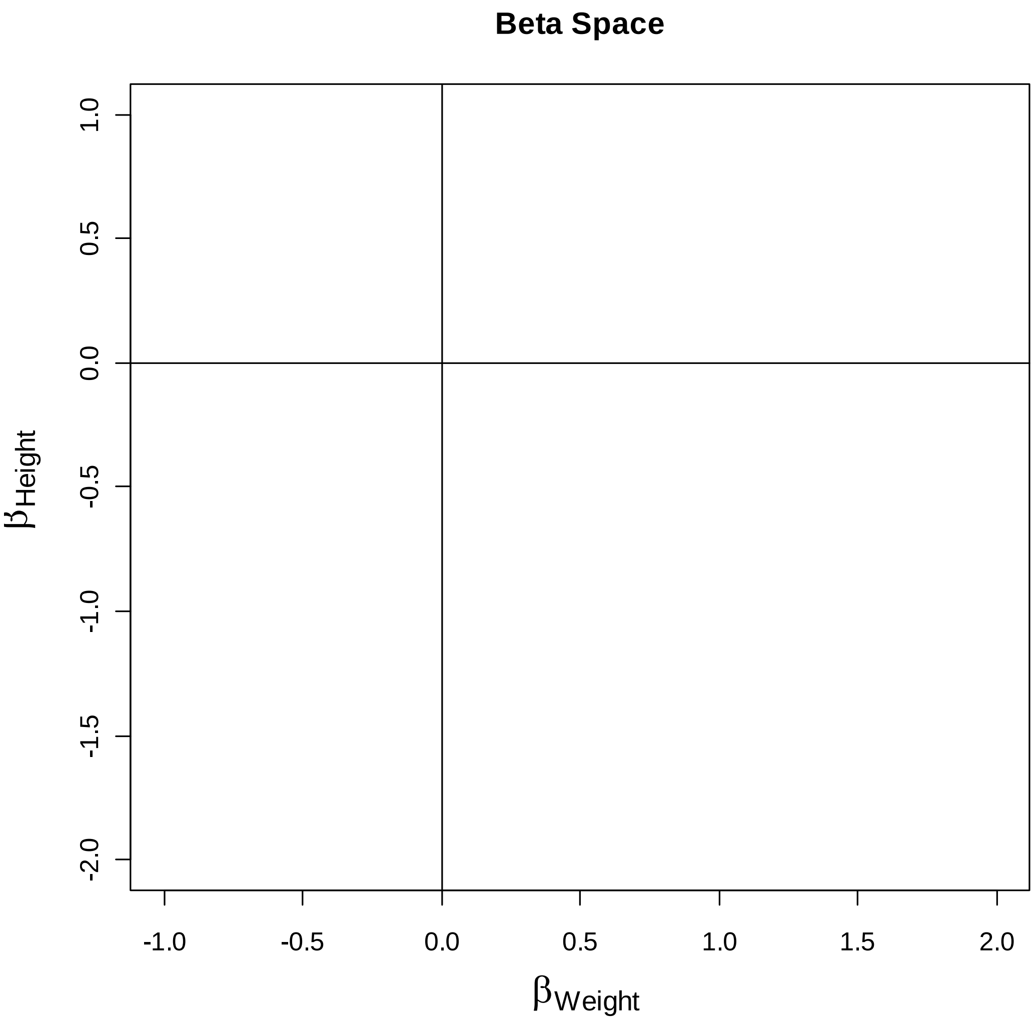
Question: Does this still mean that it is healthier to be tall?





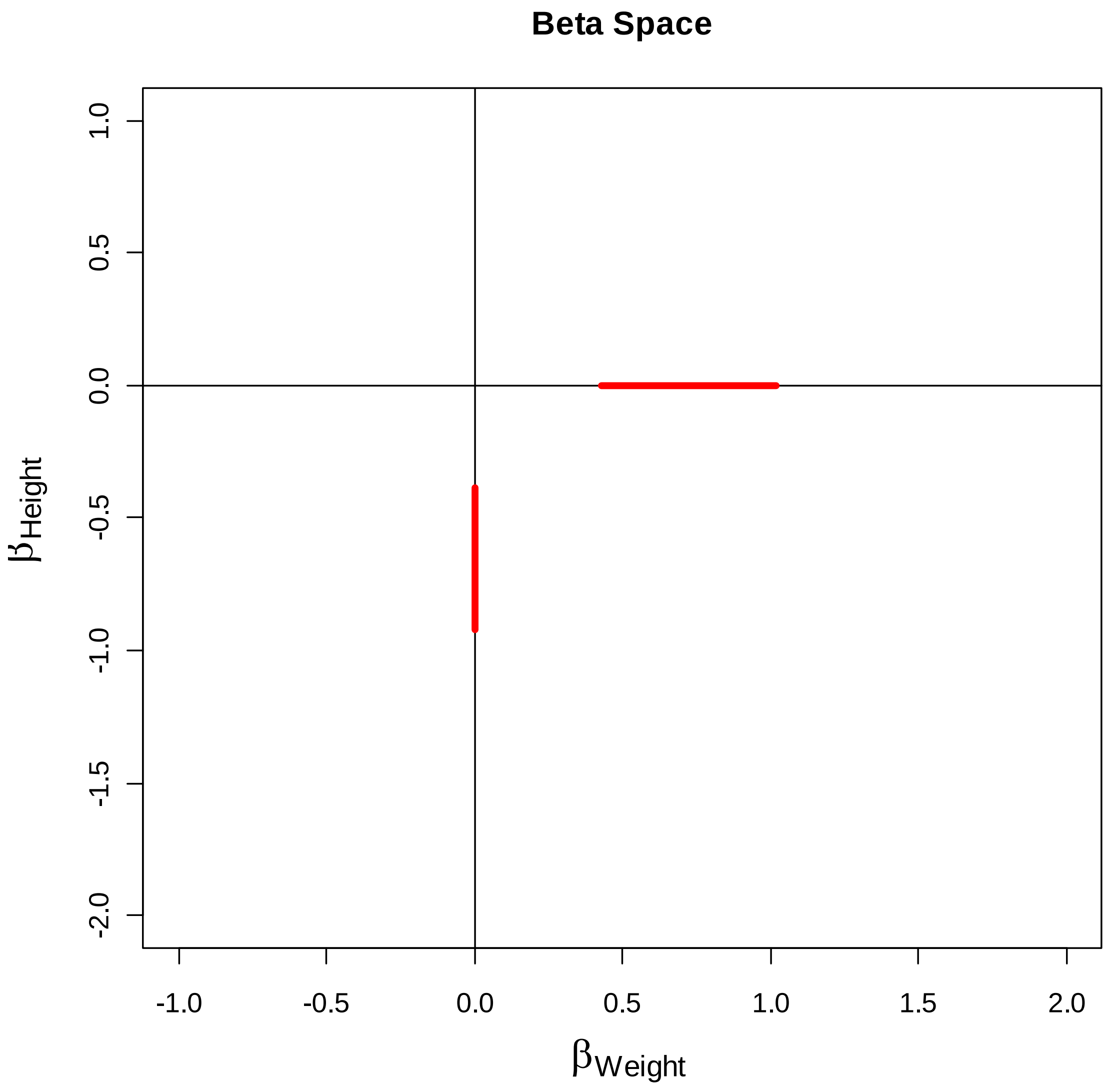
```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.02009    0.08529  11.960 5.02e-08 ***
Height       0.72500    0.13639   5.315 0.000184 ***  (Height CI)
Weight      -0.65487    0.12380  -5.290 0.000192 ***  (Weight CI)

Residual standard error: 0.1008 on 12 degrees of freedom
Multiple R-squared: 0.7182,     Adjusted R-squared: 0.6712
F-statistic: 15.29 on 2 and 12 DF,  p-value: 0.000501  (Conf. ellipse)
```

Beta space:

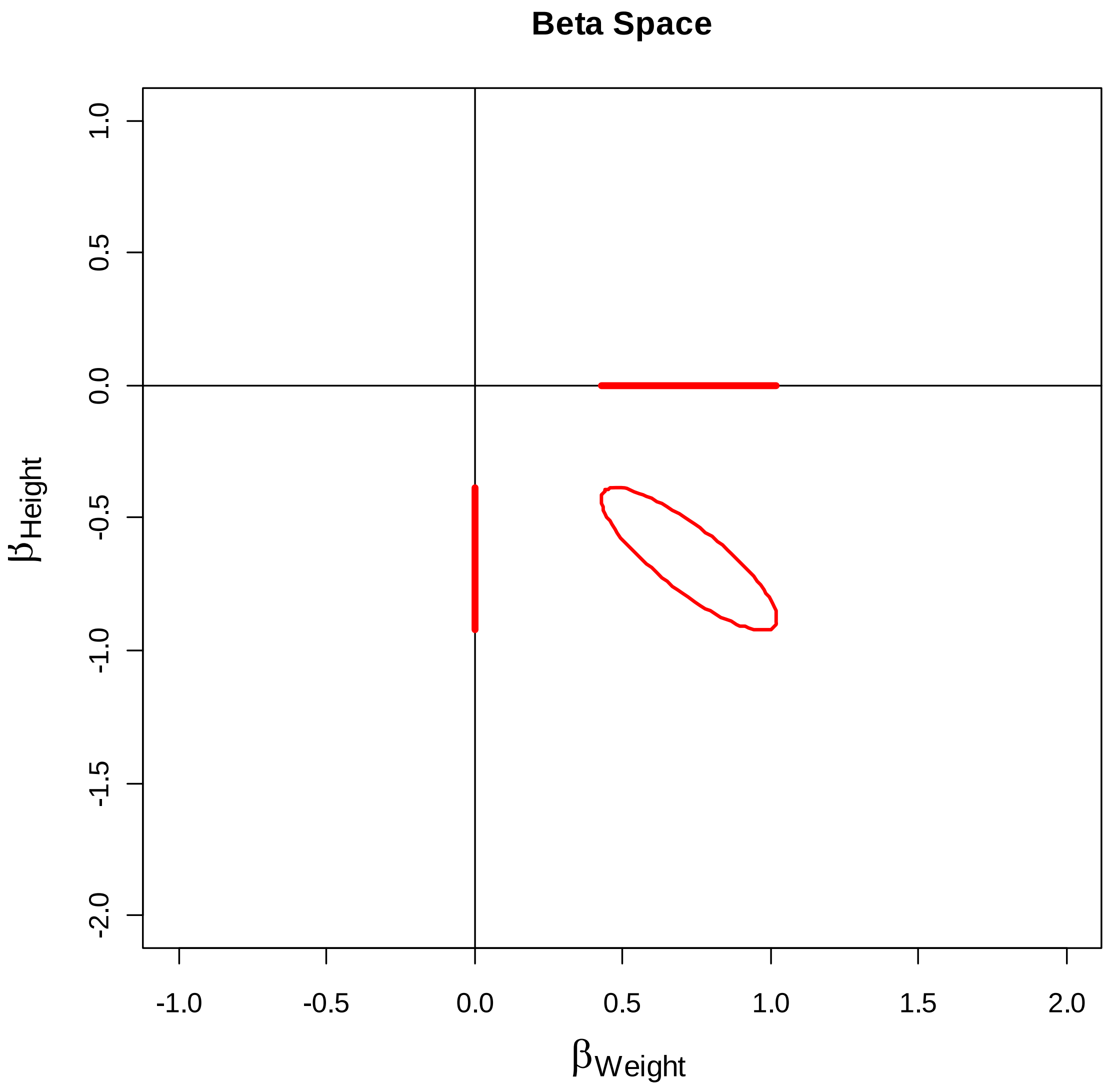


Axes are coefficients: true or estimated

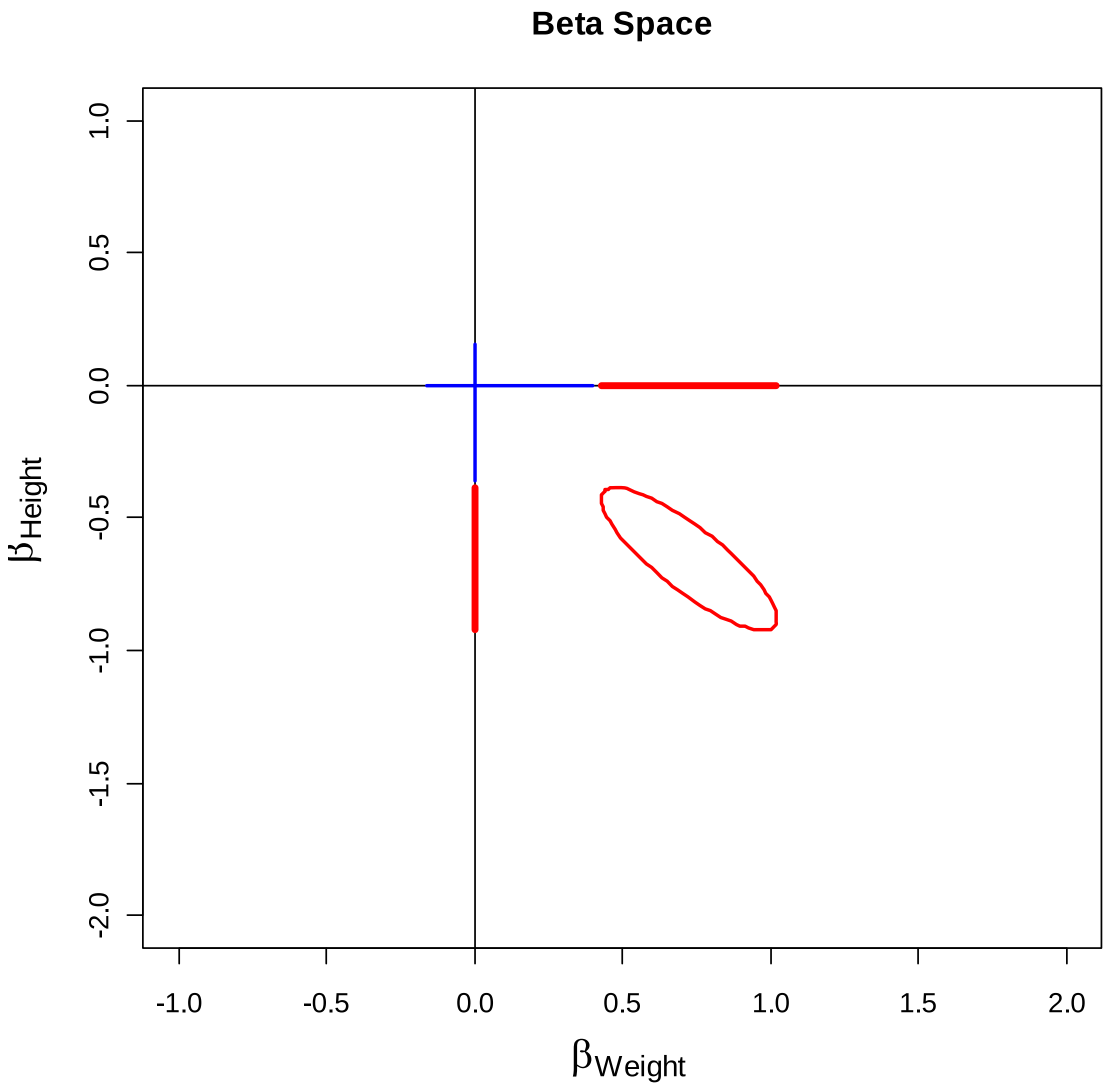
Confidence intervals



Confidence interval for weight and for height.



Confidence intervals and confidence ellipse



Blue intervals are CIs for each predictor in a simple regression.

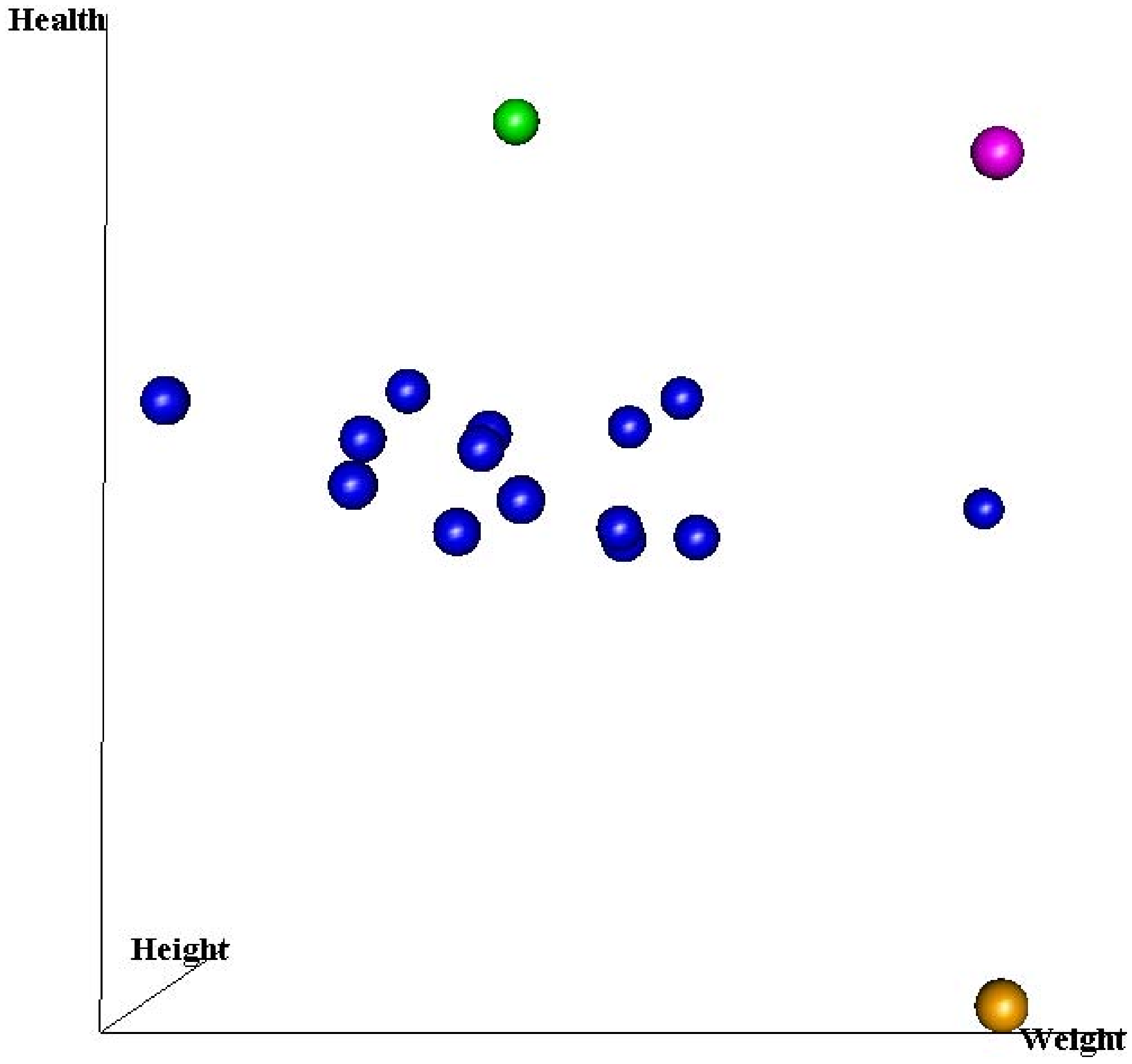
Neither simple regression yields significance yet the joint regression is highly significant.

What does this say about the use of a forward stepwise algorithm with this kind of data.

Some would call Height a *suppressor* variable for the effect of Weight.

This is an example where the traditional Venn diagrams for correlation fail: The whole is greater than the sum of its parts!

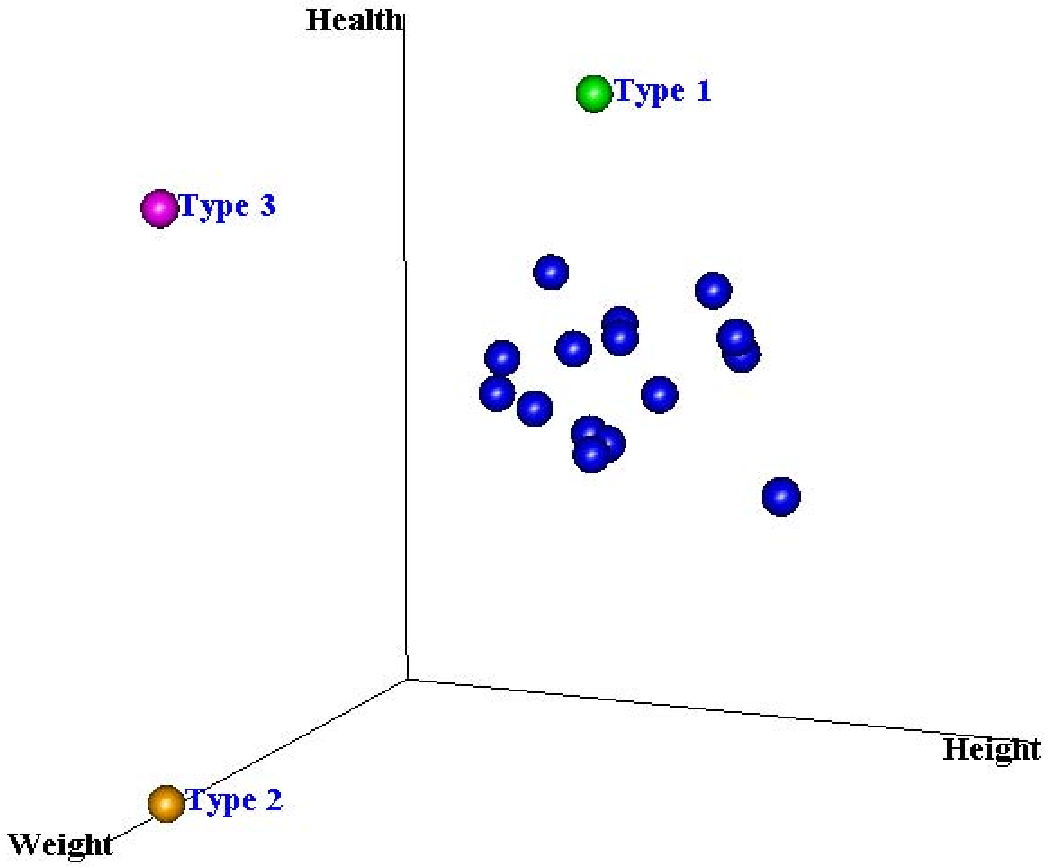
Three canonical outliers



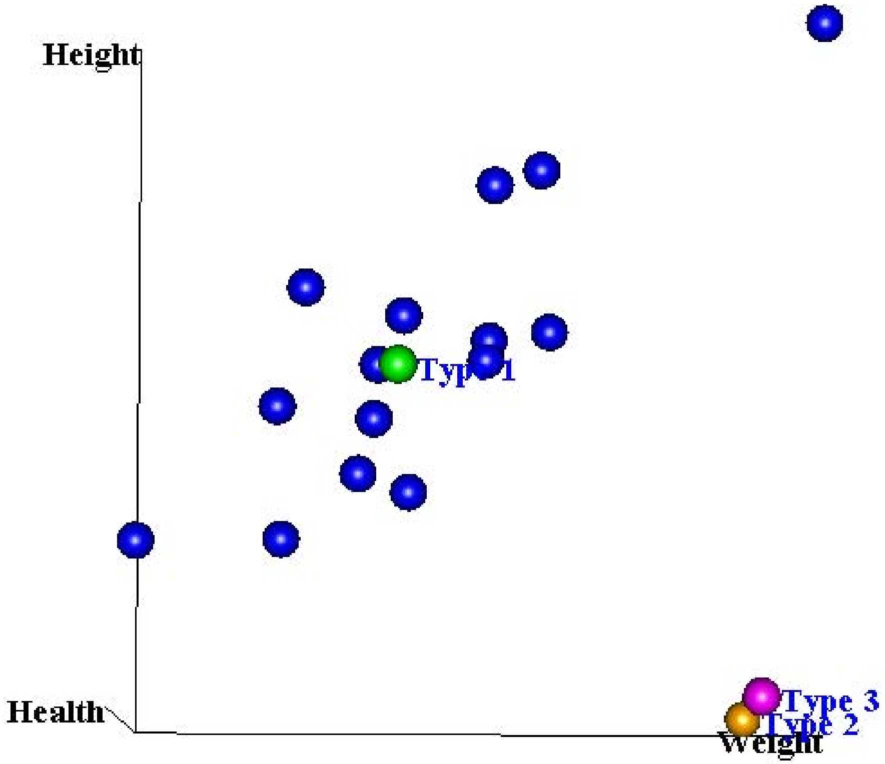
Green – Type I:
typical X, unusual Y

Orange – Type II
atypical X, conforming Y

Purple – Type III
atypical X, unusual Y

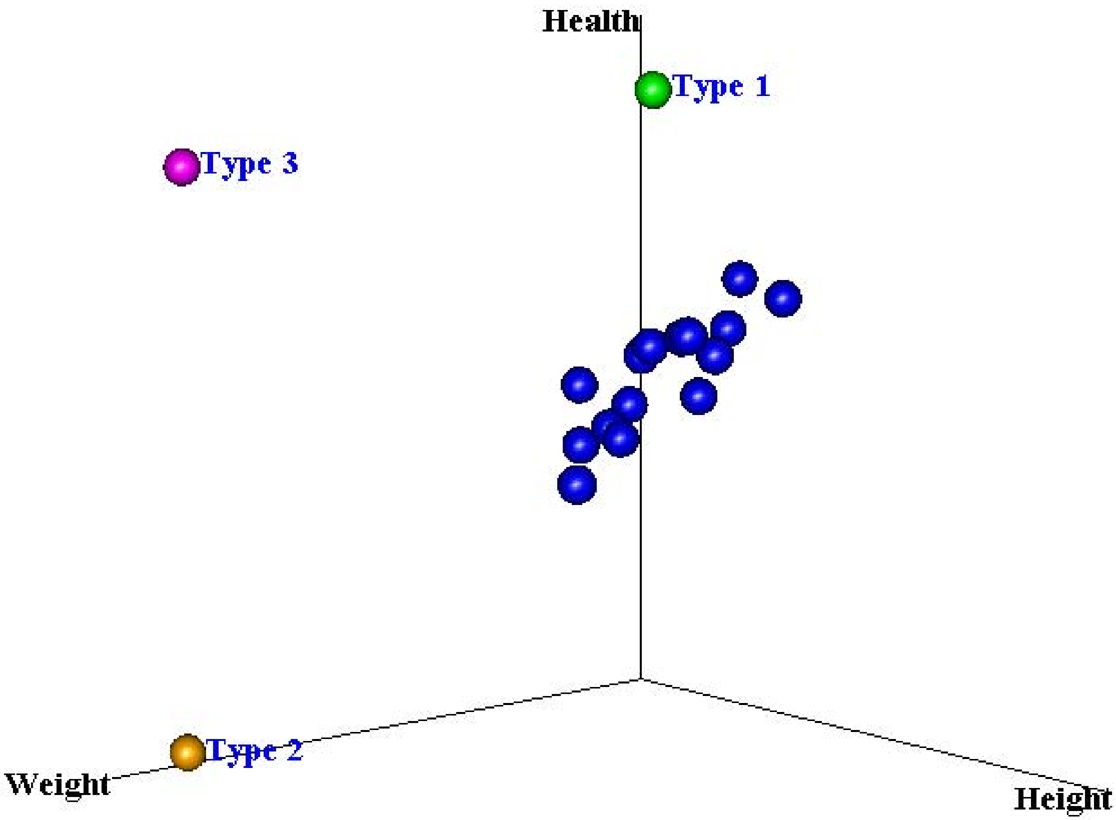

Health
Type 1
Type 3
Height
Type 2
Weight

View from above

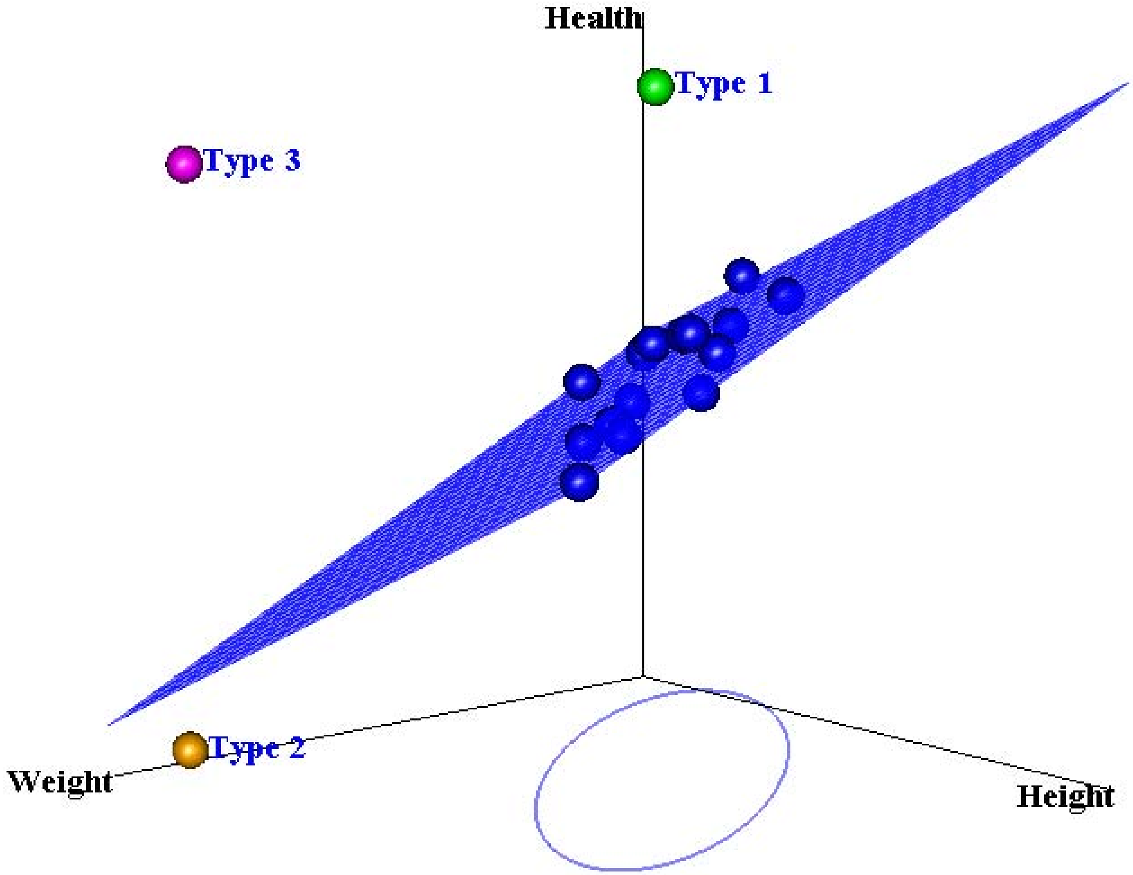


Note that orange and purple points are much farther from the data statistically than the remote blue point although the blue point is farther in Euclidean distance.

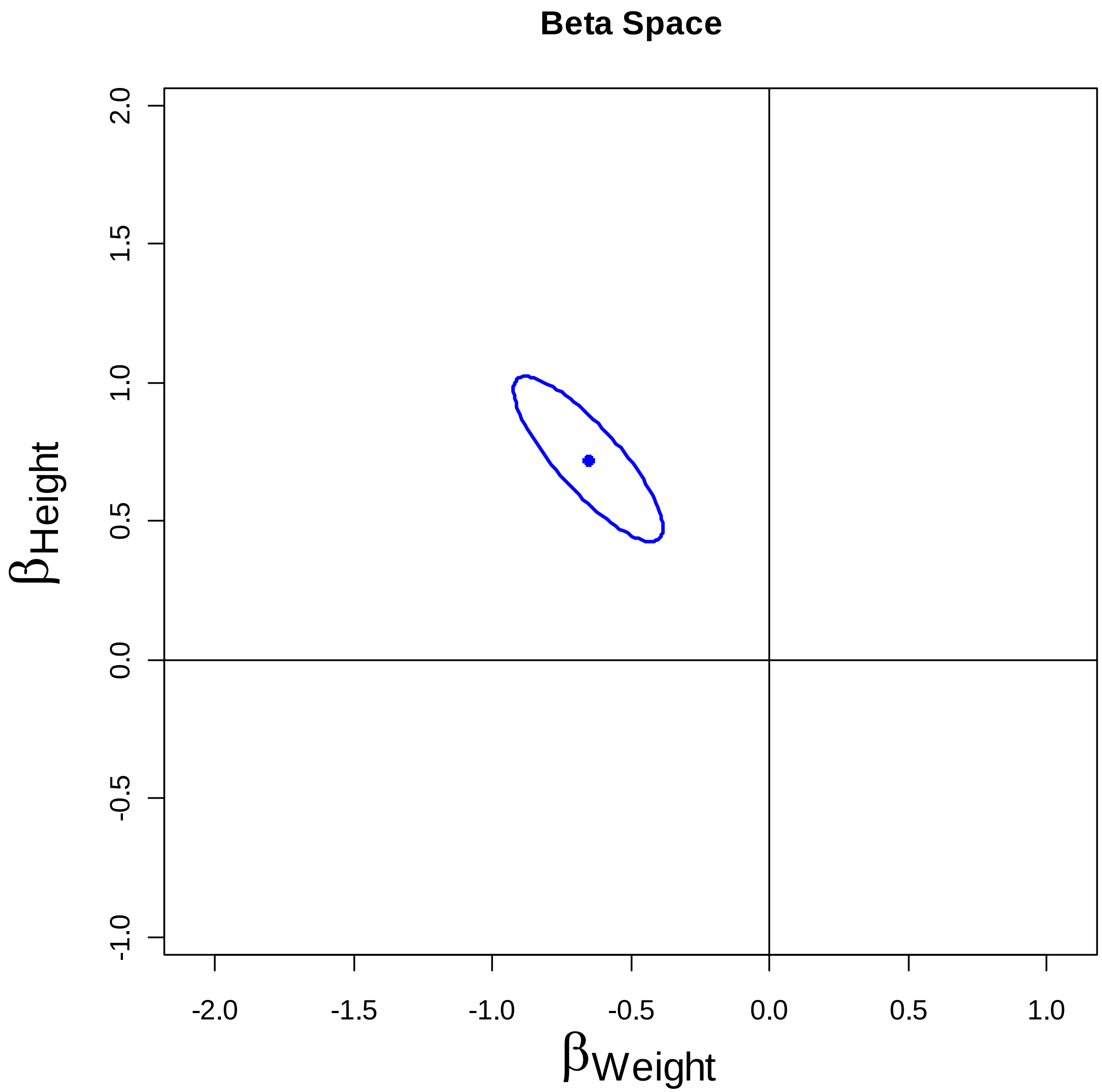
Statistical distance is measured relative to the data ellipse.


Health
Type 1
Type 3
Type 2
Weight
Height

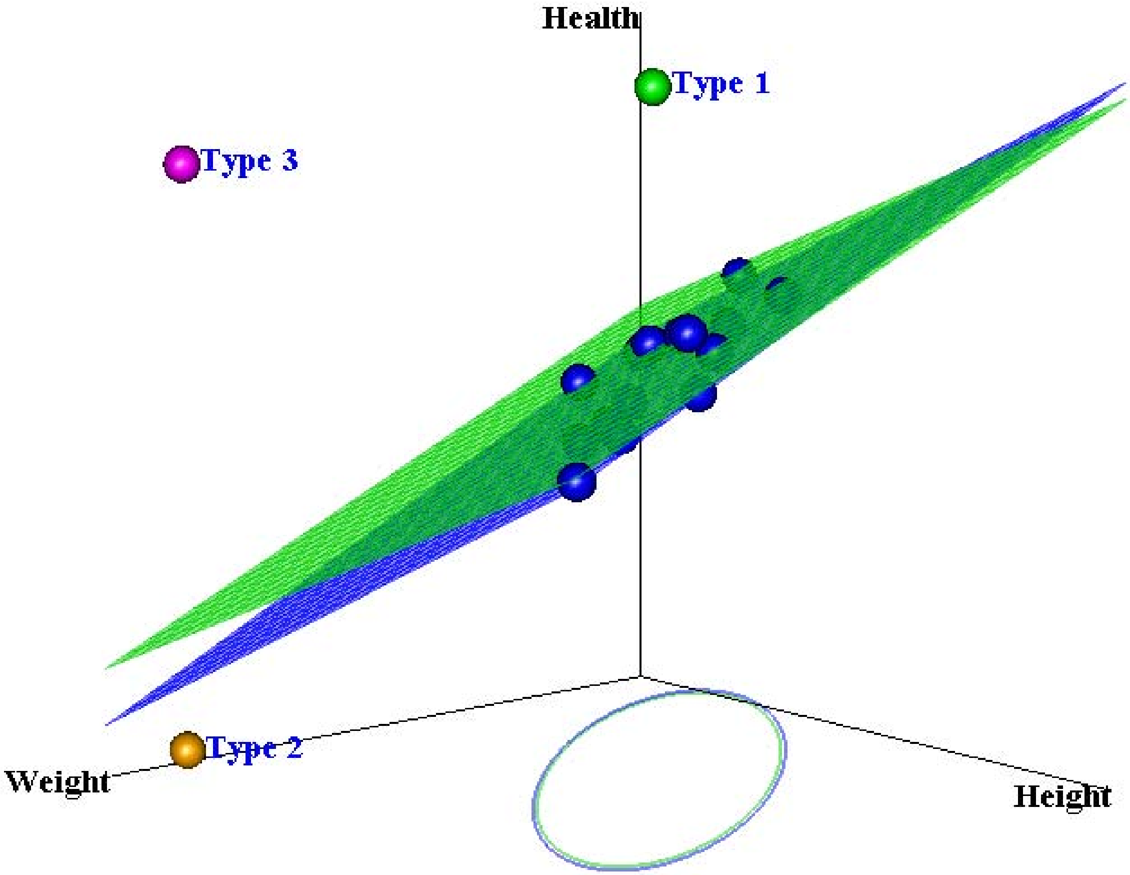
Fit without outliers:



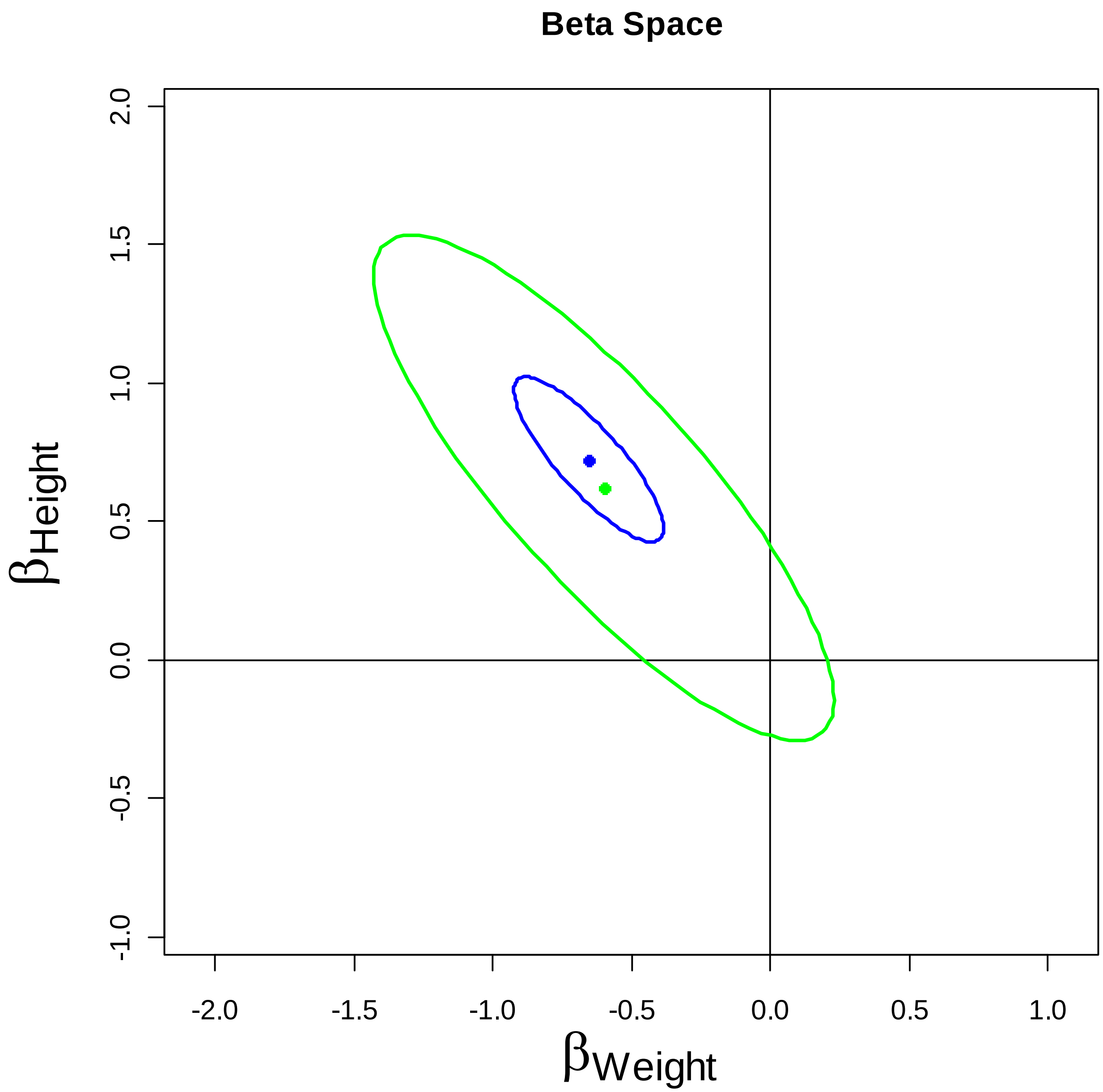
Confidence ellipse without outliers



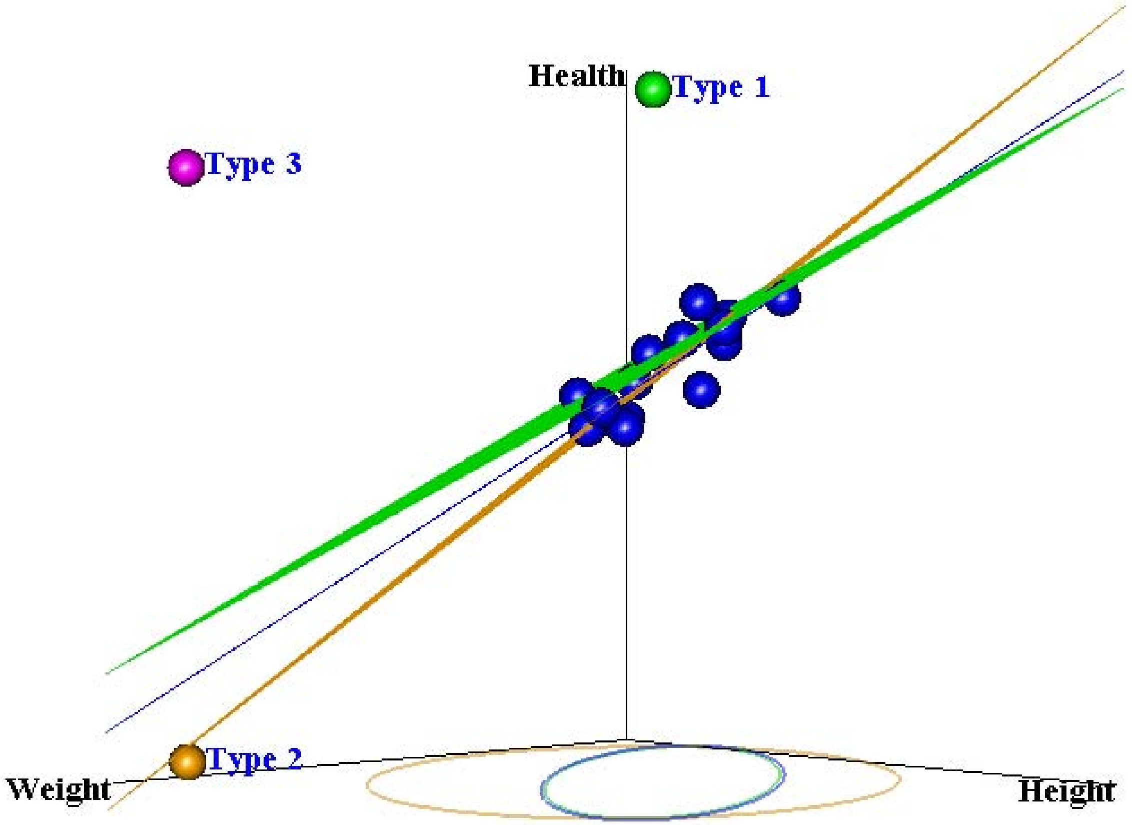
Fit with Type I outlier



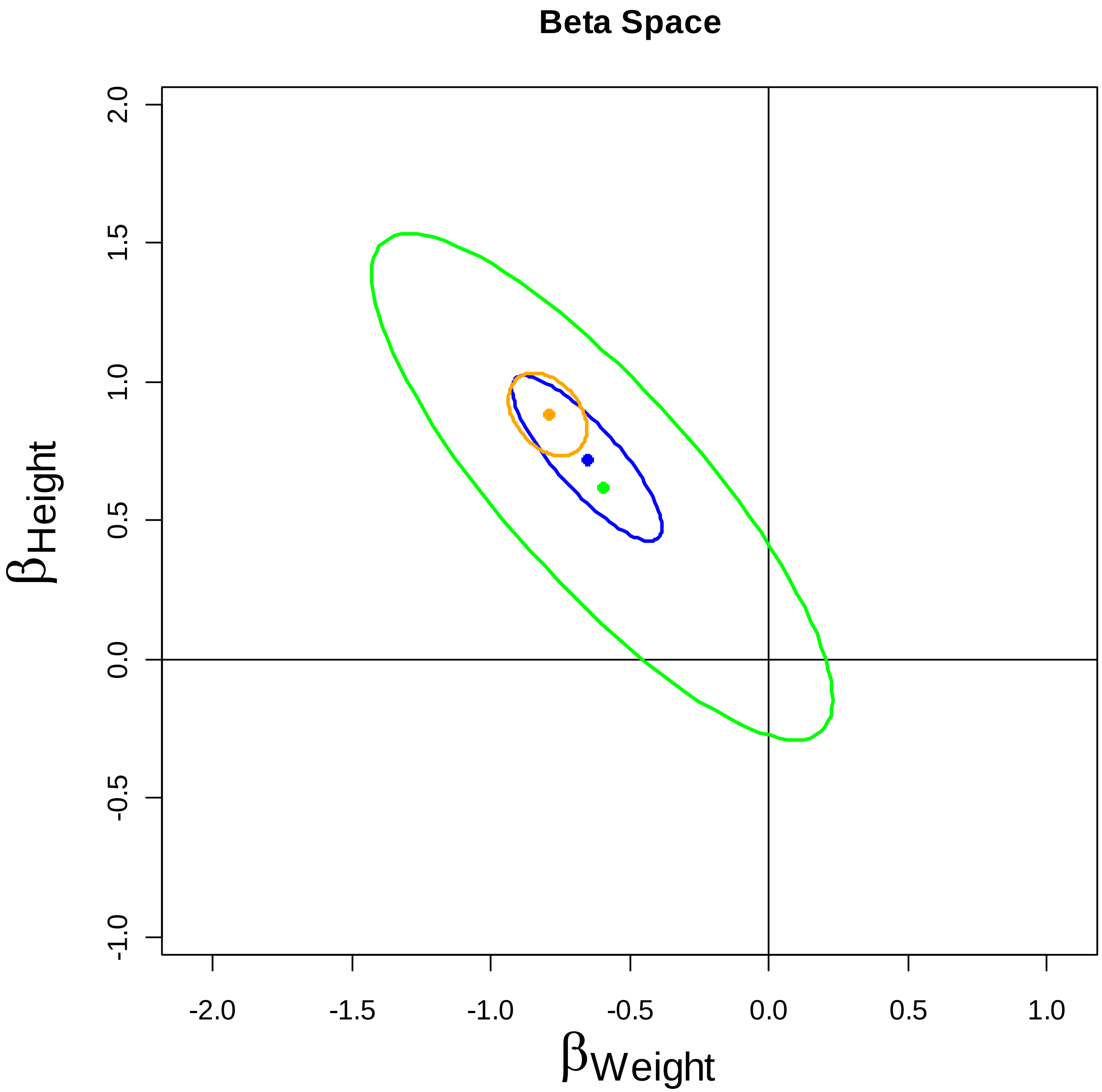
Green: Type I outlier



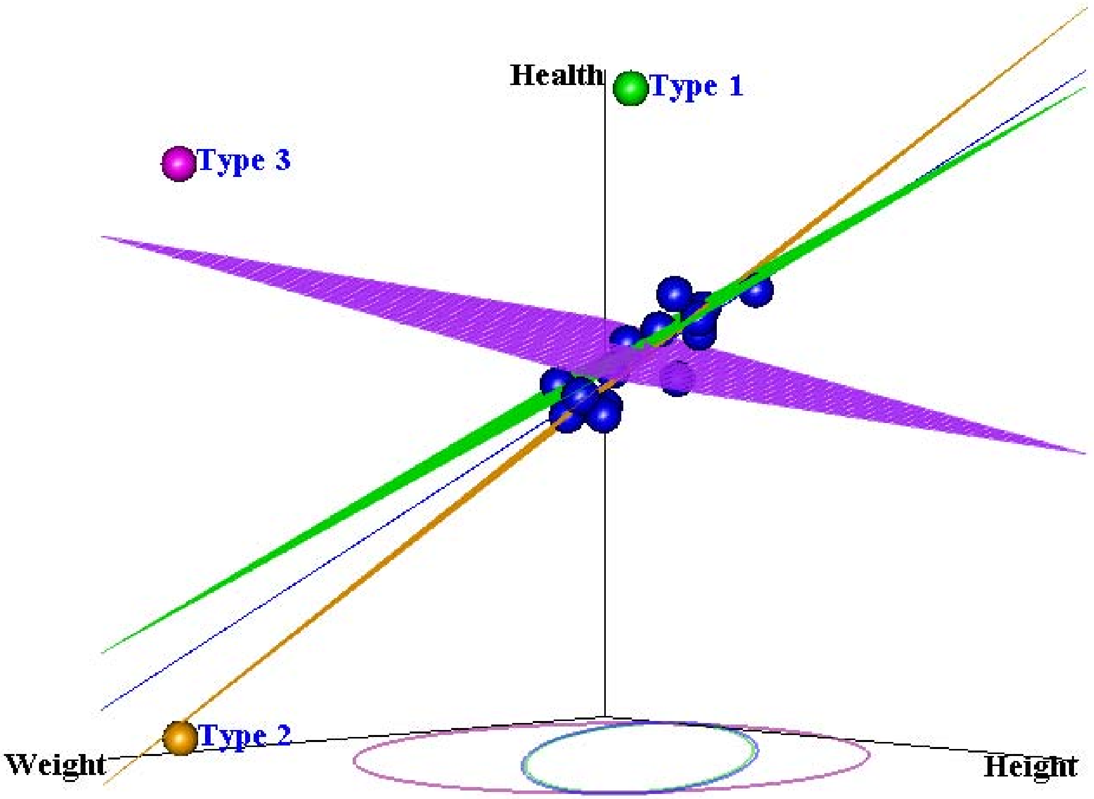
Fit with Type II outlier



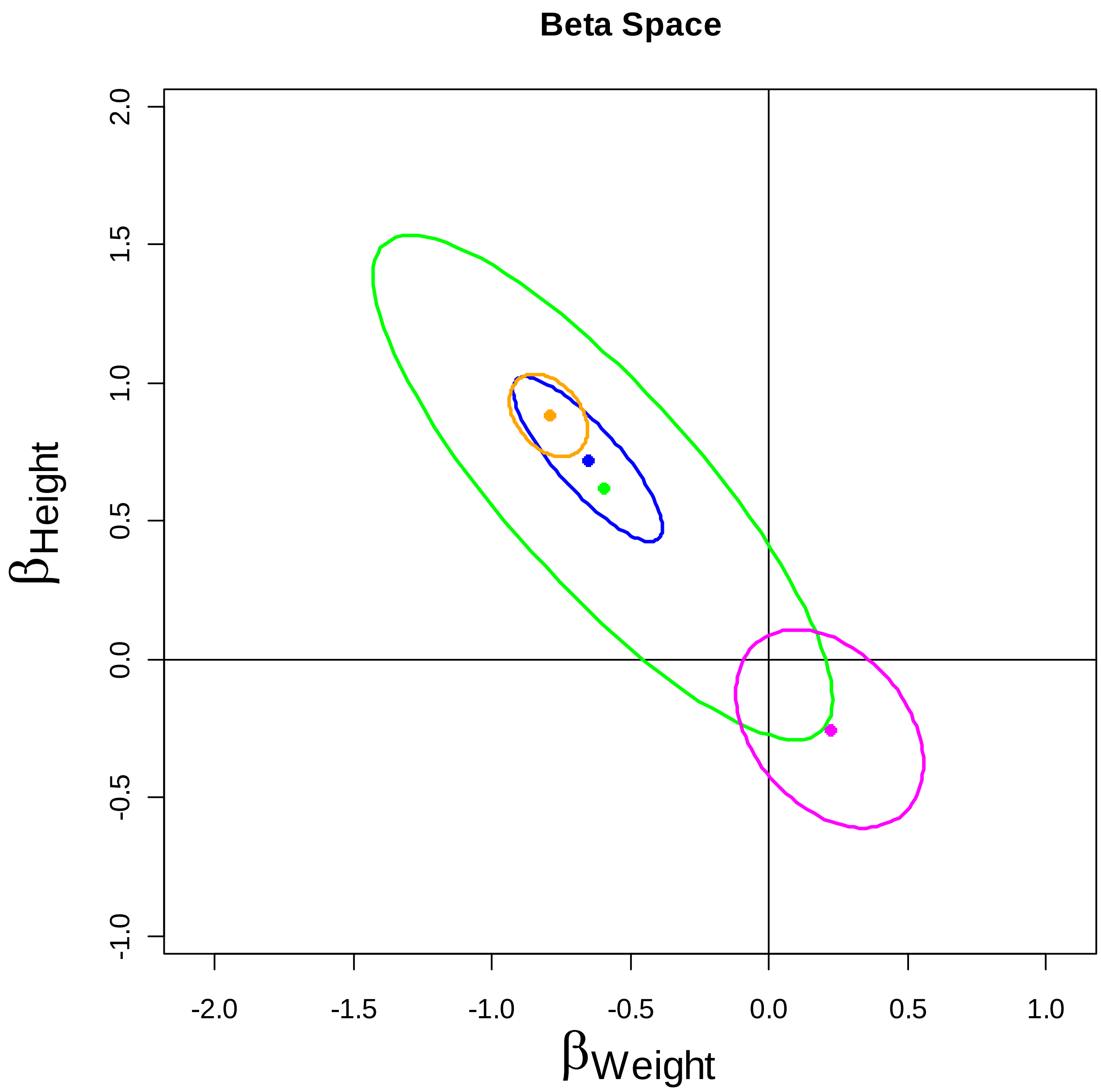
Adding CE for Type II outlier



Fit with Type III outlier



Adding CE with Type III outlier



Summary

Anatomy of a 95% confidence region (interval):

$$\hat{\beta} \oplus \sqrt{qF_{q,\nu}^{.95}} \times \frac{s_e}{\sqrt{n}} \times \sqrt{\Sigma_X^{-1}}$$

where

- q is the dimension for which we wish coverage
- ν is the degrees of freedom in the estimation of s_e
- Σ_X is the variance-covariance matrix of the predictors, n is the number of observations
- $\sqrt{\Sigma_X^{-1}}$ denotes the ellipse: $\{x : x'\Sigma_X x = 1\}$
- s_e is the standard error of regression
- $\oplus$ could just be $+$ since it denotes the addition of the vector $\hat{\beta}$ to the ellipse on the right but the use of $\oplus$ serves as a reminder that the object on the right is a transformed circle.
- The linear shadows of the confidence region on linear subspaces are also confidence regions. q can be chosen to achieve accurate coverage in the desired dimension.

$$\hat{\beta} \oplus \sqrt{qF_{q,\nu}^{.95}} \times \frac{s_e}{\sqrt{n}} \times \sqrt{\Sigma_X^{-1}}$$

Effect of invalid outliers:

Component	Type I	Type II	Type III
$\hat{\beta}$	~	~	large
s_e	increases	none	increases
$\sqrt{\Sigma_X}$ shape of data ellipse	~	stretches	stretches
$s_e \times \sqrt{\Sigma_X^{-1}}$ shape of confidence ellipse	larger	smaller	undetermined
Overall	loss of power larger p-value	false power small p-value	crazy

Notes:

At one point people focused on Type III because the other types don't affect $\hat{\beta}$ much.

But we're interested in more than unbiased estimation. We’re also interested in inference: tests of hypotheses and confidence intervals. Understanding the effect and diagnosis of Type I and Type II outliers is important for inference.

Diagnosis:

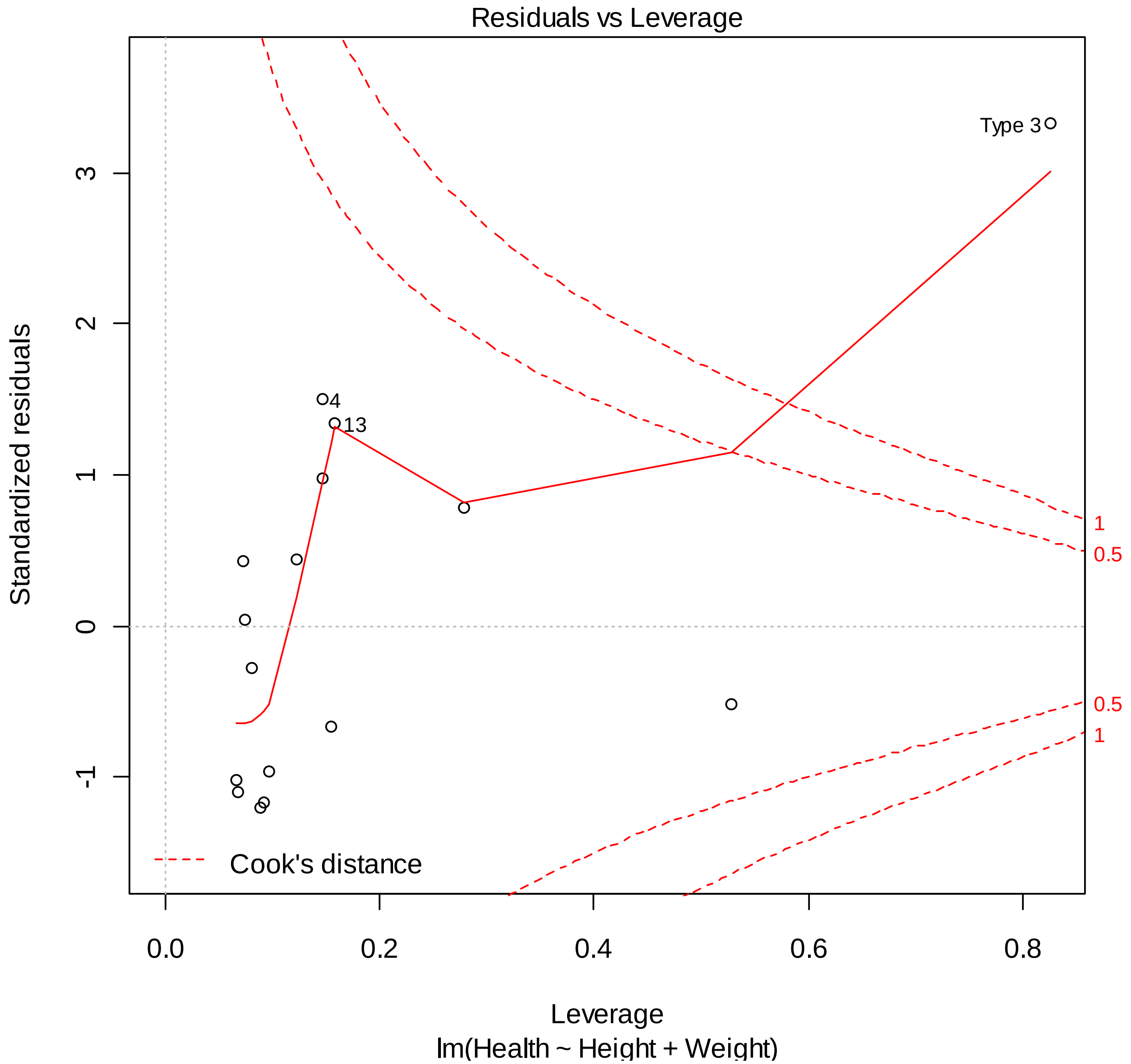
1. Normal quantile plots of residuals for normality. This can reveal outliers that produce large residual. But not all outliers produce large residuals. Type II won't and Type III might not.

2. Added variable plot: great to find outliers that matter for a particular coefficient.

3. Residual – Leverage plot: Plots standardized residual versus leverage = elliptical distance from the centre of the predictor data ellipse.

Residual vs Leverage plot:

```
> plot( fit.mult )
```

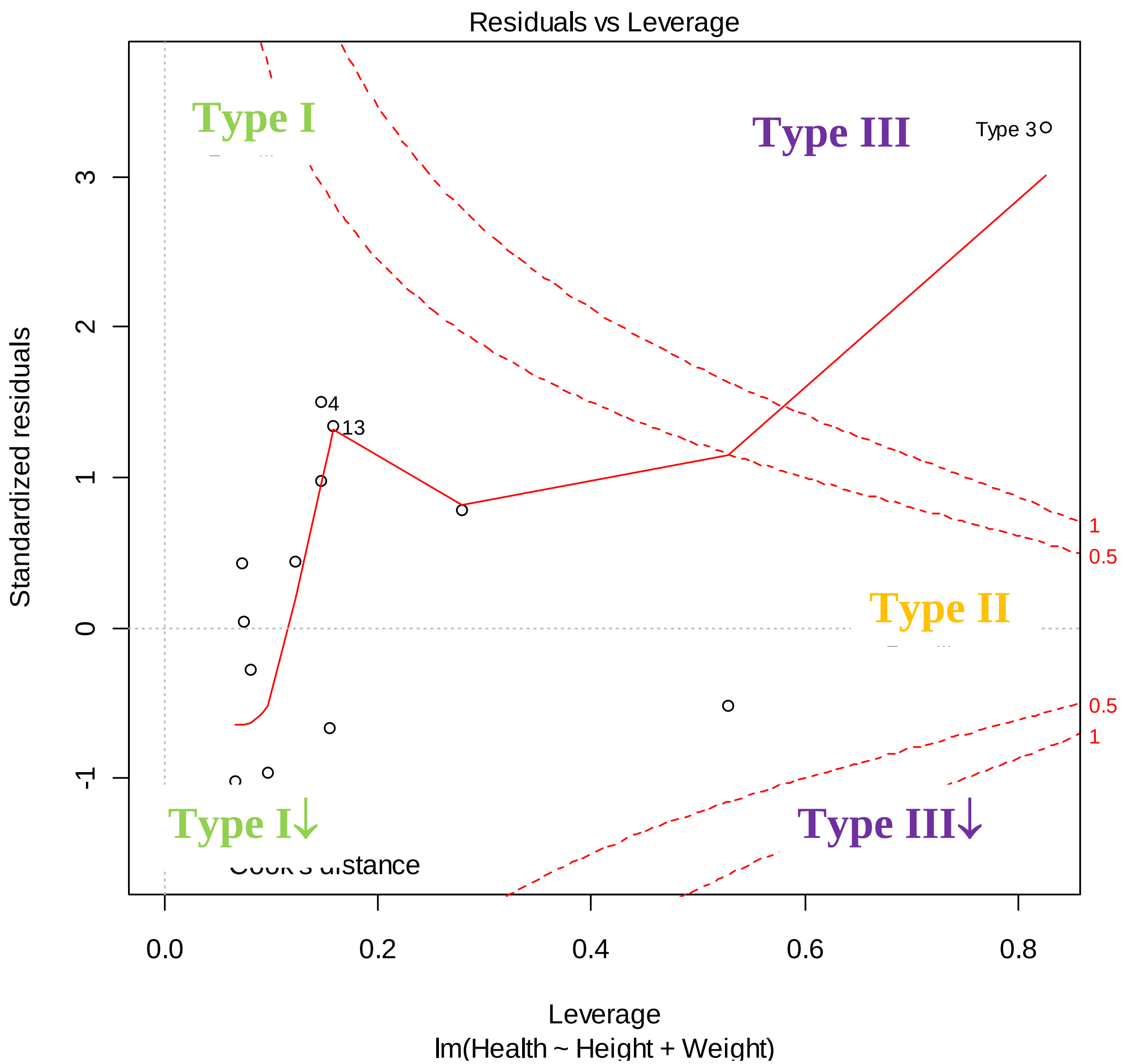


Vertical axis:
Standardized residuals

Horizontal axis
Leverage ≈ distance from centre of predictor data ellipsoid

Residual vs Leverage plot: where to look for what:

```
> plot( fit.mult )
```



Vertical axis:

Standardized residuals

= poor fit compared to pattern in the rest of the data

⇒ Type I or Type III

Horizontal axis

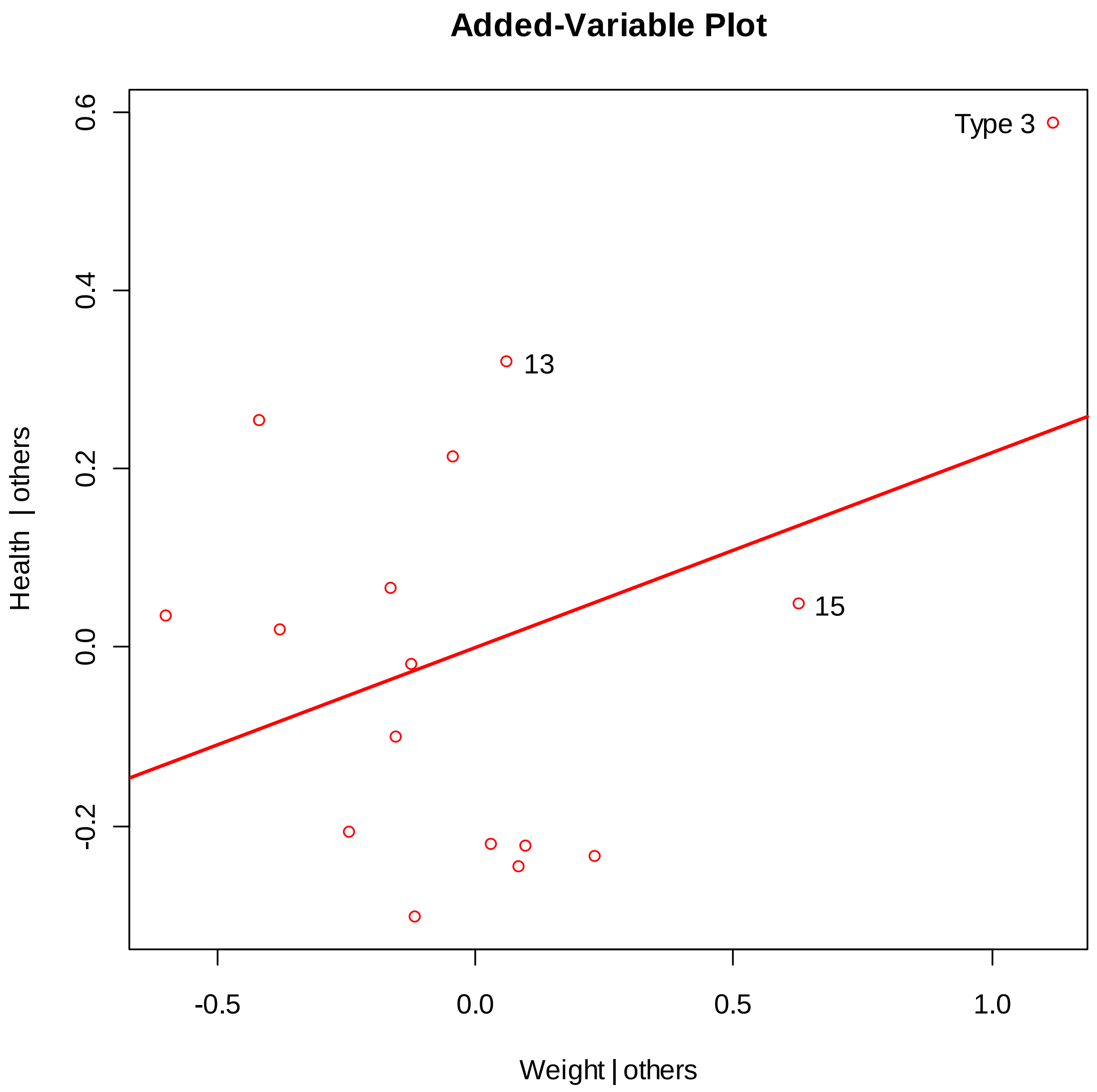
Leverage ≈ distance from centre of predictor data ellipsoid

= atypical Xs

= high potential influence on $\hat{\beta}$

⇒ Type II or Type III

Added variable plot for Weight:



Vertical axis:
Residual of Health regressed on all but Weight – i.e. Height –
= portion of Health "not explained by Height"

Horizontal axis:
Residual of Weight regression on all but Weight
= portion of Weight not explained by Height
= degree of over/underweight?
= residual from propensity score

Facts:

1) slope of AVP = $\hat{\beta}$ in mult. reg.
2) SE of AVP ≈ SE in mult. reg.
3) R^2 in AVP = partial R^2 in mult. reg.
4) CI for $\hat{\beta}$ in AVP ≈ CI for $\hat{\beta}$ in mult.reg.

The AVP provides the best insight into the relationship between Y and X in a multiple regression.

Interaction with continuous predictors

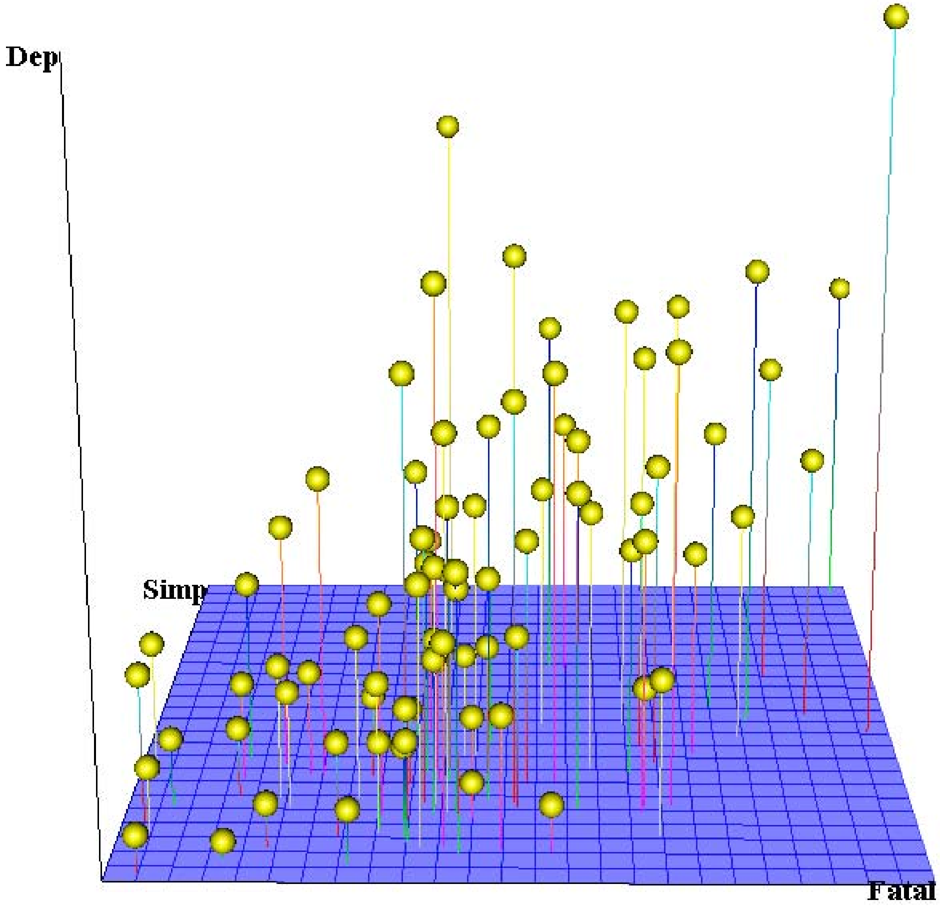
Note: Interaction is often confused with collinearity.

Collinearity refers to associations among predictors: i.e. the extent to which the predictor data ellipse is tilted and eccentric. It is not related to Y

Interaction refers to a situation in which the relationship between Y and the predictor variables is not 'additive',i.e. the effect of some variable depends on the levels of the other variable. It has nothing to do with the relationships among the Xs – only in the way they affect Y

Interaction or collinearity can exist with or without the other. The presence of either does not even suggest the likely presence of the other.

Ginzburg depression data:



```
> library(car)
> data( Ginzburg )
> dd <- Ginzburg
```

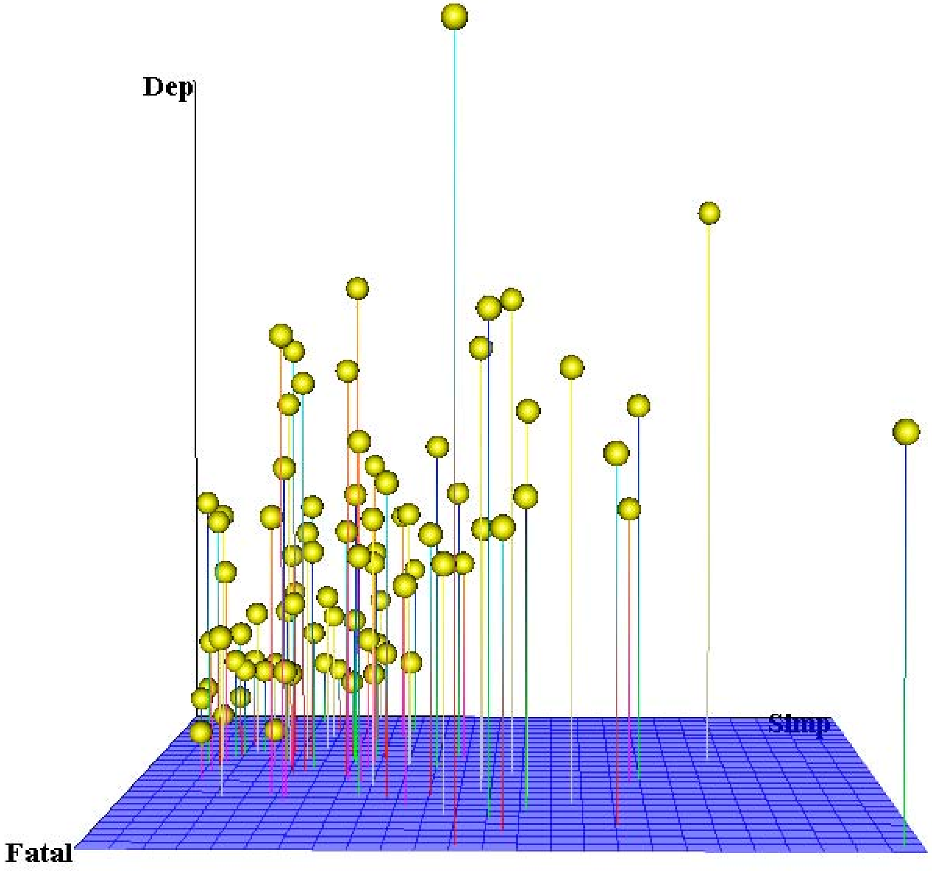
Shows 'Depression' measured by Beck scale regressed on Fatalism and Complexity (flipped and renamed Simplicity).

Subjects were from a clinical sample of depressed patients and non-psychiatric patients

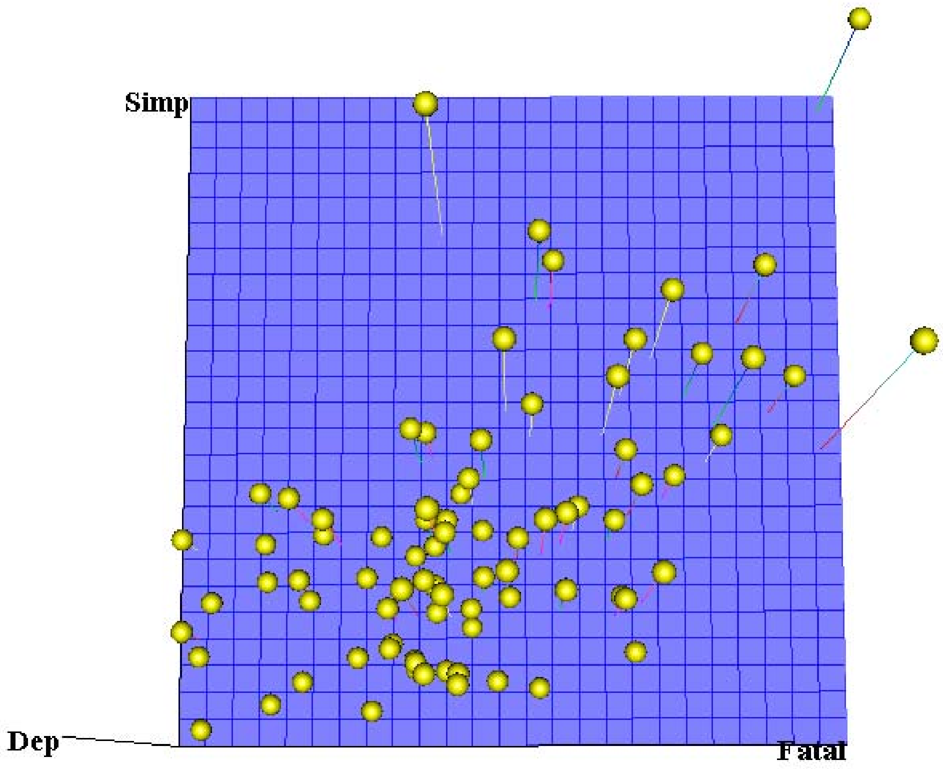
Variable shown in graph are adjusted for a number of other psychological variables. You can think of it as a 3D AVP with two predictors.

We want to see what is predicted by Fatalism and Simplicity that is not already explained by other variables.

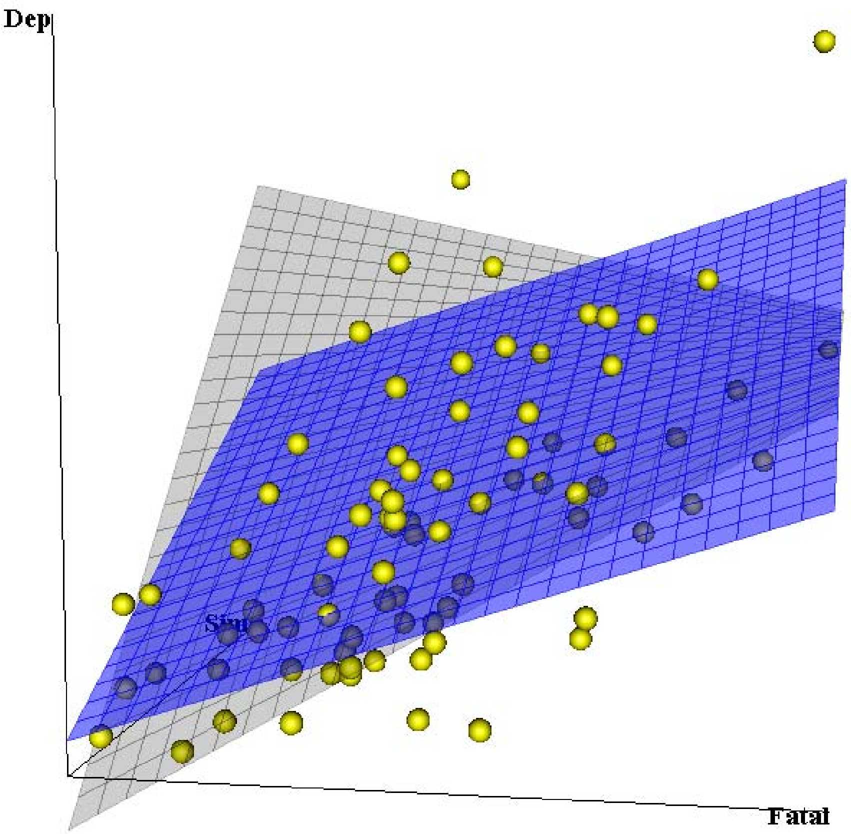
Rotating the data:



View from above



Additive model and interaction model I:



Additive model:

```
>   fit.add <- lm( Dep
          ~ Simp + Fatal, dd)
```

Interaction model (multiplicative interaction):

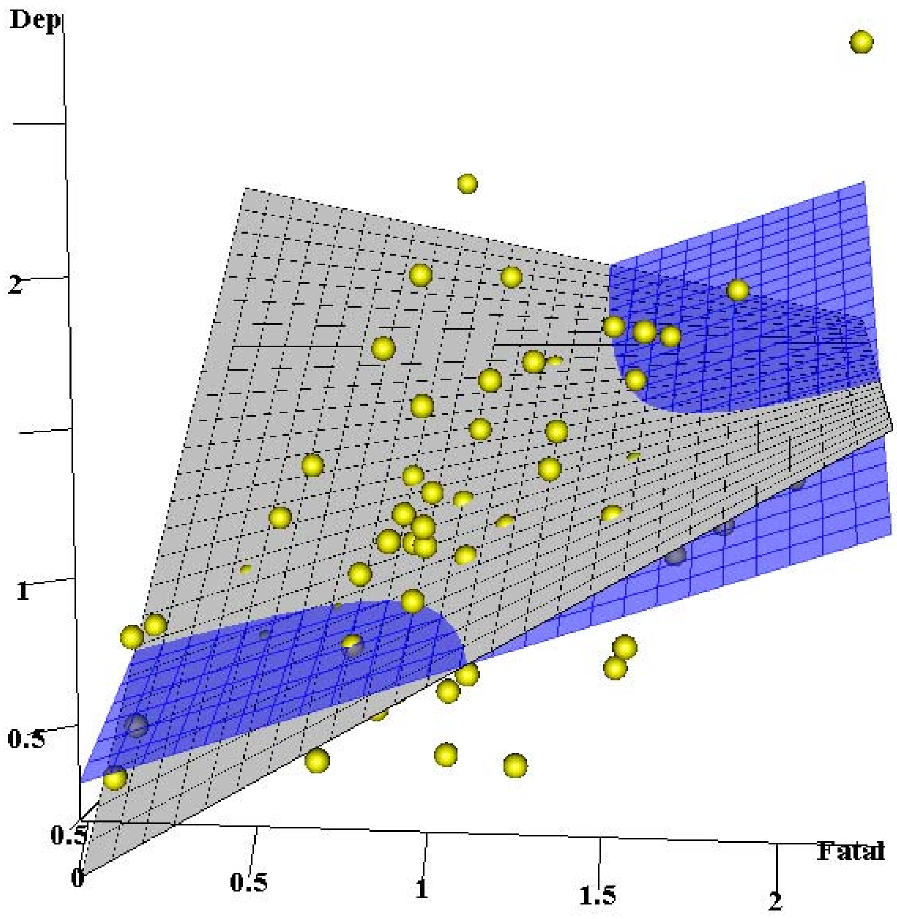
```
>  fit.int <- lm( Dep
          ~ Simp * Fatal, dd)
```

Note: Effect of Fatalism stronger with low Simplicity than with high Simplicity. Ditto the other way around.

We suspected a possible ceiling effect which motivated viewing the data in 3D

```
Additive model:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   0.2492     0.1054   2.365 0.020501 *
Simp          0.3663     0.1004   3.649 0.000471 ***
Fatal         0.3845     0.1004   3.829 0.000256 ***
```

Additive model and interaction model II:



Additive model:

```
>   fit.add <- lm( Dep
          ~ Simp + Fatal, dd)
```

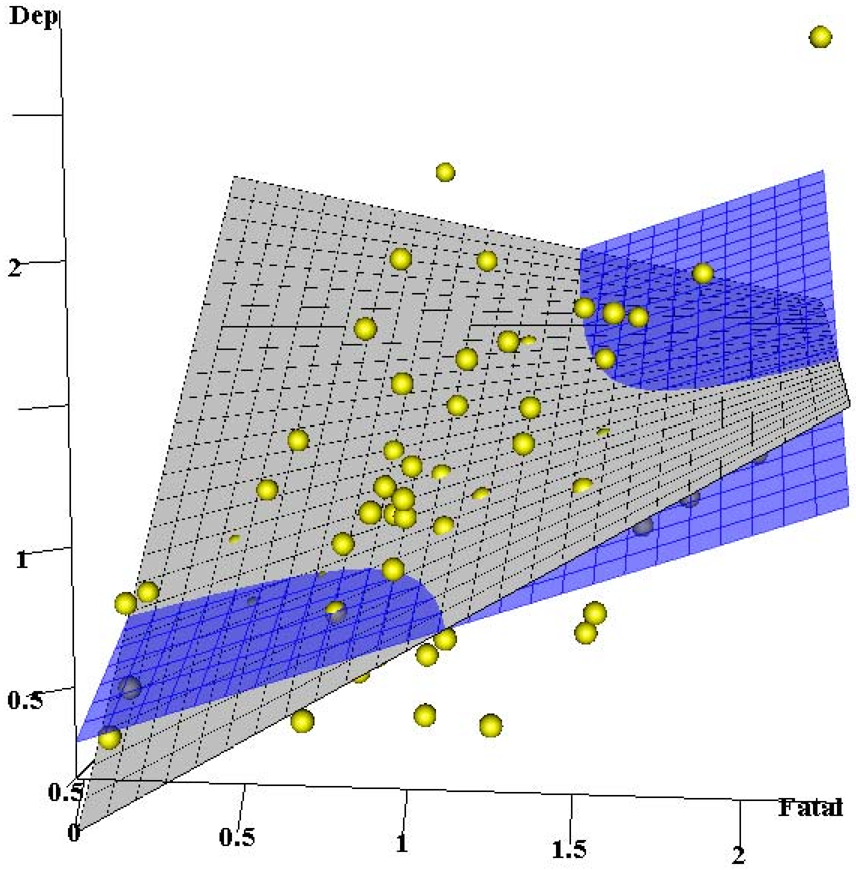
Interaction model (multiplicative interaction):

```
>  fit.int <- lm( Dep
          ~ Simp * Fatal, dd)
```

Note that main effects of Simp and Fatal are estimates of the slopes over the 0 point. Here that happens to be the 'origin' in the graph but that is not necessarily the case.

```
Model with interaction:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -0.1366     0.2023  -0.676 0.501319
Simp          0.7736     0.2083   3.714 0.000381 *** [slope over origin]
Fatal         0.7218     0.1811   3.987 0.000150 *** [slope over origin]
Simp:Fatal   -0.3168     0.1430  -2.216 0.029607 *    [Is there a twist?]
```

Additive model and interaction model III:



To estimate the slope at other points (specific effects) we can use a Wald test (equivalent of ESTIMATE or CONTRASTS in PROC GLM in SAS)

Use model formula:

$$E(Dep) = \beta_0 + \beta_{Simp} Simp + \beta_{Fatal} Fatal$$
$$+ \beta_{S\times F} Simp \times Fatal$$

Differentiate with respect to *Fatal* to get the specific effect of *Fatal:*

$$\frac{\Delta E(Dep)}{\Delta Fatal} = \beta_{Fatal} + \beta_{S\times F} Simp$$

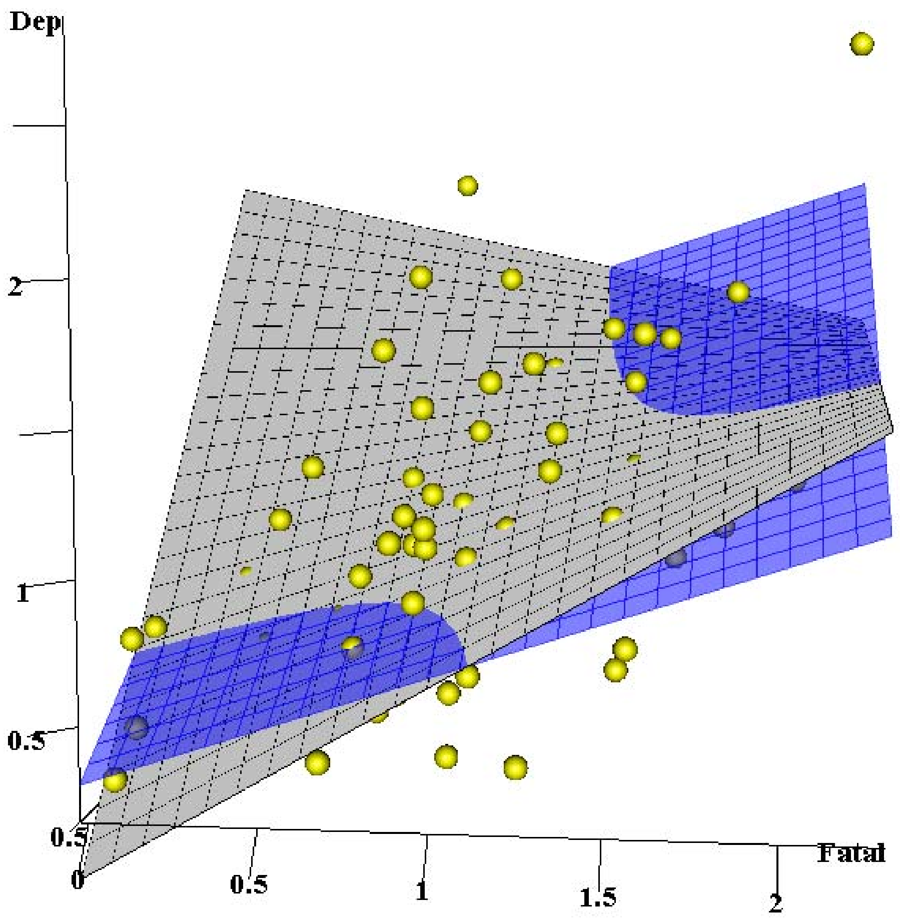
So to estimate this when $Simp = 2$ we need to multiply the coefficient vector:

```
> coef( fit.int)
(Intercept)         Simp       Fatal  Simp:Fatal
 -0.1366333   0.7736409   0.7218125  -0.3168196
```

by the matrix:

```
>  L <- rbind("Fatal|Simp=2" = c( 0,0,1,2) )
>  L
             [,1] [,2] [,3] [,4]
Fatal|Simp=2    0    0    1    2
```

Additive model and interaction model IV:



Using the 'wald' command in 'fun.R' produces an the usual output plus a confidence interval. There is also and overall F test which is useful if our are estimating more than one linear function of the coefficients.

```
> wald(fit.int, list( "Specific effect of Fatal" = L))
                         numDF denDF   F.value p.value
Specific effect of Fatal     1    78 0.2829107 0.59631

               Estimate Std.Error DF  t-value p-value Lower 0.95 Upper 0.95
  Fatal|Simp=2 0.088173  0.165773 78 0.531893 0.59631  -0.241855   0.418201
```

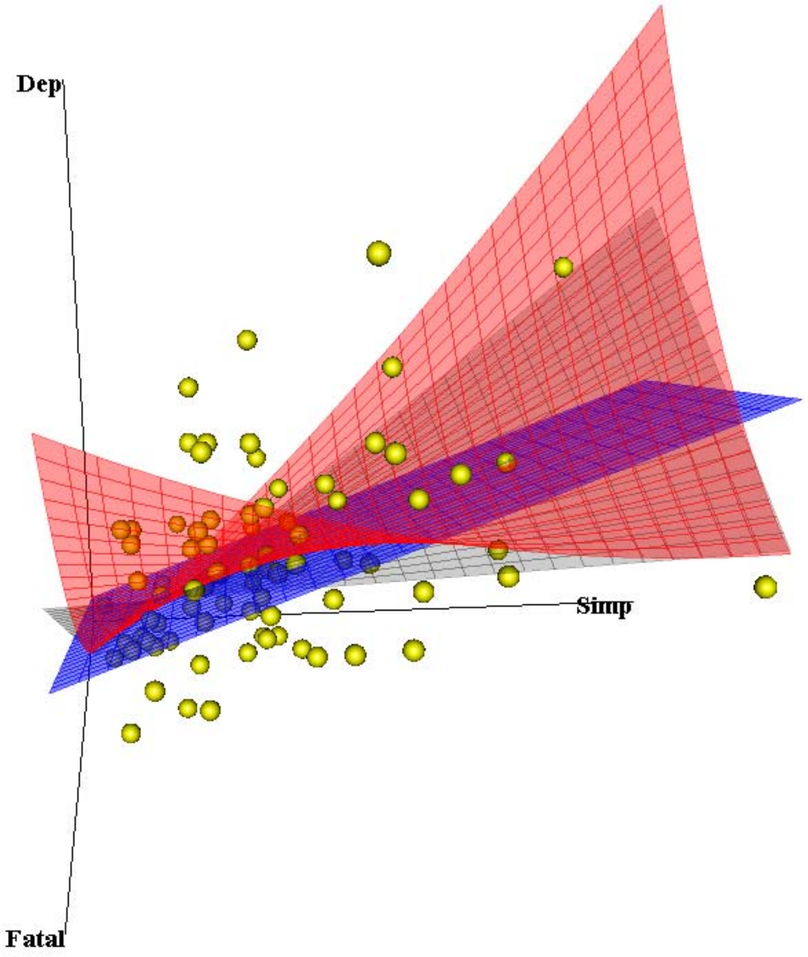
Estimating many slopes at once

```
> L <- rbind(
+            "Fatal|Simp=0" = c( 0,0,1,0),
+            "Fatal|Simp=1" = c( 0,0,1,1),
+            "Fatal|Simp=2" = c( 0,0,1,2),
+            "Simp|Fatal=0" = c( 0,1,0,0),
+            "Simp|Fatal=1" = c( 0,1,0,1),
+            "Simp|Fatal=2" = c( 0,1,0,2))
> L
             [,1] [,2] [,3] [,4]
Fatal|Simp=0    0    0    1    0
Fatal|Simp=1    0    0    1    1
Fatal|Simp=2    0    0    1    2
Simp|Fatal=0    0    1    0    0
Simp|Fatal=1    0    1    0    1
Simp|Fatal=2    0    1    0    2
> wald( fit.int, list( "Specific effects" = L))
                 numDF denDF  F.value p.value
Specific effects     3    78 22.75861 <.00001

               Estimate Std.Error DF  t-value p-value Lower 0.95 Upper 0.95
  Fatal|Simp=0 0.721813  0.181053 78 3.986748 0.00015   0.361364   1.082261
  Fatal|Simp=1 0.404993  0.098438 78 4.114213 0.00010   0.209019   0.600967
  Fatal|Simp=2 0.088173  0.165773 78 0.531893 0.59631  -0.241855   0.418201
  Simp|Fatal=0 0.773641  0.208308 78 3.713929 0.00038   0.358932   1.188350
  Simp|Fatal=1 0.456821  0.106172 78 4.302664 0.00005   0.245450   0.668193
  Simp|Fatal=2 0.140002  0.141540 78 0.989133 0.32566  -0.141782   0.421786
```

Notes:

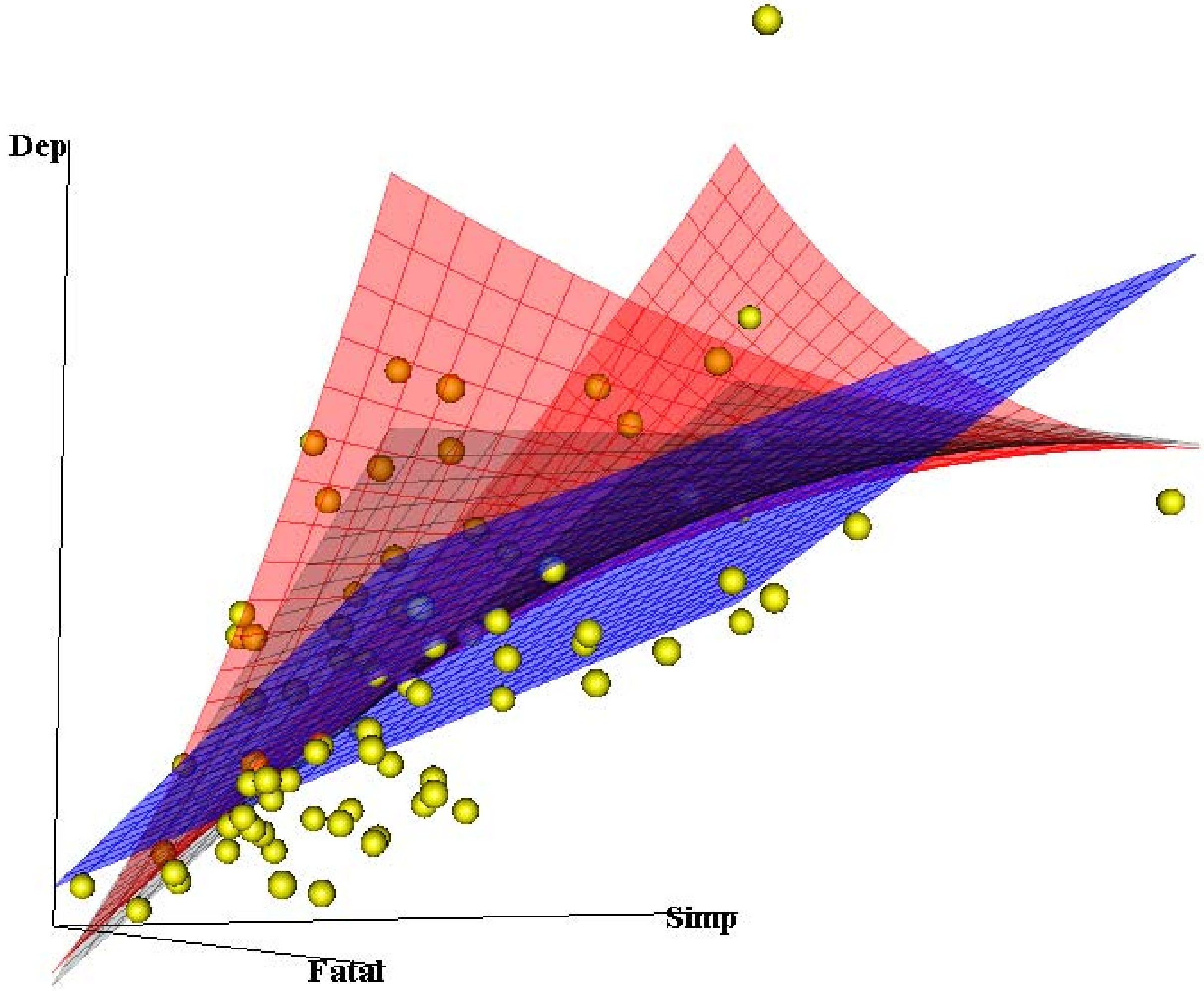
1) **SE gets larger as you move away from the centre of the data.**
2) **Slopes get flatter when the other variable is at the top of its range.**
3) **The overall F test has 3 df – not 6 – and is a test of the null hypothesis that the surface is horizontal, i.e. no regression effects**

Interaction model with quadratic surface: (rotation invariant)

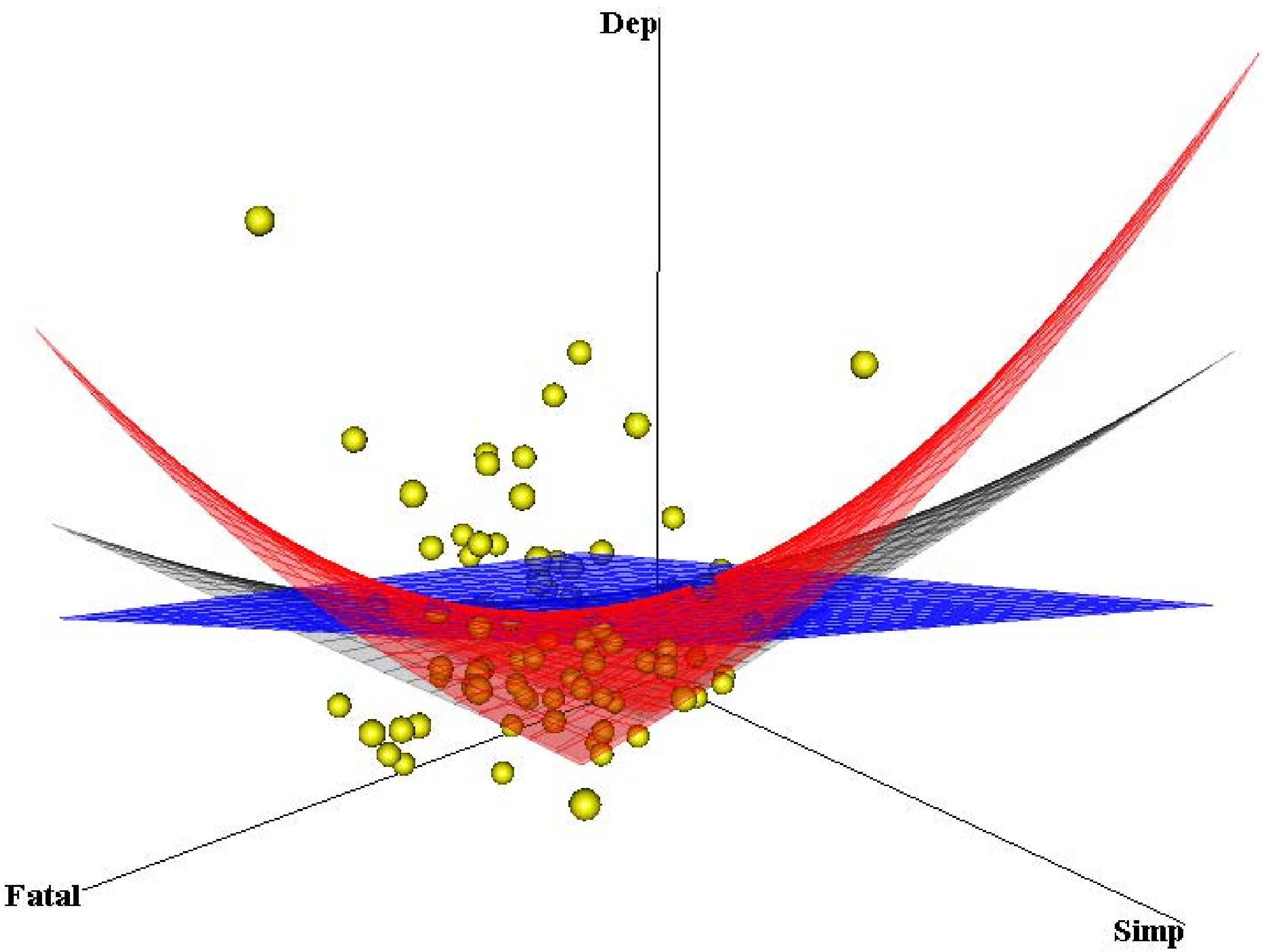


```
fit.quad <- lm(
 Dep ~ Simp * Fatal
   + I(Simp^2)
   + I(Fatal^2), dd)
```

```
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.10865    0.20687  -0.525   0.6010
Simp         0.82039    0.29702   2.762   0.0072 **
Fatal        0.55247    0.28391   1.946   0.0554 .
I(Simp^2)    0.08172    0.15374   0.532   0.5966
I(Fatal^2)   0.20927    0.18663   1.121   0.2657
Simp:Fatal  -0.55366    0.27855  -1.988   0.0505 .
```

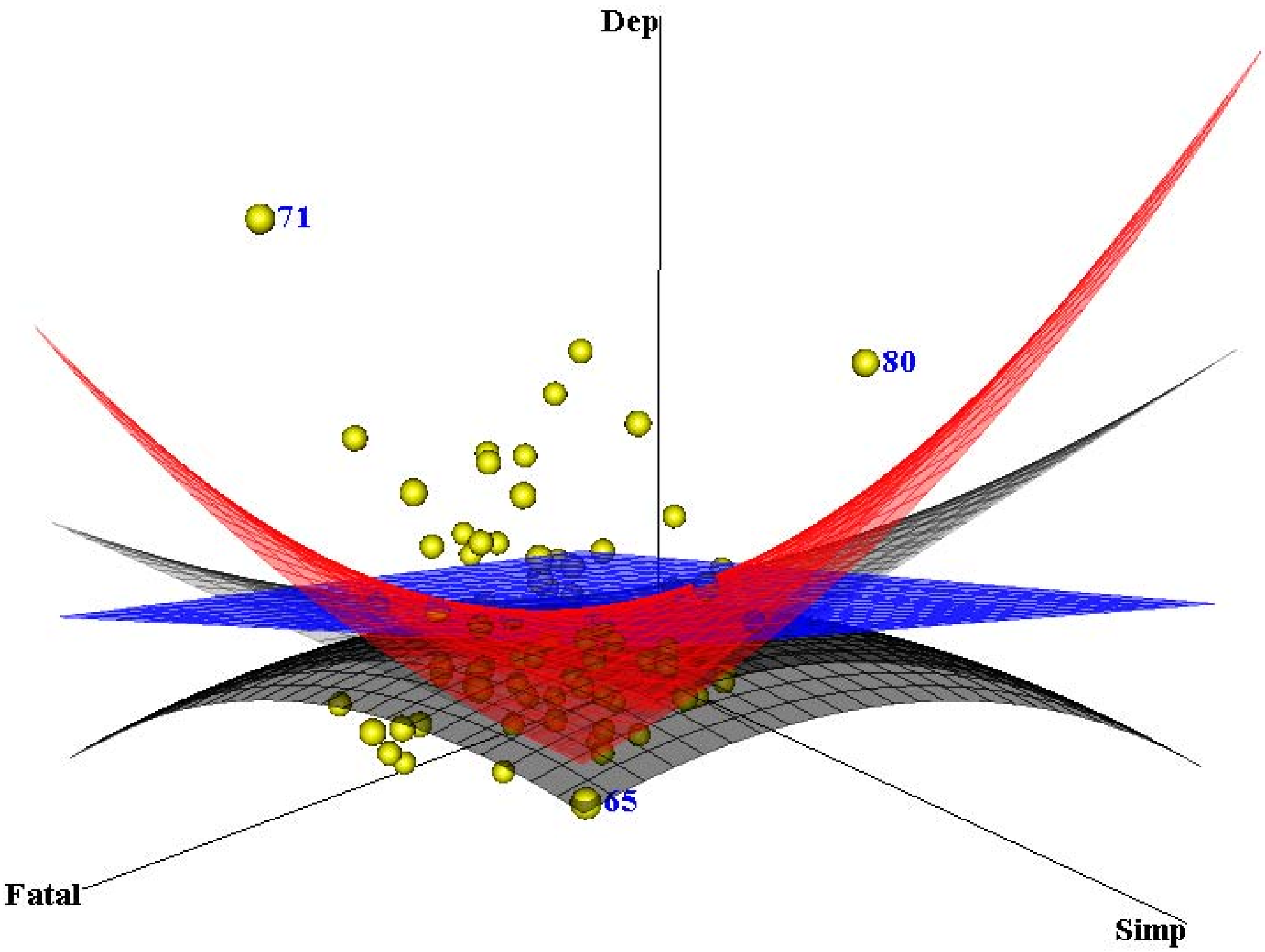


One observation seems to be pulling down on 'nose' of surface.



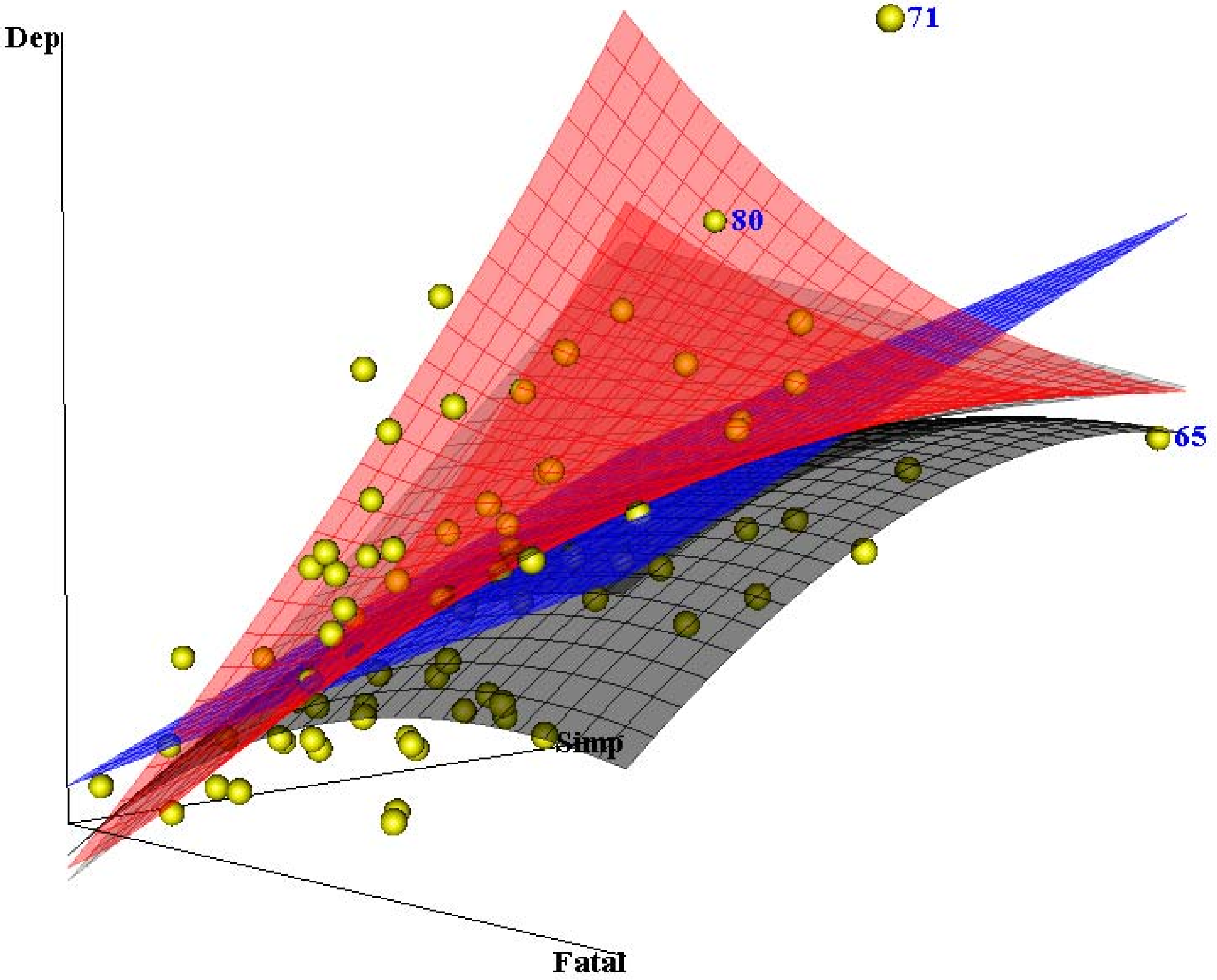
Two observations seem to be pu
wings.

What happens if we drop these o



Identifying the points and checking data revealed that one point '71' was misclassified. The study intended to cover 'reactive depression'. Patient 71 suffered from endogenous depression and was far more depressed than what would be predicted by this model.

Removing the 3 points (only 71 was removed in the final analysis) produces the dark gray regression surface. Removing only 3 points results in a dramatic change in the model.


Dep
71
80
65
Simp
Fatal

Interaction with a categorical variable

Introducing a hierarchical data set:

Subsample from U.S. public and Catholic schools studied in the 1982 High School and Beyond (HS&B) Survey.

```
> hs <- read.csv( "http://www.math.yorku.ca/~georges/Data/hs.csv")
```

40 schools: 19 Public and 21 Catholic

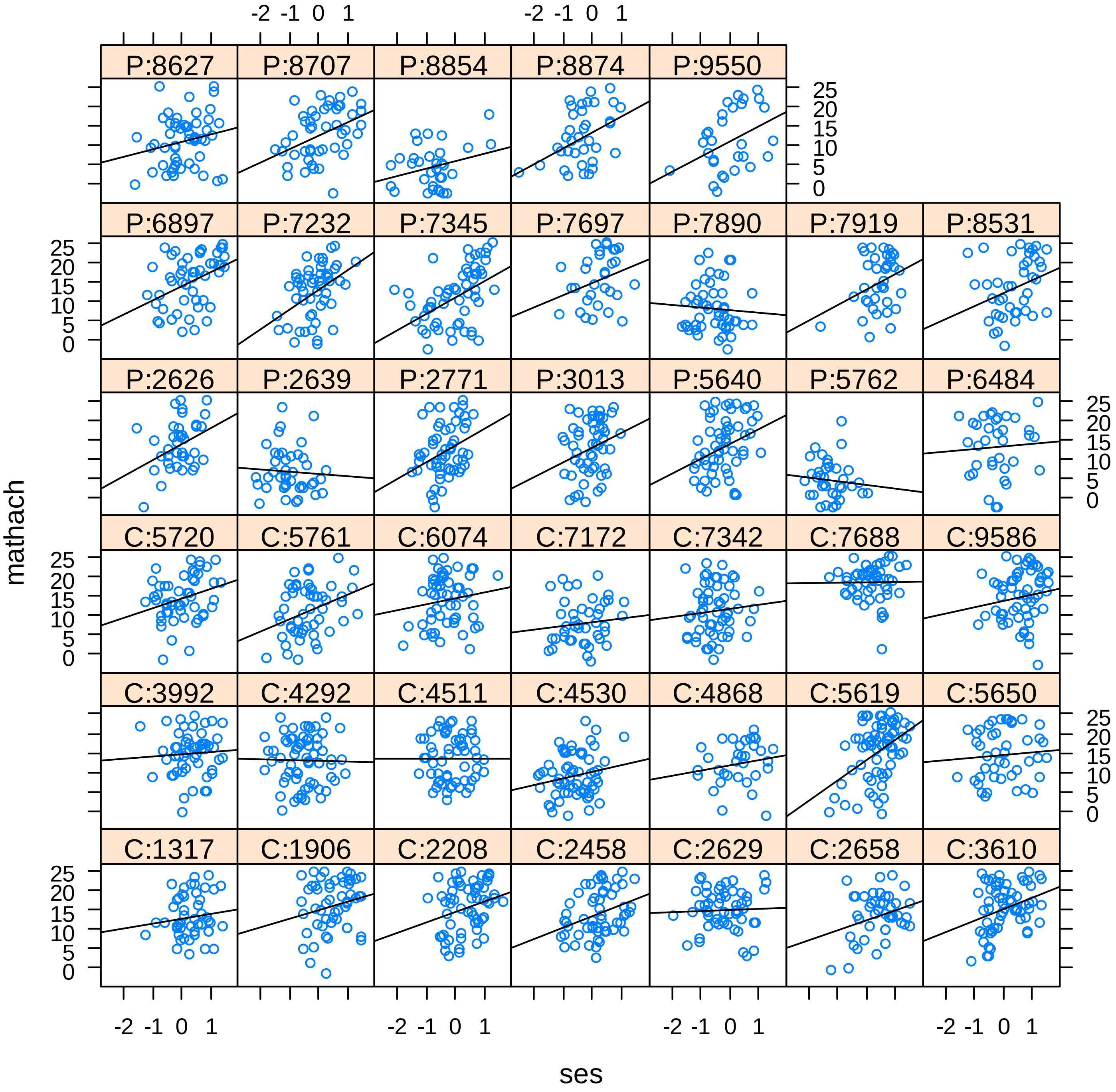
Variables (more later):

- MathAchievement of individual students
- SES of individual students

We would like to study the relationship between MathAchievement and SES

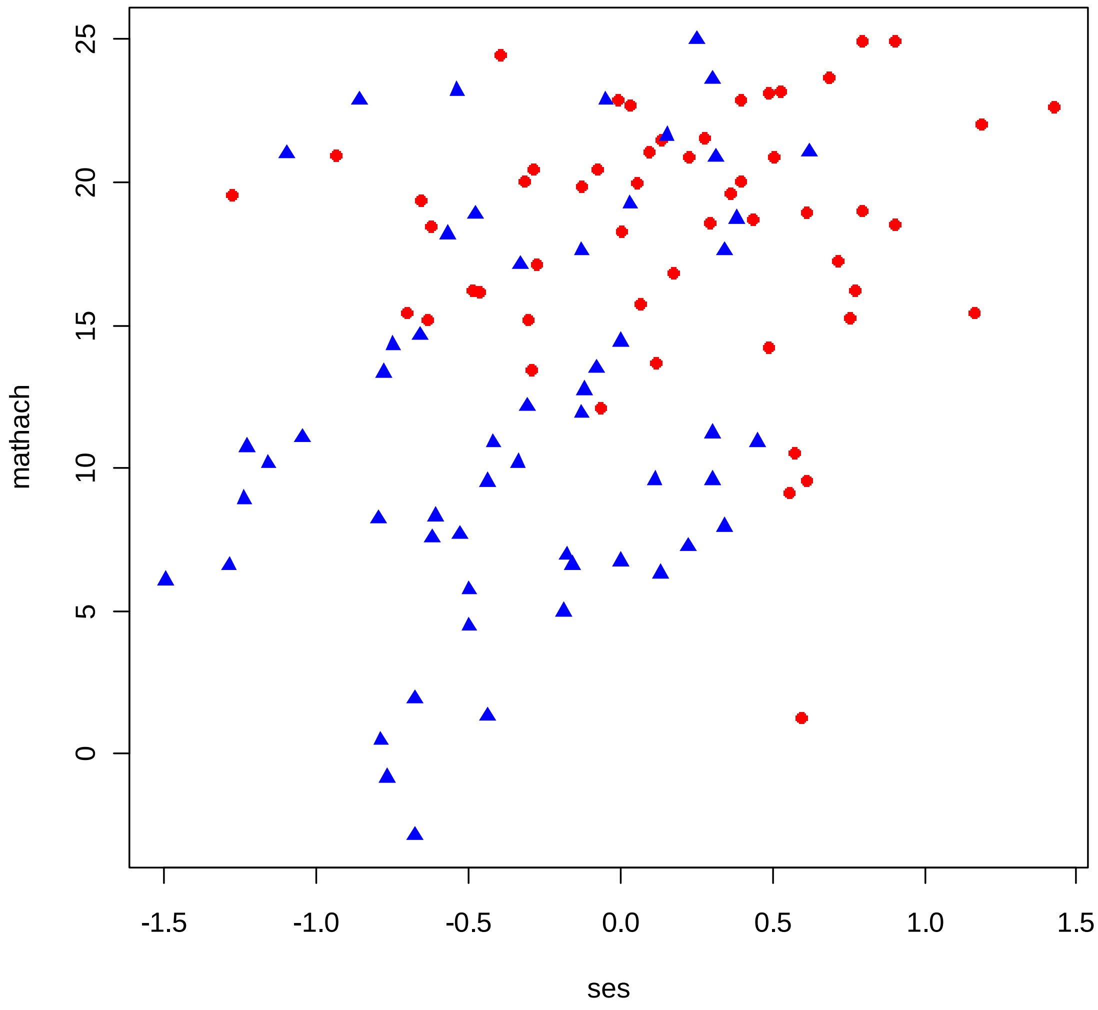
Seeing the data in each school:

```
> library( lattice )
> xyplot( mathach ~ ses | school, hs)  # plot with panel for each school
> hs$label <- factor(paste( substring(hs$Sector, 1,1), ":",
          hs$school, sep ="")) # nicer labels
> xyplot( mathach ~ ses | label, hs, # points and reg. lines in each panel
     panel = function( x, y, ...){
                panel.xyplot( x, y, ...)
                panel.lmline( x, y, ...)
    })
```

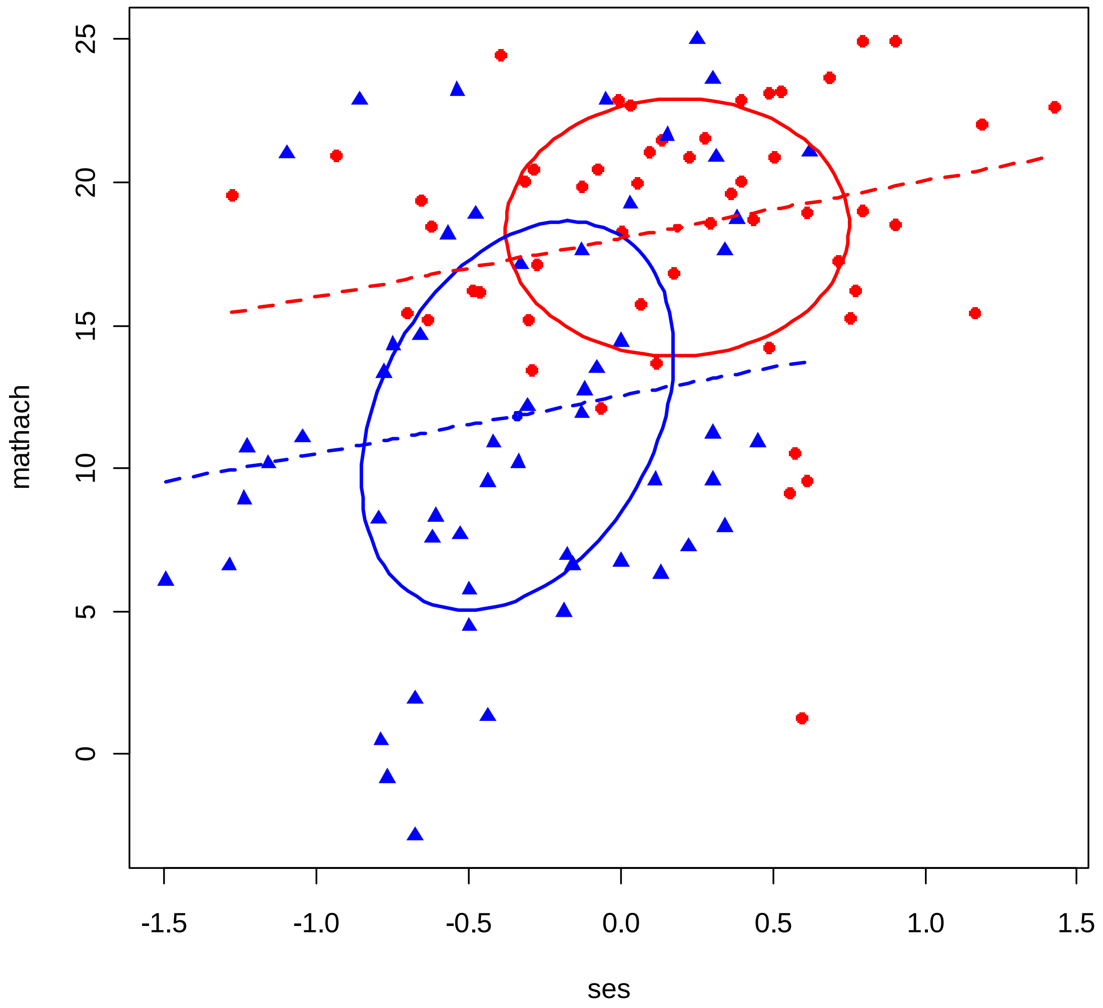

P:8627
P:8707
P:8854
P:8874
P:9550
P:6897
P:7232
P:7345
P:7697
P:7890
P:7919
P:8531
P:2626
P:2639
P:2771
P:3013
P:5640
P:5762
P:6484
C:5720
C:5761
C:6074
C:7172
C:7342
C:7688
C:9586
C:3992
C:4292
C:4511
C:4530
C:4868
C:5619
C:5650
C:1317
C:1906
C:2208
C:2458
C:2629
C:2658
C:3610
-2 -1 0 1
0
5
10
15
20
25
mathach
ses

For now we select two schools: P:2771 and C:7688

```
> hs2 <- hs[ hs$label %in% c("P:2771","C:7688"),]
```

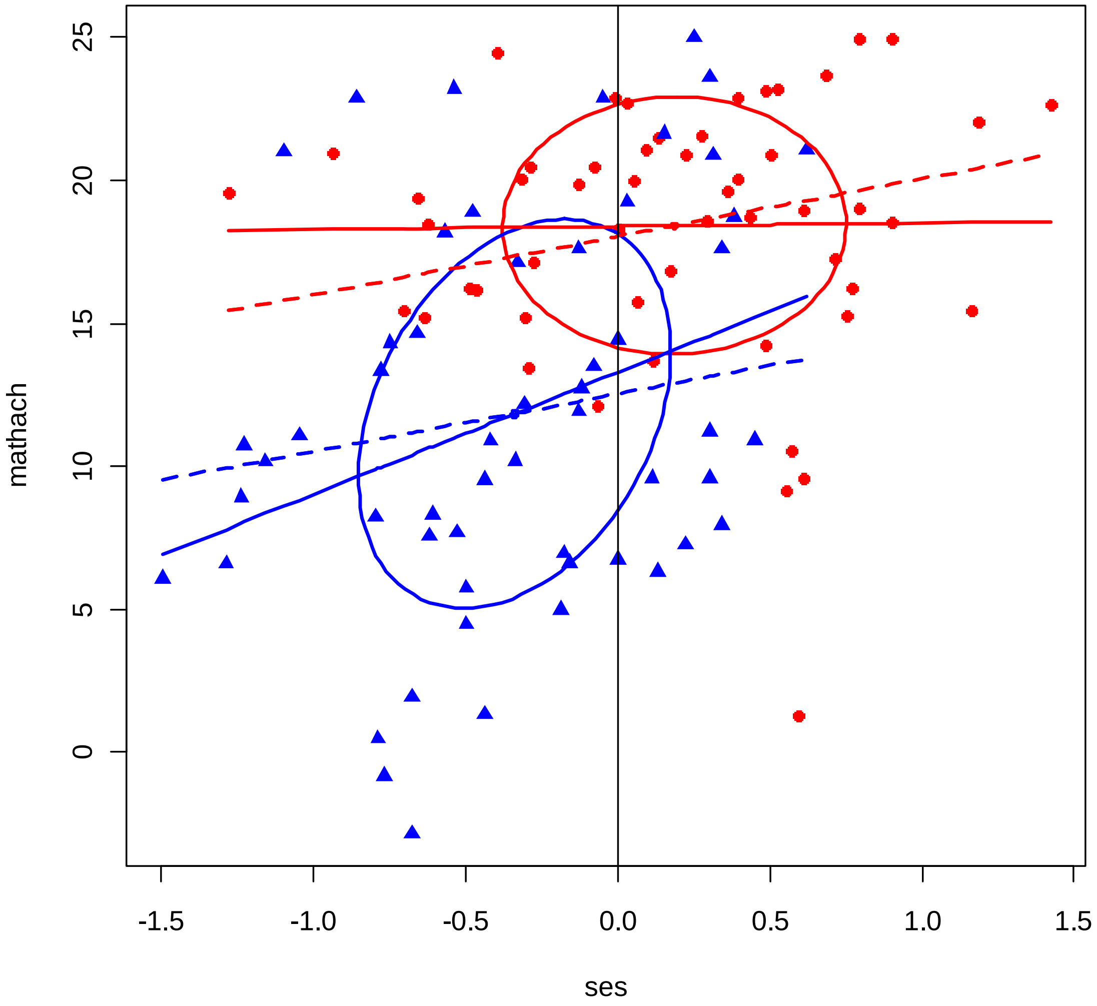


```
> plot( mathach ~ ses,
       hs2, type = 'n')

# set up the axes

> points( mathach ~ ses,
      hs2,
      subset =
        Sector == "Catholic",
      col = 'red',
      pch = 16)

> points( mathach ~ ses,
      hs2,
      subset =
        Sector == "Public",
      col = 'blue',
      pch = 17)
```



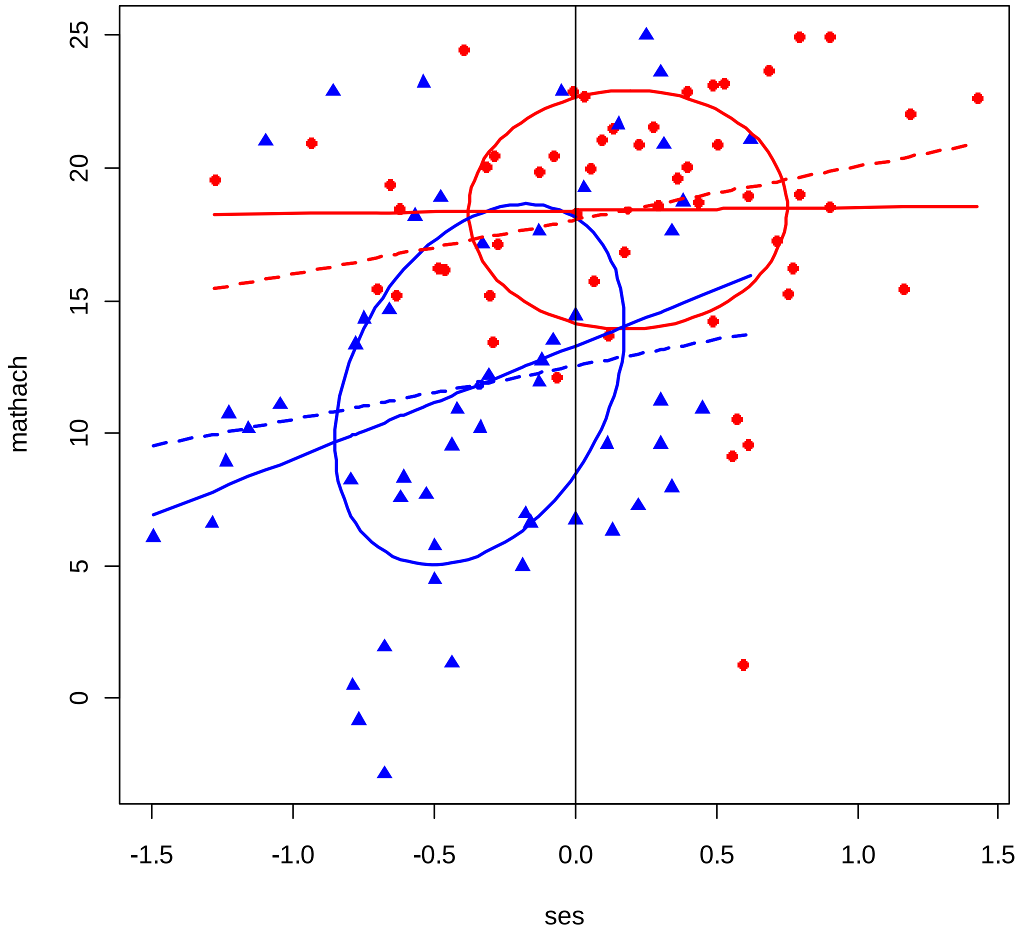
```
> lines( with( subset(hs2,
Sector == "Catholic"), dell(
ses, mathach, radius =
c(.02,1))), col = 'red', lwd
= 2)

> lines( with( subset(hs2,
  Sector == "Public"),
  dell( ses, mathach,
  radius = c(.02,1))),
  col = 'blue', lwd = 2)
> pred <- hs2[
    order(hs2$ses),]
# make dataset ordered for
plotting lines (ses in
order)
> pred$yhat.add <-
    predict( fit.add, pred)
# add yhat for additive
model
> pred$yhat.int <- predict(
fit.int, pred)   # add yhat
for interaction model
```

Additive model:



```
> summary(fit.add)
              Estimate Std. Error t value Pr(>|t|)
(Intercept)    18.0475     0.7984  22.605  < 2e-16 ***
ses             2.0166     1.0213   1.975   0.0509 .
SectorPublic   -5.5188     1.2164  -4.537 1.51e-05 ***
```

Interaction model:

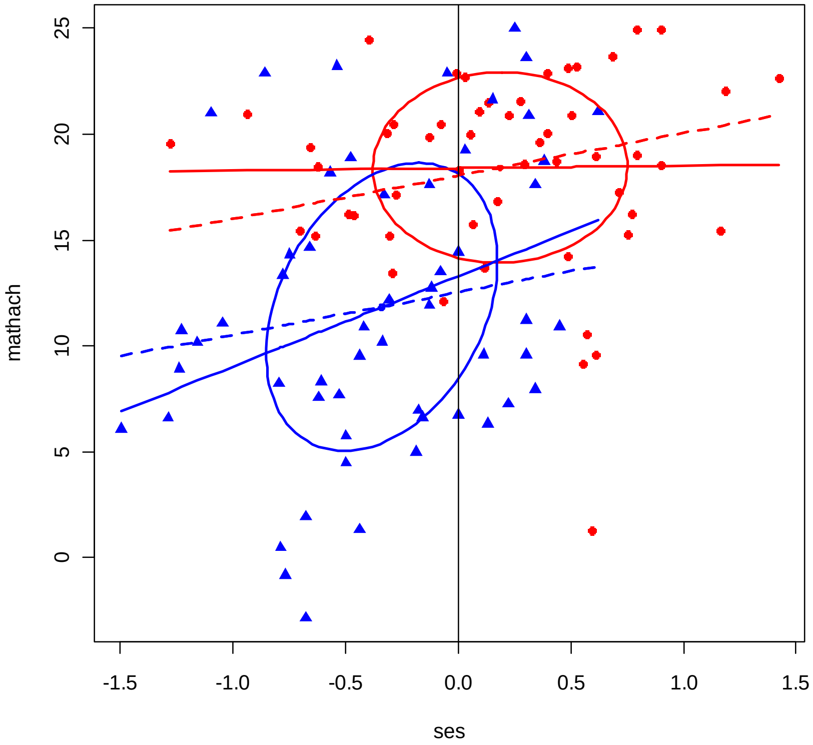


Interpreting coefficients:

```
> cbind(coef(fit.int))

[,1]
(Intercept)
18.4006876
ses
0.1163449
SectorPublic      -
5.1077227
ses:SectorPublic
4.1518431
```

```
> summary( fit.int )
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)        18.4007     0.8050  22.857  < 2e-16 ***
ses                 0.1163     1.3661   0.085   0.9323
SectorPublic       -5.1077     1.2149  -4.204 5.52e-05 ***
ses:SectorPublic    4.1518     2.0194   2.056   0.0423 *
```

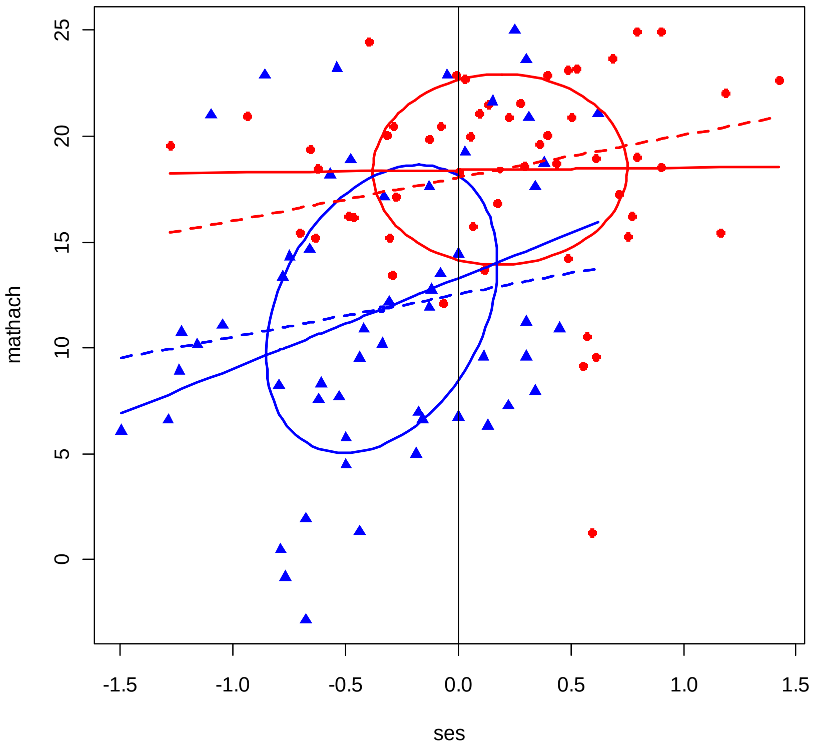
Estimating specific effects and testing specific hypotheses: Wald test for all coefficients



```
> cbind(coef(fit.int))
                          [,1]
(Intercept)         18.4006876
ses                  0.1163449
SectorPublic        -5.1077227
ses:SectorPublic     4.1518431
```

```
> wald ( fit.int)
 numDF denDF  F.value p.value
     4   105 208.6558 <.00001

Coefficients         Estimate Std.Error  DF   t-value p-value Lower 0.95 Upper 0.95
  (Intercept)       18.400688  0.805032 105 22.857085 <.00001  16.804458  19.996918
  ses                0.116345  1.366147 105  0.085163 0.93229  -2.592472   2.825162
  SectorPublic      -5.107723  1.214860 105 -4.204372 0.00006  -7.516566  -2.698880
  ses:SectorPublic  4.151843  2.019378 105  2.056001 0.04226   0.147789   8.155897
```

Estimating specific effects and testing specific hypotheses: estimating a specific level

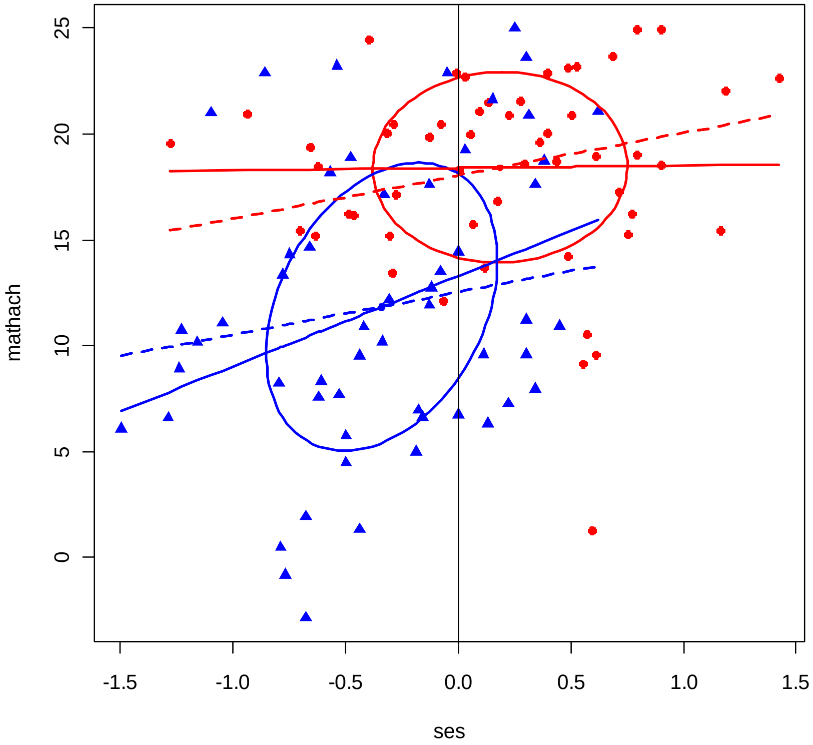


```
> cbind(coef(fit.int))
                            [,1]
(Intercept)           18.4006876
ses                    0.1163449
SectorPublic          -5.1077227
ses:SectorPublic       4.1518431
```

Level for Catholic when ses = 1.5 (Catholic is the ***REFERENCE LEVEL***)

```
> Lc <- rbind("ses=1.5|Catholic"= c(1, 1.5, 0, 0))
> Lc
                 [,1] [,2] [,3] [,4]
ses=1.5|Catholic    1  1.5    0    0
>         wald( fit.int, Lc)
  numDF denDF  F.value p.value
1     1   105 90.64267 <.00001
                 Estimate Std.Error  DF  t-value p-value Lower 0.95 Upper 0.95
  ses=1.5|Catholic 18.57521  1.951045 105 9.520644 <.00001   14.70664   22.44377
```

Estimating specific effects and testing specific hypotheses: estimating a specific level

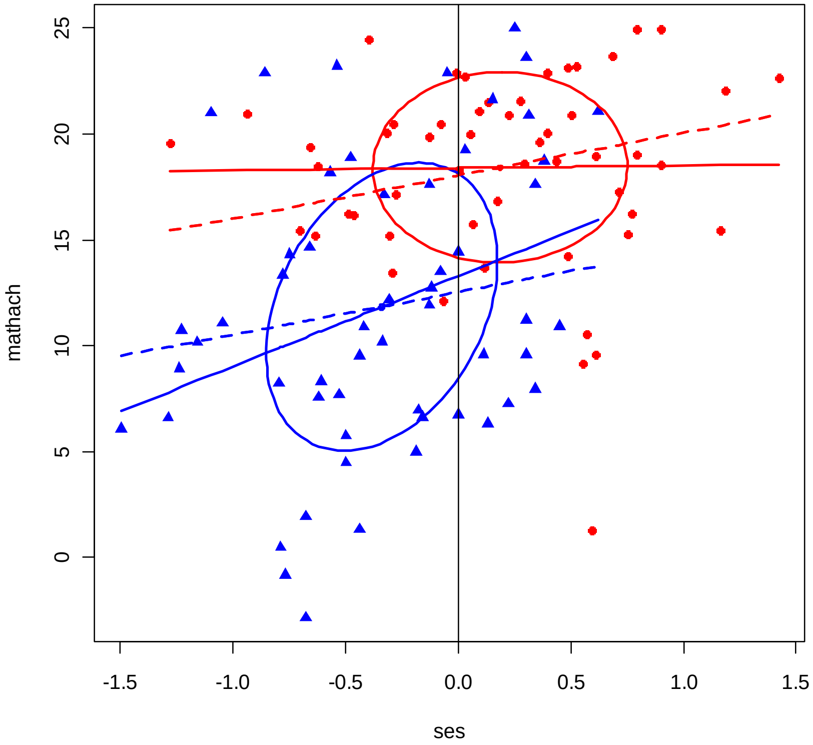


```
> cbind(coef(fit.int))
                        [,1]
(Intercept)       18.4006876
ses                0.1163449
SectorPublic      -5.1077227
ses:SectorPublic   4.1518431
```

Level for Public when ses = 1

```
> Lp <- rbind( "ses=1.5|Public" = c(1, 1.5, 1, 1.5))
> Lp
               [,1] [,2] [,3] [,4]
ses=1.5|Public    1  1.5    1  1.5
> wald( fit.int, Lp)
  numDF denDF  F.value p.value
1     1   105 48.15143 <.00001

                 Estimate Std.Error  DF  t-value p-value Lower 0.95 Upper 0.95
  ses=1.5|Public 19.69525  2.838291 105 6.939123 <.00001   14.06744   25.32305
```

Estimating specific effects and testing specific hypotheses: difference

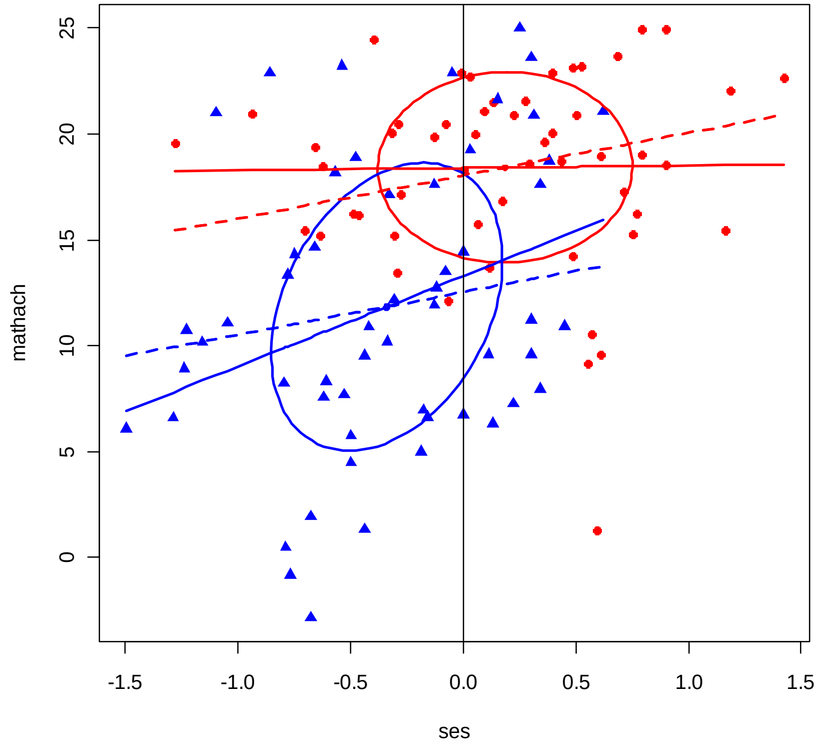


```
> cbind(coef(fit.int))
                          [,1]
(Intercept)         18.4006876
ses                  0.1163449
SectorPublic        -5.1077227
ses:SectorPublic     4.1518431
```

Difference at ses = -1.5

```
> Ld2 <- rbind(  "Cath - Pub | ses = 1.5" = c( 0,0,-1,1.5))
> wald( fit.int, Ld2 )
  numDF denDF  F.value p.value
1     1   105 13.61187 0.00036

                            Estimate Std.Error  DF  t-value p-value Lower 0.95 Upper 0.95
Cath - Pub | ses = 1.5 11.33549  3.072425 105 3.689427 0.00036   5.243436   17.42754
```

Estimating specific effects and testing specific hypotheses: simultaneous tests



```
> cbind(coef(fit.int))
                          [,1]
(Intercept)         18.4006876
ses                  0.1163449
SectorPublic        -5.1077227
ses:SectorPublic     4.1518431
```

Testing many simultaneously (note pattern of p-values)

```
> Ldm <- list( "Cath - Pub" = rbind(
+              "ses = -2" = c( 0, 0, -1, 2),
+              "ses = -1" = c( 0, 0, -1, 1),
+              "ses =  0" = c( 0, 0, -1, 0),
+              "ses =  1" = c( 0, 0, -1, -1),
+              "ses =  2" = c( 0, 0, -1, -2)))
```

```
> wald( fit.int, Ldm )    # note df for overall test
           numDF denDF  F.value p.value
Cath - Pub     2   105 12.71923   1e-05

            Estimate Std.Error  DF   t-value p-value Lower 0.95 Upper 0.95
  ses = -2 13.411409  4.021483 105  3.334941 0.00118   5.437552  21.385266
  ses = -1  9.259566  2.178581 105  4.250274 0.00005   4.939842  13.579289
  ses =  0  5.107723  1.214860 105  4.204372 0.00006   2.698880   7.516566
  ses =  1  0.955880  2.522168 105  0.378991 0.70546  -4.045113   5.956873
  ses =  2 -3.195963  4.404833 105 -0.725558 0.46972 -11.929933   5.538007
```

Testing selected parameters with pattern matching
note df for overall test and compare with previous

```
> wald( fit.int, "Sector" ) #       numDF denDF  F.value p.value
Sector     2   105 12.71923   1e-05

Coefficients        Estimate Std.Error  DF   t-value p-value Lower 0.95 Upper 0.95
  SectorPublic     -5.107723  1.214860 105 -4.204372 0.00006  -7.516566  -2.698880
  ses:SectorPublic  4.151843  2.019378 105  2.056001 0.04226   0.147789   8.155897
```

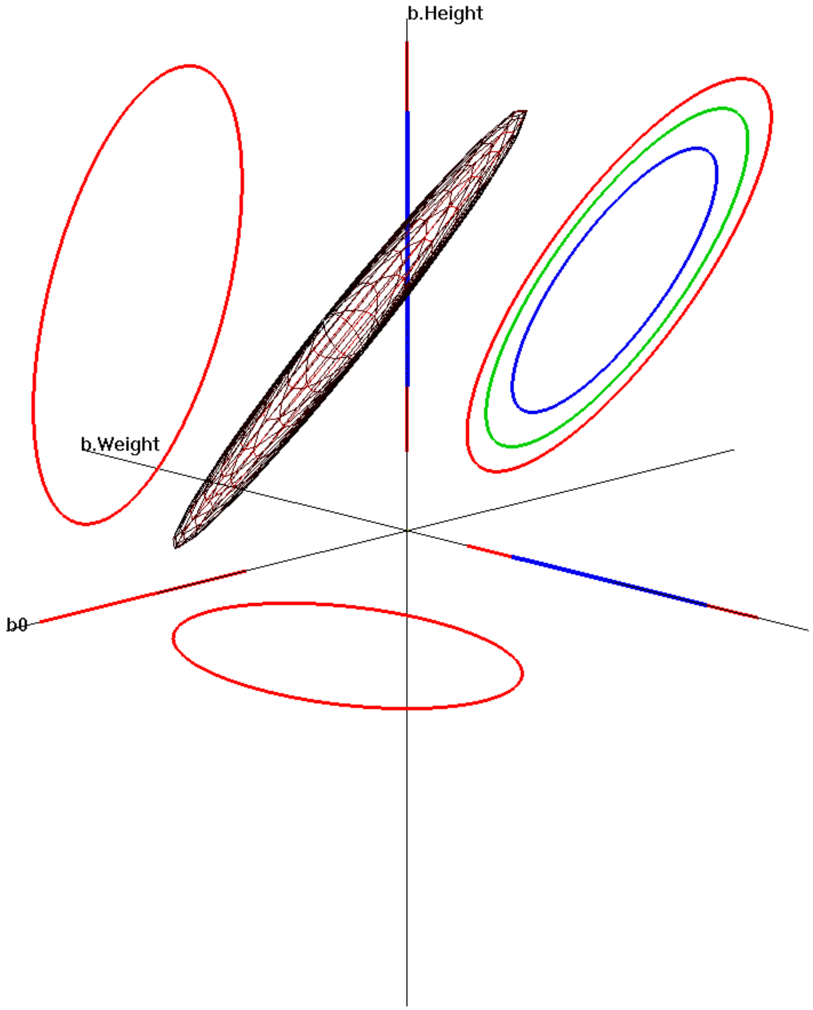
Appendix 1

Anatomy of a confidence region:

$$\hat{\beta} \oplus \sqrt{qF_{q,\nu}^{.95}} \times \frac{s_e}{\sqrt{n}} \times \sqrt{\Sigma_X^{-1}}$$

where

- q is the dimension for which we wish coverage
- ν is the degrees of freedom in the estimation of s_e
- Σ_X is the variance-covariance matrix of the predictors
- $\sqrt{\Sigma_X^{-1}}$ denotes the ellipse: $\{x : x'\Sigma_X x = 1\}$
- s_e is the standard error of regression, n is the number of observations.
- $\oplus$ could just be $+$ since it denotes the addition of the vector $\hat{\beta}$ to the ellipse on the right but the use of $\oplus$ serves as a reminder that the object on the right is a transformed circle.
- The linear shadows of the confidence region on linear subspaces are also confidence regions. q can be chosen to achieve accurate coverage in the desired dimension.



q **= 1 (blue), 2 (green) or 3 (red) produces confidence regions that have 95% coverage for a priori hypotheses of the corresponding dimension.**

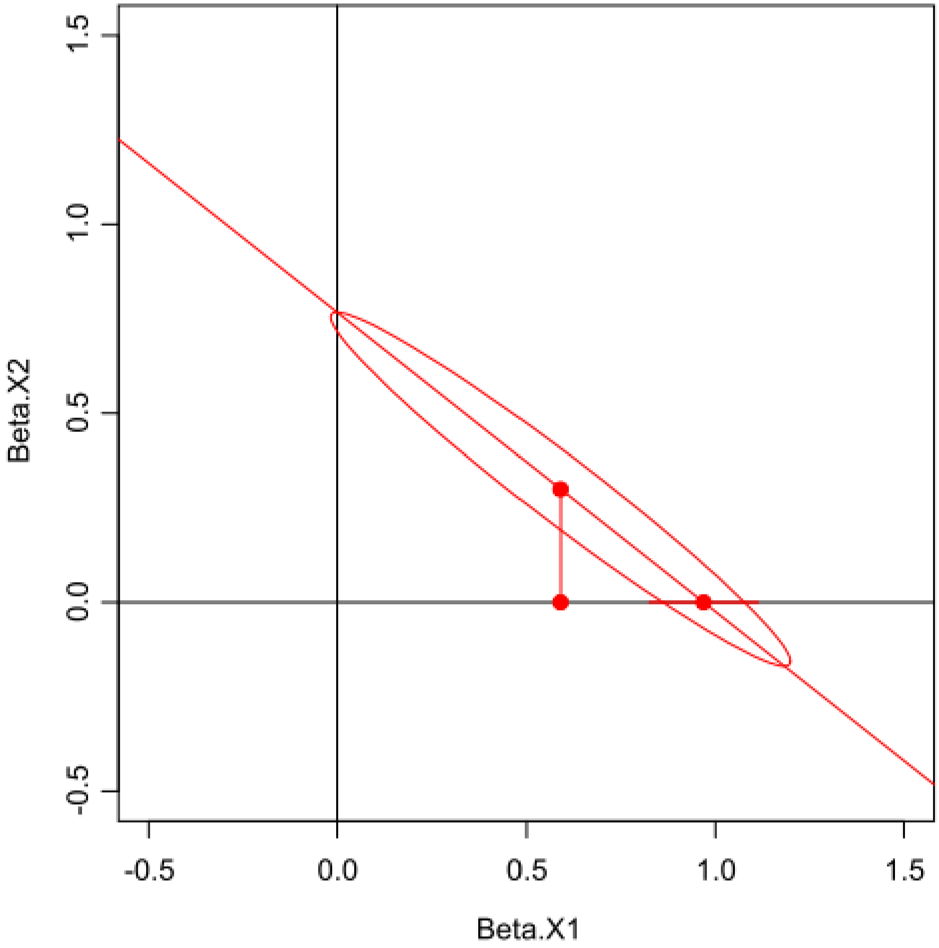
Appendix 2:
Dropping a non-significant term?

Fact or myth: “If a term is not significant, dropping it should not make much of a difference”

$$\widehat{Y} = \widehat{\beta}_0 + \widehat{\beta}_1 X_1 + \widehat{\beta}_2 X_2$$
$$\widehat{Y} = \widehat{\phi}_0 + \widehat{\phi}_1 X_1$$

If $\widehat{\beta}_2$ is not significant, what effect can dropping it have on the model?

What happens to the coefficient for X_1 in the following situation?



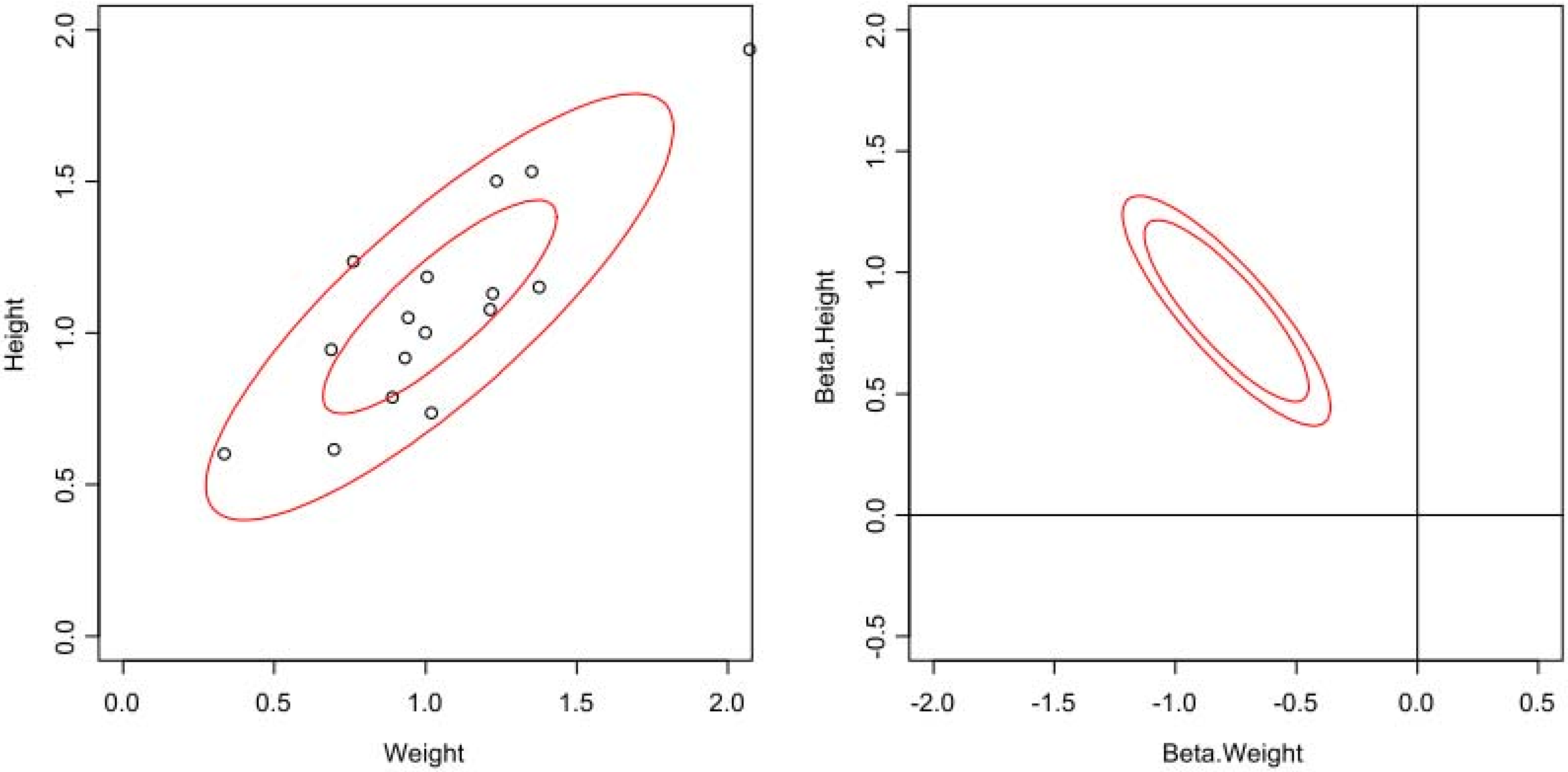
When we drop X_2 the effect of X_1 goes from appearing not to be significant to being highly significant. If it is important to control for X_2 in interpreting the coefficient of X_1 then dropping X_1 leads to a fallacious result even though the effect of X_1 is not significant.

Appendix 3:

Adding non-significant terms?

Fact or myth: If neither X_1 nor X_2 are significant in simple regressions, then they are unlikely to be significant in a multiple regression.

Consider:



What does this suggest about forward stepwise regression?