Bitte einfach nur den "Versuchsaufbau" *genau* beschreiben

Es wurden in der Universität in allen Fakultäten Plakate aufgeklebt, die unsere Studie bewerben, d.h. die Freiwilligen werden auf einen Onlinefragebogen eingeladen, mit dem wir alle Krankheiten usw. ausschließen, es ist eine Operationalisierung des Konzepts „bio-psycho-soziale Gesundheit“. Diejenigen, die die Tests überstanden haben, insgesamt *N* =12 werden nun zu zwei Sessions ins Labor eingeladen, mit einem Zeitabstand von > 1 Woche, d.h. es gibt insgesamt 24 Beobachtungen. Ein Kollege weist per Randomisierung die Versuchspersonen auf die Versuchsreihenfolge entweder Verum-Placebo oder Placebo-Verum zu; weder Versuchsleiter noch Versuchsperson wissen, welche Reihenfolge stattfindet, also double-blind. 50% erhalten so Verum-Placebo und die restlichen 50% Placebo-Verum, d.h. es ist ein *balanced design*. Wichtig ist, dass die Versuchspersonen den zu prüfenden Stimulus nicht bewusst hören können, er geht subjektiv unter einem Rauschen unter: VERUM ist Rauschen + spezielle Sinosoide und PLACEBO nur das Rauschen.

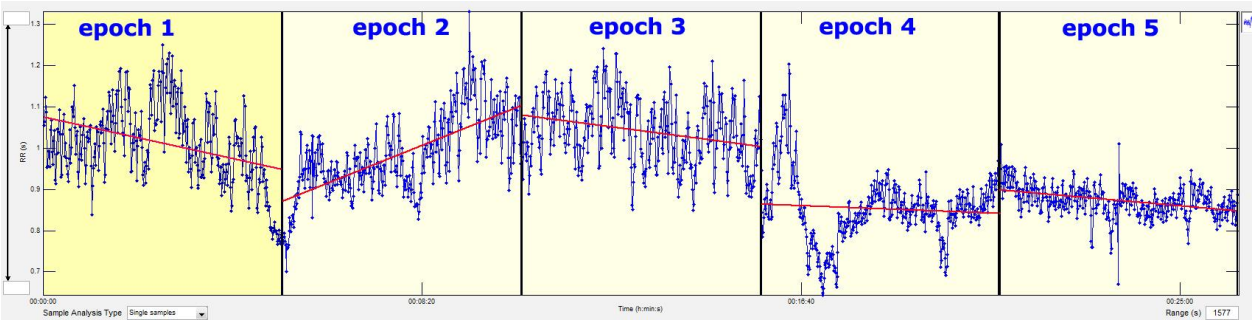
An den Versuchstagen werden den Versuchspersonen zunächst mit allen Messfühlern für die Biosignale des Körpers beklebt und die Signalqualität gecheckt. Dann fängt der Versuch an, indem über Kopfhörer bei geschlossenen Augen erstmal ca. 30min eine Entspannungsübung eingespielt wird, damit sich die Leute besser ihrem „entspannten Normalzustand“ annähern. Nun wird nahtlos anschließend der VERUM oder PLACEBO Stimulus eingespielt, er ist genau 25min lang. Nach dem Versuch wird der Person ein etablierter Fragebogen gegeben, der das subjektive Entspannungsgefühl misst und als psychometrischen Testscore ausdrückt. Als Daten haben wir also jeweils einen Testscore für VERUM und einen für PLACEBO und die zughörigen Aufzeichnungen der Biosignale über 25min, alle mit hoher Genauigkeit von 1024Hz digital gesampelt.

Als erste Analyse ergibt sich ein signifikanter Unterschied in den psychometrischen Testscores des subjektiven Entspannungsgefühls. Weil VERUM vs. PLACEBO ja immer dieselben Personen sind, handelt es sich also um *related samples.* Mit dem WILCOXON signed-rank test ergibt sich:

Z	-2,511
Asymp. Sig. (2-tailed)	,012
Exact Sig. (2-tailed)	,009
Exact Sig. (1-tailed)	,004
Point Probability	,001

Weil die Sinosoide unhörbar sind und es ein double-blind Versuch ist, kann also davon ausgegangen werden, dass es *subjektiv* einen Effekt dieser Sinosoide tatsächlich gibt. Jetzt ist die Frage, ob man auch etwas in den Biosignalen feststellen kann.

Dazu beschränke ich mich hier nur auf die Analyse der *Heart Rate Variability*. Die Herzfrequenz, gemäß Konvention gemessen in ms-Abständen zwischen zwei Herzschlägen = Interbeat-Interval (*IBI),* wird über die 25min des VERUM oder PLACEBO Stimulus *kontinuierlich* gemessen und jeweils in 5-Minuten Epochen unterteilt, also so:



Die roten Linien stellen das laut Konvention notwendige lineare Detrending per Epoche dar. Für alle 5 Epochen werden nun nach Konvention spezielle deskriptive Parameter berechnet, die das „Typische“ der Epoche ausdrücken sollen, dieser Schritt wird *Parametrisierung* genannt. Dann wird laut Konvention noch eine

Datentransformation vorgenommen, nämlich natürliche Logarithmierung dieser Parameter oder im Fall von Verhältnissen/Anteilen/Prozenten die *logit* – Transformation.

Über die wichtigsten gängigen HRV-Parameter checken wir nun mit dem QUADE-Test über alle *N = 12* Personen einmal über alle 5 VERUM – Epochen vs. über alle 5 PLACEBO – Epochen, ob es signifikante Änderungen im Zeitverlauf gibt oder nicht. Weil wir 30 QUADE-Tests ausführen, müssen wir wegen alpha-Fehler-Kumulierung mit 1.5 falsch signifikanten Ergebnissen rechnen (*false detection rate*). Wir erhalten folgendes Ergebnis:

QUADE-TEST over 5 epochs

	data transform	p-values (FstatisticApprox_PvalueTwoTailed) PLACEBO vs.	VERUM
MeanRR_ms	log	0,7908	0,3745
SDNN_ms	log	0,5527	0,3425
RMSSD_ms	log	0,2059	0,9708
pNN50_percent	logit	0,2100	0,4755
HRVtriangularIndex_dimless	log	0,6704	0,8038
AR_PeakHF_Hz	log	0,9151	0,4755
FFT_PeakHF_Hz	log	0,2714	0,8873
FFT_HFpower_ms2	log	0,9630	0,7627
FFT_HFpower_prc	logit	0,3173	0,0030
FFT_LFpower_ms2	log	0,5492	0,1189
FFT_LFpower_prc	logit	0,7750	0,7488
AR_HFpower_ms2	log	0,4841	0,6467
AR_HFpower_prc	logit	0,9935	0,0149
AR_LFpower_ms2	log	0,7464	0,1638
AR_LFpower_prc	logit	0,9212	0,9888
averaged_BreathingFREQ_Hz	log	0,1312	0,3578

FRIEDMAN test
VERUM pexact=.021 <==>
PLACEBO pexact =.849

NUR die HFpower normalisiert in Prozent zeigt signifikante Veränderungen über die Zeit und das nur unter VERUM, nicht unter PLACEBO. Dabei ist es egal, ob man die HFpower mit der auf *Fast Fourier Transform* basierten *WELCH* –Technik berechnet, oder mit dem auch oft verwendeten AR-Ansatz, beides führt zum Erfolg. Das Gegenchecken mit einem FRIEDMAN-Test zeigt, dass auch er das Ergebnis bestätigt.

Soweit, so schön. Es ist wohl so, dass unter VERUM irgendwelche Fluktuationen in irgendeiner/irgendwelchen Epochen in HFpower auftreten, so dass man die Nullhypothese, dass alle Messpunkte sehr wohl aus der gleichen Population stammen könnten, verwerfen muss. Nach Bestätigung der *Omnibus-/Globalen* Hypothese, dass es tatsächlich wohl eine *psychophysiological reactivity* auf den VERUM-Stimulus gibt, interessiert natürlich das *lokale* Geschehen innerhalb jener 5 Epochen, das wäre im ANOVA-Ansatz dann wohl die Abteilung post-hoc tests. Es schient nun so zu sein, dass mit *N* = 12 die Klärung des lokalen Geschehens ein wenig im statistischen Rauschen untergeht, bzw. es ist schwierig, hier was Konsistentes zu erkennen. Aber natürlich werde ich als erstes gefragt werden: Und was denn nun, steigt das da oder sinkt oder was passiert denn da? Da muss ich schon irgendwas zu sagen können, auch wenn ggf. der einzige Schluss dieser wäre: „Dazu ist das Versuchsdesign leider nicht geeignet, es ging nur erstmal darum *ob* es überhaupt placebo-überlegene Reaktionen gibt, alles Weitere muss in Folgestudien geklärt werden“

Aber: Du erklärst alles sehr ausführlich, nur die Kernhypothese, die ihr testen wollt und was für Daten ihr genau habt, geht irgendwie immer unter.

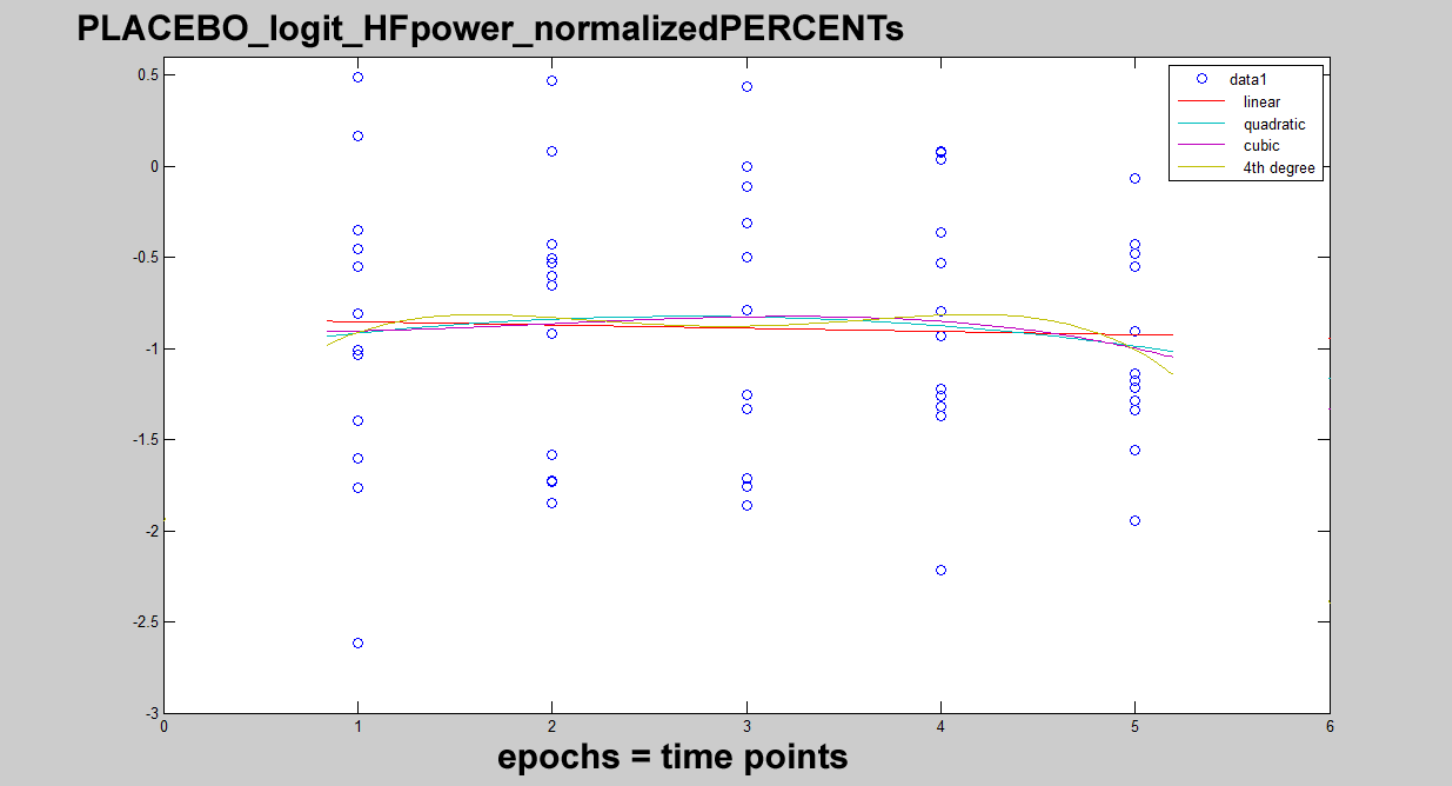
Die Kernhypothese war, ob diese Sinosoide überhaupt irgendwelche vom Placebo verschiedenen Reaktionen auslösen können, oder ob das ganze Humbug ist. Im Grunde geht es als Ziel der Studie darum: Sollte man diesen Effekt weiter und genauer und aufwändiger untersuchen, wenn es ihn denn überhaupt gibt? Die Idee war, weniger Leute zu messen, dafür aber in multiplen Domänen, also subjektive Daten, physiologische Daten in verschiedenen Ebenen, Speichelproben etc. Wenn es in vielen Domänen Hinweise gibt, dass der Effekt existiert und diese Hinweise auch theoriekonform sind, dann können die Befunde in ihrer Gesamtheit kein bloßer Zufall sein, d.h. es ist wahrscheinlich dass es diese Sinosoide tatsächlich wirksam sind. Es ist eine Wirksamkeitsstudie, genauso wie wenn man eine neue Substanz gefunden hätte und vermutet, dass sie z.B.

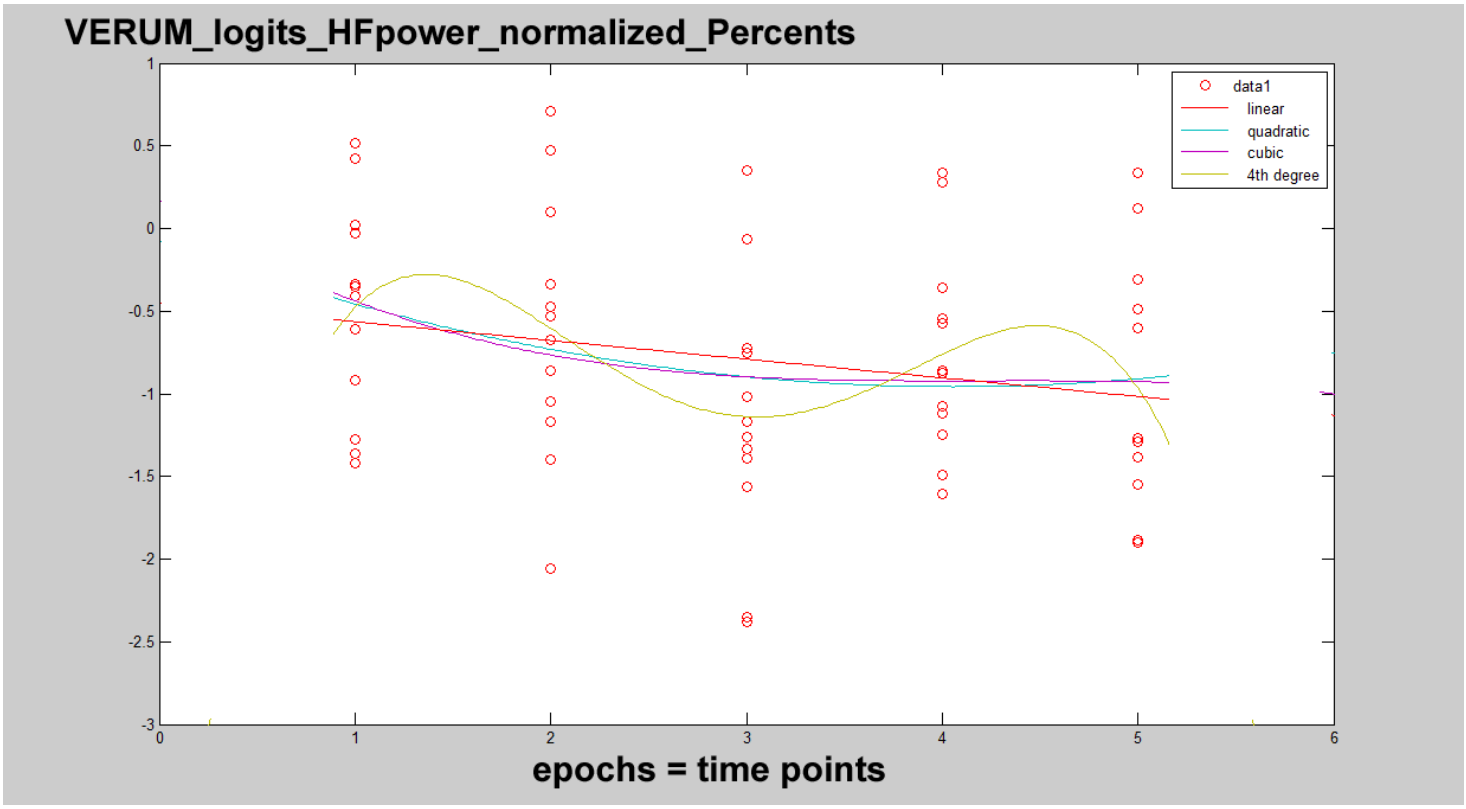
blutdrucksenkend wirkt; bevor man biochemische Details etc. aufwändig erforscht muss man ja erstmal sehen, ob diese Substanz zumindest eine kleine spezifische Blutdrucksenkung erzeugen kann.

Zu deiner Frage mit dem OLS-Schätzer: Ja, den kann man mitteln, nur das ergibt bei 12 Werten vermutlich keinen Sinn. Dazu müsste man die Ergebnisse sich genauer ansehen. Aber du kannst nicht einfach mit einem nicht-parametrischen Test anfangen, dann mit nicht-linearität kommen und am Ende einen linearen Schätzer darüber legen. Damit wirfst du alle statistischen Argumentationen durcheinander und das wird dir jeder, der halbwegs statistisch bewandert ist, auf einer Konferenz um die Ohren hauen.

Mit dieser Argumentation hast Du natürlich völlig Recht, erst non-parametrisch und dann alles linear, das geht gegen den Strich. Wenn ich es richtig verstehe, geht es hier um *curve* fitting, d.h. wir suchen diejenige Ausgleichkurve, die die *Stichprobenwerte* am besten zusammenfassend beschreibt, es geht erstmal *nicht* darum, auf die Verhältnisse in der Population inferenzstatistisch zurückzuschließen.

Deswegen habe ich mal verschiedene *non-lineare curve fitting* – Methoden ausprobiert, jeweils VERUM vs. PLACEBO:





Fit name: Smoothing Spline

X data: timepoints

Y data: PLACEBO_logits_HFpower_prc

Z data: (none)

Weights: (none)

Smoothing Spline

Smoothing Parameter

Default

Specify: < Smoother 0.99999999999999999999999 Rougher >

Center and scale

Auto fit

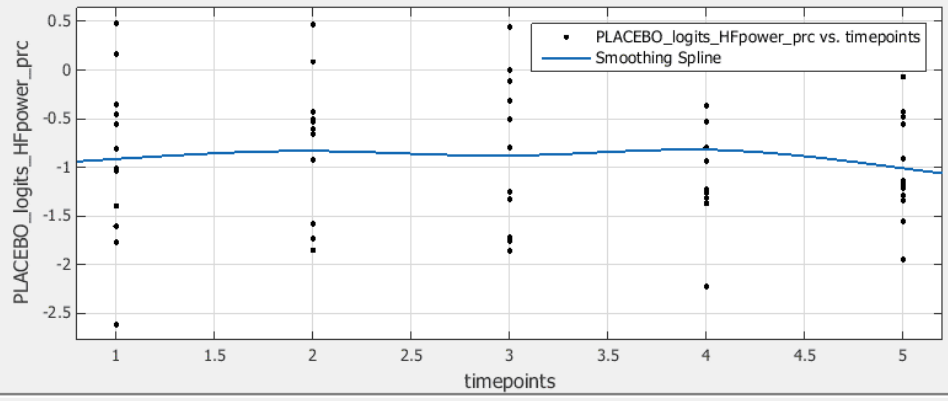
Fit

Stop

Results

Smoothing spline:
f(x) = piecewise polynomial computed from p
where x is normalized by mean 3 and std 1.426
Smoothing parameter:
p = 1

Goodness of fit:
SSE: 29.56
R-square: 0.009407
Adjusted R-square: -0.06264
RMSE: 0.7331



Fit name: Smoothing Spline

X data: timepoints

Y data: VERUM_logits_HFpower_prc

Z data: (none)

Weights: (none)

Smoothing Spline

Smoothing Parameter

Default

Specify: < Smoother 0.99999999999999999999999 Rougher >

Center and scale

Auto fit

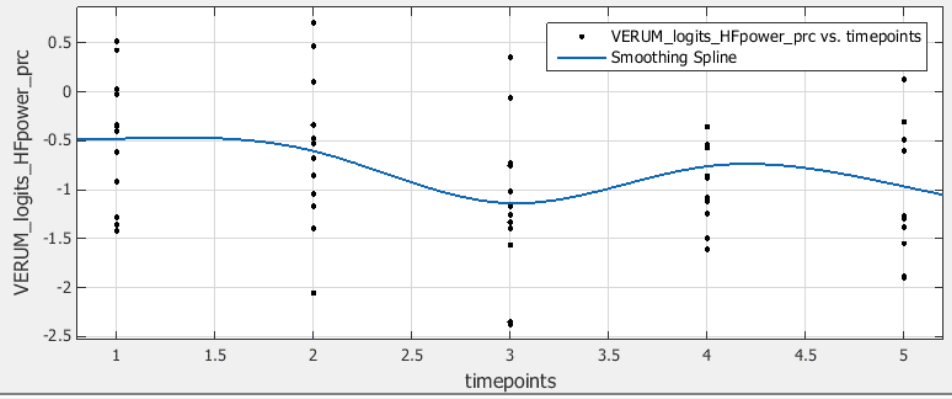
Fit

Stop

Results

Smoothing spline:
f(x) = piecewise polynomial computed from p
where x is normalized by mean 3 and std 1.426
Smoothing parameter:
p = 1

Goodness of fit:
SSE: 29.16
R-square: 0.1046
Adjusted R-square: 0.03946
RMSE: 0.7282



Fit name: Fourier_8terms

X data: timepoints

Y data: PLACEBO_logits_HFpower_prc

Z data: (none)

Weights: (none)

Fourier

Number of terms: 8

Equation: a0 + a1*cos(x*w) + b1*sin(x*w) + ... + a8*cos(8*x*w)+b8*sin(8*x*w)

Center and scale

Fit Options...

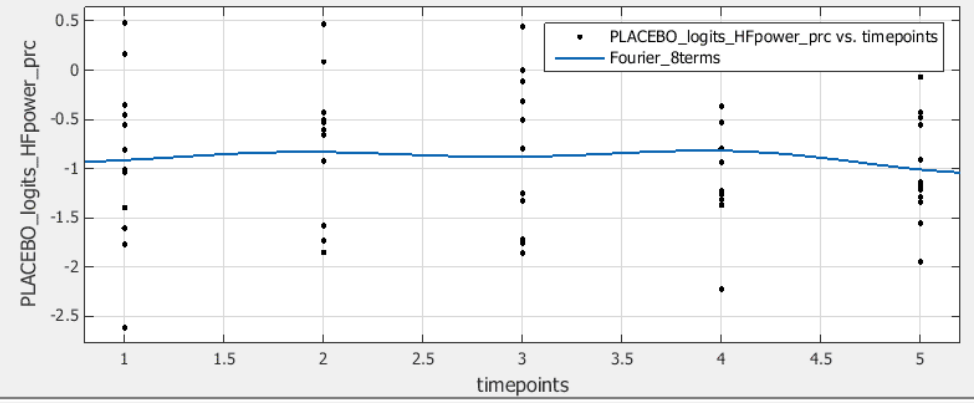
Auto fi

Fit

Stop

Results

General model Fourier8:
f(x) =
a0 + a1*cos(x*w) + b1*sin(x*w) +
a2*cos(2*x*w) + b2*sin(2*x*w) + a3*cos(3*x*w) + b3
a4*cos(4*x*w) + b4*sin(4*x*w) + a5*cos(5*x*w) + b5
a6*cos(6*x*w) + b6*sin(6*x*w) + a7*cos(7*x*w) + b7
a8*cos(8*x*w) + b8*sin(8*x*w)
Coefficients (with 95% confidence bounds):
a0 = -0.8794 (-6.266e+46, 6.266e+46)
a1 = 0 (-2.106e+47, 2.106e+47)
b1 = 0 (-1.275e+47, 1.275e+47)
a2 = 0 (-1.407e+47, 1.407e+47)
b2 = 0 (-2.172e+47, 2.172e+47)
a3 = 0 (-7.517e+46, 7.517e+46)
b3 = 0.0257 (-6.466e+46, 6.466e+46)
a4 = 0 (-1.4e+46, 1.4e+46)
b4 = 0 (-8.907e+45, 8.907e+45)
a5 = 0 (-2.099e+46, 2.099e+46)
b5 = -0.05102 (-8.095e+45, 8.095e+45)
a6 = 0 (-1.02e+45, 1.02e+45)



Fit name: Smoothing Spline

X data: timepoints

Y data: VERUM_logits_HFpower_prc

Z data: (none)

Weights: (none)

Fourier

Number of terms: 8

Equation: a0 + a1*cos(x*w) + b1*sin(x*w) + ... + a8*cos(8*x*w)+b8*sin(8*x*w)

Center and scale

Fit Options...

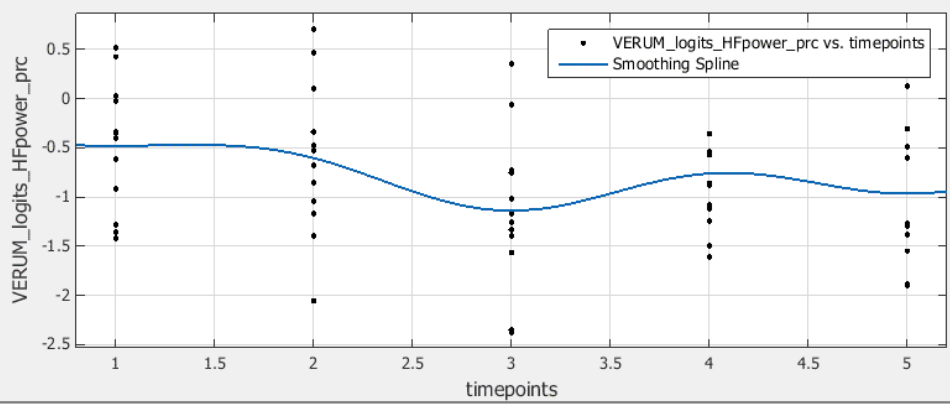
Auto fit

Fit

Stop

Results

General model Fourier8:
f(x) =
a0 + a1*cos(x*w) + b1*sin(x*w) +
a2*cos(2*x*w) + b2*sin(2*x*w) + a3*cos(3*x*w) + b3
a4*cos(4*x*w) + b4*sin(4*x*w) + a5*cos(5*x*w) + b5
a6*cos(6*x*w) + b6*sin(6*x*w) + a7*cos(7*x*w) + b7
a8*cos(8*x*w) + b8*sin(8*x*w)
Coefficients (with 95% confidence bounds):
a0 = -0.7898 (-3.917e+46, 3.917e+46)
a1 = 0 (-1.805e+46, 1.805e+46)
b1 = 0 (-2.753e+46, 2.753e+46)
a2 = 0 (-1.01e+46, 1.01e+46)
b2 = 0 (-7.018e+46, 7.018e+46)
a3 = 0 (-4.915e+46, 4.915e+46)
b3 = 0.2322 (-6.517e+45, 6.517e+45)
a4 = 0 (-4.731e+45, 4.731e+45)
b4 = 0.08565 (-1.826e+45, 1.826e+45)
a5 = 0 (-5.23e+45, 5.23e+45)
b5 = 0.1409 (-6.775e+45, 6.775e+45)
a6 = 0 (-1.943e+46, 1.943e+46)



Dann habe ich noch etwas ziemlich Spannendes ausprobiert, nämlich *Symbolische Regression*. Das ist ein *machine learning* Ansatz:

> Symbolic regression proceeds with model building by first, asking the researcher to select a set of primitive functional operators allowed in the mathematical models, second, by applying an evolutionary algorithm to evolve both model structures and model parameters, and third, by scrutinizing modeling results to identify the driving input variables, and to select the final ensemble of models.

Man gibt zugelassene Funktionsklassen vor, also z.B. trigonometrische Funktionen wie sinus, tangens, oder Exponentialfunktionen usw., gibt die Daten ein, also hier 60 Messpunkte und nun sucht der Algorithmus diejenigen Funktionen mit denjenigen Parametern heraus, die einerseits den Error minimieren, andererseits aber möglich wenig mathematisch komplex sind. Diese sehr rechenaufwändige Suche funktioniert mit dem Ansatz des *genetic programming.* Es gibt ein gutes Programm, das das erledigen kann: http://www.nutonian.com/products/eureqa/ lies mal unter dem Reiter *Technology* nach.

Wenn ich also mit diesem Ansatz ein curve fitting durchführe (Rechenzeit einen Tag mit einem Rechner i7 und 8 cores!), dann kommt folgendes raus, sieh dir bitte das *VERUMvsPLACEBO_curveFitting_SymbolicRegression.gif* an, das passt nicht so gut hier rein, deswegen habe ich es separat

Also, ich finde schon, dass man sagen kann, dass die HFpower unter Verum abgenommen hat, wenn man sich die rote Ausgleichskurve ansieht und in Betracht zieht, dass der QUADE/FRIEDMAN-Test die Globalhypothese von überzufälligen Veränderungen im Zeitverlauf bestätigt hat? Was meinst Du?

Tausend Dank für Deine Hilfe, das sind wirklich sehr viele Infos hier gewesen.

Übrigens, Frohes Neues Jahr!