Analysis of Continuous Data

Master of Statistical Data Analysis

January 13, 2014

Q1: On stunted growth (inspired by Berngard et al., 2013)

Important variables in the study of growth retardation (in children) are the length-for-age z-score (LAZ) and weight-for-age z-score (WAZ). LAZ has the following definition in a targeted population (similar for WAZ):

$$LAZ = \frac{L - E(L|A)}{\sqrt{\text{Var}(L|A)}}$$

where the conditional mean, $E(L|A)$, and variance, $\text{Var}(L|A)$, of the length, L, are considered for given age, A, in a well-defined reference population.

(a) What is the expected value and variance of LAZ when a random person is drawn form the reference population?

(b) Consider the data in the Figure of WAZ versus LAZ below, where the solid line represents the least squares line. By just looking at the figure, describe in at most 3 sentences what elements of it are pointing to possible problems with growth.

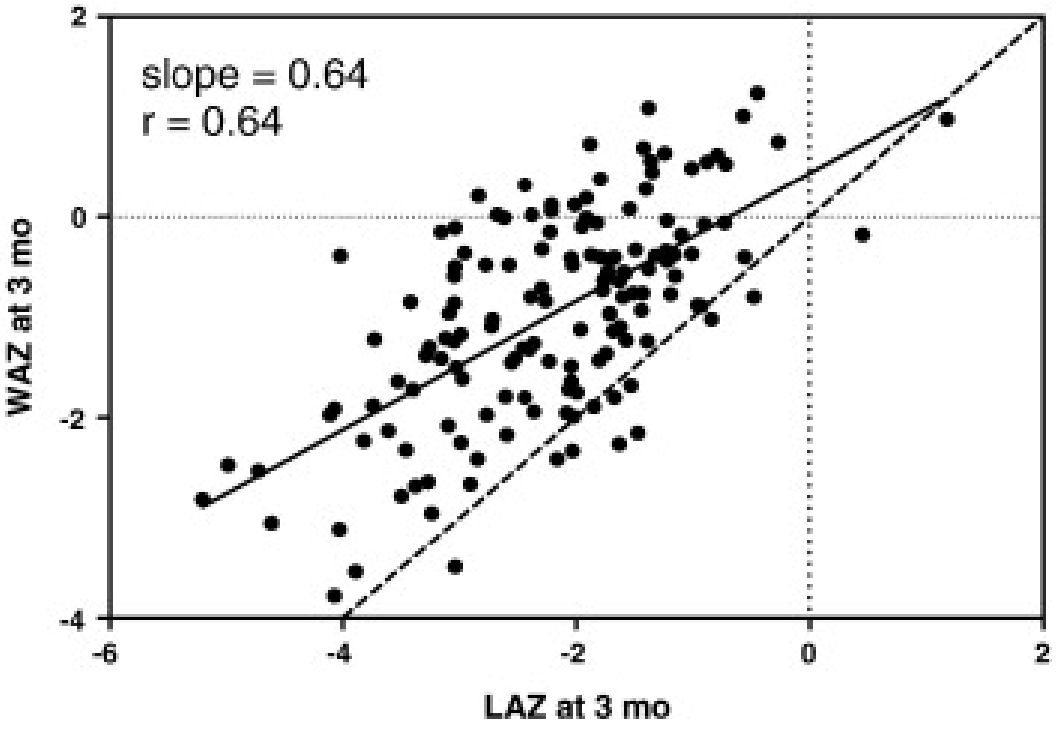
(c) Continuing with that figure, write down the linear model for the regression of WAZ against LAZ at 3 months and give estimates for all the model parameters based on the global features of the figure or any other information given there.

(d) Write down the assumptions involved in the simple linear model above and explain where possible whether they appear to be justified or not, based on the information provided.

(e) Assuming the assumptions hold, calculate an estimated 95% confidence and 95% prediction interval for WAZ at a given LAZ value of -2. Explain your derivations. If some assumptions were considered doubtful, explain briefly how this might affect the confidence and prediction interval just calculated.

(f) What percentage of the variation in WAZ is explained by the variation in LAZ according to the fitted model?

(g) One wishes to test whether the slope of the regression line through the data coincides with the slope of the identity line (dotted line on the plot) or not. Write down the null hypothesis and alternative for this test and



perform the test. Explain your calculations and use approximate estimates obtained from the global features or any other information seen in the figure.

(h) Calculate the power of this test to detect a slope equal to 0.75 (rather than 1), when the test is one-sided to the left. To this end, you may assume that the population variance is known and you may set it equal to the value estimated above.

(i) The children in this study have not only been measured at month 3 but also at month 6. Suppose the data of both time points (months 3 and 6) had been added to the figure and analyzed together by the model under (c) above. What problem would then occur for the simple linear regression model? How would that affect the estimate and 95% confidence interval for the slope?

(j) Now assume that our dataset and the figure show measurements taken from children who are 3 months old *and* different (independent) children who are 6 months old. Assume further more that the variance of W is equal at both these ages, and also the variance of L is equal at both these ages. Consider then the regression of weight, W, on both age and length. First, weight is regressed on age and next we plan to add length to the model. Explain briefly how the typical partial residual plot is constructed that helps decide whether length can simply be added to the linear model and where you can then read off the regression coefficient of the added length variable from this partial residual plot

(k) We continue under that same set-up in (j) of children at two different ages, with constant variances of L and W at both ages. Now consider again the plot provided in the Figure (but with points from months 3 and 6) and explain how it relates to the partial residual plot described under (j). Derive then from the given plot in the figure an estimate for the regression coefficient β_L in the model:

$$W = \beta_0 + \beta_A A + \beta_L L + \epsilon$$

where the standard assumptions of the multiple linear regression model are assumed to hold.

Question 2: Sugar-Sweetened Beverages

In the paper "Sugar-Sweetened Beverages and Genetic Risk of Obesity" (The New England Journal of Medicine (2012)), Qibin, Qi et al. analyse the interaction between genetic predisposition (GenPred) and the intake of sugar-sweetened beverages (SSB, considered as numeric variable) in relation to body-mass index (BMI) in 6 934 women from the Nurses' Health Study (NHS) and in 4 423 men from the Health Professionals Follow-up Study (HPFS). The genetic-predisposition score was calculated on the basis of 32 BMI-associated genetic elements (loci). The intake of sugar-sweetened beverages was examined prospectively in relation to BMI. Regression models were also fitted at baseline for each of the 2 cohorts studied. An extract of table 1, with some descriptives of the variables investigated, and an extract of table 2 in the paper, with results from these 2 cohorts, is shown below.

Table 1. Baseline Characteristics of the Participants in the Three Cohorts, According to Intake of Sugar-Sweetened Beverages.*

Characteristic	Servings of Sugar-Sweetened Beverages				P Value†
	<1 per Month	1–4 per Month	2–6 per Week	≥1 per Day	
NHS cohort					
Participants — no.	2843	1776	1496	819	
Age — yr	48.7±6.5	48.0±6.7	46.9±6.9	45.8±6.8	<0.001
BMI‡	24.5±4.6	24.5±4.7	24.9±4.8	25.2±5.7	<0.001
Alcohol consumption — g/day	7.7±11.1	5.8±9.9	5.5±8.9	5.3±9.7	<0.001
Current smoking — no. (%)	722 (25.4)	361 (20.3)	316 (21.1)	248 (30.3)	0.58
Physical activity — MET-hr/wk§	15.5±18.3	13.4±17.8	13.1±16.2	12.6±17.0	<0.001
Time spent watching television — hr/wk	13.4±12.0	13.4±11.5	13.5±11.2	13.6±12.0	0.67
Total energy intake — kcal/day	1460±457	1580±479	1690±475	1840±524	<0.001
Alternative Healthy Eating Index score¶	30.7±9.1	28.6±8.7	28.2±8.4	27.5±8.0	<0.001
Artificially sweetened beverages — servings/day	0.61±0.99	0.32±0.71	0.26±0.55	0.35±0.74	<0.001
Genetic-predisposition score‖	29.3±4.0	29.1±4.0	29.0±4.0	29.1±4.0	0.15

* Plus–minus values are means ±SD. Baseline data were from 6934 women in the Nurses' Health Study (NHS, 1980), 4423 men in the Health Professionals Follow-up Study (HPFS, 1986), and 21,740 women in the Women's Genome Health Study (WGHS, 1992). Physical activity was assessed in 1986 for the NHS cohort. Television watching was assessed in 1992 for the NHS cohort and in 1988 for the HPFS cohort.
† P values are for the trend across the four categories of intake of sugar-sweetened beverages.
‡ The body-mass index (BMI) is the weight in kilograms divided by the square of the height in meters.
§ MET denotes metabolic equivalents.
¶ Scores on the Alternative Healthy Eating Index range from 2.5 to 87.5, with higher scores indicating a healthier diet.[27]
‖ The genetic-predisposition score ranges from 0 to 64, with higher scores indicating a higher genetic predisposition to obesity.

1. We are interested in the relationship between 'Alcohol consumption - g/day' and categorised 'Intake of Sugar-Sweetened Beverages'. You are asked to answer the following questions based on the information provided in the extract of Table 1 from the paper.

   (a) Write down a linear population model for this relation of interest, imposing little or no restrictions on how the mean 'Alcohol consumption' changes with the categories of 'Intake of Sugar-Sweetened Beverages'. Use clear notation and explain all symbols.

   (b) Fill in estimated values for the regression coefficients of this model. Give a clear interpretation of each number and explain your calculations.

   (c) Discuss the model assumptions and check them. Whenever this is not possible, discuss what further information you would like to have at hand to be able to check that particular assumption. If an assumption is likely not fulfilled, describe the consequences on the estimation and p-values of the model.

   (d) Write down the population model for a linear regression function relating 'Alcohol consumption - g/day' to the different categories of SSB in a way that represents a possible trend. Define explicitly the coding you are using. Based on the information shown in Table 1, discuss the appropriateness of this new model for your data.

   (e) Assuming that the linear model assumptions are fulfilled for models (a) and (d), explain how you would perform a test to compare the 2 models.

   (f) Give a rough estimate of the slope in this model (indicate your approach), and explain how one might do this more carefully (no need for exact calculations here).

   (g) Write down the null hypothesis and alternative hypothesis for the corresponding p-value given in Table 1.

   (h) These authors further acknowledge as a limitation of their study that *"Measurement errors in the intake of sugar-sweetened beverages and other dietary factors are inevitable,..."*. Explain what the impact is of such errors on the estimates and conclusions of this latest analysis.

   (i) Calculate the average alcohol consumption (g/day) in the studied population.

2. Next, you are asked to look at the models described in the extract of Table 2. Furthermore, you are requested once more to consider the categorised covariate 'Intake of Sugar-Sweetened Beverages' (SSB).

   (a) Write down the multivariate linear population model 1 for the NHS cohort (line 1 in the table). Do not consider the adjustment variables. Use clear notation and explain all symbols.

Table 2. Increase in BMI per Increment of 10 Risk Alleles, According to Intake of Sugar-Sweetened and Artificially Sweetened Beverages in the NHS and HPFS Cohorts.*

Analysis	Increase in BMI				P Value for Interaction
	<1 Serving per Month	1–4 Servings per Month	2–6 Servings per Week	≥1 Serving per Day	
Sugar-sweetened beverages					
NHS cohort					
Model 1†	1.17±0.18	1.66±0.16	1.84±0.23	2.12±0.39	0.004
Model 2‡	1.18±0.17	1.56±0.16	1.78±0.22	2.03±0.38	0.008
HPFS cohort					
Model 1†	0.80±0.20	0.42±0.21	1.05±0.19	1.59±0.37	0.06
Model 2‡	0.77±0.19	0.43±0.20	1.08±0.19	1.54±0.37	0.02
Pooled cohorts§					
Model 1†	1.00±0.13	1.20±0.13	1.37±0.15	1.85±0.27	<0.001
Model 2‡	1.00±0.13	1.12±0.12	1.38±0.14	1.78±0.27	<0.001

* Plus–minus values are β coefficients ±SE. Data were derived from repeated-measures analysis for women in the NHS (five measures during the period from 1980 to 1998) and for men in the HPFS (three measures during the period from 1986 to 1998). Data on beverage intake were assessed 4 years before the assessment of BMI.
† Data were adjusted for age and source of genotyping data.
‡ Data were further adjusted for level of physical activity, time spent watching television, status with respect to current smoking, alcohol intake, Alternative Healthy Eating Index score, and total energy intake.
§ Results for the two cohorts were pooled by means of inverse-variance–weighted, fixed-effects meta-analyses.

(b) Based on the estimates for model 1 in the NHS cohort *as presented in Table 2*, estimate (some of) the regression parameters of your model to the extent possible.

(c) Clearly interpret the estimated regression coefficients of the previous question. Consider the proper scales of the covariates in your interpretation.

(d) Consider the p-value for interaction = 0.004. Give a short summary of how you would test for an interaction effect between SSB and GenPred in model 1 for the NHS cohort. In addition, if there are degrees of freedom involved in the calculation of the test statistic, compute their value. There is no need for further calculations.

(e) Pooled results for women and men are also shown in table 2. Explain briefly how the estimated effect sizes and corresponding standard errors in model 1 for the pooled cohorts relate to those of the NHS and HPFS cohort (no exact derivation necessary).

(f) From the descriptive analyses shown in Table 1 of the paper (only an extract of it is shown here), the authors state: *"..., as compared with participants with a lower intake of sugar-sweetened beverages, those with a higher intake were younger and tended to have lower levels of alcohol consumption, physical activity,... and a lower score on the Alternative Healthy Eating Index; they also had a higher total energy intake."* Describe and explain how these relations could affect the results of regression model

1 in the NHS cohort shown in Table 2. What phenomenon is playing here?

Good luck!