**Evaluation of a linear regression analysis model with the plasma albumin level as outcome and C-reactive protein level as censored predictor**

 Among the 82700 CRP measurements, 11155 (13.5%) had a recorded CRP-value of 0, which represents that the method (Modular P® [Roche, Mannheim, Germany]) cannot measure CRP <10 mg/L precisely. The following histogram shows the distribution of the CRP level in the recorded range of 0-20 mg/L:



 The number of specimens declined steadily from 10 through 20 mg/L, a trend that continued beyond 20 mg/L (data not shown). The distribution from 0 through 9 mg/L is an artefact, representing the lack of sensitivity for CRP <10 mg/L. Consequently, we allocated a random number between 0 and 1 to all CRP-values from 0 through 9 mg/L and allocated a distribution to the random numbers similar to the distribution in the range 10-19 mg/L:

|  |  |
| --- | --- |
| Range 10-19 mg/L | Range 0-9 mg/L (according to randomly allocated numbers) |
| CRP (mg/L) | No. specimens | Cumulative % | CRP (mg/L) | No. specimens | Cumulative % |
| 10 | 1237 | 12.6 | 0 | 2115 | 12.6 |
| 11 | 1159 | 24.5 | 1 | 2045 | 24.8 |
| 12 | 1075 | 35.5 | 2 | 1891 | 36.1 |
| 13 | 1063 | 46.3 | 3 | 1766 | 46.6 |
| 14 | 996 | 56.5 | 4 | 1667 | 56.5 |
| 15 | 964 | 66.3 | 5 | 1676 | 66.5 |
| 16 | 850 | 75.0 | 6 | 1465 | 75.3 |
| 17 | 821 | 83.4 | 7 | 1331 | 83.2 |
| 18 | 879 | 92.4 | 8 | 1622 | 92.9 |
| 19 | 747 | 100 | 9 | 1197 | 100 |

 In linear regression analysis, we assume that there is a linear relationship, that the errors are independent, that there is no presence of homoscedasticity, and that the residuals of the outcome variable levels should be normally distributed [1]. As independent variable in the linear regression we used transformed CRP-values by deriving the natural logarithm of the CRP-values, after shifting CRP to CRP + 1 in order to avoid negative Ln(CRP).

 We accepted the first assumption by visual inspection of Figure 3 and from the observed vs. predicted values (data not shown). From graphs showing residual vs. independent variables we concluded that there was no pattern present and that independence of errors was thus satisfied (data not shown). Although the last two assumptions are of minor concern when results from regressions are not used for prediction of individual data points [1] we nevertheless checked that residuals were normally distributed by inspection of QQ-plots for the residuals, see further below.

 We then performed linear regression analyses for all models and time periods on three different data sets: i) by randomly distributing CRP values for all CRP<10 g/L; ii) using a dummy variable (1 for CRP<10 g/L, 0 for CRP≥10 g/L); and iii) by excluding observations with CRP<10 g/L. Using only fully-observed data is generally preferred due to the otherwise possible introduction of expansion bias [2]. But if the results from the three models are materially the same we can proceed with the linear regression analysis in which we have allocated a random value to CRP in the range from 0 through 9. For the article’s Model 1, covering all time periods, we got the following results with plasma albumin as the outcome:

|  |  |
| --- | --- |
| Variable | Coefficient (95% confidence interval) |
| Random distribution of CRP 0-9 g/L(n = 49293) | Dummy variable for CRP 0-9 g/L vs. CRP≥10 g/L(n = 49293) | Exclusion of CRP 0-9 g/L(n = 38055) |
| Ln(CRP [g/L]) | -1.74 (-1.78/-1.71) | -1.57 (-1.62/-1.52) | -1.83 (-1.89/-1.77) |
| Hemoglobin (mmol/L) | 2.08 (2.03/2.12) | 2.05 (2.01-2.10) | 2.05 (1.99/2.11) |
| Male gender | -0.12 (-0.23/-0.03) | -0.11 (-0.21/-0.01) | -0.003 (-0.118/0.012) |
| Age (1 y) | -0.061 (-0.064/-0.060) | -0.058 (-0.061/-0.056) | -0.063 (-0.067/-0.060) |
| Charlson index |  |  |  |
|  1-2 | -0.12 (-0.31/0.08) | -0.08 (-0.29/0.11) | 0.17 (-0.06/0.40) |
|  >2 | 0.41 (0.26/0.57) | 0.44 (0.28/0.59) | 0.72 (0.55/0.90) |

 The differences between the three datasets’ coefficients were generally immaterial. The following figures evaluated whether the residuals were normally distributed in the three analyses:



 In each panel, the results from dataset i, ii, and iii are found from left to right. There were no differences, either for the Epanechnikov kernel density plot (above panel), standardized normal probability plot (middle panel), or the quantiles of the residuals against quantiles of the normal distribution (lower panel). Moreover, the residuals were approximately normally distributed, thus confirming that dataset i) could be used for the analyses.

 All reiterations of the above analyses for the remaining models and time periods showed immaterial differences between datasets i), ii), and iii) and all residuals were approximately normally distributed (data not shown), thus confirming that dataset i) could be used for all models and time periods.

**References**

1. **Gelman A, Hill J.** *Linear regression: the basics*. In: Gelman A, Hill J, eds. *Data Analysis Using Regression and Multilevel/Hierarchical Models*. 1 ed. New York, USA: Cambridge University Press, 2007: 31-52.

2. **Austin PC, Hoch JS.** Estimating linear regression models in the presence of a censored independent variable. *Stat Med* 2004; **23**: 411-429.