Statistical analysis plan: Predicting body fat proportion using anthropometric measures

Georg Heinze

July 31, 2024

1 Research question in one sentence

The percentage of body fat should be predicted using a set of available anthropometric variables.

2 Workflow

- Develop a model for percentage bodyfat on the women's data set, using at maximum 5 independent variables, allowing for nonlinear functional form
- Internally validate the model building strategy in the women's data set by bootstrapping
- Externally validate the final model by applying it to the men's data set.

3 Variables

3.1 Outcome variable

The outcome variable is Fat, which holds the percentages of bodyfat given as continuous values between 0 and 100.

3.2 Potential predictors

The potential predictors are the 18 variables as defined in the task. The predictors Waist, Height and Weight are assumed to be strongly associated with the outcome and should be preferred over all other predictors in model building. All predictors are continuous variables.

4 Model development

The multivariable fractional polynomial algorithm as implemented in the R package mfp2 should be used for model development. Fractional polynomials of order 2 should be used for prioritised predictors, and order 1 for other potential predictors as it is assumed, that predictors will most likely have a monotone association with the outcome.

1. The first model M1 will consist of Waist, Height and Weight, which will be entered into the model as FP(2) without selection.
2. Secondly, a model M2 will additionally include BMI. If this model has no better AIC than M1, do not consider BMI anymore.
3. Thirdly, starting with M2 a forward selection will be performed with all other variables, using AIC as selection criterion, to yield model M3.
4. Fourthly, if M3 now contains more than five variables, a backward selection with AIC will be conducted to reduce the model to five predictors only, yielding the final model M4.

5 Apparent validity of the model

With the final model, residual diagnostics will be performed. The ultimate goal of these diagnostics is to check whether the model could be locally biased or if the prediction error could be heterogenous.

- Check local bias: Plot residuals vs. predicted values, add LOESS smoother and 95% confidence band. Expectation if assumptions correct: LOESS smoother should be close to 0 everywhere. Possible reaction if violated: include higher-order FPs for predictors.
- Check heteroscedasticity: Plot square root of absolute residuals vs. predicted values, add LOESS smoother (with default settings) and 95% confidence band. Expectation if assumptions correct: LOESS smoother close to flat line. Possible reaction if violated: consider transformations of outcome variable or other model type (e.g. ordinal).
- Check residual nonlinear association: Plot residuals vs. each predictor of the model, add LOESS smoother and 95% confidence band. Expectation if assumptions correct: LOESS smoother close to 0 everywhere. Possible reaction if violated: consider higher-order FPs for predictors.
- Check association with excluded predictors: Plot residuals vs. each candidate predictor not in the final model, add LOESS smoother and 95% confidence band. Expectation if assumptions correct: LOESS smoother close to 0 everyhwere. Possible reaction if violated: add that predictor and re-check its possible importance for the model.

- Check normality of residuals by QQ-plots. Expectation if assumptions correct: QQ-plot indicates normal distribution of residuals. Possible reaction if violated: use robust standard errors for prediction intervals, consider transformation of outcome variable or use a different model.

Furthermore, estimate and report the prediction error (root mean squared prediction error) as $\hat{\sigma}_{app}$ and the adjusted R^2 as $\hat{R}^2_{app}$ of the model.

6 Model description

- The association of each predictor with the predicted outcome as assumed by the final model will be depicted using appropriate term plots.
- Standardized regression coefficients (given as standard deviation of partial linear predictors of each predictor) will be given to address predictor importance. The usual standardized regression coefficient is defined as $\hat{\beta}^*_j = \hat{\beta} \cdot SD(x_j)$ which is equivalent to $SD(x_j\hat{\beta}_j)$. If an FP2 was selected for a predictor, there are two parameters $(\beta_{j,1}, \beta_{j,2})$ for $x_j^{p_1}$ and $x_j^{p_2}$, and the generalized standardized regression coefficient can then be estimated as $\hat{\beta}^*_j = SD(x_j^{p_1}\hat{\beta}_{j,1} + x_j^{p_2}\hat{\beta}_{j,2})$.

7 Internal validation using bootstrapping

Compute optimism-corrected estimates of the prediction error and of the R^2:

- Set $B = 1000$.
- Draw B resamples of the training set with replacement.
- Develop a model $M4^{(b)}$ on each resample b following the four steps outlined under 'model development'.
- Compute apparent measures $\hat{\sigma}^{(b)}_{app}$ and $\hat{R}^{2(b)}_{app}$
- Apply the model $M4^{(b)}$ to the observations that were not included in the bootstrap resample b and compute $\hat{\sigma}^{(b)}_{test}$ and $\hat{R}^{2(b)}_{test}$
- Estimate the optimism as $O_{PE} = Av(\hat{\sigma}^{(b)}_{app}) - Av(\hat{\sigma}^{(b)}_{test})$ and $O_{R^2} = Av(\hat{R}^{2(b)}_{app}) - Av(\hat{R}^{2(b)}_{test})$, where $AV(\cdot)$ denotes the mean.
- Compute optimism-corrected values of the prediction error and the R^2 as $\hat{\sigma}_{app} - O_{PE}$ and $\hat{R}^{2(b)}_{app} - O_{R^2}$.

Compute shrinkage factor of the model:

- Apply each model $M4^{(b)}$ to the full training set to compute linear predictors.

- Regress the observed outcomes on the linear predictors and store the slope of this regression as $\hat{\beta}_{LP}^{(b)}$.
- Compute the shrinkage factor as $Av(\hat{\beta}_{LP}^{(b)})$ and its standard error as $SD(\hat{\beta}_{LP}^{(b)})$.
- If the shrinkage factor is lower than 0.95, the final model will be recalibrated using the computed shrinkage factor and adjusting the intercept accordingly before external validation.

Investigate model uncertainty:

- Report selection frequencies of each predictor across the bootstrap resamples, and for each predictor describe the bootstrap frequencies of the selected FP powers.
- Report selection frequency of the predictor set of the final model.
- Graphically describe the bootstrapped predictor-outcome association as spaghetti plots and as mean and 2.5th and 97.5th percentiles. In this plot, each bootstrap sample contributes one line describing the predictor-outcome association. Compare with the plots under section *Model description.*

8 External validation

Use M4 and apply it to the men's data set to obtain predictions.

- Compute the prediction error and the R^2 at external validation.
- Compute calibration slope and calibration intercept by regressing the outcome variable on the linear predictors and by regressing the outcome variable on offsets defined by the linear predictors, respectively.
- Construct a calibration plot, which plots the observed body fat percentages in men to the predicted percentages, adding a LOESS smoother (95% confidence band) and the diagonal reference line. (This plot is equivalent to plotting residuals vs. predicted values.)
- Create residual plots as in section *Apparent validity of the model* to investigate how the model fits to men's data, focusing on local bias, heteroscedasticity, residual nonlinear association and association with excluded predictors.

If the analysis reveals a strong miscalibration, restrict the age range of the men's data set to that of the women's data set and repeat all steps of external validation to check if the model would at least fit well to men of similar age.