1. Marginal and conditional fits (written): Consider an outcome vector $y$ and centered predictor vectors $x_1$ and $x_2$. Recall the conditional model is $E[y|x_1, x_2] = \alpha_0 + \alpha_1 x_1 + \alpha_2 x_2$ and the marginal model is $E[y|x_1] = \beta_0 + \beta_1 x_1$.

(a) Find a formula for the standard error of $\hat{\alpha}_1$ based on $x_1$, $x_2$ and $RSS$.

(b) For the case that $x_1$ and $x_2$ are centered and $x_1 \cdot x_2 = 0$, find a formula for the standard error of $\hat{\alpha}_1$ based on dot products of $x_1$, $x_2$ and $y$ (Hint: Use 2.(c) from Homework 4).

(c) For the same case, under what conditions will the $t$-statistic for $\beta_1$ be smaller than that of $\alpha_1$? Under what conditions will it be bigger?

(d) In what ways does the fact that $x_1$ and $x_2$ are centered matter for these calculations (Hint: Try a numerical example, and/or refer to 1.(a) from Homework 4).

2. Glucose prediction (Rmd): This exercise concerns analysis of the datasets Pima.tr and Pima.te in the MASS package. Load the data and read the help-files to understand the variables. In this exercise, you will build and analyze predictive models of glucose as a function of the other variables. You will need to convert the type variable to a zero-one numeric binary variable.

(a) Using the Pima.tr dataset, compute the correlations between glucose and each of the other variables. Comment on the magnitude of the correlations, and which variables you think might be important for prediction. Also compute the correlations among the other variables, and comment on which pairs have the highest correlations.

(b) Again using the Pima.tr dataset, fit a linear model with glucose as the outcome and all of the other variables as predictors. Make a summary table of the estimated coefficients, standard errors, $t$-values and $p$-values. Also report $\hat{\sigma}^2$ and $R^2$.

(c) One by one, iteratively remove any variables from the model that have $p$-values greater than 0.10. Report a summary table, $\hat{\sigma}^2$ and $R^2$ and compare to the results in (b).
(d) Now make predictions for glucose in the Pima.te dataset, using the estimated model coefficients you obtained in (b) and (c) using the Pima.tr dataset. Compute the mean squared error of prediction $\sum_i (y_i - x_i^T \hat{\beta})^2 / n$ for each of the two models, where $i$ is a person in the Pima.te dataset and $\hat{\beta}$ are the coefficients from one of the two models, which were estimated using the Pima.tr dataset. Which model makes better predictions, the full model from (b) or the reduced model from (c).