# A complete real-life problem on linear regression:

# A reproduction to the prostate cancer example.

- Download the prostate cancer dataset from the book website.
- Standardize the data using two different methods (centering followed by sample variance, and min-max to scale it to $[-1, 1]$) and do a scatter plot for each.
- Devide the data, randomly, to two- and one-thirds (for training and testing). Build three full linear models, one on the original data and two on the two versions of standardization above. Make sure that your prediction vector on the testing set is identical for the three models. What is the estimated error $err_{\mathbf{tr}}$?
- Re-model the problem using all the variables but with regularization using rdige regression (use the centered model). Use the testing set to select the best $\lambda$ by plotting $err_{\mathbf{tr}}(\lambda)$ vs. $\lambda$.

  Hint: for just one arbitrary value of $\lambda$, use the non-centered model with regularization (check the lecture notes) and make sure that the prediction is the same as the centered model.
- Use the centered model, the two-thirds of the dataset (as a training set), and 10-$K$ CV to choose the best $\lambda$ (and accordingly the best $\beta$). Train youTest your selected model on the testing set (the one-third of the dataset) to find a final estimate of $err$