BIOSTAT/STAT 570: Coursework 1

To be submitted to the course canvas site by 1:30pm Monday 11th October, 2021.

The data we analyze were collected on $n = 97$ men before radical prostatectomony and we take as response the log of prostate specific antigen (PSA) which was being forwarded in the paper as a preoperative marker (to predict the clinical stage of cancer). We will consider two covariates for modeling log PSA (`lpsa`) on patient i: log(weight) (`lweight`, x_{i1}) and seminal vesicle invasion (`svi`, $x_{i2} = 0/1$) and their interaction $x_{i3} = x_{i1} \times x_{i2}$. All logs are base e.

We will examine the model for PSA (ng/ml) with main effects due to log(prostate weight) (prostate weight in gm) and SVI, and their interaction.

We let Y_i represent log PSA, and (x_{i1}, x_{i2}, x_{i3}), the covariates, for individual i, $i = 1, \ldots, n = 97$. We fit the model

$$y_i = \beta_0 + \sum_{j=1}^{3} x_{ij}\beta_j + \epsilon_i,$$

$i = 1, \ldots, n$ in `R`, using least squares, and the output below (which has been edited slightly) was produced.

The computation part

1. Using `R`, reproduce every number in the handout using matrix and arithmetic operations.

The interpretation part: imagine this part will be read by a non-statistician

1. Based on the fitted model, provide an informative plot that summarizes the association between log PSA, log weight and SVI.
2. Give interpretations of each of the parameters β_j, $0 = 1, \ldots, 3$.
3. Suppose we wish to interpret the coefficients on the original scale for PSA and weight. Write down the model in terms of the variables on their original scale, and hence, provide interpretations of the associations between PSA, weight and SVI.

   You may find it useful to repeat the plot you created in the first part, for the variables on their original scale.

The assumptions part

State the assumptions that are required valid for:

1. An unbiased estimate of β_j, $j = 0, \ldots, 3$.

2. An accurate estimate of the standard error of β_j, $0 = 1, \ldots, 3$.

3. Accurate coverage probabilities for 100$(1-\alpha)$% confidence intervals of the form

   $$\widehat{\beta}_j \pm \widehat{\text{var}}(\widehat{\beta}_j)^{1/2} \times z_{1-\alpha/2},$$

   where $z_{1-\alpha/2}$ represents the $(1-\alpha/2)$ quantile of an $N(0,1)$ random variable.

4. Accurate coverage probabilities for 100$(1-\alpha)$% confidence intervals of the form

   $$\widehat{\beta}_j \pm \widehat{\text{var}}(\widehat{\beta}_j)^{1/2} \times t_{n-4}(1-\alpha/2),$$

   where $t_{n-4}(1-\alpha/2)$ represents the $(1-\alpha/2)$ quantile of a standard Student's t random variable on $n-4$ degrees of freedom.

5. An accurate prediction for an *observed* outcome at $\boldsymbol{x} = \boldsymbol{x}_0$.

```
library(lasso2)
data(Prostate)
attach(Prostate)
y <- Prostate$lpsa
lweight <- Prostate$lweight
svi <- Prostate$svi
lmod <- lm(y~svi+lweight+svi:lweight)
summary(lmod)
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -0.5965     0.7127  -0.837 0.404727
svi           4.1826     2.4149   1.732 0.086591 .
lweight       0.7541     0.1946   3.876 0.000198 ***
svi:lweight  -0.7197     0.6425  -1.120 0.265521

Residual standard error: 0.8969 on 93 degrees of freedom
```