MA 575, Linear Models : Homework 8

Problem 8.1

Question 8.1.1

Let e_i be the projection vector on the i^{th} dimension. By noticing that $h_{ij} = e_i' H e_j$, one gets :

$$h_{ij} = e_i' X(X'X)^{-1}X'e_j = (X'e_i)'(X'X)^{-1}(X'e_j) = x_i'(X'X)^{-1}x_j$$

where x_j represents all the predictor's values for the j^{th} experiment.

Using the same logic, for h_{ji} and by reminding the reader that $H = H^T$, one easily obtains :

$$h_{ij} = x_i'(X'X)^{-1}x_j = x_j'(X'X)^{-1}x_i = h_{ji}$$

Question 8.1.2

Let's prove that $\frac{1}{n} \leq h_{ii}$. To do so, it is reminded to the reader that :

$$h_{ii} = \frac{1}{n} + (x_i^* - \bar{x})'(\mathcal{X}'\mathcal{X})^{-1}(x_i^* - \bar{x})$$

The above expression can also be written as :

$$h_{ii} = \frac{1}{n} + \|A(x_i^* - \bar{x})\|^2 \qquad \text{where } A = (\mathcal{X}'\mathcal{X})^{-1/2}$$

Hence :

$$h_{ii} \geq \frac{1}{n}$$

Let's now prove that $h_{ii} \leq \frac{1}{r}$. To do so, it is reminded to the reader that : $H^2 = H$. Therefore, the diagonal term of H can be written as :

$$
\begin{aligned}
h_{ii} &= \sum_{j=1}^{n} h_{ij} h_{ji} \\
&= \sum_{j=1}^{n} h_{ij}^2 && \text{(as } h_{ij} = h_{ji}\text{)} \\
&= \sum_{j \in J_i} h_{ij}^2 + \sum_{j \notin J} h_{ij}^2 && \text{where } J_i = \{ j \in 1..n \mid x_i = x_j \} \\
&= \sharp \{J_i\} \, h_{ii}^2 + \sum_{j \notin J} h_{ij}^2 && \text{(if } x_i = x_j \text{ then } h_{ij} = h_{ji} = h_{ii}\text{)} \\
&= r h_{ii}^2 + \sum_{j \notin J} h_{ij}^2 \\
&\geq r h_{ii}^2
\end{aligned}
$$

Hence,

$$h_{ii} \leq \frac{1}{r}$$

Problem 6.13

Question 6.13.1

From Figure 1, one can see that :

- women present (globally) a lower salary than men but also that they have worked for the college a smaller amount of years in the current rank (which could explain the difference of salary);
- there will be different regression lines for men and women in the case $Salary$ vs $YSdeg$. It seems that women get less paid for an equivalent number of years since the obtention of degree. However, the $YSdeg$ is not equivalent to the number of years of experience. As one knows, women tend to get time of work when getting children. It would therefore not be surprising for women to have less years of experience than men for an equivalent $YSdeg$;
- women tends to get less advanced position for an equivalent $YSdeg$.

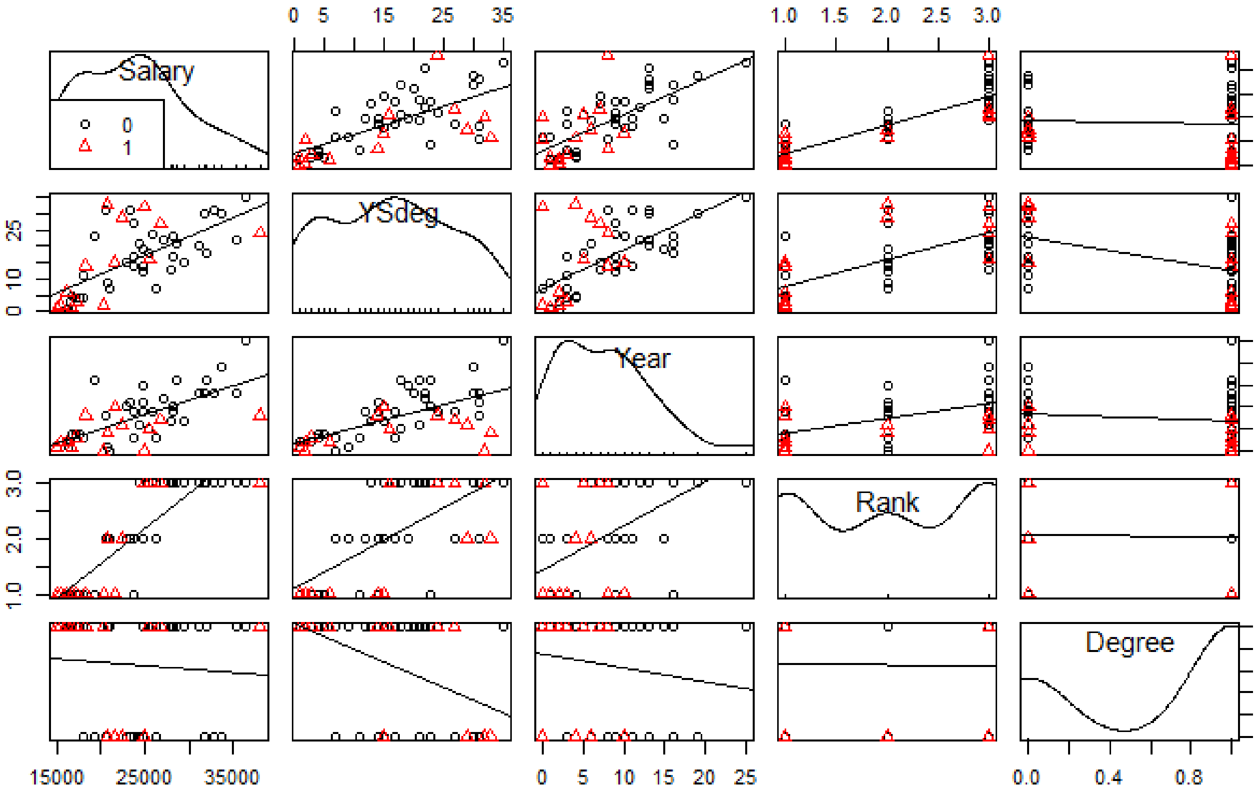

Figure 1

```
library(car)

salary = read.table(file="salary.txt",header=TRUE)
attach(salary)

scatterplotMatrix(~Salary+YSdeg+Year+Rank+Degree|Sex,smoother=FALSE)
```

Question 6.13.2

Let $Y_i = \mu + \beta Sex_i + \epsilon_i$ where $i \in 1..n$ be a model for the salary depending on the sex of the individual. An other way to write the latter expression is as follow :

$$Y_i = \begin{cases} \mu & \text{if the } i^{th} \text{ individual is a man} \\ \mu + \beta & \text{if the } i^{th} \text{ individual is a woman} \end{cases}$$

Therefore, testing the fact that men and women have the same mean salary is equivalent to testing $\hat{\beta} = 0$. The alternative hypothesis that one should test is the following : $\hat{\beta} < 0$ which is equivalent to saying that women are paid less. The one sided $p - value$ is obtained as follow :

$$
\begin{aligned}
\text{Two sided } p-value &= \mathbb{P}(T \le -|t| \cup T \ge |t|) && \text{where } T \sim \tau(n-2) \text{ and } t \text{ is equal to } \tfrac{\hat{\beta}}{\widehat{sd}(\beta)} \\
&= \mathbb{P}(T \le -|t|) + (T \ge |t|) \\
&= 2\mathbb{P}(T \le -|t|) \\
&= 2\mathbb{P}(T \le \frac{\hat{\beta}}{\widehat{sd}(\beta)}) && (\text{as } \hat{\beta} < 0) \\
&= 2(\text{one sided } p-value)
\end{aligned}
$$

By using R-code below, one obtains 0.0706 (resp. 0.0353) for the value of the two-sided (resp. one sided) $p-value$. This test tends to say that women are paid less than men in that college.

```
m1 = lm(Salary~Sex) # Other possible code m1 = lm(Salary~factor(Sex))
summary(m1)

Call:
lm(formula = Salary ~ Sex)

Residuals:
    Min      1Q  Median      3Q     Max
-8602.8 -4296.6  -100.8  3513.1 16687.9

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)    24697        938  26.330   <2e-16 ***
Sex            -3340       1808  -1.847   0.0706 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 5782 on 50 degrees of freedom
Multiple R-squared:  0.0639,Adjusted R-squared:  0.04518
F-statistic: 3.413 on 1 and 50 DF,  p-value: 0.0706
```

Question 6.13.3

In this question, it is asked to compare the impact of the rank upon the salary when adjusted to the variables $Years$, $YSdeg$ and $rank$. To do so, we propose the following model :

$$Y_{ij} = \mu + \beta_{3j} + \beta_1 Year_i + \beta_2 YSdeg_i + \epsilon_{ij}$$

where $j \in 1..3$, $i \in 1..n_j$ and n_j is the number of individuals ranked j.

By shifting the rank within the range of value $j = 0..2$ (which is what R does), our model can be written as follow :

$$Y_{ij} = \mu_0 + \alpha_j + \beta_1 Year_i + \beta_2 YSdeg_i + \epsilon_{ij}$$

where $\mu_0 = \mu + \beta_{31}$, $j \in 1..2$, $i \in 1..n_j$ and n_j is the number of individuals ranked j.

The latter model can also be presented as follow:

$$Y_i = \begin{cases} \mu_0 + \beta_1 Year_i + \beta_2 YSdeg_i + \epsilon_i & \text{if the } i^{th} \text{ individual has a rank 1} \\ \mu_0 + \alpha_1 + \beta_1 Year_i + \beta_2 YSdeg_i + \epsilon_i & \text{if the } i^{th} \text{ individual has a rank 2} \\ \mu_0 + \alpha_2 + \beta_1 Year_i + \beta_2 YSdeg_i + \epsilon_i & \text{if the } i^{th} \text{ individual has a rank 3} \end{cases}$$

Therefore, testing the hypothesis that all rank have the average salary is same as testing $\alpha_1 = \alpha_2 = 0$. This is done using an ANOVA. The $p - value$ is equal to $6.544e - 10$. The test reject the null hypothesis.

```
m2 = lm(Salary~factor(Degree) + Year + YSdeg + factor(Rank))
anova(m2)

Analysis of Variance Table

Response: Salary
               Df    Sum Sq   Mean Sq  F value    Pr(>F)
factor(Degree)  1   8681649   8681649   1.4902    0.2284
Year            1 869721395 869721395 149.2842 4.807e-16 ***
YSdeg           1 235224812 235224812  40.3754 8.512e-08 ***
factor(Rank)    2 404108665 202054333  34.6818 6.544e-10 ***
Residuals      46 267993336   5825942
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

To test that the average salary depending on sex is the same for each rank, the interaction $Sex \times Rank$ is added to the previous model. For each rank, the sex will have no impact on the model of the added parameter is equal to 0. Since the $p - values$ of the t-test are all greater than 5%, the hypothesis that there is a difference of salary between men and women for an equivalent rank is not rejected when we consider the two-sided alternative hypothesis. Because the $t - statistic$ is positive, the one sided $p - value$ can be expressed (using same logic than previously) as follow :

$$\begin{aligned} \text{Two sided } p_value &= 2\mathbb{P}\left(T \leq -\frac{\hat{\beta}}{\widehat{sd}(\beta)}\right) \\ &= 2\mathbb{P}\left(T \geq \frac{\hat{\beta}}{\widehat{sd}(\beta)}\right) \\ &= 2\left(1 - \mathbb{P}(T \leq \frac{\hat{\beta}}{\widehat{sd}(\beta)})\right) \\ &= 2\left(1 - \text{one sided } p - value\right) \end{aligned}$$

Therefore the one-sided $p - value$ is $(1 - 0.0733/2 = 0.96335)$ for the third rank seems to imply that full professor women are better paid than full professor men.

```
m3 = lm(Salary~factor(Degree) + Year + YSdeg + factor(Rank) + Sex:factor(Rank))
summary(m3)

Call:
lm(formula = Salary ~ factor(Degree) + Year + YSdeg + factor(Rank) +
    Sex:factor(Rank))
```

```
Residuals:
    Min      1Q  Median      3Q     Max
-4230.0 -1352.4  -140.4   765.4  7937.5

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)       17504.67    1284.95  13.623  < 2e-16 ***
factor(Degree)1   -1501.51    1029.80  -1.458   0.1521
Year                522.14     105.53   4.948 1.20e-05 ***
YSdeg              -148.61      86.83  -1.711   0.0942 .
factor(Rank)2      5119.00    1168.49   4.381 7.46e-05 ***
factor(Rank)3     10539.36    1426.21   7.390 3.52e-09 ***
factor(Rank)1:Sex   444.30    1153.54   0.385   0.7020
factor(Rank)2:Sex   942.58    2194.85   0.429   0.6697
factor(Rank)3:Sex  2954.54    1609.28   1.836   0.0733 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2402 on 43 degrees of freedom
Multiple R-squared:  0.8611,Adjusted R-squared:  0.8352
F-statistic: 33.32 on 8 and 43 DF,  p-value: 5.225e-16
```

Question 6.13.4

If we ignore the rank, the influence of the gender is not significant ($p - value = 0.332209$).

```
 m4 = lm(Salary~ factor(Degree) + Year + YSdeg + factor(Sex))
> summary(m4)

Call:
lm(formula = Salary ~ factor(Degree) + Year + YSdeg + factor(Sex))

Residuals:
    Min      1Q  Median      3Q     Max
-8146.9 -2186.9  -491.5  2279.1 11186.6

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)     13884.22    1639.82   8.467 5.17e-11 ***
factor(Degree)1  3299.35    1302.52   2.533 0.014704 *
Year              351.97     142.48   2.470 0.017185 *
YSdeg             339.40      80.62   4.210 0.000114 ***
factor(Sex)1    -1286.54    1313.09  -0.980 0.332209
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3744 on 47 degrees of freedom
Multiple R-squared:  0.6312,Adjusted R-squared:  0.5998
F-statistic: 20.11 on 4 and 47 DF,  p-value: 1.048e-09
```

If one adds up the rank effect, the gender still does not have a significant impact on the average salary.

```
m5 = lm(Salary~ factor(Degree) + Year + YSdeg + factor(Sex) +factor(Rank))
summary(m5)

Call:
lm(formula = Salary ~ factor(Degree) + Year + YSdeg + factor(Sex) +
    factor(Rank))

Residuals:
    Min      1Q  Median      3Q     Max
-4045.2 -1094.7  -361.5   813.2  9193.1

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)     17134.66    1197.70  14.306  < 2e-16 ***
factor(Degree)1 -1388.61    1018.75  -1.363    0.180
Year              476.31      94.91   5.018 8.65e-06 ***
YSdeg            -124.57      77.49  -1.608    0.115
factor(Sex)1     1166.37     925.57   1.260    0.214
factor(Rank)2    5292.36    1145.40   4.621 3.22e-05 ***
factor(Rank)3   11118.76    1351.77   8.225 1.62e-10 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2398 on 45 degrees of freedom
Multiple R-squared:  0.855,Adjusted R-squared:  0.8357
F-statistic: 44.24 on 6 and 45 DF,  p-value: < 2.2e-16
```

Problem 6.14

Question 6.14.1

The question is a bit unclear here but the variable Sex as to be regarded as quantitative rather than qualitative. If not, shifting the values of Sex will not affect the values of the coefficients.

To obtain the values of the new coefficients of the mean function after transforming the value of Sex ($\Rightarrow Sex_2$), one just has to notice that $Sex_2 = 2 - Sex$. One can compute them using the R-code below or derive them by replacing Sex by $2 - Sex_2$ within the mean function.

```
m6 = lm(Salary~ Year + Sex + Sex:Year)
summary(m6)

Call:
lm(formula = Salary ~ Year + Sex + Sex:Year)

Residuals:
     Min       1Q   Median       3Q      Max
-10904.0  -3150.2   -632.2   2896.8  13112.6

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  18222.6     1308.6  13.925  < 2e-16 ***
Year           741.0      126.2   5.870 3.95e-07 ***
Sex           -570.8     2297.2  -0.248    0.805
```

```
Year:Sex       169.1      387.0   0.437    0.664
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4342 on 48 degrees of freedom
Multiple R-squared:  0.4932,Adjusted R-squared:  0.4615
F-statistic: 15.57 on 3 and 48 DF,  p-value: 3.323e-07

S2 = -Sex + 2

m7 =lm(Salary~ Year + S2 + S2:Year)
summary(m7)

Call:
lm(formula = Salary ~ Year + S2 + S2:Year)

Residuals:
     Min       1Q   Median       3Q      Max
-10904.0  -3150.2   -632.2   2896.8  13112.6

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  17081.1     3996.5   4.274 9.06e-05 ***
Year          1079.1      742.4   1.454    0.153
S2             570.8     2297.2   0.248    0.805
Year:S2       -169.1      387.0  -0.437    0.664
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4342 on 48 degrees of freedom
Multiple R-squared:  0.4932,Adjusted R-squared:  0.4615
F-statistic: 15.57 on 3 and 48 DF,  p-value: 3.323e-07
```

Question 6.14.b

The same reasoning stands here by noticing that the new values Sex_3 of Sex are defined as follow : $Sex_3 = 2Sex - 1$.

```
S3 = 2*Sex - 1

m8 =lm(Salary~ Year + S3 + S3:Year)
summary(m8)

Call:
lm(formula = Salary ~ Year + S3 + S3:Year)

Residuals:
     Min       1Q   Median       3Q      Max
-10904.0  -3150.2   -632.2   2896.8  13112.6

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 17937.21    1148.62  15.616  < 2e-16 ***
Year          825.55     193.48   4.267 9.27e-05 ***
```

```
S3           -285.38    1148.62  -0.248    0.805
Year:S3        84.53     193.48   0.437    0.664
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4342 on 48 degrees of freedom
Multiple R-squared:  0.4932,Adjusted R-squared:  0.4615
F-statistic: 15.57 on 3 and 48 DF,  p-value: 3.323e-07
```