STAT-S631

Assignment 9

John Koo

Problem 1

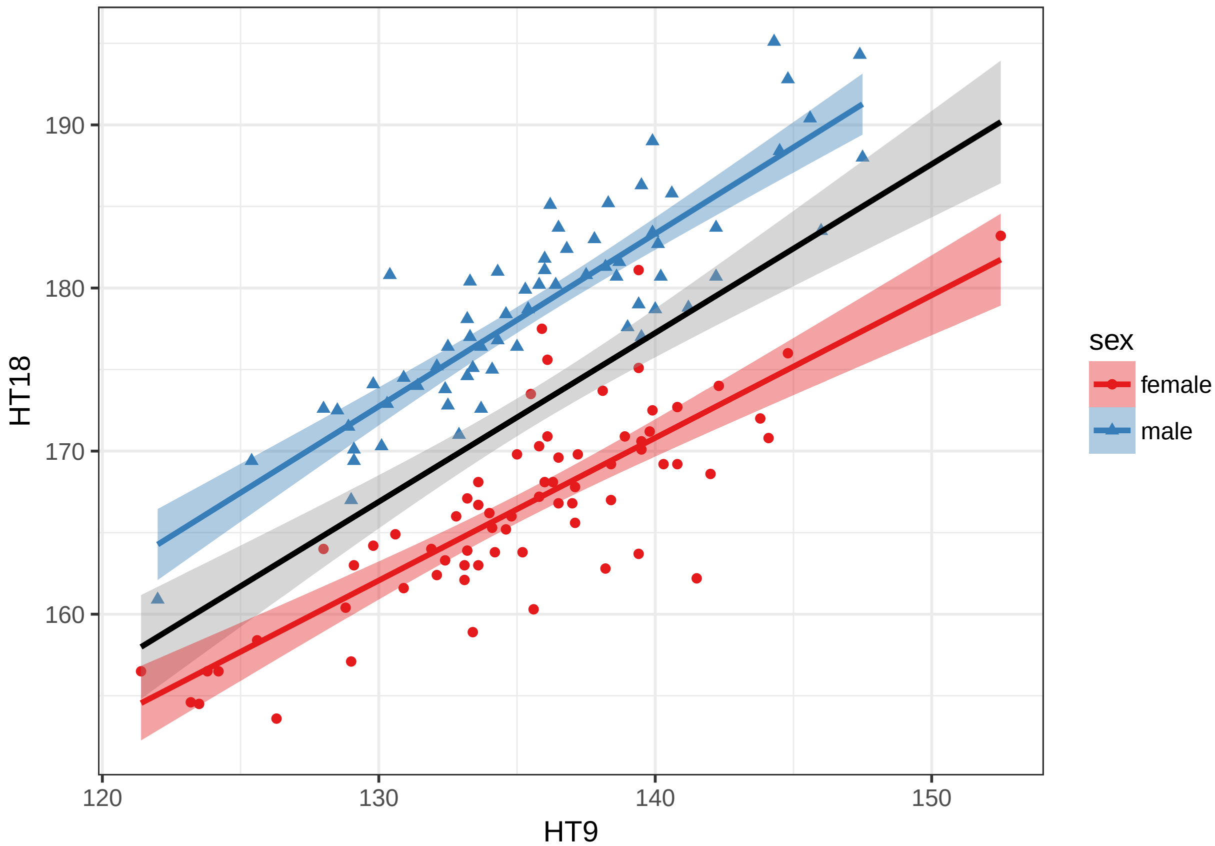
[From ALR 5.14]

```
dp <- loadNamespace('dplyr')
import::from(magrittr, `%>%`, `%<>%`)
import::from(car, Anova)
import::from(ggplot2,
             ggplot,
             geom_point, stat_smooth,
             aes,
             scale_colour_brewer, scale_fill_brewer,
             theme_set, theme_bw)
theme_set(theme_bw())

bgsall.df <- alr4::BGSall %>%
  dp$mutate(sex = dp$if_else(Sex == 0, 'male', 'female'))
```

Part 1

```
ggplot(bgsall.df) +
  geom_point(aes(x = HT9, y = HT18, shape = sex, colour = sex)) +
  stat_smooth(aes(x = HT9, y = HT18), method = 'lm', colour = 'black') +
  stat_smooth(aes(x = HT9, y = HT18, colour = sex, fill = sex),
              method = 'lm') +
  scale_colour_brewer(palette = 'Set1') +
  scale_fill_brewer(palette = 'Set1')
```



From the scatterplot, there's a fair amount of separation between the sexes, and it appears that a model containing both `HT9` and `sex` but not an interaction between the two would be the most appropriate.

Part 2

```
ht9.model <- lm(HT18 ~ HT9, data = bgsall.df)
parallel.model <- lm(HT18 ~ HT9 + sex, data = bgsall.df)
full.model <- lm(HT18 ~ HT9 * sex, data = bgsall.df)

Anova(full.model)
```

```
Anova Table (Type II tests)

Response: HT18
          Sum Sq  Df  F value   Pr(>F)
HT9       3740.5   1 322.1883 < 2e-16 ***
sex       4624.0   1 398.2872 < 2e-16 ***
HT9:sex     34.4   1   2.9638 0.08749 .
Residuals 1532.5 132
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

From the F-test (which is the same as a t-test in this case since the factor only has two levels), we obtain a p-value of 0.0875, which is significant at the $\alpha = .1$ level but not at $\alpha = .05$. On the other hand, the p-value for the intercept term is significant. This test compares the model with just `HT9` vs. the model with both `HT9` and `sex` without the interaction term.

Part 3

```r
summary(parallel.model)
```

```
Call:
lm(formula = HT18 ~ HT9 + sex, data = bgsall.df)

Residuals:
     Min       1Q   Median       3Q      Max
-10.4694  -2.0952  -0.0136   1.7101  10.4467

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 36.82147    7.29177    5.05 1.43e-06 ***
HT9          0.96006    0.05388   17.82  < 2e-16 ***
sexmale     11.69584    0.59036   19.81  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.432 on 133 degrees of freedom
Multiple R-squared:  0.8516,    Adjusted R-squared:  0.8494
F-statistic: 381.7 on 2 and 133 DF,  p-value: < 2.2e-16
```

```r
confint(parallel.model, 'sexmale')
```

```
           2.5 %   97.5 %
sexmale 10.52813 12.86355
```

Problem 2

We are given:

$X = [X_1|X_2]$
$H = X(X^TX)^{-1}X^T$
$H_R = X_1(X_1^TX_1)^{-1}X_1^T$

Part a

Show $H_RX_1 = X_1$

$H_RX_1 = X_1(X_1^TX_1)^{-1}X_1^TX_1 = X_1(X_1^TX_1)^{-1}(X_1^TX_1) = X_1$

Show $HX_1 = X_1$

Consider HX. We know that $HX = X$, so we can say:

$$HX = H[x_0, x_1, x_2, ..., x_p]$$
$$= [Hx_0, Hx_1, ..., Hx_p]$$

Where x_i is the i^{th} column vector of X.

But then $HX = X = [x_0, ..., x_p]$. Therefore:

$$[Hx_0, ..., Hx_p] = [x_0, ..., x_p]$$
$$\implies Hx_i = x_i$$

Then if we consider HX_1:

$$\begin{aligned} HX_1 &= H[x_0, ..., x_q] \\ &= [Hx_0, ..., Hx_q] \\ &= [x_0, ..., x_q] \\ &= H_1 \end{aligned}$$

Show $HH_R = H_R$

$HH_R = H\big(X_1(X_1^T X_1)^{-1} X_1^T\big) = (HX_1)(X_1^T X_1)^{-1} X_1^T = X_1(X_1^T X_1)^{-1} X_1^T = H_R$

Part b

Show $H - H_R$ **is symmetric**

$H - H_R$ is symmetric iff $H - H_R = (H - H_R)^T$.

We also know that H and H_R are symmetric.

Therefore, $(H - H_R)^T = H^T - H_R^T = H - H_R$.

Show $H - H_R$ **is idempotent**

$H - H_R$ is idempotent iff $(H - H_R)^2 = H - H_R$

We know that H and H_R are idempotent.

Therefore:

$$\begin{aligned} (H - H_R)(H - H_R) &= HH - HH_R - H_R H H_R H_R \\ &= H - H_R - H_R H + H_R \\ &= H - H_R H \end{aligned}$$

Consider that H and H_R are symmetric and $HH_R = H_R$. Therefore, $H_R = H_R^T = (HH_R)^T = H_R^T H^T = H_R H$ $\implies H_R = H_R H$.

Therefore:

$$H - H_R H = H - H_R$$
$$\implies (H - H_R)^2 = H - H_R$$

Part c

$$\frac{SSreg}{\sigma^2} = \frac{RSS_R - RSS_F}{\sigma^2}$$
$$= \frac{Y^T(I-H_R)Y - Y^T(I-H)Y}{\sigma^2}$$
$$= \frac{Y^T(H-H_R)Y}{\sigma^2}$$
$$= \frac{(Y-X_1\hat{\beta}_1)^T(H-H_R)(Y-X_1\hat{\beta}_1)}{\sigma^2}$$

We know that $Y - X_1\hat{\beta}_1 \sim \mathcal{N}(0, \sigma^2(I-H_R))$. Furthermore, we know that $rank(H-H_R) = rank(H) - rank(H_R) = p+1-(p+1-q) = q$ (assuming H and H_R are full rank).

Then $\frac{SSreg}{\sigma^2} \sim \chi^2_q$ if $(\frac{H-H_R}{\sigma^2})(\sigma^2(I-H_R)) = (H-H_R)(I-H_R)$ is idempotent. But $(H-H_R)(I-H_R) = H - HH_R - H_R + H_RH_R = H - H_R - H_R + H_R = H - H_R$ which we already know to be idempotent. Therefore,

$$\frac{SSreg}{\sigma^2} \sim \chi^2_q$$

Part d

$$\hat{\sigma}^2 = \frac{RSS}{n-p-1}$$
$$= Y^T\frac{I-H}{n-p-1}Y$$

So we have to show:

$$\left(\frac{H-H_R}{\sigma^2}\right)\left(\sigma^2(I-H)\right)\left(\frac{I-H}{n-p-1}\right) = 0$$

We know that the product of the first two components is $H - H_R$. Therefore,

$$\left(\frac{H-H_R}{\sigma^2}\right)\left(\sigma^2(I-H)\right)\left(\frac{I-H}{n-p-1}\right) = (H-H_R)(I-H)\frac{1}{n-p-1}$$
$$= (H - H - H_R + H_R)\frac{1}{n-p-1}$$
$$= 0$$

Part e

We know that $\frac{SSreg}{\sigma^2} \sim \chi^2_q$ and $\frac{RSS}{\sigma^2} \sim \chi^2_{n-p-1}$. Then we know that $\frac{\frac{SSreg}{\sigma^2}/q}{\frac{RSS}{\sigma^2}/(n-p-1)} = \frac{SSreg/q}{RSS/(n-p-1)} \sim F_{q,n-p-1}$

Problem 3

[From ALR 6.4]