S631 HW8

Erik Parker

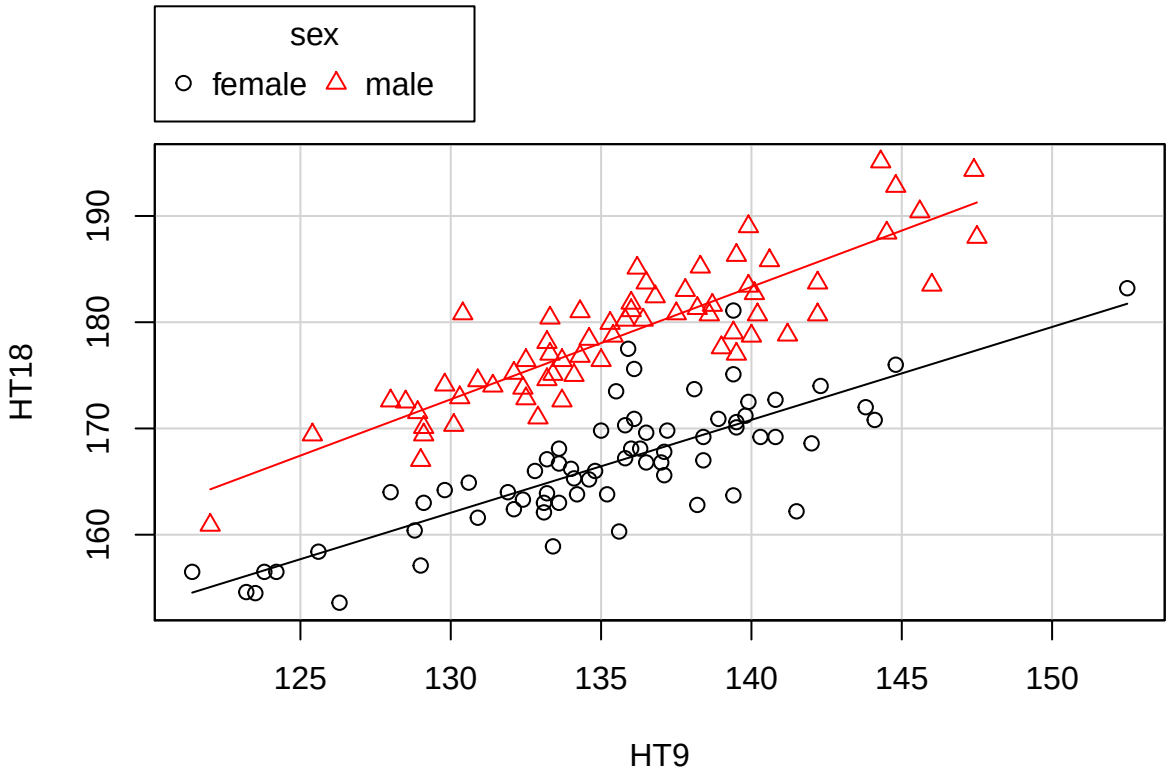
October 21, 2017

1. ALR 5.14: Using the data file *BGSall*, consider the regression of *HT18* on *HT9* and the grouping factor *Sex*.

```
rm(list = ls())

library(alr4)

Berkley <- BGSall

Berkley$sex <- ifelse(Berkley$Sex == "0", "male", "female")
```

5.14.1: Draw the scatterplot of *HT18* versus *HT9*, using a different symbol for males and females. Comment on the information in the graph about an appropriate mean function for these data.

```
scatterplot(HT18 ~ HT9 | sex, data = Berkley, smooth = FALSE, boxplots = FALSE)
```



From this plot, it seems pretty clear that there is real seperation between the male and female groups in terms of their height. The intercepts of the two lines seem to be different, with the male one higher than the female one, but the slopes of the lines appear to be the same, or very close to the same. Furthermore, there is also a clear relationship in both sexes, that as the height at age 9 increases, so too does the height at age 18. This suggests to me that a proper mean function for

these data will be one with the continuous *HT9* and the categorical *sex* as predictors, but no interaction. So, it will be of the form: $HT18\ HT9 + sex$.

5.14.2 Obtain the appropriate test for a parallel regression model.

```
mpar <- lm(HT18 ~ HT9 + sex, data = Berkley)

summary(mpar)
```

```
##
## Call:
## lm(formula = HT18 ~ HT9 + sex, data = Berkley)
##
## 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
```

Looks good! Explains quite a bit of the variation, and the coefficients for both the continuous and categorical regressors are very significant.

Assuming the parallel regression model is adequate, estimate a 95% confidence interval for the difference between males and females. For the parallel regression model, this is the difference in the intercepts of the two groups.

```
confint(mpar, level = 0.95)
```

```
##                  2.5 %    97.5 %
## (Intercept) 22.3986375 51.244301
## HT9          0.8534845  1.066628
## sexmale     10.5281335 12.863548
```

So, based on my previous model, we see here that we are 95% confident that the true coefficient obtained when we move the females to males, is within the interval of 10.528 to 12.864. This means that we are 95% confident that the true increase in height seen in 18 year olds in this study is between 10.528 cm and 12.864 cm when we move from females to males.

2. Show the following equalitites:

Work in notebook, insert picture?

ALR 6.4 and in addition, using the full model, perform the test

$$H_0 : \beta_{02} - \beta_{03} = 14 \quad and \quad \beta_{12} + \beta_{13} = 0.2$$

with H_A : at least one equality doesn't hold. Show your work. In addition, how could you interpret this test?

6.4: With the UN data, consider testing *NH: lifeExpF~log(ppgdp) + group:log(ppgdp) AH : lifeExpF~group + log(ppgdp) + group:log(ppgdp)* The AH model is the most general model given at (6.10), but the NH was not given previously.