S631 HW9

Erik Parker

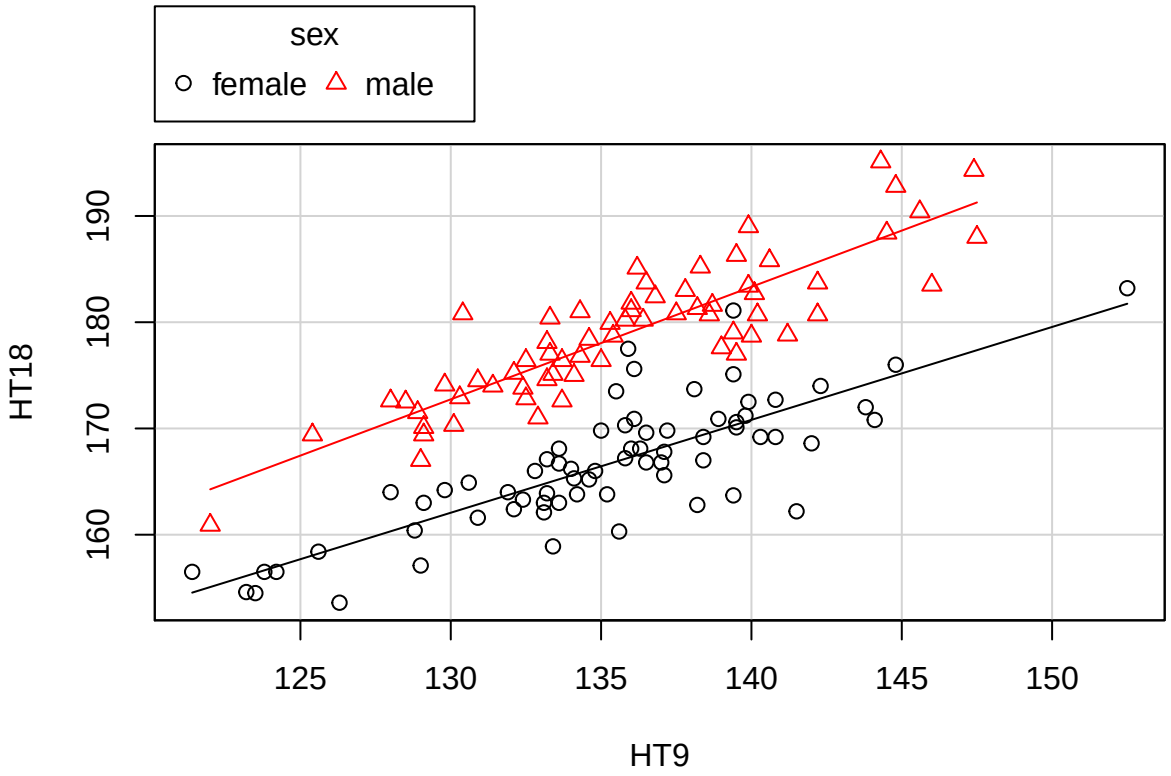
October 29, 2017

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.

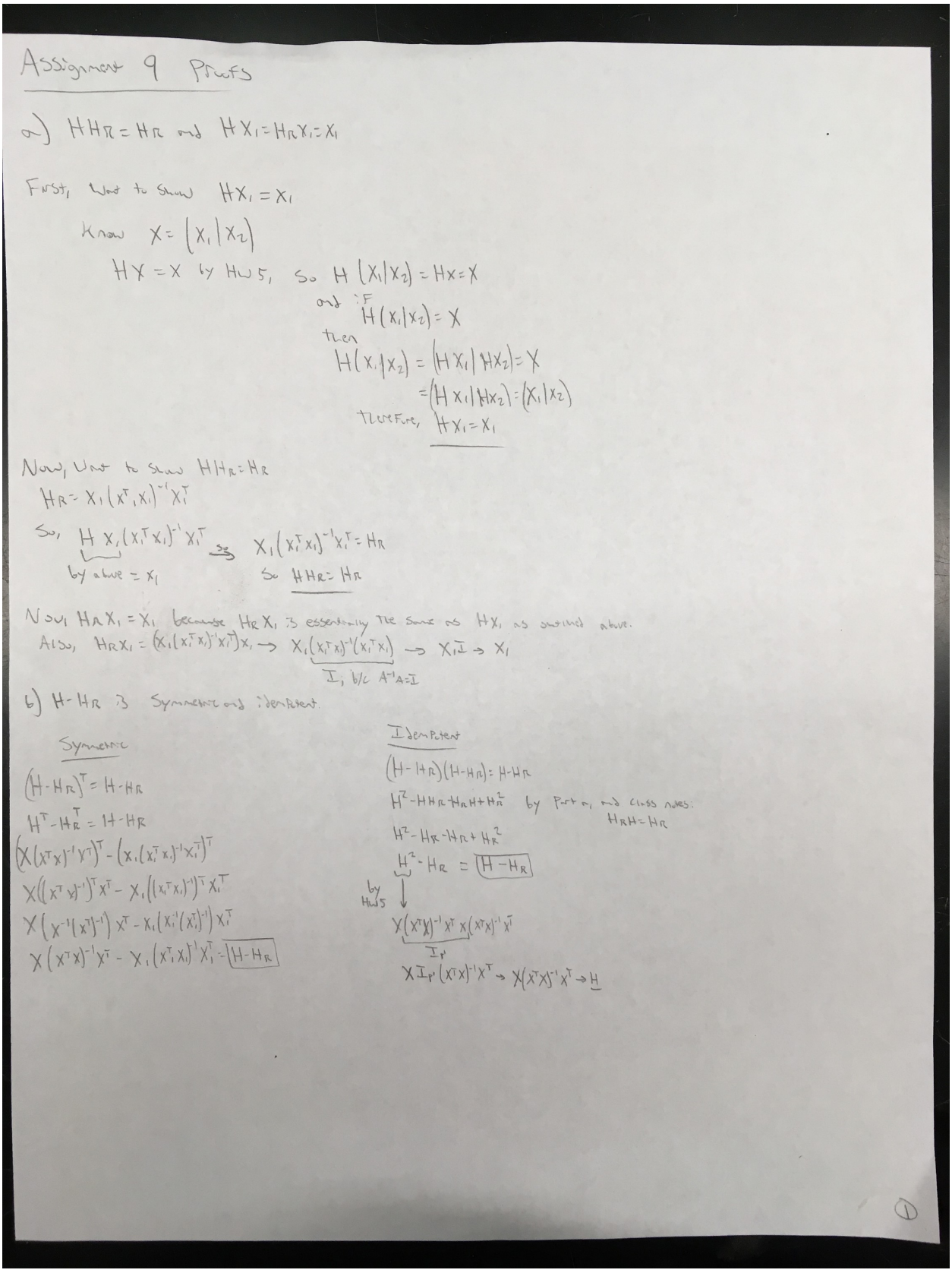


Figure 1: Handwritten proofs pg. 1

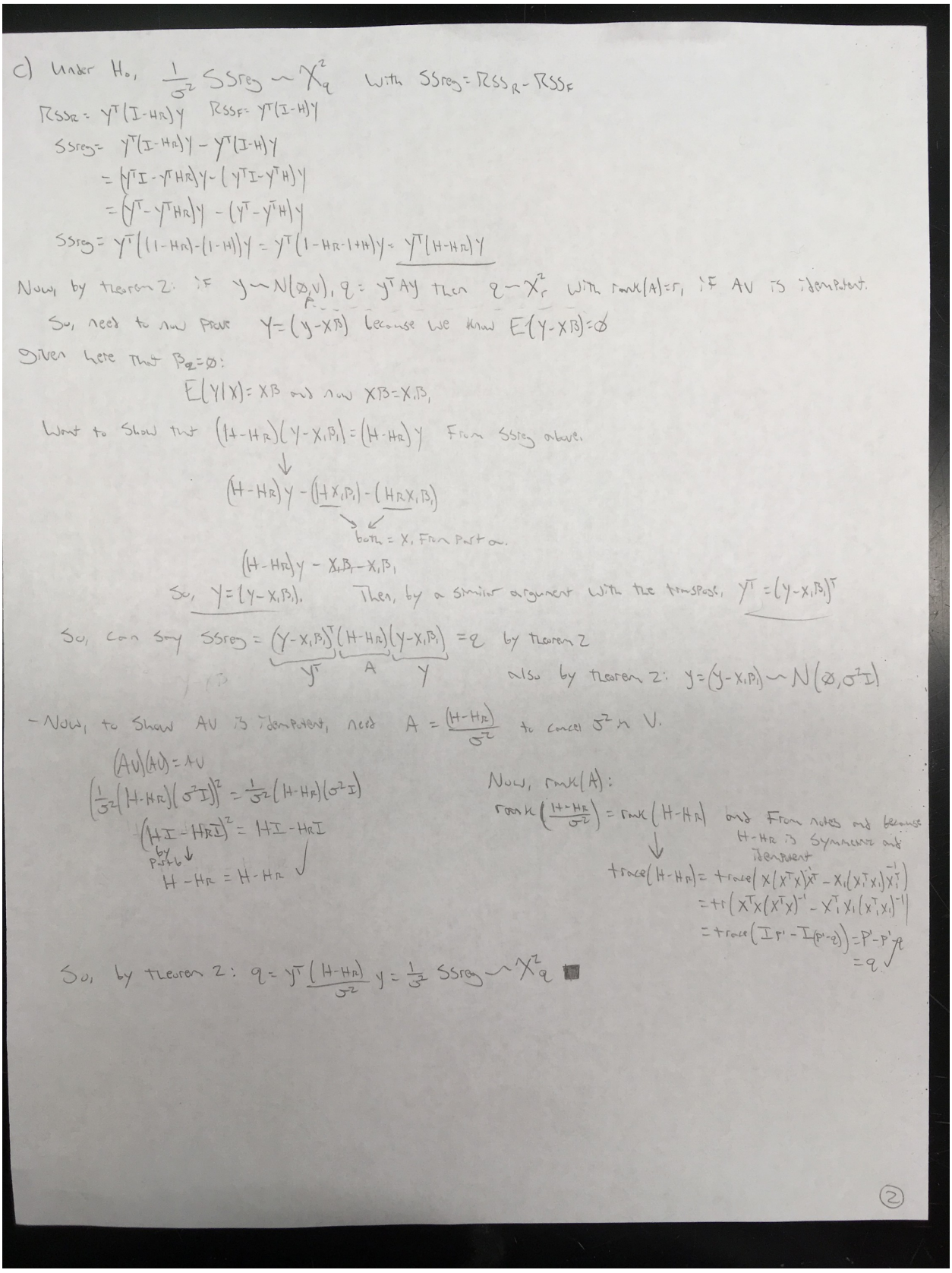


Figure 2: Handwritten proofs pg. 2

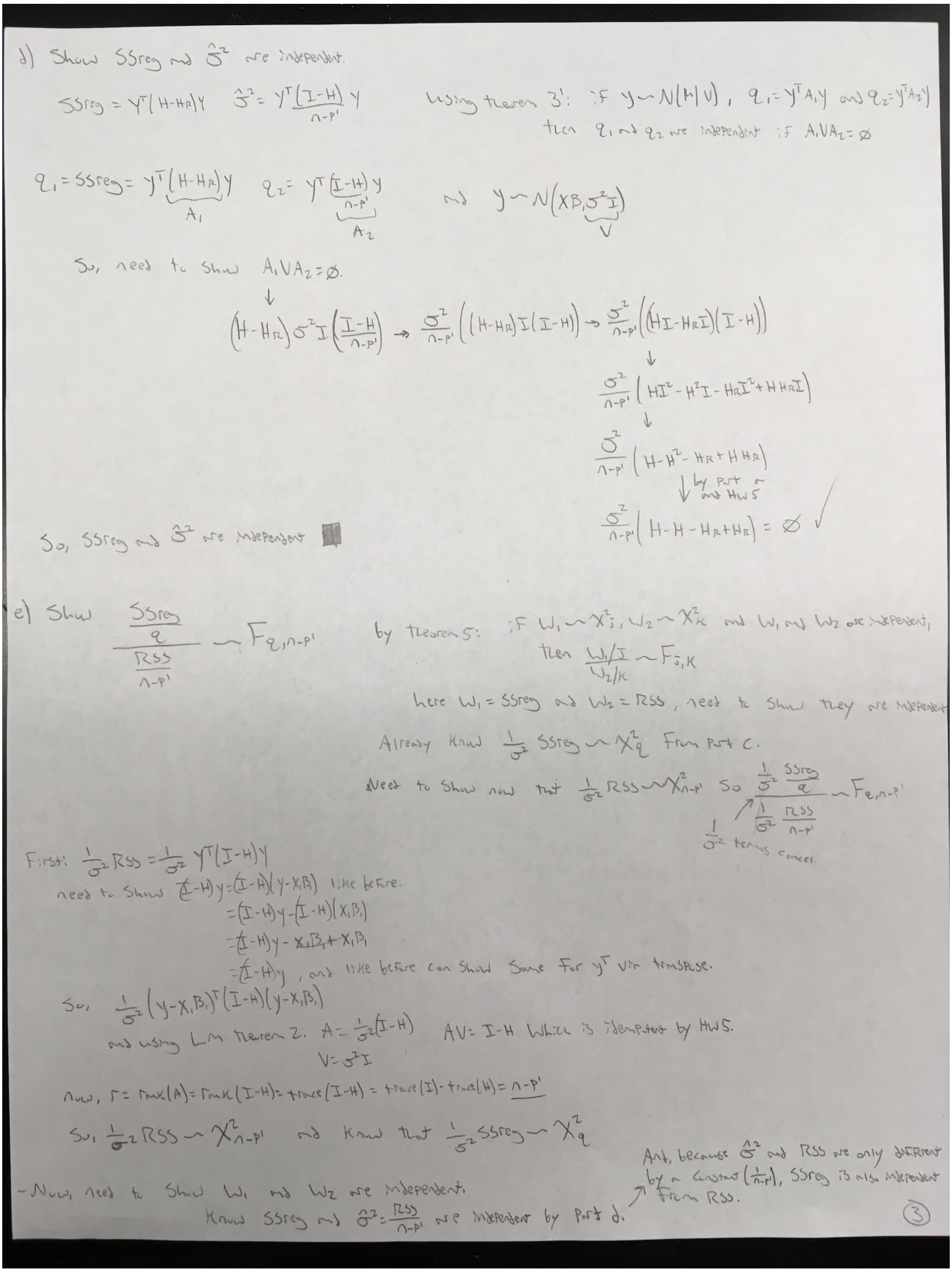


Figure 3: Handwritten proofs pg. 3

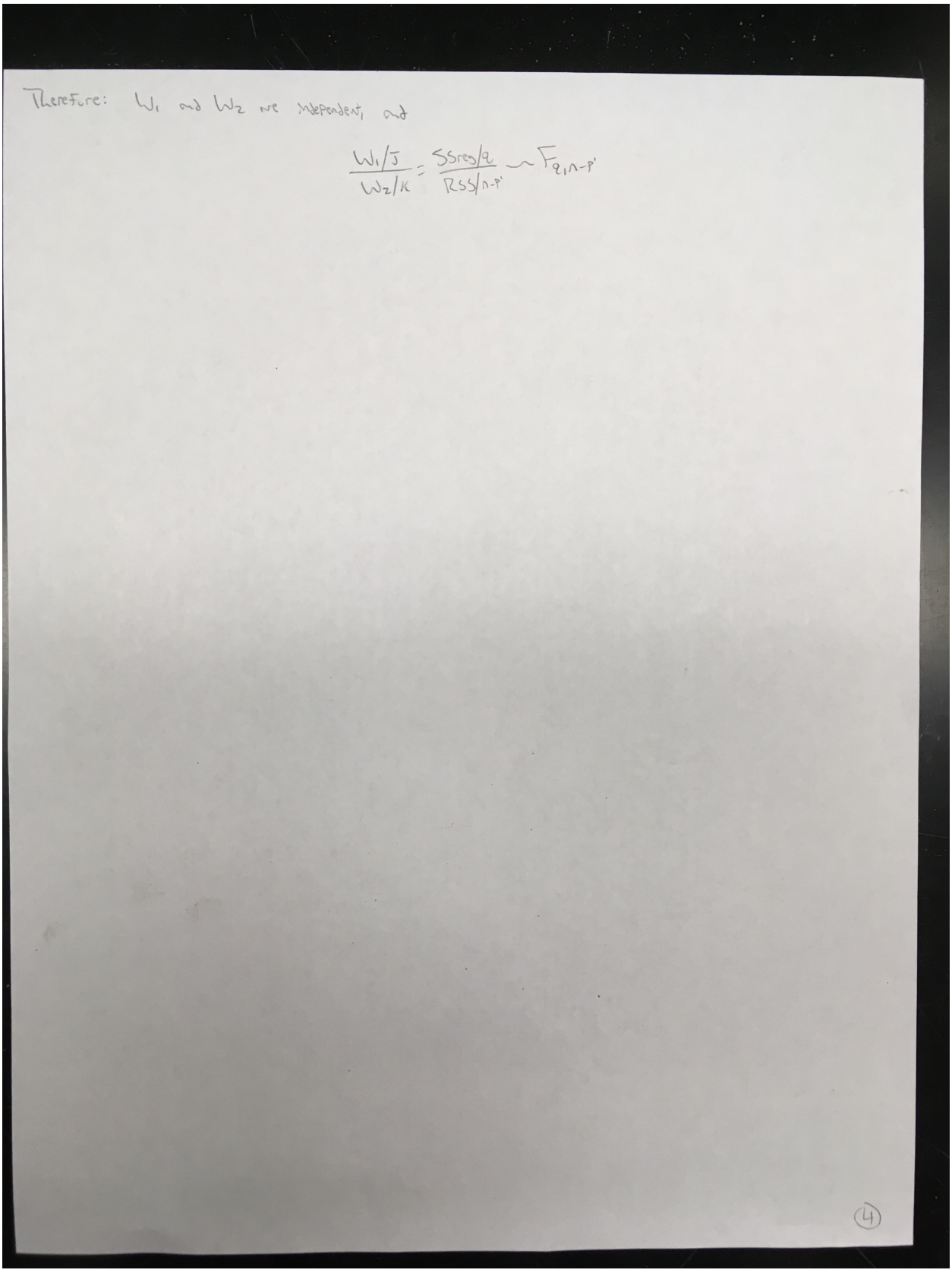


Figure 4: Handwritten proofs pg. 4

ALR 6.4

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.

```
un <- UN11

m1 <- lm(lifeExpF ~ log(ppgdp) + group:log(ppgdp), data = un)
m2 <- lm(lifeExpF ~ group + log(ppgdp) + group:log(ppgdp), data = un)
```

6.4.1: What is the meaning of the NH model?

This NH model here is stating that the mean function is fully described by this first set of regressors, *log(ppgdp)* and the interaction between *group* and *log(ppgdp)*, and that there is no influence of the main effects from *group*.

6.4.2 Perform the test and summarize the results

```
anova(m1, m2)
```

```
## Analysis of Variance Table
##
## Model 1: lifeExpF ~ log(ppgdp) + group:log(ppgdp)
## Model 2: lifeExpF ~ group + log(ppgdp) + group:log(ppgdp)
##   Res.Df    RSS Df Sum of Sq      F  Pr(>F)
## 1    195 5232.0
## 2    193 5077.7  2    154.31 2.9326 0.05564 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

From this test, we can see that we got a p-value of about 0.06 when comparing the reduced (*m1*) and full (*m2*) models outlined above. This means that, with an alpha of 0.05, we are unable to reject the null hypothesis that the mean function is fully described by the first reduced model, containing the regressors *log(ppgdp)* and the interaction *group:log(ppgdp)*. So, we can't reject the hypothesis that the estimated coefficient associated with the main effect for *group* is equal to zero, meaning that regressor does not have an effect on lifeExpF when *log(ppgdp)* and *group:log(ppgdp)* are already present in the model.

So, here we would (just barely with this p-value) conclude that the reduced null model is sufficient to explain the response, and that there is no need to include *group* in the model.

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?

```
summary(m2)
```

```
##
## Call:
## lm(formula = lifeExpF ~ group + log(ppgdp) + group:log(ppgdp),
##     data = un)
```

```
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -18.634  -2.089   0.301   2.255  14.489
##
## Coefficients:
##                         Estimate Std. Error t value Pr(>|t|)
## (Intercept)              59.2137    15.2203   3.890 0.000138 ***
## groupother              -11.1731    15.5948  -0.716 0.474572
## groupafrica             -22.9848    15.7838  -1.456 0.146954
## log(ppgdp)                2.2425     1.4664   1.529 0.127844
## groupother:log(ppgdp)     0.9294     1.5177   0.612 0.540986
## groupafrica:log(ppgdp)    1.0950     1.5785   0.694 0.488703
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.129 on 193 degrees of freedom
## Multiple R-squared:  0.7498, Adjusted R-squared:  0.7433
## F-statistic: 115.7 on 5 and 193 DF,  p-value: < 2.2e-16
```

```
# here beta_02 corresponds to groupother, beta_03 to groupafrica, beta_12 to
# groupother:log(ppgdp) and beta_13 to groupafrica:log(ppgdp)

# To compare level means like this, can use the corresponding code from
# class on 10.26.17:


L = matrix(c(0, 1, -1, 0, 0, 0, 0, 0, 0, 0, 1, 1), byrow = TRUE, nrow = 2)


c.vector = c(14, 0.2)
ht3 = linearHypothesis(m2, hypothesis.matrix = L, rhs = c.vector)
ht3
```

```
## Linear hypothesis test
##
## Hypothesis:
## groupother - groupafrica = 14
## groupother:log(ppgdp)  + groupafrica:log(ppgdp) = 0.2
##
## Model 1: restricted model
## Model 2: lifeExpF ~ group + log(ppgdp) + group:log(ppgdp)
##
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1    195 5092.6
## 2    193 5077.7  2    14.906 0.2833 0.7536
```

So, with this large p-value (F-value?) from the test performed above, I would interpret this as meaning that there is not enough support to reject the null hypothesis that both $\beta_{02} - \beta_{03} = 14$ *and* $\beta_{12} + \beta_{13} = 0.2$, and so it appears that both equalities hold.