Part One: The dataset contains body-fat measurements for 20 healthy females aged 35–44, together with three additional measurements: (i) triceps skinfold thickness, (ii) thigh circumference and (iii) midarm circumference. Obtaining accurate measurements of body-fat percentage is expensive (and often time consuming). The goal is to find cheaper and easier measurements to are able to accurately predict the value.

```
bfdata = read.table('bodyfat.txt',header=TRUE)
head(bfdata)
```

```
##   Triceps Thigh Midarm BodyFat
## 1    19.5  43.1   29.1    11.9
## 2    24.7  49.8   28.2    22.8
## 3    30.7  51.9   37.0    18.7
## 4    29.8  54.3   31.1    20.1
## 5    19.1  42.2   30.9    12.9
## 6    25.6  53.9   23.7    21.7
```

1. Fit a regression model in which the body-fat measurement bfdata[,4] is the outcome/response variable and tricep skinfold bfdata[,1] is the predictor variable, using the code m1 = lm(BodyFat ~ Triceps, data = bfdata). Obtain a summary of the analysis using sm1 =summary(m1). Use the code sm1$coef to obtain the coefficients for the model. What does the table tell us about the association between tricep skinfold and body-fat?

```
m1 = lm(BodyFat ~ Triceps, data = bfdata)
sm1 = summary(m1)
sm1$coef
```

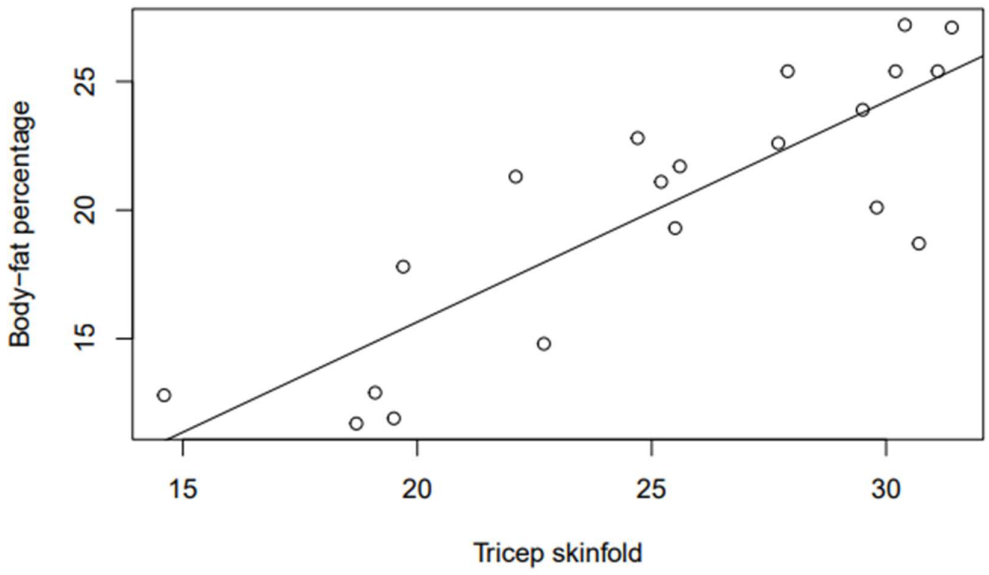
```
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   -1.496      3.319  -0.451 6.58e-01
## Triceps        0.857      0.129   6.656 3.02e-06
```

- The estimated increase in the body-fat percentage on average for a one unit increase in Triceps is 0.8571865.
- Since the p-value of Triceps is less than 0.05, there is evidence that the true effect of Triceps (i.e. in the population) is not zero.

2. Plot the data using plot(bfdata$Triceps, bfdata$BodyFat, xlab="Tricep skinfold", ylab="Body-fat percentage"). You can add the fitted regression line using abline(m1). Comment on the general trend observed in the plot.

- In general, the trend is when the Tricep skinfold increases, the BodyFat increases.There is a positive correlation between the two variables.

plot(bfdata$Triceps, bfdata$BodyFat, xlab="Tricep skinfold", ylab="Body-fat percentage") abline(m1)



3. The table obtained in question 1 includes a column of standard errors. Why are these useful?

- Standard error is an estimate of the standard deviation of sample means.
- Standard error is useful as it can be used to compute confidence interval of the coefficients. Confidence interval can be used to check if the coefficient is further from zero.

4. Find a 95% confidence interval for the effect of tricep skinfold on body-fat (see tutorial 2 or last week's lecture notes). Interpret this interval. Does it suggest that tricep skinfold is associated with body-fat in the population of 35–44-year-old healthy females?

```
ci1 = confint(m1,2)
ci1
```

```
##         2.5 % 97.5 %
## Triceps 0.587   1.13
```

- We are 95% confident that the true effect is in the interval (0.5866281, 1.127745). We again have evidence that the true effect is not zero, as 0 lies outside this interval. Therefore tricep skinfold is associated with body-fat (positive relationship) in the population of 35–44 year old healthy females,

5. Refit the regression model by adding in the predictor variable mid-arm circumference (Midarm). Assign this analysis to the variable m2, and its summary to the variable sm2. Show the table of coefficients for the model. Is there evidence of tricep skinfold and midarm circumference being important predictors of body-fat?

```
m2 = lm(BodyFat ~ Triceps+Midarm, data = bfdata)
sm2 = summary(m2)
sm2$coef
```

```
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)    6.792      4.488    1.51 1.49e-01
## Triceps        1.001      0.128    7.80 5.12e-07
## Midarm        -0.431      0.177   -2.44 2.58e-02
```

- Triceps and Midarm are all important predictors of body-fat, because the p value for both Triceps and Midarm are less than 0.05,the p value less than 0.05 means that their coefficients are not likely to be zero.

6. Interpret the effect of tricep skinfold on body-fat. Interpret the effect of midarm circumference on body-fat.

- We can estimate that when each unit of Triceps increases, the body-fat percentage will increase by 1.00585 units on average.
- We can estimate that when each unit of Midarm increases, the body-fat percentage will decrease by 0.431442 units on average.

7. Use sm2$r.sq to find the R2 value, and interpret the result.

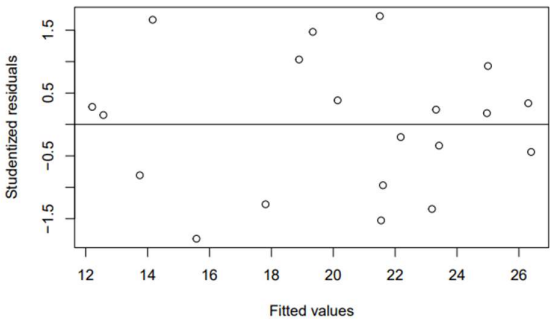
```
sm2$r.sq
```

```
## [1] 0.786
```

- This means that the regression model explained 78.61598% of the variation in the body-fat variable.

8. Use the code rm2 = rstudent(m2) to obtain the studentized residuals. Plot these against the fitted values using the code plot(m2$fitted,rm2,xlab="YYY",ylab="ZZZ"), where you replace “YYY” and “ZZZ” by appropriate labels. Then use abline(h=0) to draw a horizontal line at 0. Is there any evidence that the model assumptions have been violated?

```
rm2 = rstudent(m2)
plot(m2$fitted,rm2,xlab="Fitted values",ylab="Studentized residuals")
abline(h=0)
```



- There is no evidence that the assumptions being violated.
- In particular, there are no obvious trends, no clear evidence of non-constant variance and no obvious outliers.
- Interpretation of these plots can sometimes be subjective, but here there is no obvious problem.

Part Two Data is available on all individuals who played in both the 1991 and 1992 Major League Baseball seasons, and had at least 50 "hits" in the 1991 season. There are three variables in the dataset:

- `salary`: The logarithm of the player's salary in the 1992 season
- `value`: The number of runs scored in the 1991 season as a direct result of the player's "at-bats". The higher this number the more valuable the player.
- `eligibility`: A categorical variable with two levels. An `eligible` player was able to either (i) be a free-agent at the end of the 1991 season (and so be signed by any team) or (ii) enter into negotiations with his current team about his salary in the 1992 season. A `not_eligible` player was on a long-term contract, and had no legal means to change his team or salary for the 1992 season.

The data are available on the resource page and can be read into R as follows:

```
bbdata = read.table('baseball.txt',header=TRUE)
head(bbdata)
```

```
##   salary value eligibility
## 1   8.10   104    eligible
## 2   7.86    66    eligible
## 3   7.82    73    eligible
## 4   7.81    50    eligible
## 5   7.75    58    eligible
## 6   7.68   100    eligible
```

To start with we will ignore the continuous predictor variable value, as we want to focus on using analysis of variance (ANOVA).

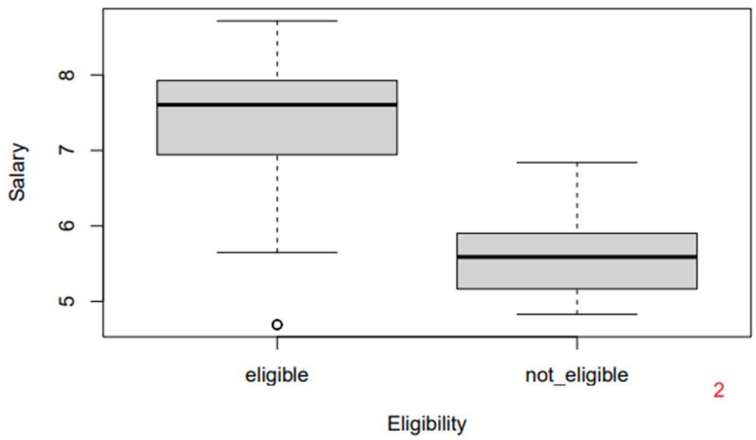
1. Describe in words the null and alternative hypotheses we are testing if we use an ANOVA F-test on these data, with salary as the outcome variable and eligibility as the categorical predictor variable

- H0: beta1(difference between the mean of eligible group and the mean of not eligible group) equal to 0
- HA: beta1(difference between the mean of eligible group and the mean of not eligible group) not equal to 0

2. Draw a boxplot with each level of the variable eligibility on the x-axis using the code boxplot(salary ~ eligibility, data = bbdata). You should include appropriate axis labels. Interpret the plot.

- There is a significant difference between the mean of the eligible group and the mean of the not eligible group. The salary range of the eligible group is much higher than the salary range of the not eligible group

boxplot(salary ~ eligibility, data = bbdata,xlab = "Eligibility", ylab = "Salary")



3. Use the lm function to fit the linear regression version of an ANOVA model, where salary is the outcome variable and eligibility is the categorical predictor variable. How many dummy variables does R need to include in the regression model for this predictor? Use the anova function to obtain the ANOVA table, and use it to carry out and interpret the hypothesis test defined in part 1.

- For a categorical predictor with k levels, We need k - 1 dummy variables
- In this case, there is 2 levels (eligible and not eligible), therefore there is 1 (1 = 2-1) dummy variable R needs to include in the regression model.

```
bbdata$eligibility=factor(bbdata$eligibility)
m2 = lm(salary ~ eligibility, data = bbdata)
anova(m2)
```

```
## Analysis of Variance Table
##
## Response: salary
##              Df Sum Sq Mean Sq F value Pr(>F)
## eligibility   1    175   175.2     391 <2e-16 ***
## Residuals   254    114     0.4
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

- P-value is less than 2.2e-16, therefore less than 0.05 (5%). So we can reject the null hypothesis H0, that is to reject the difference between mean of eligible group and not eligible group is equal to 0. This means the difference between the mean of eligible group and and the mean of not eligible group is not equal to 0 and significantly different. The boxplot above also verifies this conclusion.

Part 3. The datafile contains body-fat measurements for 20 healthy females aged 35–44, as well as their triceps skinfold thickness, thigh circumference, and midarm circumference.

The first part of the question is about how the F-test can be used for general regression models, as opposed to the special case of ANOVA. This then leads into model building in the second part. There is no correct final model in the second part. Your mark will be based on how you justify the decisions you make, more than the decisions themselves.

You may still have a copy of the datafile (`bodyfat.txt`). If not, download it again from the Resources section of the webpage and read it into R, as follows:

```
bfdata = read.table('bodyfat.txt',header=TRUE)
head(bfdata)
```

```
##   Triceps Thigh Midarm BodyFat
## 1    19.5  43.1   29.1    11.9
## 2    24.7  49.8   28.2    22.8
## 3    30.7  51.9   37.0    18.7
## 4    29.8  54.3   31.1    20.1
## 5    19.1  42.2   30.9    12.9
## 6    25.6  53.9   23.7    21.7
```

1. Use the scale function to standardise each of the predictor variables (tricep skinfold, thigh circumference, and midarm.

bfdata$zTriceps = scale(bfdata$Triceps)

bfdata$zThigh = scale(bfdata$Thigh)

bfdata$zMidarm = scale(bfdata$Midarm)

bfdata

```
##    Triceps Thigh Midarm BodyFat zTriceps  zThigh  zMidarm
## 1     19.5  43.1   29.1    11.9  -1.1556 -1.5417  0.40580
## 2     24.7  49.8   28.2    22.8  -0.1204 -0.2617  0.15903
## 3     30.7  51.9   37.0    18.7   1.0740  0.1395  2.57187
## 4     29.8  54.3   31.1    20.1   0.8948  0.5979  0.95417
## 5     19.1  42.2   30.9    12.9  -1.2353 -1.7136  0.89933
## 6     25.6  53.9   23.7    21.7   0.0587  0.5215 -1.07481
## 7     31.4  58.5   27.6    27.1   1.2134  1.4003 -0.00548
## 8     27.9  52.1   30.6    25.4   0.5166  0.1777  0.81708
## 9     22.1  49.9   23.2    21.3  -0.6380 -0.2426 -1.21191
## 10    25.5  53.5   24.8    19.3   0.0388  0.4451 -0.77321
## 11    31.1  56.6   30.0    25.4   1.1536  1.0373  0.65256
## 12    30.4  56.7   28.3    27.2   1.0143  1.0564  0.18645
## 13    18.7  46.5   23.0    11.7  -1.3149 -0.8921 -1.26674
## 14    19.7  44.2   28.6    17.8  -1.1158 -1.3315  0.26870
## 15    14.6  42.7   21.3    12.8  -2.1311 -1.6181 -1.73286
## 16    29.5  54.4   30.1    23.9   0.8351  0.6170  0.67998
## 17    27.7  55.3   25.7    22.6   0.4768  0.7890 -0.52644
## 18    30.2  58.6   24.6    25.4   0.9745  1.4194 -0.82804
## 19    22.7  48.2   27.1    14.8  -0.5186 -0.5674 -0.14258
## 20    25.2  51.0   27.5    21.1  -0.0209 -0.0325 -0.03290
```

2. Fit the following two regression models using lm , where the outcome variable is BodyFat . The first model should contain just the intercept (~ 1). Assign this fitted model to the object m0 . The second model should contain all three standardised predictor variables. Assign this fitted model to mf .

m0 = lm(BodyFat ~ 1, data = bfdata)

mf = lm(BodyFat ~ zTriceps + zThigh + zMidarm, data = bfdata)

3. Carry out an F-test to compare these two models. What are the null and alternative hypotheses being considered in this test? What can you conclude from the p-value?

anova(m0,mf)

```
## Analysis of Variance Table
##
## Model 1: BodyFat ~ 1
## Model 2: BodyFat ~ zTriceps + zThigh + zMidarm
##   Res.Df RSS Df Sum of Sq    F  Pr(>F)
## 1     19 495
## 2     16  98  3       397 21.5 7.3e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

- Null hypotheses: β1 = β2 = β3 = 0, this means the coefficients for zTriceps, zThigh, zMidarm in model mf are 0.The null hypotheses implies that the model with only intercept fits the data the same as the full model.
- Alternative hypotheses: at least one of β1, β2, β3 not equal to 0, this means at least one of the coefficients for zTriceps, zThigh, zMidarm in model mf is not 0. The alternative hypothesis implies that the full model fits the data better than the intercept-only model.
- The F-test is to check if two models, in this case, model mo and model mf, are they significantly different.
- The p-value for the test is 7.3e-06 < 0.05. This means there is evidence to reject the null hypothesis that the coefficients for zTriceps, zThigh, zMidarm are 0, so at least one of the coefficients for zTriceps, zThigh, zMidarm in model mf is not 0. The full model fit the data better than the interceptonly model.

4. Now fit three separate regression models using lm , where the outcome variable is the body-fat measurement and the predictor variable is one of the three standardised variables.

mtr = lm(BodyFat ~ zTriceps, data = bfdata)

mth = lm(BodyFat ~ zThigh, data = bfdata)

mmi = lm(BodyFat ~ zMidarm, data = bfdata)

5. Find a summary of model mf using the summary function. How does the estimate for the effect of each predictor variable compare to that found using the individual regression models in 4? Can you think of a possible reason for any differences?

summary(mf)

```
##
## Call:
## lm(formula = BodyFat ~ zTriceps + zThigh + zMidarm, data = bfdata)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -3.726 -1.611  0.392  1.466  4.128
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   20.195      0.555   36.42   <2e-16 ***
## zTriceps      21.771     15.148    1.44     0.17
## zThigh       -14.954     13.516   -1.11     0.28
## zMidarm       -7.973      5.819   -1.37     0.19
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.48 on 16 degrees of freedom
## Multiple R-squared:  0.801,  Adjusted R-squared:  0.764
## F-statistic: 21.5 on 3 and 16 DF,  p-value: 7.34e-06
```

summary(mtr)

```
##
## Call:
## lm(formula = BodyFat ~ zTriceps, data = bfdata)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -6.120 -2.190  0.674  1.938  3.852
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   20.195      0.631   32.03   <2e-16 ***
## zTriceps       4.306      0.647    6.66    3e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.82 on 18 degrees of freedom
## Multiple R-squared:  0.711,  Adjusted R-squared:  0.695
## F-statistic: 44.3 on 1 and 18 DF,  p-value: 3.02e-06
```

summary(mth)

```
##
## Call:
## lm(formula = BodyFat ~ zThigh, data = bfdata)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -4.495 -1.567  0.124  1.336  4.408
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   20.195      0.561   35.98  < 2e-16 ***
## zThigh         4.484      0.576    7.79  3.6e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.51 on 18 degrees of freedom
## Multiple R-squared:  0.771,  Adjusted R-squared:  0.758
## F-statistic: 60.6 on 1 and 18 DF,  p-value: 3.6e-07
```

summary(mmi)

```
##
## Call:
## lm(formula = BodyFat ~ zMidarm, data = bfdata)
##
## Residuals:
##    Min     1Q Median     3Q    Max
##  -8.59  -3.85   1.46   3.56   6.91
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   20.195      1.161   17.39  1.1e-12 ***
## zMidarm        0.727      1.191    0.61     0.55
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.19 on 18 degrees of freedom
## Multiple R-squared:  0.0203, Adjusted R-squared:  -0.0341
## F-statistic: 0.373 on 1 and 18 DF,  p-value: 0.549
```

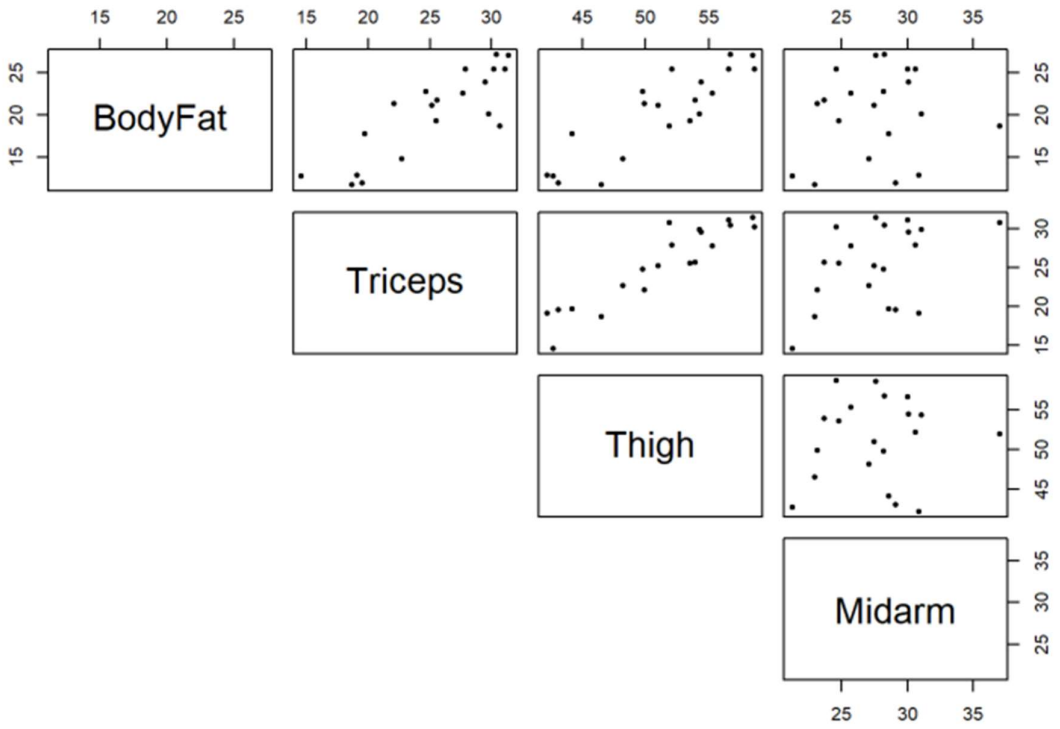
- The estimated effect of each predictor variable in model mf is very different compare to that found using the individual regression models.
- As the estimates of effects of each predictor variable change substantially depending on what terms included, this means potentially there is multicollinearity, that means there are predictor variables are correlated.

6. Compare the individual t-tests from the summary in 5 with the F-test you carried out in 3. Does this support your explanation in 5?

- The F-test result in question 3 is significant, it was the full model fit the data better than the intercept-only model. So at least one of the coefficients is not 0.
- However, in the t-test in question 5, that is the summary(mf) result, shows that all the coefficients for predictor variables are not significant as their p values are all greater than 0.05, that means the coefficients for all predictor variables are equal 0.
- These 2 results are not consistent in terms of the coefficient of the predictor variables. Significant F-test, with not significant t-test, means multicollinearity, this supports the explanation in question 5.

7. The function pairs is useful for looking at the pairwise relationships of three or more variables. Use the following command to have a look at the relationships between the three predictor variables: pairs(bfdata[,c(4,1:3)],lower.panel=NULL,pch=20) (do not worry about the details in this command for now). Correlation coefficients can be used to summarise the strength of "collinearity", and can be found using the cor function. Use the command cor(bfdata[,1:3]) to find the correlation coefficients for each pair of predictor variables. Explain how the plots and correlation coefficients indicate a strong correlation between one pair of predictor variables.

pairs(bfdata[,c(4,1:3)],lower.panel=NULL,pch=20)



cor(bfdata[,1:3])

Triceps Thigh Midarm

Triceps 1.000 0.9238 0.4578

Thigh 0.924 1.0000 0.0847

Midarm 0.458 0.0847 1.0000

- From the pairwise plot, we can see a very strong linear correlation between Triceps and Thigh.
- From the correlation coefficients, we can confirm that there is a very strong correlation (0.924) between Triceps and Thigh.

8. Based on the observed correlation between one pair of predictor variables, come up with a way to reduce the number of predictor variables. For the rest of this question, you should use just two predictor variables.

bfdata$zTrithi = (bfdata$zTriceps + bfdata$zThigh)/2

bfdata

##	Triceps	Thigh	Midarm	BodyFat	zTriceps	zThigh	zMidarm	zTrithi
## 1	19.5	43.1	29.1	11.9	-1.1556	-1.5417	0.40580	-1.3486
## 2	24.7	49.8	28.2	22.8	-0.1204	-0.2617	0.15903	-0.1911
## 3	30.7	51.9	37.0	18.7	1.0740	0.1395	2.57187	0.6067
## 4	29.8	54.3	31.1	20.1	0.8948	0.5979	0.95417	0.7464
## 5	19.1	42.2	30.9	12.9	-1.2353	-1.7136	0.89933	-1.4744
## 6	25.6	53.9	23.7	21.7	0.0587	0.5215	-1.07481	0.2901
## 7	31.4	58.5	27.6	27.1	1.2134	1.4003	-0.00548	1.3068
## 8	27.9	52.1	30.6	25.4	0.5166	0.1777	0.81708	0.3471
## 9	22.1	49.9	23.2	21.3	-0.6380	-0.2426	-1.21191	-0.4403
## 10	25.5	53.5	24.8	19.3	0.0388	0.4451	-0.77321	0.2420
## 11	31.1	56.6	30.0	25.4	1.1536	1.0373	0.65256	1.0955
## 12	30.4	56.7	28.3	27.2	1.0143	1.0564	0.18645	1.0354
## 13	18.7	46.5	23.0	11.7	-1.3149	-0.8921	-1.26674	-1.1035
## 14	19.7	44.2	28.6	17.8	-1.1158	-1.3315	0.26870	-1.2237
## 15	14.6	42.7	21.3	12.8	-2.1311	-1.6181	-1.73286	-1.8746
## 16	29.5	54.4	30.1	23.9	0.8351	0.6170	0.67998	0.7261
## 17	27.7	55.3	25.7	22.6	0.4768	0.7890	-0.52644	0.6329
## 18	30.2	58.6	24.6	25.4	0.9745	1.4194	-0.82804	1.1969
## 19	22.7	48.2	27.1	14.8	-0.5186	-0.5674	-0.14258	-0.5430
## 20	25.2	51.0	27.5	21.1	-0.0209	-0.0325	-0.03290	-0.0267

- Strong correlation between Triceps and Thigh, which may lead to problems. we can combine the variables and define a new variable that is the average of the zTriceps and zThigh. The new variable is a predictor variable.
- The other predictor variable is zMidarm

9. For each of your predictor variables, state whether you expect it to be positively or negatively associated with body-fat.

- I expect the new variable zTrithi (average of the zTriceps and zThigh) to be positively associated with body fat.
- I expect zMidarm to be positively associated with body fat from common sense, but from the pairwise plot above, Midarm does not seem to be very associate with body fat.

10. Suppose there is a potential interaction between these two predictor variables. How would you interpret this?

- If there were a interaction between two predictor variables, the effect of one predictor (say x2) on the response variable depends on another predictor (say x1)
- In this case, if there were a interaction between zTrithi and zMidarm, the effect of zMidarm on bodyfat would depend on zTrithi of the body.

11. For the rest of the question we will assume that there is no interaction between the two predictor variables. Using the Gelman and Hill approach (Lecture 5), state your choice of full model.

- My full model contains the new zTrithi variable as well as zMidarm, with no interaction between them,i.e. mnew : $y = \beta 0 + \beta 1x1 + \beta 2x2 + \epsilon$, where y = bodyfat, x1 = zTrithi, x2 = zMidarm, and $\epsilon \sim N(0, \sigma 2)$.

12. Fit this full model, and decide which variables to keep using the Gelman and Hill approach. What is your final model? Interpret the results from this model.

```
mnew = lm(BodyFat ~ zTrithi + zMidarm, data = bfdata)

summary(mnew)$coef

## Estimate Std. Error t value Pr(>|t|)

## (Intercept) 20.195 0.565 35.748 1.94e-17

## zTrithi 4.725 0.615 7.684 6.29e-07

## zMidarm -0.554 0.603 -0.919 3.71e-01

ci = confint(mnew)

ci

## 2.5 % 97.5 %

## (Intercept) 19.00 21.387

## zTrithi 3.43 6.022

## zMidarm -1.83 0.718
```

- As expected, the effect of zTrithi (average of zTriceps and zThigh) is positive and highly significant to body fat, so we keep that in the model.
- The effect of zMidarm is close to zero and not significant, we were unsure of what sign this should be by looking at the 95% confident interval, I also can see that 0 is included in the 95% confident interval of zMidarm. So we could keep it in the model or remove it; the choice we make here will probably have little effect on the analysis. In this case, I probably choose to not include zMidarm in the final model, but I will do some other tests (see below) to confirm.

13. For the AIC approach, state up to 3 candidate models you will consider.

- First model: new zTrithi variable as well as zMidarm, with no interaction between them.
- Second model: only new zTrithi variable.

- Third model: only intercept, with no predictor variables.

14. Fit these three candidate models and find their AIC values, using the AIC function. For example, AIC(m) returns the AIC value of fitted model m . Which is the best model? Compare this with the final model in 12.

```
mt = lm(BodyFat ~ zTrithi, data = bfdata)

AIC(mnew,mt,m0)

## df AIC

## mnew 4 98.6

## mt 3 97.5

## m0 2 124.9
```

- Model t (with zTrithi, the average of zTriceps and zThigh alone) is the best model as it has the lowest AIC, Although the full model (mnew) isn't far behind. Model 0, which does not include any predictor variable has a much higher AIC value than the other two models.
- AIC value confirms my decision in question 12 to remove zMidarm and only include zTrithi (the average of zTriceps and zThigh) alone in the final model.
- In additional, we also can confirm the conclusion about the statistical significance of the coefficient of zMidarm, by carrying out an F-test of the null hypothesis that the coefficient of zMidarm is 0.

```
anova(mnew, mt)

## Analysis of Variance Table

##

## Model 1: BodyFat ~ zTrithi + zMidarm

## Model 2: BodyFat ~ zTrithi

## Res.Df RSS Df Sum of Sq F Pr(>F)

## 1 17 108

## 2 18 114 -1 -5.39 0.84 0.37
```

- The p-value for the test is 0.37 > 0.05. This means there is no evidence to reject the null hypothesis that the coefficient of zMidarm is 0, so the coefficient of zMidarm is 0. This matches the result from the AIC that the best model is the model only with zTrithi.

Part 4: We will look again at the baseball data analysed in Assignment 1. The three variables are:

- salary : The logarithm of the player's salary in the 1992 season
- value : The number of runs scored in the 1991 season as a direct result of the player's "at-bats". The higher this number the more valuable the player.
- eligibility : A categorical variable with two levels. An eligible player was able to either (i) be a free-agent at the end of the 1991 season (and so be signed by any team) or (ii) enter into negotiations with his current team about his salary in the 1992 season. A not eligible player

was on a long-term contract, and had no legal means to change his team or salary for the 1992 season.

You may still have a copy of the datafile (baseball.txt). If not, download it again from the Resources section of the webpage and read it into R, as follows:

```
bbdata = read.table('baseball.txt',header=TRUE)

bbdata$eligibility = factor(bbdata$eligibility)

head(bbdata)

## salary value eligibility

## 1 8.10 104 eligible

## 2 7.86 66 eligible

## 3 7.82 73 eligible

## 4 7.81 50 eligible

## 5 7.75 58 eligible

## 6 7.68 100 eligible
```

1. Standardise the bbdata$value variable using the scale function. Fit a regression model that includes eligibility , the standardised value , and the interaction between these two variables. Present the summary of the model fit using the summary function. Write down the fitted model separately for eligible and non-eligible players.

```
bbdata$zvalue = scale(bbdata$value)

mbb = lm(salary ~ eligibility * zvalue, data = bbdata)

summary(mbb)


##

## Call:

## lm(formula = salary ~ eligibility * zvalue, data = bbdata)

##

## Residuals:

## Min 1Q Median 3Q Max

## -2.730 -0.370 0.027 0.342 1.297

##

## Coefficients:

## Estimate Std. Error t value Pr(>|t|)

## (Intercept) 7.3609 0.0424 173.51 <2e-16 ***
```

eligibilitynot_eligible -1.6792 0.0816 -20.57 <2e-16 ***

zvalue 0.4165 0.0416 10.01 <2e-16 ***

eligibilitynot_eligible:zvalue -0.2239 0.0847 -2.64 0.0087 **

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

##

Residual standard error: 0.563 on 252 degrees of freedom

Multiple R-squared: 0.724, Adjusted R-squared: 0.72

F-statistic: 220 on 3 and 252 DF, p-value: <2e-16

- E[y] = β0 + β1x1 + β2x2 + β12x1x2
- x1 = 1, if not eligible
- x1 = 0, otherwise (means eligible)
- y non-eligible = β0 + β1 + β2x2 + β12x2 = 7.36090 - 1.67917 + 0.41650x2 - 0.22393x2 = 5.68173 + 0.17257x2
- y eligible = β0 + β2x2 = 7.36090 + 0.41650x2

2. Find 95% confidence intervals for each of the two main effects and the interaction.

Interpret (i) the main effect of eligibility and (ii) the interaction. Is there evidence that the relationship between salary and standardised value is different for eligible and non-eligible players?

confint(mbb)

2.5 % 97.5 %

(Intercept) 7.277 7.4444

eligibilitynot_eligible -1.840 -1.5184

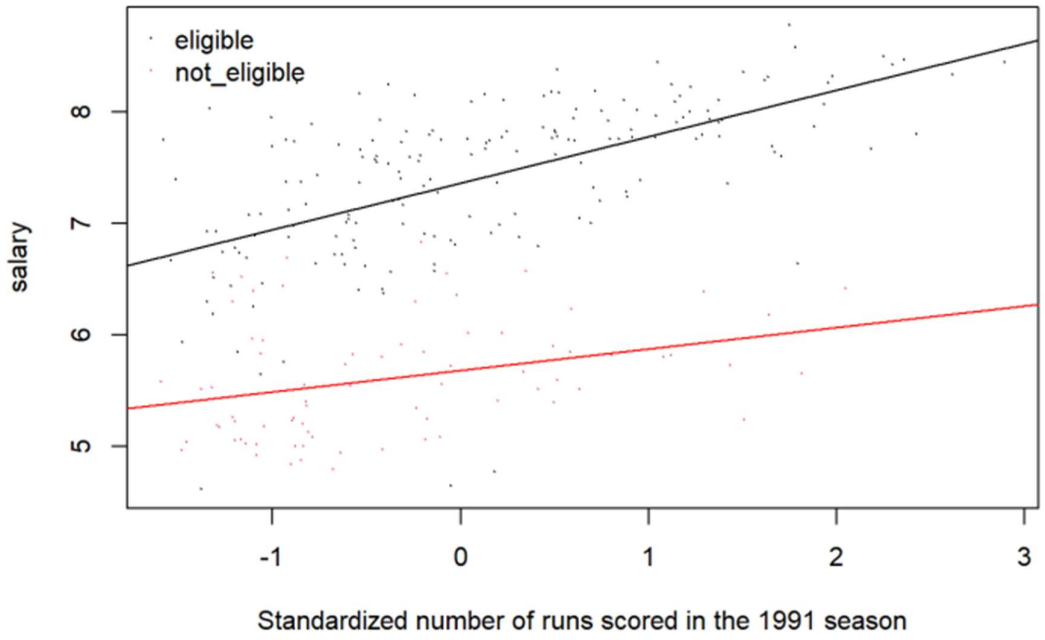
zvalue 0.335 0.4984

eligibilitynot_eligible:zvalue -0.391 -0.0571

- Main effect due to eligibility β1: We are 95% confident the expected difference in salary from not eligible and those from eligible, both taken at the mean value (number of runs scored in the 1991 season), will be between -1.8399175 and -1.51842442.
- Interaction effect β12: We are 95% confident β12 lies between -0.3907604 and -0.05709488. This means if we move 1 standard deviation further from the value (number of runs scored in the 1991 season), then we are 95% confident the expected difference in salary from not eligible and those from eligible will be between -0.3907604 and -0.05709488. Since the 95% confidence interval does not include 0, we conclude there is evidence for an interaction effect.
- As there is evidence for an interaction effect, β12 does not equal 0, therefore the relationship between salary and standardised value is different for eligible and non-eligible players because the slop of the regression lines will be different.

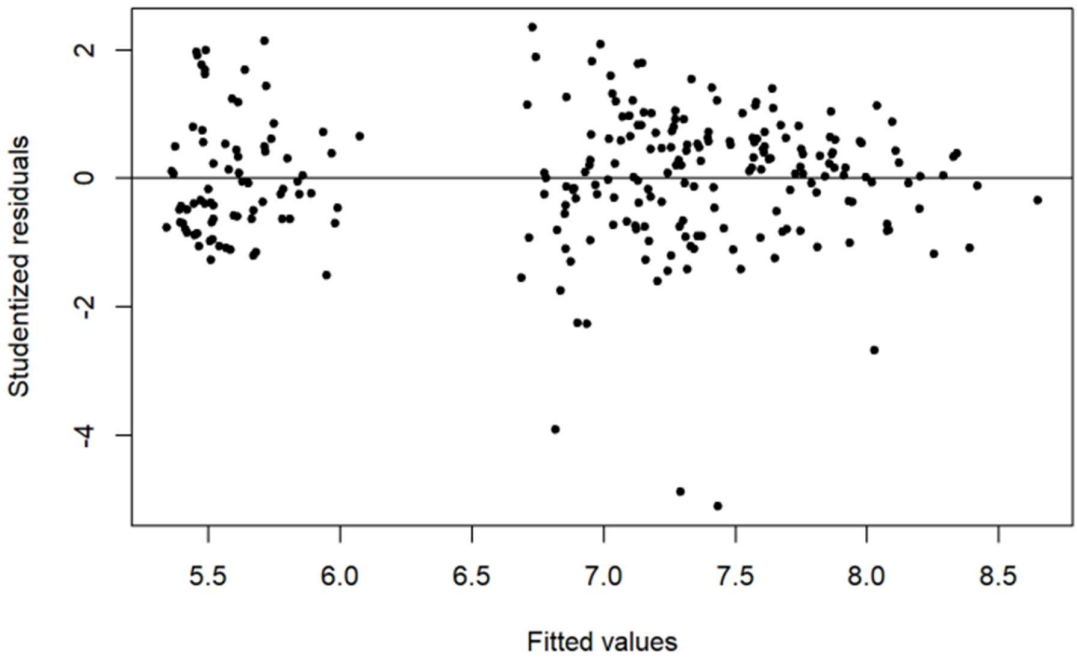
3. Plot salary (y-axis) against standardised value (x-axis). Colour the points according to eligibility . Use abline to include the fitted regression line for each level of eligibility . Include appropriate axis labels and a legend.

```
int.eligible = mbb$coef[1]

int.noteligible = mbb$coef[1]+mbb$coef[2]

slope.eligible = mbb$coef[3]

slope.noteligible = mbb$coef[3]+mbb$coef[4]

plot(jitter(bbdata$zvalue,amount=0), jitter(bbdata$salary,amount=0), pch=20,

 xlab="Standardized number of runs scored in the 1991 season", ylab="salary",

 cex=0.25, col=bbdata$eligibility)

legend("topleft", legend=levels(bbdata$eligibility), col=1:2, pch=20, pt.cex=0.25,

bty="n")

abline(int.eligible, slope.eligible, col="black")

abline(int.noteligible, slope.noteligible, col="red")
```



4. Give two qualitative conclusions from your fitted model.

```
plot(jitter(mbb$fitted, amount=0), rstudent(mbb), xlab="Fitted values",

ylab="Studentized residuals", pch=20)

abline(h=0)
```



- There is no major pattern in the residuals.
- There does not appear to be strong evidence of non-constant variance.
- There are two apparent outliers, with studentized residuals close to -5, and one outlier with studentized residuals close to -4. We can check how many extreme residuals there are, and how many we would expect, as follows. Suppose we consider those that are less then –3 or greater than +3. The following code works out how many such studentised residuals there are:

```
nr3 = sum(abs(rstudent(mbb))>3)

nr3

## [1] 3

pr3<-2*pnorm(3,lower.tail=F)

pr3

## [1] 0.0027

er3<-nrow(bbdata)*pr3

er3

## [1] 0.691
```

- So we observe 3 studentized residuals that are less then –3 or greater than +3, while we expect only 0.6911478 (less than 1!). This implies that we should investigate why these observations are so extreme and/or consider how we might improve the model