STATS 201/208 Assignment 3

Chunkai Meng / 857058772

Due Date: 3pm Thursday 29th September 2016

## Loading required package: s20x

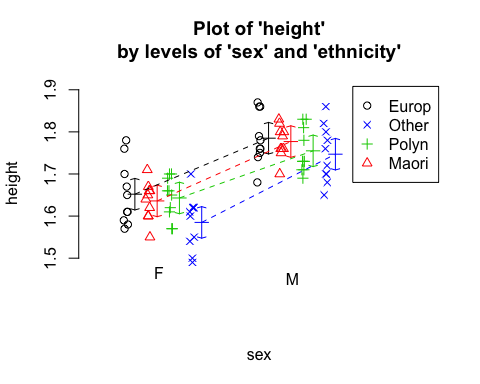
# Question 1

### Question of interest/goal of the study:

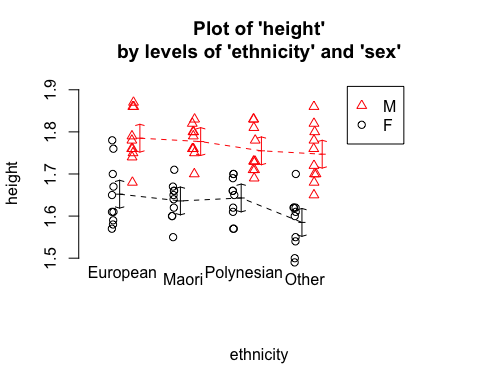
We investigated how people's height depends on their sex and their the ethnicity, and whether these factors influence each other.

### Read in and inspect the data:

height.df=read.table("height.txt",header=T)  
  
# create interaction plot(s)  
interactionPlots(height ~ sex + ethnicity, data=height.df)



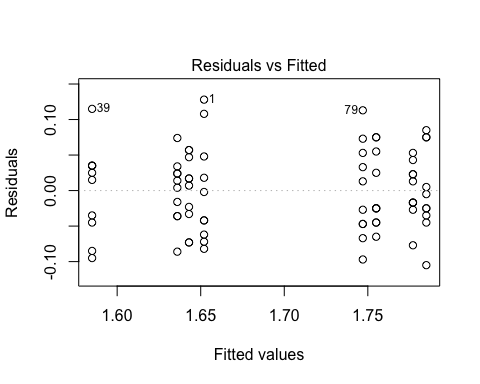
interactionPlots(height ~ ethnicity + sex, data=height.df)



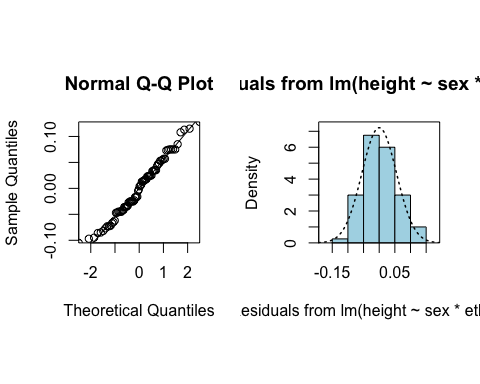
Here we see that the European males seem to be higher than all the other people on average. There are a slight difference among some of the above slops, but these is not so much interaction.  
However, for these type of data (i.e., 2 categorical explanatories) we start by fitting the model with interaction. We will see whether it can be simplifyed by Occam's razor rule.

### Fit model and check assumptions

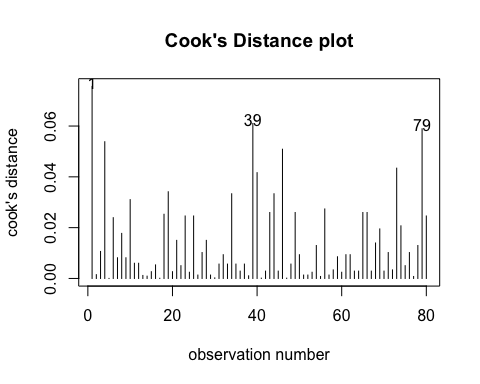
height.fit = lm(height ~ sex \* ethnicity, data = height.df)  
eovcheck(height.fit)



# showed approximately constant and not trend  
normcheck(height.fit)



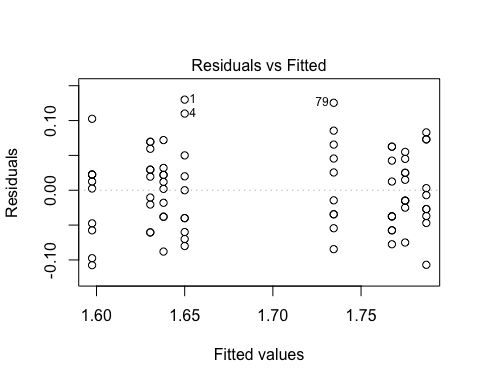
# residuals looks normal  
cooks20x(height.fit)



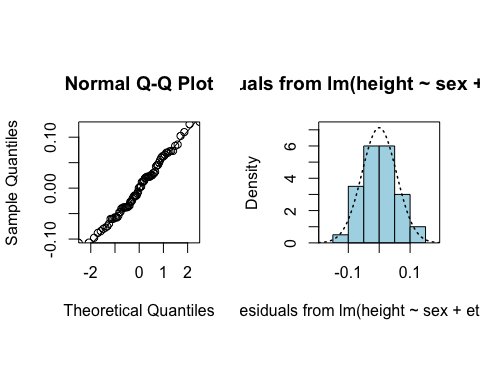
# no unduly influential points  
anova(height.fit)

## Analysis of Variance Table  
##   
## Response: height  
## Df Sum Sq Mean Sq F value Pr(>F)   
## sex 1 0.37538 0.37538 112.2492 2.427e-16 \*\*\*  
## ethnicity 3 0.03033 0.01011 3.0232 0.0351 \*   
## sex:ethnicity 3 0.00641 0.00214 0.6389 0.5924   
## Residuals 72 0.24078 0.00334   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

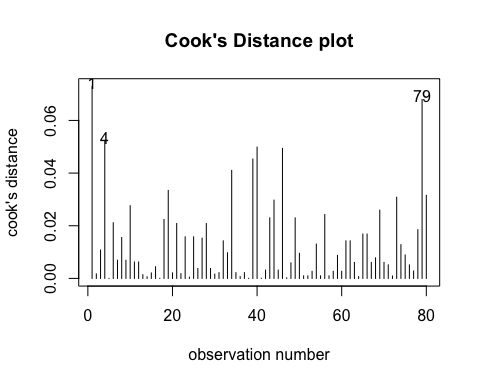
# There is no evidence of an interaction so we can simplify it to an additive model (i.e., no interaction term).  
  
height.fit2 = lm(height ~ sex + ethnicity, data = height.df)  
eovcheck(height.fit2)



# showed approximately constant and not trend  
normcheck(height.fit2)



# residuals looks normal  
cooks20x(height.fit2)



# no unduly influential points  
anova(height.fit2)

## Analysis of Variance Table  
##   
## Response: height  
## Df Sum Sq Mean Sq F value Pr(>F)   
## sex 1 0.37538 0.37538 113.8942 <2e-16 \*\*\*  
## ethnicity 3 0.03033 0.01011 3.0675 0.033 \*   
## Residuals 75 0.24719 0.00330   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

# It's ok now, let use summary to get coefficients R^2, and summary2way to get Tukey pairwise comparisons within and between the two factors.  
summary(height.fit2)

##   
## Call:  
## lm(formula = height ~ sex + ethnicity, data = height.df)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -0.10750 -0.03850 0.00100 0.03463 0.13000   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 1.65000 0.01435 114.963 < 2e-16 \*\*\*  
## sexM 0.13700 0.01284 10.672 < 2e-16 \*\*\*  
## ethnicityMaori -0.01200 0.01815 -0.661 0.51064   
## ethnicityOther -0.05250 0.01815 -2.892 0.00501 \*\*   
## ethnicityPolynesian -0.01950 0.01815 -1.074 0.28622   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.05741 on 75 degrees of freedom  
## Multiple R-squared: 0.6214, Adjusted R-squared: 0.6012   
## F-statistic: 30.77 on 4 and 75 DF, p-value: 3.695e-15

summary2way(height.fit2, page="nointeraction")

##   
##   
## sex comparisons:  
##   
## Estimate Tukey.L Tukey.U Tukey.p  
## F - M -0.137 -0.1626 -0.1114 0  
##   
##   
## ethnicity comparisons:  
##   
## Estimate Tukey.L Tukey.U Tukey.p  
## European - Maori 0.0120 -0.0357 0.0597 0.9113  
## European - Other 0.0525 0.0048 0.1002 0.0253  
## European - Polynesian 0.0195 -0.0282 0.0672 0.7063  
## Maori - Other 0.0405 -0.0072 0.0882 0.1243  
## Maori - Polynesian 0.0075 -0.0402 0.0552 0.9761  
## Other - Polynesian -0.0330 -0.0807 0.0147 0.2732

# extract the CIs where Tukey adjusted P-values are less than 0.05.  
## Estimate Tukey.L Tukey.U Tukey.p  
## F - M -0.137 -0.1626 -0.1114 0  
## European - Other 0.0525 0.0048 0.1002 0.0253

### Method and Assumption Checks

The interaction plots suggested a little interaction between sex and ethnicity. We fitted an ANOVA model with interaction but the p-value (0.5924) indicated that this interaction term was not significant. So we simplified it to additive model. Normality looks good and no influential points were detected. Model assumptions are satisfied.  
The final fitted model is:

Where is the dummy variable that takes the value 1 if is TRUE, otherwise 0, and

To assess differences between ethnicity, Tukey Confidence intervals were calculated.

### Executive Summary

Here we see the effect of sex on people's height doesn't depend on their ethnicity.  
We see that, on average, females are between 11 and 16 centimetres shorter than male without regarding their ethnicity.  
We see that, on average, European are between 0.5 centimetres to 10 centimeter taller than the people of other ethnicity (i.e., not European, Maori or Polynesian).

The fitted model explained 62% of the variability in data.

# Question 2

### Create contingency table

CD.tb=matrix(c(28, 632, 204, 9583), nrow = 2, ncol = 2, byrow = TRUE, dimnames = list(c("Cancer","Not.Cancer"), c("Diabetes","Not.Diabetes")))  
CD.tb

## Diabetes Not.Diabetes  
## Cancer 28 632  
## Not.Cancer 204 9583

### Calculate an approximate 95% confidence interval for the odds ratio manually using R.

# To calculate the OR for comparing the odds of getting cancer for people who do and do not have diabetes, we need to swap the row and column.  
DC.tb=matrix(c(28, 632, 204, 9583), nrow = 2, ncol = 2, byrow = FALSE, dimnames = list(c("Diabetes","Not.Diabetes"), c("Cancer","Not.Cancer")))  
round(digits=4)

## [1] 4

OR=28 \* 9583 / (204 \* 632)  
OR

## [1] 2.081193

# se of log(OR)  
se=sqrt(1/28 + 1/204 + 1/632 + 1/9583)  
# 95% CI of log OR  
logOR = log(OR) + c(-1,1) \* 1.96 \* se  
# back-transform  
exp(logOR)

## [1] 1.390711 3.114496

### Verify with oddsratio

# Use the oddsratio function (from the epitools package) to confirm your odds ratio calculations.  
  
# Further note: you only need the rev= parameter for the oddsratio command if you need to reverse either the rows or columns of the table to calculate the desired odds ratio.   
  
require(epitools)

## Loading required package: epitools

oddsratio(DC.tb, method="wald")

## $data  
## Cancer Not.Cancer Total  
## Diabetes 28 204 232  
## Not.Diabetes 632 9583 10215  
## Total 660 9787 10447  
##   
## $measure  
## NA  
## odds ratio with 95% C.I. estimate lower upper  
## Diabetes 1.000000 NA NA  
## Not.Diabetes 2.081193 1.390721 3.114473  
##   
## $p.value  
## NA  
## two-sided midp.exact fisher.exact chi.square  
## Diabetes NA NA NA  
## Not.Diabetes 0.001027494 0.0008910293 0.0002710024  
##   
## $correction  
## [1] FALSE  
##   
## attr(,"method")  
## [1] "Unconditional MLE & normal approximation (Wald) CI"

### Interpret your confidence interval for the odds ratio (i.e. write a sentence explaining what it means).

We estimate that the odds of getting cancer for people who have diabetes is between 1.391 and 3.115 times more than the odds for people who do not have diabetes.

### Use chisq.test to do the test for contingency tables. You may use the default form with Yates correction.

chisq.test(DC.tb)

##   
## Pearson's Chi-squared test with Yates' continuity correction  
##   
## data: DC.tb  
## X-squared = 12.286, df = 1, p-value = 0.0004565

### Interpret the test.

This test does a continuity correction to values, so as to correct for the fact that the cell counts are integer values.  
The p-value (0.0004565 < 0.05) indicating a strong evidence against the null hypothesis of no association between getting cancer and diabetes

### Is the test reliable in this case? Justify your answer.

DC.chisq = chisq.test(DC.tb, correct = FALSE)  
DC.chisq

##   
## Pearson's Chi-squared test  
##   
## data: DC.tb  
## X-squared = 13.261, df = 1, p-value = 0.000271

DC.chisq$expected

## Cancer Not.Cancer  
## Diabetes 14.65684 217.3432  
## Not.Diabetes 645.34316 9569.6568

DC.tb

## Cancer Not.Cancer  
## Diabetes 28 204  
## Not.Diabetes 632 9583

The test here is reliable because the sampling method of given data can be considered as simple random sampling. Second, the variable under study was categorical, and each level of the categorical variable had an expected frequency count of at least 5. Finally, The result from corrected and standard chi-squared tests makes no meaninful difference.