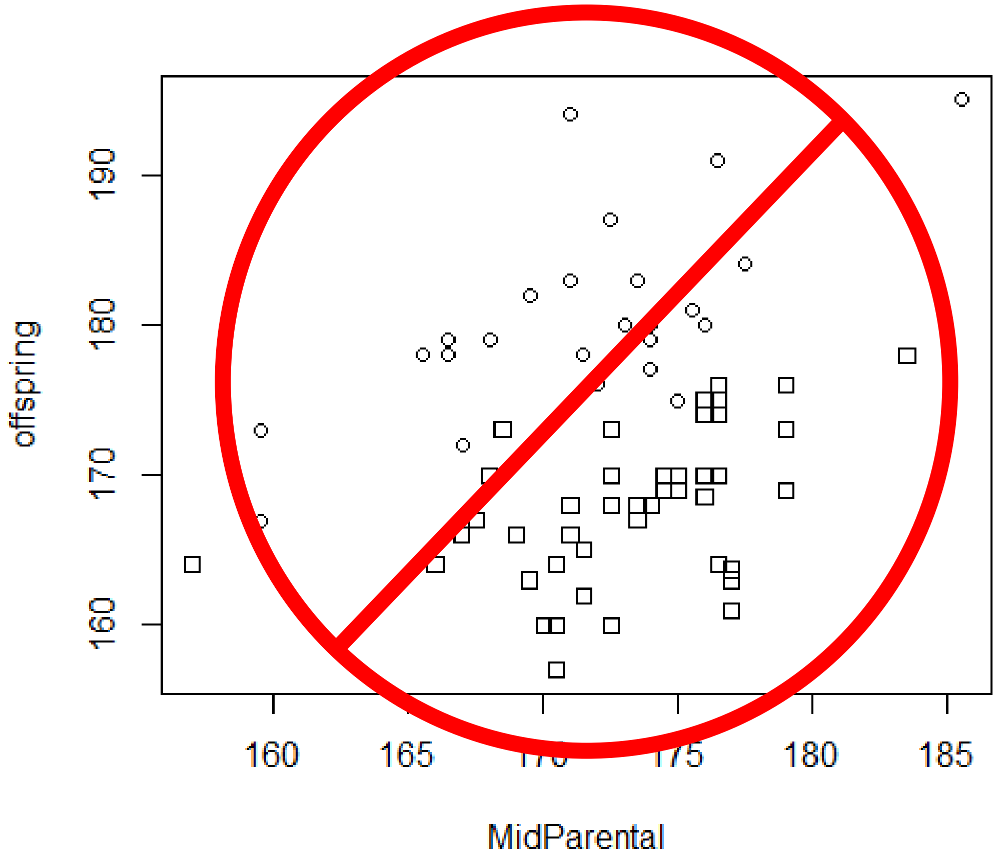
What to do when the assignment is given on ItsLearning

Give the assignment a brief glance – what will this be about? If your thoughts are along the lines of:

SAY WHAT NOW?! Ô_ô

You probably should attend the practicals where you can ask the assistants for help – also ask other questions you might be wondering about. Instead of commenting on the hand-in that you didn't understand what we were asking for.

A plot should be understood on its own, without having to read the text.



Missing:

- Units on axis
- Figure text: what does the figure show, what are the differences between the point symbols?

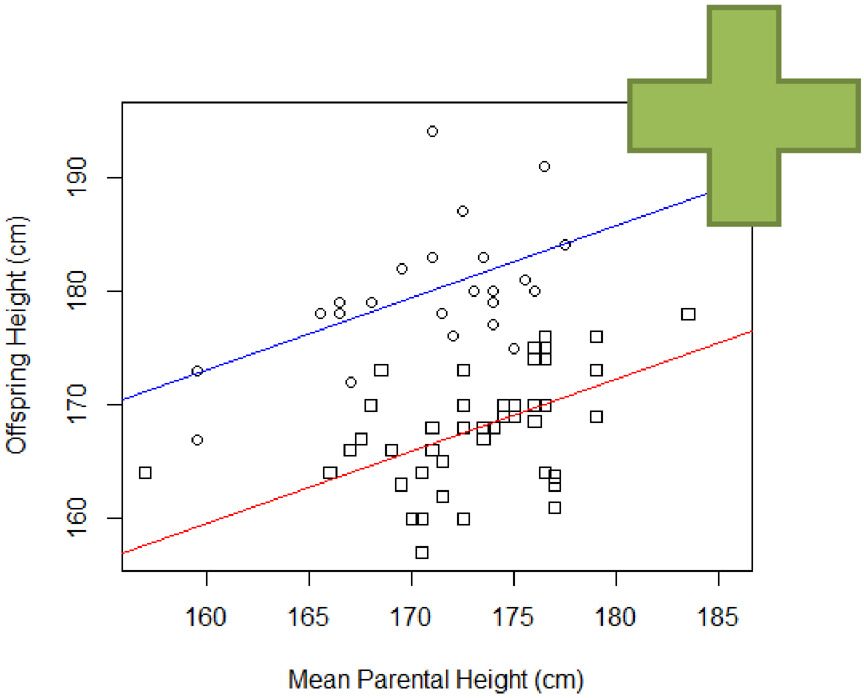


Figure 1: Relationship between offspring height and mean parental height (males:circles, females:squares). Solid lines are the regression lines from a linear model (males: blue, females: red).

Present what’s necessary and present it nicely

What NOT to do:

```
> summary(lm(offspring~sex, data=heights))

Call:
lm(formula = offspring ~ sex, data = heights)

Residuals:
     Min       1Q   Median       3Q      Max 
-13.4000  -3.5191  -0.1383   2.1234  14.6000 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 167.8766     0.7882 212.983  < 2e-16 ***
sex          12.5234     1.3376   9.362 5.75e-14 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 5.404 on 70 degrees of freedom
Multiple R-squared:  0.556,    Adjusted R-squared:  0.5496 
F-statistic: 87.65 on 1 and 70 DF,  p-value: 5.749e-14

> 
> summary(lm(offspring~MidParental, data=heights))

Call:
lm(formula = offspring ~ MidParental, data = heights)

Residuals:
    Min      1Q  Median      3Q     Max 
-14.349  -4.889  -1.809   6.203  22.443 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)   
(Intercept) 100.3828    33.2833   3.016  0.00357 **
MidParental   0.4162     0.1928   2.159  0.03425 * 
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 7.852 on 70 degrees of freedom
Multiple R-squared:  0.06245,  Adjusted R-squared:  0.04906 
F-statistic: 4.663 on 1 and 70 DF,  p-value: 0.03425

> summary(lm(offspring~sex+MidParental, data=heights))

Call:
lm(formula = offspring ~ sex + MidParental, data = heights)

Residuals:
    Min      1Q  Median      3Q     Max 
-9.3030 -2.5560  0.2545  2.5900 13.9421 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  58.1637    19.3822   3.001  0.00374 ** 
sex          13.5562     1.1280  12.018  < 2e-16 ***
MidParental   0.6336     0.1119   5.664 3.14e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.497 on 69 degrees of freedom
Multiple R-squared:  0.6969,   Adjusted R-squared:  0.6881 
F-statistic: 79.32 on 2 and 69 DF,  p-value: < 2.2e-16
```

What to DO

Table XX – The estimates (±SE) for linear regression models with different sets of explanatory variables. All estimates are significant different from zero (p-value<0.05)/ or include p-values in parantheses with the estimate

	Explanatory variables included		
Estimates	Sex + midparental	Sex	Midparental
Sex	13.5562±1.1280	12.5234±1.3376	-
Midparental	0.6336±0.1119	-	0.4162±0.1928

The answer should include an explanation/reasoning.

e.g.

How does the estimate change when including/excluding other variables?

It decreases/decreases…

BECAUSE…

- The SE increases due to more uncertainty
- Collinearity between explanatory variables
- There are demonic intrusions in our data set, so we better call on an excorcist (NOT a part of the curriculum in this course – or any other course in Norway for that matter)

What is all the fuss about significance and p-value?

Well, we know that the differences in average height of males and females are 12.5 cm, but given that this is only few of all the males and female in the world can we be really sure that this is the "true" difference - or did we simply choose an unrepresentative part of the population?

Along with estimate (differences between the means, estimate of a slope etc.) we are given an estimate of standard error (SE). Standard error is a an estimate telling us about the precision of the our estimate and will decrease with the our sample size/degree of freedom. SE = SD/sqrt(n), n = sample size. So the larger our sample size is, the more certain is our estimate.

The estimate and its SE are used to calculate a t-statistic, where beta0 is a value we want to compare our estimate with (usually defaults to zero in most models):

$$T = \frac{\hat{\beta}-\beta_0}{SE\ (\hat{\beta})}$$

Given our T-statistic and the degree of freedom (n-k, k = number of parameters we are estimating), we can from known tables/functions find the p-value. *The p-value is the probability of obtaining our*

T-statistic (or a more extreme) given our null-hypothesis (that the difference in means is zero, or the slope is zero)

The usual threshold/significance level is p-value=0.05,. This equals to an acceptable rate of Type 1 errors (http://en.wikipedia.org/wiki/Type_I_and_type_II_errors). So even though we say that the estimates are different a p-value = 0.05 or lower, they can still be similar by pure chance in 1 out of 20 trials.

The sharp-minded of you may now have seen that the SE is simply a function of sample size, causing p-value again to a function of sample size. So given a large enough sample, then even the tiniest of differences in means will be statistically significant. Which begs the question “would it make sense to distinguish between the heights of males and females if was only 0.000000001 cm, even though the p-value<0.05”?

How to report test results?

Keep it short and informative ☺

“Boys were significantly taller than girls (Estimated difference = 12.5, t-value=ZZ, df=YY, p-value = XXX).”