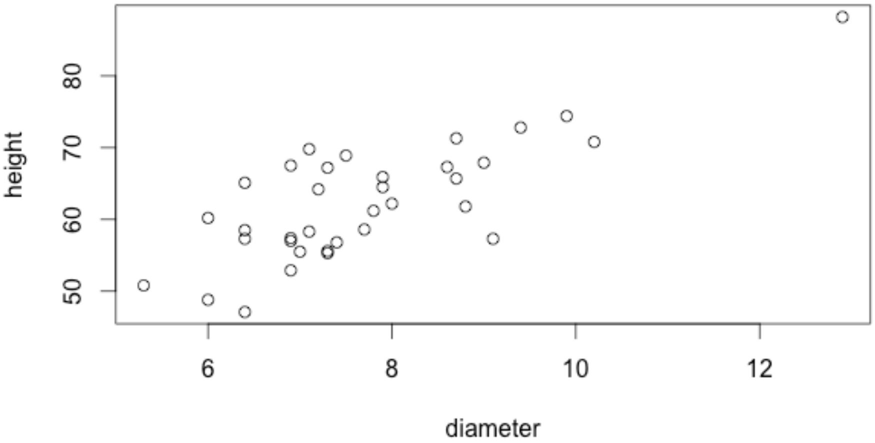
Jiayuan Shi
MA684 Midterm Project

1A. Find the scatter plot showing the association between height and diameter for this sample of trees. Given this scatter plot, do you think these data meet the assumptions of linear regression?
Solution:



I think these data meets the assumptions of linear regression. The assumptions of linear regression $Y_i = \beta_0 + \beta_1 X_i + E_i$ are,
i) Independent, random sample from underlying population:
True, the sample was gathered independently.
ii) Linearity, the means of Y|X fall on a straight line:
True, the scatterplot shows a straight line association.
iii) Homoscedasticity, the variance of Y|X is the same for all X(variance of E_i is the same for all X):
True, the scatterplot shows equal variance about the regression line, and variance of E_i is constant in the sample.
iv) Normality, the distribution of Y|X follows a normal distribution for all X:
True, when we plot a "Normal Q-Q" plot, we can see a straight line, which means the distribution of Y|X follows a normal distribution for all X.
v) Existence, the model holds for valid values of X:
True, we can see all the values of diameter.
Therefore, these data meets all the assumptions of linear regression.

1B. Find the regression formula to predict the height of a tree from its diameter, reporting the slope and intercept, standard errors, t-statistics, and p-values from the regression. Is there a significant association between the diameter and height? Explain. How well does the diameter predict the height of a tree?
Solution: I conduct a regression in R to predict the height of a tree from its diameter.

	Parameter Estimate	Standard Error	t-value (33 df)	p-value
Intercept	28.327	4.807	5.893	1.32e-06
Diameter	4.412	0.612	7.210	2.89e-08

Regression formula: height = 28.327 + 4.412*diameter

There is a significant association between the diameter and height.
The p-value for slope is testing the null hypothesis that there is no association between the diameter and height. Since the p-value is 2.89e-08, below 0.05, we can reject the null hypothesis. Therefore, there is a significant association between the diameter and height.

The multiple R-squared is 0.6117, which means diameter explains 61.17% of the variability in height of a tree. It is a good model to predict tree height.

1C. Give an interpretation of the slope from this regression line. Find the 95% confidence interval for this slope. Explain the relationship between this confidence interval and the p-value for the slope.
Solution: The slope is 4.412, which means for each inch increase of diameter, on average, height is expected to increase by 4.412 feet.

95% confidence interval for the slope is, 4.412±2.0345 (0.612) or (3.1669, 5.6571).

The p-value for slope is testing the null hypothesis that there is no association between the diameter and height. Since the p-value is 2.89e-08, below 0.05, we can reject the null hypothesis. Therefore, there is a significant association between the diameter and height. From the confidence interval (3.1669, 5.6571), we are 95% confident that the true population slope for diameter falls between 3.1669 and 5.6571. Since this confidence interval does not contain the null value of 0, it shows a significant association between the diameter and height.

1D (6 points). Based on this regression, what is the predicted height for a tree with a diameter of 10 inches? For a tree with a diameter of 12 inches? How close do you expect the true height of the tree to be to this estimate – that is, find and interpret an appropriate interval estimate for the height of a tree with a diameter of 10 inches, and for a tree with a diameter of 12 inches.
Solution: From the 'predict()' command in R, a tree with a diameter of 10 inches is expected to be 72.45 feet tall. We are 95% confident that 95% of trees with a diameter of 10 inches will be between 61.38 and 83.52 feet tall.

From the 'predict()' command in R, a tree with a diameter of 12 inches is expected to be 81.28 feet tall. We are 95% confident that 95% of trees with a diameter of 12 inches will be between 69.32 and 93.23 feet tall.

2A. As a description of the study sample, complete the following table:
Solution: Table 1A. Description of the study sample

Variable	Mean	Standard Deviation	Minimum	Maximum
Age	72.32	8.4317	55	85
IQ	99.66	15.7015	58	130
Hippo Change	-0.7433	0.9493	-3.1	1.8

Table 1B. Description of the study sample

Variable	N	%
Sex	300	1
Male	159	0.53
Female	141	0.47

2B. As a preliminary analysis, find and interpret the correlations between hippochange (the change in hippocampus volume) and age, and between hippochange and IQ. Based on these correlations, are age and/or IQ significantly associated with hippochange?
Solution: From the 'cor.test(hippochange,age)' command in R, there is a significant (p-value=0.0020<0.01) but very weak negative correlation between change in hippocampus volume and age (r=-0.1776).
From the 'cor.test(hippochange,iq)' command in R, there is a non-significant (p-value=0.199>0.05) and very weak negative correlation between change in hippocampus volume and IQ (r=0.0743).

2C. Complete the following tables summarizing the results of this multiple regression.
Solution: I conduct a regression in R to predict hippochange from age, sex, and IQ.
Table 2a. Multiple regression predicting percent change in hippocampus volume

Variable	Slope	SE of Slope	p-value
Intercept	0.310	0.601	0.6069
Age	-0.020	0.006	0.0027
Sex Female	-0.046	0.108	0.6708
IQ	0.004	0.003	0.2724

Table 2b. The ANOVA for the regression predicting change in hippocampus volume

Source	df	Sum of Squares	Mean Squares	F	p-value
Model	3	9.73	3.2433	3.698	0.0122
Residual	296	259.72	0.877		
Total	299	269.45	0.9012		

Report and interpret the R^2 for this regression model. Report the F-statistic and p-value from the ANOVA table for the regression. What can you conclude from this p-value?
Solution: R^2 is 9.73/ 269.45= 0.361, which means 36.1% of the variability in change in hippocampus volume can be explained by age, sex, and IQ.

F-statistic is 3.698 with 3 and 296 df, which gives the p-value is 0.0122.

The F-statistic is testing the null hypothesis that none of the variables in the regression are related to change in hippocampus volume. Because the p-value is less than 0.05, we can reject the null hypothesis, so at least one of age, sex, and IQ is significantly related to change in hippocampus volume.

2D. What can you say about the associations between age, sex, and IQ and change in hippocampus volume, based on this regression? Explain.
Solution: From the p-values in the Table 2a, age (p=0.0027<0.05) is significantly associated with change in hippocampus volume, after controlling for the other variables in the model, while sex (p=0.6708>0.05) and IQ (p=0.2724>0.05) are not significantly associated with change in hippocampus volume.

2E. Find and interpret the standardized slopes from this regression. Include in your response an explanation of what the standardized slopes from a multiple regression measure.
Solution: Using the QuantPsyc package and the lm.beta( ) command, we find that with the highest standardized slope of -0.1732, age is most strongly associated with change in hippocampus volume. Both IQ and sex, with standardized slopes of 0.0629 and -0.0243, are not significantly associated with quality of life, controlling for the other variables in the model.

2F. Provide a Table 3, similar to Table 2a from Question 2C, reporting slopes, standard errors, and p-values from this regression. Report and interpret the R^2 from this regression. Which of these independent variables are significantly related to hippochange in this analysis?
Solution: Including the 'relevel(factor(exercisegroup),'1'))' command in the regression to set the No Exercise group as a reference variable, so that we can predict hippochange from age, sex, IQ, Walking exercise and Yoga exercise.

Table 3. Multiple regression predicting percent change in hippocampus volume

Variable	Slope	SE of Slope	p-value
Intercept	-0.570	0.515	0.2699
Age	-0.017	0.006	0.0019
Sex Female	-0.091	0.092	0.3200
IQ	0.007	0.003	0.0224
relevel(factor(exercisegroup), "1")2	0.153	0.113	< 0.001
relevel(factor(exercisegroup), "1")3	0.199	0.113	0.0798

R^2 is 0.3146, which means 31.46% of the variability in change in hippocampus volume can be explained by the model.

From the p-values in the Table 3, age (p=0.0019<0.05), IQ(p=0.0224<0.05) and Walking exercise(p<0.001), are significantly related with change in hippocampus volume, after controlling for the other variables in the model, while sex (p=0.32>0.05), and Yoga exercise(p=0.0798>0.05) are not significantly related with change in hippocampus volume.

2G. Our primary interest in this analysis is in whether or not either Walking exercise or Yoga exercise has a positive benefit on change in hippocampus volume, compared to the No Exercise group. Interpret the results of this multiple regression analysis with a focus on this question - include as part of your answer an interpretation of slopes relating to exercise, and confidence intervals for these slopes.
Solution:

- Slope:

The slope for Walking exercise is 1.153, so we can say the mean change in hippocampus volume for Walking exercise group is 1.153% higher than that of No Exercise group, controlling for age, sex and IQ.

The slope for Yoga exercise is 0.199, so we can say the mean change in hippocampus volume for Yoga exercise group is 0.199% higher than that of No Exercise group, controlling for age, sex and IQ.

- Confidence Interval:

95% confidence interval for the slope of Walking exercise is 1.153 ±1.9679 (0.113) or (0.9306, 1.3754). We are 95% confident that the true population slope for Walking exercise falls between 0.9306 and 1.3754. Since this confidence interval does not contain the null value of 0, it shows a significant association between Walking exercise and change in hippocampus volume. We can also prove this through their p-values, which the p-value for slope of Walking exercise is less than 0.001.

95% confidence interval for the slope of Yoga exercise is 0.199 ±1.9679 (0.113) or (-0.0234, 0.4214). We are 95% confident that the true population slope for Yoga exercise falls between -0.0234 and 0.4214. Since this confidence interval contains the null value of 0, it does not show a significant association between Yoga exercise and change in hippocampus volume. We can also prove this through their p-values, which the p-value for slope of Yoga exercise is 0.0798, larger than 0.05.

Therefore, both Walking exercise and Yoga exercise has a positive benefit on change in hippocampus volume, compared to the No Exercise group. The association between Walking exercise and change in hippocampus volume is significant, but the association between Yoga exercise and change in hippocampus volume is not significant.

2H. As another way of describing the effect of exercise, controlling for age, sex, and IQ, give a (Type II) partial R^2 for the addition of exercise group, after controlling for age, sex, and IQ. Also give and interpret the partial F test for exercise group, after controlling for age, sex, and IQ (report the F statistic, degrees freedom, and p-value from this test).
Solution: Using the Anova( ) command, and do the Type II tests for exercise group, after controlling for age, sex, and IQ.
Below I cut-and-paste the Anova Table from R,

```
Response: hippochange
                          Sum Sq  Df F value    Pr(>F)
age                          6.191   1  9.8563  0.001865 **
```

```
sexf                          0.623   1  0.9926  0.319931
iq                          3.313   1  5.2735  0.022355 *
relevel(factor(exercisegroup), "1")  75.045   2 59.7343 < 2.2e-16 ***
Residuals                      184.678 294
---
```

$$SStotal = SSmodel + SSres = 6.191 + 0.623 + 3.313 + 75.045 + 184.678 = 269.85$$

$$Type\ II\ partial\ R^2 for\ exercise\ group = \frac{75.045}{269.85} = 0.278$$

So adding exercise to a model that already contains age, sex and IQ explains an additional 27.8% of the variability in change in hippocampus volume.

Type II partial F statistic is 59.734, df is (2, 294), and p-value is less than 0.001. The p-value for test is testing the null hypothesis the Full model(with exercise variable) explains no more variability than the Reduced model(without exercise variable). Since the p-value is less than 0.001, we can reject the null hypothesis. Therefore, exercise group is significantly related to change in hippocampus volume, after controlling for age, sex, and IQ.

2I. As a final analysis, one investigator wants to look at the interaction between age and exercise group. Run a multiple regression predicting hippochange from age, sex, IQ, exercise group, and the interaction between exercise group and age. Give a Table 4, similar to Tables 2a and 3, summarizing this analysis.

Solution: I conduct a regression in R to predict hippochange from age, sex, IQ, exercise group, and the interaction between exercise group and age(exercise group* age).

Table 4. Multiple regression predicting percent change in hippocampus volume

Variable	Slope	SE of Slope	p-value
Intercept	-1.913	0.730	0.2121
Age	-0.013	0.009	0.1651
Sex Female	-0.091	0.092	0.3264
IQ	0.007	0.002	0.0219
relevel(factor(exercisegroup), "1")2	1.280	0.938	0.1732
relevel(factor(exercisegroup), "1")3	1.289	1.012	0.2037
age:relevel(factor(exercisegroup), "1")2	-0.002	0.013	0.8890
age:relevel(factor(exercisegroup), "1")3	-0.015	0.014	0.2816

R^2 is 0.3177, so 31.77 % of the variability in change in hippocampus volume can be explained by age, sex, IQ, exercise group, and the interaction between exercise group and age.

From the p-values in the Table 4, IQ (p=0.0219<0.05) is significantly associated with change in hippocampus volume, after controlling for the other variables in the model, while age, sex, both Walking and Yoga exercise, and the interaction term, with p-values larger than 0,05, are not significantly associated with change in hippocampus volume.

What is being tested by the interaction between exercise group and age?

By the interaction between exercise group and age, the effect of the exercise group on change in hippocampus volume, differing by the level of the age is being tested.

What can you conclude about the effects of exercise from this analysis?

To sum up,

- From 2F, in the model of predict hippochangeing from age, sex, IQ, Walking exercise and Yoga exercise, after controlling for the other variables in the model, Walking exercise($p<0.001$) is significantly related with change in hippocampus volume, after controlling for the other variables in the model, while Yoga exercise($p=0.0798>0.05$) is not significantly related with change in hippocampus volume. We prove this in 2G. Although compared to the No Exercise group, both Walking and Yoga exercise have a positive benefit on change in hippocampus volume, the association between Walking exercise and change in hippocampus volume is significant, and the association between Yoga exercise and change in hippocampus volume is not significant.

- From 2H, we get p-value less than 0.05 from Type II tests, and the association between exercise group and change in hippocampus volume is significant in this model, after controlling for age, sex, and IQ.

- When we include an interaction between exercise group and age into the regression, because of the influence of age, this time Walking exercise is not significantly associated with change in hippocampus volume, after controlling for the other variables in the model. I think this is because the influence of the age variable. The interaction term is also not significantly associated with change in hippocampus volume. I think it means that interaction term does not affect the change in hippocampus volume, and we can exclude the interaction term.