Code ▾

36-309 Homework 3: Can It Be All So Simple (https://www.youtube.com/watch?v=7m148vZDwJA) Linear Regression?

Due Wednesday, September 28, 11:59pm on Gradescope

Sanaz Saadatifar

Question 1: Using Linear Regression to Watch Dogs Run (70 points)

Professor S. Dogg (https://en.wikipedia.org/wiki/Snoop_Dogg) is studying the relationship between dogs' weight and running speed. In his study, Professor Dogg weighs each dog (measured in pounds) and then times how long it takes them to run a 100-meter dash (where time is measured in seconds). For the sake of this question, assume that the dogs are trained to run the 100-meter dash as fast as they can on command.

Here is the dataset resulting from Professor Dogg's study:

Hide

```
dogData = read.csv("https://raw.githubusercontent.com/zjbranson/stat309fall2022/master/dogData.csv")
```

There are only two variables in this dataset, `weight` and `time` .

a. (13pts) Before running statistical models, let's first consider some initial exploration of this dataset. For this part, answer the following questions.

- (3pts) In this study, which variable is the outcome variable, and which is the explanatory variable?

[weight is the explanatory variable and time is the outcome variable]

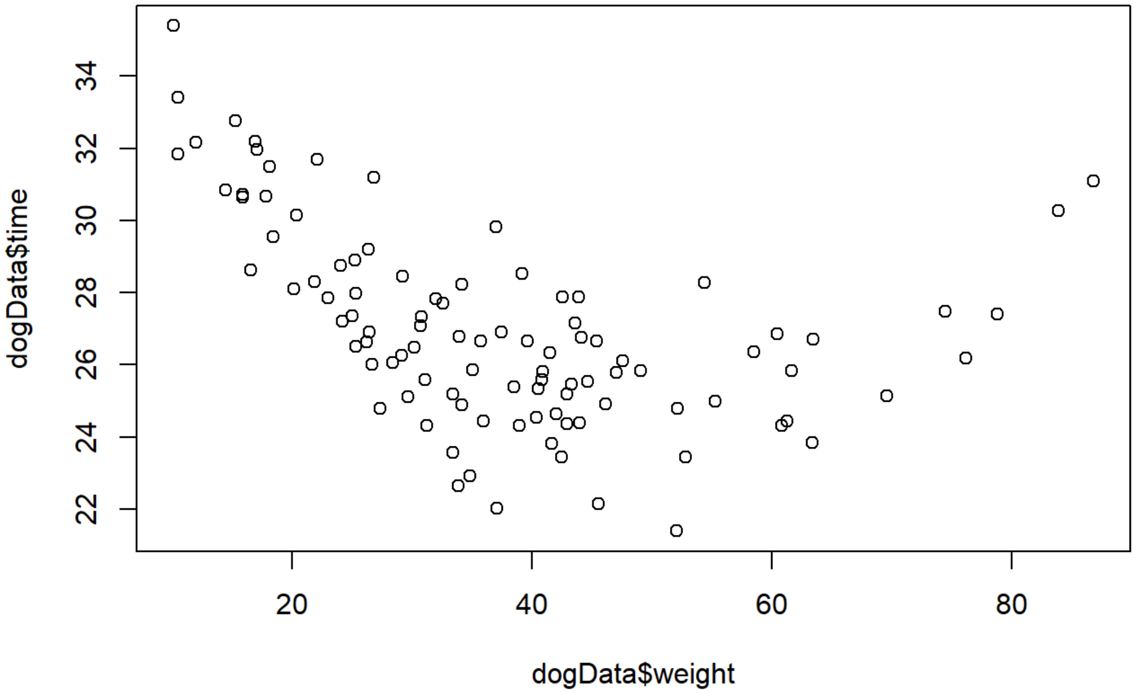
- (4pts) Given the setup of this problem, what is an appropriate form of bivariate graphical EDA for this dataset? Explain in 1-2 sentences why that EDA would be appropriate for this dataset in particular. For this part, just submit a written answer.

[scatterplot.because both variables (explanatory and outcome) are quantitative. so the most common graphical EDA for two quantitative variables is a scatterplot]

- (6pts) Now make the bivariate graphical EDA you mentioned above. After doing this, answer the following: Does there seem to be a relationship between a dog's weight and its running time? If so, do these variables seem *positively* correlated or *negatively* correlated? Explain your answer in 1-3 sentences.

Hide

```
plot(dogData$weight, dogData$time)
```



[there is not a linear relation ship as we see a smile-like pattern. in terms of the weights less than 50 there seems a negative correlation (as the dog weights increases the run time decreases) and for the dogs with weights above 50 there seems to be a positive relation (as dog weights increases, the run times increases)]

b. (20pts) In this part we’ll run the appropriate linear regression for this dataset. Answer the following questions.

- (5pts) What is the **statistical model** for the appropriate linear regression for this problem? You may use mathematical notation or words to state your answer. Be sure that your answer mentions the variable names for this dataset, rather than just “X” and “Y”.

[time iid~ N (β0 +β1weight ,σ2)] in this formula, σ2 is Residual Standard Error squared. the formula in other words is runtime = β0 +β1weight where β0 is intercept and β1 is slope

- (6pts) Now perform the appropriate linear regression analysis for this dataset using the `lm()` function. In your answer, be sure to include the code and `summary()` output for the linear regression. After doing this, answer the following: What value in the output is the estimate of the common error variance in the linear model (i.e., “sigma squared”)?

Hide

```
summary(lm(time~weight, data = dogData))
```

```
##
## Call:
## lm(formula = time ~ weight, data = dogData)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.0212 -1.4474 -0.2798  1.4120  7.7420
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 29.79000    0.60810  48.988  < 2e-16 ***
## weight      -0.07421    0.01486  -4.993 2.59e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.453 on 98 degrees of freedom
## Multiple R-squared:  0.2028, Adjusted R-squared:  0.1946
## F-statistic: 24.93 on 1 and 98 DF,  p-value: 2.591e-06
```

[Residual Standard Error is s , not s^2. hence our estimate of σ2 is Residual Standard Error^2 (2.453^2 =6.017209)] + (5pts) Write your interpretation of the **estimated slope** for this linear regression. Then, state your scientific conclusions about the slope within the context of this study. Be sure to mention a p-value in your conclusions.

[our estimate of the slope is -0.07421, meaning that average run time of a dog is estimated to decrease by 0.07421 for every one-pound increase in dog weight. the null hypothesis for the slop of this study is that H0:β1=0 (slope is 0), and the alternative hypothesis that HA: β1≠0 (none-zero slop). since the P-value is less than 0.05(2.59e-06), then we have enough evidence to reject the null hypothesis. so this test in this study shows there is s significant correlation between weight of a dog and its run time of a 100-meter dash]

- (4pts) Does the intercept (constant) estimate have a scientifically meaningful interpretation within the context of this study? Explain your answer in 1-2 sentences.

[We see that our estimate of the intercept is 29.79000, meaning that the average run time of a dog who has no weight is estimated to be 29.79000. However, NO, the intercept does not have a scientifically meaningful interpretation, because it is not possible for a dog to weight 0, and to be able to have run time of 29.79000 for a 100-meter dash. if a dog wights 0 then it means it does not exists. and data set also does not have such a data for dog weight. Also scientific conclusions shows that considering that the null hypothesis for the intercept of this study is that H0:β0=0 (intercept is 0), and the alternative hypothesis that HA: β0≠0 (none-zero intercept). since the P-value is less than 0.05(< 2e-16), then we have enough evidence to reject the null hypothesis.]

c. (12pts) Now we’ll consider some implications of the linear regression model you ran in Part B. For this part, answer the following questions.

- (8pts) Consider two dogs: (1) An Australian cattle dog (https://i.pinimg.com/originals/50/a8/71/50a871d3bef52e878fa9cb28605146c0.jpg) that’s 32 pounds, and (2) a Bernese mountain dog (https://media1.popsugar-assets.com/files/thumbor/3YgNuQGQmhenu9ajrk2uO2ikjHE/0x125:2003x2128/fit-in/2048xorig/filters:format_auto-!!-:strip_icc-!!-/2020/01/15/847/n/1922243/cdea0de05e1f661d6c6a54.93383595_/i/photos-of-bernese-mountain-dogs.jpg) that’s 100 pounds. Based on your linear regression in Part B, what is the estimated mean running time for each of these dogs? Be sure to show the equation you used to arrive at your answer. After you do that, also mention (for each dog) whether your estimate is an interpolation or an extrapolation of your linear regression model.

[The prediction equation is runtime = 29.79 -0.07421×weight. This means that the estimated mean runtime for An Australian cattle dog that’s 32 pounds is: 29.79 -0.07421×32= 27.41528 And the estimated mean runtime for a Bernese mountain dog that’s 100 pounds is: 29.79 -0.07421×100= 22.369 Since our data set ranges from 10 to 86 pounds, then for the Australian cattle dog that’s 32 pounds it can be considered an interpolation and for the Bernese mountain dog that’s 100 pounds it can be considered a extrapolation.]

- (4pts) Looking at your linear regression in Part B, Professor Dogg gets excited: He says, “I think your results suggest that heavier dogs tend to run faster. I want my dog Juelz Broadus (https://www.instagram.com/juelzbroadus/?hl=en) to run faster, so I’m going to have him gain weight.” Do you think Professor Dogg’s conclusion is a reasonable interpretation of your results in Part B? Explain in 1-3 sentences. (**Hint**: Don’t worry about any modeling assumption violations - for the sake of this particular part, assume that they are not violated.)

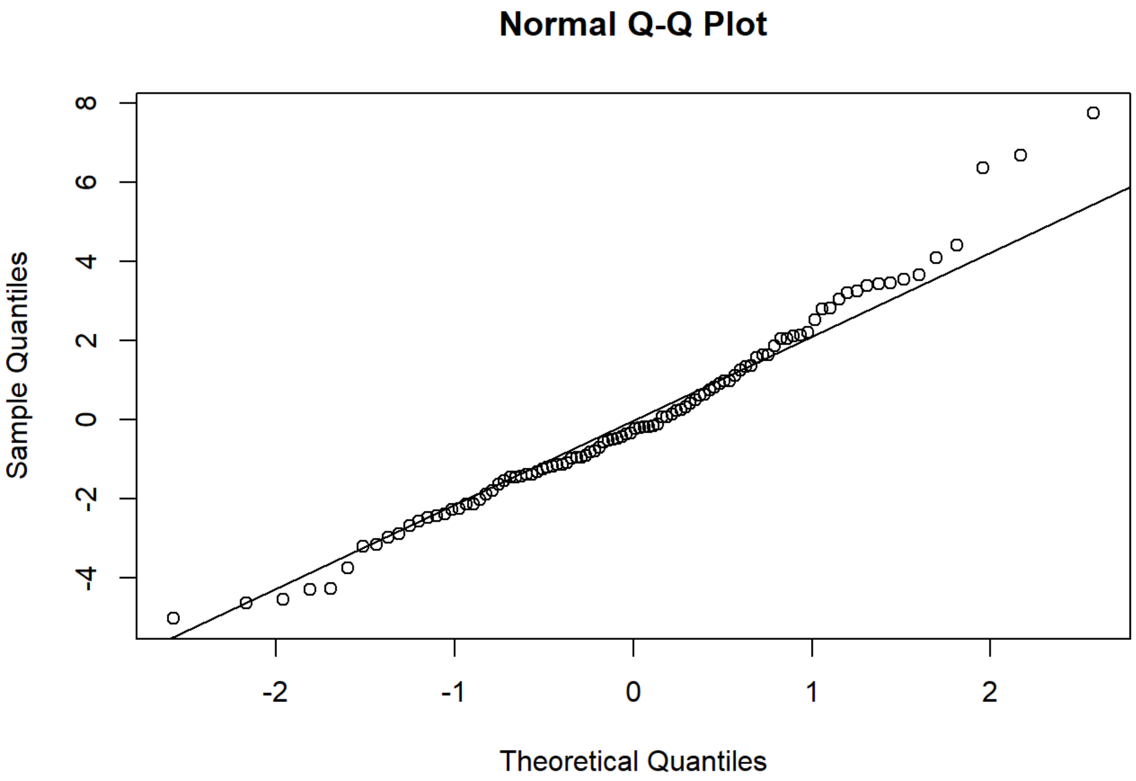
[I don’t think Professor Dogg’s conclusion is a reasonable interpretation because we cannot make a causal statement here, as only explanatory variable considered is the weight of the dog. However, there are other factors that might affect the runtime of a dog such as its age, race, etc., that none of the are considered in this study. Therefore, there is not high internal validity. Also, the study is observation not experiment where dogs are assigned specifically to a specific group.]

d. (10pts) Now we’re going to assess some of the modeling assumptions for the linear regression you ran in Part B. Answer the following questions:

- (5pts) Create a quantile-normal plot of residuals from the linear regression model you ran in Part B. (**Hint**: You’ll have to use the `residuals()` function.) Interpret this plot in relation to the appropriate linear model assumption(s). In particular, state whether you think each assumption(s) is “Very plausible,” “Plausible,” “Not plausible,” or “Very not plausible,” and explain your reasoning in 1-2 sentences.

Hide

```
res = residuals(lm(time~weight, data = dogData))
qqnorm(res)
qqline(res)
```



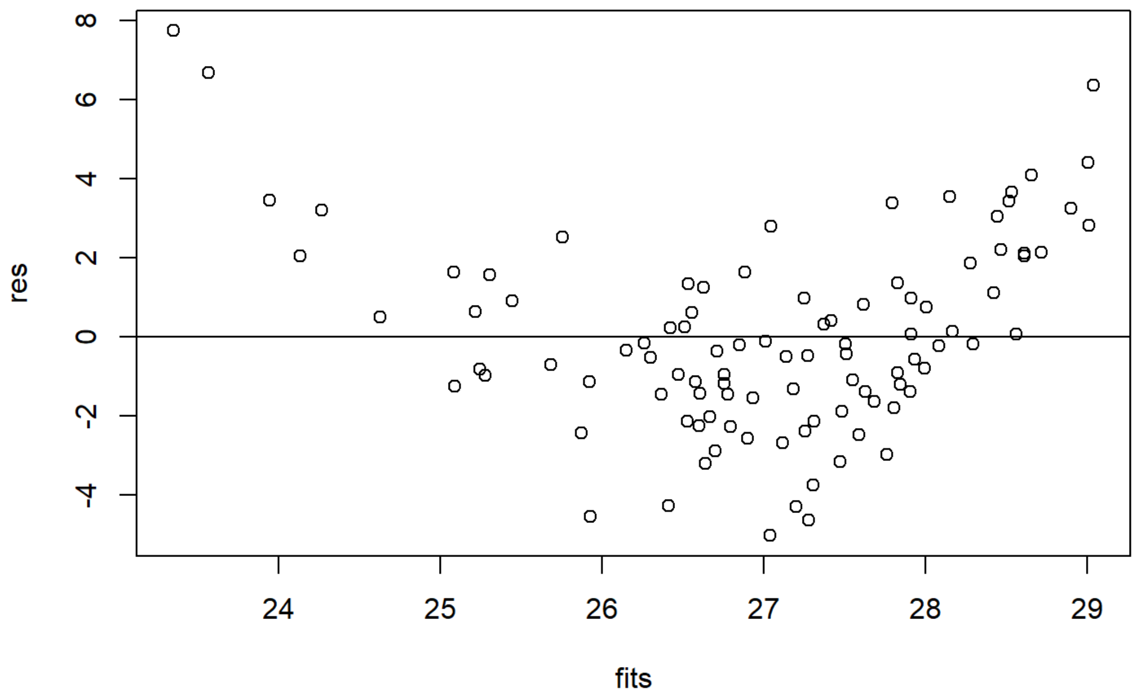
[This plot shows the normality. The points pretty much fall on the line except few points on the right, so the Normality assumption appears “plausible”.]

- (5pts) Create a residual vs. fit plot from the linear regression model you ran in Part B. (**Hint**: You’ll have to use the `residuals()` and `fitted()` functions.) Interpret this plot in relation to the appropriate linear model assumption(s).

In particular, state whether you think each assumption(s) is “Very plausible,” “Plausible,” “Not plausible,” or “Very not plausible,” and explain your reasoning in 1-2 sentences.

Hide

```
#making the residuals
res = residuals(lm(time~weight, data = dogData))
#making the fitted values:
fits = fitted(lm(time~weight, data = dogData))
#making the residual-vs-fit plot:
plot(fits, res)
abline(h=0)
```



[The residual vs. fit plot is used to check the equal variance and linearity. However there seems to be a strong violation of the equal variance assumption: The points appear to be not equally vertical spread from left-to-right, there are more points on the right. There also seems to be a strong violation of the linearity assumption: The dots appear to have a clear trend of U-shape (like a smile). Therefore, I say they are “very not plausible”.]

e. (15pts) Professor Dogg thinks for a bit and says, “I have a hypothesis. I think there is a significantly negative linear relationship between racing time and weight for small dogs (less than 40 pounds) but a significantly positive linear relationship for large dogs (greater than 40 pounds). To assess this hypothesis, he creates the following two datasets:

Hide

```
#subset of the data with just small dogs
smallDogs = subset(dogData, weight < 40)
#subset of the data with just large dogs
largeDogs = subset(dogData, weight > 40)
```

For **each** of these two datasets, do the following things:

- (3pts) Run the appropriate linear regression model (similar to Part B) and include the `summary()` output. (Remember, you need to run two models, one for each dataset.)

Hide

```
summary(lm(time~weight, data = smallDogs))
```

```
##
## Call:
## lm(formula = time ~ weight, data = smallDogs)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.2952 -1.0952 -0.1945  1.2270  4.7719
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 35.29226    0.79679  44.293  < 2e-16 ***
## weight      -0.27649    0.02855  -9.684  1.2e-13 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.788 on 57 degrees of freedom
## Multiple R-squared:  0.622,  Adjusted R-squared:  0.6153
## F-statistic: 93.78 on 1 and 57 DF,  p-value: 1.204e-13
```

Hide

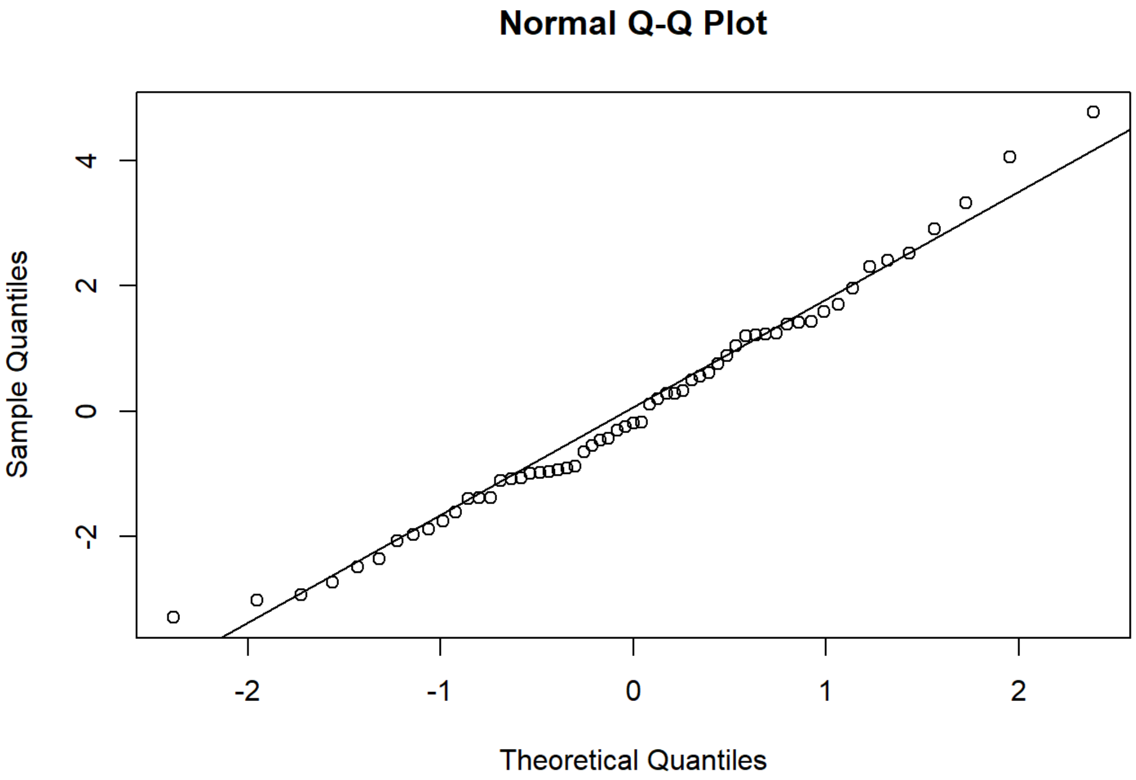
```
summary(lm(time~weight, data = largeDogs))
```

```
##
## Call:
## lm(formula = time ~ weight, data = largeDogs)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.2634 -0.8985  0.2508  0.7637  3.0624
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 22.10193    1.11414  19.838  < 2e-16 ***
## weight       0.06831    0.02047   3.337  0.00187 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.675 on 39 degrees of freedom
## Multiple R-squared:  0.2221, Adjusted R-squared:  0.2021
## F-statistic: 11.13 on 1 and 39 DF,  p-value: 0.001872
```

- (7pts) Assess the Normality, Equal Variance, and Linearity assumptions for each linear regression model you ran (similar to Part D). Include any relevant graphs you used to do this. State whether you think each assumption(s) is “Very plausible,” “Plausible,” “Not plausible,” or “Very not plausible,” and explain your reasoning in 1-2 sentences. (Again, remember to do this for each dataset.)

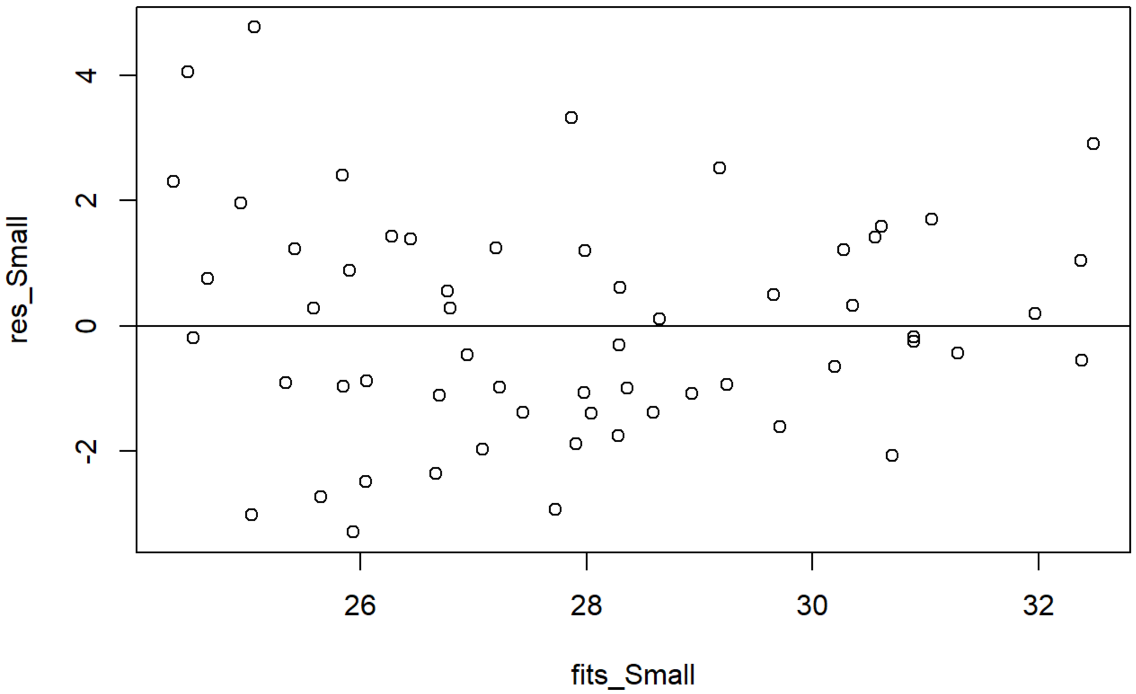
Hide

```
#making the residuals
res_Small = residuals(lm(time~weight, data = smallDogs))
qqnorm(res_Small)
qqline(res_Small)
```



Hide

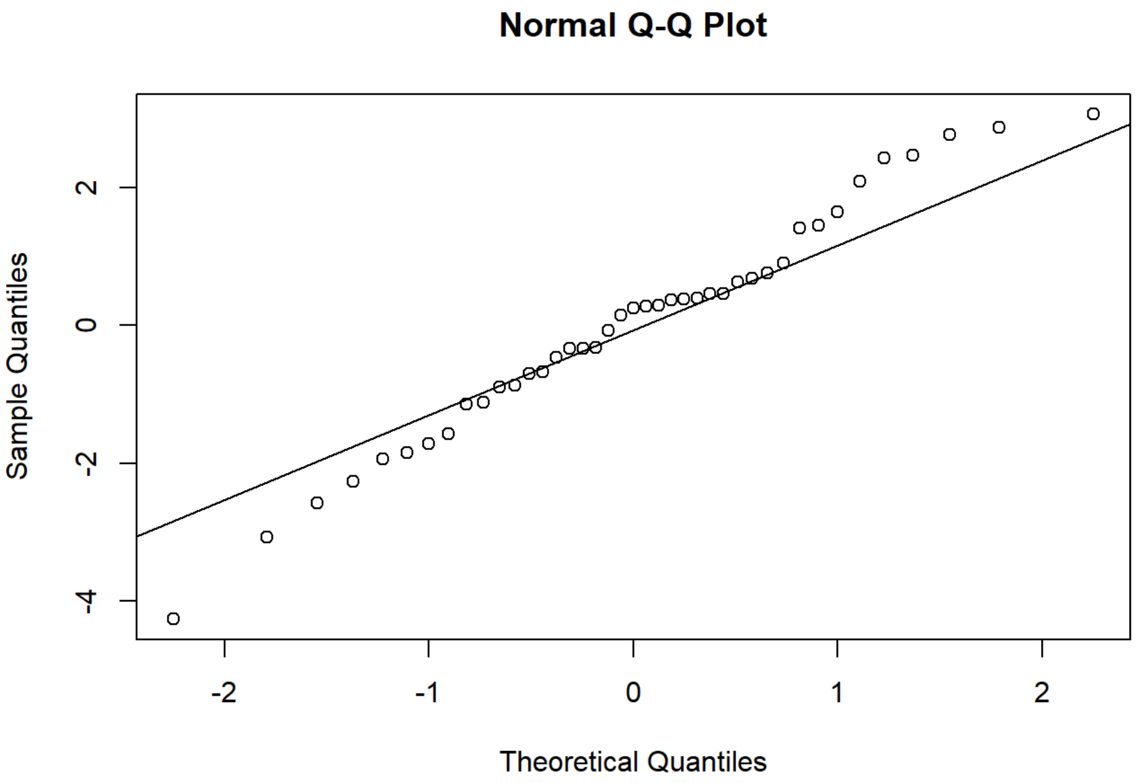
```
#making the fitted values:
fits_Small = fitted(lm(time~weight, data = smallDogs))
#making the residual-vs-fit plot:
plot(fits_Small, res_Small)
abline(h=0)
```



[For the small dog, all Normality, Equal Variance, and Linearity assumptions are “Very plausible”, because on the normal q-q plot The points pretty much fall on the line so normality is “Very plausible”. There does not appear to be a strong violation of the equal variance assumption: The points appear to be equally vertical spread from left-to-right. So Equal Variance is “Very plausible”. There also does not appear to be a strong violation of the linearity assumption: The dots appear to be randomly scattered around the line with no clear trend. So linearity is “Very plausible”.]

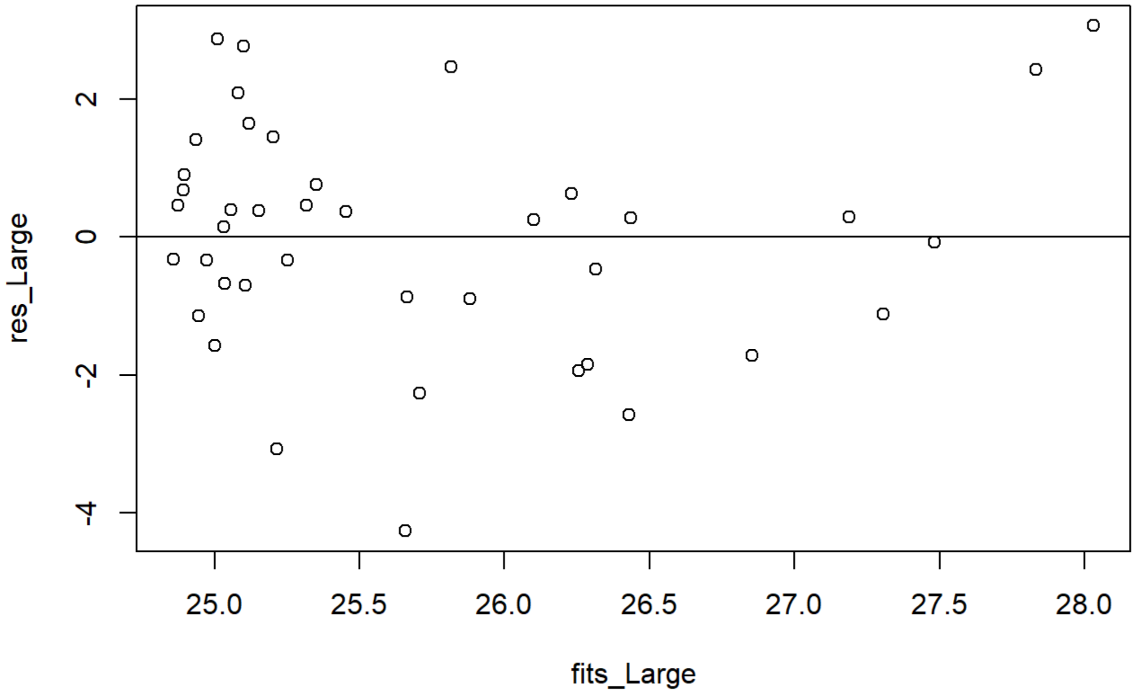
Hide

```
#making the residuals
res_Large = residuals(lm(time~weight, data = largeDogs))
qqnorm(res_Large)
qqline(res_Large)
```



Hide

```
#making the fitted values:
fits_Large = fitted(lm(time~weight, data = largeDogs))
#making the residual-vs-fit plot:
plot(fits_Large, res_Large)
abline(h=0)
```



[For the large dog, on the normal q-q plot The points don't really fall on the line especially on the right and left so normality is " Not plausible ". There does not appear to be a strong violation of the equal variance assumption: The points appear to be approximately equally vertical spread from left-to-right. So Equal Variance is" plausible". There seems to appear a slight violation of the linearity assumption: The dots appear to be slightly focused on the left and in some parts there are very far from the line 0. So linearity is "not plausible".]

- (5pts) Based on your work above, state whether you "Strongly Agree," "Agree," "Disagree," or "Strongly Disagree" with Professor Dogg's hypothesis. Explain your reasoning in 2-4 sentences.

[I disagree overall, Because of multiple reasons. although the p-value for both these new data sets are less than 0.05 and slope for the small dog data set is negative and slop for the large data set is positive, some assumptions such as normality and linearity for the large data set were not very plausible. Moreover, for the internal validity issue still remains, because as discussed before, there are other factors that might affect the runtime of a dog such as its age, race, etc., that none of them are considered in this study]

Question 2: Using Linear Regression to Measure the Effects of Studying, More or Less (30 points)

In this problem, we'll focus on this dataset:

Hide

```
study = read.csv("https://raw.githubusercontent.com/zjbranson/stat309fall2022/master/study.csv")
```

This dataset is based on a (simulated) experiment where students are randomized to study more or fewer hours for the next (third of seven) test in a class. For example, if a student studied for 4 hours for the second test, and they were randomized to study for 2 more hours, they would study for 6 hours for the third test. (To be extremely unrealistic, let's assume the students actually followed their randomized change in studying.) The explanatory variable, `deltaH` , is the change in the number of hours studied (after minus before), and the outcome variable, `deltaTS` , is the change in test score (third test score minus second test score).

a. (14pts) For this part, answer the following questions.

- (4pts) First, run the appropriate linear regression analysis for this problem. For this part, you just need to include your code and summary output.

Hide

```
summary(lm(deltaTS~deltaH, data = study))
```

```
##
## Call:
## lm(formula = deltaTS ~ deltaH, data = study)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -25.166  -3.175   1.810   3.834  17.842
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   9.1711     1.4404   6.367 3.29e-07 ***
## deltaH        3.9947     0.5771   6.922 6.59e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8.519 on 33 degrees of freedom
## Multiple R-squared:  0.5921, Adjusted R-squared:  0.5798
## F-statistic: 47.91 on 1 and 33 DF,  p-value: 6.586e-08
```

- (5pts) What is the interpretation of the intercept **estimate** for this problem, within the context of this dataset?

[We see that our estimate of the intercept is 9.1711, meaning that the average score change of a student who has not changed his/her study hours(not increased or decreased the duration of his/her study) is estimated to be 9.1711. The intercept does have a scientifically meaningful interpretation, because it is possible for a student not to change his or her study hours, and to be able to have a better score of 9.1711 (the test might be easier). And the data set also has two people doing that where deltaH is 0. In terms of scientific conclusions, the null hypothesis for the intercept of this study is that H0:β0=0 (intercept is 0), and the alternative hypothesis that HA: β0≠0 (none-zero intercept). since the P-value is less than 0.05(< 3.29e-07), then we have enough evidence to reject the null hypothesis]

- (5pts) What is the interpretation of the slope **estimate** for this problem, within the context of this dataset?

[our estimate of the slope is 3.9947, meaning that average change in test score is estimated to increase by 3.9947 for every one-hour increase in number of study hours. the null hypothesis for the slop of this study is that H0:β1=0 (slope is 0), and the alternative hypothesis that HA: β1≠0 (none-zero slop). since the P-value is less than 0.05(6.59e-08), then we have enough evidence to reject the null hypothesis. so this test in this study shows there is a significant correlation between the change in the number of hours of a student and change in his/her number of hours studied]

b. (6pts) For this dataset, what conditions must be met to make the interpretation of the intercept meaningful **and** trustworthy? Do we meet those conditions? Explain in 1-3 sentences.

[Other than linear regression conditions (linearity, fixed X, independence, equal variance, and normality); There are multiple factors that we should meet. first the students should be randomly assigned. Which is done. Second, the test's difficulty level should be similar to the previous test which we have no data in description about that, however the data set shows an intercept of 9.17 which might be used to say that the test after randomization was easier than the previous test, therefore we don't meet this condition. Third, other factors can affect the results such as students IQ, amount of sleep they got the night before the test, their stress level, etc., and students who participated in the study should have the same conditions in terms of the mentioned factors such as IQ, etc. although students are randomly assigned, this is not enough for internal validity therefore there

is no enough data showing that whether we meet this condition or not, considering that it would be so hard to find students with all same mentioned factors, I assume that we do not meet this condition. Fourth, I do not see a control group, however I see two treatment groups (fewer vs more hours of study). There should be another group with students who did not change their number of study hours to better compare the results of treatments. So far we only have 2 subjects in the data set with 0 deltaH, which is not enough at all to be considered as the control group. Hence we also do not meet this condition.]

c. (10pts) Now let's consider the original setup for this problem: In this experiment, students were randomized to study more or fewer hours for the third test, compared to the second test. Consider the following claims:

1. The third test was easier than the second test.
2. Studying more led to a bigger change in test scores.

Use what you found in the previous parts to assess these claims. In particular, for each claim, state whether you Agree or Disagree, and explain your reasoning in 1-3 sentences. (For the sake of this problem, assume that all of the modeling assumptions for the linear regression model hold.)

[Because of the reasons mentioned in the last part (B), there are many conditions that did not meet. so although the statical results shows that intercept is 9.17 with the p-value of 3.29e-07 (as a sign of The third test was easier than the second test) and the slope is positive 3.9947 with the p-value of 6.59e-08 (as a sign of Studying more led to a bigger change in test scores), I do not trust the results. And I disagree with both of these claims. Specifically for the first claim there should be a separate control group with students who did not change their number of study hours which there is not. And specifically for the second claim, we cannot make this claim because other factors can affect the results such as students IQ, amount of sleep they got the night before the test, their stress level, etc., and students who participated in the study should have the same conditions in terms of the mentioned factors such as IQ, etc. although students are randomly assigned, this is not enough for internal validity therefore there is no enough data showing that whether we meet this condition or not, considering that it would be so hard to find students with all same mentioned factors, I assume that we do not meet this condition]

d. (5pts, ONLY REQUIRED FOR 36-749 STUDENTS; BONUS QUESTION FOR 36-309 STUDENTS) For this dataset, subjects were indeed randomized to study more or fewer hours for the next test. In the last homework, we got practice assessing if a randomization was "done well" by quantifying *internal validity*. However, we've discussed internal validity within the context of a discrete number of treatment groups. In this case, the "treatment" that was randomized to subjects (the change in number of hours studied) is a continuous variable, such that our previous discussion of internal validity doesn't immediately apply. In this problem we'll consider how to measure internal validity with a continuous treatment variable like `deltaH` . Answer the following questions.

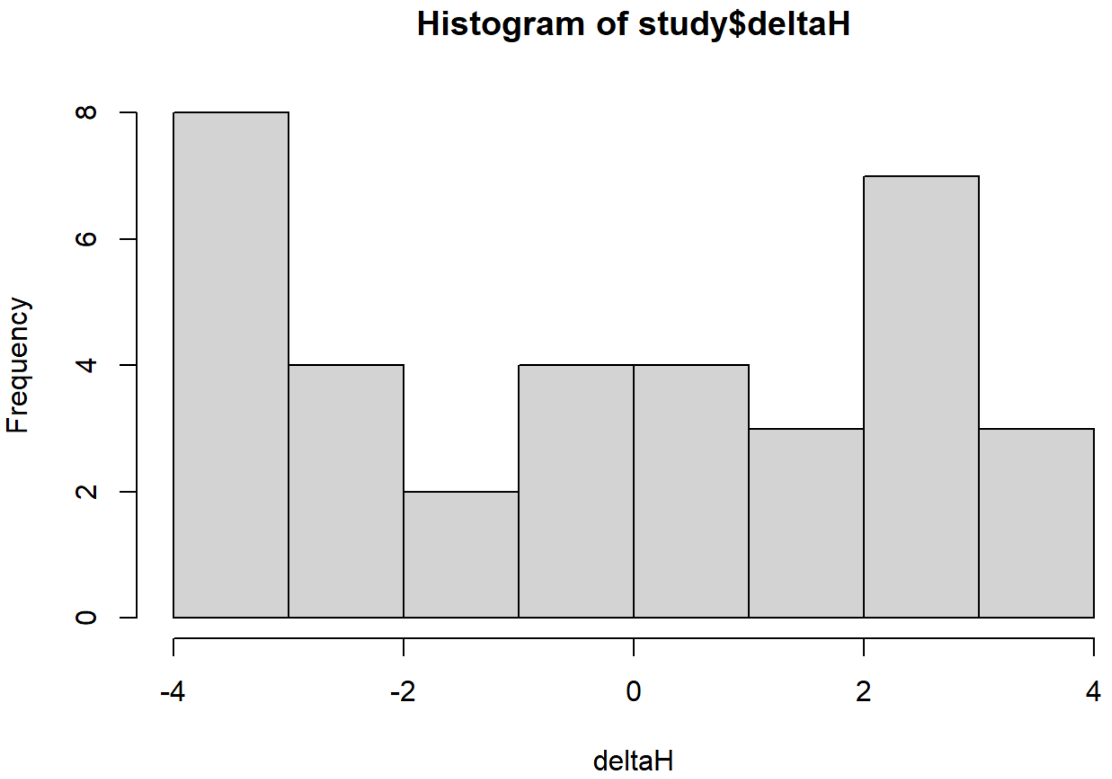
- (2pts) Here is a further description of the randomization protocol for this experiment: "Each subject was randomized to some continuous value of `deltaH` , such that each subject was equally likely to have a positive or negative change in hours assigned; furthermore, each subject was just as likely to have a large change in hours assigned as a small change in hours, ranging from -4 to 4." Given this description, if the randomization was done well in this study, what we would expect the *distribution* of `deltaH` to look like across subjects? Explain in 1-2 sentences.

[I would expect to see A random distribution, a distribution that lacks an apparent pattern and can have several peaks, becasue each subject was randomly assigned and was just as likely to have a large change in hours assigned as a small change in hours, ranging from -4 to 4.]

- (3pts) Given your answer above, do you think the randomization was done well in this study? State Yes or No, and explain your answer in 1-3 sentences. Use some kind of EDA (graphical or non-graphical) to answer this question.

Hide

```
hist(study$deltaH,
xlab ="deltaH",
breaks = 10)
```



[Yes, I think the randomizatoion was done well. beacsue the histogarm shows that a set of random numbers are selecetd to be used, and the histogram reveals a random distribution that lacks an apparent pattern and has several peaks]