James Scruggs

Group Work 3

4 Regression

This week we will consider regression analysis. We will examine relationships between paired variables. For example, suppose we are interested in comparing age and income. Each person in the study has one number for age and one number for income. Since the numbers came from the same person, those numbers are paired together.

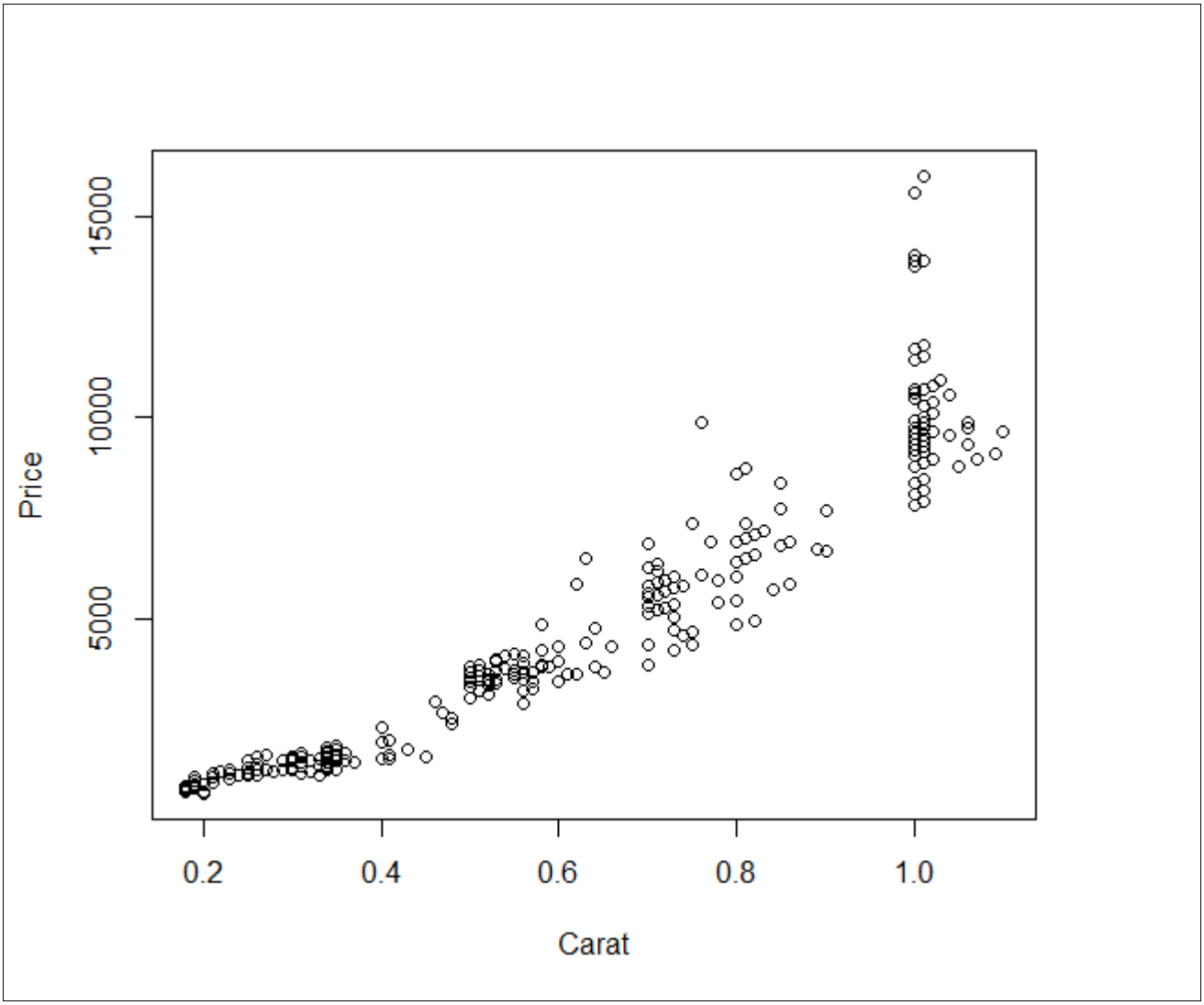
First, we will examine a graph to display the data. Then we will address how to tell if there is a relationship between the variables. Then we will see how such a relationship can be used to predict one variable based on the other.

Scatter plots

1. Load and attach the file 4c.csv. This file contains information on a sample of diamonds. We want to determine if there is a relationship between the number of carats and the cost (in Singapore dollars). We will call the number of carats the ***predictor*** and the cost the ***response***. Create a scatter plot of this data.

plot(Price ~ Carat)

Paste your graph below.

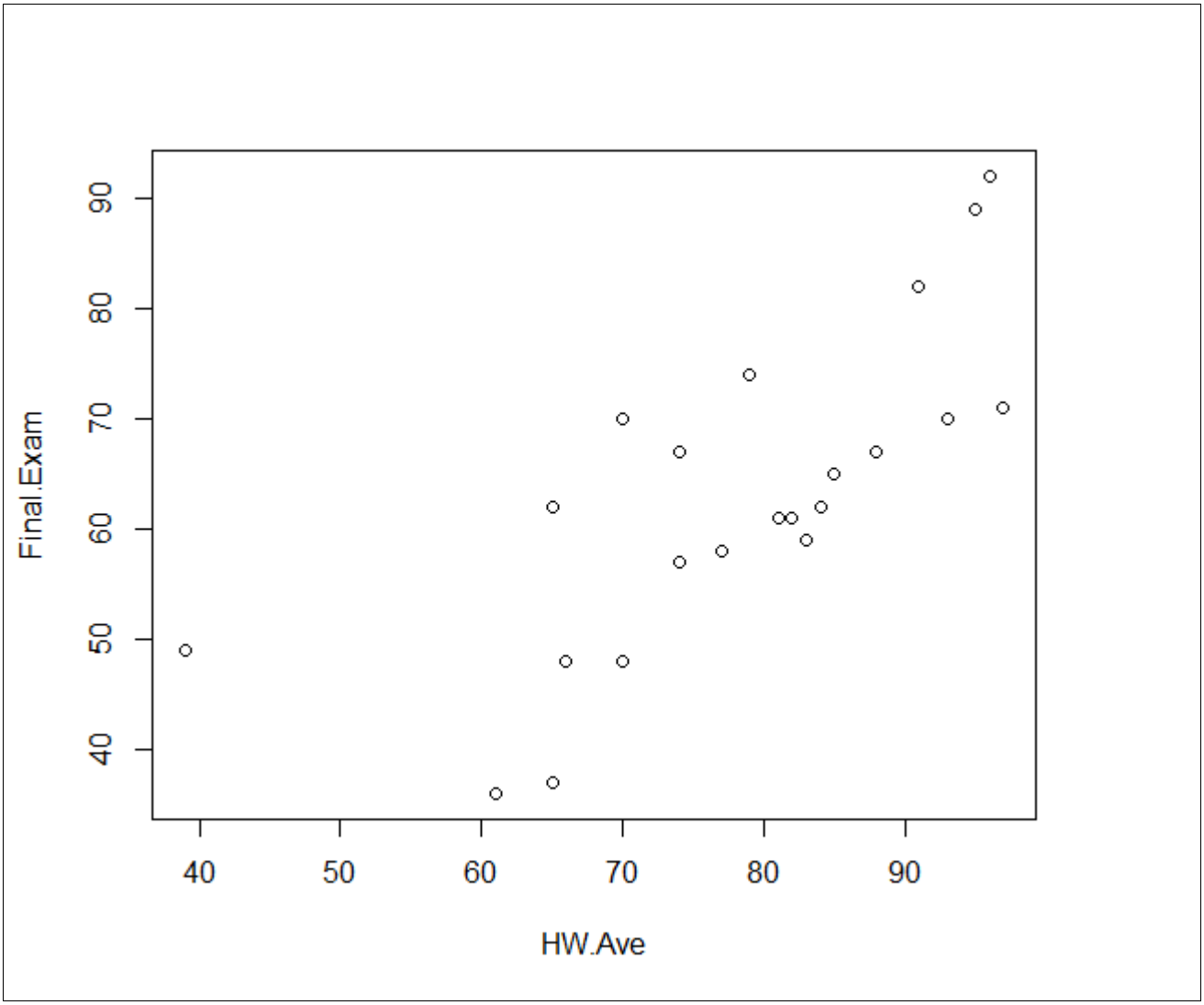


Each dot in the graph represents one diamond. The x-value of the dot (it's shadow on the horizontal axis) is the number of carats. The y-value (vertical axis) is the cost.

2. Does it look like there is a relationship between the values? If so, would you describe the relationship as linear. ("Linear" meaning that the dots look like they fall along a straight line with no curves.) What do you can say about correlation coefficient.

Yes, there is a relationship between the values. I would describe it as linear. The correlation coefficient is close to one.

3. Load the file grades.csv. (Note: These grades are not from any one class, I changed numbers to protect privacy.) Create a scatter plot in which the predictor (x variable) is the HW Ave, and the response (y variable) is the Final, and paste it below.



This graph shows a relationship, and since there is no obvious curving behavior, we will model it as a linear relationship.

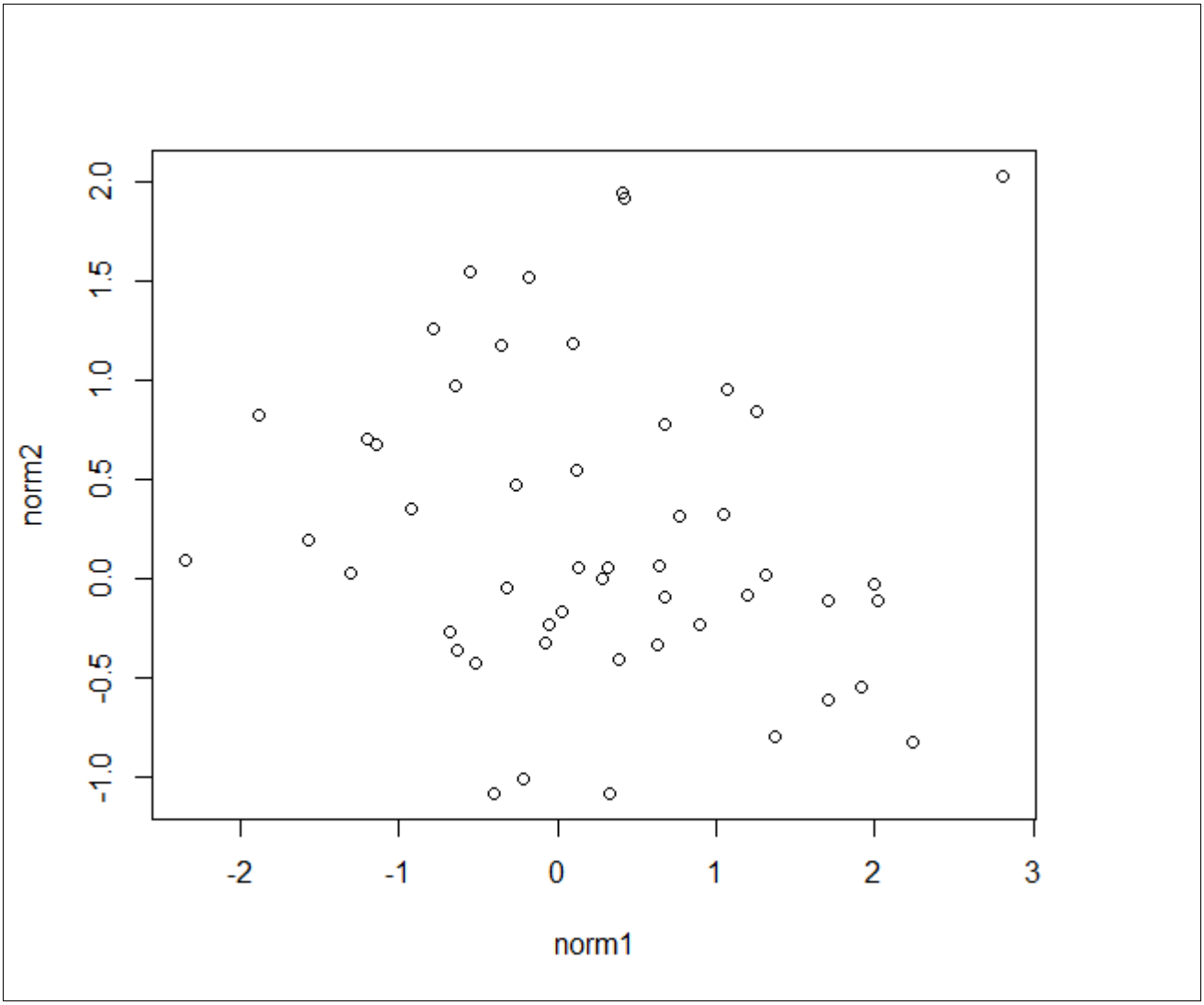
4. Let's look at what a graph would look like if there was no relationship.

norm1 <- rnorm(50,0,1)

norm2 <- rnorm(50,0,1)

plot(norm2 ~ norm1)

We know that the two columns of numbers actually have nothing to do with each other. Paste your plot below. How would you describe the picture?



Correlation

If the data in the scatter plot shows a linear trend, we may want to quantify the strength of the trend. We will use the ***correlation coefficient***. This value, r, ranges from -1 to 1. If a data set had a correlation coefficient of +1, then the scatter plot would show dots on a perfect increasing line, with no variation. If r = -1, then the points are on a perfect line which slopes down. If r = 0, then the variables in the dataset have no relationship.

5. Using the grades data to determine the correlation coefficient.

cor(Final.Exam, HW.Ave) # Note that cor(HW.Ave,Final.Exam) gives the same result.

0.7406325

6. Using your faked data, determine the correlation coefficient.

-0.1121631

Notice that where there was a positive linear relationship, the correlation coefficient was positive and closer to 1 than zero. Where there was no relationship, the coefficient was close to zero.

Linear Regression

7. We have determined that there is a linear relationship between HW Ave and Final. Let's describe that relationship with a mathematical function that can be used to predict the cost of future diamonds.

Check all the conditions to satisfy linear regression analysis and give comments.

lm.grades<-lm(Final.Exam~ HW.Ave) #creates a linear regression model

#lm stands for linear model

summary(lm.grades))

Paste the output below.

```
Call:
lm(formula = Final.Exam ~ HW.Ave)

Residuals:
      Min        1Q    Median        3Q       Max
-16.037  -5.747  -3.934   9.034  15.869

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   3.2735    12.2891   0.266    0.793
HW.Ave        0.7656     0.1553   4.930 8.08e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 9.892 on 20 degrees of freedom
Multiple R-squared:  0.5485,    Adjusted R-squared:  0.526
F-statistic:  24.3 on 1 and 20 DF,  p-value: 8.083e-05
```

The regression equation is Final Exam = 3.2735 + 0.7656 HW Ave. There are two numbers in this equation. The first, 3.2735, is the intercept. You can see it in the "Intercept" row of the table labeled "Coefficients." It is interpreted as the value of the response if the predictor is zero. In this example, if you got a zero homework grade, we predict that your final exam grade would be 3.3. In some

examples, the intercept makes no sense. For example, a diamond with zero carats doesn't exist. So the interpretation of the intercept depends on what your data look like.

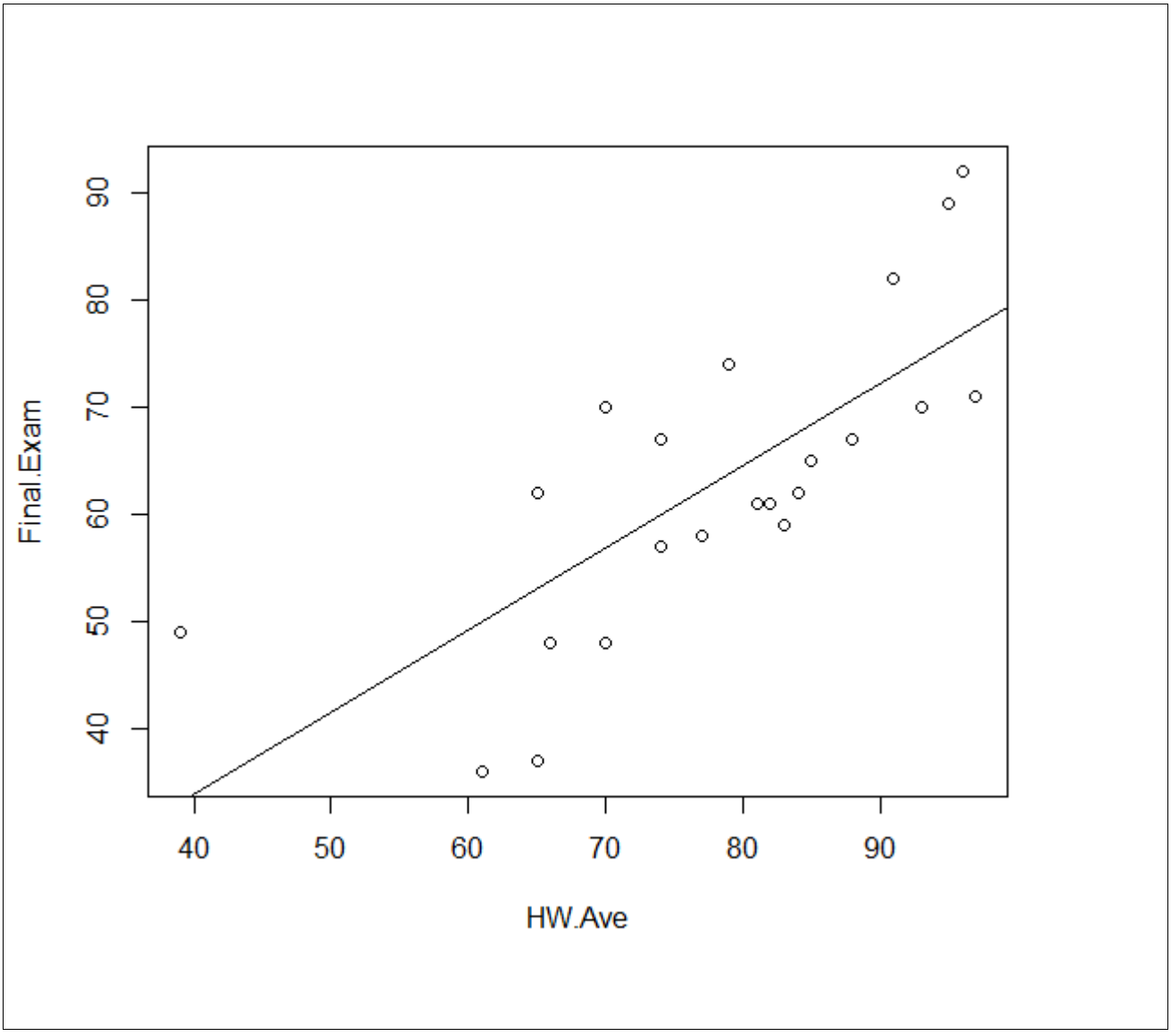
The second number, 0.7656, is the slope. You can see it in the "HW.Ave" row of the same table. It is interpreted as the amount the response changes if the predictor changes by 1. In this example, if we increase the HW Ave by one point, we expect the Final Exam grade to increase by 0.77 points.

8. Let's look at a graph of the regression equation on the scatter plot of the data.

plot(Final.Exam~ HW.Ave)

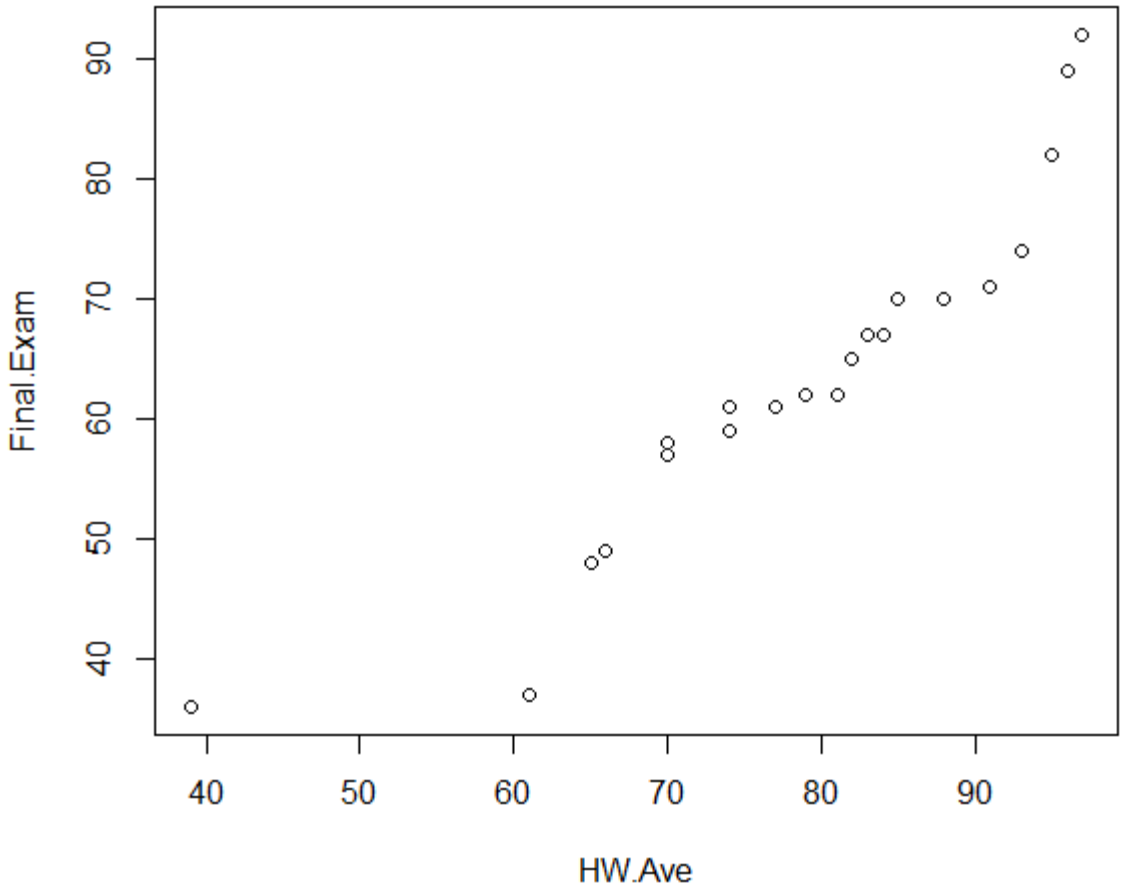
abline(lm.grades)

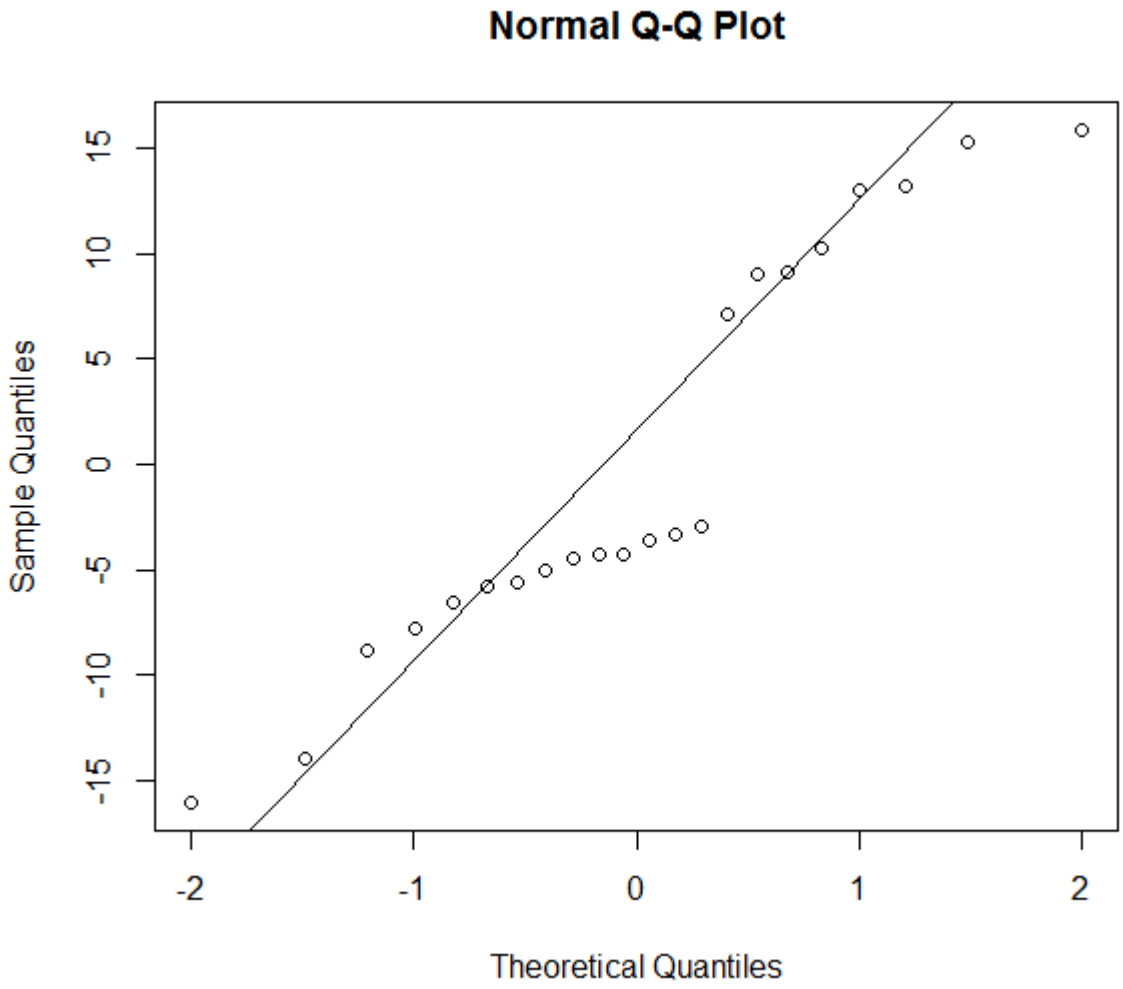
Paste the graph below.



We can use the regression equation to make predictions. If we want to predict the final exam grade for someone who made a 70 on the homework, we replace the HW Ave variable with 70, getting Final Exam = 3.2735033 + 0.7655877 * 70 = 56.86.

9. Check normality condition and constant covariance condition of error and attach plots below.





10. What final exam grade do we expect someone who got an 85 homework average to get?

68.35

Of course, you can have the computer do the predicting for you.

predict(lm.grades,data.frame(HW.Ave=70),level=0.95,interval="prediction")
 # the predict function requires that the new input be in a data frame, so the function
 # data.frame forces the HW.Ave=70 to be interpreted as a data frame.

Notice we get the value of 56.86 that we got above, as well as an interval that describes the uncertainty in that prediction. The interval is called a "prediction interval", and we are 95% sure that the actual value will be within this interval.

11. What is the prediction interval for the final exam grade for someone who got an 85 on homework?

lwr: 47.12704 upr: 89.56988

There are a lot of other numbers in the summary of the linear model, most of which are used to quantify how good a fit to the data this function is. MTSU offers a full semester on regression analysis if you are interested in learning more.