Aman Patel

December 11, 2020

STAT-S 350

Problem Set 6

1. ISIR 13.4 Question 2

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Brown | Yellow | Red | Blue | Orange | Green |
| 0.13 | 0.14 | 0.13 | 0.24 | 0.20 | 0.16 |

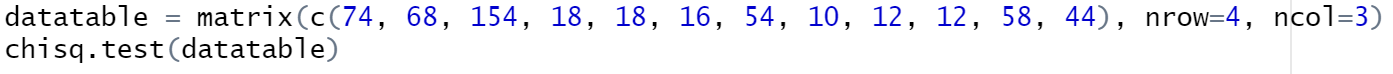
* 1. Claimed proportions

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Brown | Yellow | Red | Blue | Orange | Green |
| 121 | 84 | 118 | 226 | 226 | 123 |

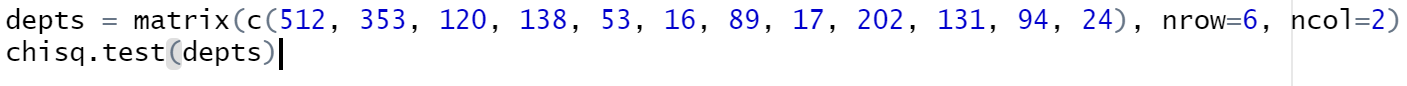
* 1. Observed counts
  2. Expected counts

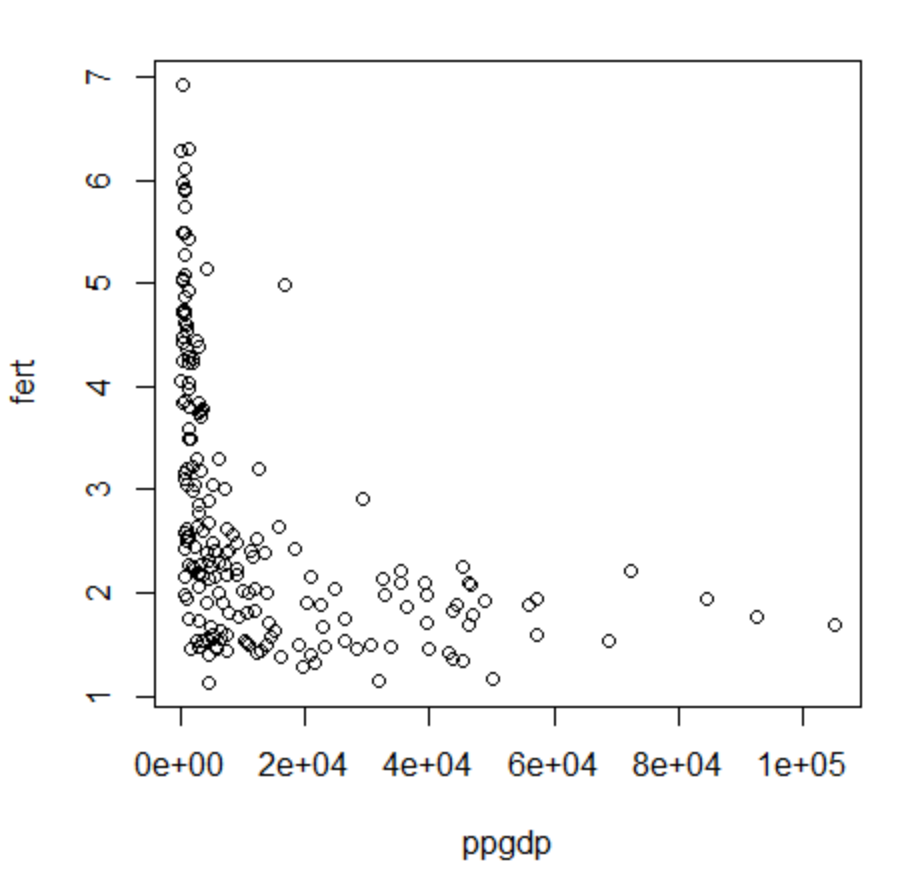
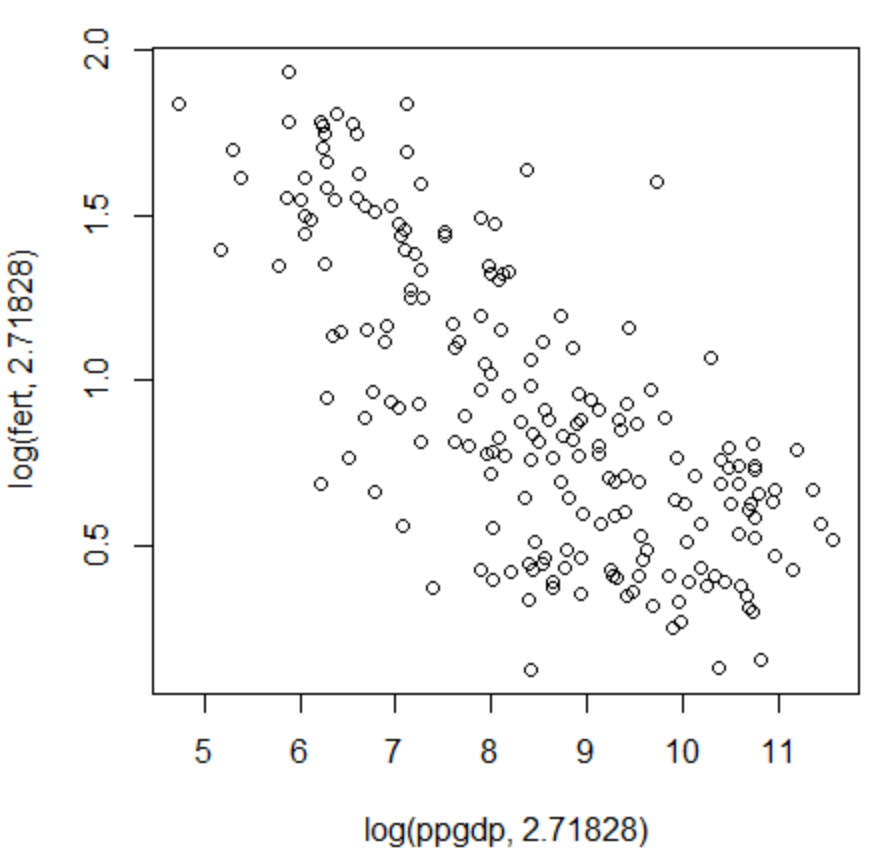
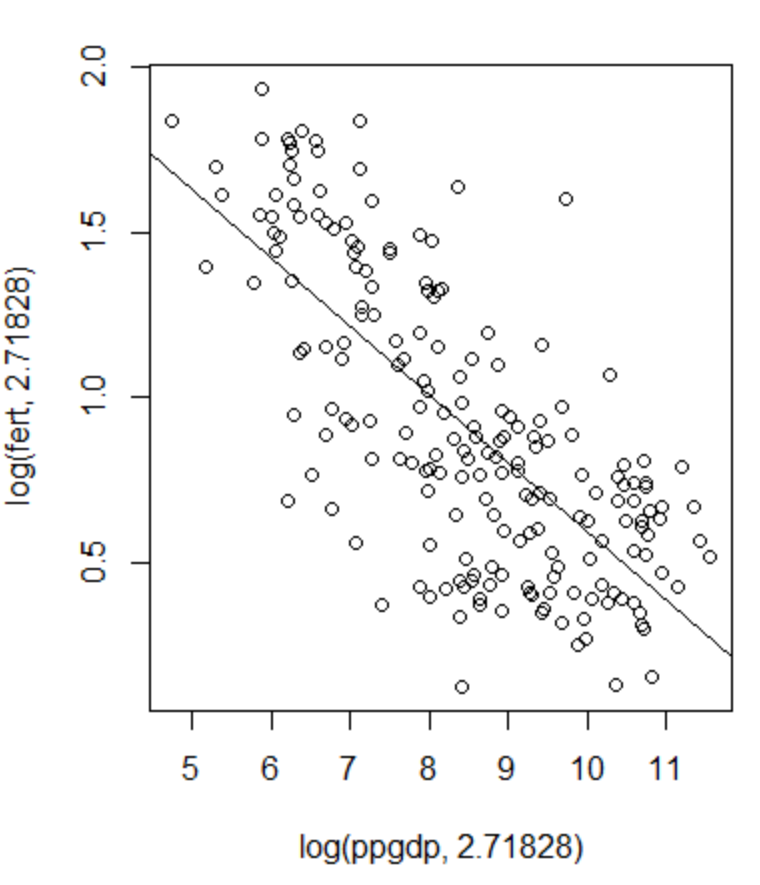
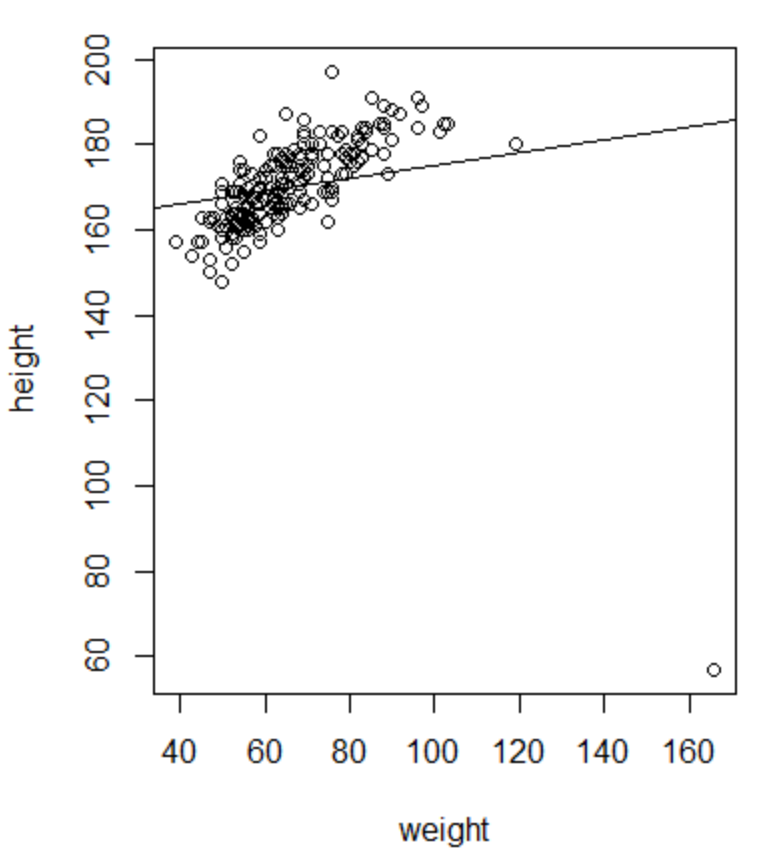
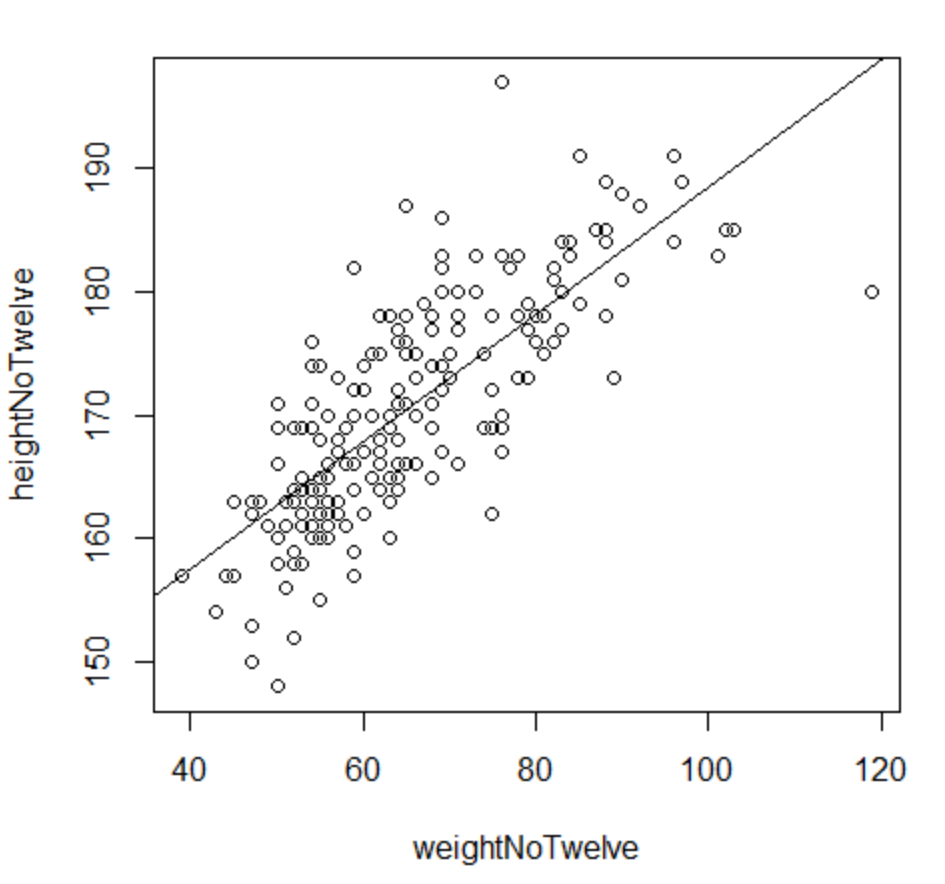
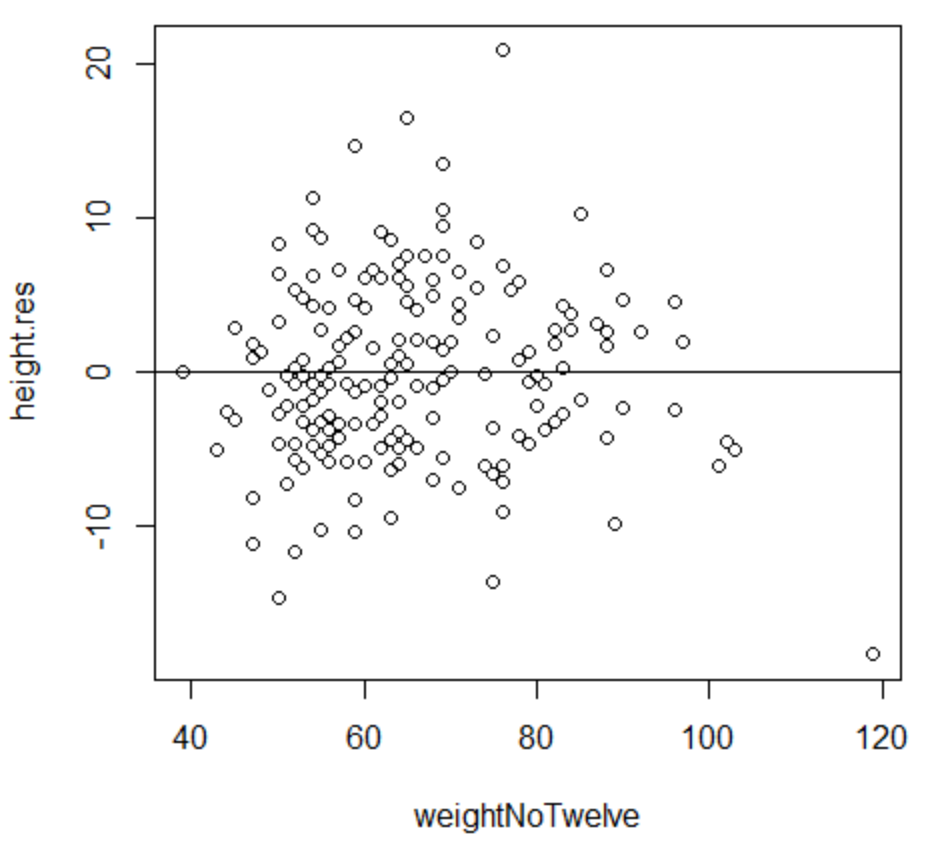
|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Brown | Yellow | Red | Blue | Orange | Green |
| 116.74 | 125.72 | 116.74 | 215.52 | 179.6 | 143.68 |

* 1. Chi-squared
     1. The p-value found is very small, suggesting the claimed proportions are not credible.

1. ISIR 13.4 Question 11
   1. R code
   2. Results
      1. With an incredibly small p-value, the data do not provide evidence that a patient’s response to treatment varies by histological type.
2. Graduate students
   1. Dept and gender are independent
      1. Assuming the same total admissions for each department, the following values are expected:

|  |  |  |
| --- | --- | --- |
| Department | Men admitted | Women admitted |
| A | 300.5 (301) | 300.5 (300) |
| B | 185 | 185 |
| C | 161 | 161 |
| D | 134.5 (135) | 134.5 (134) |
| E | 73.5 (74) | 73.5 (73) |
| F | 20 | 20 |

* + 1. Chi-squared tests should not be used if any value is less than 5, but all values in these tables are greater than 5.
  1. Pearson chi-square test using R
  2. The p-value is the probability of obtaining a result that is at least as extreme as the observed results. This p-value is compared to a predetermined significance level to determine whether to reject or accept the null hypothesis, which is the claim. If the p-value is larger than the significance level, the null hypothesis is accepted, and it is rejected if the p-value is smaller than the significance level. Assuming a significance level of 0.05, this p-value is much smaller, and the null hypothesis can be rejected.

1. *UN* from *carData*
   1. Scatterplot  
      1. The graph represents an exponential curve mirrored about the y-axis
      2. A straight-line mean does not seem plausible
   2. Ln Scatterplot  
      1. A linear regression model would most likely work with this data, but the data would not be concentrated around the line
   3. Linear regression (see attached R)
   4. Plot w/ linear regression
   5. Significance Test
      1. Results: Reject the null, p-value is much smaller than significance level of 0.05
   6. Coefficient of determination
      1. The coefficient of determination is a measure of the degree of correlation between two variables. Its values range from 0 to 1, where 0 represents no correlation and 1 represents perfect correlation. A value of represents mediocre correlation.
      2. Point prediction:
2. *Davis* from *carData*
   1. Linear regression (see attached R)
   2. Scatterplot w/ linear regression  
      1. The line does not fit the data well
   3. Significance test
      1. The p-value is smaller than 0.05, we can reject the null hypothesis that the slope is 0 and we can assume changing the weight does influence the height
   4. Not all the height variation is explained, as the slope of the linear regression is shallower than the ideal line of best fit
   5. Remove value 12
      1. Linear regression
      2. Scatterplot w/ linear regression  
           
           
           
           
           
           
           
         1. The linear regression fits much better after the removal
      3. Significance test
         1. The p-value is significantly smaller than it was before, furthering the assessment that weight does have an influence on height
      4. The results prior to removal still led to a rejection of the null hypothesis, but after removal, the p-value was much smaller, indicating a stronger correlation than before removal
   6. Residual plot  
        
        
        
        
        
        
        
      1. The residual plot seems to have a fairly even distribution above and below the x-axis. There is slightly more clustering on the left side of the graph as opposed to the right.