Q1.

One measure of a university's effectiveness is how many of its students graduate. Universities can try to improve rates of completion by offering services, such as advising, reduced class size, and better faculty; adjusting incentives, such as cost to students; and by selective admission, such as percent of students admitted. You can investigate some of these factors, because the federal government made this data available here (MERGED2013_PPv2.csv).

a) Which of the following variables is related to a University’s completion rate (C150_4): number of undergraduates in degree programs (UGDS), average faculty salary (AVGFACSAL), average cost per year (COSTT4_A), average SAT scores (SAT_AVG), percent of students with PELL grants (low income families, PCTPELL), admission rate (ADM_RATE)? Provide summary of statistical results for all tests (r, df, p) and describe the relationships between the predictors and outcome in sentences (including direction of relationship).

- data: UGDS and C150_4
  t = 6.177, df = 2446,
  p-value = 7.628e-10
  alternative hypothesis: true correlation is not equal to 0
  95 percent confidence interval:
  0.08473246 0.16275130
  sample estimates:
  cor = 0.1239334
    - For undergraduates in degree programs and University’s completion rate, the r = 0.1239334, df = 2446, and p = 7.628e-10. We reject the null hypothesis and report that university completion rate is related to the number of undergraduates in degree programs with a correlation coefficient of 0.124. This is a weak positive relationship.
- data: AVGFACSAL and C150_4
  t = 30.25, df = 2428, p-value < 2.2e-16
  alternative hypothesis: true correlation is not equal to 0
  95 percent confidence interval:
  0.4936900 0.5514738
  sample estimates:
  cor = 0.523183
    - For average faculty salary and University’s completion rate, the r = 0.523183, df = 2428, and p = 2.2e-16. We reject the null hypothesis and report that university completion rate is related to the average faculty salary with a correlation coefficient of 0.523. This is a positive relationship.

-
  ```
  data:  COSTT4_A and C150_4
  t = 31.191, df = 2385, p-value < 2.2e-16
  alternative hypothesis: true correlation is not equal to 0
  95 percent confidence interval:
   0.5091345 0.5661538
  sample estimates:
        cor = 0.5382598
  ```
    - For average cost per year and University’s completion rate, the r = 0.5382598, df = 2385, and p = 2.2e-16. We reject the null hypothesis and report that university completion rate is related to the average cost per year with a correlation coefficient of 0.538. This is a positive relationship.

-
  ```
  data:  SAT_AVG and C150_4
  t = 51.934, df = 1375, p-value < 2.2e-16
  alternative hypothesis: true correlation is not equal to 0
  95 percent confidence interval:
   0.7952032 0.8309441
  sample estimates:
       cor = 0.813842
  ```
    - For average SAT scores and University’s completion rate, the r = 0.813842, df = 1375, and p = 2.2e-16. We reject the null hypothesis and report that university completion rate is related to the average SAT scores with a correlation coefficient of 0.814. This is a strong positive relationship.

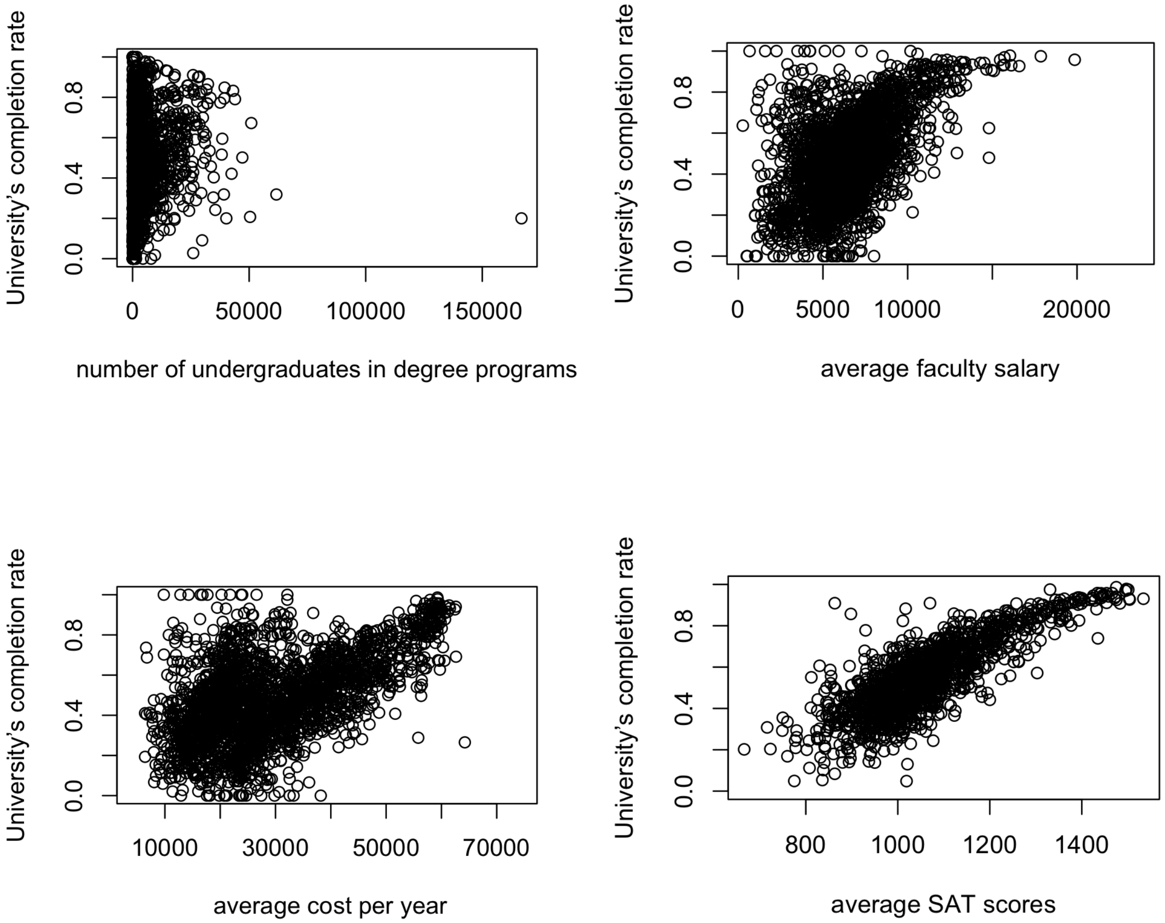
-
  ```
  data:  PCTPELL and C150_4
  t = -31.149, df = 2444, p-value < 2.2e-16
  alternative hypothesis: true correlation is not equal to 0
  95 percent confidence interval:
   -0.5608720 -0.5041065
  sample estimates:
         cor = -0.5330889
  ```
    - For percent of students with PELL grants and University’s completion rate, the r = -0.5330889, df = 2444, and p = 2.2e-16. We reject the null hypothesis and report that university completion rate is related to the percent of students with PELL grants with a correlation coefficient of -0.533. This is a negative relationship.

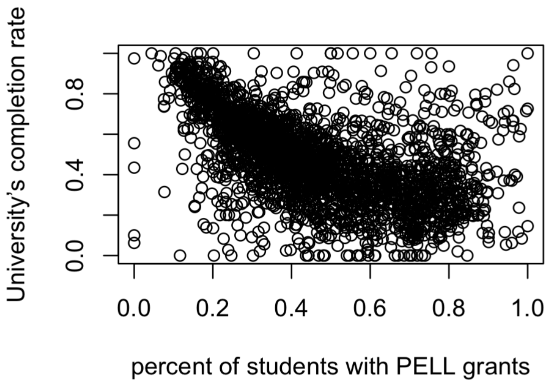
-
  ```
  data:  ADM_RATE and C150_4
  t = -14.432, df = 1794, p-value < 2.2e-16
  alternative hypothesis: true correlation is not equal to 0
  95 percent confidence interval:
   -0.3633545 -0.2804510
  sample estimates:
  ```

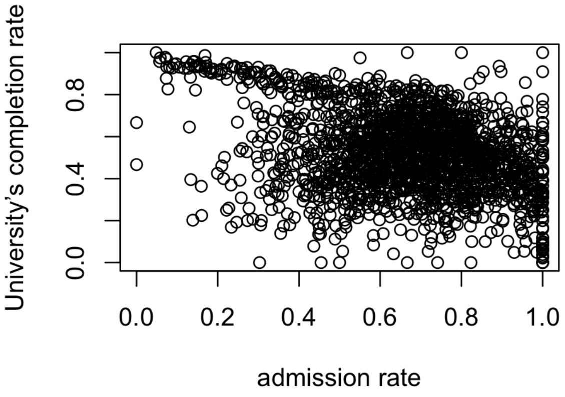
cor = -0.3225211

- For admission rate and University's completion rate, the r = -0.3225211, df = 1794, and p = 2.2e-16. We reject the null hypothesis and report that university completion rate is related to the admission rate with a correlation coefficient of -0.323. This is a negative relationship.

b) Identify and plot the strongest relationship between the predictor variables and the outcome variable. Label your axes with understandable labels (e.g. "Percent of students receiving PELL grants" not "d$PCTPELL").







- plot(SAT_AVG, C150_4, xlab="average SAT scores", ylab="University's completion rate")
    - Average SAT (the predictor variable) has a strong positive correlation with University completion rate (outcome variable).

c) Did all the correlation tests have the same degrees of freedom? Discuss why or why not.

- Not all the correlation tests have the same degrees of freedom. This is likely due to varying numbers of data entries affecting the maximum number of logically independent values.

d) Run a regression model using the two variables from question 2. Report the coefficient for the predictor variable and its p value.

-
  ```
  Residuals:
      Min       1Q   Median       3Q      Max
  -0.45471 -0.06395  0.00423  0.06498  0.57566

  Coefficients:
               Estimate Std. Error t value
  (Intercept) -6.052e-01  2.238e-02  -27.04
  SAT_AVG      1.088e-03  2.094e-05   51.93
              Pr(>|t|)
  (Intercept)   <2e-16 ***
  SAT_AVG       <2e-16 ***
  ---
  Signif. codes:
  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
  ```

```
Residual standard error: 0.1017 on 1375 degrees of freedom
  (6427 observations deleted due to missingness)
Multiple R-squared:  0.6623,	Adjusted R-squared:  0.6621 
F-statistic:  2697 on 1 and 1375 DF,  p-value: < 2.2e-16
```

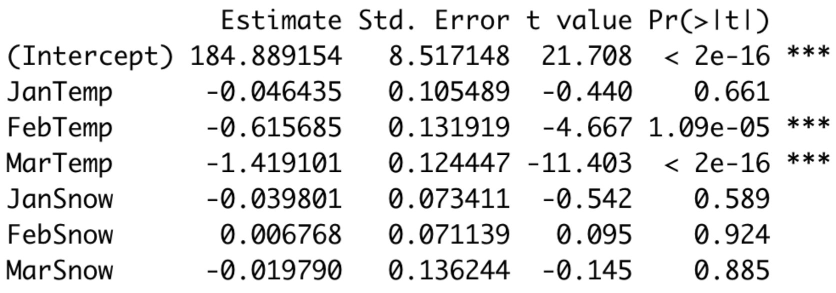
- The regression coefficient for the predictor variable is 1.088e-03, t(1375) = -27.04, and its p value is less than 2e-16.

Q2.

For this question you will use the cherry blossom data set

The goal of this assignment is to build the best possible linear model to explain the day of peak bloom. Using the data from 1921 to 2015 (*** note do not include 2016 data) build a multiple linear regression model to predict the Peak Bloom Date.

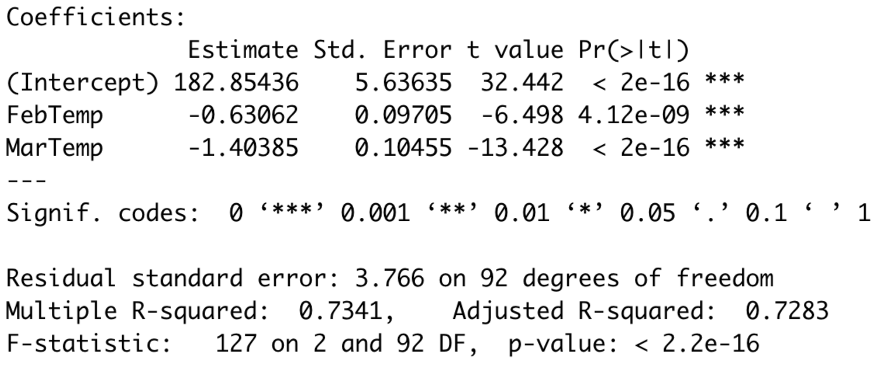
a) Construct a multiple linear regression model including all predictors for snow and temperature for January, February, and March. Identify the predictors that are significantly related to Peak Bloom Date.



```
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) 184.889154   8.517148  21.708  < 2e-16 ***
JanTemp      -0.046435   0.105489  -0.440    0.661    
FebTemp      -0.615685   0.131919  -4.667 1.09e-05 ***
MarTemp      -1.419101   0.124447 -11.403  < 2e-16 ***
JanSnow      -0.039801   0.073411  -0.542    0.589    
FebSnow       0.006768   0.071139   0.095    0.924    
MarSnow      -0.019790   0.136244  -0.145    0.885    
```

- Looking at the p values, it looks that February and March temperatures are significantly related to Peak Bloom Date.

b) Rerun the regression model only including significant predictors. What is the estimated regression equation? Describe the relationship between each predictor and the outcome (remember to consider, *significance*, *sign*, and *size*). How much variance is explained by this model?



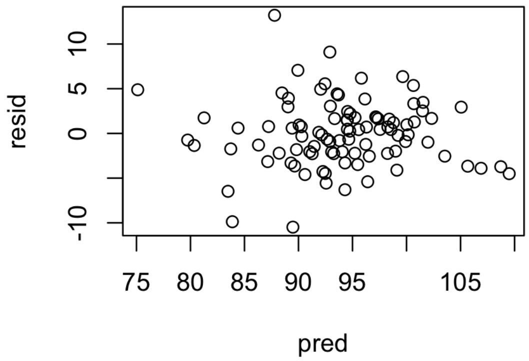
```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) 182.85436    5.63635  32.442  < 2e-16 ***
FebTemp      -0.63062    0.09705  -6.498 4.12e-09 ***
MarTemp      -1.40385    0.10455 -13.428  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.766 on 92 degrees of freedom
Multiple R-squared:  0.7341,	Adjusted R-squared:  0.7283 
F-statistic:   127 on 2 and 92 DF,  p-value: < 2.2e-16
```

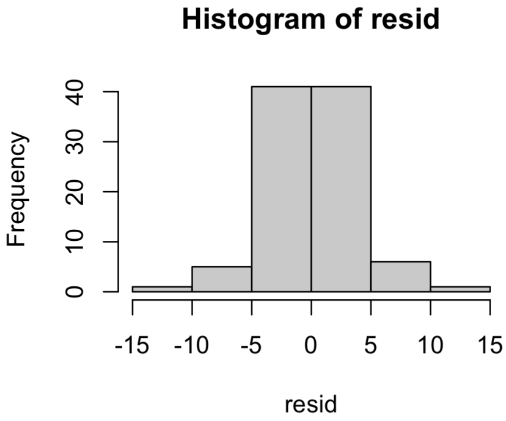
- The estimated regression equation is y = -0.63*FebTemp + -1.40*MarTemp + 182.85.
- February temperature significantly predicted peak bloom b = -0.63, t(92) = -6.498, p < 0.01. March temperature significantly predicted peak bloom as well, b = -1.40, t(92) = -13.428, p < 0.01.

- Using this estimated equation, we see that any increase in either of the temperatures results in the peak bloom date moving up earlier. A higher temperature in March results in a larger effect on the peak bloom date.
- The multiple R squared represents the percent of the variance, and adjusted R squared takes into account the number of samples and variables used. R^2 must land between 0 and 1, with 1 being a perfect fit. Both R squared values sit around 0.73, indicating that the regression function fits fairly well and has a little amount of variance.

c) Plot the relationship between your predicted values and your residuals using a scatterplot (see the lecture outline for R code). Do the residuals seem to vary consistently at all predicted values?



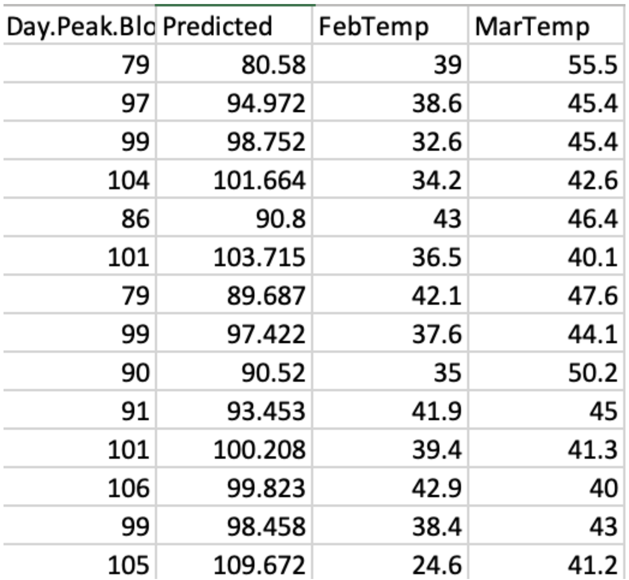
- The residuals seem to vary more in the 90-100 range of predicted values.



- Plotting the residuals on a histogram shows that the distribution is normal.

d) Use your estimated regression equation from *part b*to predict the Peak Bloom Date for 2016. How accurate is the prediction? The actual Peak Bloom Date from 2016 is available in this dataset

- This prediction is fairly accurate. Running the FebTemp and MarTemp into the formula “y = -0.63*FebTemp + -1.40*MarTemp + 182.85” resulted in predictions that were only a few days to a week off of the actual peak bloom day.



Day.Peak.Blo	Predicted	FebTemp	MarTemp
79	80.58	39	55.5
97	94.972	38.6	45.4
99	98.752	32.6	45.4
104	101.664	34.2	42.6
86	90.8	43	46.4
101	103.715	36.5	40.1
79	89.687	42.1	47.6
99	97.422	37.6	44.1
90	90.52	35	50.2
91	93.453	41.9	45
101	100.208	39.4	41.3
106	99.823	42.9	40
99	98.458	38.4	43
105	109.672	24.6	41.2