Robert Bruffey
INST 314
<div align="center">Project 5: Multiple Linear Regression</div>

**Step 1:** Construct a multiple linear regression model including all predictors for snow and temperature for January, February, and March. Identify which if any predictors are significantly related to Peak Bloom Date

```
cherry = read.csv("~/Downloads/CherryData.csv") #load data
c = subset(cherry, cherry$Year < 2016) #subset data to not include 2016
View(c)
```

**Hypothesis**:

$H_0$ = There is no difference in the mean Peak Bloom Dates for the snow and temperatures of January, February, and March.

$H_1$ = There is at least one difference in the means for the Peak Bloom Dates for the snow and temperatures of January, February, and March.

**Test Statistic:**

**RCode:** summary(lm(formula = c$Day.Peak.Bloom ~ c$JanTemp + c$FebTemp + c$MarTemp + c$JanSnow + c$FebSnow + c$MarSnow))

Call:
lm(formula=c$Day.Peak.Bloom ~ c$JanTemp+c$FebTemp+c$MarTemp+c$JanSnow+ c$FebSnow+c$MarSnow)

Residuals:

| Min | 1Q | Median | 3Q | Max |
|---|---|---|---|---|
| -10.8356 | -2.2316 | 0.0441 | 1.9986 | 13.1526 |

Coefficients:

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

---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.843 on 88 degrees of freedom
Multiple R-squared:  0.7352,  Adjusted R-squared:  0.7171
F-statistic: 40.71 on 6 and 88 DF,  p-value: < 2.2e-16

**Model:**
Y = 194.889154 - .046435*JanTemp - 0.615685*FebTemp - 1.419101*MarTemp - 0.039801*JanSnow - .006768*FebSnow - 0.019790*MarSnow

Based on our results above there are two predictors that are significantly related to Peak Bloom Date, FebTemp and MarTemp. We know this because the p-value for FebTemp is 1.09e-05, which is less than our significant level of .05. Also, the p-value for MarTemp is <2e-16, which is less than .05. Therefore, we can reject our null hypothesis and say that FebTemp and MarTemp have different means for the peak bloom date compared to the other temps and snow,

**Step 2:** Rerun the regression model only including significant predictors. Report the results of this model. Which predictors were not significant from Step 1? Which predictors were significant? Describe the relationship between the predictors and the outcome. How much variance is explained by this model? Test the assumptions of this test. Interpret the findings in light of the statistical results, effect size, and limitations

**New Regression Model:**
        **R Code:** summary((lm(Day.Peak.Bloom ~ FebTemp + MarTemp, data = c)))

Call:
lm(formula = Day.Peak.Bloom ~ FebTemp + MarTemp, data = c)

Residuals:
    Min      1Q     Median     3Q      Max
-10.4823  -2.2295   0.1214   1.7676  13.2150

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

**Model:**
Y = 182.85436 - 0.63062*FebTemp - 1.40385*MarTemp

**Analysis:**
      Again, from this data we see that FebTemp and MarTemp are still significant and have a p-value less than .05. From our previous data in step 1 we saw that JanTemp, JanSnow, FebSnow, and MarSnow are all not significant because they had p-values greater than .05. From our model above, the relationship between our significant predictors and the outcome is a negative relationship. As the FebTemp and MarTemp rises, the outcome for Bloom Peak Day decreases. The effect size of this model has an Multiple R-squared value of .7341 and an adjusted R-squared of .7283, which means there is a 73.41% and 72.83% of variance in the DV (Peak Bloom Day) explained by the IV's (FebTemp and MarTemp). One limitation about this data is that we are not sure where the location of the data is taken. If could have been gathered from many different locations over the years, where some areas are colder than others.

```
> pred = d$fitted.values
> resid = d$residuals
> plot(pred, resid)
```
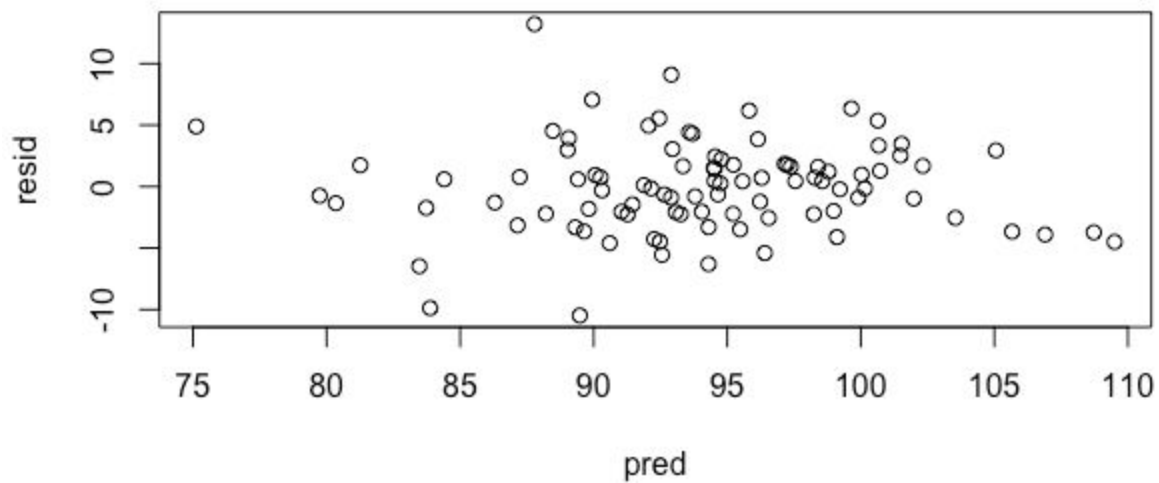
> cov(c$FebTemp+c$MarTemp, c$Day.Peak.Bloom)
[1] -34.72798
    This covariance value shows that the variables are not related too much in the negatvie manner.

> cor.test(c$FebTemp+c$MarTemp, c$Day.Peak.Bloom)

    Pearson's product-moment correlation

data:  c$FebTemp + c$MarTemp and c$Day.Peak.Bloom
t = -13.441, df = 93, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 -0.8713571 -0.7305978
sample estimates:
     cor
-0.8125024
            The correlation value of -.8125 is a strong negative correlation, in that as
        Day Peek Bloom goes down, FebTemp and MarTemp go up.

**Step 3:** Use your model from Step 2 to predict the Peak Bloom Date for 2016. How accurate is the prediction compared to the actual value for 2016?

Model from step 2: $Y = 182.85436 - 0.63062*FebTemp - 1.40385*MarTemp$

FebTemp(2016) = 39.9

MarTemp(2016) = 53.5

$Y = 182.85436 - .63062*39.9 - 1.40385*53.5 = 82.5866$

Actual 2016 Day.Peak.Bloom = 84

Difference = 84-82.5866 = 1.4133

Our predicted value for 2016 is relatively close to the actual value for 2016, in that they are only 1.4133 days apart.