Math 338 **Lab Assignment #26** Fall 2019

For this lab, we will attempt to predict ice cream consumption (IC) from price, income, temperature, and/or year. Download the *icecream.txt* file from Titanium and import it into RStudio. If it's not the default, set the **Delimiter** or **Separator** to "Tab" so that RStudio can correctly parse the file.

Let's start by doing some exploratory data analysis on our data set. Since we only have a few variables, let's make a scatterplot matrix. The command to create this is found in the GGally package, which you'll have to install. Once you install the package, run the code:

```
> library(GGally)
> ggpairs(icecream, lower = list(continuous = "smooth"), upper =
list(continuous = "smooth"))
```

If you're having trouble installing GGally, you can get a barebones scatterplot matrix using the command:

```
> pairs(icecream)
```

Question #1 Copy and scatterplot matrix below. The top row contains scatterplots of IC (response) vs. one of the four explanatory variables. Looking across the top row, do you see any issues with nonlinearity, or any outliers? If so, which variables have the problems?

In temperature, price, and income there is one outlier for each. Please see first graph attached.

Question #2 Looking at the other rows of the scatterplot matrix, which pair(s) of explanatory variables appear to be linearly related? Why might we have an issue if both variables in the pair are included in the model?

Income and temperature appear to be linearly related. They might shift in sync with each other, where we want one to change and the other not to.

Now let's create our multiple regression model. Let's start by just unthinkingly throwing all our explanatory variables in a linear model. This is usually a very bad idea – typically we use a combination of real-world sense and exploratory data analysis to do feature engineering, but we don't have time for real feature engineering in today's lab):

```
> lm_all <- lm(IC ~ price + income + temp + Year, data = icecream)
> summary(lm_all)
```

Question #3 Insert the entire output of the `summary()` command below. Use the parameter estimates given in the output to write out the fitted regression equation.

```
Call:
lm(formula = IC ~ price + income + temp + Year, data = icecream)

Residuals:
     Min       1Q   Median       3Q      Max
-0.052572 -0.016208 -0.003463  0.010333  0.075354

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.7140368  0.2887794   2.473  0.02057 *
price       -1.2873605  0.7292161  -1.765  0.08971 .
income      -0.0023718  0.0021099  -1.124  0.27165
temp         0.0031511  0.0003998   7.882 3.08e-08 ***
Year         0.0508220  0.0165386   3.073  0.00506 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.032 on 25 degrees of freedom
Multiple R-squared:  0.796,  Adjusted R-squared:  0.7634
F-statistic: 24.39 on 4 and 25 DF,  p-value: 2.564e-08
```

Question #4 Interpret the meaning of the slope estimate for *price* in context of the problem.

-1.2873605 is the slope for the variable price. This means it has negative correlation.

Question #5 What is the (null) hypothesis for a Multiple Linear Regression ANOVA test using this data?

Population mean is the same and is unaffected by anything.

Question #6 What test statistic do you obtain for an ANOVA test, what is its distribution under your hypothesis in **Question #5** (include the degrees of freedom), what is its observed value for this test, and what is the corresponding p-value? [Hint: you can find the relevant values in the very bottom line of the summary output]

We obtain a f-statistic which means we have a f-distribution. Our observed value is 24.39, and we have a corresponding p-value of 2.564e-08.

Question #7 Looking at the `Coefficients` table, which explanatory variable is the least significant predictor?

Our least significant predictor is going to be income and price.

Now we are going to **update** our model to remove that predictor from the model. For example, if *price* is the least significant predictor, we would run:

```
> lm2 <- update(lm_all, .~. -price)  # don't actually run this - replace price with the actual variable to be removed
> summary(lm2)
```

Then, if we still had non-significant predictors in **lm2** and *Year* was the least significant predictor, we would run:

```
> lm3 <- update(lm2, .~. -Year)  # don't actually run this - replace Year with the actual variable to be removed
> summary(lm3)
```

Continue to update the formula and check the new model until all remaining explanatory variables are significant at the 5% significance level. This is about the worst way to do backward selection, but better ways of doing it are beyond the scope of the class.

Question #8 Copy and paste the **Coefficients:** table for each new model you created.

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 0.1608478  0.0220541   7.293 7.61e-08 ***
temp        0.0033925  0.0003796   8.936 1.50e-09 ***
Year        0.0355695  0.0082154   4.330 0.000184 ***
```

Question #9 What is the regression equation for your final model?

IC = 0.1608478 + 0.0033925*temp + 0.0355695*Year

Question #10 Using your final model, predict the icecream consumption ($\widehat{IC}$) when price is at 0.280, income is at 85, temperature is at 68, and Year is 1. Ignore the values of any variables that are not in your final model.

0.427109

Optionally, check your answer using the following code:

```
> new_data_frame <- data.frame(price = 0.280, income = 85, temp = 68, Year = 1)  # new values of all predictor variables
> predict(lm_final, newdata = new_data_frame)  # replace lm_final with the variable name of your final model
```