Applied LM Homework #1

Blain Morin

February 16, 2018

Question 1: Wine

The data in the file wine.csv (in the datasets folder on Canvas) give the average wine consumption rates (in liters per person) and number of ischemic heart attack deaths (per 1000 men aged 55 to 64 years) for 18 industrialized countries.

Analyze the data and write a brief report that includes a summary of findings, a graphical display and a section describing the methods used to answer the questions of interest.

First, lets explore the wine data.

```
dim(Wine)
```

```
## [1] 18  3
```

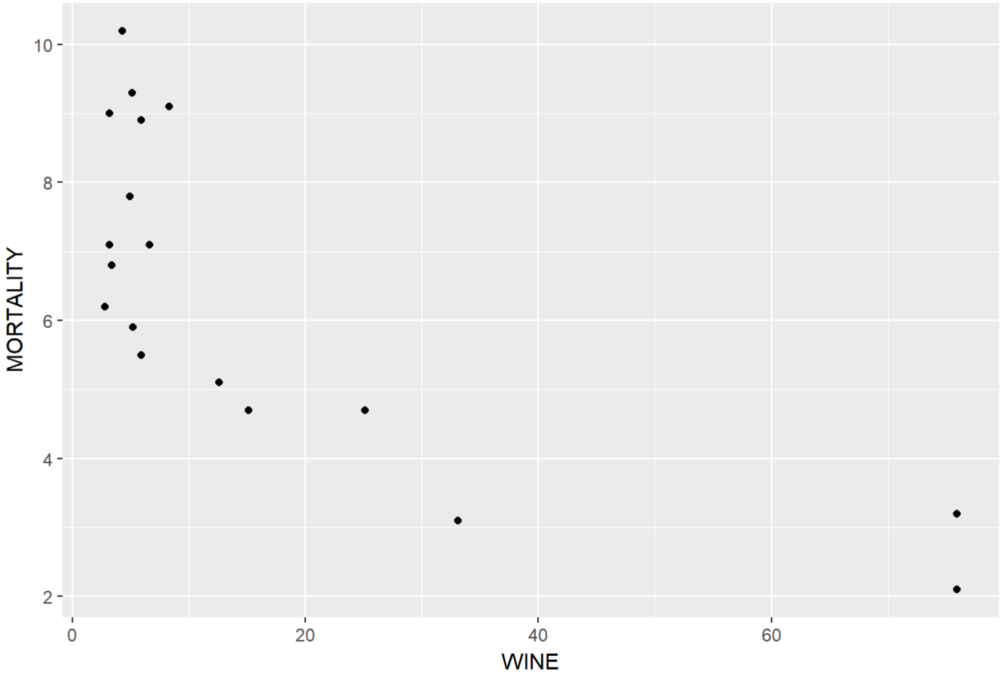
```
summary(Wine)
```

```
##    COUNTRY               WINE         MORTALITY
##  Length:18          Min.   : 2.80   Min.   : 2.100
##  Class :character   1st Qu.: 4.45   1st Qu.: 4.800
##  Mode  :character   Median : 5.90   Median : 6.500
##                     Mean   :16.47   Mean   : 6.433
##                     3rd Qu.:14.47   3rd Qu.: 8.625
##                     Max.   :75.90   Max.   :10.200
```

We see that the data set contains wine consumption and heart disease mortality rates for 18 different countries. The min, max, median, mean, and quartiles are shown in the above table.

Next, let’s visualize the data using a scatter plot:

Figure 1



Do these data suggest that heart disease death rates are associated with average wine consumption? If so, how can that be described?

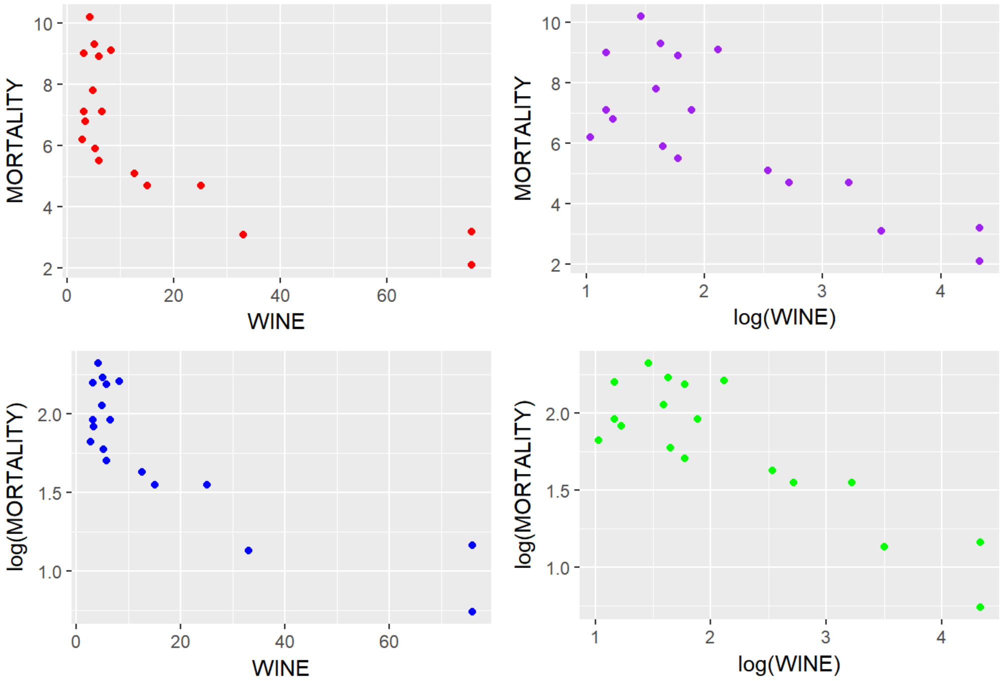
We can see from Figure 1 that there seems to be a negative correlation between wine consumption and heart disease death rates. Computing the correlation confirms the inverse relationship:

```
cor(WINE, MORTALITY)
```

```
## [1] -0.7455682
```

Although there is a strong negative correlation, the relationship does not seem to be linear. We can try to make a stronger linear relationship by using log transformations:

Figure 2

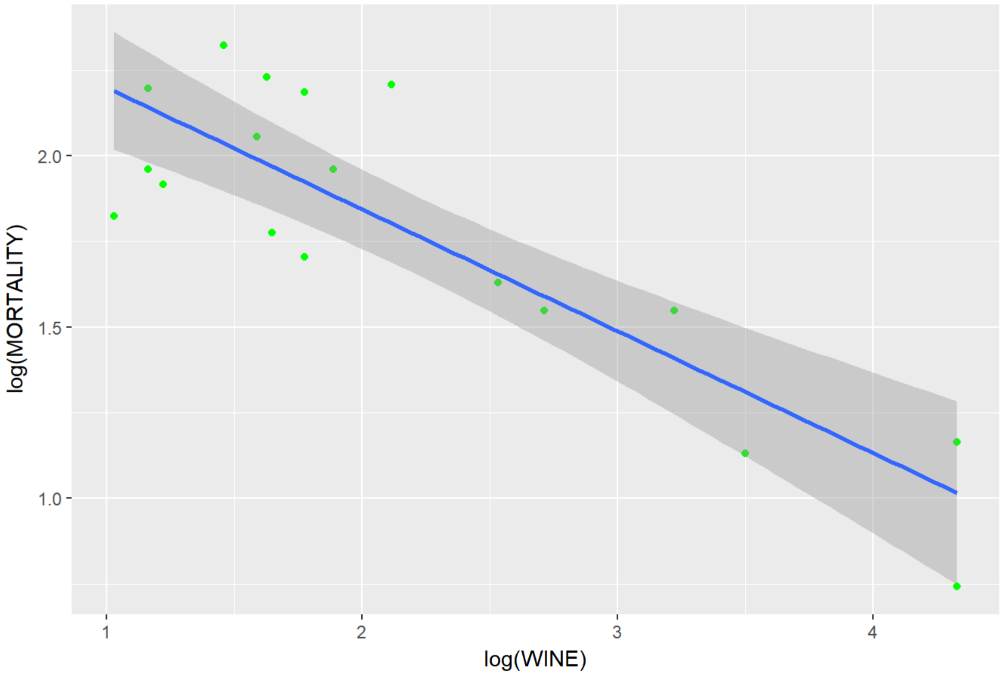


Visually, it seems that using log wine consumption and using both log wine and log mortality have the best linear appearance. Next, lets run a linear model on the set of transformed data.

	MORTALITY			MORTALITY			log(MORTALITY)			log(MORTALITY)		
	B	*CI*	*p*	*B*	*CI*	*p*	*B*	*CI*	*p*	*B*	*CI*	*p*
(Intercept)	7.69	6.68 – 8.69	<.001	10.28	8.52 – 12.04	<.001	2.05	1.90 – 2.19	<.001	2.56	2.29 – 2.82	<.001
WINE	-0.08	-0.11 – -0.04	<.001				-0.02	-0.02 – -0.01	<.001			
log(WINE)				-1.77	-2.51 – -1.04	<.001				-0.36	-0.47 – -0.24	<.001
Observations	18			18			18			18		
R^2 / adj. R^2	.556 / .528			.620 / .596			.718 / .700			.738 / .722		

We see from the table that the model with the best linear fit (from the R squared and adjusted R squared) is from the model that uses log transforms for both wine consumption and mortality. Here is the linear model plotted on the scatter plot:

Figure 3



Do any countries have substantially higher or lower death rates than others with similar wine consumption rates?

We see from Figure 2 that there seems to be significant variation in log mortality rates among countries with similar log wine consumption rates, especially when log wine consumption is between 1 and 2. On the figure above, the gray bar represents the 95% confidence interval. We can see that some points lie above the gray bar and some below the bar. Since the residuals seem to be higher at the low end of log wine consumption, they may be heteroskedastic. This would violate our linear model assumptions and should be further explored.

Question 2: Flowers

Meadowfoam is a small plant that grows in Pacific Northwest and is domesticated for its seed oil. A study was set up to determine if meadowfoam can be made into a profitable crop. In a controlled growth chamber, the plant was grown at 6 different light intensities and two different timings of onset of light treatment. The outcome of interest is the number of flowers per plant which was measured by averaging numbers of flowers produced by 10 seedlings in each group. Growth was replicated at each combination of time and light intensity.

a. First put the data into a dataset with four variables: number of flowers, light intenisty, timing and replicate.

```
flowers <- read.csv('flowers.csv')
flowers[,4] <- rep(c(1,2))
names(flowers)[4] <- "REPLICATE"
```

b. Create a categorical form of the light intensity with 6 categories.

```
flowers = flowers %>%
  mutate(CAT_INTENS = as.factor(INTENS))

flowers$TIME = as.factor(flowers$TIME)
```

c. The research questions are: What are effects of intensity and timing? Is there an interaction between the two factors?

First, lets run 4 linear model:

Model 1 is number of flowers regressed on categorical intensity and time.

Model 2 is number of flowers regressed on categorical intensity, time, and their interaction.

Model 3 is number flowers regressed on continuous intensity and time.

Model 4 is number of flowers regressed on continuous intensity, time, and their interaction.

Their model summaries are:

	FLOWERS			FLOWERS			FLOWERS			FLOWERS		
	B	*CI*	*p*	*B*	*CI*	*p*	*B*	*CI*	*p*	*B*	*CI*	*p*
(Intercept)	67.20	59.54 – 74.85	<.001	69.85	58.46 – 81.24	<.001	71.31	64.50 – 78.11	<.001	71.62	62.56 – 80.68	<.0
TIME2	12.16	6.37 – 17.95	<.001	6.85	-9.26 – 22.96	.372	12.16	6.69 – 17.63	<.001	11.52	-1.29 – 24.34	.07
CAT_INTENS												
CAT_INTENS300	-9.13	-19.15 – 0.90	.072	-15.10	-31.21 – 1.01	.064						
CAT_INTENS450	-13.38	-23.40 – -3.35	.012	-14.10	-30.21 – 2.01	.081						
CAT_INTENS600	-23.23	-33.25 – -13.20	<.001	-27.30	-43.41 – -11.19	.003						
CAT_INTENS750	-27.75	-37.77 – -17.73	<.001	-31.75	-47.86 – -15.64	.001						
CAT_INTENS900	-29.35	-39.37 – -19.33	<.001	-30.50	-46.61 – -14.39	.001						
TIME2:CAT_INTENS300				11.95	-10.83 – 34.73	.275						
TIME2:CAT_INTENS450				1.45	-21.33 – 24.23	.892						
TIME2:CAT_INTENS600				8.15	-14.63 – 30.93	.451						
TIME2:CAT_INTENS750				8.00	-14.78 – 30.78	.459						
TIME2:CAT_INTENS900				2.30	-20.48 – 25.08	.830						
INTENS							-0.04	-0.05 – -0.03	<.001	-0.04	-0.06 – -0.03	<.0
TIME2:INTENS										0.00	-0.02 – 0.02	.91
Observations	24			24			24			24		
R^2 / adj. R^2	.823 / .761			.849 / .710			.799 / .780			.799 / .769		

d. First create an analysis of variance using timing and the categorical form of the light intensity variable. Determine if there is an effect of each factor.

```
##             Df Sum Sq Mean Sq F value   Pr(>F)    
## TIME         1  887.0   887.0   19.65 0.000365 ***
## CAT_INTENS   5 2683.5   536.7   11.89 4.63e-05 ***
## Residuals   17  767.5    45.1                     
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

From the above anova table, there is evidence to suggest that both time and categorical intensity have an effect on the number of flowers. The large f-statistics and small p values mean that there is a difference between the mean number of flowers between early and late and also that at least on of the light intensity categories has a significant difference in the number of flowers.

e. Then create an interaction between light intensity and timing by multiplying the two variables and test for the presence of an interaction.

```
##                 Df Sum Sq Mean Sq F value   Pr(>F)    
## TIME             1  887.0   887.0  16.227 0.001675 ** 
## CAT_INTENS       5 2683.5   536.7   9.819 0.000639 ***
## TIME:CAT_INTENS  5  111.5    22.3   0.408 0.834157    
## Residuals       12  655.9    54.7                     
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

From the above anova table, there is not evidence to suggest that there is a significant interaction between time and categorical light intensity.

f. Now repeat the process but using light intensity as a continuous variable.

The ANOVA table for time and intensity as a continuous variable is as follows:

```
##             Df Sum Sq Mean Sq F value   Pr(>F)    
## TIME         1  887.0   887.0   21.38 0.000146 ***
## INTENS       1 2579.8  2579.8   62.18 1.04e-07 ***
## Residuals   21  871.2    41.5                     
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The ANOVA table for time, intensity as a continuous variable, and their interaction is as follows:

```
##             Df Sum Sq Mean Sq F value   Pr(>F)    
## TIME         1  887.0   887.0  20.374 0.000212 ***
## INTENS       1 2579.8  2579.8  59.260  2.1e-07 ***
## TIME:INTENS  1    0.6     0.6   0.013 0.909567    
## Residuals   20  870.7    43.5                     
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Similar to the model where light intensity was coded as categorical, we see that the data does not provide evidence of an interaction between time and intensity. Like the categorical models, continuous light intensity and time have a statisticall significant effect of the number of flowers.

g. Then perform F-tests to compare the four model you have created (light as continuous and categorical with and without the interaction)

Here is the summary of an f-test with light as a categorical variable, between the model without the interaction term and the model with the interaction term:

	Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>
1	17	767.4721	NA	NA	NA	NA
2	12	655.9250	5	111.5471	0.4081457	0.8341569

2 rows

Here is the summary of an f-test with light as a continuous variable, between the model without the interaction term and the model with the interaction term:

	Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>
1	21	871.2358	NA	NA	NA	NA
2	20	870.6598	1	0.5760357	0.01323217	0.9095675

2 rows

Both ANOVA tables suggest that there is not enough evidence to suggest that we should include an interaction term. We can see that the p-value for the f-statistic for the comparison of the models is the same as the p-value for their corresponding interaction term.

h. Predict the number of flowers grown at each combination of light and timing for each of the four models.

Here is the code we use to set up the predictions and calculate the mean squared errors:

```
m1 = lm(FLOWERS ~ TIME + CAT_INTENS, data = flowers)
p1 = predict(m1, se.fit = TRUE, interval = "conf")
resi_m1 = p1$fit[,1] - flowers$FLOWERS
mse_m1 = sum(resi_m1^2) / p1$df

m2 = lm(FLOWERS ~ TIME + CAT_INTENS + TIME*CAT_INTENS, data = flowers)
p2 = predict(m2, se.fit=TRUE, interval = "conf")
resi_m2 = p2$fit[,1] - flowers$FLOWERS
mse_m2 = sum(resi_m2^2) / p2$df

m3 = lm(FLOWERS ~ TIME + INTENS, data = flowers)
p3 = predict(m3, se.fit = TRUE, interval = "conf")
resi_m3 = p3$fit[,1] - flowers$FLOWERS
mse_m3 = sum(resi_m3^2) / p3$df


m4 = lm(FLOWERS ~ TIME + INTENS + TIME*INTENS, data = flowers)
p4 = predict(m4, se.fit = TRUE, interval = "conf")
resi_m4 = p4$fit[,1] - flowers$FLOWERS
mse_m4 = sum(resi_m4^2) / p4$df
```

Predictions for each of the models are shown in the table below:

Time	Intesity	m1	m2	m3	m4
1	150	67.19583	69.85	65.23512	65.46190
1	150	67.19583	69.85	65.23512	65.46190
1	300	58.07083	54.75	59.16440	59.30048
1	300	58.07083	54.75	59.16440	59.30048
1	450	53.82083	55.75	53.09369	53.13905
1	450	53.82083	55.75	53.09369	53.13905
1	600	43.97083	42.55	47.02298	46.97762
1	600	43.97083	42.55	47.02298	46.97762
1	750	39.44583	38.10	40.95226	40.81619
1	750	39.44583	38.10	40.95226	40.81619
1	900	37.84583	39.35	34.88155	34.65476
1	900	37.84583	39.35	34.88155	34.65476
2	150	79.35417	76.70	77.39345	77.16667
2	150	79.35417	76.70	77.39345	77.16667
2	300	70.22917	73.55	71.32274	71.18667
2	300	70.22917	73.55	71.32274	71.18667
2	450	65.97917	64.05	65.25202	65.20667
2	450	65.97917	64.05	65.25202	65.20667
2	600	56.12917	57.55	59.18131	59.22667
2	600	56.12917	57.55	59.18131	59.22667
2	750	51.60417	52.95	53.11060	53.24667
2	750	51.60417	52.95	53.11060	53.24667
2	900	50.00417	48.50	47.03988	47.26667

Time	Intesity	m1	m2	m3	m4
2	900	50.00417	48.50	47.03988	47.26667

i. Compare each prediction to the observed number of flowers and calculate the difference (observed – predicted). This is the residual. Calculate the residual mean squared error for each model by adding the squared residuals together and dividing by the number of residual degrees of freedom. This should equal the mean squared error in each ANOVA table.

Our MSE calculated in part h are displayed in the table below. Each of their values matches the residual MSE from their corresponding anova tables.

mse_m1 <dbl>	mse_m2 <dbl>	mse_m3 <dbl>	mse_m4 <dbl>
45.14542	54.66042	41.48742	43.53299

1 row

j. Now plot the residuals vs. the predicted for each model and see if there are any patterns. If you see any, what might you do to remove them?

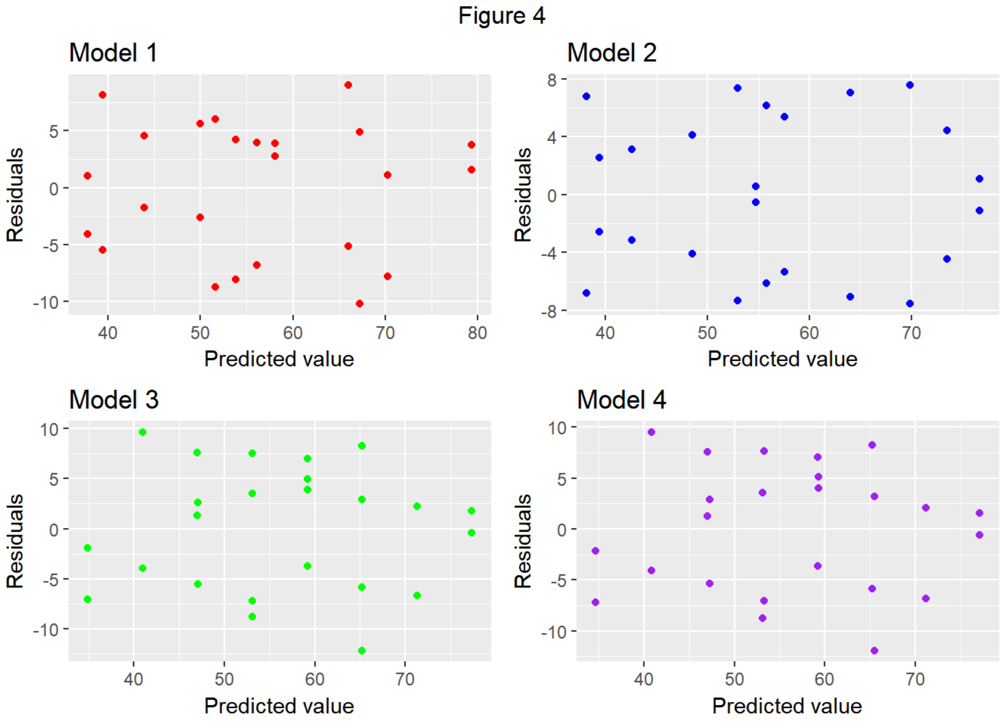


Figure 3 shows the residuals vs the predicted values for each of the models. For models 1 and 2, there seems to be some symmetry around the zero line. This is due to the categorical variable. The symmetry can be fixed by changing intensity to a continuous variable, as seen in model 3 and 4. However, the symmetry does not necessarily violate the assumptions of the linear model because the residuals maintain homoskedasticity.

k. Finally, take the model you think describes the data the best and write a short report for your grandmother who would like to grow these flowers carefully explaining to her how she should best grow them and why. Note that your grandmother is curious about how much changes in light and timing might affect her flowers and how sensitive her results will be to the settings she makes.

I think model 1 best describes the data. Although it does not have the highest adjusted R^2, I think it provides the most insight. If I were to explain it to Grandma I would say:

Dear Grandma,

Both light intensity and the timing of planting play a role in the number of flowers your plants will have. You should plant them late. Late planting results in an average increase of 12 flowers per plant. Moreover, your plants will do better in low intensity light. You should be careful to keep the light intensity under 600 lumens. Plants receiving lights at the 600 lumen level average 23 less plants than a plant receiving 150 lumens.

Love, Your grandson