

Chapter 2

Making Connections: Extending the Statistical Model to More Than One Explanatory Variable

In the previous Making Connections lab the t-test, ANOVA, and regression provided identical results when comparing two types of paper towels. There was one factor, brand, that was used to explain the differences in paper towel strength. However, modeling real-world phenomena often requires more than just one factor to explain changes in the response. Clearly the wetness of the paper towel will also impact strength.

There are many possible ways to conduct an experiment to determine the effects of brand and/or amount of water on strength. One simple plan, as shown in the first lab, would be to control the effect of water by applying the same amount, 15 drops, to all towels. However, it is rather unlikely in day-to-day activities that exactly 15 drops of water will be absorbed by a paper towel as it is used.

Another possible study, which will be discussed in this lab, would be to let the amount of water be considered as a second factor in the experiment. This lab will extend the paper towel study to two explanatory variables. In this lab we will simultaneously test for the effect of three levels of water and two brands of paper towels.

The t-test is limited to data sets with 1 factor and 2 levels¹, however ANOVA and regression will be used to analyze this experiment with multiple factors and two or more levels of each factor. Section 2.1 discusses hypothesis tests, Section 2.2 develops a statistical model, and Section 2.3 describes the model assumptions. Using ANOVA to analyze data with categorical explanatory variables will be discussed in Section 2.4 and regression in Section 2.5.

¹Multiple comparison t-tests can be conducted, however, the level of confidence in correctly rejecting the null hypothesis is dramatically reduced

2.1 Revisiting the Paper Towel study

The first Making Connections lab only discussed a part of the paper towel study. The strength of Comfort (Brand C) and Decorator (Brand D) paper towels were actually compared at varying levels of water; 0 drops, 5, drops and 15 drops. Thus the design consisted of the following:

Factors : brand and amount of water

Levels :

- Brand C and 0 drops of water
- Brand C and 5 drops of water
- Brand C and 15 drops of water
- Brand D and 0 drops of water
- Brand D and 5 drops of water
- Brand D and 15 drops of water

Response variable: Breaking Strength in grams. Breaking strength is defined as the last weight that the paper towel successfully held before the sheet broke. Small weights were added to the center of a paper towel until it broke.

Experimental Units: 26 sheets were tested at each of the 6 levels.

In this design there are six conditions which are called **factor-level combinations** or **factorial combinations**. The 2-sample t-test suggested in the Making Connections lab has only two meaningful groups: the Brand C group and the Brand D group. This two factor design has many more meaningful groups: the Brand C group, the Brand D group, the 0 drops of water group, the 5 drops of water group, the 15 drops of water group, as well as each of the six factorial combinations.

Since more factors are included in this experiment, there are also more hypotheses to be tested. The three null hypotheses corresponding to this two factor design are:

1. H_{01} : there is no difference in strength between the two brands of paper towel vs.
 H_{a1} : the two brand means are different
2. H_{02} : there is no difference in strength between the 0 drops, 5 drops, and 15 drops of water amounts vs. H_{a2} : at least one water amount group is different
3. H_{03} : brand has no influence on how the water amount affects strength, or equivalently,
 H_{03} : the amount of water has no influence on how brand affect strength vs.
 H_{a3} : amount of water impacts how brand affects strength

2.2 Developing a Statistical Model

Table 2.1 represents some of the data for the paper towel study. Each of the 6 cells has 26 observations. The complete data set is in the file `PaperTowelData2`. While not all observations are shown, Table 1.1 helps us view the data structure. Notice that in this

Breaking Strength of Paper Towels (grams)		Amount of Water		
		0 drops	5 drops	15 drops
Brand of Towel	Brand C	3200	2000	375
		3400	1800	475
		2800	1700	500
		⋮	⋮	⋮
	Brand D	3100	1800	325
		2400	875	400
2400		600	450	
	2000	825	325	
	⋮	⋮	⋮	
	1700	700	300	

Table 2.1: Strength of Paper Towels (grams of weight added before the towel broke)

experiment the two factors are **crossed**: every possible row and column combination is included in the study. Crossing has a visual representation in Table 2.1: every row extends all the way from the left to the right and crosses each of the columns; every column goes from the top to bottom and crosses all the rows.

As the data structure becomes more complex, a statistical model becomes more useful in appropriately defining the population. Figure 2.1 is useful in visualizing the meaningful groups within this data set: the overall average, the two brand groups, the amount of water groups, and the six level-combination groups. The diagram labels (i.e. the grand mean, effects, and residuals) are described below and used to develop a model.

Extensions Figure 2.1: 2-way Factorial Diagram.[SEE LAST PAGE]

The symbols in Figure 2.1 represent:

- y_{ijk} k^{th} observed breaking strength ($k = 1, 2, \dots, 26$) for brand i and water level j
- μ overall mean breaking strength of the entire population of paper towels
- α_i brand effect ($i = 1, 2$)
- β_j amount of water effect ($j = 1, 2, 3$)
- $(\alpha\beta)_{ij}$ interaction effect
- ϵ_{ijk} residual error (difference between the observed and expected values)

Notice that each of the $2 \times 3 \times 26 = 156$ observed strength measurements represent one of the 156 equations in this diagram. The structure of the data is now used to create a statistical model.

$$y_{ijk} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \epsilon_{ijk} \quad i = 1, 2 \quad j = 1, 2, 3 \text{ and } k = 1, 2, \dots, 26 \quad (2.2.1)$$

If it would be appropriate to assume that water was a quantitative variable, it may be appropriate to test if there is a linear relationship between water and breaking strength. An alternative form of the above statistical model would be:

$$y_m = \beta_0 + \beta_1 x_{1m} + \beta_2 x_{2m} + \beta_{12} x_{1m} x_{2m} + \epsilon_m \quad m = 1, 2, \dots, 156 \quad (2.2.2)$$

Where x_1 represents the brand factor (1 is Brand C, 0 represents Brand D) and x_2 represent the appropriate water level (0, 5, or 15)². While there are many similarities between these two models, take a minute to understand that β_1 has very different meanings depending on the model that is used. In equation 2.2.1, β_1 represents the effect, or the change in the expected value, from the overall mean to the mean of the 0 water group. In equation 2.2.2, β_1 represents the slope, or the change in the expected value when moving from the the Brand D group to the Brand C group.

Another alternative to writing the statistical model for this data structure is to just leave the subscripts assumed:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_{12} x_1 x_2 + \epsilon \quad (2.2.3)$$

[KEY CONCEPT] Identifying the meaningful groups within each data set (the data structure) is the first step in developing a statistical model

[On Your Own:]

1. What numbers in Table 2.1 are represented by y_{213} and y_{122} ?

²If x_2 is treated as a quantitative value the test statistics and p-values will not be identical to treating x_2 as a categorical value.

2. Give the proper algebraic notation for the observation representing the 3rd paper towel with Brand D and 15 drops of water.
3. Why doesn't μ have any subscripts?
4. Draw a diagram similar to Figure 2.1 for a 2-way factorial design with 4 levels of the first factor, 3 levels of the second factor and two observations per level combination.
5. Identify the deterministic model and the random error terms in equation (2.2.1).

2.3 Exploratory Data Analysis and Model Assumptions

Before any formal analysis is conducted, it is often beneficial to visually explore the data.

[On Your Own:]

6. Draw dot plots and side by side boxplots of the 6 factor combinations. [In Minitab select **Graph** \rightarrow **Dotplots** \rightarrow **With Groups**. Click OK. Select **Strength** for Graph variables **Water** and **Brand** for Categorical variables. Click OK. Do the same for boxplots.]. Recall that in a boxplot, the 3 horizontal lines on each box represent the 25th, 50th and 75th percentiles of the data. The whiskers of each box represent the most extreme values (maximum and minimum that fall within the expected data range).
7. Identify any outliers. Any asterisks beyond the “fences”, or expected data range, may be called outliers. Unless there is convincing evidence that the outliers represent incorrect data, do NOT automatically delete all outliers. In fact, in some situations the outliers may be the most interesting aspect of the data. It is often reasonable to analyze the data twice; once with the outlier and once without. If the results are similar with or without the outlier, there is no problem. If the results vary, it is important to be cautious about interpreting any results.
8. Describe the data. Without doing any statistical calculations, do you expect to reject the three null hypotheses in Section 2.1? Justify your answer by visually comparing the variation (i.e. spread) between groups to the variation within groups.

In each form of the model in Section 2.2, it is assumed that the parameters (the α 's and the β 's) are constant. In addition several assumptions need to be validated about the error term, ϵ , before a hypothesis test can be developed:

- the error terms are independent and identically distributed
- the population variances are equal ($\sigma_{11}^2 = \sigma_{12}^2 = \sigma_{13}^2 = \sigma_{21}^2 = \sigma_{22}^2 = \sigma_{23}^2 = \sigma^2$) and

- the error terms follow a normal probability distribution: denoted as $\epsilon \sim N(0, \sigma^2)$.

[On Your Own:]

9. The **independence** assumption implies that there is no relationship between one observation and the next. The **identically distributed** assumption means that each observation sampled within each brand/water combination is from a “population” with the same mean and variance. If all 26 paper towels were sampled from one roll to assess the Brand C, 5 Drops factor combination, would you be concerned about violating the independence and/or the identically distributed assumptions? Why or why not?
10. Calculate the sample standard deviations of all six factor combinations. If the largest sample standard deviation is more than twice the size of the smallest sample standard deviation, the equal variance assumption should be questioned. Clearly the variation within each group increases as the means increase. To address this issue, a transformation of the data can often be conducted that will allow us to assume equal variances. Create a new variable `LNStrength` and again test for equal variances. [In Minitab select `Calc` → `Calculator`. Store result in variable `LNStrength`. In Expression: type `LOGE(Strength)` [or select `Natural log` in Functions]. Click `OK`. Select `Stat` → `Basic Statistics` → `Display Descriptive Statistics`. Select `LNStrength` in Variables and `Water` and `Brand` in By variables. Click `Statistics` and select only the mean and Standard deviation.] What is the ratio of the largest and smallest sample standard deviations? Also calculate the square root of `Strength`, by calculating `SQRT(Strength)`. Then compare the sample standard deviations of `SQRTStrength`. What can we conclude about the transformed data?

To fit the model assumptions, for the rest of this lab, the transformed response variable `SQRTStrength` should be used instead of `Strength`.

The only model assumption left to check is that the random errors follow a normal distribution. Since we have a relatively large sample size in each group, it is enough to check for extreme skewness or outliers. In addition, empirical tests have shown that ANOVA is reliable even if the normality assumption is violated.

2.4 ANOVA to Compare Population Means

Analysis of variance (ANOVA) is a statistical technique used to investigate and model the relationship between a response variable and one or more factors, where each factor consists of two or more levels. One specific type of ANOVA, **factorial designs**, are typically the most efficient method of collecting and analyzing data with multiple factors where each factor has two or more levels.

A **full factorial design** is simply a design where all possible combinations of the conditions are tested. This is still a completely randomized design, the only change is that there are now more conditions. Just as it is in a 2-sample t-test, it is also preferable (but not required) in factorial designs to have a **balanced design** which has the same number of experimental units assigned to each condition.

Factorial designs are efficient because much more information can be calculated, such as means and p-values for multiple hypothesis tests, without requiring more experimental units than the typical two-sample t-test. Factorial designs are very beneficial in situations where experimental units are expensive or difficult to obtain. The next sections will discuss how to organize and draw conclusions for each of the above hypotheses in a full factorial design with two factors, also called a **2-way factorial design**. In a 2-way factorial design, it is often beneficial to display the results in a table where each cell represents a specific treatment combination.

2.4.1 Calculating Effects

After the data are collected, the averages for all meaningful groups of the data can be calculated. These groups correspond to the diagram in Figure 2.1. The **grand mean** (often written $\bar{y}_{...}$) in this example is 35.39, which is the overall average of all 156 observations.

SQRT(Strength)		Amount of Water			Brand Avg
		0 drops	5 drops	15 drops	
Brand of towel	Brand C	56.56			39.27
	Brand D	47.02	26.49	21.03	31.51
Water Avg		51.79		20.49	35.39

Table 2.2: Average SQRTStrength of Paper Towels (grams)

The following mathematical notation is often used to represent the calculated sample averages:

$\bar{y}_{1..}$ = 39.27 the average breaking strength of the Brand C.

$\bar{y}_{2..}$ = 31.51 the average breaking strength of Brand D.

$\bar{y}_{.3}$ = 20.49 the average breaking strength of the 15 drops of water group.

$\bar{y}_{22.}$ = 26.49 the average breaking strength corresponding to the 5 drops of water and Brand D group.

[On Your Own:]

11. Finish calculating the averages in Table 2.2.
12. What are the values of $\bar{y}_{.2}$ and \bar{y}_{21} ?

Table 2.2 reveals that the **SQRTStrength** average changes from 39.27 to 31.51 when changing brands. This difference is smaller than the changes due to the water amount factor. Main effects are calculated to measure the impact of changing the levels of each factor in the model. A **main effect** is the difference between the factor level average and the grand mean.

The effect of Brand C is: Brand C mean – grand mean = $\bar{y}_{.1} - \bar{y}_{...} = 39.27 - 35.39 = 3.88$

The effect of 0 Drops of Water is: $\bar{y}_{.1} - \bar{y}_{...} = 51.79 - 35.39 = 16.40$

[On Your Own:]

13. Calculate the effects of Brand D with 5 Drops of Water, and of Brand D with 15 Drops of Water. Explain any symmetry that you find.
14. Visualization of the main effects is done through a **main effects plot**. This graph simply plots the averages of each factor level. [In Minitab select **Stat** → **ANOVA** → **Main Effects Plot**.] To properly visualize the effect size of each factor, it is essential to keep the vertical axis the same for all factors. Does changing the water level factor or changing the brand factor impact the results more?

The next section will discuss the significance (find the p-value) of each factor based on these effect calculations. The p-values will determine if brand (or water) effect sizes are large enough to be confident that brands (or water amounts) truly impact paper towel strength.

In addition to determining the main effects for each factor, it is often critical to identify how multiple factors interact in affecting the results. An **interaction** occurs when one factor affects the results differently depending on a second factor. To calculate the effect of the brand and water interaction, take the average for a particular factor combination minus the grand mean and the corresponding main effects.

$$\begin{aligned} &\text{The interaction effect of Brand C and 15 Drops of Water} \\ &= \text{Average of the Brand C, 15 Drops group} \\ &\quad - (\text{effect of Brand C} + \text{effect of 15 Drops of Water} + \text{the grand mean}) \\ &= 19.95 - (3.88 + -14.9 + 35.39) \\ &= -4.42 \end{aligned}$$

[**KEY CONCEPT**] Effect sizes determine which factors have the most significant impact on the results. All effects are calculated by subtracting the partial fit from the appropriate

average. **Partial fit** is the effect of all the influencing factors. The influencing factor for main effects is the grand mean. The influencing factors for interaction effects are the grand mean and the appropriate main effects.

[On Your Own:]

15. Calculate the other 5 interaction effects. Hand draw Figure 2.1 and fill out the effect sizes of Figure 1 with actual values (replace μ , α_i , β_j , and $\alpha_i\beta_j$, with actual values). Explain any symmetry that you find.
16. Create an interaction plot. [In Minitab select **Stat** → **ANOVA** → **Interactions Plot**.] **Interaction plots** use level averages to visualize the effect size of interactions. The non-parallel lines show an interaction; the effect of brand changes depending on the amount of water. When interaction effects are 0, the lines are parallel. More perpendicular lines represent stronger interaction effects. This interaction plot shows that the effect of brand depends on the amount of water. Comfort appears stronger with 5 Drops of Water but with 15 Drops of Water Decorator is slightly better.
17. **Residuals** are the difference between the actual observation and the expected value (i.e. observed minus expected). In this 2-way factorial design, the expected value is the sum of the grand mean, the two main effects, and the interaction effect; this is equivalent to the appropriate factor level combination. Calculate the residual values for y_{213} and y_{122} .
18. Show that $\bar{y}_{ij} - \bar{y}_{i..} - \bar{y}_{.j.} + \bar{y}_{...}$ is equivalent to the ij^{th} interaction effect.

Notice that the residual values sum to zero within each group of interest. This will always be true whenever calculating effects. This is not surprising since effects measure the unit deviation of an observed value from the expected value.

2.4.2 Calculations for ANOVA

Main effect and interaction plots help visualize the impact of each factor combination and identify which factors are most influential. However, a statistical hypothesis test is needed in order to determine if any of these effects are significant. ANOVA consists of simultaneous hypothesis tests to determine if any of the effects are significant. ANOVA tests the null hypothesis that the population means of each level are equal, versus them not all being equal. The statement “factor effects are zero” is equivalent to “the means are equal for all levels of a factor”.

Sum of Squares (SS) for brand is identical in the one factor and multi-factor ANOVA. SS for brand is found by summing over all squared brand effects. This is mathematically

written by:

$$SS_{brand} = \sum_{i=1}^2 78 * (\bar{y}_{i..} - \bar{y}_{...})^2 \quad (2.4.1)$$

This equation can be generalized. Instead of 78 elements in each brand group we can substitute n_i elements. Instead of 2 levels of brand we can substitute I levels. The first factor, Brand, is labeled Factor A, the second factor, Water Amount, is labeled Factor B, etc

$$SS_{brand} = SS_A = \sum_{i=1}^I n_i * (\bar{y}_{i..} - \bar{y}_{...})^2 \quad (2.4.2)$$

In the two factor case, SS_B and SS_{AB} will also need to be calculated and compared to MSE.

$$SS_{water} = SS_B = \sum_{j=1}^3 52 * (\bar{y}_{.j.} - \bar{y}_{...})^2 \quad (2.4.3)$$

$$= \sum_{j=1}^J n_j * (\bar{y}_{.j.} - \bar{y}_{...})^2 \quad (2.4.4)$$

The general equation that calculates the interaction SS, SS_{AB} , is:

$$SS_{AB} = \sum_{i=1}^I \sum_{j=1}^J n_{ij} (ij^{th} \text{ level effect})^2 = \sum_{i=1}^2 \sum_{j=1}^3 26 (\bar{y}_{ij.} - \bar{y}_{i..} - \bar{y}_{.j.} - \bar{y}_{...})^2 \quad (2.4.5)$$

Degrees of Freedom (df) is the number of free units of information. In this example, there are 3 levels of factor B. Since the effects sum to 0, knowing the 0 and 5 drops of water effect forces a known 15 drops of water effect. If there are J levels for factor B, one level is fixed if we know the other $J - 1$ levels. Thus, when there are J levels for a main factor of interest, there are $J - 1$ free pieces of information.

For the AB interaction term there are I*J effects that are calculated. In this example I*J = 2*3 = 6. Each effect is a piece of information. In addition:

- All the AB effects sum to 0 (1 free piece of information is taken-only 5 effects are free)
- All the AB effects within Brand C add up to the Brand C effect (the same for Brand D). This requires I = 2 pieces of information (however 1 piece of information is already fixed from the requiring that the sum of all brand effects is 0). Thus 2-1, (I-1), pieces of information are required to fit this requirement.

- All the AB effects within 0 Water Amount add up to the 0 Water effect (the same for 5 and 15 drops of water). This requires $J = 3$ pieces of information (however 1 piece of information is already fixed from the requiring that the sum of all water effects is 0). Thus $3-1, (J-1)$, pieces of information are required to fit this requirement.

Thus the degrees of freedom for the interaction effect is:

$$\begin{aligned} df_{AB} &= \# \text{ of effects} - [df_A + df_B + 1] \\ &= IJ - [(I - 1) + (J - 1) + 1] \\ &= (I - 1)(J - 1) \end{aligned}$$

Mean Square(MS) = SS/df for each appropriate factor and interaction. Mean Square Error (MSE) is a pooled measure of the variability within each level combination. While many texts give specific formulas for Sum of Squares Error (SSE) and degrees of freedom Error (dfE), they can be most easily calculated by subtracting all other SS from the Total SS = $(N-1)(\text{Overall variance})$. In this example, $N = 156$, the total number of samples.

To determine if the difference between brand means is significant (i.e. at least one $\alpha_i \neq 0$), we compare the between-level variation of brand (MS_A) to the MSE.

If the MS_A is much larger than MSE, it is reasonable to conclude that there truly is a difference between brand level means and the differences observed in the sample data was not simply due to random chance.

F-statistic = MS for factor/MSE. The F-statistic is a ratio of the **between variability** (variation between factor level averages) over the **within variability** (pooled variability within each factor combination).

For each hypothesis test, if the F-statistic is large, it seems unlikely that the population means of that level combination are truly equal. These calculations are summarized in the ANOVA table below.

Mathematical theory proves that if the appropriate assumptions hold, the F-statistic follows an F distribution with df_{AB} (if testing factor AB) and df_E degrees of freedom. The **p-value** is looked up in an F table and gives the likelihood of observing an F statistic at least this extreme (at least this large), assuming that the true population factor has equal level means. Thus, when the p-value is small (e.g. less than 0.05 or 0.1), the effect size of that factor is statistically significant.

[On Your Own:]

19. Use Minitab to conduct an analysis on the paper towel data to test for differences between brands, water levels, and interactions. Provide appropriate graphs. Check the equal variance assumptions in Minitab. [Stat → ANOVA → 2-way ANOVA]. Use the p-values, SS and F statistics to summarize the results.

Source	df	SS	MS	F-Statistic
A	$I - 1$	$\sum_{i=1}^I n_i (\bar{y}_{i..} - \bar{ybar})^2$	$\frac{SS_A}{df_A}$	$\frac{MS_A}{MSE}$
B	$J - 1$	$\sum_{j=1}^J n_j (\bar{y}_{.j.} - \bar{ybar})^2$	$\frac{SS_B}{df_B}$	$\frac{MS_B}{MSE}$
AB	$(I - 1)(J - 1)$	$\sum_{j=1}^J \sum_{i=1}^I n_{ij} (\bar{y}_{ij.} - \bar{y}_{i..} - \bar{y}_{.j.} - \bar{ybar})^2$	$\frac{SS_{AB}}{df_{AB}}$	$\frac{MS_{AB}}{MSE}$
Error	<i>subtraction</i>	<i>subtraction</i>	$\frac{SS_{Error}}{df_{Error}}$	
Total	$N - 1$	$\sum_{j=1}^J \sum_{i=1}^I \sum_{k=1}^K (\bar{y}_{ijk} - \bar{ybar})^2$		

Table 2.3: 2-Way ANOVA

20. Figure 2 below gives a 3-way ANOVA table. Fill in the formulas for MS_C and the F statistic for the BC interaction. [See Factorial Design Figure 2: 3-way ANOVA on last page]
21. All paper towels for this experiment were purchased as one local store. How does this impact conclusions that can be made about this experiment? The students did all towels with 15 drops of water first, then all paper towels with 5 drops and finally did the 0 drops. How does this impact conclusions that can be made about this experiment?

2.5 ANOVA for Regression

Section 2.2 showed that a regression model for the paper towel data structure could be:

$$y_m = \beta_0 + \beta_1 x_{1m} + \beta_2 x_{2m} + \beta_{12} x_{1m} x_{2m} + \epsilon_m \quad m = 1, 2, \dots, 156 \quad (2.5.1)$$

Where x_1 is the brand categorical variable and x_2 is a quantitative variable with values of 0, 5, or 15 drops of water. Each of the beta coefficients $(\beta_1, \beta_2, \beta_{12})$, are the unknown parameters that will be estimated from the data. As in the single explanatory variable case, each of the 156 ϵ 's, error terms, are assumed to be independent and have a $N(0, \sigma^2)$ distribution.

The calculation of each of the β estimates uses the same procedure as when we have only one explanatory variable: find β estimates that minimize the sum of squared residuals. The

mathematical and computational work to find these estimates typically uses matrix algebra to simplify the notation and to speed up computations. In this lab, we will use software to calculate the coefficients instead of calculating them by hand. Once the estimates have been calculated, they can be plugged into equation (2.5.1) to find the expected breaking strength, \hat{y} .

The following ANOVA table is used to calculate R^2 , the percent of the total variation that can be explained by the regression model.

Source	df	SS	MS	F-Statistic
Regression Model	$p - 1$	$\sum_{m=1}^N (\hat{y}_m - \bar{y})^2$	$\frac{SS_{Model}}{df_{Model}}$	$\frac{MS_{Model}}{MSE}$
Error	$N - p$	$\sum_{m=1}^N (y_m - \hat{y}_m)^2$	$\frac{SS_{Error}}{df_{Error}}$	
Total	$N - 1$	$\sum_{m=1}^N (y_m - \bar{y})^2$		

Table 2.4: ANOVA for Regression

Here N is 156, the total number of samples. p is 4, the total number of β s that are estimated in the model. \hat{y}_m is the predicted strength value based on the model, \bar{y} is the overall mean strength, and y_m is the observed strength. $R^2 = SS_{Model}/SS_{Total}$, the variation described by the model, over the overall variability.

[On Your Own:]

22. Create a regression model using **Strength** as the response and **Brand** and **Water** as the explanatory variables. Create **residual vs brand** and **residual vs water** plots. Calculate the regression line, the R^2 value, the t-statistics and p-values for the coefficients. [In Minitab select **Stat** → **Regression** → **Regression**. The response is **Strength** The Predictors are **Water** and the dummy variable **BrandC**. In Graphs select **Histogram of residuals** and in **Residuals versus the variables**: select **Water** and **BrandC**.] Do the residual plots show any reason to doubt that the errors are normally distributed with a constant variance?
23. Create a regression model using **SQRTStrength** as the response and **Brand** and **Water** as the explanatory variables. Create the appropriate residual plots. Calculate the regression line, the R^2 value, the t-statistics and p-values for the coefficients. Do the residual plots show any reason to doubt that the errors are normally distributed with a constant variance?
24. Compare the test statistics in the two regression calculations in the previous questions.

Describe any relationship. Does transforming the response variable affect R^2 or the significance of the β s? Explain why or why not?

25. Use the regression line (using `SQRTStrenth`) to predict breaking strength for Brand C with 10 Drops of Water. Do you believe this is a fairly accurate prediction? Describe why. Provide a predicted breaking strength for Brand C with 50 Drops of Water (using the `SQRTStrenth` model). Do you believe this is a fairly accurate prediction? Describe why.

In equation 2.5.1, amount of water is assumed to be a quantitative variable. Treating amount of water as a quantitative variable may not be considered valid by some since there are only three levels. Typically it would be best to have a minimum of 5 levels (often much more) of the explanatory variable to explore linear relationships between an explanatory and response variable.

A second regression analysis, which follows our original design, would be to treat amount of water as a categorical variable. Creating **dummy variables** is a process of mapping the one column of categorical data into columns of 0 and 1 data. The 3 water levels (**0 drops**, **5 drops**, and **15 drops**) can be recoded using dummy variables, also called indicator variables. Instead of one column for amount of water, there will be 3 new columns, **0 drops**, **5 drops**, and **15 drops**. The **5 drops** column will have a 1 when the water amount is 5 and have a 0 value otherwise. The **15 drops** column will have a 1 when the water amount is 15 and have a 0 value otherwise. The same for the **0 drops** column.

Any of the three dummy variables can be included into a regression model, but only two should be included at any one time. This is because there is complete redundancy in the third the dummy variable. For example, knowing **0 drops** and **5 drops** has a value of 0 automatically tells us that there are 15 drops of water. We will leave the **15 drops** column out of our model. The slope coefficient for a dummy variables is an estimate of the average amount of change (of the response variable) when there is a “1” for that dummy variable rather than a “0.” Using this from of the regression model, it is not appropriate to attempt to make predictions for any “10 drops of water” group.

The statistical model treating water amount as a categorical variable would be:

$$y_m = \beta_0 + \beta_1 x_{1m} + \beta_2 x_{2m} + \beta_3 x_{3m} + \beta_{12} x_{1m} x_{2m} + \beta_{13} x_{1m} x_{3m} + \epsilon_{ijk} \quad m = 1, 2, \dots, 156 \quad (2.5.2)$$

Where x_1 is the brand dummy variable, x_2 is the 0 drops dummy variable and x_3 is the 5 drops dummy variable.

[On Your Own:]

26. Create columns of indicator variables. [In Minitab, Calc → Make Indicator Variables → Indicator Variable for Water → Store Results into c6-c8 (select any 3

unused columns here) \rightarrow OK]. Name the columns, in order Zerodrops, Fivedrops, and Tendrops.

27. Create a regression model using **SQRTStrength** as the response and **Brand**, **Zerodrops** and **Fivedrops** as the explanatory variables. Create the appropriate residual plots. Calculate the least squares regression line, the R^2 value, the t-statistics and p-values for the coefficients. Do the residual plots show any reason to doubt the normality or equal variance assumptions? Compare the test statistics in this regression analysis to the F statistics in ANOVA. Describe any relationship.

This lab has introduced the basics of very powerful statistical techniques. Multivariate ANOVA and multivariate regression. The ability to simultaneously test for the impact of multiple variables upon a response allows statisticians to better model real world situations. If used properly, these techniques efficiently and reliably help us better understand the world we live in. In future labs, you will have the opportunity to experience for yourself how these techniques are used in biology, chemistry, engineering, psychology and many other disciplines.