I. Review of ANOVA vs. regression:
   a) Same
      i) Context is an observational study or an experiment
      ii) Outcome variable: quantitative
      iii) All subjects with the same level of the explanatory variable (“in the same group”) are assumed to have the same mean and vary around that mean according to a Normal distribution (bell shaped curve)
      iv) All subjects with the same level of the explanatory variable (“in the same group”) are assumed to have a common variance, $\sigma^2$.
      v) Errors (deviations from the group mean) are assumed to be independent across subjects.
      vi) Group assignments are assumed to be clear cut (fixed x assumption).
   b) Different
      i) One-way ANOVA
         (1) Explanatory variable: Categorical
         (2) Mean parameters are $\mu_1$ through $\mu_k$. Best predictions are $\bar{x}_1$ through $\bar{x}_k$.
      ii) Regression
         (1) Explanatory variable: Quantitative
         (2) Coefficient parameters are $\beta_0$ and $\beta_1$.
         (3) Mean outcome at $X$ is $\mu(Y|X)=\beta_0 + \beta_1 X$ (linearity assumption).
            Best prediction is $\hat{\mu}(Y|X)=\hat{\beta}_0 + \hat{\beta}_1 X$ (a.k.a. $b_0 + b_1 X$).
II. **Multiple Regression**

a) **New idea #1**: In *multiple* regression we extend to multiple explanatory variables in a natural way

i) If $X_1$, $X_2$ are two explanatory variables, then the model states that

   the mean of the outcome for any group of subjects with any particular set of values for the explanatory variables is

   $\mu(Y|X_1, X_2) = \beta_0 + \beta_1 X_1 + \beta_2 X_2$.  

ii) The best prediction is

   $\hat{\mu}(Y|X_1, X_2) = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2$.  

iii) E.g. Lab 6, marigold data: consider time of exposure to radiation as an additional explanatory variable

b) Consequences of this new idea:

i) $\beta_0$ is the mean of the outcome when *both* explanatory variables are zero.

ii) $\beta_1$ is the change in outcome associated with (or caused by) a one unit increase in the first explanatory variable when the second explanatory variable is held constant.

iii) $\beta_2$ is the change in outcome associated with (or caused by) a one unit increase in the second explanatory variable when the first explanatory variable is held constant.

iv) At any fixed level of $X_1=c$, the mean of the outcome for various levels of $X_2$ is

   $\mu(Y|X_1=c, X_2) = \beta_0 + \beta_1 c + \beta_2 X_2 = (\beta_0 + \beta_1 c) + \beta_2 X_2$, i.e. it is a straight line with intercept $(\beta_0 + \beta_1 c)$ and slope $\beta_2$. 

III. Dummy or Indicator Variables

a) *New idea #2:* In multiple regressions, categorical explanatory variables can be used if they are coded as indicator (dummy) variables.
   i) Regression with at least one quantitative and at least one categorical variable is called ANCOVA (analysis of covariance)
   ii) Especially when the categorical variable is of primary interest.

b) For *binary* nominal variables, the most frequently used codes are 0/1 and -1/+1.

c) For nominal variables with *k levels*, create *k*-1 explanatory variables that are 0/1 where any case can have at most one of these explanatory variables equal to 1. E.g. to code red vs. blue vs. green, create a “red” variable and a “blue” variable, then code cases as follows:

<table>
<thead>
<tr>
<th>Color</th>
<th>“Red” explanatory variable</th>
<th>“Blue” explanatory variable</th>
</tr>
</thead>
<tbody>
<tr>
<td>Red</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>Blue</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Green</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

Here, green is the “baseline”, but our conclusions about the experiment will be the same no matter how we code the color variable, although the meaning of specific parameters will change.

d) For *ordinal* variables, code like nominal, but analyze differently.

e) *Important idea:* Simplify regression equations, especially when there are categorical explanatory variables.
   i) Color example: \( \mu(Y|X_1, \text{Red, Blue}) = \beta_0 + \beta_1 X_1 + \beta_{\text{red}} \text{Red} + \beta_{\text{blue}} \text{Blue} \)
   becomes
   (1) For Green: \( \beta_0 + \beta_1 X_1 \)
   (2) For Red: \( (\beta_0 + \beta_{\text{red}}) + \beta_1 X_1 \)
   (3) For Blue: \( (\beta_0 + \beta_{\text{blue}}) + \beta_1 X_1 \)

f) Conclusion: In regression, dummy variables’ “slopes” become different *intercepts*. 
IV. Diaper Example
A psychologist wants to study the relationship between urinary serotonin excretion and various psychological traits in infants. Wary of the potential sample collection problems, she conducts a pilot experiment to determine whether squeezing urine out of diapers affects the accuracy of serotonin measurement. Her two explanatory variables are diaper type and time the urine sits in the diaper. Her outcome is serotonin concentration. The diaper types are control (no diaper, just a standard plastic urine container), cloth and disposable (paper). Her times are immediate (coded as 0 hrs), 1, 2 and 3 hours. The basic experimental procedure follows. Collect and pool urine from 10 adults. Pour 10 mL (milliliters) of urine into each of 12 cups for the controls. Pour 30 mL urine onto each of 12 cloth and 12 disposable diapers. Incubate the diapers and cups at 90° C for the designated times then, for a group of 3 diapers, squeeze about 10 mL out of the diapers into cups and refrigerate. Analyze all of the urine samples using standard laboratory techniques for serotonin.

a) Experiment or observational study? Experimental units?
   Interpretability? Generalizability? Power?

b) Null hypotheses?

c) Your guess at the outcomes?
d) Regression output:

```
The REG Procedure
Model: MODEL1
Dependent Variable: SERTONIN

Analysis of Variance

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>Sum of Squares</th>
<th>Mean Square</th>
<th>F Value</th>
<th>Pr &gt; F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td>3</td>
<td>13165</td>
<td>4388.45185</td>
<td>79.85</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>Error</td>
<td>32</td>
<td>1758.64444</td>
<td>54.95764</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Corrected Total</td>
<td>35</td>
<td>14924</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Root MSE 7.41334  R-Square 0.8822
Dependent Mean 114.33333  Adj R-Sq 0.8711
Coeff Var 6.48397

Parameter Estimates

| Variable | DF | Parameter Estimate | Standard Error | t Value | Pr > |t| |
|----------|----|--------------------|----------------|---------|------|---|
| Intercept| 1  | 138.70000          | 2.70697        | 51.24   | <.0001 |
| TIME     | 1  | -5.35556           | 1.10512        | -4.85   | <.0001 |
| CLOTH    | 1  | -7.33333           | 3.02648        | -2.42   | 0.0212 |
| PAPER    | 1  | -41.66667          | 3.02648        | -13.77  | <.0001 |
```

e) Prediction equations:
f) Residual plots:

Scatterplot
Dependent Variable: Serotonin Conc.
Regression Standardized Predicted Value

Regression Standardized Residual

g) EDA we should not have skipped.
V. Interaction

a) **New idea #3**: An interaction between explanatory variable A and explanatory variable B indicates that the effects of A on the outcome will depend on the specific level of B and vice versa. (Contrast with 1c above.)

b) Interaction can be coded as a new explanatory variable that is the product of two old explanatory variables.

c) ANCOVA with interaction:
\[ \mu(Y|X_1, X_2) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_{12} (X_1 X_2). \]
The best prediction is
\[ \hat{\mu}(Y|X_1, X_2) = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2 + \hat{\beta}_{12} (X_1 X_2). \]

d) Diaper example: \( \mu(\text{Serotonin}|\text{time, cloth, paper}) = \beta_0 + \beta_{\text{time time}} + \beta_{\text{cloth Cloth}} + \beta_{\text{paper Paper}} + \beta_{\text{time cloth}(\text{time*Cloth})} + \beta_{\text{time paper}(\text{time*Paper})}. \)
Which simplifies as:
\[ \mu(\text{Serotonin}|\text{time, nodiaper}) = \beta_0 + \beta_{\text{time time}} \]
\[ \mu(\text{Serotonin}|\text{time, cloth}) = (\beta_0 + \beta_{\text{cloth}}) + (\beta_{\text{time}} + \beta_{\text{time cloth}}) \text{time} \]
\[ \mu(\text{Serotonin}|\text{time, paper}) = ? \]
VI. Re-visit Diapers

The REG Procedure

Model: MODEL1
Dependent Variable: SERTONIN

Analysis of Variance

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>Sum of Squares</th>
<th>Mean Square</th>
<th>F Value</th>
<th>Pr &gt; F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td>5</td>
<td>13830</td>
<td>2766.04000</td>
<td>75.87</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>Error</td>
<td>30</td>
<td>1093.80000</td>
<td>36.46000</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Corrected Total</td>
<td>35</td>
<td>14924</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Root MSE 6.03821  R-Square 0.9267
Dependent Mean 114.33333  Adj R-Sq 0.9145
Coeff Var 5.28123

Parameter Estimates

| Variable | Parameter DF | Standard Parameter | Standard Error | t Value | Pr > |t| |
|----------|-------------|--------------------|----------------|---------|------|---|
| Intercept| 1           | 132.66667          | 2.91673        | 45.48   | <.0001 |
| TIME     | 1           | -1.33333           | 1.55906        | -0.86   | 0.3992 |
| CLOTH    | 1           | -3.03333           | 4.12488        | -0.74   | 0.4678 |
| PAPER    | 1           | -27.86667          | 4.12488        | -6.76   | <.0001 |
| TIMECLTH | 1           | -2.86667           | 2.20484        | -1.30   | 0.2034 |
| TIMEPAPR | 1           | -9.20000           | 2.20484        | -4.17   | 0.0002 |
Model equations:

\[ \mu(Y|\text{time},\text{paper},\text{cloth}) = \beta_0 + \beta_{\text{time}} \text{time} + \beta_{\text{cloth}} \text{cloth} + \beta_{\text{paper}} \text{paper} + \beta_{\text{time}\text{*cloth}} \text{time}\text{*cloth} + \beta_{\text{time}\text{*paper}} \text{time}\text{*paper} \]

The separate prediction equations are:

\[ \hat{\mu}(Y | \text{time, control}) = \hat{\beta}_0 + \hat{\beta}_{\text{time}} \text{Time} \] because all other terms are zero.

\[ \hat{\mu}(Y | \text{time, paper}) = \hat{\beta}_0 + \hat{\beta}_{\text{time}} \text{Time} + \beta_{\text{paper}} \text{Paper} + \beta_{\text{time}\text{*paper}} \text{Time}\text{*Paper} \]

which simplifies as \((\hat{\beta}_0 + \hat{\beta}_{\text{paper}}) + (\hat{\beta}_{\text{time}} + \hat{\beta}_{\text{time}\text{*paper}}) \text{Time}\)

So, e.g. \(\hat{\beta}_{\text{paper}}\) is the estimate of the difference in intercept between control and paper. And \(\hat{\beta}_{\text{time}\text{*cloth}}\) is the estimate of the difference in slope between control and cloth.

Numeric equations:

\[ \hat{\mu}(Y | \text{time, control}) = 132.7 - 1.33\text{Time} \]

\[ \hat{\mu}(Y | \text{time, paper}) = 104.8 - 10.53\text{Time} \]

Interpretation:

My interpretation is that both control diapers initially read 132.7, on average, which is presumably the true serotonin concentration, and maintains the correct concentration over the 3 hours of incubation. The paper diaper seems to absorb an average of 27.9 units of serotonin immediately. In addition an additional 9.20 units per hour are lost or destroyed during 1 to 3 hours of incubation in paper at 90° C.

\[ \mu(Y|\text{time, cloth})=? \]
Residual Plots:

Scatterplot
Dependent Variable: Serotonin Conc.

Regression Standardized Predicted Value

Sitting Time (Hours)

Unstandardized Residual

Sitting Time (Hours)