This presentation covers the topics of Simple Linear Regression (SLR), Polynomial Regression, Multiple Regression (including using categorical variables as predictors), and Logistic Regression using a binomial response variable.

Section 1: Simple Linear Regression

In Simple Linear Regression we attempt to construct a functional model from data. The data includes a single response variable, $Y$, which is numerical and a single explanatory or predictor variable, $X$, which is also numerical. We use the least squares regression technique to "fit" a model that captures both the appropriate structure on the means presented by the data and an error structure that allows us to do inference on the parameters in our model.

For a SLR Model, the mean response is a simple linear function of the explanatory (predictor) variable. The individual responses, $Y_i$, vary around the mean. Symbolically, we have:

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i,$$

$$\epsilon_i \sim N(0, \sigma^2)$$

The structure on the means is $E(Y_i \mid X_i) = \beta_0 + \beta_1 X_i$. For every $x$ there is a distribution of possible $y$'s centered at $E(Y \mid X_i)$. The error structure is $\epsilon_i \sim N(0, \sigma^2)$. Notice that the variance is the same for all $X_i$. The figure below illustrates the basic model for Simple Linear Regression:

![Figure 1: Graphical Representation of $Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$.](image)
**SLR for Data Exploration:** If we are exploring data looking for possible models, we are most interested in the structure on the means. We examine fits and residuals and look at $R^2$, the percentage of variability in the response that is explained by the model. To investigate the structure on the means requires no assumptions about the distribution of the errors.

**SLR for Inference:** If we wish to evaluate the statistical significance of regression coefficients, we need the error structure. We assume several things about errors.
- they are independent (random)
- they are normally distributed
- they have mean zero and constant variance, $\sigma^2$.

These assumptions about the error structure can be checked by analyzing the residuals. The residuals are considered to be a random sample from the distribution of the errors, that is, a random sample from $N(0,\sigma^2)$. If these assumptions on the errors are violated, the probability values and confidence levels used in inference will not be correct.

Note that there are no assumptions about the distribution of $X$.

**Residuals and Errors**
Note that the errors are theoretical, they go with $\beta_0$ and $\beta_1$. There is a distinction to be made between the theoretical errors and the residuals. The residuals are sample versions of the errors. The residuals go with $\hat{\beta}_0$ and $\hat{\beta}_1$. The residuals give information needed to estimate the errors and the properties of the errors. The residuals come from data. The errors are what you have theoretically before you look at the data.

The residuals give you information about both the supposed structure of the mean and the structure of the errors. Consider the two plots below:

**Figure 2:** Two residual plots
The residual plot on the left indicates that the structure of the mean is not being captured by the model used, while the residual plot on the right indicates that the structure of the means is OK, but the assumption about the constant variance of the errors is incorrect.

**Modeling the Growth of an Epidemic, HIV/AIDS example:**

The September 13, 1990 *Iowa State Daily* reported deaths each year from 1980 to 1988 attributable to HIV/AIDS in women aged 15 to 44 years. The *Daily* reported the total number of deaths. Since the number of women at risk was approximately 60,000,000 throughout the period, the total deaths provide the same information as incidence rates, *i.e.* deaths per 100,000 women. Below are the data.

**HIV/AIDS Data:**

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>X</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>6</td>
<td>7</td>
<td>8</td>
<td>9</td>
</tr>
<tr>
<td>Deaths</td>
<td>18</td>
<td>30</td>
<td>36</td>
<td>92</td>
<td>198</td>
<td>360</td>
<td>631</td>
<td>1016</td>
<td>1430</td>
<td>?</td>
</tr>
</tbody>
</table>

The response, or predicted variable, *Y* represents the number of deaths attributable to HIV/AIDS in women aged 15 to 44. The explanatory variable, *X* represents the year since 1980 (e.g. *Year* = 1982, *X* = 2). Since this is time series data, the response values cannot be considered independent. We cannot do inference with this data. Consequently, we will consider only the structure on the means.

Questions:
- Is there a strong linear correlation between Deaths and Year (or *X* = *Year* − 1980)?
- Will a simple linear regression model provide a good fit to these data?

**HIV/AIDS SLR:**
- Correlation: *r* = 0.909
- Coefficient of determination: $R^2 = 82.7\%$
- Prediction equation: $\hat{Y} = -247.5 + 167.7X$

With a correlation of *r* = 0.909 and coefficient of determination $R^2 = 82.7\%$, the answer to the question "Is there a strong linear correlation between Deaths and Year?" is "Yes. With a correlation of 0.909, there is a strong linear correlation between Deaths and Year." As the predictor variable *Year* increases, the response variable *Deaths* increases.

A look at the scatterplot gives additional information about the relationship between *Year* and *Deaths*. 
If you looked only at the values of $r$ and $R^2$, you might think that the Simple Linear Regression is not a bad fit. The slope coefficient indicates that each year, on average, there are 168 more deaths. In interpreting the slope, we understand that a unit change in the predictor variable corresponds to an *average* change in the response variable, since we are trying to establish the structure on the means.

However, as is made evident from the graph above, there are a few problems with this regression. The intercept is negative, and the prediction for 1989 ($Deaths = 1262$) is less than the observed value for 1988. A look at the residual plot offers further evidence that while there is a strong linear correlation, a linear model is insufficient to capture the structure on the means.
The residuals are the difference between the observed data and the predicted mean response of the model, the fitted data. There is a definite curved pattern to the residuals. Consequently, we reject the linear structure on the means as a sufficient model. Perhaps a model with a squared term will improve the fit and capture more of the structure on the means.

Simple Quadratic Model

The simple regression model with a quadratic term has the following structure on the mean:

$$E(Y_i \mid X_i) = \alpha_0 + \alpha_1 X_i^2.$$

Our prediction equation is $\hat{Y} = -72.8 + 21.9 X^2$. For this model, the correlation factor is $r = 0.987$, and the coefficient of determination is $R^2 = 97.4\%$, indicating a very strong linear correlation between Deaths and Year$^2$.

As with the Simple Linear Regression with the predictor value $X$, there are problems with the line using $X^2$ as the predictor variable. Again, we have a negative intercept, so the model still predicts negative deaths in 1980. Also, our fitted values over-predict from
'83 through '86 and under predict from '87 through '88, as illustrated in the residual plot below. Notice that the range of the residuals is reduced in half from the previous model (See Figure 5).

![Residual plot](image)

Figure 7  HIV/AIDS residuals

Although the model with $X^2$ is better than the model with $X$, it still does not fit the data as well as we would like. So, we will try a model with both $X$ and $X^2$ in it. Such a model will have the following structure on the means.

$$E(Y_i | X_i) = \beta_0 + \beta_1 X_i + \beta_2 X_i^2.$$  

Regression Analysis  (JMP-IN for the quadratic model.)

Polynomial Fit degree=2  
Deaths = 64.4909 – 99.6779 X + 33.4264 X^2

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>Sum of Squares</th>
<th>Mean Square</th>
<th>F Ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td>2</td>
<td>2032204.3</td>
<td>1016102</td>
<td>666.0087</td>
</tr>
<tr>
<td>Error</td>
<td>6</td>
<td>9154.0</td>
<td>1526</td>
<td>Prob&gt;F</td>
</tr>
<tr>
<td>C Total</td>
<td>8</td>
<td>2041358.2</td>
<td></td>
<td>&lt;.0001</td>
</tr>
</tbody>
</table>

Parameter Estimates

| Term    | Estimate | Std Error | t Ratio | Prob>|t| |
|---------|----------|-----------|---------|--------|
| Intercept | 64.490909 | 31.7468 | 2.03 | 0.0885 |
| X       | -99.67792 | 18.50534 | -5.39 | 0.0017 |
| X^2     | 33.426407 | 2.225632 | 15.02 | <.0001 |
From the computer output, we see that the coefficient of determination is $R^2 = 99.6\%$ and the prediction equation is

$$\hat{Y} = 64.49 - 99.68X + 33.43X^2.$$  

![Figure 8 Quadratic Model Plot]

Notice there is no computed value of $r$, the correlation coefficient. While $R^2$ is defined for any number of predictor variables, $r$ is defined for only 1 predictor variable. Consequently, $r^2 = R^2$ only for one predictor variable.

Having $R^2 = 99.6\%$ is a big improvement over the 82.7% from the simple model with $X$ and also over the 97.4% from the simple model with $X^2$. However, with this quadratic model, the coefficients are strange. We know that as the years progressed the number of deaths increased. However, the coefficient for $X$ is negative. In this model, $X$ and $X^2$ are highly related. This causes problems with the interpretation of coefficients in the model. Students should understand that the coefficient of $X$ cannot be interpreted the same way as in the linear model, since you cannot separate a change in $X$ from a change in $X^2$.

**Multicollinearity**

When predictor variables are highly correlated, they are called multicollinear. A change in one predictor variable is accompanied by a change in the other. The usual interpretation of a slope coefficient (the average change in response for a unit change in $X$, holding all other variables constant) does not hold true for multicollinear variables. Correlated predictor variables also cause problems with interpreting test statistics, and tend to produce misleading standard errors.

**Questions for the Quadratic Model.**

What is the average change in deaths for a unit (one year) change in $X$? The answer depends on where the one unit change occurs.
Fitted Values:

1983: \( X = 3, \hat{Y} = 66.32 \)  
1984: \( X = 4, \hat{Y} = 200.65 \)

\( \text{Change} = 134.33 \)

1987: \( X = 7, \hat{Y} = 1004.8 \)  
1988: \( X = 8, \hat{Y} = 1406.57 \)

\( \text{Change} = 401.77 \)

Derivative:

\[-99.7 + 66.8X\]  
Evaluated at 3.5: 134.1  
Evaluated at 7.5: 401.3

One way to make the coefficients more easily interpreted is to re-code the data by centering the \( x \)'s. Instead of having 1980 as 0, use 1984 as 0. This puts the 0 point in the center of the predictor variable. If we code the \( x \)'s by subtracting 4, so that \( x_c = x - 4 \), a Simple Linear Regression on \( x_c \) yields the following equation:

\[ \hat{Y} = 423.4 + 167.7x_c \]

Recall that without re-coding the equation was

\[ \hat{Y} = -247.5 + 167.7x \].

Notice that there is no change in slope, yet this fit hides the problem of the negative \( y \)-intercept.

With a little algebra, it is easy to see that the two equations are the same:

\[ \hat{Y} = 423.4 + 167.7x_c \]
\[ = 423.4 + 167.7(x - 4) \]
\[ = -247.4 + 167.7x \]

Including the quadratic term in the re-coded form yields the following:

\[ \hat{Y} = 200.6 + 167.7x_c + 33.4x_c^2 \]

The coefficient for \( x_c \) does not change from the linear model. It still represents the idea that on average the number of deaths go up 168 each year. Note that with this re-coding, the correlation between \( x \) and \( x^2 \) is now 0 since the predictor variables \( x_c \) and \( x_c^2 \) are orthogonal.

<table>
<thead>
<tr>
<th>( x )</th>
<th>-4</th>
<th>-3</th>
<th>-2</th>
<th>-1</th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
</tr>
</thead>
<tbody>
<tr>
<td>( x^2 )</td>
<td>16</td>
<td>9</td>
<td>4</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>4</td>
<td>9</td>
<td>16</td>
</tr>
</tbody>
</table>
Note that the terms in the dot-product cancel each other,

\[ \sum x \cdot x^2 = (-4)(16) + (-3)(9) + \cdots + (3)(9) + (4)(16) = 0. \]

The coding on \( x \) removed the multicollinearity between \( x \) and \( x^2 \). If adding a variable does not cause a change in the coefficients you already have in a model, then the variables are probably independent. So with re-coding, the purely linear component can be separated from the purely quadratic component.

What implication does this have for experimental design? When we get to design experiments and pick our own values of \( x \), we can pick them so that the coding will work to produce orthogonal predictor variables. As shown in this example, the coefficients of the predictor variables are easier to interpret when they are orthogonal.

With coding, the predictions are the same, the value of \( R^2 \) remains the same, and the Analysis of Variance (ANOVA) tables are the same. As before, if we replace \( x_c \) with \( x - 4 \), we have the original quadratic fit.

\[
\hat{Y} = 200.6 + 167.7x_c + 33.4x_c^2 \\
\hat{Y} = 200.6 + 167.7(x - 4) + 33.4(x - 4)^2 \\
\hat{Y} = 64.5 - 99.7x + 33.4x^2
\]

Note that the range of the residuals has been cut in half again (compare to Figure 7), but a cubic pattern remains. The quadratic model still has some problems; the fitted values for 1981 and 1982 are negative and there appears to be a double curve in the plot of residuals (two turning points). Again, the structure on the means appears insufficient.
We could add a cubic term to the model, however, even with the centered $x$’s, there will be multicollinearity among the predictor variables.

**Cubic Regression (JMP-IN Output)**

Polynomial Fit degree=3  
Deaths = 200.602 + 140.726 $X^*$ + 33.4264 $X^{*^2}$ + 2.28872 $X^{*^3}$  

Summary of Fit  
- RSquare: 0.999174  
- R Square Adj: 0.998678  
- Root Mean Square Error: 18.36473  
- Mean of Response: 423.4444  
- Observations (or Sum Wgts): 9

Analysis of Variance  
<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>Sum of Squares</th>
<th>Mean Square</th>
<th>F Ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td>3</td>
<td>2039671.9</td>
<td>679891</td>
<td>2015.904</td>
</tr>
<tr>
<td>Error</td>
<td>5</td>
<td>1686.3</td>
<td>337</td>
<td>Prob&gt;F</td>
</tr>
<tr>
<td>C Total</td>
<td>8</td>
<td>2041358.2</td>
<td></td>
<td>&lt;.0001</td>
</tr>
</tbody>
</table>

Parameter Estimates  
| Term     | Estimate   | Std Error | t Ratio | Prob>|t| |
|----------|------------|-----------|---------|-----|
| Intercept| 200.60173  | 9.28121   | 21.61   | <.0001 |
| $X^*$    | 140.72643  | 6.209827  | 22.66   | <.0001 |
| $X^{*^2}$| 33.426407  | 1.046428  | 31.94   | <.0001 |
| $X^{*^3}$| 2.2887205  | 0.486391  | 4.71    | 0.0053 |

**Exponential Model**  
Continuing to add variables to the polynomial fit will give increasing values of $R^2$ until eventually, the $R^2$ is 1. However, we may not have a useful model. The residual plots indicate that the polynomials still miss important characteristics on the structure on the means. Perhaps, the mean response exhibits exponential growth. In this case, we have  
$$E(Y_i|X_i) = \gamma_0 e^{\alpha_1 X_i}$$  
The logarithm of the mean is linear.  
$$\ln(E(Y_i|X_i)) = \alpha_0 + \alpha_1 X_i$$

Now we can fit a simple linear function of $\ln(Deaths)$ on $X$.  
Fitted $\ln(Deaths) = 2.790 + 0.586 X$ with $R^2=98.7\%$

To compare $R^2$ values for different models, we have:  
- SLR( $X$ ): $R^2=82.7\%$  
- SLR( $X^2$ ): $R^2=97.4\%$  
- Quadratic Model: $R^2=99.6\%$  
- Cubic Model: $R^2=99.9\%$  
- Exponential: $R^2=98.7\%$
Note that the exponential model’s $R^2$ value is on the re-expressed data, and is therefore different from and not comparable to the others since the residuals are on a logarithmic scale. It’s okay to compare the $R^2$ values of the first and second models, $X$ and $X^2$, since they are both based on one variable.

Obviously, the higher the value of $R^2$, the better the fit. If we just look at the numbers, the cubic model is the winner. Note, however, that adding a variable to a model can never decrease $R^2$, even if the variable has nothing to do with the situation. Therefore, it is often helpful to look at $R^2_{\text{adjusted}}$.

$$R^2 = \frac{SS_{\text{Regression}}}{SS_{\text{Total}}} = 1 - \frac{SS_{\text{Error}}}{SS_{\text{Total}}}$$

$R^2_{\text{adjusted}}$ adjusts $R^2$ for the number of variables and number of observations by using the Mean Square Errors.

$$\text{Adjusted } R^2 = 1 - \frac{MS_{\text{Error}}}{MS_{\text{Total}}}$$

The sum of squares total is the unscaled variance of $y$. $SS_{\text{Total}} = (n-1)s^2 = \sum (y_i - \bar{y})^2$. No matter what model we fit, the $SS_{\text{Total}}$ never changes.

The $SS_{\text{Error}}$ measures the spread of the residuals, $\sum resid^2$. Adding terms to the model always decreases $\sum resid^2$ whether the additional term is significant or not. However, $MS_{\text{Error}} = \frac{\sum resid^2}{df}$ may not decrease, since both $\sum resid^2$ and $df$ go down whenever you add a term to the model. Determining whether the loss in $\sum resid^2$ is overshadowed by the loss of degrees of freedom is a question of inference.

<table>
<thead>
<tr>
<th>Source</th>
<th>SLR df</th>
<th>$x$ and $x^2$</th>
<th>$x$, $x^2$ and $x^3$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Regression</td>
<td>1</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>Error</td>
<td>$n-2$</td>
<td>$n-3$</td>
<td>$n-4$</td>
</tr>
<tr>
<td>Total</td>
<td>$n-1$</td>
<td>$n-1$</td>
<td>$n-1$</td>
</tr>
</tbody>
</table>

Adding variables never decreases $R^2$, but the adjusted $R^2$ could easily go down—“spending a degree of freedom” doesn’t give us much back when this happens. Adjusted $R^2$ actually can be less than 0; adjusted $R^2$ is always less than $R^2$.

Also, $MS_{\text{Total}} \left(1-adj R^2\right) = MS_{\text{Error}}$. 

11
SLR(\(X\)): \(\text{adj } R^2 = 80.2\%\)
SLR(\(X^2\)): \(\text{adj } R^2 = 97.0\%\)
Quadratic: \(\text{adj } R^2 = 99.4\%\)
Cubic: \(\text{adj } R^2 = 99.9\%\)

In the exponential growth model, the response variable is on a different scale than the other models. Convert fitted log deaths to deaths before looking at residuals.

To compare this \(R^2\) value to others, we must first convert back to the original scale.

Figure 10: Re-expressed semi-log graph for HIV/AIDS Data

Figure 11: Exponential graph on original scale

Note that the original scale shows (year, Deaths) instead of (year, ln Deaths).

The exponential growth model also has its problems; the fit for 1988 is way off, and the value of \(R^2\) is not comparable to the values for the other models.
<table>
<thead>
<tr>
<th>X</th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
</tr>
</thead>
<tbody>
<tr>
<td>Deaths</td>
<td>Y</td>
<td>18</td>
<td>30</td>
<td>36</td>
<td>92</td>
<td>198</td>
<td>360</td>
<td>631</td>
<td>1016</td>
</tr>
<tr>
<td>Pred Deaths</td>
<td>Ŷ</td>
<td>16.3</td>
<td>29.3</td>
<td>52.6</td>
<td>94.5</td>
<td>169.7</td>
<td>305.0</td>
<td>547.9</td>
<td>984.5</td>
</tr>
<tr>
<td>(Y - Ŷ)²</td>
<td>2.94</td>
<td>0.55</td>
<td>274.7</td>
<td>6.06</td>
<td>799.5</td>
<td>3030</td>
<td>6900</td>
<td>991.3</td>
<td>114,886</td>
</tr>
</tbody>
</table>

The table above gives the predictions transformed to the original scale and the squared residuals. The sum is not exact due to round-off in presenting the table. For this model, we have

\[ SS_{Error} = \sum (Y - \hat{Y})^2 = 126,890 \, . \]

So \( R^2 = 1 - \frac{SS_{Error}}{SS_{Total}} \) gives a coefficient of determination \( R^2 = 1 - \frac{126,890}{2,041,358} = 93.8\% \). This \( R^2 \) is more comparable to the others.

So which is the best model? The fit provided by a model that includes linear, quadratic, and cubic terms is quite good. Models of this type are used frequently to model the growth of epidemics.

Note that if you have theory supporting a particular function type, it’s often better to start with a model of that type. The process described/discussed above would more likely be used in situations where you are investigating/exploring the data.

We have not mentioned statistical inference or discussed \( t \)-tests for the statistical significance of variables. Even though a statistical analysis package will automatically calculate values of test statistics and \( p \)-values, they may not be applicable. Inference is the process of using random sample data to make generalizations about the population. In our HIV/AIDS example, we do not have a random sample of data and so do not have an error structure. Without a random sample (error structure) we have nothing to infer. We are simply looking at the fit (or lack of fit) of various models and trying to judge what "fits" the best.

**Other examples to pursue:** U.S. Population data for census years 1790-1990: Students should consider why the population figures for 1940 and 1950 are much lower than an exponential growth model predicts.