Multivariate and Non-linear Regression Models

Non-linear Models - Introduction

- To model non-linear relationships with OLS regression, the data must first be transformed in a way that makes the relationship linear
- All the steps for linear regression may then be performed on the transformed data
- The most common forms of non-linear models are:
  - Logarithmic
  - Exponential
  - Power
Linear Transformations

<table>
<thead>
<tr>
<th>12</th>
<th>Unit Space</th>
<th>Model</th>
<th>Log Space</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Logarithmic</td>
<td></td>
</tr>
<tr>
<td>20</td>
<td></td>
<td>( y = a + b \ln x )</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>( \ln y = \ln a + b x )</td>
<td></td>
</tr>
</tbody>
</table>

Exponential

\( y = ae^{bx} \)

\( \ln y = \ln a + bx \)

Power

\( y = ax^b \)

\( \ln y = \ln a + b \ln x \)

Example: Exponential Model

- We will use the same data as before, but apply an exponential model to it
  - Recall that the data failed the White test for homoscedasticity \((p = 0.047)\)
  - In practice, a Power model (linear when take logs of both sides) might be called for here, but this is shown in Module 7 (Learning Curves), so Exponential (linear when take log of y) is demonstrated

- The next step is to conduct linear regression analysis on the data in semi-log space

- After the analysis is complete, we will transform the parameters of the linear equation back to unit space

Tip: Exponential is rare in practice
Example: Exponential Model

- First we run the regression:
  - Then we check the residual plot:
    - The residual plot is ambiguous; we expand the White test...
    - ...for a formal determination on homoscedasticity
      \[ \ln y = \ln a + bx \]
      \[ y = ae^{bx} \]
  - Finally, we find the parameters for the unit space equation:

\[ a = e^{\ln a} = e^{1.34} = 3.81 \]
\[ b = 0.07 \]
\[ \hat{Y} = 3.81e^{0.07x} \]

Tip: LOGEST() produces same output, however LOGEST coefficients are the exponentials of the LINEST coefficients.

The regression is nearly statistically significant at \( \alpha = 0.10 \) with semi-log space \( R^2 = 0.62 \).

Unit-space data showing exponential trend.

Example: Exponential Model (Expanded White Test)

White Test Conclusions
- Homoscedasticity still rejected at \( \alpha = 0.10 \) (now not rejected at \( \alpha = 0.05 \))
- In practice, could use MLE or Power Model (or the 5% significance level), but we will proceed as if OLS assumptions were validated

Next Steps
- Calculate unit-space goodness of fit statistics for apples-to-apples model comparisons
Unit-Space Goodness of Fit Comparison

<table>
<thead>
<tr>
<th>Statistic</th>
<th>Linear</th>
<th>Exponential</th>
<th>How to Calculate/Comments</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fit Space R^2</td>
<td>0.62</td>
<td>0.62</td>
<td>From LINEST()</td>
</tr>
<tr>
<td>Unit Space R^2</td>
<td>0.62</td>
<td>0.64</td>
<td>1-SSE/SST = 1-SUMSQ(y)/DEVSQ(y) in unit space</td>
</tr>
<tr>
<td>Unit Space Adj R^2</td>
<td>0.50</td>
<td>0.52</td>
<td>1-(((1-R^2)*(n-1))/df)</td>
</tr>
<tr>
<td>Fit Space SEE</td>
<td>2.46</td>
<td>0.30</td>
<td>From LINEST()</td>
</tr>
<tr>
<td>Unit Space SEE</td>
<td>2.46</td>
<td>2.40</td>
<td>SQRT(SSE/DF) in unit space</td>
</tr>
<tr>
<td>Fit Space CV</td>
<td>31%</td>
<td>45%</td>
<td>SEE/y-bar in fit space</td>
</tr>
<tr>
<td>Unit Space CV</td>
<td>31%</td>
<td>30%</td>
<td>SEE/y-bar in unit space</td>
</tr>
</tbody>
</table>

- These differences are not overwhelming, but the routine serves as a reference for comparison of more complicated, multivariate models across types

Warning: It is unusual for a power or exponential model to have better unit space than fit space statistics; generally the unit space conversion causes these stats to worsen.

Minding the Intercept

- One common mistake when performing OLS regression is the omission of a y-intercept in power and exponential models when one exists
  - This has the effect of causing higher than necessary error in the regression
  - Fortunately, it can easily be detected by examining the relation of the trendline to the data
  - It can also be corrected by adding (or subtracting) a constant value to (from) the y-values in the data and examining the change in the trendline (“the simple way”) or by using Solver or other packages (“the elegant way”)

- The example on the next page assumes the data follows a power curve with a non-zero y-intercept

**“To b or not to b” The y-intercept in Cost Estimation, R. L. Coleman, J. R. Summerville, P. J. Braxton, B. L. Cullis, E. R. Druker, SCEA, 2007.**
Minding the Intercept - Example

- The plot to the top-right shows the actual data as well as the data that has been adjusted to take the intercept into account.
- The data was adjusted by subtracting a constant from all y-values until the optimal $R^2$ was achieved.
  - This constant is the best guess for the y-intercept.
  - A "bowing" of the data in relation to the trendline is the symptom that led to the belief in an intercept.
- The plot to the bottom-right shows this same graph in log-space.
- Adding the intercept greatly increases the $R^2$ of the regression.
  - Without intercept: $.9162$
  - With intercept: $1$

Power Curve in Log-Space

\[ y = -0.2467x + 1.4555 \]
\[ R^2 = 0.9162 \]
\[ y = -x - 2 \times 10^{-16} \]
\[ R^2 = 1 \]

Minding the Intercept - Example

- The ANOVA statistics for the two regressions are shown to the right.
- Notice the decrease in standard error and increase in $R^2$.

\[ \begin{array}{l}
\text{Regression Statistics} \\
\text{Multiple R} \quad 0.957159023 \\
\text{R Square} \quad 0.916153395 \\
\text{Adjusted R Square} \quad 0.910164352 \\
\text{Standard Error} \quad 0.060953166 \\
\text{Observations} \quad 16 \\
\end{array} \]

\[ \begin{array}{l}
\text{ANOVA} \\
\text{df} \quad \text{SS} \quad \text{MS} \quad F \quad \text{Significance F} \\
\text{Regression} \quad 1 \quad 0.568333545 \quad 0.568334 \quad 152.9716 \quad 6.33998 \times 10^{-9} \\
\text{Residual} \quad 14 \quad 0.052014039 \quad 0.003715 \\
\text{Total} \quad 15 \quad 0.620347583 \\
\end{array} \]

\[ \begin{array}{l}
\text{Coefficients} \\
\text{Standard Error} \quad t \quad \text{P-value} \\
\text{Intercept} \quad 1.455511028 \quad 0.018553102 \quad 78.45109 \quad 6.5 \times 10^{-20} \\
\text{ln x} \quad -0.246655725 \quad 0.019942786 \quad -12.36817 \quad 6.34 \times 10^{-9} \\
\end{array} \]

\[ \begin{array}{l}
\text{Regression Statistics} \\
\text{Multiple R} \quad 1 \\
\text{R Square} \quad 1 \\
\text{Adjusted R Square} \quad 1 \\
\text{Standard Error} \quad 3.68462 \times 10^{-16} \\
\text{Observations} \quad 16 \\
\end{array} \]

\[ \begin{array}{l}
\text{ANOVA} \\
\text{df} \quad \text{SS} \quad \text{MS} \quad F \quad \text{Significance F} \\
\text{Regression} \quad 1 \quad 9.341591866 \quad 9.341592 \quad 6.88 \times 10^{10} \quad 3.024 \times 10^{-216} \\
\text{Residual} \quad 14 \quad 1.9007 \times 10^{-30} \quad 1.36 \times 10^{-31} \\
\text{Total} \quad 15 \quad 9.341591866 \\
\end{array} \]

\[ \begin{array}{l}
\text{Coefficients} \\
\text{Standard Error} \quad t \quad \text{P-value} \\
\text{Intercept} \quad 0 \quad 1.12154 \times 10^{-16} \quad 0 \quad 1 \\
\text{ln x} \quad -1 \quad 1.20554 \times 10^{-16} \quad -8.3 \times 10^{15} \quad 3 \times 10^{-216} \\
\end{array} \]

Warning: These perfect results are from a toy problem using "cooked" data.

NEW!
Non-linear Model Summary

- The same process performed on the exponential example applies to other non-linear model types
  - The only difference lies in which piece of the data set gets transformed
    - i.e. Logarithmic ↔ take the log of the x data
    - Exponential ↔ take the log of the y data
    - Power ↔ take the log of both x and y
  - Other functions can be used to transform data (e.g., $\sqrt{x}$, sin $x$, etc.) but logarithms are the most common

Tip: Power models are used to analyze learning curves - they are probably the most common use of non-linear regression in cost analysis

Multivariate Regression

- Basics
- ANOVA Revisited
- Adjusted $R^2$
- t and F Summary
Multivariate Regression

- If there is more than one independent variable in linear regression, we call it multivariate regression.
- The general equation is as follows:
  \[ y = a + b_1 x_1 + b_2 x_2 + \ldots + b_k x_k + \epsilon \]
  - So far, we have seen that for one independent variable, the equation forms a line in 2-dimensions.
  - For two independent variables, the equation forms a plane in 3-dimensions.
  - For three or more variables, we are working in higher dimensions which are difficult to display visually in Excel.
- The math is more complicated, but the results can be easily obtained from a regression tool or simple formula (LINEST()) as found in Excel.

In general, the underlying math is similar to the simple model, but matrices are used to represent the coefficients and variables.
- Understanding the math requires background in Linear Algebra.
- Demonstration is beyond the scope of the module, but can be obtained from the references.

Some key points to remember for multivariate regression include:
- Perform residual analysis between each X variable and Y.
- Avoid multicollinearity, i.e., the situation in which high correlation among (2 or more) X variables inflates standard errors and therefore biases significance tests.
- Use the "Goodness of Fit" metrics and significance tests to guide you toward a good model.
Identifying a Multivariate Regression

\[ y = a + b_1x_1 + b_2x_2 + \ldots + b_kx_k + \varepsilon \]

- In general, theory and sound reasoning should be used to determine which variables to include in a multivariate model
  - Choose variables that are correlated with the dependent variable and can be justified; i.e. show correlation and causation
  - It is hard to 'prove' that a model is correctly identified, but with correlation statistics and well developed reasoning, a model can be shown to be robust
- If a relevant variable is omitted, it may cause \( b \) estimates to be biased and will increase SSE ("omitted variable bias")

Coefficients in Multivariate Regression

\[ y = a + b_1x_1 + b_2x_2 + \ldots + b_kx_k \]

- The Excel output gives the predicted coefficients

**Example Multivariate Regression**

\[
\hat{Y} = 1.4 + 0.3 X_1 - 1.1 X_2 - 0.2 X_3
\]

Note: \( \text{LINEST()} \) outputs numbers in gray box. Analyst adds labels and other calculations.
Analysis of Variance (ANOVA)

- **Mean Measures of Variation**
  - Mean Squared Error (or Residual) (MSE):
    \[
    \text{MSE} = \frac{\text{SSE}}{n - k - 1}
    \]
  - Mean of Squares of the Regression (MSR):
    \[
    \text{MSR} = \frac{\text{SSR}}{k}
    \]

The denominator for each of the above is called the degrees of freedom, or \( df \), associated with each type of variation.

Excel Demo: ANOVA

Note: \( df_{\text{resid}} \) is provided. \( df_{\text{reg}} \) must be calculated using \( df_{\text{reg}} = n - df_{\text{resid}} \).
### Adjusted R²

- Adjusted R², or $R^2_a$, adjusts for degrees of freedom
  - Can be used to compare coefficients of determination between models with different numbers of variables including in the same model when a variable is considered for elimination due to lack of significance
  - Can be used as justification for including near-significant variables in models if those variables improve the model's performance

\[
R^2_a = 1 - \left( \frac{SSR}{SST} \right) \left( \frac{n-1}{n-k-1} \right)
\]

Tip: SSR=SSE=SST is true only in OLS. In general, we have $R^2 = 1 - SSE/SST$ but not $R^2_a = SSR/SST$. Note also that $R^2_a$ can be negative.

**Warning:** negative values of $R^2_a$ may occur when fitting non-OLS trends to data

% unexplained penalty (> 1)

### t statistics in Multivariate Regression

- In multivariate regression, a t test is conducted for each coefficient
- The results provide insight as to which variables add the most value to the prediction of cost
  - Adding additional variables will always decrease SSE and increase (unadjusted) $R^2$
  - An insignificant t statistic makes a variable a candidate to be eliminated from the regression (can compare nested vs. full model using SEE, CV, adjusted $R^2$ and F statistics)
    - A variable whose p value is greater than 0.5 should almost certainly be eliminated because it is more likely than not that its (nonzero) coefficient was observed by chance

### Note

Correlation between the independent variables may affect results. High correlation among independent variables is often associated with multicollinearity and should be avoided. A correlation matrix is a good first step to check for multicollinearity. The example is expanded later as an Advanced Topic.

The p-values suggest that $x_1$ is highly significant (as is the intercept, which is generally retained regardless of significance results). The remaining variables are candidates for elimination.

1. There are several methods, such as stepwise regression for determining the best subset of independent variables. See the references for more details.
Calculation of t statistic

- As before, the t statistic for each variable may be calculated as ratio of the estimated coefficient to the corresponding standard error:
  \[ t = \frac{\hat{b}_i}{s_e b_i} \]

- t is also the square root of the partial F statistic, \( F^* \)
  \[ t = \sqrt{F^*} \]

\[ F^* = \frac{SSR_{FullModel} - SSR_{ReducedModel}}{SSE_{FullModel}} = \frac{SS(b_i | a, b_1, \ldots, b_{i-1}, b_{i+1}, \ldots, b_k)}{SSE_{FullModel}} \]

Partial sum of squares - captures the value of adding the variable in question

The F statistic

- An F test is used to determine whether the coefficients of all the independent variables are zero
  - Depends on the ratio of the MSR to the MSE, called an F statistic

\[ y = a + b_1x_1 + b_2x_2 + b_3x_3 \]

Are all coefficients = 0?

<table>
<thead>
<tr>
<th>Coefficients</th>
<th>b1</th>
<th>b2</th>
<th>b3</th>
<th>a</th>
</tr>
</thead>
<tbody>
<tr>
<td>Standard Errors</td>
<td>0.150</td>
<td>1.125</td>
<td>0.251</td>
<td>1.872</td>
</tr>
<tr>
<td>R^2, SE(y)</td>
<td>71.4%</td>
<td>0.170</td>
<td></td>
<td></td>
</tr>
<tr>
<td>F, DF</td>
<td>5.140</td>
<td>11</td>
<td></td>
<td></td>
</tr>
<tr>
<td>SSR, SSE</td>
<td>0.867</td>
<td>0.308</td>
<td></td>
<td></td>
</tr>
<tr>
<td>MSR, MSE</td>
<td>0.269</td>
<td>0.032</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ADJ R^2, SE, CV</td>
<td>63.6%</td>
<td>0.170</td>
<td>16.1%</td>
<td></td>
</tr>
<tr>
<td>T stats</td>
<td>-0.92</td>
<td>-5.14</td>
<td>1.12</td>
<td>8.41</td>
</tr>
<tr>
<td>P values</td>
<td>0.3750</td>
<td>0.0003</td>
<td>0.2977</td>
<td>0.0000</td>
</tr>
<tr>
<td>Significance F</td>
<td>0.0025</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

We conclude the regression is a good model as a whole. Note, the results from the t test should still be addressed.

Example Setup:
Set \( \alpha = 0.05 \)
Hypothesis:
- \( H_0: b_1 = b_2 = b_3 = 0 \)
- \( H_a: \) at least one \( b_i \neq 0 \)

Test Statistic:
- \( F = \frac{MSR}{MSE} = \frac{0.289}{0.032} = 9.140 \)
- P-value: 0.0025
- Decision: We reject \( H_0 \) if the p-value is less than the chosen significance level (0.05)

Since 0.0025 < 0.05
We reject \( H_0 \)
This regression as a whole is statistically significant

We conclude the regression is a good model as a whole.
Note, the results from the t test should still be addressed.
### Paring Down the Multivariate Regression Model

- You may have a model in which some of the coefficients are significant, and some not
  - Note: If only the y-intercept is significant, then it is not really a linear model, it devolves down to a simple average
- If the F statistic is significant, but only some of the t statistics are, then you may be able to achieve a better model by removing the non-significant variables
  - Re-run the model with the least significant variable excluded
  - Compare SEE, CV, F stat, and $R^2_a$ for the two models
  - Continue the above step until all coefficients are significant
  - Compare goodness of fit and significance statistics across all models you’ve seen. Using these, and sound engineering judgment, select the final model.
- Only the y-intercept may be non-significant … in practice, it is used “as is” even if it is not significant. This is important because without the y-intercept, OLS estimators are **not** Best Linear Unbiased Estimators (BLUE).

**Tip:** Given logical relationship, near significance, and explained variation, it may be beneficial to keep non-significant variables in a model. Such variables should only be retained if they improve d.f.- adjusted metrics.

### t and F Summary

- The t statistics tell us if each independent variable is a good predictor
- The F statistic tells us if the regression as a whole is a good model

**Note:** In a regression with one independent variable, the F test and t test will yield the same results

- In our example, the model was found to be significant (large F), but two of the three variables were not (small t)

**Tip:** If possible, test the resulting model on an independent data set.
Selecting the Best Model

Choosing a Model

- We have seen what the linear model is, and explored it in depth
- We have looked briefly at how to generalize the approach to non-linear models
- You may, at this point, have several significant models from regressions
  - One or more linear models, with one or more significant variables
  - One or more non-linear models
- Now we will learn how to choose the “best model”
Steps for Selecting the “Best Model”

- You should already have rejected all non-significant models first
  - If the F statistic is not significant
- You should already have stripped out all non-useful variables and made the model “minimal”
  - Variables that do not incrementally contribute to goodness of fit, overall model significance, (adjusted) variation explained, etc. were already removed
- Select “within type” based on (adjusted) \( R^2 \)
  - When comparing multivariate regression models, select based on adjusted \( R^2 \), which compensates for the number of independent variables
- Select “across type” based on SSE (SEE for multivariate models)

We will examine each in more detail...

Selecting “Within Type”

- Start with only significant, “minimal” models
- In choosing among “models of a similar form”, \( R^2 \) is the criterion
- “Models of a similar form” means that you will compare
  - e.g., linear models with other linear models
  - e.g., power models with other power models

Tip: If a model has a lower \( R^2 \), but has variables that are more useful for decision makers, retain these, and consider using them for CAIV trades and the like
Selecting “Across Type”

- Start with only significant, “minimal” models
- In choosing among “models of a different form”, the SSE in unit space is the criterion (SEE if degrees of freedom change; CV if dependent variables changes)
- “Models of a different form” means that you will compare:
  - e.g., linear models with non-linear models
  - e.g., power models with logarithmic models
- We must compute the SSE by:
  - Computing \( \hat{Y} \) in unit space for each data point
  - Subtracting each \( \hat{Y} \) from its corresponding actual \( Y \) value
  - Sum the squared values, this is the SSE
- An example follows...

**Warning:** We cannot use \( R^2 \) to compare models of different forms because the \( R^2 \) from the regression is computed on the transformed data, and thus is distorted by the transformation.

Selecting “Across Type” Example

- Suppose we want to choose between the following models for a method of estimating cost:
  - Option 1. Power Model
    \( \hat{Y} = 1.14 \times 0.35 \)
    \( R^2 = 0.86 \)
    \( \text{SSE}_1 = \sum (Y - \hat{Y})^2 = 7.2 \)
  - Option 2. Linear Model
    \( \hat{Y} = 0.39 \times 2.1 \)
    \( R^2 = 0.80 \)
    \( \text{SSE}_2 = \sum (Y - \hat{Y})^2 = 8.3 \)

We choose the power model because it has the lower unit-space SSE (SEE if the two had different number of vars.)
Comparing Nested Models Example

- Suppose we want to choose between the following models for a method of estimating cost:

<table>
<thead>
<tr>
<th>Coefficients</th>
<th>b3</th>
<th>b2</th>
<th>b1</th>
<th>a</th>
</tr>
</thead>
<tbody>
<tr>
<td>Standard Errors</td>
<td>0.373</td>
<td>0.211</td>
<td>0.228</td>
<td>0.163</td>
</tr>
<tr>
<td>R^2, SE?(y)</td>
<td>71.4%</td>
<td>0.178</td>
<td></td>
<td></td>
</tr>
<tr>
<td>F, DF</td>
<td>9.340</td>
<td>11</td>
<td></td>
<td></td>
</tr>
<tr>
<td>SSR, SSE</td>
<td>6.867</td>
<td>0.340</td>
<td></td>
<td></td>
</tr>
<tr>
<td>MSR, MSE</td>
<td>0.339</td>
<td>0.067</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Adj R^2, SEE, CV</td>
<td>61.0%</td>
<td>0.178</td>
<td></td>
<td></td>
</tr>
<tr>
<td>T stats</td>
<td>-5.57</td>
<td>-5.34</td>
<td>1.12</td>
<td>8.41</td>
</tr>
<tr>
<td>P values</td>
<td>0.0005</td>
<td>0.0003</td>
<td>0.2677</td>
<td>0.0006</td>
</tr>
<tr>
<td>Significance F</td>
<td>0.0005</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

- Reduced model eliminates the least significant variable (b3). We can see by removing the least significant variable R^2, SEE, CV, significance F and adjusted R^2 all improve when b3 is removed. A (possible) next step would be to also eliminate b1 and compare again.

Regression Summary

- Regression analysis is a powerful tool in cost analysis, particularly for developing CERs
- Two of the most important results of OLS Regression are:
  - Statistical significance
  - Uncertainty
- This module has covered:
  - The basic math behind the analysis
  - How to interpret the results from a regression tool such as Excel
  - How to apply the results and choose among models
- Many other regression techniques extend beyond the scope of this module, but can be found in the resources provided
# Resources - Textbooks

- *Applied Linear Regression Models*, Neter et al., Irwin Inc., 1996

---

# Resources - Papers

- “The Multicollinearity Problem: Coping with the Persistent Beast” Kevin Cincotta, David Lee, LMI, February 2007
- “The Minimum-Unbiased-Percentage Error (MUPE) Method In CER Development” Shu-Ping Hu, DoDCAS, 2001
- “Why ZMPE When You Can MUPE?” Dr. Shu-Ping Hu, Alfred Smith, SCEA/ISPA, 2007
Related and Advanced Topics
- Geometric Interpretations
- Derivation of Formulae
- White Test
- ANOVA Redux
- The Bivariate Normal Distribution and the Geometry of Regression
- Correction Factors
- Multicollinearity
- Non-OLS Models
- Maximum Likelihood Estimation

Geometric Interpretations
- Means = “center of gravity”

\[ \bar{Y} = 83.33 \]

\[ \bar{X} = 23.89 \]
Geometric Interpretations

- Deviations from mean sum to zero

\[ Y_i - \bar{Y} \]

\[ X_i - \bar{X} \]

\[ \sum_{i=1}^{n} (X_i - \bar{X})(Y_i - \bar{Y}) = \sum_{i=1}^{n} X_iY_i - n\bar{X}\bar{Y} \]

\[ \bar{Y} = 83.33 \]

\[ \bar{X} = 23.89 \]
Geometric Interpretations

- Deviations from mean sum to zero

Deriving the Equations

\[
SSE(a, b) = \sum_{i=1}^{n} (\hat{Y}_i - Y_i)^2 = \sum_{i=1}^{n} (a + bX_i - Y_i)^2
\]

\[
\frac{\partial SSE}{\partial a} = 2 \sum_{i=1}^{n} (a + bX_i - Y_i) = 0 \implies \sum_{i=1}^{n} Y_i = an + b \sum_{i=1}^{n} X_i
\]

\[
\frac{\partial SSE}{\partial b} = 2 \sum_{i=1}^{n} X_i (a + bX_i - Y_i) = 0
\]

\[
a \sum_{i=1}^{n} X_i + b \sum_{i=1}^{n} X_i^2 - \sum_{i=1}^{n} X_i Y_i = 0
\]
Deriving the Equations (cont’d.)

\[ \bar{y} = a + b \bar{x} \Rightarrow a = \bar{y} - b \bar{x} \]

\[ (\bar{y} - b \bar{x}) \sum_{i=1}^{n} x_i + b \sum_{i=1}^{n} x_i^2 - \sum_{i=1}^{n} x_i y_i = 0 \]

\[ b \left( \sum_{i=1}^{n} x_i^2 - \bar{x} \sum_{i=1}^{n} x_i \right) = \sum_{i=1}^{n} x_i y_i - \bar{y} \sum_{i=1}^{n} x_i \]

\[ b = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \]

White Test

- Perform the regression as usual to generate squared errors (\( \varepsilon^2 \))
- Regress \( \varepsilon^2 \) on each regressor, squared regressor, pairwise crossproduct, and an intercept
  - For 1 x: Regress on intercept, \( x, x^2 \)
  - For 2 x’s: Regress on intercept, \( x_1, x_2, x_1^2, x_2^2 \), and \( x_1 x_2 \)
  - For 3 x’s: Regress on intercept, \( x_1, x_2, x_3, x_1^2, x_2^2, x_3^2, x_1 x_2, x_1 x_3, \) and \( x_2 x_3 \)
  - For \( k \) x’s: \( m+1 = C(k+2,2) = (k+2)(k+1)/2 \) (including intercept)
- Calculate the \( R^2 \) from the auxiliary regression
- White statistic = \( nR^2 \) follows a chi square distribution with \( (m-1) \) degrees of freedom where \( m = \) number of estimated parameters (not including intercept) from auxiliary regression
- Reject the null hypothesis of homoscedasticity and conclude that OLS cannot be used if p-value is less than a specified critical value \( \alpha \) (say, 0.10)
Sums of Squares Shortcuts

- These formulae are more computationally efficient:
  - Total Sum of Squares (SST): \[ \sum_{i=1}^{n} Y_i^2 - \bar{Y} \sum_{i=1}^{n} Y_i \]
  - Residual or Error Sum of Squares (SSE): \[ \sum_{i=1}^{n} Y_i^2 - a \sum_{i=1}^{n} Y_i - b \sum_{i=1}^{n} X_i Y_i \]
  - Regression Sum of Squares (SSR): \[ b \left( \sum_{i=1}^{n} X_i Y_i - \bar{X} \sum_{i=1}^{n} Y_i \right) \]

- Can you verify the identity using these?

\[ \text{SST} = \text{SSE} + \text{SSR} \]
R² and Reducing CV

- We said that one of the goals of running CERs is to reduce CV, and that R² is the percent explained variation.
  - But how are the two related?
  - We can show that the reduction of CV is a function of R²

\[
CV_{\text{old}} = \frac{s_y}{\bar{Y}} = \frac{1}{\bar{Y}} \sqrt{\frac{\text{SST}}{n-1}} \quad CV_{\text{new}} = \frac{\text{SEE}}{\bar{Y}} = \frac{1}{\bar{Y}} \sqrt{\frac{\text{SSE}}{n-k}} \quad R^2 = \frac{\text{SSR}}{\text{SST}} = 1 - \frac{\text{SSE}}{\text{SST}}
\]

\[
CV_{\text{redux}} = \frac{CV_{\text{new}}}{CV_{\text{old}}} = \sqrt{\frac{n-1}{n-k}} \left( \frac{\text{SSE}}{\text{SST}} \right) = \frac{n-1}{n-k} \sqrt{1 - R^2} \approx \sqrt{1 - R^2}
\]

\[
CV_{\text{redux}} = \frac{\text{SEE}}{s_y} = \frac{2.46}{3.36} = \sqrt{\frac{5-1}{5-2}} \sqrt{1 - 0.62} = 0.71 = 30.8\% \quad \frac{43.3\%}{100\%}
\]

Zero-Intercept R²

Warning: When the regression line is forced through the origin (0), R² and R², in the trendline can be different than in the LINEST or macro output.

\[
R^2 = 1 - \frac{\text{SEE}}{\text{SST}} \quad \text{is preferred to} \quad R^2 = \frac{\text{SSR}}{\text{SST}}, \quad \text{because}
\]

\[
R^2 = \frac{\text{SSR}}{\text{SST}} \quad \text{can only be used when} \quad \text{SSR} + \text{SSE} = \text{SST}
\]

Only the case for OLS regression in fit space.
Leverage

- Leverage is a measure of how far an observation is from the average values of all the independent variables in the equation.

\[
\hat{Y} \pm t_{\alpha/2, df} \times \text{SEE} \approx \frac{1}{n} \left( \frac{(X - \overline{X})^2}{n \sum X^2 - n \overline{X}^2} \right)^{1/2} = \text{Leverage}
\]

An observation is considered a potential outlier with respect to X if its leverage value is greater than \(2(p/n)\), where \(p\) is the number of parameters and \(n\) is the number of observations.

Regression Distributions

Univariate
- Cost data
- Sums of squares
- Chi square
- Variance
- Std dev

Bivariate
- Slope
- Std error
- Data points
- Variance
- Std dev
- Chi square
- SSR
- MSR
- SSE
- F(df1, df2)

The assumption of normality drives all the other distributions.
The Geometry of Regression

• The following charts show the geometry of regression by building up a picture
  - The picture provides a mental image that aids in understanding the regression equation
  - This visual framework has potential applications in risk analysis

• The below facts enable us to derive the picture
  - For any two jointly distributed variables, there is a regression line
    • The slope is: \( b = \rho \left( \frac{\sigma_y}{\sigma_x} \right) \)
    • The y intercept is: \( a = \mu_y - \rho \left( \frac{\sigma_y}{\sigma_x} \right) * \mu_x \)
  - If the variables are joint bivariate normal, then \( \rho \) is the correlation coefficient

Let's look at the graph...

The Geometry of Bivariate Normality

and the implications for Regression

First construct a box 2\( \sigma \) by 2\( \sigma \) centered at the means
**The Geometry of Bivariate Normality**

*and the implications for Regression*

- **Intercept** $a$ varies with slope, $\mu_x$, and $\mu_y$.
- **Slope** $m$ varies with $\rho$, $\sigma_x$, $\sigma_y$.
- Dispersion varies with $\rho$.

**Equation:**

$$y = \rho(\sigma_y / \sigma_x) (x - \mu_y) + \mu_y$$

- When $\rho = 1$, the line has the "unseen" slope.
- The slope that would be true if $\rho = 1$.
- When $\rho = 0$, the line has a different slope.
- When $\rho = -1$, the line is flipped.

- **Range of intercepts**
- **Range of slopes**

**The Geometry of Bivariate Normality**

*and the implications for Regression*

- **Intercept** $a$ varies with slope, $\mu_x$, and $\mu_y$.
- **Slope** $m$ varies with $\rho$, $\sigma_x$, $\sigma_y$.
- Dispersion varies with $\rho$.

**Equation:**

$$y = \rho(\sigma_y / \sigma_x) (x - \mu_y) + \mu_y$$

- When $\rho = 1$, the line has the "unseen" slope.
- The slope that would be true if $\rho = 1$.
- When $\rho = 0.75$, the line has a different slope.
- When $\rho = -0.75$, the line is flipped.

- **Range of intercepts**
- **Range of slopes**

ICEAA 2014 Professional Development & Training Workshop
Correction Factors

- When converting CERs developed in log-linear space to not log-linear space, the CER will predict closer to the median than the mean

Goldberger Factor \( (GF) \approx e^{\left(1-r_0\right)\frac{s^2}{2}} \)

PING Factor \( (PF) \approx e^{\left(1-\frac{p}{n}\right)\frac{s^2}{2}} \)

- \( s \) = standard error of the estimate
- \( r_0 \) = leverage value in log space if \( x_0 \) is a vector of independent variables in the data matrix
- \( P \) = # of estimated coefficients
- \( n \) = sample size
- \( s \) = standard error of the estimate

Multicollinearity

- Multicollinearity occurs when there is a strong linear relationship among two or more independent variables
  - The model form of this linear relationship must match the model form of the regression in order for multicollinearity to occur
- Some symptoms that multicollinearity may be occurring are:
  - Large changes in the values of the regression coefficients when another variable is added or deleted
  - Regression coefficients having an opposite sign from what intuition predicts
  - Two (independent) variables thought to be similar have large (in absolute value) but “opposite” signs
  - Variables expected to be significant are not
  - A high overall \( R^2 \) with several non-significant independent variables
  - High Variance Inflation Factors (VIFs) or Variance Amplification Factors (VAFs)
- The existence of Multicollinearity has a couple of adverse effects on the results of OLS:
  - Biases coefficients and inflates their standard errors
    - This in turn biases t-tests and p-values; also makes them imprecise
  - Makes it difficult to understand the effect each independent variable has on predicting the outcome
    - It is important to note that multicollinearity does not affect the reliability of the model predictions; it simply biases individual coefficient values and their estimated significance.
Multicollinearity

- One of the most robust methods to address multicollinearity is to find the VIF of each independent variable.
- The VIF of an independent variable is defined as \( \frac{1}{1 - R_{y,s}^2} \) where \( R_{y,s}^2 \) is the coefficient of determination when the dependent variable \( y \) is regressed against all other dependent variables.
  - The VIF of \( \beta \) is the multiplicative factor (\( \geq 1 \)) by which the variance of \( \beta \) is increased due to correlation among the regressors.
  - All information needed to compute the VIF for each variable is found as part of Excel’s Linest() function.
- In general, a VIF of over 10 indicates a severe enough problem to take a second look. If no VIF exceeds 4, you may reasonably conclude that there is no issue with multicollinearity.
- If multicollinearity is clouding the results of a regression model, consider removing the variable with the largest VIF and re-running the model, understanding that the variable to be removed may be the intercept.
- Continue until no VIF exceeds 4 (ideally) or 10 (if desperate).

Note: VIFs may be calculated in this manner in standard OLS regression. For zero-intercept regression, \( R_{y,s}^2 \) can’t be used because it assumes a constant term. However, the VIF can also be calculated as \( \frac{\text{SE}_{\beta_j}^2}{\text{SE}_{\beta_j}^2_{\text{native}}} \) where \( \text{SE}_{\beta_j}^2_{\text{native}} = \frac{\text{SSE}}{(n-1)\text{Var}(X_j)} \).

Multicollinearity - Example

<table>
<thead>
<tr>
<th>Regression Details</th>
<th>( R^2 )</th>
<th>Adjusted ( R^2 )</th>
<th>Standard Error</th>
<th>Observations</th>
</tr>
</thead>
<tbody>
<tr>
<td>Before removing multicollinearity</td>
<td>( 0.956 )</td>
<td>( 0.934 )</td>
<td>( 3.4 )</td>
<td>( 10 )</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>ANOVA</th>
<th>df</th>
<th>SS</th>
<th>MS</th>
<th>F</th>
<th>Significance F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Regression</td>
<td>3</td>
<td>1488.03</td>
<td>496.01</td>
<td>43.14</td>
<td>( 0.0001 )</td>
</tr>
<tr>
<td>Residual</td>
<td>6</td>
<td>68.99</td>
<td>11.49</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>9</td>
<td>1557.02</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Coefficients</th>
<th>Standard Error</th>
<th>( t ) Stat</th>
<th>P-value</th>
<th>VIF</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>1.62</td>
<td>2.70</td>
<td>0.60</td>
<td>0.57</td>
</tr>
<tr>
<td>( x_1 )</td>
<td>3.65</td>
<td>3.06</td>
<td>1.20</td>
<td>0.28</td>
</tr>
<tr>
<td>( x_2 )</td>
<td>-0.86</td>
<td>3.23</td>
<td>-0.27</td>
<td>0.79</td>
</tr>
<tr>
<td>( x_3 )</td>
<td>2.15</td>
<td>0.31</td>
<td>7.01</td>
<td>0.00</td>
</tr>
</tbody>
</table>

\( \text{VIF}_{\text{native}} = \frac{\text{SE}_{\beta_j}^2}{\text{SE}_{\beta_j}^2_{\text{native}}} \).

After removing multicollinearity

<table>
<thead>
<tr>
<th>Regression Details</th>
<th>( R^2 )</th>
<th>Adjusted ( R^2 )</th>
<th>Standard Error</th>
<th>Observations</th>
</tr>
</thead>
<tbody>
<tr>
<td>Before removing multicollinearity</td>
<td>( 0.985 )</td>
<td>( 0.942 )</td>
<td>( 3.2 )</td>
<td>( 10 )</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>ANOVA</th>
<th>df</th>
<th>SS</th>
<th>MS</th>
<th>F</th>
<th>Significance F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Regression</td>
<td>2</td>
<td>1487.22</td>
<td>743.61</td>
<td>74.58</td>
<td>( 1.90 \times 10^{-5} )</td>
</tr>
<tr>
<td>Residual</td>
<td>7</td>
<td>69.80</td>
<td>9.97</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>9</td>
<td>1557.02</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Coefficients</th>
<th>Standard Error</th>
<th>( t ) Stat</th>
<th>P-value</th>
<th>VIF</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>1.49</td>
<td>2.48</td>
<td>0.60</td>
<td>0.57</td>
</tr>
<tr>
<td>( x_1 )</td>
<td>2.84</td>
<td>0.35</td>
<td>8.02</td>
<td>( 8.94 \times 10^{-5} )</td>
</tr>
<tr>
<td>( x_3 )</td>
<td>2.15</td>
<td>0.29</td>
<td>7.54</td>
<td>( 0.00 )</td>
</tr>
</tbody>
</table>

\( \text{VIF}_{\text{native}} = \frac{\text{SE}_{\beta_j}^2}{\text{SE}_{\beta_j}^2_{\text{native}}} \).

Notice neither independent variable is significant.

- The VIFs for \( x_1 \) and \( x_2 \) are high, indicating that multicollinearity is present.
- This is further verified by scatter plotting them together (see below).
- By re-running the regression with each of the variables removed, the best regression is found.
Multicollinearity - Ridge Regression

- Ridge regression is one of several ways of regularizing a regression model so that all the independent variables may remain in the analysis
  - Should mainly be used when there is extremely high correlation between the independent variables
- In ridge regression, a ridge variable is added to the SSE expression, such that solutions with inflated $SE_{\beta}$ values are no longer optimal
- Advantages of Ridge Regression
  - Reduces the standard errors of the estimated coefficients
  - No independent variable is removed from the analysis
- Disadvantages of Ridge Regression
  - Model estimates will be biased!
  - Coefficients lose some of their interpretability
  - Set the ridge too high, and estimates are biased beyond recognition (recall: multicollinearity does not bias overall model estimates). Set it too low, and the multicollinearity problem is not remedied.

Warning: Ridge regression is trial-and-error intensive!

A Very Brief Overview of Two Non-OLS Regression Techniques

Weighted Least Squares Regression
Multiplicative Error Regression

*These slides are meant as a top-level overview of these techniques, not an instruction guide. For more detailed information, seek the resources provided in the Resources section.
Non-OLS Regression Techniques

Weighted Least Squares

- Weighted Least Squares regression is similar to Ordinary Least Squares in that it still works by minimizing the sum squared error
  - The difference is instead of treating the errors associated with the data points equally, certain points are weighted
  - \[ \sum_{i=1}^{n} (y_i - f(x_i))^2 \text{ vs. } \sum_{i=1}^{n} w_i(y_i - f(x_i))^2 \]
  - One common method is setting \( w_i = \frac{1}{\sigma_i^2} \)
    - Thus giving higher weight to points with lower variance in measurement of the x’s. When (and only when) this weighting convention is used, WLS estimators are BLUE.

- WLS regression is useful in many cases
  - To compensate for a violation of the homoscedasticity assumption of OLS (funnel-shaped residual plots)
  - When certain data points are believed to be more correct or applicable than other data points

Non-OLS Regression Techniques

Multiplicative Error Regression

- OLS seeks to minimize the additive error of the regression
  - \( Y = ax + \epsilon \)
- However, non-linear functions may exhibit multiplicative error instead
  - \( Y = (ax)^\epsilon \)
- When this is the case, multiplicative error techniques must be used
  - Examples include MUPE and ZMPE
  - Papers on these techniques are listed in a special section on the resources slide
- When prediction intervals around OLS-transformed regressions are produced
  - they demonstrate a multiplicative error pattern as well
Non-OLS Regression Techniques: Maximum Likelihood Estimation

- If you believe the function to be linear but still have issues with heteroscedasticity, maximum likelihood estimation (MLE) may be in order
- This *generalization of OLS* accounts for non-constant error term variances
- MLE solutions reduce to OLS solutions when variance is held constant, so MLE estimators are OLS estimators when OLS assumptions hold
- MLE estimators are asymptotically BLUE for large data sets; even with heteroskedasticity, as long as other OLS assumptions hold (e.g. zero-mean, normal i.i.d. error term)
- If $s$ is constant, log likelihood objective function $= \sum \{ \ln(1/\sigma) - \varepsilon^2/2\sigma^2 \}$ is *maximized* when SSE is minimized
- Must specify $\sigma^2$ as a function of $x$
- Other remedies for heteroscedasticity include generalized least squares (GLS, a generalization of WLS) and transformation to log space