Diagnostics and Remedial Measures: An Overview

- Residuals
- Model diagnostics
  - Graphical techniques
  - Hypothesis testing
- Remedial measures
  - Transformation
- Later: more about all this for multiple regression
Model Assumptions

Recall simple linear regression model.

\[ Y_i = \beta_0 + \beta_1 X_i + \epsilon_i, \quad \epsilon_i \sim \text{iid } \mathcal{N}(0, \sigma^2), \]

for \( i = 1, \ldots, n \).

- A linear-line relationship between \( E(Y) \) and \( X \):

- Homogeneous variance:

- Independence:

- Normal distribution:
Ramifications If Assumptions Violated

Recall simple linear regression model.

- **Nonlinearity**
  - Linear model will fit poorly
  - Parameter estimates may be meaningless

- **Non-independence**
  - Parameter estimates are still unbiased
  - Standard errors are a problem and thus so is inference

- **Nonconstant variance**
  - Parameter estimates are still unbiased
  - Standard errors are a problem

- **Non-normality**
  - Least important, why?
  - Inference is fairly robust to non-normality
  - Important effects on prediction intervals
Model Diagnostics

- Reliable inference hinges on reasonable adherence to model assumptions
- Hence it is important to evaluate the **FOUR** model assumptions, that is, to perform model “diagnostics”.
- The main approach to model diagnostics is to examine the residuals (thanks to the additive model assumption)
- Consider two approaches.
  - Graphical techniques: More subjective but quick and very informative for an expert.
  - Hypothesis tests: More objective and comfortable for amateurs, but outcomes depend on assumptions, sensitivity. Tendency to use as a crutch.
Graphical Techniques

- At this point in the analysis, you have already done EDA.
  - 1D exploration of $X$ and $Y$.
  - 2D exploration of $X$ and $Y$.
  - Not very effective for model diagnostics except in drastic cases

- Recall the definition of residual

\[ e_i = Y_i - \hat{Y}_i, \text{ where } i = 1, \ldots, n \]

- $e_i$ can be treated as an estimate of the true error

\[ e_i = Y_i - E(Y_i) \sim \text{iid } N(0, \sigma^2) \]

- $e_i$ can be used to check normality, homoscedasticity, linearity, and independence.
Properties of Residuals

- Mean:
  \[ \bar{e} = \]

- Variance:
  \[ MSE = \frac{SSE}{n - 2} = \frac{\sum_{i=1}^{n} e_i^2}{n - 2} = \frac{\sum_{i=1}^{n} (e_i - \bar{e})^2}{n - 2} = s^2. \]

- Nonindependence:

  When the sample size \( n \) is large, however, residuals can be treated as independent.
Standardized Residuals

- For diagnostics there are superior choices to the ‘ordinary residuals’
  - Standardized (KNNL: ‘semi-studentized’) residuals:
    \[ \text{Var}(\epsilon_i) = \sigma^2 \]
    therefore is it is natural to apply the standardization
    \[ e_i^* = \]

- But each \( e_i \) has a different variance.
  - Use this fact to derive superior type of residuals below
Hat Values

\[ \hat{Y}_i = \]

The \( h_{ij} \) are called hat values.
Deriving the Variance of Residuals

Using \( \hat{Y}_i = \sum_j h_{ij} Y_j \) we obtain

\[ e_i = \]

Therefore (since the Y’s are independent)

\[ Var\{e_i\} = \]
Continuing to Derive Variance of Residuals

Using \( \text{Var}\{e_i\} = \sigma^2 \left[ (1 - h_{ii})^2 + \sum_{j \neq i} h_{ij}^2 \right] \), we have

\[
\sum_j h_{ij}^2 = h_{ii}
\]

(show it in HW.) Finally,

\[
\begin{align*}
\text{Var}\{e_i\} &= \sigma^2 \left( (1 - h_{ii})^2 + \sum_{j \neq i} h_{ij}^2 \right) \\
&= \sigma^2 \left( 1 - 2h_{ii} + h_{ii}^2 + \sum_{j \neq i} h_{ij}^2 \right) \\
&= \sigma^2 \left( 1 - 2h_{ii} + \sum_j h_{ij}^2 \right) \\
&= \sigma^2 (1 - 2h_{ii} + h_{ii}) \\
&= \sigma^2 (1 - h_{ii})
\end{align*}
\]
Studentized Residuals

- Now we may scale each residual separately by its own standard deviation
- The (internally) studentized residual is

\[ r_i = \frac{e_i}{\sqrt{MSE(1 - h_{ii})}} \]

- There is still a problem: Imagine that \( Y_i \) is a severe outlier
  - \( Y_i \) will strongly ‘pull’ the regression line toward it
  - \( e_i \) will understate the distance between \( Y_i \) and the ‘true’ regression line
- The solution is to use ‘externally studentized residuals’…
Studentized Residuals

- Now we may scale each residual separately by its own standard deviation.
- The (internally) studentized residual is

\[ r_i = \frac{e_i}{\sqrt{MSE(1 - h_{ii})}} \]

- There is still a problem: Imagine that \( Y_i \) is a severe outlier.
  - \( Y_i \) will strongly ‘pull’ the regression line toward it.
  - \( e_i \) will understate the distance between \( Y_i \) and the ‘true’ regression line.
- The solution is to use ‘externally studentized residuals’...
Studentized Deleted Residuals

- To eliminate the influence of $Y_i$ on the misfit at the $i$th point, fit the regression line based on all points except the $i$th.
- Define the prediction at $X_i$ using this deleted regression as $\hat{Y}_{i(i)}$.
- The ‘deleted residual’ is $d_i = Y_i - \hat{Y}_{i(i)}$.
- The studentized deleted residual is
  
  $$t_i = d_i / \hat{s}\{d_i\} = \frac{Y_i - \hat{Y}_{i(i)}}{\sqrt{MSE(i)}/(1 - h_{ii})}$$

- No need to fit $n$ deleted regressions, we can show that
  
  $$d_i = e_i / (1 - h_{ii})$$

  $$(n - 2)MSE = (n - 3)MSE(i) + e_i^2 / (1 - h_{ii})$$

- Also, $t_i$ has a t-distribution: $t_i \sim t_{n-3}$
Residual Plots

Residual plot is a primary graphic diagnostic method.

- Departures from model assumptions can be difficult to detect directly from \( X \) and \( Y \).
- Use the externally standardized residuals

Some key residual plots:

- Plot \( t_i \) against predicted values \( \hat{Y}_i \) (Not \( Y_i \))
  - detect nonconstant variance
  - detect nonlinearity
  - detect outliers
- Plot \( t_i \) against \( X_i \).
  - In simple linear regression this is same as above (Why?)
  - In multiple regression will be useful to detect partial correlation
- Plot \( t_i \) versus other possible predictors (e.g., time)
  - Detect important lurking variable
- Plot \( t_i \) versus lagged residuals
  - Detect correlated errors
- QQ-plot or normal probability (PP-) plot of \( t_i \).
  - Detect non-normality
Nonlinearity of Regression Function

- Plot $t_i$ against $\hat{Y}_i$ (and $X_i$ for multiple linear regressions).
  - Random scatter indicates no serious departure from linearity.
  - Banana indicates departure from linearity.
  - Could fit nonparametric smoother to residual plot to aid detection

- Example: Curved relationship (KNNL Figure 3.4(a)).

- Plotting $Y$ vs. $X$ is not nearly as effective for detecting nonlinearity because trend has not been removed
  - Logically, you are investigating model assumptions not “marginal effect”.
Nonlinearity of Regression Function

- Plot $t_i$ against $\hat{Y}_i$ (and $X_i$ for multiple linear regressions).
  - Random scatter indicates no serious departure from linearity.
  - Banana indicates departure from linearity.
  - Could fit nonparametric smoother to residual plot to aid detection.

- Example: Curved relationship (KNNL Figure 3.4(a)).

- Plotting $Y$ vs. $X$ is not nearly as effective for detecting nonlinearity because trend has not been removed.
  - Logically, you are investigating model assumptions not “marginal effect”.

W. Zhou (Colorado State University)
Nonconstant Error Variance

- Plot $t_i$ against $\hat{Y}_i$ (and $X_i$ for multiple linear regressions).
  - Random scatter indicates no serious departure from constant variance.
  - Could fit nonparametric smoother to this plot to aid detection
- Funnel indicates non-constant variance.
- Example: KNNL Figure 3.4(c).
- Often both nonconstant variance and nonlinearity exist.
Nonindependence of Error Terms

- Possible causes of nonindependence.
  - Observations collected over time and/or across space.
  - Study done on sets of siblings.
- Departure from independence. For example,
  - Trend effect (KNNL Figure 3.4(d), 3.8(a)).
  - Cyclical nonindependence (KNNL Figure 3.8(b)).
- Plot $t_i$ against other covariate, such as time.
- Autocorrelation function plot (acf())
Nonnormality of Error Terms

- Box plot, histogram, stem-and-leaf plot of $t_i$.
- QQ (quantile-quantile) plot.
  1. Order the residuals: $t_{(1)} \leq t_{(2)} \leq \cdots \leq t_{(n)}$.
  2. Find the corresponding "rankits": $z_{(1)} \leq z_{(2)} \leq \cdots \leq z_{(n)}$, where for $k = 1, \ldots, n$,

$$z_{(k)} = \sqrt{MSE} \times z \left( \frac{k - 0.375}{n + 0.25} \right)$$

is an approximation of the expected value of the $k$th smallest observation in a normal random sample.
  3. Plot $t_{(k)}$ against $z_{(k)}$.

- QQ plot should be approximately linear if normality holds
  - ‘S’ shape means distribution of residuals has light (‘short’) tails
  - Backwards ‘S’ means heavy tails
  - ‘C’ or backwards ‘C’ means skew

- It is a good idea to examine other possible problems first.
Presence of Outliers

- An outlier refers to an extreme observation.

- Some diagnostic methods
  - Box plot of $t_i$.
  - Plot $t_i$ against $\hat{Y}_i$ (and $X_i$).
  - $t_i$ which are very unlikely compared to the reference t-distribution could be called outliers
  - Modern cluster analysis methods

- Outliers may convey important information.
  - An error.
  - A different mechanism is at work.
  - A significant discovery.

- Temptation to throw away outliers because they may strongly influence parameter estimates.
  - Doesn’t mean that the model is right and the data point is wrong
  - The data point is right and the model is wrong
Graphical Techniques: Remarks

- We generally do not plot residuals \((t_i)\) against response \((Y_i)\). Why?
- Residual plots may provide evidence against model assumptions, but do not generally validate assumptions.
- For data analysis in practice:
  - Fit model and check model assumptions (an iterative process).
  - Generally do not include residual plots in a report, but include a sentence or two such as “Standard diagnostics did not indicate any violations of the assumptions for this model.”
- For this class, always include residual plots for homework assignments so you can learn the methods
- No magic formulas.
- Decision may be difficult for small sample size.
- As much art as science.
Diagnostic Methods Based on Hypothesis Testing

- Tests for linearity: $F$ test for lack of fit (Section 3.7).
- Tests for constancy of variance (Section 3.6):
  - Brown-Forsythe test.
  - Breusch-Pagan test.
  - Levene’s test.
  - Bartlett’s test.
- Tests for independence (Chapter 12):
  - Runs test.
  - Durbin-Watson test.
- Tests for normality (Section 3.5):
  - $\chi^2$ test.
  - Kolmogorov-Smirnov test.
- Tests for outliers (Chapter 10).
Residual plots can be used to assess the adequacy of a simple linear regression model. A more formal procedure is a test for lack of fit using “pure error”.

Need ‘repeat groups’

For a given data set, suppose we have fitted a simple linear regression model and computed regression error sum of squares

\[ SSE = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2. \]

These deviations \( Y_i - \hat{Y}_i \) could be due to either random fluctuations around the linear line or an inadequate model.
Pure Error and Lack of Fit

- The main idea is to take several observations on $Y$ for the same $X$, independently, to distinguish the error due to random fluctuations around the linear line and the error due to lack of fit of the simple linear regression model.
- The variation among the repeated measurements is called “pure error”.
- The remaining error variation is called “lack of fit”.
- Thus we can partition the regression SSE into two parts:
  \[\text{SSE} = \text{SSPE} + \text{SSLF}\]
  where SSPE = SS Pure Error and SSLF = SS Lack of Fit.
- Actually, we are comparing a “Linear Function” with a “Simple function.”
Pure Error and Lack of Fit

- One possibility is that pure error is comparatively large and the linear model seems adequate. That is, pure error is a large part of the SSE.

- The other possibility is that pure error is comparatively small and linear model seems inadequate. That is, pure error is a small part of the regression error and error due to lack of fit is then a large part of the SSE.

- If the latter case holds, there may be significant evidence of lack of fit.
Notation

- Models (R notation):
  - Null (N): $Y \sim 1$, common mean model
  - Linear regression is Reduced (R): $Y \sim X$, regression model
  - ANOVA is Full (F): $Y \sim \text{factor}(X)$, separate mean model

- Notation:
  - $Y_{ij}$ are the data, where $j$ indexes groups and $i$ indexes individuals. (Sums will be taken over all available indices).
  - $\bar{Y}$ is the grand mean
  - $\bar{Y}_j$ is the $j$th group mean
  - $\hat{Y}_{ij}$ are the fitted values using the regression line.
  - Note that $\bar{Y}_j$ are the fitted values under the ANOVA model that fits group means, $Y \sim \text{factor}(X)$
Sums of Squares

Recall: All sums are over both \( i \) and \( j \) except as noted.

- \( SSTO = \sum(Y_{ij} - \bar{Y})^2 \)
- \( SSR_R = \sum(\hat{Y}_{ij} - \bar{Y})^2 \)
- \( SSE_R = \sum(Y_{ij} - \hat{Y}_{ij})^2 \)
- \( SSTO = SSR_R + SSE_R \)
- \( SSPE = SSE_F = \sum(Y_{ij} - \bar{Y}_j)^2 \)
- \( SSLF = \sum(\bar{Y}_j - \hat{Y}_{ij})^2 = \sum_j n_j (\bar{Y}_j - \hat{Y}_{ij})^2 \)
- \( SSE_R = SSPE + SSLF \)
## LOF ANOVA Table

One way to summarize the LOF test is by ANOVA:

<table>
<thead>
<tr>
<th>Source</th>
<th>df</th>
<th>SS</th>
<th>MS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Regression</td>
<td>1</td>
<td>SSR</td>
<td>SSR/1</td>
</tr>
<tr>
<td>Lack of Fit</td>
<td>$r - 2$</td>
<td>SSLF</td>
<td>$MS_{LF} = SSLF / (r - 2)$</td>
</tr>
<tr>
<td>Pure Error</td>
<td>$n - r$</td>
<td>SSPE</td>
<td>$MS_{PE} = SSPE / (n - r)$</td>
</tr>
<tr>
<td>Total</td>
<td>$n - 1$</td>
<td>SSTO</td>
<td></td>
</tr>
</tbody>
</table>

- $E(MS_{PE}) = \sigma^2$ and $E(MS_{LF}) = \sigma^2 + \frac{\sum_{i=1}^{r} n_i (\mu_i - (\beta_0 + \beta_1 x_i))^2}{r - 2}$
- F-test for lack of fit is therefore:
LOF as model comparison

- In fact, the above lack of fit test is doing model comparison
  - our desired model \( Y \sim X \) to the potentially better model \( Y \sim \text{factor}(X) \)
    which would be required if the linear model fit poorly.

- Apply the GLT to compare these two models (are they nested?)

\[
F_{LOF} = \frac{SSE_R - SSE_F}{df_R - df_F} / \frac{SSE_F}{df_F}
\]

- Notice \( SSE_R - SSE_F = SSE_R - SSPE = SSLF \) and \( SSE_F = SSPE \) so
\[
F_{LOF} = MSLF/MSPE
\]
and LOF ANOVA F-test is same as model comparison by \( F_{LOF} \).
Lack of Fit in R

\[
\text{anova}(\text{reduced.lm})
\]

\[
\begin{array}{lcccc}
\text{Df} & \text{Sum Sq} & \text{Mean Sq} & \text{F value} & \text{Pr(>F)} \\
\hline
\text{x} & 1 & 60.95 & 60.950 & 193.07 & 1.395e-09 \\
\text{Residuals} & 14 & 4.42 & 0.316 & \\
\end{array}
\]

\[
\text{full.lm}=\text{lm}(y \sim \text{factor(x)}, \text{purerr})
\]

\[
\text{anova}(\text{full.lm})
\]

\[
\begin{array}{lcccc}
\text{Df} & \text{Sum Sq} & \text{Mean Sq} & \text{F value} & \text{Pr(>F)} \\
\hline
\text{factor(x)} & 7 & 65.272 & 9.3245 & 758.6 & 1.198e-10 \\
\text{Residuals} & 8 & 0.098 & 0.0123 & \\
\end{array}
\]

\[
\text{anova}(\text{reduced.lm}, \text{full.lm})
\]

\[
\begin{array}{lcccc}
\text{Model 1:} & y \sim x \\
\text{Model 2:} & y \sim \text{factor(x)} \\
\text{Res.Df} & \text{RSS} & \text{Df Sum of Sq} & \text{F} & \text{Pr(>F)} \\
1 & 14 & 4.4196 & & \\
2 & 8 & 0.0983 & 6 & 4.3213 & 58.594 & 3.546e-06 \\
\end{array}
\]

Therefore

\[
F_{LOF} = \frac{(SSE_R - SSE_F)/(df_{SSE,R} - df_{SSE,F})}{SSE_F/df_{SSE,F}}
\]

\[
= \frac{(4.42 - 0.098)/(14 - 8)}{0.098/8} = \frac{4.3213/6}{0.0983/8} = 58.594
\]
LOF p-value

- Compare $F = 58.594$ with $F(6, 8)$, p-value $= P(F(6, 8) \geq 58.594) < 0.001$.
- There is very strong evidence of a lack of fit.

**ANVOA Table for LOF**

<table>
<thead>
<tr>
<th>Source</th>
<th>df</th>
<th>SS</th>
<th>MS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Regression</td>
<td>1</td>
<td>SSR = 60.950</td>
<td>MSR = 60.950</td>
</tr>
<tr>
<td>Lack of Fit</td>
<td>6</td>
<td>SSLF = 4.322</td>
<td>MSLF = 0.720</td>
</tr>
<tr>
<td>Pure Error</td>
<td>8</td>
<td>SSPE = 0.098</td>
<td>MSPE = 0.0123</td>
</tr>
<tr>
<td>Total</td>
<td>15</td>
<td>SSTO = 65.370</td>
<td>-</td>
</tr>
</tbody>
</table>
LOF by hand

The data consist of 16 observations with $X$ repeated at several values:

<table>
<thead>
<tr>
<th>$X$</th>
<th>4.1</th>
<th>5.1</th>
<th>5.1</th>
<th>5.1</th>
<th>6.3</th>
<th>6.3</th>
<th>7.0</th>
<th>7.9</th>
</tr>
</thead>
<tbody>
<tr>
<td>$Y$</td>
<td>6.3</td>
<td>7.3</td>
<td>7.4</td>
<td>7.4</td>
<td>7.8</td>
<td>7.7</td>
<td>8.4</td>
<td>10.8</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>$X$</th>
<th>7.9</th>
<th>7.9</th>
<th>8.6</th>
<th>9.4</th>
<th>9.4</th>
<th>9.4</th>
<th>10.2</th>
<th>10.2</th>
</tr>
</thead>
<tbody>
<tr>
<td>$Y$</td>
<td>10.6</td>
<td>10.9</td>
<td>11.0</td>
<td>11.1</td>
<td>10.9</td>
<td>11.0</td>
<td>12.6</td>
<td>12.8</td>
</tr>
</tbody>
</table>

![Graph showing the relationship between X and Y.](image-url)
Computing SS Pure Error by hand

- There are 5 repeat groups out of 8 groups:

<table>
<thead>
<tr>
<th>i</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>–</th>
<th>–</th>
<th>–</th>
</tr>
</thead>
<tbody>
<tr>
<td>X</td>
<td>5.1</td>
<td>6.3</td>
<td>7.9</td>
<td>9.4</td>
<td>10.2</td>
<td>4.1</td>
<td>7.0</td>
<td>8.6</td>
</tr>
<tr>
<td>Y</td>
<td>7.3</td>
<td>7.8</td>
<td>10.8</td>
<td>11.1</td>
<td>12.6</td>
<td>6.3</td>
<td>8.4</td>
<td>11.0</td>
</tr>
<tr>
<td></td>
<td>7.4</td>
<td>7.7</td>
<td>10.6</td>
<td>10.9</td>
<td>12.8</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>7.4</td>
<td>10.9</td>
<td>11.0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>n_i</td>
<td>3</td>
<td>2</td>
<td>3</td>
<td>3</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>

- Compute SSPE as

\[
\sum_{i=1}^{3}(Y_{i1} - \bar{Y}_1)^2 + \sum_{i=1}^{2}(Y_{i2} - \bar{Y}_2)^2 + \sum_{i=1}^{3}(Y_{i3} - \bar{Y}_3)^2 \\
+ \sum_{i=1}^{3}(Y_{i4} - \bar{Y}_4)^2 + \sum_{i=1}^{2}(Y_{i5} - \bar{Y}_5)^2
\]
Computing SS Pure Error by hand

- $Y_{ij}$: the $i$th observation for the $j$th group.
- $n_j$: the number of observations in the $j$th group.
- $c$: the number of repeat groups.

In general,

$$SSPE = \sum_{j=1}^{c} \sum_{i=1}^{n_j} (Y_{ij} - \bar{Y}_j)^2,$$

$$df\ PE = \sum_{j=1}^{c} (n_j - 1) = n - r$$

In the example, $n_1 = 3, n_2 = 2, n_3 = 3, n_4 = 3, n_5 = 2, c = 5$ and hence $df\ PE$ is $2 + 1 + 2 + 2 + 1 = 8$ and $SSPE = 0.098$. 
Computing SS Lack of Fit by Hand

By subtraction. In the example,

- \( SSE = 4.42 \) from the regression ANOVA table on df = 14.
- \( SSPE = 0.098 \) on df = 8.

Thus
\[
SSLF = SSE - SSPE = 4.42 - 0.098 = 4.322.
\]

- df LF = 14 − 8 = 6.
- Note
\[
SSLF = \sum_{j=1}^{c^*} \sum_{i=1}^{n_j} (\bar{Y}_j - \hat{Y}_{ij})^2
\]
where \( k^* \) denotes the number of groups (here \( c^* = 8 \)).
Lack of Fit Test by Hand

- For testing $H_0$: No lack of fit (here, simple linear regression is adequate) versus $H_a$: Lack of fit (here, simple linear regression is inadequate).

- Use the fact that, under $H_0$,
  \[
  F = \frac{SSLF/df_{LF}}{SSPE/df_{PE}} \sim F(df_{LF}, df_{PE}).
  \]

- In the example, the observed $F$ test statistic is
  \[
  F^* = \frac{4.332/6}{0.098/8} = 58.62.
  \]
Lack of Fit Test: Remarks

- Note that the $R^2 = 93.2\%$ is high, but according to the LOF test, the model is still inadequate.
- Possible remedy is to use polynomial regression.
- The repeats need to be independent measurements.
  - If there are no repeats at all, some consider approximate repeat groups by binning the $X$’s close to one another into groups. In this case, the LOF test is an approximate test.
Remedial Measures

- For simple linear regression, consider two basic approaches.
  - Abandon the current model and look for a better one.
  - Transform the data so that the simple linear regression model is appropriate.

- Nonlinearity of regression function:
  - Transformation (X or Y or both)
  - Polynomial regression.
  - Nonlinear regression.

- Nonconstancy of error variance:
  - Transformation (Y)
  - Weighted least squares.
Remedial Measures

- Simultaneous nonlinearity and nonconstant variance
  - Sometimes works to...
  - Transform Y to fix variance then
  - Transform X to fix linearity

- Nonindependence of error terms:
  - First-order differencing.
  - Models with correlated error terms.

- Nonnormality of error terms.
  - Transformation.
  - Generalized linear models.

- Presence of outliers:
  - Removal of outliers (with extreme caution).
  - Analyze both with and without outliers
  - Robust estimation.
  - New model.
Example: bacteria

Data consist of number of surviving bacteria after exposure to X-rays for different periods of time. Let \( t \) denote time (in number of 6-minute intervals) and let \( n \) denote number of surviving bacteria (in 100s) after exposure to X-rays for \( t \) time.

<table>
<thead>
<tr>
<th>( t )</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>( n )</td>
<td>355</td>
<td>211</td>
<td>197</td>
<td>166</td>
<td>142</td>
<td>166</td>
<td>104</td>
<td>60</td>
</tr>
<tr>
<td>( t )</td>
<td>9</td>
<td>10</td>
<td>11</td>
<td>12</td>
<td>13</td>
<td>14</td>
<td>15</td>
<td></td>
</tr>
<tr>
<td>( n )</td>
<td>56</td>
<td>38</td>
<td>36</td>
<td>32</td>
<td>21</td>
<td>19</td>
<td>15</td>
<td></td>
</tr>
</tbody>
</table>
Example: bacteria

We fit a simple linear regression model to the data.

- It appears that the linear-line model is not adequate.
- The assumption of correct model seems to be violated.
- What to do?
Example: bacteria

- Consider nonlinear model (why nonlinear?)

\[ n_t = n_0 e^{\beta t}, \]

where \( t \) is time, \( n_t \) is the number of bacteria at time \( t \), \( n_0 \) is the number of bacteria at \( t = 0 \), and \( \beta < 0 \) is a decay rate.

- Take natural logs of both sides of the model, we have,

\[
\ln(n_t) = \ln(n_0) + \ln(e^{\beta t}) = \ln(n_0) + \beta t
\]

\[
= \alpha + \beta t,
\]

by setting \( \alpha = \ln(n_0) \).

- That is, we log-transformed \( n_t \) and the result is a usual linear-line model!
Example: bacteria

The transformed data are as follows.

<table>
<thead>
<tr>
<th>$t$</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>$\ln(n)$</td>
<td>5.87</td>
<td>5.35</td>
<td>5.28</td>
<td>5.11</td>
<td>4.96</td>
<td>5.11</td>
<td>4.64</td>
<td>4.09</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>$t$</th>
<th>9</th>
<th>10</th>
<th>11</th>
<th>12</th>
<th>13</th>
<th>14</th>
<th>15</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\ln(n)$</td>
<td>4.03</td>
<td>3.64</td>
<td>3.58</td>
<td>3.47</td>
<td>3.04</td>
<td>2.94</td>
<td>2.71</td>
</tr>
</tbody>
</table>

It appears that the linear-line model is now adequate. The residual plot now shows a random scatter.
**Example: bacteria**

Based on the log-transformed counts, we can fit the model to get the LS estimates

\[ \hat{\alpha} = 6.029, \quad \hat{\beta} = -0.222, \quad s_{\ln(N) \cdot t} = 0.1624, \quad R^2 = 0.9757. \]

- Inference for \( \beta \) is straightforward. [Unit? Interpretation?]
- Inference for \( \alpha \) is straightforward. [Unit? Interpretation?]
- Inference for \( n_0 \) is not straightforward.
  - Since \( \alpha = \ln(n_0) \), \( n_0 = e^\alpha \).
  - Given \( \hat{\alpha} = 6.029 \), we obtain an estimate of \( n_0 \)
    \[ \hat{n}_0 = e^{\hat{\alpha}} = 415.30. \]
  - But the estimate \( \hat{n}_0 \) is biased (i.e., \( E(\hat{n}_0) \neq n_0 \)).
The purpose of transformation is to meet the assumptions of the linear regression analysis.

<table>
<thead>
<tr>
<th>Linear Models</th>
<th>Nonlinear Models</th>
</tr>
</thead>
<tbody>
<tr>
<td>(L1) $Y = \beta_0 + \beta_1 X$</td>
<td>(N1) $Y = \alpha e^{\beta X} \rightarrow \ln(Y) = \ln(\alpha) + \beta X$</td>
</tr>
<tr>
<td>(L2) $Y = \beta_0 + \beta_1 X^2$</td>
<td>(N2) $Y = \alpha X^{\beta} \rightarrow \ln(Y) = \ln(\alpha) + \beta \ln(X)$</td>
</tr>
<tr>
<td>(L3) $Y = \beta_0 + \beta_1 e^X$</td>
<td>(N3) $Y = \alpha + e^{\beta X}$</td>
</tr>
</tbody>
</table>

- In (L2) and (L3), the relationship between $X$ and $Y$ is not linear, but the model is linear in the parameters and hence the model is linear.
- In (N1) and (N2), the model is nonlinear, but can be log-transformed to a linear model (i.e., linearized).
- In (N3), the model is nonlinear and cannot be linearized.
Transformation: Remarks

- Transformation could be for $X$, or $Y$, or both. Common transformations are $\log_{10}\{Z\}$, $\ln\{Z\}$, $\sqrt{Z}$, and $Z^2$. Less common transformations include $1/Z$, $1/Z^2$, $\arcsin\sqrt{Z}$, and $\log_2\{Z\}$.

- Another use of transformation is to control unequal variance.
  - Example: If $Y$ are counts, then often larger variances are associated with larger counts. In this case, $\sqrt{Y}$ transformation can help stabilize variance.
  - Example: If $Y$ are proportions (of successes among trials), then $\text{Var}(Y) = \pi(1 - \pi)/n$, which depends on the true success rate $\pi$. Residual plots would reveal the unequal variance problem. In this case, $\arcsin(\sqrt{Y})$ transformation can help stabilize variance.

- Rule of thumbs: positive data - use $\log Y$; data are proportions - use $\arcsin\sqrt{Y}$; data are counts - use $\sqrt{Y}$
Transformation: Remarks

- Ideally, theory should dictate what transformation to use as in the bacteria count example. But in practice, transformation is usually chosen empirically.
- Transforming $Y$ can affect both linearity and variance homogeneity, but transforming $X$ can affect only linearity.
- Sometimes solving one problem can create another. For example, transforming $Y$ to stabilize variance causes curved relationship.
- Usually it is best to start with a simple transformation and experiment. It happens often that a simple transformation allows the use of the linear regression model. When needed, use more complicated methods such as nonlinear regression.
- Transformations are useful not only for simple linear regression, but also for multiple linear regression and design of experiment.
Variance Stabilizing Transformations

- If $\text{Var}(Y) = g(\mathbb{E}(Y))$, then a variance stabilizing transform is

$$h(y) \propto \int (g(z))^{-1/2} dz$$

- Example:
  - if var $\propto$ mean, then $g(z) = z$ and $h(y) = \sqrt{y}$
  - if var $\propto$ mean$^2$, then $g(z) = z^2$ and $h(y) = \ln(y)$
  - if var $\propto$ mean(1-mean), then $g(z) = z(1-z)$, then $h(y) = \sin^{-1} \sqrt{y}$
Box-Cox Transformation

- Consider a transformation ladder for $Z = X$ or $Y$.

\[
\begin{array}{cccccccc}
\lambda & \cdots & -2 & -1 & -0.5 & 0 & 0.5 & 1 & 2 & \cdots \\
\hline
Z' & \cdots & \frac{1}{Z^2} & \frac{1}{Z} & \frac{1}{\sqrt{Z}} & \log(Z) & \sqrt{Z} & Z & Z^2 & \cdots
\end{array}
\]

- Moving up or down the ladder (starting at 1) changes the residual plots in a consistent manner. Use only these choices for manual search, too!
- Box-Cox method is a formal approach to selecting $\lambda$ to transform $Y$.

- The idea is to consider $Y_i^\lambda = \beta_0 + \beta_1 X_i + \epsilon_i$.
- Estimate $\lambda$ (along with $\beta_0, \beta_1, \sigma^2$) using maximum likelihood.
- Box-Cox method may give $\hat{\lambda} = -0.512$. Round to the nearest interpretable value $\hat{\lambda} = -0.5$.
- If $\hat{\lambda} \approx 1$, do not transform.
- In R, boxcox gives a 95% CI for $\lambda$. Choose an interpretable $\hat{\lambda}$ within the CI.
Box-Cox Transformation

- Box-Cox family of transformations

\[ Z = Y^\lambda I(\lambda \neq 0) + \ln(Y) \]

- Estimate \( \lambda \) using maximum likelihood or use the variance - mean relationship

- Using variance-mean relationship to estimate \( g(Y) \)
  - works for 2 groups or many groups (ANOVA presented in near future)
  - Compute \( \bar{Y} \) and \( S_Y \) for each group
  - Regress \( \log(S_Y) \) on \( \log(\bar{Y}) \) and estimate the slope \( \beta \)
  - Use transformation \( Y^\lambda \) with \( \lambda = 1 - \beta \)
Box-Cox Transformation

- When does this work
  - model for variability
    \[ \sigma = \sqrt{\text{Var}(Y)} = k\mu^\beta \]
  - or
    \[ \text{Var}(Y) = \sigma^2 = [k\mu^\beta]^2 := g(\mu) \]
  - Use the delta method to obtain the transformation: \( Z = g(Y) = Y^\lambda \)

- Consider the Taylor expansion
  \[ Z = g(Y) \approx g(\mu) + (Y - \mu)g'(\mu) \]
  then an approximation for \( \text{Var}(g(\mu)) \) is
  \[ \text{Var}(g(Y)) \approx [g'(\mu)]^2 \text{Var}(Y) \]
  which is the famous Delta Method
Box-Cox Transformation

- For $Z = g(Y) = Y^\lambda$ we have

$$\frac{dZ}{dY} = g'(Y) = \lambda Y^{\lambda-1}$$

- From the Delta method

$$\text{Var}(Z) = (\lambda \mu^{\lambda-1})^2 (k \mu^\beta)^2 = k^2 \lambda^2 \mu^2 (\lambda-1+\beta)$$

- When $\lambda = 1 - \beta$, $\text{Var}(Z) \approx k^2 \lambda^2$ is approximately constant

- Analyze the transformed data: e.g.,

  $Z_{11} = \ln(Y_{11}), Z_{12} = \ln(Y_{12}), \ldots, Z_{2,n_2} = \ln(Y_{2,n_2})$

- Usually round to a reasonable value if $\beta$ is not an integer as discussed above

- Caution: Some researchers estimate the slope from the regression of

  $\ln(\text{Var}(Y))$ on $\ln(\bar{Y})$ then use the transform $Z = Y^\lambda$ with $\lambda = 1 - \beta/2$. 