# Simple Linear Regression
## Part 7: Model Diagnostics

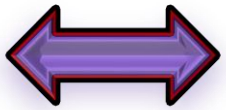
STAT 705: Regression and Analysis of Variance

# Possible Difficulties

- Diagnostic plots (normal probability plot and various residual plots) may reveal that the model does not accurately fit the data
- This could indicate one or more of
  - Errors may not be normal
  - Errors may not have constant variance
  - Relationship between X and Y may not be linear

# Consequences

- If the model assumptions are violated
  - the F statistic (in the ANOVA table) may not follow an F distribution
  - $\hat{\beta}_0$ and $\hat{\beta}_1$ may not follow t distributions
- This mean all hypothesis tests, confidence intervals, and prediction intervals are not valid
- We must check the adequacy of the model before performing inference
- An inadequate model is one example of

**GARBAGE IN**  **GARBAGE OUT**

# Detecting Model Inadequacies

- Graphical methods and/or formal hypothesis tests
- Formal hypothesis tests are limited in scope
  - They each have their own assumptions that must be verified
- We will concentrate on graphical techniques
  - These result in subjective determinations
  - Different analysts may come to different conclusions
- We are looking for extreme inadequacies
  - Minor deviations could be the result of sampling variability

# Diagnostic Plots

- Scatterplot of Y vs. X ⇨ Before fitting the model
  - Is it reasonable to model a linear relationship between X and Y?
- Normal probability plot ⇨ After fitting the model
  - Do the residuals appear to follow a normal distribution?
- Residual plot(s) ⇨ After fitting the model
  - Is the relationship linear?
  - Do the residuals appear to have constant variance?
  - There are several types of residual plots (more about this later)

# Information from Scatterplots

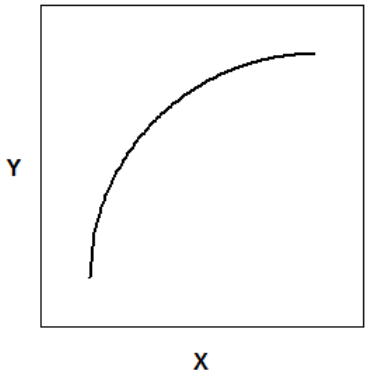
- A nonlinear pattern may be corrected by transforming of one or both of the variables
- Typical shapes and suggested transformations are on the next slide
- When transformed variables are used
  - the new model must still be checked for adequacy
  - the transformed variables must be back-transformed before interpreting the results

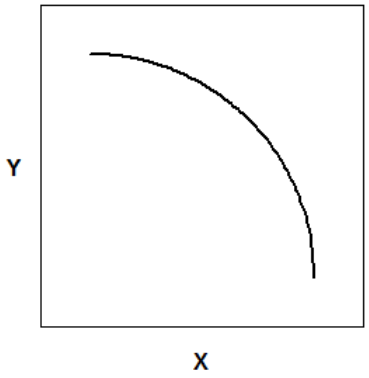
# Suggested Transformations

Transform x
- $x' = \log(x)$
- $x' = 1/x$

-- OR --

Transform y
- $y' = y^2$
- $y' = y^3$




Transform x
- $x' = x^2$
- $x' = x^3$

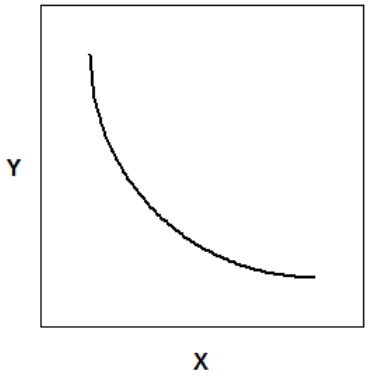
-- OR --

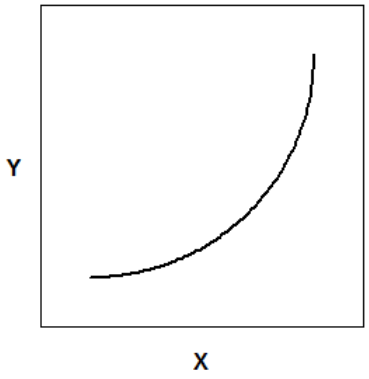
Transform y
- $y' = y^2$
- $y' = y^3$

Transform x
- $x' = \log(x)$
- $x' = 1/x$

-- OR --

Transform y
- $y' = \log(y)$
- $y' = 1/y$




Transform x
- $x' = x^2$
- $x' = x^3$

-- OR --

Transform y
- $y' = \log(y)$
- $y' = 1/y$

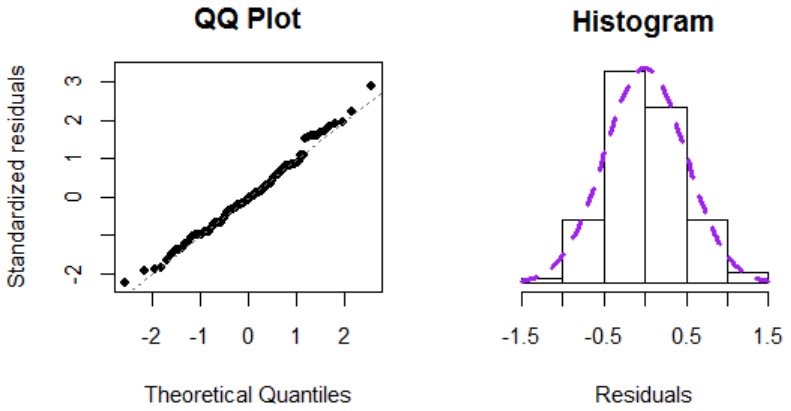
# After the Transformation

- Fit a linear model using the transformed variables
- Assess the adequacy of the fit using the methods on the next few slides
- A complete example is shown in the file ‘TransformXorY’ on the course website

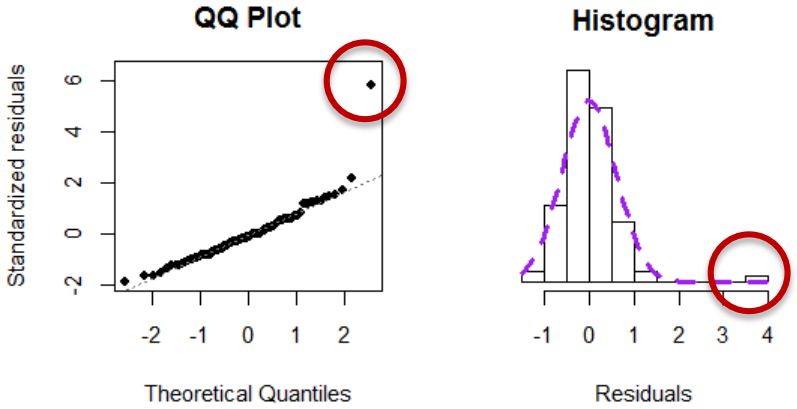
# Information from Normal Plot

- QQ plot and normal probability plot convey same information
  - x-axis has quantiles from theoretical (normal) distribution
  - y-axis has quantiles of the residuals
- If the residuals follow a normal distribution, the points should follow the line
  - Small samples can be deceptive (can erroneously appear non-normal)

# Example QQ plots



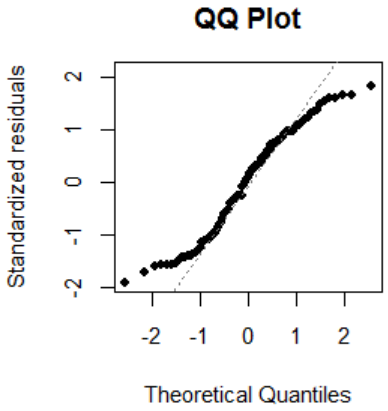

- This is what the QQ plot should look like
- Points follow the line
- Histogram is symmetric and centered at 0
- These graphs use sample size 100
- For smaller samples, graphs may not be quite so 'nice'

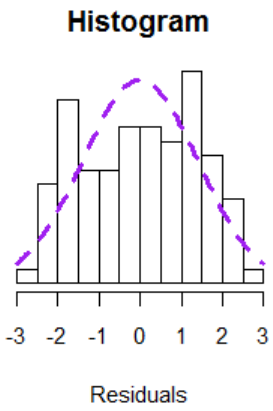



- Potential outlier, circled in red
- This could be a mistake in the data or just an unusual (x, y) pair
- This point should be investigated before proceeding

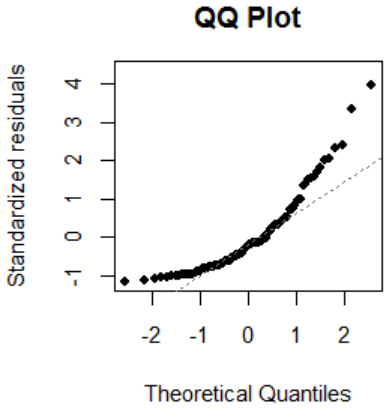
(purple dotted line is the normal density curve)

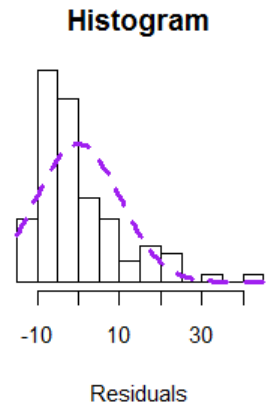
# More QQ plots



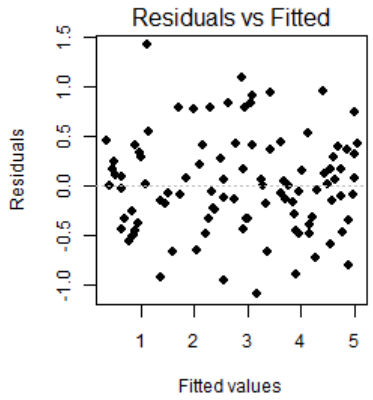

- Distinct 'S' shape indicates a heavy-tailed distribution
- Histogram shows taller than expected bars on both the left and right sides
- However, distribution is symmetric and centered at 0 (this is good)
- Inference (using F and t) will be approximate



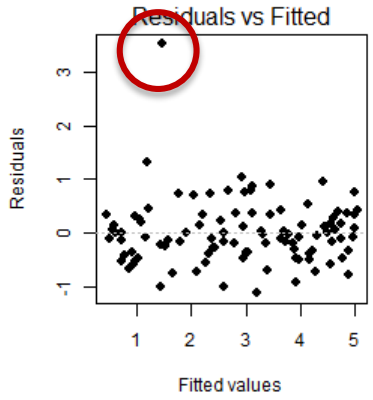




- Extreme example of non-normal residuals
- QQ plot shows 'U' shape
- Histogram is highly skewed
- Clear violation of the normality assumption
- Do not proceed with the analysis until this issue is resolved

# Example Residual Plots

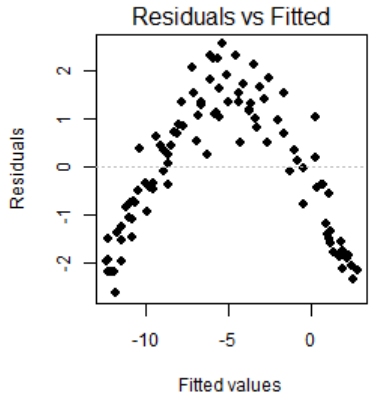



- This is what the residual plot should look like
- Points are scattered, with no obvious pattern
- Histogram is symmetric and centered at 0

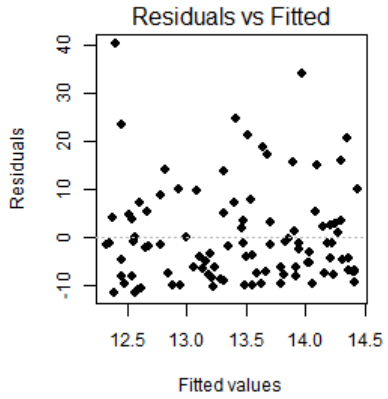



- Potential outlier, circled in red
- This could be a mistake in the data or just an unusual (x, y) pair
- This point should be investigated before proceeding

# More Residual Plots




- This is an absolutely clear and un-ambiguous violation of the assumption of independent errors
- The quadratic shape indicates that including an $x^2$ term in the model may improve the fit




- At first glance, this plot may appear to be okay
- But look at the scale on the y-axis, and note that y=0 is near the bottom of the graph
- Points above y=0 are much more dispersed than point below y=0
- This pattern indicates the distribution of residuals is not symmetric, so they cannot follow a normal distribution

# Residual Plots

- Residual plots on previous slides had fitted values (the y-hats) on the x-axis and the residuals on the y-axis
- Sometimes
  - x-axis contains the observed values for X
  - y-axis contains standardized or studentized residuals (more about this later)
- These graphs are interpreted the same way

# Does the Model Fit the Data?

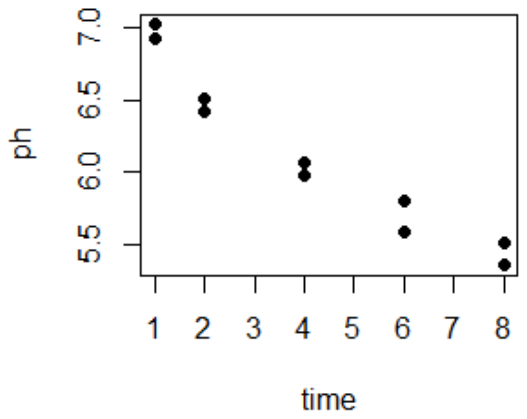
- Diagnostic plots may all look okay, but the model may be a poor fit to the data
- Possible reasons
  - X and Y have a nonlinear relationship
  - X explains only a small portion of the total variability in Y
- We consider numeric summaries to quantify the goodness of fit

# Assessing Goodness of Fit

- Coefficient of determination, $R^2$
  - proportion of variability in Y that is explained by the regression model
  - is always between 0 and 1
- Root mean square error, RMSE
  - Square root of Mean Square Error (MSE)
  - Measures the variability around the regression line
- To compare two competing models that are fitted to the same data
  - Verify that the diagnostic plots look okay for both models
  - Compare the two values of $R^2$; larger is better
  - Compare the two values of RMSE; smaller is better

# Example

- Following a hazardous waste accident, soil at two locations was periodically tested to determine its pH level
- We want to model the change in pH level over time
- X = time since accident; Y = measured pH of soil
- Start with a scatterplot



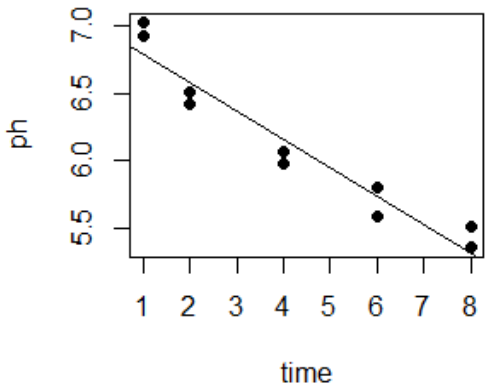

- Scatterplot show curvature
- Compare the fit of two models

Model 1: $Y = \beta_0 + \beta_1 X + \varepsilon$

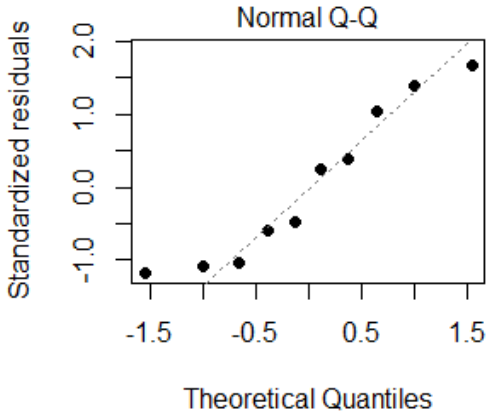
Model 2: $Y = \beta_0 + \beta_1 \log(X) + \varepsilon$

# Model 1: $Y = \beta_0 + \beta_1 X + \varepsilon$

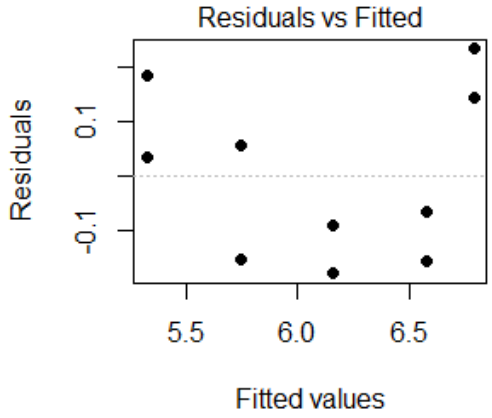
**Scatterplot with Least Squares line**




**QQ Plot 'S' Shape (not good)**



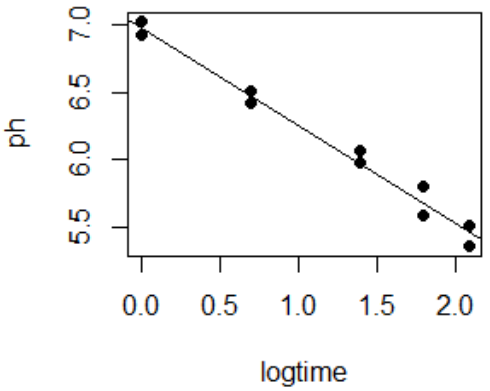

**Residual Plot 'U' Shape (very bad)**



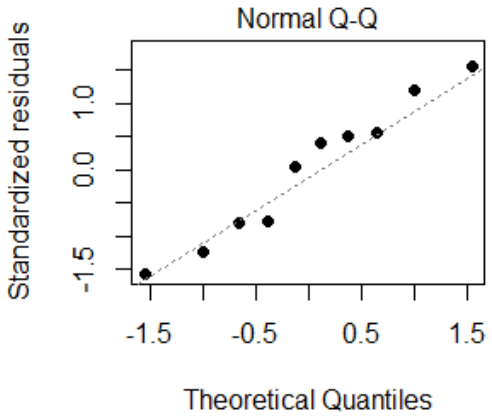

- $R^2 = 0.9324$ ⇨ 93.24% of the variability in pH is explained by this model. (This is very high.)
- Root MSE = 0.16092 ⇨ Cannot interpret this directly; we compare it to other models

# Model 2: $Y = \beta_0 + \beta_1 \log(X) + \varepsilon$

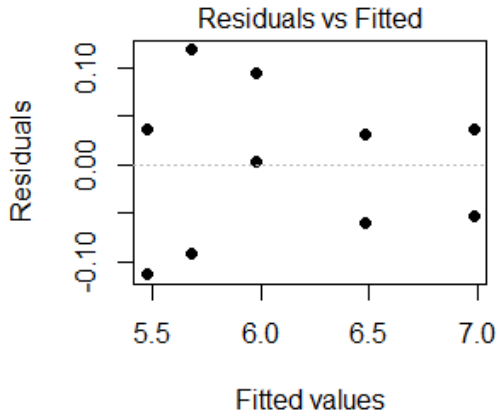
**Scatterplot with Least Squares line**




**QQ Plot Looks MUCH better**



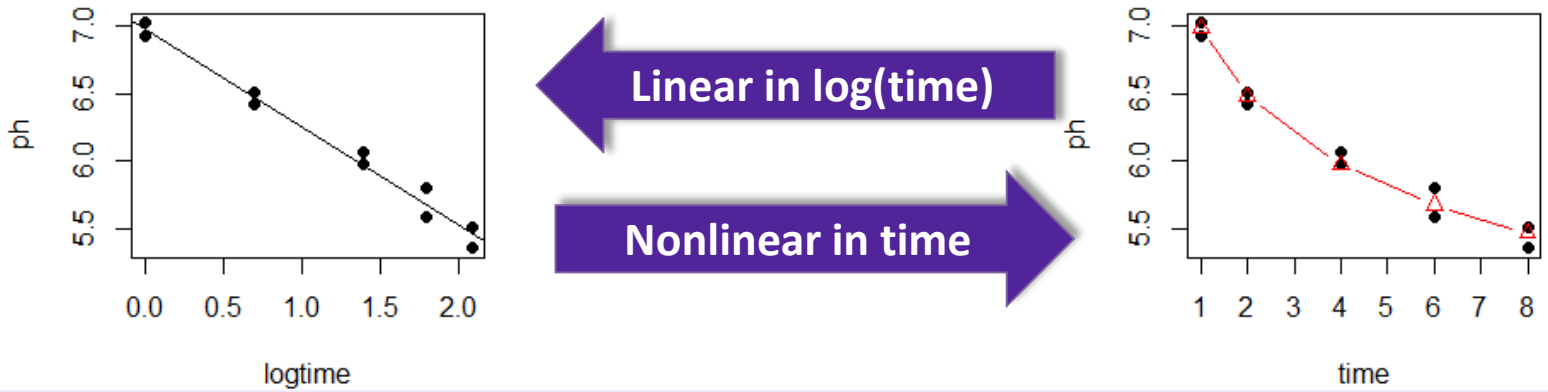

**Residual Plot Looks MUCH better**




- $R^2 = 0.9824$ ⇨ Better (larger) than previous.
- Root MSE = 0.08214 ⇨ Better (smaller) than previous.
- Model 2 is preferred over Model 1

# Back-Transform Log(Time)

- When transformed variables are used to fit the model, they should be back-transformed before reporting the results
- Transformation: x' = log(x)
- Back-transformation is inverse function: x = exp(x')




**Additional details for fitting these models, including the SAS code, is provided on the course website. There are two files: a PDF document and a SAS program (both named 'TransformXorY'). You should run the code and read the explanations as part of this lesson.**

# Things You Should Know

- How to generate fitted models and their diagnostic plots
- Interpret the output to decide
  - if any assumptions appear to be violated
  - if any transformations seem warranted
- Perform transformations and back-transformations
- Compare models and justify which model is a better fit to the data