Chapter 5: Model Post-Fitting for Inference

5.1 Examining Residuals

5.2 Influential Observations

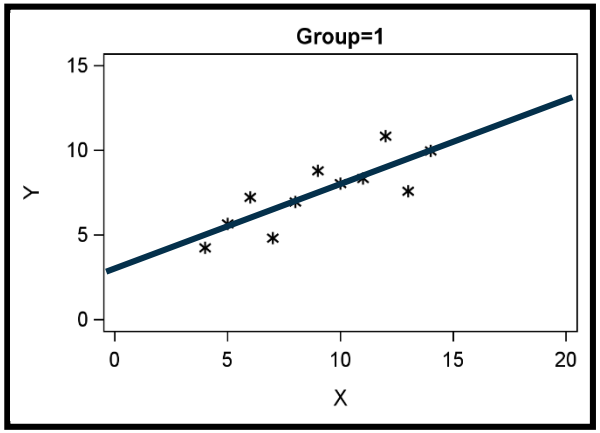
5.3 Collinearity

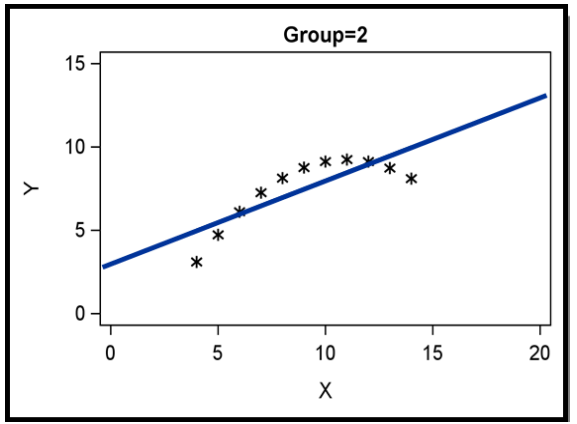
5.01 Poll

Predictor variables are assumed to be normally distributed in linear regression models.

- ❍ True
- ❍ False

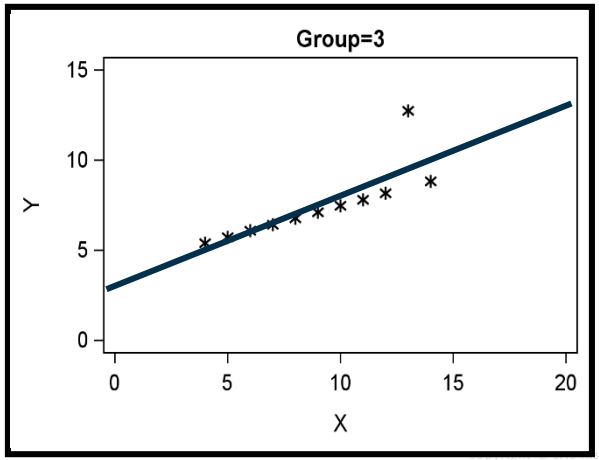
Importance of Plotting Data

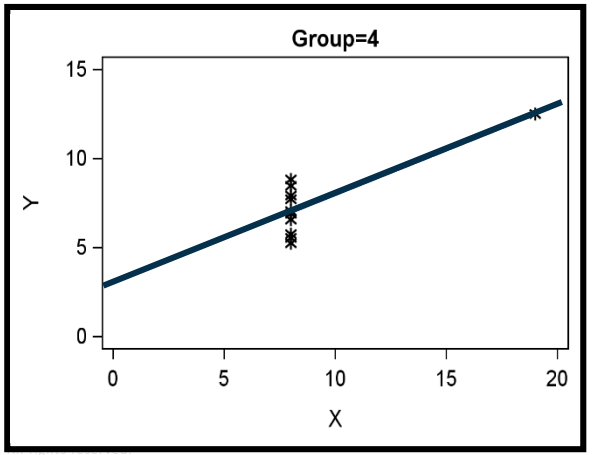




Y-hat=3+0.5X

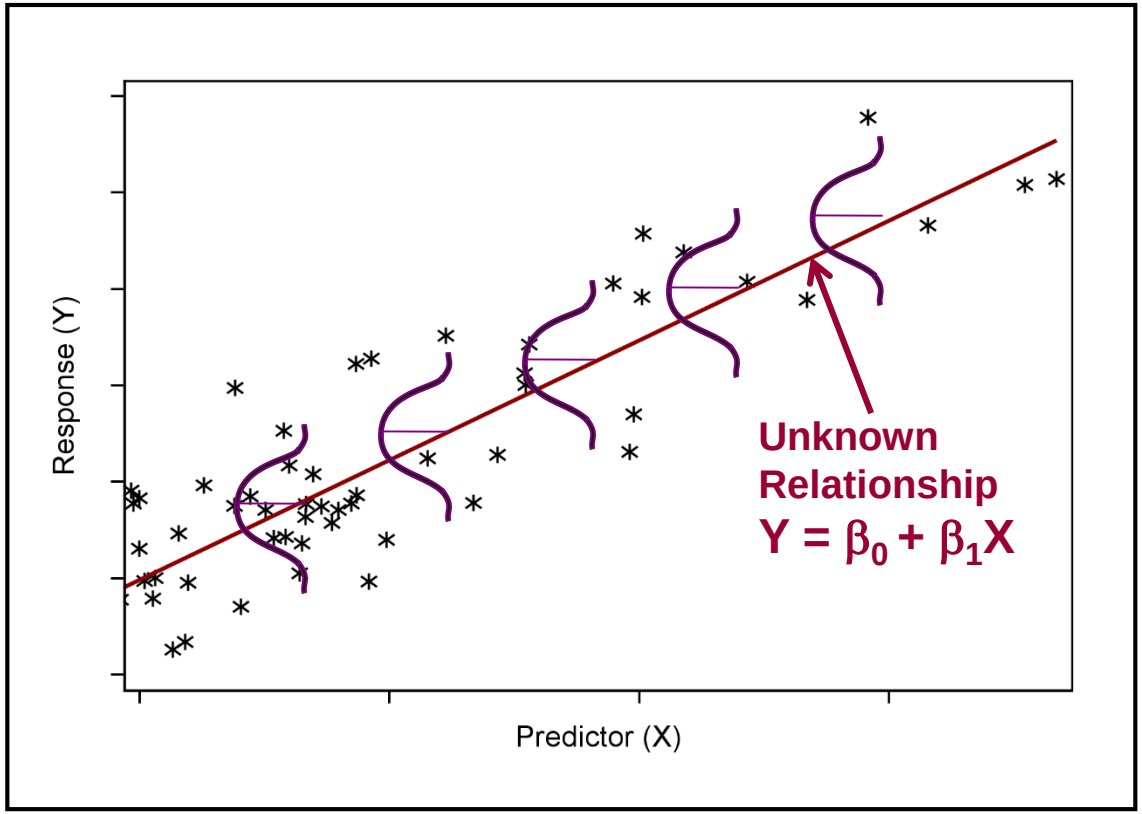
R^2 = 0.67





5.1 Examining Residuals

Assumptions for Regression



Assumptions for Linear Regression

- The mean of the Ys is accurately modeled by a **linear** function of the Xs.
- The random error term, ε, is assumed to have a **normal** distribution with a mean of zero.
- The random error term, ε, is assumed to have a **constant variance**, σ^2.
- The errors are **independent**.

- **No perfect collinearity**

Violation of Model Assumptions

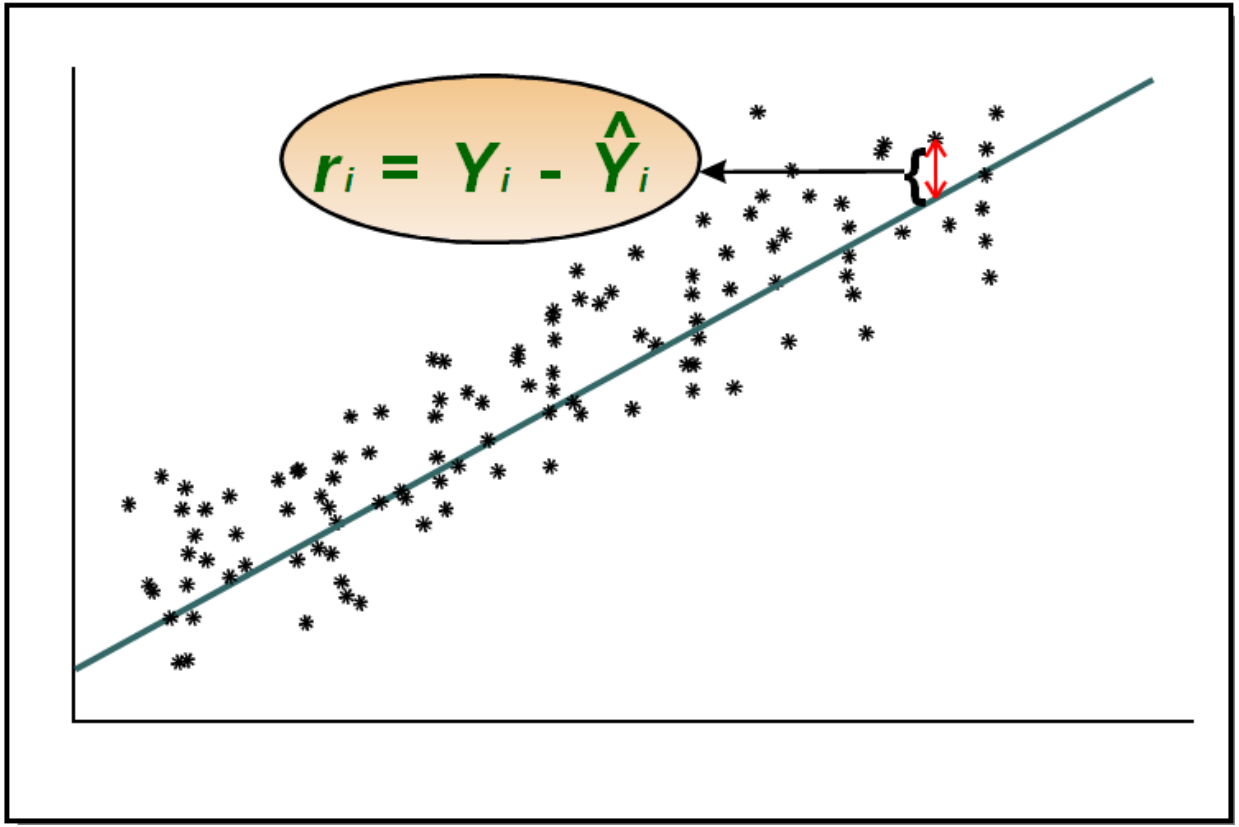
Linear in the parameters – indicates a ***misspecified model***, and therefore the results are not meaningful.

Constant Variance – does not affect the parameter estimates, but the standard errors are compromised.

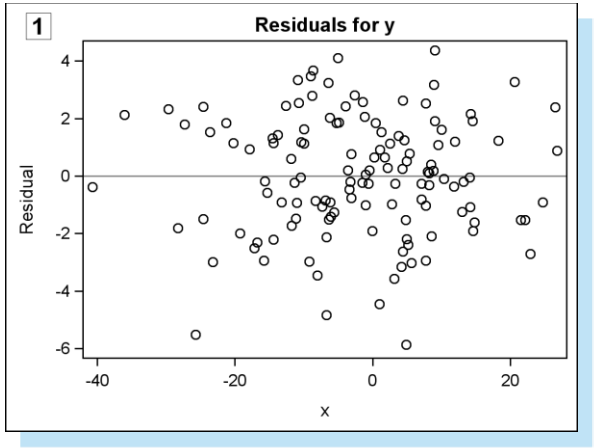
Normality – does not affect the parameter estimates, but it affects the test results.

Independent observations – does not affect the parameter estimates, but the standard errors are compromised.

Verifying Assumptions



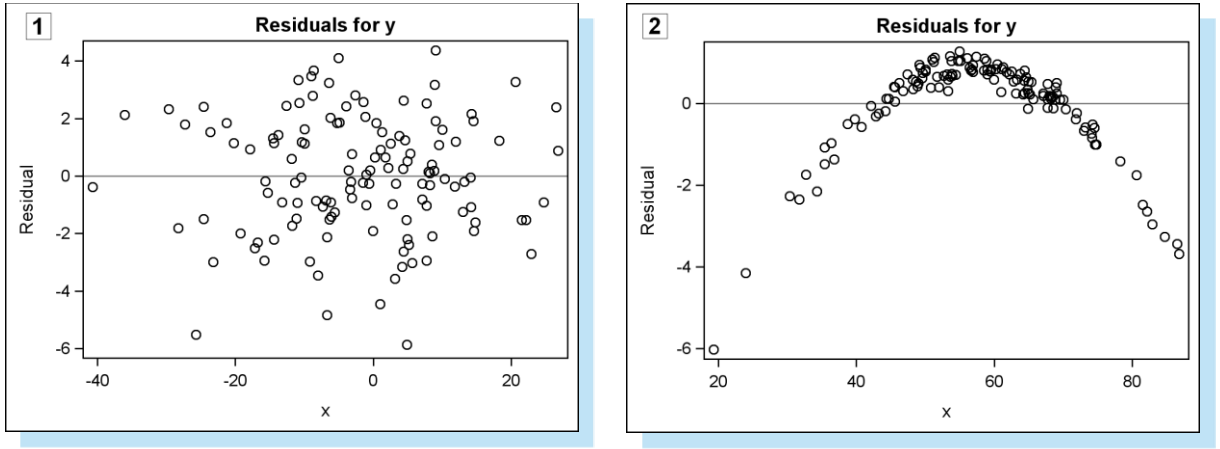
Examining Residual Plots (good residual plot)



- Plot residuals(y-axis) versus each x(x-axis) (or residuals(y-axis) versus predicted value(x-axis))
  - Residuals are randomly scattered about zero reference line.
  - No patterns found.
  - Model form appears to be adequate.

Misspecified Model

Examining Residual Plots-Misspecified Model



- Curvilinear pattern detected in residuals.
- Model form is incorrect.
- Possible remedies, depending on pattern, include polynomial terms, interactions, splines, and so on.

Polynomial Regression Models

- Quadratic Polynomial Model

$$Y_j = \beta_0 + \beta_1 X_j + \beta_2 X_j^2 + \varepsilon_j$$

- Cubic Polynomial Model

$$Y_j = \beta_0 + \beta_1 X_j + \beta_2 X_j^2 + \beta_3 X_j^3 + \varepsilon_j$$

- Polynomial Model with a Cross-Product Term

$$Y_j = \beta_0 + \beta_1 X_{1j} + \beta_2 X_{2j} + \beta_3 X_{1j} X_{2j} + \varepsilon_j$$

Example for Polynomial regression

A researcher is interested in studying the effect of a chemical additive on paper strength. Data can be found in the SAS code document. The response variable is the amount of force required to break the paper (strength) and the explanatory variable is the amount of chemical additive (amount).

SAS Code

```
proc sgplot data=paper;
  scatter x=amount y=strength;
  title “Paper data ”;
run;
proc sgplot data=paper;
  reg x=amount y=strength/lineatrs=(color=blue pattern=solid);
  title “Paper data-Linear ”;
run;
proc reg data=paper;
  model strength=amount;
run;
quit;
```

SAS Code

```
proc glmselect data=paper outdesign=paper2;
  effect p_effect=polynomial (amount/degree=4);
  model strength=p_effect/selection=backward select=SL slstay=0.05
    hierarchy=single showpvalues;
  title “Paper data: backward selection”;
run;
quit;
```

SAS Code

```
proc reg data=paper2 plots=all;
  model strength=&_GLSMOD/lackfit;
run;
quit;
```

SAS Code for interactions

```
proc glmselect data=stat1.ameshousing3_train outdesign=ames3;
  effect p_effect=polynomial (age_sold/degree=4);
  model
saleprice=Gr_Liv_Area|Basement_Area|Garage_Area|Deck_Porch_Area|Lot
_Area|Age_Sold|Bedroom_AbvGr|Total_Bathroom|p_effect
@2/selection=backward select=SL slstay=0.05
    hierarchy=single showpvalues;
run;
quit;
```

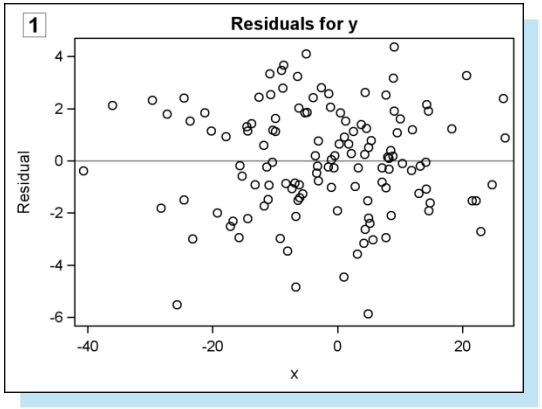
When a straight line is inappropriate

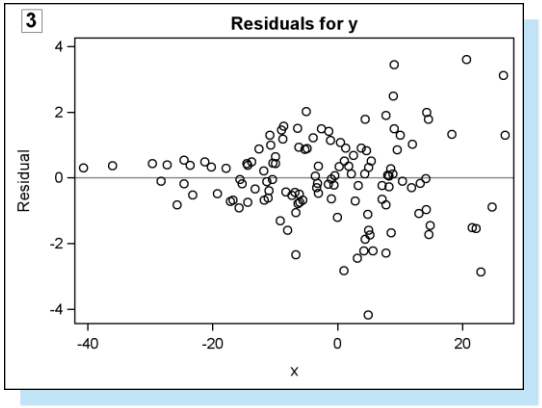
Consider the following options:

- Fit a polynomial/more complex regression model.
- Transform the dependent and/or independent variables to obtain linearity.
- Fit a nonlinear regression model using PROC NLIN if appropriate.
- Fit a nonparametric regression model using PROC LOESS.

Lack of constant variance

Examining Residual Plots-Variance is not constant



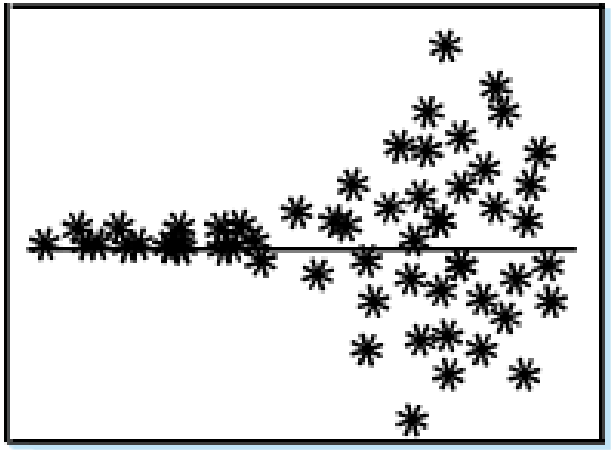


- Constant variance assumption is violated.
- Possible remedy is transforming variables to stabilize the variance.
- Procedures that model the non-constant variance can be used. (GENMOD, GLIMMIX)

Homoscedasticity

The random error term, ε, is assumed to have a constant variance, σ^2 – **homoscedasticity**.

This is an example of **heteroscedasticity**.



Does ***not*** affect the calculation of the parameter estimates.

Does affect the standard errors of the parameter estimates.

Heteroscedasticity

Any inferences under the traditional assumptions will be incorrect.

Hypothesis tests and confidence intervals based on the t, F, χ^2 distributions will not be valid.

Detecting Heteroscedasticity

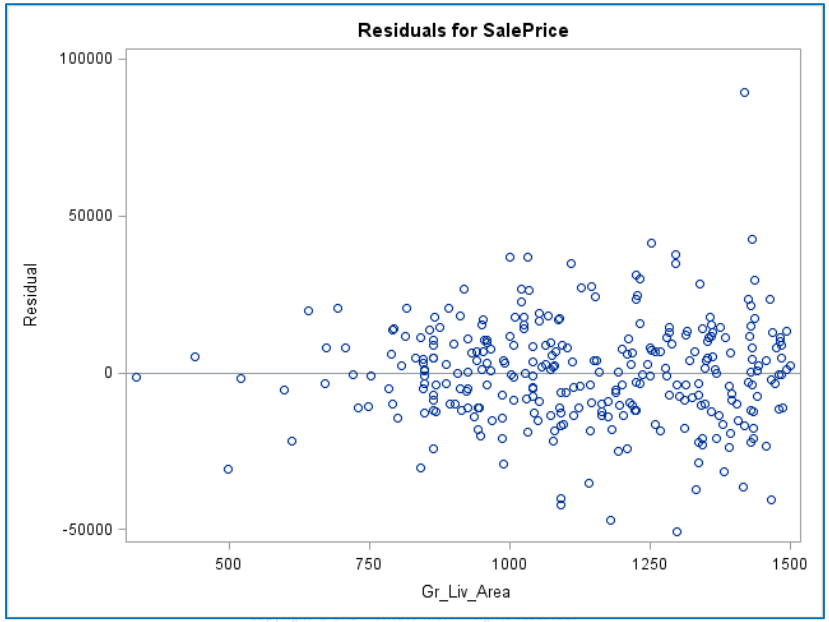
There are a couple of approaches to detecting heteroscedasticity in a data set.

1. Plotting residuals and looking for patterns.
2. Formal tests for heteroscedasticity.
3. Spearman Rank Correlation

Detecting Heteroscedasticity

There are a couple of approaches to detecting heteroscedasticity in a data set.

1. Plotting residuals and looking for patterns.



Detecting Heteroscedasticity

There are a couple of approaches to detecting heteroscedasticity in a data set.

1. Plotting residuals and looking for patterns.
2. Formal tests for heteroscedasticity- (self study).

   Can be very sensitive. One test is the White's general test and another is the Breusch-Pagan test (this one is looking at each independent variable individually).

   H_0: Variance is homoscedastic

   H_A: Variance is heteroscedastic
3. Spearman Rank Correlation

Detecting Heteroscedasticity

There are a couple of approaches to detecting heteroscedasticity in a data set.

1. Plotting residuals and looking for patterns.
2. Formal tests for heteroscedasticity.
3. Spearman Rank Correlation

   The Spearman correlation uses ranks of the data (still between -1 and 1)

Spearman Rank Correlation

If the Spearman rank correlation coefficient between the ***absolute value of the residuals*** and the ***predicted values*** is

- close to zero, then the variance is potentially homoscedastic
- positive, then the variance increases as the mean increases
- negative, then the variance decreases as the mean increases
- Can perform a test: H_0: variance is homoscedastic

  H_A: variance is heterscedastic
- If there is a relationship between the absolute value of residuals and predicted value but it is not linear, this test will NOT discover it

Detecting Heteroscedasticity

```
data check;
  set check;
  abserror=abs(residual);
run;
proc corr data=check spearman nosimple;
  var abserror pred;
run;
```

Accounting for Heteroscedasticity

There are a couple of approaches to account for heteroscedasticity:

1. Use Weighted Least Squares (WLS) or iteratively reweighted least squares (IRLS).
2. Transform data (more on this in next section).

Lack of Normality

Detecting Lack of Normality

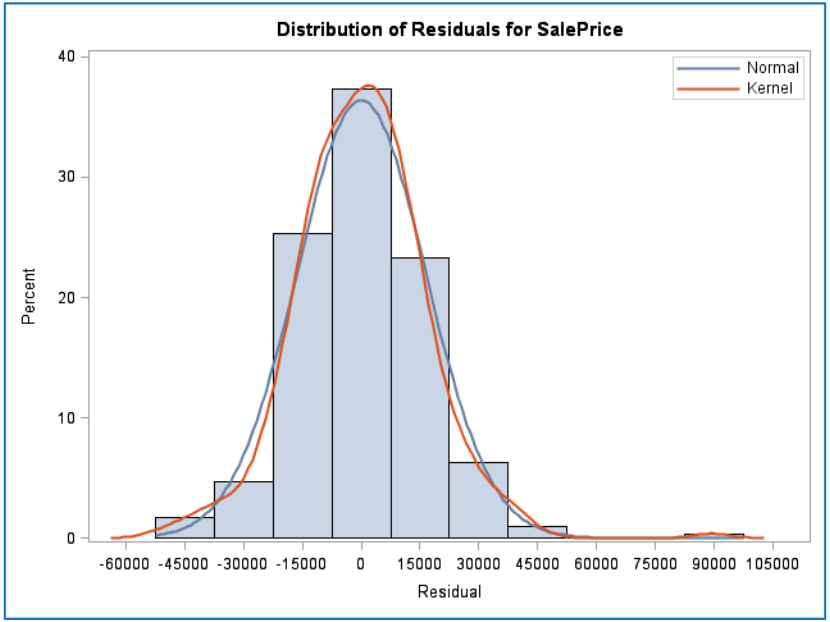
Check that the error terms are Normally distributed by examining:

1. Histogram of the residuals
2. Normal probability plot of the residuals (QQ-plot)
3. Formal tests for Normality

Detecting Lack of Normality

Check that the error terms are Normally distributed by examining:

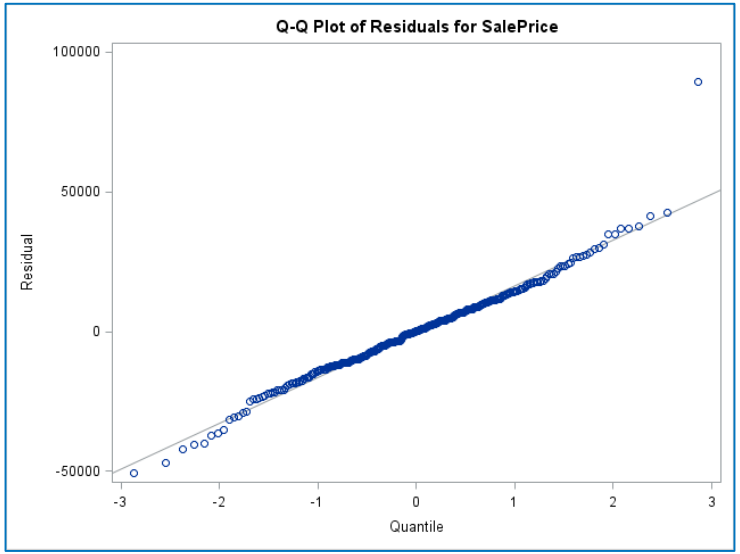
1. Histogram of the residuals



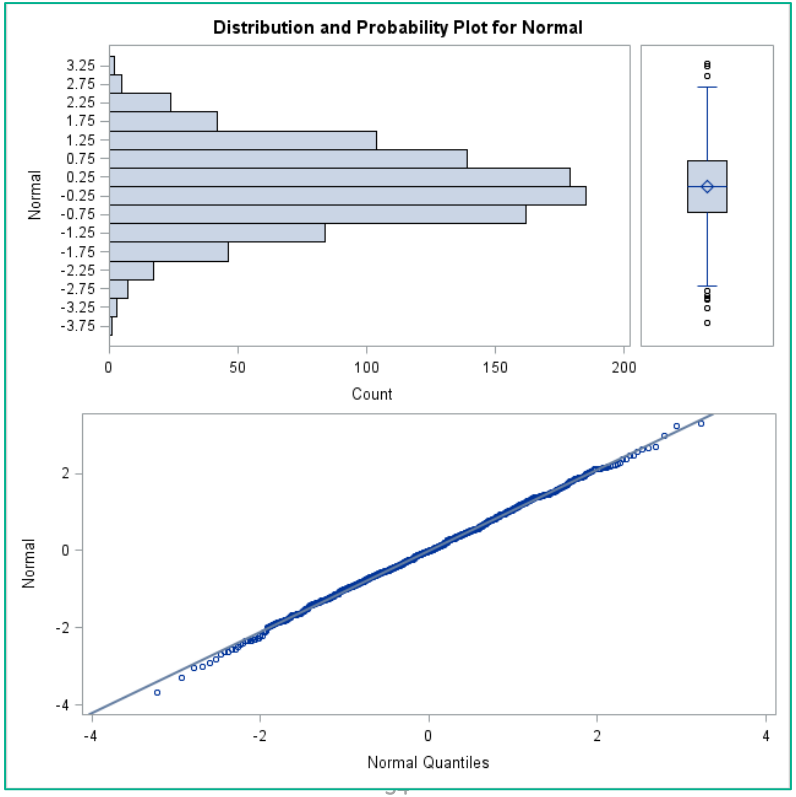
Detecting Lack of Normality

Check that the error terms are Normally distributed by examining:

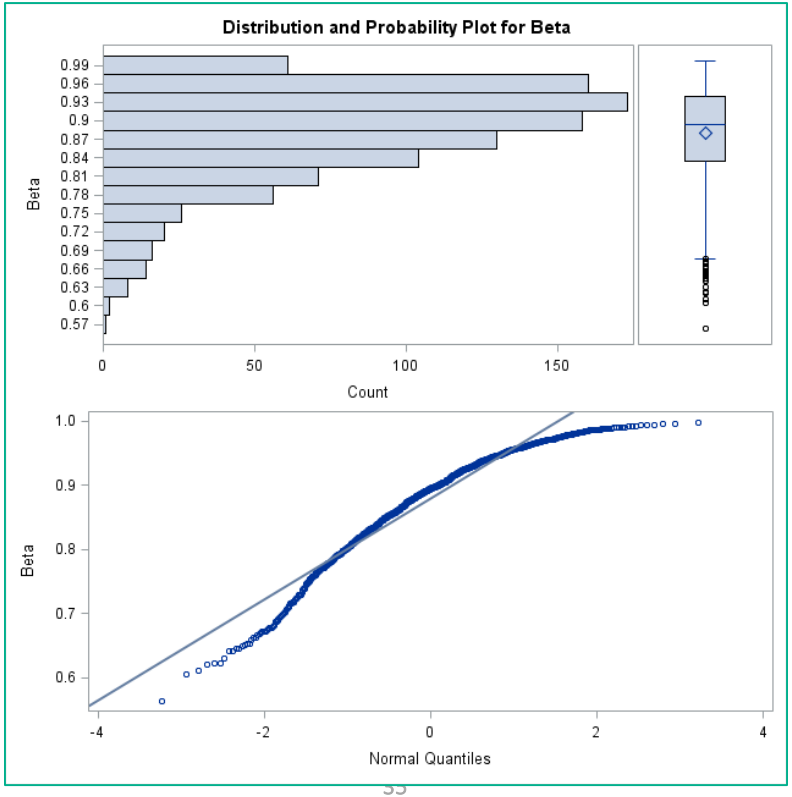
1. Histogram of the residuals
2. Normal probability plot of the residuals (QQ-plot)



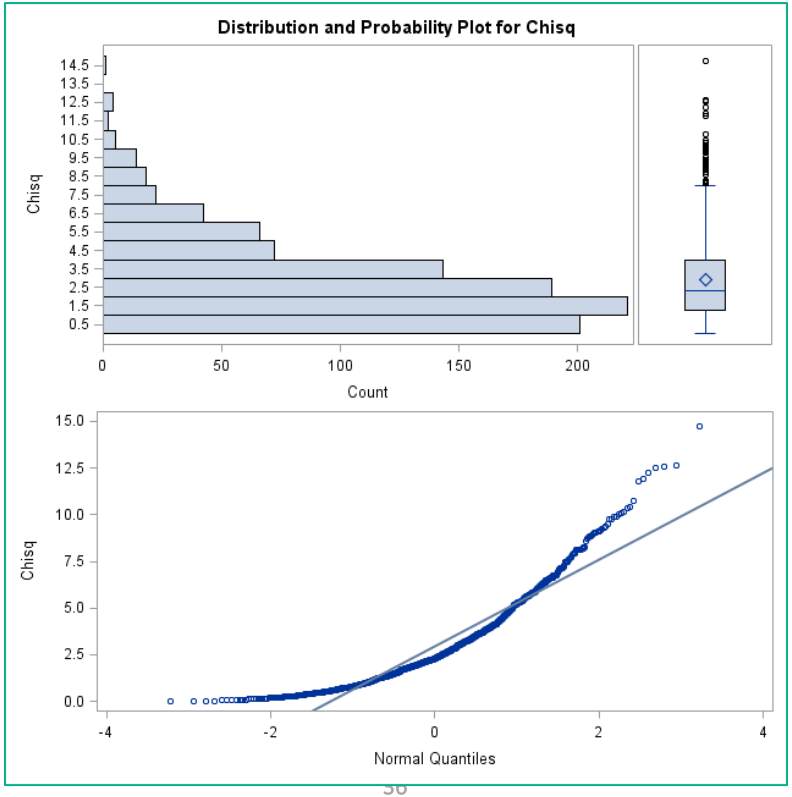
QQ-plots – Normal



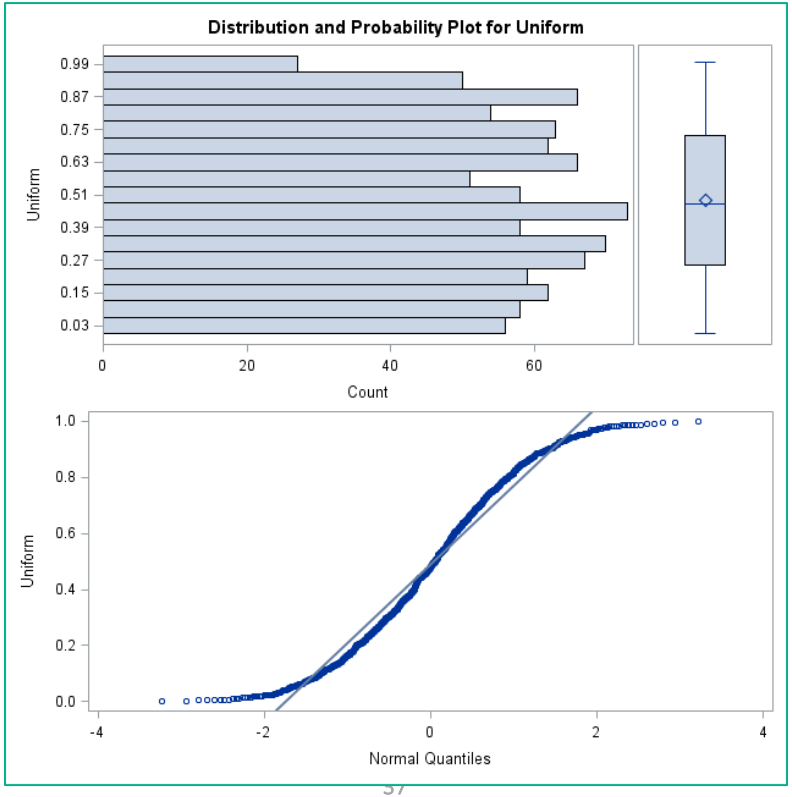
QQ-plots – Left-Skewed



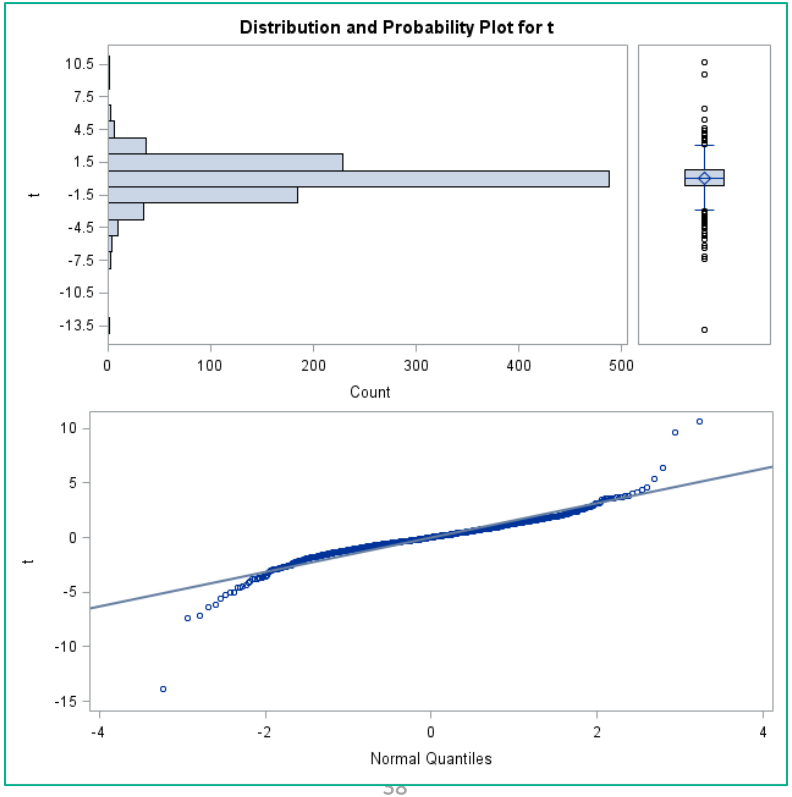
QQ-plots –Right-Skewed



QQ-plots – Platykurtic



QQ-plots – Leptokurtic



Detecting Lack of Normality

Check that the error terms are Normally distributed by examining:

1. Histogram of the residuals
2. Normal probability plot of the residuals (QQ-plot)
3. Formal tests for Normality

H_0: Normality

H_A: Not Normality

Detecting Lack of Normality

Check that the error terms are Normally distributed by examining:

1. Histogram of the residuals
2. Normal probability plot of the residuals (QQ-plot)
3. Formal tests for Normality (Shapiro-Wilk, K-S, Cramer-von Mises and Anderson-Darling are all provided with the code below).

```
proc univariate data=check normal plots;
 var residual;
 probplot;
run;
```

Accounting for Lack of Normality

Depends on why the lack of Normality occurred:

- Outliers → Robust Regression
- Relationship between variables → Transformation Needed
- Can try Box-Cox transformation

Box-Cox transformation

- Box-Cox (1964) developed a method to determine the best transformation to induce normality.
- The Box-Cox transformation has the following form:

$$\begin{cases} (y^{\lambda} - 1)/\lambda & \lambda \neq 0 \\ \log(y) & \lambda = 0 \end{cases}$$

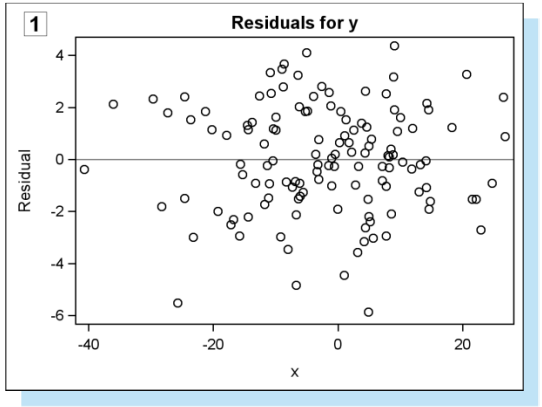
λ is the power in which the response variable y is raised to (so, if $\lambda = 2$, then we would square y). The exception is when $\lambda=0$ (this the log transformation).

How to do Box-Cox in SAS

```
proc transreg data=trans2;
  model BoxCox(y2)=identity(x);
  run;
```

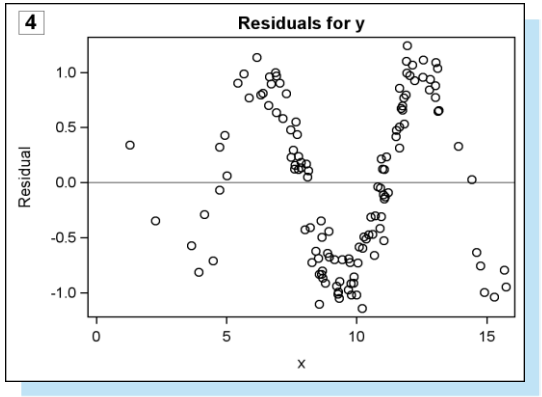
Correlated error terms

Examining Residual Plots



- Observations not independent.
- Residuals follow cyclic pattern.
- Most evident when collected over time.

- Remedy is to analyze using PROC AUTOREG or time series



Independence

Know the source of your data:

- Clustered/Grouped data
- Observations connected in some way
- Complex survey designs
- Repeated measures
- Data gathered over time

Independence

Know the source of your data:

- Clustered data
- Complex survey designs
- Repeated measures
- Data gathered over time

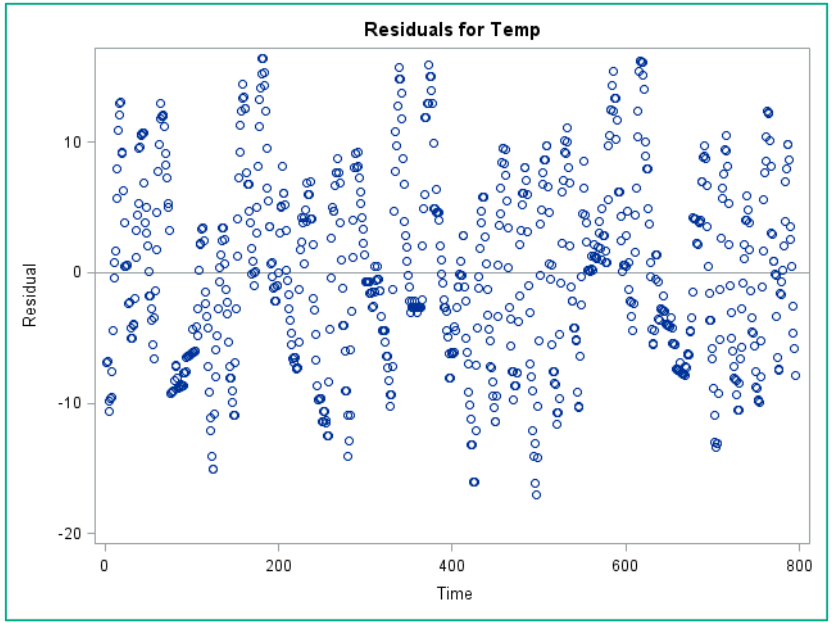
For time-series data, check that the errors are independent by examining:

- Plots of residuals versus time or other ordering component
- Durbin-Watson statistic or the first-order autocorrelation statistic for time-series data

Independence

For time-series data:

1. Plots of residuals versus time or other ordering component



Independence

For time-series data:

1. Plots of residuals versus time or other ordering component
2. Durbin-Watson statistic or the first-order autocorrelation statistic for time-series data

$$H_0: \text{No Residual Correlation}$$

$$H_A: \text{Residual Correlation}$$

Independence

For time-series data:

1. Plots of residuals versus time or other ordering component
2. Durbin-Watson statistic or the first-order autocorrelation statistic for time-series data

$$d = \frac{\sum_{t=2}^{T}(e_t - e_{t-1})^2}{\sum_{t=1}^{T} e_t^2}$$

Independence

For time-series data:

1. Plots of residuals versus time or other ordering component
2. Durbin-Watson statistic or the first-order autocorrelation statistic for time-series data

```
proc reg data=stat1.minntemp;
 model temp=time timesq/dwprob;
run;
quit;
```

Independence

For time-series data:

1. Plots of residuals versus time or other ordering component
2. Durbin-Watson statistic or the first-order autocorrelation statistic for time-series data

```
proc autoreg data=bootcamp.Minntemp;
     model Temp = Time TimeSq / dw=24 dwprob;
run;
```

Chapter 5: Model Post-Fitting for Inference

5.1 Examining Residuals

5.2 Influential Observations/Outliers

5.3 Collinearity

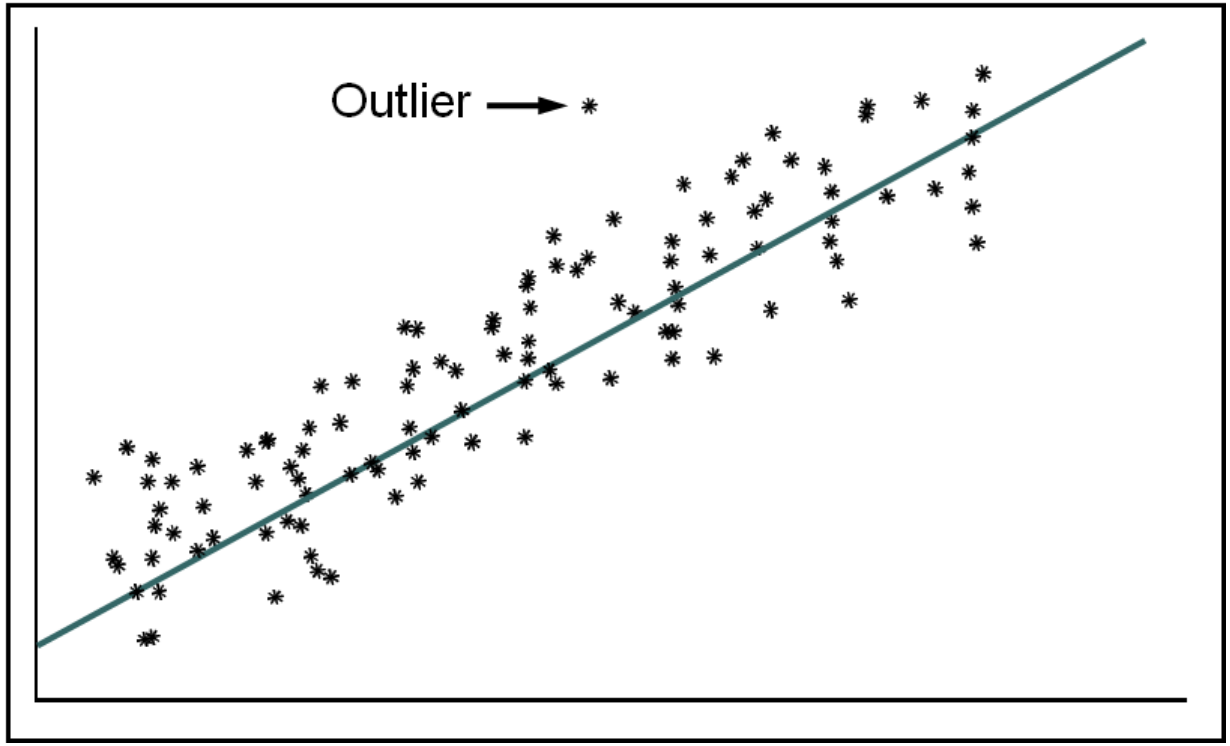
Anomalous Observations

There are two types of anomalous observations that will be discussed:

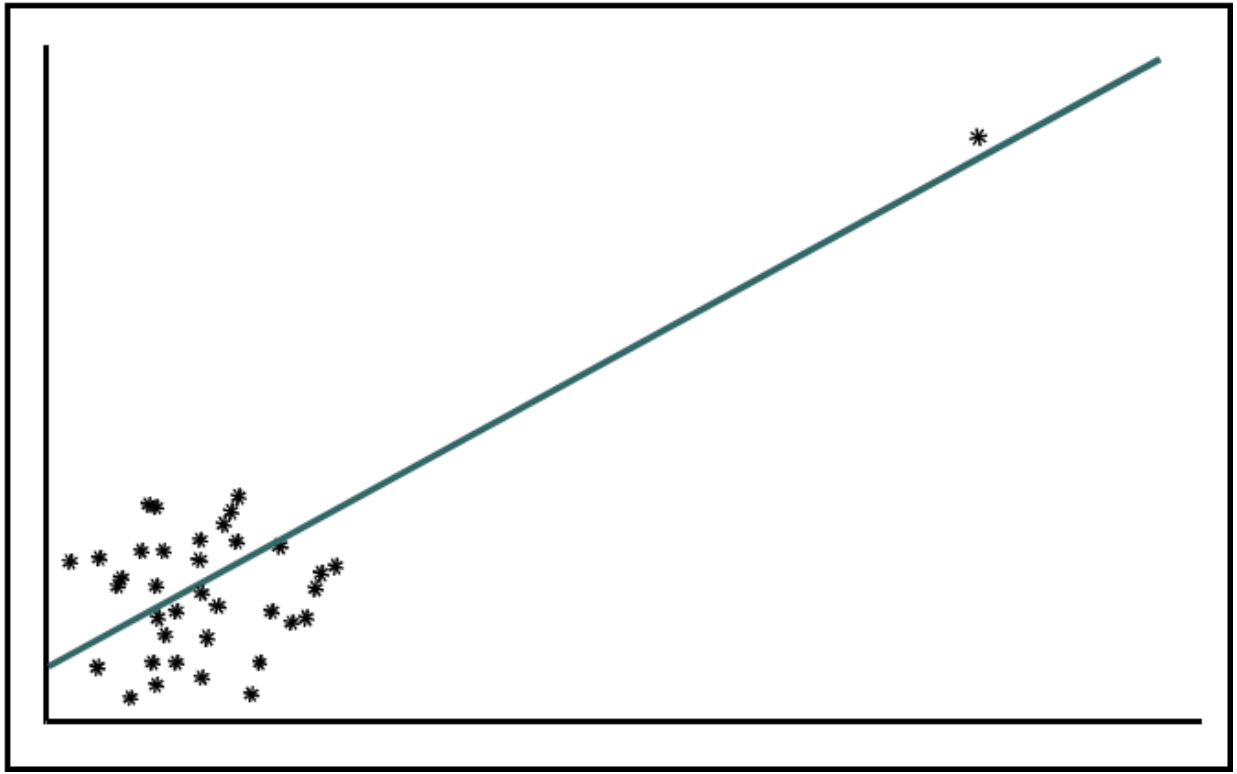
1. Outliers – point with a large standardized residual (lie far away from the fitted line in the Y-direction).
2. Leverage Points – point that falls outside the normal range (far from the mean) in the X-space (possible values of the predictors) and have a large “influence” on the regression line.

Observations could be one or both of these.

Detecting Outliers



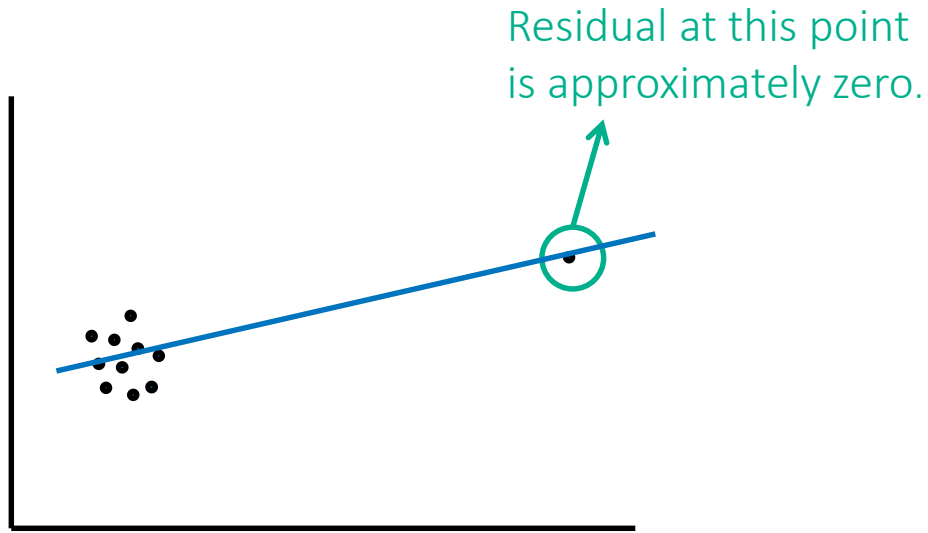
Influential Observations



Residual Analysis

Don’t only focus efforts on residuals of data.

Residual analysis only tends to discover outliers instead of leverage points.



Diagnostic Statistics

Statistics that help identify influential observations are the following:

- Studentized residuals (good for detecting outliers)
- RSTUDENT residuals (outliers)
- Cook's *D* (good for detecting influential observations)
- DFFITS (good for detecting influential observations)
- DFBETAS (good for detecting influential observations)

Studentized (Standardized) Residuals

Studentized residuals (SR) are obtained by dividing the residuals by their standard errors.

Suggested cutoffs are as follows:

- |SR| > 2 for data sets with a relatively small number of observations
- |SR| > 3 for data sets with a relatively large number of observations

5.02 Multiple Choice Poll

Given the properties of the standard normal distribution, you would expect about 95% of the studentized residuals to be between which two values?

a. -3 and 3

b. -2 and 2

c. -1 and 1

d. 0 and 1

e. 0 and 2

f. 0 and 3

Cook's D Statistic

Cook's D statistic is a measure of the simultaneous change in the parameter estimates when the i^{th} observation is deleted from the analysis.

A suggested cutpoint for influence is shown below:

$$Cook's\ D_i > \frac{4}{n}$$

DFFITS

DFFITSi measures the impact that the i^{th} observation has on the predicted value.

A suggested cutoff for influence is shown below:

$$|\mathbf{DFFITS_i}| > 2\sqrt{\frac{p}{n}}$$

DFBETAS

- Measure of change in the j^{th} parameter estimate with deletion of the i^{th} observation
- One DFBETA per parameter per observation
- Helpful in explaining on which parameter coefficient the influence most lies

- A suggested cutoff for influence is shown below:

$$|\mathbf{DFBETA_{ij}}| > 2\sqrt{\frac{1}{n}}$$

SAS Code

```
proc glmselect data=ameshousing3_train plots=all outdesign=glmameshousing3AIC2;
  STEPWISEAIC: model SalePrice=&interval/selection=stepwise details=steps
select=AIC;
  title "Stepwise model for SalePrice - AIC";
  store out=glmameshousing3AIC;
run;

proc reg data=glmameshousing3AIC2 plots(unpack) = all;
  model SalePrice=&_GLSMOD/influence spec partial;
output out=check r=residual p=pred rstudent=rstudent h=leverage;
run;
quit;
```

SAS Code

```
proc reg data=stat1.ameshousing3 plots (only)=(Rstudentbypredicted cooksd
dffits dfbetas);
  model SalePrice = &_GLSIND/influence spec partial;
  id PID;
  output out=check r=residual p=pred rstudent=rstudent h=leverage;
run;
quit;
```

How to Handle Influential Observations

1. Recheck the data to ensure that no transcription or data entry errors occurred.
2. If the data is valid, one possible explanation is that the model is not adequate.
   - A model with higher-order terms, such as polynomials and interactions between the variables, might be necessary to fit the data well.
3. Determine the robustness of the inference by running the analysis both with and without the influential observations.
4. Robust Regression (Covered Later in Program)
5. Weighted Least Squares (WLS)

Chapter 5: Model Post-Fitting for Inference

5.1 Examining Residuals

5.2 Influential Observations

5.3 Collinearity

Illustration of Collinearity

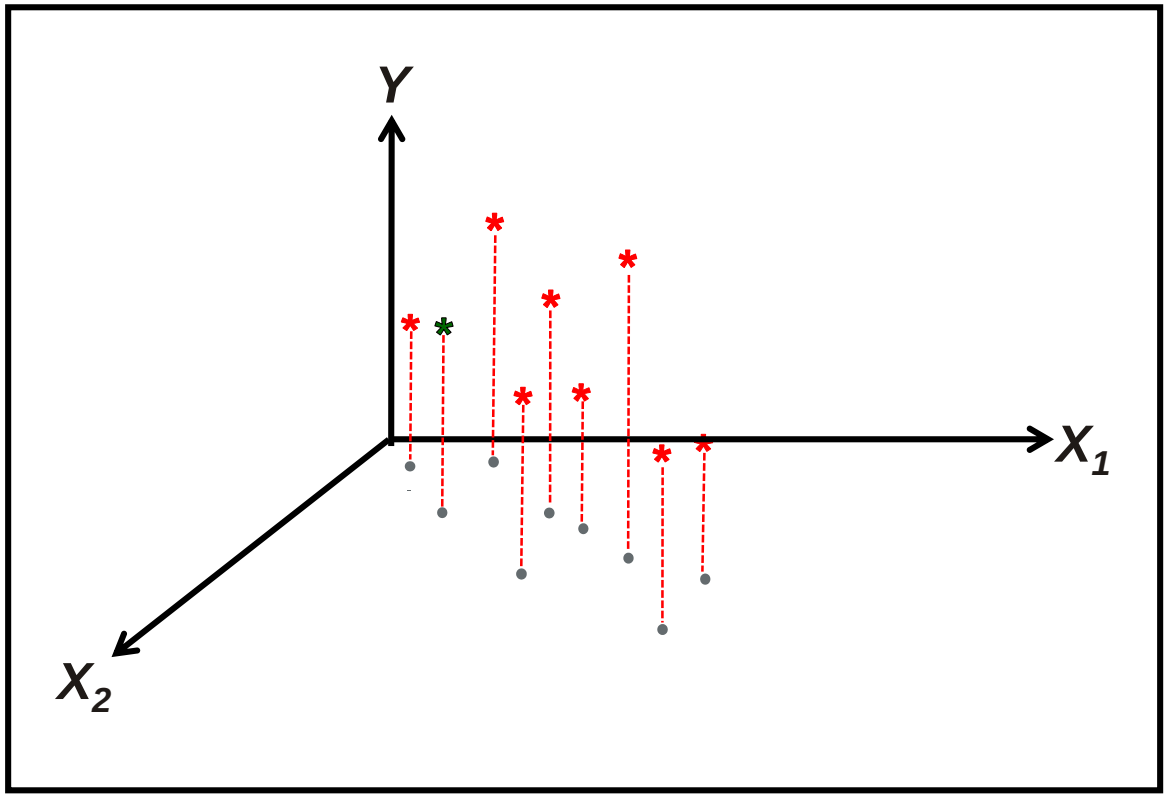


Illustration of Collinearity

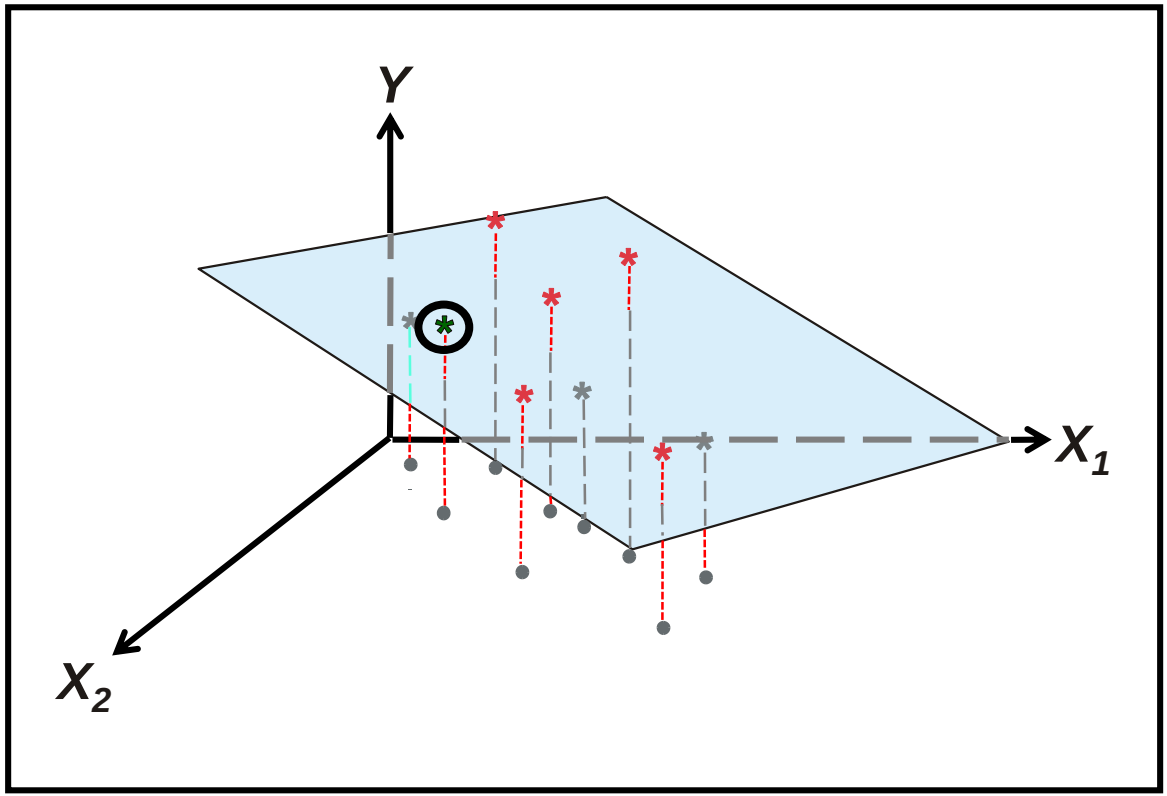
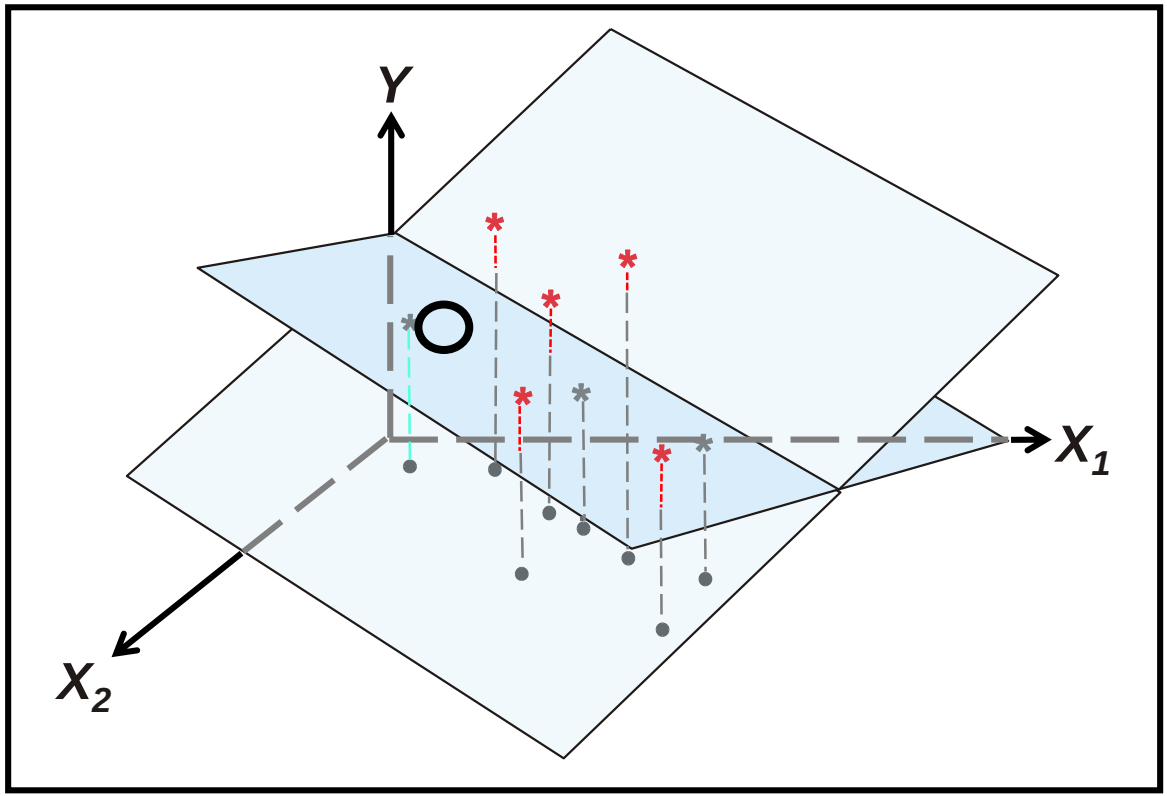


Illustration of Collinearity



Collinearity Diagnostics

PROC REG offers these tools that help quantify the magnitude of the collinearity problems and identify the subset of Xs that is collinear:

- VIF
- COLLIN
- COLLINOINT

This course focuses on VIF.

Variance Inflation Factor (VIF)

(1) The *VIF* is a relative measure of the increase in the variance because of collinearity. It can be thought of as this ratio:

$$VIF_i = \frac{1}{1 - R_i^2}$$

A $VIF_i > 10$ indicates that collinearity is potentially a problem.

(2) Correlation statistics:

Close to 1 or -1 indicate a high degree of linear relationship

Close to 0 indicate no clear linear relationship

COLLIN and COLLINOINT

The COLLIN and COLLINOINT are calculating the same information. The difference is that COLLIN includes the intercept in the calculations and the COLLINOINT excludes the intercept. This option in SAS produces the conditional index value.

- If the conditional index is between 10 and 30, this suggests weak dependencies
- If the conditional index is between 30 and 100, this suggests moderate dependencies
- If the conditional index is larger than 100, this suggests strong collinearity (also if the proportion of variation explained by PC is greater than 0.5 for a large conditional index

SAS Code

```
data ames3;
set stat1.ameshousing3;
BA2=Basement_Area**2;
BA3=Basement_Area**3;
run;
proc reg data=ames3;
  model SalePrice=Basement_Area BA2 BA3/vif collin collinoint;
run;
quit;
```

Dealing with Multicollinearity

Exclude redundant independent variables.

Redefine variables.

Use biased regression techniques (for example, LASSO).

Center the independent variables in polynomial regression models.

Center Variables

First take a look at VIF when variables are NOT centered

Use the STDIZE procedure with the METHOD=mean option.

Use SAS DATA steps to subtract the means from the variables.

```
proc stdize data=stat1.ameshousing3 method=mean
     out=ameshousing3_center (rename=(Basement_Area= Center_BA));
     var Basement_Area;
run;
data ameshousing3_center;
     set ameshousing3_center;
     center_BA2=Center_BA**2;
     center_BA3=Center_BA**3;
run;
```

Center Variables continued

```
proc reg data=ameshousing3_center;
   model SalePrice = Center_BA Center_BA2 Center_BA3 /
                     vif;
   title 'Centered Cubic Model';
run;
quit;
```

5.04 Multiple Choice Poll

Which of the following assumptions does collinearity violate?

a. Independent errors

b. Constant variance

c. Normally distributed errors

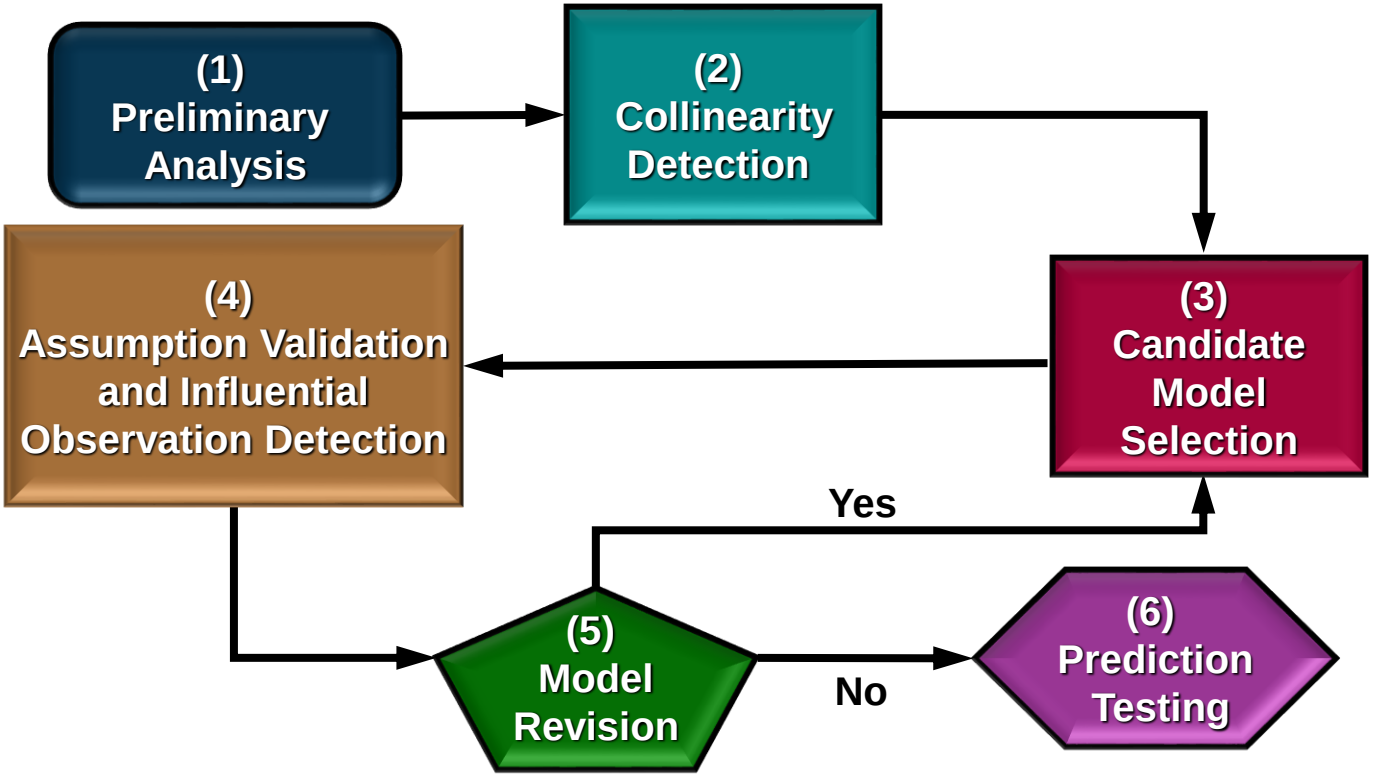
d. None of the above

5.05 Poll

If there is no correlation among the predictor variables, can there still be collinearity in the model?

- Yes
- No

An Effective Modeling Cycle



Scoring Recipe

- The model results in a formula or rules.
- The data require modifications.
  - Derived inputs
  - Transformations
  - Missing value imputation
- The scoring code is deployed.
  - To score, you do not rerun the algorithm; apply score code (equations) obtained from the final model to the scoring data.

Ways to Score Using PROC GLMSELECT

There are several options:

- Use a SCORE statement in PROC GLMSELECT.
- Use a STORE statement in PROC GLMSELECT and then a SCORE statement in PROC PLM.
- Use a STORE statement in PROC GLMSELECT and then a CODE statement in PROC PLM to output SAS code and then score in a DATA step.

SAS Code

```
proc glmselect data=ameshousing3_train valdata=ameshousing3_valid;
  class &categorical/param=glm ref=first;
  model SalePrice= &interval &categorical/selection=backward select=sbc
choose=validate;
  store out=amesstore;
  title “Selecting the Best Model using Validation data”;
run;
```

****Note: IF you use validation data to select model (whether to tune model or select the best model, then it is best to have a test data to truly see how well the model predicts.

SAS Code

```
proc plm restore=amesstore;
  score data=stat1.ameshousing4 out=scored;
  code file="&homefolder\scoring.sas";
run;

data scored2;
  set stat1.ameshousing4;
  %include "&homefolder\scoring.sas";
run;

proc compare base=scored compare=scored2 criterion=0.0001;
var Predicted;
with P_SalePrice;
run;
```

Scoring data

- In PROC GLM, you can also use the “store” command and then use PROC PLM.
- In PROC REG, you need to put outset=SASDATASET in the PROC REG statement. Then you can use PROC SCORE.

```
proc reg data=ameshousing3_train outest=estforscore;
  model SalePrice=&interval;
run;
quit;

proc score data=stat1.ameshousing4 score=estforscore type=parms out=scored3;
var &interval;
  run;
```

Questions?