USC
School of Business

Data Driven Decision Making: Multiple Linear Regression Analysis II

GSBA 545, Fall 2021

Professor Dawn Porter

Multiple Linear Regression II

- Dummy (Indicator) Variables
- Ordinal & Nominal Variables
- Heteroscedasticity
- Appendix
  - Residual Plots & Analysis
  - Influence Analysis & Extreme Values
  - Heteroscedasticity Example
  - Interaction Effects
  - Non-linearity

Qualitative Predictors

1. Involves categorical X variable with 2 Levels
   - e.g., Manager vs Non-manager, College-No College etc.
2. Variable levels are coded 0 & 1
3. Assumes only intercept is different between two
   - Slopes are constant across categories
4. Dummy-Variable (or Indicator-Variable) model
   - One or more of the *X* variables are dichotomous:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 \text{ College} + \ldots + \beta_k X_k + \varepsilon$$

Interpretation

$$\widehat{Salary} = b_0 + b_1 Exp + b_2 Mgt$$

Mgt = 1 if employee is a manager and 0 if not

Non-managers (*Mgt* = 0):

$$\widehat{Salary}_{Mgt=0} = b_0 + b_1 Exp + b_2(0) \rightarrow \widehat{Salary}_{Mgt=0} = b_0 + b_1 Exp$$

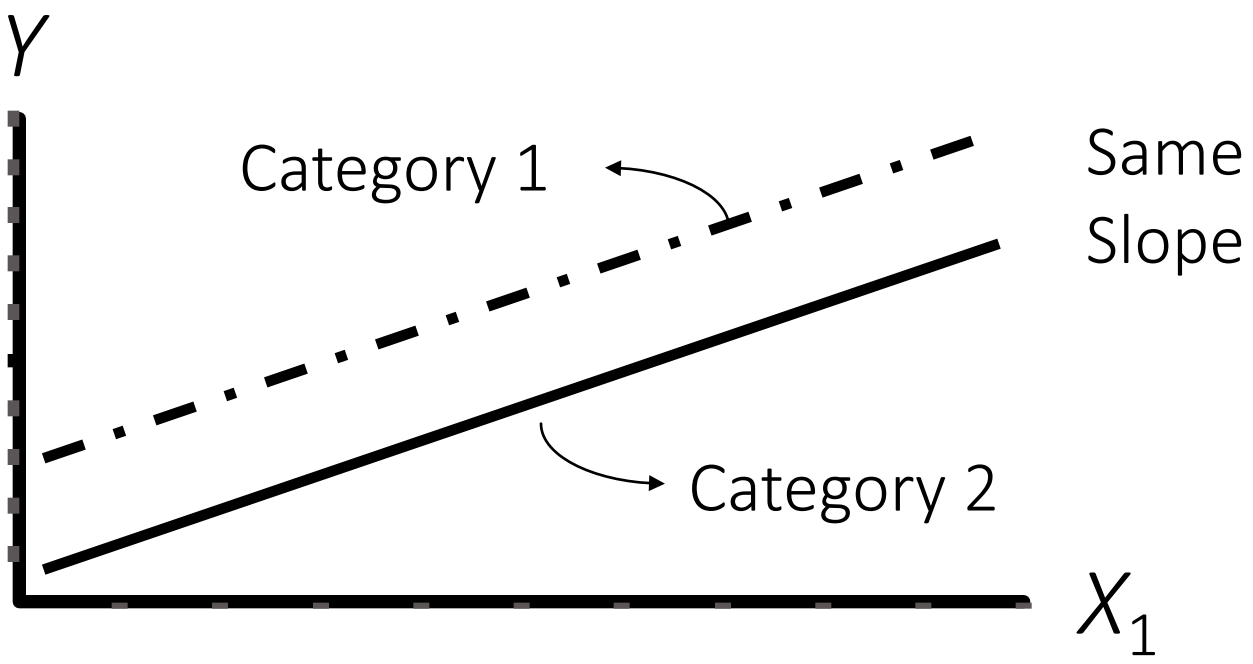
Same Slopes

Manager (*Mgt* = 1):

$$\widehat{Salary}_{Mgt=1} = b_0 + b_1 Exp + b_2(1) \rightarrow \widehat{Salary}_{Mgt=1} = (b_0 + b_2) + b_1 Exp$$

*Note: *Exp* refers to the number of years of experience in a particular industry.

Qualitative Predictors



Different
Y-Intercepts

Salary Data: Experience & Mgt

VARIABLES

Salary (Numerical):	Annual salary
Exp (Numerical):	Years of experience in industry
Mgt (Categorical):	Manager or not (1/0)
Educ (Categorical):	Education level (1=HS, 2=College, 3=Grad)

Although years of experience should be useful for predicting salary, can we get a better model by incorporating whether an employee is a manager or not? This is accomplished by using a dummy variable for Mgt.

$$Salary = \beta_0 + \beta_1 Exp + \beta_2 Mgt + \varepsilon$$

USC School of Business

Salaries: Exp & Mgt (Excel & JMP)

SUMMARY OUTPUT					
Regression Statistics					
Multiple R	0.9302				
R Square	0.8653				
Adj R Square	0.8591				
Standard Error	17,706.06				
Observations	46				
ANOVA					
	df	*SS*	*MS*	*F*	*Significance F*
Regression	2	86,629,061,467.54	43,314,530,734	138.1624	1.8992E-19
Residual	43	13,480,696,256.38	313,504,564		
Total	45	1.0011E+11			
	Coefficients	*Standard Error*	*t Stat*	*P-value*	
Intercept	102,103.567	5,259.995	19.4113	6.7485E-23	
Experience	5,271.081	511.063	10.3140	3.3497E-13	
Management	71,450.151	5,273.204	13.5497	4.0541E-17	

Summary of Fit

RSquare	0.865
RSquare Adj	0.859
Root Mean Square Error	17706.06
Mean of Response	172702.0
Observations (or Sum Wgts)	46.000

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	2	8.6629e+10	4.331e+10	138.1624
Error	43	1.3481e+10	313504564	Prob > F
C. Total	45	1.0011e+11		<.0001

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	102103.57	5259.995	19.41	<.0001
Experience	5271.081	511.063	10.31	<.0001
Management	71450.151	5273.204	13.55	<.0001

Salary Data: Experience & Mgt

$$\widehat{Salary} = 102103.57 + 5271.08Exp + 71450.15Mgt$$

Mgt = 1 if employee is a manager and 0 if not

Non-managers (*Mgt* = 0): $\widehat{Salary} = 102103.57 + 5271.08Exp + 71450.15(0)$

$$\rightarrow \widehat{Salary}_{Mgt=0} = 102{,}103.57 + 5271.08Exp$$

Different Intercepts

Same Slopes

Manager (*Mgt* = 1): $\widehat{Salary} = 102103.57 + 5271.08Exp + 71450.15(1)$

$$\rightarrow \widehat{Salary}_{Mgt=1} = 173{,}553.72 + 5271.08Exp$$

Nominal Variables: Hybrid Car MPG

VARIABLES

MPG (Numerical):	Average miles per gallon
MSRP (Numerical):	Retail price in 2013
Accelrate (Numerical):	Acceleration rate (km/hr per sec)
Car Class (Categorical):	Compact (C), Midsize (M), Two-seater (TS), Large (L), Pickup Truck (PT), Minivan (MV), SUV

- *MSRP* and *Accelrate* should be useful for predicting *MPG*, but can we do better by incorporating which *Class* a car is?
- This is accomplished by creating 7 dummy variables for *Car Class* and using 7 – 1 = 6 of them explicitly.

$$Hybrid\ MPG = \beta_0 + \beta_1 MSRP + \beta_2 Accelrate + \beta_3 C + \beta_4 M + \beta_5 TS + \beta_6 L + \beta_7 PT + \beta_8 MV + \varepsilon$$

Nominal Variables: Hybrid Car MPG

$$Hybrid\ MPG = \beta_0 + \beta_1 MSRP + \beta_2 Accelrate + \beta_3 C + \beta_4 M + \beta_5 TS + \beta_6 L + \beta_7 PT + \beta_8 MV + \varepsilon$$

- Indicator variables uncover "base" levels (when indicators are 0).
- Difference between coefficients of *Car Class* variables shows the value, with respect to *MPG*, of each class of car.

If the coefficient of any *Car Class* variable is statistically significant, there is evidence of a significant additive difference between that type and the base level (*SUV*).

*Note: This is only true for *fixed levels* of the other variables.

Nominal Vars: Hybrid Car MPG (JMP)

It actually appears that all categories of *Car Class* are statistically significant.

Pickup Truck, though, has a negative slope (with *SUV* being the baseline), so is predicted to have lower *MPG* than *SUV*s.

Summary of Fit

RSquare	0.622
RSquare Adj	0.601
Root Mean Square Error	6.934
Mean of Response	34.797
Observations (or Sum Wgts)	153.000

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	8	11376.766	1422.10	29.5779
Error	144	6923.463	48.08	Prob > F
C. Total	152	18300.229		<.0001*

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	39.089	3.245	12.05	<.0001*
MSRP(1000s)	-0.119	0.042	-2.81	0.0056*
Accelrate	-0.573	0.295	-1.94	0.0537
Compact	13.416	1.828	7.34	<.0001*
Midsize	8.417	1.498	5.62	<.0001*
Two Seater	17.624	2.860	6.16	<.0001*
Large	6.281	2.976	2.11	0.0366*
Pickup Truck	-9.296	3.078	-3.02	0.0030*
Minivan	18.217	3.833	4.75	<.0001*

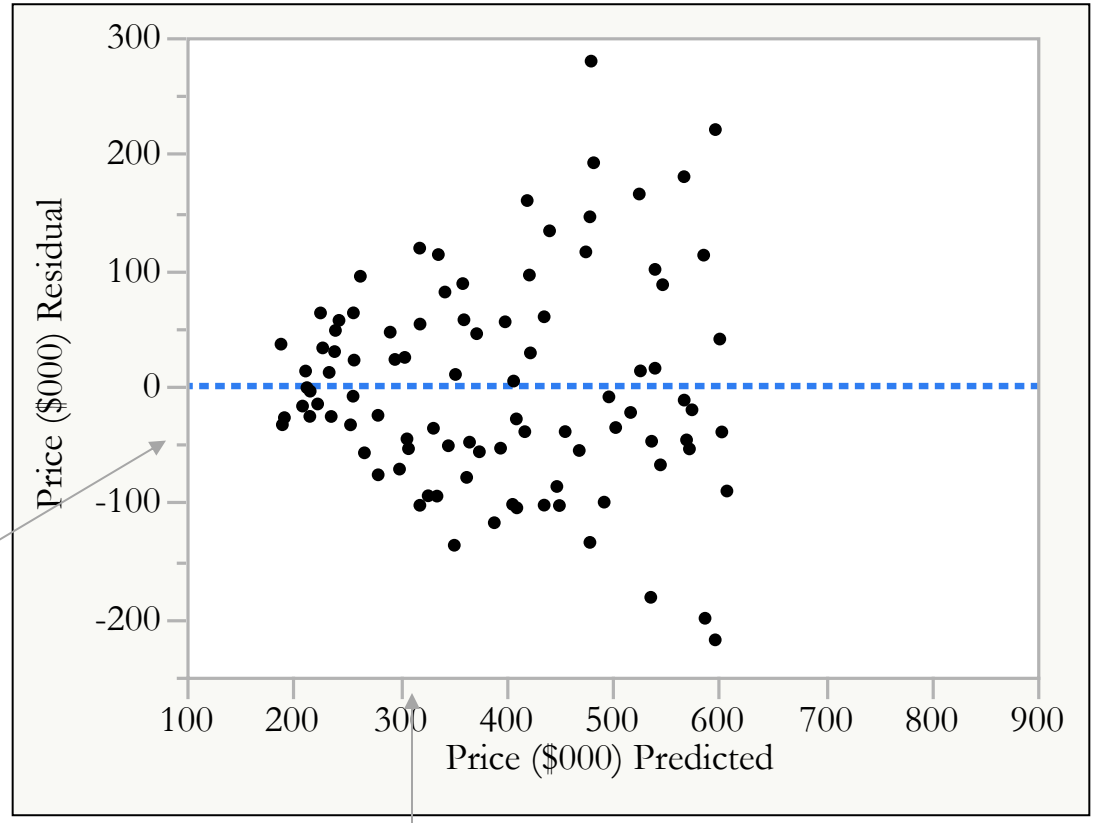
Nominal Variables: Hybrid Cars

$Compact$:	$MPG = 52.505 - 0.119MSRP - 0.573Accelrate$
$Midsize$:	$MPG = 47.506 - 0.119MSRP - 0.573Accelrate$
Two Seater:	$MPG = 56.713 - 0.119MSRP - 0.573Accelrate$
$Large$:	$MPG = 44.370 - 0.119MSRP - 0.573Accelrate$
$Pickup\ Truck$:	$MPG = 29.793 - 0.119MSRP - 0.573Accelrate$
$Minivan$:	$MPG = 57.306 - 0.119MSRP - 0.573Accelrate$
SUV:	$MPG = 39.089 - 0.119MSRP - 0.573Accelrate$

Residual Plots

Residual by Predicted plots:

- Display potential model violations.
- Simplify multiple variables into two dimensions by converting to errors and predicted values instead.



Y-axis represents the error, or residual value, for each observation.

X-axis represents the predicted, or fitted value, for each observation.

Numerical Output

Regression Analysis 1

RSquare	0.667
RSquare Adj	0.629
Root Mean Square Error	1.237
Mean of Response	7.501
Observations (or Sum Wgts)	11.000

F Ratio	Prob > F
17.990	0.0022*

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	3.000	1.125	2.67	0.0257*
X1	0.500	0.118	4.24	0.0022*

Regression Analysis 2

RSquare	0.666
RSquare Adj	0.629
Root Mean Square Error	1.237
Mean of Response	7.501
Observations (or Sum Wgts)	11.000

F Ratio	Prob > F
17.966	0.0022*

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	3.001	1.125	2.67	0.0258*
X2	0.500	0.118	4.24	0.0022*

Regression Analysis 3

RSquare	0.666
RSquare Adj	0.629
Root Mean Square Error	1.236
Mean of Response	7.500
Observations (or Sum Wgts)	11.000

F Ratio	Prob > F
17.972	0.0022*

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	3.002	1.124	2.67	0.0256*
X3	0.500	0.118	4.24	0.0022*

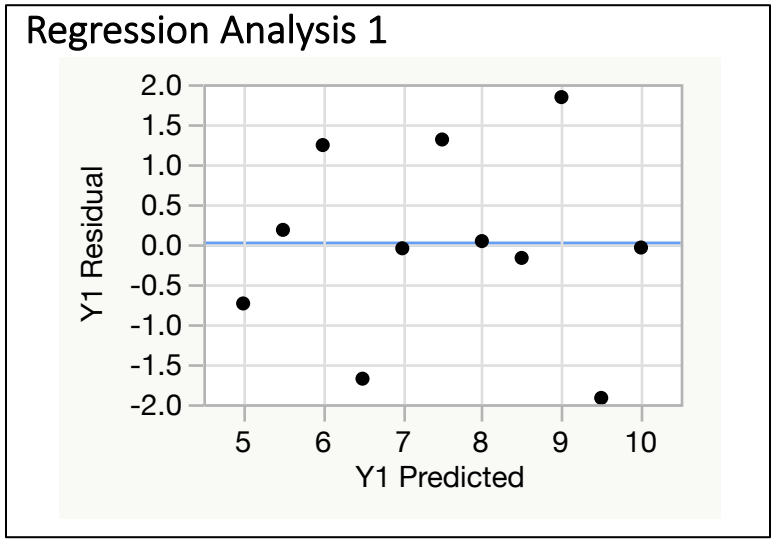
Regression Analysis 4

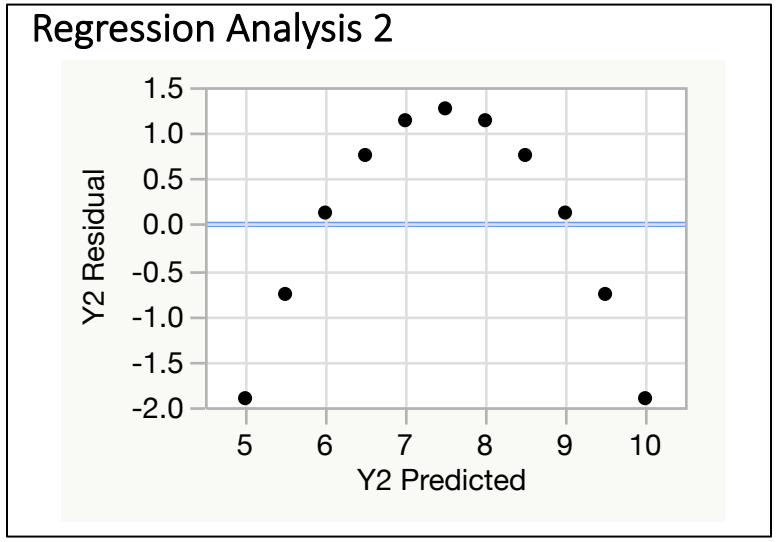
RSquare	0.667
RSquare Adj	0.630
Root Mean Square Error	1.236
Mean of Response	7.501
Observations (or Sum Wgts)	11.000

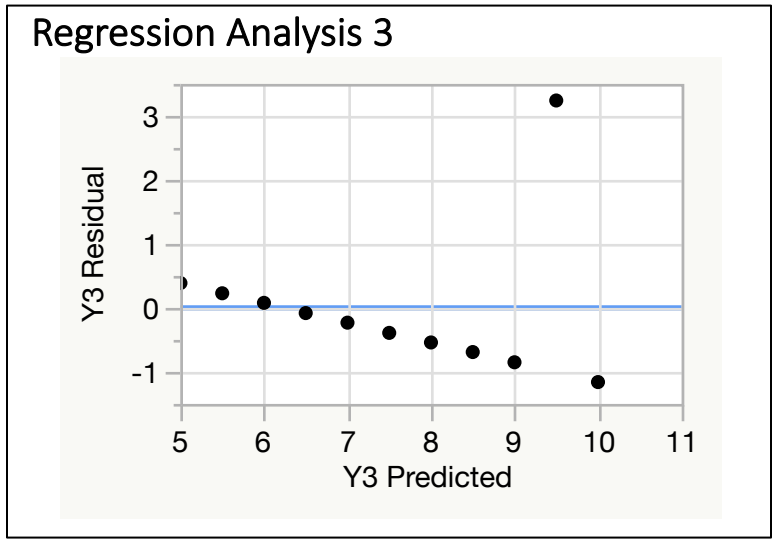
F Ratio	Prob > F
18.003	0.0022*

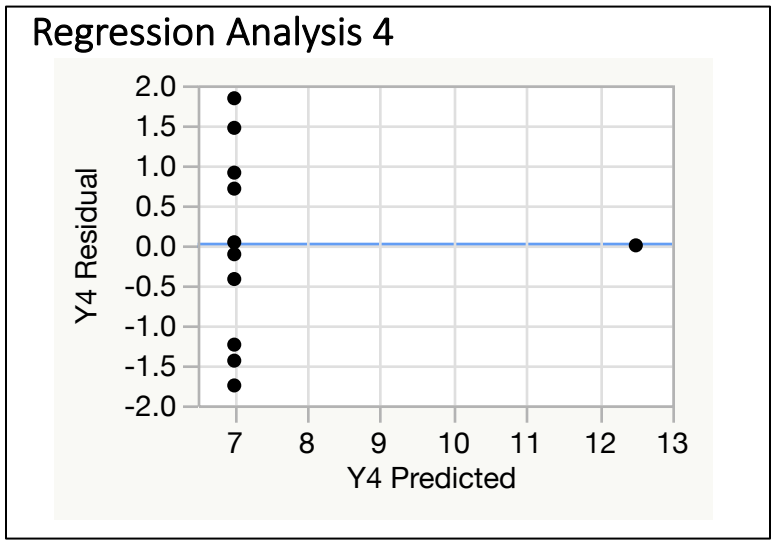
Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	3.002	1.124	2.67	0.0256*
X4	0.500	0.118	4.24	0.0022*

Graphical Output









Residual Plots: Graphs

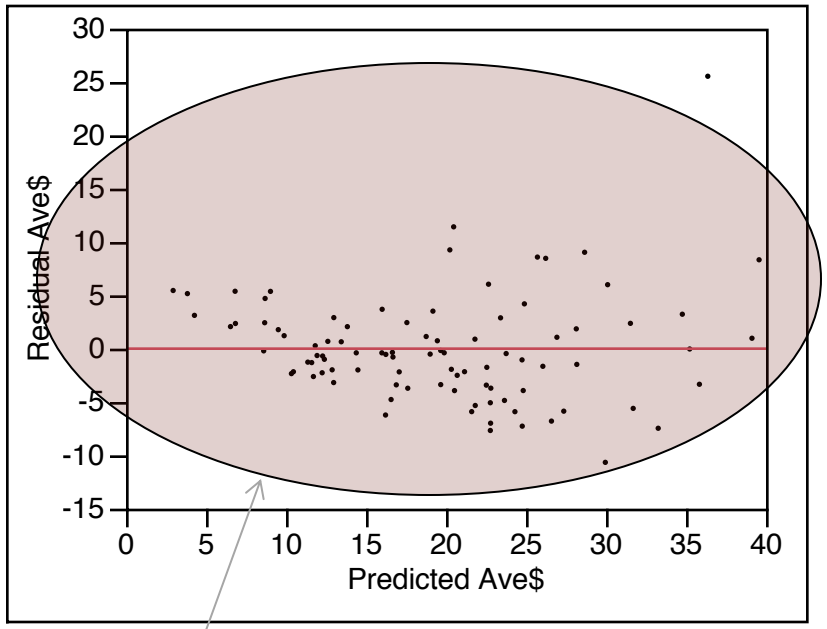
What to look for in **Residual by Predicted plots**:

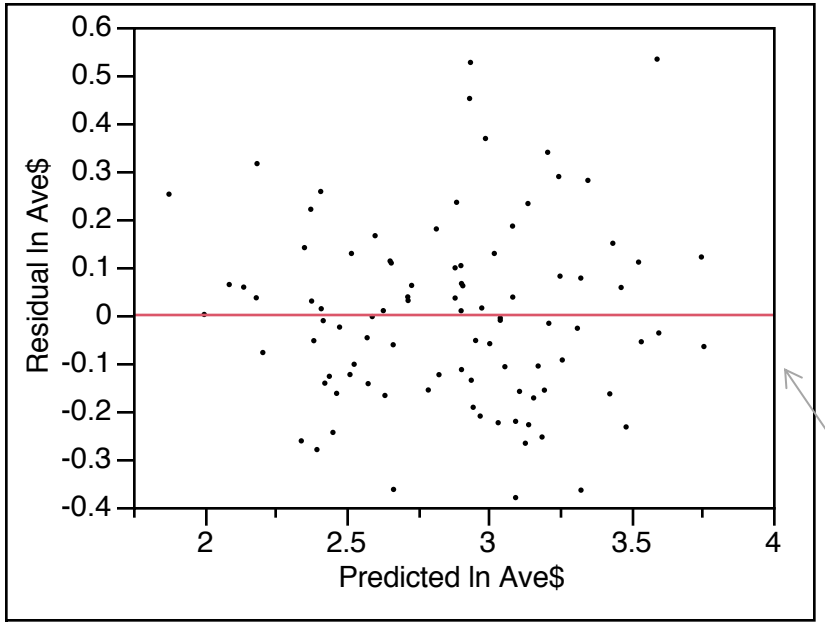
1. **Heteroscedasticity**
   Violation of constant variance assumption

2. **Potential outliers**
   If some are noticed, check the *standardized residuals* and/or *Cook's Distance* to verify

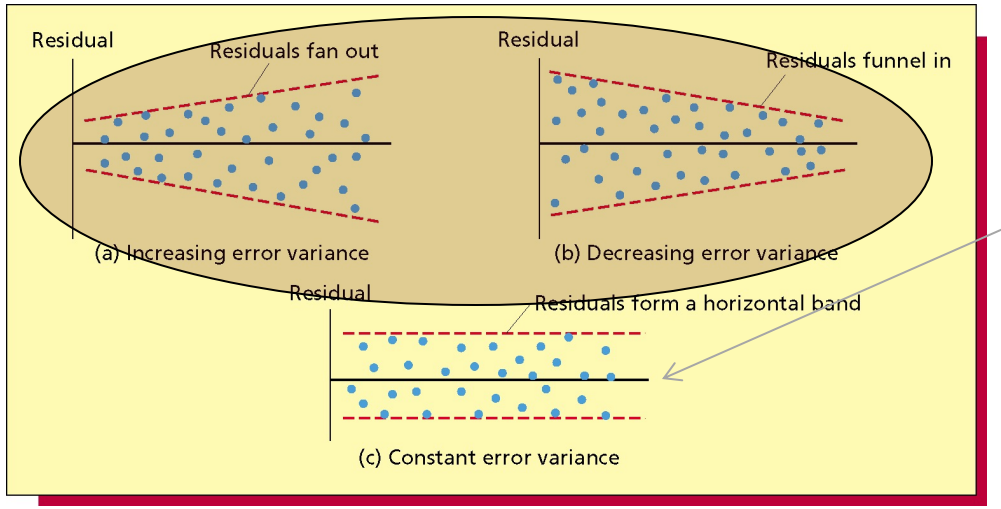
3. **Non-linear patterns**
   Examine individual scatterplots to see where non-linearity is apparent
   Should some variables be transformed to better represent relationships?

Non-Constant Variance





Heteroscedastic (non-constant variance)



Homoscedastic (constant variance)

Non-Constant Variance

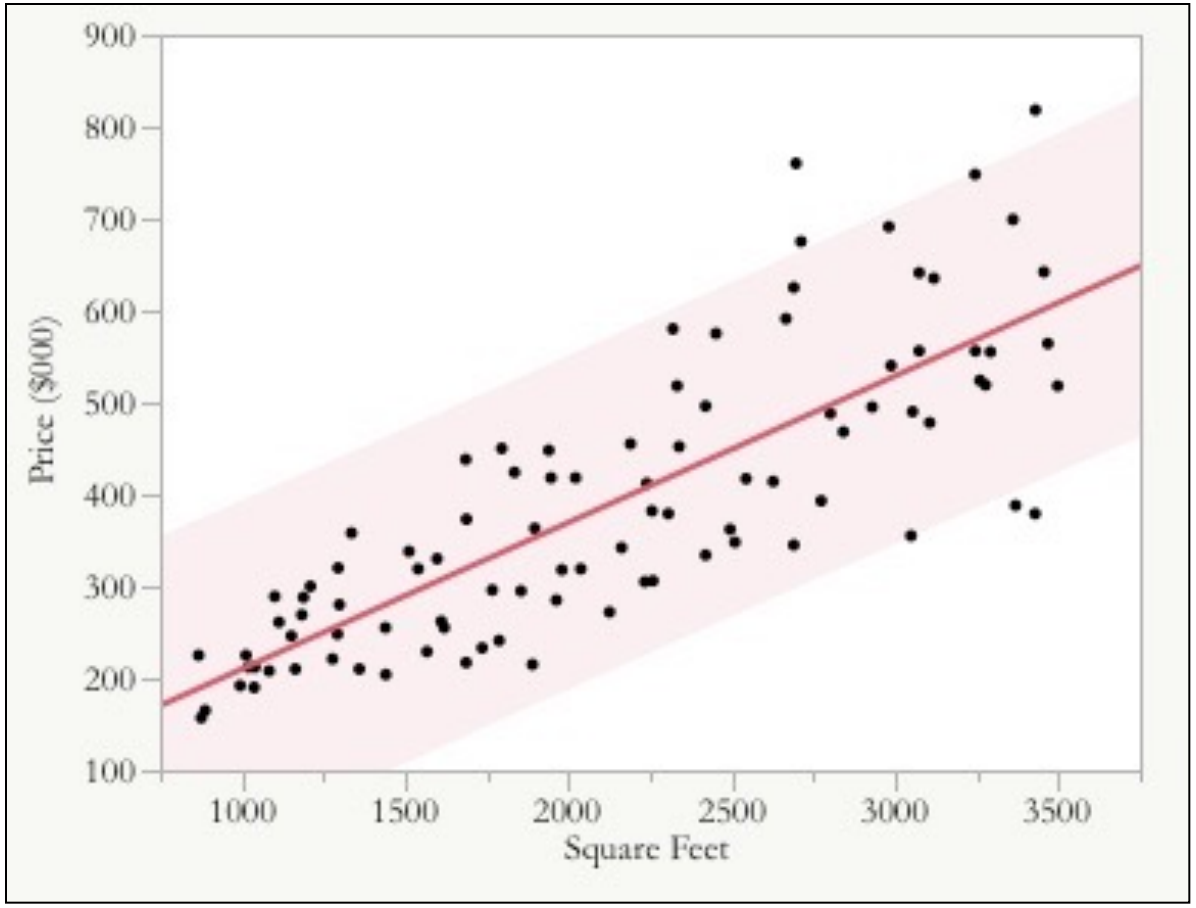
Why is non-constant variance, or heteroscedasticity, a problem, and where do we see the consequences?

1. **Prediction Intervals for Estimates**
   Can be too narrow or too wide at various levels of X

2. **Confidence Intervals for Slopes**
   Not reliable because the slope standard errors are possibly inaccurate

3. **Slope and Intercept Testing**
   T-statistics for slopes and the intercept are not reliable

Heteroscedasticity



As noted, the 95% PI is too narrow in some places and too wide in others.

Appendix:

- Residual Plots
- Influence Analysis & Extreme Values
- Heteroscedasticity Example
- Non-linear patterns
  - Examining individual scatterplots to check for curvature
  - Considering transformations to better represent relationships

Residual Plots: Graphs

What to look for in **Residual by Predicted plots**:

1. **Heteroscedasticity**
   Violation of constant variance assumption

2. **Potential outliers**
   If some are noticed, check the *standardized residuals* and/or *Cook's Distance* to verify

3. **Non-linear patterns**
   Examine individual scatterplots to see where non-linearity is apparent
   Should some variables be transformed to better represent relationships?

Residual Plots: Graphs

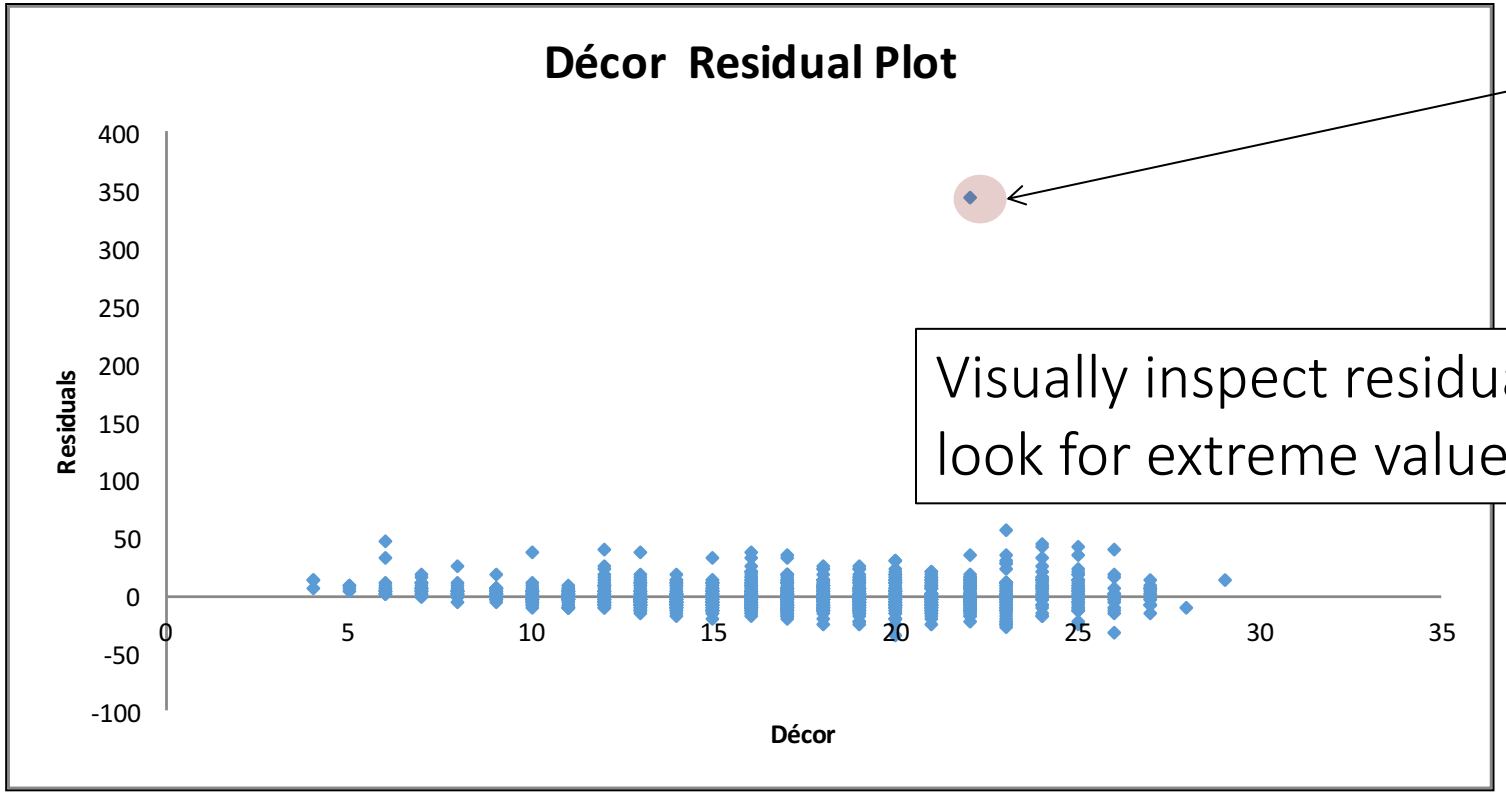
1. Graphically examine observations that may be strongly affecting coefficients
   - Some outliers may not have a large influence on output.
   - It's best to rely on measures such as **standardized residuals** (Excel) and (JMP) to assess if extreme values have a significant impact on the regression model results.

2. Try to understand why influential observations occurred
   - Are potential outliers unlikely to occur again?
   - Example: Data collected over a period when a union strike occurred.

3. **Cautiously** consider deleting one observation at a time

Influence Analysis: Excel

LA Zagat restaurant data: Is a restaurant's *Décor* rating significantly related to its average *Price* per meal?



Influence Analysis: Excel

Guideline: Examine residuals that are more than 2 or 2.5 standard deviations away from their predicted values.

In **Data Analysis – Regression** window, ask for *Standardized Residuals* to get a list of normalized error values.

Urasawa is clearly an extreme value!

RESIDUAL OUTPUT

Observation	*Predicted Price*	*Residuals*	*Standard Residuals*
1	30.198	-2.198	-0.154
2	42.147	10.853	0.761
3	25.418	11.582	0.813
4	37.368	-2.368	-0.166
5	34.978	10.022	0.703
6	37.368	-12.368	-0.868
7	37.368	-11.368	-0.797
8	30.198	-6.198	-0.435
331	51.707	45.293	3.177
332	49.317	5.683	0.399
333	6.298	8.702	0.610
790	18.248	-2.248	-0.158
791	54.097	42.903	3.010
792	11.078	6.922	0.486
793	30.198	-1.198	-0.084
813	32.588	16.412	1.151
814	49.317	55.683	3.906
815	25.418	-2.418	-0.170
816	20.638	-11.638	-0.816
1158	37.368	6.632	0.465
1159	46.927	343.073	24.067
1160	30.198	-11.198	-0.786
1161	34.978	-12.978	-0.910
1162	51.707	20.293	1.424
1163	37.368	-11.368	-0.797
1164	15.858	0.142	0.010
1165	30.198	-15.198	-1.066
1166	49.317	-19.317	-1.355
1167	44.537	-1.537	-0.108
1168	15.858	3.142	0.220

Influence Analysis: Excel

With Urasawa				
Regression Statistics				
Multiple R	0.628			
R Square	0.394			
Adj R Square	0.394			
Standard Error	14.261			
Observations	1242			
	Coefficients	***Standard Error***	***t Stat***	***P-value***
Intercept	-5.651	1.487	-3.801	0.000
Décor	2.390	0.084	28.416	0.000

Without Urasawa				
Regression Statistics				
Multiple R	0.733			
R Square	0.537			
Adj R Square	0.537			
Standard Error	10.410			
Observations	1241			
	Coefficients	***Standard Error***	***t Stat***	***P-value***
Intercept	-4.914	1.086	-4.526	0.000
Décor	2.330	0.061	37.937	0.000

The results of removing the outlier in this case (**Urasawa**) are rather dramatic.

- R^2 increased 14%
- Standard Error decreased by almost $4
- The slope standard error decreased by about 25%

Cook's Distance: JMP

Focus on the effect an influential point has on the regression.

- If removal of a point significantly changes the regression results, the observation has high influence and we *may* choose to take it out.

Common measure of influence is Cook's distance

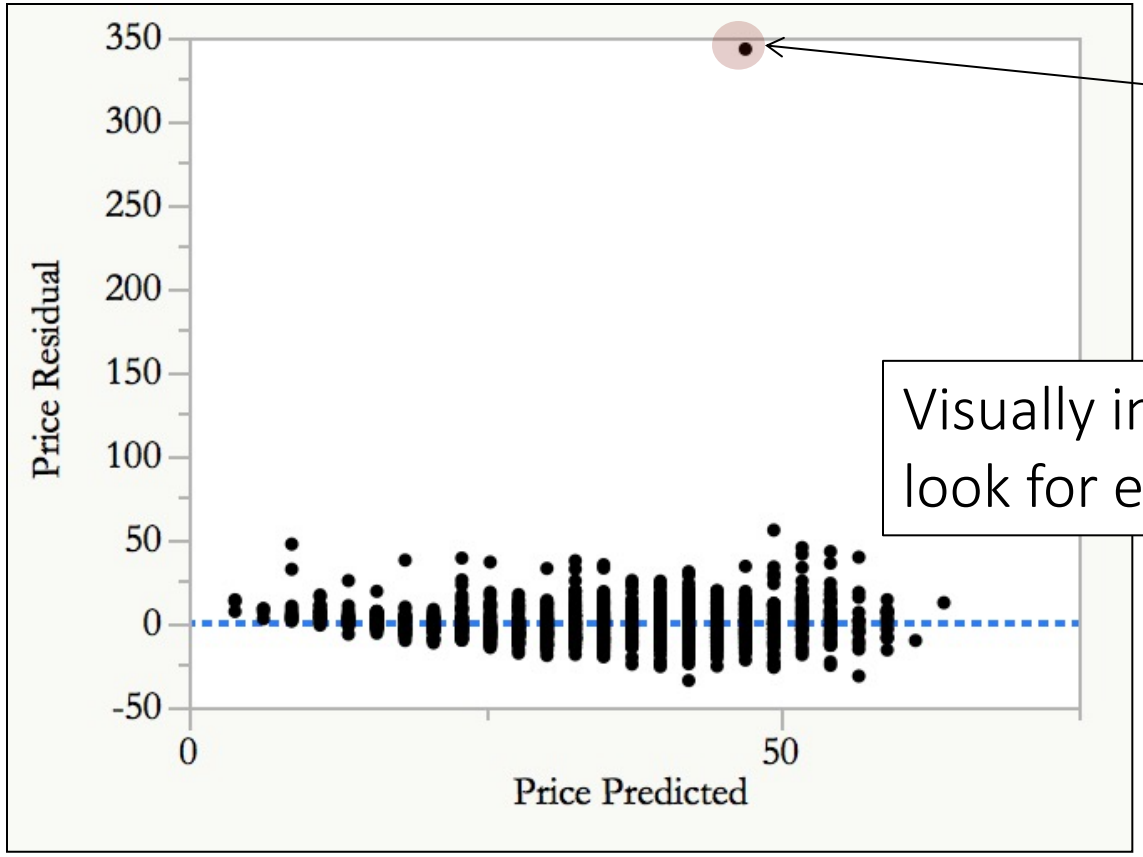
- Measures the scaled change in fitted regression coefficients after removal.
- Takes into consideration distances on both the *X*-axis and *Y*-axis.
- A value of Cook's D > 4/*n* (some use *D* > 1) indicates a problem.

Cook's D in JMP: After the regression is run, go to the red triangle and select the *Save Columns* option. There, choose *Cook's Distance*.

*Note: Omit observations one at a time and then re-assess the regression output.

Influence Analysis: JMP

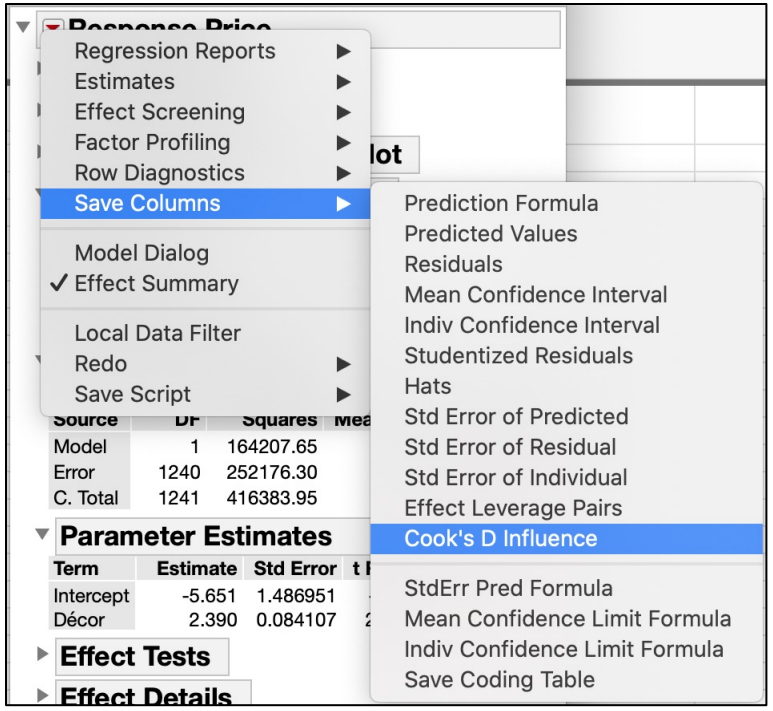
LA Zagat restaurant data: Is a restaurant's *Décor* rating significantly related to its average *Price* per meal?



Potential outlier

Visually inspect residual plots to look for extreme values.

Cook's Distance: JMP



Guideline: Examine residuals that have *Cook's D* values larger than 4/*n*, which is 4/1242 = 0.0032.

	Name	Décor	Price	Cook's D Influence Price
1151	Tulipano	13	32	0.0001
1152	Tuscany Il Risto	21	50	0.0001
1153	Twelve + Highl	20	46	0.0000
1154	Twin Palms	20	35	0.0001
1155	Typhoon	20	37	0.0001
1156	U-Zen	12	32	0.0003
1157	Ugo an Italian B	15	24	0.0001
1158	Ugo an Italian C	15	24	0.0001
1159	Ulysses Voyage	16	30	0.0000
1160	Ummba Grill	11	21	0.0000
1161	Uncle Bill's Pan	12	14	0.0003
1162	Uncle Darrow's	8	14	0.0000
1163	Upstairs 2	18	44	0.0001
1164	Urasawa	22	390	0.4850
1165	Urth Caffé	15	19	0.0003
1166	uWink	17	22	0.0003
1167	Valentino	24	72	0.0025

Urasawa is clearly an extreme value!

Influence Analysis: JMP

Summary of Fit

RSquare	0.394
RSquare Adj	0.394
Root Mean Square Error	14.261
Mean of Response	35.006
Observations (or Sum Wgts)	1242.000

With Urasawa

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	-5.651	1.487	-3.801	0.0002*
Décor	2.390	0.084	28.416	<.0001*

The results of removing the outlier in this case (**Urasawa**) are rather dramatic.

- R^2 increased 14%
- Standard Error decreased by almost $4
- The slope standard error decreased by about 25%

Summary of Fit

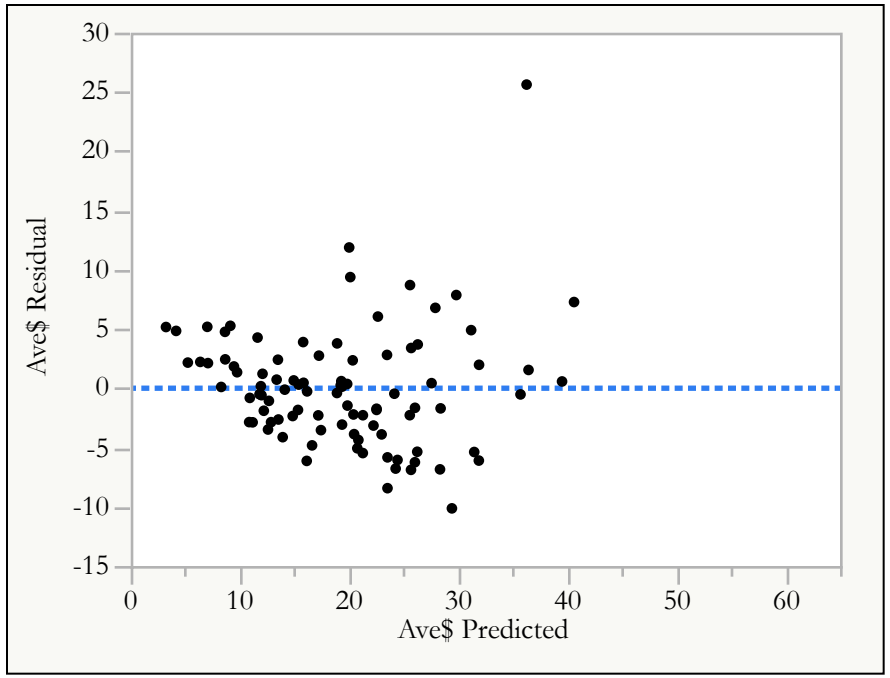
RSquare	0.537
RSquare Adj	0.537
Root Mean Square Error	10.410
Mean of Response	34.720
Observations (or Sum Wgts)	1241.000

Without Urasawa

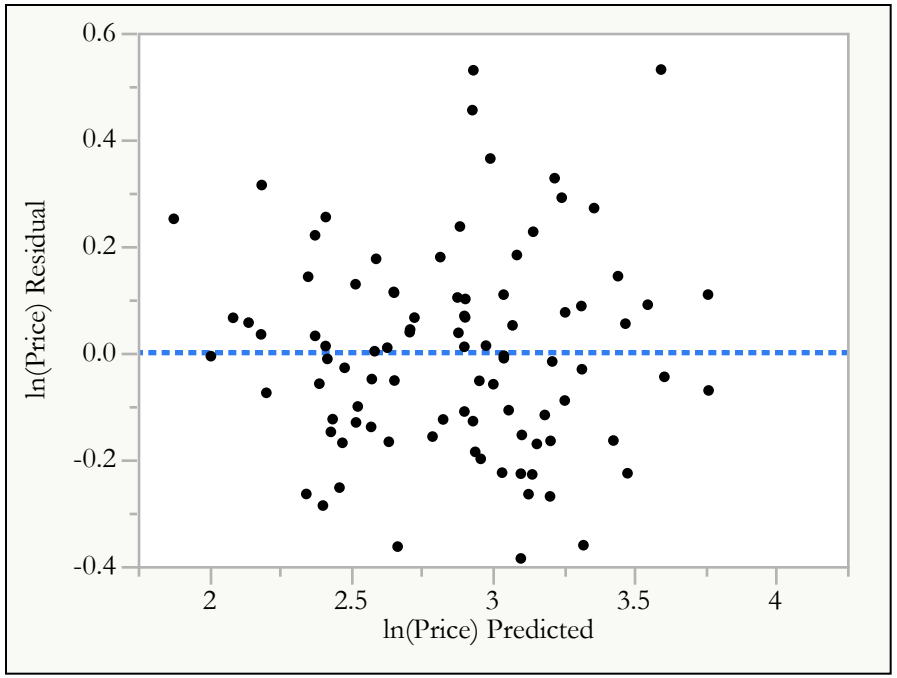
Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	-4.914	1.086	-4.526	0.000
Décor	2.330	0.061	37.937	0.000

Residual Analysis: Example



This pattern indicates **heteroscedasticity**, which is a violation of the constant variance assumption.



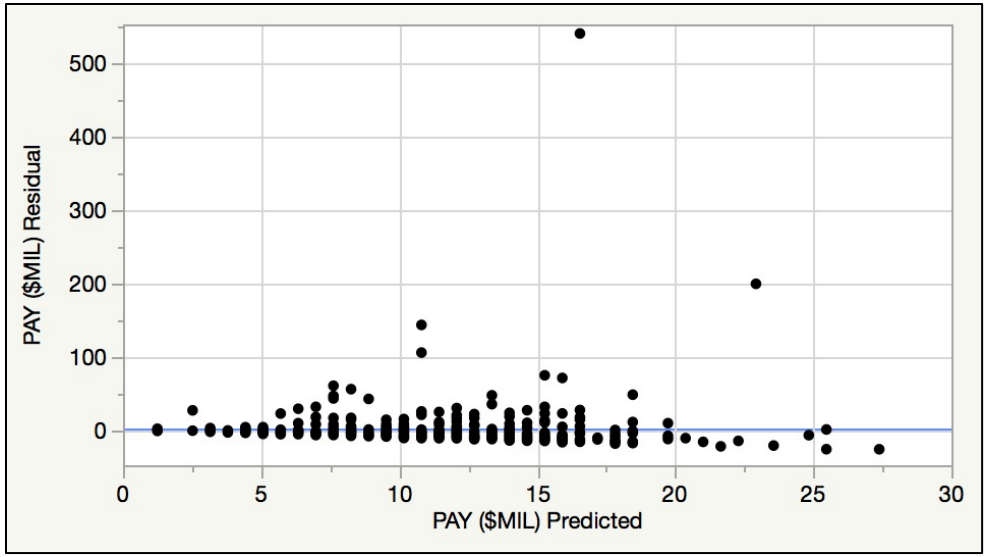
If the *Y* variable is right-skewed or some kind of financial variable, taking the *natural log* will often alleviate **heteroscedasticity**. The slope estimate can then be translated as a *percent change* in *Y.*

Original Model: CEO Pay

Original Model: The slope relates to a *dollar response ($Mil)* in *Y*.

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	-24.2569	11.9172	-2.04	0.0423*
AGE	0.6375	0.2116	3.01	0.0027*



$$\widehat{Pay} = -24.257 + 0.638Age$$

A 1 yr increase in *Age* should correspond to a **$638,000 increase** in *Pay*.

* **Nonconstant variance,** however, is quite apparent and needs to be remedied.

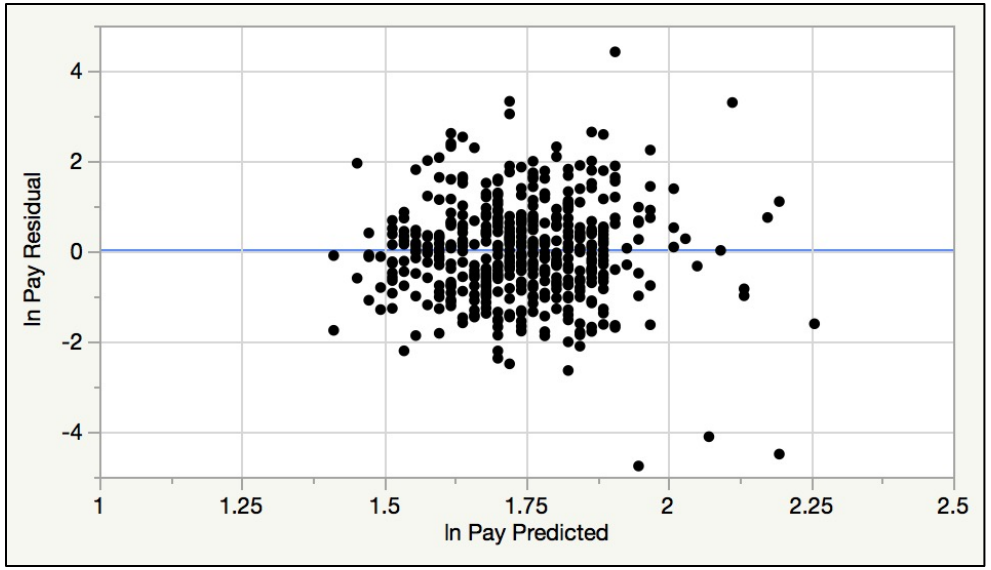
Natural Log Model: CEO Pay

Natural Log Model: The slope relates to a *percent change* in *Y*.

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	0.5876	0.4376	1.34	0.1800
AGE	0.0206	0.0078	2.65	0.0083*

$$\widehat{\ln Pay} = 0.5876 + 0.0206 Age$$

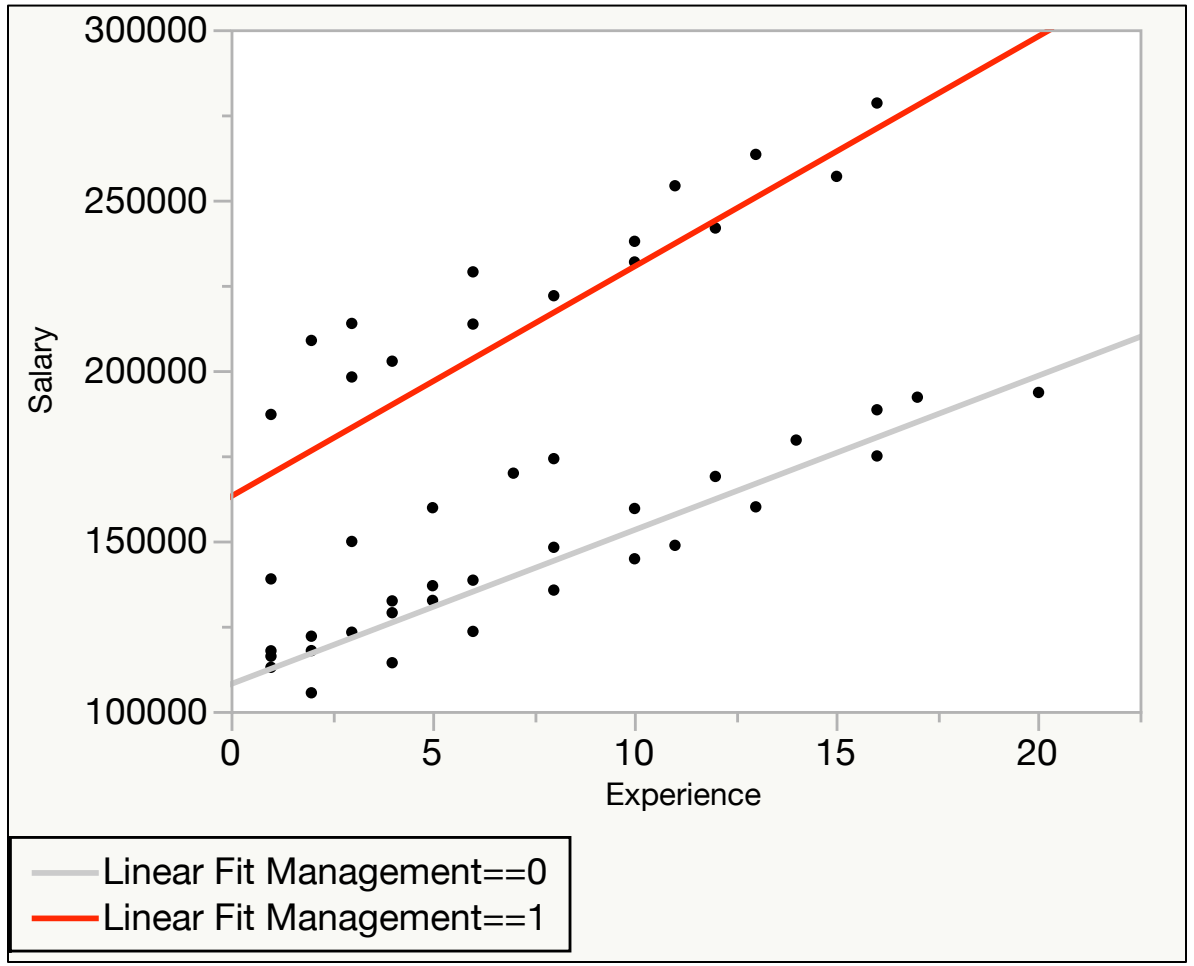


A 1 yr increase in *Age* should correspond to a **2.06% increase** in *Pay*.

* Now the residual plot exhibits **homoscedasticity** and the results are reliable.

Interaction Effects

Salary vs Experience & Management



Interactions: Experience & Mgt

The original model relied on an additive difference between groups forming the indicator variable. What if there is a more complex relationship?

- Significant indicator variables are equivalent to a constant shift in the model – two models have different intercepts.
- Perhaps the effect isn't additive, but a shift in the slope coefficient. In this situation, examine the two groups to assess each one's steepness of relationship to *Y.*

Is there an effect involving **both** *Management* position AND years of *Experience*?

$$Salary = \beta_0 + \beta_1 Exp + \beta_2 Mgt + \beta_3 (Exp \times Mgt) + \varepsilon$$

Interactions: Experience & Mgt

$$\widehat{Salary} = b_0 + b_1 Exp + b_2 Mgt + b_3 (Exp \times Mgt)$$

Non-managers (*Mgt* = 0):

$$\widehat{Salary} = b_0 + b_1 Exp + b_2 (0) + b_3 (Exp \times 0) \rightarrow \widehat{Salary} = b_0 + b_1 Exp$$

Different Intercepts

Different Slopes

Managers (*Mgt* = 1):

$$\widehat{Salary} = b_0 + b_1 Exp + b_2 (1) + b_3 (Exp \times 1) \rightarrow \widehat{Salary} = (b_0 + b_2) + (b_1 + b_3) Exp$$

Interactions: Experience & Mgt (JMP)

Summary of Fit

RSquare	0.879
RSquare Adj	0.870
Root Mean Square Error	17012.27
Mean of Response	172702.0
Observations (or Sum Wgts)	46.000

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	3	8.7954e+10	2.932e+10	101.3004
Error	42	1.2156e+10	289417279	Prob > F
C. Total	45	1.0011e+11		<.0001*

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	107859.11	5725.083	18.84	<.0001*
Experience	4526.582	601.807	7.52	<.0001*
Management	55017.708	9200.203	5.98	<.0001*
Exp*Mgt	2227.401	1040.938	2.14	0.0382*

Interactions: Experience & Mgt

$$\widehat{Salary} = 107859.11 + 4526.58Exp + 55017.71Mgt + 2227.40(Exp \times Mgt)$$

Non-managers ($Mgt = 0$):

$$\widehat{Salary} = 107859.11 + 4526.58Exp + 55017.71(0) + 2227.40(Exp \times 0)$$

$$\rightarrow \widehat{Salary} = 107859.11 + 4526.58\ Exp$$

Different Intercepts

Different Slopes

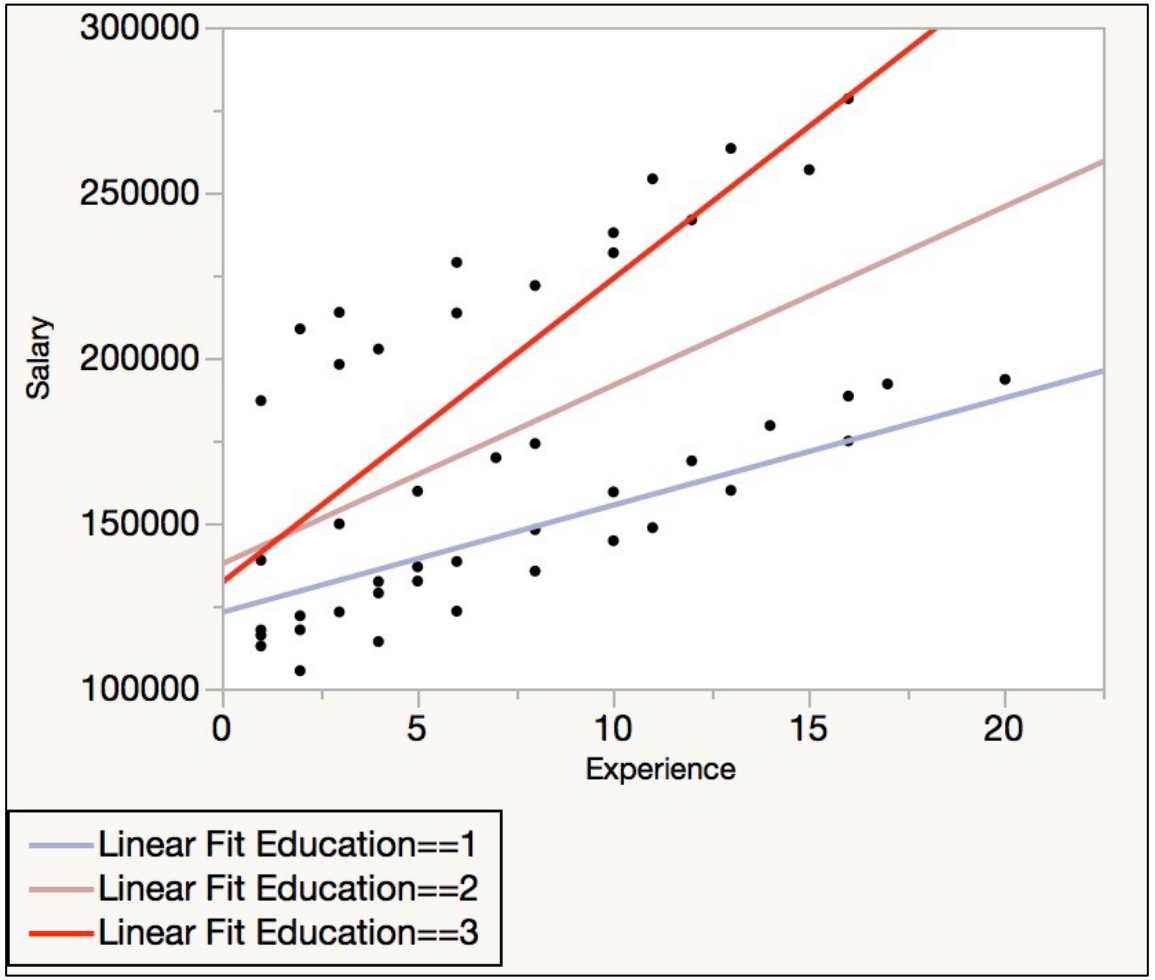
Managers ($Mgt = 1$):

$$\widehat{Salary} = 107859.11 + 526.58Exp + 55017.71(1) + 2227.40(Exp \times 1)$$

$$\rightarrow \widehat{Salary} = 162876.82 + 6753.98Exp$$

Interactions: Experience & Education

Salary vs Experience & Education Level



Interaction Interpretations

What are the possible combinations of additive and multiplicative relationships?

Underlying relationship:	*Variable significance*:
1. Intercepts are same, slopes are different.	1. Dummy term: No Interaction term: Yes
2. Intercepts are different, slopes are the same.	2. Dummy term: Yes Interaction term: No
3. Intercepts and slopes are different. Two separate Regressions.	3. Dummy term: Yes Interaction term: Yes
4. Intercepts and slopes are the same. Equivalent to a pooled model where the groups are indistinguishable.	4. Dummy term: No Interaction term: No

Residual Plots: Graphs

What to look for in **Residual by Predicted plots**:

1. **Heteroscedasticity**
   Violation of constant variance assumption

2. **Potential outliers**
   If some are noticed, check the *standardized residuals* and/or *Cook's Distance* to verify

3. **Non-linear patterns**
   Examine individual scatterplots to see where non-linearity is apparent
   Should some variables be transformed to better represent relationships?

Example: Car Price Data

Predicting Car Prices

VARIABLES

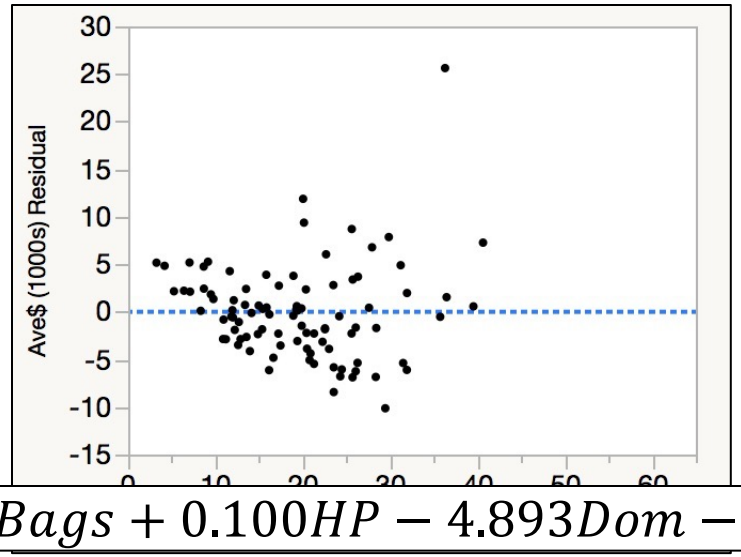
Ave$:	Average price for car model in $thousands
CityMPG:	Average MPG for car model while city driving
Air Bags:	Number of standard air bags for car model
HP:	Horsepower for car model
Domestic:	1 = domestic car, 0 = foreign car
Trans:	Transmission type: 1 = manual, 0 = automatic

What variables have the strongest relationship to *Ave$*?

Slope Values: Car Prices

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	12.873	4.995	2.58	0.0117
City MPG	-0.280	0.140	-2.01	0.0480
Air Bags	3.374	0.876	3.85	0.0002
HP	0.100	0.016	6.31	<.0001
Domestic	-4.893	1.230	-3.98	0.0001
Trans	-2.461	1.393	-1.77	0.0807



$$Ave\$ = 12.873 - 0.280CityMPG + 3.374AirBags + 0.100HP - 4.893Dom - 2.461Trans$$

For a 1 *City MPG* increase, we expect a **$280 decrease** in *Ave$*.

For an extra *air bag* in the car, we expect a **$3374 increase** in *Ave$*.

For one extra *HP* in the car, we expect a **$100 increase** in *Ave$*.

If the car is *domestic*, we expect a **$4893 decrease** in *Ave$*.

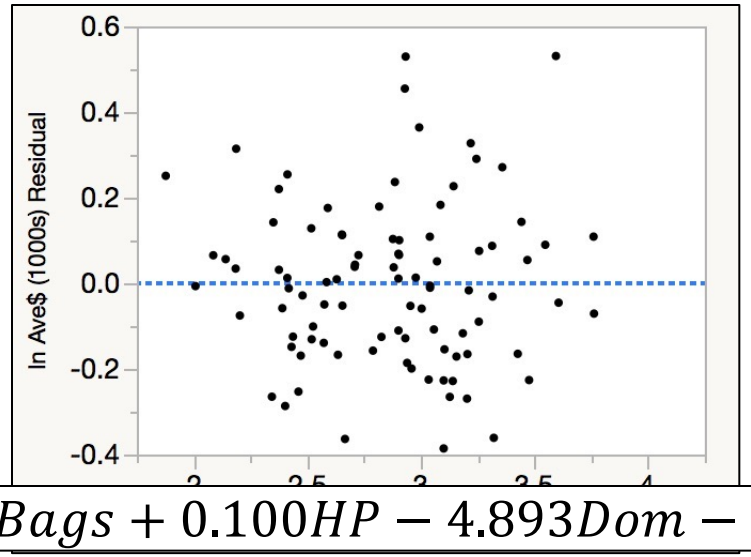
If the car is *manual*, we expect a **$2461 decrease** in *Ave$*.

*These estimates are accounting for the other *X*s in the model.

Slope Values: ln Car Prices

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	2.929	0.189	15.51	0.000
City MPG	-0.027	0.005	-5.02	0.000
Air Bags	0.173	0.033	5.21	0.000
HP	0.004	0.001	6.64	0.000
Domestic	-0.213	0.047	-4.58	0.000
Trans	-0.107	0.053	-2.03	0.045



$$Ave\$ = 12.873 - 0.280CityMPG + 3.374AirBags + 0.100HP - 4.893Dom - 2.461Trans$$

For a 1 *City MPG* increase, we expect a **2.75% decrease** in *Ave$*.

For an extra *air bag* in the car, we expect a **17.3% increase** in *Ave$*.

For one extra *HP* in the car, we expect a **0.4% increase** in *Ave$*.

If the car is *domestic,* we expect a **21.3% decrease** in *Ave$*.

If the car is *manual*, we expect a **10.7% decrease** in *Ave$*.

*These estimates are accounting for the other *X*s in the model.

Residual Plots: Graphs

What to look for in Residual by Predicted plots:

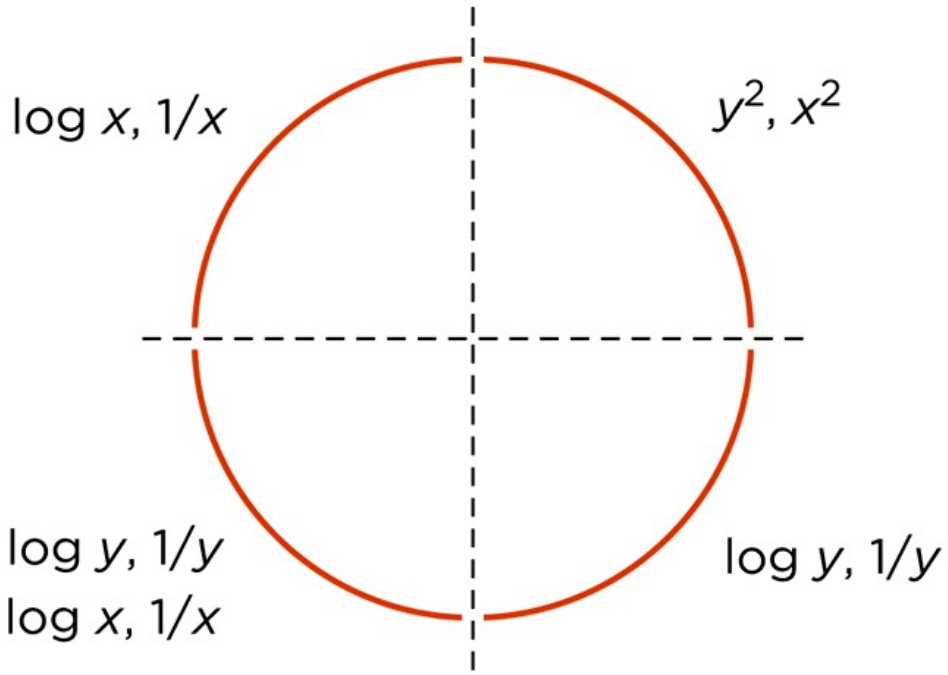
1. **Heteroscedasticity**
   Violation of constant variance assumption

2. **Potential outliers**
   If some are noticed, check the *standardized residuals* and/or *Cook's Distance* to verify

3. **Non-linear patterns**
   Examine individual scatterplots to see where non-linearity is apparent
   Should some variables be transformed to better represent relationships?

Potential Transformations

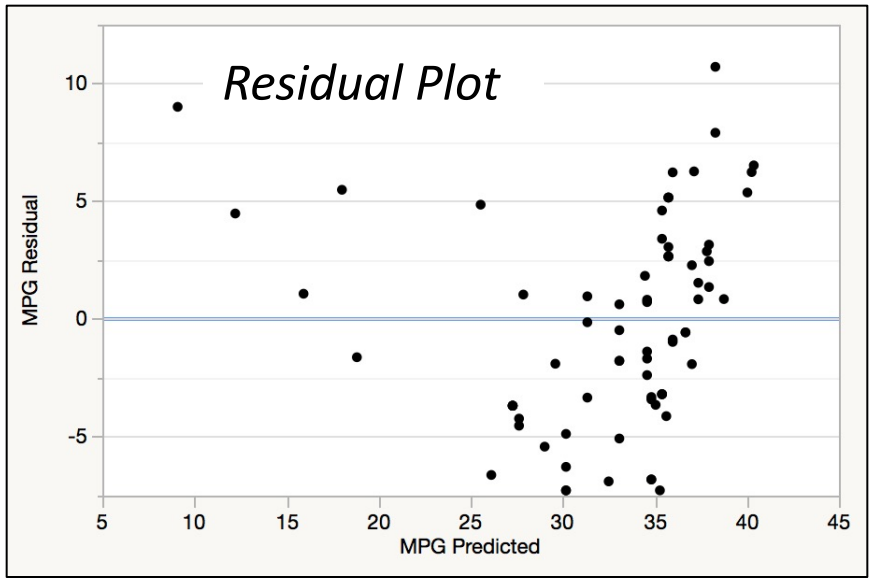
If ***individual scatterplots*** look like one of these curves, try the suggested transformations.

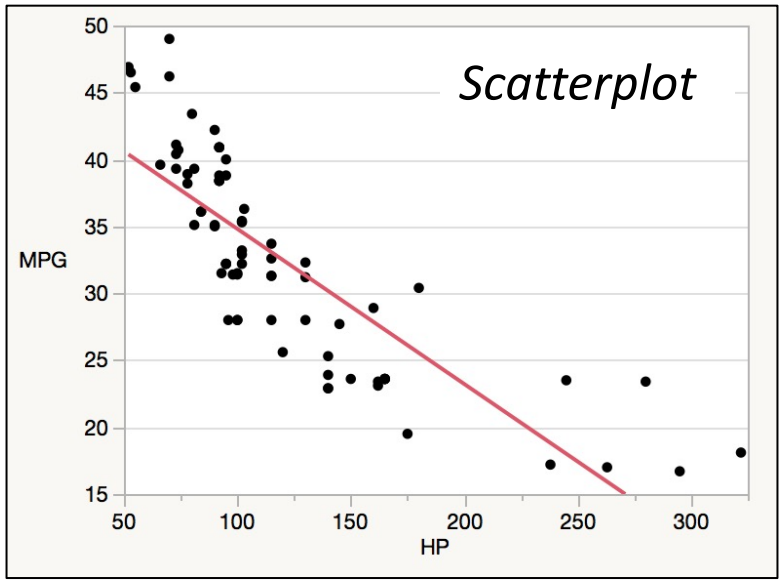


X Transformations

Want: *random scattering.*

Note any patterns or nonconstant variance in a *Residual Plot.*



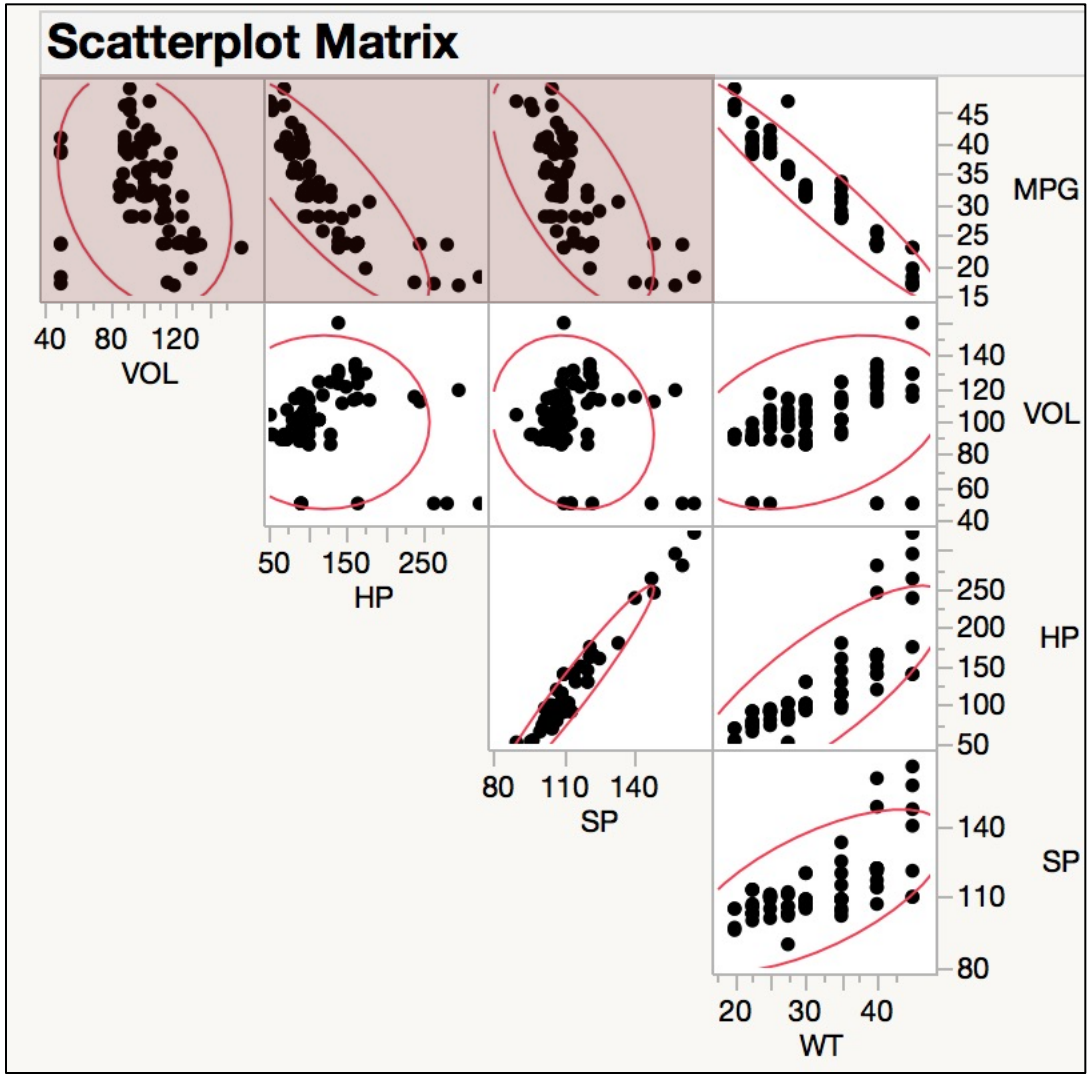


Residual Plot shows a pattern → Examine **individual scatterplots**.

- A scatterplot of *MPG* vs *HP* shows curvature.

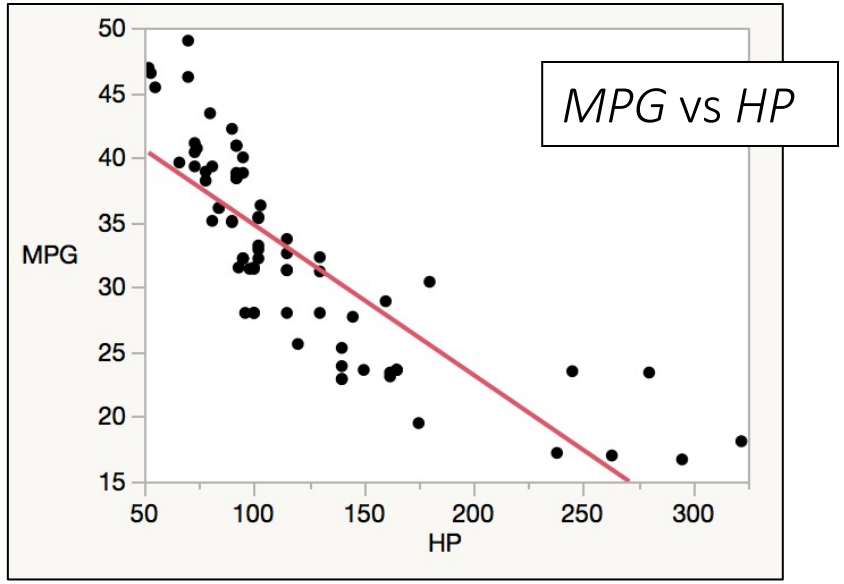
Potential Transformations

There appears to be some curvature in the data. In fact, 3 out of 4 independent variables seem to have an inverse relationship to *MPG*.

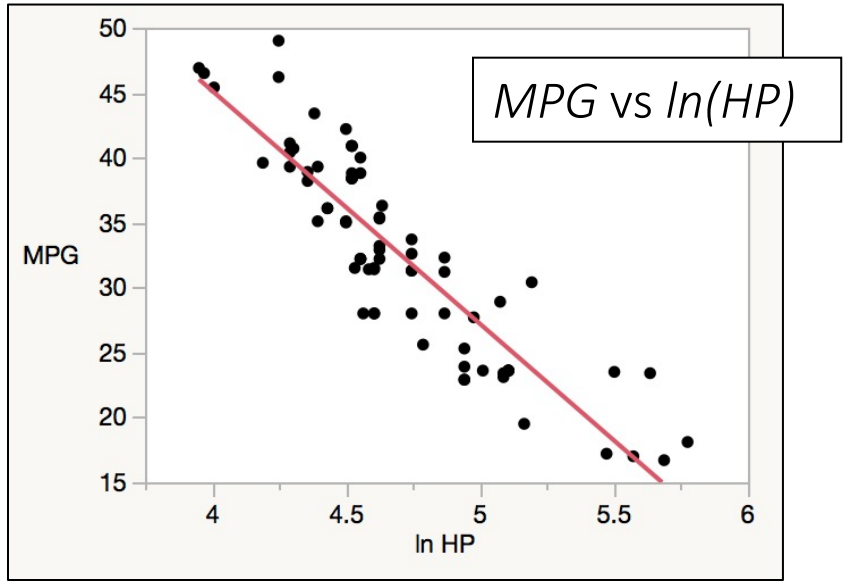


Ln(X) Transformation

It seems reasonable that there would be an asymptote at some level of *MPG* as the *HP* increases; this indicates using the *natural log of X*.



Summary of Fit	
RSquare	0.6747
RSquare Adj	0.6703
Root Mean Square Error	4.4942
Mean of Response	32.5013
Observations (or Sum Wgts)	75.0000



Summary of Fit	
RSquare	0.7978
RSquare Adj	0.7950
Root Mean Square Error	3.5433
Mean of Response	32.5013
Observations (or Sum Wgts)	75.0000

Ln(X) Interpretation

Ln(X) Model. If the left-hand side variable is **not** transformed but the right-hand side is in a **natural log** form:

- A **1–percent** increase in *X* relates to a $(1/100) \cdot b_1$ **unit change** in *Y*.

MPG vs ln(HP):

Parameter Estimates

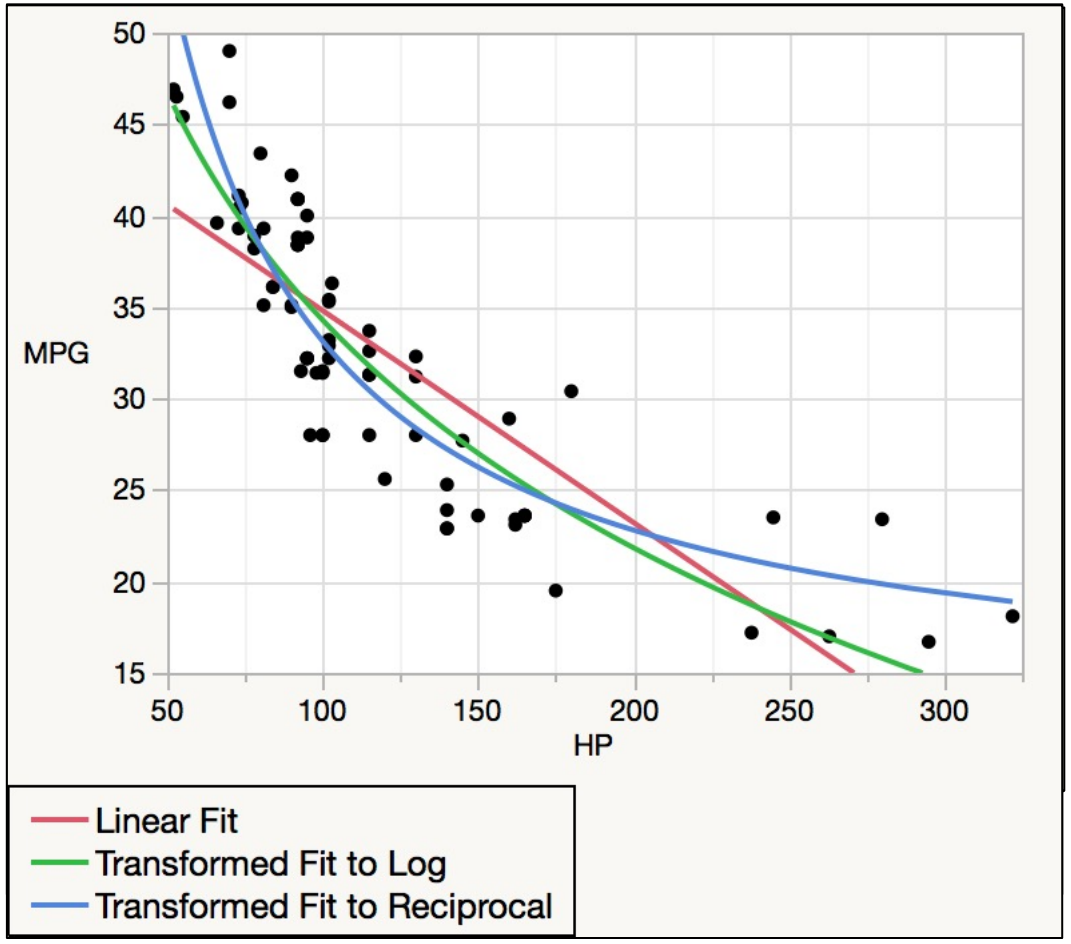
Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	116.9881	4.9949	23.42	<.0001*
ln HP	-17.9574	1.0581	-16.97	<.0001*

$$\widehat{MPG} = 116.9881 - 17.9574\ lnHP$$

For **1% increase** in *HP*, we expect a **0.1796-unit decrease** in *MPG*.

X Transformations

Would a different nonlinear model fit better? How do we decide?



Summary of Fit	
RSquare	0.675
RSquare Adj	0.670
Root Mean Square Error	4.494
Mean of Response	32.501
Observations (or Sum Wgts)	75.000

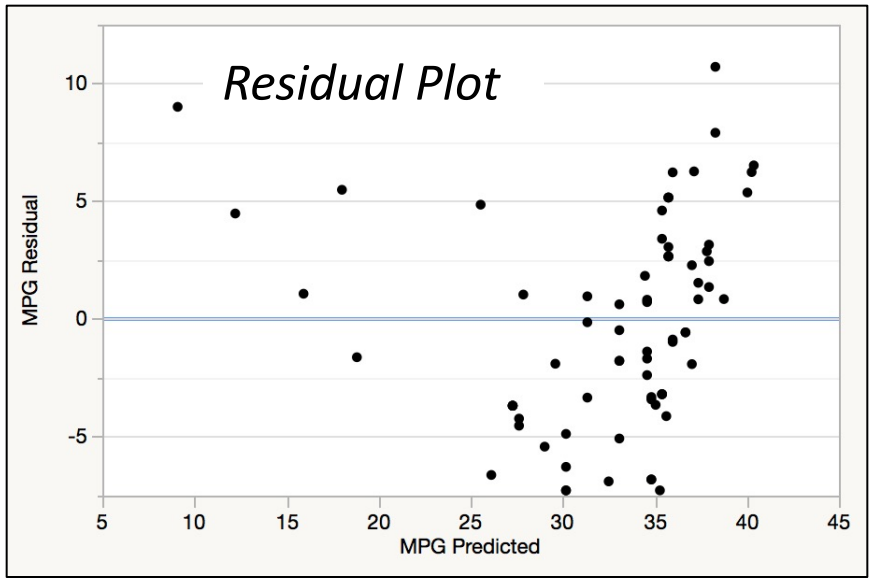
Summary of Fit	
RSquare	0.798
RSquare Adj	0.795
Root Mean Square Error	3.543
Mean of Response	32.501
Observations (or Sum Wgts)	75.000

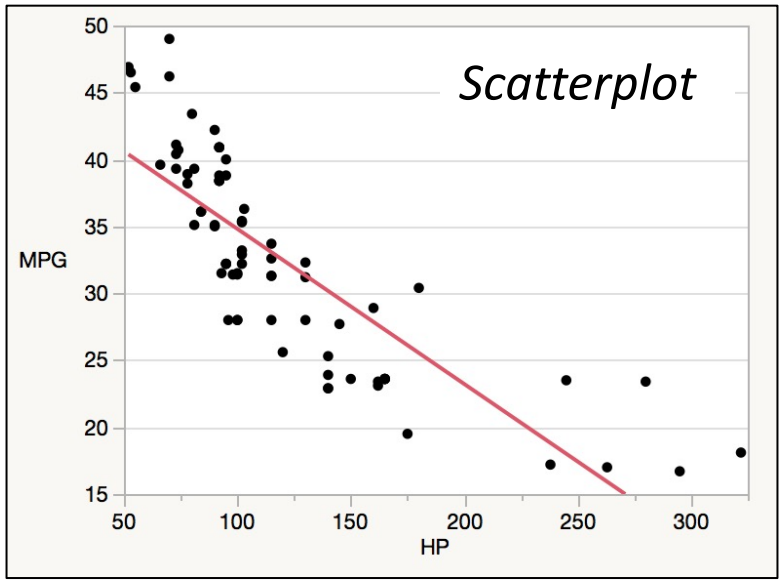
Summary of Fit	
RSquare	0.806
RSquare Adj	0.804
Root Mean Square Error	3.468
Mean of Response	32.501
Observations (or Sum Wgts)	75.000

Y Transformations

Want: *random scattering.*

Note any patterns or nonconstant variance in a *Residual Plot.*



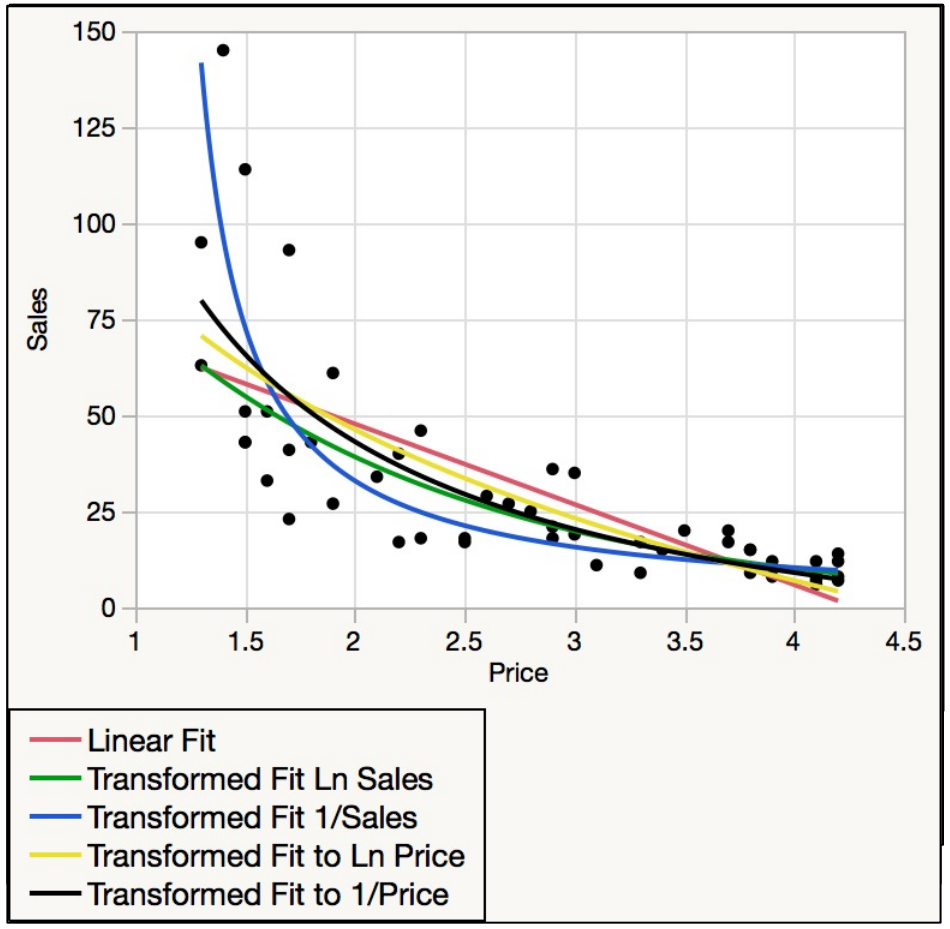


Residual Plot shows a pattern → Examine **individual scatterplots**.

- A scatterplot of *Juice Sales* vs *Price* shows a potential nonlinear pattern.

Y Transformations

How do we decide which model is best?



Summary of Fit

RSquare	0.517
RSquare Adj	0.507
Root Mean Square Error	20.112
Mean of Response	
Observations (or Sum	

Summary of Fit

RSquare	0.575
RSquare Adj	0.566
Root Mean Square Error	18.882
Mean of Response	31.360
Observations (or Sum Wgts)	50.000

Summary of Fit

RSquare	0.747
RSquare Adj	0.742
Root Mean Square Error	0.392
Mean of Response	3.136
Observations (or Sum	

Summary of Fit

RSquare	0.616
RSquare Adj	0.608
Root Mean Square Error	17.948
Mean of Response	31.360
Observations (or Sum Wgts)	50.000

Summary of Fit

RSquare	0.679
RSquare Adj	0.673
Root Mean Square Error	0.023
Mean of Response	0.056
Observations (or Sum Wgts)	50.000

Ln(Y) Interpretation

Ln(Y), or *Log-lin* Model. The left-hand side variable is in natural log form, but the right-hand side is not:

- A **1–unit** increase in *X* relates to a $100 \cdot b_1$ **percent change** in *Y.*

ln(Juice Sales) vs Price:

Parameter Estimates

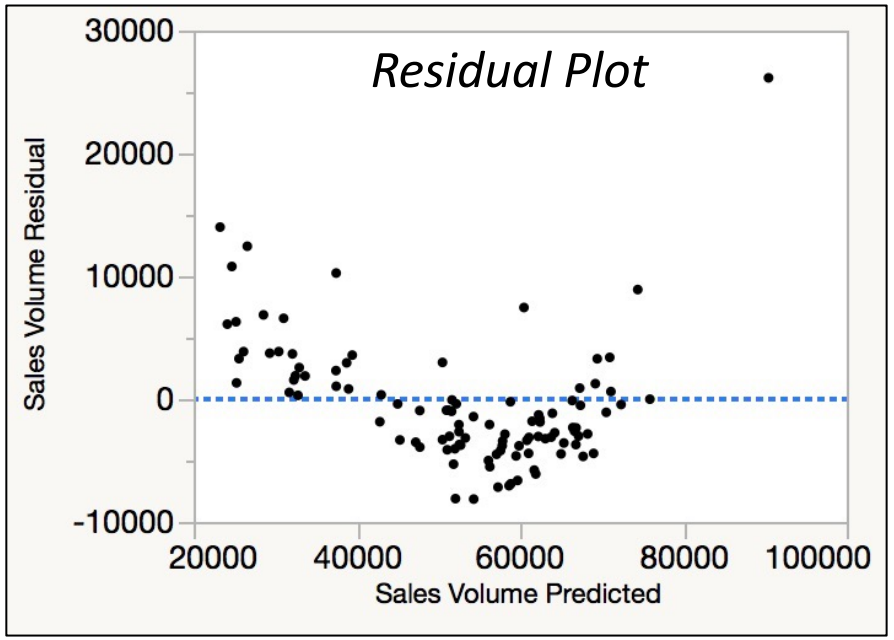
Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	5.0241	0.168	29.93	<.0001*
Price	-0.6780	0.057	-11.92	<.0001*

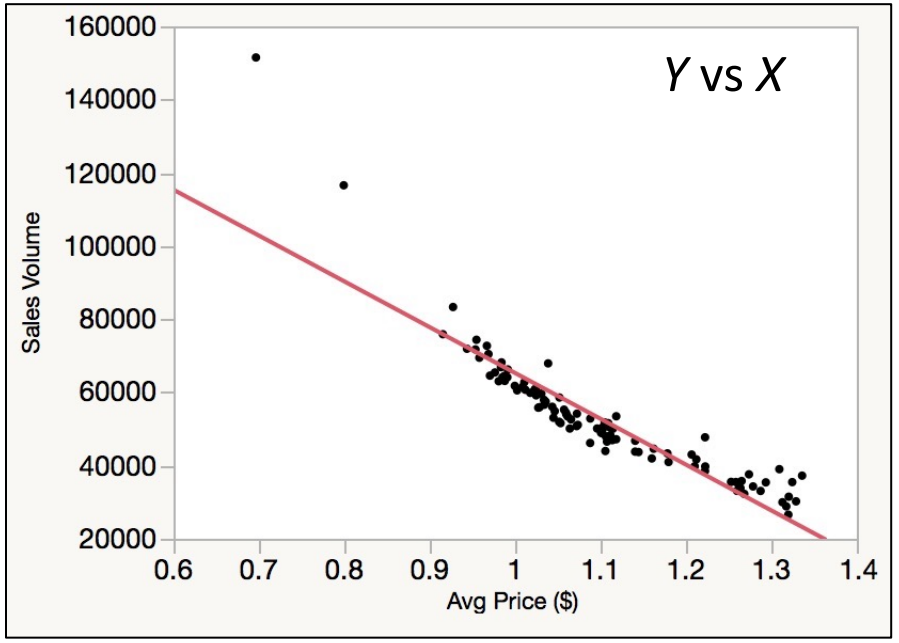
$$\widehat{ln\,Sales} = 5.0241 - 0.678\,Price$$

For a $1 increase in *Price*, we expect a **67.8% decrease** in *Sales*.

Multiple Transformations

Does it appear that the *Average Price* of a can of pet food is linearly related to average *Sales*? Or would a curve fit better?

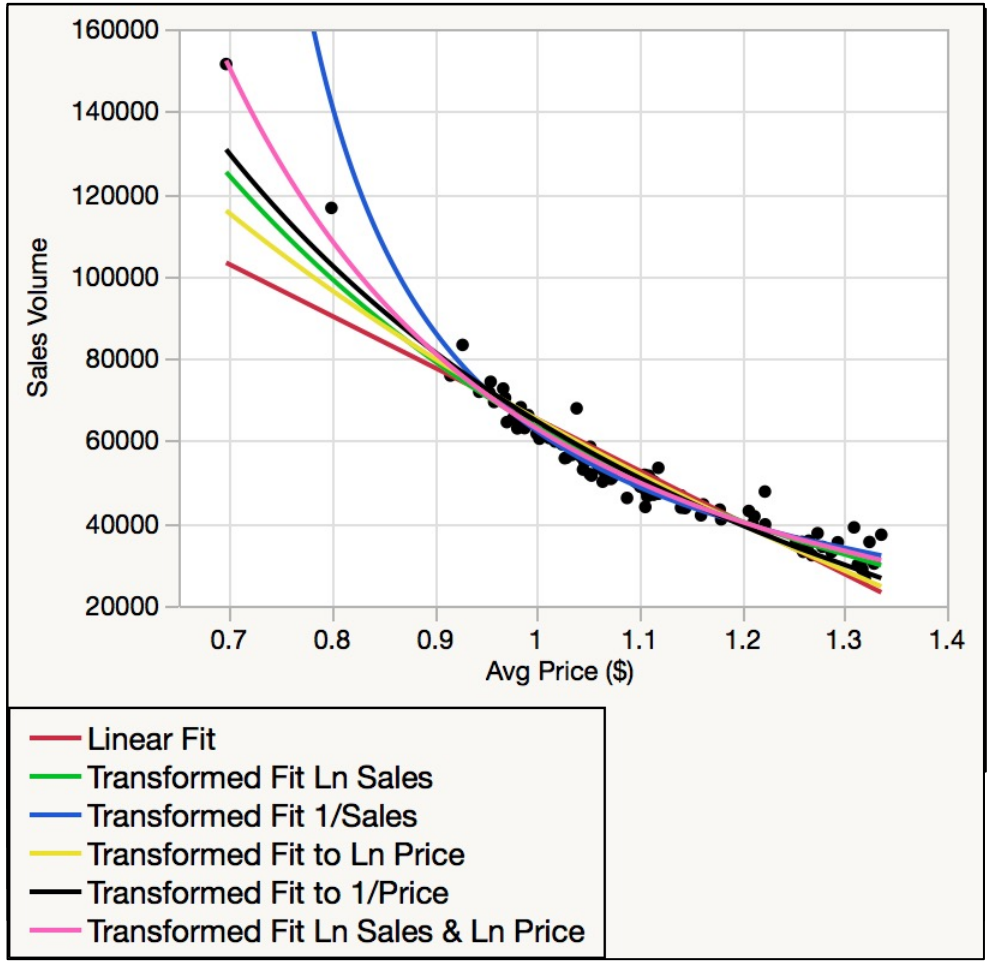




Obviously, there is some curvature here. The scatterplot of Pet Food *Sales* vs *Average Price* shows a potential nonlinear pattern.

Multiple Transformations

How do we decide which model is best?



Summary of Fit

RSquare	0.828
RSquare Adj	0.827
Root Mean Square Error	6991.410
Mean of Response	53135.07
Observations (or Sum Wgts)	

Summary of Fit

RSquare	0.887
RSquare Adj	0.886
Root Mean Square Error	5681.776
Mean of Response	53135.07
Observations (or Sum Wgts)	104.000

Summary of Fit

RSquare	0.943
RSquare Adj	0.942
Root Mean Square Error	0.068
Mean of Response	10.839
Observations (or Sum Wgts)	

Summary of Fit

RSquare	0.936
RSquare Adj	0.936
Root Mean Square Error	4255.023
Mean of Response	53135.07
Observations (or Sum Wgts)	104.000

Summary of Fit

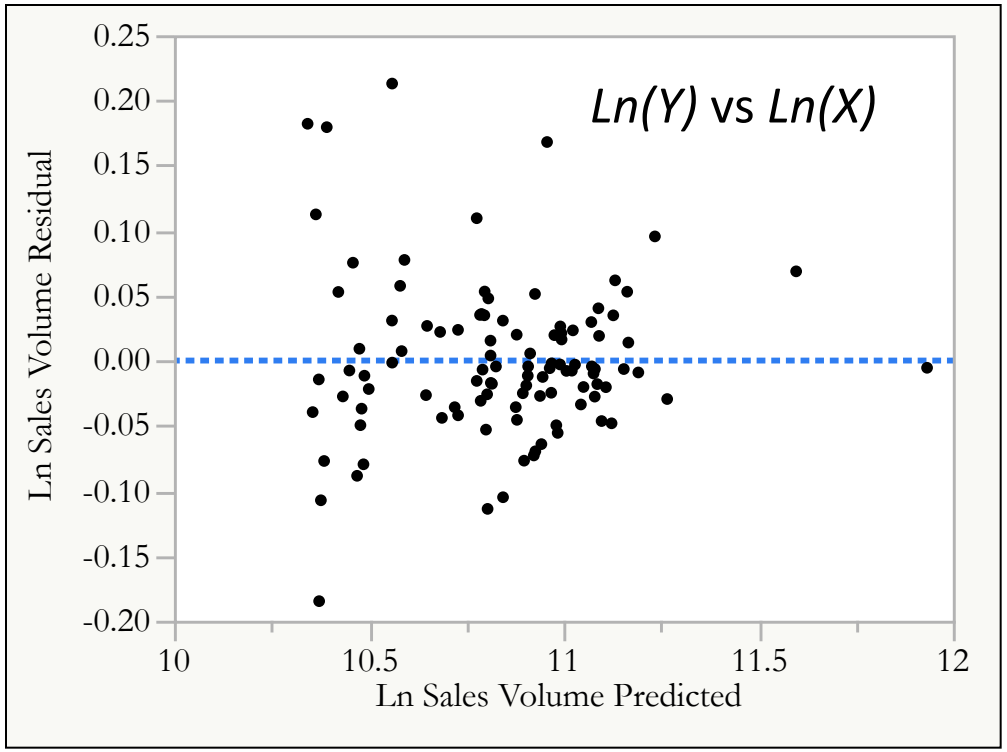
RSquare	0.921
RSquare Adj	0.920
Root Mean Square Error	0.000
Mean of Response	0.000
Observations (or Sum Wgts)	

Summary of Fit

RSquare	0.954554
RSquare Adj	0.954109
Root Mean Square Error	0.060453
Mean of Response	10.8394
Observations (or Sum Wgts)	104

Log-Linear Transformations

This model has a reasonable *Residual by Predicted* plot.



- When the *natural log* is used, that variable is interpreted in **percent changes** instead of the original units.
- Because both sides use the *natural log* in this model, it is called a *log-linear, or log-log* model and both variables are interpreted in **percent changes.**

Log-Linear Interpretation

Log-Linear Model. If both sides of the model use the natural log transformation, the coefficient estimates represent elasticities:

- A **1–percent** increase in *X* relates to a b_1–**percent change** in *Y*.

ln(*Sales*) vs *ln*(*AvgPrice*):

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	11.0506	0.0075	1477.4	<.0001*
Ln Avg Price	-2.4420	0.0528	-46.29	<.0001*

$$\widehat{ln\,Sales} = 22.297 - 2.442\ lnAvgPrice$$

For a 1% increase in *AvgPrice*, we expect a **2.44% decrease** in *Sales*.