CHAPTER 3

Multiple Regression

3.4 New Predictors from Old

3.4 New Predictors from Old:

In this section, we can illustrate two of the more common methods for combining predictors, using a product of two predictors in an interaction model and using one or more powers of a predictor in a polynomial regression.

Regression Model with Interaction:

For two predictors, X_1 and X_2, a multiple regression model with **interaction** has the form

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2 + \in$$

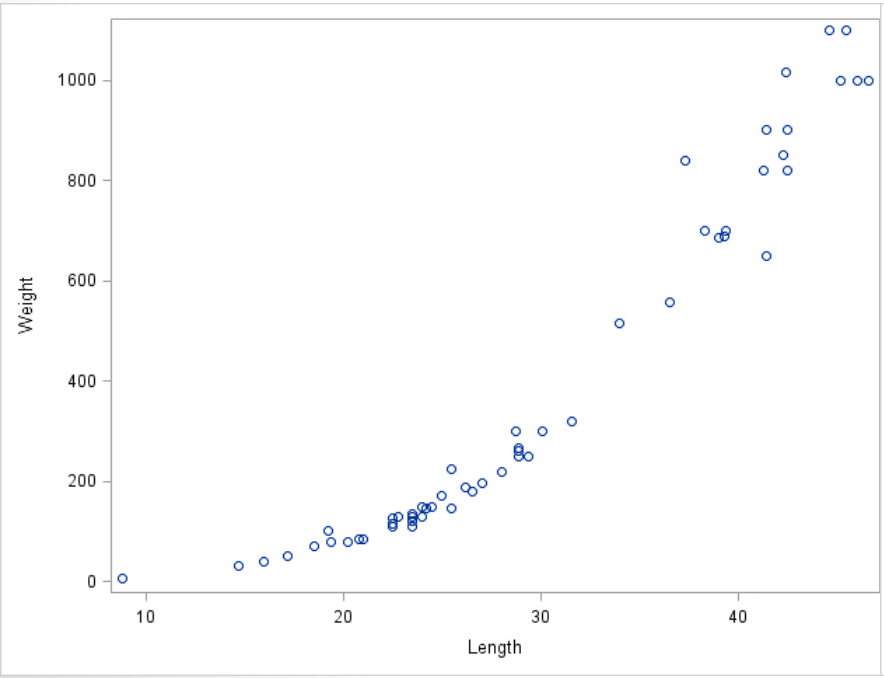
The interaction product term, $X_1 X_2$, allows the slope with respect to one predictor to change for value of the second predictor.

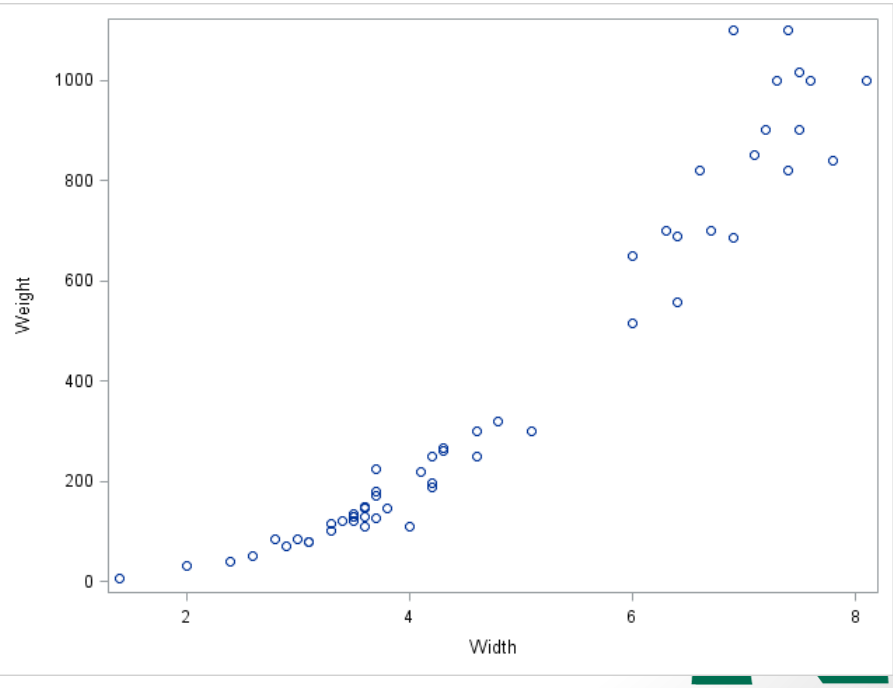
Interaction effects occur when the effect of one variable depends on the value of another variable.

Example 1: *(Perch weight)*

The dataset *Prech.csv* represents a large sample of perch caught in a lake in Finland. This dataset contains the weight (in grams), length (in centimeters), and width (in centimeters) for 56 of these fish.

1. Graph the scatterplot of the weight vs. each predictor?
2. Do you think the linearity assumption will met? Why?





3. Fit the linear regression model for weight vs. length, width. Find ANOVA table and interpret

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	6229332	3114666	396.09	<.0001
Error	53	416762	7863.43265		
Corrected Total	55	6646094			

$$H_0: \beta_{Length} = \beta_{Width} = 0$$

$$H_a: at\ least\ one\ of\ \beta_i's \neq 0$$

Decision: Since p-value is very small (< 0.0001), so we reject H_0.

Conclusion: At least one of the explanatory variables has a significant relationship with the weight.

4. Write the linear regression model for weight with length, width and test each predictor.

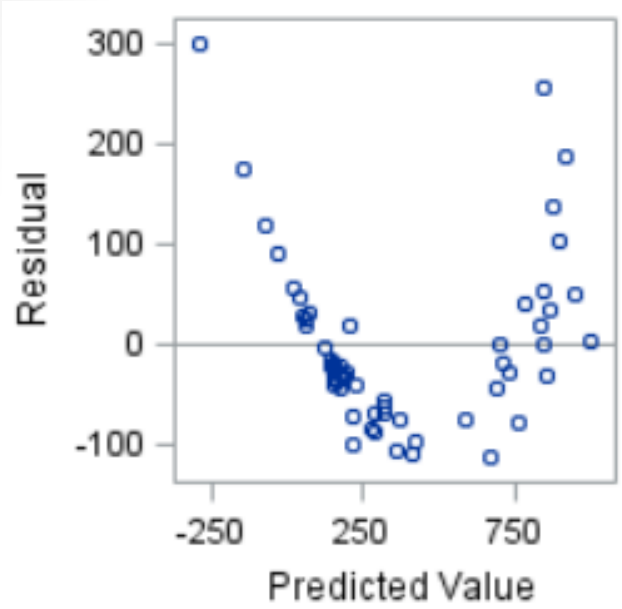
Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	-578.75777	43.66725	-13.25	<.0001
Length	1	14.30738	5.65880	2.53	0.0145
Width	1	113.49966	30.26474	3.75	0.0004

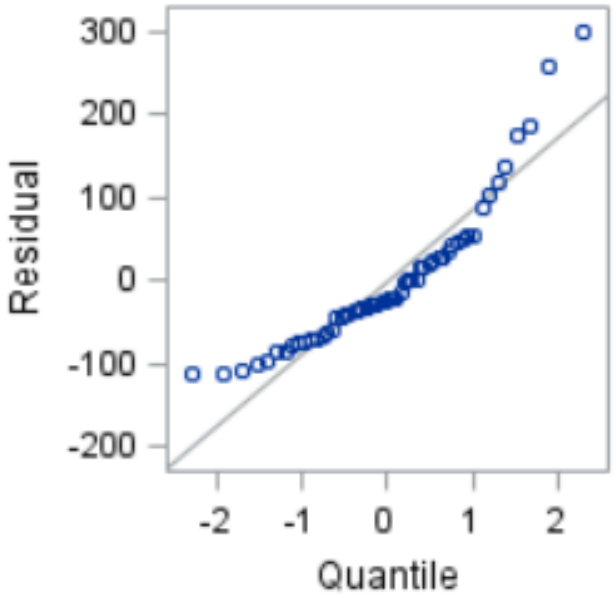
The linear regression model is

$$\widehat{Weight} = -578.76 + 14.31\ Length + 113.5\ Width$$

There is a significant relationship between the weight and each one of length and width.

5. Check the regression assumptions.





- Since the residual vs. predicted graph shows U shape, so the linearity assumption didn't met.
- The Q-Q plot also shows that the normality assumption didn't met.

6. Fit the linear regression model for weight with length, width, and the **interaction** between both of them as predictors. (create a new dataset.)

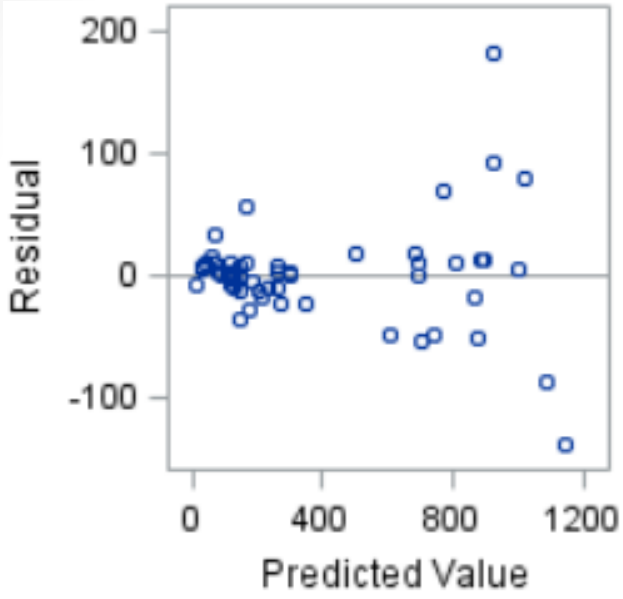
```
data perch1;
set perch;
LW = Length * Width;
run;
```

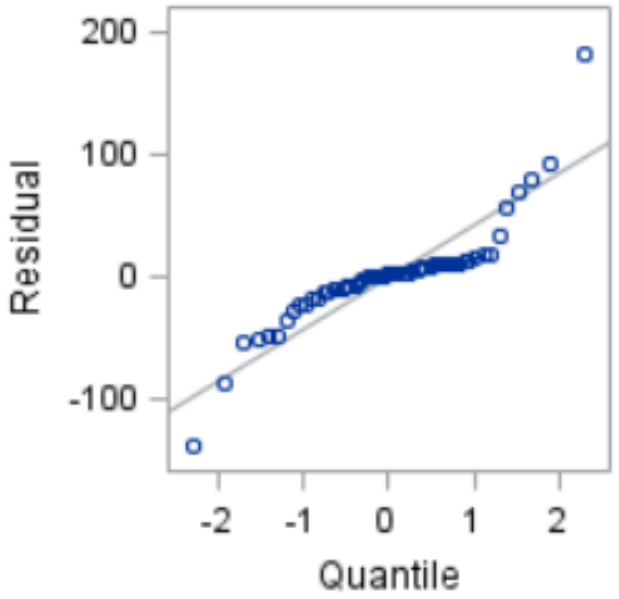
Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	113.93490	58.78439	1.94	0.0580
Length	1	-3.48269	3.15210	-1.10	0.2743
Width	1	-94.63090	22.29543	-4.24	<.0001
LW	1	5.24124	0.41312	12.69	<.0001

The linear regression model is

$$\widehat{Weight} = 113.93 - 3.48\,Length - 94.63\,Width + 5.24\,Length \times Width$$

7. Check the residual assumptions.





- Since the predicted vs. residual graph shows cone shape, so the equal (constant) variance assumption didn't met.
- The Q-Q plot also shows that the normality assumption didn't met.

8. Compare between the two models, which one is better. Use root mean square error $\left(\sqrt{MSE}\right)$ and $Adj(R^2)$.

Model	$\sqrt{MSE}$	$Adj(R^2)$
No-interaction	88.67	0.9394
Interaction	44.23	0.9838

The interaction model is the best model because it has the lowest $\sqrt{MSE}$ and the highest $Adj(R^2)$.

9. Find the relationship between the weight and the width for perch that are $25\ cm$ long.

$$\widehat{Weight} = 113.93 - 3.48\,(25) - 94.63\,Width$$
$$+5.24\,(25) \times Width$$

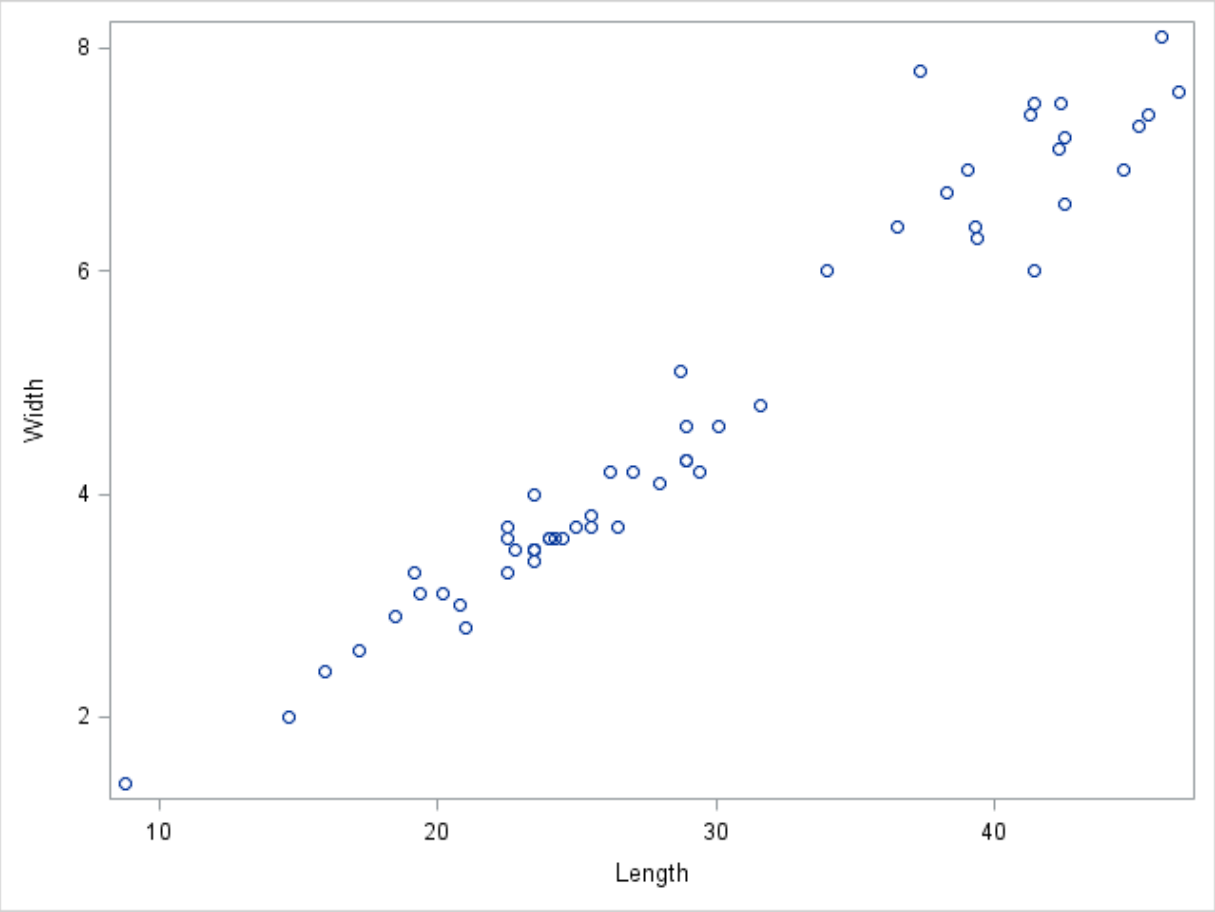
$$\widehat{Weight} = 26.93 + 36.37\,Width$$

10. Find the relationship between the weight and the length for perch that are $2\ cm$ wide.

$$\widehat{Weight} = 113.93 - 3.48\,Length - 94.63(2)$$
$$+5.24\,Length \times (2)$$

$$\widehat{Weight} = -75.33 + 7\,Length$$

10. Graph the scatterplot of length vs. width. What did you observe?



There is a **strong positive linear relationship** between the length and width; this relationship is known as multicollinearity problem. (Section 3.5)

Polynomial Regression Model:

1. For a single **quantitative** predictor X, a **quadratic (second order) regression model** has the form

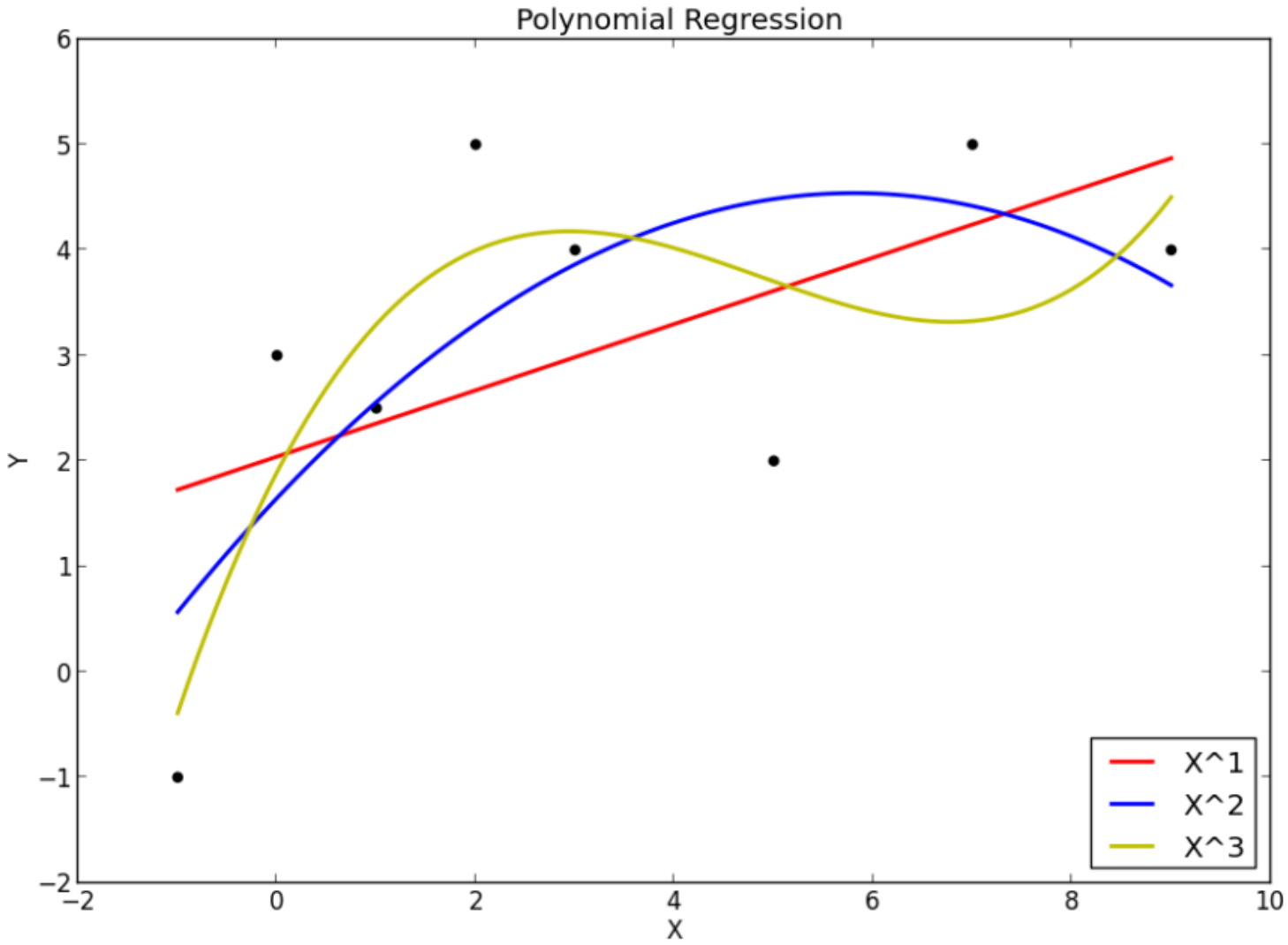
$$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \in$$

2. For a single **quantitative** predictor X, a **cubic regression model** has the form

$$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \beta_3 X^3 + \in$$

3. In general, for a single **quantitative** predictor X, a **polynomial regression model** has the form

$$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \cdots + \beta_k X^k + \in$$



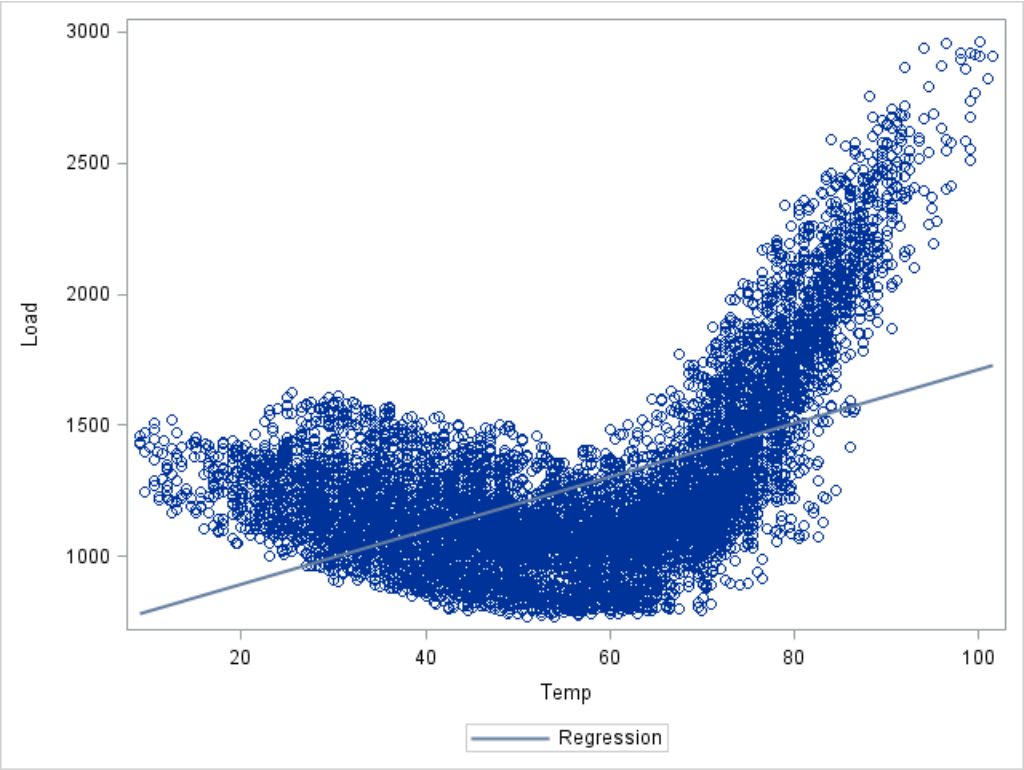
Red line shows the linear regression line, blue line shows the quadratic regression line, and yellow line shows the cubic regression line.

Example 2: *(ElectricityLoad)*

ElectricityLoad.csv represents 2 variables which are the hourly electricity load and temperature over a year.

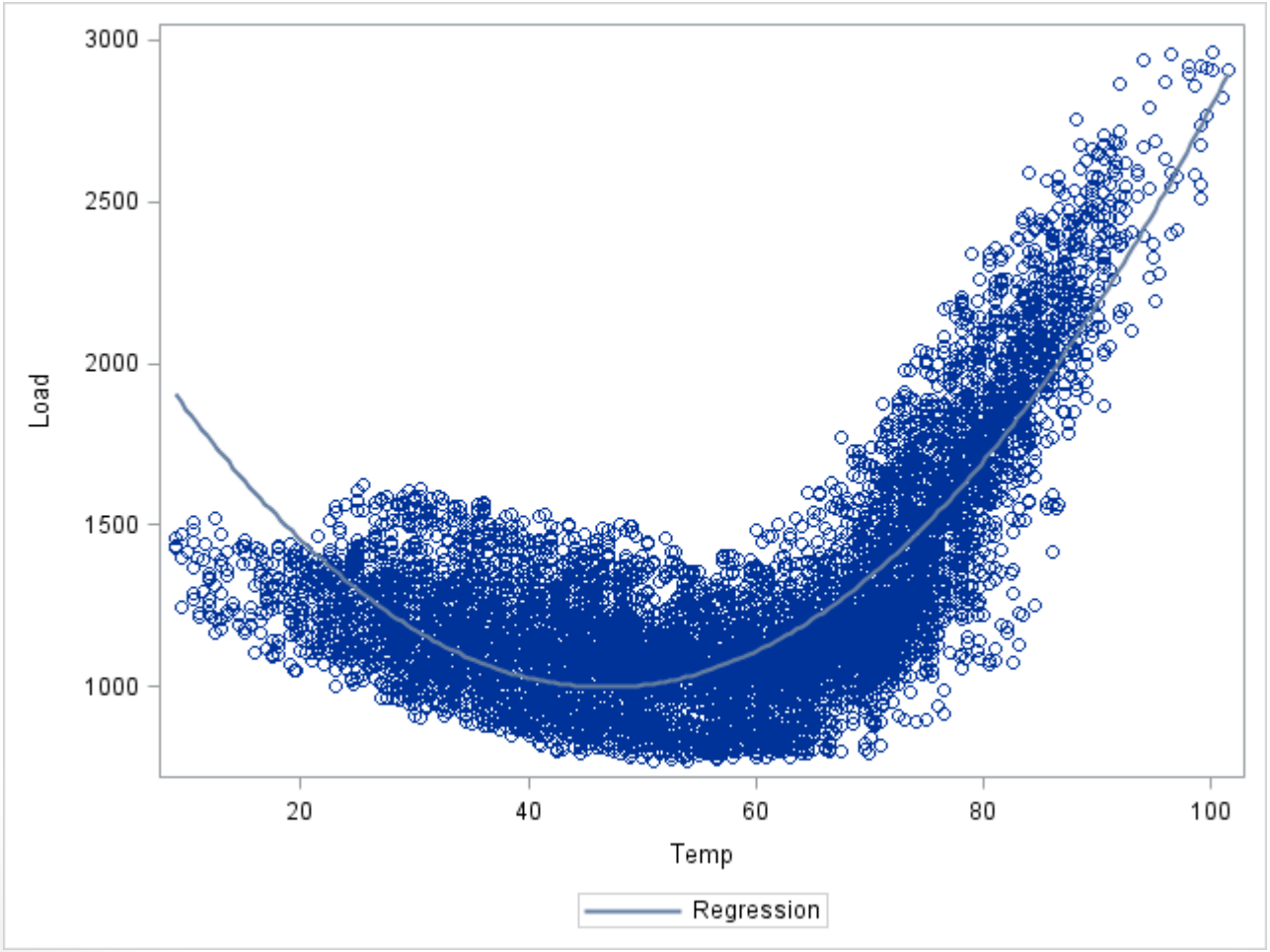
1. Identify the response and explanatory variable.
2. Graph the scatterplot with the regression line.

```
proc sgplot data=ElectricityLoad;
reg x = temp y = load;
run;
```



3. Graph the scatterplot with the quadratic regression line. Compare it with the linear regression.

```
proc sgplot data=ElectricityLoad;
reg x = temp y = load / DEGREE=2;
run;
```



4. Fit the simple linear regression model for the load and temperature.

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	691.99673	10.40677	66.49	<.0001
Temp	1	10.23369	0.17500	58.48	<.0001

Root MSE	298.86899	R-Square	0.2808
Dependent Mean	1271.20042	Adj R-Sq	0.2807
Coeff Var	23.51077		

The simple linear regression model is

$$\widehat{Load} = 691.99 + 10.23\, Temperature$$

4. Fit the multiple linear regression model for the load, temperature, and second degree of temperature.

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	2384.67703	17.24320	138.30	<.0001
Temp	1	-59.29041	0.66005	-89.83	<.0001
temp2	1	0.63407	0.00593	106.97	<.0001

Root MSE	196.78939	R-Square	0.6882
Dependent Mean	1271.20042	Adj R-Sq	0.6882
Coeff Var	15.48059		

The simple linear regression model is

$$\widehat{Load} = 2384.68 - 69.29\,Temp + 0.63\,Temp^2$$

5. Compare between the two models, which one is better.

Model	$\sqrt{MSE}$	$Adj(R^2)$
Linear model	298.87	0.2807
Quadratic model	196.79	0.6882

The quadratic model better fits the data.

Complete Second-order Model:

For two **quantitative** predictor X_1 and X_2, a **complete second-order regression model** has the form

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2 + \beta_4 X_1^2 + \beta_5 X_2^2 + \in$$

Note: we will work on the complete second order model in the lab tonight.

Note: in **proc reg**, if you used a large dataset (> 5000), the residual plots may not appear, so you need to add this option "**plots = maxpoints (# of the observations)**"

```
proc reg data=ElectricityLoad PLOTS(MAXPOINTS=100000);
model load = temp  ;
output out=new p=predcted;
run;
```

Reading Assignment

Read section 3.4