Multiple Regression

Part 4: General Linear Regression Model

STAT 705: Regression and Analysis of Variance

Kansas State University

General Linear Regression Model

$$Y_i = \beta_0 + \beta_1 Z_{1i} + \beta_2 Z_{2i} + \ldots + \beta_p Z_{p,i} + \varepsilon_i \quad i = 1, 2, \ldots n$$

- Number of terms (p) is limited only by the sample size ($p < n - 1$)
- "General" model because
  - The Z's can be X's (observed variables in the data set)
  - The Z's can be **functions** of the X's
  - The Z's can represent non-numeric data (e.g., gender)
- The model is still "linear" because it is linear in the β's

Kansas State University

Z’s can be Functions of the X’s

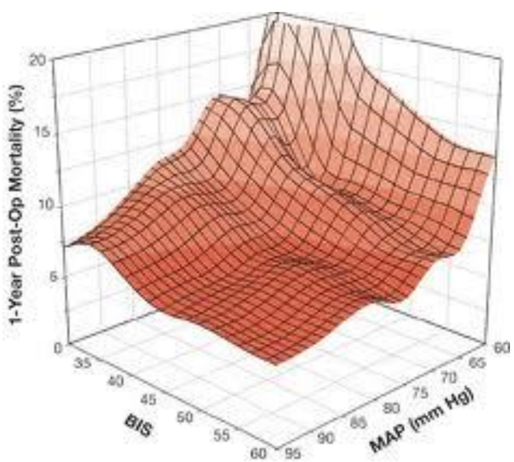
- Suppose there are two measured variables (X_1 and X_2) in the data set.
- The Z’s could be
  - $Z_1 = X_1$; $Z_2 = X_2$, $Z_3 = X_1^2$, $Z_4 = X_2^2$, $Z_5 = X_1X_2$
  - Then the model is

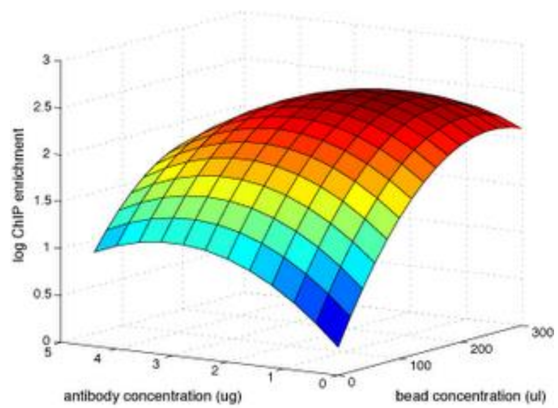
$$Y_i = \beta_o + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{1i}^2 + \beta_4 X_{2i}^2 + \beta_5 X_{1i} X_{2i} + \varepsilon_i \quad i = 1, 2, \ldots n$$

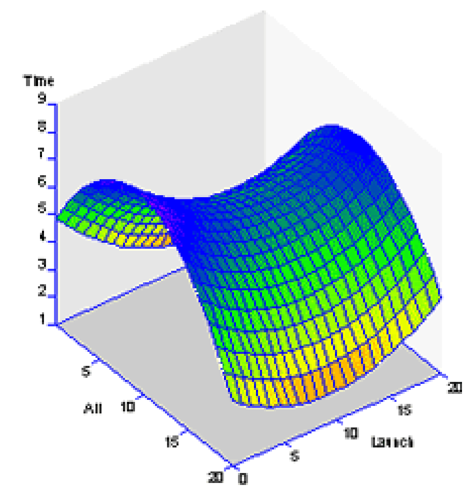
- The additional terms allow the least squares *plane* to become a least squares *surface*

Response Surfaces

- Although these surfaces appear complicated, the model is still linear in the β's, and all our work with multiple regression models is still valid and applicable.
  - We use Least Squares method to estimate the β's
  - For specified values of the predictors, the expected value of Y is the corresponding point on the response surface.







Some Terminology

- A ‘first order’ model has no squared terms
  - e.g. a first order model with two predictors

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \varepsilon_i$$

- A ‘second order’ model has squared terms
  - e.g. a second order model with one predictor

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{1i}^2 + \varepsilon_i$$

- An ‘interaction’ term is the product of two predictors
- An ‘additive’ model has no interaction terms
- A ‘nonlinear’ model is not linear in the β’s
  - e.g. $Y_i = \beta_0 \exp\left(\beta_1 X_i\right) + \varepsilon_i$ or $Y_i = \beta_0 X_i^{\beta_1} + \varepsilon_i$

KANSAS STATE UNIVERSITY

"All Models are Wrong…"

- It can be quite tricky to decide which, if any, transformed predictor variables should be considered for inclusion in the model
  - Knowledge of the subject matter might provide clues
  - There are an infinite number of possibilities (e.g., X^2, X^3, log(X), 1/X, square root of X, …)
  - Limit the scope to something that is reasonable
- Remember: "All models are wrong, but some are useful."
- We are looking for a "useful" model

Example

- A data set contains two measured predictors (X_1 and X_2) and a response (Y)
- Data and SAS program are in the file 'FitInteraction.sas'
- We will fit a model to these data
- This is fake data, so there is no story to put the data in context.

First Model

- Fit a first order model with no interaction
- This is the simplest model that uses all the data.

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \varepsilon_i$$

```
data fake;
input x1 x2 y @@;
x1sq = x1**2;
x2sq = x2**2;
datalines;

. . . data goes here . . .
```

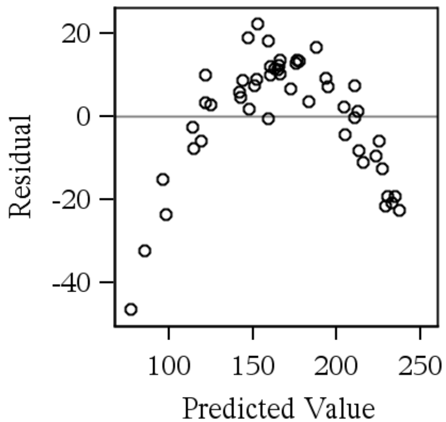
Create new variables to hold X_1^2 and X_2^2.

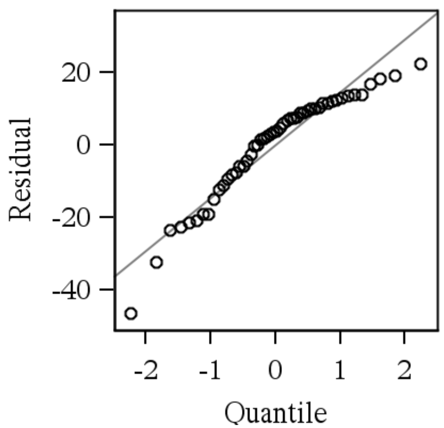
We will use these in the second model.

```
proc reg data=fake plots=residuals;
 model y = x1 x2;
 run;
```

The “plots=residuals” option generates a new kind of residual plot

Diagnostic Plots for First Model

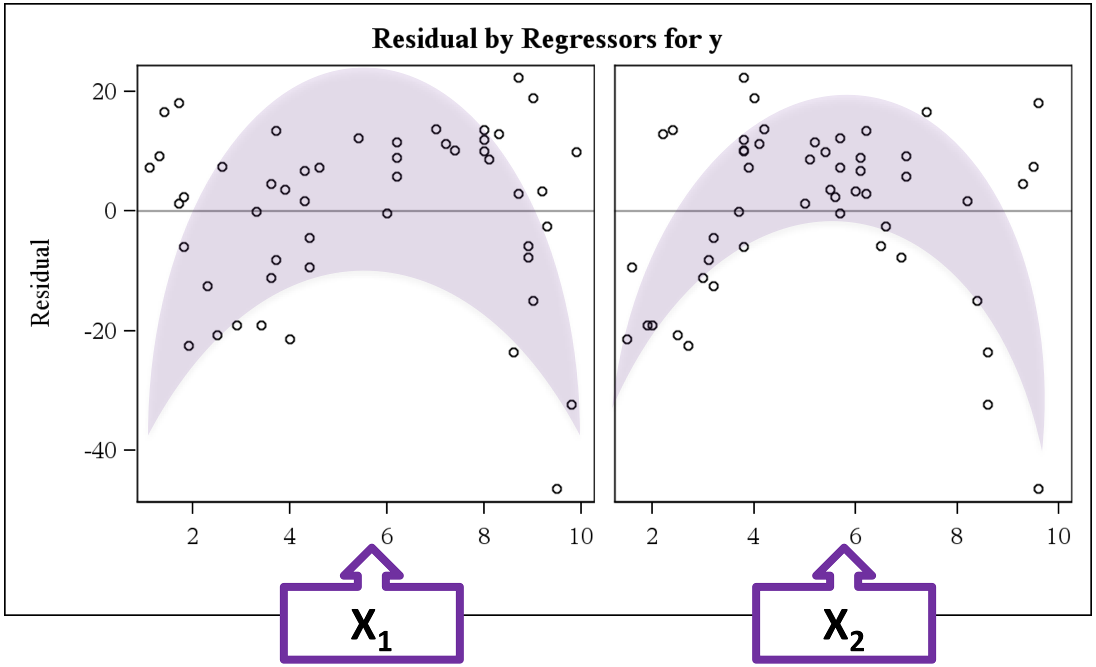




- Normal probability plot (on the bottom) could be okay
- The residual plot (on the top) has a very distinct quadratic shape
- We <u>must</u> modify this model
- Add a squared term to the model
- Since there are two predictors, we must decide which one (or both) we should square

Partial Residual Plots

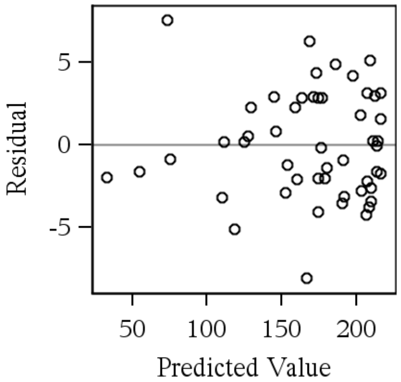
Plot the residuals against each of the **original predictors.**
Generated by the 'plots=diagnostics' option on PROC REG.

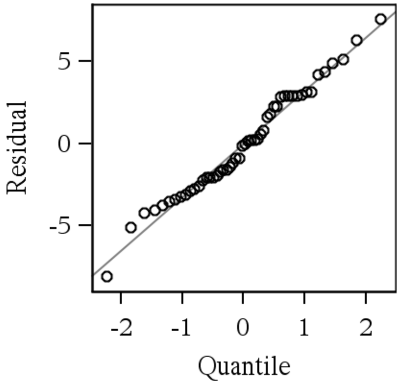


Both graphs show a quadratic pattern, but it is more distinct for X_2.
We will include squared terms for <u>both</u> of these variables.

Diagnostic Plots for Second Model

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{1i}^2 + \beta_4 X_{2i}^2 + \varepsilon_i$$





- Both of these graphs look very good
- MUCH better than the first model
- We will use this model to proceed with the analysis

Results of Second Model

Analysis of Variance					
Source	**DF**	**Sum of Squares**	**Mean Square**	**F Value**	**Pr > F**
Model	4	91624	22906	139.09	<.0001
Error	45	7410.81752	164.68483		
Corrected Total	49	99035			

This model is highly significant.

Root MSE	12.83296	**R-Square**	0.9252
Dependent Mean	169.76200	**Adj R-Sq**	0.9185
Coeff Var	7.55938		

This is an extraordinarily large R^2.
(This rarely happens with 'real' data.)

Results of Second Model

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	242.73456	13.36723	18.16	<.0001
x1	1	-5.84354	3.67487	-1.59	0.1188
x2	1	4.00068	3.96260	1.01	0.3181
x1sq	1	-0.46071	0.32827	-1.40	0.1673
x2sq	1	-1.36271	0.34632	-3.93	0.0003

- The estimated model is

  $$Y = 242.73 - 5.84X_1 + 4.00X_2 - 0.46X_1^2 - 1.36X_2^2$$

- Only the intercept and X_2^2 are significant predictors
- We should keep X_2^2 and its lower-order term (X_2) in the model
- Can we remove X_1 and X_1^2?

A Question

- Should we use the ‘full’ model or will the smaller (‘reduced’) model suffice?
- Full model: $Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{1i}^2 + \beta_4 X_{2i}^2 + \varepsilon_i$
- Reduced: $Y_i = \tau_0 + \tau_1 X_{2i} + \tau_2 X_{2i}^2 + \varepsilon_i$

  (The τ’s are simply the coefficients. We use a different Greek letter because they can have different values than the β’s in the full model.)

- This is equivalent to testing the hypotheses

  H_0: $\beta_1 = \beta_3 = 0$ vs. H_a: β_1 or β_3 is not 0

A New Hypothesis Test

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{1i}^2 + \beta_4 X_{2i}^2 + \varepsilon_i$$

- Hypotheses: H_0: $\beta_1 = \beta_3 = 0$ vs. H_a: β_1 or β_3 is not 0
- These hypotheses are different than the ones we have previously encountered
  - The ANOVA F test examines <u>all</u> the slopes (β_1 through β_4)
  - The t tests examine <u>one</u> slope at a time
  - We want to test <u>two</u> of the four slopes
- If the null hypothesis is true (i.e., ‘under H_0’), then the model becomes the reduced model
  - Reduced model is sometimes called the null model

F Test for Nested Models

- The full model must have all the terms that are in the reduced model (i.e., the models must be nested)
  - Test is also called a ‘comparison of models’ F test or a ‘partial’ F test
- This test will require some hand calculations

1. Fit the full model (in SAS) and record SS and df *for Error*
2. Fit the reduced model (in SAS) and record SS and df *for Error*
3. Calculate the test statistic (by hand, formula on next slide)
4. Under H_0, the test statistic follows an F distribution
5. Find critical value in F table and make the conclusion

F Test for Nested Models

Analysis of Variance – FULL MODEL			
Source	DF	Sum of Squares	Mean Square
Model	4	91624	22906
Error	45	7410.81752	164.68483
Corrected Total	49	99035	

Analysis of Variance – REDUCED MODEL			
Source	DF	Sum of Squares	Mean Square
Model	2	46402	23201
Error	47	52633	1119.84083
Corrected Total	49	99035	

Test statistic:

$$F = \frac{\dfrac{SSE(Red) - SSE(Full)}{dfE(Red) - dfE(Full)}}{\dfrac{SSE(Full)}{dfE(Full)}} = \frac{\dfrac{52633 - 7410.81752}{47 - 45}}{\dfrac{7410.81752}{45}} = \frac{\dfrac{45222.18}{2}}{\dfrac{7410.81752}{45}}$$

$F = 137.3$

Reference distribution: F, with df 2 and 45

Nested Model F Test, Continued

- Find the critical value in the F table
  - df numerator = 2 (because we are testing 2 parameters)
  - df denominator = 45 (this dfE in full model)
  - Significance level $\alpha = 0.05$
  - Critical value is between 3.23 and 3.15
- Compare the test statistic to critical value
  - 137.3 is much greater than 3.2
  - We strongly reject H_0
- Conclusion: We should use the full model.

Is This a Contradiction?

- From the individual t tests, neither X_1 nor its square are significant in the full model (So it seems we should be able to remove them)
- Nested model F test indicates we should use the full model
- This is NOT a contradiction
  - The t tests are testing each parameter <u>assuming the other terms are kept in the model</u>
  - For testing X_1: p-value = .1188 ⇨ X_1 can be removed from the model, provided X_2, X_2^2 **and X_1^2** stay in the model
  - For testing X_1^2: p-value = .1673 ⇨ X_1^2 can be removed from the model, provided X_2, X_2^2 **and X_1** stay in the model
  - The nested model F test indicates we cannot remove both X_1 and X_1^2

What You Should Know

- Use SAS to fit general linear models
  - Investigate when transformed variables may be beneficial to the model
  - Create transformed variables in SAS data step
- When to use and how to interpret
  - Overall ANOVA F test
  - Individual t tests
  - Nested F test
- Perform the nested F test by hand