Multiple Regression
Part 5: Qualitative Predictors

STAT 705: Regression and Analysis of Variance

General Linear Regression Model

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

- "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)
- In the last lesson, we worked with functions of the X's (e.g. X^2)
- We now explore non-numeric predictors
  - a.k.a., categorical or qualitative predictors

Qualitative Predictors

- Examples
  - Gender

    → Two levels: Male or Female
  - Education

    → Three levels: No college, Some college, College graduate
  - Seasons

    → Four levels: Spring, Summer, Fall, Winter
- There are no units associated with qualitative predictors

Indicator ('Dummy') Variables

- Convert qualitative predictors to indicator variables
- For a qualitative predictor with k levels, there are k-1 indicator variables
- Indicator variables can have only the values 0 or 1
- Example: Indicator variable for Gender could be

$$X_1 = \begin{cases} 1 & \text{if Gender is Male} \\ 0 & \text{otherwise} \end{cases}$$

- Example: Seasons would have 3 indicator variables

$$X_1 = \begin{cases} 1 & \text{if Season is Spring} \\ 0 & \text{otherwise} \end{cases} \quad X_2 = \begin{cases} 1 & \text{if Season is Summer} \\ 0 & \text{otherwise} \end{cases} \quad X_3 = \begin{cases} 1 & \text{if Season is Fall} \\ 0 & \text{otherwise} \end{cases}$$

The Missing Level

- Example: Seasons
  - Four levels: Spring, Summer, Fall, Winter
  - Three indicator variables (for Spring, Summer, Fall)
- What happened to Winter?

Partial Data Set				
Obs	Season	X_1	X_2	X_3
1	Fall	0	0	1
2	Spring	1	0	0
3	Summer	0	1	0
4	Winter	0	0	0

Observation 1 (Fall) : only X_3 is 1

Observation 2 (Spring): only X_1 is 1

Observation 3 (Summer): only X_2 is 1

Observation 4 (Winter): **ALL indicators are 0**

Example

- We want to compare two over-the-counter pain relievers: acetaminophen and ibuprofen
- Which one works best on headaches?
- Fifty chronic headache sufferers volunteered
  - 25 were randomly assigned to take acetaminophen
  - the other 25 took ibuprofen
  - the volunteers did not know what they were assigned
- The next time they had a headache, they took their assigned medication and recorded how long (in minutes) it took to get pain relief

(continued)

The Data

- The volunteers also recorded the severity of the headache
  - on a scale from 1 to 10
  - 1 is least severe
  - 10 is most severe
  - the data set looks like this

Subject	Drug	Severity	Time
1	acetaminophen	3	8.4
2	ibuprofen	7	21.5
3	ibuprofen	9	29.1
4	acetaminophen	5	4.8
…	…	…	…

- We want to model Time as function of Drug and Severity

The Model

- Define X_1 = Severity and $X_2 = \begin{cases} 1 & \text{if Drug is acetaminophen} \\ 0 & \text{otherwise} \end{cases}$
- Then the model can be written $Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \varepsilon_i$
- But this model looks just like the models we have been working with . . . how is this model different?
  - If the drug is ibuprofen ($X_2 = 0$), the model reduces to
    $$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 \cdot 0 + \varepsilon_i \Rightarrow Y_i = \beta_0 + \beta_1 X_{1i} + \varepsilon_i$$
  - If the drug is acetaminophen ($X_2 = 1$), the model is
    $$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 \cdot 1 + \varepsilon_i \Rightarrow Y_i = (\beta_0 + \beta_2) + \beta_1 X_{1i} + \varepsilon_i$$
- Both models have the same slope (β_1), but possibly different intercepts (β_0 vs. β_0+β_2)

Fit the Model in SAS

```
data headaches;
input Subject Drug $ Severity Time;
datalines;
1  acetaminophen  2 15.4
2  acetaminophen  3  8.4
... more datalines here ...
;
proc glm data=headaches plots=all;
 class Drug;
 model Time = Drug Severity / solution;
 run;
```

The dollar sign tells SAS that the preceding variable is character (not numeric).

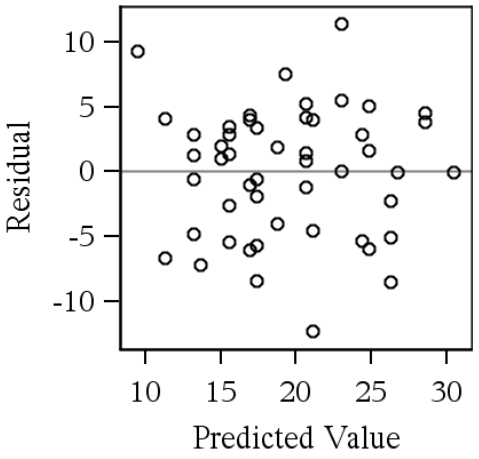
The class statement tells SAS that Drug is a categorical variable. SAS automatically defines the indicator variables.

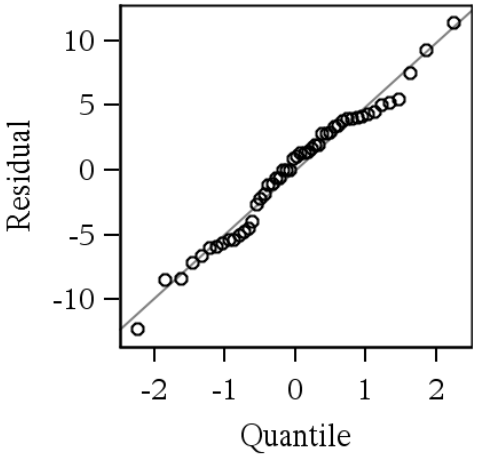
We are using Proc GLM (General Linear Model) because Proc Reg does not permit categorical predictors.

The 'solution' option on the model statement generates the individual t tests. These are not automatically printed with Proc GLM.

The complete code for this example is in the file 'headaches.sas'

Diagnostic Plots





- Both plots look very good
- We have no reason to suspect the assumptions might be violated

Proc GLM Output

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

The ANOVA table should look familiar. The F test is testing the null hypothesis $H_0: \beta_1 = \beta_2 = 0$.

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	1250.15	625.075	24.49	<.0001
Error	47	1199.76	25.527		
Corrected Total	49	2449.91			

Parameter	Estimate		Standard Error	t Value	Pr > |t|
Intercept	11.81	B	1.668	7.08	<.0001
Drug acetamin	-4.20	B	1.446	-2.90	0.0056
Drug ibuprofe	0.00	B	.	.	.
Severity	1.86		0.276	6.74	<.0001

This table is the result of the 'solution' option on the model statement. It is analogous to the 'Parameter Estimates' table from Proc Reg.

Note the estimate for 'Drug ibuprofe' is 0, with undefined standard error. This occurs because this term is not actually in the model .

The error message shown below can be ignored.

Note: The X'X matrix has been found to be singular, and a generalized inverse was used to solve the normal equations. Terms whose estimates are followed by the letter 'B' are not uniquely estimable.

Fitted Model

- The fitted model is

  $E(Y) = 11.81 + 1.86\,(\text{Severity}) - 4.20\,(\text{Indicator})$

- For acetaminophen, the expected time until pain relief is

  $E(Y) = (11.81 - 4.20) + 1.86\,(\text{Severity})$, or

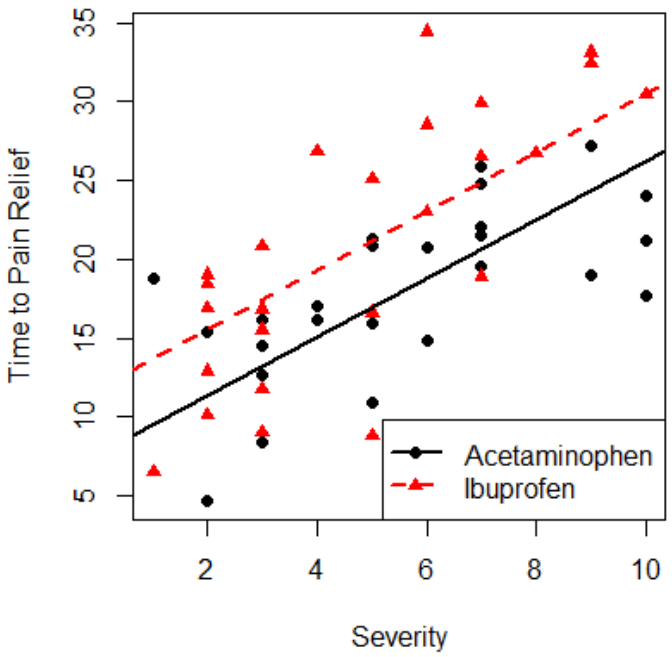
  $E(Y) = 7.61 + 1.86\,(\text{Severity})$

- For ibuprofen, the expected time until pain relief is

  $E(Y) = 11.81 + 1.86\,(\text{Severity})$

- For both drugs, the least squares estimate is a line
  - the two lines have the same slope
  - the two lines have different intercepts

Scatterplot



- This is an additive model
- The two lines are parallel
- For any specified value of severity, if the two drugs have the same mean time to relief then the lines will have the same intercept
  - H_0: $\beta_2 = 0$ vs. H_a: $\beta_2 \neq 0$
- From parameter estimates table
  - t = -2.90, p-value = 0.0056

We conclude that the mean time to pain is different for the different drugs (and this is true for any level of severity).

Interaction

- The additive model forces the two lines to have parallel slopes
- This excludes the possibility that one drug works better for moderate (low severity) headaches, while the other drug works better for chronic (high severity) headaches
- To incorporate this possible, we need to use an interaction model

Interaction Model

- Create interaction variable
  - X1 is severity (continuous) and X2 is indicator
  - interaction variable is X1*X2
- Interaction model: $Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{1i} X_{2i} + \varepsilon_i$
  - If the drug is ibuprofen ($X_2 = 0$), the model reduces to
    $$Y_i = \beta_0 + \beta_1 X_{1i} + \varepsilon_i$$
  - If the drug is acetaminophen ($X_2 = 1$), the model is
    $$Y_i = (\beta_0 + \beta_2) + (\beta_1 + \beta_3) X_{1i} + \varepsilon_i$$
- The intercepts are the same if $\beta_2 = 0$
- The slopes are the same if $\beta_3 = 0$
- The drugs have the same mean response if $\beta_2 = \beta_3 = 0$

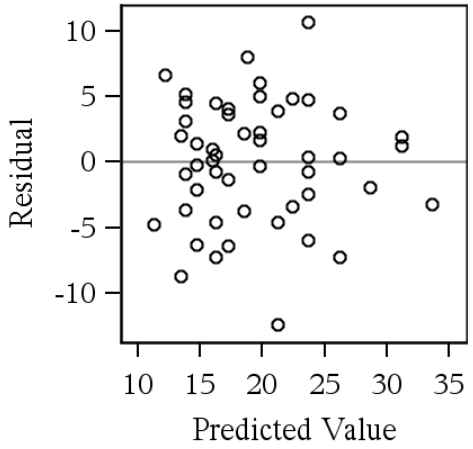
Interaction Model Solution

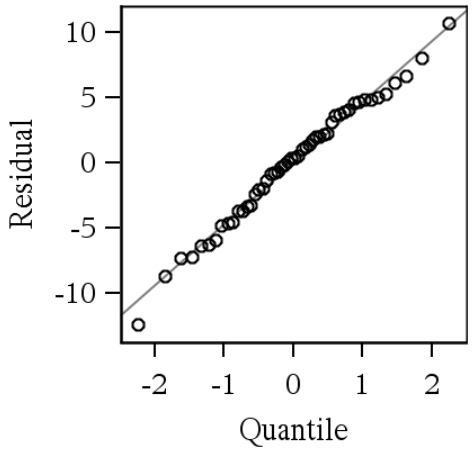
```
proc glm data = headaches plots=all;
class Drug;
model Time = Drug Severity Drug*Severity / solution;
run;
```

Parameter		Estimate		Standard Error	t Value	Pr > |t|
Intercept		8.836	B	2.068	4.27	<.0001
Drug	acetamin	2.042	B	3.081	0.66	0.5108
Drug	ibuprofe	0.000	B	.	.	.
Severity		2.482	B	0.381	6.52	<.0001
Severity*Drug	acetamin	-1.203	B	0.530	-2.27	0.0281
Severity*Drug	ibuprofe	0.000	B	.	.	.

For acetaminophen : Time = 10.878 + 1.279 Severity

For ibuprofen: Time = 8.836 + 2.482 Severity

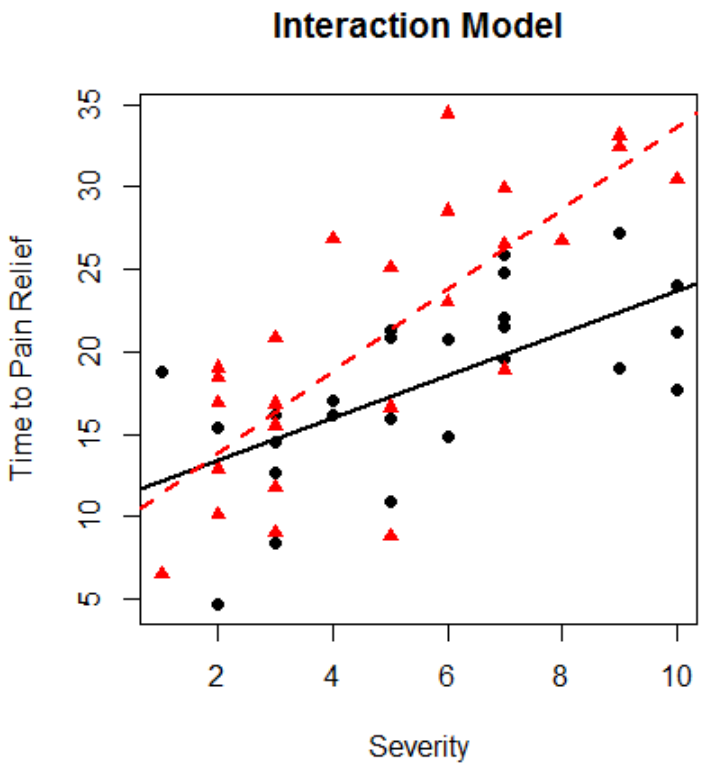




Test for Significant Interaction

- A test for interaction depends on the model
- In a single test, test all terms in the model that are involved with the interaction
- The current model: $Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{1i} X_{2i} + \varepsilon_i$
  - The interaction is a single term
  - Test $H_0: \beta_3 = 0$ vs. $H_a: \beta_3 \neq 0$
- Results (on previous slide):
  - $t = -2.27$, $p = 0.0281$ ⇨ Reject H_0
  - Interaction is significant

Visualize the Interaction



Estimated models

For acetaminophen :

Time = 10.878 + 1.279 Severity

For ibuprofen:

Time = 8.836 + 2.482 Severity

- Intercepts are different
- Slopes are different.
- The lines intersect.
- Drugs are the same for some values of severity and are different for others.

What You Should Know

- Construct indicator variables for qualitative predictors
- Interpret regression lines (or surfaces) based on qualitative predictors
- Test for interaction