Today, we have a bit of a challenge. At a minimum you should fit an Analysis of Covariance to the data (description on the next page). The dataset is ex0920.csv. You may use PROC GLM or PROC REG. If you use PROC REG you will need to prepare the indicator variable and interactions in the data step.

```
data Derby; length Winner $ 16;
            infile input1 missover DSD dlm="," firstobs=2;
   input year winner $ condition $ speed;
     label Winner = 'Name of the Winner'
           Speed = 'Speed (mph)'
           Year = 'Year (1896-2000)'
           Condition = 'Track condition (slow, good, fast)';
   if condition eq 'slow' then do; Cond1 = 1; Cond2 = 0;
                YC1 = year*cond1; YC2 = year*cond2; end;
   if condition eq 'good' then do; Cond1 = 0; Cond2 = 1;
                YC1 = year*cond1; YC2 = year*cond2; end;
   if condition eq 'fast' then do; Cond1 = 0; Cond2 = 0;
                YC1 = year*cond1; YC2 = year*cond2; end;
datalines;
run;
```

Note that everything between the “DO” and the “END” is done if the conditions (IF …) is true.

If you decide to use PROC GLM you need only put the group variable (condition) in the “CLASS” statement. Interactions can be done in the MODEL statement with an asterisk (e.g. year*condition).

Your program should determine if there is a relationship between the dependent variable “speed” and the potential independent variables, year and condition and their interaction. You program should have the following elements.

a) Plots of the dependent variable “speed” on year and on condition.

b) Arrange for the usual comments and title. No HTML needed today.

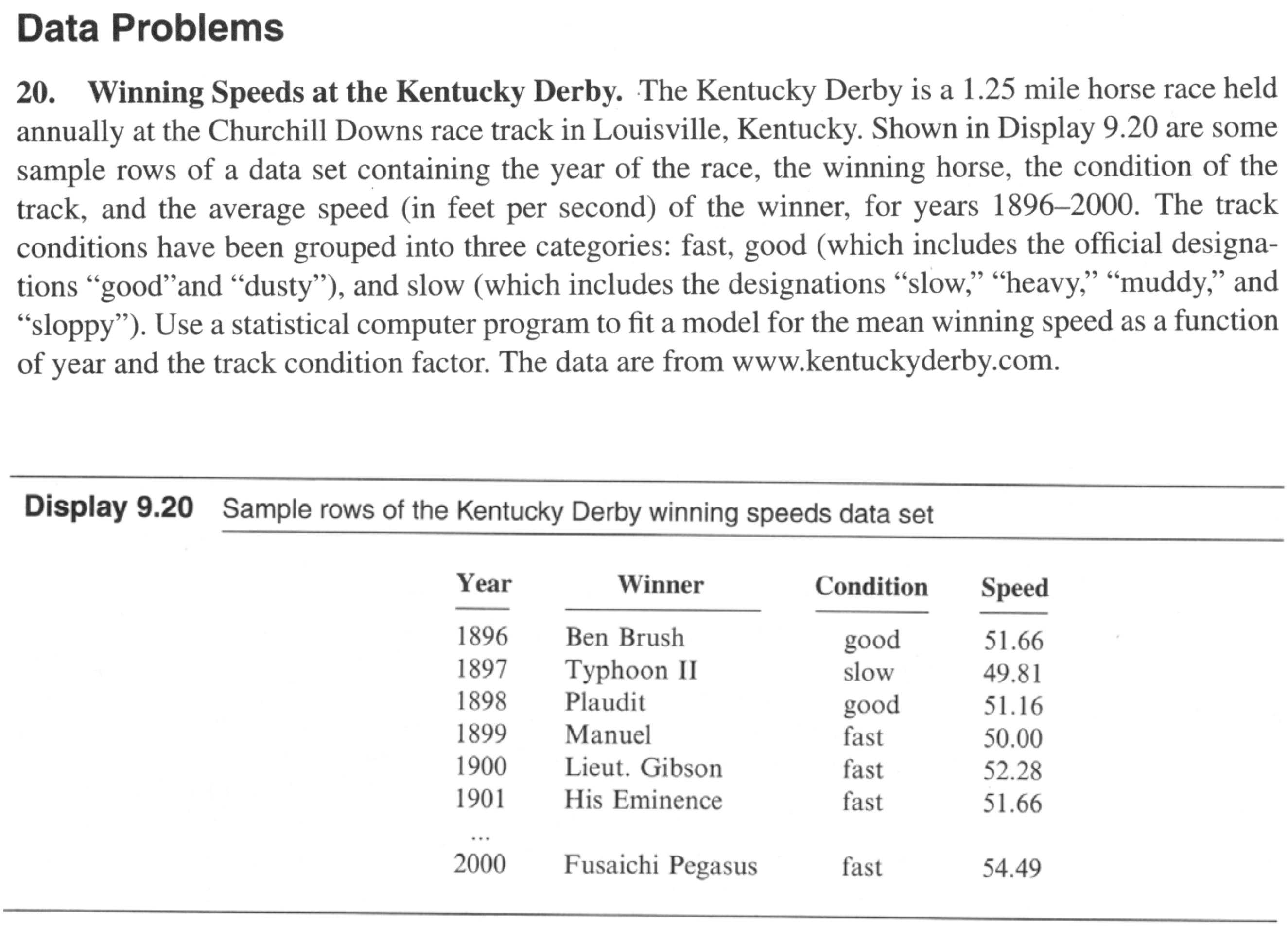
c) Conduct the analysis with both quantitative and qualitative variables and their interaction.

d) For the model described in part “c”, (1) plot the residuals on the predicted value, (2) plot the residuals on year, (3) plot the residuals on condition, and (4) test the residuals for normality. Consider these diagnostics carefully and determine if they provide any information on the suitability of the model?

e) From the analysis in part “c” and “d”, determine if all terms are needed for the model. Also, does it appear that the model can be improved or that it requires additional terms? Finish by fitting the best model as your final model.

For this exercise I am interested in what you determine is the best, final model. If you run several models make the last one the best one and title it “BEST MODEL”. There are no additional questions to be addressed.

OVER



Data Problems

20. Winning Speeds at the Kentucky Derby. The Kentucky Derby is a 1.25 mile horse race held annually at the Churchill Downs race track in Louisville, Kentucky. Shown in Display 9.20 are some sample rows of a data set containing the year of the race, the winning horse, the condition of the track, and the average speed (in feet per second) of the winner, for years 1896–2000. The track conditions have been grouped into three categories: fast, good (which includes the official designations "good"and "dusty"), and slow (which includes the designations "slow," "heavy," "muddy," and "sloppy"). Use a statistical computer program to fit a model for the mean winning speed as a function of year and the track condition factor. The data are from www.kentuckyderby.com.

Display 9.20 Sample rows of the Kentucky Derby winning speeds data set

Year	Winner	Condition	Speed
1896	Ben Brush	good	51.66
1897	Typhoon II	slow	49.81
1898	Plaudit	good	51.16
1899	Manuel	fast	50.00
1900	Lieut. Gibson	fast	52.28
1901	His Eminence	fast	51.66
...			
2000	Fusaichi Pegasus	fast	54.49

OVER