36-402/608 Homework 9 Solutions: SAS March 25

Problem 1 (50 points)

Your code (30 points) should always include a title. The infile statement includes "DSD" to handle comma-separated-values and "firstobs=2" to skip the header line in the file. The "player=1" and "if" statements are one way to create the indicator variable. You always need to print at least some of the file to verify correct file reading and variable creation.

ALWAYS check the log output; it showed no problems for my code.

```
title "Violin Data (HW9, problem 1)";
data violin;
  infile "ex0730.csv" dsd firstobs=2;
  input years activity;
  player=1;
  if years=0 then player=0;
run;
proc print;
run;
```

Standard (non-graphical) EDA is frequencies for categorical variables and much of the stuff included under "univariate" for quantitative variables. Graphical EDA in the form of a barplot for player and/or a scatter-plot for years vs. activity earns you 2 bonus point each.

```
proc freq;
  tables player;
run;
proc univariate;
  var years activity;
run;
```

If we treat years of playing as categorical (player or not) we can analyze these data as an ANOVA (or independent samples t-test). The "class" statement tells "proc anova" that player is a categorical variable. If you made a residual vs. fit plot you would see that we violated the equal variance assumption, which is why the regression is a better analysis.

```
proc anova;
  class player;
  model activity=player;
  means player;
run;
```

You can use "proc reg" or "proc glm" to do the simple regression.

```
proc reg;
  model activity=years;
  plot residual.*predicted.;
run;
```

Results:

Obs	years	activity	player
1	0	5.0	0
2	0	6.0	0
3	0	7.5	0
4	0	9.0	0
5	0	9.5	0
6	0	11.0	0
7	5	16.0	1
8	6	16.5	1
9	8	11.5	1
10	10	16.0	1
11	12	25.0	1
12	13	25.5	1
13	17	25.5	1
14	18	23.0	1
15	19	26.5	1

The FREQ Procedure

player	Frequency	Percent	Cumulative Frequency	Cumulative Percent
0	6	40.00	6	40.00
1	9	60.00	15	100.00

The UNIVARIATE Procedure
Variable: years

Moments

N	15	Sum Weights	15
Mean	7.2	Sum Observations	108
Std Deviation	7.24273035	Variance	52.4571429
Skewness	0.41655124	Kurtosis	-1.3695395
Uncorrected SS	1512	Corrected SS	734.4
Coeff Variation	100.593477	Std Error Mean	1.87006493

Basic Statistical Measures

Location		Variability	
Mean	7.200000	Std Deviation	7.24273
Median	6.000000	Variance	52.45714
Mode	0.000000	Range	19.00000
		Interquartile Range	13.00000

The UNIVARIATE Procedure
Variable: activity

Moments

N	15	Sum Weights	15
Mean	15.5666667	Sum Observations	233.5
Std Deviation	7.78245891	Variance	60.5666667
Skewness	0.22191434	Kurtosis	-1.5604338
Uncorrected SS	4482.75	Corrected SS	847.933333
Coeff Variation	49.9943827	Std Error Mean	2.00942225

Basic Statistical Measures

Location		Variability	
Mean	15.56667	Std Deviation	7.78246
Median	16.00000	Variance	60.56667
Mode	16.00000	Range	21.50000
		Interquartile Range	16.00000

Note: The mode displayed is the smallest of 2 modes with a count of 2.

The ANOVA Procedure

Class Level Information

Class	Levels	Values
player	2	0 1

Number of Observations Read	15
Number of Observations Used	15

Dependent Variable: activity

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	572.5444444	572.5444444	27.03	0.0002
Error	13	275.3888889	21.1837607		
Corrected Total	14	847.9333333			

R-Square	Coeff Var	Root MSE	activity Mean
0.675223	29.56691	4.602582	15.56667

Source	DF	Anova SS	Mean Square	F Value	Pr > F
player	1	572.5444444	572.5444444	27.03	0.0002

Level of player	N	-----------activity---------- Mean	Std Dev
0	6	8.0000000	2.25831796
1	9	20.6111111	5.58892755

We see a significant difference in activity level (p=0.0002) with the mean difference estimated to be about 12.6 units higher for players.

The REG Procedure
Model: MODEL1
Dependent Variable: activity

Number of Observations Read	15
Number of Observations Used	15

Analysis of Variance

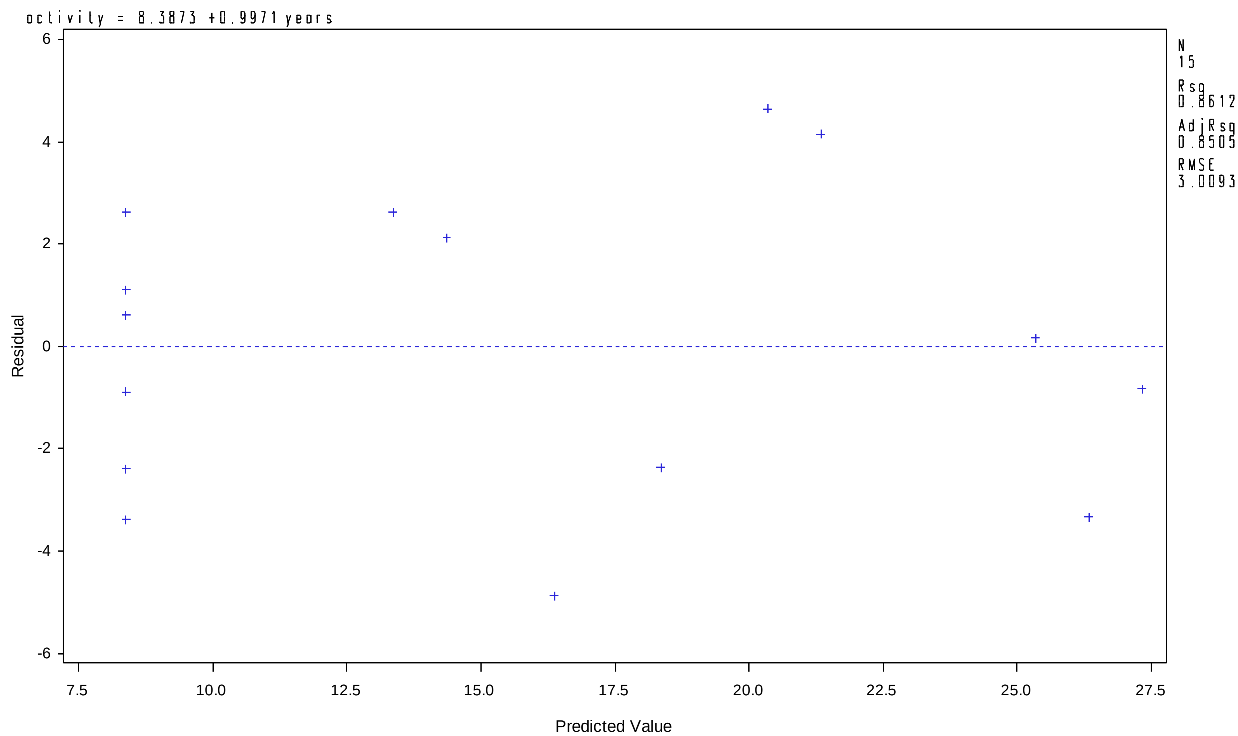
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	730.20600	730.20600	80.63	<.0001
Error	13	117.72733	9.05595		
Corrected Total	14	847.93333			

Root MSE	3.00931	R-Square	0.8612
Dependent Mean	15.56667	Adj R-Sq	0.8505
Coeff Var	19.33176		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	8.38725	1.11489	7.52	<.0001
years	1	0.99714	0.11105	8.98	<.0001

Violin Data (HW9, problem 1)



We see a slope (p<0.0001) indicating an estimate rise in mean activity of 0.997 per year with a residual standard error of 3.01 around the mean.

Answers to the questions (20 points):

There is strong evidence of a difference in neuron activity between the stringed musicians and the controls by ANOVA (p=0.0002, F=27.03 with 1,13 df). (The p-value is not fully reliable due to unequal variance of activity when compared between the two groups, but any correction would not change the value enough to change the overall conclusion.) "Different" is insufficient for any conclusion, you must (2 points) state that the higher activity is in the string players.

There is strong evidence that the amount of activity is associated with the number of years of string playing (not "caused by", because there is no randomization of treatment). Using (simple) regression we see strong evidence that the amount of neuronal activity (in the brain region studied here) rises (- 2 points for just "changes") by 1.00 units for each additional year of string playing (p<0.0001, t=8.98, df=13 (not 1!!), SE=0.11, approximate 95% CI = [0.78, 1.22]). You really should include the SE and/or the CI because point estimates are not interpretable scientifically.

Problem 2 (50 points)

Again, your code should start with a title. Although the “informat” statement is not required in this problem, the winner name is cutoff at 8 characters without it (and space is wasted for the condition).

```
title "Kentucky Derby (HW9, problem 2)";
  data derby;
  infile "ex0920.csv" dsd firstobs=2;
  informat winner $20. condition $4.;
  input year winner condition speed;
run;
proc print data=derby(obs=7);
run;
```

The EDA is similar to the previous problem.

```
proc freq;
  tables condition;
run;
  proc univariate;
  var year speed;
run;
```

The code here explicitly models condition as categorical and year as quantitative, as required.

```
proc glm;
  class condition;
  model speed = year condition / solution;
  output out=derbydiag residual=res predicted=fit;
run;
proc gplot data=derbydiag;
  plot res*fit / vref=0;
run;
```

You get 2 bonus points for including a quantile-normal plot (with or without the reference line that requires the extra “univariate” step on the residuals to obtain the residual sd of 0.641.

```
proc univariate;
  var res;
run;
proc univariate noprint;
  var res;
  qqplot / normal (mu=0 sigma=0.641 color=red);
run;
```

You get 5 bonus points for noting the evidence of non-linearity on the residual vs. fit plot and attempting to deal with it in any way. I added a square term for year, which worked well.

```
title2 "Trying square term for year";
  data derby;
  set derby;
  year2 = year*year;
run;
proc glm;
  class condition;
  model speed = year year2 condition / solution;
  output out=derbydiag residual=res predicted=fit;
run;
proc gplot data=derbydiag;
  plot res*fit / vref=0;
run;
proc univariate;
  var res;
run;
proc univariate noprint;
  var res;
  qqplot / normal (mu=0 sigma=0.538 color=red);
run;
```

Here are the results:

Kentucky Derby (HW9, problem 2) 1

08:52 Thursday, March 18, 2010

Obs	winner	condition	year	speed
1	Ben Brush	good	1896	51.6634
2	Typhoon II	slow	1897	49.8113
3	Plaudit	good	1898	51.1628
4	Manuel	fast	1899	50.0000
5	Lieut. Gibson	fast	1900	52.2772
6	His Eminence	fast	1901	51.6634
7	Alan-a-Dale	fast	1902	51.2621

The FREQ Procedure

condition	Frequency	Percent	Cumulative Frequency	Cumulative Percent
fast	75	71.43	75	71.43
good	10	9.52	85	80.95
slow	20	19.05	105	100.00

The UNIVARIATE Procedure
Variable: year

Moments

N	105	Sum Weights	105
Mean	1948	Sum Observations	204540
Std Deviation	30.4548847	Variance	927.5
Skewness	0	Kurtosis	-1.2
Uncorrected SS	398540380	Corrected SS	96460
Coeff Variation	1.56339244	Std Error Mean	2.97209242

Basic Statistical Measures

Location		Variability	
Median	1948.000	Variance	927.50000
Mode	.	Range	104.00000
		Interquartile Range	52.00000

The UNIVARIATE Procedure
Variable: speed

Moments

N	105	Sum Weights	105
Mean	53.0424326	Sum Observations	5569.45543
Std Deviation	1.33169903	Variance	1.77342232
Skewness	-0.9319577	Kurtosis	0.33624777
Uncorrected SS	295601.9	Corrected SS	184.435921
Coeff Variation	2.51062964	Std Error Mean	0.12996052

Basic Statistical Measures

Location		Variability	
Mean	53.04243	Std Deviation	1.33170
Median	53.39806	Variance	1.77342
Mode	53.92157	Range	6.45981
		Interquartile Range	1.73259

The GLM Procedure

Class Level Information

Class	Levels	Values
condition	3	fast good slow

Number of Observations Read	105
Number of Observations Used	105

Dependent Variable: speed

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	141.7030990	47.2343663	111.64	<.0001
Error	101	42.7328218	0.4230972		
Corrected Total	104	184.4359208			

R-Square	Coeff Var	Root MSE	speed Mean
0.768305	1.226300	0.650459	53.04243

Source	DF	Type I SS	Mean Square	F Value	Pr > F
year	1	103.5684162	103.5684162	244.79	<.0001
condition	2	38.1346829	19.0673414	45.07	<.0001

Source	DF	Type III SS	Mean Square	F Value	Pr > F
year	1	64.26378114	64.26378114	151.89	<.0001
condition	2	38.13468286	19.06734143	45.07	<.0001

Parameter	Estimate	Standard Error	t Value	Pr > |t|
Intercept	-0.679883986 B	4.22339304	-0.16	0.8724
year	0.026931522	0.00218523	12.32	<.0001
condition fast	1.615054231 B	0.17041109	9.48	<.0001
condition good	1.114072805 B	0.25289631	4.41	<.0001
condition slow	0.000000000 B	.	.	.

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.

Variable: res

Moments

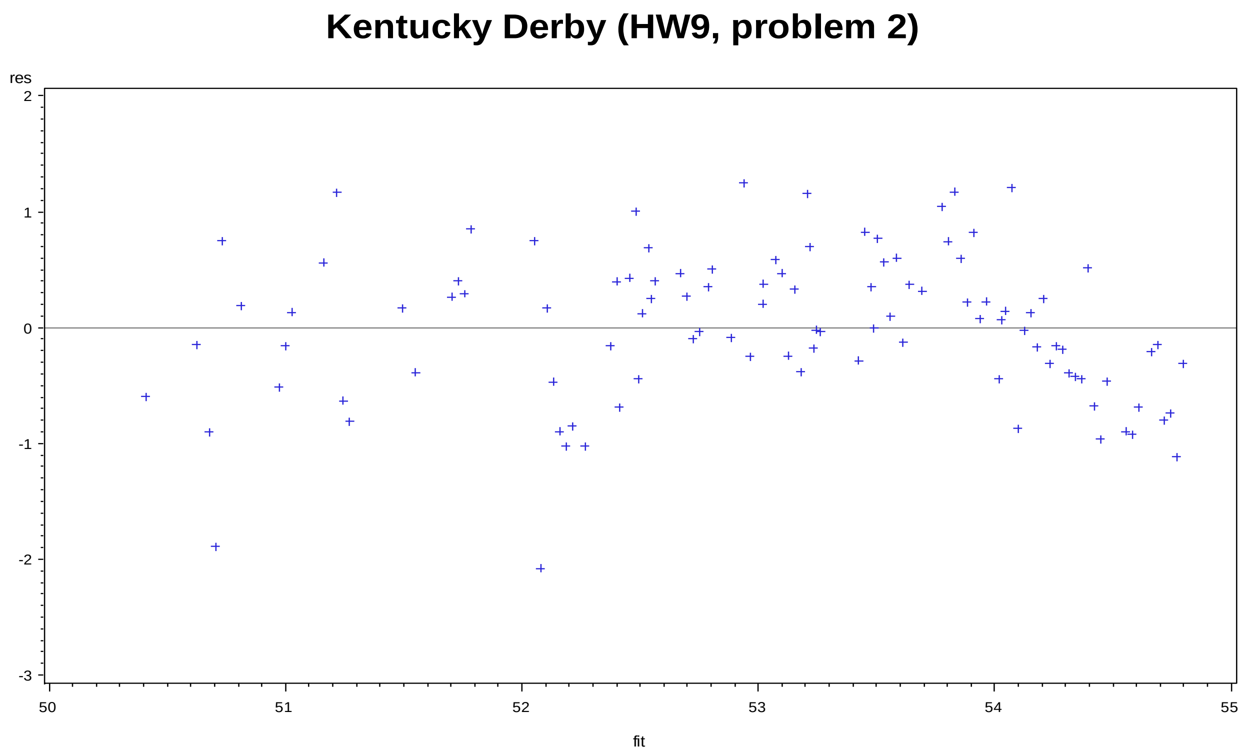
Std Deviation	0.64100898	Variance	0.41089252

Trying square term for year 08:52 Thursday, March 18, 2010

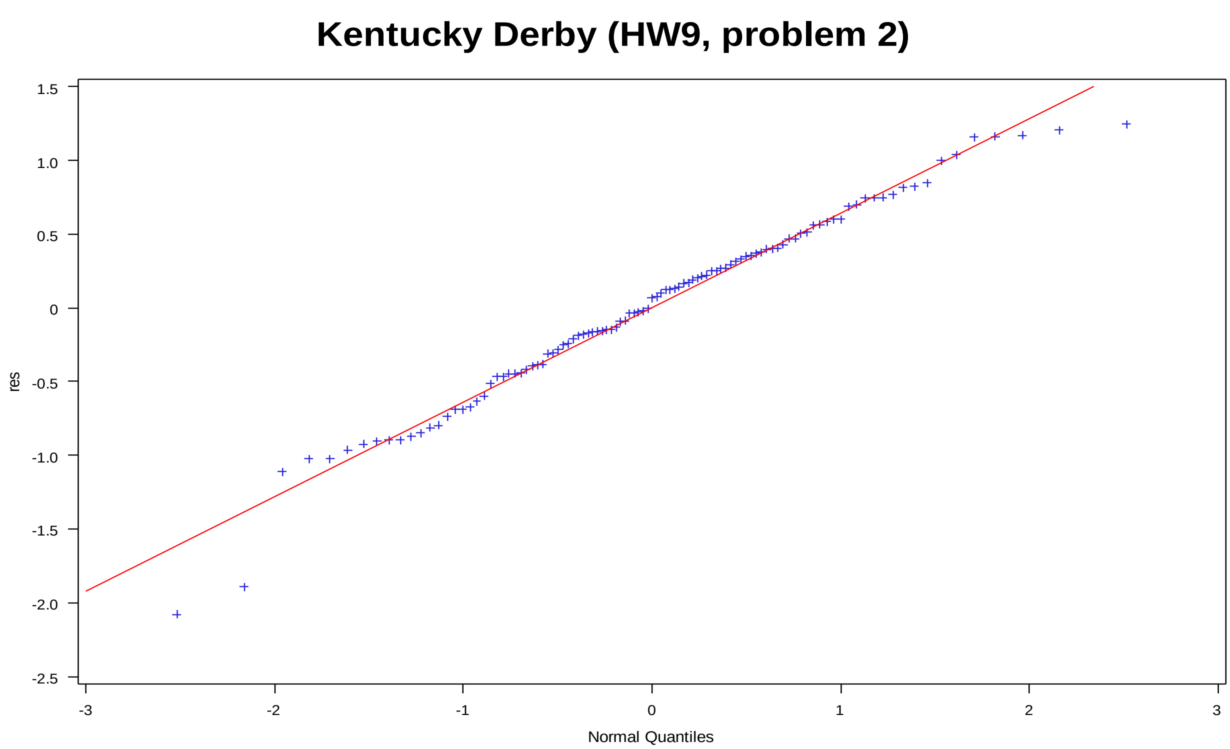
R-Square	Coeff Var	Root MSE	speed Mean
0.836479	1.035346	0.549173	53.04243

Parameter	Estimate	Standard Error	t Value	Pr > |t|
Intercept	-1599.247282 B	247.6012549	-6.46	<.0001
year	1.668578	0.2542541	6.56	<.0001
year2	-0.000421	0.0000653	-6.46	<.0001
condition fast	1.609853 B	0.1438778	11.19	<.0001
condition good	1.077923 B	0.2135899	5.05	<.0001
condition slow	0.000000 B	.	.	.

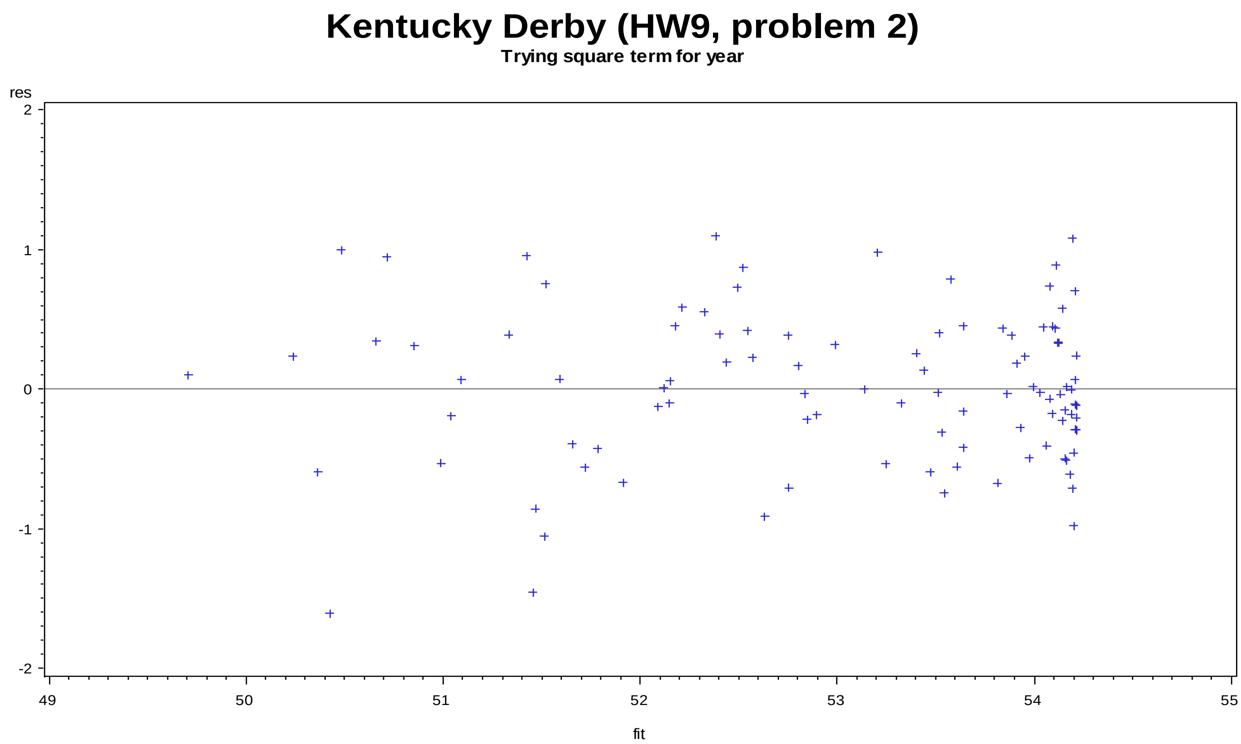
Note: The "singular matrix" message does not indicate a problem; it only means that the two "condition" estimates must be interpreted compared to an arbitrary baseline (as in R).



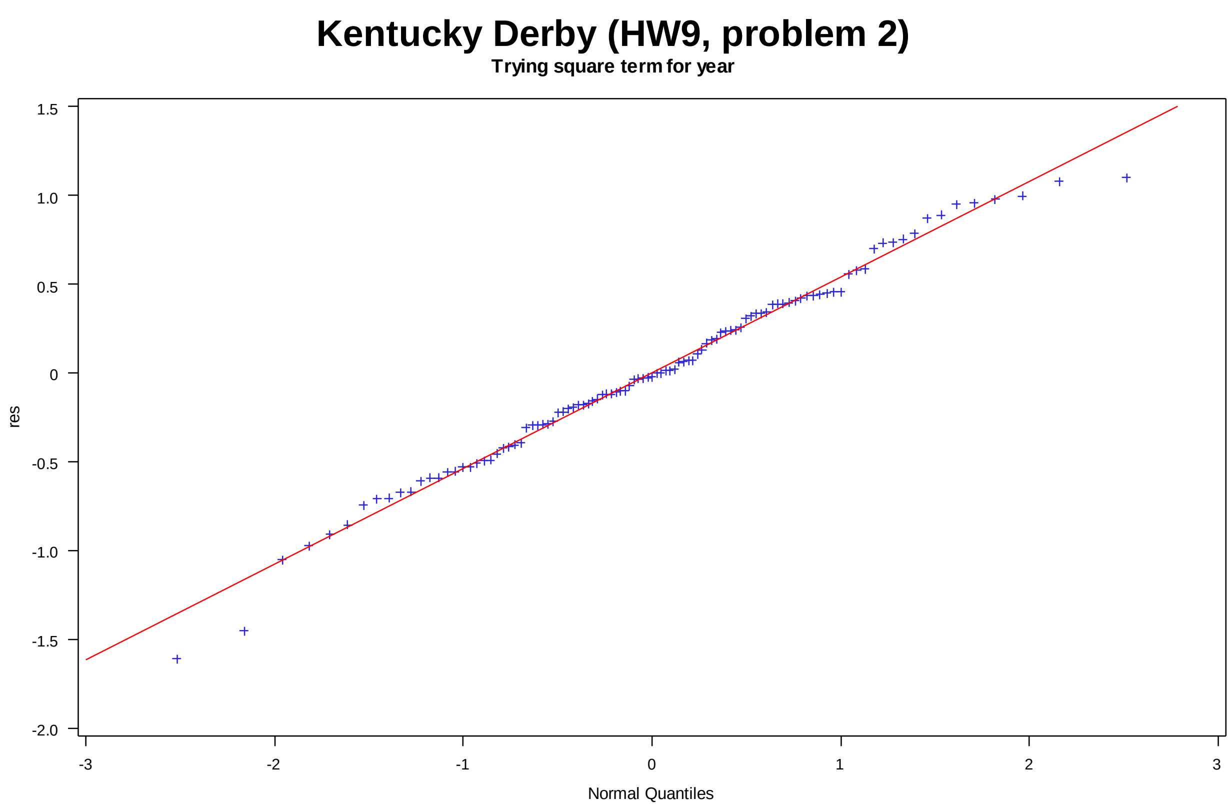
I see evidence of non-linearity in the above plot.



This is really quite consistent with Normal errors plus about 4 outliers (or a very slight skew left).



The non-linearity is fixed.



From the table plus a few other points of output:

Parameter	Estimate	Standard Error	t Value	Pr > |t|
Intercept	-1599.247282 B	247.6012549	-6.46	<.0001
year	1.668578	0.2542541	6.56	<.0001
year2	-0.000421	0.0000653	-6.46	<.0001
condition fast	1.609853 B	0.1438778	11.19	<.0001
condition good	1.077923 B	0.2135899	5.05	<.0001
condition slow	0.000000 B	.	.	.

we could summarize our conclusions as follows: 84% of the variability in speed is explained by allowing three parallel quadratic relationships between speed and year of the race, separate for fast , good, and slow track conditions. The quadratic curve is convex upward with an extrapolate peak at year=3962 [from setting $d(1599.25 +1.668Y -0.000421Y^2)/dY = 0$]. So the trend is upward with a very slight "leveling off" direction of curvature. Good conditions increase speed by 1.08 mph (95%CI = [0.60,1.50]) compared to slow conditions. Fast conditions increase speed by 1.61 mph (95%CI = [0.32,1.90]) compared to slow conditions. Over the 1896 to 2000, the estimate rise in speed is 2.89 mph.

Ask if you don't know how to do any of these calculations. There are 5 bonus points available for a good summary.