Exercise 1

a) From the tabulation, it looks like 4 cylinder cars may be more fuel efficient than 6 cylinders. Sedans might be more efficient than sports cars at least for 4 cylinders. There does not appear to be a big difference across origin. We also see that the data is not balanced, so we will need to use proc glm.

			MPG (Highway)		
			Mean	Std	N
Cylinders	Origin	Type			
4	Asia	Sedan	33.35	4.27	49
		Sports	27.88	3.18	8
	USA	Sedan	32.69	3.31	29
6	Asia	Sedan	26.56	1.84	41
		Sports	26.33	1.51	6
	USA	Sedan	27.27	2.90	45
		Sports	27.00	2.83	2

b) In the main effects model, we see that Cylinders and Type are significant but Origin is not.

Dependent Variable: MPG_Highway MPG (Highway)

Source	DF	Type I SS	Mean Square	F Value	Pr > F
Cylinders	1	1470.787732	1470.787732	137.50	<.0001
Origin	1	8.564346	8.564346	0.80	0.3721
Type	1	108.057489	108.057489	10.10	0.0018

Source	DF	Type III SS	Mean Square	F Value	Pr > F
Cylinders	1	1453.170429	1453.170429	135.85	<.0001
Origin	1	0.841224	0.841224	0.08	0.7795
Type	1	108.057489	108.057489	10.10	0.0018

Removing Origin we get our best main effects model. The model is highly significant, both Cylinders and Type are highly significant, and the model describes 45.7% of the variation in highway fuel efficiency.

Dependent Variable: MPG_Highway MPG (Highway)

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	1586.568342	793.284171	74.55	<.0001
Error	177	1883.492769	10.641202		
Corrected Total	179	3470.061111			

R-Square	Coeff Var	Root MSE	MPG_Highway Mean
0.457216	11.01023	3.262086	29.62778

Source	DF	Type I SS	Mean Square	F Value	Pr > F
Cylinders	1	1470.787732	1470.787732	138.22	<.0001
Type	1	115.780611	115.780611	10.88	0.0012

Source	DF	Type III SS	Mean Square	F Value	Pr > F
Cylinders	1	1481.993512	1481.993512	139.27	<.0001
Type	1	115.780611	115.780611	10.88	0.0012

c) The Cylinders*Type interaction is also significant when added to the model, so the final model has Cylinders, Type, and their interaction. The model is highly significant as are the three terms. The model describes 48.14% of the variation in fuel efficiency.

Dependent Variable: MPG_Highway MPG (Highway)

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	1670.425229	556.808410	54.45	<.0001
Error	176	1799.635883	10.225204		
Corrected Total	179	3470.061111			

R-Square	Coeff Var	Root MSE	MPG_Highway Mean
0.481382	10.79287	3.197687	29.62778

Source	DF	Type I SS	Mean Square	F Value	Pr > F
Cylinders	1	1470.787732	1470.787732	143.84	<.0001
Type	1	115.780611	115.780611	11.32	0.0009
Cylinders*Type	1	83.856886	83.856886	8.20	0.0047

Source	DF	Type III SS	Mean Square	F Value	Pr > F
Cylinders	1	207.5516175	207.5516175	20.30	<.0001
Type	1	116.6363540	116.6363540	11.41	0.0009
Cylinders*Type	1	83.8568863	83.8568863	8.20	0.0047

The following differences of least squares means tell us that 4 cylinder cars are expected to be 3.77 mpg more efficient than 6 cylinders with a confidence interval of (2.12, 5.43). Sedans are 2.83 mpg more efficient than sports cars with a confidence interval of (1.18, 4.48), and 4 cylinder sedans are significantly more fuel efficient than the other types of cars.

Least Squares Means
Adjustment for Multiple Comparisons: Tukey-Kramer

Cylinders	MPG_Highway LSMEAN	H0:LSMean1=LSMean2	
		t Value	Pr > |t|
4	30.4887821	4.51	<.0001
6	26.7151163		

Least Squares Means for Effect Cylinders				
i	j	Difference Between Means	Simultaneous 95% Confidence Limits for LSMean(i)-LSMean(j)	
1	2	3.773666	2.120634	5.426697

Least Squares Means
Adjustment for Multiple Comparisons: Tukey-Kramer

Type	MPG_Highway LSMEAN	H0:LSMean1=LSMean2	
		t Value	Pr > |t|
Sedan	30.0163983	3.38	0.0009
Sports	27.1875000		

Least Squares Means for Effect Type				
i	j	Difference Between Means	Simultaneous 95% Confidence Limits for LSMean(i)-LSMean(j)	
1	2	2.828898	1.175867	4.481930

Least Squares Means
Adjustment for Multiple Comparisons: Tukey-Kramer

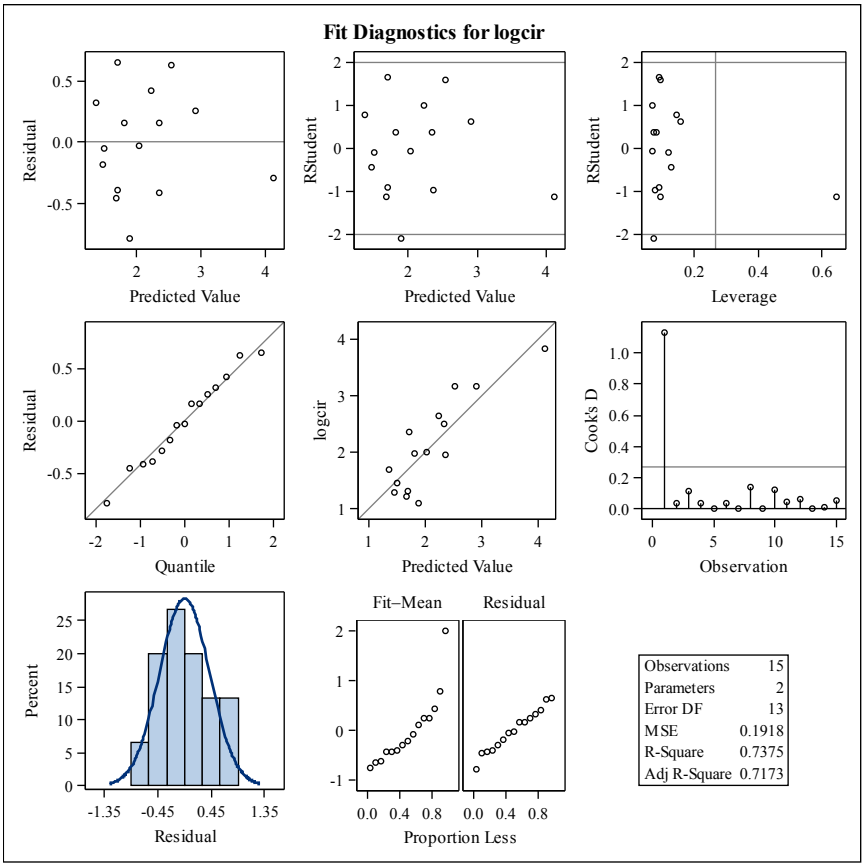
Cylinders	Type	MPG_Highway LSMEAN	LSMEAN Number
4	**Sedan**	33.1025641	1
4	**Sports**	27.8750000	2
6	**Sedan**	26.9302326	3
6	**Sports**	26.5000000	4

Least Squares Means for Effect Cylinders*Type				
			Simultaneous 95% Confidence Limits for LSMean(i)-LSMean(j)	
i	j	Difference Between Means		
1	**2**	5.227564	2.148489	8.306639
1	**3**	6.172332	4.875485	7.469178
1	**4**	6.602564	3.523489	9.681639
2	**3**	0.944767	-2.120956	4.010491
2	**4**	1.375000	-2.771993	5.521993
3	**4**	0.430233	-2.635491	3.495956

Exercise 2

a) The following diagnostics for the log-cirrhosis rate linear regression indicate that the first observation with a large Cook's distance is unduly influential and so we will remove it and refit the model. After doing so, diagnostics for the model (shown in part b) look fine.

Model: MODEL1
Dependent Variable: logcir



b) For the final model, we see that the ANOVA test is significant, so the model is better than an error only model. The model describes 64.4% of the variation in log cirrhosis death rates. The intercept and alcohol terms are both significant. The coefficient of .1578 for alcohol tells us that the expected cirrhosis death rate would be multiplied by $e^{.1578} = 1.171$ when the alcohol consumption level increases by 1, so again we see an increase in cirrhosis related death rate as alcohol consumption increases. For 0 alcohol consumption the model would predict a cirrhosis rate of $e^{.7536} = 2.125$.

We see no remaining issues in the diagnostics. While the distribution of residuals was acceptable in the regular linear model, they look even better in this case. We cannot compare the R^2 values directly because the responses are different, but they are fairly similar in magnitude so we still explain a pretty good amount of the variation in cirrhosis related death rates. This model also has the benefit of always predicting non-negative values. This model appears to be slightly better than the regular linear model.

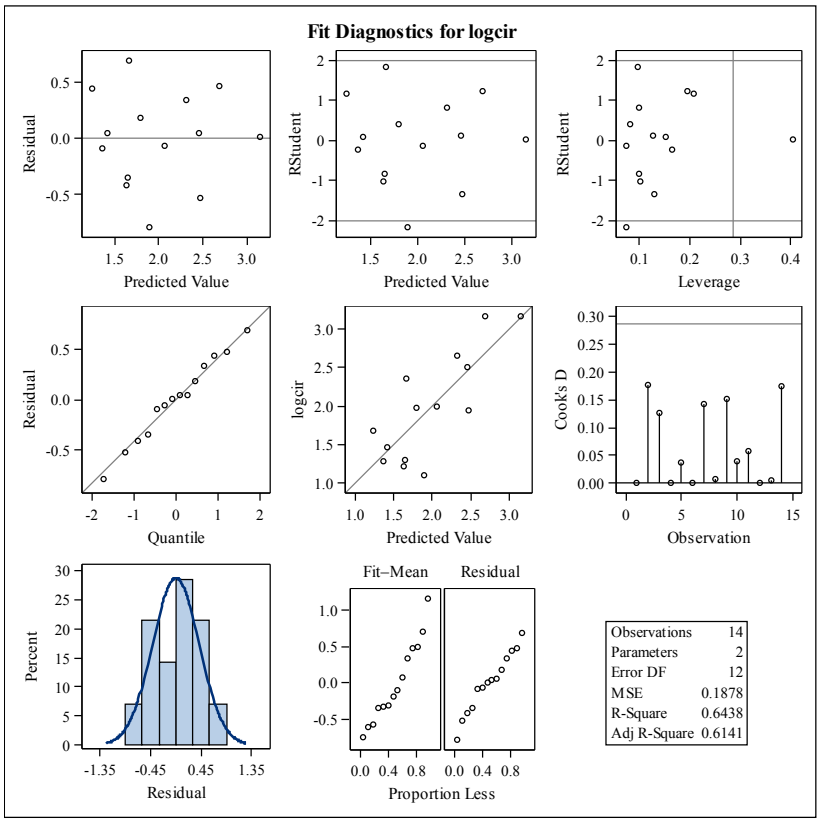
Model: MODEL1
Dependent Variable: logcir

Analysis of Variance					
Source	**DF**	**Sum of Squares**	**Mean Square**	**F Value**	**Pr > F**
Model	1	4.07318	4.07318	21.69	0.0006
Error	12	2.25370	0.18781		
Corrected Total	13	6.32689			

Root MSE	0.43337	**R-Square**	0.6438
Dependent Mean	1.98775	**Adj R-Sq**	0.6141
Coeff Var	21.80194		

Parameter Estimates					
Variable	**DF**	**Parameter Estimate**	**Standard Error**	**t Value**	**Pr > |t|**
Intercept	1	0.75356	0.28922	2.61	0.0230
alcohol	1	0.15780	0.03388	4.66	0.0006

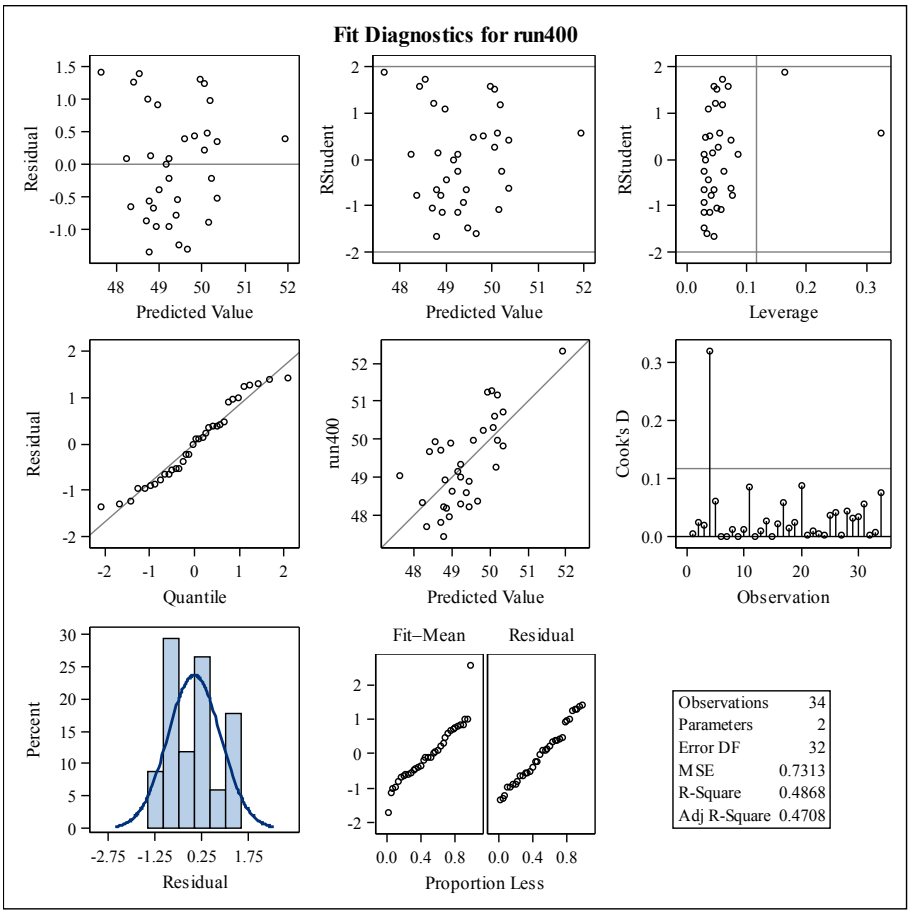
Model: MODEL1
Dependent Variable: logcir



Exercise 3

a) First we check the diagnostics for run400 as a function of run100. There is one Cook's distance which is a fair amount larger than the others, so we should remove that observation.

Model: MODEL1
Dependent Variable: run400



After removing the observation, we get the following model. The model is significantly better than error, describes 53.85% of the variation in 400 meter dash time, and predicts that for a 1 second increase in 100 meter dash time we would expect a 3.23 second increase in 400 meter dash time.

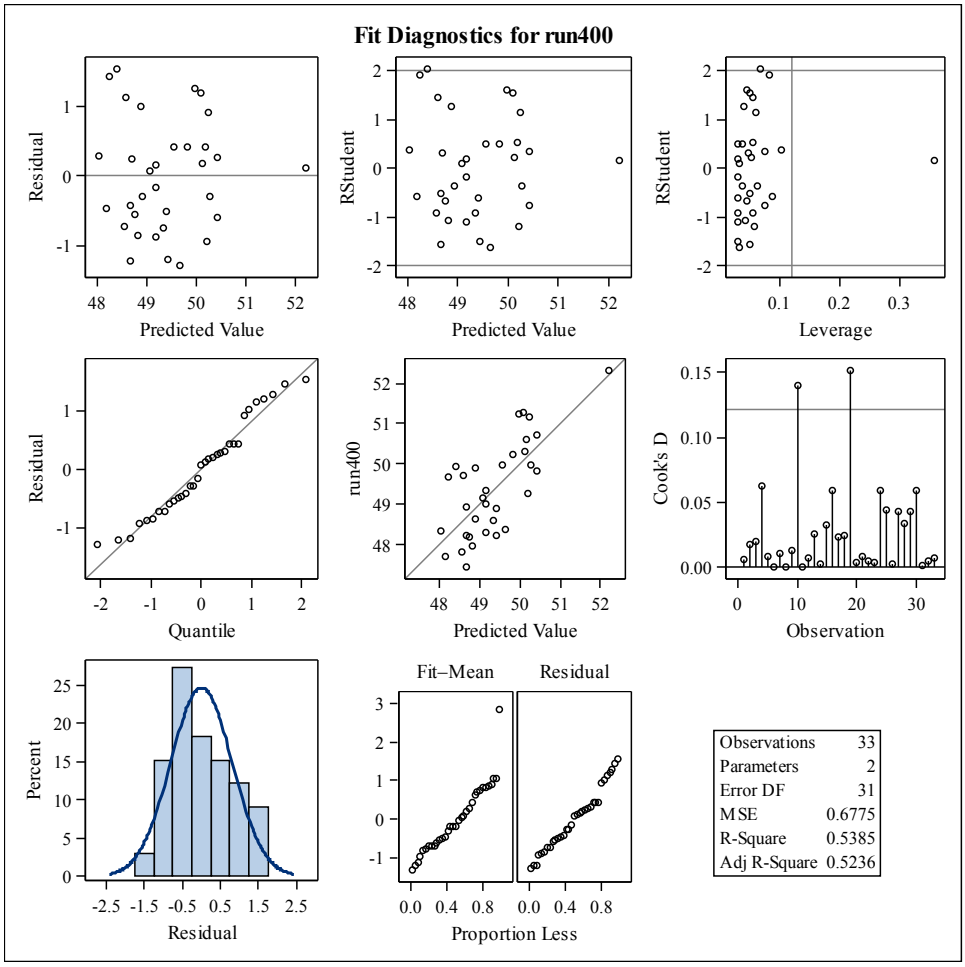
Model: MODEL1
Dependent Variable: run400

Analysis of Variance					
Source	**DF**	**Sum of Squares**	**Mean Square**	**F Value**	**Pr > F**
Model	1	24.50471	24.50471	36.17	<.0001
Error	31	21.00171	0.67747		
Corrected Total	32	45.50642			

Root MSE	0.82309	**R-Square**	0.5385
Dependent Mean	49.37545	**Adj R-Sq**	0.5236
Coeff Var	1.66700		

Parameter Estimates					
Variable	**DF**	**Parameter Estimate**	**Standard Error**	**t Value**	**Pr > |t|**
Intercept	1	13.05278	6.04118	2.16	0.0386
run100	1	3.23103	0.53723	6.01	<.0001

Model: MODEL1
Dependent Variable: run400



b) Using the 1500 meter time as the predictor, there are no significant issues in the diagnostics, so there is no need to refit. The model is significantly better than error, describes 30.74% of the variation in 400 meter dash time, and predicts that for a 1 second increase in 1500 meter dash time we would expect a .04836 second increase in 400 meter dash time.

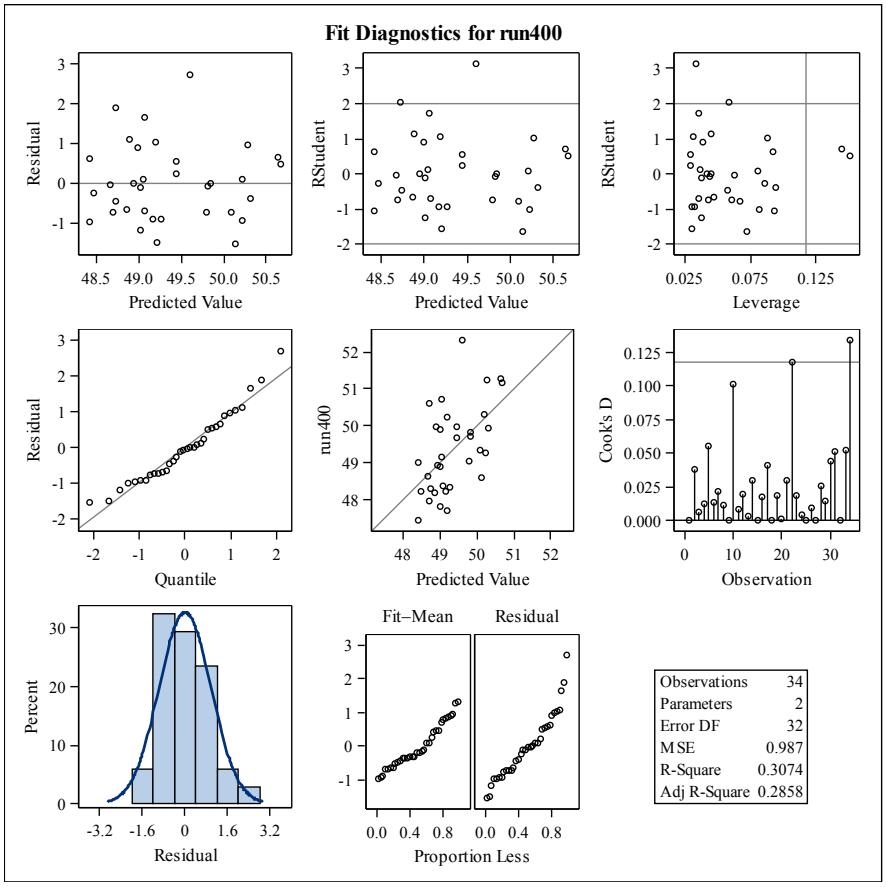
Model: MODEL1
Dependent Variable: run400

Analysis of Variance					
Source	**DF**	**Sum of Squares**	**Mean Square**	**F Value**	**Pr > F**
Model	1	14.01932	14.01932	14.20	0.0007
Error	32	31.58368	0.98699		
Corrected Total	33	45.60300			

Root MSE	0.99347	**R-Square**	0.3074
Dependent Mean	49.36618	**Adj R-Sq**	0.2858
Coeff Var	2.01246		

Parameter Estimates					
Variable	**DF**	**Parameter Estimate**	**Standard Error**	**t Value**	**Pr > |t|**
Intercept	1	36.00985	3.54798	10.15	<.0001
run1500	1	0.04836	0.01283	3.77	0.0007

Model: MODEL1
Dependent Variable: run400



c) Neither model has issues in the diagnostics, so we would choose the model based on 100 meter dash time because it has a better R^2 value, predicting almost 54% of the variation while the 1500 meter model describes less than 31% of the variation.