Exercise1

The GENMOD Procedure

Model Information	
Data Set	WORK.AUTO
Distribution	Gamma
Link Function	Log
Dependent Variable	highwaympg

Criteria For Assessing Goodness Of Fit			
Criterion	**DF**	**Value**	**Value/DF**
Deviance	190	1.9470	0.0102
Scaled Deviance	190	197.3240	1.0385
Pearson Chi-Square	190	1.9643	0.0103
Scaled Pearson X2	190	199.0784	1.0478
Log Likelihood		-494.2043	
Full Log Likelihood		-494.2043	
AIC (smaller is better)		1004.4087	
AICC (smaller is better)		1005.1747	
BIC (smaller is better)		1030.6743	

Analysis Of Maximum Likelihood Parameter Estimates								
Parameter		**DF**	**Estimate**	**Standard Error**	**Wald 95% Confidence Limits**		**Wald Chi-Square**	**Pr > ChiSq**
Intercept		1	4.1811	0.1992	3.7908	4.5715	440.76	<.0001
weight		1	-0.0003	0.0000	-0.0003	-0.0002	65.27	<.0001
height		1	0.0012	0.0039	-0.0065	0.0090	0.10	0.7511
horsepower		1	-0.0034	0.0004	-0.0041	-0.0026	77.74	<.0001
enginesize		1	0.0015	0.0004	0.0006	0.0023	11.22	0.0008
price		1	-0.0000	0.0000	-0.0000	0.0000	0.08	0.7830
ndoors	four	1	0.0074	0.0174	-0.0267	0.0416	0.18	0.6695
ndoors	two	0	0.0000	0.0000	0.0000	0.0000	.	.
Scale		1	101.3477	10.1949	83.2125	123.4352		

Note: The scale parameter was estimated by maximum likelihood.

LR Statistics For Type 1 Analysis				
Source	**2*LogLikelihood**	**DF**	**Chi-Square**	**Pr > ChiSq**
Intercept	-1306.0078			
weight	-1079.8538	1	226.15	<.0001
height	-1062.4948	1	17.36	<.0001
horsepower	-1000.7707	1	61.72	<.0001
enginesize	-988.6775	1	12.09	0.0005
price	-988.5907	1	0.09	0.7684
ndoors	-988.4087	1	0.18	0.6696

LR Statistics For Type 3 Analysis			
Source	**DF**	**Chi-Square**	**Pr > ChiSq**
weight	1	56.36	<.0001
height	1	0.10	0.7511
horsepower	1	65.05	<.0001
enginesize	1	10.90	0.0010
price	1	0.08	0.7831
ndoors	1	0.18	0.6696

a) According to the results of the gamma model with log link, the type 1 analysis tell us that the weight, height, horsepower, and enginesize are significant, while price and ndoors are insignificant. Type 3 analysis tells us that weight, horsepower, enginesize are significant, whereas height, price, and ndoors are insignificant. The parameter estimate table indicates that two doors is a constant term, so significant parameters would indicate significant differences from the baseline value. However, difference between two doors and four doors is insignificant. The parameter estimate of weight and horsepower are negative, so increases in weight and horsepower would cause the highyway mpg to decrease. The parameter estimate of enginesize is positive, so increase in enginesize would increase highyway mpg.

Model Information	
Data Set	WORK.AUTO
Distribution	Gamma
Link Function	Log
Dependent Variable	highwaympg

Criteria For Assessing Goodness Of Fit			
Criterion	**DF**	**Value**	**Value/DF**
Deviance	191	1.9477	0.0102
Scaled Deviance	191	197.3241	1.0331
Pearson Chi-Square	191	1.9619	0.0103
Scaled Pearson X2	191	198.7543	1.0406
Log Likelihood		-494.2423	
Full Log Likelihood		-494.2423	
AIC (smaller is better)		1002.4845	
AICC (smaller is better)		1003.0771	
BIC (smaller is better)		1025.4669	

Analysis Of Maximum Likelihood Parameter Estimates								
Parameter		**DF**	**Estimate**	**Standard Error**	**Wald 95% Confidence Limits**		**Wald Chi-Square**	**Pr > ChiSq**
Intercept		1	4.1951	0.1926	3.8175	4.5727	474.19	<.0001
weight		1	-0.0003	0.0000	-0.0003	-0.0002	69.10	<.0001
height		1	0.0011	0.0039	-0.0065	0.0088	0.08	0.7753
horsepower		1	-0.0034	0.0004	-0.0041	-0.0027	87.06	<.0001
enginesize		1	0.0014	0.0004	0.0006	0.0022	12.57	0.0004
ndoors	four	1	0.0076	0.0174	-0.0265	0.0418	0.19	0.6604
ndoors	two	0	0.0000	0.0000	0.0000	0.0000	.	.
Scale		1	101.3088	10.1910	83.1806	123.3878		

LR Statistics For Type 3 Analysis			
Source	**DF**	**Chi-Square**	**Pr > ChiSq**
weight	1	58.87	<.0001
height	1	0.08	0.7753
horsepower	1	71.03	<.0001
enginesize	1	12.22	0.0005
ndoors	1	0.19	0.6605

Model Information	
Data Set	WORK.AUTO
Distribution	Gamma
Link Function	Log
Dependent Variable	highwaympg

Criteria For Assessing Goodness Of Fit			
Criterion	**DF**	**Value**	**Value/DF**
Deviance	192	1.9486	0.0101
Scaled Deviance	192	197.3242	1.0277
Pearson Chi-Square	192	1.9628	0.0102
Scaled Pearson X2	192	198.7717	1.0353
Log Likelihood		-494.2830	
Full Log Likelihood		-494.2830	
AIC (smaller is better)		1000.5660	
AICC (smaller is better)		1001.0081	
BIC (smaller is better)		1020.2652	

Analysis Of Maximum Likelihood Parameter Estimates								
Parameter		**DF**	**Estimate**	**Standard Error**	**Wald 95% Confidence Limits**		**Wald Chi-Square**	**Pr > ChiSq**
Intercept		1	4.2490	0.0387	4.1732	4.3247	12081.5	<.0001
weight		1	-0.0003	0.0000	-0.0003	-0.0002	83.45	<.0001
horsepower		1	-0.0035	0.0004	-0.0041	-0.0028	96.66	<.0001
enginesize		1	0.0014	0.0004	0.0006	0.0022	12.55	0.0004
ndoors	four	1	0.0095	0.0162	-0.0223	0.0412	0.34	0.5593
ndoors	two	0	0.0000	0.0000	0.0000	0.0000	.	.
Scale		1	101.2670	10.1868	83.1463	123.3369		

LR Statistics For Type 3 Analysis			
Source	**DF**	**Chi-Square**	**Pr > ChiSq**
weight	1	68.80	<.0001
horsepower	1	77.72	<.0001
enginesize	1	12.20	0.0005
ndoors	1	0.34	0.5595

The GENMOD Procedure

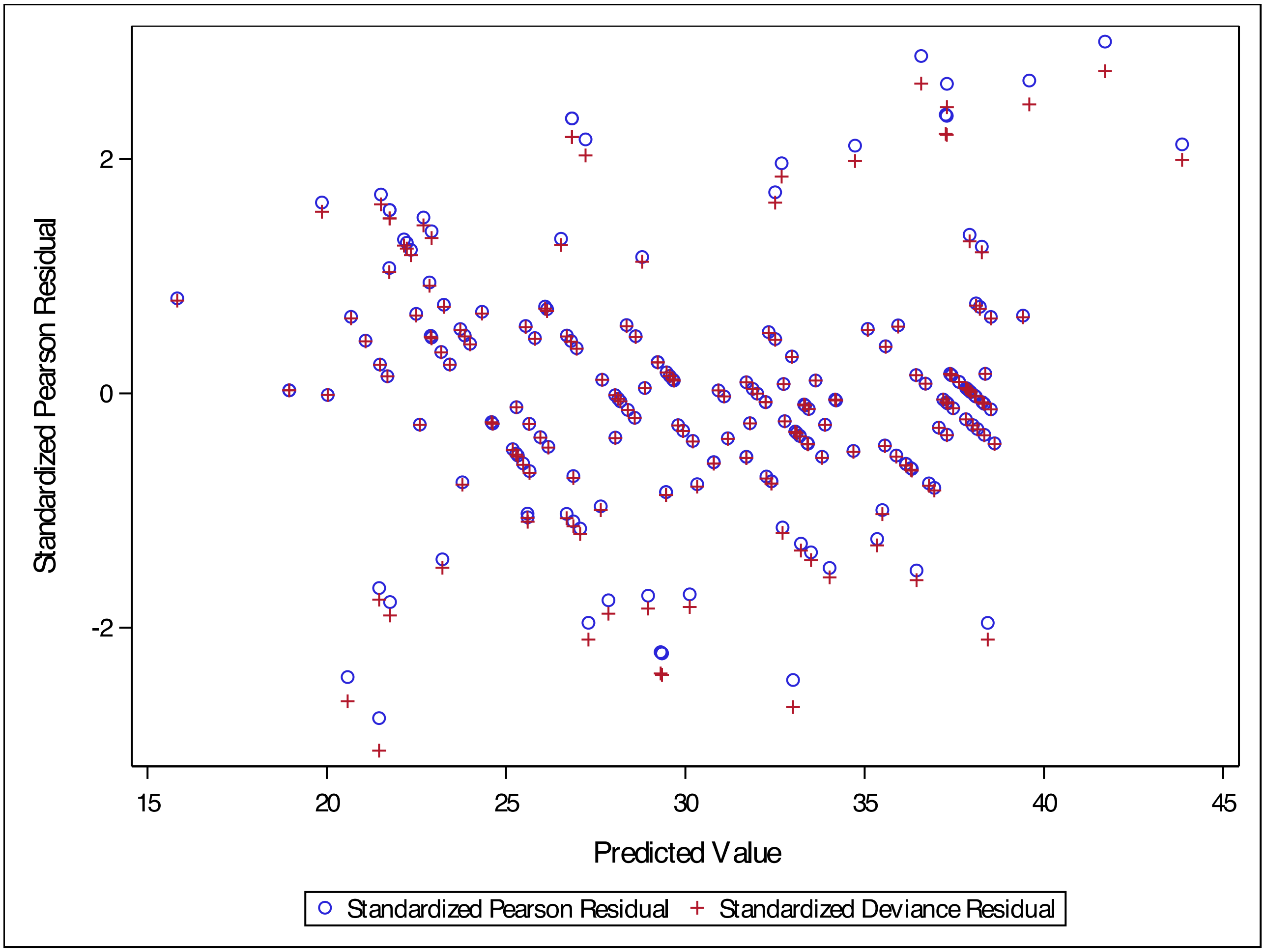
Model Information	
Data Set	WORK.AUTO
Distribution	Gamma
Link Function	Log
Dependent Variable	highwaympg

Criteria For Assessing Goodness Of Fit			
Criterion	**DF**	**Value**	**Value/DF**
Deviance	193	1.9519	0.0101
Scaled Deviance	193	197.3248	1.0224
Pearson Chi-Square	193	1.9632	0.0102
Scaled Pearson X2	193	198.4671	1.0283
Log Likelihood		-494.4533	
Full Log Likelihood		-494.4533	
AIC (smaller is better)		998.9066	
AICC (smaller is better)		999.2207	
BIC (smaller is better)		1015.3226	

Analysis Of Maximum Likelihood Parameter Estimates							
Parameter	**DF**	**Estimate**	**Standard Error**	**Wald 95% Confidence Limits**		**Wald Chi-Square**	**Pr > ChiSq**
Intercept	1	4.2460	0.0384	4.1708	4.3212	12253.3	<.0001
weight	1	-0.0003	0.0000	-0.0003	-0.0002	96.13	<.0001
horsepower	1	-0.0035	0.0003	-0.0042	-0.0029	110.51	<.0001
enginesize	1	0.0014	0.0004	0.0006	0.0021	12.22	0.0005
Scale	1	101.0927	10.1692	83.0032	123.1245		

LR Statistics For Type 3 Analysis			
Source	**DF**	**Chi-Square**	**Pr > ChiSq**
weight	1	77.33	<.0001
horsepower	1	86.33	<.0001
enginesize	1	11.89	0.0006

b) From the Type 3 analysis of a, price has the largest p-value of 0.7831, so we remove it from the full model in the first step. According to the Type 3 analysis after removing price, AIC decreases from 1004.4087 to 1002.4845, and predictor height has the largest p-value of 0.7753, so we remove it in the second step. After removing height, AIC decreases from 1002.4845 to 1000.5660, and the result of Type 3 analysis shows that only ndoors has p-value greater than 0.05, which is 0.5595, so we remove it in the third step. Now, there are three predictors left in the model: weight, horsepower, and enginesize, which have p-value less than 0.05. Also, AIC decreases from 1000.5660 to 998.9066. Therefore, we keep weight, horsepower, and enginesize in our final model.



c) The residual plot shows no reason for concern. The standardized Pearson and deviance residuals are pretty evenly distributed above and below 0, and are all pretty well bounded by -2 and 2. Looking at the plot versus predicted values, we don’t see problematic trends, so the assumptions are fine.

Exercise 2

Observations	197
Variables	7

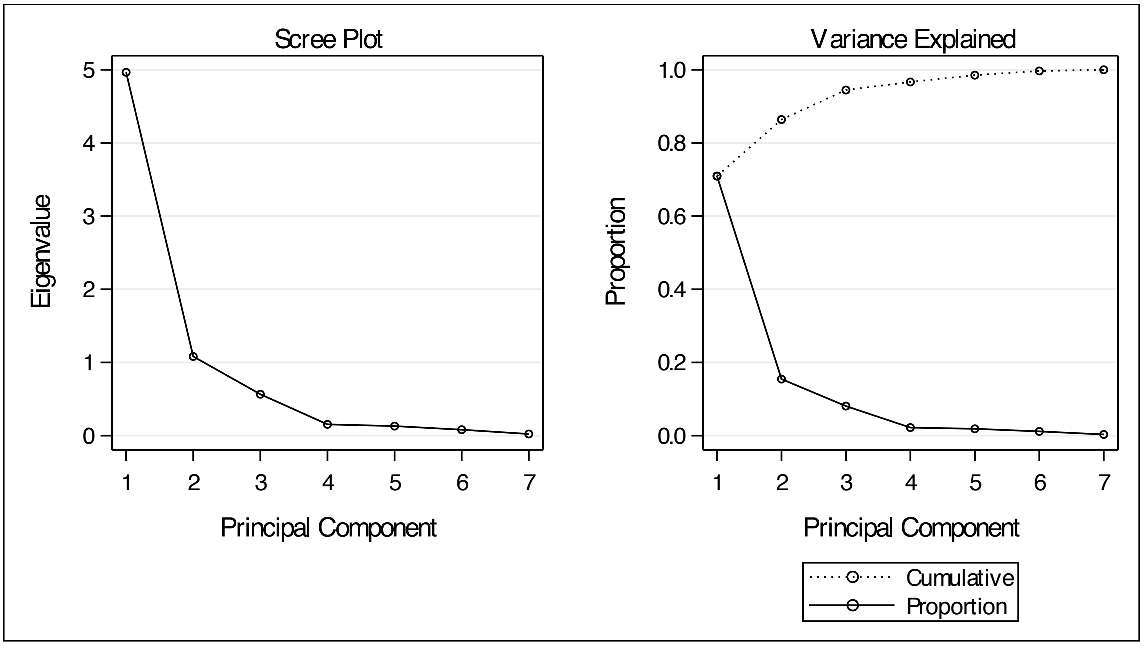
Simple Statistics							
	highwaympg	**citympg**	**weight**	**height**	**horsepower**	**enginesize**	**price**
Mean	30.62944162	25.15228426	2558.456853	53.78324873	103.6040609	126.9949239	13279.64467
StD	6.83625884	6.43786292	521.782047	2.44589903	37.6392053	41.9131144	8010.33422

Correlation Matrix							
	highwaympg	**citympg**	**weight**	**height**	**horsepower**	**enginesize**	**price**
highwaympg	1.0000	0.9724	-.8001	-.1131	-.8037	-.6847	-.7087
citympg	0.9724	1.0000	-.7556	-.0593	-.8220	-.6557	-.6929
weight	-.8001	-.7556	1.0000	0.3061	0.7599	0.8489	0.8347
height	-.1131	-.0593	0.3061	1.0000	-.0846	0.0719	0.1331
horsepower	-.8037	-.8220	0.7599	-.0846	1.0000	0.8253	0.8120
enginesize	-.6847	-.6557	0.8489	0.0719	0.8253	1.0000	0.8737
price	-.7087	-.6929	0.8347	0.1331	0.8120	0.8737	1.0000

Eigenvalues of the Correlation Matrix				
	Eigenvalue	**Difference**	**Proportion**	**Cumulative**
1	4.96574192	3.88401983	0.7094	0.7094
2	1.08172209	0.51633227	0.1545	0.8639
3	0.56538982	0.41160858	0.0808	0.9447
4	0.15378124	0.02348892	0.0220	0.9667
5	0.13029232	0.04957439	0.0186	0.9853
6	0.08071793	0.05836326	0.0115	0.9968
7	0.02235467		0.0032	1.0000

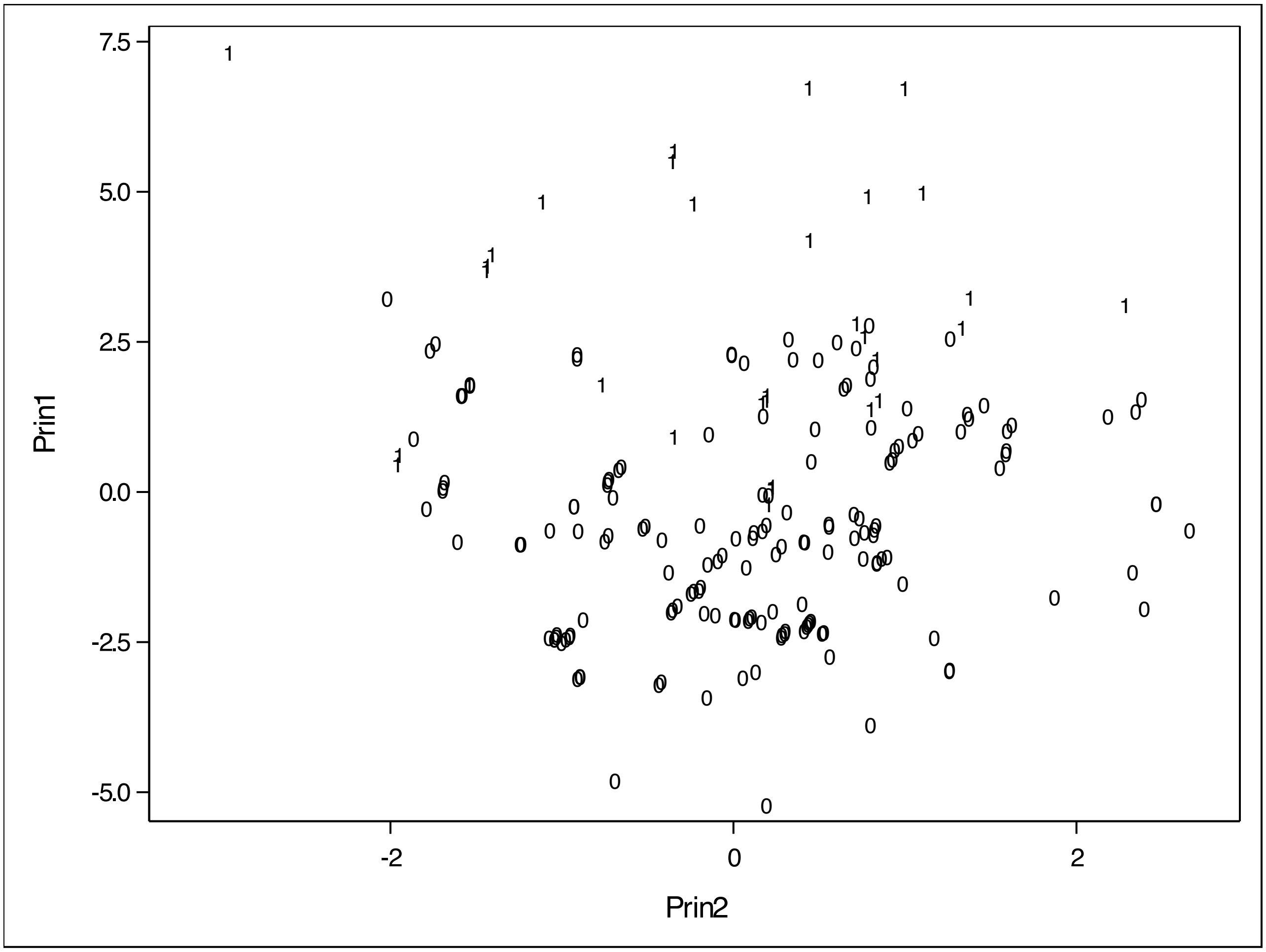
The PRINCOMP Procedure

Eigenvectors							
	Prin1	**Prin2**	**Prin3**	**Prin4**	**Prin5**	**Prin6**	**Prin7**
highwaympg	-.409235	0.028342	0.511068	0.150919	0.109474	0.218549	0.698576
citympg	-.402703	0.089059	0.550741	-.050223	0.090639	0.167909	-.698321
weight	0.414221	0.210014	0.130190	-.631395	0.023680	0.600966	0.083573
height	0.061796	0.943549	-.080867	0.266949	0.147097	-.080348	0.001495
horsepower	0.411111	-.235116	0.044947	0.574930	0.529655	0.387670	-.111001
enginesize	0.401916	-.029780	0.496168	-.227877	0.359968	-.636843	0.065722
price	0.405475	0.026612	0.404724	0.349726	-.739926	0.023848	-.026699



a) The first two principal components explain 86.39% of the total variation in the data, so two principal components should be kept in order to explain a minimum total variance of 85%. Based on the average eigenvalue test, two principal components should be kept because only the first two have eigenvalues greater than 1 (the average eigenvalue). Based on the scree plot, three principal components should be kept since after the third principal component the eigenvalues become relatively constant.

b) The large positive coefficient values of first principal component are weight, height, horsepower, enginesize, and price. The large negative coefficient values of first principal component are highwaympg and citympg. The positive values are car's characteristics, and the negative values are mileage variables. So PC1 is a contrast of car's characteristics and mileage variables. The large positive coefficient values of second principal component are weight and height, and the large negative coefficient value of second principal component is horsepower. So PC2 is a contrast of car's physical feature and car's power.



c) Type 1 has positive values for principal component 1 and 2, which indicates luxury cars have larger than average value of mileage variables compared to car's characteristics and have larger than average value of physical features compared to power. Since type 0 has negative values for principal component 1, this implies non-luxury cars have lower than average value of mileage variables compared to car's characteristics. Type 0 has positive values for principal component 2, this implies non-luxury cars have larger than average value of physical features compared to power.

Exercise 3

The PRINCOMP Procedure

Observations	197
Variables	7

Simple Statistics

	highwaympg	citympg	weight	height	horsepower	enginesize	price
Mean	30.62944162	25.15228426	2558.456853	53.78324873	103.6040609	126.9949239	13279.64467
StD	6.83625884	6.43786292	521.782047	2.44589903	37.6392053	41.9131144	8010.33422

Covariance Matrix

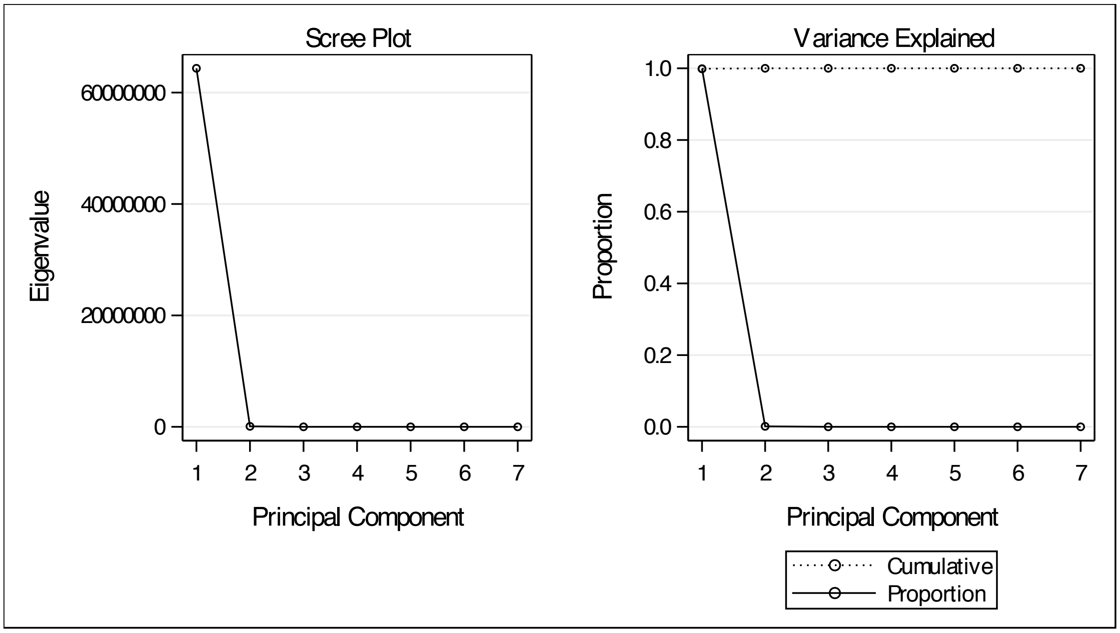
	highwaympg	citympg	weight	height	horsepower	enginesize	price
highwaympg	46.73	42.80	-2854.03	-1.89	-206.79	-196.18	-38806.69
citympg	42.80	41.45	-2538.04	-0.93	-199.17	-176.94	-35734.94
weight	-2854.03	-2538.04	272256.50	390.72	14924.52	18565.73	3488885.73
height	-1.89	-0.93	390.72	5.98	-7.79	7.37	2607.82
horsepower	-206.79	-199.17	14924.52	-7.79	1416.71	1301.95	244806.04
enginesize	-196.18	-176.94	18565.73	7.37	1301.95	1756.71	293336.94
price	-38806.69	-35734.94	3488885.73	2607.82	244806.04	293336.94	64165454.29

Total Variance	64440978.378

Eigenvalues of the Covariance Matrix

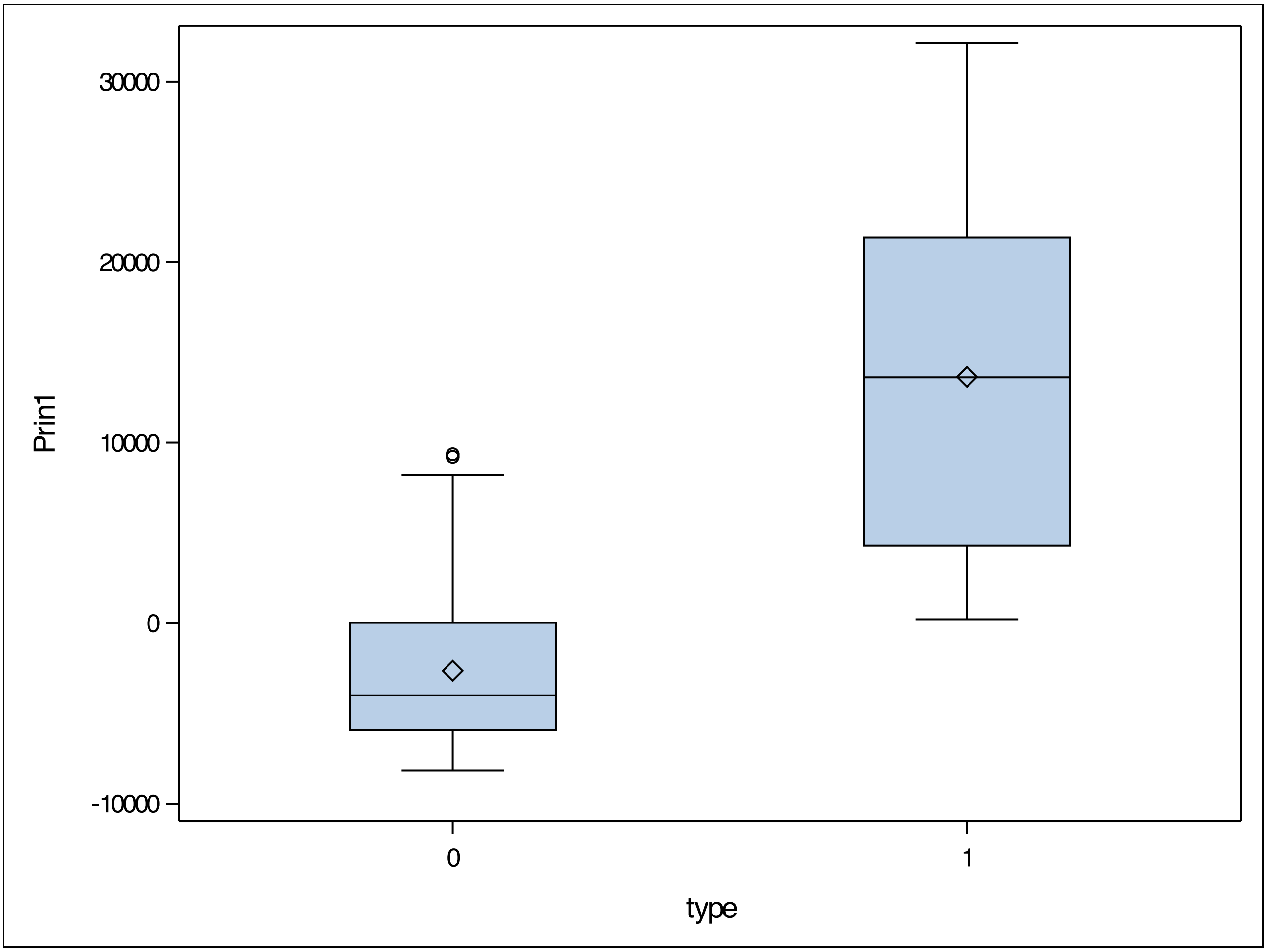
	Eigenvalue	Difference	Proportion	Cumulative
1	64357720.6	64275285.1	0.9987	0.9987
2	82435.5	81893.0	0.0013	1.0000
3	542.5	287.1	0.0000	1.0000
4	255.4	236.0	0.0000	1.0000
5	19.4	15.5	0.0000	1.0000
6	3.9	2.9	0.0000	1.0000
7	1.0		0.0000	1.0000

Eigenvectors							
	Prin1	**Prin2**	**Prin3**	**Prin4**	**Prin5**	**Prin6**	**Prin7**
highwaympg	-.000605	-.008997	-.068205	0.119471	0.714344	-.020062	-.685787
citympg	-.000557	-.007196	-.078850	0.135288	0.668186	0.009924	0.727225
weight	0.054362	0.997748	-.034375	-.014904	0.010815	-.004620	-.000915
height	0.000041	0.003005	-.048282	0.005741	0.001811	0.998578	-.021564
horsepower	0.003811	0.019518	0.842386	-.504439	0.182596	0.043608	0.017007
enginesize	0.004567	0.031619	0.525347	0.844221	-.098886	0.020433	-.009197
price	0.998503	-.054550	-.003830	-.000977	-.000029	-.000056	0.000018



a) The first principal components explain 99.87% of the total variation in the data, so one principal components should be kept in order to explain a minimum total variance of 85%. Based on the average eigenvalue test, one principal components should be kept because only the first has eigenvalues greater than the average eigenvalue. Based on the scree plot, one principal components should be kept since after the first principal component the eigenvalues become relatively constant.

b) The large positive coefficient value of first principal component is price. The negative coefficient values are tiny, so we can ignore it. Therefore, PC1 represents price of cars. Price is the most predominant feature in principal component 1. Since price decides the car's type (luxury or non-luxury), so covariance-based PCA choose out this predominant feature.



c) Type 0 has negative values for principal component 1, which indicates non-luxury cars have lower than average price of cars. Type 1 has positive values for principal component 1, which indicates luxury cars have larger than average price of cars.

d) Since a correlation matrix is a covariance matrix of the standardized data, covariance-based PCA is more sensitive to the magnitude of variables' variances.