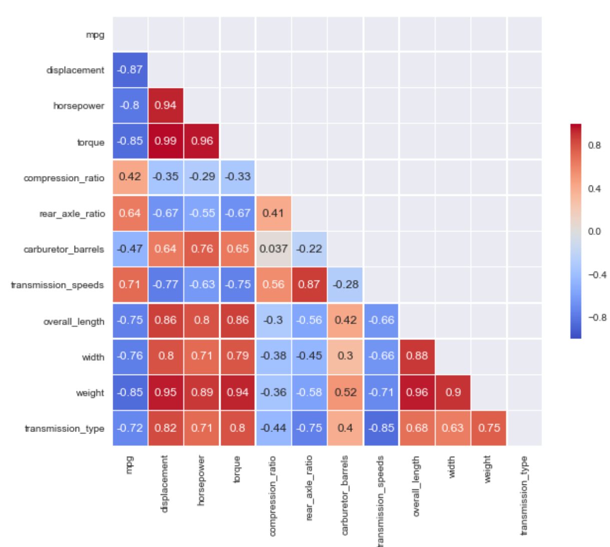
Introduction

For this assignment, we're revisiting the Chatterjee dataset used in Assignment 5. We will summarize our findings from the previous exercise, and add new material focusing on the use of Principal Component Analysis (PCA) and Principal Component Regression (PCR) as ways to provide new insights into the data. We will use PCA to identify principal components (PCs) that account for 50, 70 and 90% of the variation in the predictors. We will then use the identified PCs in ordinary least squares regression models. The resultant models will be compared to each other and to the reduced-variable model from assignment 5.

Review of Sample Data and Exploratory Data Analysis

The data set is composed of 30 observations of 12 total variables. All observations were retained. The miles-per-gallon (MPG) variable is used as the response we are trying to predict. In Assignment 5 we saw that the data are not normally distributed over the response variable. A scatterplot matrix of the variables revealed a number of predictor variables showing strong linear relationships to each other. There were both positively and negatively related predictor variables according to the scatterplots. A correlation matrix of the data, including the correlation coefficients, confirm the strong collinearity in the data. There are a number of pairs with correlation coefficients of >= 90%

Figure 1



Summary of Multiple Linear Regression Work – Full and Subset Models

The Full multiple regression model from Assignment 5 has an R^2 value of .835, which means, 83% of the variation of the response variable is accounted for by the model. There is difference of 10% between the R^2 and adjusted-R^2 values, which indicates the possible presence of variables which are not contributing much to the model. The QQ plot and Residual versus Predicted plots for this model were within an acceptable range of normalcy.

For the subset model, I had landed on a model with 5 predictor variables. The summary for which is shown below:

Table 1

OLS Regression Results

Dep. Variable:	mpg	R-squared:	0.787
Model:	OLS	Adj. R-squared:	0.743
Method:	Least Squares	F-statistic:	17.74
Date:	Thu, 20 Jul 2017	Prob (F-statistic):	2.27e-07
Time:	16:00:06	Log-Likelihood:	-73.899
No. Observations:	30	AIC:	159.8
Df Residuals:	24	BIC:	168.2
Df Model:	5		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	22.0313	25.656	0.859	0.399	-30.920	74.982
rear_axle_ratio	1.0229	1.634	0.626	0.537	-2.351	4.396
width	-0.1857	0.267	-0.694	0.494	-0.738	0.366
displacement	-0.0341	0.021	-1.631	0.116	-0.077	0.009
weight	-9.607e-05	0.003	-0.030	0.976	-0.007	0.007
compression_ratio	2.1929	2.419	0.907	0.374	-2.799	7.185

Omnibus:	1.451	Durbin-Watson:	2.105
Prob(Omnibus):	0.484	Jarque-Bera (JB):	0.985
Skew:	0.442	Prob(JB):	0.611
Kurtosis:	2.931	Cond. No.	1.67e+05

It was pointed out to me in the feedback for Assignment 5, that my model had high p-values for the predictors. Rather than use that model, I re-visited the assignment, and found that models with 2 or more predictors have high p-values for 1 or more of the predictors. All the revised models I tried showed one variable with a significant p-value and the other variables had p-values that were high. I found both Ridge, and Lasso selected the same, single predictor as significant in a 1- or 2-variable model. So, working with single, significant predictor, I produced a new model, its' fitted values and residual plots are shown below.

Table 2 – Revised Assignment 5 Reduced Predictor Model

OLS Regression Results

Dep. Variable:	mpg	R-squared:	0.760
Model:	OLS	Adj. R-squared:	0.751
Method:	Least Squares	F-statistic:	88.70
Date:	Sat, 29 Jul 2017	Prob (F-statistic):	3.55e-10
Time:	13:43:30	Log-Likelihood:	-75.687
No. Observations:	30	AIC:	155.4
Df Residuals:	28	BIC:	158.2
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	33.4878	1.537	21.786	0.000	30.339	36.636
displacement	-0.0471	0.005	-9.418	0.000	-0.057	-0.037

Omnibus:	0.664	Durbin-Watson:	1.702
Prob(Omnibus):	0.717	Jarque-Bera (JB):	0.438
Skew:	0.288	Prob(JB):	0.803
Kurtosis:	2.867	Cond. No.	830.

Figure 2

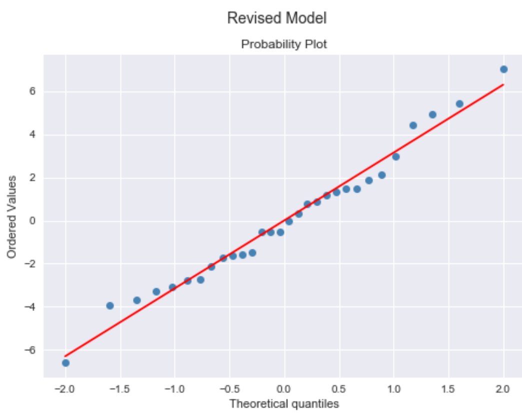
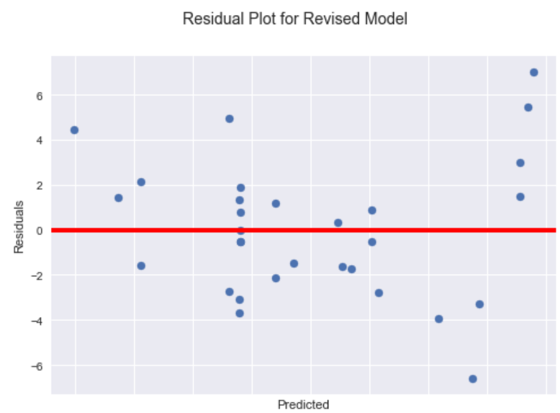


Figure 3



The revised model in Table 2 has a better adjusted-R^2 than my original assignment 5 model, and the p-value is significant; my p-values for the . The plots indicate a model which is somewhat close to normal, and are comparable to the model I used in assignment 5. For purpose of this discussion, I will move forward using the revised model as my assignment 5 baseline for discussions and comparisons.

Principal Components Analysis

I ran a Principal Component Analysis (PCA) using all 11 predictor variables and the response variable given in the original case. We run PCAs using computer techniques, but the process is based on linear algebra.

1. To compute the PCA, we start by standardizing the data such that each variable *X* has a 0 mean and the variance for *X* is 1. You do this by subtracting the sample mean for each variable from each observation of that variable, and divide each observation by the sample standard deviation. For a matrix, you would iterate over each column and perform the transformation.
2. Next, we compute the covariance matrix for the sample observations; since we standardized the data, the covariance matrix is the same as the correlation matrix.

   For purposes of illustration, let's confine this example to the covariance between just 2 of our variables, weight and width. The sample covariance for weight & width would be

$$\sum_{i=1}^{n} ((weight_i - sample\ mean\ of\ weight)(width_i - sample\ mean\ width) \Big/ n-1$$

   A matrix can be computed for the pairs of variables taken together, for our full 12 variable set, we get a 12 x 12 sample covariance matrix. Note: the major diagonal of the matrix (variable *X* is the row and the column) gives the variance of that specific variable.

3. We find the eigenvalues and eigenvectors. The formula for doing that is:

$$(A - \lambda I)x = 0$$

   A is the matrix of the observations of the variables, *I* is the identity matrix (1's on the major diagonal, 0 elsewhere) and λ is what you are solving for. When you have the values for λ, you have found the eigenvalues, you use them to plug into the following formula:

$$A * vector_i = \lambda_i * vector_i$$

$$(A - \lambda_i) * vector_i = 0$$

to find each of the eigenvectors. The coefficients for the ith Principal Component are the elements of the ith eigenvector. The ith λ value is the variance for the Principal Component.

In my calculations, I found the following set of predictor variances for our 11 Principal Components:

Table 3

```
PC 1 accounts for 70% of variation; cummulative variation is: 70%
PC 2 accounts for 12.8% of variation; cummulative variation is: 82.8%
PC 3 accounts for 7% of variation; cummulative variation is: 89.8%
PC 4 accounts for 5.2% of variation; cummulative variation is: 95%
PC 5 accounts for 1.9% of variation; cummulative variation is: 96.9%
PC 6 accounts for 1.3% of variation; cummulative variation is: 98.2%
PC 7 accounts for 0.9% of variation; cummulative variation is: 99.1%
PC 8 accounts for 0.5% of variation; cummulative variation is: 99.6%
PC 9 accounts for 0.3% of variation; cummulative variation is: 99.9%
PC 10 accounts for 0.1% of variation; cummulative variation is: 100%
PC 11 accounts for 0% of variation; cummulative variation is: 100%
```

The first Principal Component 1 (PC1) accounts for 70% of the variance right out of the gate. If we allow rounding, we can reach 90% of the variance with PC1, PC2 and PC3.

Principal Components Regression

Before performing PCR, I took at quick look at the correlations between the first 3 PCs to verify they are unrelated.

Figure 4

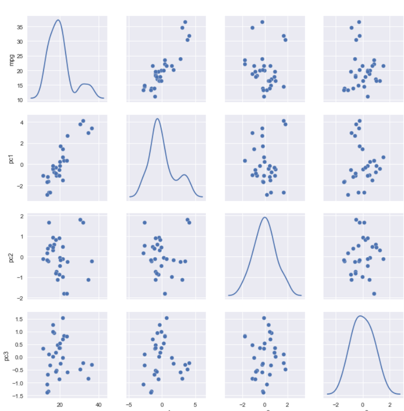


Table 4

	pc1	pc2	pc3
pc1	1.000000e+00	-8.287267e-17	-5.730827e-17
pc2	-8.287267e-17	1.000000e+00	2.608069e-16
pc3	-5.730827e-17	2.608069e-16	1.000000e+00

The scatter plot shows that there are no observable patterns of relationship between the PCs, the printed correlations values confirm that interpretation of the graphs.

Running an OLS regression on PC1 (the PC which accounts for 70% of variance) yielded:

Table 5 – 70% Variance

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                    mpg   R-squared:                       0.770
Model:                            OLS   Adj. R-squared:                  0.762
Method:                 Least Squares   F-statistic:                     93.61
Date:                Sat, 29 Jul 2017   Prob (F-statistic):           1.99e-10
Time:                        13:14:49   Log-Likelihood:                -75.069
No. Observations:                  30   AIC:                             154.1
Df Residuals:                      28   BIC:                             156.9
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept     20.0433      0.558     35.896      0.000      18.900      21.187
pc1            2.9143      0.301      9.675      0.000       2.297       3.531
==============================================================================
Omnibus:                        0.881   Durbin-Watson:                   2.291
Prob(Omnibus):                  0.644   Jarque-Bera (JB):                0.685
Skew:                           0.355   Prob(JB):                        0.710
Kurtosis:                       2.792   Cond. No.                         1.85
==============================================================================
```

Figure 5

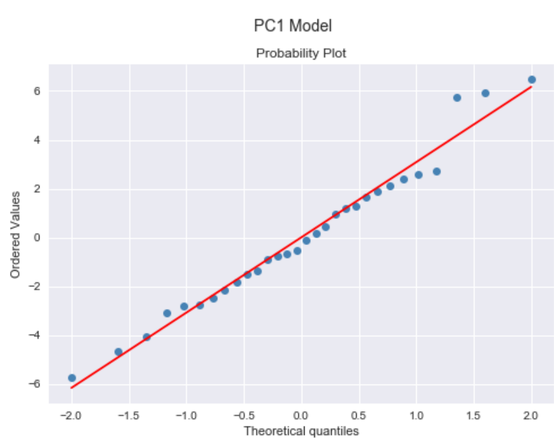


Figure 6

The residuals form a pretty good looking QQ-plot and the plot of residuals has a nice, random looking distribution. Both plots are indicative of a good fit.

To get to 90% of variance, I ran an OLS regression using PC1+PC2+PC3, which produced:

Table 6 – 90% Variance

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                    mpg   R-squared:                       0.774
Model:                            OLS   Adj. R-squared:                  0.748
Method:                 Least Squares   F-statistic:                     29.72
Date:                Sat, 29 Jul 2017   Prob (F-statistic):           1.47e-08
Time:                        13:14:49   Log-Likelihood:                -74.775
No. Observations:                  30   AIC:                             157.5
Df Residuals:                      26   BIC:                             163.2
Df Model:                           3
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept     20.0433      0.574     34.932      0.000      18.864      21.223
pc1            2.9143      0.310      9.415      0.000       2.278       3.551
pc2           -0.4382      0.631     -0.695      0.494      -1.735       0.859
pc3           -0.1390      0.766     -0.181      0.857      -1.714       1.436
==============================================================================
Omnibus:                        0.737   Durbin-Watson:                   2.181
Prob(Omnibus):                  0.692   Jarque-Bera (JB):                0.533
Skew:                           0.315   Prob(JB):                        0.766
Kurtosis:                       2.830   Cond. No.                         2.48
==============================================================================
```

Figure 7

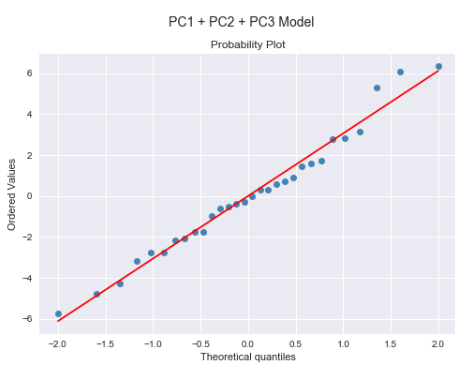
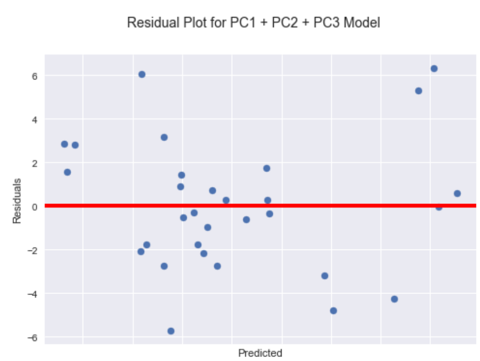


Figure 8



As with the first plot, the views of the residuals show the model fits well.

Model Comparison and Recommendation

Comparing the 2 PCR models first, we can see that the model using PC1-only performs better that the one using 3 PCs. The first model has a better adjusted-R^2, a lower Condition Number, and the single p-value is significant. In the second model, we can see that the p-values for PC2and PC3 are quite high. Of the two models produced through PCR, the first is the better model.

Looking at the scree plot for the PCs, we see that there is a sharp drop after the first PC.

If we use the rule of finding the scree plot “elbow” and using that to select our PCs, using only PC1 is the correct decision.

However, if we look at the eigenvalues for the set of PCs, we see:

Table 7 - Eigenvalues

```
PC 1 has an eigen value of 7.703
PC 2 has an eigen value of 1.403
PC 3 has an eigen value of 0.773
PC 4 has an eigen value of 0.577
PC 5 has an eigen value of 0.211
PC 6 has an eigen value of 0.142
PC 7 has an eigen value of 0.095
PC 8 has an eigen value of 0.05
PC 9 has an eigen value of 0.033
PC 10 has an eigen value of 0.008
PC 11 has an eigen value of 0.003
```

Kaiser’s rule however, would have us use PC1 and PC2. The fitted model for PC selection via Kaiser’s Rule looks like:

Table 8 – Using Kaiser's Rule for PC Selection

OLS Regression Results

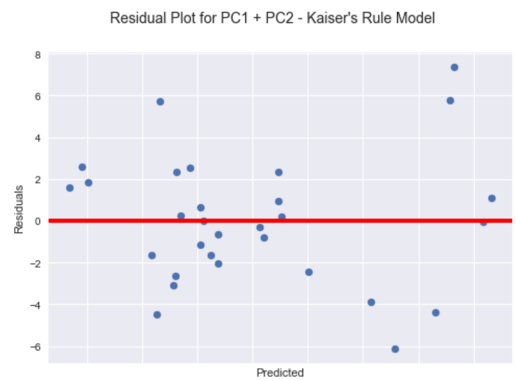
Dep. Variable:	mpg	R-squared:	0.755
Model:	OLS	Adj. R-squared:	0.736
Method:	Least Squares	F-statistic:	41.49
Date:	Sun, 30 Jul 2017	Prob (F-statistic):	5.83e-09
Time:	15:52:13	Log-Likelihood:	-76.031
No. Observations:	30	AIC:	158.1
Df Residuals:	27	BIC:	162.3
Df Model:	2		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	20.0433	0.587	34.137	0.000	18.839	21.248
pc1	1.9270	0.212	9.109	0.000	1.493	2.361
pc2	0.0492	0.496	0.099	0.922	-0.968	1.066

Omnibus:	1.211	Durbin-Watson:	2.021
Prob(Omnibus):	0.546	Jarque-Bera (JB):	0.661
Skew:	0.363	Prob(JB):	0.719
Kurtosis:	3.051	Cond. No.	2.78

Figure 9

Figure 10



The selection made by Kaiser's Rule, in this case, does not offer an improvement over the visual selection by scree plot. The residual plots are comparable for the two models, and the PC1 model has a better adjusted-R^2 than does the PC1+PC2 model. We also see in the PC1+PC2 model, the p-value for PC2 is .922, which is very high indeed.

So, using Kaiser's rule for the selection of the number of Principal Components did not improve over the first model. Of the PCR models, the PC1-only model is still the preferred one.

Comparing the reduced-variable model (Figures 2 & 3) from assignment 5 to the PC1 model (Figures 5 & 6), we can see that the QQ plot and distribution of residuals are slightly better for the PC1 model. Additionally, it has a higher adjusted-R^2 value.

Table 9

Figure 2

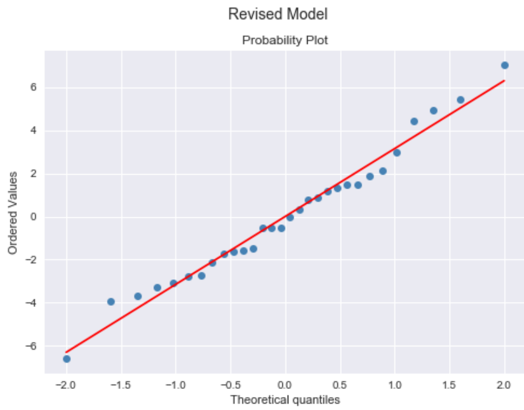


Figure 3

Figure 5

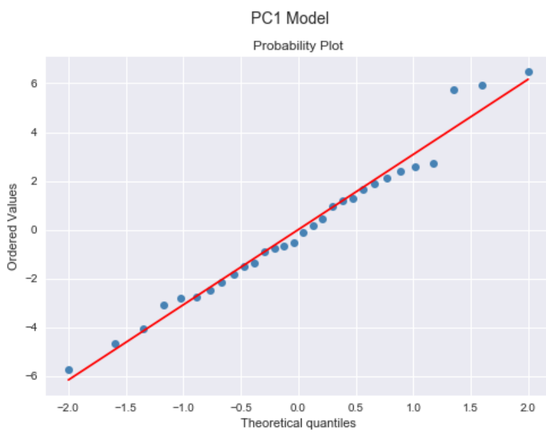
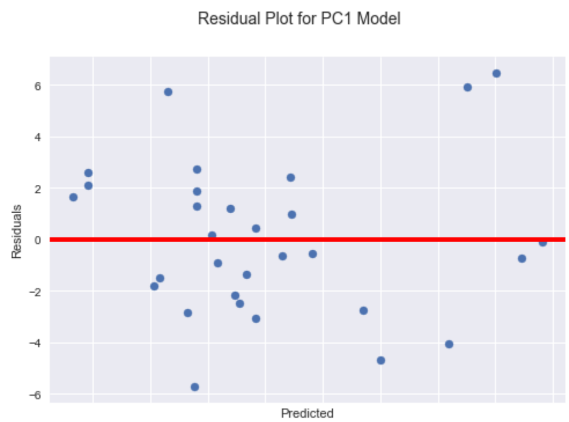


Figure 6



The PC1 model hits a good balance between accounting for the variation in the response variable (adjusted-R^2 of .762) and in the predictor-variable explained variance of 70%. Given this choice of models, I would recommend moving forward with PC1. Management can use the model with data that fall within range of values the original observations had. It can be used to infer miles-per-gallon, based on the 11 predictor variables.

Conclusion

Overall, the OLS model using a single Principal Component out-performed my best predictor-variable OLS models. The math behind calculating a PC is daunting to anyone unfamiliar with these sorts of calculations, but first and second Principal Components can be explained visually (to a degree) making them a viable option in a workplace setting. The use of PCA to select a subset of PCs to work with has a clear potential for reducing huge datasets down to something smaller, which can then be regressed via standard ordinary least squares.

In this data set, there is a high degree of collinearity. The use of PCA allows us to formulate models which address the instability and wrong inferences that can happen working with the collinear data directly.