HW#2 - MATH 585

Nate Islip

Due Date: 01/14/2021 @ 11:59 PM

Note: Worked with Jasen

Chapter 7

Problem 7.9:

Consider the following data on one predictor variable z_1 and two responses Y_1, and Y_2: Determine the least squares estimates of the parameters in the bi variate straight-line regression model

$$Y_{j1} = \beta_{01} + \beta_{11} z_{j1} + \epsilon_{j1}$$
$$Y_{j1} = \beta_{02} + \beta_{12} z_{j1} + \epsilon_{j2}$$

Also, calculate the matrices of fitted values $\hat{\mathbf{Y}}$ and residuals $\hat{\epsilon}$ with $\mathbf{Y} = [y_1|y_2]$. Verify the sum of squares and cross-products decomposition

$$\mathbf{Y}'\mathbf{Y} = \hat{\mathbf{Y}}'\hat{\mathbf{Y}} + \hat{\epsilon}'\hat{\epsilon}$$

To determine the least squares estimates of the parameters

$$\mathbf{Z}'\mathbf{y}_{(2)} = \begin{bmatrix} 15 \\ -9 \end{bmatrix} \qquad \mathbf{Z}'\mathbf{y}_{(2)} = \begin{bmatrix} 0 \\ 15 \end{bmatrix}$$
$$\hat{\beta}_{(1)} = \begin{bmatrix} 3.0 \\ -0.9 \end{bmatrix} \qquad \hat{\beta}_{(2)} = \begin{bmatrix} 0.0 \\ 1.5 \end{bmatrix}$$

$$\hat{\beta} = \begin{bmatrix} 3.0 & 0.0 \\ -0.9 & 1.5 \end{bmatrix} \qquad \hat{\mathbf{Y}} = \begin{bmatrix} 4.8 & -3.0 \\ 3.9 & -1.5 \\ 3.0 & 0.0 \\ 2.1 & 1.5 \\ 1.2 & 3.0 \end{bmatrix} \qquad \hat{\epsilon} = \begin{bmatrix} 0.2 & 0.0 \\ -0.9 & 0.5 \\ 1.0 & -1.0 \\ -0.1 & 0.5 \\ -0.2 & 0.0 \end{bmatrix}$$

$$\mathbf{Y}'\mathbf{Y} = \begin{bmatrix} 55 & -15 \\ -15 & 24 \end{bmatrix} \qquad \hat{\mathbf{Y}}'\hat{\mathbf{Y}} = \begin{bmatrix} 53.1 & -13.5 \\ -13.5 & 22.5 \end{bmatrix} \qquad \hat{\epsilon}'\hat{\epsilon} = \begin{bmatrix} 1.9 & -1.5 \\ -1.5 & 1.5 \end{bmatrix}$$

Problem 7.15:

Use the real-estate data in Table 7.1 and the linear regression model in Example 7.4. Data used for these results is presented in the Data Appendix at the end of the document.

a) Verify the results in Example 7.4. First we generate the design matrix with the variables z_1, z_2 using `Z <- matrix(c(rep(1,n), df$z1, df$z2), ncol = 3)` where *df* is our data frame. Calculating the least squares estimator, $\hat{\beta} = (\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{y}$ using `ZZinv <- inv(t(Z) \%*\% Z)` and `Beta <- ZZinv \%*\% t(Z) \%*\% Y` we get,

$$(\mathbf{Z}'\mathbf{Z})^{-1} = \begin{bmatrix} 5.152 & 0.254 & -0.146 \\ 0.254 & 0.051 & -0.017 \\ -0.146 & -0.017 & 0.007 \end{bmatrix} \qquad \hat{\beta} = (\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{y} = \begin{bmatrix} 30.96653012 \\ 2.63439114 \\ 0.04543286 \end{bmatrix}$$

and therefore the fitted equation is as follows,

$$\hat{y} = 30.96653012 + 2.63439114z_1 + 0.04543286z_2$$

Using the R command `linear_model <- lm(df$Y ~ df$z1 + df\$z2, data=df)` and summarizing the model with `summary(linear_model)` we produce the following output,

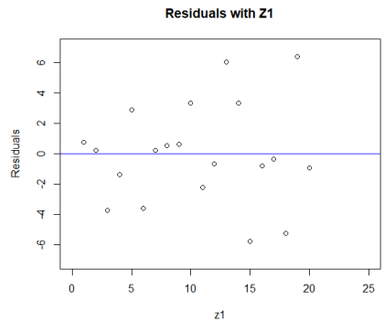
Table 1

	Dependent variable:
	y
z1	2.634***
	(0.786)
z2	0.045
	(0.285)
Constant	30.967***
	(7.882)
Observations	20
R^2	0.834
Adjusted R^2	0.815
Residual Std. Error	3.473 (df = 17)
F Statistic	42.828*** (df = 2; 17)
Note:	*p<0.1; **p<0.05; ***p<0.01

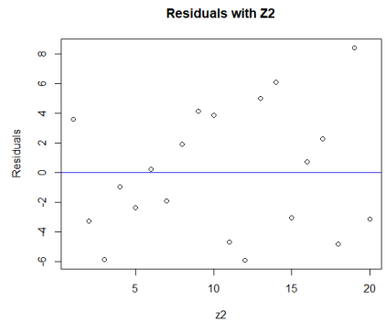
thus, the results in 7.4 are verified.

b) Analyze the residuals to check the adequacy of the model (see section 7.6).

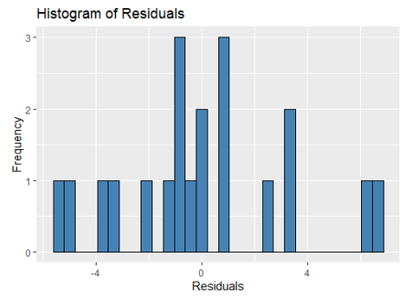
According to the textbook, there are multiple ways to detect/analyze possible anomalies. For instance, below are graphs of the predictor variables against their residuals.



(a) Residuals versus z_1



(b) residuals versus z_2



(c) residuals versus z_2

Figure 1: Histograms with the linear model containing z_1 and z_2

c) Generate a 95% prediction interval for the selling price (Y_0) corresponding to total dwelling size $z_1 = 17$ and assessed value $z_2 = 46$. (Seems mini tab was the quickest way to calculate this so I used a free trial)

The 95% prediction interval for Y_0 is $(64.12, 91.54)$

d) Carry out a likelihood ratio test of $H_0 : \beta_2 = 0$ with a significance level of $\alpha = 0.05$. Should the original model be modified? Discuss.

Since the t value for z_2 is 0.158 and has a p-value of 0.87598, the 5% level of significance will be accepted. Therefore, z_2 is not need in the fitted model.

$$\hat{\beta}_2 \pm t_{17}(0.05/2)\sqrt{\hat{\text{VAR}}(\hat{\beta}_2)} = 0.45 \pm 2.110(2.85)$$

Chapter 8

Problem 8.2:

$$\mathbf{\Sigma} = \begin{bmatrix} 5 & 2 \\ 2 & 2 \end{bmatrix}$$

Convert the covariance matrix to a correlation matrix ρ. Using the R command `Rho <- cov2cor(E)` where $E = \mathbf{\Sigma}$.

$$\rho = \begin{bmatrix} 1.0 & 0.6324555 \\ 0.6324555 & 1.0 \end{bmatrix}$$

(a) Determine the principal components Y_1 and Y_2 from ρ and compute the proportion of total population variance explained by Y_1.

Computing the eigenvalue-eigenvector pairs using `eigen(Rho)`

$$\lambda_1 = 1.6324555 \qquad \mathbf{e}_1 = \begin{bmatrix} 0.7071068 & -0.7071068 \end{bmatrix}$$
$$\lambda_2 = 0.3675445 \qquad \mathbf{e}_2 = \begin{bmatrix} 0.7071068 & 0.7071068 \end{bmatrix}$$

Therefore the principal components from ρ are,

$$Y_1 = \mathbf{e}_1'\mathbf{X} = 0.7071068\mathbf{X}_1 - 7071068\mathbf{X}_2$$
$$Y_2 = \mathbf{e}_2'\mathbf{X} = 0.7071068\mathbf{X}_1 + 0.7071068\mathbf{X}_2$$

After calculating the total population variance explained by Y_1,

$$\frac{\lambda_1}{\lambda_1 + \lambda_2} = \frac{1.6324555}{1.6324555 + 0.3675445} = 0.8162275 = 81.62\%$$

is 81.62 %.

(b) Compare the components calculated in Part a with those obtained in Exercise 8.1. Are they the same? Should they be?

The results obtained in 8.1 are not the same as the results obtained in part (a) because the variances in part a are not the same as the variances in the exercise. Therefore, the principal components will not be the same.

(c) compute the correlations $\rho_{Y_1}, z_1, \rho_{Y_1}, z_2$ and ρ_{Y_2}, z_1

Using (Result 8.4) we can compute the correlation coefficients between the components $\mathbf{Y}_i$ and $\mathbf{X}_k$,

$$\rho_{Y_1,z_1} = e_{11}\sqrt{\lambda_1} = 0.7071068 \cdot \sqrt{1.6324555} \approx 0.9034532602$$
$$\rho_{Y_1,z_1} = e_{21}\sqrt{\lambda_1} = 0.7071068 \cdot \sqrt{1.6324555} \approx 0.9034532602$$
$$\rho_{Y_2,z_1} = e_{11}\sqrt{\lambda_1} = 0.7071068 \cdot \sqrt{0.3675445} \approx 0.4286866685$$

Problem 8.6:

Data on sales and profits for the 10 largest companies in the worlds were listed in Exercise 1.4 of Chapter 1.

$$\bar{\mathbf{x}} = \begin{bmatrix} 155.60 \\ 14.70 \end{bmatrix} \qquad \mathbf{S} = \begin{bmatrix} 7476.45 & 303.62 \\ 303.62 & 26.19 \end{bmatrix}$$

a) Determine the sample principal components and their variances for these data. (You may need the quadratic formula to solve for the eigenvalues of $\mathbf{S}$.)

Using `eigen(S)` to compute the eigenvalues and eigenvectors of the matrix $\mathbf{S}$, `S <- matrix(c(7476.45, 303.62, 303.62, 26.19), nrow = 2, ncol=2 )`, we compute,

$$\hat{\lambda}_1 = 7488.80293 \qquad \hat{\lambda}_2 = 13.83707$$
$$\hat{\mathbf{e}}_1 = \begin{bmatrix} -0.99917337 & 0.04065185 \end{bmatrix}^T \qquad \hat{\mathbf{e}}_2 = \begin{bmatrix} -0.04065185 & -0.99917337 \end{bmatrix}^T$$

for the ith component, the sample principal components are,

$$\hat{y}_1 = -0.99917337(x_1 - 155.60) + 0.04065185(x_2 - 14.70)$$
$$\hat{y}_2 = -0.04065185(x_1 - 155.60) - 0.99917337(x_2 - 14.70)$$

and the Sample variance $(\hat{y}_k) = \hat{\lambda}_k$ for $k = 1, 2, \ldots, p$ (8-20) is given by

$$\hat{y}_1 = 7488.80293$$
$$\hat{y}_2 = 13.83707$$

b) Find the proportion of the total sample variance explained by $\hat{y}_1$.

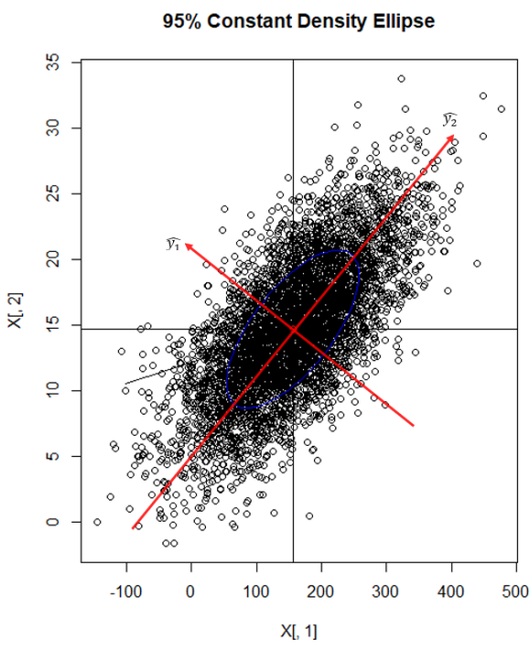
Using equation (8-7) to calculate the proportion of total population variance due to *kth* principal component,

$$\frac{\hat{\lambda}_1}{\hat{\lambda}_1 + \hat{\lambda}_2} = \frac{7488.80293}{7488.80293 + 13.83707} = 0.9981557 = 99.82\%$$

c) Sketch the constant density ellipse $(\mathbf{x} - \bar{\mathbf{x}})'\mathbf{S}^{-1}(\mathbf{x} - \bar{\mathbf{x}}) = 1.4$, and indicate the principal components $\hat{y}_1$ and $\hat{y}_2$ on your graph.

The center of the constant density ellipse will be at (x_1, x_2) and the axes will be defined by $\hat{y}_1$ and $\hat{y}_2$.

Figure 2: Scree Plot of PC for 5 variables



d) Compute the correlation coefficients $r_{\hat{y}_1, x_k}, i, k = 1, 2$. What interpretation, if any, can you give to the first principal component?

Use the formula,

$$r_{\hat{y}_1, x_k} = \frac{\hat{e}_{ik}\sqrt{\hat{\lambda}_i}}{\sqrt{s_{kk}}}$$

to calculate the correlation coefficients.

$$r_{\hat{y}_1, x_1} = 1.00000$$
$$r_{\hat{y}_1, x_k} = 0.686136$$

Problem 8.10:

a) Construct the sample covariance matrix **S** and find the sample principal components in (8-20).

Computing the Sample Covariance matrix using `S_WRR <- cov(WRR)` where `WRR` is the data matrix of Weekly rates of returns.

Table 2: Sample Covariance matrix **S**

	JP Morgan	Citi Bank	Wells Fargo	Royal Dutch Shell	Exxon Mobile
JP Morgan	0.0004332695	0.0002756679	0.0001590265	0.0000641193	0.0000889662
Citi Bank	0.0002756679	0.0004387172	0.0001799737	0.0001814512	0.0001232623
Wells Fargo	0.0001590265	0.0001799737	0.0002239722	0.0000734135	0.0000605461
Royal Dutch Shell	0.0000641193	0.0001814512	0.0000734135	0.0007224964	0.0005082772
Exxon Mobile	0.0000889662	0.0001232623	0.0000605461	0.0005082772	0.0007656742

To find the sample principal components of the data matrix, we first compute the eigenvalues,

Table 3: Eigenvalues of the matrix **S**

$\hat{\lambda_1}$	$\hat{\lambda_1}$	$\hat{\lambda_1}$	$\hat{\lambda_1}$	$\hat{\lambda_1}$
0.0013677	0.0007012	0.0002538	0.0001426	0.0001189

and eigenvectors,

Table 4: Eigenvectors of the Matrix **S**

$\hat{\mathbf{e}}_1$	0.2228228	0.6252260	0.3261122	0.6627590	0.1176595
$\hat{\mathbf{e}}_2$	0.3072900	0.5703900	−0.2495901	−0.4140935	-0.5886080
$\hat{\mathbf{e}}_3$	0.1548103	0.3445049	−0.0376393	−0.4970499	0.7803043
$\hat{\mathbf{e}}_4$	0.6389680	−0.2479475	−0.6424974	0.3088689	0.1484555
$\hat{\mathbf{e}}_5$	0.6509044	−0.3218478	0.6458606	−0.2163758	−0.0937178

of the sample covariance matrix **S**. Therefore, the sample Principal components are,

$$\begin{aligned}
\hat{y}_1 &= 0.2228228x_1 + 0.6252260x_2 + 0.3261122x_3 + 0.6627590x_4 + 0.1176595x_5 \\
\hat{y}_2 &= 0.3072900x_1 + 0.5703900x_2 - 0.2495901x_3 - 0.4140935x_4 - 0.5886080x_5 \\
\hat{y}_3 &= 0.1548103x_1 + 0.3445049x_2 - 0.0376393x_3 + -0.4970499x_4 + 0.7803043x_5 \\
\hat{y}_4 &= 0.1548103x_1 + 0.3445049x_2 - 0.0376393x_3 - 0.4970499x_4 + 0.7803043x_5 \\
\hat{y}_5 &= 0.6509044x_1 - 0.3218478x_2 + 0.6458606x_3 - 0.2163758x_4 - 0.0937178x_5
\end{aligned}$$

b) Determine the proportion of the total sample variance explained by the first three principal components. Interpret these components.

The proportion of the total sample variance is given as,

$$\frac{\sum_{i=1}^{3} \hat{\lambda}_i}{\sum_{i=1}^{5} \hat{\lambda}_i} = 0.8988 = 89.8\%$$

c) Construct Bonferroni simultaneous 90% confidence intervals for the variances λ_1, λ_2 and λ_3 of the first three population components Y_1, Y_2 and Y_3.

Using (8-33) we can compute the Bonferroni simultaneous confidence intervals,

$$\frac{\hat{\lambda}_i}{(1 + z(\alpha/2)\sqrt{2/n})} \leq \lambda_i \leq \frac{\hat{\lambda}_i}{(1 - z(\alpha/2)\sqrt{2/n})} \tag{1}$$

with $n = 103$, $\alpha = 0.10$, and the eigenvalues $\hat{\lambda}_1 = 0.0013677$, $\hat{\lambda}_2 = 0.0007012$, and $\hat{\lambda}_3 = 0.0002538$. The intervals are defined below,

$$\begin{aligned}
\lambda_1 &: (0.001112652, 0.00177437) \\
\lambda_2 &: (0.000570417, 0.00090965) \\
\lambda_3 &: (0.000206476, 0.00032927)
\end{aligned}$$

d) Given the results in Parts a-c, do you feel that the stock rates of return data can be summarized in fewer than five dimensions?

Based off the results presented below in table 5, and the values from above that the stock rates of return data can be summarized in fewer than five dimensions, or two dimensions (80%).

Table 5: Summary of for the principal components and their variation

	Comp.1	Comp.2	Comp.3	Comp.4	Comp.5
Standard deviation	0.03680217	0.02635056	0.01585365	0.01188352	0.01085046
Proportion of Variance	0.52926066	0.27133298	0.09821584	0.05518400	0.04600652
Cumulative Proportion	0.52926066	0.80059364	0.89880948	0.95399348	1.00000000

Problem 8.26:

(Note: Only do PCA (a) - (d) for correlation matrix $\mathbf{R}$)

Consider the psychological profile data in Table 4.6. Using the five variables, Indep, Supp, Benev, Conform and Leader, performs a principal component analysis using the correlation matrix $\mathbf{R}$.

a) Determine the appropriate number of components to effectively summarize the variability. Construct a scree plot to aid in your determination.

Figure 3: Scree Plot of PC for 5 variables

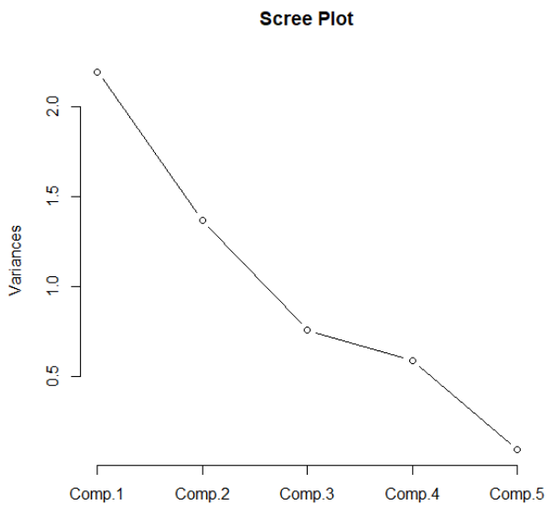


Table 6: Correlation matrix $\mathbf{R}$ generate by `R <- cor(X)`

1	−0.173	−0.561	−0.471	0.187
−0.173	1	0.018	−0.327	−0.401
−0.561	0.018	1	0.298	−0.492
−0.471	−0.327	0.298	1	−0.333
0.187	−0.401	−0.492	−0.333	1

Table 7: Eigen-values of the correlation matrix $\mathbf{R}$

$\hat{\lambda_1}$	$\hat{\lambda_2}$	$\hat{\lambda_3}$	$\hat{\lambda_4}$	$\hat{\lambda_5}$
2.197	1.368	0.756	0.589	0.090

Table 8: Eigen-vectors of the correlation matrix $\mathbf{R}$

$\mathbf{e}_1$	0.521	0.087	0.667	0.253	0.460
$\mathbf{e}_2$	−0.121	0.788	−0.187	−0.351	0.454
$\mathbf{e}_3$	−0.548	−0.008	−0.115	0.733	0.386
$\mathbf{e}_4$	−0.439	−0.491	0.295	−0.525	0.451
$\mathbf{e}_5$	0.469	−0.361	−0.648	−0.007	0.480

Table 9: Summary of for the principal components and their variation

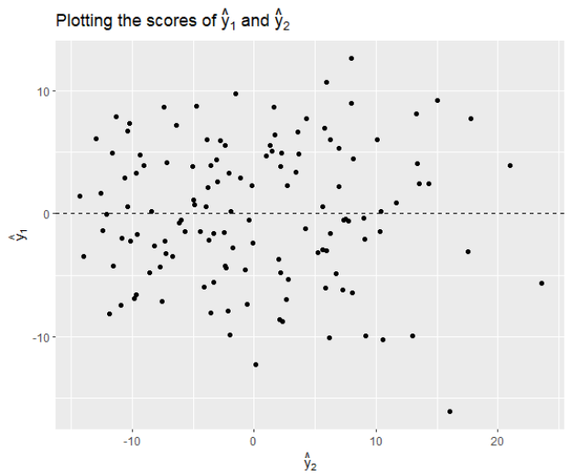
	Comp.1	Comp.2	Comp.3	Comp.4	Comp.5
Standard deviation	1.4821014	1.1697220	0.8694038	0.7673239	0.30079386
Proportion of Variance	0.4393249	0.2736499	0.1511726	0.1177572	0.01809539
Cumulative Proportion	0.4393249	0.7129748	0.8641474	0.9819046	1.00000000

b) Interpret the sample principal components.

According to the scree plot, there is not a clearly definitive elbow, however, we could infer that four sample principal components could effectively summarize the total sample variance. The total or cumulative proportion of the principal components is about 92.2%.

c) Using the values for the first two principal components, plot the data in a two-dimensional space with $\hat{y}_1$ along the vertical axis and $\hat{y}_2$ along the horizontal axis. Can you distinguish groups representing the two socioeconomic levels and/or the two genders? are there any outliers?

Figure 4: Plot of the $\hat{y}_1$ and $\hat{y}_2$ generated using `pcal <- princomp(X)` and computing the scores



After observing the scatter plot above, there seems to be 1 outlier.

d) Construct a 95% confidence interval for λ_1, the variance of the first population principal component from the covariance matrix.

Using Equation (8-33), with $n = 130$, $\alpha = 0.05$, and $\lambda_1 = 2.197$ we compute the 95% confidence interval as,

$$\frac{\hat{\lambda}_1}{(1 + z(\alpha/2)\sqrt{2/n})} \leq \lambda_1 \leq \frac{\hat{\lambda}_1}{(1 - z(\alpha/2)\sqrt{2/n})}$$

$$\lambda_1 : (1.767049, 2.902147)$$

Data

Chapter 7:

z_1	z_2	Y
15.31	57.3	74.8
15.2	63.8	74
16.25	65.4	72.9
14.33	57	70
14.57	63.8	74.9
17.33	62.2	76
14.48	60.2	72
14.91	57.7	73.5
15.25	56.4	74.5
13.89	55.6	73.5
15.18	62.6	71.5
14.44	63.4	71
14.87	60.2	78.9
18.63	67.2	86.5
15.2	57.1	68
25.76	89.6	102
19.05	68.6	84
15.37	60.1	69
18.06	66.3	88
16.35	65.8	76

Table 10: Table 7.1: Real-Estate Data