STAT5120, Regression Model Building, pt. 2, Allen Baumgarten (nominal Marlins fan)

1. Perform PCA on the Iris data. http://www.instantr.com/2012/12/18/performing-a-principal-component-analysis-in-r/

To view the dataset, simply type *iris* at the R prompt. We will not attempt to build a regression model for this dataset because the response (Species) is categorical, and so linear regression won't work. Let's just explore the nature of the relationships between the predictor variables. Run PCA (not PCR) on the variables Sepal.Length, Sepal.Width, Petal.Length, and Petal.Width.

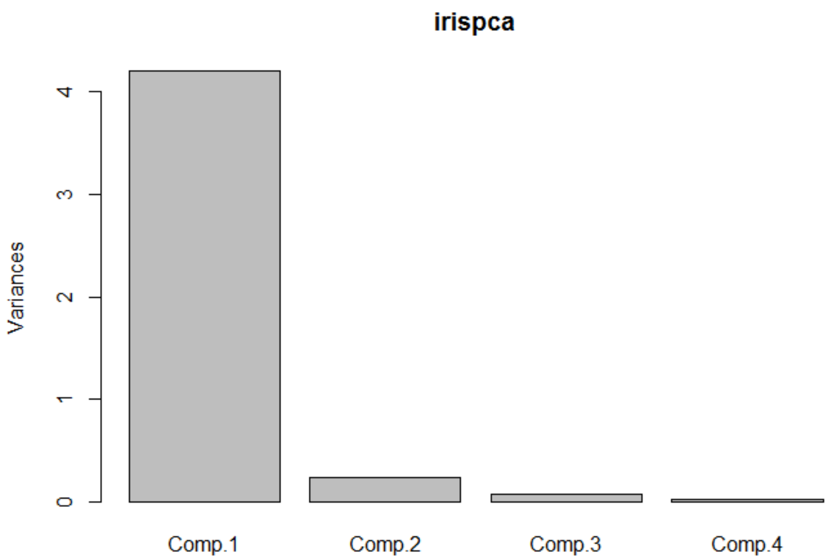
(a) List the eigenvalues in order from highest to lowest, along with the percentage of variation captured by each principle component.

Importance of components:

	Comp.1	Comp.2	Comp.3	Comp.4
Standard deviation	2.0494032	0.49097143	0.27872586	0.153870700
Proportion of Variance	0.9246187	0.05306648	0.01710261	0.005212184
Cumulative Proportion	0.9246187	0.97768521	0.99478782	1.000000000

(b) What is the total variation captured by the first component? 92.4% What is the total variation captured by the first two components? 97.8% The first three? 99.5% All four? 100%

(c) Make a scree plot. How many principle components do you think are enough to adequately describe the variation in the data? I would propose that the first principle component is adequate to describe the variation in the data with 92.4% of the variation captured. One could, of course, add in the second principle component to take that percentage up.



(d) What do the loadings for the components indicate? Be specific. The loadings are "weights that are used to multiply the original coordinates of the variables to get the new ones (called scores) on the principle components,"[1] and these in particular indicate that there is a strong correlation between the Petal Length with the weight assigned to it in the first principal component.

[1] Jones, Matthew O., "Chapter 11: Model Building II, Shrinkage Methods," 2006-Present, 163.

Loadings:

	Comp.1	Comp.2	Comp.3	Comp.4
Sepal.Length	0.361	-0.657	-0.582	0.315
Sepal.Width	-0.730	0.598	-0.320	
Petal.Length	0.857	0.173	-0.480	
Petal.Width	0.358	0.546	0.754	

(e) What do the scores for the observations tell you? The scores are shown below and indicate that component #1 indeed captures most of the variation:

	Comp.1	Comp.2	Comp.3	Comp.4
[1,]	-2.684125626	-0.319397247	-0.027914828	0.0022624371
[2,]	-2.714141687	0.177001225	-0.210464272	0.0990265503
[3,]	-2.888990569	0.144949426	0.017900256	0.0199683897

2. Another way to describe the lasso method is that it estimates the regression coefficients by choosing them to be the values of the b_j, j ∈ {0, 1, …, p-1} by minimizing

$$\sum_{i=1}^{n}\left(y_i - b_0 - \sum_{j=1}^{p-1} b_j x_{ij}\right)^2 \quad \text{subject to} \quad \sum_{k=1}^{p-1} |\beta_k| \leq s$$

for some number s. For parts (a) through (f), indicate which of the following occurs and justify your answer.

i. remain constant.

ii. monotonically increase.

iii. monotonically decrease.

iv. initially increase, then decrease.

v. initially decrease, then increase.

(a) As s increases from 0, the training SSE will monotonically increase

(b) As s increases from 0, the training R^2 will initially increase, then decrease increase

(c) As s increases from 0, the test or validation SSE will monotonically increase

(d) As s increases from 0, the test or validation R^2 will initially increase, then decrease

(e) As s increases from 0, the squared bias will initially decrease, then increase

(f) As s increases from 0, the variance will remain constant

3. Consider estimating regression coefficients by choosing the b_j, j ∈ {0, 1, …, p-1} that minimizes

$$\sum_{i=1}^{n}\left(y_i - b_0 - \sum_{j=1}^{p-1} b_j x_{ij}\right)^2 + \lambda \sum_{j=1}^{p-1} b_j^2$$

for fixed λ. For parts (a) through (f), indicate which of the following occurs and justify your answer.

i. remain constant.

ii. monotonically increase.

iii. monotonically decrease.

iv. initially increase, then decrease.

v. initially decrease, then increase.

(a) As λ increases from 0, the training SSE will remain constant

(b) As λ increases from 0, the training R^2 will monotonically increase

(c) As λ increases from 0, the test or validation SSE will monotonically decrease

(d) As λ increases from 0, the test or validation R^2 will initially increase, then decrease

(e) As λ increases from 0, the squared bias will initially decrease, then increase

(f) As λ increases from 0, the variance will remain constant

4. Load and read the documentation for the *College* data set from the ISLR package. We want to build a model to predict the number of applications received using the other variables.

(a) Split the data set into a training set and a validation/test set, approximately 70%, 30%, respectively. Split data into training and test groups (see code below).

(b) Fit a linear least-squares regression model on the training set. Compute the test MSE and test R^2.

```
Residuals:
   Min     1Q  Median     3Q    Max
-5235.2 -343.5     5.7  284.5 7185.2

Coefficients:
               Estimate   Std. Error   t value  Pr(>|t|)
(Intercept) -613.21658   462.63041  -1.325    0.18558
PrivateYes  -323.84919   169.29370  -1.913    0.05630 .
Accept         1.70689     0.04854  35.164  < 2e-16 ***
Enroll        -1.35509     0.22586  -6.000   3.68e-09 ***
Top10perc     45.42084     6.57841   6.905   1.46e-11 ***
Top25perc    -15.83576     5.27942  -3.000   0.00283 **
```

```
F.Undergrad    0.09912   0.03885     2.551   0.01101 *
P.Undergrad    0.01581   0.05051     0.313   0.75440
Outstate      -0.09220   0.02185    -4.220   2.88e-05 ***
Room.Board     0.11873   0.05396     2.200   0.02821 *
Books         -0.03743   0.25967    -0.144   0.88545
Personal       0.05974   0.07197     0.830   0.40686
PhD           -5.59724   5.12251    -1.093   0.27504
Terminal      -5.29911   5.53622    -0.957   0.33892
S.F.Ratio     21.40193  15.11700     1.416   0.15744
perc.alumni    1.97445   4.65425     0.424   0.67158
Expend         0.10761   0.01487     7.238   1.63e-12 ***
Grad.Rate      8.15148   3.29431     2.474   0.01366 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 992.5 on 526 degrees of freedom
Multiple R-squared:  0.9257,	Adjusted R-squared:  0.9232
F-statistic: 385.2 on 17 and 526 DF,  p-value: < 2.2e-16
```

(c) Fit a ridge regression model on the training set. Use cross-validation to choose the tuning parameter λ. Give the test MSE and test R^2.

```
       Length  Class  Mode
coef    357    -none- numeric
scales   17    -none- numeric
Inter     1    -none- numeric
lambda   21    -none- numeric
ym        1    -none- numeric
xm       17    -none- numeric
GCV      21    -none- numeric
kHKB      1    -none- numeric
kLW       1    -none- numeric
```

(d) Fit a lasso regression model on the training set. Use cross-validation to choose the tuning parameter λ. Give the test MSE and test R^2. Attempted a lasso regression model but unable to get this to work…

(e) Fit a principle components regression model on the training set and use cross-validation to choose the number of principle components. Give the test MSE and test R^2, and the number of principle components. Attempted a PC regression model but unable to get this to work…

(f) Fit a partial least squares regression model on the training set and use cross-validation to choose the number of new model features. Give the test MSE and test R^2, and the number of new features used in the model.

(g) Compare the five models. Which ones seem better? Is there much difference between the test R^2 and test MSE values? How well do these models predict the number of college applications?

5. Prove the form of the ridge regression coefficients:

$$\hat{\beta} = (X^T X + \lambda I^*)^{-1} X^T Y.$$

Proving this mathematically is a little over my head at this point, regrettably. I did find the proof itself mapped out as follows and can make some sense of it. The following, however, is NOT my work but is from a lecture given at Stanford (author acknowledged and footnoted below).[2]

Proving that $\hat{\beta}_\lambda^{\text{ridge}}$ is biased

- Let $\mathbf{R} = \mathbf{Z}^\top \mathbf{Z}$
- Then:

$$\begin{aligned}
\hat{\beta}_\lambda^{\text{ridge}} &= (\mathbf{Z}^\top \mathbf{Z} + \lambda \mathbf{I}_p)^{-1} \mathbf{Z}^\top \mathbf{y} \\
&= (\mathbf{R} + \lambda \mathbf{I}_p)^{-1} \mathbf{R} (\mathbf{R}^{-1} \mathbf{Z}^\top \mathbf{y}) \\
&= [\mathbf{R}(\mathbf{I}_p + \lambda \mathbf{R}^{-1})]^{-1} \mathbf{R} [(\mathbf{Z}^\top \mathbf{Z})^{-1} \mathbf{Z}^\top \mathbf{y}] \\
&= (\mathbf{I}_p + \lambda \mathbf{R}^{-1})^{-1} \mathbf{R}^{-1} \mathbf{R} \hat{\beta}^{\text{ls}} \\
&= (\mathbf{I}_p + \lambda \mathbf{R}^{-1}) \hat{\beta}^{\text{ls}}
\end{aligned}$$

- So:

$$\begin{aligned}
\mathbb{E}(\hat{\beta}_\lambda^{\text{ridge}}) &= \mathbb{E}\{(\mathbf{I}_p + \lambda \mathbf{R}^{-1}) \hat{\beta}^{\text{ls}}\} \\
&= (\mathbf{I}_p + \lambda \mathbf{R}^{-1}) \beta \\
&\overset{(\text{if } \lambda \neq 0)}{\neq} \beta.
\end{aligned}$$

[2] "Regularization: Ridge Regression and the LASSO" Lecture: Statistics 305, Autumn Quarter 2006/2007. Accessed on 4/8/18 at http://statweb.stanford.edu/~tibs/sta305files/Rudyregularization.pdf

APPENDIX: R SCRIPTS

Question 1:

```
> head(iris)
  Sepal.Length  Sepal.Width  Petal.Length  Petal.Width  Species
1        5.1          3.5           1.4           0.2       setosa
2        4.9          3.0           1.4           0.2       setosa
3        4.7          3.2           1.3           0.2       setosa
4        4.6          3.1           1.5           0.2       setosa
5        5.0          3.6           1.4           0.2       setosa
6        5.4          3.9           1.7           0.4       setosa

> irispca<-princomp(iris[-5])
> summary(irispca)
Importance of components:
                          Comp.1     Comp.2      Comp.3      Comp.4
Standard deviation      2.0494032  0.49097143  0.27872586  0.153870700
Proportion of Variance  0.9246187  0.05306648  0.01710261  0.005212184
Cumulative Proportion   0.9246187  0.97768521  0.99478782  1.000000000
> irispca$loadings

Loadings:
              Comp.1  Comp.2  Comp.3  Comp.4
Sepal.Length  0.361   -0.657  -0.582   0.315
Sepal.Width  -0.730    0.598  -0.320
Petal.Length  0.857    0.173  -0.480
Petal.Width   0.358    0.546   0.754

               Comp.1 Comp.2 Comp.3 Comp.4
SS loadings      1.00   1.00   1.00   1.00
Proportion Var   0.25   0.25   0.25   0.25
Cumulative Var   0.25   0.50   0.75   1.00
> irispca$scores
          Comp.1        Comp.2        Comp.3        Comp.4
 [1,] -2.684125626  -0.319397247  -0.027914828  0.0022624371
 [2,] -2.714141687   0.177001225  -0.210464272  0.0990265503
 [3,] -2.888990569   0.144949426   0.017900256  0.0199683897
> screeplot(irispca)
```

Question 4:

```
> library(ISLR)
Warning message:
package 'ISLR' was built under R version 3.4.4
> head(College)
                             Private Apps Accept Enroll Top10perc Top25perc F.Undergrad
Abilene Christian University     Yes 1660   1232    721        23        52        2885
Adelphi University               Yes 2186   1924    512        16        29        2683
Adrian College                   Yes 1428   1097    336        22        50        1036
Agnes Scott College              Yes  417    349    137        60        89         510
Alaska Pacific University        Yes  193    146     55        16        44         249
```

```
Albertson College              Yes  587    479    158       38       62        678
                          P.Undergrad Outstate Room.Board Books Personal PhD Terminal
Abilene Christian University        537     7440      3300   450     2200  70      78
Adelphi University               1227    12280      6450   750     1500  29      30
Adrian College                    99    11250      3750   400     1165  53      66
Agnes Scott College                63    12960      5450   450      875  92      97
Alaska Pacific University          869     7560      4120   800     1500  76      72
Albertson College                  41    13500      3335   500      675  67      73
                          S.F.Ratio perc.alumni Expend Grad.Rate
Abilene Christian University      18.1         12   7041       60
Adelphi University             12.2         16  10527       56
Adrian College                 12.9         30   8735       54
Agnes Scott College             7.7         37  19016       59
Alaska Pacific University       11.9          2  10922       15
Albertson College               9.4         11   9727       55

> college_training <- College[1:544,]
> college_test <- College[545:777,]
> regmod_collegetrain <- lm(college_training$Apps ~.,college_training)
> summary(regmod_collegetrain)

Call:
lm(formula = college_training$Apps ~ ., data = college_training)

Residuals:
    Min      1Q  Median      3Q     Max
-5235.2  -343.5     5.7   284.5  7185.2

Coefficients:
                Estimate      Std. Error     t value   Pr(>|t|)
(Intercept) -613.21658    462.63041   -1.325     0.18558
PrivateYes   -323.84919   169.29370   -1.913     0.05630 .
Accept          1.70689    0.04854     35.164   < 2e-16 ***
Enroll         -1.35509    0.22586     -6.000    3.68e-09 ***
Top10perc      45.42084    6.57841      6.905    1.46e-11 ***
Top25perc     -15.83576    5.27942     -3.000    0.00283 **
F.Undergrad     0.09912    0.03885      2.551    0.01101 *
P.Undergrad     0.01581    0.05051      0.313    0.75440
Outstate       -0.09220    0.02185     -4.220    2.88e-05 ***
Room.Board      0.11873    0.05396      2.200    0.02821 *
Books          -0.03743    0.25967     -0.144    0.88545
Personal        0.05974    0.07197      0.830    0.40686
PhD            -5.59724    5.12251     -1.093    0.27504
Terminal       -5.29911    5.53622     -0.957    0.33892
S.F.Ratio      21.40193   15.11700      1.416    0.15744
perc.alumni     1.97445    4.65425      0.424    0.67158
Expend          0.10761    0.01487      7.238    1.63e-12 ***
Grad.Rate       8.15148    3.29431      2.474    0.01366 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
Residual standard error: 992.5 on 526 degrees of freedom
Multiple R-squared:  0.9257,	Adjusted R-squared:  0.9232
F-statistic: 385.2 on 17 and 526 DF,  p-value: < 2.2e-16

> library(MASS)
> regmod_ridge <- lm.ridge(college_training$Apps ~.,college_training, lambda = seq(0, 5e-8, len=21))
> summary(regmod_ridge)


       Length Class  Mode   
coef   357    -none- numeric
scales  17    -none- numeric
Inter    1    -none- numeric
lambda  21    -none- numeric
ym       1    -none- numeric
xm      17    -none- numeric
GCV     21    -none- numeric
kHKB     1    -none- numeric
kLW      1    -none- numeric

> lasso_regmod <- lars(college_training$Apps, college_training$Enroll)
Error in rep(1, n) : invalid 'times' argument
> head(college_training)
                             Private Apps Accept Enroll Top10perc Top25perc F.Undergrad
Abilene Christian University     Yes 1660   1232    721        23        52        2885
Adelphi University               Yes 2186   1924    512        16        29        2683
Adrian College                   Yes 1428   1097    336        22        50        1036
Agnes Scott College              Yes  417    349    137        60        89         510
Alaska Pacific University        Yes  193    146     55        16        44         249
Albertson College                Yes  587    479    158        38        62         678
                             P.Undergrad Outstate Room.Board Books Personal PhD Terminal
Abilene Christian University         537     7440       3300   450     2200  70       78
Adelphi University                  1227    12280       6450   750     1500  29       30
Adrian College                        99    11250       3750   400     1165  53       66
Agnes Scott College                   63    12960       5450   450      875  92       97
Alaska Pacific University            869     7560       4120   800     1500  76       72
Albertson College                     41    13500       3335   500      675  67       73
                             S.F.Ratio perc.alumni Expend Grad.Rate
Abilene Christian University      18.1          12   7041        60
Adelphi University                12.2          16  10527        56
Adrian College                    12.9          30   8735        54
Agnes Scott College                7.7          37  19016        59
Alaska Pacific University         11.9           2  10922        15
Albertson College                  9.4          11   9727        55
> lasso_regmod <- lars(college_training[,-2], college_training$Enroll)
Error in one %*% x : requires numeric/complex matrix/vector arguments
> regmod_plsr <- plsr(college_training$Apps ~ ., data=college_training, ncomp=50, validation="CV")
Error in plsr(college_training$Apps ~ ., data = college_training, ncomp = 50,  : 
  could not find function "plsr"
```