Homework 4
Psychology 312

1. *Multiple Regression. The Model and its Estimates.* In standard multiple regression, we fit the model

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{e} \tag{1}$$

where $\boldsymbol{y}$ is an $n \times 1$ vector of independent criterion scores, $\boldsymbol{X}$ an $n \times p'$ matrix of scores on p' predictor terms, including an intercept term. So in the typical case where there are p predictors and an intercept, $p' = p + 1$. The predictor scores in $\boldsymbol{X}$ are considered fixed constants. The errors in $\boldsymbol{e}$ are independent random variables, with zero means and constant variance σ^2. So we say that

$$\mathrm{Var}(\boldsymbol{e}) = \sigma^2(\boldsymbol{I}) \tag{2}$$

Note that since both $\boldsymbol{X}$ and $\boldsymbol{\beta}$ contain only constants, it immediately follows that

$$\mathrm{Var}(\boldsymbol{y}) = \mathrm{Var}(\boldsymbol{e}) = \sigma^2(\boldsymbol{I}) \tag{3}$$

and

$$\mathrm{E}(\boldsymbol{y}) = \boldsymbol{X}\boldsymbol{\beta} \tag{4}$$

We'll assume a parameterization $\boldsymbol{X}$ of full column rank, including a column of 1s for the intercept if necessary. So $(\boldsymbol{X}'\boldsymbol{X})^{-1}$ exists. The $p' \times 1$ vector $\boldsymbol{\beta}$ contains the regression parameters, including the intercept. The vector $\boldsymbol{e}$ consists of errors.

Note: these errors are *unobservable*, because $\boldsymbol{\beta}$ is itself not known. We can *estimate* $\boldsymbol{\beta}$, and use the estimate to construct *estimates* of errors. Note again, these estimates are the model *residuals*. The residuals are not the errors — they are estimates of the errors. Although the errors are assumed to be uncorrelated, the residuals, as we shall see, are not.

The least squares estimates of the regression parameters in $\boldsymbol{\beta}$ are calculated as

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y} \tag{5}$$

The *predicted scores* are

$$\begin{aligned} \hat{\boldsymbol{y}} &= \boldsymbol{X}\hat{\boldsymbol{\beta}} && (6)\\ &= \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y} && (7)\\ &= \boldsymbol{P}_x\boldsymbol{y} && (8) \end{aligned}$$

The *residuals* are

$$\begin{aligned}\hat{\boldsymbol{e}} &= \boldsymbol{y} - \hat{\boldsymbol{y}} \\ &= \boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}} \\ &= \boldsymbol{y} - \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y} \\ &= \boldsymbol{y} - \boldsymbol{P}_x\boldsymbol{y} \\ &= (\boldsymbol{I} - \boldsymbol{P}_x)\boldsymbol{y} \\ &= \boldsymbol{Q}_x\boldsymbol{y} \end{aligned} \tag{9}$$

(a) **(10 points)**. While the elements of $\boldsymbol{\beta}$ are constants, the elements of $\hat{\boldsymbol{\beta}}$ are random variables, and have an expected value and a variance. All regression programs compute and report estimated *standard errors* for the elements of $\hat{\boldsymbol{\beta}}$, i.e., the estimates of regression parameters. These are obtained as follows: (a) Compute an estimate of the variance-covariance matrix of the random vector $\hat{\boldsymbol{\beta}}$; (b) Since the diagonal elements of this estimated variance-covariance matrix are estimated variances, take the square roots of these diagonal elements to produce estimated standard errors.

Given the above, and remembering that $\boldsymbol{P}_x$ and $\boldsymbol{Q}_x$ are matrices of constants, **prove the following two results**.

i.

$$\mathrm{E}(\hat{\boldsymbol{\beta}}) = \boldsymbol{\beta} \tag{10}$$

Hint. Combine Equations 4 and 5.

ii.

$$\boldsymbol{\Sigma}_{\hat{\boldsymbol{\beta}}\hat{\boldsymbol{\beta}}} = \mathrm{Var}(\hat{\boldsymbol{\beta}}) = \sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1} \tag{11}$$

Hint. Use your results from expected value algebra and compute $\mathrm{Var}(\hat{\boldsymbol{\beta}})$ as $\mathrm{Var}(\boldsymbol{A}'\boldsymbol{y})$, where $\boldsymbol{A}' = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'$, remembering the result on the variance of a linear combination that I described as essential to multivariate analysis theory, and using the result of Equation 3.

Note that, since σ^2 is not known, it must be estimated. The unbiased estimator is

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{n} \hat{e}_i^2}{n - p'} \tag{12}$$

The estimate for the variance-covariance matrix of $\hat{\boldsymbol{\beta}}$ is therefore

$$\hat{\boldsymbol{\Sigma}}_{\hat{\boldsymbol{\beta}}\hat{\boldsymbol{\beta}}} = \widehat{\mathrm{Var}}(\hat{\boldsymbol{\beta}}) = \hat{\sigma}^2(\boldsymbol{X}'\boldsymbol{X})^{-1} \tag{13}$$

(b) **(10 points)**. Start up R, and load in the file *KidIqData.csv*, using the `read.csv` command. Download the R Utility Functions from the R section of the website. They are accessible under the *R Utility Functions* link.

Load in the functions as shown below. The actual filename is *Steiger R Library Functions.txt*. You will probably find it convenient to `attach` the data file. Treat the variable `kid.iq` as the dependent variable $\boldsymbol{y}$, and treat the variables `mom.iq` and `mom.grad` as the predictors in a matrix $\boldsymbol{X}$. Verify that the file has 500 observations on 3 variables. Add an intercept variable (a column of 1's) to $\boldsymbol{X}$. Here is an example of how to do this.

```
data <- read.csv("KidIqData.csv")
attach(data)
dim(data)

## [1] 500   3

names(data)

## [1] "kid.iq"   "mom.iq"   "mom.grad"

## Create an intercept variable
source("Steiger R Library Functions.txt")
one <- UnitVector(500)
X <- cbind(one,mom.iq,mom.grad)
y <- kid.iq
```

The variable `mom.grad` is a binary variable coded 0-1, depending on whether the mother graduated from high school. Using the `lm` function, predict `kid.iq` from `mom.iq` and `mom.grad` and save the result as a `lm` object. (*Note*, if you don't know how to do this, examine the online lecture notes on multiple regression and chapters 3 and 4 from Gelman and Hill.) Apply the `summary` function to the object. You should see values for the intercept and regression coefficients corresponding to $\boldsymbol{\beta}' = [32.83, 0.66, 14.99]$. Notice also that there are standard errors reported for the coefficients. The standard errors are, respectively, 4.891,0.049, 1.948. Notice also that the program prints a "Residual Standard Error" of 14.46. This is the square root of the quantity defined in Equation 12. Here is the actual output.

```
##
## Call:
## lm(formula = kid.iq ~ mom.iq + mom.grad)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -37.183  -9.282  -0.201  10.247  37.993
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 32.83107    4.89084   6.713 5.22e-11 ***
```

```
## mom.iq       0.66067    0.04934  13.390  < 2e-16 ***
## mom.grad    14.98590    1.94786   7.694 7.76e-14 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 14.46 on 497 degrees of freedom
## Multiple R-squared:  0.4631,Adjusted R-squared:  0.4609
## F-statistic: 214.3 on 2 and 497 DF,  p-value: < 2.2e-16
```

Calculate $\hat{\boldsymbol{\beta}}$ from Equation 5, and verify that it agrees with the R output.

(c) **(10 points)**. Once you have the fit object, it is easy to extract the residuals. Just apply the `residuals` function to the fit object you have saved. **Calculate the estimated residual variance** defined in Equation 12. Check that its square root is equal to the "residual standard error" shown in the R output. (*Hint.* Remember the definition of p'.)

The predicted scores can be generated in R by applying the `predict` function to the linear model object. Of course, we can also calculate them as $\boldsymbol{P}_x\boldsymbol{y}$. My support routines contain the functions `P` and `Q` to allow you to create projectors easily and directly. Compute the predicted scores using `predict`, and extract the first predicted score. Compare it to the first element of the array calculated in R using the as `y.hat <- P(X) %*% y`. Then compare the value of the first element of the vector of residuals produced by the `residuals` function, and verify that `(Q(X) %*% y)[1]` is in fact equal to the first residual.

(d) **(10 points)**. Using the estimated residual variance you computed in the preceding part, and also using the result of Equation 13, **compute the estimated standard errors** of the elements of $\hat{\boldsymbol{\beta}}$, and verify that they agree with the R output.

(e) **(10 points)**. The squared multiple correlation R^2 between $\boldsymbol{y}$ and the non-intercept variables in $\boldsymbol{X}$ can be calculated in a number of ways. One way is from a formula analogous to the population formula given in the lecture slides on Key Regression Algebra, i.e.

$$R^2 = \frac{\boldsymbol{s}_{yx}\boldsymbol{S}_{xx}^{-1}\boldsymbol{s}_{xy}}{s_y^2} = \frac{\boldsymbol{s}'_{xy}\boldsymbol{S}_{xx}^{-1}\boldsymbol{s}_{xy}}{s_y^2} \tag{14}$$

Another way is to use the linear model fit object, compute the correlation between $\boldsymbol{y}$ and $\hat{\boldsymbol{y}}$, and square it, i.e.

$$R^2 = r_{y,\hat{y}}^2 \tag{15}$$

Compute R^2 both ways with R (remember, the first calculation will require that you drop the intercept column from $\boldsymbol{X}$), and verify that the result agrees with the value of 0.4631 printed in the summary table from the `lm` object.

(f) **(5 points)**. Verify in one line of R that the predicted and error scores from the linear model analysis have a correlation of zero.

2. *Eigenvalues and Eigenvectors.* Load in the data from the *AthleticsData.csv* file.

(a) **(10 points)**. Compute the correlation matrix for the AthleticsData. Call it $\boldsymbol{R}_{yy}$. Then compute the eigenvectors and eigenvalues of the correlation matrix using the `eigen` command. Create a matrix $\boldsymbol{V}$ containing the eigenvectors and a diagonal matrix $\boldsymbol{D}$ containing the eigenvalues on the diagonal. Verify in R that

$$\boldsymbol{R}_{yy} = \boldsymbol{V}\boldsymbol{D}\boldsymbol{V}' = \boldsymbol{F}\boldsymbol{F}'$$

up to rounding error by computing the square root of the sum of squared differences between $\boldsymbol{R}_{yy}$ and $\boldsymbol{V}\boldsymbol{D}\boldsymbol{V}'$.

(b) **(10 points)**. Compute $\boldsymbol{F} = \boldsymbol{V}\boldsymbol{D}^{1/2}$ The matrix is sometimes called the "principal component pattern." Install the `psych` library on your system, load the library, and compute the raw principal component pattern with the command

```
principal(AthleticsData,nfactors=9,rotate="none").
```

Compare the output with your $\boldsymbol{F}$. Explain any discrepancy.

(c) **(10 points)**. Extract the first 3 columns of your $\boldsymbol{F}$. Call the extracted columns $\boldsymbol{G}$. Compute $\hat{\boldsymbol{R}} = \boldsymbol{G}\boldsymbol{G}'$, and compare it to $\boldsymbol{R}_{yy}$ as follows. Extract the elements of $\boldsymbol{R}_{yy}$ into an 81×1 vector, and, likewise, extract the elements of $\hat{\boldsymbol{R}}$ as an 81×1 vector. Compute the correlation between the two vectors and do a scatterplot. Add an identity line (slope 1, intercept 0) to the plot to aid interpretation. Notice anything interesting in the scatterplot? Is there some way you could "clean up" the scatterplot to improve it as a way of comparing the two matrices?

3. **(15 points)**. In the class notes on Key Regression Algebra, we discussed a "regression component" system of the form

$$\boldsymbol{y} = \boldsymbol{F}\boldsymbol{x} + \boldsymbol{e}$$

in which all variables are in deviation score form. We define $\boldsymbol{\Sigma} = \mathrm{E}(\boldsymbol{y}\boldsymbol{y}')$, and assume $\boldsymbol{x} = \boldsymbol{B}'\boldsymbol{y}$, and $\boldsymbol{F}$ is a set of least squares linear regression

weights for predicting $\boldsymbol{y}$ from $\boldsymbol{x}$. In the notes, we showed that $\boldsymbol{F}$ can be computed as

$$\boldsymbol{F} = \boldsymbol{\Sigma}\boldsymbol{B}(\boldsymbol{B}'\boldsymbol{\Sigma}\boldsymbol{B})^{-1} \tag{16}$$

This means that given $\boldsymbol{B}$ and $\boldsymbol{\Sigma}$, the system is completely determinate.

(a) Using simple (but careful and involved) substitution, show that, given $\boldsymbol{F}$ and $\boldsymbol{\Sigma}$, $\boldsymbol{B}$ is completely determined and may be calculated as

$$\boldsymbol{B} = \boldsymbol{\Sigma}^{-1}\boldsymbol{F}(\boldsymbol{F}'\boldsymbol{\Sigma}^{-1}\boldsymbol{F})^{-1} \tag{17}$$

(*Hint*: simply substitute Equation 16 into the right side of Equation 17 and show it reduces to $\boldsymbol{B}$.)

(b) Principal components analysis is a special case of regression component analysis, so we may apply the rules of the latter to the former. Earlier, we demonstrated that if $\boldsymbol{\Sigma} = \boldsymbol{V}\boldsymbol{D}\boldsymbol{V}'$ is an Eckart-Young decomposition of $\boldsymbol{\Sigma}$ into eigenvectors and eigenvalues, that the principal components pattern may be written as $\boldsymbol{F} = \boldsymbol{V}\boldsymbol{D}^{1/2}$. Using Equation 17, and your knowledge of matrix factorization and symmetric square roots, prove that, if $\boldsymbol{F} = \boldsymbol{V}\boldsymbol{D}^{1/2}$, then $\boldsymbol{B} = \boldsymbol{V}\boldsymbol{D}^{-1/2}$.

(c) Prove that, if $\boldsymbol{x} = \boldsymbol{B}'\boldsymbol{y}$, and $\boldsymbol{B} = \boldsymbol{V}\boldsymbol{D}^{-1/2}$, then $\mathrm{Var}(\boldsymbol{x}) = \mathrm{E}(\boldsymbol{x}\boldsymbol{x}') = \boldsymbol{I}$.