Assignment 3 - Econometrics

Victoria Milani

1. Predicting Wages in USA

(a) Load, prepare and summarize data.

```
# US census data from the CPS in the year 2012.

df <- cps2012
describe(df)
```

```
# y is the dependent variable: lnw <- log_wage
y <- df %>% select(lnw) %>% data.matrix()

# x is the predictors' matrix , using the 16 variables
x <- df %>% select(female, widowed, divorced, separated, nevermarried, hsd08, hsd911,
    hsg, cg, ad, mw, so, we, exp1, exp2, exp3) %>% data.matrix()

# Splitting the data in order to make the tests MSE
set.seed(6697)
## Defining training set
train <- df %>% sample_frac(0.7)
## Defining test set
test <- df %>% setdiff(train)

## Dependent variable y: lnw
train_y <- train %>% select(lnw) %>% unlist() %>% data.matrix()
test_y <- test %>% select(lnw) %>% unlist() %>% data.matrix()
## Matrix of predictors
train_x <- train %>% select(female, widowed, divorced, separated, nevermarried, hsd08,
    hsd911, hsg, cg, ad, mw, so, we, exp1, exp2, exp3) %>% data.matrix()
test_x <- test %>% select(female, widowed, divorced, separated, nevermarried, hsd08,
    hsd911, hsg, cg, ad, mw, so, we, exp1, exp2, exp3) %>% data.matrix()
```

(b) Apply Ridge-Regression with cross-validation (CV):

(i) Apply ridge regression to the previous dataset for the the default grid of values of lambda. Plot the 10-fold CV MSE as a function of lambda.

```
### RIDGE REGRESSION

# Fit ridge regression using CV
ridge <- cv.glmnet(x, y, alpha = 0)

# Plot the 10-fold as function of lambda
plot(ridge)
```

(ii) Then, select the optimal lambda (λ) by cross-validation. How many variables are used in the Ridge fit?

```
# Finding the best lambda by CV
(best_lambda <- ridge$lambda.min)
```

```
## [1] 0.01845918
```

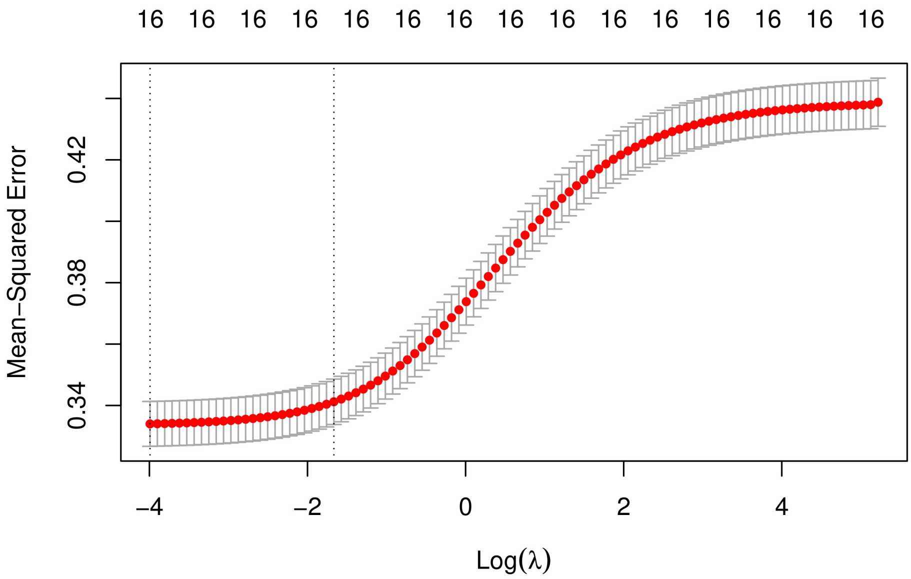

Figure 1: MSE for the ridge regression predictions, as a function of λ.

```
# Fitting the model for the best lambda
(ridgereg = glmnet(x, y, alpha = 0, lambda = best_lambda))
```

```
##
## Call:  glmnet(x = x, y = y, alpha = 0, lambda = best_lambda)
##
##   Df  %Dev  Lambda
## 1 16 23.97 0.01846
```

The optimal lambda by CV is the one that minimizes the training Mean Squared Error (MSE) with value equal 0.0185. The numbers of variables used was 16 and it is possible to know since Ridge does not make any variable selection.

(iii) Why is the test MSE for Ridge often smaller than for OLS when lambda is not zero?

```
# Calculating the MSE test for Ridge
training_ridgereg = glmnet(train_x, train_y, alpha = 0, lambda = best_lambda)
(mse_ridge <- mean(((ridge_pred = predict(training_ridgereg, s = best_lambda, newx = test_x,
    newdata = test)) - test_y)^2))
```

```
## [1] 0.3285392
```

```
# Calculating the MSE test for OLS_1
training_olsreg1 = glmnet(train_x, train_y, alpha = 0, lambda = 0)
(mse_ols1 <- mean(((ols1_pred = predict(training_olsreg1, newx = test_x, newdata = test)) -
    test_y)^2))
```

```
## [1] 0.3278118
```

```
# Calculating the MSE test for OLS_2
training_olsreg2 = lm(train_y ~ female + widowed + divorced + separated + nevermarried +
```

```r
    hsd08 + hsd911 + hsg + cg + ad + mw + so + we + exp1 + exp2 + exp3, data = train)
(mse_ols2 <- mean(((ols2_pred = predict(training_olsreg2, newx = test_x, newdata = test)) -
    test_y)^2))
```

```
## [1] 0.3278162
```

Usually the test MSE for Ridge is smaller because of the optimization method that is used. It will increase the bias, but decrease the variance, and since $MSE = bias + \sigma^2$, we can obtains some prediction gains by compensating the effects.

More intuitively, Ridge drecrease the relevance on the regression from the variables with less explanatory power, making their coefficient shrinks to zero, what rises the explanatory power to outside-sample data. The OLS describes better inside-sample data, but has in overfitting on the own dataset. So, depending on the objective, you can pick with method fits better.

(iv) What is the optimal value of lambda? Is unrestricted OLS optimal here, in a test MSE sense?

Considering the fact that the optimal lambda is very close to zero(0.0185), when comparing both methods, the tests MSE have similar results, but OLS seems to be better on this case, with a MSE value of 0.3278, compared to 0.3285 from Ridge.

(c) Apply LASSO with cross-validation (CV):

(i) Apply Lasso regression to the previous dataset for the the default grid of values of lambda. Plot the 10-fold CV MSE as a function of lambda.

```r
# y (dependent variable) and x (predictors' matrix) remains the same from part b)

### LASSO REGRESSION

# Fit LASSO regression using CV
lasso <- cv.glmnet(x, y, alpha = 1)
coef(lasso)
```

```
## 17 x 1 sparse Matrix of class "dgCMatrix"
##                         1
## (Intercept)   2.688330434
## female       -0.237743587
## widowed       .
## divorced      .
## separated     .
## nevermarried -0.097881013
## hsd08        -0.243151434
## hsd911       -0.239070677
## hsg          -0.137901576
## cg            0.288622837
## ad            0.520915658
## mw           -0.032472027
## so            .
## we            .
## exp1          0.006015065
## exp2          .
## exp3          .
```

```r
# Plot the 10-fold as function of lambda
plot(lasso)
```

(ii) Then, select the optimal lambda (λ) by cross-validation. What is the optimal lambda?

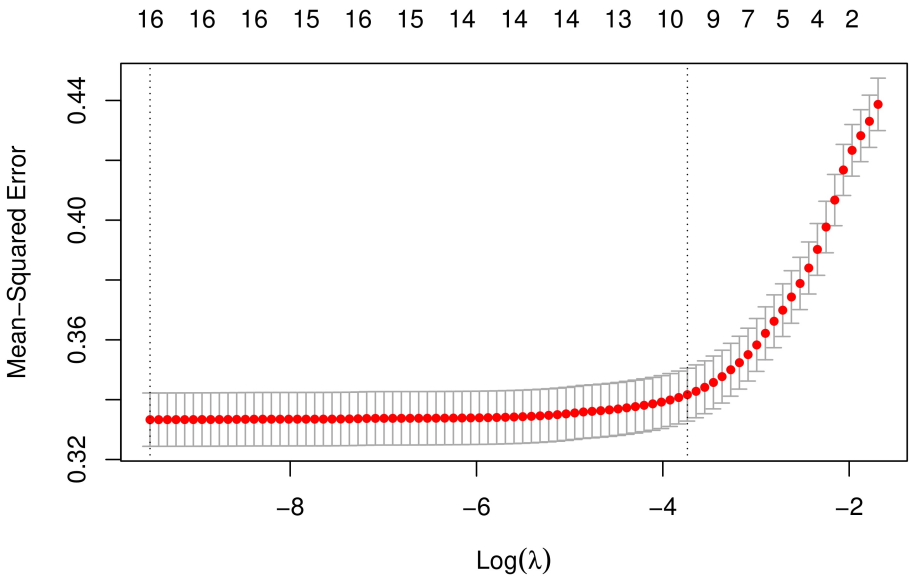

Figure 2: MSE for the LASSO regression predictions, as a function of λ.

```r
# Finding lambda by CV
(minimal_lambda_lasso <- lasso$lambda.min)
```

```
## [1] 7.452004e-05
```

The minimum lambda by CV is equal to 7.4520045×10^{-5}.

(iii) How many variables are used in the optimal Lasso fit? What are their coefficients? Is there a big difference here between Ridge and Lasso (in terms of test MSE)?

```r
# Fitting the model
lassoreg_min = glmnet(x, y, alpha = 1, lambda = minimal_lambda_lasso)
```

For the minimal value of lambda, lasso does not make variables selection, so it uses the 16 variables, with coefficients showed above.

```r
# Calculating the MSE test for LASSO
training_lasso = glmnet(train_x, train_y, alpha = 1, lambda = minimal_lambda_lasso)
(mse_lasso <- mean(((lasso_pred = predict(training_lasso, s = minimal_lambda_lasso,
    newx = test_x, newdata = test)) - test_y)^2))
```

```
## [1] 0.327846
```

Remembering that MSE for Ridge was 0.3285, now we compare to the MSE for LASSO that has a value of 0.3278. Both have a pretty close value, but the one from LASSO is still a little bit smaller.

(iv) Which method of prediction would you choose and why? Is gender an important factor in the prediction model? Interpret the coefficient of female.

```r
lassofinal = glmnet(x, y, alpha = 1, lambda = minimal_lambda_lasso)
coef(lassofinal)
```

```
## 17 x 1 sparse Matrix of class "dgCMatrix"
```

```
##                      s0
## (Intercept)   2.47732797
## female       -0.27953288
## widowed      -0.14203369
## divorced     -0.07796664
## separated    -0.10818413
## nevermarried -0.13131962
## hsd08        -0.61290619
## hsd911       -0.39086942
## hsg          -0.17319759
## cg            0.35189072
## ad            0.59926504
## mw           -0.10546213
## so           -0.05371613
## we           -0.01092662
## exp1          0.04141845
## exp2         -0.12645027
## exp3          0.01342990
```

I would choose LASSO method, since it is the one with smaller MSE. Also, it is a better method for predicting than OLS, even if both had almost the same MSE value.

Considering the `female` variable, it is possible to say that it is relevant for understanding the wages, meaning that gender affects the wage that a person can have. An interpretation would be that if you have `female = 1`, that is, if you are a woman, you receive 28% less than if you were a man, ceteris paribus.

(d) Repeat (b) and (c) for a more flexible speciffication: You would like to analyse the effect of gender and interaction e§ects of other variables with gender on wage jointly. The dependent variable is still the logarithm of the wage.

```
# x_int is the predictors' matrix , using the 16 variables and interactions
x_int <- model.matrix(~female * (widowed + divorced + separated + nevermarried +
    hsd08 + hsd911 + hsg + cg + ad + mw + so + we + exp1 + exp2 + exp3), df)

# Total matrix with y and x_int
int <- cbind(df[, "lnw"], x_int) %>% as.data.frame()

# Splitting the data in order to make the tests MSE
set.seed(6697)
## Defining training set
int_train <- int %>% sample_frac(0.7)
## Defining test set
int_test <- int %>% setdiff(int_train)

## Dependent variable y: lnw
int_train_y <- int_train %>% select(V1) %>% unlist() %>% data.matrix()
int_test_y <- int_test %>% select(V1) %>% unlist() %>% data.matrix()
## Matrix of predictors
int_train_x <- int_train[, -1:-2] %>% data.matrix()
int_test_x <- int_test[, -1:-2] %>% data.matrix()
```

b) (i)

```
### RIDGE REGRESSION

# Fit ridge regression using CV
```

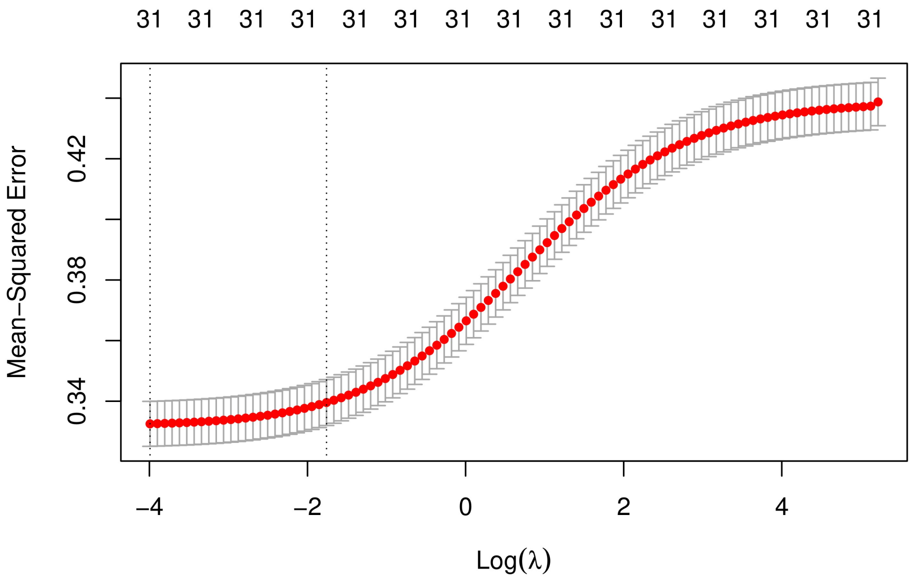

Figure 3: MSE for the ridge regression predictions, as a function of λ.

```
ridge_int <- cv.glmnet(x_int, y, alpha = 0)

# Plot the 10-fold as function of lambda
plot(ridge_int)
```

(ii)

```
# Finding the best lambda by CV
(best_lambda_int <- ridge_int$lambda.min)
```

```
## [1] 0.01845918
```

```
# Fitting the model for the best lambda
(ridgereg_int = glmnet(x_int, y, alpha = 0, lambda = best_lambda_int))
```

```
##
## Call:  glmnet(x = x_int, y = y, alpha = 0, lambda = best_lambda_int)
##
##   Df  %Dev  Lambda
## 1 31 24.37 0.01846
```

The optimal lambda by CV is the one that minimizes the training Mean Squared Error (MSE) with value equal 0.0185. The numbers of variables used was 16 and it is possible to know since Ridge does not make any variable selection.

(iii)

```
# Calculating the MSE test for Ridge
training_ridgereg_int = glmnet(int_train_x, int_train_y, alpha = 0, lambda = best_lambda_int)
(mse_ridge_int <- mean(((ridge_pred_int = predict(training_ridgereg_int, s = best_lambda_int,
    newx = int_test_x, newdata = int_test)) - int_test_y)^2))
```

```
## [1] 0.3505775
```

```r
# Calculating the MSE test for OLS_1
training_olsreg1_int = glmnet(int_train_x, int_train_y, alpha = 0, lambda = 0)
(mse_ols1_int <- mean(((ols1_pred_int = predict(training_olsreg1_int, newx = int_test_x,
    newdata = int_test)) - int_test_y)^2))
```

```
## [1] 0.3498975
```

Same justify as in part b)

(iv)

Considering the fact that the optimal lambda is very close to zero (0.0185), when comparing both methods, the tests MSE have similar results, but OLS seems to be better on this case, with a MSE value of 0.3499, compared to 0.3506 from Ridge.

(c) (i)

```r
# y (dependent variable) and x (predictors' matrix) remains the same from part b)

### LASSO REGRESSION

# Fit LASSO regression using CV
lasso_int <- cv.glmnet(x_int, y, alpha = 1)
coef(lasso_int)
```

```
## 33 x 1 sparse Matrix of class "dgCMatrix"
##                                  1
## (Intercept)          2.677153e+00
## (Intercept)          .
## female              -2.401581e-01
## widowed              .
## divorced            -1.035886e-02
## separated            .
## nevermarried        -1.055414e-01
## hsd08               -3.108115e-01
## hsd911              -2.589873e-01
## hsg                 -1.366031e-01
## cg                   2.992926e-01
## ad                   5.343762e-01
## mw                  -4.070119e-02
## so                   .
## we                   .
## exp1                 6.798627e-03
## exp2                 .
## exp3                 .
## female:widowed       .
## female:divorced      .
## female:separated     .
## female:nevermarried  .
## female:hsd08         .
## female:hsd911       -2.686264e-02
## female:hsg          -2.151171e-02
## female:cg            .
## female:ad            .
## female:mw            .
```

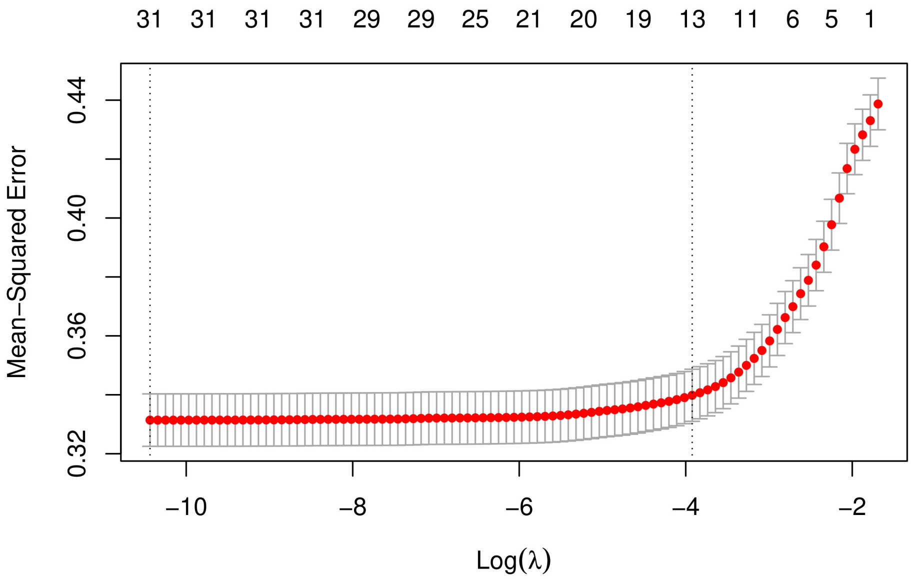

Figure 4: MSE for the LASSO regression predictions, as a function of λ.

```
## female:so            .
## female:we            .
## female:exp1          .
## female:exp2          .
## female:exp3          -5.619678e-05
```

```
# Plot the 10-fold as function of lambda
plot(lasso_int)
```

(ii)

```
# Finding lambda by CV
(minimal_lambda_lasso_int <- lasso_int$lambda.min)
```

```
## [1] 2.939224e-05
```

The minimum lambda by CV is equal to 2.9392241×10^{-5}.

(iii)

```
# Fitting the model
lassoreg_min_int = glmnet(x_int, y, alpha = 1, lambda = minimal_lambda_lasso_int)
```

For the minimal value of lambda, lasso does not make variables selection, so it uses the 16 variables, with coefficients showed above.

```
# Calculating the MSE test for LASSO
training_lasso_int = glmnet(int_train_x, int_train_y, alpha = 1, lambda = minimal_lambda_lasso_int)
(mse_lasso_int <- mean(((lasso_pred_int = predict(training_lasso_int, s = minimal_lambda_lasso_int,
    newx = int_test_x, newdata = int_test)) - int_test_y)^2))
```

```
## [1] 0.3498872
```

Remembering that MSE for Ridge was 0.3506, now we compare to the MSE for LASSO that has a value of 0.3499. Both have a pretty close value, but the one from LASSO is still a little bit smaller.

(iv)

```
lassofinal_int = glmnet(x_int, y, alpha = 1, lambda = minimal_lambda_lasso_int)
coef(lassofinal_int)
```

```
## 33 x 1 sparse Matrix of class "dgCMatrix"
##                               s0
## (Intercept)          2.411280618
## (Intercept)          .
## female              -0.199768644
## widowed             -0.223885142
## divorced            -0.145807258
## separated           -0.121916724
## nevermarried        -0.203013592
## hsd08               -0.585074504
## hsd911              -0.348249945
## hsg                 -0.160983553
## cg                   0.349802303
## ad                   0.605030390
## mw                  -0.103557574
## so                  -0.050664392
## we                  -0.009044125
## exp1                 0.049831849
## exp2                -0.150275151
## exp3                 0.015312780
## female:widowed       0.140673017
## female:divorced      0.136599480
## female:separated     0.038388246
## female:nevermarried  0.170898479
## female:hsd08        -0.124887405
## female:hsd911       -0.153760941
## female:hsg          -0.024558589
## female:cg            0.006297287
## female:ad           -0.017944727
## female:mw           -0.002341415
## female:so           -0.007092895
## female:we           -0.003529873
## female:exp1         -0.005635932
## female:exp2         -0.024508250
## female:exp3          0.009187785
```

I would choose LASSO method, since it is the one with smaller MSE. Also, it is a better method for predicting than OLS, even if both had almost the same MSE value.

Considering the `female` variable, it is possible to say that it is relevant for understanding the wages, meaning that gender affects the wage that a person can have. An interpretation would be that if you have `female = 1`, that is, if you are a woman, you receive 28% less than if you were a man, ceteris paribus.

(e) Based on the variable selection provided by Lasso, study the Gender Pay Gap (GPG) with this data set, including an Oaxaca-Blinder decomposition for the GPG.

The variable selection used is the one related to the first LASSO regression made on part 1-c)(i)

```
## 17 x 1 sparse Matrix of class "dgCMatrix"
```

```
##                         1
## (Intercept)   2.688330434
## female       -0.237743587
## widowed       .
## divorced      .
## separated     .
## nevermarried -0.097881013
## hsd08        -0.243151434
## hsd911       -0.239070677
## hsg          -0.137901576
## cg            0.288622837
## ad            0.520915658
## mw           -0.032472027
## so            .
## we            .
## exp1          0.006015065
## exp2          .
## exp3          .
```

```
## oaxaca: oaxaca() performing analysis. Please wait.
```

```
##
## Bootstrapping standard errors:
```

```
## 1 / 100 (1%)
```

```
## 10 / 100 (10%)
```

```
## 20 / 100 (20%)
```

```
## 30 / 100 (30%)
```

```
## 40 / 100 (40%)
```

```
## 50 / 100 (50%)
```

```
## 60 / 100 (60%)
```

```
## 70 / 100 (70%)
```

```
## 80 / 100 (80%)
```

```
## 90 / 100 (90%)
```

```
## 100 / 100 (100%)
```

```
## $beta.A
##  (Intercept) nevermarried        hsd08       hsd911          hsg
##   2.61778795  -0.20733103  -0.61716540  -0.37035190  -0.17223115
##           cg           ad           mw         exp1
##   0.35214787   0.61340856  -0.08269741   0.01163037
##
## $beta.B
##  (Intercept) nevermarried        hsd08       hsd911          hsg
##  2.366237186 -0.039686183 -0.720482860 -0.511183860 -0.190639825
##           cg           ad           mw         exp1
##  0.353628947  0.584711659 -0.077631410  0.008598067
##
## $beta.diff
##  (Intercept) nevermarried        hsd08       hsd911          hsg
##  0.251550769 -0.167644852  0.103317463  0.140831958  0.018408680
```

```
##          cg          ad          mw        exp1
## -0.001481078  0.028696905 -0.005065998  0.003032307
##
## $beta.R
##                 (Intercept) nevermarried      hsd08     hsd911        hsg
## [1,]  0.0000000    2.366237  -0.03968618 -0.7204829 -0.5111839 -0.1906398
## [2,]  1.0000000    2.617788  -0.20733103 -0.6171654 -0.3703519 -0.1722311
## [3,]  0.5000000    2.492013  -0.12350861 -0.6688241 -0.4407679 -0.1814355
## [4,]  0.5712428    2.509934  -0.13545209 -0.6614635 -0.4307346 -0.1801240
## [5,] -1.0000000    2.508277  -0.13872910 -0.5906664 -0.3608969 -0.1591640
## [6,] -2.0000000    2.636407  -0.13425668 -0.6454025 -0.4092403 -0.1809899
##             cg        ad          mw        exp1
## [1,] 0.3536289 0.5847117 -0.07763141 0.008598067
## [2,] 0.3521479 0.6134086 -0.08269741 0.011630374
## [3,] 0.3528884 0.5990601 -0.08016441 0.010114220
## [4,] 0.3527829 0.6011046 -0.08052533 0.010330250
## [5,] 0.3552537 0.5996944 -0.08687275 0.010205479
## [6,] 0.3535837 0.6044337 -0.08106076 0.010265653
```

Assignment 3

VICTORIA MILANI LEWASCHIW

CHAPTER 2 -

Ex 105: Let $X_1, \dots, X_n$ be iid from a population with the Lebesgue p.d.f. $f_\theta(x) = 2^{-1}(1+\theta x) I_{(-1,1)}(x)$, where $\theta \in (-1,1)$ is an unknown parameter. Find a consistent estimator of θ. Is your estimator $\sqrt{n}$-consistent?

$$E[X_i] = \int_{-1}^{1} x \cdot 2^{-1} \cdot (1+\theta x)\, dx = \frac{1}{2}\int_{-1}^{1} (x+\theta x^2)\, dx$$

$$\Rightarrow E[X_i] = \frac{1}{2}\left[\frac{x^2}{2} + \theta\frac{x^3}{3}\right]\Bigg|_{-1}^{1} = \frac{1}{2}\left[\frac{1}{2}+\frac{\theta}{3}-\frac{1}{2}+\frac{\theta}{3}\right] = \frac{1}{2}\cdot\frac{2\theta}{3} = \frac{\theta}{3}$$

so, by the law of Large numbers, $\bar{X} \xrightarrow{a.s.} E[X_i]$

let $E[X_i]=\mu$, by CMT $3\bar{X} \xrightarrow{a.s.} 3E[X_i] = \theta$, so $3\bar{X}$ is a strongly consistent estimator of θ. And since $3\bar{x}$ is unbiased, when we compute $Var(3\bar{x})$:

$$Var(3\bar{x}) = \frac{9}{n} Var(X_i) = \frac{3-\theta^2}{n}$$

$$\hookrightarrow Var(X_i) = E[X_i^2] - E[X_i]^2 = \frac{1}{2}\int_{-1}^{1}(x^2+\theta x^3)dx - \frac{\theta^2}{9} =$$

$$\Rightarrow Var(X_i) = \frac{1}{2}\left[\frac{x^3}{3}+\theta\frac{x^4}{4}\right]\Bigg|_{-1}^{1} - \frac{\theta^2}{9} = \frac{1}{2}\left[\frac{1}{3}+\frac{\theta}{4}+\frac{1}{3}-\frac{\theta}{4}\right] - \frac{\theta^2}{9} = \frac{3-\theta^2}{9}$$

and we have that $3\bar{x}$ is consistent in MSE.

By CLT, $\sqrt{n}\left(\bar{x}-\frac{\theta}{3}\right) \xrightarrow{d} N\left(0, \frac{3-\theta^2}{9}\right)$. and also $\sqrt{n}(3\bar{x}-\theta) \xrightarrow{d} N(0, 3-\theta^2)$,

implying that $\sqrt{n}(3\bar{x}-\theta) = O_p(1) \Rightarrow (3\bar{x}-\theta) = O_p(\sqrt{n})$.

$\hookrightarrow$ the estimator is $\sqrt{n}$-consistent

CHAPTER 3 -

Ex 301: Let $X_1, \dots, X_n$ be iid with mean μ, variance σ^2, and finite $\mu_j = E[X^j]$, $j = 2,3,4$. The sample coefficient of variation is defined to be $\frac{S}{\bar{X}}$, where S is the squared root of the sample variance S^2

a) if $\mu \neq 0$, show that $\sqrt{n}\left(\frac{S}{\bar{X}} - \frac{\sigma}{\mu}\right) \xrightarrow{d} N(0,\tau)$ and obtain an explicit formula of τ in terms of μ, σ^2 and μ_j.

by CLT,
$$\sqrt{n}\begin{pmatrix}\bar{X}-\mu \\ S-\sigma^2\end{pmatrix} \xrightarrow{d} N_2(0, \Sigma)$$

$$\hookrightarrow \Sigma = \begin{pmatrix}\sigma^2 & \sigma_3 \\ \sigma_3 & \sigma_4-\sigma^4\end{pmatrix}$$

considering $h(x_1,x_2) = \frac{\sqrt{x_2}}{x_1} \Rightarrow \nabla h(x_1,x_2) = \left(-\frac{\sqrt{x_2}}{x_1^2}, \frac{1}{2}\frac{1}{x_1\sqrt{x_2}}\right)$, and $\nabla(\mu,\sigma^2)^T = \left(\frac{-\sigma}{\mu^2}, \frac{1}{2\mu\sigma}\right)$

so, by the Delta method, $\sqrt{n}\left(\frac{S}{\bar{x}} - \frac{\sigma}{\mu}\right) \xrightarrow{d} N(0,\tau)$, where $\tau = \nabla(\mu,\sigma^2)^T \Sigma \nabla(\mu,\sigma^2)$

$$\tau = \underbrace{\begin{pmatrix} -\frac{\sigma}{\mu^2} & \frac{1}{2\mu\sigma} \end{pmatrix} \cdot \begin{pmatrix} \sigma^2 & \sigma_3 \\ \sigma_3 & \sigma_4 - \sigma^4 \end{pmatrix}}_{\begin{pmatrix} -\frac{\sigma^3}{\mu^2} + \frac{\sigma_3}{2\mu\sigma} & -\frac{\sigma\sigma_3}{\mu^2} + \frac{\sigma_4-\sigma^4}{2\mu\sigma} \end{pmatrix}} \begin{pmatrix} -\frac{\sigma}{\mu^2} \\ \frac{1}{2\mu\sigma} \end{pmatrix}$$

$$\tau = \frac{\sigma^4}{\mu^4} + \frac{\sigma_3}{2\mu^3} - \frac{\sigma^3}{2\mu^3} + \frac{\sigma_4 - \sigma^4}{4\mu^2} = \frac{\sigma^4}{\mu^4} + \frac{\sigma_4 - \sigma^4}{4\mu^2} \quad //$$

b) if $\mu = 0$, show that $n^{-1/2} \frac{S}{\bar{X}} \to_d [N(0,1)]^{-1}$.

again, by CLT

$\sqrt{n}(\bar{X} - 0) \to_d N(0, \sigma^2)$, and $S^2 \to_p \sigma^2 \Rightarrow S \to_p \sigma$, so by Slutsky Theorem

$\sqrt{n}\left(\frac{\bar{X}}{S}\right) \to_d \sigma^{-1} N(0, \sigma^2) = N(0,1)$

let $g(x) = x^{-1}$. Applying continuous mapping theorem,

$$\frac{1}{\sqrt{n}} g\left(\frac{\bar{X}}{S}\right) \to_d g(N(0,1)) \Rightarrow \frac{1}{\sqrt{n}}\left(\frac{S}{\bar{X}}\right) \to_d [N(0,1)]^{-1}$$

Ex 103: Let $X_1, \dots, X_n$ be iid from P in a parametric family. Obtain moment estimators of parameters in the following cases.

a) P is the gamma distribution $\Gamma(\alpha, \gamma)$, $\alpha > 0$, $\gamma > 0$.

$$\left.\begin{aligned} E[X_i] &= \mu_1 = \alpha\gamma \\ Var[X_i] &= E[X_i^2] - E[X_i]^2 = \mu_2 - \mu_1^2 = \alpha\gamma^2 \end{aligned}\right\} \quad \left.\begin{aligned} \hat{\mu}_1 &= \frac{\sum X_i}{n} \\ \hat{\mu}_2 &= \frac{\sum X_i^2}{n} \end{aligned}\right\} \text{SAMPLE MOMENTS}$$

by moments method,

$$\left.\begin{aligned} \hat{\mu}_1 &= \alpha\gamma \\ \hat{\mu}_2 &= \alpha\gamma^2 + \alpha^2\gamma^2 = \alpha\gamma^2(\alpha+1) \end{aligned}\right\} \frac{\hat{\mu}_2}{\hat{\mu}_1^2} = \frac{\alpha+1}{\alpha} \Rightarrow \alpha\left(\frac{\hat{\mu}_2}{\hat{\mu}_1^2} - 1\right) = 1$$

$$\hat{\alpha} = \left(\frac{\hat{\mu}_2}{\hat{\mu}_1^2} - 1\right)^{-1} = \frac{\hat{\mu}_1^2}{\hat{\mu}_2 - \hat{\mu}_1^2}$$

$$\hat{\gamma} = \frac{\hat{\mu}_2 - \hat{\mu}_1^2}{\hat{\mu}_1}$$

b) P is the exponencial distribution $E(a, \theta)$, $a \in \mathbb{R}$, $\theta > 0$

$$\left.\begin{aligned} E[X_i] &= \mu_1 = \theta + a \\ Var[X_i] &= E[X_i^2] - E[X_i]^2 = \mu_2 - \mu_1^2 = \theta^2 \end{aligned}\right\} \quad \left.\begin{aligned} \hat{\mu}_1 &= \frac{\sum X_i}{n} \\ \hat{\mu}_2 &= \frac{\sum X_i^2}{n} \end{aligned}\right\} \text{SAMPLE MOMENTS}$$

$\hookrightarrow \theta^2 + (\theta + a)^2$

by M.M.

$\hat{\mu}_1 = \theta + a$

$\hat{\mu}_2 = \theta^2 + \hat{\mu}_1^2 \to \hat{\theta} = \sqrt{\hat{\mu}_2 - \hat{\mu}_1^2}$

hence, $\hat{a} = \hat{\mu}_1 - \sqrt{\hat{\mu}_2 - \hat{\mu}_1^2}$

c) P is the beta distribution $B(\alpha, \beta)$, $\alpha>0$, $\beta>0$.

$$E[X_i] = \mu_1 = \frac{\alpha}{\alpha+\beta} \quad (*)$$

$$Var[X_i] = E[X_i^2] - E[X_i]^2 = \mu_2 - \mu_1^2 = \frac{\alpha\beta}{(\alpha+\beta+1)(\alpha+\beta)^2}$$

$$\mu_2 = \frac{\alpha(\alpha+1)}{(\alpha+\beta+1)(\alpha+\beta)}$$

$$\left.\begin{aligned} \hat{\mu}_1 &= \frac{\sum X_i}{n} \\ \hat{\mu}_2 &= \frac{\sum X_i^2}{n} \end{aligned}\right\} \text{SAMPLE MOMENTS}$$

$$(*)\ 1 + \frac{\beta}{\alpha} = \mu_1^{-1} \Rightarrow \mu_2 = \frac{\mu_1(1+\alpha^{-1})}{(\mu_1^{-1}+\alpha^{-1})}$$

by MM: $\hat{\mu}_1 = \dfrac{\alpha}{\alpha+\beta} \Rightarrow \alpha = \dfrac{\beta\hat{\mu}_1}{1-\hat{\mu}_1}$

$$\hat{\mu}_2 = \frac{\hat{\mu}_1(1+\alpha^{-1})}{(\hat{\mu}_1^{-1}+\alpha^{-1})}$$

$$\hat{\mu}_2 = \frac{\hat{\mu}_1\left(1+\frac{1-\hat{\mu}_1}{\beta\hat{\mu}_1}\right)}{\frac{1}{\hat{\mu}_1}+\frac{1-\hat{\mu}_1}{\beta\hat{\mu}_1}} = \frac{\frac{\beta\hat{\mu}_1+1-\hat{\mu}_1}{\cancel{\beta}}}{\frac{\beta+1-\hat{\mu}_1}{\cancel{\beta}\hat{\mu}_1}} = \frac{[\beta\hat{\mu}_1+1-\hat{\mu}_1]\hat{\mu}_1}{\beta+1-\hat{\mu}_1} = \hat{\mu}_2$$

$$(\beta+1-\hat{\mu}_1)\cdot\hat{\mu}_2 = \beta\hat{\mu}_1^2 + \hat{\mu}_1 - \hat{\mu}_1^2 \Rightarrow \beta\hat{\mu}_2 - \beta\hat{\mu}_1^2 = \hat{\mu}_1 - \hat{\mu}_1^2 + \hat{\mu}_2 - \hat{\mu}_1\hat{\mu}_2$$

$$\hat{\beta} = \frac{\hat{\mu}_1(1-\hat{\mu}_1) + \mu_2(1-\hat{\mu}_1)}{(\hat{\mu}_2-\hat{\mu}_1^2)} = \frac{(1-\hat{\mu}_1)(\hat{\mu}_1+\hat{\mu}_2)}{\hat{\mu}_2-\hat{\mu}_1^2}$$

hence, $\hat{\alpha} = \dfrac{\hat{\mu}_1(\hat{\mu}_1-\hat{\mu}_2)}{\hat{\mu}_2-\hat{\mu}_1^2}$

d) P is the log-normal distribution $LN(\mu, \sigma^2)$, $\mu \in \mathbb{R}$, $\sigma>0$

$$E[X_i] = e^{\mu+\sigma^2/2} = \mu_1$$

$$Var[X_i] = E[X_i^2] - E[X_i]^2 = \mu_2 - \mu_1^2 = e^{2\mu+\sigma^2}(e^{\sigma^2}-1)$$

$$\mu_2 = e^{2\mu+2\sigma^2}$$

$$\left.\begin{aligned} \hat{\mu}_1 &= \frac{\sum X_i}{n} \\ \hat{\mu}_2 &= \frac{\sum X_i^2}{n} \end{aligned}\right\} \text{SAMPLE MOMENTS}$$

by M.M:

$$\begin{aligned} \hat{\mu}_1 &= e^{\mu+\sigma^2/2} \\ \hat{\mu}_2 &= e^{2\mu+2\sigma^2} \end{aligned} \Rightarrow \begin{aligned} \hat{\mu}_1^2 &= e^{2\mu+\sigma^2} \\ \hat{\mu}_2 &= \hat{\mu}_1^2 e^{\sigma^2} \end{aligned} \Rightarrow \begin{aligned} \hat{\sigma} &= \sqrt{\ln\left(\frac{\hat{\mu}_2}{\hat{\mu}_1^2}\right)} \\ \hat{\mu} &= \ln\left(\frac{\hat{\mu}_1}{\sqrt{\hat{\mu}_2}}\right) \end{aligned}$$

Ex 507: Let $X_1, \dots, X_n$ be iid random variables having the Lebesgue pdf $f_{\alpha,\beta}(x) = \alpha\beta^{-\alpha}x^{\alpha-1}I_{(0,\beta)}(x)$, where $\alpha>0$ and $\beta>0$ are unknown.

a) Obtain moment estimators of α and β

$$E[X_i] = \int_0^\beta x\,\alpha\beta^{-\alpha}x^{\alpha-1}\,dx = \alpha\beta^{-\alpha}\int_0^\beta x^\alpha\,dx = \alpha\beta^{-\alpha}\left.\frac{x^{\alpha+1}}{\alpha+1}\right|_0^\beta = \frac{\alpha\beta^{-\alpha}\beta^{\alpha+1}}{\alpha+1} = \frac{\alpha\beta}{\alpha+1}$$

$$E[X_i^2] = \int_0^\beta x^2 \alpha \beta^{-\alpha} x^{\alpha-1} dx = \alpha\beta^{-\alpha}\int_0^\beta x^{\alpha+1} dx = \alpha\beta \frac{x^{\alpha+2}}{\alpha+2}\Big|_0^\beta = \frac{\alpha\beta^2}{\alpha+2}$$

* SAMPLE MOMENTS:

$$\hat{\mu}_1 = \frac{\sum X_i}{n} = \frac{\alpha\beta}{\alpha+1} \qquad \hat{\mu}_2 = \frac{\sum X_i^2}{n} = \frac{\alpha\beta^2}{\alpha+2}$$

$$\frac{\hat{\mu}_2}{\hat{\mu}_1^2} = \frac{(\alpha+1)^2}{\alpha(\alpha+2)} = \frac{\alpha^2+2\alpha+1}{\alpha^2+2\alpha} = 1 + \frac{1}{\alpha^2+2\alpha} \Rightarrow \alpha^2+2\alpha = \frac{\hat{\mu}_1^2}{\hat{\mu}_2-\hat{\mu}_1}$$

SOLVING THE QUADRATIC EXPRESSION, WE ARIVE IN $\alpha_1 > 0$ AND $\alpha_2 < 0$, BUT SINCE $\alpha > 0$, THEN

$$\hat{\alpha} = \frac{\hat{\mu}_1^2}{\hat{\mu}_2 - \hat{\mu}_1^2 + \sqrt{\hat{\mu}_2^2 - \hat{\mu}_1\hat{\mu}_2}}$$

$$\hat{\beta} = \frac{\hat{\mu}_2 + \sqrt{\hat{\mu}_2^2 - \hat{\mu}_1\hat{\mu}_2}}{\hat{\mu}_1}$$

b) Obtain the asymptotic distribution of the moment estimators of α and β derived in (a).

* Solving using delta method:

By CLT: $\sqrt{n}\begin{pmatrix}\hat{\mu}_1 - \mu_1 \\ \hat{\mu}_2 - \mu_2\end{pmatrix} \xrightarrow{d} N(0, \Sigma)$ $\qquad \Sigma = \begin{pmatrix}\mu_2 - \mu_1^2 & \mu_3 - \mu_1\mu_2 \\ \mu_3 - \mu_1\mu_2 & \mu_4 - \mu_2^2\end{pmatrix}$

setting $\alpha(x,y) = \dfrac{x^2}{y - x^2 + \sqrt{y^2 - xy}}$

$\beta(x,y) = \dfrac{y + \sqrt{y^2 - xy}}{x}$

then we can find $\dfrac{\partial(\alpha,\beta)}{\partial(x,y)}$, that, when evaluated at $x = \mu_1$, $y = \mu_2$, we have $\Lambda = \dfrac{\partial(\alpha,\beta)}{\partial(x,y)}\Big|_{\substack{x=\mu_1 \\ y=\mu_2}}$

So, by the delta method:

$$\sqrt{n}(\hat{\theta} - \theta) \xrightarrow{d} N(0, \Lambda\Sigma\Lambda^T)$$

where $\hat{\theta} = (\hat{\alpha}, \hat{\beta})$

$$* \frac{\partial(\alpha,\beta)}{\partial(x,y)} = \begin{pmatrix} \dfrac{2x}{y - x^2 + \sqrt{y^2 - xy}} + \dfrac{x^2\left(4x + \frac{y}{\sqrt{y^2-xy}}\right)}{2\left(y - x^2 + \sqrt{y^2-xy}\right)^2} & -\dfrac{x^2\left(1 + \frac{y - \frac{x}{2}}{\sqrt{y^2-xy}}\right)}{\left(y - x^2 + \sqrt{y^2-xy}\right)^2} \\ -\dfrac{y}{2x\sqrt{y^2-xy}} - \dfrac{y + \sqrt{y^2-xy}}{x^2} & \dfrac{1}{x} + \dfrac{2y - x}{2x\sqrt{y^2-xy}} \end{pmatrix}$$

CHAPTER 4-

Ex 96: Let $X_1, \dots, X_n$ be iid with pdf f_θ. Find an MLE of θ in each of the following cases.

b) $f_\theta(x) = e^{-(x-\theta)} I_{(\theta,\infty)}(x)$, $\theta > 0$

let $X_{(1)}$ be the smallest order statistic.

LIKELIHOOD FUNCTION: $\ell(\theta) = e^{-\Sigma(x_i-\theta)} I_{(0,X_{(1)})}(\theta)$ $\begin{cases} 0, \text{ when } \theta > X_{(1)} \\ \text{increasing on } (0, X_{(1)}) \end{cases}$ → maximum

MLE is $X_{(1)}$

d) $f_\theta(x) = \frac{\theta}{1-\theta} x^{(2\theta-1)/(1-\theta)} I_{(0,1)}(x)$, $\theta \in (\frac{1}{2}, 1)$

$$\frac{\partial \log \ell(\theta)}{\partial \theta} = 0 \Rightarrow \frac{n}{\theta(1-\theta)} + \frac{1}{(1-\theta)^2} \sum_i \log X_i = 0$$

$$\hat{\theta} = \left(1 - \frac{\Sigma \log X_i}{n}\right)^{-1}$$

if $\theta > \hat{\theta} \rightarrow \frac{\partial \log \ell(\theta)}{\partial \theta} < 0$

if $\theta < \hat{\theta} \rightarrow \frac{\partial \log \ell(\theta)}{\partial \theta} > 0$

⇒ Knowing that, the MLE of θ is $\max\{\hat{\theta}, \frac{1}{2}\}$

f) $f_\theta(x) = \theta x^{-2} I_{(\theta,\infty)}(x)$, $\theta > 0$

$\ell(\theta) = \theta^n \prod_i X_i^{-2} I_{(0,X_{(1)})}(\theta)$, where $X_{(1)}$ is the smaller order statistic.

since $\ell(\theta) = 0$ if $\theta > X_{(1)}$ and is increasing on $(0, X_{(1)})$, the MLE of θ is $X_{(1)}$

h) $f_{(\theta)}(x)$ is the pdf of $N(\theta, \theta^2)$, $\theta \in \mathbb{R}$, $\theta \neq 0$

$$\frac{\partial \log \ell(\theta)}{\partial \theta} = -\frac{1}{\theta^3}\left(n\theta^2 + \theta \Sigma X_i - \Sigma X_i^2\right)$$

when we set $\frac{\partial \log \ell(\theta)}{\partial \theta}$, we can find two solutions

$$\theta_{1,2} = \frac{-\Sigma X_i \pm \sqrt{(\Sigma X_i)^2 + 4n \Sigma X_i^2}}{2n}$$

$\theta_1 = \dots + \dots$
$\theta_2 = \dots - \dots$

both are maximum points.

So, the MLE of θ is

$$\hat{\theta} = \begin{cases} \theta_1 & \text{if } \ell(\theta_1) < \ell(\theta_2) \\ \theta_2 & \text{if } \ell(\theta_1) \geq \ell(\theta_2) \end{cases}$$

j) $f_\theta(x)$ is the pdf of the log normal distribution $LN(\mu, \sigma^2)$, $\theta = (\mu, \sigma^2) \in \mathbb{R} \times (0, \infty)$

Let $Y_i = \log X_i$, $i = 1, \dots, n$

then $\ell(\theta) = \frac{1}{(\sigma\sqrt{2\pi})^n} e^{\wedge}\left\{-\frac{1}{2\sigma^2} \Sigma(Y_i - \mu)^2 - \Sigma Y_i\right\}$

for $\frac{\partial \log \ell(\theta)}{\partial \theta} = 0$ we obtain the MLE for μ and σ:

- μ: $\bar{Y} = \frac{\Sigma Y_i}{n}$

$$- \sigma^2 : \frac{\sum (y_i - \bar{y})^2}{n}$$

l) $f_\theta(x) = \beta^{-\alpha} \alpha x^{\alpha-1} I_{(0,\beta)}(x)$, $\theta = (\alpha, \beta) \in (0,\infty) \times (0,\infty)$

$\ell(\theta) = \alpha^n \beta^{-n\alpha} \prod X_i^{\alpha-1} I_{(X_{(n)},\infty)}(\beta)$

$$\ell(\theta) = \begin{cases} 0 & \text{when } \beta < X_{(n)} \\ \text{decreasing} & \text{o.w} \end{cases} \Rightarrow \text{MLE of } \beta \text{ is } X_{(n)}$$

MLE of α is $n \left[\sum \log \left(\frac{X_{(n)}}{X_i} \right) \right]^{-1}$

found by substituting $\beta = X_{(n)}$ into $\ell(\theta)$

n) $f_\theta(x) = \frac{1}{2}(1-\theta^2) e^{\theta x - |x|}$, $\theta \in (-1, 1)$

with $\bar{X} = \frac{\sum X_i}{n}$:

$$\frac{\partial \log \ell(\theta)}{\partial \theta} = n\bar{X} - \frac{2n\theta}{1-\theta^2} = 0 \Rightarrow \theta_{1,2} = -1 \pm \sqrt{1 + \bar{X}^2}$$

cannot be a solution

so the MLE of θ is $\sqrt{1+\bar{X}} - 1$

Ex 144: Let $X_1, \dots, X_n$ be iid bivariate normal random vectors with mean 0 and an unknown covariance matrix whose diagonal elements are σ_1^2 and σ_2^2 and off-diagonal element is $\rho\sigma_1\sigma_2$. Let $\theta = (\sigma_1^2, \sigma_2^2, \rho)$.
Obtain $I_n(\theta)$ and $[I_n(\theta)]^{-1}$ and derive a non degenerated asymptotic distribution of the MLE of θ.

$$\ell(\theta) = \left(2\pi\sqrt{\sigma_1^2\sigma_2^2(1-\rho)}\right)^{-n} \exp\left\{ -\frac{1}{2(1-\rho^2)} \left(\frac{\sum x_{1i}^2}{\sigma_1^2} + \frac{\sum x_{2i}^2}{\sigma_2^2} - \frac{2\rho \sum x_{1i} x_{2i}}{\sigma_1\sigma_2} \right) \right\}$$

if considering $\mu_1 = \mu_2 = 0$, we can reach a log-likelihood:

$$\log \ell(\theta) = -n\log(2\pi) - \frac{n}{2}\left[\log(\sigma_1^2) + \log(\sigma_2^2) + \log(1-\rho^2)\right] - \frac{1}{2(1-\rho^2)} \cdot \left(\frac{\sum x_{1i}^2}{\sigma_1^2} + \frac{\sum x_{2i}^2}{\sigma_2^2} - \frac{2\rho\sum x_{1i} x_{2i}}{\sigma_1\sigma_2} \right)$$

and to find the score: $\frac{\partial \log \ell(\theta)}{\partial \theta}$

$$\frac{\partial \log \ell(\theta)}{\partial \sigma_1^2} = \frac{-n}{2\sigma_1^2} + \frac{1}{2(1-\rho^2)} \left(\frac{\sum x_{1i}^2}{(\sigma_1^2)^2} + \frac{\cancel{2}\rho \sum x_{1i} x_{2i} \frac{1}{\cancel{2}} \sigma_1^2 \sigma_2}{\sigma_1^2 \sigma_2^2} \right) = 0$$

$$\frac{\partial \log \ell(\theta)}{\partial \sigma_2^2} = \frac{-n}{2\sigma_2^2} + \frac{1}{2(1-\rho^2)} \left(\frac{\sum x_{2i}^2}{(\sigma_2^2)^2} + \frac{\cancel{2}\rho \sum x_{1i} x_{2i} \frac{1}{\cancel{2}} \sigma_1 \sigma_2^2}{\sigma_1^2 \sigma_2^2} \right) = 0$$

$$\frac{\partial \log \ell(\theta)}{\partial \rho} = \frac{2n\rho}{2(1-\rho^2)} - \left(\frac{-2\sum x_{1i} x_{2i}}{\sigma_1^2 \sigma_2^2} \cdot \frac{1}{2(1-\rho^2)} - \left(\frac{\sum x_{1i}^2}{\sigma_1^2} + \frac{\sum x_{2i}^2}{\sigma_2^2} - \frac{2\rho \sum x_{1i} x_{2i}}{\sigma_1\sigma_2} \right) \frac{-4\rho}{(2(1-\rho^2))^2} \right)$$

(*)

from (*) it is possible to find the MLE estimators for θ. And we know they are consistent, so $\hat{\theta} \xrightarrow{p} \theta$

By CLT, since data is iid and $EX^4 < \infty$, $\sqrt{n}(\hat{\theta} - \theta) \xrightarrow{d} N(0, [I(\theta)]^{-1})$

So, $\hat{\theta}$ is an optimal estimator, since the asymptotic variance is the smallest one to obtain $I(\theta)$:

$$I = \text{Var}(\text{score}) = -E\left[\frac{\partial S\theta}{\partial\theta}\right] \Rightarrow I_n = -E_n\left[\frac{\partial S\theta}{\partial\theta}\right] = -E_n\left[\nabla_{\theta\theta} \log \ell(\theta)\right]$$

and $[I_n(\theta)]^{-1} = [-E_n\,[\nabla_{\theta\theta} \log \ell(\theta)]]^{-1} \Rightarrow [I(\theta)]^{-1} = [-E\,[\nabla_{\theta\theta} \log \ell(\theta)]]^{-1}$