STAT 8003 FINAL EXAM, FALL 2013
SOLUTION

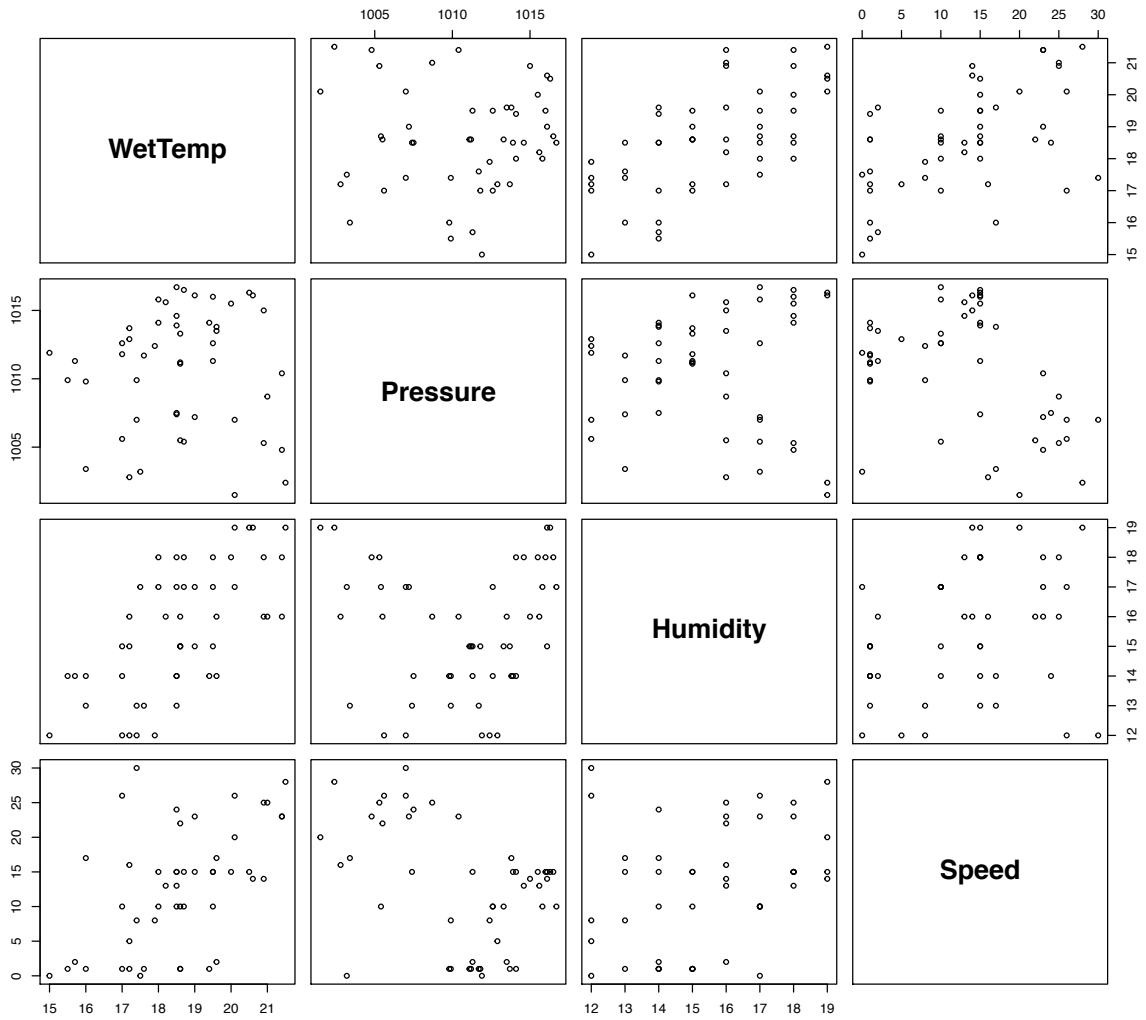
Problem 1. Scientists measured the weather and wind data at 3 hour intervals for Gabo Island, off the eastern tip of Victoria, during 1989. Observations were made 7 times a day for 7 days. Thus, in total there are 49 observations.

Variable	Description
WetTemp	Wet bulb temperature
Pressure	Barometer pressure
Humidity	Relative humidity in percent
Speed	Wind speed in knots

Now suppose the outcome is $\mathbf{Y} = (Y_1, \ldots, Y_n)^{\mathrm{T}}$, where Y_i is the wet bulb temperature of the ith observation. Suppose $\mathbf{X}_1$, $\mathbf{X}_2$ and $\mathbf{X}_3$ are the observations for pressure, humidity and speed, respectively.

a) Plot a scatterplots for the data. Describe the pattern you observe. Fit a linear model for the data. Interprete your results.

Thanks to Hsiang-Chieh Yang for the solution!



From the plot matrix, I observed that

1. There is a positive correlation between WetTemp and Humidity.
2. There is a positive correlation between WetTemp and Speed.
3. The other covariates seem to be less collinear than the variables mentioned above.

Fit a linear model for the data.

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3} + \epsilon_i, \quad i = 1, ..., 49$$

Where

$$\begin{aligned} Y_i &= \text{WetTemp: Wet bulb temperature} \\ X_{i1} &= \text{Pressure: Barometer pressure} \\ X_{i2} &= \text{Humidity: Relative humidity in percent} \\ X_{i3} &= \text{Speed: Wind speed in knots} \\ \beta_k &= \text{the effect on WetTemp of each covariate, } k = 1, 2, 3 \end{aligned}$$

Let

$$\mathbf{Y} = \begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_{49} \end{pmatrix}, \mathbf{X} = \begin{pmatrix} 1 & X_{1,1} & X_{1,2} & X_{1,3} \\ 1 & X_{2,1} & X_{2,2} & X_{2,3} \\ \vdots & \vdots & \vdots & \\ 1 & X_{49,1} & X_{49,2} & X_{49,3} \end{pmatrix}, \boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_3 \end{pmatrix}, \boldsymbol{\epsilon} = \begin{pmatrix} \epsilon_0 \\ \epsilon_1 \\ \vdots \\ \epsilon_{49} \end{pmatrix}$$

The model can written in matrix form:

$$\begin{aligned} &\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}, \\ &\text{where } E(\boldsymbol{\epsilon}) = 0 \text{ and } \text{Var}(\boldsymbol{\epsilon}) = \boldsymbol{\sigma}^2\mathbf{I} \end{aligned}$$

The LSE

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}$$

Variance of LSE

$$\begin{aligned} \hat{\boldsymbol{\sigma}}^2 &= \frac{(\mathbf{Y} - \hat{\mathbf{Y}})^\intercal(\mathbf{Y} - \hat{\mathbf{Y}})}{n - (p+1)} \\ &= \frac{\boldsymbol{\epsilon}^\intercal\boldsymbol{\epsilon}}{49 - 4} = \frac{RSS}{45} \end{aligned}$$

Thus,

$$\text{Var}(\hat{\boldsymbol{\beta}}) = \hat{\boldsymbol{\sigma}}^2(\mathbf{X}^\intercal\mathbf{X})^{-1}$$

By using lm function, we can get the results as follow:

```
##
## Call:
## lm(formula = weather$WetTemp ~ weather$Pressure + weather$Humidity +
##     weather$Speed)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.9369 -0.8866  0.0796  0.9089  2.1980
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)      -53.2058    39.3342   -1.35  0.18292
## weather$Pressure   0.0640     0.0390    1.64  0.10782
## weather$Humidity   0.3889     0.0774    5.02  8.6e-06 ***
## weather$Speed      0.0842     0.0204    4.14  0.00015 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.07 on 45 degrees of freedom
## Multiple R-squared:  0.583,Adjusted R-squared:  0.555
## F-statistic:   21 on 3 and 45 DF,  p-value: 1.2e-08
```

From the results shown above, we can see that Wet bulb temperature is significantly correlated to relative humidity and wind speed. The p-value of their coefficients are nearly zero.On average, the one percent increase of relative humidity can lead wet bulb temperature increase by 0.3889 degrees, and the wind speed can increase 0.0842 degrees of the temperature for every 1 knot.On the other hand, the Barometer pressure is not linearly correlated with wet bulb temperature significantly. From the previous part of plot matrix analysis, we can also deduce that the effects of Humidity and Speed still exist when they are adjusted by pressure. In sum, the F-statistics is significant with a nearly zero p-value, implying that the 3 covariates in whole have significant effect on the temperature.

Problem 2. Consider the same example in Problem 1. Since the observation are recorded subsequently, it is resonable to assume that the unmeasured the effects of the temperature are correlated. We consider the following model:

$$Y_k = \beta_0 + \beta_1 X_{k1} + \beta_2 X_{k2} + \beta_3 X_{k3} + \epsilon_k. \tag{1}$$

Here $\epsilon_k \sim \mathrm{N}(0, \sigma^2)$, but they are not independent from each other. We assume that $\boldsymbol{\epsilon} = (\epsilon_1, \ldots, \epsilon_n) \sim \mathrm{N}(\mathbf{0}, \boldsymbol{\Sigma})$, with $\boldsymbol{\Sigma}$ known.

a) When the error terms are correlated, we cannot use ordinary least squares to find the solution. But instead we can use a modified version. Define the new RSS function:

$$RSS = (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})^{\mathrm{T}} \boldsymbol{\Sigma}^{-1} (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}), \tag{2}$$

with $\mathbf{X}$ is the design matrix with the intercept. The general least square estimator $\hat{\boldsymbol{\beta}}$ is obtained by minimizing (2). Please show that

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^{\mathrm{T}} \boldsymbol{\Sigma}^{-1} \mathbf{X})^{-1} (\mathbf{X}^{\mathrm{T}} \boldsymbol{\Sigma}^{-1} \mathbf{Y}).$$

Also show $\mathtt{E}(\hat{\boldsymbol{\beta}})$ and $\mathtt{Var}(\hat{\boldsymbol{\beta}})$.

b) Suppose a statistician ignores the effect of $\boldsymbol{\Sigma}$, so that under model (1) he uses the ordinary least square estimators $\tilde{\boldsymbol{\beta}} = (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}(\mathbf{X}^{\mathrm{T}}\mathbf{Y})$ to estimate $\boldsymbol{\beta}$. For any given vector $\mathbf{a}$, suppose we are interested in estimating $\mathbf{a}^{\mathrm{T}}\boldsymbol{\beta}$. Now the statistician uses $\mathbf{a}^{\mathrm{T}}\tilde{\boldsymbol{\beta}}$ to estimate $\mathbf{a}^{\mathrm{T}}\boldsymbol{\beta}$. Is the estimator unbiased? And what's the variance? If you are the statistician, will you use $\mathbf{a}^{\mathrm{T}}\tilde{\boldsymbol{\beta}}$ or $\mathbf{a}^{\mathrm{T}}\hat{\boldsymbol{\beta}}$ to estimate $\mathbf{a}^{\mathrm{T}}\boldsymbol{\beta}$? Why? (Hint: Let $A = \boldsymbol{\Sigma}^{-1/2}$. Consider a equivalent linear model $\tilde{\mathbf{Y}} = A\mathbf{Y}$ and $\tilde{\mathbf{X}} = A\mathbf{X}$, and obtain its LSE.)

Thanks Bruce Scott for his solution!

Problem 2. Consider the same example in Problem 1. Since the observations are recorded subsequently, it is reasonable to assume the unmeasured effects of temperature are correlated. We consider the following model:

$$Y_k = \beta_0 + \beta_1 X_{k1} + \beta_2 X_{k2} + \beta_3 X_{k3} + \epsilon_k \tag{1}$$

Here $\epsilon_k \sim N(0, \sigma^2)$, but they are not independent from each other. We assume that $\boldsymbol{\epsilon} = (\epsilon_1, \ldots, \epsilon_n) \sim N(\mathbf{0}, \boldsymbol{\Sigma})$, with $\boldsymbol{\Sigma}$ known.

a) When the error terms are correlated, we cannot use ordinary least squares to find the solution. But instead we can use a modified version. Define the new RSS function:

$$RSS = (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})^T \boldsymbol{\Sigma}^{-1} (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}) \tag{2}$$

with $\mathbf{X}$ is the design matrix with the intercept. The general least square estimator $\hat{\beta}$ is obtained by minimizing (2). Please show that

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^T \boldsymbol{\Sigma}^{-1} \mathbf{X})^{-1} (\mathbf{X}^T \boldsymbol{\Sigma}^{-1} \mathbf{Y})$$

Also find E($\hat{\boldsymbol{\beta}}$) and Var($\hat{\boldsymbol{\beta}}$).

In order to find the estimator of β that minimizes the new RSS function, we can take the derivative with respect to β, set it equal to 0, and solve for β:

$$\begin{aligned}
&\frac{\partial RSS}{\partial \boldsymbol{\beta}} = (-\mathbf{X}^T)(2\boldsymbol{\Sigma}^{-1}(\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}})) = 0 \\
&\rightarrow \mathbf{X}^T \boldsymbol{\Sigma}^{-1} \mathbf{X} \hat{\boldsymbol{\beta}} = \mathbf{X}^T \boldsymbol{\Sigma}^{-1} \mathbf{Y} \\
&\rightarrow \hat{\boldsymbol{\beta}} = (\mathbf{X}^T \boldsymbol{\Sigma}^{-1} \mathbf{X})^{-1} \mathbf{X}^T \boldsymbol{\Sigma}^{-1} \mathbf{Y}
\end{aligned}$$

This follows as $\boldsymbol{\Sigma}$ is a positive definite matrix assuming linear independence among independent variables and therefore $\boldsymbol{\Sigma}^{-1}$ is also positive definite. Also rank($\mathbf{X}$) = $p \leq n$ where p is the number of parameters in the linear model, so $\mathbf{X}^T \boldsymbol{\Sigma}^{-1} \mathbf{X}$ is positive definite and thus invertible. However, we must check the second derivative for positive definiteness to ensure the result minimizes the new RSS function:

$$\frac{\partial^2 RSS}{\partial \boldsymbol{\beta}^2} = 2\mathbf{X}^T \boldsymbol{\Sigma}^{-1} \mathbf{X}$$

Since we know $\mathbf{X}^T \boldsymbol{\Sigma}^{-1} \mathbf{X}$ is positive definite (see previous paragraph), then the second derivative is positive definite and therefore $\hat{\boldsymbol{\beta}} = (\mathbf{X}^T \boldsymbol{\Sigma}^{-1} \mathbf{X})^{-1} (\mathbf{X}^T \boldsymbol{\Sigma}^{-1} \mathbf{Y})$ minimizes the new RSS function accordingly.

The expectation and variance of $\hat{\boldsymbol{\beta}}$ are as follows:

$$\begin{aligned}
\mathrm{E}[\hat{\boldsymbol{\beta}}] &= \mathrm{E}[(\mathbf{X}^T\boldsymbol{\Sigma}^{-1}\mathbf{X})^{-1}(\mathbf{X}^T\boldsymbol{\Sigma}^{-1}\mathbf{Y})] \\
&= (\mathbf{X}^T\boldsymbol{\Sigma}^{-1}\mathbf{X})^{-1}(\mathbf{X}^T\boldsymbol{\Sigma}^{-1}\mathrm{E}[\mathbf{Y}]) \\
&= (\mathbf{X}^T\boldsymbol{\Sigma}^{-1}\mathbf{X})^{-1}(\mathbf{X}^T\boldsymbol{\Sigma}^{-1}\mathbf{X})\boldsymbol{\beta} \\
&= \boldsymbol{\beta} \\
\mathrm{Var}[\hat{\boldsymbol{\beta}}] &= \mathrm{Var}[(\mathbf{X}^T\boldsymbol{\Sigma}^{-1}\mathbf{X})^{-1}(\mathbf{X}^T\boldsymbol{\Sigma}^{-1}\mathbf{Y})] \\
&= (\mathbf{X}^T\boldsymbol{\Sigma}^{-1}\mathbf{X})^{-1}(\mathbf{X}^T\boldsymbol{\Sigma}^{-1})\mathrm{Var}[\mathbf{Y}][(\mathbf{X}^T\boldsymbol{\Sigma}^{-1}\mathbf{X})^{-1}(\mathbf{X}^T\boldsymbol{\Sigma}^{-1})]^T \\
&= (\mathbf{X}^T\boldsymbol{\Sigma}^{-1}\mathbf{X})^{-1}\mathbf{X}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\Sigma}\boldsymbol{\Sigma}^{-1}\mathbf{X}(\mathbf{X}^T\boldsymbol{\Sigma}^{-1}\mathbf{X})^{-1} \\
&= (\mathbf{X}^T\boldsymbol{\Sigma}^{-1}\mathbf{X})^{-1}\mathbf{X}^T\boldsymbol{\Sigma}^{-1}\mathbf{X}(\mathbf{X}^T\boldsymbol{\Sigma}^{-1}\mathbf{X})^{-1} \\
&= (\mathbf{X}^T\boldsymbol{\Sigma}^{-1}\mathbf{X})^{-1}
\end{aligned}$$

This follows since $\mathrm{E}[\mathbf{Y}] = \mathrm{E}[\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}] = \mathbf{X}\boldsymbol{\beta} + \mathrm{E}[\boldsymbol{\epsilon}] = \mathbf{X}\boldsymbol{\beta}$ as the expectation of the error term is still $\mathbf{0}$, and $\mathrm{Var}[\mathbf{Y}] = \boldsymbol{\Sigma}$.

b) Suppose a statistician ignores the effect of $\boldsymbol{\Sigma}$, so that under model (1) he uses the ordinary least square estimators $\tilde{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X})^{-1}(\mathbf{X}^T\mathbf{Y})$ to estimate $\boldsymbol{\beta}$. For any given vector $\mathbf{a}$, suppose we are interested in estimating $\mathbf{a}^T\boldsymbol{\beta}$. Now the statistician uses $\mathbf{a}^T\tilde{\boldsymbol{\beta}}$ to estimate $\mathbf{a}^T\boldsymbol{\beta}$. Is the estimator unbiased? And what's the variance? If you are the statistician, will you use $\mathbf{a}^T\tilde{\boldsymbol{\beta}}$ or $\mathbf{a}^T\hat{\boldsymbol{\beta}}$ to estimate $\mathbf{a}^T\boldsymbol{\beta}$? Why? (Hint: Let $A = \boldsymbol{\Sigma}^{-1/2}$. Consider an equivalent linear model $\tilde{\mathbf{Y}} = A\mathbf{Y}$ and $\tilde{\mathbf{X}} = A\mathbf{X}$, and obtain its LSE.)

Here is the expectation and variance of $\mathbf{a}^T\tilde{\boldsymbol{\beta}}$ and $\mathbf{a}^T\hat{\boldsymbol{\beta}}$ for comparison:

$$\begin{aligned}
\mathrm{E}[\mathbf{a}^T\tilde{\boldsymbol{\beta}}] &= \mathbf{a}^T\mathrm{E}[\tilde{\boldsymbol{\beta}}] \\
&= \mathbf{a}^T\mathrm{E}[(\mathbf{X}^T\mathbf{X})^{-1}(\mathbf{X}^T\mathbf{Y})] \\
&= \mathbf{a}^T(\mathbf{X}^T\mathbf{X})^{-1}(\mathbf{X}^T\mathrm{E}[\mathbf{Y}]) \\
&= \mathbf{a}^T(\mathbf{X}^T\mathbf{X})^{-1}(\mathbf{X}^T\mathbf{X})\boldsymbol{\beta} \\
&= \mathbf{a}^T\boldsymbol{\beta} \\
\mathrm{Var}[\mathbf{a}^T\tilde{\boldsymbol{\beta}}] &= \mathbf{a}^T\mathrm{Var}[\tilde{\boldsymbol{\beta}}]\mathbf{a} \\
&= \mathbf{a}^T\mathrm{Var}[(\mathbf{X}^T\mathbf{X})^{-1}(\mathbf{X}^T\mathbf{Y})]\mathbf{a} \\
&= \mathbf{a}^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathrm{Var}[\mathbf{Y}]\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{a} \\
&= \mathbf{a}^T(\mathbf{X}^T\mathbf{X})^{-1}(\mathbf{X}^T\boldsymbol{\Sigma}\mathbf{X})(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{a} \\
\mathrm{E}[\mathbf{a}^T\hat{\boldsymbol{\beta}}] &= \mathbf{a}^T\mathrm{E}[\hat{\boldsymbol{\beta}}] = \mathbf{a}^T\boldsymbol{\beta} \\
\mathrm{Var}[\mathbf{a}^T\hat{\boldsymbol{\beta}}] &= \mathbf{a}^T\mathrm{Var}[\hat{\boldsymbol{\beta}}]\mathbf{a} \\
&= \mathbf{a}^T(\mathbf{X}^T\boldsymbol{\Sigma}^{-1}\mathbf{X})^{-1}\mathbf{a}
\end{aligned}$$

You can see that both estimators are unbiased, so we need to find a way to determine which has smaller variance (not immediately clear in comparing the variance terms above). Following the hint, if we take our original linear model and multiply it by $\mathbf{\Sigma}^{-1/2}$, we get an equivalent linear model:

$$\mathbf{\Sigma}^{-1/2}\mathbf{Y} = \mathbf{\Sigma}^{-1/2}\mathbf{X}\boldsymbol{\beta} + \mathbf{\Sigma}^{-1/2}\boldsymbol{\epsilon}$$
$$\tilde{\mathbf{Y}} = \tilde{\mathbf{X}}\boldsymbol{\beta} + \tilde{\boldsymbol{\epsilon}}$$

What's special about this linear model is the distribution of the error terms $\tilde{\boldsymbol{\epsilon}}$. Since $\mathbf{\Sigma}^{-1/2}$ is constant and known, $\tilde{\boldsymbol{\epsilon}}$ is also multivariate normal with mean and variance:

$$\mathrm{E}[\tilde{\boldsymbol{\epsilon}}] = \mathbf{\Sigma}^{-1/2}\mathrm{E}[\boldsymbol{\epsilon}] = \mathbf{0}$$
$$\mathrm{Var}[\tilde{\boldsymbol{\epsilon}}] = \mathbf{\Sigma}^{-1/2}\mathrm{Var}[\boldsymbol{\epsilon}]\mathbf{\Sigma}^{-1/2} = \mathbf{\Sigma}^{-1/2}\mathbf{\Sigma}^{1/2}\mathbf{\Sigma}^{1/2}\mathbf{\Sigma}^{-1/2} = \mathbf{I}$$

The error terms in this equivalent linear model are independent and normally distributed with a constant variance of 1. In this setting, the estimator for $\mathbf{a}^T\boldsymbol{\beta}$ using the ordinary least squares estimator ($\mathbf{a}^T\hat{\boldsymbol{\beta}}_{OLS}$) has a variance that is less than or equal to the variance of all other unbiased linear estimators according to the Gauss-Markov Theorem. Therefore, the LSE for the equivalent linear model will have minimum variance in estimating $\mathbf{a}^T\boldsymbol{\beta}$.

LSE estimate for $\tilde{\mathbf{Y}} = \tilde{\mathbf{X}}\boldsymbol{\beta} + \tilde{\boldsymbol{\epsilon}}$ can be found as follows:

$$\rightarrow \boldsymbol{\beta}_{OLS} = (\tilde{\mathbf{X}}^T\tilde{\mathbf{X}})^{-1}\tilde{\mathbf{X}}^T\tilde{\mathbf{Y}}$$
$$\rightarrow \boldsymbol{\beta}_{OLS} = (\mathbf{X}^T\mathbf{\Sigma}^{-1/2}\mathbf{\Sigma}^{-1/2}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{\Sigma}^{-1/2}\mathbf{\Sigma}^{-1/2}\mathbf{Y}$$
$$\rightarrow \boldsymbol{\beta}_{OLS} = (\mathbf{X}^T\mathbf{\Sigma}^{-1}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{\Sigma}^{-1}\mathbf{Y}$$
$$\rightarrow \boldsymbol{\beta}_{OLS} = \hat{\boldsymbol{\beta}}$$

Therefore, I would choose $\mathbf{a}^T\hat{\boldsymbol{\beta}}$ to estimate $\mathbf{a}^T\boldsymbol{\beta}$ since it has the minimum variance by the Gauss-Markov theorem.

Problem 3. In many studies, a large number of independent variables, $X_1, \ldots, X_p$, are measured. However, it may be impractical to include all these variables in a linear regression model. One way to reduce the dimensionality of the model is via principal components; a linear combination of the variables. The ith principal component, Z_i, is given by

$$Z_i = \mathbf{a}_i^{\mathrm{T}}\mathbf{X} = a_{i1}X_1 + \cdots + a_{ip}X_p$$

such that

$$\mathbf{a}_i^{\mathrm{T}}\mathbf{X} \text{ maximizes } \mathtt{Var}(\mathbf{a}_i^{\mathrm{T}}\mathbf{X})$$
$$\text{subject to } \mathbf{a}_i^{\mathrm{T}}\mathbf{a}_i = 1, \text{ and } \mathtt{Cov}(Z_i, Z_k) = 0, \text{ for } k \neq i.$$

Since principal component analysis forcusses on maximizing the variance of the independent variables, the theorems for matrices are useful for understanding the properties of the principal components. Once such theorem is the maximization for quadratic forms. That is, given a positive definite matrix $\mathbf{B}$

$$\max_{\mathbf{x}\neq 0,\ \mathbf{x}\perp\mathbf{e}_1,\ldots,\mathbf{e}_{k-1}} \frac{\mathbf{x}^{\mathrm{T}}\mathbf{B}\mathbf{x}}{\mathbf{x}^{\mathrm{T}}\mathbf{x}} = \lambda_k, \quad \text{and the maximization achieves when } \mathbf{x} = \mathbf{e}_k$$

where $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p \geq 0$ are the eigenvalues and $\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_p$ are the associated normalized eigenvectors of $\mathbf{B}$.

a) Using the maximization theorem for quadratic forms mentioned above, or otherwise, show that (i) $Z_i = \mathbf{e}_i^{\mathrm{T}}\mathbf{X}$, (ii) $\mathtt{Var}(Z_i) = \lambda_i$, (iii) $\sum_{k=1}^{p} \mathtt{Var}(X_k) = \sum_{i=1}^{p} \mathtt{Var}(Z_i)$, where e_i is the ith eigenvector of $\mathtt{Var}(X)$, corresponding to the ith largest eigenvalues of $\mathtt{Var}(X)$.

b) In practice, the linear model

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p + \epsilon,$$

can be replaced with

$$Y = \alpha_0 + \alpha_1 Z_1 + \cdots + \alpha_k Z_k + \epsilon, \quad k \leq p.$$

Explain how you would determine k.

Thanks Jianyun Zhu for the solution!

a)(i)

The ith principle component Z_i is given by

$$Z_i = \mathbf{a}_i^T\mathbf{X} = a_{i1}X_1 + a_{i2}X_2 + \cdots + a_{ip}X_p, \qquad i = 1, 2, \cdots, p$$

such that

$$\mathbf{a}_i^T\mathbf{X} \text{ maximize } \mathrm{Var}(\mathbf{a}_i^T\mathbf{X})$$
$$\text{subject to } \mathbf{a}_i^T\mathbf{a}_i = 1 \text{ and } \mathrm{Cov}(Z_i, Z_k) = 0 \text{ for } k \neq i$$

Let $\mathbf{\Sigma}$ be the covariance matrix associate with the random $\mathbf{X}^T = [X_1, X_2, \cdots, X_p]$.Then $\mathrm{Var}(\mathbf{a}_i^T\mathbf{X}) = \mathbf{a}_i^T\mathrm{Var}(\mathbf{X})\,\mathbf{a}_i = \mathbf{a}_i^T\mathbf{\Sigma}\,\mathbf{a}_i$

To find out Z_i, we need to find $\mathbf{a}_i^T\mathbf{X}$ maximize $\mathbf{a}_i^T\mathbf{\Sigma}\,\mathbf{a}_i$

$$\max \mathbf{a}_i^T\mathbf{\Sigma}\mathbf{a}_i = \max \frac{\mathbf{a}_i^T\mathbf{\Sigma}\mathbf{a}_i}{\mathbf{a}_i^T\mathbf{a}_i}$$

Using the maximization theorem for quadratic form, we get

$$\max_{\mathbf{a}_i \neq 0, \mathbf{a}_i \perp \mathbf{e}_1, \mathbf{e}_2, \cdots \mathbf{e}_{k-1}} \frac{\mathbf{a}_i^T \mathbf{\Sigma} \mathbf{a}_i}{\mathbf{a}_i^T \mathbf{a}_i} = \lambda_i$$

and the maximization achieves when $\mathbf{a}_i = \mathbf{e}_i$ where $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p \geq 0$ are ordered eigenvalues of $\mathbf{\Sigma}$ and $e_1, e_2, \cdots e_p$ are the associated normalized eigenvectors of $\mathbf{\Sigma}$.
Clearly $\mathbf{e}_i$ satisfied the condition for principle component Z_i, that is $Z_i = \mathbf{e}_i^T \mathbf{X}$.

(ii)
$\mathrm{Var}(Z_i) = \mathrm{Var}(\mathbf{e}_i^T \mathbf{X}) = \mathbf{e}_i^T \mathrm{Var}(\mathbf{X}) \mathbf{e}_i = \mathbf{e}_i^T \mathbf{\Sigma} \mathbf{e}_i = \mathbf{e}_i^T (\lambda_i \mathbf{e}_i) = \lambda_i \mathbf{e}_i^T \mathbf{e}_i = \lambda_i$

(iii)
Using eigenvalue decomposition, we get $\mathbf{\Sigma} = \mathbf{P} \Lambda \mathbf{P}^T$ where Λ is diagonal matrix of eigenvalues and $\mathbf{P} = (\mathbf{e}_1, \mathbf{e}_2, \cdots, \mathbf{e}_p)$ So that $\mathbf{P}\mathbf{P}^T = \mathbf{P}^T \mathbf{P} = \mathbf{I}$ Thus:

$$\sum_{i=1}^{p} \mathrm{Var}(X_i) = \mathrm{tr}(\mathbf{\Sigma}) = \mathrm{tr}(\mathbf{P} \Lambda \mathbf{P}^T) = \mathrm{tr}(\Lambda \mathbf{P} \mathbf{P}^T) = \mathrm{tr}(\Lambda) = \sum_{i=1}^{p} \lambda_i = \sum_{i=1}^{p} \mathrm{Var}(Z_i)$$

(b)
(1)From part a)(1),

$$\begin{aligned} \text{Total population variance} &= \textstyle\sum_{i=1}^{p} \sigma_{ii}^2 \\ &= \textstyle\sum_{i=1}^{p} \lambda_i \end{aligned}$$

and then the proportion of total variance explained by the kth principle component is

$$\frac{\lambda_k}{\sum_{i=1}^{p} \lambda_i}, k = 1, , \cdots, p$$

If most (might be 80% to 90%) of total population variance can be explained by first k components, then those components can replace the original p independent variables without loss lot of information.
So one way to determine k is to find out cumulative percentage of total variation. Using $0.8 - 0.9$ as a threshold to determined k principle components retained in the model. Where

$$\text{cumulative percentage of total variation } = \frac{\sum_{i=1}^{k} \lambda_i}{\sum_{i=1}^{p} \lambda_i}$$

(2) another simple way is Kaiser's Rule, We would like to choose the principle components with eigenvalues over 1, because any component should account for at least as much as a single variable.
(3) Chi-square test is third way to determine k. where Null hypothesis is that k number of principle components is sufficient. non-significant result is expected to decide k is enough to explain most of information of data.

(c)

all p components are required to reproduce the total independent variable's variability. The model fitted with all p principle components is actually the full model fitted by all centered and standardized independent variables.

I Construct following Hypothesis:

$H_0:$ truncated components model is enough to cover the information from original data set

$H_a:$ full components model is required to cover the information from original data set

The test statistics is

$$F = \frac{(SSE_{\text{reduced}} - SSE_{\text{full}})/(df_{\text{reduced}} - df_{\text{full}})}{MSE_{\text{full}}}$$

Table 1 is the results from full model; Table 2 is the results from reduced model.

We have

$SSE_{\text{reduced}} = 6.94 * 46 = 319.24$

$SSE_{\text{full}} = 8.05 * 25 = 201.25$

$df_{\text{reduced}} = 46$

$df_{\text{full}} = 25$

$MSE_{\text{full}} = 8.05$

Therefore

$$F = \frac{(319.24 - 201.25)/(46 - 25)}{8.05} = 0.6979$$

corresponding p-value is $F_{21,25}(0.6979) = 0.2029 > 0.05$

We should not reject the null hypothesis, therefore truncated components model is enough. That means linear regression model using the first 3 principal components has cover the most variation for the dataset. other components are not important.

```
> pf(0.6979,21,25)
[1] 0.2029545
```

Problem 4. One market monitoring organization would like to compare the life time of two brands of bulbs, Brand A and Brand B. They design the experiment in this way. Let X_i and Y_i be the life time of ith bulb in Brand A and Brand B respectively, which can be approximated by independent random variables with exponential distributions with expectations λ and μ

respectively. They pair X_i and Y_i. In the ith experiment, instead of letting both two bulbs burn until they die out, they stop when one of the bulbs burn out, and record the burning time Z_i and indicator W_i of which one burns out. They repeat the experiment n times. Mathematically, Z_i and W_i can be defined as

$$Z_i = \min(X_i, Y_i) \text{ and } W_i = \begin{cases} 1 & \text{if } Z_i = X_i, \\ 0 & \text{if } Z_i = Y_i; \end{cases} \quad i = 1, \ldots, n$$

a) Find closed form expressions for the maximum likelihood estimators of λ and μ. Note: Justify that your estimator is in fact the global maximizer of the likelihood. Use the data "bulb.txt" to compute the MLE.

b) Consider applying the EM algorithm with the complete data taken to be $(X_1, Y_1), \ldots, (X_n, Y_n$ Show that the EM sequence is given by

$$\hat{\lambda}^{(k+1)} = \frac{\sum_{i=1}^n W_i Z_i + \sum_{i=1}^n (1 - W_i)(Z_i + \lambda^{(k)})}{n}$$
$$\hat{\mu}^{(k+1)} = \frac{\sum_{i=1}^n (1 - W_i) Z_i + \sum_{i=1}^n W_i (Z_i + \mu^{(k)})}{n}$$

Use the data "bulb.txt" to find a solution, using the starting value $\mu^{(0)} = 1$ and $\lambda^{(0)} = 1$.

Note: For the model in Problem 4, the EM algorithm is not needed because a closed form expression for the MLE is available. But for other related models with censored data, no closed form expression is available and the EM algorithm is useful.

Thanks Qi Xia for the solution!

Problem 4. One market monitoring organization would like to compare the life time of two brands of bulbs, Brand A and Brand B. They design the experiment in this way. Let X_i and Y_i be the life time of ith bulb in Brand A and Brand B respectively, which can be approximated by independent random variables with exponential distributions with expectations λ and μ respectively. They pair X_i and Y_i. In the ith experiment, instead of letting both two bulbs burn until they die out, they stop when one of the bulbs burn out, and record the burning time Z_i and indicator Wi of which one burns out. They repeat the experiment n times. Mathematically, Z_i and W_i can be defined as

$$Z_i = \min(X_i, Y_i) \text{ and } W_i = \begin{cases} 1 & \text{if } Z_i = X \\ 0 & \text{if } Z_i = Y \end{cases} \quad i = 1, ..., n$$

a). Find closed form expressions for the maximum likelihood estimators of λ and μ. Note: Justify that your estimator is in fact the global maximizer of the likelihood. Use the data "bulb.txt" to compute the MLE.

Answer: First of all, to find the likelihood function, we need to find the joint pdf of W and Z.

$$\begin{aligned} & P(Z \le z, W = 1) = P(X \le z, Y \ge X) \\ =& \int_0^z f_X(x) \int_x^\infty f_Y(y) \, \mathrm{d}x\mathrm{d}y \\ =& \int_0^z \frac{1}{\lambda} e^{\frac{-x}{\lambda}} \int_x^\infty \frac{1}{\mu} e^{\frac{-y}{\mu}} \, \mathrm{d}x\mathrm{d}y \\ =& \frac{\mu}{\lambda+\mu} \left[1 - \exp(-\frac{\lambda+\mu}{\lambda\mu} z)\right]. \end{aligned}$$

Similarly, we have

$$\begin{aligned} & P(Z \le z, W = 0) = P(Y \le z, X \ge Y) \\ =& \int_0^z f_Y(y) \int_y^\infty f_X(x) \, \mathrm{d}x\mathrm{d}y \\ =& \int_0^z \frac{1}{\mu} e^{\frac{-y}{\mu}} \int_y^\infty \frac{1}{\lambda} e^{\frac{-x}{\lambda}} \, \mathrm{d}y\mathrm{d}x \\ =& \frac{\lambda}{\lambda+\mu} \left[1 - \exp(-\frac{\lambda+\mu}{\lambda\mu} z)\right]. \end{aligned}$$

$$f_{Z_i,W_i}(z_i, w_i) = \begin{cases} \frac{1}{\lambda} \exp(-\frac{\lambda+\mu}{\lambda\mu} z) & \text{if } W_i = 1 \\ \frac{1}{\mu} \exp(-\frac{\lambda+\mu}{\lambda\mu} z) & \text{if } W_i = 0 \end{cases} \quad i = 1, ..., n$$

$$\Rightarrow f(z, w) = (\frac{1}{\lambda})^w (\frac{1}{\mu})^{1-w} \exp\left[-(\frac{1}{\lambda} + \frac{1}{\mu}) z\right]$$

Then we have our Likelihood function:

$$Like(\lambda,\mu)=\prod_{i=1}^{n}(\frac{1}{\lambda})_i^w(\frac{1}{\mu})^{1-w_i}\exp\left[-(\frac{1}{\lambda}+\frac{1}{\mu})z_i\right]$$
$$=(\frac{1}{\lambda})^{\sum_{i=1}^n w_i}(\frac{1}{\mu})^{n-\sum_{i=1}^n w_i}\exp\left[-(\frac{1}{\lambda}+\frac{1}{\mu})\sum_{i=1}^{n}z_i\right]$$
$$\Rightarrow l(\lambda,\mu)=-\sum_{i=1}^{n}w_i\log\lambda-(n-\sum_{i=1}^{n}w_i)\log\mu-(\frac{1}{\lambda}+\frac{1}{\mu})\sum_{i=1}^{n}z_i$$

By calculating the first partial derivative and setting to 0, we have:

$$\frac{\partial l}{\partial\lambda}=-\sum_{i=1}^{n}w_i/\lambda+\sum_{i=1}^{n}z_i/\lambda^2=0$$
$$\frac{\partial l}{\partial\mu}=-(n-\sum_{i=1}^{n}w_i)/\mu+\sum_{i=1}^{n}z_i/\mu^2=0$$
$$\Rightarrow\hat{\lambda}=\frac{\sum_{i=1}^n Z_i}{\sum_{i=1}^n W_i},\ \hat{\mu}=\frac{\sum_{i=1}^n Z_i}{n-\sum_{i=1}^n W_i}$$

By checking the second order derivatives matrix

$$\left.\begin{pmatrix}\frac{\partial^2 l}{\partial\lambda^2} & \frac{\partial^2 l}{\partial\lambda\partial\mu}\\ \frac{\partial^2 l}{\partial\lambda\partial\mu} & \frac{\partial^2 l}{\partial\mu^2}\end{pmatrix}\right|_{\lambda=\hat{\lambda},\mu=\hat{\mu}}=\begin{pmatrix}\frac{\sum_{i=1}^n W_i}{\lambda^2}-\frac{2\sum_{i=1}^n W_i}{\lambda^3} & \mathbf{0}\\ \mathbf{0} & \frac{n-\sum_{i=1}^n W_i}{\mu^2}-\frac{2(n-\sum_{i=1}^n W_i)}{\mu^3}\end{pmatrix}<0$$

Therefore, $\hat{\lambda}$ and $\hat{\mu}$ are MLE for λ and μ. By plugging the data, $\hat{\lambda}=0.8058$ and $\hat{\mu}=1.8802$.

b). Consider applying the EM algorithm with the complete data taken to be $(X_1,Y_1),...,(X_n,Y_n)$.Show that the EM sequence is given by

$$\hat{\lambda}^{(k+1)}=\frac{\sum_{i=1}^n W_iZ_i+\sum_{i=1}^n(1-W_i)(Z_i\lambda^{(k)})}{n}$$
$$\hat{\mu}^{(k+1)}=\frac{\sum_{i=1}^n(1-W_i)Z_i+\sum_{i=1}^n W_i(Z_i\mu^{(k)})}{n}$$

Use the data"bulb.txt" to find a solusion, using starting value $\mu^{(0)}=1$ and $\lambda^{(0)}=1$.

Answer: Since the complete data is $(X_1, Y_1), ..., (X_n, Y_n)$, but we did not observe X when X>Y and also we did not observe Y when X>Y. Therefore in our case X and Y are censored data. Since X and Y are independent, we can write their joint desity function as

$$
\begin{aligned}
f(x, y; X, Y) &= \frac{1}{\lambda} \exp(\frac{-x}{\lambda}) \frac{1}{\mu} \exp(\frac{-y}{\mu}) \\
\Rightarrow l(x, y) &= -n \log \lambda - n \log \mu - \sum_{i=1}^{n} x_i/\lambda - \sum_{i=1}^{n} y_i/\mu \\
\Rightarrow \hat{\lambda}_{mle} &= \sum_{i=1}^{n} x_i/n \\
\hat{\mu}_{mle} &= \sum_{i=1}^{n} y_i/n
\end{aligned}
$$

By EM Algorithm,

Expectation step. We calculate $Q(\lambda, \mu, \lambda^{(k)}, \mu^{(k)})$ by

$$
\begin{aligned}
Q(\lambda, \mu, \lambda^{(k)}, \mu^{(k)}) &= E\left\{ l((\mu, \lambda), (X, Y)) | (Z, W), (\lambda^{(k)}, \mu^{(k)}) \right\} \\
= &- n \log \lambda - n \log \mu - \sum_{i=1}^{n} E(X_i | (Z, W), (\lambda^{(k)}, \mu^{(k)}))/\lambda - \sum_{i=1}^{n} E(Y_i | (Z, W), (\lambda^{(k)}, \mu^{(k)}))/\mu
\end{aligned}
$$

By minimizing function Q, we have

$$
\begin{aligned}
\hat{\lambda}_{mle} &= \sum_{i=1}^{n} E(X_i | (Z, W), (\lambda^{(k)}, \mu^{(k)})/n = \hat{E}(X_i | (Z, W), (\lambda^{(k)}, \mu^{(k)}) \\
\hat{\mu}_{mle} &= \sum_{i=1}^{n} E(Y_i | (Z, W), (\lambda^{(k)}, \mu^{(k)})/n = \hat{E}(Y_i | (Z, W), (\lambda^{(k)}, \mu^{(k)})
\end{aligned}
$$

To complete E step, we need to know $\hat{E}(X_i | (Z, W), (\lambda^{(k)}, \mu^{(k)})$ and $\hat{E}(Y_i | (Z, W), (\lambda^{(k)}, \mu^{(k)})$. To calculate, we can divide X_i's into two groups: Z_i and U_i where $U_i = Z_i + T_{xi} > Y_i$. And by memorylessness property of exponential distribution, we know that $T_{xi} \sim Exp(\lambda)$ thus

$$\sum_{i=1}^{n} X_i = \sum_{i=1}^{n} W_i Z_i + \sum_{i=1}^{n} (1 - W_i)(Z_i + T_i)$$

$$\Rightarrow \hat{E}\left\{ X_i | (Z, W), (\lambda^{(k)}, \mu^{(k)}) \right\}$$

$$= \frac{\sum_{i=1}^{n} W_i Z_i + \sum_{i=1}^{n} (1 - W_i)(Z_i + E(T_{xi} | (Z, W), (\lambda^{(k)}, \mu^{(k)})))}{n}$$

$$= \frac{\sum_{i=1}^{n} W_i Z_i + \sum_{i=1}^{n} (1 - W_i)(Z_i + \lambda^{(k)})}{n}$$

Similarly, we can divide Y_i's into two groups: Z_i and V_i where $U_i = Z_i + T_{yi} > X_i$. And by memorylessness property of exponential distribution, we know that $T_{yi} \sim Exp(\mu)$ thus

$$\sum_{i=1}^{n} Y_i = \sum_{i=1}^{n} W_i Z_i + \sum_{i=1}^{n} (1 - W_i)(Z_i + T_{yi})$$

$$\Rightarrow \hat{E}\left\{ X_i | (Z, W), (\lambda^{(k)}, \mu^{(k)}) \right\}$$

$$= \frac{\sum_{i=1}^{n} W_i Z_i + \sum_{i=1}^{n} (1 - W_i)(Z_i + E(T_{yi} | (Z, W), (\lambda^{(k)}, \mu^{(k)})))}{n}$$

$$= \frac{\sum_{i=1}^{n} W_i Z_i + \sum_{i=1}^{n} (1 - W_i)(Z_i + \mu^{(k)})}{n}$$

By running EM algorithm in R,we find the solution $\hat{\lambda} = 0.8058$ and $\hat{\mu} = 1.8802$ which is almost the same as our mle result.

iteration = 24
lambda = 0.8058154
mu = 1.880067
conv.err 5.223367e-09

Problem 5. The incidence of a rare disease seems to be decreasing. In successive years, the number of new cases is $y_1, \ldots, y_n$. We assume that $y_1, \ldots, y_n$ are independent random variables from Poisson distributions with means $\theta, \theta^2, \ldots, \theta^n$ respectively.

a) Forumlate a likelihood ratio test for testing $\mathrm{H}_0 : \theta = 1$ versus $\mathrm{H}_a : \theta < 1$. For $(y_1, y_2) = (2, 0)$, would such test with size 0.20 test accept or reject H_0?

b) Describe a procedure for forming a level 0.95 one-sided confidence interval of the form $(0, \theta_u)$ [you do not need to come up with a closed form expression and can express that you would need to calculate the quantiles of certain distributions and do a numerical search to form the confidence interval]. Use your procedure to find (approximately) a realized confidence interval of the form $(0, \theta_u)$ for the sample $(y_1, y_2) = (2, 0)$ (you may want to write a computer program for this).