

HW-3

Dennis Goldenberg

2024-02-09

Homework 3 - Predictive Modeling in Finance and Insurance

1. Likelihood Function for mean of normal distribution

a. Joint Density Function

Note that Y_1, Y_2 , and Y_3 are independent. Therefore, their joint probability density function (p.d.f) is a product of their marginal probability density functions:

$$\begin{aligned} f_{(Y_1, Y_2, Y_3)}(y_1, y_2, y_3) &= f_{Y_1}(y_1)f_{Y_2}(y_2)f_{Y_3}(y_3) \\ &= \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2\sigma^2}(y_1-\mu_1)^2} * \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2\sigma^2}(y_2-\mu_2)^2} * \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2\sigma^2}(y_3-\mu_3)^2} \\ &= \frac{1}{(2\pi\sigma^2)^{\frac{3}{2}}} * e^{-\frac{1}{2\sigma^2}(\sum_{i=1}^3 (y_i-\mu_i)^2)} \end{aligned}$$

b. Likelihood function and Log-Likelihood

The likelihood function is just the joint p.d.f, given parameter of interest $\vec{\mu} = (\mu_1, \mu_2, \mu_3)$:

$$L(\vec{\mu}) = f_{(Y_1, Y_2, Y_3)}(y_1, y_2, y_3; \mu) = \frac{1}{(2\pi\sigma^2)^{\frac{3}{2}}} * e^{-\frac{1}{2\sigma^2}(\sum_{i=1}^3 (y_i-\mu_i)^2)}$$

The log-likelihood is just the natural log of this function:

$$\begin{aligned} \ell(\vec{\mu}) = \ln(L(\mu)) &= \ln\left(\frac{1}{(2\pi\sigma^2)^{\frac{3}{2}}}\right) + \ln\left(e^{-\frac{1}{2\sigma^2}(\sum_{i=1}^3 (y_i-\mu_i)^2)}\right) \\ &= -\frac{3}{2}\ln(2\pi\sigma^2) - \frac{1}{2\sigma^2}\left(\sum_{i=1}^3 (y_i - \mu_i)^2\right) \end{aligned}$$

c. Score function, Observed Information, Expected Information

The score function is simply the derivative of the log likelihood with respect to the parameter of interest, $\vec{\mu}$. Note that the function is actually a matrix, as I takt the derivative with respect to μ_1, μ_2 , and μ_3 :

$$S(\vec{\mu}) = \frac{d}{d\mu} \ell(\mu) = \begin{bmatrix} \frac{d}{d\mu_1} \left(-\frac{3}{2}\ln(2\pi\sigma^2) - \frac{1}{2\sigma^2} \left(\sum_{i=1}^3 (y_i - \mu_i)^2 \right) \right) \\ \frac{d}{d\mu_2} \left(-\frac{3}{2}\ln(2\pi\sigma^2) - \frac{1}{2\sigma^2} \left(\sum_{i=1}^3 (y_i - \mu_i)^2 \right) \right) \\ \frac{d}{d\mu_3} \left(-\frac{3}{2}\ln(2\pi\sigma^2) - \frac{1}{2\sigma^2} \left(\sum_{i=1}^3 (y_i - \mu_i)^2 \right) \right) \end{bmatrix} = \begin{bmatrix} \frac{y_1 - \mu_1}{\sigma^2} \\ \frac{y_2 - \mu_2}{\sigma^2} \\ \frac{y_3 - \mu_3}{\sigma^2} \end{bmatrix}$$

Note that the observed information matrix is a matrix of second derivatives of the log-likelihood function.

Since we have 3 variables to differentiate with respect to, it is a 3×3 matrix, multiplied by -1:

$$j(\vec{\mu}; Y) = -1 * \begin{bmatrix} \frac{d^2 \ell(\mu)}{d\mu_1^2} & \frac{d^2 \ell(\mu)}{d\mu_2 d\mu_1} & \frac{d^2 \ell(\mu)}{d\mu_3 d\mu_1} \\ \frac{d^2 \ell(\mu)}{d\mu_1 d\mu_2} & \frac{d^2 \ell(\mu)}{d\mu_2^2} & \frac{d^2 \ell(\mu)}{d\mu_3 d\mu_2} \\ \frac{d^2 \ell(\mu)}{d\mu_1 d\mu_3} & \frac{d^2 \ell(\mu)}{d\mu_2 d\mu_3} & \frac{d^2 \ell(\mu)}{d\mu_3^2} \end{bmatrix} = \begin{bmatrix} \frac{1}{\sigma^2} & \frac{1}{\sigma^2} & \frac{1}{\sigma^2} \\ \frac{1}{\sigma^2} & \frac{1}{\sigma^2} & \frac{1}{\sigma^2} \\ \frac{1}{\sigma^2} & \frac{1}{\sigma^2} & \frac{1}{\sigma^2} \end{bmatrix} = \frac{1}{\sigma^2} \mathbf{1}_{3 \times 3}$$

The expected information matrix is simply the expectation with respect to our observations of our observed information matrix:

$$i(\vec{\mu}) = \mathbb{E}[j(\vec{\mu}; Y)] = \frac{1}{\sigma^2} \mathbb{E}[\mathbf{1}_{3 \times 3}] = \frac{1}{\sigma^2} \mathbf{1}_{3 \times 3}$$

Given the observations, these matrices take on the values:

$$S(\vec{\mu}; Y) = \begin{bmatrix} \frac{4}{\sigma^2} & \frac{6.5}{\sigma^2} & \frac{5}{\sigma^2} \end{bmatrix}^T \text{ and } i(\vec{\mu}) = j(\vec{\mu}; Y) = \frac{1}{\sigma^2} \mathbf{1}_{3 \times 3}$$

2. Fun with Distributions

a. Distribution of Y_1^2

Since $Y_1 \sim N(0, 1)$, $Y_1^2 \sim \chi^2(1)$, or the chi-squared distribution with 1 degree of freedom.

b. Combination of Y_1 and Y_2

Note $\frac{Y_2 - \mu_2}{\sigma_2} = \frac{Y_2 - 3}{2} \sim N(0, 1)$; therefore:

$$\left(\frac{Y_2 - 3}{2}\right)^2 \sim \chi^2(1)$$

Using the independence of Y_1 and Y_2 and Cochran's Theorem:

$$y^T y = \begin{bmatrix} Y_1 & \frac{Y_2 - 3}{2} \end{bmatrix} * \begin{bmatrix} Y_1 \\ \frac{Y_2 - 3}{2} \end{bmatrix} = Y_1^2 + \left(\frac{Y_2 - 3}{2}\right)^2 = \chi^2(1 + 1) = \chi^2(2)$$

So, $y^T y$ has the chi-squared distribution with 2 degrees of freedom.

c. Multivariate Normal

Note that V in this case is the Variance-Covariance matrix. Since Y_1 and Y_2 are independent, the off-diagonal elements, which represent covariance, are 0. There diagonal elements are just $\sigma_1^2 = 1$ and $\sigma_2^2 = 4$, respectively, so:

$$V = \begin{bmatrix} 1 & 0 \\ 0 & 4 \end{bmatrix}$$

I find the inverse of this 2 by 2 matrix:

$$V^{-1} = \frac{1}{1(4) - 0(0)} \begin{bmatrix} 4 & 0 \\ 0 & 1 \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & \frac{1}{4} \end{bmatrix}$$

Therefore:

$$\begin{aligned} y^T V^{-1} y &= \begin{bmatrix} Y_1 & Y_2 \end{bmatrix} * \begin{bmatrix} 1 & 0 \\ 0 & \frac{1}{4} \end{bmatrix} * \begin{bmatrix} Y_1 \\ Y_2 \end{bmatrix} \\ &= \begin{bmatrix} Y_1 & \frac{Y_2}{4} \end{bmatrix} * \begin{bmatrix} Y_1 \\ Y_2 \end{bmatrix} \\ &= Y_1^2 + \left(\frac{Y_2}{2}\right)^2 \end{aligned}$$

4. Linear Regression

a. Fitting model B

```
library(ggplot2)
library(readxl)
```

I first import the data:

```
carbData <- read_excel("Table 6.3 Carbohydrate diet-1.xls", skip = 2, sheet = "Sheet1")
```

Then, I fit the model:

```
mod_B <- lm("carbohydrate ~ age + protein", data = carbData)
summary(mod_B)
```

```
##
## Call:
## lm(formula = "carbohydrate ~ age + protein", data = carbData)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -11.2692  -5.9968   0.9902   5.7952   9.5474
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  15.08848    12.16239   1.241  0.2316
## age          -0.09167     0.12818  -0.715  0.4842
## protein       1.68189     0.73693   2.282  0.0356 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7.002 on 17 degrees of freedom
## Multiple R-squared:  0.2372, Adjusted R-squared:  0.1475
## F-statistic: 2.643 on 2 and 17 DF,  p-value: 0.1001
```

The 95% confidence interval for β_1 , using the fact that $\hat{\beta}_1 \sim N(\beta, \sigma^2(x^T x)^{-1})$:

$$\left(\hat{\beta}_1 + \hat{se}(\hat{\beta}_1) z_{\frac{1-.95}{2}}, \hat{\beta}_1 + \hat{se}(\hat{\beta}_1) z_{\frac{1-.95}{2}} \right) = \left(\hat{\beta}_1 - 1.96 * \hat{se}(\hat{\beta}_1), \hat{\beta}_1 + 1.96 * \hat{se}(\hat{\beta}_1) \right)$$

I plug in the values from the summary to get the 95% confidence interval:

```
beta_hat <- mod_B$coefficients['age']
se_beta_hat <- sqrt(diag(vcov(mod_B)))['age']
CI_bh <- c(beta_hat - 1.96*se_beta_hat, beta_hat + 1.96*se_beta_hat)
names(CI_bh) <- c("Lower Bound", "Upper Bound")
CI_bh
```

```
## Lower Bound Upper Bound
## -0.3428922  0.1595565
```

Note that, in the model, the probability that the t distribution with $20 - 1 = 19$ has a greater absolute value the t-statistic generated by $\hat{\beta}_1$ is $.4842 > .05$, so **we fail to reject H_0** ; thus there is evidence that the response does not depend on age.

I can show this manually as well by generating the t-statistic and showing the probability that the t-distribution is further from 0 than this value:

```
t_stat <- beta_hat/se_beta_hat
prob_t <- 2*pt(t_stat, df = dim(carbData) - 3)[1]
prob_t
```

```
## [1] 0.4842089
```

b. Prediction Interval

I first fit Model A, and find the summary:

```
mod_A <- lm('carbohydrate ~ protein', data = carbData)
summary(mod_A)

##
## Call:
## lm(formula = "carbohydrate ~ protein", data = carbData)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -12.4979  -5.9829   0.9019   4.8870  10.6620
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  12.4787    11.4435   1.090  0.2899
## protein       1.5800     0.7131   2.216  0.0399 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.907 on 18 degrees of freedom
## Multiple R-squared:  0.2143, Adjusted R-squared:  0.1706
## F-statistic: 4.909 on 1 and 18 DF,  p-value: 0.03986
```

For the 95% prediction interval, I note:

$$\hat{se}(\text{pred}) = s * \sqrt{1 + \frac{1}{N} + \frac{(x_a - \bar{x})^2}{(N-1)s_X^2}}$$

The interval itself can be represented by the equation:

$$\hat{y} \pm t_{\frac{1+.95}{2}, n-1} \hat{se}(\text{pred})$$

Here, s is the residual standard error. So, I obtain all of these quantities, and generate the prediction interval:

```
pred_data <- data.frame(protein = c(21))
yhat <- predict.lm(mod_A, pred_data)
n <- dim(carbData)[1]
se_yhat <- summary(mod_A)$sigma *
sqrt(1 + 1/n + ((21 - mean(carbData$protein))^2)/((n-1)*var(carbData$protein)))
PI <- c(yhat - qt(0.975, df = 18)*se_yhat, yhat + qt(0.975, df = 18)*se_yhat)
names(PI) <- c("PI lower Bound", "PI upper Bound")
PI
```

```
## PI lower Bound PI upper Bound
##      28.94052      62.37504
```

c. Testing General Significance of Age

Model B is the full model, and Model A is the reduced model; I can calculate an F-statistic relating to the significance of *Age* by scaling the difference in SSE's through division of number of predictors being tested (1 in this case) and dividing it by the Mean Squared error:

$$F_{stat} = \frac{SSE_{full} - SSE_{reduced}}{q * s^2} = \frac{SSE_{full} - SSE_{reduced}}{s^2} \sim F(ndf = 1, ddf = N - 3)$$

I get these values by calculating SST , then obtaining the value for SSE from the R^2 values from the models. Then, I get the F-stat:

```
SST <- var(carbData$carbohydrate)*(n - 1)
SSE_full <- SST * (1 - summary(mod_B)$r.squared)
SSE_reduced <- SST * (1 - summary(mod_A)$r.squared)
MSE <- (summary(mod_B)$sigma)^2
F_stat <- (SSE_reduced - SSE_full)/MSE
F_stat
```

```
## [1] 0.5114732
```

Using this F-stat, I calculate the p-value:

```
pf(F_stat, df1 = 1, df2 = n - 3, lower.tail = FALSE)
```

```
## [1] 0.4842089
```

Note that $0.4842 > 0.05$, so the age predictor is not significant at the 5% significance level.

5. National Life Expectancies

I import the data:

```
UN_data <- read.csv("UNLifeExpectancy-3.csv", header = TRUE)
```

a. Fitting Regression model

I fit the model specified:

```
reg_UN <- lm("LIFEEXP ~ FERTILITY + PUBLICEDUCATION + log(PRIVATEHEALTH)",
             data = UN_data)
summary(reg_UN)
```

```
##
## Call:
## lm(formula = "LIFEEXP ~ FERTILITY + PUBLICEDUCATION + log(PRIVATEHEALTH)",
##     data = UN_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -23.013  -4.090   1.218   4.597  12.589
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)    85.6264     2.0033  42.742  <2e-16 ***
## FERTILITY      -5.3993     0.3308 -16.324  <2e-16 ***
## PUBLICEDUCATION -0.1846     0.2685  -0.688    0.493
## log(PRIVATEHEALTH) -1.0296     0.9431  -1.092    0.277
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.645 on 148 degrees of freedom
## (33 observations deleted due to missingness)
## Multiple R-squared:  0.645, Adjusted R-squared:  0.6378
## F-statistic: 89.64 on 3 and 148 DF, p-value: < 2.2e-16
```

i. The coefficient on public education states that - all else equal - for every 1 unit increase in the public education variable, life expectancy is expected to decrease by -0.1846 years.

ii. For a significance test, I would have the following null and alternative hypotheses:

$$H_0 : \beta_{PE} = 0 \text{ and } H_1 : \beta_{PE} \neq 0$$

I would use the t-statistic and decision boundary:

$$t = \frac{\hat{\beta}_{PE}}{\hat{se}(\hat{\beta}_{PE})} \sim \text{T-dist}(df = n - 3 - 1) \text{ and } \mathbb{P}(T \geq |t|) \leq 0.05$$

Here, $T \sim \text{T-dist}(df = n - 3 - 1)$, and $\alpha = 0.05$; if the probability is less than 5% that a t-statistic has a greater absolute value under H_0 , I reject H_0 and say that the variable is statistically significant. Based on the summary of the fit, however, that coefficient has a very high p-value for its t-statistic at 0.493, meaning that we would conclude it not to be statistically significant at the significance level .05.

iii. My null alternative hypothesis are:

$$H_0 : \beta_{PE} = \beta_{lnH} = 0 \text{ and } H_a : \beta_{PE} \neq 0 \text{ and/or } \beta_{lnH} \neq 0$$

My test statistic and decision boundary is:

$$F_{stat} = \frac{SSE_{full} - SSE_{reduced}}{2 * s^2} \sim \text{F-dist}(ndf = 2, ddf = N - 4) \text{ and } \mathbb{P}(F \geq F_{stat}) \leq 0.05$$

If, under H_0 , there is less than a 5% chance that a value in this F-distribution exceeds the generated f-statistic, I reject H_0 and say that the two variables are jointly statistically significant. To get all of the values, I run the reduced model (without PE and $\ln H$), get the SSE's, and calculate the F-stat (note that the SST calculation is trickier due to the fact that I have different observations omitted due to missingness):

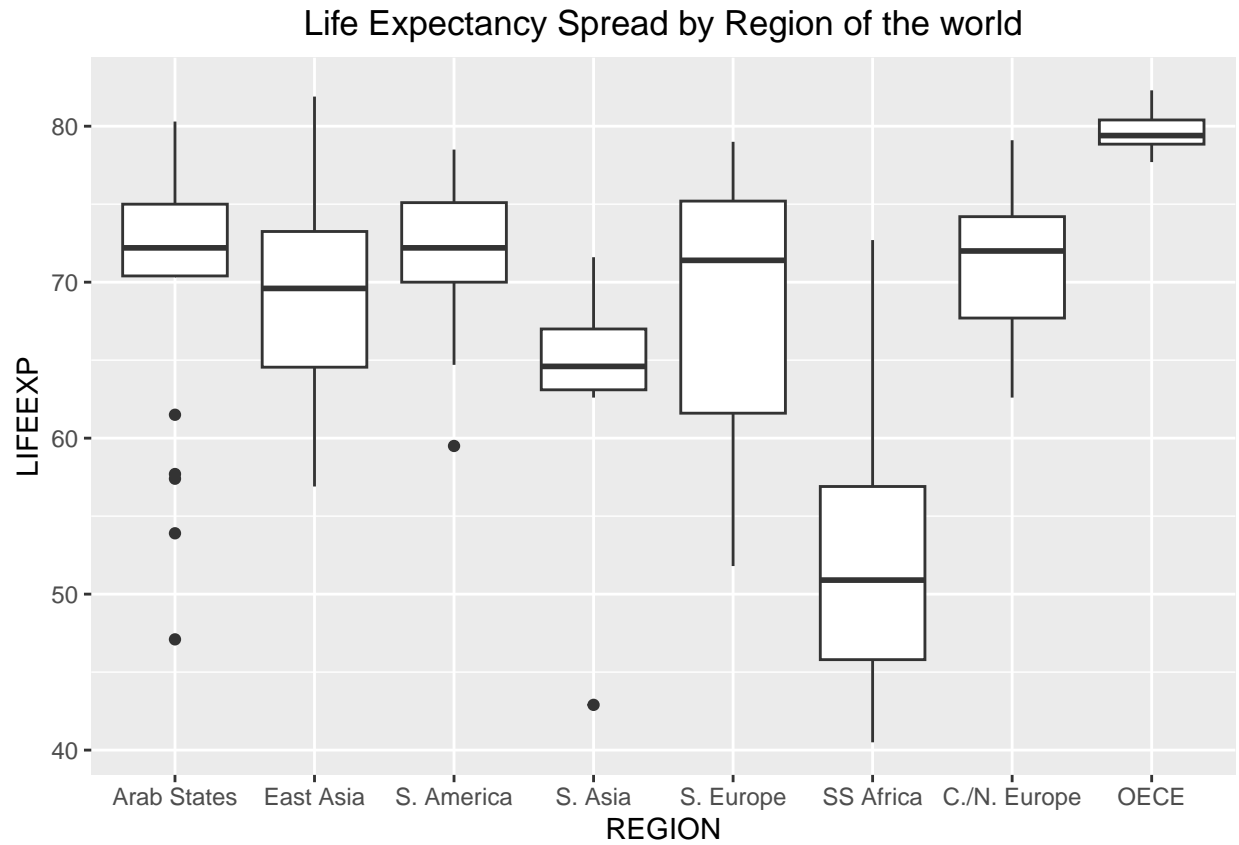
```
#reg_UN_red <- lm("LIFEEXP ~ FERTILITY", data = UN_data)
#SST_UN <- var(UN_data$LIFEEXP)*(n_UN - 1)
#SSE_full_UN <- SST_UN * (1 - summary(reg_UN)$r.squared)
#SSE_reduced_UN <- SST_UN * (1 - summary(reg_UN_red)$r.squared)
#MSE_UN <- (summary(reg_UN)$sigma)^2
#F_stat_UN <- (SSE_reduced_UN - SSE_full_UN)/(2*MSE_UN)
#F_stat_UN
#attributes(summary(reg_UN_red))
#summary(reg_UN_red)$df
#summary(reg_UN)$df
#summary(reg_UN_red)
```

Then, I calculate the p-value of this F_{stat} ...

b. Generating Box Plot

I turn region into a categorical variable, and then generate the boxplot:

```
UN_data$REGION <- factor(UN_data$REGION, levels = 1:8, ordered = TRUE)
RegionLabels <- c("Arab States", "East Asia", "S. America", "S. Asia",
                  "S. Europe", "SS Africa", "C./N. Europe",
                  "OECE")
ggplot(UN_data) + geom_boxplot(aes(x = REGION, y = LIFEEXP)) +
  scale_x_discrete(labels= RegionLabels) +
  ggtitle("Life Expectancy Spread by Region of the world") +
  theme(plot.title = element_text(hjust = 0.5))
```

Here, most regions hover in the 65-70 range for mean life expectancy, but Sub-Saharan Africa and South Asia are particularly low in terms of life expectancy, with High-Income OECE countries being particularly high.

c. Regression Model with Region