When Is the Best Time to Look for a Job for Women?
-- A time series analysis

1. Introduction

In this project, our goal is to identify a model best fitting the time series monthly unemployment data for female aged 16-19 years from 1948-1981[1]. The data is a summary of monthly employment survey of 60,000 households, conducted by the U.S. Bureau of Labor. People are classified as unemployed if they did not work during the survey week, or if they make efforts to find a job in the previous few weeks, or if they were available for work during the survey week. Figure 1 shows the rising of unemployment rate of 408 months. The moving average smoothing line emphasizes the upward trend in the series throughout these 30 years. The seasonal pattern of adult female unemployment shows twin peaks across the year. It increases at the beginning of the year, then declines through the first six months, and rises again and declines after Christmas.[2] In order to analyze the unemployment rate, we first transformed the data to stabilize the series. Then, we built several time series regression models to fit the data and then evaluated the models by R^2, Akaike's Information Criterion (AIC) and Bayesian Information Criterion (BIC).

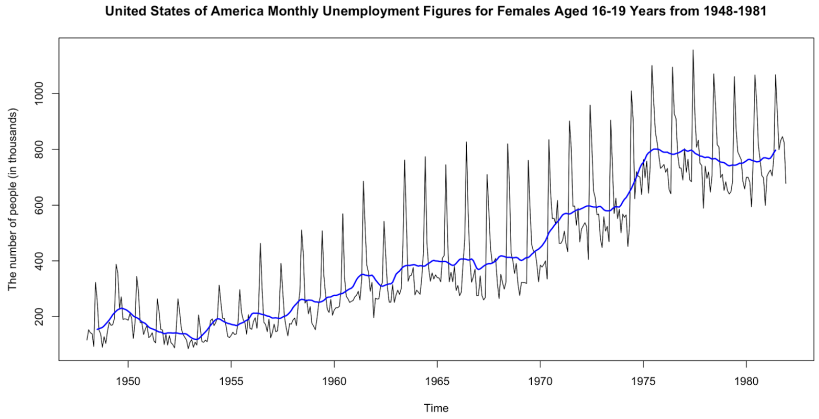

Figure 1. Female Aged 16-19 Unemployment from 1948 to 1981

2. Material and Methods

To understand the unemployment data further, exploratory data analysis was conducted.

Figure 1 shows the unemployment series, say x_t, exhibits heteroscedasticity. To stabilize the variance over the series, log-transformation or Box-Cox transformation is useful. In this project, we used log-transformation $y_t = \log x_t$, which also improved the approximation to normality.

As shown in Figure 2, y_t is not stationary either. To make it stationary, detrending or differencing could have been used. In this project, we wanted to fit several regression models and compare them. Therefore, detrending was used since it was easier to interpret for the unemployment data. The time has been centralized to avoid drastic intercepts. Here we supposed the model was of the form (notice all following t was centered):

$$y_t = \beta_0 + \beta_1 t + w_t \quad (1)$$

Next we built several linear regression models based on the results of ACF and PACF plots of the residuals of the proposed model (1). The specific experiments were regression with lagged variables and regression with indicator variables. We compared these models by computing model selection criteria including R^2, AIC and BIC. We also conducted F test in ANOVA analysis to evaluate the full model and reduced models.

3. Result

3.1 Data Preprocessing

Before log-transformation, the variance of the first half of data is 13877.32 and that of the second half data is 43372.48. After transformation, the variances become 0.2044 and 0.1349. The variability over the length of the data becomes more stable.

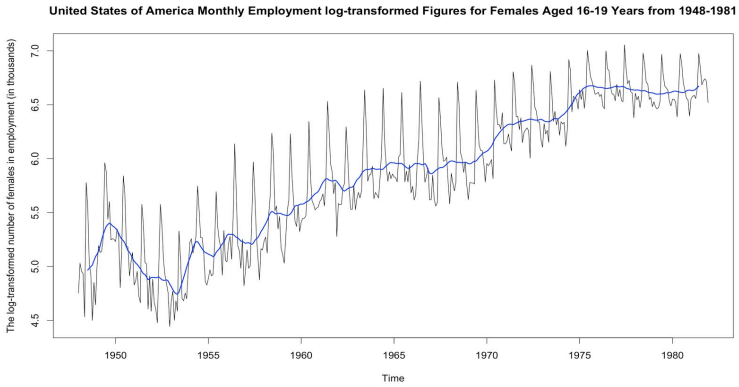

Figure 2. Log Female Aged 16-19 Unemployment from 1948 to 1981

From the time series (Figure 2) and the ACF plot (Figure 3), we know that the data exists a particular uptrend and is highly autocorrelated. Furthermore, some lags of PACF values are large. Hence, we need to do detrending.

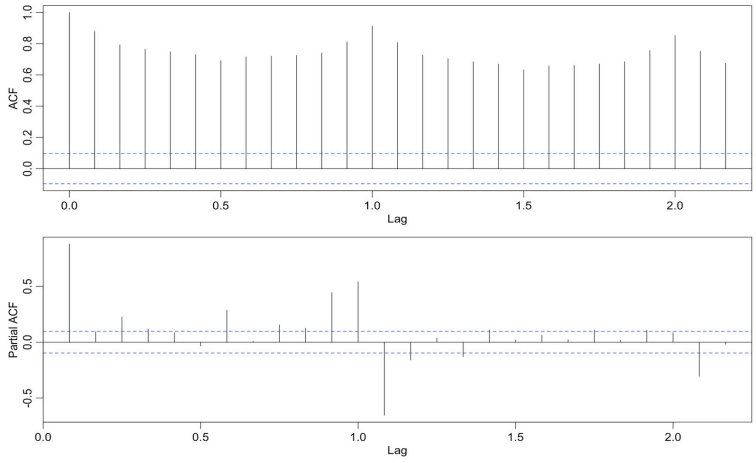

Figure 3. ACF and PACF of Unemployment Data

3.2 Model Candidates

- Model 1:

Obviously, the number of unemployment female increases as time goes on. So it is reasonable to regress y_t on time:

$$\text{Model 1:} \quad y_t = 5.8511 + 0.0586t$$

The independent variable t is significant from t test and R^2 is 0.7897, which means 78.97% variation of y_t can be explained by t. In order to know whether there are other features that may influence y_t, we plotted the ACF and PACF of residuals.

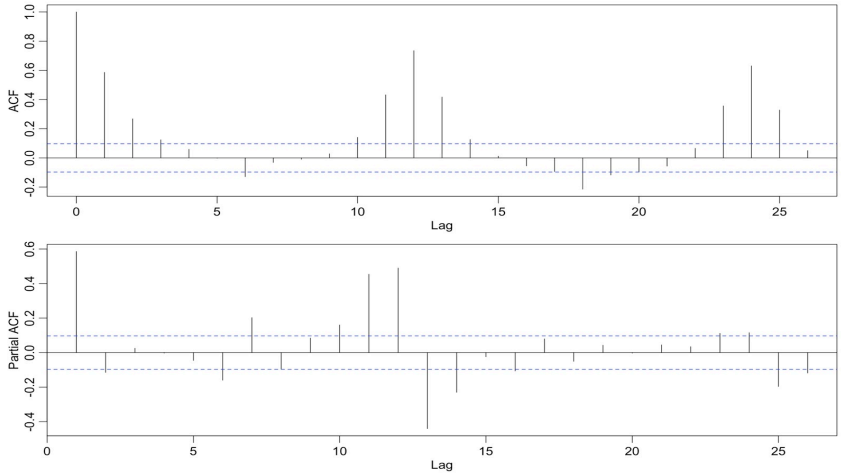

Figure 4. ACF and PACF of Model 1's Residuals

Figure 4 shows residuals are not white noise. It is necessary to add more features to explain y_t. The ACF and PACF values of lag1 and lag12 are both significant, so it is possible that y_{t-1} and y_{t-12} are correlated with y_t, which may be helpful for explaining the variation of y_t. Therefore, we separately added y_{t-1}, y_{t-12}, y_{t-1} and y_{t-12} to model 1.

- Model 2:

In model 2, we added y_{t-1} to model 1.

$$\text{Model 2:} \quad y_t = 2.3198 + 0.0232t + 0.6039y_{t-1}$$

The results of this model 2 shows that t and y_{t-1} are both significant from t test and R^2 is 0.8633, larger than that in model 1. And y_{t-1} is also significant from F test. Therefore, it is necessary to include y_{t-1} in model 2.

- Model 3:

In model 3, we added y_{t-12} to model 1.

$$\text{Model 3:} \quad y_t = 1.5370 + 0.0148t + 0.7439y_{t-12}$$

The result of model 3 shows that t and y_{t-12} are both significant from t test and R^2 is 0.9082, larger than that in model1. And y_{t-12} is also significant from F test. Therefore, it is necessary to include y_{t-12} in model3.

- Model 4:

In model 4, we added both y_{t-1} and y_{t-12} to model 1.

$$\text{Model 4:} \quad y_t = 0.4280 + 0.0037t + 0.3338y_{t-1} + 0.5986y_{t-12}$$

The result of model 4 shows that y_{t-1} and y_{t-12} are both significant from t test and R^2 is 0.9274, larger than that in model 1. And y_{t-1} and y_{t-12} are also significant from F test. Therefore, it is necessary to include both y_{t-1} and y_{t-12} in model4.

3.3 Structural Model

Every year has 12 months' records of unemployment rate data. So this satisfies the basic structural model's condition, which consists of three parts: trend component, month component and irregular component. And then we have 12 indicator variables. The model is:

$$y_t = \beta t + \alpha_1 Q_1(t) + \alpha_1 Q_1(t) + \alpha_2 Q_2(t) + \cdots + \alpha_{12} Q_{12}(t) + w_t$$

We had fitted this model in R. From the output given by R, we can see all the variables are significant. And the R^2 is 0.9989. It is higher than the previous model, and performs best at interpreting the variance of data.

3.4 Model Selection

We have built 5 different models, 4 regression models with different variables and one structural model. The following table summarizes the important criteria from each model which are important in choosing the models.

Table 1. Criteria of Candidate Models

Model	R^2	AIC	BIC
Model 1	0.7897	-1.4161	-1.1387
Model 2	0.8633	-8.7980	-1.8408
Model 3	0.9082	-2.2661	-2.2271
Model 4	0.9274	-2.4894	-2.4406
Model 5	0.9989	-2.2106	-2.0729

R^2 can demonstrate how much variance the model explains. But a higher R^2 alone does not guarantee a better model. Some model has over-fitting issues. So we computed corresponding AIC, BIC values to help us select. After considering all three values for each model, we decided to use model4 as our final model.

4. Conclusion and Discussion

4.1 Final Model

Based on our results, we can express our final model as:

$$y_t = 0.4280 + 0.0037\mathrm{t} + 0.3338y_{t-1} + 0.5986y_{t-12}$$

And we can check the residuals by plotting the histogram:

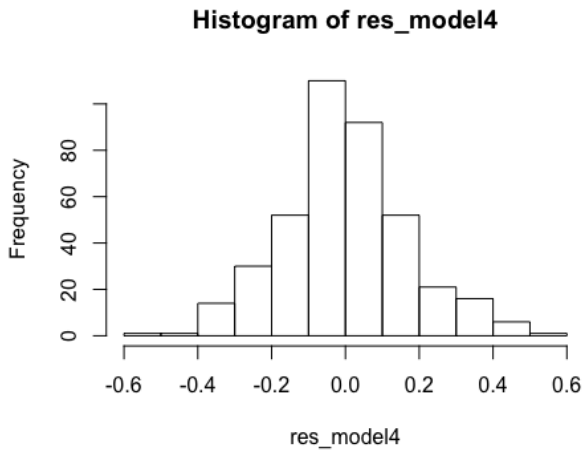

Figure 5. Histogram of Model 4's Residuals

The residuals are normally distributed, which is consistent with our model's white noise assumption. So we can trust our final model.

4.2 Discussion of the Final Model

There are two lagged variables in the final model, one being the value measured at time (t-1), and another at time (t-12). So the time series value is associated with the value from a month ago and the value from 12 months ago. It makes sense, since these two variables represent short term effect and long term effect from the human resources market respectively, which is related to the unemployment rate.

Based on our analysis on the past unemployment data above, we concluded that it is important for women to prepare early when it comes to job hunting, and avoiding unemployment peak in summer might be a good choice. To predict the unemployment data in a certain month, we can take last month and this month of last year's data into consideration.

Appendix

```
# Import data

unemployment = read.csv("monthly-us-female-1619-years-une.csv",
                        stringsAsFactors = FALSE)
unemployment = unemployment[-409, ]
colnames(unemployment) = c("month", "unemployment")

# Transform data into time series data
ue = ts(unemployment$unemployment, frequency = 12, start = c(1948, 1))
# Plot time series data and moving average line
wue = c(0.5, rep(1,11), 0.5)/12
unemploy = filter(ue, sides = 2, filter = wue)
plot(ue,main = "United States of America Monthly Employment Figures for Females Aged 16-
19 Years from 1948-1981", ylab = "The number of females in employment (in thousands)")
lines(unemploy, lwd = 2, col = 4)
# Check the two half parts of data's variance
var(ue[1:204])
var(ue[205:408])
# Do the log-transformation for data
log_ue = log(ue)
# Plot time series data and moving average line
wlog_ue = c(0.5, rep(1,11), 0.5)/12
log_unemploy = filter(log_ue, sides = 2, filter = wlog_ue)
plot(log_ue,main = "United States of America Monthly Employment log-transformed Figures for
Females Aged 16-19 Years from 1948-1981", ylab = "The log-transformed number of females in
employment (in thousands)")
lines(log_unemploy, lwd = 2, col = 4)
# Check the two half parts of data's variance
var(log_ue[1:204])
var(log_ue[205:408])
# Plot the acf and pacf of the data
par(mfrow = c(2,1))
acf(ue, main = "ACF of ue")
pacf(ue, main = "PACF of ue")
par(mfrow = c(1,1))
times = time(log_ue)
mean(times)  #1965
# Centralize time
times_central = times - mean(times)
# Fit the first model that contains only t as variable
model1 = lm(log_ue ~ times_central)
summary(model1)
R1 = summary(model1)$r.squared
# Calculate the AIC and BIC
```

```
AIC1 = AIC(model1) / length(log_ue) - log(2 * pi)
BIC1 = BIC(model1) / length(log_ue) - log(2 * pi)
res_model1 = residuals(model1)
# Plot acf of the residual in model1
par(mfrow = c(2,1))
acf(res_model1, main = "ACF of residuals of model1")
pacf(res_model1, main = "PACF of residuals of model1")
par(mfrow = c(1,1))
# Add 1 time lag variable
data_lag1_12 = ts.intersect(log_ue, times_central, lag1 = lag(log_ue, -1), lag12 = lag(log_ue, -
12))
# Fit the second model that contains time t and value at time (t-1) as variable
model2 = lm(log_ue ~ times_central + lag1,data = data_lag1_12)
summary(model2)
R2 = summary(model2)$r.squared
# calculate the AIC and BIC
AIC2 = AIC(model2) / length(log_ue) - log(2 * pi)
BIC2 = BIC(model2) / length(log_ue) - log(2 * pi)
# F-test
anova(model2)
# Fit the third model that contains time t and value at time (t-12) as variable
model3 = lm(log_ue ~ times_central + lag12,data = data_lag1_12)
summary(model3)
R3 = summary(model3)$r.squared
# calculate the AIC and BIC
AIC3 = AIC(model3) / length(log_ue) - log(2 * pi)
BIC3 = BIC(model3) / length(log_ue) - log(2 * pi)
# F-test
anova(model3)
# Fit the third model that contains time t, value at time (t-1) and value at time (t-12) as variable
model4 = lm(log_ue ~ times_central + lag1 + lag12, data = data_lag1_12)
summary(model4)
R4 = summary(model4)$r.squared
# calculate the AIC and BIC
AIC4 = AIC(model4) / length(log_ue) - log(2 * pi)
BIC4 = BIC(model4) / length(log_ue) - log(2 * pi)
# F-test
anova(model4)
# make monthly factors
M = factor(cycle(ue) )
# fit a structural model
model5 = lm(log(ue)~0 + times_central + M, na.action=NULL)
R5 = summary(model5)$r.squared
# calculate the AIC and BIC
AIC5 = AIC(model5) / length(log_ue) - log(2 * pi)
BIC5 = BIC(model5) / length(log_ue) - log(2 * pi)
```

```
# F-test
anova(model5)
data.frame(R1,R2,R3,R4,R5)
data.frame(AIC1,AIC2,AIC3,AIC4,AIC5)
data.frame(BIC1,BIC2,BIC3,BIC4,BIC5)
# Check the residual of model4 from the histogram to see whether it is satisfy the white noise condition
res_model4 = residuals(model4)
hist(res_model4, xlab = "residuals", main = "Histogram of residuals of model4")
```