Stat 506 HW7

Jacob Oard

April 6, 2023

Introduction

We are using observational data to inspect ice cream consumption. In the dataset our response is labeled "IC" containing ice cream consumption with the following covariates: "Date" containing time period 1-30 of the study (from 03/18/51 to 07/11/53), "Price" containing price of ice cream per pint in dollars, "Income" contains weekly family income in dollars, and "Temp" containing mean temperatures in Fahrenheit. We are interested in modeling ice cream consumption as a function of these date, price, income, and temperature.

Exploratory Data Analysis

During my initial exploratory data analysis, I visually inspected the relationship between ice cream consumption and the covariates. Figure [1] displays the most noticeable aspects in the data. In plot (a) on the left, there appears to be some moderate correlation between temperature and ice cream consumption. This was expected as Summer time, when temperatures are higher, is a traditional time to eat ice cream. The date also captures this as well. In the line plot (plot (b)) on the right, the time periods (1-30) appear to be correlated with each other. That is, ice cream consumption in one time period is roughly nearby that of another time period. Intuitively, this might be from seasons changing from hotter time periods to colder time periods.

Figure [2] contains the less informative plots. In plot (a) on the left, it appears as though the family income is independent of ice cream consumption. Hopefully this is an indication that people of all classes had access to ice cream in the early 1950s. Plot (b) on the right displays ice cream consumption for different prices. There appears to be some loose negative correlation, but that could also be from noise.

The Model

While trying to fit a model, I first checked multiple linear regression with and without interactions. Among the many models during this phase, the model that stood apart with the minimum Akaike's Information Criterion (AIC), only contained the single covariate of temperature. However, knowing that there may be temporal auto-correlation, I checked the ACF plot which illustrates auto-correlation of the normalized residuals at each time lag (1 time period).

In figure [3], there is clear temporal auto-correlation of about 0.6 from the first time period to one lag later at the second time period. This is a clear violation of independence in the data, so I explored other models that accounted for that.

Several models were built using different correlation functions and were then compared using AIC. The model with the auto-regressive structure of order 1 (AR-1) for the correlation had the lowest AIC value. And among all models that I checked with AR-1, the model with only the temperature as the covariate had the lowest AIC value. The final model is the following:

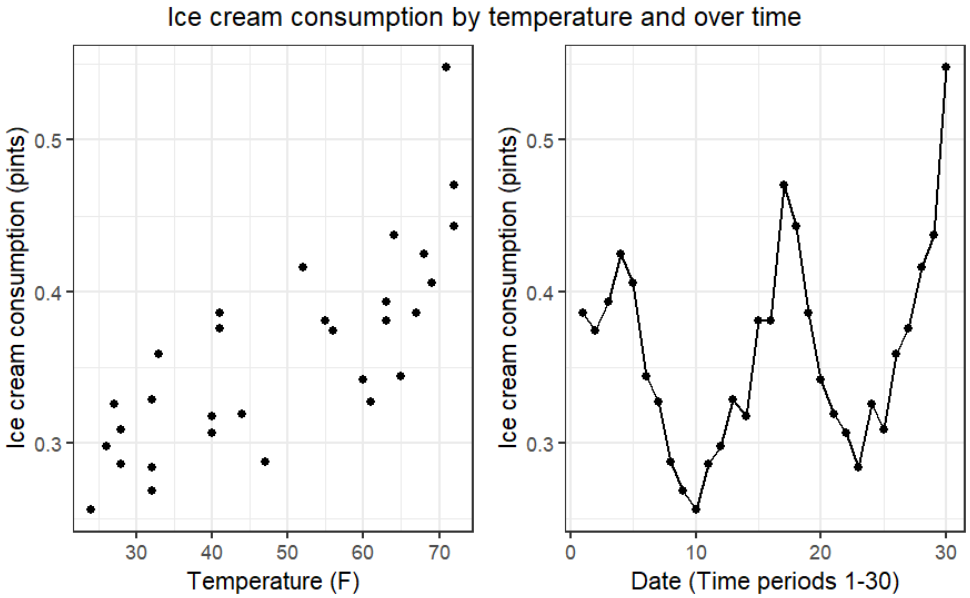

Figure 1: (a) Scatter-plot of ice cream consumption in pints versus temperature in Fahrenheit. (b) Trend of ice cream consumption over the 30 time periods.

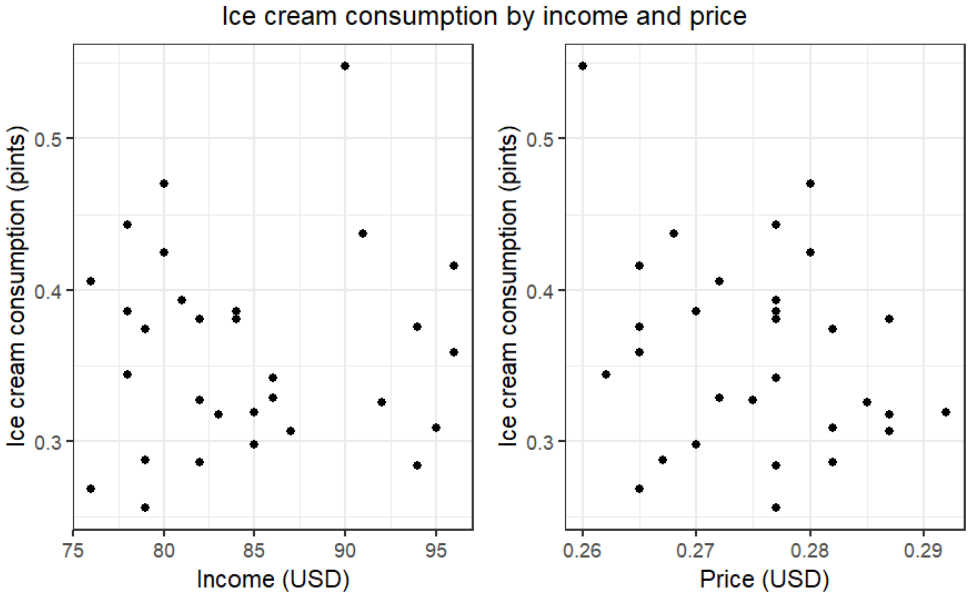

Figure 2: (a) Scatter-plot of ice cream consumption in pints versus family weekly income in dollars. (b) Scatter-plot of ice cream consumption in pints versus price per pint of ice cream in dollars.

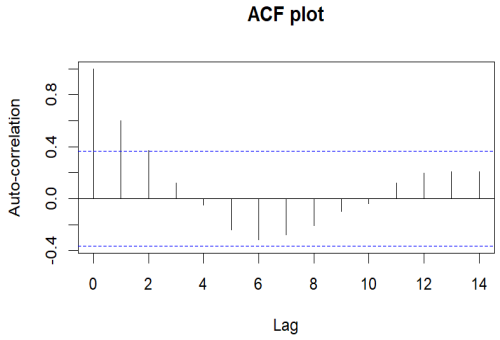

Figure 3: ACF plot with auto-correlation on the y-axis and time lag (1 time period) on the x-axis.

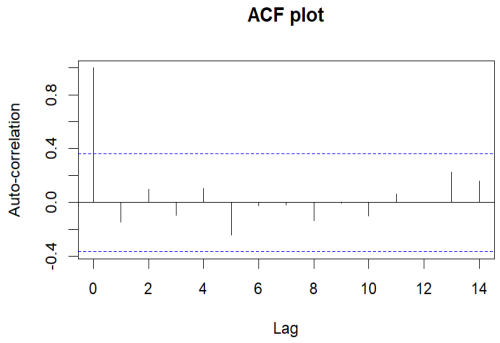

Figure 4: ACF plot with auto-correlation on the y-axis and time lag (1 time period) on the x-axis.

$$
\begin{aligned}
y_i &= \beta_0 + \beta_1 \cdot temperature_i + \epsilon_i \\
\epsilon_i &= \rho\epsilon_{i-1} + \eta_i \\
\epsilon_{i-1} &\sim N(0, \sigma^2) \\
\eta_i &\overset{\text{iid}}{\sim} N(0, \sigma_\eta^2)
\end{aligned}
$$

where $i = \{1, \ldots, 29\}$ and $temp_i$ is the number of degrees Fahrenheit above or below the average temperature.

After accounting for the auto-correlation from the date, I inspected the ACF plot again to see whether there was any improvement.

As seen in figure [4], there does not appear to be any more auto-correlation from time. I also inspected the residuals versus fitted plot as well as the QQ-plot, and there were no issues present.

Results

The estimated model is the following:

$$\hat{y}_i = 0.361 + 0.003 \cdot temp_i$$

	Value	Std. Error	p-value
Intercept	0.361	0.019	0e+00
temp	0.003	0.001	2e-04

Table 1: Intercept and slope coefficient values, standard errors, and corresponding p-values.

where $i = \{1, \ldots, 29\}$ and $temp_i$ is the number of degrees away from the average recorded temperature of about 48 degrees Fahrenheit. The estimated values along with their associated standard errors and p-values are displayed in table [1]. So, the estimated ice cream consumption at this average temperature is about 0.361 pints, and for every degree increase from the average temperature, the estimated ice cream consumption increases by an estimated 0.003 pints. With a p-value of $2e-04$, there is strong evidence that the slope coefficient associated with the temperature is not zero.

Confidence Interval

To test this models performance, I calculated the point estimate for ice cream consumption for a 30^{th} time period with a temperature of 71 degrees Fahrenheit. Then, I calculated the 90% confidence interval. To do this, I generated the fitted values and its corresponding standard error. I calculated the radius of this interval using the multiplier of 1.645 from the standard normal approximation times the aforementioned standard error. The following is the interval:

$$\hat{y} \pm 1.645\hat{SE}$$

For this 30^{th} time period, I found the ice cream consumption upper bound to be 0.455 pints and the lower bound to be 0.381 pints with 90% confidence. Unfortunately, the observed ice cream consumption was 0.548 pints for this time period and temperature, so my model fails in this way.

Appendix

```
dat <- suppressMessages(read_csv("icecream.csv"))
dat0 <- dat[-30,]
dat0$temp_c <- dat0$temp - mean(dat0$temp)

fig1 <- dat %>% ggplot(aes(y = IC, x = temp)) +
  geom_point() +
  theme_bw() +
  ylab("Ice cream consumption (pints)") +
  xlab("Temperature (F)")

fig2 <- dat %>% ggplot(aes(y = IC, x = date)) +
  geom_line() +
  geom_point() +
  theme_bw() +
  ylab("Ice cream consumption (pints)") +
  xlab("Date (Time periods 1-30)")

grid.arrange(fig1, fig2, ncol=2, top=textGrob("Ice cream consumption by temperature and over time"))

fig3 <- dat %>% ggplot(aes(y = IC, x = income)) +
  geom_point() +
  theme_bw() +
  ylab("Ice cream consumption (pints)") +
```

```
  xlab("Income (USD)")

fig4 <- dat %>% ggplot(aes(y = IC, x = price)) +
  geom_point() +
  theme_bw() +
  ylab("Ice cream consumption (pints)") +
  xlab("Price (USD)")

grid.arrange(fig3, fig4, ncol=2, top=textGrob("Ice cream consumption by income and price"))

lm1a <- gls(IC ~ date + price + income + temp,data = dat0)
lm1b <- gls(IC ~ date + price + income * temp,data = dat0)
lm1c <- gls(IC ~ date + price * income + temp,data = dat0)
lm1d <- gls(IC ~ date * price + income + temp,data = dat0)
lm1e <- gls(IC ~ date *temp + price + income ,data = dat0)
lm2a <- gls(IC ~ price + income + temp,data = dat0)
lm2b <- gls(IC ~ price + income * temp,data = dat0)
lm2c <- gls(IC ~ price * income + temp,data = dat0)
lm2d <- gls(IC ~ price * temp + income,data = dat0)
lm3a <- gls(IC ~ income + temp,data = dat0)
lm3b <- gls(IC ~ income * temp,data = dat0)
lm4 <- gls(IC ~ temp,data = dat0)

AIC(lm1a, lm1b, lm1c, lm1d, lm1e, lm2a, lm2b, lm2c, lm2d, lm3a, lm3b, lm4)
acf(residuals(lm4, type = "normalized"),
    main="ACF plot",
    ylab = "Auto-correlation")

m1 <- gls(IC ~ date + price + income + temp,
    data = dat0,
    correlation = corCompSymm(form = ~date))
m2 <- gls(IC ~ date + price + income + temp,
    data = dat0,
    correlation = corAR1(form = ~date))
m3 <- gls(IC ~ date + price + income + temp,
    data = dat0,
    correlation = corAR1(form = ~date))
m4 <- gls(IC ~ date + price + income + temp,
    data = dat0,
    correlation = corARMA(0.3, p = 0, q = 1, form = ~date))
m5 <- gls(IC ~ date + price + income + temp,
    data = dat0,
    correlation = corARMA(c(0.3, 0.2, 0.1), p = 0, q = 3))
m6 <- gls(IC ~ date + price + income + temp,
    data = dat0,
    correlation = corARMA(c(0.3, 0.2, 0.1, 0.3, 0.2, 0.1), p = 3, q = 3))


acf(residuals(m3, type = "normalized"))
AIC(m1, m2, m3, m4, m5, m6)

m3_1a <- gls(IC ~ temp + price + income + date,
    data = dat0,
    correlation = corAR1(form = ~date))
```

```
m3_1b <- gls(IC ~ temp + price + income*date,
    data = dat0,
    correlation = corAR1(form = ~date))
m3_1c <- gls(IC ~ temp + price*income + date,
    data = dat0,
    correlation = corAR1(form = ~date))
m3_1d <- gls(IC ~ temp*price + income + date,
    data = dat0,
    correlation = corAR1(form = ~date))
m3_2b <- gls(IC ~ temp * price + income,
    data = dat0,
    correlation = corAR1(form = ~date))
m3_2a <- gls(IC ~ temp + price + income,
    data = dat0,
    correlation = corAR1(form = ~date))
m3_3a <- gls(IC ~ temp  + price,
    data = dat0,
    correlation = corAR1(form = ~date))
m3_3b <- gls(IC ~ temp  * price,
    data = dat0,
    correlation = corAR1(form = ~date))
m3_4 <- gls(IC ~ temp,
    data = dat0,
    correlation = corAR1(form = ~date))
m3_5 <- gls(IC ~ temp*price,
    data = dat0,
    correlation = corAR1(form = ~date))

AIC(m3_1a, m3_1b, m3_1c, m3_1d, m3_2a, m3_2b, m3_3a, m3_3b, m3_4, m3_5)

m3_4 <- gls(IC ~ temp_c,
    data = dat0,
    correlation = corAR1(form = ~date))

acf(residuals(m3_4, type = "normalized"))

plot(x = predict(m3_4),
     y = resid(m3_4),
     pch = 16,
     xlab = "Fitted values",
     ylab = "Residuals",
     main = "Residuals vs Fitted")

qqnorm(m3_4,
       pch = 16,
       col = "black",
       main = "QQ Plot")

new.dat <- dat[-c(1:29),-c(2:4)]

pred <- predictSE.gls(m3_4, new.dat)
pred$fit + c(-1,1)*1.645*pred$se.fit
```