Task: Linear models

Hongjin Lyu B00978648

2025-02-25

In this task you will practice fitting linear models and adding them to visualizations.

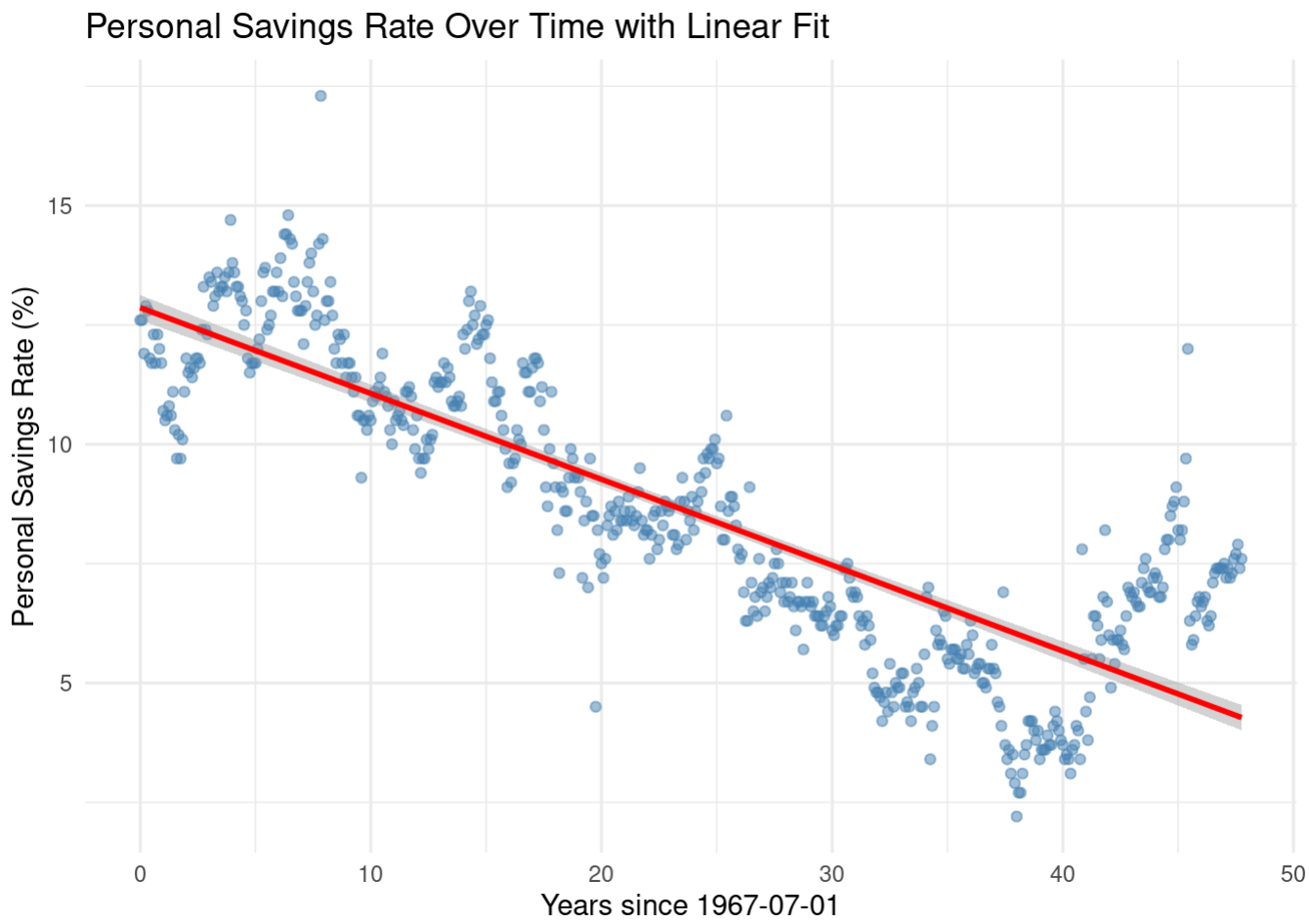
Question 1

The `economics` dataset in the `ggplot2` package reports several economic indicators for the USA monthly from 1967-07-01 to 2015-04-01. The data include `pop` (total population, in thousands), `unemploy` (the number of unemployed persons, in thousands), `uempmed` (median duration of unemployment in weeks), and `psavert` (the personal savings rate). For more information, see the help page.

Plot the personal savings rate as a function of time. Add a regression line (straight line, `y ~ x`).

```
my_econ <- economics %>%
  mutate(
    dd = decimal_date(date),
    years = dd - min(dd)
  )


ggplot(my_econ, aes(x = years, y = psavert)) +
  geom_point(alpha = 0.5, color = "steelblue") +
  geom_smooth(method = "lm", formula = y ~ x,
              color = "red", se = TRUE) +
  labs(
    x = "Years since 1967-07-01",
    y = "Personal Savings Rate (%)",
    title = "Personal Savings Rate Over Time with Linear Fit"
  ) +
  theme_minimal()
```



I have computed the time since the start of the time series and called it `years`. Use `lm` to create a straight line regression line. Display a table showing the intercept and slope of the regression line, with a confidence interval on these parameters.

```
my_econ <- economics %>% mutate(dd = decimal_date(date),
                                years = dd - min(dd))
lm_fit <- lm(psavert ~ years, data = my_econ)

tidy(lm_fit, conf.int = TRUE) %>%
  kable(
    digits = 3,
    caption = "Regression coefficients and 95% confidence intervals",
    col.names = c("Term", "Estimate", "Std. Error", "Statistic", "p-value", "CI Lower", "CI Upp
er")
  ) %>%
  kable_styling(
    bootstrap_options = "striped",
    full_width = FALSE,
    position = "center"
  )
```

Regression coefficients and 95% confidence intervals

Term	Estimate	Std. Error	Statistic	p-value	CI Lower	CI Upper
(Intercept)	12.862	0.135	95.474	0	12.598	13.127
years	-0.180	0.005	-36.828	0	-0.189	-0.170

How well does a straight line represent this data? Describe any concerns you have about this representation of the data.

Write your response here. The straight line model captures fluctuations (e.g., peaks around 1975 and 2008) that a straight line cannot capture. But, a few extreme values (e.g., the spike around 2012) may disproportionately influence the slope estimate.

Question 2

Sometimes economists report a “seasonally adjusted” rate of unemployment, in recognition that there are seasonal patterns of employment in some sectors of the economy that cause regular, predictable fluctuations in employment and unemployment.

Below I compute the day of the year for each observation (discarding the actual year) and the unemployment rate. Plot the unemployment rate as a function of the day of the year. Add quantile regression lines for the 25, 50, and 75th quantiles.

```
my_econ2 <- my_econ |> mutate(yday = yday(date),
                              urate = unemploy/pop)
ggplot(my_econ2, aes(x = yday, y = urate)) +
  geom_point(alpha = 0.3, color = "steelblue") +
  geom_quantile(quantiles = c(0.25, 0.5, 0.75),
                formula = y ~ x,
                color = "red", linewidth = 1) +
  labs(x = "Day of the Year", y = "Unemployment Rate") +
  theme_minimal()
```