5.10 Exercises

1. Daily electricity demand for Victoria, Australia, during 2014 is contained in `elecdaily` . The data for the first 20 days can be obtained as follows.

```
daily20 <- head(elecdaily,20)
```

a. Plot the data and find the regression model for Demand with temperature as an explanatory variable. Why is there a positive relationship?

b. Produce a residual plot. Is the model adequate? Are there any outliers or influential observations?

c. Use the model to forecast the electricity demand that you would expect for the next day if the maximum temperature was $15°$ and compare it with the forecast if the with maximum temperature was $35°$. Do you believe these forecasts?

d. Give prediction intervals for your forecasts. The following R code will get you started:

```
autoplot(daily20, facets=TRUE)
daily20 %>%
  as.data.frame() %>%
  ggplot(aes(x=Temperature, y=Demand)) +
  geom_point() +
  geom_smooth(method="lm", se=FALSE)
fit <- tslm(Demand ~ Temperature, data=daily20)
checkresiduals(fit)
forecast(fit, newdata=data.frame(Temperature=c(15,35)))
```

e. Plot Demand vs Temperature for all of the available data in `elecdaily` . What does this say about your model?

2. Data set `mens400` contains the winning times (in seconds) for the men's 400 meters final in each Olympic Games from 1896 to 2016.

   a. Plot the winning time against the year. Describe the main features of the plot.

   b. Fit a regression line to the data. Obviously the winning times have been decreasing, but at what *average* rate per year?

   c. Plot the residuals against the year. What does this indicate about the suitability of the fitted line?

   d. Predict the winning time for the men's 400 meters final in the 2020 Olympics. Give a prediction interval for your forecasts. What assumptions have you made in these calculations?

3. Type `easter(ausbeer)` and interpret what you see.

4. An elasticity coefficient is the ratio of the percentage change in the forecast variable (yy) to the percentage change in the predictor variable (xx). Mathematically, the elasticity is defined as $(dy/dx) \times (x/y)$ $(dy/dx) \times (x/y)$. Consider the log–log model,

$$\log y = \beta_0 + \beta_1 \log x + \varepsilon.$$

Express yy as a function of xx and show that the coefficient $\beta_1 \beta_1$ is the elasticity coefficient.

5. The data set `fancy` concerns the monthly sales figures of a shop which opened in January 1987 and sells gifts, souvenirs, and novelties. The shop is situated on the wharf at a beach resort town in Queensland, Australia. The sales volume varies with the seasonal population of tourists. There is a large influx of visitors to the town at Christmas and for the local surfing festival, held every March since 1988. Over time, the shop has expanded its premises, range of products, and staff.

   a. Produce a time plot of the data and describe the patterns in the graph. Identify any unusual or unexpected fluctuations in the time series.

   b. Explain why it is necessary to take logarithms of these data before fitting a model.

   c. Use R to fit a regression model to the logarithms of these sales data with a linear trend, seasonal dummies and a "surfing festival" dummy variable.

   d. Plot the residuals against time and against the fitted values. Do these plots reveal any problems with the model?

   e. Do boxplots of the residuals for each month. Does this reveal any problems with the model?

   f. What do the values of the coefficients tell you about each variable?

   g. What does the Breusch–Godfrey test tell you about your model?

h. Regardless of your answers to the above questions, use your regression model to predict the monthly sales for 1994, 1995, and 1996. Produce prediction intervals for each of your forecasts.

i. Transform your predictions and intervals to obtain predictions and intervals for the raw data.

j. How could you improve these predictions by modifying the model?

6. The `gasoline` series consists of weekly data for supplies of US finished motor gasoline product, from 2 February 1991 to 20 January 2017. The units are in "million barrels per day." Consider only the data to the end of 2004.

a. Fit a harmonic regression with trend to the data. Experiment with changing the number Fourier terms. Plot the observed gasoline and fitted values and comment on what you see.

b. Select the appropriate number of Fourier terms to include by minimising the AICc or CV value.

c. Check the residuals of the final model using the `checkresiduals()` function. Even though the residuals fail the correlation tests, the results are probably not severe enough to make much difference to the forecasts and prediction intervals. (Note that the correlations are relatively small, even though they are significant.)

d. To forecast using harmonic regression, you will need to generate the future values of the Fourier terms. This can be done as follows.

```
fc <- forecast(fit, newdata=data.frame(fourier(x,K,h)))
```

where `fit` is the fitted model using `tslm()`, `K` is the number of Fourier terms used in creating `fit`, and `h` is the forecast horizon required.

Forecast the next year of data.

e. Plot the forecasts along with the actual data for 2005. What do you find?

7. Data set `huron` gives the water level of Lake Huron in feet from 1875 to 1972.

a. Plot the data and comment on its features.

b. Fit a linear regression and compare this to a piecewise linear trend model with a knot at 1915.

    c. Generate forecasts from these two models for the period up to 1980 and comment on these.

8. *(For advanced readers following on from Section 5.7).*

Using matrix notation it was shown that if $y = X\beta + \varepsilon$, where e has mean 0 and variance matrix $\sigma^2 I$, the estimated coefficients are given by $\hat{\beta} = (X'X)^{-1}X'y$ and a forecast is given by $\hat{y} = x^* \hat{\beta} = x^*(X'X)^{-1}X'y$ where x^* is a row vector containing the values of the regressors for the forecast (in the same format as X), and the forecast variance is given by $var(\hat{y}) = \sigma^2 \left[1 + x^*(X'X)^{-1}(x^*)'\right]$.

Consider the simple time trend model where $y_t = \beta_0 + \beta_1 t$. Using the following results,

$$\sum_{t=1}^{T} t = \frac{1}{2}T(T+1), \quad \sum_{t=1}^{T} t^2 = \frac{1}{6}T(T+1)(2T+1)$$

derive the following expressions:

a. $X'X = \dfrac{1}{6}\begin{bmatrix} 6T & 3T(T+1) \\ 3T(T+1) & T(T+1)(2T+1) \end{bmatrix}$

b. $(X'X)^{-1} = \dfrac{2}{T(T^2-1)}\begin{bmatrix} (T+1)(2T+1) & -3(T+1) \\ -3(T+1) & 6 \end{bmatrix}$

c. $\hat{\beta}_0 = \dfrac{2}{T(T-1)}\left[(2T+1)\sum_{t=1}^{T} y_t - 3\sum_{t=1}^{T} ty_t\right]$

$\hat{\beta}_1 = \dfrac{6}{T(T^2-1)}\left[2\sum_{t=1}^{T} ty_t - (T+1)\sum_{t=1}^{T} y_t\right]$

d. $\text{Var}(\hat{y}_t) = \hat{\sigma}^2\left[1 + \dfrac{2}{T(T-1)}\left(1 - 4T - 6h + 6\dfrac{(T+h)^2}{T+1}\right)\right]$