Prelims Stats: Rdemo-2 TT 2023

Old Faithful data

https://en.wikipedia.org/wiki/Old_Faithful

Old Faithful is a geyser in Yellowstone National Park, USA. To look at the first few rows of the dataset `faithful`, which is built-in to R, use

```
head(faithful)
```

The two columns in the dataset are:

- `eruptions`, the duration of eruptions of the geyser (in minutes) – take this as explanatory variable x
- `waiting`, the waiting time until the next eruption of the geyser (in minutes) – take this as response variable y. [To see the help page for the dataset which gives these details, use `?faithful`]

Suppose we are interested in using the duration of the current eruption (i.e. `eruptions`) to predict the time until next eruption takes place (i.e. `waiting`).

Plot the data, fit a simple linear regression, and draw the regression line on the plot:

```
plot(waiting ~ eruptions, data=faithful)
fit1 <- lm(waiting ~ eruptions, data=faithful)
fit1

# add the fitted line to the plot
abline(fit1, col="red")
```

The fitted line is $y = 33.47 + 10.73x$. If the most recent eruption lasted 4 minutes, how long does the model predict until the next eruption?

The function `fitted()` extracts the fitted values $\widehat{y}_i$. The function `resid()` extracts the residuals e_i. The function `rstandard()` extracts the residuals r_i, sometimes called "standardised" residuals, which are called "studentised" residuals in the synopses and in JWHT (2013).

Residual plots:

```
plot(resid(fit1) ~ fitted(fit1))
# plot(resid(fit1) ~ eruptions, data=faithful)
# the same plot but different horizontal scale

# similar plot but use the standardised/studentised residuals
plot(rstandard(fit1) ~ fitted(fit1))
```

Auto data

The `Auto` dataset used in lectures is part of the `ISLR` package. To install this package, use

```
install.packages("ISLR")
```

If you are using RStudio this should download and install the package in a few seconds. [If you are using R, not RStudio, then you may get pop-up box asking you to choose a mirror site to download the package from – choose e.g. "UK (Bristol)" and then the package should download and install.]

Load the package and look at the first few rows of the dataset:

```
library(ISLR)
head(Auto)
```

Plot the data, fit a straight line regression, and draw the regression line on the plot:

```
plot(mpg ~ horsepower, data=Auto)
fit2 <- lm(mpg ~ horsepower, data=Auto)
```

```
fit2
abline(fit2, col="red")
```

The curved nature of the following residual plot indicates that a straight line model isn't really appropriate:

```
plot(rstandard(fit2) ~ fitted(fit2))
```

Try a model of the form $y = \beta_0 + \beta_1 x + \beta_2 x^2 + \epsilon$:

```
fit3 <- lm(mpg ~ horsepower + I(horsepower^2), data=Auto)
fit3
```

Note: for the quadratic term to be included as we intend, the identity function `I()` is needed above. Note also: when using the `lm()` function above, we have specified the linear term `horsepower` and the quadratic term `I(horsepower^2)`, and R automatically includes an intercept term. [We can omit an intercept term by using a "-1", say using `lm(y ~ -1 + x)`, but we rarely want to do this.]

Plot the data, the fitted straight line, and the fitted quadratic:

```
plot(mpg ~ horsepower, data=Auto)
abline(fit2, col="red")
curve(56.90 - 0.4662*x + 0.0012*x^2, col="blue", add=TRUE)
```

Repeat the previous residual plot but for the quadratic model to see how/if the quadratic term has improved the behaviour of the residuals:

```
plot(rstandard(fit3) ~ fitted(fit3))
```

Forbes data

Install the package `MASS`, load it, and look at the `forbes` dataset:

```
install.packages("MASS")
library(MASS)
forbes
?forbes
```

The 19th-century physicist James Forbes had a theory that suggested that log(pressure) (i.e. the log of `pres`) was linearly related to temperature (`bp`).

Try a linear model fit – before plotting the fitted line, does a linear model look like a good idea?

```
plot(log(pres) ~ bp, data=forbes)
fit4 <- lm(log(pres) ~ bp, data=forbes)
fit4
abline(fit4, col="red")
```

Once the fitted line is included the problem should be apparent. The following plot makes the problem point (data point #12) even more obvious:

```
plot(rstandard(fit4) ~ fitted(fit4))
```

Olympics data

The following will read-in data (from Rogers and Girolami, 2012) of winning times in the 100 metres at Olympics in 2008 and earlier. Try fitting a linear model to each dataset and hence predicting the winning times for 2012 and 2016.

You should be able to cut-and-paste this into R:

```
femaleyear <- c(1928, 1932, 1936, 1948, 1952, 1956, 1960, 1964, 1968,
                1972, 1976, 1980, 1984, 1988, 1992, 1996, 2000, 2004, 2008)

female100 <- c(12.2, 11.9, 11.5, 11.9, 11.5, 11.5, 11, 11.4, 11, 11.07,
```

```
             11.08, 11.06, 10.97, 10.54, 10.82, 10.94, 11.12, 10.93, 10.78)

maleyear <- c(1896, 1900, 1904, 1906, 1908, 1912, 1920, 1924, 1928,
              1932, 1936, 1948, 1952, 1956, 1960, 1964, 1968, 1972,
              1976, 1980, 1984, 1988, 1992, 1996, 2000, 2004, 2008)

male100 <- c(12, 11, 11, 11.2, 10.8, 10.8, 10.8, 10.6, 10.8,
             10.3, 10.3, 10.3, 10.4, 10.5, 10.2, 10, 9.95, 10.14,
             10.06, 10.25, 9.99, 9.92, 9.96, 9.84, 9.87, 9.85, 9.69)

plot(female100 ~ femaleyear)
plot(male100 ~ maleyear)
```

Matlab

In each case you will first need to read-in the data. Download `faithful.txt`, `Auto.txt`, `forbes.txt`, `olympicsf.txt`, `olympicsm.txt` from `http://www.stats.ox.ac.uk/~laws/data/`

I saved them to folder `D:\Prelims`. Then, for example:

```
%%%%% Old Faithful data
cd D:\Prelims
faithful = readtable('faithful.txt');
y = faithful.waiting;
x = faithful.eruptions;
scatter(x, y)

% find coefficients b of linear model
X = [ones(length(x),1) x];
b = X\y

residuals = y - (33.47 + 10.73*x);
scatter(33.47 + 10.73*x, residuals)

%%%%% Auto data
cd D:\Prelims
Auto = readtable('Auto.txt');
y = Auto.mpg;
x = Auto.horsepower;
scatter(x, y)

% find coefficients of quadractic model
X = [ones(length(x),1) x x.^2];
b = X\y

%%%%% Forbes data
cd D:\Prelims
forbes = readtable('forbes.txt')
% etc

%%%%% Olympics data
cd D:\Prelims
olympicsf = readtable('olympicsf.txt');
olympicsm = readtable('olympicsm.txt');
% etc
```