Case: PCB

Story

The concentration of polychlorinated biphenyl (PCB) residues in a series of lake trout from Cayuga Lake, NY, were reported in Bache et al. (1972). The ages of the fish were accurately known, because the fish were annually stocked as yearlings and distinctly marked as to year class. Each whole fish was mechanically chopped, ground, and thoroughly mixed, and 5-gram samples taken. The samples were treated and PCB residues in parts per million (ppm) were estimated using column chromatography. Data are available in the file `troutpcb.txt`.

The main objective of this case is to obtain a reasonable description (model) of the relation between age of trouts in lake Cayuga and PCB concentration. This includes model fitting, model diagnostics, quantification of uncertainty, communication of the model, and illustration of the model on the data.

Data

Variable	Description
PCB	PCB concentration in ppm
Age	Age of the trout in years

Exercise

1. Download `troutpcb.txt` from Campus Net and save it in a new empty folder. Open RStudio or another R editor of your choice and change the *working directory* to the directory where you saved `troutpcb.txt`. In RStudio you go to `Session` → `Set Working Directory` and navigate to relevant directory.

   Now open a new R script (In RStudio go to `File` → `New File` → `R Script`). Now save the file in the same place as `troutpcb.txt`.

2. Read the data into R with something like

```
pcb <- read.table("troutpcb.txt", header=TRUE, sep="\t")
```

   Now inspect the data and see that they have been read in correctly:

```
head(pcb)
str(pcb)
summary(pcb)
```

3. Plot the data using something like

```
plot(PCB ~ Age, data=pcb)
```

4. Fit the linear regression model and add the fit to the plot:

```r
fm <- lm(PCB ~ Age, data=pcb)
abline(fm)
```

Now `fm` holds the *linear model* object and you can for instance extract the coefficients, i.e. parameter estimates with `coef(fm)`.

What is your impression of the model fit — do you think you have found a reasonable model?

5. Perform model diagnostics and plot the residuals:

```r
par(mfrow=c(2, 2)) ## split plotting region in 4
plot(fm, which=1:4)
```

Realize that there is a serious problem with this fit: Argue that the variance is not constant and that a log-transform could remedy the problem.

6. Take the logarithm of PCB concentration and save the variable in the data frame:

```r
pcb$log10PCB <- log10(pcb$PCB)
## Even better R code:
## pcb <- within(pcb, log10PCB <- log10(PCB))
```

Argue that $\log_{10}$ is better than $\log_e$ (Hint: consider interpretation of the axes in a plot of age versus log PCB concentration).

7. Illustrate log PCB versus age, fit the model using log PCB and check the residuals again:

```r
par(mfrow=c(1, 1))
plot(log10PCB ~ Age, data=pcb)
fm2 <- lm(log10PCB ~ Age, data=pcb)
abline(fm2)
```

```r
par(mfrow=c(2, 2))
plot(fm2, which=1:4)
```

Argue that the residuals are much better behaved. Is there evidence of variance heterogeneity? There is one observation which might be influential; which one?

8. The relation between PCB and age now has the following form:

$$\log_{10} PCB = \hat{\beta}_0 + \hat{\beta}_1 Age$$

Taking the anti-log on both sides, we obtain

$$\begin{aligned} PCB &= 10^{\hat{\beta}_0 + \hat{\beta}_1 Age} \\ &= 10^{\hat{\beta}_0} \cdot 10^{\hat{\beta}_1 Age} \\ &= \hat{\beta}_0' \cdot \hat{\beta}_1'^{Age} \quad \text{where} \quad \hat{\beta}_0' = 10^{\hat{\beta}_0}, \ \hat{\beta}_1' = 10^{\hat{\beta}_1} \end{aligned}$$

Compute the coefficients and insert the numbers for $\hat{\beta}_0'$ and $\hat{\beta}_1'$. How many decimals do you think are appropriate here?

9. Observe that while the model for $\log PCB$ is additive, the resulting model for PCB is *multiplicative*!

   The results of a multiplicative model can be interpreted as

   - PCB concentration increases with XX% per year
   - The rate of increase is between YY and ZZ percent per year based on a 95% confidence interval.

   Compute the right numbers for XX, YY and ZZ.

   Note that if L and U are the lower and upper confidence limits for $\hat{\beta}_j$, then 10^L and 10^U are appropriate confidence limits for $10^{\hat{\beta}_j}$. Observe that while the CI is symmetric for $\hat{\beta}_j$, it is *asymmetric* for $10^{\hat{\beta}_j}$ — uncertainty is not symmetric in the multiplicative model.

10. Suppose you are writing a paper based on these analyses and want to include a table of your findings. You compute the following:

```
B <- coef(summary(fm2))
df <- data.frame(Estimate = round(10^B[, 1], 2),
                 Lower = round(10^confint(fm2)[1, ], 2),
                 Upper = round(10^confint(fm2)[2, ], 2),
                 "p-value" = format.pval(B[, 4], digits=3, eps=1e-3))
df
```

   What does the `format.pval` function do?

   The p-value measures evidence against the null hypothesis: $H_0 : \ \beta_1' = X$. What is the value of X here?

   To communicate your results verbally, you write that "There is xxxx evidence that the rate of increase in PCB different from X." Choose xxxx among *no*, *little*, *some* or *strong*.

   You may also want to add something like this to your paper:

   - The rate of increase is significantly different from WW (p-value$?KKK$; $t = JJJ$; $df = NN$)

   Substitute the relevant numbers. Here "?" can be one of "<", "=", or ">".

11. Illustrate the model on the original scale of measurement.

   First plot PCB versus age:

```
plot(PCB ~ Age, data=pcb, ylim=c(0, 35),
     xlab="Age [years]", ylab="PCB concentration [ppm]", bty="n",
     las=1)
```

   Then compute the model fit and confidence intervals for the fitted model:

```
xval <- seq(1, 12, length=500)
pred <- predict(fm2, newdata=data.frame(Age=xval),
                interval="confidence")
lines(xval, 10^pred[, "fit"], lwd=2)
lines(xval, 10^pred[, "lwr"], col="red", lwd=2, lty=2)
lines(xval, 10^pred[, "upr"], col="red", lwd=2, lty=2)
```

You can also add *prediction* intervals to the figure:

```
pred <- predict(fm2, newdata=data.frame(Age=xval),
                interval="prediction")
lines(xval, 10^pred[, "lwr"], col="blue", lty=3, lwd=2)
lines(xval, 10^pred[, "upr"], col="blue", lty=3, lwd=2)
```

Finally you may want to add a legend:

```
legend("topleft", legend=c("Fit", "95% Confidence interval",
                "95% Prediction interval"),
       lwd=2, col=c("black", "red", "blue"), lty=1:3, bty="n")
```

12. Save the code that performs your analysis in a script and add plenty of comments to your code.

Extra exercises

1. Investigate the sensitivity to the outlier identified previously. Do that by excluding this point from the data, then refit the model and compare the estimates of β_1' and their 95% confindence intervals for the two models. How large is the change in the coefficient relative to the statistical uncertainty?

   Perform model diagnostics on this model; do the residuals behave better or worse?

2. Another remedy is to transform `Age` using `Age`$^{\texttt{c}}$. We could choose `c` from the data, but here we will use the cube root transform, i.e. $Age^{1/3} = \sqrt[3]{Age}$. Fit the model for $log_{10}PCB$ using $\sqrt[3]{Age}$ instead of just Age and inspect the residuals. What is your assessment of the residual diagnostics now?

3. Visualize the model uncertainty by displaying the four different model fits on the same graph. The four models are:

   (a) The linear fit: $PCB \propto Age$

   (b) The log-linear model: $\log_{10} PCB \propto Age$

   (c) The log-linear model with the influential observation excluded.

   (d) The log-linear model using the transformed age: $\log_{10} PCB \propto \sqrt[3]{Age}$

   Which model would you place most trust in?

4. (Difficult) Interpretation of the model with $\log_{10} PCB \propto \sqrt[3]{Age}$ is hampered by the fact that rate of increase in PCB with Age is not constant. Derive and compute the rate of increase in PCB as a function of age and plot the relation. Hint: we can interpret the rate of increase as the slope of the relation between log PCB and age.

   Illustrate the rate of increase in Age as a function of PCB for the log-linear models with Age and $\sqrt[3]{Age}$