5 Regression

5.1 Simple and multiple linear regression: Cherry data

Prerequisites: Exercises 3.1 and 3.3 (reading and working with the cherry data)

Consider the dataset `cherry` from Exercise 3.1. We will fit an examine a linear regression modelling the expected value of volume as a linear function of girth.

1. (*Fit of model and plots*) Use the following commands:

   ```
   plot(cherry$Girth, cherry$Volume)
   lm(Volume ~ Girth, data=cherry)
   ```

   This gives you the estimated intercept and slope. Notice how we specify the dataset; then it is not necessary to attach the data or use the $ syntax (cf. Exercise 3.3.) It is often convenient to give the object with the fit a name. Try

   ```
   linreg1 <- lm(Volume ~ Girth, data=cherry)
   linreg1
   abline(linreg1)
   ```

2. (*Estimates, standard errors, tests*) We need more information than just the estimated coefficients. Try

   ```
   summary(linreg1)
   ```

   Some explanation of the output is probably needed. The "Coefficients" part is the most important, and it is important to understand its structure. It has two lines — one for each parameter in the model — and four columns. The first line is about the intercept, the second line is about the slope. The columns are

   - `Estimate`: The estimated value of the parameter (intercept or slope)
   - `Std. Error`: The standard error (estimated standard deviation) associated with the estimate in the first column
   - `t value`: The t-test statistic for the hypothesis that the corresponding parameter is zero. Computed as the estimate divided by the standard error.
   - `Pr(>|t|)`: The p-value associated with the hypothesis just mentioned. In particular the p-value in the second line is for the hypothesis that there is no effect of girth on volume.

   Below the "Coefficients" part you find, among others, the "Residual standard error", *i.e.* the estimated standard deviation for the observations.

   Identify the estimates and corresponding standard errors in the output. Is there a significant effect of girth on volume? What is the estimated standard deviation for the observations?

3. (*Confidence intervals*) Try the command `confint(linreg1)`, which gives you 95% confidence intervals for the intercept and slope.

4. (*Model validation*) The easiest way to make model validation plots for the model fit is as follows:

   ```
   par(mfrow=c(2,2))
   plot(linreg1)
   ```

   Notice how the command `par(mfrow=c(2,2))` splits the plot window into 4 subplots.

   Does the model seem to be appropriate for the data?

   Fitted values, raw residuals and standardized residuals from the fit are extracted with the functions `fitted`, `residuals` and `rstandard`, respectively, so the classical model validation plots can also be obtained as follows (the commands with `abline` add relevant lines to the plots):

   ```
   plot(fitted(linreg1), residuals(linreg1))
   abline(h=0)
   plot(fitted(linreg1), rstandard(linreg1))
   abline(h=0)
   qqnorm(rstandard(linreg1))
   abline(0,1)
   ```

5. (*Transformation*) Fit a new linear regression model where you use `log(Volume)` as the response variable and `log(Girth)` as covariate.

   Is the model appropriate for the data? Is the effect of `log(Girth)` significant? What is the estimated relation between `log(Girth)` and `log(Volume)`? Which relation between `Girth` and `Volume` does this correspond to?

6. (*Multiple linear regression*) Fit a multiple linear regresison model where `log(Girth)` as well as `log(Height)` are used as covariate as follows:

   ```
   linreg3 <- lm(log(Volume) ~ log(Girth) + log(Height), data=cherry)
   ```

   Are both covariates significant (use `summary`)? Finally try the command

   ```
   anova(linreg3, linreg2)
   ```

   which carries out the F-test for comparison of the two models. Did you see the p-value before (explain where and why)?

5.2 Simple linear regression, prediction: Heart and body weights*

Prerequisites: Exercises 3.3 (working with datasets) and 5.1 (linear regression)

1. In the R package `MASS` there is a dataset called `cats`. Run the following commands:

```
library(MASS)
data(cats)
```

Have a look at the dataset. The variables `Bwt` and `Hwt` give the weight of the body (kg) and the heart (g), respectively. There are both male and female cats. Make a dataset with the data from males only.

2. Make a scatterplot of the data for the male cats (`Bwt` on x-axis, `Hwt` on y-axis). Does it look reasonable to use a linear regression model for the data?

3. Fit a linear regresison model for the male cats, that allows for prediction of the heart weight given the body weight. Add the fitted regression line to the scatterplot from the previous question.

4. Find the coefficients of the fitted line. How large is the expected difference in heart weight for two cats with a difference of 1 kg in bodyweight? Find a confidence interval for this difference? How large is the expected difference in heart weight for two cats with a difference of 100 g in bodyweight?

5. Use model validation plot to examine if the model is appropriate for the data.

6. Use the estimates to find the expected heart weight for a male cat that weighs 3 kg. Then try the commands (where you replace the name `regModel` with whatever name you gave the the model fit in question 2).

```
newObs <- data.frame(Bwt=3)
newObs
predict(regModel, newObs)
predict(regModel, newObs, interval="predict")
```

5.3 Multiple linear regression: Toxicity of dissolutions*

Prerequisites: Exercises 3.1, 3.3 (reading and working with data) and 5.1 (linear regression)

Data from 24 chemical dissolutions have been collected in order to examine the association between the toxicity of the dissolution on the one side and three explanatory variables on the other side. The data are saved in the files `lser.xlsx` and `lser.csv` with the following variables:

- `tox`: Toxicity of the dissolution
- `base`: Ablility to accept hydrogen ions
- `acid`: Ablility to liberate hydrogen ions
- `colour`: Ability to change colour

1. Make an R dataset called `lser` with the data, and use the command `plot(lser)` to get an overview of the data.

2. Fit a multiple linear regression model with `tox` as response and `base`, `acid` and `colour` as explanatory variables. Is the model appropriate for the data? What is the interpretation of the parameter estimates?

3. Are all three explanatory variables significant? Remove insignificant variables (one at a time) until all terms are significant.

4. Calculate the expected toxicity for a solvent which has `base=0.60`, `acid=0.95`, and `colour=0.52`.

5.4 Nonlinear regression: Enzyme kinetics

Prerequisites: Exercises 3.1, 3.3 (reading and working with data), and 4.1 (scatter plot)

In a chemical experiment the enzyme activity has been measured for different concentrations of the substrate and different concentrations of an inhibitor. The enzyme activity is measured as a reaction rate. There are three measurement series corresponding to three concentrations of the inhibitor (no inhibitor, 50 μM, 100 μM). For each series the reaction rate has been measured for 6 different substrate concentrations ranging from 10 μM to 600 μM. There are two replications and therefore 36 observations in total.

Data are available in the files `inhib.xlsx` and `inhib.csv` with the following variables:

- `S`: Concentration of substrate
- `I`: Concentration of inhibitor
- `R`: Reaction rate

1. Make a scatterplot of the data with `S` on the x-axis and `R` on the y-axis. Why is this plot not very illustrative? Try the commands

```
grp <- c(rep(1,times=12), rep(2,times=12), rep(3,times=12))
plot(S, R, col=grp)
plot(S, R, pch=grp)
```

When there is no inhibitor, the association between substrate concentration and reaction rate is most often described by the so-called Michaelis-Menten relation:

$$R \approx \frac{V_{\max} \cdot S}{K + S} \tag{1}$$

where $V_{\max}$ and K are parameters that should be estimated from the data. This is typically done with least squares. The function `nls` can do this, but needs starting values for the parameters.

2. Make a dataset, `dat0`, which only contains the data with no inhibitor (`I`=0). Then try the commands

```
mm0 <- nls(R ~ Vmax * S / (K + S), start = list(Vmax=3, K=100), data=dat0)
summary(mm0)
```

Make sure you understand the output.

3. It is not possible to extract standardized residuals from a `nls` fit, so we have to evaluate the raw residuals instead. The residualplot with raw residuals against fitted values is made in the usual way:

   ```
   plot(fitted(mm0), residuals(mm0))
   ```

   Does the model seem to be appropriare for the data?

4. Make a scatterplot for the data with no inhibitor. Then try the following commands:

   ```
   f <- function(S) 2.9811 * S / (35.802 + S)
   plot(f, from=0, to=620, add=T)
   ```

   The first command defines the fitted function, and the second adds a graph to the scatterplot.

Now we want to make fit a model to the complete dataset. Consider the association

$$R \approx \frac{V_{\max} \cdot S}{K_1 \cdot (1 + I/K_2) + S} \tag{2}$$

between inhihitor concentration, substrate concentration and reaction rate. Here $V_{\max}$, K_1, and K_2 are parameters to be estimated from the data.

5. Fit the model with `nls`. You can for example use $V_{\max} = 3$, $K_1 = 100$, and $K_2 = 25$ as starting values. What are the estimated coefficients?

6. Make the scatterplot from question 1 again, with different colours for different inhihitor concentrations. Add three fitted curves to the plot, one for each inhibitor concentration. Use the same colours for the graphs as you did for the datapoints.

7. Alternatively, we could use separate Michaelis-Menten models for each inhibitor concentration, *i.e.* use (1), but with three different values of $V_{\max}$ and K. This model can be fitted with the command

   ```
   mm2 <- nls(R ~ Vmax[grp] * S / (K[grp] + S),
              start = list(Vmax=c(3,3,3), K=c(100,100,100)), data=inhib)
   ```

   Notice that starting values are needed for each parameter in the model. Fit the model.

8. The model given by (2) is nested in the model just fitted (explain why), and the models can be compared with an F-test using `anova`. If `mm1` is the model from question 5, then try

   ```
   anova(mm1, mm2)
   ```

   Is the model given by (2) appropriate for the data?

5.5 Logistic regression: Pneumoconiosis among coalminers

Prerequisites: None (although Exercise 5.1 is perhaps useful for understanding the output)

Binary variables are variables with two possible outcomes, for example dead or alive, healthy or sick, germinated or not. Logistic regression is relevant when the response variable is binary. The aim is to relate the probabilities of the two outcomes to one or more explanatory variables. In the simplest case with one continuous covariate, x, and binary response, y, the logistic regression model assumes that log-odds is a linear function of x. Let $p = P(y = 1)$, and recall that the odds is defined as $P(y = 1)/P(y = 0) = p/(1 - p)$. Then the assumption is

$$\log\left(\frac{p}{1-p}\right) = \alpha + \beta x. \tag{3}$$

Just as in ordinary linear regression (with continuous response) the first aim is to estimate the parameters α and β.

This exercise is about pneumoconiosis among coalface workers. Data were collected in order to examine the relationship between exposure time (years) and risk of disease. Severity of disease was originally rated into three categories, but here we will use only two (normal and diseased):

Exposure time	Normal	Diseased
5.8	98	0
15	51	3
21.5	34	9
27.5	35	13
33.5	32	19
39.5	23	15
46	12	16
51.5	4	7

The data are saved in the files `coalworker1.xlsx` and `coalworker1.csv`.

Notice how, for each person, it has been observed whether he was diseased or not. This corresponds to a binary variable. In the following we will consider a model with log-exposure as covariate, *i.e.* with `log(exposure)` as the x-variable in (3).

1. Read the data into R; call the dataset `coalworker1`. Then try the following commands and make sure you understand what happens:

   ```
   total <- normal + diseased
   relativeFreq <- diseased/total
   logOdds <- log(relativeFreq / (1-relativeFreq))
   plot(log(exposure), logOdds)
   ```

   What does the plot tell you about the appropriateness of the logistic regression model with log-exposure as explanatory variable?

2. Fit the logistic regression model with the following commands:

```
status <- matrix(c(diseased,normal), ncol=2)
status
logreg1 <- glm(status~log(exposure), family=binomial)
```

Notice the option `family=binomial`. It tells R to interpret values in the matrix `status` as outcomes from a binomial distribution.

3. Try the commands

```
summary(logreg1)
abline(logreg1)
```

What are the estimated values of α and β? Does the model seem to fit the data reasonably well?

Above, the data was represented by the numbers of diseased and normal for each level of exposure with 8 observations in total. The data could also be represented with one observation per person (371 observations in total). The files `coalworker2.xlsx` and `coalworker2.csv` contain the data in this representation with the variable `y` telling whether the person was diseased (`y=1`) or normal (`y=0`). Apart from this there is as variable `exposure2` with the exposure time.

4. Read the data into R in the new form; call the dataset `coalworker2`. Make sure you understand the structure of the new dataset, and make sure you understand that `coalworker1` and `coalworker2` contain the same information.

5. When the data is represented as in `coalworker2`, the logistic regression model is fitted as follows:

```
logreg2 <- glm(y ~ log(exposure2), family=binomial, data=coalworker2)
summary(logreg2)
```

Fit the model and make sure that you get the same estimates as you did with `logreg1` before.

The last questions are about the interpretation of the estimates and the models.

6. Consider a person with exposure time equal to 30 years. What is the estimated log-odds of this person being diseased? What is the estimated probability that the person is diseased?

Hints: For the estimation of log-odds, remember that the explanatory variable is the log-transformed exposure time. For the probability you should solve expression (3) for p.

7. How many years of exposure gives a 50% risk of having developed the disease?

Hint: We are looking for the exposure time corresponding to $p = 0.5$. What is the corresponding value of log-odds? Use the estimates to compute the (logarithmic) exposure time.

8. What happens with the odds if the exposure time is doubled?

   Hint: What happens to log-exposure if exposure time is doubled? What is the effect on log-odds? Which effect on odds does that correpsond to?