

Tutorial week 11

1. In a study of the relationship between the death rate from lung cancer and the per capita consumption of cigarettes twenty years earlier, data from 9 countries yielded a correlation of $r = 0.73$. Test the null hypothesis that the population coefficient of correlation is 0.5 against the alternative that it is more than 0.5.
2. In a study of the relationship between the available heat of green wood and air-dried wood, data from 13 pairs yielded $r = 0.94$. Calculate a 95% confidence interval for the population correlation coefficient ρ , stating any assumptions you need to make.
3. The *ethanol* data contains 88 measurements on E and NOx from an experiment where ethanol was burned in a single cylinder engine. Consider the kernel regression of NOx on E with kernel *box* and three different bandwidths: 0.01, 0.1 and 0.3. Consider the following output, in which the column *c1* gives the original numbering of the points, e.g., the point (0.568, 0.374) appearing in the third row is actually (x_{26}, y_{26}) ; the last three columns give the fitted values for the three fits with bandwidth 0.01, 0.1 and 0.3, respectively.

```
> c1<-order(E)
> cbind(c1, x, y, f1, f2, f3)
```

```
numeric matrix: 88 rows, 6 columns.
```

```
      c1    x    y    f1    f2    f3
[1,]  87 0.535 0.530 0.530000 0.4880000 0.9188334
[2,]  86 0.562 0.370 0.370000 0.6614286 1.0339413
[3,]  26 0.568 0.374 0.374000 0.6614286 1.0553334
[4,]  85 0.584 0.678 0.678000 0.6488751 1.1306001
[5,]  14 0.601 1.192 1.057500 0.6540000 1.1539048
[6,]  72 0.602 0.923 1.057500 0.6540000 1.1539048
[7,]  84 0.608 0.563 0.563000 0.7924445 1.1539048
[8,]  83 0.629 0.561 0.561000 1.0206250 1.2704799
[9,]  45 0.637 0.571 0.571000 1.1848888 1.2704799
[10,] 88 0.655 1.900 1.900000 1.2281818 1.4594282
[11,] 52 0.676 1.777 1.777000 1.3060000 1.7455877
[12,] 56 0.684 1.587 1.588500 1.4517500 1.7455877
[13,] 16 0.686 1.590 1.588500 1.4517500 1.7824237
[14,] 32 0.693 1.369 1.240250 1.5318182 1.7824237
```

```

[15,] 73 0.694 1.527 1.240250 1.5318182 1.7824237
[16,] 15 0.696 0.926 1.240250 1.5318182 1.7824237
[17,] 69 0.696 1.139 1.240250 1.5318182 1.7824237
[18,] 47 0.715 1.419 1.419000 1.5768461 1.8903327
[19,] 57 0.729 1.397 1.808000 1.5944999 2.0615706
.....
.....
.....
[86,] 65 1.230 0.672 0.6040000 0.6933636 0.9104165
[87,] 7 1.231 0.638 0.6040000 0.6933636 0.9104165
[88,] 33 1.232 0.542 0.6040000 0.6661000 0.8752609

```

- (a) Describe the kernel smoothing procedure with kernel "box".
 - (b) How many points are used to compute the fitted value $f_{126} = 0.374$?
 - (c) How many points are used to compute the fitted value $f_{226} = 0.6614286$?
 - (d) How many points are used to compute the fitted value $f_{326} = 1.0553334$?
 - (e) Verify for the second fit (bandwidth = 0.1) that $f_{226} = 0.6614286$, by performing the calculations directly.
4. Define Cook's distance for observation i as $D_i = \sum_{j=1}^n (\hat{y}_j - \hat{y}_{j(i)})^2 / (2\text{MSE})$, where $\hat{y}_{j(i)}$ is the j -th fitted value when the i -th observation is deleted. Show that this is algebraically equivalent to $D_i = (\hat{\beta} - \hat{\beta}_{(i)})^\top \mathbf{X}^\top \mathbf{X} (\hat{\beta} - \hat{\beta}_{(i)}) / (2\text{MSE})$.

Computer Exercises week 11

1. The *ethanol* data contains 88 measurements on E and NOx from an experiment where ethanol was burned in a single cylinder engine.
 - (a) Load the data and let `x <- sort(ethanol$E)` and `y <- ethanol$NOx[order(ethanol$E)]`.
 - (b) Split your graphics window into a 3×1 display. Consider the kernel regression of NOx on E with kernel `box` and three different bandwidths: 0.05, 0.1 and 0.3. Plot the data and add the smoothed curve to the plot.
 - (c) Consider the kernel regression of NOx on E with kernel `normal` and three different bandwidths: 0.05, 0.1 and 0.3, and obtain similar graphs as in part (b).
 - (d) Compare the graphs given in parts (b) and (c). Which bandwidth gives the best fit? Comment on the claim that the choice of bandwidth is much more important than the choice of kernel.
 - (e) Consider the locally weighted regression of NOx on E with span 0.5. Obtain the scatter plot of y against x with the regression curve on it. From your plot use the smoothed curve to estimate the y value when $x = 1.0$.
 - (f) Test the hypothesis that NOx is normally distributed with the Shapiro-Wilk test.