MAST20005/MAST90058: Week 6 Lab

Goals: (i) Fit and interpret the simple regression model in R; (ii) Check model assumptions; (iii) Explore more general estimation methods beyond the least squares method.

Data: In this lab, we will use the `crime` data from Agresti and Finlay (1997), describing crime rates in the United States. The variables are state ID (`sid`), state name (`state`), violent crimes per 100,000 people (`crime`), murders per 1,000,000 people (`murder`), the percentage of the population living in metropolitan areas (`pctmetro`), the percentage of the population that are white (`pctwhite`), the percentage of the population with a high school education or above (`pcths`), the percentage of the population living under the poverty line (`poverty`), and the percentage of the population that are single parents (`single`). Data can be obtained from the shared folder in the computer labs, from the LMS, or downloaded from https://stats.idre.ucla.edu/stat/data/crime.dta. The data file is in the Stata binary format (`.dta`) so you will need to install the `foreign` package to import it into R.

1 Explore pairwise associations

R can import and export data in different formats. For example, the `crime.dta` data file is produced by Stata, a commercial statistical software package. The R package `foreign` defines useful functions to manipulate data formats different from `.csv` or `.txt`. First we need to install it.

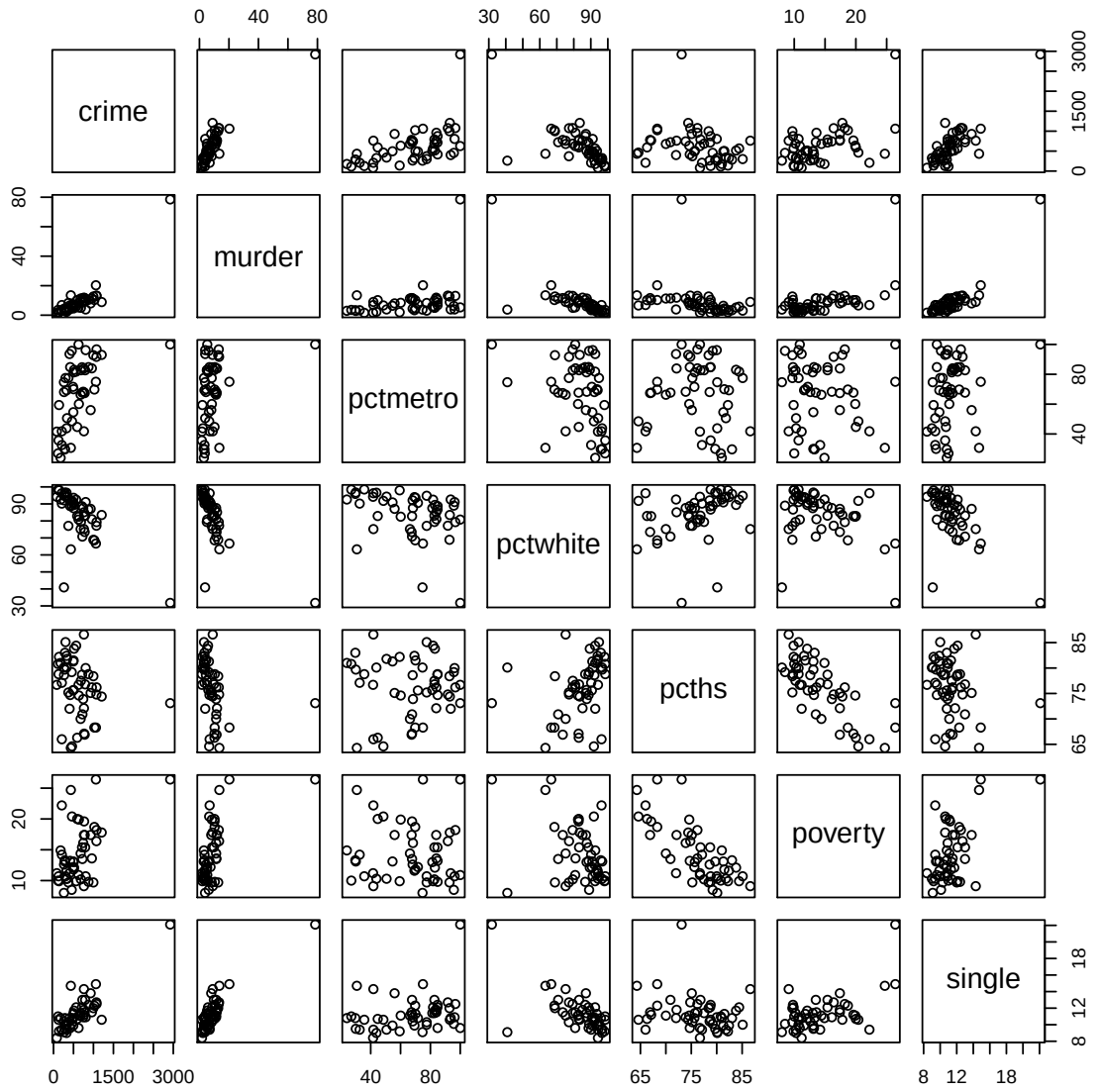
```
install.packages("foreign")
```

Now we can load the data and compute some basic summary statistics.

```
library(foreign)
cdata <- read.dta("crime.dta")
summary(cdata)
```

1. Explore pairwise association between variables by constructing a scatter plot matrix. We are interested in finding variables related to crime.

```
plot(cdata[, -c(1, 2)])
```



We excluded the first two columns since they do not contain numerical variables. Often transforming the variables by the `log` function helps us see linear association between variables.

```
plot(log(cdata[, -c(1, 2)]))
```

2. Linear association between variables is estimated by the sample correlation coefficient:

$$\hat{\rho} = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_i (x_i - \bar{x})^2 \sum_i (y_i - \bar{y})^2}}.$$

Recall that when $\hat{\rho}$ is close to ± 1 we have perfect linear association, while $\hat{\rho} \approx 0$ indicates a lack of linear association. A matrix containing pairwise correlation coefficients can be obtained using the following command:

```
cor(cdata[, -c(1, 2)])
```

This is a bit hard to read. Let's multiply by 100 and round to the nearest integer (e.g. 0.54 becomes 54, etc.):

```
round(100 * cor(cdata[, -c(1, 2)]))
```

```
##          crime murder pctmetro pctwhite pcths poverty single
## crime      100     89       54      -68   -26      51     84
## murder      89    100       32      -71   -29      57     86
## pctmetro    54     32      100      -34     0      -6     26
## pctwhite   -68    -71      -34      100    34     -39    -66
## pcths      -26    -29        0       34   100     -74    -22
## poverty     51     57       -6      -39   -74     100     55
## single      84     86       26      -66   -22      55    100
```

Are there any variables showing linear association with crime?
What about other types of association?

2 Fit a linear model

Fit a simple linear regression model and carry out some diagnostics on the fitted model.

1. The following command fits the model,

$$\text{crime}_i = \alpha + \beta \times \text{single}_i + \varepsilon_i, \quad \varepsilon_i \sim \text{N}(0, \sigma^2). \tag{1}$$

```
m1 <- lm(crime ~ single, data = cdata) # fit model (compute estimates)
summary(m1)  # show results
```

```
##
## Call:
## lm(formula = crime ~ single, data = cdata)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -767.42 -116.82  -20.58  125.28  719.70
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1362.53     186.23  -7.316 2.15e-09 ***
## single        174.42      16.17  10.788 1.53e-14 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 242.5 on 49 degrees of freedom
## Multiple R-squared:  0.7037,Adjusted R-squared:  0.6977
## F-statistic: 116.4 on 1 and 49 DF,  p-value: 1.529e-14
```

2. Report estimates for α, β and give confidence intervals. Then interpret the value of $\hat{\beta}$ in the context of the problem.

   From the previous output, the estimates are $\hat{\alpha} = -1362.53$ and $\hat{\beta} = 174.42$. Confidence intervals can be computed using:

```
confint(m1)
```

```
##                   2.5 %    97.5 %
## (Intercept) -1736.7818 -988.2831
## single        141.9278  206.9093
```

   States with a larger proportion of single parents tend to have high violent crime rates. In particular, we expect about 174 additional violent crimes per 100,000 people for each 1 percentage point increase in the proportion of single parents. Since the 95% confidence interval for β is clearly above 0, we have good evidence for this positive association.

3. Obtain a 95% confidence interval of the mean crime rate when the percent of population that are single parents is 11.5. In other words, compute a confidence interval for $\mu(11.5) = \alpha + \beta \times 11.5$.

```
newdata <- data.frame(single = 11.5)
predict(m1, newdata, interval = "confidence")
```

```
##        fit     lwr      upr
## 1 643.2809 574.796 711.7658
```

   Then obtain a 95% prediction interval for Y for the same scenario.

```
predict(m1, newdata, interval = "prediction")
```

```
##        fit      lwr     upr
## 1 643.2809 151.0919 1135.47
```

   Why is the prediction interval wider than the confidence interval?

4. Estimate the error standard deviation, σ.

   From the previous output we obtain, $\hat{\sigma} = 242.5$.

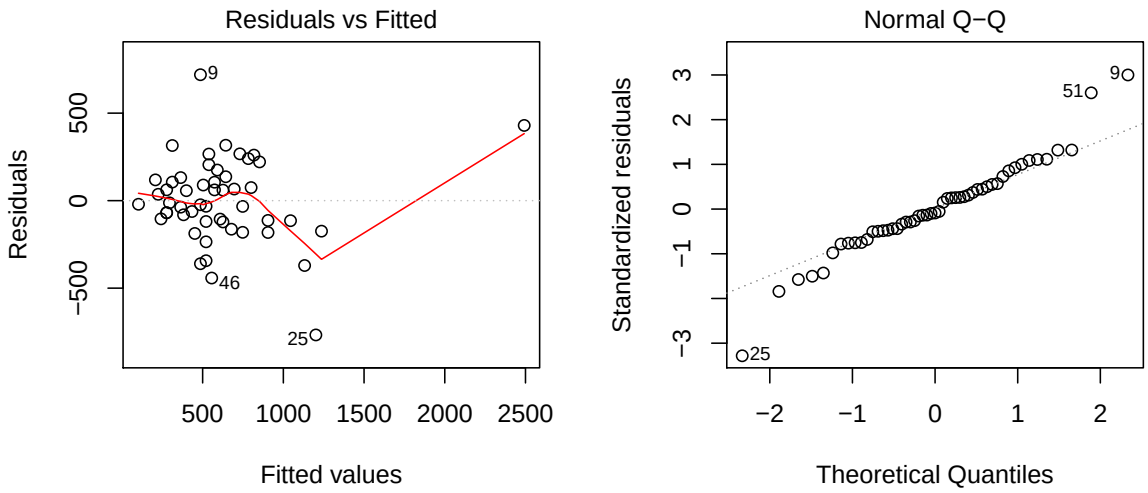
   Note that R computes $\hat{\sigma} = \sqrt{d^2/(n-2)}$, where $d^2 = \sum_i r_i^2$ is the sum of squared residuals, with $r_i = y_i - (\hat{\alpha} + \hat{\beta} x_i)$. This estimate is different from the MLE of σ, which is computed as $\hat{\sigma}_{\text{MLE}} = \sqrt{d^2/n}$. If we needed the latter (which we rarely do in practice), we could calculate it in R as follows:

```
n <- nrow(cdata) # sample size
RSS <- sum(m1$residuals^2) # extract, square and sum the residuals
sqrt(RSS / n)  # MLE
```

```
## [1] 237.7361
```

5. Let us have a closer look at the residuals, $r_i = \text{crime}_i - (\hat{\alpha} + \hat{\beta}\,\text{single}_i)$. These represent the left-over, unexplained variation under model (1). Model diagnostic plots can be obtained as follows:

```
par(mfrow = c(1, 2))  # set up a 1*2 array for drawing plots
plot(m1, 1:2)  # this will draw 2 plots, filling up the 1*2 array
```



The first plot shows the residuals against fitted values. The red line is a smooth approximation of the data helping us detect anomalous trends in the residuals. Ideally, the residuals should be scattered randomly around the zero line. The second plot is a QQ plot to check how close the residuals follow a normal distribution. While the central part of the distribution looks normal, a few of the extreme observations deviate a little from the normality assumption.

Overall the model looks reasonably appropriate for the data except for a few points that might be incompatible with the model assumptions. The most extreme data points are numbered in the plots, for convenience. In this case, it is worth looking closely at observations 9, 25 and 51, to see if there is any reason to suspect they might be not representative of the other states.

```
cdata[c(9, 25, 51),]
```

```
##    sid state crime murder pctmetro pctwhite pcths poverty single
## 9    9    fl  1206    8.9      93.0     83.5  74.4    17.8   10.6
## 25  25    ms   434   13.5      30.7     63.3  64.3    24.7   14.7
## 51  51    dc  2922   78.5     100.0     31.8  73.1    26.4   22.1
```

3 Comparisons with the null model

1. Consider the model

$$\text{crime}_i = \alpha + \varepsilon_i, \quad \varepsilon_i \sim \text{N}(0, \sigma^2) \tag{2}$$

and give an estimate of σ. Model (2) is sometimes called the *null model* since it assumes that there is no linear relationship between the predictor(s) and `crime`, by assuming that

the mean of `crime` is constant. An estimate of σ in this case is just the sample standard deviation $s_y = \sum_i (y_i - \bar{y})^2/(n-1)$. In R:

```
y <- cdata$crime
sd(y)

## [1] 441.1003
```

The estimated variance for this model is much larger than that for model (1), suggesting that model (1) is superior to model (2) in terms of explaining the variability in `crime`.

By the way, you can also fit the null model in R using `lm()` as follows:

```
m0 <- lm(crime ~ 1, data = cdata)
summary(m0)

##
## Call:
## lm(formula = crime ~ 1, data = cdata)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -530.84 -286.34  -97.84  160.16 2309.16
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   612.84      61.77   9.922 2.09e-13 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 441.1 on 50 degrees of freedom
```

Can you spot the estimates of α and σ in the output?

2. The improvement in terms of explained variability of model (1) compared to the null model (2) is typically measured by the *coefficient of determination*, also known as "R-squared":

$$R^2 = 1 - \frac{\text{residuals sum of squares}}{\text{total sum of squares}} = 1 - \frac{\sum_i (y_i - \hat{\alpha} - \hat{\beta} x_i)^2}{\sum_i (y_i - \bar{y})^2}.$$

```
TSS <- sum((y - mean(y))^2)
1 - sum(RSS / TSS)

## [1] 0.7037109
```

This number is also given in the summary output of `lm()`, labelled as "`Multiple R-squared`" (can you spot it in the earlier output for the model `m1`?). If $R^2 = 1$, the linear relationship between the predictor and the response explains all the variability in the response (i.e.

the points are exactly on a straight line). In this case $R^2 = 0.7037$, meaning that about 70% of the variability in `crime` is explained by the linear relationship between `crime` and `single`.

3. The coefficient of determination can be used to compare models with the same number of predictors. For example, let us fit another model:

```
m2 <- lm(crime ~ poverty, data = cdata)
summary(m2)
```

Clearly the linear model including `single` as a predictor does a better job in explaining the variability in `crime` compared to the model that with `poverty`.

4 Robust regression

There are a number of estimation methods that may be used to fit the regression model and the choice of the method ultimately depends on the goal of the analysis. In this section we explore use of the weighted least squares (WLS) method to compute estimates for α and β. WLS finds α and β by solving the weighted least squares problem,

$$\min_{\alpha,\beta} \sum_i w_i (y_i - \alpha - \beta x_i)^2,$$

where w_i are user-defined weights. When the goal is to reduce the influence of potential outliers we can use the Huber weights, which have the form:

$$w_i = \begin{cases} 1 & \text{if } |e_i| \leqslant k \\ k/|e_i| & \text{if } |e_i| > k \end{cases}$$

where $k > 0$ is a tuning constant[1] and $e_i = y_i - \alpha - \beta x_i$. In Huber weighting, observations with small residuals get a weight of 1 and the larger the residual, the smaller the weight.

1. In the following example, we use the `rlm` fuction from the `MASS` package to carry out WLS estimation with Huber weights (there are several other weighting functions that can be used). The default value of the tuning constant is $k = 1.345\hat{\sigma}$. Theory suggests that this value leads to a considerable increase in robustness but with only a small increase in the variance of the estimator.

```
library(MASS)
m3 <- rlm(crime ~ single, data = cdata)
summary(m3)

##
## Call: rlm(formula = crime ~ single, data = cdata)
## Residuals:
##      Min       1Q   Median       3Q      Max
## -797.489 -130.956   -9.108  127.019  716.664
```

[1] A *tuning constant* or *tuning parameter* is parameter in a model that usually needs to be set to a value before we can estimate the other parameters. If so, then it is not a parameter that we estimate. Hence, we can think of it as one of the (many) assumptions we need to make in order to do inference.

```
##
## Coefficients:
##             Value      Std. Error t value
## (Intercept) -1429.3999   164.1623    -8.7072
## single        181.0128    14.2519    12.7010
##
## Residual standard error: 192.9 on 49 degrees of freedom
```

2. The following gives a table with Huber weights (in increasing order) and the corresponding residuals:

```
hweights <- data.frame(state  = cdata$state,
                       resid  = m3$resid,
                       weight = m3$w)
hweights2 <- hweights[order(m3$w), ]
hweights2[1:15, ]
```

```
##    state      resid    weight
## 25    ms -797.48851 0.3254235
## 9     fl  716.66395 0.3621087
## 46    vt -447.74111 0.5795995
## 1     ak -398.08345 0.6519515
## 21    me -363.33605 0.7142369
## 51    dc  351.01652 0.7390823
## 26    mt -347.53858 0.7467066
## 31    nj  318.67677 0.8143642
## 14    il  307.75248 0.8432232
## 19    ma  261.36024 0.9929159
## 2     al  127.75248 1.0000000
## 3     ar   85.56277 1.0000000
## 4     az  -45.85528 1.0000000
## 5     ca  244.73966 1.0000000
## 6     co -193.85528 1.0000000
```

We can see that as the absolute residual tends to go down, the weight goes up. In other words, cases with a large residuals tend to be down-weighted. This output shows us that the observation for Mississippi will be down-weighted the most. Florida will also be substantially down-weighted. All observations not shown above have a weight of 1. In ordinary least squares regression, all cases have a weight of 1.

Exercises

1. (a) Draw a plot of the data and an overlayed model fit for model (1).
   (b) Is there a data point that's very different to the others? Which one is it?
   (c) Do you think this point has a strong impact on the model fit, in the sense of having a larger influence on the parameter estimates than the other observations?
   (d) Check whether this is a concern by excluding this data point and re-fitting the model. How do the estimates compare to the original ones?

2. The President of the United States asks you to find out how the murder rate relates to the proportion of single parents.

   (a) Fit a simple linear regression model that relates these two variables.

   (b) Draw a standard set of diagnostic plots. What do you see?

   (c) Remove one of the data points that seems to be problematic and re-fit the model. (Note: this now means the model fit is representative of all of the remaining states but not the one that has been omitted.) How does your model fit compare to the original?

   (d) The President would like you predict the murder rate a state would have if 12% of its population were single parents. Calculate an appropriate 90% prediction interval. *Hint:* You might need to run `help(predict.lm)` to find out something relevant.

   (e) Repeat this for a state with only 8% single parents. What do you notice?

   (f) Since the measurements are all positive, a standard way to fix the problem you have just observed is to first take the logarithm of all of the data values and work with those instead. Do this and, re-fit the model and re-compute the confidence interval. (Remember to transform your interval back to the original scale at the end.)

   (g) What model is this procedure equivalent to? Can you write its model equation? How do you interpret the 'slope' parameter in this model?

   (h) Draw a plot of the data and a 90% confidence band using this latest model.

3. Do question 7 from the tutorial problems.