Inference Tutorial 3

This sheet covers some slightly less basic linear modelling and inference in R.

1. An online advertising company is interested in modelling the relationship between the number of adverts clicked by customers (a standardized monthly rate), x_i, and their on line spending, y_i (per year), on a set of websites that they are involved with. They also have (inferred) data on customers office for national statistics socio-economic status (categories 1-7). They want to estimate a model

$$y_i = \alpha_{\mathrm{se}(i)} + \gamma x_i + \epsilon_i$$

under the usual linear modelling assumptions, where $\alpha_1, \ldots \alpha_7$ and γ are parameters and $\mathrm{se}(i)$ is the socio-economic class of customer i. The dataset contains wildly different numbers in the different social classes, so it is suggested to even things out by estimating the model parameters to minimise the weighted least squares objective

$$\sum_{i=1}^{n} (y_i - \mu_i)^2/(n_{\mathrm{se}(i)})$$

where n_j is the total number of customers in socio-economic group j.

   (a) Show that, for an appropriate choice of diagonal matrix $\mathbf{W}$, the weighted least squares objective is equivalent to an un-weighted least squares objective $\|\tilde{\mathbf{y}} - \tilde{\mathbf{X}}\boldsymbol{\beta}\|^2$, where $\tilde{\mathbf{y}} = \mathbf{W}\mathbf{y}$ and $\tilde{\mathbf{X}} = \mathbf{W}\mathbf{X}$. ($\boldsymbol{\beta}$ is the vector of all model parameters.)

   (b) Is $\hat{\boldsymbol{\beta}}$, obtained by minimizing the weighted least squares objective, a linear unbiased estimator?

   (c) After fitting the model by weighted least squares, careful residual checking suggests that the model is a good fit and the residual assumptions are met, overall and within each socio-economic group. Would you recommend the weighted least squares estimator here? Explain your answer.

   Solution

   (a) Let $W_{ii} = n_{se(i)}^{-1/2}$. Then

$$\sum_{i=1}^{n} (y_i - \mu_i)^2/(n_{\mathrm{se}(i)}) = \sum_{i=1}^{n} W_{ii}^2 (y_i - \mu_i)^2 = \|\mathbf{W}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})\|^2 = \|\mathbf{W}\mathbf{y} - \mathbf{W}\mathbf{X}\boldsymbol{\beta}\|^2.$$

   (b) It's a linear estimator, since the weighted least squares estimator will be linear in $\tilde{\mathbf{y}}$ from the previous part, and $\tilde{\mathbf{y}}$ is linear in $\mathbf{y}$. It's also unbiased, as $E(\mathbf{W}\mathbf{y}) = \mathbf{W}\mathbf{X}\boldsymbol{\beta}$, so unbiasedness follows by the same argument as in the unweighted case.

   (c) Not really. The Gauss Markov theorem tells us that the un-weighted least squares estimator would have lower variance than (or at worst equal variance to) the weighted least squares estimator. The companies reasoning only really makes sense if they know that the model is wrong, but in that case it might be better to try a better model, for example allowing a different slope parameter within each socio-economic group.

2. A 'quant' working for prestigious international financial firm Soldman Craps, proposes the following model for predicting the half hour return rate, r_i, on a particular short term trade, based on simple market index and volatility data:

$$\begin{aligned} r_i = \beta_0 + \beta_1 \cos(2t_i\pi/48) + \beta_2 \sin(2t_i\pi/48) + \beta_3 \mathtt{DJ}_{i-1} + \beta_4 \mathtt{DAX}_{i-1} + \beta_5 \mathtt{FT}_{i-1} \\ + \beta_6(\mathtt{DJ}_{i-1} - \mathtt{DJ}_{i-2})^2 + \beta_7(\mathtt{DAX}_{i-1} - \mathtt{DAX}_{i-2})^2 + \beta_8(\mathtt{FT}_{i-1} - \mathtt{FT}_{i-2})^2 + \epsilon_i \end{aligned} \quad (1)$$

where the ϵ_i are i.i.d. $N(0, \sigma^2)$, r_i is the return in the i^{th} half hour period, t_i is the time measured in half hours from some arbitrary start point, $\mathtt{DJ}_i$, $\mathtt{DAX}_i$ and $\mathtt{FT}_i$ are the state of the Dow Jones, DAX and FT100 indices in the i^{th} half hour period. The $\sin$ and $\cos$ terms are there to model daily fluctuations. The model is to be estimated by least squares using data for the fortnight leading up to the trading half hour of interest. There is money to be made if r_i can be predicted along with accurate assessment of the prediction uncertainty. In test runs the prediction accuracy seems reasonable.

   What basic thing is likely to be wrong with this model, and how might you plot the model residuals to investigate the problem?

   Solution

   The ϵ_i will almost certainly not be independent, since there will be short-term carry-over in r from one time point to the next, not captured by the model. Plotting residuals, $\hat{\epsilon}_i$, against lagged residuals, $\hat{\epsilon}_{i-1}$ would be likely to show up the effect (via an obvious correlation).

3. (a) Consider the R dataset `PlantGrowth` from §3.1.1 of the notes. Often people would like to use the model

$$\texttt{weight}_i = \mu + \beta_{g(i)} + \epsilon_i,$$

where μ is the overall population mean weight, and $\beta_{g(i)}$ represents the departure from the overall mean as a result of being in group $g(i)$. What is the problem with estimating this model from data, and how will it manifest itself in the rank of the model matrix?

(b) R deals with the problem identified in part 1 automatically. Try fitting the model using R...

```
m1 <- lm(weight ~ group,data=PlantGrowth)
```

and now use `model.matrix(m1)` to see how R has dealt with the problem. R's solution changes the interpretation of the model parameters. Explain the new interpretation.

(c) Suppose we want to test whether `group` is really needed in the model. How might you formulate an appropriate hypothesis test for this? For this simple model `anova(m1)` conducts such a test. Which result from the notes is it using? And what do you conclude about the hypothesis?

(d) An alternative approach to the test is to formulate a null model

$$\texttt{weight}_i = \mu + \epsilon_i$$

without the group effect, and to compare its fit to that of the original model.

```
m0 <- lm(weight ~ 1,data=PlantGrowth)
anova(m0,m1)
```

Which result from the notes do you think is being used now? Note that the p-value has not changed!

(e) Look at the results from `summary(m1)`. What hypotheses are being tested here. Comment on the p-values obtained here compared to the previous test.

(f) Find a 90% confidence interval for the difference between treatment 2 and the control.

Solution

(a) There is not a unique mapping from the parameters to $E(y_i)$. For any set of values for μ and β_j, the same $E(y_i)$ would result from $\mu + k$ and $\beta_j - k$ for any finite constant k. Hence we can not uniquely estimate the parameters. Writing out the model matrix we have:

$$\mathbf{X} = \begin{pmatrix} 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ . & . & . & . \\ 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ . & . & . & . \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 \\ . & . & . & . \\ 1 & 0 & 0 & 1 \end{pmatrix}$$

which is clearly rank 3 and not rank 4 (any column is a linear combination of the other 3). In consequence the factor $\mathbf{R}$ in its QR decomposition is only rank 3, not 4, and hence $\mathbf{R}^{-1}$ does not exist and we can not compute the parameter estimates.

(b)
```
> m1 <- lm(weight ~ group,data=PlantGrowth)
> model.matrix(m1)
   (Intercept) grouptrt1 grouptrt2
1            1         0         0
2            1         0         0
...
9            1         0         0
10           1         0         0
11           1         1         0
12           1         1         0
```

```
13           1         1         0
...
20           1         1         0
21           1         0         1
22           1         0         1
...
```

R has just dropped the column of the model matrix corresponding to β_0, the effect for the control group. This is equivalent to constraining the control group effect to be zero ($\beta_0 = 0$). This restores full column rank to the model matrix and ensures that we can uniquely estimate the remaining parameters. The parameter interpretations have changed, however. Clearly in the new model, $E(y_i) = \mu$ for the control group. So μ is now the control group mean (expected value). $E(y_i) = \mu + \beta_{g(i)}$ for the two treatment groups, so β_1 and β_2 are now the differences between the expected values for treatment 1 and the control group, and between treatment 2 and the control group.

(c) Appropriate null hypothesis is $H_0 : \beta_1 = \beta_2 = 0$.

```
> anova(m1)
Analysis of Variance Table

Response: weight
          Df  Sum Sq Mean Sq F value  Pr(>F)
group      2  3.7663  1.8832  4.8461 0.01591 *
Residuals 27 10.4921  0.3886
...
```

This is using result (7) from section 3.4.4 of the notes. The fairly low p-value is evidence against the null hypothesis — at least one of the treatments seems to have an effect on weight.

(d)
```
> m0 <- lm(weight ~ 1,data=PlantGrowth)
> anova(m0,m1)
Analysis of Variance Table

Model 1: weight ~ 1
Model 2: weight ~ group
  Res.Df    RSS Df Sum of Sq      F  Pr(>F)
1     29 14.258
2     27 10.492  2    3.7663 4.8461 0.01591 *
...
```

This is using the alternative formulation of the test statistic (7) in terms of the residual sums of squares of alternative models. The fact that it amounts to a re-writing of the same test statistic is why the p-value is identical.

(e)
```
> summary(m1)
...
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   5.0320     0.1971  25.527   <2e-16 ***
grouptrt1    -0.3710     0.2788  -1.331   0.1944
grouptrt2     0.4940     0.2788   1.772   0.0877 .
...
Residual standard error: 0.6234 on 27 degrees of freedom
Multiple R-squared:  0.2641,Adjusted R-squared:  0.2096
F-statistic: 4.846 on 2 and 27 DF,  p-value: 0.01591
```

The p-values in the coefficients table are testing the individual hypotheses $H_0 : \beta_1 = 0$ and $H_0 : \beta_2 = 0$. The p-values are substantially larger than in the previous part. We don't seem to have enough evidence to reject either individual null hypothesis. So the combined test is more powerful (more able to reject) than the individual hypotheses. We can say with some confidence that at least one of the treatment groups differs from the control, but which one is more difficult (probably treatment 2). This sort of effect is part of the reason for testing the effects of factor variables all in one go, rather than starting with the tests about the factor's individual coefficients.

(f)
```
> beta2 <- coef(m1)[3]
> sigb2 <- diag(vcov(m1))[3]^.5
```

```
> crit <- qt(.95,df=27)
> c(beta2 - crit*sigb2,beta2 + crit*sigb2)
 grouptrt2  grouptrt2
0.01915451 0.96884549
```