Coursework - Applied Statistics

CID: 02091191

November 15, 2021

Abstract

In this document we are presenting the work conducted during the second Applied Statistics Coursework of the year.

1 Question 1

(a) Relationship between AGE and NOX variables

In this Coursework, we are working on the same data set as during the first coursework about Boston Housing. Please refer to the documentation [1] or to the previous Coursework where we made a full data exploratory analysis for further details. We then assume that all variables are known and we will now try to extend the previous linear model by categorising the AGE variable.

```
plot(dfm$age, dfm$nox, pch=16, xlab="AGE", ylab="NOX")
```

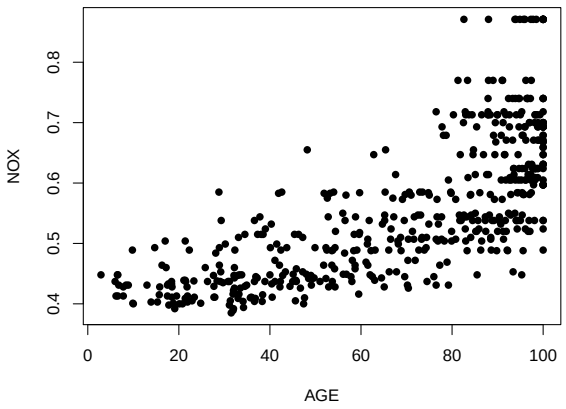

Figure 1.1: Scatter plot of NOX against AGE

The scatter plot shown in Figure 1.1 shows the relationship between the nitric oxides concentration (NOX) and the proportion of houses built prior to 1940 (AGE). The nitric oxides concentration increases with the age of houses which seems consistent. However, the relationship between both variables doesn't seem clearly linear as the scatter plot is becomes more and more wide spread as AGE increases (heteroscedasticity). Hence, the relationship between NOX and AGE shall not be linear globally but may be linear by parts on some intervals. We could think that the relationship between NOX and AGE is linear on $\texttt{AGE} \in [0, 25]$ with some parameters, linear on $\texttt{AGE} \in [25, 80]$ with some others parameters and non linear on $\texttt{AGE} \in [80, 100]$.

(b) Creation of the new covariate AGEBAND

In **R**, we then create a new factor variable called AGEBAND in which we store three levels: "low", "medium" and "high" corresponding to AGE values in $(0, 25]$, $(25, 80]$ and $(80, 100]$ respectively. The number of observations in each category is written below.

```
create.category <- function(x){
  if(x>0 & x<=25) return("low")
  if(x>25 & x<=80) return("medium")
  if(x>80 & x <=100) return("high")
}
dfm$ageband <- as.factor(sapply(dfm$age, create.category))
summary(dfm$ageband)
```

```
medium   high    low
   217    240     49
```

(c) Comparison of NOX levels between medium and high age categories

To augment the previous linear model with this AGEBAND variable, we can introduce the following notation. Respectively denoting the categories "low", "medium" and "high" as 1,2 and 3, let y_{ij} denote the NOX level for the i^{th} observation in the data from AGEBAND category $j, i = 1, \ldots, n_j, j = 1, 2, 3$, with n_j denoting the number of observations in category j. The revised linear model we now consider is

$$y_{ij} = \alpha_j + \sum_{k=1}^{4} \beta_k x_{ijk}, \quad i = 1, \ldots, n_j; j = 1, 2, 3$$

with $\beta \in \mathbb{R}^4$. For observation i from AGEBAND category $j, x_{ijk}, k = 1, 2, 3, 4$, denotes the respective value of the variables INDUS, RAD, TAX, AGE. The remaining terms $\alpha \in \mathbb{R}^3$ represent the main effects for each AGEBAND category. Assume we are interested in whether there is a change in NOX levels between between the medium and high age categories when also conditioning on the other variables in the model.

We need to choose a way to encode the categorical (factor) variable AGEBAND in our new linear model. As there is a natural order to the AGEBAND categories, we will use the Helmert contrast coding which will make easier comparisons between populations from different categories while conditioning on the other variables. We also need to pay attention to which level we use as a baseline in order to compute the good comparison between medium and high age categories. We will then use "medium" as our baseline here. Table 1 summarises the encoding as:

	X_1	X_2
medium	-1	-1
high	1	-1
low	0	2

Table 1: Encoding for comparing NOX levels between medium and high age categories

```
dfm$ageband <- relevel(dfm$ageband, ref="medium")
contrasts(dfm$ageband) <- "contr.helmert"
mod <- lm(nox ~ indus + rad + tax + age + ageband, data=dfm)
summary(mod)
```

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  3.759e-01  1.701e-02  22.092  < 2e-16 ***
indus        6.172e-03  6.868e-04   8.986  < 2e-16 ***
rad          2.457e-03  7.939e-04   3.095  0.00208 **
tax          1.345e-05  4.745e-05   0.283  0.77696
age          1.106e-03  2.529e-04   4.373 1.49e-05 ***
ageband1     1.844e-02  5.729e-03   3.219  0.00137 **
ageband2    -6.282e-03  5.586e-03  -1.125  0.26127
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

The parameter of interest here is the estimate of `ageband1` coefficient as it corresponds to

$$\gamma := \frac{\mu_{\text{high}} - \mu_{\text{medium}}}{2} \text{ where } \mu_{\text{high}} = \mathbb{E}[Y|\text{high}] \text{ and } \mu_{\text{medium}} = \mathbb{E}[Y|\text{medium}]$$

The **R** code above gives an estimate of this coefficient of 0.01844, with a t-statistic of 3.219 and a significance p-value of 0.00137. This means that there is enough evidence to reject the null hypothesis H_0: $\gamma = 0$ (with a 99% level given by two stars **). Hence, there is a significant change in NOX levels between the medium and high age categories.

We also notice that the estimated coefficient of `ageband2` can be interpreted as equal to zero according to the t-statistic and the p-value (formally, there is not enough evidence to reject the null hypothesis that this coefficient is equal to zero). This would mean that $\mu_{\text{low}} = \frac{\mu_{\text{medium}} + \mu_{\text{high}}}{2}$ and then that there is no significant difference in NOX levels between the low category and the "mean reunited" high and medium categories.

(d) Comparison of NOX levels between low and medium age categories

We proceed as for (c) to compare the low and medium age categories just by changing the baseline of our contrast: we will here use the low age category as our reference. Table 2 summarises the encoding:

	X_1	X_2
low	-1	-1
medium	1	-1
low	0	2

Table 2: Encoding for comparing NOX levels between low and medium age categories

```
dfm$ageband <- relevel(dfm$ageband, ref="low")
contrasts(dfm$ageband) <- "contr.helmert"
mod <- lm(nox ~ indus + rad + tax + age + ageband, data=dfm)
summary(mod)
```

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 3.759e-01  1.701e-02  22.092  < 2e-16 ***
indus       6.172e-03  6.868e-04   8.986  < 2e-16 ***
rad         2.457e-03  7.939e-04   3.095  0.00208 **
tax         1.345e-05  4.745e-05   0.283  0.77696
age         1.106e-03  2.529e-04   4.373 1.49e-05 ***
```

```
ageband1    2.033e-04  6.741e-03   0.030  0.97595
ageband2    1.236e-02  5.196e-03   2.379  0.01774 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Here, the parameter of interest is the estimate of `ageband1` coefficient as it corresponds to

$$\gamma := \frac{\mu_{\text{low}} - \mu_{\text{medium}}}{2} \text{ where } \mu_{\text{low}} = \mathbb{E}[Y|\text{low}] \text{ and } \mu_{\text{medium}} = \mathbb{E}[Y|\text{medium}]$$

This estimate coefficient is roughly 0.0002 with t-statistic of 0.030 which gives a p-value of 0.97595. Hence, there is no sufficient evidence to reject the null hypothesis H_0: $\gamma = 0$. Therefore, there is no significant change in NOX levels between the low age and medium age categories. Also, with the same reasoning as in (c), the p-value associated to the `ageband2` coefficient makes us reject the null hypothesis that this coefficient is equal to zero which means that there is significant change in NOX levels between the low and medium age reunited categories and the high age category.

These results are consistent with what we could have expected from Figure 1.1: even if the relationship is not strictly linear as we are working on categorical variables, we could see that the mean NOX values for low and medium age categories are roughly the same and are very different from the high age category, which makes sense. Doing this inference using a linear model is however different from just comparing the means of NOX values on each parts of the plot as it takes into account all the other variables of the linear model.

(e) Confidence interval for α_3

We can construct a confidence interval for the main effect parameter for the high age category α_3 in **R**. To proceed, we use the "high" level in AGEBAND as our baseline contrast and use a treatment contrast (dummy coding). Hence, we obtain an estimate of α_3 which will be in this case the intercept term of our linear model.

```
dfm$ageband <- relevel(dfm$ageband, ref="high")
contrasts(dfm$ageband) <- "contr.treatment"
mod <- lm(nox ~ indus + rad + tax + age + ageband, data=dfm)
confint(mod, level=0.99)
```

```
                    0.5 %        99.5 %
(Intercept)    0.3350667937  0.4661293809
```

Thus, $\alpha_3 \in [0.335, 0.466]$ at a 99% level.

2 Question 2

(a) Exponential family member

Consider the probability mass function for a non-negative integer y

$$\mathbb{P}(y \mid p) = \binom{y+r-1}{y} p^y (1-p)^r, \quad y = 0, 1, 2, \ldots \tag{1}$$

with $0 < p < 1$ an unknown parameter and $r \geq 0$ assumed known. We can re-write it as

$$\begin{aligned}\mathbb{P}(y \mid p) &= \exp\left\{\log\binom{y+r-1}{y} + y\log p + r\log(p-1)\right\} \\ &= \exp\left\{\frac{\theta y - b(\theta)}{a(\phi)} + c(y,\phi)\right\}\end{aligned}$$

where

$$a(\phi) = \phi = 1, \quad \theta = \log p, \quad c(y, \phi) = \log \begin{pmatrix} y + r - 1 \\ y \end{pmatrix}, \quad b'(\theta) = \frac{re^\theta}{1 - e^\theta}, \quad b''(\theta) = \frac{re^\theta}{(1 - e^\theta)^2}$$

$$\mu = \frac{rp}{1 - p}, \quad b'^{-1}(\mu) = \log \frac{\mu}{\mu + r}, \quad V(\mu) = \frac{\mu(\mu + r)}{r}$$

Here, μ is the mean and $V(\mu)$ is the variance function. We are here assuming that $r > 0$. If $r = 0$, then $\mathbb{P}(y \mid p) = 0$ and the distribution is not a member of exponential family.

(b) Expectation and variance of this random variable

Using that

$$\mathbb{E}[y \mid \theta, \phi] = b'(\theta) = \mu = \frac{rp}{1 - p}$$

$$\mathbb{V}[y \mid \theta, \phi] = b''(\theta)a(\phi) = \frac{\mu(\mu + r)}{r} = \frac{rp}{(1 - p)^2}$$

Therefore,

$$\frac{\mathbb{E}[y \mid \theta, \phi]}{\mathbb{V}[y \mid \theta, \phi]} = \frac{rp}{(1 - p)} \frac{(1 - p)^2}{rp} = 1 - p < 1$$

Hence, the variance is larger than the mean for this distribution (always assuming $r > 0$).

(c) Limiting distribution as $r \to \infty$

Re-writing the quantity $\mathbb{P}(y \mid p)$ with $p = \frac{\mu}{\mu + r}$ gives the following

$$\begin{aligned} \mathbb{P}(y \mid p) &= \begin{pmatrix} y + r - 1 \\ y \end{pmatrix} \left(\frac{\mu}{\mu + r} \right)^y \left(1 - \frac{\mu}{\mu + r} \right)^r \\ &= \frac{(y + r - 1)!}{(r - 1)!\,(\mu + r)^y} \cdot \frac{\mu^y}{y!} \cdot \left(1 - \frac{\mu}{\mu + r} \right)^r \end{aligned}$$

When $r \to \infty$, $\frac{(y+r-1)!}{(r-1)!\,(\mu+r)^y} \to 1$ as suggested in the question and $\left(1 - \frac{\mu}{\mu+r}\right)^r \to e^{-\mu}$. Hence,

$$\mathbb{P}(y \mid p) \xrightarrow{r \to \infty} e^{-\mu} \cdot \frac{\mu^y}{y!}$$

which corresponds to $\mathbb{P}[Y = y]$ where Y follows a Poisson distribution with parameter $\mu > 0$. Hence, the distribution defined in 1 converges in distribution to a Poisson distribution with parameter μ.

(d) Should we prefer a Poisson distribution?

We have shown above that when $r \to \infty$ then the distribution defined in (1) converges to a Poisson distribution with parameter μ. Hence, we could consider using a Poisson distribution instead of the more complex distribution given in (1) under some circumstances. Both of these distributions may be used for count data. A Poisson distribution with parameter μ has mean μ and variance μ whereas the distribution defined in (1) has larger variance than expectation. However, when $r \to \infty$

$$\frac{\mathbb{E}[y \mid \theta, \phi]}{\mathbb{V}[y \mid \theta, \phi]} = 1 - p = 1 - \frac{\mu}{\mu + r} \to 1$$

Hence, when $r \to \infty$, the expectation and the variance of the distribution (1) converge to the same value, and therefore it is wise to use a Poisson distribution to approximate the other distribution. The Poisson distribution would be easier to handle on calculations.

(e) Link function

Before commenting the choice of any link function, let specify precisely the generalised linear model.

- **Random component:** the components of the random response vector $y = (y_1, \dots, y_n)$ are independent and have the same distribution as (1). We checked in (a) that this distribution was member of the exponential family with $\mathbb{E}[y] = \mu$.
- **Systematic component:** using the covariates $x_i = (x_{i1}, \dots, x_{ip}) \in \mathbb{R}^p$ we construct the matrix X and define the linear predictor $\eta = X\beta$ with $\beta \in \mathbb{R}^p$
- **Link function:** the link between the random and systematic component is $\eta = g(\mu)$, where g must be a monotonic increasing function. Here, we are using the canonical link such that $\eta_i = \theta_i$ for all $i = 1, \dots, n$

Now that the generalised linear model assumptions have been defined, let us focus on the link function. We chose as link function $\eta_i = \theta_i = \log p_i$. Thus, $p_i = e^{\eta_i}$ and this requires that $\eta_i < 0$ as $0 < p < 1$. This can be hard to verify, especially when predictors x_{ij} are sometimes positive and sometimes negative. If for issue we know in advance that all predictors are positive, we would need to have β coefficients negative in order to have $\eta < 0$. Moreover, the link function can be re-written as $\eta_i = \log p_i = \log \frac{\mu_i}{\mu_i + r}$. We are ensured that $\mu_i > 0$ for all i as $\mu_i = \frac{rp}{1-p}$ except for $r = 0$. Also, if $r = 0$ then the link function becomes constant breaking the monotonic increasing assumption.

(f) Deriving the quantities for IWLS algorithm

Assuming the canonical link function, we compute the following quantities which will be required for the Iterative Weighted Least Squares algorithm (IWLS).

- $\eta_i = \log \frac{\mu_i}{\mu_i + r}$. Inverting this function easily gives $\mu_i = \frac{re^{\eta_i}}{1 - e^{\eta_i}}$
- $\frac{\partial \eta_i}{\partial \mu_i} = \frac{\partial}{\partial \mu_i} \left(\log \frac{\mu_i}{\mu_i + r} \right) = \frac{r}{\mu_i(\mu_i + r)}$
- $w_{ii}^{-1} = \left(\frac{\partial \eta_i}{\partial \mu_i} \right)^2 V(\mu_i) = \frac{r}{\mu_i(\mu_i + r)} \iff w_{ii} = \frac{\mu_i(\mu_i + r)}{r}$

(g) R code to perform IWLS

By using the same structure as in the lecture notes [2] (p.90), we write in **R** the following

```
iwls <- function(x, y, init, r){
  beta <- init
  for(i in 1:25){
    eta <- cbind(1, x) %*% beta
    mu <- r * exp(eta)/(1-exp(eta))
    z <- eta + (y-mu) * r/(mu*(mu+r))
    w <- (mu*(mu+r))/r
    lmod <- lm(z~x, weights=w)
    beta <- as.numeric(lmod$coeff)
  }
  return(beta)
}
```

As it is required to write a code that could work for every value of r, we decided to implement a function which takes as arguments the response data `y`, the predictors `x`, an initial guess of the beta coefficients `init` and a value of `r`. The function returns the β coefficients of this generalised linear model.

(h) Running this function on an example

```
x <- dfrm$x
y <- dfrm$y
beta <- iwls(x, y, c(-1,-2), 2)
```

```
[1] -0.5156141 -2.5392669
```

Running the **R** code above to the data in `glmxy.R` gives the results above. We hence estimate the intercept $\beta_1 = -0.5156$ and the slope $\beta_2 = -2.5393$. We notice that this code only works if initial values of β are negative, because as all predictors x are positive and $\eta < 0$ then we need $\beta < 0$ (as discussed in (e)). We may then plot the data and the fitted model to check if the model has converged correctly.

```
plot(x,y,pch=18)
xs <- seq(-1,1,by=0.01)
lines(x=xs,y=2*exp(beta[1]+beta[2]*xs)/(1-exp(beta[1]+beta[2]*xs)),
      lwd=3, col="red")
```

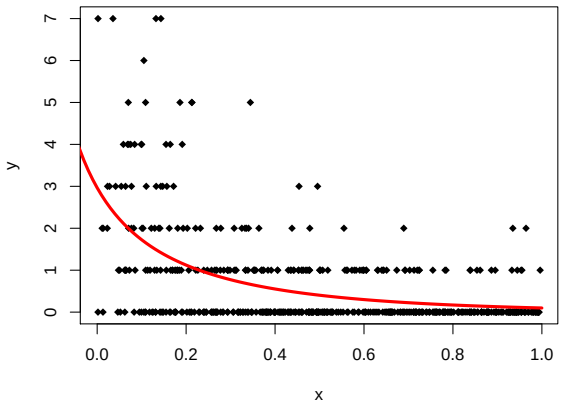

Figure 2.1: Scatter plot of y against x and fitted GLM in red

As shown on Figure 2.1, the fitted GLM curve in red seems to have correctly modelled the data y against x. Implementing manually the IWLS was here necessary as the distribution defined in (1) is not already implemented in **R**, so we could not have used directly the `glm()` function.

References

[1] **R** Documentation. Boston Housing data set description. `https://www.rdocumentation.org/packages/mlbench/versions/2.1-3/topics/BostonHousing`

[2] Dr Din-Houn Lau (Autumn 2020) *MATH70071 - Applied Statistics lecture notes*, Imperial College London MSc Statistics resources