Applied Statistics Assignment 2

CID: 01389741

November 15, 2021

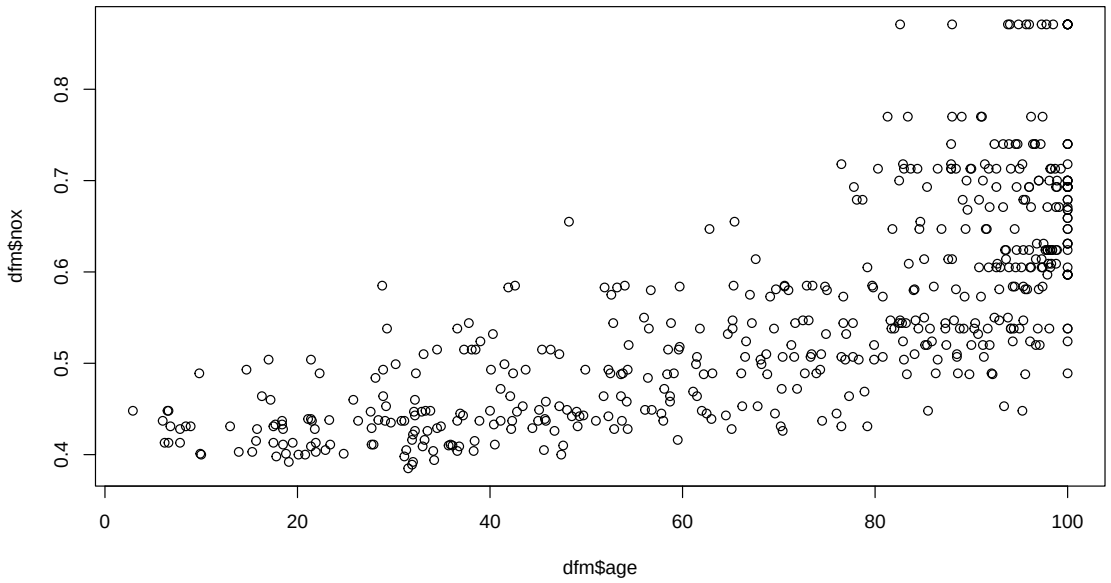
Question 1

In this question we are exploring the Boston Housing Dataset published by Harrison et al [1].

a)

We create a scatter plot between NOX (nitric oxides concentration (parts per 10 million)) and AGE (proportion of owner-occupied units built prior to 1940) to explore possible relationship between the two variables. The dataframe is saved under the name `dfm`.

```
plot(dfm$age,dfm$nox)
```



We can see from the scatterplot that a linear relationship does exists between the variables for AGE more than 25 and less than 80. It seems that for AGE 0-25 there is a roughly constant number of NOX. Partitioning the data for only AGE of 0-25 we see an average NOX value of 0.429.

```
> dfm_25 <- dfm[dfm$age <=25,]
> mean(dfm_25$nox)
[1] 0.4293449
```

For AGE 80-100 we identify that the data might have been right censored since there is a large range of NOX values corresponding to a small number of AGE values. With AGE 25-80 there is a clear linear relationship. These 3 intervals suggest that we should break the AGE variable up by creating three factor variables and see the kind of dependence we have on NOX.

b)

Now we split the AGE variable up corresponding to three intervals, *low*, *medium* and *high*. We store this factor object under the name `ageband` in our original dataframe.

```
> dfm$ageband = cut(dfm$age,c(0,25,80,100), labels = c('low', 'medium', 'high'))
> length(dfm$ageband[dfm$ageband == 'low'])
[1] 49
> length(dfm$ageband[dfm$ageband == 'medium'])
[1] 217
> length(dfm$ageband[dfm$ageband == 'high'])
[1] 240
```

We have 49 observations in the *low* category, 217 in the *medium* category and 240 in the *high* category. The lengths of each interval as shown above add up to 506, the total number of observations we have, meaning that each observation was included in one of the three categories.

c)

Now we will consider a model of the form:

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

where k =1,2,3,4, denotes the respective value of the variables INDUS, RAD, TAX, AGE, β_k are the coefficients of each of those variables. j is referring to the category of the AGEBAND variable with 1,2,3 corresponding to *low*, *medium* and *high* respectively. n_j is referring to the total number of observations in category j and a_j are the coefficients corresponding to each category. This means that intuitively the intercept of the model is going to be affected from the coefficients of the categorical variables. y_{ij} are the observed NOX levels for the ith observation and the jth variable.

To fit such a linear model in `R`, we will use coding for the categorical variable where we create dummy regressors, one for each possible value taking values of 0 and 1. There are multiple ways of doing coding like deviation, helmet and treatment codings, all of which are equivalent when fitting a model, they just have different definitions of assigment to their values. For this coursework we will use `R`'s default, the treatment coding, which orders the levels of characters and defines dummy coding levels in alphabetical order.

```
> contrasts(dfm$ageband)
       medium high
low         0    0
medium      1    0
high        0    1
```

This means that we have two dummy regressors named as *medium* and *high*. Calling these Z_1 and Z_2 respectively we can see that focusing only on the regression with this factor, the model we are trying to regress is:

$$\text{NOX} = \alpha_1 + \alpha_2 Z_1 + \alpha_3 Z_2 + \epsilon \tag{2}$$

where ϵ are the normally distributed errors. By looking the above table we infer that our reference level is *low* corresponding to coefficient α_1. Since we now want to inspect the change in NOX levels between *medium* and *high* age categories we need to change our reference category to be *medium*. This can be done by:

```
dfm$ageband_new = factor(dfm$ageband, levels = c("medium","high","low"))
```

Now performing a linear regression:

```
> lm1 = lm(nox ~  ageband_new + indus + rad + tax + age  , data = dfm)
> summary(lm1)

Call:
lm(formula = nox ~ ageband_new + indus + rad + tax + age, data = dfm)

Residuals:
      Min        1Q    Median        3Q       Max
-0.148538 -0.036610 -0.008176  0.025893  0.240504

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)      3.637e-01  1.672e-02  21.752  < 2e-16 ***
ageband_newhigh  3.688e-02  1.146e-02   3.219  0.00137 **
ageband_newlow  -4.066e-04  1.348e-02  -0.030  0.97595
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 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.06235 on 499 degrees of freedom
Multiple R-squared:  0.7139,        Adjusted R-squared:  0.7105
F-statistic: 207.6 on 6 and 499 DF,  p-value: < 2.2e-16
```

There is multiple information to unpack from here. First we see that from the F statistic it is clear that creating a model with the above estimators is a meaningful test since the null hypothesis of having a model with just an intercept is rejected with a p value very close to zero. Moreover, we see that the intercept (ie. our reference level corresponding to the medium AGE) is statistically significant and it should be included in the model. Now the parameter of interest for concluding whether the change of NOX levels between the medium and high categories is the first estimate namely `ageban_newhigh`. We can see that it has a t value of 3.219 and comparing this to a Student's t-distribution we obtain a p value of 0.00137. This gives us sufficient evidence to reject the null hypothesis (no difference in NOX levels between medium and high AGE) at the 1% significance level.

d)

Now performing the same analysis but this time with *low* as the reference category:

```
> dfm$ageband_new = factor(dfm$ageband, levels = c("low","high","medium"))
> lm2 = lm(nox ~  ageband_new + indus + rad + tax + age  , data = dfm)
> summary(lm2)

Call:
lm(formula = nox ~ ageband_new + indus + rad + tax + age, data = dfm)

Residuals:
      Min        1Q    Median        3Q       Max
-0.148538 -0.036610 -0.008176  0.025893  0.240504

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)       3.633e-01  1.412e-02  25.726  < 2e-16 ***
```

```
ageband_newhigh   3.729e-02  2.111e-02   1.766  0.07793 .
ageband_newmedium 4.066e-04  1.348e-02   0.030  0.97595
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 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.06235 on 499 degrees of freedom
Multiple R-squared:  0.7139,        Adjusted R-squared:  0.7105
F-statistic: 207.6 on 6 and 499 DF,  p-value: < 2.2e-16
```

In this case we see that comparing the *low* and *high* age categories there is no sufficient evidence to reject the Null Hypothesis (no change in NOX levels when comparing low and high age levels) with a t value of 1.766 and p value 0.07793.

This analysis shows that there is change in NOX levels between medium and high categories but no change in NOX levels between low and high categories. This reinforces what we observed in the scatter plot since the low age category was qualitatively thought of as constant and the high age category right censored. So a combination of these two would not give intuitive change in the NOX variable since both seem to have a an attribute that does not coincide with the normal linear regression model. However the combination of medium with high compares a linear relationship with a possible right censored one which suggests that if it was not right censored the high category would also follow a linear relationship.

e)

To obtain a confidence interval for the coefficient of the *high* category we can set *high* as the reference category and observe the confidence interval of the intercept (which corresponds to a_3):

```
> dfm$ageband_new = factor(dfm$ageband, levels = c("high","medium","low"))
> lm3 = lm(nox ~  ageband_new + indus + rad + tax + age  , data = dfm)
> confint(lm3, interval='prediction', level=0.99)
                          0.5 %        99.5 %
(Intercept)        0.3350667937  0.4661293809
ageband_newmedium -0.0665104442 -0.0072525342
ageband_newlow    -0.0918700344  0.0172939001
indus              0.0043960701  0.0079478427
rad                0.0004039671  0.0045097823
tax               -0.0001092510  0.0001361508
age                0.0004520221  0.0017597629
```

So the 99% confidence interval for a_3 is (0.335,0.466).

Question 2

$$\mathrm{p}(y \mid p) = \binom{y+r-1}{y} p^y (1-p)^r, \quad y = 0, 1, 2, \ldots \tag{3}$$

a)

We can rewrite 3 as:

$$\begin{aligned} & p(y \mid p) = \binom{y+r-1}{y} p^y (1-p)^r \\ =& \frac{(y+r-1)!}{y!(r-1)!} \cdot p^y (1-p)^r \\ =& \exp\left[y \log(p) + r \log(1-p) + \log\left(\frac{(y+r-1)!}{y!(r-1)!} \right) \right] \\ =& \exp\left[\frac{\log(p) y - (-r \log(1-p)}{1} + \log\left(\frac{(y+r-1)!}{y!(r-1)!} \right) \right] \end{aligned} \tag{4}$$

From this we can identify:

$$\begin{aligned} & \theta = \log(p) \\ & a(\varphi) = \varphi = 1 \\ & c(y, \varphi) = \log\left(\tfrac{(y+r-1)!}{y!(r-1)!} \right) \\ & b(\theta) = -r \log(1-p) = r \log\left(\tfrac{1}{1-e^\theta} \right) \\ & b'(\theta) = \tfrac{r e^\theta}{1-e^\theta} \\ & b''(\theta) = \tfrac{r e^\theta}{\left(1-e^\theta\right)^2} \\ & \mu = b'(\theta) = \tfrac{r e^\theta}{1-e^\theta} \\ & \mu - \mu e^\theta = r e^\theta \rightarrow e^\theta (r+\mu) = \mu \\ & \therefore b'^{-1}(\theta) = \log\left(\tfrac{\mu}{r+\mu} \right) \\ & V(\mu) = b''(\theta) = \tfrac{r e^\theta}{\left(1-e^\theta\right)^2} = \tfrac{\mu(\mu+r)}{r} \end{aligned} \tag{5}$$

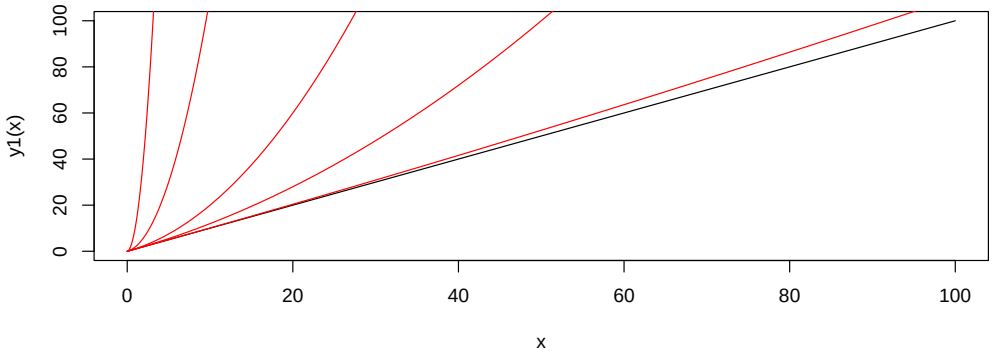
b)

We know from the notes [2] that the equations for the mean and variance are:

$$\begin{aligned} & \mathbb{E}[y \mid \theta, \varphi] = b'(\theta) = \tfrac{r e^\theta}{1-e^\theta} = \mu \\ & \mathrm{Var}(y \mid \theta, \varphi) = a(\varphi) b''(\theta) = \tfrac{\mu(\mu+r)}{r} \end{aligned} \tag{6}$$

We can observe that for $r = 0$ the Variance diverges and for $r > 0$ the Variance will always be larger than the mean μ of this distribution therefore it is over dispersed. This is the case because as r goes to infinity $\frac{\mu(\mu+r)}{r} = \frac{\mu^2}{r} + \frac{\mu r}{r}$ which goes to μ. We can also see this from a simple plot where the red lines corresponding to the variance can never :

```
y1 <- function(x) x
y2 <- function(x, r) x*(x+r)/r
x <- seq(0,100,0.01)
plot(x,y1(x), type='l')
lines(x,y2(x,r=0.1), col='red')
lines(x,y2(x,r=1), col='red')
lines(x,y2(x,r=10), col='red')
lines(x,y2(x,r=50), col='red')
lines(x,y2(x,r=1000), col='red')
```



c)

$$\begin{aligned} \mu &= \tfrac{re^{\theta}}{1-e^{\theta}} = \tfrac{rp}{1-p} \\ \mu - \mu p &= rp \\ p(r+\mu) &= \mu \\ p &= \tfrac{\mu}{r+\mu} \end{aligned} \tag{7}$$

Now the distribution can be written as:

$$p(y \mid p) = \frac{(y+r-1)!}{(r-1)!y!}\left(\frac{\mu}{r+\mu}\right)^{y}\left(\frac{r}{r+\mu}\right)^{r} \tag{8}$$

$$p(y \mid p) = \frac{(y+r-1)!}{(r-1)!(r+\mu)^{y}} \cdot \frac{(\mu)^{y}}{y!} \cdot \left(\frac{1}{1+\frac{\mu}{r}}\right)^{r} \tag{9}$$

And now taking the limit of r $\to \infty$:

$$p(y \mid p) = 1 \cdot \frac{\mu^{y}}{y!} \cdot \left(\frac{1}{e^{\mu}}\right) = \frac{\mu^{y}}{y!}e^{-\mu} \tag{10}$$

d)

Note that the Poisson probability mass function (pmf) with rate μ is defined as:

$$\frac{\mu^{y}}{y!}e^{-\mu} \tag{11}$$

Comparing this with 10, we recognise that they are identical. This means that for a large value of r the two distributions have the same pmf and would therefore be a good idea to approximate this distribution as Poisson which has a smaller variance (Variance of Poisson is just μ). The value of r basically controls the deviation of this distribution to the Poisson.

e)

Under the canonical link we have that:

$$\theta_i = \log(p_i) = \eta_i = \beta x_i \tag{12}$$

Problems might arise sometimes because the link function covers a very limited range. Because p_i can only take values from 0 to 1, the range of the link function is always negative. So when we are trying to find an estimate for β we should make sure we never have $\beta x_i > 0$. Also when p_i is very small the logarithm will diverge to -∞.

f)

$$\mu_i = \frac{re^{\theta_i}}{1-e^{\theta_i}} = \frac{re^{\eta_i}}{1-e^{\eta_i}} \tag{13}$$

$$\begin{array}{c} \eta = \log\left(\frac{\mu}{r+\mu}\right) \\ \frac{\partial n}{\partial_\mu} = \frac{r}{\mu^2+\mu r} \\ \frac{\partial \mu}{\partial n} = \frac{\mu(\mu+r)}{r} \end{array} \tag{14}$$

$$\begin{aligned} \tilde{w}_{ii} &= \frac{1}{V(\mu)}\left(\frac{\partial \mu}{\partial \eta_i}\right)^2_{\mu=\hat{\mu}_i} = \frac{r}{\mu(\mu+r)}\frac{r^2 e^{\eta_i}}{(1-e^{\eta_i})^2} \\ &= \frac{r}{\mu(\mu+r)}\frac{\left(\mu^2+\mu r\right)^2}{r^2} = \frac{\mu(\mu+r)}{r} \end{aligned} \tag{15}$$

g)

For the IWLS algorithm we also need z_i which is:

$$z_i = \hat{\eta}_i + (y_i - \hat{\mu}_i)\frac{r}{\hat{\mu}_i(\hat{\mu}_i + r)} \tag{16}$$

where the hats on each parameter mean that the parameter is the estimate at each iteration.

Now to estimate the coefficients β we will use the quantities μ, z and w derived above. So in `R` code we have:

```
> r=2
> beta <- c(-0.1,-0.5) #initial guess
> for (i in 1:25){
+   eta <- cbind(1,x)%*%beta #estimated linear predictor
+   mu <- r * exp(eta)/(1-exp(eta))  #estimated mean response
+   z <- eta +(y-mu)*r/(mu*(mu+r)) #form the adjusted variate
+   w <- mu*(mu+r)/r #weights
+   lmod <- lm(z~x, weights=w) #regress z on x with weights w
+   beta <- as.numeric(lmod$coeff) #new beta
+   print(beta) #print out the beta estimate every iteration
+ }
```

h)

Adding the file `glmxy.R` in the `R` workspace and adding the x and y variables, we get the following estimates for the regression coefficients with $r = 2$:

```
> beta <- c(-0.1,-0.5) #initial guess
> for (i in 1:10){
+   eta <- cbind(1,x)%*%beta #estimated linear predictor
+   mu <- r * exp(eta)/(1-exp(eta))  #estimated mean response
+   z <- eta +(y-mu)*r/(mu*(mu+r)) #form the adjusted variate
+   w <- mu*(mu+r)/r #weights
+   lmod <- lm(z~x, weights=w) #regress z on x with weights w
+   beta <- as.numeric(lmod$coeff) #new beta
+   print(beta) #print out the beta estimate every iteration
+ }
[1] -0.1842767 -0.8647846
[1] -0.3101147 -1.3609327
[1] -0.4432252 -1.9058417
[1] -0.511774 -2.342236
[1] -0.5167772 -2.5212924
```

```
[1] -0.5156263 -2.5391447
[1] -0.5156141 -2.5392669
[1] -0.5156141 -2.5392669
[1] -0.5156141 -2.5392669
[1] -0.5156141 -2.5392669
```

We can see that after 6 iterations, the ILWS algorithm found estimates of -0.5156 for the intercept and -2.5393 for the slope. Note that if the initial guess for βx_i was positive, the algorithm would not converge. The reason for this is that the link function can only take negative values as explained in part e). For this reason, negative values were used as initial guesses.

References

[1] David Harrison and Daniel L Rubinfeld. Hedonic housing prices and the demand for clean air. *Journal of Environmental Economics and Management*, 5(1):81–102, 1978. ISSN 0095-0696. doi: https://doi.org/10.1016/0095-0696(78)90006-2. URL https://www.sciencedirect.com/science/article/pii/0095069678900062.

[2] Nick Heard. Normal linear model. *Imperial College London (Lecture)*, page 67, October 2021.