625.661 - Homework Six

Eric Niblock

April 9, 2022

1. **A study was conducted attempting to relate home ownership to family income. Twenty households were selected and family income was estimated, along with information concerning home ownership ($y = 1$ indicates yes and $y = 0$ indicates no). Randomly select 15 rows and complete the following.**

   (a) **Fit a logistic regression model to the response variable y. Use a simple linear regression model as the structure for the linear predictor.**

   The attached PDF contains solutions to this problem. A logistic regression model was fit and the following parameters were determined: $\beta_0 = -8.9731, \beta_1 = 0.0002$.

   (b) **Does the model deviance indicate that the logistic regression model from part (a) is adequate?**

   The attached PDF contains solutions to this problem. The model deviance does indicate that the logistic regression model is adequate. The deviance divided by $n - p$ is close to unity (1.206), and at an α-level of 0.05 for χ^2_{14}, we have that $p = 0.26$. Therefore, we cannot reject the null hypothesis that the model is adequate.

   (c) **Provide an interpretation of the parameter β_1 in this model.**

   The attached PDF contains solutions to this problem. Since the model only contains one regressor, we have that,

$$\hat{O}_R = e^{\hat{\beta}_1} = e^{0.000207} = 1.000207 \tag{1}$$

In other words, for every additional dollar earned (everytime x_1 increases by one), there is a 0.000207% increase in the odds of home ownership. β_1 describes the relationship between changes in income to changes in the odds of home ownership.

(d) Expand the linear predictor to include a quadratic term in income. Is there any evidence that this quadratic term is required in the model?

The attached PDF contains solutions to this problem. There is no evidence that the quadratic term is required in this model. The partial deviance, $D(\beta_2|\beta_1)$, was calculated at 2.971. This value is smaller than the critical chi-squared statistic given by $\chi^2_{0.05,1} = 3.841$. This suggests that at a significance level of $\alpha = 0.05$, we cannot reject the null hypothesis that $\beta_2 = 0$. Therefore, there is no evidence that the quadratic term is needed in the model.

2. Myers [1990] presents data on the number of fractures (y) that occur in the upper seams of coal mines in the Appalachian region of western Virginia. Four regressors were reported: x_1 = inner burden thickness (feet), the shortest distance between seam floor and the lower seam; x_2 = percent extraction of the lower previously mined seam; x_3 = lower seam height (feet); and x_4 = time (years) that the mine has been in operation. Randomly select only 30 rows of the data. Complete the following.

(a) Fit a Poisson regression model to these data using the log link.

The attached PDF contains solutions to this problem. A Poisson regression was fit to the data using a log link function.

(b) **Does the model deviance indicate that the model from part (a) is satisfactory?**

The attached PDF contains solutions to this problem. The model deviance does indicate that the logistic regression model is adequate. The deviance divided by $n - p$ is close to unity (0.823), and at an α-level of 0.05 for χ^2_{26}, we have that $p = 0.72$. Therefore, we cannot reject the null hypothesis that the model is adequate.

(c) **Perform a type 3 partial deviance analysis of the model parameters. Does this indicate that any regressors could be removed from the model?**

The attached PDF contains solutions to this problem. A Type 3 partial deviance analysis was performed by finding $D(\beta_i|\beta_{j\neq i})$ for $i \in \{1, 2, 3, 4\}$. Each partial deviance was compared to the critical χ^2_1 value of 3.841 for $\alpha = 0.05$. Since every value fell above the critical value, no regressor was determined to be insignificant, and every regressor should remain in the model.

(d) **Compute Wald statistics for testing the contribution of each regressor to the model. Interpret the results of these test statistics.**

The attached PDF contains solutions to this problem. The Wald statistics are given in the z column of the model summary. The Wald statistic for x_3 suggests that the regressor x_3 is insignificant.

(e) **Find approximate 95% Wald confidence intervals on the model parameters.**

The attached PDF contains solutions to this problem. The 95% confidence intervals are provided in the model summary. The confidence interval for x_3 contains 0, as expected.

In [2]:
```
import pandas as pd
import numpy as np
import statsmodels.api as sm
from statsmodels.genmod.generalized_linear_model import GLM
from statsmodels.genmod import families
```

Problem 1(a)

In [3]:
```
data = pd.DataFrame({'Income': [38000, 51200, 39600, 43400, 47700, 53000, 41500,
                                38000, 42000, 54000, 51700, 39400, 40900, 52800],
                    'Owner Status': [0,1,0,1,0,0,1,0,1,1,1,0,1,0,1,1,1,0,0,1]})
```

In [4]:
```
n = 15
sample = data.sample(n)
sample = data.loc[[3,11,7,10,0,2,13,14,5,8,1,18,12,9,19]]
X = np.array(sample['Income'])
y = np.array(sample['Owner Status'])
sample
```

Out[4]:

	Income	Owner Status
3	43400	1
11	40100	0
7	40800	0
10	38700	1
0	38000	0
2	39600	0
13	38000	0
14	42000	1
5	53000	0
8	45400	1
1	51200	1
18	40900	0
12	49500	1
9	52400	1
19	52800	1

In [5]:
```
X = sm.add_constant(X)
res = GLM(y,X,family=families.Binomial()).fit()
print(res.summary())
```

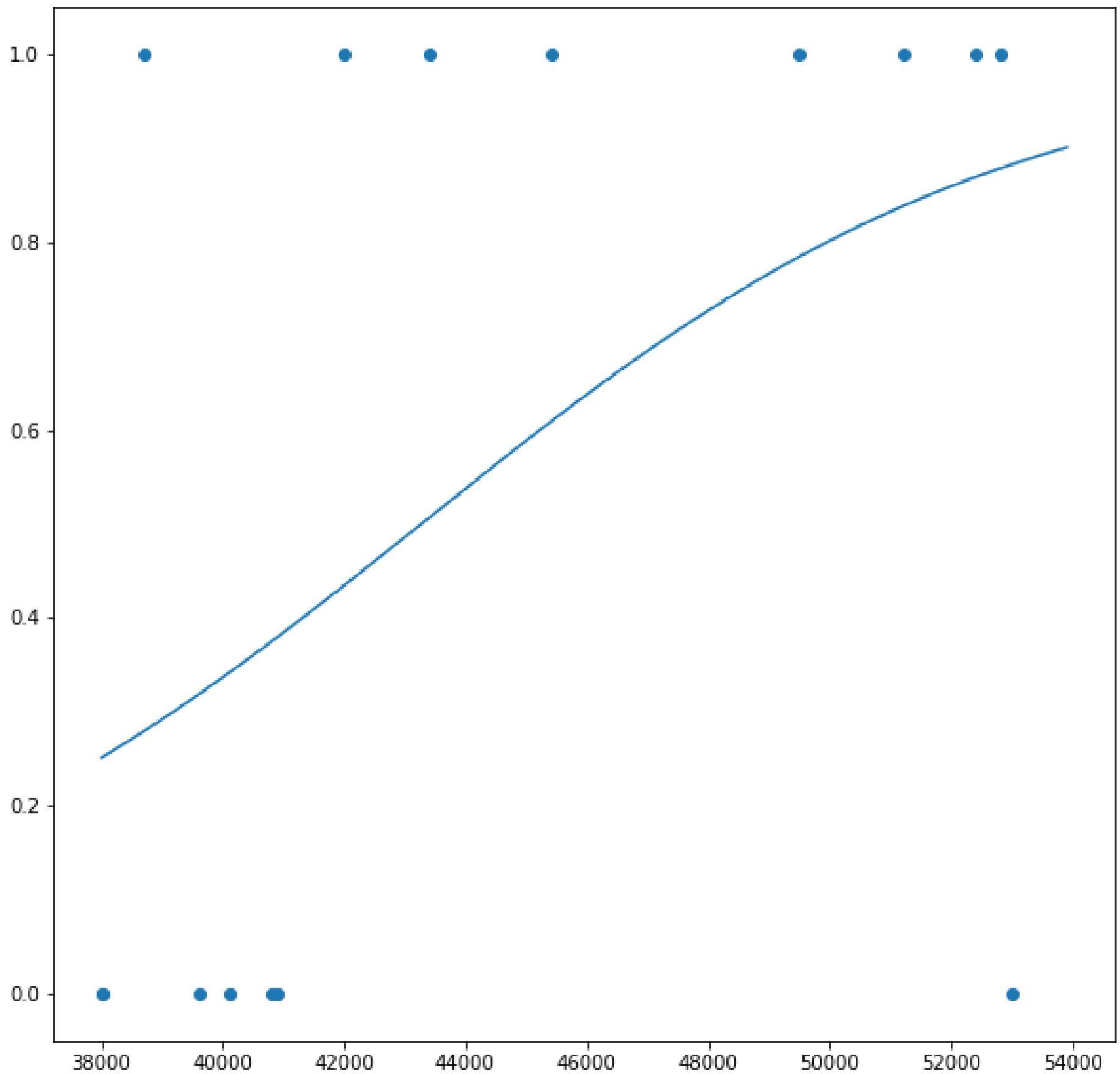
```
                 Generalized Linear Model Regression Results
==============================================================================
Dep. Variable:                      y   No. Observations:                   15
Model:                            GLM   Df Residuals:                       13
Model Family:                Binomial   Df Model:                            1
Link Function:                  logit   Scale:                          1.0000
Method:                          IRLS   Log-Likelihood:                -8.4428
Date:                Wed, 13 Apr 2022   Deviance:                       16.886
Time:                        09:17:55   Pearson chi2:                     16.6
No. Iterations:                     4
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
const         -8.9731      5.270     -1.703      0.089     -19.302       1.356
x1             0.0002      0.000      1.712      0.087   -3.01e-05       0.000
==============================================================================
```

In [7]:
```
import matplotlib.pyplot as plt
%matplotlib inline

sx = np.arange(38000,54000,100)
ps = res.params
sy = [1/(1+np.exp(-1*(ps[0]+i*ps[1]))) for i in sx]

plt.figure(figsize=(10,10))
plt.scatter(X[:,1],y)
plt.plot(sx,sy)
```

Out[7]: [<matplotlib.lines.Line2D at 0x1f3a20e7be0>]



Problem 1(b)

In [104]:
```
print('Deviance/(n-p):    ', 16.886/(15-1))
```

```
Deviance/(n-p):     1.2061428571428572
```

Problem 1(d)

In [14]:
```
newX = np.zeros((len(X),3))
newX[:,:2] = X
newX[:,2] = X[:,1]**2
```

In [103]:
```
newX[:,:2] = X
newX[:,2] = X[:,1]**2
newX
```

Out[103]:
```
array([[1.00000e+00, 4.34000e+04, 1.88356e+09],
       [1.00000e+00, 4.01000e+04, 1.60801e+09],
       [1.00000e+00, 4.08000e+04, 1.66464e+09],
       [1.00000e+00, 3.87000e+04, 1.49769e+09],
       [1.00000e+00, 3.80000e+04, 1.44400e+09],
       [1.00000e+00, 3.96000e+04, 1.56816e+09],
       [1.00000e+00, 3.80000e+04, 1.44400e+09],
       [1.00000e+00, 4.20000e+04, 1.76400e+09],
       [1.00000e+00, 5.30000e+04, 2.80900e+09],
       [1.00000e+00, 4.54000e+04, 2.06116e+09],
       [1.00000e+00, 5.12000e+04, 2.62144e+09],
       [1.00000e+00, 4.09000e+04, 1.67281e+09],
       [1.00000e+00, 4.95000e+04, 2.45025e+09],
       [1.00000e+00, 5.24000e+04, 2.74576e+09],
       [1.00000e+00, 5.28000e+04, 2.78784e+09]])
```

In [17]:
```
res = GLM(y,newX,family=families.Binomial()).fit()
print(res.summary())
```

```
                 Generalized Linear Model Regression Results
==============================================================================
Dep. Variable:                      y   No. Observations:                   15
Model:                            GLM   Df Residuals:                       12
Model Family:                Binomial   Df Model:                            2
Link Function:                  logit   Scale:                          1.0000
Method:                          IRLS   Log-Likelihood:                -6.9573
Date:                Wed, 13 Apr 2022   Deviance:                       13.915
Time:                        11:09:42   Pearson chi2:                     16.4
No. Iterations:                     5
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
const       -144.7134     95.517     -1.515      0.130    -331.924      42.497
x1             0.0062      0.004      1.484      0.138      -0.002       0.014
x2         -6.541e-08   4.52e-08     -1.447      0.148   -1.54e-07    2.32e-08
==============================================================================
```

In [19]:
```
print('D(B1) - D(B):   ', 16.886 - 13.915)
```

```
D(B1) - D(B):    2.971
```

Problem 2(a)

In [98]:
```
data = pd.DataFrame(np.array([[1,    2,    50,    70    , 52   ,   1.0],
[2,    1,    230,    65   ,  42   ,  6.9  ],
[3,    0,    125,    70   ,  45   ,  1.0 ],
[4,    4,    75,    65   ,  68   ,  0.5 ],
[5,    1,    70,    65   ,  53   ,  0.5 ],
[6,    2,    65,    70   ,  46   ,  3.0  ],
[7,    0,    65,    60   ,  62   ,  1.0  ],
[8,    0,    350,    60   ,  54   ,  0.5    ],
[9,    4,    350,    90   ,  54   ,  0.5     ],
[10,    4,    160,    80   ,  38   ,  0.0     ],
[11,    1,    145,    65   ,  38    ,  10.0     ],
[12,    4,    145,    85   ,  38   ,  0.0     ],
[13,    1,    180,    70   ,  42   ,  2.0     ],
[14,    5,    43,    80   ,  40    ,  0.0     ],
[15,    2,    42,    85   ,  51    ,  12.0     ],
[16,    5,    42,    85   ,  51    ,  0.0     ],
[17,    5,    45,    85   ,  42    ,  0.0     ],
[18,    5,    83,    85   ,  48    ,  10.0     ],
[19,    0,    300,    65   ,  68   ,  10.0     ],
[20,    5,    190 ,    90   ,  84   ,  6.0     ],
[21,    1,    145 ,    90   ,  54   ,  12.0     ],
[22,    1,    510  ,  80   ,  57   ,  10.0     ],
[23,    3,    65,    75   ,  68    ,  5.0     ],
[24,    3,    470,    90   ,  90   ,  9.0     ],
[25,    2,    300,    80   ,  165  ,   9.0     ],
[26,    2,    275,    90    ,  40   ,  4.0     ],
[27,    0,    420 ,    50   ,  44   ,  17.0     ],
[28,    1,    65   ,  80   ,  48    ,  15.0   ],
[29,    5,    40   ,  75   ,  51    ,  15.0     ],
[30,    2,    900 ,    90   ,  48   ,  35.0     ],
[31,    3,    95    ,  88   ,  36    ,  20.0     ],
[32,    3,    40   ,  85   ,  57    ,  10.0     ],
[33,    3,    140   ,  90   ,  38   ,  7.0     ],
[34,    0,    150   ,  50   ,  44   ,  5.0     ],
[35,    0,    80    ,  60   ,  96    ,  5.0     ],
[36,    2,    80   ,  85   ,  96    ,  5.0     ],
[37,    0,    145   ,  65   ,  72    ,  9.0     ],
[38,    0,    100   ,  65   ,  72   ,  9.0     ],
[39,    3,    150   ,  80   ,  48   ,  3.0     ],
[40,    2,    150   ,  80   ,  48   ,  0.0     ],
[41,    3,    210   ,  75   ,  42   ,  2.0     ],
[42,    5,    11    ,  75    ,  42    ,  0.0     ],
[43,    0,    100   ,  65 ,    60   ,  25.0     ],
[44,    3,    50    ,  88 ,    60    ,  20.0]]), columns=['Obs','y','x1','x2','x3','x4

n = 30
sample = data.sample(n)
y = np.array(sample[['y']])
X = np.array(sample[['x1','x2','x3','x4']])
sample
```

Out[98]:

	Obs	y	x1	x2	x3	x4
37	38.0	0.0	100.0	65.0	72.0	9.0
41	42.0	5.0	11.0	75.0	42.0	0.0

	Obs	y	x1	x2	x3	x4
17	18.0	5.0	83.0	85.0	48.0	10.0
3	4.0	4.0	75.0	65.0	68.0	0.5
25	26.0	2.0	275.0	90.0	40.0	4.0
7	8.0	0.0	350.0	60.0	54.0	0.5
30	31.0	3.0	95.0	88.0	36.0	20.0
0	1.0	2.0	50.0	70.0	52.0	1.0
16	17.0	5.0	45.0	85.0	42.0	0.0
33	34.0	0.0	150.0	50.0	44.0	5.0
31	32.0	3.0	40.0	85.0	57.0	10.0
32	33.0	3.0	140.0	90.0	38.0	7.0
26	27.0	0.0	420.0	50.0	44.0	17.0
10	11.0	1.0	145.0	65.0	38.0	10.0
4	5.0	1.0	70.0	65.0	53.0	0.5
12	13.0	1.0	180.0	70.0	42.0	2.0
9	10.0	4.0	160.0	80.0	38.0	0.0
20	21.0	1.0	145.0	90.0	54.0	12.0
35	36.0	2.0	80.0	85.0	96.0	5.0
14	15.0	2.0	42.0	85.0	51.0	12.0
15	16.0	5.0	42.0	85.0	51.0	0.0
22	23.0	3.0	65.0	75.0	68.0	5.0
38	39.0	3.0	150.0	80.0	48.0	3.0
34	35.0	0.0	80.0	60.0	96.0	5.0
19	20.0	5.0	190.0	90.0	84.0	6.0
36	37.0	0.0	145.0	65.0	72.0	9.0
2	3.0	0.0	125.0	70.0	45.0	1.0
21	22.0	1.0	510.0	80.0	57.0	10.0
18	19.0	0.0	300.0	65.0	68.0	10.0
5	6.0	2.0	65.0	70.0	46.0	3.0

In [99]:
```
X = sm.add_constant(X)
res = GLM(y,X,family=families.Poisson()).fit()
print(res.summary())
```

```
                 Generalized Linear Model Regression Results
==============================================================================
Dep. Variable:                      y   No. Observations:                   30
Model:                            GLM   Df Residuals:                       25
Model Family:                 Poisson   Df Model:                            4
Link Function:                    log   Scale:                          1.0000
Method:                          IRLS   Log-Likelihood:                -41.682
Date:                Thu, 14 Apr 2022   Deviance:                       21.401
Time:                        10:14:58   Pearson chi2:                     19.5
No. Iterations:                     5
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
const         -3.4214      1.355     -2.525      0.012      -6.077      -0.766
x1            -0.0038      0.002     -2.264      0.024      -0.007      -0.001
x2             0.0656      0.016      4.173      0.000       0.035       0.096
x3            -0.0039      0.008     -0.485      0.627      -0.020       0.012
x4            -0.0592      0.027     -2.185      0.029      -0.112      -0.006
==============================================================================
```

Problem 2(b)

In [100]:
```
print('Deviance/(n-p):    ', 21.401/(30-4))
```

```
Deviance/(n-p):     0.8231153846153846
```

Problem 2(c)

In [101]:
```
c=1
for ind in [[2,3,4],[1,3,4],[1,2,4],[1,2,3]]:
    res = GLM(y,X[:,ind],family=families.Poisson()).fit()
    print('D(B'+str(c)+'|B):  ', res.deviance - 21.401)
    c+=1
```

```
D(B1|B):   16.1012971451708
D(B2|B):   36.410134503819975
D(B3|B):   10.586867186596827
D(B4|B):   9.767909881361216
```