**HAP 823 | Spring 2018 | Assignment 7: Stratified Regression | March 27, 2018**

**Student name:** Jehanzeb Cheema

**Question 1**

* In the first step I constructed strata based on the 41 comorbidities in the data file. This was simply achieved by concatenating the values of the 41 variables as all of these variables are dummy coded in the source data file (see Figure 1).

Next, I aggregated the data file based on stratum and Cancer. This gave me an aggregated data file with each stratum occupying two rows, one row for which Cancer = 1 and a second row for the same stratum with Cancer = 0. This aggregated data file also contained the probability of mortality and number of cases for each combination of stratum and Cancer. It should be noted that for strata with 0 cases or 0 controls there was only one corresponding row of data in this aggregated data file (see Figure 2).

In the next step I restructured the data in such a way that each row represented a single stratum with separate columns for (1) number of cases in the stratum [i.e. observations with Cancer = 1], (2) mortality rate for cases in the stratum, (3) number of controls in the stratum [i.e. observations with Cancer = 0], and (4) mortality rate for controls in the stratum. The total number of strata obtained this way was 203,150 (see Figure 3).

In order to get robust estimates I dropped all strata with less than 10 cases [i.e. observations with Cancer = 1]. In addition I also dropped strata with mortality rate of 1 in the cases sub-group. This left a total of 154 strata in the truncated file (see Figure 4).

* In the final step I regressed the mortality rate of cases on mortality rate of controls in the 154 strata using number of cases with cancer as weight for each stratum. The use of observed *n*'s as weights has the effect of inflating the sample size. Since there were a total of 6,108 cases with cancer in the 154 strata (see Figure 5), this means inflating the sample size from 154 to 6,108. Although this increase in sample size does not affect parameter estimates and some effect size measures such as *R* square, it does affect standard errors of parameter estimates by suppressing them and by extension inflating the observed test statistic values and corresponding *p* values. An alternative to avoid this problem of artificially suppressed standard errors and inflated test statistic values is to calculate normalized weights. For *k* strata, normalized weight *w* for a stratum *i* can be calculated using the following expression:

 (1)

Normalized weights guarantee that the weighted and unweighted sample sizes remain the same, and that the mean normalized weight equals 1 (see Figure 6).

* The simple regression results regressing mortality rate for cases on mortality rate for controls are presented in Figure 7. For comparison regression results weighted by number of cases with cancer are presented in Figure 8. It can be seen that although regression parameter estimates remain unchanged the standard errors and observed test statistic values do get affected. The regression equation is:

 (2)

The intercept of this regression equation, 0.52 represents the effect of Cancer on mortality when no other comorbidity is present.

The results from unweighted regression model are presented in Figure 9. The unweighted regression equation is:

 (3)

The weighted and unweighted intercepts in (2) and (3) are very similar to each other. In order to see how effect sizes as estimated from the stratified regression approach compare with a traditional approach such as logistic regression, a binary logistic regression model predicting mortality from cancer and the 41 comorbidities was estimated. The effect size of each individual variable in this model is the difference in log odds (or the odds ratio). These log odds are plotted against the effect size estimates from stratified regression in Figure 10. The strong linear relationship (*r* = 0.9) supports similarity of the two approaches. The outlier in the upper right hand corner of Figure 10 is the effect of Cancer. Data used for construction of Figure 10 are presented in Table 1.

* Based on the figures presented in Table 1 the equation for predicting mortality from lung cancer and comorbidities can be specified as follows:

Mortality rate = 1 – (1 – 0.52 V1) (1 – 0.05 V2) (1 – 0.11 V3) (1 – 0.03 V4) (1 – 0.02 V5)

(1 – 0.16 V6) (1 – 0.01 V7) (1 – 0.07 V8) (1 – 0.08 V9) (1 – 0.11 V10)

(1 – 0.05 V11) (1 – 0.01 V12) (1 – 0.19 V13) (1 – 0.14 V14) (1 – 0.16 V15)

(1 – 0.06 V16) (1 – 0.05 V17) (1 – 0.13 V18) (1 – 0.17 V19) (1 – 0.04 V20)

(1 – 0 V21) (1 – 0.16 V22) (1 – -0.01 V23) (1 – 0.15 V24) (1 – 0.06 V25)

(1 – 0.02 V26) (1 – 0.01 V27) (1 – 0.06 V28) (1 – 0.01 V29) (1 – 0.02 V30)

(1 – 0.06 V31) (1 – 0.05 V32) (1 – 0.27 V33) (1 – 0.13 V34) (1 – 0 V35)

(1 – 0.04 V36) (1 – 0.1 V37) (1 – 0.02 V38) (1 – 0.09 V39) (1 – 0.02 V40)

(1 – 0.11 V41) (1 – 0.02 V42)

where V1–V42 are as defined in Table 1.

* For the purpose of this exercise I selected the first three comorbidities in order to calculate the prognosis of a patient with lung cancer. The prediction equation thus becomes:

Mortality rate = 1 – (1 – 0.51) (1 – 0.05) (1 – 0.11) (1 – 0.03) (1 – 0)38 = 0.6

The multi-linear form of the equation is:

Mortality rate = 0.52 (V1) + 0.05 (V2) + 0.11 (V3) + 0.03 (V4) – 0.03 (V1 x V2) – 0.06 (V1 x V3)

– 0.01 (V1 x V4) – 0.01 (V2 x V3) – 0.001 (V2 x V4) – 0.003 (V3 x V4)

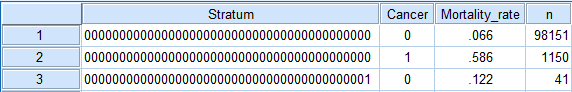
– 0.003 (V1 x V2 x V3) – 0.001 (V1 x V2 x V4) – 0.001 (V1 x V3 x V4)

– 0.0001 (V2 x V3 x V4) – 0.0001 (V1 x V2 x V3 x V4) = 0.6

where V1, V2, V3, and V4 are as defined in Table 10.

|  |  |  |
| --- | --- | --- |
|  | . . . |  |
| .  .  . |  | .  .  . |
|  | . . . |  |

Figure 1



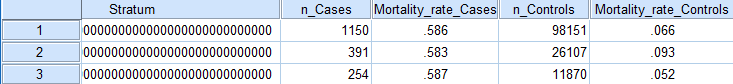
.

.

.



Figure 2



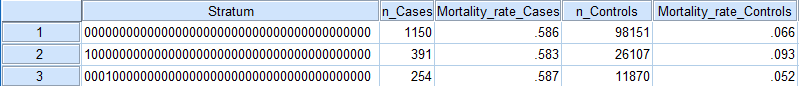
.

.

.



Figure 3



.

.

.

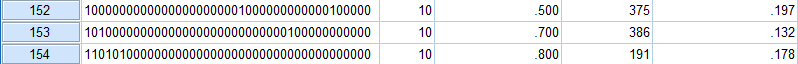


Figure 4

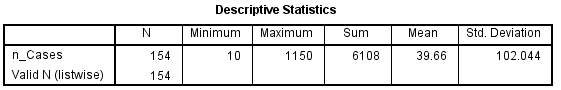


Figure 5: Descriptive statistics for number of cases

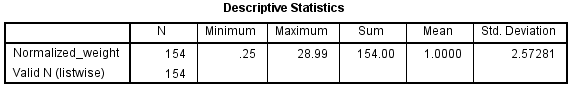


Figure 6: Descriptive statistics for normalized weight

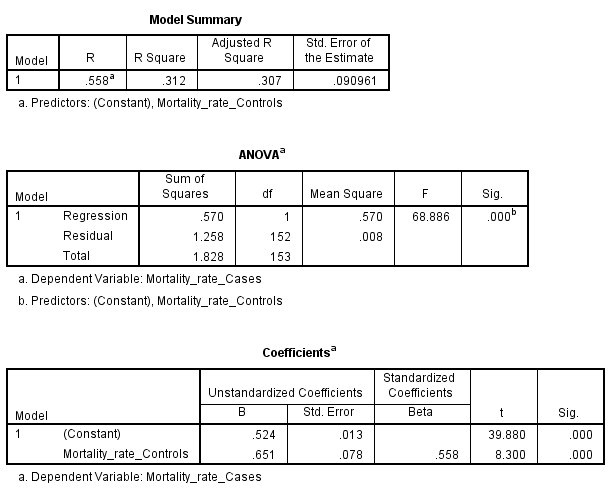


Figure 7: Regression results based on normalized weights

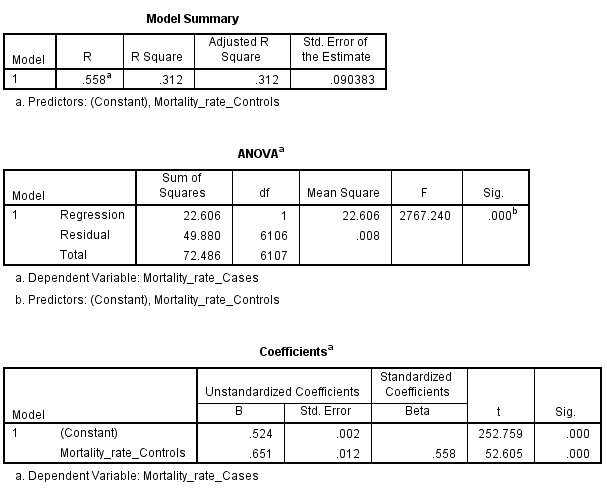


Figure 8: Regression results based on number of cases in the stratum as weight



Figure 9: Unweighted regression results

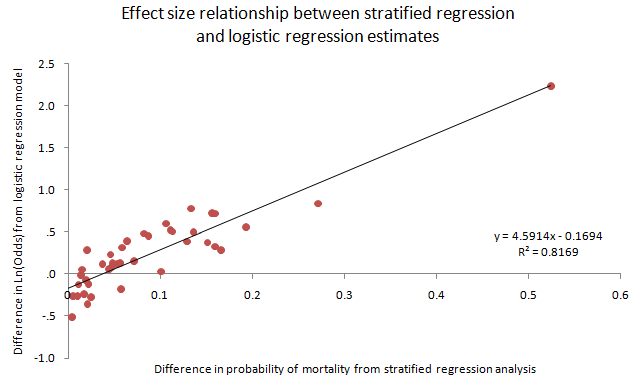


Figure 10

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Diagnosis code** | **Variable ID** | **Description** | **Stratified regression intercept** | **ΔLn(Odds)** |
|  | V1 | Cancer | 0.52 | 9.32 |
| I401.9 | V2 | Essential Primary Hypertension | 0.05 | 1.12 |
| I496 | V3 | Chronic Obstructive Pulmonary Disease with Acute Bronchitis | 0.11 | 1.69 |
| I272.4 | V4 | Other hyperlipidemia | 0.03 | 0.76 |
| I305.1 | V5 | Tobacco use disorder. | 0.02 | 0.79 |
| I486 | V6 | Pneumonia, unspecified organism | 0.16 | 2.06 |
| I530.81 | V7 | Gastro-esophageal reflux disease with esophagitis | 0.01 | 0.88 |
| I414.01 | V8 | Coronary atherosclerosis of native coronary artery | 0.07 | 1.17 |
| I285.9 | V9 | Anemia, unspecified | 0.08 | 1.61 |
| I427.31 | V10 | Atrial fibrillation | 0.11 | 1.65 |
| I600.00 | V11 | Hypertrophy (benign) of prostate without urinary obstruction and other lower urinary tract symptom | 0.05 | 1.14 |
| I311 | V12 | Major depressive disorder, single episode, unspecified | 0.01 | 0.98 |
| I491.21 | V13 | Asthma with Chronic Obstructive Pulmonary Disease | 0.19 | 1.74 |
| I276.1 | V14 | Hypo-osmolality and hyponatremia | 0.14 | 1.65 |
| I428.0 | V15 | Congestive heart failure, unspecified | 0.16 | 2.05 |
| I276.51 | V16 | Dehydration | 0.06 | 1.48 |
| I276.8 | V17 | Hypokalemia | 0.05 | 1.08 |
| I599.0 | V18 | Urinary tract infection, site not specified | 0.13 | 2.17 |
| I403.90 | V19 | Hypertensive chronic kidney disease with stage 1 through stage 4 chronic kidney disease, or unspecified chronic kidney disease. | 0.17 | 1.32 |
| IE849.7 | V20 | Unspecified place in other specified residential institution as the place of occurrence of the external cause | 0.04 | 1.06 |
| I309.81 | V21 | Posttraumatic stress disorder | 0.00 | 0.60 |
| I585.9 | V22 | Chronic kidney disease, unspecified | 0.16 | 1.38 |
| I300.00 | V23 | Anxiety state, unspecified | -0.01 | 0.85 |
| I414.00 | V24 | Coronary atherosclerosis of unspecified type of vessel, native or graft | 0.15 | 1.45 |
| I443.9 | V25 | Peripheral vascular disease | 0.06 | 1.37 |
| I244.9 | V26 | Hypothyroidism, unspecified | 0.02 | 1.33 |
| I724.2 | V27 | Lumbago | 0.01 | 0.77 |
| IV58.61 | V28 | Long term (current) use of anticoagulants | 0.06 | 0.84 |
| I250.00 | V29 | Diabetes mellitus without complications | 0.01 | 0.77 |
| I427.89 | V30 | Other specified cardiac dysrhythmias | 0.02 | 0.89 |
| I788.20 | V31 | Retention of urine, unspecified | 0.06 | 1.14 |
| I280.9 | V32 | Iron deficiency anemia, unspecified | 0.05 | 1.26 |
| I786.6 | V33 | Swelling, mass, or lump in chest | 0.27 | 2.30 |
| I518.89 | V34 | Other diseases of lung, not elsewhere classified | 0.13 | 1.47 |
| I786.59 | V35 | Other chest pain | 0.00 | 0.52 |
| I787.91 | V36 | Diarrhea | 0.04 | 1.12 |
| IV45.81 | V37 | Aortocoronary bypass status | 0.10 | 1.03 |
| IE849.0 | V38 | Home accidents | 0.02 | 1.05 |
| I070.54 | V39 | Chronic hepatitis C without mention of hepatic coma | 0.09 | 1.57 |
| I303.90 | V40 | Other and unspecified alcohol dependence | 0.02 | 0.93 |
| I287.5 | V41 | Thrombocytopenia, unspecified | 0.11 | 1.82 |
| IV45.82 | V42 | Percutaneous transluminal coronary angioplasty status | 0.02 | 0.70 |
| Constant |  |  |  | 0.12 |

Table 1