**Chapter 18**

**Stratified Regression: Rethinking Regression Coefficients**

with Amr ElRafey

**[H1] Learning Objectives**

**[INSERT NL]**

1. Rewrite regression equations so coefficients show unconfounded impact

2. Estimate the unconfounded impact of all independent variables on the outcome

3. Estimate the parameters of stratified regression using structured query language

**[END NL]**

# [H1] Key Concepts

**[INSERT BL]**

* Stratified regression
* Multilinear form
* Multiplicative form
* Correction factors
* Interaction
* Monotone relationship

**[END BL]**

# [H1] Chapter at a Glance

When interactions are present, regression coefficients have no well-defined meaning. They are not, for example, the rate of change in outcome for one unit in the independent variable. In this chapter, we show how stratified covariate balancing can be used to replace regression coefficients with the unconfounded impact of independent variables on the outcome. As a consequence of these steps, the regression equation and its parameters have well-defined meaning.

**[H1] Not in Widespread Use**

This chapter focuses on the interpretation of regression coefficients, when interaction terms are present. Interaction terms are rarely used in regression or other statistical analysis, even though most analysts are encouraged to examine interaction terms. It is not clear why analysts do not use interaction terms in their models. Perhaps it is because of concerns with interpretation of the findings. Some analysts use pair-wise interactions and ignore higher-order interactions. It is possible that higher-order interaction terms are not used because it is impractical to do so. The combination of variables grows exponentially and radically increases the need for data. Managers and analysts working for healthcare organizations may have concluded that interactions do not matter in practical situations.

This is a fallacy. In every hospital discharge, up to 15 diagnostic codes describe the diagnoses that were addressed in the hospital. Many of these diagnostic codes interact with each other. For example, diabetes, kidney, and renal disease often interact with each other. Ignoring these interactions would radically reduce accuracy. Nevertheless, the practice continues with some disregard for accuracy. This chapter was included in the book to help change the practice of statistical modeling for the better, so that every analyst routinely examines interactions, even in high-dimensional data.

# [H1] Background

Regression coefficients do not have a meaningful interpretation when independent variables have interactions—a near guarantee in high-dimensional data. Sure, in regression of *y* on a single independent variable, the coefficient of the independent variable can be interpreted as how the rate of change in the independent variable affects the dependent variable. This interpretation does not make sense if the analyst is examining multiple variables and their interactions. When dealing with multiple variables, the regression coefficients become meaningless (McElreath 2016). Consider the regression of a dependent variable *y* on two independent variables, and :

**[INSERT EQUATION]**

**[END EQUATION]**

In this equation, we are regressing on, and the interaction of the two variables captured by the product of the two variables,. In this situation, none of the three coefficients, have meaning as an effect of the variable on *y*, which is troubling; none can be interpreted as how the rate of change in an independent variable affects the dependent variable. These coefficients were derived by minimizing the sum of the square of residuals; they are essentially parameters that improve the fit; they have no corresponding real-world meaning. In this chapter, we describe stratified regression, a regression procedure that estimates coefficients that have real-world meaning.

# [H1] Multilinear Regression

Regression equations are typically written in a multilinear form, which looks something like

**[INSERT EQUATION]**

.

**[END EQUATION]**

The first term is a constant. If we want the coefficients to be meaningful, we can assume that parameters display the unconfounded impact of independent variables on the outcome *y*. If we want to display the impact of independent variables as a sum of the impact of several separate unconfounded variables (i.e., ), we need to introduce a series of corrections for various combinations of the variables that do not work as a sum of independent variables. Because we have *n* binary variables, we will have 2*n* possible combinations. For each of these combinations, we need to introduce a separate correction factor. The formula, which we call the multilinear form of stratified regression, is

**[INSERT EQUATION]**

**[END EQUATION]**

In the formula, are the correction factors, and the parameter measures the unconfounded impact of the independent variable *r* on *y*. The coefficient is the correction when none of the independent variables are present (the product shows that this correction is applied only when all independent variables are absent); the coefficient is the correction when the variables are present by themselves; the coefficients are the corrections for pair-wise interactions; are the corrections when three independent variables interact; and so on. The parameter shows the average unconfounded impact of the *r*th independent variable on outcome. It is calculated as

**[INSERT EQUATION]**

.

**[END EQUATION]**

In this equation, *s* is an index to the *S* unique combinations of indicates the average of the dependent variable in the stratum *s*; indicates the highest, and indicates the lowest values for The part of the stratified multilinear equation makes intuitive sense and corrects the problem in standard regression in which the average impact of the variable is not reported. The correction factors are there to adjust the impact of the interaction of the independent variables, and each of these correction factors corrects a specific combination of variables. A negative correction shows how much the combined effect is adjusted downward. A positive correction shows how much the combined effect is adjusted upward. The correction factors can be estimated sequentially by starting withand moving up to higher combinations of independent variables using the following six equations:

**[INSERT NL]**

1. = the predicted value from stratified regression,
2. ,
3. ,
4. . . . and so on for higher interaction parameters.

**[END NL]**

# [H1] Example: Predicting Cost of Insurance

An example can demonstrate the calculation of parameters for the stratified multilinear equation. Suppose we want to evaluate the impact of age (above 65, below 65), gender (male, female), and copay (high, low) on cost of insurance. As you review exhibit 18.1, assume that the data are given for 33 observations.

**[INSERT EXHIBIT]**

**Exhibit 18.1** Simulated Data for Impact of Age, Gender, and Copay on Cost

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| ID | Age | Gender | Copay | Cost |
| 1 | 0 | 1 | 1 | $15,539.44 |
| 2 | 0 | 1 | 1 | $15,523.63 |
| 3 | 1 | 0 | 1 | $13,278.28 |
| 4 | 1 | 1 | 1 | $36,591.23 |
| 5 | 1 | 1 | 1 | $36,166.80 |
| 6 | 1 | 0 | 1 | $13,229.83 |
| 7 | 0 | 1 | 1 | $15,298.82 |
| 8 | 0 | 0 | 1 | $5,509.25 |
| 9 | 1 | 0 | 1 | $13,814.60 |
| 10 | 0 | 0 | 1 | $5,887.46 |
| 11 | 0 | 1 | 1 | $15,524.12 |
| 12 | 0 | 0 | 1 | $5,145.90 |
| 13 | 1 | 0 | 1 | $13,438.76 |
| 14 | 1 | 0 | 1 | $13,177.20 |
| 15 | 1 | 1 | 1 | $36,847.78 |
| 16 | 0 | 1 | 1 | $15,104.37 |
| 17 | 1 | 1 | 0 | $11,982.91 |
| 18 | 1 | 0 | 0 | $3,316.66 |
| 19 | 0 | 1 | 0 | $4,039.65 |
| 20 | 1 | 0 | 0 | $3,909.32 |
| 21 | 1 | 0 | 0 | $3,313.63 |
| 22 | 0 | 1 | 0 | $4,034.86 |
| 23 | 0 | 0 | 0 | $1,354.93 |
| 24 | 1 | 1 | 0 | $11,312.31 |
| 25 | 1 | 0 | 0 | $3,674.83 |
| 26 | 1 | 0 | 0 | $3,241.37 |
| 27 | 1 | 0 | 0 | $3,075.76 |
| 28 | 1 | 0 | 0 | $3,634.44 |
| 29 | 1 | 1 | 0 | $11,343.33 |
| 30 | 0 | 1 | 0 | $4,753.82 |
| 31 | 1 | 1 | 0 | $11,027.31 |
| 32 | 0 | 0 | 0 | $1,245.29 |
| 33 | 0 | 1 | 0 | $4,946.03 |
| *Note*: Simulated so that | | | | |

**[END EXHIBIT]**

If we regress cost on independent variables (see exhibit 18.2), the estimated regression coefficient for age is $2,165. This coefficient cannot be interpreted as the rate of increase in cost for a change from young to old. These rates also depend on the interaction of age and gender; age and copay; and age, gender, and copay interaction. From any single regression coefficient, we do not know how much insurance cost goes up. To calculate the average effect of age, we would need to know not only the other regression parameters but also the frequency of occurrence of males and females, as well as the frequency of occurrence of low and high copays. The fact that we cannot read the impact of *x* on *y* from the regression equation is somewhat bizarre, as the very purpose of regression is to estimate the effect of the variable, yet multivariate regression does not allow this. It seems that the effort to fit an equation to the data has been futile, as the equation does not answer our original question. Now, let us look at whether we could rewrite the equation to display the unconfounded impact of independent variables.

**[INSERT EXHIBIT]**

**Exhibit 18.2** Regression of Cost on Age, Gender, and Interaction of Age and Gender

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | Coefficients | Standard Error | *t*-Statistic | *p*-Value | Lower 95% | Upper 95% |
| Intercept | 1374.07 | 169.59 | 8.10 | 0.00 | 1024.80 | 1723.35 |
| Age | 2165.59 | 182.17 | 11.89 | 0.00 | 1790.41 | 2540.77 |
| Gender | 3198.28 | 234.13 | 13.66 | 0.00 | 2716.08 | 3680.48 |
| Copay | 3831.89 | 234.13 | 16.37 | 0.00 | 3349.69 | 4314.09 |
| Age and gender | 4970.88 | 250.86 | 19.82 | 0.00 | 4454.23 | 5487.53 |
| Age and copay | 6150.30 | 273.63 | 22.48 | 0.00 | 5586.74 | 6713.86 |
| Gender and copay | 6947.95 | 310.66 | 22.37 | 0.00 | 6308.14 | 7587.76 |
| Age, gender, and copay | 7861.53 | 382.50 | 20.55 | 0.00 | 7073.74 | 8649.31 |

**[END EXHIBIT]**

## [H1] Estimation of Impact of Independent Variables

To estimate the unconfounded impact, we must estimate the average impact of each variable while holding all other variables constant. We do so through a structured query language (SQL) code that organizes the data into cases (having the high value of the independent variable) and controls (having the low value of the independent variable) across strata defined with the high and low values of the remaining independent variables. For example, the following code shows how the variable *X1* is used to define cases and controls, and variables *X2* and *X3* are used to create strata.

**[LIST FORMAT]**

SELECT Avg(D.Y) AS AvgOfY, D.X2, D.X3

FROM D

INTO CasesX1

WHERE D.X1=1

**GROUP BY** D.X2, D.X3

SELECT Avg(D.Y) AS AvgOfY, D.X2, D.X3

INTO ControlsX1  
 FROM D

WHERE D.X1=0

GROUP BY D.X2, D.X3

SELECT Avg([CasesX1]![AvgOfY]-[ControlsX1]![AvgOfY]) AS k1

INTO k1

FROM ControlsX1 INNER JOIN CasesX1

ON (ControlsX1.X3 = CasesX1.X3) AND (ControlsX1.X2 = CasesX1.X2);

**[END LIST]**

The strata for estimating the unconfounded impact of each variable are shown in exhibit 18.3. From these data, the average impact of each independent variable across the strata is calculated as

**[INSERT EQUATION]**

.

**[END EQUATION]**

Notice that the estimated impact of the independent variable is radically different from the coefficient of the same variable in the equation that generated the data. These estimates reflect not only the main effect of the variable but also its effect while interacting with other variables.

**[INSERT EXHIBIT]**

**Exhibit 18.3** Average Impact While Holding Other Variables Constant

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Gender | Copay | Cases (Old Age) | Controls (Young Age) | Impact of Age |
| Female | High | $3,527.52 | $1,416.58 | $2,110.93 |
| Female | Low | $13,521.85 | $5,205.96 | $8,315.89 |
| Male | High | $11,730.08 | $5,793.82 | $5,936.26 |
| Male | Low | $36,500.49 | $15,352.20 | $21,148.29 |
|  | | | **Average** | $9,377.84 |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Age | Copay | Cases (Males) | Controls (Females) | Impact of Gender |
| Young | High | $5,793.82 | $1,416.58 | $4,377.24 |
| Young | Low | $15,352.20 | $5,205.96 | $10,146.24 |
| Old | High | $11,730.08 | $3,527.52 | $8,202.57 |
| Old | Low | $36,500.49 | $13,521.85 | $22,978.64 |
|  | | | **Average** | $11,426.17 |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Age | Gender | Cases (Low  Copay) | Controls (High  Copay) | Impact  Copay |
| Young | Female | $5,205.96 | $1,416.58 | $3,789.38 |
| Young | Male | $15,352.20 | $5,793.82 | $9,558.38 |
| Old | Female | $13,521.85 | $3,527.52 | $9,994.33 |
| Old | Male | $36,500.49 | $11,730.08 | $24,770.41 |
|  | | | **Average** | $12,028.13 |

**[END EXHIBIT]**

## [H1] Estimation of Correction Factors

If we wanted to express the equation using the stratified multilinear form, the correction factors can be calculated sequentially, starting with the parameter. This parameter is calculated from the average of the observed values for the situation in which all independent variables are at their lowest level. The following SQL code estimates that the correction parameter is $1,416.58:

**[LIST FORMAT]**

SELECT Avg(D.Y) AS C0

INTO C0

FROM D

WHERE D.X1=0 AND D.X2=0 AND D.X3=0;

**[END LIST]**

SQL code can also be used to calculate the interaction correction for a high value for one of the independent variables and low values for the remaining variables. Here we see the calculation of the interaction correction parameter *C*1:

**[LIST FORMAT]**

SELECT Avg([Y]-[X1]\*[k1]) AS C1

INTO k1

FROM D, k1

WHERE D.X1=1 AND D.X2=0 AND D.X3=0;

**[END LIST]**

Similar SQL code can be used to calculate the interaction correction parameter for pairs of independent variables. Here we show the SQL code for the pair *X1* and *X2*:

**[LIST FORMAT]**

SELECT Avg([Y]-[X1]\*([k1]+[c1])-[X2]\*([k2]+[c2])-[X3]\*([k3]+[c3])) AS C12

INTO C12

FROM D, k1, k2, C1, C2

WHERE D.X1=1 AND D.X2=1 AND D.X3=0;

**[END LIST]**

Finally, the correction parameter for the interaction between all three independent variables can be estimated using the following SQL code:

**[LIST FORMAT]**

SELECT Avg([Y] - ([k1]+[C1])\*[X1] - ([k2]+[C2])\*[X2] - ([k3]+[c3])\*[X3] –

[C12]\*[X1]\*[X2] - [C13]\*[X1]\*[X3] - [k23]\*[X2]\*[X3])) AS k123

FROM D, k1, k2, C1, C2, k12, k13, k23

WHERE D.X1=1 AND D.X2=1 AND D.X3=1;

**[END LIST]**

## [H1] Final Write-Up of the Equation

If the average impacts of the independent variables are shown in bold and all correction factors in normal font, then the stratified regression equation is

**[INSERT EQUATION]**

Cost **= $9,377.84** Age + **$11,426.17** Gender + **$12,028.13** Copay

+ $1,416.58 (1 − Age) (1 − Gender) (1 − Copay)

− $5,850.33 Age − $5,632.35 Gender − $6,822.16 Copay

+ $2,408.75 Age × Gender + $4,788.37 Age × Co-pay + $4,352.42 Gender × Copay

+ $10,423.66 Age × Gender × Copay.

**[END EQUATION]**

In this equation, the first three parameters give the average stratified impact of the independent variables; the remaining variables are correction factors.

# [H1] Replacing the Multilinear Model with a Multiplicative Model

Under certain circumstances, the multilinear regression equation (with interaction terms and related correction factors) can be written as a simple multiplicative equation, looking like

**[INSERT EQUATION]**

.

**[END EQUATION]**

In this equation, *y* values are transformed outcome variables now ranging from 0 to 1; are transformed independent variables now ranging between 0 and 1; is the estimated unconfounded impact of on *y*; and *k* is a constant between −1 and 1, which is calculated as . The transformation of dependent and independent variables can be done with the formula

**[INSERT EQUATION]**

**[END EQUATION]**

If we can replace the multilinear model with the multiplicative model, there is no need for the correction factors. A comparison of the two equations shows that the multiplicative model is the same as the multilinear model, if interaction coefficients are replaced with the product of the main effect coefficients:

**[INSERT EQUATION]**

**[END EQUATION]**

The multiplicative model displays the unconfounded impact of an independent variable and does not require correction factors. It radically simplifies our task. The replacement of a multilinear with a multiplicative equation is reasonable if each independent variable is monotonely related to the dependent variable in any subset of data (Keeney and Raiffa 1976; Alemi and Elrafey 2018). A monotone relationship says that the direction of the impact of independent variables on the dependent variable is not reversed. So, if the independent variable has a positive impact, there is no subset of data where it has a negative impact. In essence, the monotone requirement says that there are no surprise reversals in any subset of the data: the estimate of the effect may change but the direction does not.

When the independent variables do not have a monotone relationship, the analyst may be able to transform the variables so that they do. A variable can be divided into several regions, each of which has a monotone relationship with outcome. Suppose age increases the risk associated with hypertension in general but not among nonagenarians. Then, “Age up to 90” can be one independent variable and “Age after 90” can be another, both of which are monotonely related to risk of mortality from hypertension. “Age up to 90” increases the risk; “Age after 90” decreases it. In the regions in which these variables are defined, the relationship between them and risk of mortality is never reversed. Given that it is possible to arrange situations where all independent variables are monotonely related to the dependent variable, there are many situations in which we could replace a multilinear equation with the simpler multiplicative equation.

## [H1] Estimation of Parameters in a Multiplicative Model

To specify the multiplicative function fully, the analyst estimates *n* different *ki* parameters. The *ki* parameters can be estimated by using a corner stratum. A corner stratum is a set of cases in which one—and only one—of the independent variables is at its maximum value. All remaining variables are at their minimum. In a corner stratum, one variable is present and all remaining variables are absent. Consider a situation in which and all remaining *n* − 1 independent variables are zero, that is, . Then, substituting these values into the multiplicative model forms gives us the following relationship:

**[INSERT EQUATION]**

**[END EQUATION]**

Simplifications of these data show that

**[INSERT EQUATION]**

.

**[END EQUATION]**

In short, is estimated as the average value of the outcome in the corner stratum. We can stratify our data and select the corner stratum to estimate the parameters of the multiplicative model.

Unfortunately, corner strata are not always present in data, and if few cases fall into the corner stratum, it may contain errors and therefore cannot be relied on. A strategy is needed to estimate the value of *y* at a corner stratum—even when these cases are not present or infrequent in the data. One approach is to simulate the value of the outcome in the corner stratum from the values of outcome in other strata. Multilevel modeling can be used to remove the effect of all variables and therefore estimate the outcome for the corner stratum.

To accomplish the multilevel modeling, regress the outcome in cases on the outcome in controls. In this regression, each stratum is one data point. Each data point reflects the influence of a combination of independent variables. In the corner stratum, all variables are absent; hence, the average outcome for the controls is zero. Therefore, the intercept to this regression measures the value of the outcome for the corner stratum of cases, where no variables are present.

Suppose we have divided the data into cases and controls at different strata. Cases are composed of the combination of the stratum *s* plus the highest value of the independent variable—that is, , where is the combination of all variables except the variable *i*. Controls are composed of the same stratum *s* but now combined with the lowest value of the independent variables (i.e., . The regression equation has the following form

**[INSERT EQUATION]**

.

**[END EQUATION]**

If this regression is evaluated at the situation where *s* = 0, then all variables are absent, that is, When evaluated at  
s = 0, the regression equation simplifies to

**[INSERT EQUATION]**

.

**[END EQUATION]**

The point is that the outcome for the corner stratum, the estimate for is equal to the intercept of the equation.

Exhibit 18.4 shows the calculation of the corner stratum for the impact of copays on health insurance prices. The first row is a corner stratum, as “young” and “female” set the gender and age variables to their lowest value of zero. The parameter can be estimated as 0.12, which is the difference in the transformed case and control costs. There are only three cases and two controls in this stratum. Therefore, we may feel uncomfortable relying on this stratum only.

**[INSERT EXHIBIT]**

**Exhibit 18.4** Corner Case for Impact of Copay on Insurance Prices

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Strata for Age and Gender | Transformed Case Cost | Transformed Control Cost | Weights | Number of Cases | Number of Controls |
| Young and female | 0.12 | 0.00 | 1.50 | 3 | 2 |
| Young and male | 0.40 | 0.09 | 1.25 | 5 | 4 |
| Old and female | 0.34 | 0.06 | 0.71 | 5 | 7 |
| Old and male | 0.99 | 0.29 | 0.75 | 3 | 4 |

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| We can also estimate the outcome for the corner stratum from the intercept of the regression of the transformed case costs on transformed control costs. The estimated intercept (see regression results in the following table) is 0.13, which is close to what we had estimated from simply looking at the young-and-female stratum.  Summary Output for Regression of Case (copay = 1) Costs on Control (copay = 0) Costs | | | | | | | | |
|  |  |  |  |  |  |  |  |  |
| *Regression Statistics* | |  |  |  |  |  |  |  |
| Multiple R | 0.999 |  |  |  |  |  |  |  |
| R-Squared | 0.998 |  |  |  |  |  |  |  |
| Adjusted R-Squared | 0.997 |  |  |  |  |  |  |  |
| Standard error | 0.020 |  |  |  |  |  |  |  |
| Observations | 4 |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |
| ANOVA |  |  |  |  |  |  |  |  |
|  | *df* | *SS* | *MS* | *F* | Sig *F* |  |  |  |
| Regression | 1 | 0.415 | 0.415 | 1040.722 | 0.001 |  |  |  |
| Residual | 2 | 0.001 | 0.000 |  |  |  |  |  |
| Total | 3 | 0.416 |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |
|  | Coefficients | Standard Error | *t-*Statistic | *p*-Value | Lower 95% | Upper 95% | Lower 95.0% | Upper 95.0% |
| Intercept | 0.130 | 0.014 | 9.086 | 0.012 | 0.069 | 0.192 | 0.069 | 0.192 |
| Control cost | 3.026 | 0.094 | 32.260 | 0.001 | 2.622 | 3.429 | 2.622 | 3.429 |

**[END EXHIBIT]**

It is difficult to conduct multiple regressions in SQL. Luckily, our situation is different and the regression has only one variable. A formula is available to calculate the intercept for single variable regression of *y* on *x*:

**[INSERT EQUATION]**

.

**[END EQUATION005D**

In this equation, *n* is the number of data points, *y* is the dependent variable, and *x* the independent variable. The intercept regression produced the following estimates for the three variables that affected insurance prices: .

## [H1] Determination of Overall Constant *k*

In the multiplicative model, the overall constant *k* can be determined through repeated trial of different values for *k* in its nonlinear formula: . The constant *k*, by definition, ranges from −1 to 1. If the effect of independent variables is less than their sum, one would expect the constant *k* to be a negative number between 0 and −1. If more than the sum, *k* needs to be a positive number between 0 and 1. For rare cases in which the effect of the independent variables is exactly the same as the sum, *k* is 0.

When more than 20 independent variables are present, the constant *k* is guessed to be at its most extreme value. If *k* is negative, it is −1; if it is positive, it is 1. These steps simplify the estimation of *k*. In a later section, we provide the SQL code for estimating the constant.

# [H1] Application of Stratified Multiplicative Regression to Prognosis of Lung Cancer

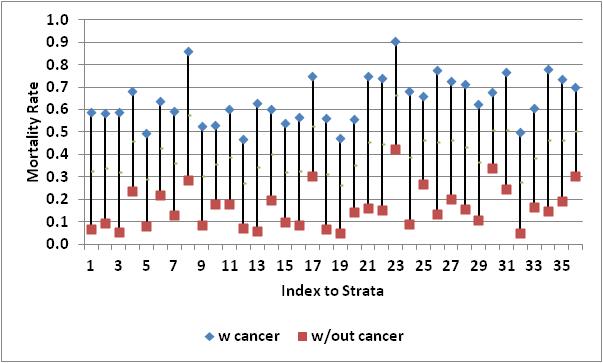
Next, we will examine the prognosis of patients hospitalized for the treatment of lung cancer. The prognosis of cancer patients depends on both the stage of cancer and the presence of various comorbidities (Søgaard et al. 2013; Lee et al. 2010)—comorbidities and their treatment may interfere with cancer therapy. For example, a depressed cancer patient, at any stage of cancer, may abandon cancer treatment prematurely and therefore have a poor prognosis.

We relied on data available through the Veterans Affairs Informatics and Computing Infrastructure (VINCI) for the years 2006 through 2016. We focused on 829,827 unique veterans who (a) had at least two primary care visits and (b) and had been hospitalized at least once during this period. These patients had 17,443,442 diagnoses, 5 to 15 diagnoses per hospitalization. Removing duplicated records and errors in data entry (patients who had visits prior to birth and patients who had a visit after date of death) reduced the number of unique patients to 818,028.

The dependent variable in this study was the probability of mortality within six months. We included 41 comorbidities that occurred at least 1,000 times with lung cancer. Exhibit A.1 in the appendix shows the rate of mortality for lung cancer at different strata. For example, in the first row of the table, we see a stratum where no diagnosis or comorbidity is present. In this stratum, there are 1,150 cases of lung cancer and 98,151 controls with no cancer. The impact of lung cancer is large; it increases the probability of mortality from 0.066 to 0.586, an increase of 0.52 points. The impact of lung cancer changes in each stratum. It ranges from a high of 0.828 points to a low of 0.012 points. In patients with serious comorbidities, the impact of lung cancer on mortality is smaller than in patients with no comorbidities.

Exhibit 18.5 visually shows the data reported in the appendix for the most common strata. The blue diamond indicates the probability of mortality for cases. These are patients who have the comorbidities in the stratum and lung cancer. The square red marker indicates the probability of mortality in the same stratum but without the lung cancer. The *x*-axis indicates the index to the strata. Keep in mind that each stratum is a combination of comorbidities and the index is simply pointing to a specific set of comorbidities. The *y*-axis shows the six-month mortality rate. The impact of cancer is shown by the length of the vertical lines, which show the difference between cases and controls in the same stratum. Notice the variation in the impact of cancer. In general, one expects that in strata with severe comorbidities there will be a smaller impact for cancer, as the patient is already at high risk of death. The data seem to have a ceiling effect—the rate of mortality cannot exceed 1, and if the comorbidities in the strata put the patient at high rates, then the addition of cancer will have no room for a large effect. We are already too close to certain death.

**[INSERT EXHIBIT]**

**Exhibit 18.5** Impact of Lung Cancer at Selected Common Strata  
****

**[END EXHIBIT]**

Surprisingly, in a small number of strata, we have a negative impact for lung cancer. It is logical to assume that every disease, especially lung cancer, will worsen a patient’s prognosis. When it does not, it could be a data anomaly, an estimation error resulting from small sample size, or an unusual pattern among diseases where one disease protects the patient from another. No matter what the explanation is for these anomalies, one must exclude these situations from analysis. In these strata, the multiplicative model is no longer valid, and the analyst should assess a different model. A quick review of the data in the appendix shows that these problems occur rarely (in fewer than 0.025 percent of strata) and can be safely ignored.

To construct the multiplicative model, we need to estimate the impact of cancer when no comorbidities are present (the corner stratum). This stratum is in our data and *k*lung cancer can be estimated at 0.586. Even though it is one of the strata in our data, it is merely a point estimate and subject to random error. It is better if we estimate the value for this corner case by examining the variation in impact of cancer in multiple layers. Across the strata, we regress the cases with cancer on controls without cancer. Then, the intercept to this equation indicates the mortality rate for the corner stratum. The confidence interval and error in estimation of the intercept is also available as an output of the regression equation. Exhibit 18.6 indicates the result of the weighted regression, where each stratum is weighted by the number of cancer cases in the strata. Using the intercept value, the *k*lung cancer = 0.531, which is close to our initial estimate.

**[INSERT EXHIBIT]**

**Exhibit 18.6** Regression of Cases on Controls Without Lung Cancer in Same Strata

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| ANOVA |  |  |  |  |  |  |  |  |
|  | *df* | *SS* | *MS* | *F* | Significance *F* |  |  |  |
| Regression | 1 | 0.373 | 0.373 | 16.485 | 0.000 |  |  |  |
| Residual | 779 | 17.631 | 0.023 |  |  |  |  |  |
| Total | 780 | 18.004 |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |
|  | Coefficients | Standard Error | *t-*Statistic | *p-*Value | Lower95% | Upper95% | Lower 95.0% | Upper 95.0% |
| Intercept | 0.531 | 0.010 | 55.405 | 0.000 | 0.512 | 0.550 | 0.512 | 0.550 |
| Without cancer | 0.160 | 0.040 | 4.060 | 0.000 | 0.083 | 0.238 | 0.083 | 0.238 |

Adjusted R2 = 0.02, *n* = 781.

**[END EXHIBIT]**

A similar analysis can be done for other variables in the multiplicative model. Stratified regression requires us to organize the data into cases composed of “strata plus the variable” and controls composed of “strata alone without the variable,” then the intercept in the regression of cases on controls provides the *ki* constant. Exhibit 18.7 shows the estimation of *ki* parameters for all of the variables, each estimated from the intercept of a separate intercept regression.

**[INSERT EXHIBIT]**

**Exhibit 18.7** Estimated Impact of Different Variables When No Other Variable Is Present

|  |  |  |
| --- | --- | --- |
| Diagnosis Code | Description | Corner Case *Ki* |
|  | Cancer | 0.56 |
| 401.9 | Essential primary hypertension | 0.21 |
| 496 | Chronic obstructive pulmonary disease with acute bronchitis | 0.26 |
| 272.4 | Other hyperlipidemia | 0.30 |
| 305.1 | Tobacco use disorder | 0.24 |
| 486 | Pneumonia, unspecified organism | 0.3 |
| 530.81 | Gastroesophageal reflux disease with esophagitis | 0.25 |
| 414.01 | Coronary atherosclerosis of native coronary artery | 0.22 |
| 285.9 | Anemia, unspecified | 0.24 |
| 427.31 | Atrial fibrillation | 0.29 |
| 600.00 | Hypertrophy (benign) of prostate without urinary obstruction and other lower-urinary-tract symptom | 0.20 |
| 311 | Major depressive disorder, single episode, unspecified | 0.18 |
| 491.21 | Asthma with chronic obstructive pulmonary disease | 0.31 |
| 276.1 | Hypo-osmolality and hyponatremia | 0.27 |
| 428.0 | Congestive heart failure, unspecified | 0.34 |
| 276.51 | Dehydration | 0.23 |
| 276.8 | Hypokalemia | 0.22 |
| 599.0 | Urinary tract infection, site not specified | 0.29 |
| 403.90 | Hypertensive chronic kidney disease with stage 1 through stage 4 chronic kidney disease, or unspecified chronic kidney disease | 0.39 |
| E849.7 | Unspecified place in other specified residential institution as the place of occurrence of the external cause | 0.14 |
| 309.81 | Posttraumatic stress disorder | 0.23 |
| 585.9 | Chronic kidney disease, unspecified | 0.39 |
| 300.00 | Anxiety state, unspecified | 0.19 |
| 414.00 | Coronary atherosclerosis of unspecified type of vessel, native or graft | 0.37 |
| 443.9 | Peripheral vascular disease | 0.23 |
| 244.9 | Hypothyroidism, unspecified | 0.17 |
| 724.2 | Lumbago | 0.21 |
| V58.61 | Long-term (current) use of anticoagulants | 0.28 |
| 250.00 | Diabetes mellitus without complications | 0.21 |
| 427.89 | Other specified cardiac dysrhythmias | 0.18 |
| 788.20 | Retention of urine, unspecified | 0.18 |
| 280.9 | Iron deficiency anemia, unspecified | 0.21 |
| 786.6 | Swelling, mass, or lump in chest | 0.34 |
| 518.89 | Other diseases of lung, not elsewhere classified | 0.23 |
| 786.59 | Other chest pain | 0.27 |
| 787.91 | Diarrhea | 0.22 |
| V45.81 | Aortocoronary bypass status | 0.35 |
| E849.0 | Home accidents | 0.14 |
| 070.54 | Chronic hepatitis C without mention of hepatic coma | 0.19 |
| 303.90 | Other and unspecified alcohol dependence | 0.17 |
| 287.5 | Thrombocytopenia, unspecified | 0.28 |
| V45.82 | Percutaneous transluminal coronary angioplasty status | 0.28 |

**[END EXHIBIT]**

Given the large number of variables, many with a large impact, the overall *k* constant is guessed to be −1. When we check −1 in the equation, we see that it fits. We can now write the equation that predicts mortality rate from lung cancer and its common comorbidities; the comorbidities are shown as a code within brackets:

**[LIST FORMAT]**

Mortality rate = 1 – (1 – 0.56 [Lung Cancer]) (1 – 0.21 [401.9]) (1 – 0.26 [496]) (1 – 0.3 [272.4])

(1 – 0.24 [305.1]) (1 – 0.3 [486]) (1 – 0.25 [530.81]) (1 – 0.22 [414.01]) (1 – 0.24 [285.9])

(1 – 0.29 [427.31]) (1 – 0.2 [600.00]) (1 – 0.18 [311]) (1 – 0.31 [491.21]) (1 – 0.27 [276.1])

(1 – 0.34 [428.0]) (1 – 0.23 [276.51]) (1 – 0.22 [276.8]) (1 – 0.29 [599.0]) (1 – 0.39 [403.90])

(1 – 0.14 [E849.7]) (1 – 0.23 [309.81]) (1 – 0.39 [585.9]) (1 – 0.19 [300.00]) (1 – 0.37 [414.00])

(1 – 0.23 [443.9]) (1 – 0.17 [244.9]) (1 – 0.21 [724.2]) (1 – 0.28 [V58.61]) (1 – 0.21 [250.00])

(1 – 0.18 [427.89]) (1 – 0.18 [788.20]) (1 – 0.21 [280.9]) (1 – 0.34 [786.6]) (1 – 0.23 [518.89])

(1 – 0.27 [786.59]) (1 – 0.22 [787.91]) (1 – 0.35 [V45.81]) (1 – 0.14 [E849.0])

(1 – 0.19 [070.54]) (1 – 0.17 [303.90]) (1 – [287.5]) (1 – 0.28 [V45.82]).

**[END LIST]**

When a diagnosis is present, the variable has the value of 1. When absent, it has a value of 0. For example, if a patient has lung cancer and with unspecified chronic kidney disease (code [585.9]), then these two variables are set to 1, and all other 40 variables are set to 0. The mortality rate is calculated as

**[INSERT EQUATION]**

Mortality rate = 1 – (1 – 0.56) (1 – 0.39) (1 – 0)40 = 0.73.

**[END EQUATION]**

For the 40 diagnoses that are not present, the score is 0, which does not change the predicted mortality rate, which is 0.73. This rate is higher than patients who have only lung cancer and nothing else. Their rate is calculated as

**[INSERT EQUATION]**

Mortality rate = 1 – (1 – 0.56) (1 – 0)41 = 0.56.

**[END EQUATION]**

It is also higher than the mortality rate for patients who just have chronic kidney disease and no cancer or other diseases

**[INSERT EQUATION]**

Mortality rate = 1 – (1 – 0.56 × 0) (1 – 0.39 × 1) (1 – 0)40 = 0.39.

**[END EQUATION]**

Note, however, that it is lower than the sum of these two risk factors (0.95). If we were to write the equation for these two diseases out in a multilinear form of stratified regression, the correction factor of −0.22 would have to be included, like so:

**[INSERT EQUATION]**

Mortality rate = 0.56 (Lung cancer) + 0.39 (Unspecified chronic kidney disease)  
 – 0.22 (Interaction of lung cancer and unspecified chronic kidney disease) = 0.73.

**[END EQUATION]**

The stratified multilinear and multiplicative models produce the same prediction. The parameters of the multiplicative model accomplish both the mathematical optimality and the requirement that these parameters should be conceptually meaningful. Each parameter means something; it indicates the impact of the independent variable by itself when nothing else is present.

# [H1] Structured Query Language Code for Stratified Regression

The estimation of parameters of stratified regression requires repeated calculations that can be facilitated with SQL. The following code shows how the parameters for the stratified multiplicative regression were calculated from the lung cancer data. We start with concatenating all independent variables into a new variable we call *Strata*. This is done to simplify the code and avoid repeatedly referring to the individual comorbidities.

**[LIST FORMAT]**

-- Concatenate the variables into strata

DROP TABLE #Data

SELECT [ID]

,[Dead]

,'C'+ str(Cancer,1)+str([401.9],1)+str([496.],1)+str([272.4],1)+str([305.1],1)+str([486.],1)+str([530.81],1)+str([414.01],1)+str([285.9],1)+str([427.31],1)+str([600.00],1)+str([311.],1)+str([491.21],1)+str([276.1],1)+str([428.0],1)+str([276.51],1)+str([276.8],1)+str([599.0],1)+str([403.90],1)+str([E849.7],1)+str([309.81],1)+str([585.9],1)+str([300.00],1)+str([414.00],1)+str([443.9],1)+str([244.9],1)+str([724.2],1)+str([V58.61],1)+str([250.00],1)+str([427.89],1)+str([788.20],1)+str([280.9],1)+str([786.6],1)+str([518.89],1)+str([786.59],1)+str([787.91],1)+str([V45.81],1)+str([E849.0],1)+str([070.54],1)+str([303.90],1)+str([287.5],1)+str([V45.82],1) AS Strata

INTO #Data

FROM [dbo].[lung]

Go

**[END LIST]**

Next, we create an index that points to the target variable *i*, for which we wish to calculate the *ki*parameter. This parameter is calculated by stratifying the remaining variables and counting cases and controls in the same strata. Note that there are 41 comorbidities and, with cancer itself, there are 42 predictors of mortality. The index variable starts with the first variable, which is set to be lung cancer.

**[LIST FORMAT]**

-- Start an index and repeat for each variable

DECLARE @Index INT

SET @index = 1

WHILE (@Index <=42) – 42 is total number of variables

BEGIN  
**[END LIST]**

For each independent variable, the code must calculate the mortality rate for the combination of various strata with cases (where the target variable is present) and controls (where the target variable is absent). The following prepares the cases at various strata.

**[LIST FORMAT]**

-- Calculate mortality for cases

DROP TABLE #Cases

SELECT SUM(Dead) as cDead -- number dead in cases

, SUM(dead)+SUM(1-dead) AS cCases -- number of cases

-- Set strata to all variables except case/control variable

, STUFF(Strata,@Index,1, '\_') AS cStrata

INTO #Cases

FROM #data

-- Set cases to variable number @index being 1

WHERE SUBSTRING(Strata, @Index, 1)='1'

-- Group by all variables except the case/control variable

GROUP BY STUFF(Strata,@Index,1, '\_')

**[END LIST]**

For the same independent variable, the code calculates the mortality rate for controls at various strata, where the target variable is absent.

**[LIST FORMAT]**

-- Calculate mortality for controls

DROP TABLE #Controls

SELECT SUM(dead) as mDead -- number dead in controls

, SUM(Dead)+SUM(1-Dead) AS mCases -- number of controls

-- set strata to all variables except case/control variable

, STUFF(Allcomorbidties,@Index,1, '\_') AS mStrata

INTO #Controls

FROM #data

-- Set control to variable number @index being not present

WHERE SUBSTRING(Strata, @Index, 1) <> '1'

-- Group by all variables except the case/control variable

GROUP BY STUFF(Strata,@Index,1, '\_')

**[END LIST]**

Next, cases and controls are matched to have the same stratum. In addition, the code drops situations in which cases have a lower mortality rate than controls. Each case has an additional diagnosis, and we are assuming that no diagnosis could reduce the patients’ risk of mortality (i.e., there is no diagnosis that protects the patient from mortality). These regions correspond to situations in which the monotone relationship between the independent variable and mortality is reversed. The multiplicative model is not appropriate for these regions. The good news is that these situations seldom occur (fewer than 0.025 percent of the time). Therefore, it is safe to ignore these few.

**[LIST FORMAT]**

-- Match cases and controls

DROP TABLE #matched

Select @Index as Variable

, Round(CAST(mDead as float)/CAST(mCases as Float),2) AS mProb

, Round(CAST(cDead as float)/CAST(cCases as Float),2) AS cProb

, cCases, mCases

, cStrata AS Strata

INTO #matched

FROM #cases full join #controls on cStrata = mStrata

WHERE cCases>9 and mCases>9 -- Ignore rare strata

**[END LIST]**

In the next step, the intercept regression is carried out and *ki* is estimated:

**[LIST FORMAT]**

-- Calculates Intercept and overlap between cases and controls

Declare @TotalCases as Float

SET @TotalCases = (SELECT SUM(CAST(cCases as FLOAT)) FROM #Cases)

INSERT INTO #Intercept -- Save ki parameters in temporary file

SELECT Max(@index) As [Variable Number]

, (SUM(cProb)\*SUM(mProb\*mProb)-SUM(mProb)\*SUM(cProb\*mProb))

/ (COUNT(mProb)\*SUM(mProb\*mProb)-SUM(mProb)\*SUM(mProb)) AS Intercept

, SUM(CAST(cCases AS Float))/@TotalCases AS Overlap

, SUM(cCases) AS [Cases Matched]

FROM #matched

**[END LIST]**

In the next portion of the code, the procedure continues to the next variable in the multiplicative model. The loop continues until all variables have been classified into cases and controls with their related strata.

**[LIST FORMAT]**

SET @Index = @Index + 1

END

GO

**[END LIST]**

The final portion of the code shows how the overall constant *k* can be estimated using trial and error methods. The estimates of parameters are available in a table called #Intercept. The code first creates a table of possible *k* values between −1 and 1, then tries these values in the equation:

**[INSERT EQUATION]**

.

**[END EQUATION]**

It tries all possible *k* values and selects the *k* values that make the two sides of the equation equal to each other. Because several values may fit the equation, the code randomly selects one.

**[LIST FORMAT]**

-- Estimate overall *k* by trial and error

DROP TABLE #Possible-K, #K

CREATE TABLE #Possible-K (K decimal (3,2))

-- insert possible *k* values into table #PossibleK

INSERT INTO #Possible-K VALUES (-1.), (-.95),(-.9),(-.85), (-.8),

(-.75), (-.7), (-.65), (-.6), (-.55), (-.5), (-.45), (-.4),(-.35),

(-.3),(-.25), (-.2), (-.15), (-.1), (-.05),(.05),(+.1), (.15), (+.2), (.25), (+.3), (.35), (+.4), (.45), (+.5), (.55), (+.6), (.65), (+.7), (.75), (+.8), (.85), (+.9), (.95), (+1.)

SELECT top 1 K -- several *k* values may fit the equation, select top one

INTO #K -- Save the optimal *k* value

FROM #Intercept cross join #Possible-K -- try different *k* values

Group by K

-- right side of equation divided by left side should be near 1

HAVING (-1+ Exp(sum(Log(1+ ki\*k)))/k between 0.99 and 1.01

ORDER BY RAND() -- select among correct *k* values randomly

**[END LIST]**

# [H1] Summary

Stratified multiplicative regression is appropriate in many situations where the relationships between the dependent and the independent variables are monotone in every subset of the data. In these situations, a multiplicative equation replaces the multilinear regression equation. The result is a regression equation that is not only as accurate as the ordinary regression, but also its parameters display the impact of the variables. These parameters have a real-world meaning; they show the impact of the independent variable, by itself and when all other variables are absent, on the dependent variable.

# [H1] Supplemental Resources

A problem set, solutions to problems, multimedia presentations, SQL code, and other related material are in the course website.

**[H1] References**

Alemi, F., and A. Elrafey. 2018. “Estimating Parameters of Multiplicative Utility Models.” Working paper, George Mason University.

Keeney, R. L., and H. Raiffa. 1976. *Decisions with Multiple Objectives*. New York: Wiley.

Lee, L., W. Y. Cheung, E. Atkinson, and M. K. Krzyzanowska. 2011. “Impact of Comorbidity on Chemotherapy Use and Outcomes in Solid Tumors: A Systematic Review.” *Journal of Clinical Oncology* 29 (1): 106–17.

McElreath, R. 2016*. Statistical Rethinking: A Bayesian Course with Examples in R and Stan.* Chapman & Hall/CRC Text in Statistical Science Series. New York: CRC Press.

Søgaard, M., R. W. Thomsen, K. S. Bossen, H. T. Sørensen, and M. Nørgaard. 2013. “The Impact of Comorbidity on Cancer Survival: A Review.” *Clinical Epidemiology* 5 (Suppl 1): 3–29.

# [H1] Appendix

Exhibit A.1 shows 154 strata composed of 41 comorbidities. These strata are useful in predicting mortality for lung cancer patients.

**[INSERT EXHIBIT]**

**Exhibit A.1** Survival from Lung Cancer for Most Common Combinations of Comorbidities Index to Strata

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | Strata  (1 indicates the presence of a lung cancer comorbidity; 0 indicates absence) | Strata and Lung Cancer | | Strata and No Lung Cancer | |
| Number | Death Rate | Number | Death Rate |
| 1 | S00000000000000000000000000000000000000000 | 1,150 | 0.586 | 98,151 | 0.066 |
| 2 | S10000000000000000000000000000000000000000 | 391 | 0.583 | 26,107 | 0.093 |
| 3 | S00010000000000000000000000000000000000000 | 254 | 0.587 | 11,870 | 0.052 |
| 4 | S01000000000000000000000000000000000000000 | 247 | 0.680 | 2,080 | 0.235 |
| 5 | S10100000000000000000000000000000000000000 | 214 | 0.495 | 14,752 | 0.079 |
| 6 | S11000000000000000000000000000000000000000 | 162 | 0.636 | 1,829 | 0.217 |
| 7 | S01010000000000000000000000000000000000000 | 127 | 0.591 | 1,119 | 0.128 |
| 8 | S00001000000000000000000000000000000000000 | 122 | 0.861 | 1,168 | 0.283 |
| 9 | S10010000000000000000000000000000000000000 | 122 | 0.525 | 5,011 | 0.083 |
| 10 | S11100000000000000000000000000000000000000 | 98 | 0.531 | 1,007 | 0.177 |
| 11 | S11010000000000000000000000000000000000000 | 85 | 0.600 | 807 | 0.176 |
| 12 | S10110000000000000000000000000000000000000 | 79 | 0.468 | 3,054 | 0.070 |
| 13 | S00000000100000000000000000000000000000000 | 78 | 0.603 | 1,556 | 0.195 |
| 14 | S00100000000000000000000000000000000000000 | 78 | 0.628 | 7,358 | 0.057 |
| 15 | S00000000000000000010000000000000000000000 | 74 | 0.541 | 2,906 | 0.096 |
| 16 | S10000100000000000000000000000000000000000 | 67 | 0.567 | 4,466 | 0.084 |
| 17 | S00000000000010000000000000000000000000000 | 64 | 0.750 | 819 | 0.302 |
| 18 | S10100100000000000000000000000000000000000 | 59 | 0.559 | 3,722 | 0.064 |
| 19 | S00000100000000000000000000000000000000000 | 55 | 0.473 | 5,622 | 0.050 |
| 20 | S11110000000000000000000000000000000000000 | 52 | 0.558 | 494 | 0.142 |
| 21 | S00000001000000000000000000000000000000000 | 48 | 0.750 | 2,234 | 0.162 |
| 22 | S00000000000000100000000000000000000000000 | 42 | 0.738 | 1,147 | 0.150 |
| 23 | S01001000000000000000000000000000000000000 | 42 | 0.905 | 151 | 0.424 |
| 24 | S10100010000000000000000000000000000000000 | 41 | 0.683 | 3,845 | 0.088 |
| 25 | S10001000000000000000000000000000000000000 | 41 | 0.659 | 463 | 0.268 |
| 26 | S10000000100000000000000000000000000000000 | 40 | 0.725 | 1,385 | 0.201 |
| 27 | S10000010000000000000000000000000000000000 | 40 | 0.775 | 1,884 | 0.133 |
| 28 | S11100010000000000000000000000000000000000 | 38 | 0.711 | 410 | 0.156 |
| 29 | S00000000010000000000000000000000000000000 | 37 | 0.622 | 2,112 | 0.107 |
| 30 | S00000000000100000000000000000000000000000 | 37 | 0.676 | 572 | 0.336 |
| 31 | S00000000000000001000000000000000000000000 | 34 | 0.765 | 1,908 | 0.246 |
| 32 | S00110000000000000000000000000000000000000 | 34 | 0.500 | 2,086 | 0.049 |
| 33 | S01100000000000000000000000000000000000000 | 33 | 0.606 | 487 | 0.162 |
| 34 | S10000000010000000000000000000000000000000 | 32 | 0.781 | 1,819 | 0.147 |
| 35 | S11001000000000000000000000000000000000000 | 30 | 0.700 | 116 | 0.302 |
| 36 | S01000100000000000000000000000000000000000 | 30 | 0.733 | 338 | 0.192 |
| 37 | S10000000000000000000000010000000000000000 | 29 | 0.621 | 1,240 | 0.133 |
| 38 | S10100000000000000000000000010000000000000 | 28 | 0.393 | 3,010 | 0.060 |
| 39 | S11100100000000000000000000000000000000000 | 28 | 0.607 | 316 | 0.117 |
| 40 | S11000010000000000000000000000000000000000 | 28 | 0.643 | 284 | 0.271 |
| 41 | S00000000000000000000000000000000000001000 | 27 | 0.519 | 1,603 | 0.183 |
| 42 | S00000000000000000001000000000000000000000 | 27 | 0.593 | 12,409 | 0.028 |
| 43 | S00011000000000000000000000000000000000000 | 27 | 0.778 | 228 | 0.132 |
| 44 | S01000000100000000000000000000000000000000 | 27 | 0.704 | 119 | 0.403 |
| 45 | S00000000000000000000000000000100000000000 | 26 | 0.538 | 2,348 | 0.103 |
| 46 | S00010100000000000000000000000000000000000 | 26 | 0.500 | 1,392 | 0.053 |
| 47 | S00000010000000000000000000000000000000000 | 26 | 0.654 | 2,128 | 0.125 |
| 48 | S11000000010000000000000000000000000000000 | 26 | 0.654 | 176 | 0.210 |
| 49 | S10000000000100000000000000000000000000000 | 26 | 0.731 | 394 | 0.266 |
| 50 | S00000000000000000000000010000000000000000 | 25 | 0.760 | 1,727 | 0.104 |
| 51 | S00000000000000000000000000000000100000000 | 25 | 0.520 | 339 | 0.195 |
| 52 | S01000010000000000000000000000000000000000 | 25 | 0.880 | 164 | 0.311 |
| 53 | S10000000000000000000000000010000000000000 | 25 | 0.600 | 3,173 | 0.069 |
| 54 | S00000000001000000000000000000000000000000 | 25 | 0.720 | 4,219 | 0.066 |
| 55 | S10010100000000000000000000000000000000000 | 24 | 0.667 | 913 | 0.066 |
| 56 | S11000100000000000000000000000000000000000 | 24 | 0.625 | 439 | 0.169 |
| 57 | S00001000000010000000000000000000000000000 | 24 | 0.875 | 88 | 0.420 |
| 58 | S00001000100000000000000000000000000000000 | 24 | 0.875 | 87 | 0.552 |
| 59 | S00100100000000000000000000000000000000000 | 23 | 0.391 | 1,945 | 0.051 |
| 60 | S11000000100000000000000000000000000000000 | 23 | 0.522 | 144 | 0.375 |
| 61 | S00000000000000010000000000000000000000000 | 22 | 0.818 | 1,000 | 0.113 |
| 62 | S11100000010000000000000000000000000000000 | 22 | 0.545 | 157 | 0.236 |
| 63 | S00000000000000000100100000000000000000000 | 21 | 0.714 | 1,075 | 0.229 |
| 64 | S00000000000000000000000000001000000000000 | 21 | 0.667 | 1,230 | 0.092 |
| 65 | S10100000100000000000000000000000000000000 | 21 | 0.619 | 956 | 0.128 |
| 66 | S10000000000000000001000000000000000000000 | 21 | 0.524 | 2,858 | 0.050 |
| 67 | S00000000000000000000010000000000000000000 | 20 | 0.750 | 2,180 | 0.058 |
| 68 | S10000001000000000000000000000000000000000 | 20 | 0.700 | 1,169 | 0.159 |
| 69 | S10000000000010000000000000000000000000000 | 19 | 0.789 | 462 | 0.208 |
| 70 | S00010000000000000001000000000000000000000 | 19 | 0.579 | 4,295 | 0.027 |
| 71 | S01110000000000000000000000000000000000000 | 19 | 0.632 | 296 | 0.088 |
| 72 | S10100001000000000000000000000000000000000 | 18 | 0.722 | 626 | 0.157 |
| 73 | S01010100000000000000000000000000000000000 | 18 | 0.444 | 190 | 0.147 |
| 74 | S01100100000000000000000000000000000000000 | 18 | 0.722 | 167 | 0.120 |
| 75 | S00000000000000000000000000010000000000000 | 18 | 0.611 | 1,886 | 0.076 |
| 76 | S10101000000000000000000000000000000000000 | 17 | 0.882 | 228 | 0.219 |
| 77 | S00100000100000000000000000000000000000000 | 17 | 0.235 | 280 | 0.104 |
| 78 | S11000000000000000000000010000000000000000 | 17 | 0.706 | 99 | 0.394 |
| 79 | S01000001000000000000000000000000000000000 | 17 | 0.824 | 115 | 0.435 |
| 80 | S00000000000000000000000000000000000010000 | 16 | 0.688 | 2,754 | 0.072 |
| 81 | S11000001000000000000000000000000000000000 | 16 | 0.750 | 99 | 0.394 |
| 82 | S00010000000000000010000000000000000000000 | 16 | 0.250 | 302 | 0.083 |
| 83 | S10000000000000000000000000000100000000000 | 16 | 0.500 | 665 | 0.144 |
| 84 | S10000000001000000000000000000000000000000 | 16 | 0.375 | 1,831 | 0.094 |
| 85 | S11100000100000000000000000000000000000000 | 16 | 0.500 | 87 | 0.241 |
| 86 | S10000000000000000010000000000000000000000 | 16 | 0.188 | 850 | 0.111 |
| 87 | S10100000010000000000000000000000000000000 | 15 | 0.400 | 1,499 | 0.091 |
| 88 | S11101000000000000000000000000000000000000 | 15 | 0.600 | 58 | 0.362 |
| 89 | S00000000000000000000000000100000000000000 | 15 | 0.467 | 785 | 0.094 |
| 90 | S00010001000000000000000000000000000000000 | 15 | 0.867 | 285 | 0.168 |
| 91 | S10100110000000000000000000000000000000000 | 15 | 0.667 | 894 | 0.081 |
| 92 | S00010000000000000000000000000000000001000 | 15 | 0.533 | 611 | 0.151 |
| 93 | S01000000000000000000000000010000000000000 | 15 | 0.533 | 108 | 0.213 |
| 94 | S01000000001000000000000000000000000000000 | 15 | 0.733 | 138 | 0.239 |
| 95 | S00000000000000000000000000000000010000000 | 14 | 0.571 | 1,854 | 0.036 |
| 96 | S01011000000000000000000000000000000000000 | 14 | 0.643 | 94 | 0.245 |
| 97 | S10000000000000000000000000001000000000000 | 14 | 0.429 | 699 | 0.134 |
| 98 | S10110100000000000000000000000000000000000 | 14 | 0.643 | 747 | 0.050 |
| 99 | S00000000000000000000000000000001000000000 | 14 | 0.643 | 188 | 0.319 |
| 100 | S10100000000000000000000010000000000000000 | 14 | 0.500 | 988 | 0.087 |
| 101 | S10100000000000000010000000000000000000000 | 14 | 0.357 | 469 | 0.096 |
| 102 | S01000000000000000010000000000000000000000 | 14 | 0.643 | 88 | 0.295 |
| 103 | S10100000000000000000001000000000000100000 | 14 | 0.714 | 747 | 0.107 |
| 104 | S11100000000000000000000100000000000000000 | 14 | 0.643 | 82 | 0.305 |
| 105 | S01000000000010000000000000000000000000000 | 13 | 0.769 | 75 | 0.387 |
| 106 | S01010010000000000000000000000000000000000 | 13 | 0.692 | 82 | 0.110 |
| 107 | S00010000100000000000000000000000000000000 | 13 | 0.231 | 126 | 0.095 |
| 108 | S10000000000000001000000000000000000000000 | 13 | 0.923 | 740 | 0.301 |
| 109 | S10010000000000000000000000010000000000000 | 13 | 0.615 | 430 | 0.058 |
| 110 | S01000000000000000000000010000000000000000 | 13 | 0.692 | 87 | 0.425 |
| 111 | S10110010000000000000000000000000000000000 | 13 | 0.538 | 1,029 | 0.071 |
| 112 | S11100000000000000000000000010000000000000 | 13 | 0.692 | 166 | 0.102 |
| 113 | S00000000000000000000000000000000000000010 | 12 | 0.750 | 447 | 0.215 |
| 114 | S00000000100000000010000000000000000000000 | 12 | 0.333 | 106 | 0.217 |
| 115 | S10010010000000000000000000000000000000000 | 12 | 0.583 | 439 | 0.116 |
| 116 | S10100100010000000000000000000000000000000 | 12 | 0.667 | 594 | 0.094 |
| 117 | S00000000000000000000000100000000000000000 | 12 | 0.583 | 506 | 0.140 |
| 118 | S00010000000100000000000000000000000000000 | 12 | 0.583 | 309 | 0.178 |
| 119 | S10000000000000000000010000000000000000000 | 12 | 0.833 | 827 | 0.086 |
| 120 | S10000000100000000000000000100000000000000 | 12 | 0.750 | 396 | 0.182 |
| 121 | S10011000000000000000000000000000000000000 | 12 | 0.667 | 90 | 0.178 |
| 122 | S10100000000000000000000100000000000000000 | 12 | 0.667 | 321 | 0.137 |
| 123 | S00000000000000000000001000000000000100000 | 12 | 0.667 | 345 | 0.226 |
| 124 | S01000000010000000000000000000000000000000 | 12 | 0.583 | 173 | 0.243 |
| 125 | S01010000000000000001000000000000000000000 | 12 | 0.583 | 137 | 0.139 |
| 126 | S10010000000000000001000000000000000000000 | 12 | 0.417 | 921 | 0.062 |
| 127 | S11000000000010000000000000000000000000000 | 12 | 0.750 | 66 | 0.318 |
| 128 | S11011000000000000000000000000000000000000 | 12 | 0.917 | 53 | 0.283 |
| 129 | S10000000000000000000000000000000000001000 | 12 | 0.333 | 917 | 0.188 |
| 130 | S10000000000000000000000000000000000010000 | 12 | 0.667 | 896 | 0.107 |
| 131 | S11010000000010000000000000000000000000000 | 12 | 0.583 | 44 | 0.318 |
| 132 | S11110100000000000000000000000000000000000 | 12 | 0.500 | 163 | 0.153 |
| 133 | S01000000000000000000000000000000000001000 | 11 | 0.545 | 115 | 0.209 |
| 134 | S10110000000000000000000100000000000000000 | 11 | 0.545 | 156 | 0.103 |
| 135 | S00010000000000000000000000000000100000000 | 11 | 0.455 | 82 | 0.146 |
| 136 | S10000000000000010000000000000000000000000 | 11 | 0.545 | 831 | 0.124 |
| 137 | S10010000000010000000000000000000000000000 | 11 | 0.636 | 147 | 0.259 |
| 138 | S10010000001000000000000000000000000000000 | 11 | 0.545 | 601 | 0.073 |
| 139 | S11000000000000000000010000000000000000000 | 11 | 0.727 | 67 | 0.239 |
| 140 | S00000000000000000000000000000010000000000 | 11 | 0.727 | 802 | 0.107 |
| 141 | S10000000000000100000000000000000000000000 | 11 | 0.727 | 582 | 0.192 |
| 142 | S00000000000000000000000001000000000000000 | 11 | 0.727 | 3,061 | 0.039 |
| 143 | S10000100010000000000000000000000000000000 | 11 | 0.636 | 594 | 0.111 |
| 144 | S10100010000000000000000000000000000000001 | 11 | 0.455 | 1,291 | 0.058 |
| 145 | S10000000000000000000001000000000000100000 | 10 | 0.500 | 375 | 0.197 |
| 146 | S00000000100000000000000000100000000000000 | 10 | 0.400 | 407 | 0.174 |
| 147 | S00010000000000000000000001000000000000000 | 10 | 0.600 | 1,092 | 0.050 |
| 148 | S00100000000000000000000010000000000000000 | 10 | 0.600 | 674 | 0.079 |
| 149 | S01010000001000000000000000000000000000000 | 10 | 0.800 | 115 | 0.113 |
| 150 | S10100000000000000000000000000100000000000 | 10 | 0.700 | 386 | 0.132 |
| 151 | S00001000000000010000000000000000000000000 | 10 | 0.800 | 57 | 0.228 |
| 152 | S00010000000010000000000000000000000000000 | 10 | 0.800 | 154 | 0.195 |
| 153 | S01010000000000000000000000000001000000000 | 10 | 0.800 | 24 | 0.333 |
| 154 | S11010100000000000000000000000000000000000 | 10 | 0.800 | 191 | 0.178 |

# *Note*: In strata, comorbidities are 1 or 0 for the following order of International Classification of Disease Version 9 codes: 401.9, 496, 272.4, 305.1, 486, 530.81, 414.01, 285.9, 427.31, 600.00, 311, 491.21, 276.1, 428.0, 276.51, 276.8, 599.0, 403.90, E849.7, 309.81, 585.9, 300.00, 414.00, 443.9, 244.9, 724.2, V58.61, 250.00, 427.89, 788.20, 280.9, 786.6, 518.89, 786.59, 787.91, V45.81, E849.0, 070.54, 303.90, 287.5, and V45.82.

**[END EXHIBIT]**