Contents

Modified Insurance data dictionary

Source:

https://www.kaggle.com/datasets/fibonamew/insurance-data

This data was found originally on the web from Kaggle. Copyright provisions are not well defined as this is publicly available data once you create a Kaggle account.

Insurance data for my final project in biostatistics 1 with Dr. Gaddis. The data will be used in hypothesis testing.

This dataset is useful in understanding variables important in determining insurance premium(s) for prospective insurees.

File: insurance_data_final.csv (Health_insurance_data.xlsx)

About this file: The insurance.csv (insurance.xlsx) dataset contains 1033 tuples (rows) and 7 attributes (columns). The dataset contains 4 numerical attributes (age, bmi, children and expenses) and 3 nominal attributes (sex, smoker and region) that were converted into factors with numerical value designated for each level. The data is collected from 4-part regions in the United States. It has 1033 tuples and 7 attributes originally but I added an 8th field unique_id for ease of use in a table joining if needed. Attributes alongside my derived attributes are named in this final dataset:

1. unique_id (unique identifier),
2. age (in years),
3. sex (male vs female),

4. bmi = body mass index (kg/cm^2),
5. children = number of children (no unit),
6. smoker (smoking status reported as yes or no),
7. region (region of origin as in southeast, northwest etc.),
8. expenses (patient health expenses in dollars)
9. BMI_codeverbose is a transformation of BMI numeric data into categories using the CDC guidelines for adult BMI interpretation,
10. BMI_codenum is my conversion of BMI_codeverbose into numeric identifiers for the BMI_codeverbose,
11. Family_size is my interpretation of the numeric variable 5 above (children) into categories for purposes of analysis in SPSS. Family_size encodes the numeric variable children into 3 levels of "small family size" for patients with 0 kids, "normal family size" for patients with 1<=kids<2," large family size" for patients with >3 kids,
12. Famsize_num is the conversion of Family_size into numeric identifiers for analysis in SPSS. 1 = small family size, 2= normal family size and 3=large family size,
13. Region_num is conversion of the string region into numeric identifiers for analysis in SPSS. 1 = southwest,2=northwest,3=southeast,4=northeast,
14. Family_size_recode is my interpretation of the numeric variable 5 above (children) into categories for purposes of analysis in SPSS. Family_size_recode encodes the numeric variable children into 2 levels of "normal family size" for patients with 0<=kids<2, and" large family size" for patients with >3 kids,
15. Family_size_recode_sex is very much identical to Family_size (variable 11) but this time there are 4 levels resulting from separating the small family size into male and female subgroups. The aim of this variable was to get to the root of the bimodal distribution of the small family size,
16. Family_size_recode_sex_num is simply the assignment of numeric identifiers to the levels encoded in Family_size_recode_sex,
17. Family_size_recode_smoker is very much identical to Family_size (variable 11) but this time there are 4 levels resulting from separating the small family size into

smoker and non-smoker subgroups. The aim of this variable was to get to the root of the bimodal distribution of the small family size,

18. Family_size_recode_smoker_num is simply the assignment of numeric identifiers to the levels encoded in Family_size_recode_smoker,
19. Expense_verbose is the segmentation of expenses by “above average” vs “below average”,
20. Expense_code_num is the assignment of numeric identifiers to levels of Expense_verbose with 1= “above average” and 0=” below average”,
21. Fam_size_exp is used to divide the small family size into those who spend above average vs those who spend below average. This dilemma is the most interesting one in the project.
22. Fam_size_exp_num assigns numeric identifiers to attribute 21 above

My final dataset consists of 22 attributes and 1033 tuples. This data is modelled for hypothesis testing using ANOVA and Chi-square test for goodness fit/association.

Excel Code For Key data transformations in my analysis

Attribute	**Excel Code**
BMI_codeverbose	=IFS(D2<18.5,"Underweight”, AND(D2>=18.5,D2<=24.9),"Normal weight”, AND(D2>=25,D2<=29.9),"Overweight",D2>=30,"Obese")
Family_size[1]	=IFS(E2=0,"Small Family Size”, AND(E2>=1,E2<=2),"Normal Family Size",E2>=3,"Large Family Size")
Famsize_num	=IFS(K2="Small Family Size",1,K2="Normal Family Size",2,K2="Large Family Size",3)
Family_size_recode_sex[2]	=IFS(AND(E2=0, C2="female"),"Small Family Size Female”, AND(E2=0,C2="male"),"Small Family Size Male”, AND(E2>=1,E2<=2),"Normal Family Size",E2>=3,"Large Family Size")

[1] My definition of family size for the purpose of this analysis.

[2] Code used in an attempt to filter the bimodal distribution found in the small family size by sex (male vs female)

Family_size_recode_smoker[3]	=IFS(AND(E2=0, F2="yes"),"Small Family Size smoker”, AND(E2=0,F2="no"),"Small Family Size non-smoker”, AND(E2>=1,E2<=2),"Normal Family Size",E2>=3,"Large Family Size")
Expense_verbose[4]	=IF(H2>7610.86,"Above Average”, “Below Average")
Expense_code_num[5]	=IF(T2="Above Average",1,0)
Chi-sq Code Demo for small above average (saa)[6]	=COUNTIFS(L2:L1034,"=1",U2:U1034,"=1")
Small below average (sba)	=COUNTIFS(L2:L1034,"=1",U2:U1034,"=0")
Normal above average (naa)	=COUNTIFS(L2:L1034,"=2",U2:U1034,"=1")
Normal below average (nba)	=COUNTIFS(L2:L1034,"=2",U2:U1034,"=0")
Small family size bimodal decoupling	=IFS(AND(K2="Small Family Size",U2=1),"SFS_AA",AND(K2="Small Family Size",U2=0),"SFS_BA",AND(K2="Normal Family Size",OR(U2=1,U2=0)),"NFS",AND(K2="Large Family Size",OR(U2=1,U2=0)),"LFS")
Assign numeric identifiers to Fam_size_exp_num	=IFS(insurance_data_final!V2="SFS_BA",1,insurance_data_final!P2=" SFS_AA ",2,insurance_data_final!P2=" NFS",3,insurance_data_final!P2="LFS",4)

[3] Code used in an attempt to filter the bimodal distribution found in the small family size by smoker (smoker vs non-smoker)

[4] Expense_verbose divides participants into above average and below average health spending

[5] Assign numeric identifiers to levels of Expense_verbose

[6] U is for the Expense_code_num so above average in this case and L is for Famsize_num so small family size

Proposed Tasks:

1. hypothesis testing
2. Statistical Modeling
3. Exploratory Data Analytics