DNSC 6303 - Programming for Analytics I

Optional Individual Assignment

Source of collected data

I have taken the data from:

The dataset used in Quiz 1.
Name: payor(1).csv

This excellent numeric dataset shows independent and dependent variables are:

- Health premium received. (The only dependent variable)

The dependent variables are:

- Patient Age
- Diabetes
- Blood pressure problem
- Any transplants
- Any Chronic disease
- Patient’s weight
- Patient’s height
- Known allergies

Purpose

Here we will build a predictive model for multiple regression analysis. Here y = Health premium and all others are independent variables (x).

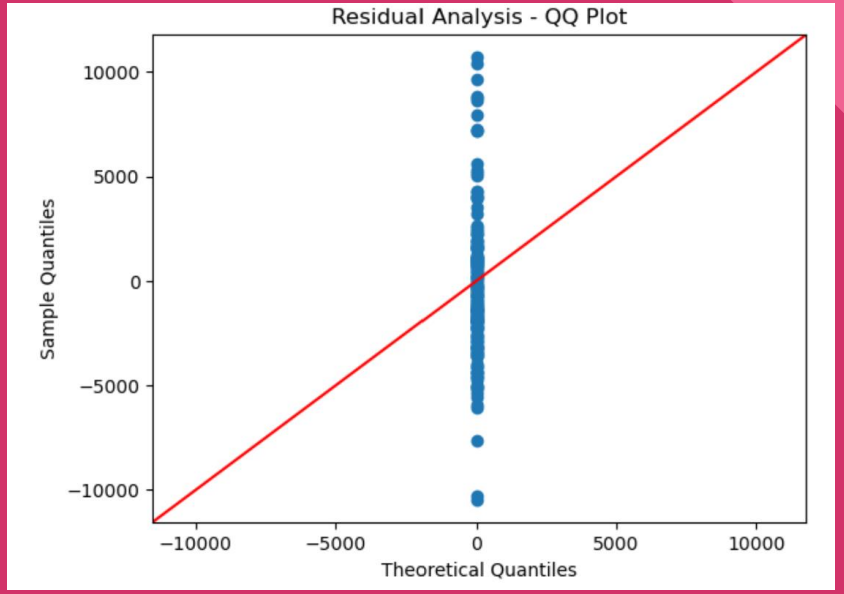
In the regression analysis we will focus into followings:

- **Loading the datasets**
- **Importing libraries**
- **Separating independent and dependent variables.**
- **Splitting the dataset into ‘Train’ and ‘Test’ in to 80:20 ratio.**
- **Model fitting**
- **Estimating parameter results**
- **Model diagnostics**
- **Residual Analysis with QQ plot**

Results

```
(array([ 3.14433510e+02, -5.78919890e+02, -1.63671604e+01,  7.90092116e+03,
         3.14274720e+03, -1.99716614e+00,  6.68128951e+01, -9.03833967e+00]),
 5718.476521736637,
 0.6942621409734047,
 <class 'statsmodels.iolib.summary.Summary'>
"""
                            OLS Regression Results
==============================================================================
Dep. Variable:          HealthPremium   R-squared:                       0.634
Model:                            OLS   Adj. R-squared:                  0.630
Method:                 Least Squares   F-statistic:                     161.9
Date:                Sat, 21 Sep 2024   Prob (F-statistic):          1.59e-157
Time:                        16:08:22   Log-Likelihood:                -7297.1
No. Observations:                 756   AIC:                         1.461e+04
Df Residuals:                     747   BIC:                         1.465e+04
Df Model:                           8
Covariance Type:            nonrobust
==============================================================================================
                             coef    std err          t      P>|t|      [0.025      0.975]
----------------------------------------------------------------------------------------------
const                   5718.4765   2424.738      2.358      0.019     958.365    1.05e+04
PatientAge               314.4335     10.374     30.310      0.000     294.068     334.799
Diabetes                -578.9199    290.947     -1.990      0.047   -1150.092      -7.748
BloodPressureProblems    -16.3672    286.683     -0.057      0.954    -579.167     546.433
AnyTransplants          7900.9212    578.010     13.669      0.000    6766.203    9035.639
AnyChronicDiseases      3142.7472    356.979      8.804      0.000    2441.945    3843.549
PatientHeight             -1.9972     13.767     -0.145      0.885     -29.024      25.030
PatientWeight             66.8129      9.446      7.073      0.000      48.269      85.356
KnownAllergies            -9.0383    337.770     -0.027      0.979    -672.130     654.053
==============================================================================
Omnibus:                      164.577   Durbin-Watson:                   1.834
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              767.603
Skew:                           0.912   Prob(JB):                     2.08e-167
Kurtosis:                       7.587   Cond. No.                     3.34e+03
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 3.34e+03. This might indicate that there are
strong multicollinearity or other numerical problems.
""")
```



Insights & Conclusion

- Age, Transplants, Chronic Diseases, and Weight are key drivers of health premiums.
- Some variables like Blood Pressure Problems and Height do *not* significantly impact premiums.
- The residual analysis shows that the residuals roughly follow a normal distribution.
- The model's R-squared value of 0.694, a reasonable fit, for the test set indicates that 69.4% of the variation in health premiums can be explained by the independent variables.
- Blood Pressure Problems, Patient Height, and Known Allergies have insignificant coefficients.