Strategy and Performance Management

Regression - Timp Health - Group Pre-Assignment

Group 4 : Kirtesh Patel, Nils Marthiensen, Neelesh Bhalla, Chia-Jung Chang

This assignment is available on GitHub under following link:

https://github.com/nilsmart96/fs-spm-assignments/tree/main

In [28]:
```
# Import of necessary libraries
import pandas as pd
import statsmodels.formula.api as sm
```

In [29]:
```
# Import of the data and removal of non-relevent columns
data_url = 'https://github.com/nilsmart96/fs-spm-assignments/blob/main/Regression-
df = pd.read_csv(data_url, delimiter='\t')
df.drop(['RecordID','MemberID'],axis=1,inplace=True)
df
```

Out[29]:

	Month	GrossDrugCost	NLISDummy	LISCHOSERDummy	RiskScore	SpecialtyDummy	Adj
0	6	1242.17	0	0	668.4	1	
1	1	625.86	0	0	290.0	1	
2	6	27.91	0	0	477.2	1	
3	6	46451.23	0	0	2135.6	0	
4	6	6.47	0	0	602.8	0	
...	...	...	...	...	...	...	
30155	2	161.26	1	0	245.6	0	
30156	3	337.39	1	0	245.6	0	
30157	4	358.93	1	0	245.6	1	
30158	5	73.92	1	0	245.6	0	
30159	6	771.96	1	0	245.6	0	

30160 rows × 13 columns

In [30]:
```
# Perform the one-predictor regression
# using the sm.ols() function, with ~ telling us what
# the dependent variable varies over and a fixed
# intercept at 1.
reg_model = sm.ols(formula = 'GrossDrugCost ~ RiskScore + 1', data = df).fit()

print(reg_model.summary())
```

```
                            OLS Regression Results
==============================================================================
Dep. Variable:          GrossDrugCost   R-squared:                       0.075
Model:                            OLS   Adj. R-squared:                  0.075
Method:                 Least Squares   F-statistic:                     2458.
Date:                Tue, 12 Sep 2023   Prob (F-statistic):               0.00
Time:                        20:41:01   Log-Likelihood:            -2.6096e+05
No. Observations:               30160   AIC:                         5.219e+05
Df Residuals:                   30158   BIC:                         5.219e+05
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept    -73.6626     14.748     -4.995      0.000    -102.570     -44.755
RiskScore      1.1117      0.022     49.574      0.000       1.068       1.156
==============================================================================
Omnibus:                    60338.277   Durbin-Watson:                   1.992
Prob(Omnibus):                  0.000   Jarque-Bera (JB):        269266253.564
Skew:                          16.226   Prob(JB):                         0.00
Kurtosis:                     464.754   Cond. No.                     1.22e+03
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly s
pecified.
[2] The condition number is large, 1.22e+03. This might indicate that there are
strong multicollinearity or other numerical problems.
```

Answer of Tasks

1)

Coefficient: This represents the estimated change in GrossDrugCost for a one-unit change in RiskScore. A positive coefficient indicates that as RiskScore increases, GrossDrugCost is expected to increase.

Standard Error: This measures the variability or uncertainty in the coefficient estimate. Smaller standard errors indicate a more precise estimate.

T-statistic (T-stat): It is calculated by dividing the coefficient by its standard error. It measures how many standard errors the coefficient estimate is away from zero. A higher T-stat indicates that the coefficient is more statistically significant.

P-value: It represents the probability of observing the T-statistic (or more extreme values) if the null hypothesis were true. In this case, the null hypothesis would be that there is no relationship between RiskScore and GrossDrugCost. A low P-value (typically < 0.05) suggests that the coefficient is statistically significant.

Interpretation for senior management:

The coefficient for the RiskScore variable is 1.11. It indicates that, on average, for every one-unit increase in RiskScore, we can expect an estimated increase of 1.11 in GrossDrugCost. This means that as the RiskScore of patients increases, their drug costs tend to increase as well.

The standard error of the RiskScore is 0.02. This value measures the uncertainty in our estimate of the relationship between RiskScore and GrossDrugCost. A smaller standard error suggests that our estimate is more precise.

The T-statistic of 49.57 indicates the significance of the RiskScore variable in explaining variations in GrossDrugCost. A larger T-statistic, relative to its standard error, suggests that the relationship between RiskScore and GrossDrugCost is statistically significant.

The P-value of the RiskScore is 0. It represents the probability that the relationship between RiskScore and GrossDrugCost is purely due to random chance. A small P-value (< 0.05) suggests that the relationship is statistically significant.

In summary, the RiskScore variable appears to be a statistically significant predictor of GrossDrugCost. As RiskScore increases, we can expect a corresponding increase in drug costs, and this relationship is not likely to be due to random chance. This information can be valuable for decision-making and resource allocation in managing healthcare costs.

2)

Coefficient: The coefficient for the intercept represents the estimated value of the dependent variable (GrossDrugCost) when all independent variables (in this case, RiskScore) are set to zero. In many practical scenarios, this interpretation may not be meaningful, as it's unlikely that RiskScore can be zero. However, it serves as the estimated starting point for GrossDrugCost when all other factors are absent or negligible.

Standard Error: The standard error associated with the intercept measures the uncertainty or variability in the estimate of the intercept coefficient. A smaller standard error indicates a more precise estimate of the intercept.

T-statistic (T-stat): The T-statistic for the intercept measures how many standard errors the estimated intercept is away from zero. In the case of the intercept, the T-statistic helps determine whether the intercept is significantly different from zero. A higher T-statistic relative to its standard error suggests that the intercept is more statistically significant.

P-value: The P-value associated with the intercept represents the probability of observing the T-statistic (or a more extreme value) if the null hypothesis were true. In this context, the null hypothesis would be that the intercept is zero (i.e., there is no intercept effect). A low P-value (typically < 0.05) suggests that the intercept is statistically significant and not likely to be zero.

Practical Explanation to Senior Management:

The intercept in our regression model represents the estimated starting point for GrossDrugCost when the RiskScore is set to zero. However, in reality, it might not make practical sense for RiskScore to be zero, so the intercept should be interpreted with caution. It essentially gives us a baseline estimate of GrossDrugCost in the absence of any risk factors or predictors.

The standard error associated with the intercept tells us how much uncertainty there is in this baseline estimate. A smaller standard error means that our estimate is more precise and reliable.

The T-statistic for the intercept helps us assess whether the baseline estimate is statistically different from zero. In our analysis, a higher T-statistic suggests that the intercept is more likely to be different from zero, indicating that there is a significant starting point for GrossDrugCost even without considering the risk score.

The P-value associated with the intercept tells us the probability of observing such an intercept value by random chance if there were no actual intercept effect. A low P-value indicates that the intercept is statistically significant, which means that it is highly likely that there is a meaningful starting point for GrossDrugCost.

In summary, the intercept provides us with a baseline estimate of GrossDrugCost when all other factors are negligible, and our analysis suggests that this baseline estimate is statistically significant. This information is important for understanding the fundamental starting point for drug costs, which can be useful for various business decisions and resource allocation in healthcare management.

3)

Adjusted R-Squared is 0.075, which is close to 0. It suggests that the independent variables in the model are not effective in explaining the variance in the dependent variable. This may indicate that the model is a poor fit for the data, and the independent variable(s) may not be good predictors. We would advice the management to use more variables to determine GrossDrugCost.

In [31]:
```
# Calculation for part 4)
round(reg_model.params['RiskScore']*510+reg_model.params['Intercept'],1)
```

Out[31]:
```
493.3
```

4)

The estimated cost with a riskscore of 510 comes out to be 493.3, as shown above