Multiple Regression Model

Analysis

OLS Regression Results

Dep. Variable:	CanDepSymptoms	R-squared:	0.024
Model:	OLS	Adj. R-squared:	0.024
Method:	Least Squares	F-statistic:	60.34
Date:	Sun, 14 Apr 2019	Prob (F-statistic):	1.17e-14
Time:	23:09:29	Log-Likelihood:	-3240.0
No. Observations:	2412	AIC:	6484.
Df Residuals:	2410	BIC:	6496.
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	0.3917	0.021	18.772	0.000	0.351	0.433
MAJORDEP12	0.3812	0.049	7.768	0.000	0.285	0.477

Omnibus:	1485.483	Durbin-Watson:	1.934
Prob(Omnibus):	0.000	Jarque-Bera (JB):	13635.724
Skew:	2.869	Prob(JB):	0.00
Kurtosis:	13.137	Cond. No.	2.70

OLS Regression Results

Dep. Variable:	CanDepSymptoms	R-squared:	0.047
Model:	OLS	Adj. R-squared:	0.044
Method:	Least Squares	F-statistic:	16.88
Date:	Sun, 14 Apr 2019	Prob (F-statistic):	6.77e-22
Time:	20:04:20	Log-Likelihood:	-3212.0
No. Observations:	2412	AIC:	6440.
Df Residuals:	2404	BIC:	6486.
Df Model:	7		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	0.3713	0.021	17.740	0.000	0.330	0.412
MAJORDEP12	0.2599	0.052	5.025	0.000	0.158	0.361
agebeganuse_c	-0.0036	0.003	-1.317	0.188	-0.009	0.002
numberjosmoked_c	0.0038	0.001	2.665	0.008	0.001	0.007
canuseduration_c	-9.436e-06	1.63e-05	-0.579	0.563	-4.14e-05	2.25e-05
GENAXDX12	0.1839	0.102	1.799	0.072	-0.017	0.384
DYSDX12	0.4399	0.100	4.382	0.000	0.243	0.637
SOCPDX12	0.3123	0.085	3.675	0.000	0.146	0.479

Omnibus:	1434.768	Durbin-Watson:	1.947
Prob(Omnibus):	0.000	Jarque-Bera (JB):	12460.255
Skew:	2.761	Prob(JB):	0.00
Kurtosis:	12.669	Cond. No.	7.43e+03

The multiple regression analysis aims to evaluate multiple predictors of the quantitative response variable, number of cannabis dependence symptoms (**CanDepSymptoms**).

The primary explanatory variable is major depression diagnosis, in the last 12 months (**MAJORDEP12**), while the confounding factors are:

- **agebeganuse_c**: Centered quantitative variable, which represents the age when individuals began using cannabis the most (values 5-64. Age).
- **numberjosmoked_c**: Centered quantitative variable, which represents the number of joints smoked per day when using cannabis the most (values 1-98. Joints).
- **canuseduration_c**: Centered quantitative variable, which represents the duration (in weeks) individuals used cannabis the most (values 1-2818. Weeks).
- **GENAXDX12**: Categorical variable, which represents the diagnosis of general anxiety in the last 12 months (0.="No", 1.="Yes").
- **DYSDX12**: Categorical variable, which represents the diagnosis of dysthymia in the last 12 months (0.="No", 1.="Yes").
- **SOCPDX12**: Categorical variable, which represents the diagnosis of social phobia in the last 12 months (0.="No", 1.="Yes").

After adjusting the potential confounding factors, major depression (Beta=0.25, p=0.0001) was significantly and positively associated with number of cannabis dependence symptoms. The R-squared value was extremely small at 0.047 and F-statistic value is equal to 16.88. For the confounding variables:

- Age when began using cannabis the most was not significantly associated with cannabis dependence symptoms and the null hypothesis cannot be rejected (Beta=-0.03, p=0.18).
- Number of joints smoked per day was significantly associated with cannabis dependence symptoms, such that the larger quantity reported a greater number of cannabis dependence symptoms (Beta= 0.003, p=0.008).
- Duration of cannabis use was not significantly associated with cannabis dependence symptoms and the null hypothesis cannot be rejected (Beta=9.4e-06, p=0.56).
- General anxiety diagnosis was not significantly associated with cannabis dependence symptoms and the null hypothesis cannot be rejected (Beta=0.18, p=0.07).
- Dysthymia diagnosis was significantly associated with cannabis dependence symptoms (Beta= 0.43, p=0.0001).
- Social phobia diagnosis was significantly associated with cannabis dependence symptoms (Beta= 0.31, p=0.0001).

Report

To evaluate if the additional explanatory variables confounded the relationship between major depression diagnosis (primary explanatory variable) and cannabis dependence symptoms (response variable), I added the variables to my model one at a time. As a result, none of this variables confounded the association, since every time I added each predictor the p-value of major depression remained significant, at 0.0001. Therefore, the results of the multiple regression analysis for these adjusted potential confounding variables, supported my initial hypothesis.

Polynomial Regression

OLS Regression Results

Dep. Variable:	CanDepSymptoms	R-squared:	0.003
Model:	OLS	Adj. R-squared:	0.002
Method:	Least Squares	F-statistic:	6.511
Date:	Sun, 14 Apr 2019	Prob (F-statistic):	0.0108
Time:	23:09:29	Log-Likelihood:	-3266.6
No. Observations:	2412	AIC:	6537.
Df Residuals:	2410	BIC:	6549.
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	0.4606	0.019	24.122	0.000	0.423	0.498
numberjosmoked_c	0.0037	0.001	2.552	0.011	0.001	0.006

Omnibus:	1511.551	Durbin-Watson:	1.953
Prob(Omnibus):	0.000	Jarque-Bera (JB):	14218.489
Skew:	2.928	Prob(JB):	0.00
Kurtosis:	13.353	Cond. No.	13.3

OLS Regression Results

Dep. Variable:	CanDepSymptoms	R-squared:	0.099
Model:	OLS	Adj. R-squared:	0.098
Method:	Least Squares	F-statistic:	132.5
Date:	Sun, 14 Apr 2019	Prob (F-statistic):	2.59e-55
Time:	23:09:29	Log-Likelihood:	-3144.0
No. Observations:	2412	AIC:	6294.
Df Residuals:	2409	BIC:	6311.
Df Model:	2		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	0.6519	0.022	30.023	0.000	0.609	0.694
numberjosmoked_c	0.0970	0.006	16.244	0.000	0.085	0.109
I(numberjosmoked_c ** 2)	-0.0011	6.69e-05	-16.055	0.000	-0.001	-0.001

Omnibus:	1457.988	Durbin-Watson:	1.970
Prob(Omnibus):	0.000	Jarque-Bera (JB):	14231.206
Skew:	2.767	Prob(JB):	0.00
Kurtosis:	13.535	Cond. No.	1.46e+03

OLS Regression Results

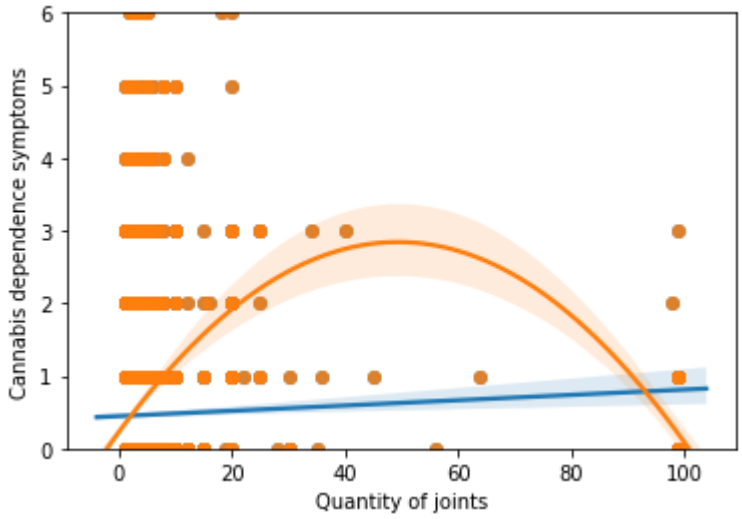
Dep. Variable:	CanDepSymptoms	R-squared:	0.189
Model:	OLS	Adj. R-squared:	0.188
Method:	Least Squares	F-statistic:	185.2
Date:	Sun, 14 Apr 2019	Prob (F-statistic):	5.83e-108
Time:	21:18:44	Log-Likelihood:	-3000.8
No. Observations:	2394	AIC:	6010.
Df Residuals:	2390	BIC:	6033.
Df Model:	3		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	0.5658	0.021	26.510	0.000	0.524	0.608
numberjosmoked_c	0.0516	0.006	8.147	0.000	0.039	0.064
I(numberjosmoked_c ** 2)	-0.0006	7.07e-05	-8.263	0.000	-0.001	-0.000
CUFREQ_c	0.0943	0.006	16.282	0.000	0.083	0.106

Omnibus:	1376.030	Durbin-Watson:	2.013
Prob(Omnibus):	0.000	Jarque-Bera (JB):	12858.709
Skew:	2.601	Prob(JB):	0.00
Kurtosis:	13.092	Cond. No.	1.39e+03

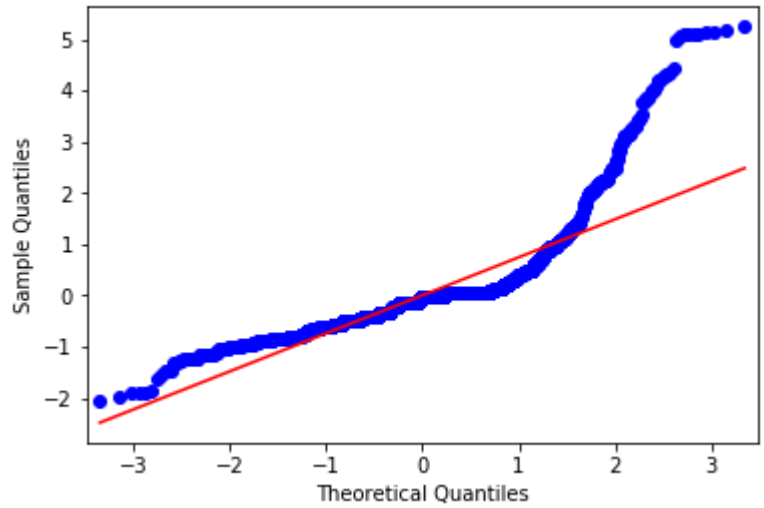
The second multiple regression analysis examines the association between the quantitative response variable, number of cannabis dependence symptoms (**CanDepSymptoms**) and the centered quantitative explanatory variable, number of joints smoked per day when using the most (**numberjosmoked_c**). A second order polynomial of number of joints variable (**'numberjosmoked_c^2**) was added to the regression equation in order to improve the fit of the model and capture the curve of linear relationship that was evident in the scatter plot. In addition, the recoded variable (**CUFREQ**) which represents the frequency of cannabis use (values 1-10, 1.="Once a year", 10.="Every day"), was included to the model as a potential confounding factor. There is also a show that coefficients for the linear, and quadratic variables, remain significant after adjusting for frequency of cannabis use rate.

If we look at the results, it is noticeable that the value for the linear term for number of joints is 0.05, and the p value is less than 0.0001. In addition, the quadratic term is negative (-0.0006) and the p value is also significant (0.0001). The R square increases from 0.003 to 0.18., which means that adding the quadratic term for cannabis joints, increase the amount of variability in cannabis dependence symptoms that can be explained by cannabis use quantity from 0.3% to 18%. For the frequency of cannabis use the coefficient is equal to 0.09 and the p-value is significantly small, at 0.0001.



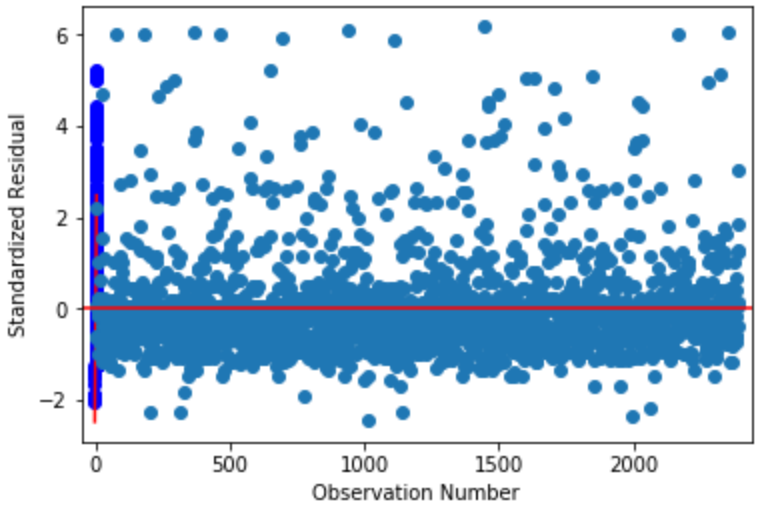
Diagnostic Plots

Q to Q Plot



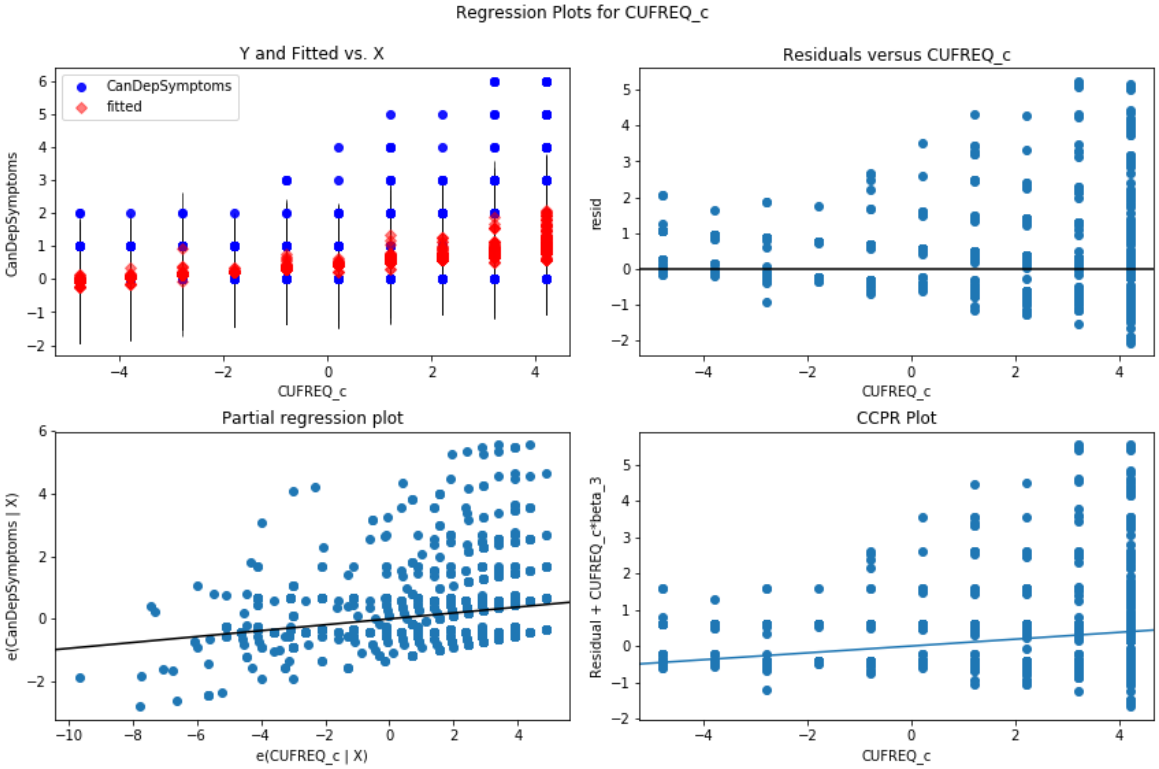
This qq-plot evaluates the assumption that the residuals from our reggression model are normally distributed. A qq-plot, plots the quantiles of the residuals that we would theoretically see if the residuals followed a normal distribution, against the quantiles for the residuals estimated from the reggression model. It is noticeable that the residuals generally deviate from a straight line, especially at higher quantiles. This indicates that our residuals did not follow perfect normal distribution. This could mean that the curvilinear association that we observed in our scatter plot may not be fully estimated by the quadratic cannabis joints variable. Therefore, there might be other explanatory variables that could improve estimation of the observed curvilinearity.

Plot of standardized residuals for all observations



To evaluate the overall fit of the predicted values of the response variable to the observed values and to look for outliers, I created a plot for the standardized residuals of each observation. As we can see, most of the residuals fall between -2 and 2, but many of them fall also above 2. This indicates that we have several outliers, basically above the mean of 0. Furthermore, some of these outliers fall above 4 (extreme outliers) which suggests that the fit of the model is relatively poor and could be improved.

Regression plots for frequency of cannabis use

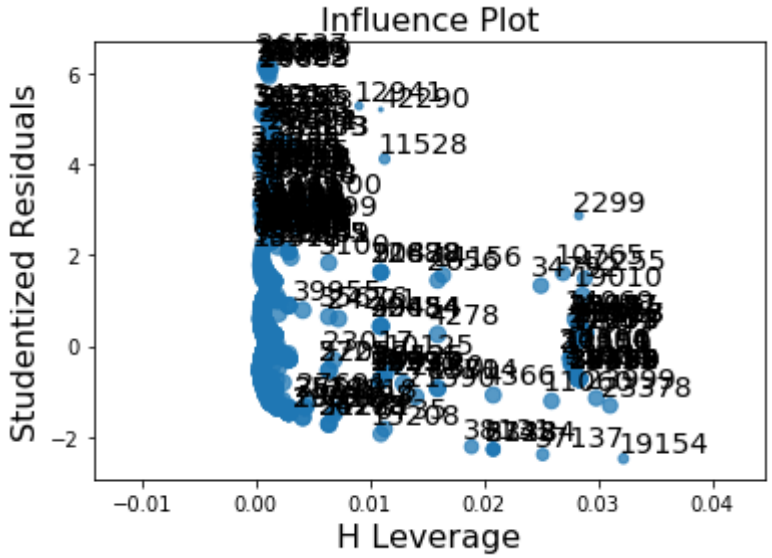


The plot in the upper right hand corner shows the residuals for each observation at different values of cannabis use frequency rate. There is clearly a funnel shaped pattern

to the residuals where we can see that the absolute values of the them are small at lower values of frequency use rate, but then they start to get larger at higher levels. This indicates that this model does not predict cannabis dependence symptoms as well for individuals who have either high or low levels of cannabis use frequency rate. But is particularly worse predicting dependence symptoms for individuals with high frequency of cannabis use.

The partial regression residual plot, in the lower left hand corner, attempts to show the effect of adding cannabis use frequency rate as an additional explanatory variable to the model. For the frequency use rate variable, the values in the scatter plot are two sets of residuals. The residuals from a model predicting the cannabis dependence symptoms response from the other explanatory variables, excluding frequency of use, are plotted on the vertical access, and the residuals from the model predicting frequency of use from all the other explanatory variables are plotted on the horizontal access.What this means is that the partial regression plot shows the relationship between the response variable and specific explanatory variable, after controlling for the other explanatory variables. The residuals are spread out in a random pattern around the partial regression line and many of the residuals are pretty far from this line, indicating a great deal of cannabis dependence symptoms prediction error. Although cannabis use frequency rate shows a statistically significant association with cannabis dependence symptoms, this association is pretty weak after controlling for the number of joints smoked.

Leverage plot



The leverage plot attempts to identify observations that have an unusually large influence on the estimation of the predicted value of the response variable, cannabis dependence symptoms, or that are outliers, or both. We can see in the leverage plot is that we have a several outliers, contents that have residuals greater than 2 or less than -

2. We've already identified some of these outliers in previous plots, but the plot also tells us which outliers have small or close to zero leverage values, meaning that although they are outlying observations, they do not have an undue influence on the estimation of the regression model.