D208 Task 1: Linear Regression Modeling

Part I: Research Question

A1. Research Question

The research question I would like to analyze is, "What variables most significantly contribute to a customer's monthly charge?"

A2. Data Analysis Goals

The objective of this data analysis is to use a predictive model, in this case, a multiple linear regression model, to gain more insight to determine which variables from the churn dataset contribute to the amount a customer is charged monthly. In this case, the variables from the dataset are the independent or explanatory variables and the monthly charge would be the dependent or target variable. Once the outcome has been determined, we should have an idea of the cost the customer could be paying monthly for the provided services. As a result, this could have an effect on customer churn if the customer feels they may be getting overcharged for the services.

Part II: Multiple Linear Regression

B1. Assumptions

Multiple linear regression is a statistical method used to predict the relationship between multiple independent (explanatory) variables and a single dependent (response) variable. Before conducting this method, we have to check for the following assumptions (Bobbitt 2021):

- Linearity: There has to be a linear relationship between the explanatory variables and response variable.
- No Multicollinearity: None of the explanatory variables are correlated with each other.
- Independence: The observations are independent of each other.
- Homoscedasticity: At every point in the model, residuals should have constant variance.
- Multivariate Normality: All residuals have normal distribution.

The multiple linear regression equation is:
$Y = \beta0+\beta1X1+\beta2X2+\beta3X3+\beta4X4+\cdots+\beta nXn+ \in$
Y = Predicted value of the dependent variable
$\beta0$ = Y-Intercept
$\beta1X1$ = Regression coefficient ($\beta1$) of independent variable (X1)
$\in$ = Model error

B2. Tool Benefits

I will be using Python to perform this data analysis. Python is a great tool to clean data and perform linear regression because of the consistent syntax that makes it easy to learn and follow along, the flexibility to create and learn new things, and all of the libraries and packages that it has to offer. For example, I will be using the following libraries and packages for my analysis (R or Python 2023):

- pandas- to load datasets
- NumPy- to work with arrays
- Sci-kit Learn- for machine learning and to transform our data
- SciPy- for mathematical problems like checking for multicollinearity
- Matplotlib- for basic plotting generally consisting of bars, lines, pies, scatter plots, and graphs
- Seaborn- for a variety of visualization patterns

B3. Justification

Multiple linear regression is the appropriate technique to use to analyze the research question because the target variable, monthly charge, is a continuous dependent variable. The multiple explanatory variables can be continuous and categorical though. However, if the target variable was categorical, then we would have to perform logistic regression. Performing multiple linear regression will help to figure out if the explanatory variables have a positive or negative impact on the chosen target variable. As shown in the multiple linear regression equation, each independent variable has a coefficient that it is multiplied with. This predictive model will indicate how strong the relationships between the X variables and Y variable are.

Part III: Data Preparation & Manipulation

C1. Data Cleaning Goals & Steps

The goal of data cleaning is to find any null or duplicated values in the dataset, correct any error or inconsistencies, and to get rid of any unnecessary variables that we will not be using for the regression analysis. I dropped the following columns since they are not important for the analysis question I have chosen: 'CaseOrder', 'Customer_id', 'Interaction', 'UID', 'City', 'State', 'County', 'Zip', 'Lat', 'Lng', 'Population', 'Area', 'TimeZone', 'Marital', 'Job', 'Techie', 'PaperlessBilling', 'PaymentMethod', 'Item1', 'Item2', 'Item3', 'Item4', 'Item5', 'Item6', 'Item7', and 'Item8'. The reason I dropped these variables is because the customer service interactions, response from the questionnaires, location, and their identification numbers do not impact the amount a customer would be charged monthly. I also do not think customer demographics should affect a customer's monthly charge, but I left age in the analysis just in case there was some type of significance to an age group. I have the general steps I performed in order to clean and prepare the data for testing written below (see code attached as an ipynb file of my Jupyter Notebook):

1. Import any necessary libraries and packages.
2. Load dataset into pandas data frame using read_csv command. The data frame is named "df".
3. Calculate the total null values and total duplicate values in the dataset. If there are not any, the values will be shown as 0.
4. Check for the number of unique values in each column.
5. Print the columns with less than 100 unique values. This can help determine what variables I would like to drop from the analysis.
6. Drop columns that are unnecessary for the analysis.
7. Use the head() command to look at what data is left.

C2. Summary Statistics

The variables I will be using for the analysis include MonthlyCharge (the continuous dependent variable) and the following explanatory variables are listed below:

- Categorical variables – Gender, Churn, Contract, Port_modem, Tablet, InternetService, Phone, Multiple, OnlineSecurity, OnlineBackup, DeviceProtection, TechSupport, StreamingTV, StreamingMovies

- Numerical variables – Children, Age, Income, Outage_sec_perweek, Email, Contacts, Yearly_equip_failure, Tenure, Bandwidth_GB_Year

C3. Univariate & Bivariate Visualizations

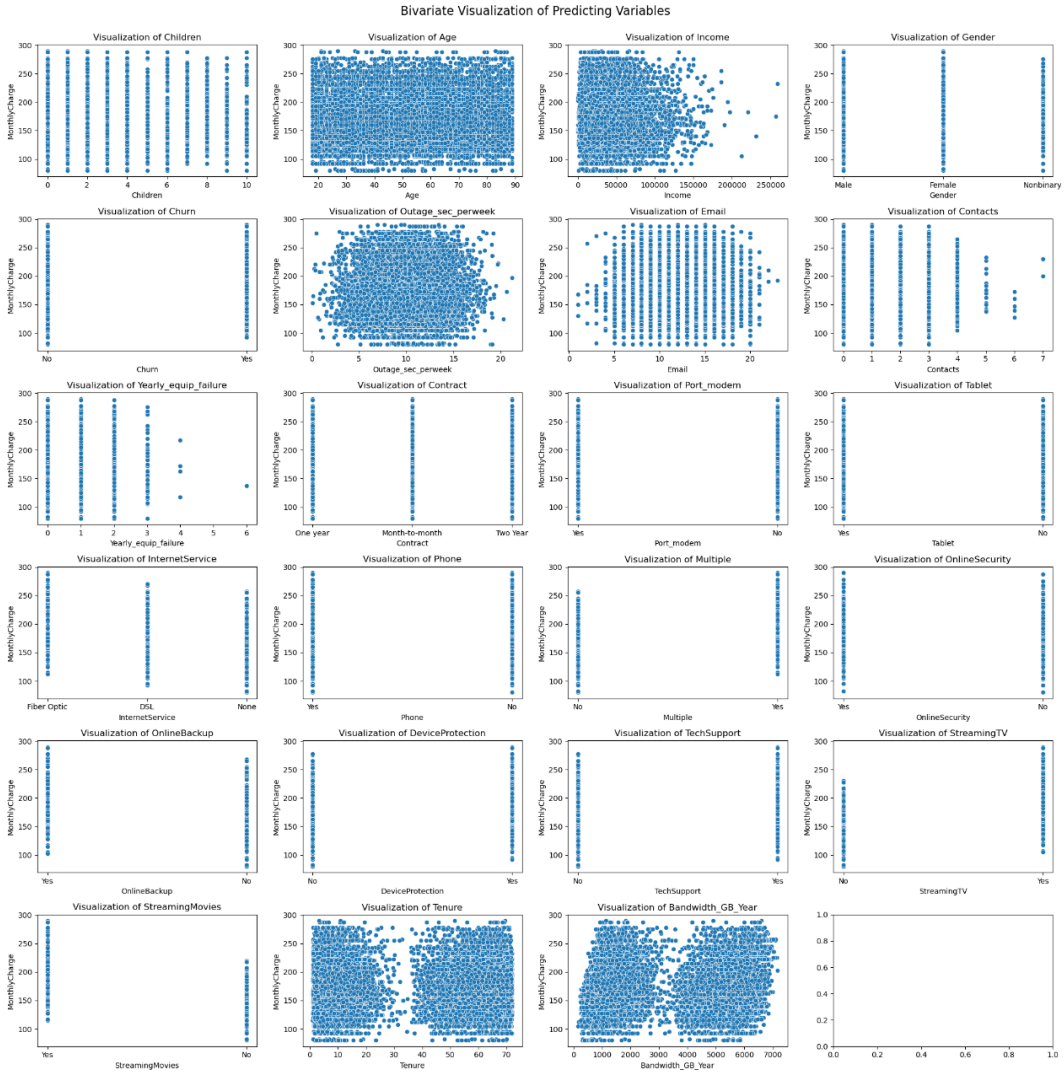
Univariate Visualizations: I have provided a screenshot of the univariate visualizations of all 23 independent (predicting) variables and the 1 dependent (target) variable, MonthlyCharge.

```
# Find the summary statistics for numerical variables
df.describe()
```

	Children	Age	Income	Outage_sec_perweek	Email	Contacts	Yearly_equip_failure	Tenure	MonthlyCharge	Bandwidth_GB_Year
count	10000.0000	10000.000000	10000.000000	10000.000000	10000.000000	10000.000000	10000.000000	10000.000000	10000.000000	10000.000000
mean	2.0877	53.078400	39806.926771	10.001848	12.016000	0.994200	0.398000	34.526188	172.624816	3392.341550
std	2.1472	20.698882	28199.916702	2.976019	3.025898	0.988466	0.635953	26.443063	42.943094	2185.294852
min	0.0000	18.000000	348.670000	0.099747	1.000000	0.000000	0.000000	1.000259	79.978860	155.506715
25%	0.0000	35.000000	19224.717500	8.018214	10.000000	0.000000	0.000000	7.917694	139.979239	1236.470827
50%	1.0000	53.000000	33170.605000	10.018560	12.000000	1.000000	0.000000	35.430507	167.484700	3279.536903
75%	3.0000	71.000000	53246.170000	11.969485	14.000000	2.000000	1.000000	61.479795	200.734725	5586.141370
max	10.0000	89.000000	258900.700000	21.207230	23.000000	7.000000	6.000000	71.999280	290.160419	7158.981530

```
PROPORTION OF EACH CATEGORICAL VARIABLE
----------------------------------------
       Gender  Proportion
0      Female      0.5025
1        Male      0.4744
2  Nonbinary       0.0231
----------------------------------------
   Churn  Proportion
0     No       0.735
1    Yes       0.265
----------------------------------------
         Contract  Proportion
0  Month-to-month      0.5456
1        Two Year      0.2442
2        One year      0.2102
----------------------------------------
   Port_modem  Proportion
0          No      0.5166
1         Yes      0.4834
----------------------------------------
   Tablet  Proportion
0      No      0.7009
1     Yes      0.2991
----------------------------------------
   InternetService  Proportion
0      Fiber Optic      0.4408
1              DSL      0.3463
2             None      0.2129
----------------------------------------
   Phone  Proportion
0  Yes      0.9067
1   No      0.0933
----------------------------------------
   Multiple  Proportion
0        No      0.5392
1       Yes      0.4608
----------------------------------------
   OnlineSecurity  Proportion
0              No      0.6424
1             Yes      0.3576
----------------------------------------
   OnlineBackup  Proportion
0            No      0.5494
1           Yes      0.4506
----------------------------------------
   DeviceProtection  Proportion
0                No      0.5614
1               Yes      0.4386
----------------------------------------
   TechSupport  Proportion
0           No       0.625
1          Yes       0.375
----------------------------------------
   StreamingTV  Proportion
0           No      0.5071
1          Yes      0.4929
----------------------------------------
   StreamingMovies  Proportion
0               No       0.511
1              Yes       0.489
```

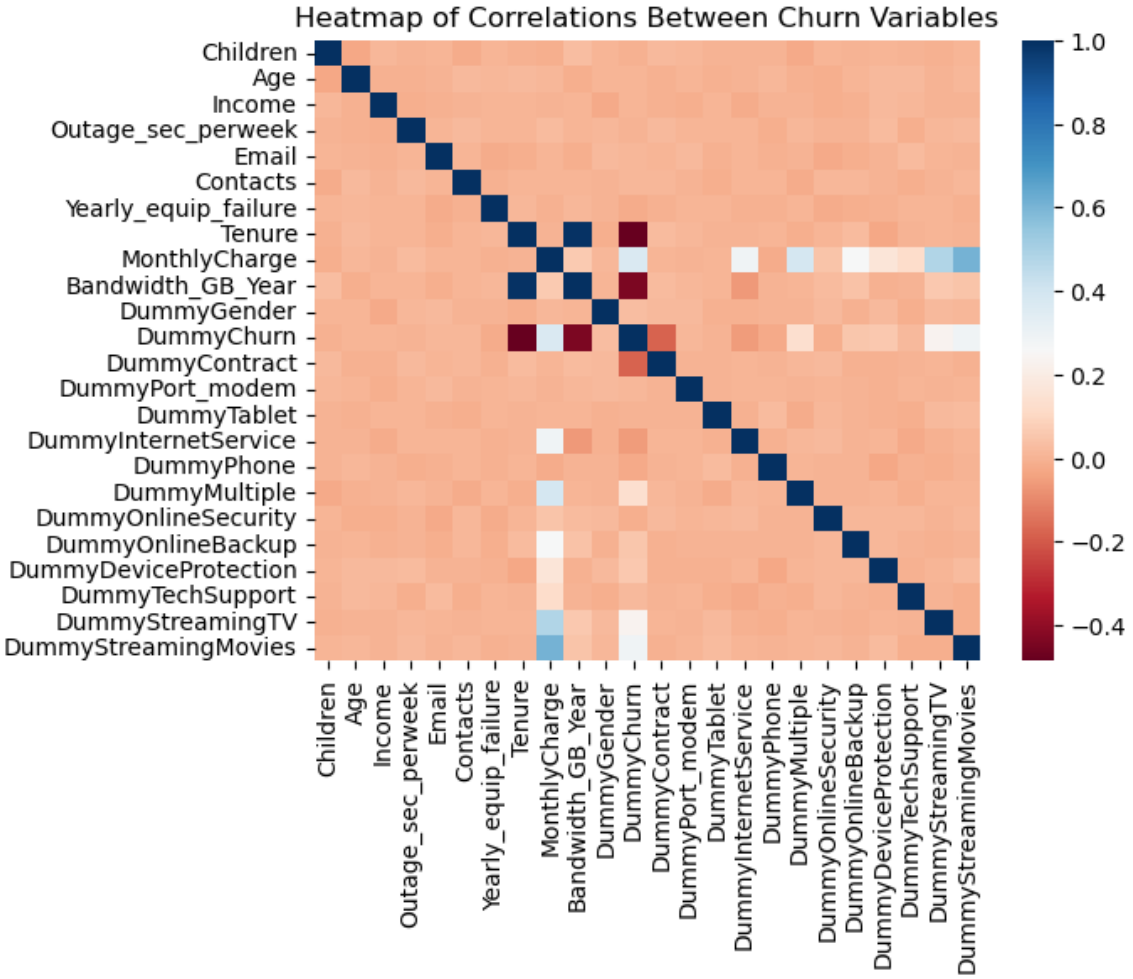
Bivariate Visualizations: I have provided a screenshot of the bivariate visualizations of all 23 independent (predicting) variables compared to the target variable, MonthlyCharge. I chose to do scatterplots to get an idea of the linearity between each predicting variable and the target variable. Based on the scatterplots, there are no linear relationships which means we technically have not met the assumption of linearity.



C4. <u>Data Transformation Goals & Steps</u>

Since my analysis does contain categorical variables, I have to re-express these variables as "dummy variables". Dummy variables represent categorical variables as values of 0 or 1. For example, churn is categorized into 2 groups (yes or no). Therefore, 1 would represent "yes" and 0 would represent "no". Another example would be gender. In this analysis, gender is categorized into 3 groups (female, male, or nonbinary). For linear regression, we would have to do one-hot encoding. One-hot encoding follows the k-1 rule, so 3-1=2 groups. As a result, 1 would represent "male" and 0 would represent everything else; in this case, "female" and "nonbinary". If we were to add an extra dummy, this would cause a multicollinearity problem. The data transformation steps can be found below:

1. Reformat the columns to have 3 decimal places
2. Create dummy variables where "Yes" is represented by 1 & "No" is represented by 0.
3. Drop the original categorical columns so only dummy categorical columns are left.
4. Check the new data frame to make sure all the data transferred correctly.
5. Check for multicollinearity by calculating the correlation matrix (corr() command).
6. Plot the correlation heatmap.
   – The heatmap helps to easily identify if there are any correlations between any of the independent variables by simply looking for the dark blue color. Based on the correlation matrix and heatmap, Tenure and Bandwidth_GB_Year are highly correlated with a value of 0.991495. This means these two variables are at risk of being dropped.



7. Extract the cleaned & wrangled dataset.

C5. Prepared Data Set

I will provide an attached copy of the prepared data set, named “df_clean.csv”.

Part IV: Model Comparison & Analysis

D1. Initial Multiple Linear Regression

Analysis of initial model summary:
In my Jupyter Notebook, I ran a model of all of the independent variables identified in part C5. I constructed this model by using the statsmodel library with the OLS() function and printed the initial model summary using the summary() function.
The initial model has a R^2 value of 0.967 and an adjusted R^2 value of 0.966, both being close to 1. If the value were 1, that would indicate a perfect fit of the model to the data.
According to the Notes section of the initial model summary, the smallest eigenvalue is 4.08e+05 which might indicate that there are strong multicollinearity problems. In part C4, I tested for the assumption of multicollinearity. I came to the conclusion that both Tenure and Bandwidth_GB_Year are highly correlated according to the heatmap.

OLS Regression Results

Dep. Variable:	MonthlyCharge	R-squared:	0.967
Model:	OLS	Adj. R-squared:	0.966
Method:	Least Squares	F-statistic:	1.254e+04
Date:	Mon, 24 Jun 2024	Prob (F-statistic):	0.00
Time:	21:37:13	Log-Likelihood:	-34795.
No. Observations:	10000	AIC:	6.964e+04
Df Residuals:	9976	BIC:	6.981e+04
Df Model:	23		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Children	-1.1881	0.040	-29.779	0.000	-1.266	-1.110
Age	0.1307	0.004	31.320	0.000	0.123	0.139
Income	2.424e-06	2.79e-06	0.869	0.385	-3.05e-06	7.89e-06
DummyGender	-2.8175	0.161	-17.458	0.000	-3.134	-2.501
DummyChurn	2.4748	0.237	10.431	0.000	2.010	2.940
Outage_sec_perweek	-0.0038	0.026	-0.145	0.884	-0.056	0.048
Email	-0.0039	0.026	-0.150	0.881	-0.055	0.047
Contacts	-0.0660	0.080	-0.828	0.407	-0.222	0.090
Yearly_equip_failure	-0.0998	0.124	-0.807	0.420	-0.342	0.143
DummyContract	0.4934	0.188	2.629	0.009	0.125	0.861
DummyPort_modem	-0.2075	0.157	-1.319	0.187	-0.516	0.101
DummyTablet	-0.1224	0.172	-0.712	0.476	-0.459	0.215
DummyInternetService	34.8883	0.206	169.097	0.000	34.484	35.293
DummyPhone	-0.3560	0.271	-1.315	0.188	-0.887	0.175
DummyMultiple	29.4535	0.164	179.178	0.000	29.131	29.776
DummyOnlineSecurity	-0.2866	0.169	-1.691	0.091	-0.619	0.046
DummyOnlineBackup	18.7519	0.166	113.292	0.000	18.427	19.076
DummyDeviceProtection	9.1302	0.164	55.548	0.000	8.808	9.452
DummyTechSupport	12.3016	0.163	75.620	0.000	11.983	12.620
DummyStreamingTV	32.7764	0.199	164.470	0.000	32.386	33.167
DummyStreamingMovies	43.4846	0.197	220.791	0.000	43.099	43.871
Tenure	-3.1807	0.043	-74.121	0.000	-3.265	-3.097
Bandwidth_GB_Year	0.0390	0.001	75.272	0.000	0.038	0.040
const	63.3174	0.640	98.929	0.000	62.063	64.572

Omnibus:	38941.451	Durbin-Watson:	1.998
Prob(Omnibus):	0.000	Jarque-Bera (JB):	1374.539
Skew:	0.024	Prob(JB):	3.33e-299
Kurtosis:	1.184	Cond. No.	4.08e+05

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 4.08e+05. This might indicate that there are
strong multicollinearity or other numerical problems.

D2. Justification of Model Reduction

To reduce the model, I first had to find the VIF of all of the independent variables to see which variables should be eliminated due to high multicollinearity. After I found the VIF, both Bandwidth_GB_Year and Email were removed one by one since they had VIF values greater than 10. According to the Notes section, the multicollinearity was taken care of.
After that, I performed backward stepwise elimination. This is an important reduction method because it not only reduces the number of predictors, but also helps to resolve overfitting and remove the predictors that do not significantly affect the target variable. For this reduction method, I had to remove the least significant features based on their p-values one at a time. Therefore, if a variable had a p-value greater than 0.05, it was removed.
The following variables are what is left in the reduced linear regression model after the removal due to their p-values and insignificance to the analysis:
'Churn', 'Contract', 'InternetService', 'Multiple', 'OnlineSecurity', 'OnlineBackup', 'DeviceProtection', 'TechSupport', 'StreamingTV', 'StreamingMovies', and 'Tenure'

D3. Reduced Linear Regression Model

OLS Regression Results

Dep. Variable:	MonthlyCharge	R-squared:	0.948
Model:	OLS	Adj. R-squared:	0.947
Method:	Least Squares	F-statistic:	1.640e+04
Date:	Mon, 24 Jun 2024	Prob (F-statistic):	0.00
Time:	21:45:57	Log-Likelihood:	16431.
No. Observations:	10000	AIC:	-3.284e+04
Df Residuals:	9988	BIC:	-3.275e+04
Df Model:	11		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
DummyChurn	0.0221	0.001	15.757	0.000	0.019	0.025
DummyContract	0.0047	0.001	4.211	0.000	0.003	0.007
DummyInternetService	0.1189	0.001	125.601	0.000	0.117	0.121
DummyMultiple	0.1536	0.001	161.200	0.000	0.152	0.155
DummyOnlineSecurity	0.0136	0.001	13.887	0.000	0.012	0.015
DummyOnlineBackup	0.1062	0.001	112.446	0.000	0.104	0.108
DummyDeviceProtection	0.0586	0.001	61.972	0.000	0.057	0.060
DummyTechSupport	0.0597	0.001	61.667	0.000	0.058	0.062
DummyStreamingTV	0.1961	0.001	200.012	0.000	0.194	0.198
DummyStreamingMovies	0.2434	0.001	242.745	0.000	0.241	0.245
Tenure	0.0130	0.001	8.734	0.000	0.010	0.016
const	-0.0120	0.001	-8.132	0.000	-0.015	-0.009

Omnibus:	13219.468	Durbin-Watson:	2.007
Prob(Omnibus):	0.000	Jarque-Bera (JB):	646.060
Skew:	-0.055	Prob(JB):	5.13e-141
Kurtosis:	1.760	Cond. No.	7.63

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.

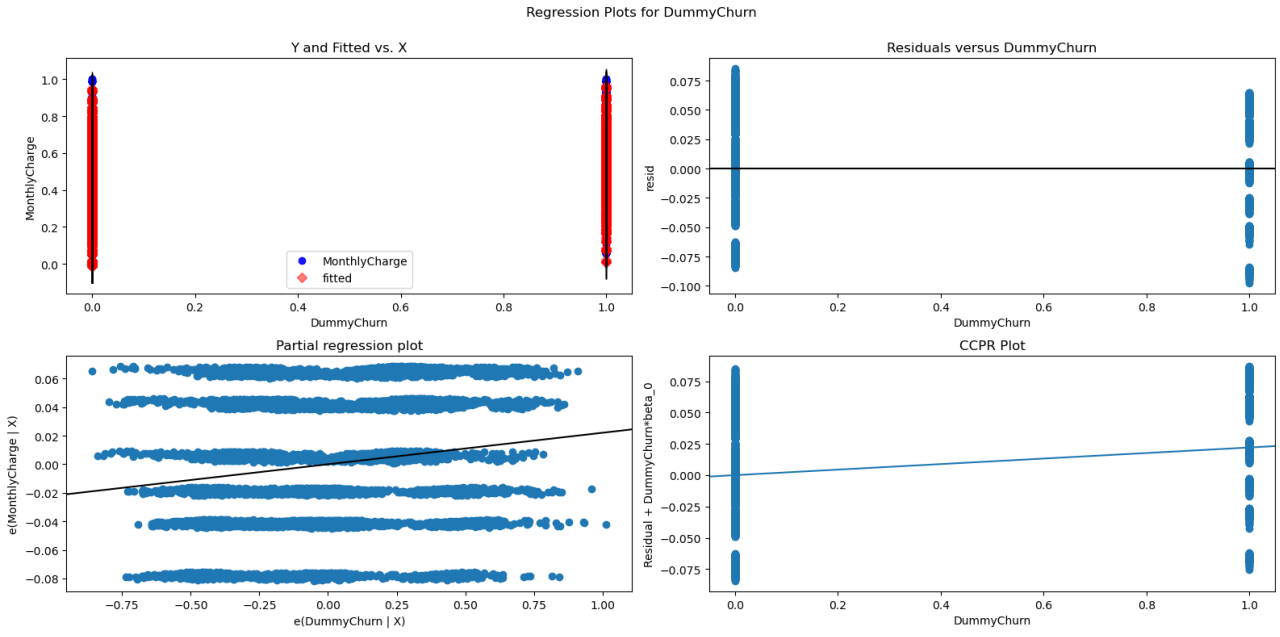
E1. Model Comparison (using a model evaluation metric)

Model Evaluation Metrics	Initial Model	Reduced Model
R-squared	0.967	0.948
Adjusted R-squared	0.966	0.947
Prob (F-statistic)	0.00	0.00
Residual Standard Error	7.86	0.05
Number of Independent Variables	23	11

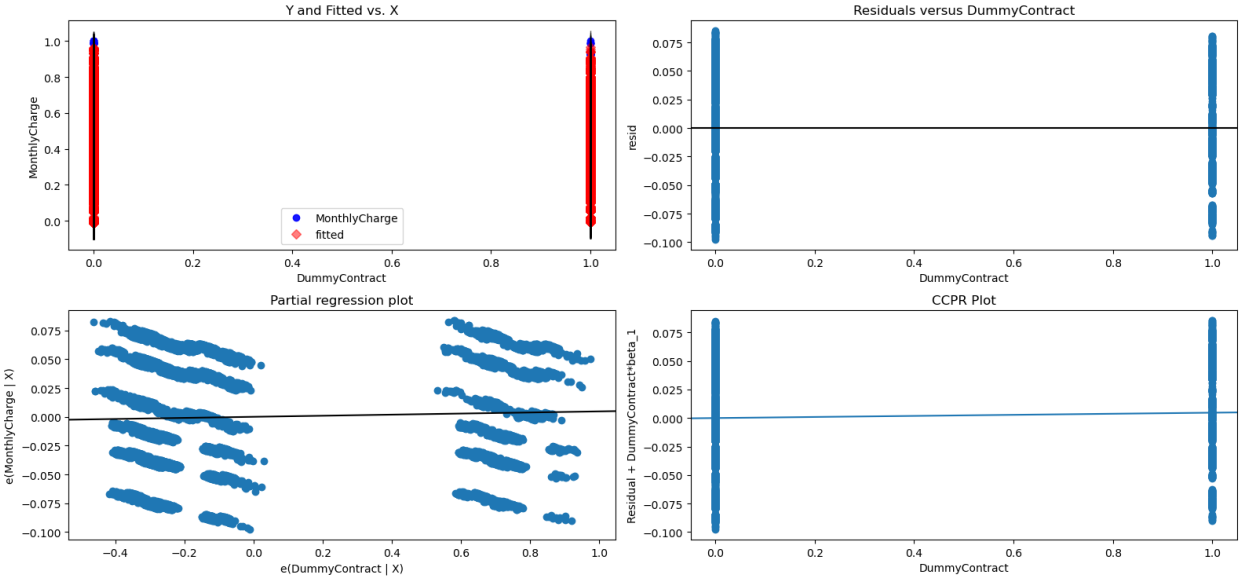
As stated in part D2, I found the variance inflation factors of all of the independent variables to determine which variables had the highest VIF that was causing multicollinearity. After removing one of the expected variables, it affected another variable's VIF, so it had to be removed from the analysis too. Then I performed backward stepwise elimination to remove several more variables with p-values greater than 0.05. All in all, these reduction methods reduced the number of independent variables from 23 to 11. Removing about half of these variables caused the R-squared value to decrease, but it is still fairly close to 1. The adjusted R-squared also decreased which may imply that some of the explanatory variables perhaps should not have been removed. The probability of the F-statistics tells us the significance of the regression. Between the initial and reduced model, this value stayed at 0, implying that the regressions are meaningful. Lastly the residual standard error decreased significantly from 7.86 to 0.05. The closer this value is to 0, the better fitting the model is.

E2. Output & Calculations

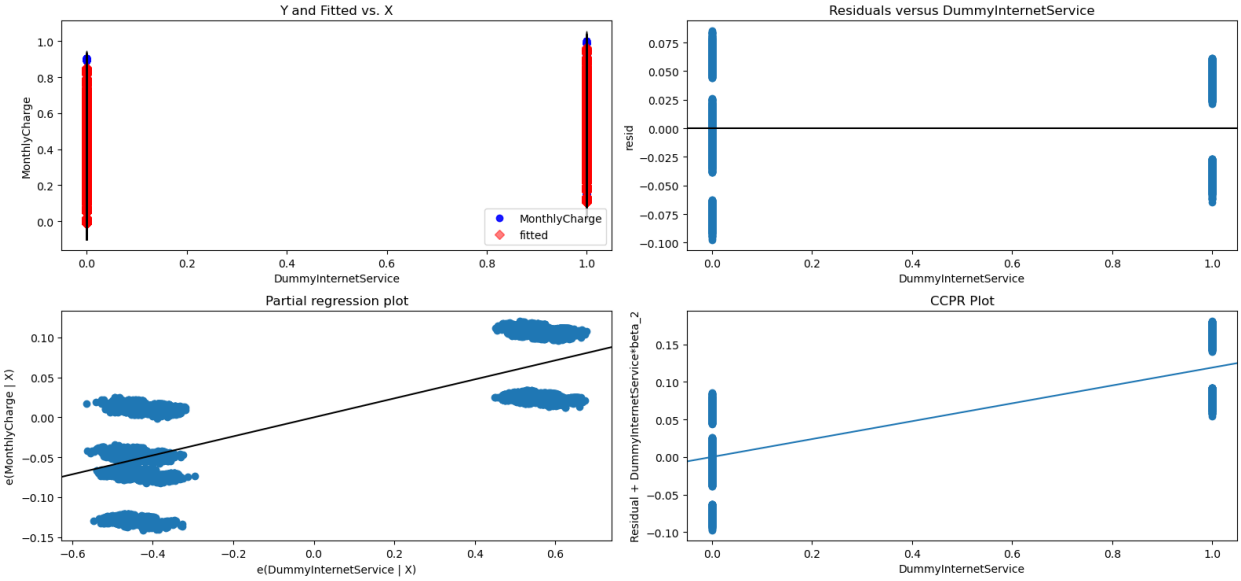
Regression plots for residuals:



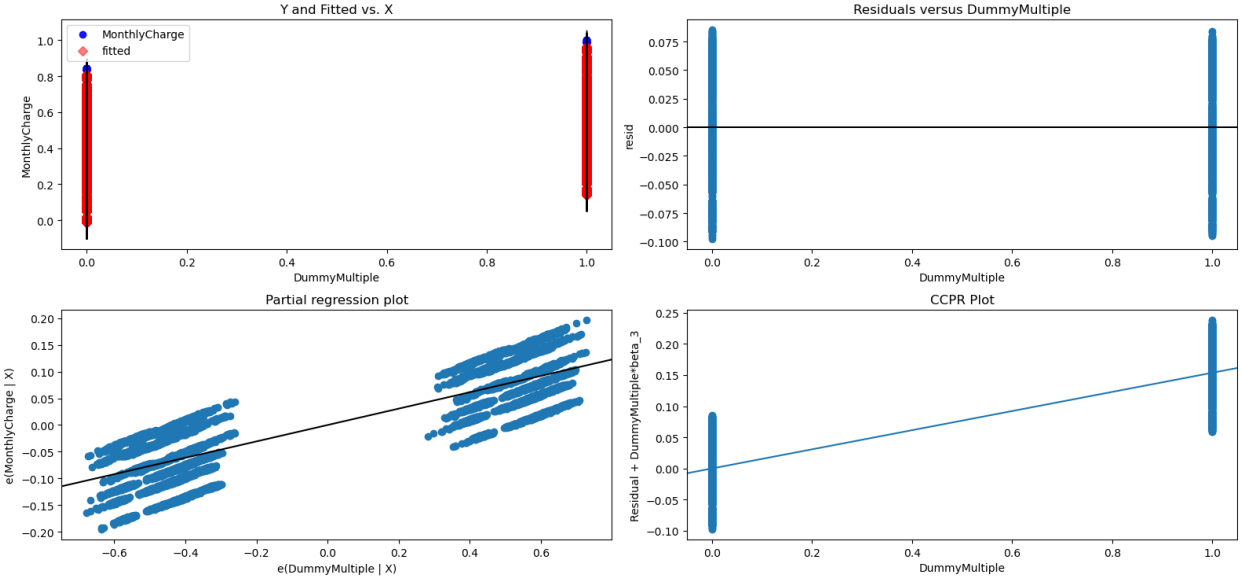
Regression Plots for DummyContract



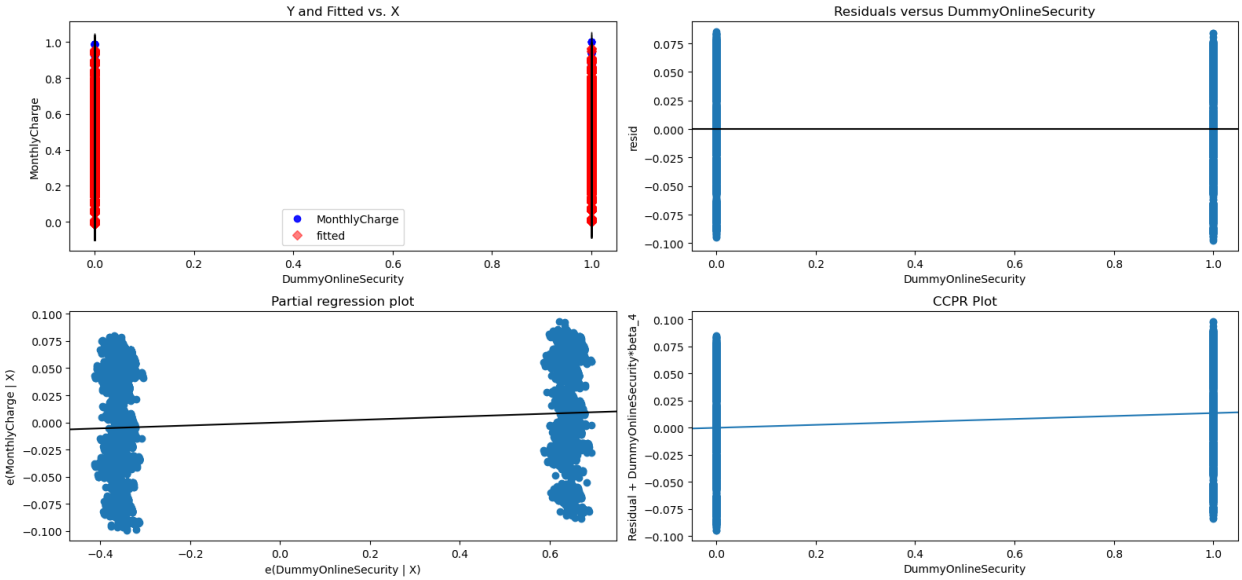
Regression Plots for DummyInternetService



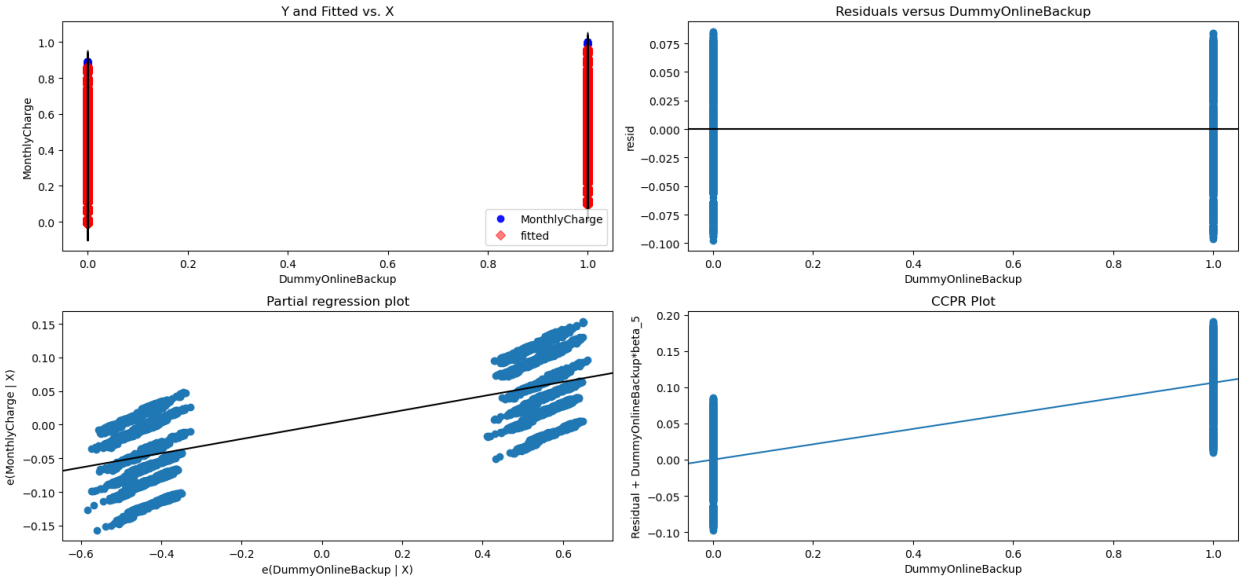
Regression Plots for DummyMultiple



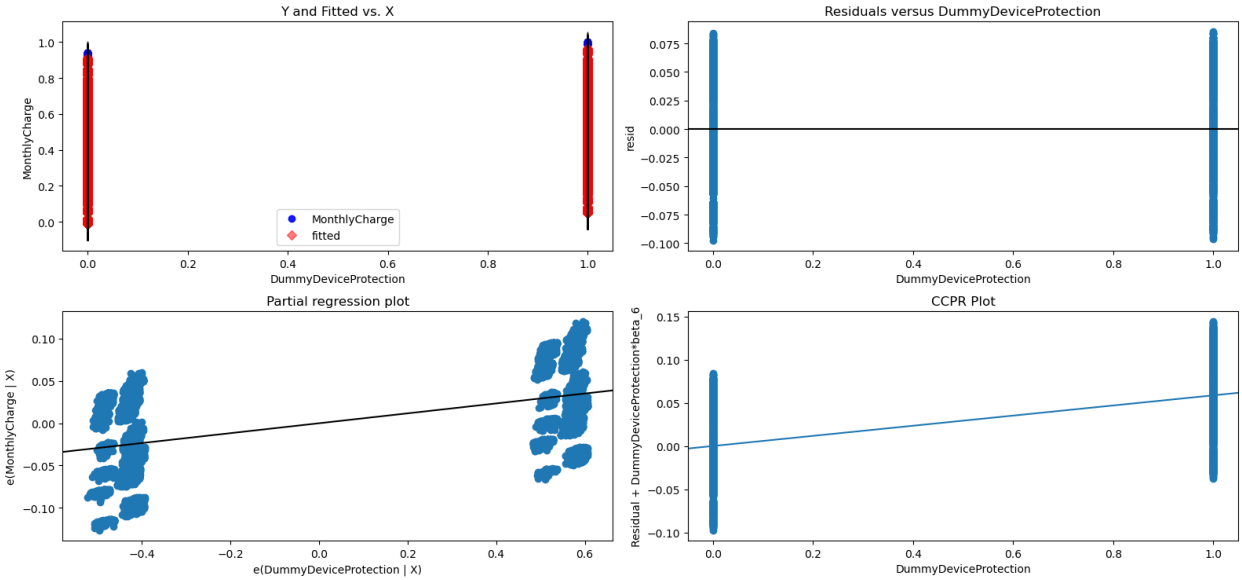
Regression Plots for DummyOnlineSecurity



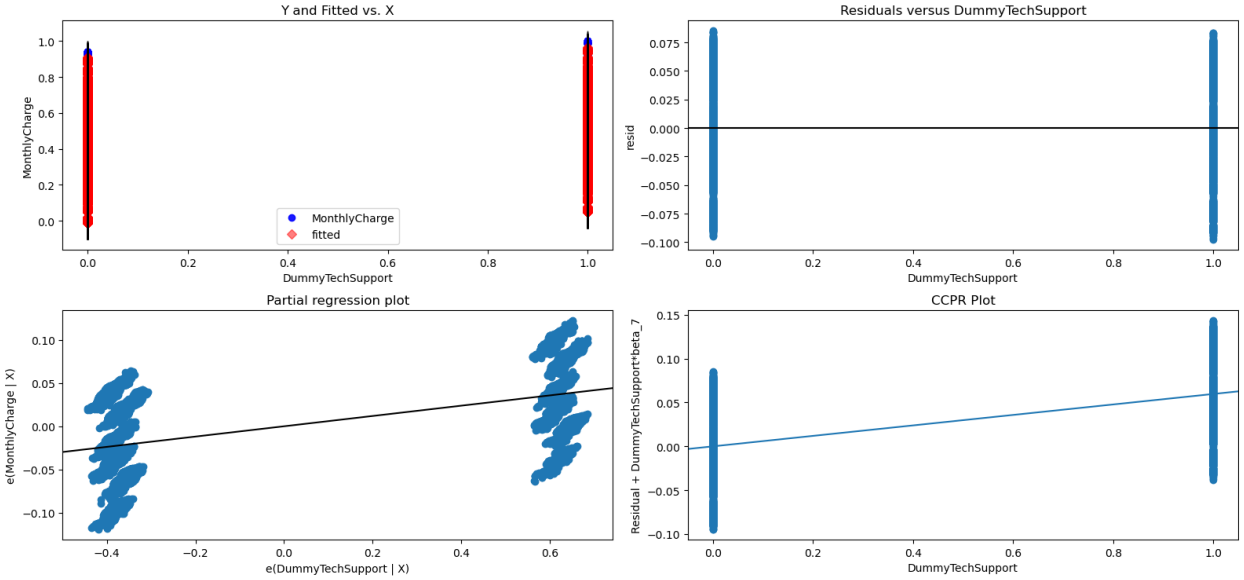
Regression Plots for DummyOnlineBackup



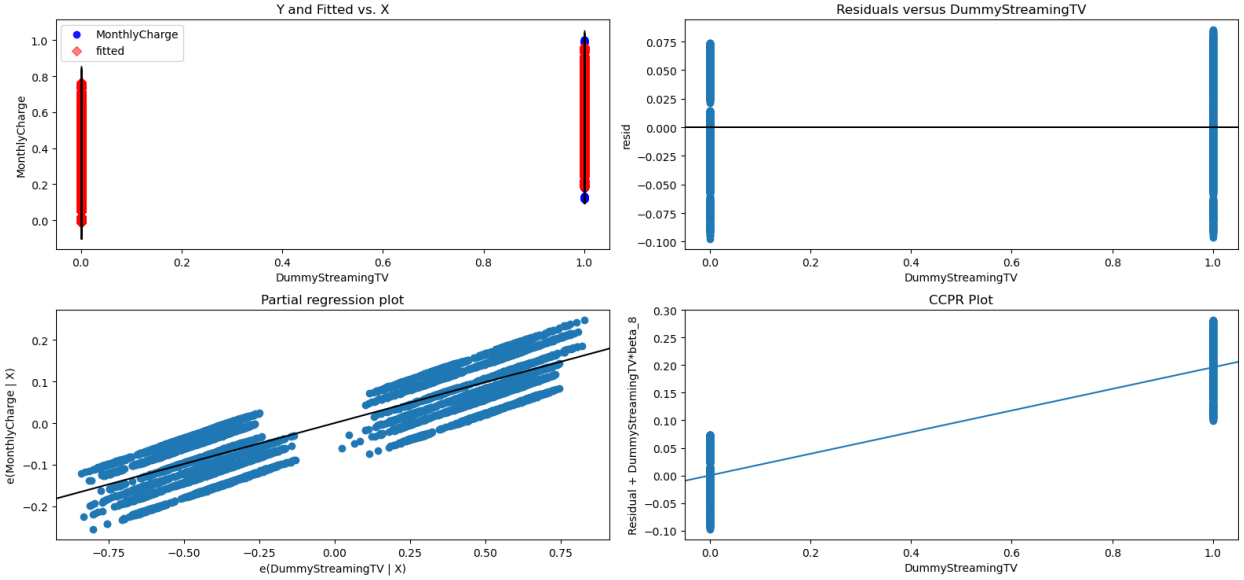
Regression Plots for DummyDeviceProtection



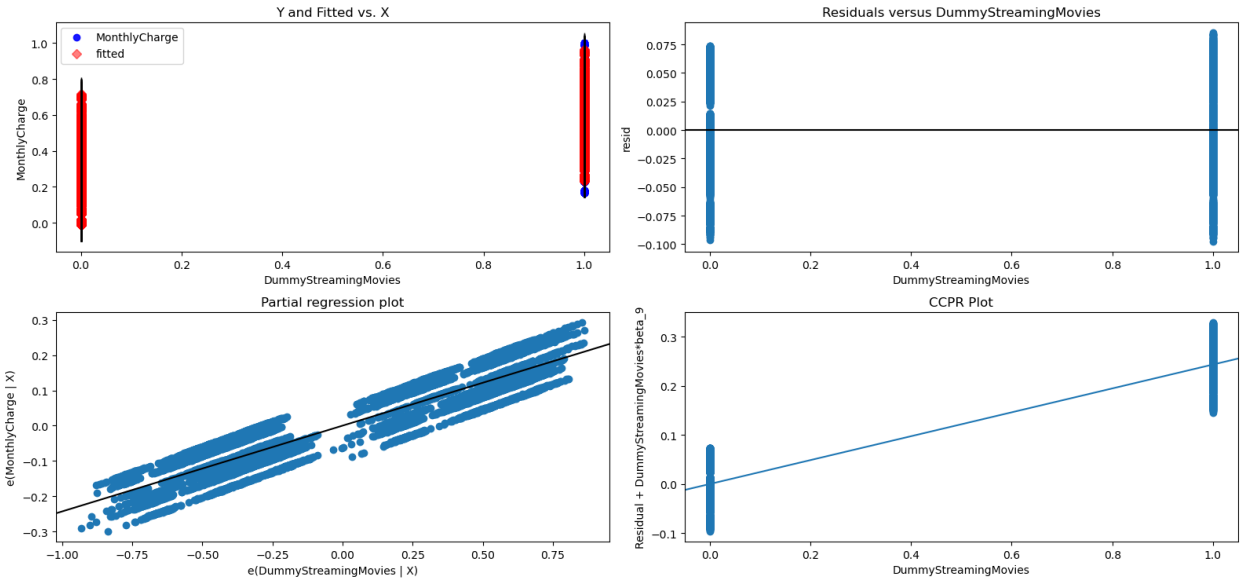
Regression Plots for DummyTechSupport

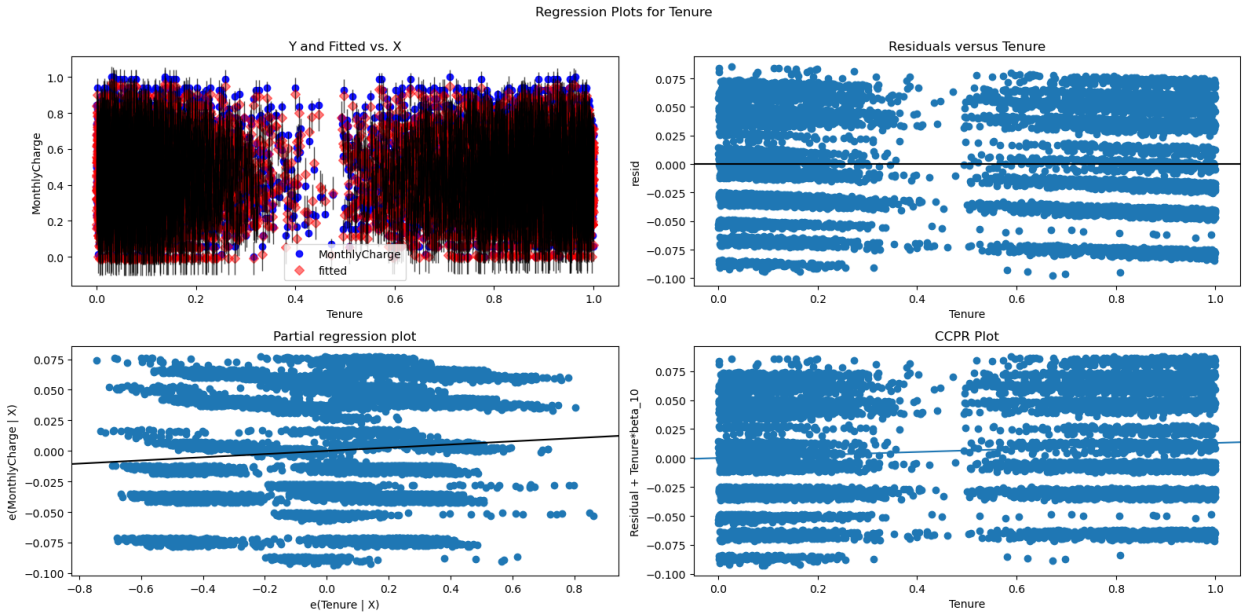


Regression Plots for DummyStreamingTV



Regression Plots for DummyStreamingMovies





Reduced model's residual standard error:

```
# Calculate the residual standard error of the final reduced model
results.resid.std(ddof=X.shape[1])
```

```
0.0468201691404279
```

E3. Copy of Code

Provided as the file "D208 Task 1.ipynb".

Part V: Data Summary & Implications

F1. Data Analysis Results

The regression equation for the reduced model:
y = -0.012 + 0.0221(Churn) + 0.0047(Contract) + 0.1189(InternetService) + 0.1536(Multiple) + 0.0136(OnlineSecurity) + 0.1062(OnlineBackup) + 0.0586(DeviceProtection) + 0.0597(TechSupport) + 0.1961(StreamingTV) + 0.2434(StreamingMovies)

An interpretation of the coefficients of the reduced model:

- y represents the monthly charges a customer has to pay.
- Keeping all things constant, customer churn increases monthly charges by 2.21%.
- Keeping all things constant, customers with a two-year contract pay 0.47% more monthly than the other contract options.
- Keeping all things constant, customers with fiber optic internet service pay 11.89% more monthly than other internet service options.
- Keeping all things constant, customers with multiple services pay 15.36% more monthly than customers with a single service.

- Keeping all things constant, customers with online security pay 1.36% more monthly than customers without online security.
- Keeping all things constant, customers with online backup pay 10.62% more monthly than customers without online backup.
- Keeping all things constant, customers with device protection pay 5.86% more monthly than customers without device protection.
- Keeping all things constant, customers who need technical support pay 5.97% more monthly than customers who did not need technical support.
- Keeping all things constant, customers who have streaming TV pay 19.61% more monthly than customers who do not have streaming TV.
- Keeping all things constant, customers who have streaming movies pay 24.34% more monthly than customers who do not have streaming movies.
- Keeping all things constant, one unit increase in tenure is associated with a 1.3% increase in monthly charges.

Statistical and practical significance of the reduced model:
The reduced model and results are statistically significant according to not only the probability of the f-statistic being 0, but also the residual standard error being 0.05. Both of these values being so close to 0 means that the reduced model is a good fitting model. This also means that there is much confidence that the results are not based on luck. However, I do not think the reduced model is practically significant. Based on the interpretation of the coefficients, some of the results don't make sense. For example, why would a customer paying for the longer two-year contract option be paying 0.47% more than a customer paying month to month or for a one-year contract? Normally, in the real world, if a customer selects to opt in for a longer contract, the company will give you a slight discount for paying the price in full. Another example would be tenure. The longer a customer has been with a company, in some cases, the lower they could get charged due to loyalty to the company. Otherwise, tenure should not affect how much a customer is charged. However, according to the model, tenure is associated with a 1.3% increase in monthly charges. Therefore, the reduced model is not practically significant.

Disadvantages of the data analysis:
During the data reduction process, I feel as if some of the variables that were dropped from the analysis due to their p-values or variance inflation factors, would have been important to use for the analysis in the real world. For example, whether a customer has a phone service or not was insignificant to the analysis because it had a p-value greater than 0.05. However, all of the other services that a customer could sign up for was included in the analysis.
Another disadvantage of using the backward stepwise elimination as the reduction method is the more explanatory variables used, the less effective this method is. But even then, the more explanatory variables used, the longer the process takes because you have to remove each variable one by one after running the model every time.
As mentioned above, the reduced regression model seems to be a good fitting model according to the model evaluation metrics compared to the initial model. But realistically, it is not practical.

F2. Recommendations

Regarding the practical significance of the model, I do believe it is an impractical model based on some of the results from the model summary and some of the explanatory variables that were removed due to high p-values or VIFs. However, as mentioned above, it is statistically significant. This model is a decent starting point and most of the data in the summary results makes sense regarding the services the customers signed up for and their add-ons. It makes sense for the security add-ons to be decently more costly than without the add-ons. It also makes sense for customers with streaming TV and movies to be significantly higher percentages than the other variables. But some of the data could be throwing off some of the values. For example, the variable "Bandwidth_GB_Year" is the average amount of data used by a customer. However, customers who are newer than a year have an estimated number based on other customers in their demographic profile. I recommend not including these customers in future analyses. This analysis only includes 10,000 customer records which I am assuming is just a small portion of a larger set. That would mean there are other customers' records that can be factored into analyses rather than these newer customers with an "estimated number". The same thing goes for "Contacts", "MonthlyCharge", and "Yearly_equip_failure", all of which are estimated values if the customer is new. More efficient data means more efficient results.

Part VI: Demonstration

G. Panopto Video

H. Third-Party Code Sources

N/A

I. Sources

Sharma, A. (2021, May 2). Predictive analysis using multiple linear regression. Medium. https://medium.com/data-science-on-customer-churn-data/predictive-analysis-using-multiple-linear-regression-b6b3b79b36b6

Bobbitt, Z. (2021, November 16). The five assumptions of multiple linear regression. Statology. https://www.statology.org/multiple-linear-regression-assumptions/

R or python. Western Governors University. (2023, July 7). https://www.wgu.edu/online-it-degrees/programming-languages/r-or-python.html

J. Professional Communication

Demonstrate professional communication in the content and presentation of your submission.