625.661 - Final Project

Eric Niblock

May 6th, 2022

1 A Quantitative Approach

The landscape of sports has quickly changed from being driven by expert tacticians and play-makers, to statisticians and data professionals. In the age of big data, football is no exception. By focusing on data collected across a season of play, coaches and other decision makers can benefit from the analysis this data drives. Winning games in football is strongly tied with the profitability of the team - this analysis attempts to determine which factors are most relevant in driving team success, and therefore profit.

2 The Data

A small dataset of containing 28 observations (one for each American football team) was used in this analysis. As such, the report represents a proof-of-concept. Should the results prove valuable to management, more data could be obtained to expand the capability and accuracy of the model.

This model uses the number of wins by team during the 1976 season as the dependent variable, y. As for the regressors, the following metrics were assigned to the corresponding potential independent variables:

Regressor	Description
x_1	Rushing yards (season)
x_2	Passing yards (season)
x_3	Punting average (yards/punt)
x_4	Field goal percentage (FGs made/FGs attempted)
x_5	Turnover differential (turnovers acquired – turnovers lost)
x_6	Penalty yards (season)
x_7	Percent rushing (rushing plays/total plays)
x_8	Opponents' rushing yards (season)
x_9	Opponents' passing yards (season)

In attempting to find the best subset of regressors in relation to our predictive model, we will simultaneously find the regressors which most heavily influence a team's ability to win, and therefore produce profit.

3 Model Building

Given that there are nine regressors, testing all-possible-regressions is feasible, though costly. In total, we would evaluate 512 distinct models. Therefore, backwards elimination was chosen in order to build our model. The determination was made that a value of $p_{out} = 0.05$ would be used, that is, at every round the regressor with the highest p-value will be eliminated from the model, until all p-values fall below 0.05.

The following shows the value of the coefficients after full model construction,

	coef	std err	t	P>|t|	[0.025	0.975]
const	-7.2919	12.813	-0.569	0.576	-34.211	19.627
x1	0.0008	0.002	0.405	0.690	-0.003	0.005
x2	0.0036	0.001	4.318	0.000	0.002	0.005
x3	0.1222	0.259	0.472	0.643	-0.422	0.666
x4	0.0319	0.042	0.767	0.453	-0.056	0.119
x5	1.511e-05	0.047	0.000	1.000	-0.098	0.098
x6	0.0016	0.003	0.490	0.630	-0.005	0.008
x7	0.1544	0.152	1.015	0.324	-0.165	0.474
x8	-0.0039	0.002	-1.898	0.074	-0.008	0.000
x9	-0.0018	0.001	-1.264	0.222	-0.005	0.001

Turnover differential, x_5, is clearly not statistically significant within the model and has the highest p-value of 1.000. It therefore is the first regressor to be removed from the model. It should be noted that this practice is is not completely theoretically sound - given that no residual analysis has been done, it is unclear whether the statistical information presented above is accurate (because it relies upon the normality assumption). In practice, however, it is reasonable to remove coefficients in this fashion in order to achieve a more manageable model, and then deal with the normality assumption when the model becomes more feasible.

The next model is then given by,

	coef	std err	t	P>|t|	[0.025	0.975]
const	-7.2937	11.336	-0.643	0.528	-31.020	16.433
x1	0.0008	0.002	0.417	0.681	-0.003	0.005
x2	0.0036	0.001	4.594	0.000	0.002	0.005
x3	0.1222	0.251	0.486	0.632	-0.404	0.648
x4	0.0319	0.040	0.804	0.431	-0.051	0.115
x6	0.0016	0.003	0.508	0.618	-0.005	0.008
x7	0.1544	0.140	1.103	0.284	-0.139	0.447
x8	-0.0039	0.002	-1.952	0.066	-0.008	0.000
x9	-0.0018	0.001	-1.299	0.210	-0.005	0.001

Without the use of x_5, the value of the other coefficients changes, and now x_1, rushing yards, becomes the least significant, with a p-value of 0.681. Continual model construction and regressor removal occurred until we reached the following model,

	coef	std err	t	P>|t|	[0.025	0.975]
const	-1.8084	7.901	-0.229	0.821	-18.115	14.498
x2	0.0036	0.001	5.177	0.000	0.002	0.005
x7	0.1940	0.088	2.198	0.038	0.012	0.376
x8	-0.0048	0.001	-3.771	0.001	-0.007	-0.002

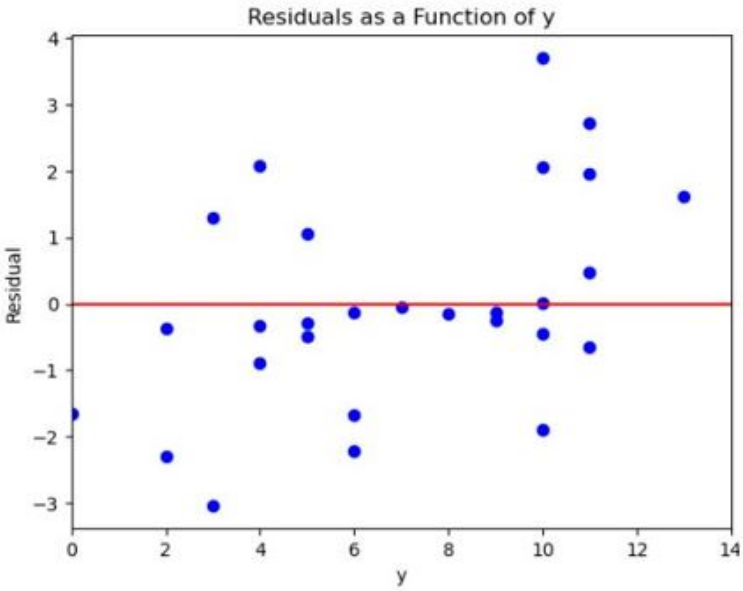
At this point, backwards elimination was halted because all three of the coefficients had p-values that were less than $p_{out} = 0.05$. Among these three regressors, interaction effects were tested by including x_2x_7, x_2x_8, x_7x_8, and $x_2x_7x_8$, and then performing another pass-through of backward elimination. The results are included in the code appendix, though none of the interactions were significant[1]. Therefore, the final model contains just x_2, passing yards, x_7, percent rushing, and x_8, opponents' rushing yards.

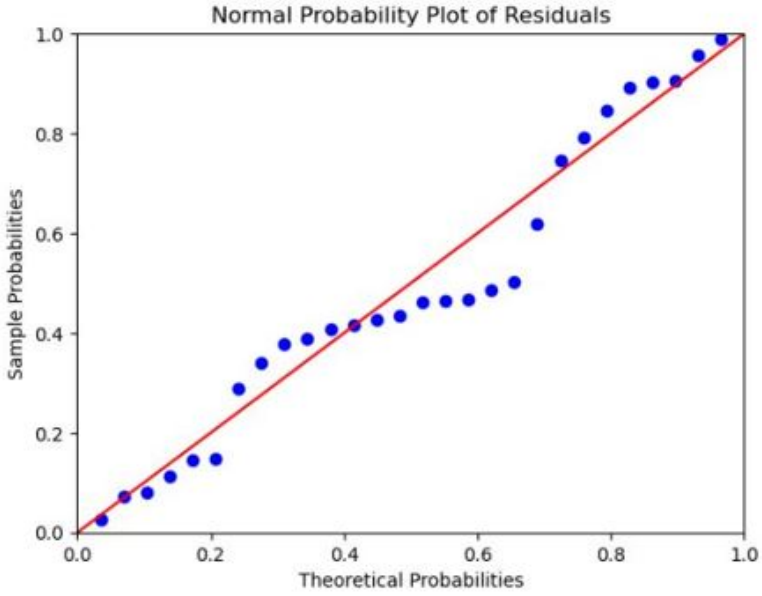
[1] An in depth examination shows that the interaction terms could be used as opposed to the individual regressors, while achieving similar success in terms of coefficient significance. Such a practice should be avoided if possible, in order to maintain the explanatory nature of the individual regressors.

4 Residual Analysis

In order to run significance tests on our produced model, we must ensure that (1) the residuals are approximately normally distributed around a mean of zero and that (2) the variance of the residuals is approximately constant.

Using the fitted model above, a normal probability plot of the residuals was constructed, as well as a plot of the residuals as a function of y. Both can be seen below,



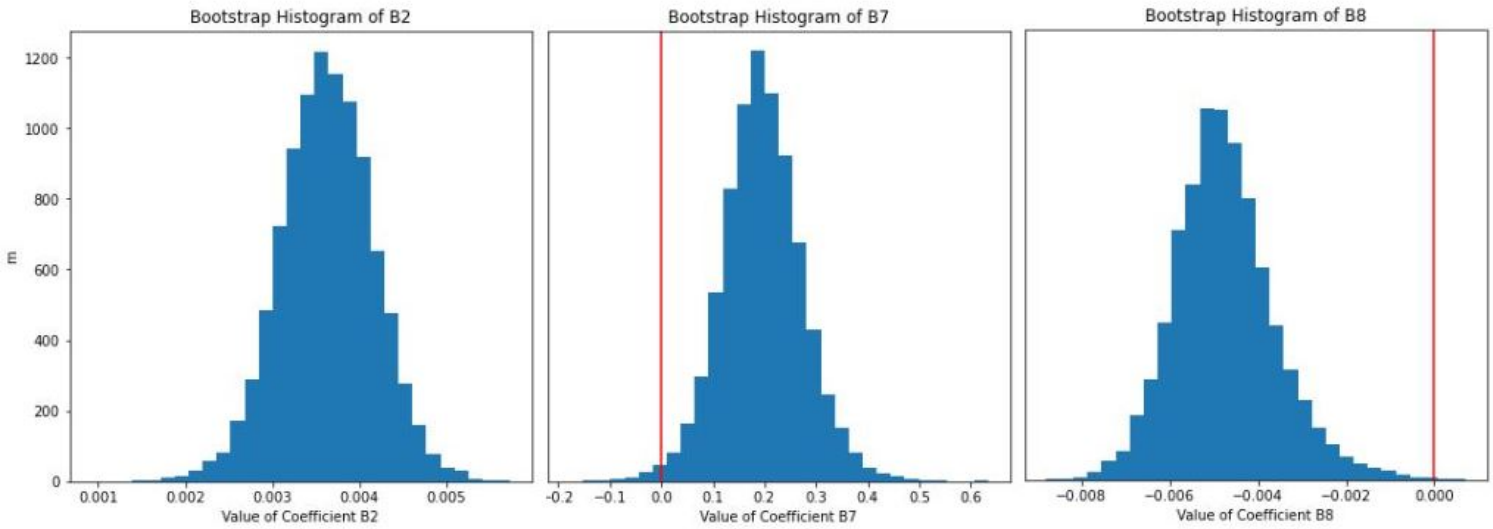


When viewing the normal probability plot, one should expect to see the data hugging the unit line, with any deviation from the line appearing to be random and without pattern. This behavior would suggest normality. However, in our normal probability plot, the residuals deviate modestly from the unit line (in certain sections) and also appear to have a distinct, wave-like pattern. This calls the normality assumption into question, and implies that significance testing of the linear coefficients, or the model as a whole, is not permissible.

Furthermore, when viewing the residuals as a function of y, we hoped to see the residuals as randomly scattered above and below the line. Though not severe, it does appear that there is a slightly positive correlation between the residuals and y. Again, this impacts our ability to conduct significance testing.

5 Bootstrapping

In the absence of normality, we move away from traditional statistical testing and instead employ bootstrapping. We drew $10,000$ random samples of size $n = 28$ from the data, with replacement and constructed a model using regressors x_2, x_7, and x_8 for each of these random samples. This provides $10,000$ values for the three linear coefficients, that are then bucketed, and plotted on histograms. In this approach, we can avoid having to assert the normality assumption, and instead rely on numerical estimation of the likelihood that a specific coefficient is zero.



In the above histograms, the red lines represent the cutoff regions regarding our bootstrap significance tests. Values of the coefficients that fall past these boundaries suggest that the coefficient could be zero. The p-value regarding a bootstrap test is simply the proportion of samples that produce a coefficient falling in these regions. Regarding β_2, we find that $p = 0$, since no samples produced a value of β_2 that fell below zero. The p-values for β_7 and β_8 are 0.0089 and 0.0008, respectively. All of these p-values are highly significant, and suggest that none of the coefficients are zero.

Furthermore, an F-test for the overall significance of regression cannot be conducted because of the normality condition. However, and equivalent test can be conducted by testing examining the probability that all three coefficients are jointly non-zero,

$$p_{overall} = (1 - 0)(1 - 0.0089)(1 - 0.0008) = 0.9903 \tag{1}$$

This value is obviously significant, and implies that the overall regression is significant.

6 Model Selection and Validation

Having shown that the model is significant, and that each individual regression coefficient is significant, it seems that the reduced model containing x_2, x_7, and x_8 is the ideal candidate model. All that remains is model validation.

The PRESS-statistic is a form of leave-one-out cross-validation that helps evaluate a model's predictive capability. In essence, $n = 28$ models are constructed, each using $n - 1$ values of the data. The individual values that are removed from each model construction are then used as a point of validation - the residual associated with this point helps describes the model's efficacy as related to prediction. The PRESS-statistic is merely the sum of these out of sample predictions. The press statistic for each step of the backwards elimination process is included below,

Regressors Included	PRESS-statistic
$x_1, x_2, x_3, x_4, x_5, x_6, x_7, x_8, x_9$	158.97
$x_1, x_2, x_3, x_4, x_6, x_7, x_8, x_9$	145.76
$x_2, x_3, x_4, x_6, x_7, x_8, x_9$	126.43
$x_2, x_3, x_4, x_7, x_8, x_9$	113.73
x_2, x_4, x_7, x_8, x_9	107.27
x_2, x_7, x_8, x_9	97.37
x_2, x_7, x_8	87.46

The table of PRESS-statistics confirms our previous selection of regressors for the model. We can see that as regressors are removed, the PRESS-statistic continues to go down, suggesting that our out of sample predictive capability improves. Furthermore, the lowest PRESS-statistic is held by our chosen model, meaning that it has the best predictive capability of all the models tested.

Therefore, having shown that the model is significant, and that it has predictive capability, the final model is given by,

$$\hat{y} = -1.8084 + 0.0036x_2 + 0.1940x_7 - 0.0048x_8 \tag{2}$$

7 Conclusions

Having provided the formula of the final model, it remains to be explained. What does this model mean in the context of football? Positive coefficients describe metrics that, when increased, improve the team's overall record. So, x_2, passing yards, and x_7, percent rushing, when increased, also tend to increase a team's standing. Negative coefficients imply that decreasing the associated metric tends to improve a team's record. Regressor x_8, opponents' rushing yards, should be limited in order to improve a team's performance.

These metrics have significant implications regarding what should be a team's focus. Regressors x_2 and x_7 provide insights as to a team's offensive strategy - winning teams should aim for high passing yardage while frequently running the ball. Regressor x_8 relates to defensive strategy. Limiting an opponent's ability to run the ball is significant if you hope to win.

As mentioned previously, profit is driven by success, and success depends on an improvement mentality. This regression helped provide insights as to what the key focus areas should be for a team's coaching staff, which ultimately helps drive business for the industry.

8 Appendix

Attached are the numerical and statistical calculations employed in this analysis.

Data Import and Explanation

In [1]:
```
# y : Games won (per 14 - game season)
# x1 : Rushing yards (season)
# x2 : Passing yards (season)
# x3 : Punting average (yards/punt)
# x4 : Field goal percentage (FGs made/FGs attempted 2season)
# x5 : Turnover differential (turnovers acquired – turnovers lost)
# x6 : Penalty yards (season)
# x7 : Percent rushing (rushing plays/total plays)
# x8 : Opponents ’  rushing yards (season)
# x9 :  Opponents ’   passing  yards  (season)
```

In [2]:
```
import pandas as pd
import numpy as np
import statsmodels.api as sm
import statsmodels.api as sm
import scipy.stats as stats
import matplotlib.pyplot as plt
from tqdm import tqdm
```

In [3]:
```
# Load Data

df = pd.read_excel(r'C:\Users\Eric\Downloads\linear_regression_5e_data_sets\linea
```

```
WARNING *** OLE2 inconsistency: SSCS size is 0 but SSAT size is non-zero
```

In [4]:
```
y = np.array(df['y'])
X = np.array(df[['x1','x2','x3','x4','x5','x6','x7','x8','x9']])
```

Backward Elimination and PRESS

In [55]:
```
X = sm.add_constant(X)
mod = sm.OLS(y, X)
results = mod.fit()
print(results.summary())

Xmod = X[:,:]
H = Xmod@np.linalg.inv(Xmod.T@Xmod)@Xmod.T
lev = np.diag(H)
PRESS_res = (res/(1-lev))**2
print('--------------')
print(sum(PRESS_res))
```

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                      y   R-squared:                       0.816
Model:                            OLS   Adj. R-squared:                  0.723
Method:                 Least Squares   F-statistic:                     8.846
Date:                Sat, 07 May 2022   Prob (F-statistic):           5.30e-05
Time:                        15:44:05   Log-Likelihood:                -50.468
No. Observations:                  28   AIC:                             120.9
Df Residuals:                      18   BIC:                             134.3
Df Model:                           9
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const         -7.2919     12.813     -0.569      0.576     -34.211      19.627
x1             0.0008      0.002      0.405      0.690      -0.003       0.005
x2             0.0036      0.001      4.318      0.000       0.002       0.005
x3             0.1222      0.259      0.472      0.643      -0.422       0.666
x4             0.0319      0.042      0.767      0.453      -0.056       0.119
x5          1.511e-05      0.047      0.000      1.000      -0.098       0.098
x6             0.0016      0.003      0.490      0.630      -0.005       0.008
x7             0.1544      0.152      1.015      0.324      -0.165       0.474
x8            -0.0039      0.002     -1.898      0.074      -0.008       0.000
x9            -0.0018      0.001     -1.264      0.222      -0.005       0.001
==============================================================================
Omnibus:                        0.114   Durbin-Watson:                   1.746
Prob(Omnibus):                  0.945   Jarque-Bera (JB):                0.143
Skew:                          -0.118   Prob(JB):                        0.931
Kurtosis:                       2.742   Cond. No.                     1.60e+05
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctl
y specified.
[2] The condition number is large, 1.6e+05. This might indicate that there are
strong multicollinearity or other numerical problems.
--------------
158.9737975242984
```

In [56]:
```
mod = sm.OLS(y, X[:,[0,1,2,3,4,6,7,8,9]])
results = mod.fit()
print(results.summary(xname=['const','x1','x2','x3','x4','x6','x7','x8','x9']))

Xmod = X[:,[0,1,2,3,4,6,7,8,9]]
H = Xmod@np.linalg.inv(Xmod.T@Xmod)@Xmod.T
lev = np.diag(H)
PRESS_res = (res/(1-lev))**2
print('--------------')
print(sum(PRESS_res))
```

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                      y   R-squared:                       0.816
Model:                            OLS   Adj. R-squared:                  0.738
Method:                 Least Squares   F-statistic:                     10.50
Date:                Sat, 07 May 2022   Prob (F-statistic):           1.54e-05
Time:                        15:44:30   Log-Likelihood:                -50.468
No. Observations:                  28   AIC:                             118.9
Df Residuals:                      19   BIC:                             130.9
Df Model:                           8
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const         -7.2937     11.336     -0.643      0.528     -31.020      16.433
x1             0.0008      0.002      0.417      0.681      -0.003       0.005
x2             0.0036      0.001      4.594      0.000       0.002       0.005
x3             0.1222      0.251      0.486      0.632      -0.404       0.648
x4             0.0319      0.040      0.804      0.431      -0.051       0.115
x6             0.0016      0.003      0.508      0.618      -0.005       0.008
x7             0.1544      0.140      1.103      0.284      -0.139       0.447
x8            -0.0039      0.002     -1.952      0.066      -0.008       0.000
x9            -0.0018      0.001     -1.299      0.210      -0.005       0.001
==============================================================================
Omnibus:                        0.114   Durbin-Watson:                   1.746
Prob(Omnibus):                  0.944   Jarque-Bera (JB):                0.143
Skew:                          -0.118   Prob(JB):                        0.931
Kurtosis:                       2.742   Cond. No.                     1.46e+05
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctl
y specified.
[2] The condition number is large, 1.46e+05. This might indicate that there are
strong multicollinearity or other numerical problems.
--------------
145.76288263628717
```

In [57]:
```
mod = sm.OLS(y, X[:,[0,2,3,4,6,7,8,9]])
results = mod.fit()
print(results.summary(xname=['const','x2','x3','x4','x6','x7','x8','x9']))

Xmod = X[:,[0,2,3,4,6,7,8,9]]
H = Xmod@np.linalg.inv(Xmod.T@Xmod)@Xmod.T
lev = np.diag(H)
PRESS_res = (res/(1-lev))**2
print('--------------')
print(sum(PRESS_res))
```

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                      y   R-squared:                       0.814
Model:                            OLS   Adj. R-squared:                  0.749
Method:                 Least Squares   F-statistic:                     12.50
Date:                Sat, 07 May 2022   Prob (F-statistic):           4.45e-06
Time:                        15:45:10   Log-Likelihood:                -50.596
No. Observations:                  28   AIC:                             117.2
Df Residuals:                      20   BIC:                             127.9
Df Model:                           7
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const         -9.1299     10.230     -0.893      0.383     -30.468      12.209
x2             0.0036      0.001      4.693      0.000       0.002       0.005
x3             0.1671      0.222      0.751      0.461      -0.297       0.631
x4             0.0370      0.037      1.001      0.329      -0.040       0.114
x6             0.0015      0.003      0.476      0.639      -0.005       0.008
x7             0.1891      0.110      1.716      0.102      -0.041       0.419
x8            -0.0042      0.002     -2.336      0.030      -0.008      -0.000
x9            -0.0017      0.001     -1.263      0.221      -0.004       0.001
==============================================================================
Omnibus:                        0.043   Durbin-Watson:                   1.627
Prob(Omnibus):                  0.979   Jarque-Bera (JB):                0.204
Skew:                          -0.078   Prob(JB):                        0.903
Kurtosis:                       2.612   Cond. No.                     1.17e+05
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctl
y specified.
[2] The condition number is large, 1.17e+05. This might indicate that there are
strong multicollinearity or other numerical problems.
--------------
126.42767436409402
```

In [58]:
```
mod = sm.OLS(y, X[:,[0,2,3,4,7,8,9]])
results = mod.fit()
print(results.summary(xname=['const','x2','x3','x4','x7','x8','x9']))

Xmod = X[:,[0,2,3,4,7,8,9]]
H = Xmod@np.linalg.inv(Xmod.T@Xmod)@Xmod.T
lev = np.diag(H)
PRESS_res = (res/(1-lev))**2
print('--------------')
print(sum(PRESS_res))
```

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                      y   R-squared:                       0.812
Model:                            OLS   Adj. R-squared:                  0.758
Method:                 Least Squares   F-statistic:                     15.10
Date:                Sat, 07 May 2022   Prob (F-statistic):           1.19e-06
Time:                        15:45:17   Log-Likelihood:                -50.754
No. Observations:                  28   AIC:                             115.5
Df Residuals:                      21   BIC:                             124.8
Df Model:                           6
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const         -7.6949      9.594     -0.802      0.432     -27.647      12.257
x2             0.0036      0.001      4.762      0.000       0.002       0.005
x3             0.1675      0.218      0.767      0.451      -0.287       0.622
x4             0.0350      0.036      0.972      0.342      -0.040       0.110
x7             0.1930      0.108      1.790      0.088      -0.031       0.417
x8            -0.0044      0.002     -2.500      0.021      -0.008      -0.001
x9            -0.0017      0.001     -1.284      0.213      -0.004       0.001
==============================================================================
Omnibus:                        0.052   Durbin-Watson:                   1.618
Prob(Omnibus):                  0.975   Jarque-Bera (JB):                0.064
Skew:                          -0.036   Prob(JB):                        0.968
Kurtosis:                       2.777   Cond. No.                     1.10e+05
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctl
y specified.
[2] The condition number is large, 1.1e+05. This might indicate that there are
strong multicollinearity or other numerical problems.
--------------
113.73520210752422
```

In [59]:
```
mod = sm.OLS(y, X[:,[0,2,4,7,8,9]])
results = mod.fit()
print(results.summary(xname=['const','x2','x4','x7','x8','x9']))

Xmod = X[:,[0,2,4,7,8,9]]
H = Xmod@np.linalg.inv(Xmod.T@Xmod)@Xmod.T
lev = np.diag(H)
PRESS_res = (res/(1-lev))**2
print('--------------')
print(sum(PRESS_res))
```

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                      y   R-squared:                       0.807
Model:                            OLS   Adj. R-squared:                  0.763
Method:                 Least Squares   F-statistic:                     18.34
Date:                Sat, 07 May 2022   Prob (F-statistic):           3.42e-07
Time:                        15:45:26   Log-Likelihood:                -51.141
No. Observations:                  28   AIC:                             114.3
Df Residuals:                      22   BIC:                             122.3
Df Model:                           5
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const         -4.6269      8.640     -0.536      0.598     -22.544      13.291
x2             0.0037      0.001      5.127      0.000       0.002       0.005
x4             0.0264      0.034      0.778      0.445      -0.044       0.097
x7             0.2347      0.092      2.542      0.019       0.043       0.426
x8            -0.0037      0.001     -2.481      0.021      -0.007      -0.001
x9            -0.0018      0.001     -1.399      0.176      -0.004       0.001
==============================================================================
Omnibus:                        0.192   Durbin-Watson:                   1.562
Prob(Omnibus):                  0.908   Jarque-Bera (JB):                0.122
Skew:                           0.134   Prob(JB):                        0.941
Kurtosis:                       2.821   Cond. No.                     9.98e+04
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctl
y specified.
[2] The condition number is large, 9.98e+04. This might indicate that there are
strong multicollinearity or other numerical problems.
--------------
107.2760613090066
```

In [60]:
```
mod = sm.OLS(y, X[:,[0,2,7,8,9]])
results = mod.fit()
print(results.summary(xname=['const','x2','x7','x8','x9']))

Xmod = X[:,[0,2,7,8,9]]
H = Xmod@np.linalg.inv(Xmod.T@Xmod)@Xmod.T
lev = np.diag(H)
PRESS_res = (res/(1-lev))**2
print('--------------')
print(sum(PRESS_res))
```

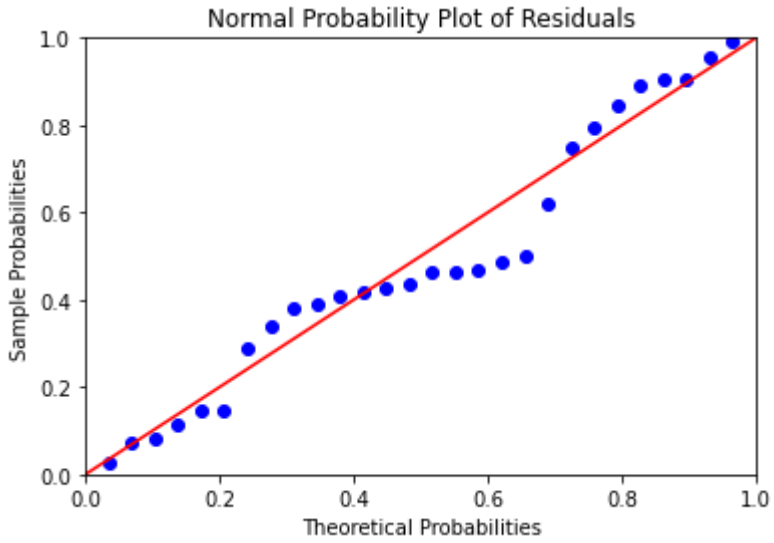
```
                            OLS Regression Results
==============================================================================
Dep. Variable:                      y   R-squared:                       0.801
Model:                            OLS   Adj. R-squared:                  0.767
Method:                 Least Squares   F-statistic:                     23.17
Date:                Sat, 07 May 2022   Prob (F-statistic):           8.74e-08
Time:                        15:45:34   Log-Likelihood:                -51.522
No. Observations:                  28   AIC:                             113.0
Df Residuals:                      23   BIC:                             119.7
Df Model:                           4
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const         -1.8217      7.785     -0.234      0.817     -17.926      14.282
x2             0.0038      0.001      5.416      0.000       0.002       0.005
x7             0.2169      0.089      2.446      0.023       0.033       0.400
x8            -0.0040      0.001     -2.871      0.009      -0.007      -0.001
x9            -0.0016      0.001     -1.312      0.202      -0.004       0.001
==============================================================================
Omnibus:                        0.392   Durbin-Watson:                   1.677
Prob(Omnibus):                  0.822   Jarque-Bera (JB):                0.007
Skew:                           0.001   Prob(JB):                        0.997
Kurtosis:                       3.076   Cond. No.                     9.07e+04
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctl
y specified.
[2] The condition number is large, 9.07e+04. This might indicate that there are
strong multicollinearity or other numerical problems.
--------------
97.37173955060527
```

In [68]:
```
mod = sm.OLS(y, X[:,[0,2,7,8]])
results = mod.fit()
print(results.summary(xname=['const','x2','x7','x8']))

Xmod = X[:,[0,2,7,8]]
H = Xmod@np.linalg.inv(Xmod.T@Xmod)@Xmod.T
lev = np.diag(H)
PRESS_res = (res/(1-lev))**2
print('--------------')
print(sum(PRESS_res))
```

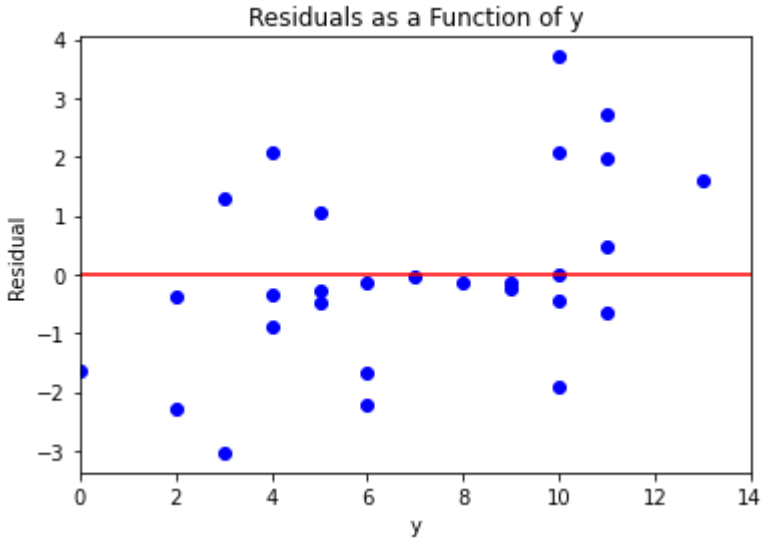
```
                            OLS Regression Results
==============================================================================
Dep. Variable:                      y   R-squared:                       0.786
Model:                            OLS   Adj. R-squared:                  0.760
Method:                 Least Squares   F-statistic:                     29.44
Date:                Sat, 07 May 2022   Prob (F-statistic):           3.27e-08
Time:                        16:51:54   Log-Likelihood:                -52.532
No. Observations:                  28   AIC:                             113.1
Df Residuals:                      24   BIC:                             118.4
Df Model:                           3
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const         -1.8084      7.901     -0.229      0.821     -18.115      14.498
x2             0.0036      0.001      5.177      0.000       0.002       0.005
x7             0.1940      0.088      2.198      0.038       0.012       0.376
x8            -0.0048      0.001     -3.771      0.001      -0.007      -0.002
==============================================================================
Omnibus:                        0.665   Durbin-Watson:                   1.492
Prob(Omnibus):                  0.717   Jarque-Bera (JB):                0.578
Skew:                           0.321   Prob(JB):                        0.749
Kurtosis:                       2.712   Cond. No.                     7.42e+04
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctl
y specified.
[2] The condition number is large, 7.42e+04. This might indicate that there are
strong multicollinearity or other numerical problems.
--------------
87.4612306618114
```

Residual Analysis

In [69]:
```
import statsmodels.api as sm
import scipy.stats as stats
pplot = sm.ProbPlot(res, stats.t, fit=True)
fig = pplot.ppplot(line="45")
h = plt.title("Normal Probability Plot of Residuals")
plt.show()
```



In [71]:
```
plt.scatter(y,res, c='b')
plt.plot([0,14],[0,0],c='r')
plt.xlim(0,14)
plt.xlabel('y')
plt.ylabel('Residual')
h = plt.title("Residuals as a Function of y")
plt.show()
```

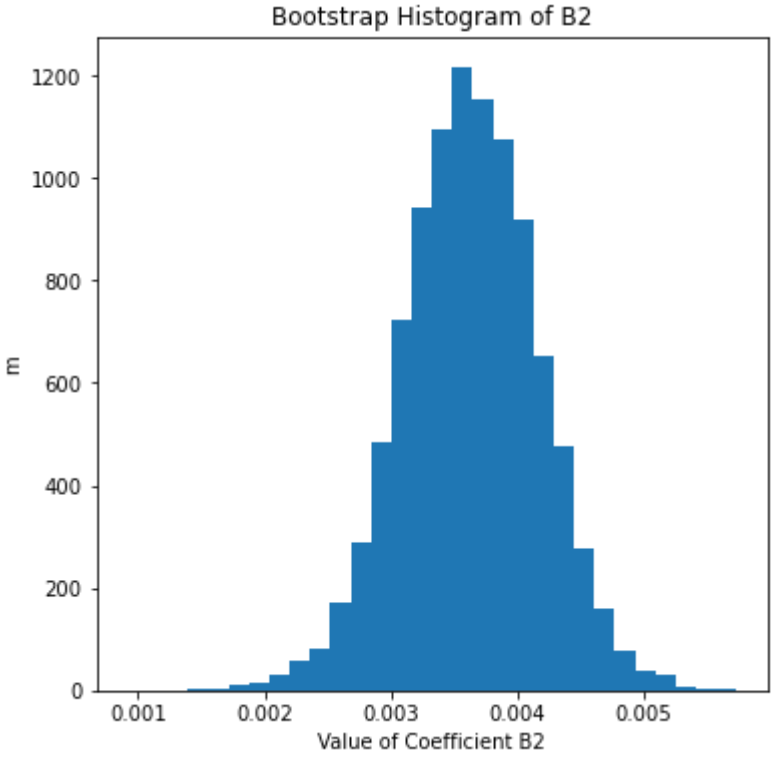


Bootstrapping

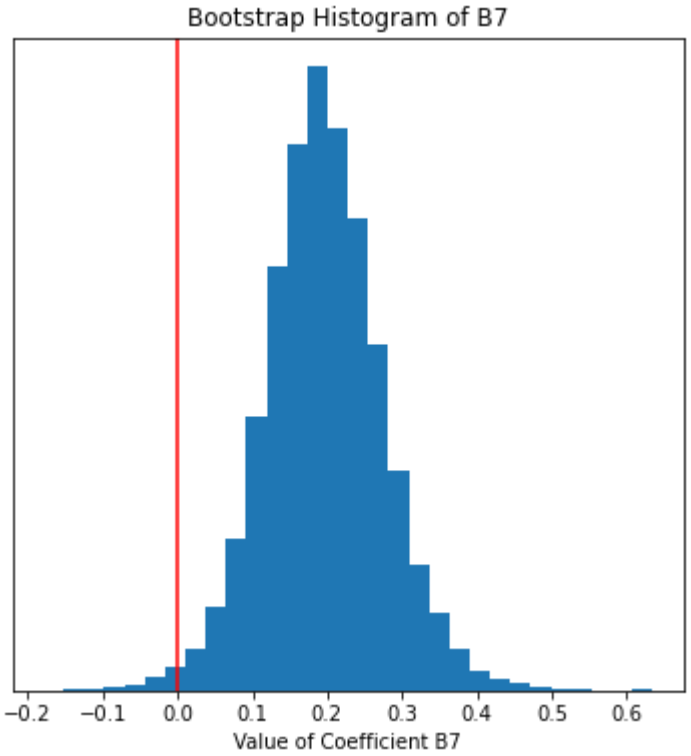
In [13]:
```
all_params = np.zeros((10000,4))
for t in tqdm(range(10000)):
    sample = df.sample(len(df), replace=True)
    y_ = np.array(sample['y'])
    X_ = np.array(sample[['x2','x7','x8']])
    X_ = sm.add_constant(X_)
    mod = sm.OLS(y_, X_)
    results = mod.fit()
    all_params[t,:] = results.params
```

```
100%|██████████████████████████████████████████████████████████████
██| 10000/10000 [00:11<00:00, 845.85it/s]
```

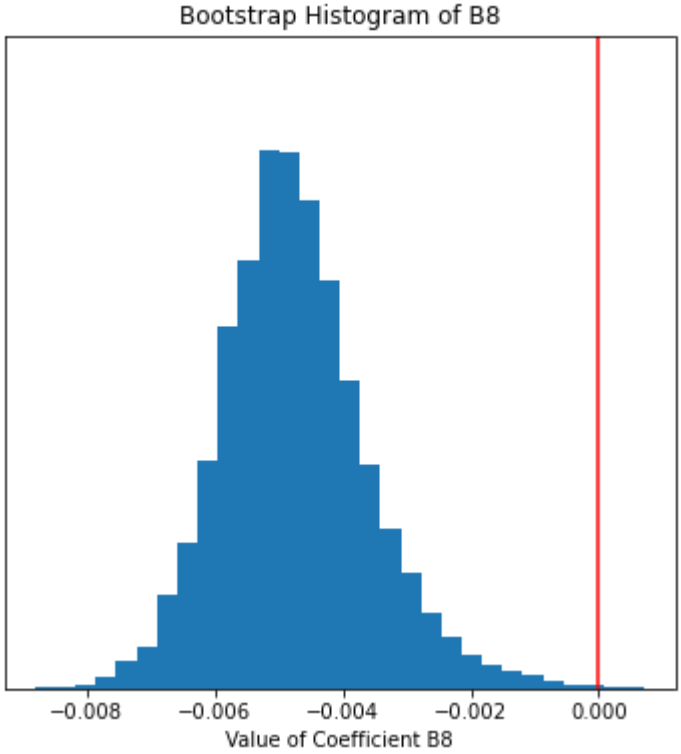
In [33]:
```
plt.figure(figsize=(6,6))
plt.hist(all_params[:,1], bins=30)
plt.ylabel('m')
plt.xlabel('Value of Coefficient B2')
plt.title('Bootstrap Histogram of B2')
plt.show()
```



In [38]:
```
plt.figure(figsize=(6,6))
plt.hist(all_params[:,2], bins=30)
plt.plot([0,0],[0,1600],c='r')
plt.ylim(0,1600)
ax = plt.gca()
ax.get_yaxis().set_visible(False)
plt.xlabel('Value of Coefficient B7')
plt.title('Bootstrap Histogram of B7')
plt.show()
```



In [39]:
```
plt.figure(figsize=(6,6))
plt.hist(all_params[:,3], bins=30)
plt.plot([0,0],[0,1500],c='r')
plt.ylim(0,1500)
ax = plt.gca()
ax.get_yaxis().set_visible(False)
plt.xlabel('Value of Coefficient B8')
plt.title('Bootstrap Histogram of B8')
plt.show()
```



In [32]:
```
print('', len([i for i in all_params[:,1] if i < 0])/10000)
print('', len([i for i in all_params[:,2] if i < 0])/10000)
print('', len([i for i in all_params[:,3] if i > 0])/10000)
```

```
0.0
0.0089
0.0008
```

Interaction Effect Testing

In [56]:
```
mod = sm.OLS(y, X[:,[0,2,7,8]])
results = mod.fit()
print(results.summary(xname=['const','x2','x7','x8']))
```

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                      y   R-squared:                       0.786
Model:                            OLS   Adj. R-squared:                  0.760
Method:                 Least Squares   F-statistic:                     29.44
Date:                Fri, 06 May 2022   Prob (F-statistic):           3.27e-08
Time:                        11:07:51   Log-Likelihood:                -52.532
No. Observations:                  28   AIC:                             113.1
Df Residuals:                      24   BIC:                             118.4
Df Model:                           3
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const         -1.8084      7.901     -0.229      0.821     -18.115      14.498
x2             0.0036      0.001      5.177      0.000       0.002       0.005
x7             0.1940      0.088      2.198      0.038       0.012       0.376
x8            -0.0048      0.001     -3.771      0.001      -0.007      -0.002
==============================================================================
Omnibus:                        0.665   Durbin-Watson:                   1.492
Prob(Omnibus):                  0.717   Jarque-Bera (JB):                0.578
Skew:                           0.321   Prob(JB):                        0.749
Kurtosis:                       2.712   Cond. No.                     7.42e+04
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctl
y specified.
[2] The condition number is large, 7.42e+04. This might indicate that there are
strong multicollinearity or other numerical problems.
```

In [61]:
```
Xnew = np.zeros((len(X),14))
Xnew[:,:10] = X
Xnew[:,10] = X[:,2]*X[:,7]
Xnew[:,11] = X[:,2]*X[:,8]
Xnew[:,12] = X[:,7]*X[:,8]
Xnew[:,13] = X[:,2]*X[:,7]*X[:,8]
```

In [64]:
```
mod = sm.OLS(y, Xnew[:,[0,2,7,8,10,11,12,13]])
results = mod.fit()
print(results.summary(xname=['const','x2','x7','x8','x2x7','x2x8','x7x8','x2x7x8'
```

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                      y   R-squared:                       0.801
Model:                            OLS   Adj. R-squared:                  0.732
Method:                 Least Squares   F-statistic:                     11.51
Date:                Fri, 06 May 2022   Prob (F-statistic):           8.31e-06
Time:                        11:13:14   Log-Likelihood:                -51.521
No. Observations:                  28   AIC:                             119.0
Df Residuals:                      20   BIC:                             129.7
Df Model:                           7
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const       -168.5370    152.514     -1.105      0.282    -486.676     149.602
x2             0.0806      0.067      1.211      0.240      -0.058       0.219
x7             2.8785      2.515      1.145      0.266      -2.367       8.124
x8             0.0625      0.067      0.937      0.360      -0.077       0.202
x2x7          -0.0012      0.001     -1.130      0.272      -0.004       0.001
x2x8       -3.062e-05   2.83e-05     -1.080      0.293   -8.98e-05    2.85e-05
x7x8          -0.0011      0.001     -0.966      0.346      -0.003       0.001
x2x7x8      4.979e-07   4.82e-07      1.033      0.314   -5.08e-07     1.5e-06
==============================================================================
Omnibus:                        1.735   Durbin-Watson:                   1.580
Prob(Omnibus):                  0.420   Jarque-Bera (JB):                0.957
Skew:                           0.448   Prob(JB):                        0.620
Kurtosis:                       3.137   Cond. No.                     1.18e+11
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctl
y specified.
[2] The condition number is large, 1.18e+11. This might indicate that there are
strong multicollinearity or other numerical problems.
```

In [66]:
```
mod = sm.OLS(y, Xnew[:,[0,2,7,8,10,11,13]])
results = mod.fit()
print(results.summary(xname=['const','x2','x7','x8','x2x7','x2x8','x2x7x8']))
```

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                      y   R-squared:                       0.792
Model:                            OLS   Adj. R-squared:                  0.732
Method:                 Least Squares   F-statistic:                     13.32
Date:                Fri, 06 May 2022   Prob (F-statistic):           3.27e-06
Time:                        11:14:50   Log-Likelihood:                -52.160
No. Observations:                  28   AIC:                             118.3
Df Residuals:                      21   BIC:                             127.6
Df Model:                           6
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const        -26.4725     40.420     -0.655      0.520    -110.531      57.586
x2             0.0202      0.023      0.888      0.384      -0.027       0.067
x7             0.4966      0.495      1.003      0.327      -0.533       1.526
x8            -0.0016      0.007     -0.250      0.805      -0.015       0.012
x2x7          -0.0002      0.000     -0.700      0.491      -0.001       0.000
x2x8       -3.673e-06   5.05e-06     -0.727      0.475   -1.42e-05    6.83e-06
x2x7x8      3.822e-08   7.76e-08      0.493      0.627   -1.23e-07       2e-07
==============================================================================
Omnibus:                        0.414   Durbin-Watson:                   1.616
Prob(Omnibus):                  0.813   Jarque-Bera (JB):                0.230
Skew:                           0.213   Prob(JB):                        0.891
Kurtosis:                       2.873   Cond. No.                     3.14e+10
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctl
y specified.
[2] The condition number is large, 3.14e+10. This might indicate that there are
strong multicollinearity or other numerical problems.
```

In [68]:
```
mod = sm.OLS(y, Xnew[:,[0,2,7,8,10,11]])
results = mod.fit()
print(results.summary(xname=['const','x2','x7','x8','x2x7','x2x8']))
```

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                      y   R-squared:                       0.790
Model:                            OLS   Adj. R-squared:                  0.742
Method:                 Least Squares   F-statistic:                     16.50
Date:                Fri, 06 May 2022   Prob (F-statistic):           8.40e-07
Time:                        11:16:43   Log-Likelihood:                -52.320
No. Observations:                  28   AIC:                             116.6
Df Residuals:                      22   BIC:                             124.6
Df Model:                           5
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const        -23.2883     39.208     -0.594      0.559    -104.600      58.024
x2             0.0137      0.018      0.752      0.460      -0.024       0.051
x7             0.4290      0.467      0.918      0.369      -0.540       1.398
x8            -0.0013      0.006     -0.200      0.844      -0.014       0.012
x2x7          -0.0001      0.000     -0.507      0.617      -0.001       0.000
x2x8       -1.645e-06   2.88e-06     -0.572      0.573   -7.62e-06    4.32e-06
==============================================================================
Omnibus:                        0.551   Durbin-Watson:                   1.512
Prob(Omnibus):                  0.759   Jarque-Bera (JB):                0.377
Skew:                           0.271   Prob(JB):                        0.828
Kurtosis:                       2.829   Cond. No.                     5.48e+08
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctl
y specified.
[2] The condition number is large, 5.48e+08. This might indicate that there are
strong multicollinearity or other numerical problems.
```

In [71]:
```
mod = sm.OLS(y, Xnew[:,[0,2,7,8]])
results = mod.fit()
print(results.summary(xname=['const','x2','x7','x8']))
```

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                      y   R-squared:                       0.786
Model:                            OLS   Adj. R-squared:                  0.760
Method:                 Least Squares   F-statistic:                     29.44
Date:                Fri, 06 May 2022   Prob (F-statistic):           3.27e-08
Time:                        11:17:27   Log-Likelihood:                -52.532
No. Observations:                  28   AIC:                             113.1
Df Residuals:                      24   BIC:                             118.4
Df Model:                           3
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const         -1.8084      7.901     -0.229      0.821     -18.115      14.498
x2             0.0036      0.001      5.177      0.000       0.002       0.005
x7             0.1940      0.088      2.198      0.038       0.012       0.376
x8            -0.0048      0.001     -3.771      0.001      -0.007      -0.002
==============================================================================
Omnibus:                        0.665   Durbin-Watson:                   1.492
Prob(Omnibus):                  0.717   Jarque-Bera (JB):                0.578
Skew:                           0.321   Prob(JB):                        0.749
Kurtosis:                       2.712   Cond. No.                     7.42e+04
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctl
y specified.
[2] The condition number is large, 7.42e+04. This might indicate that there are
strong multicollinearity or other numerical problems.
```