Oil Production Model

STAT 656, Applied Analytics, Homework #1

E. Lee Rainwater

Mays Business School, Texas A&M University, College Station, Texas, USA, lee.rainwater@tamu.edu

1 SAS Enterprise Miner Assignment

1.1 General Modeling and Results

The import, preprocessing, modeling, and reporting processes were set up according to the following Process Flow Diagram in Figure 1.

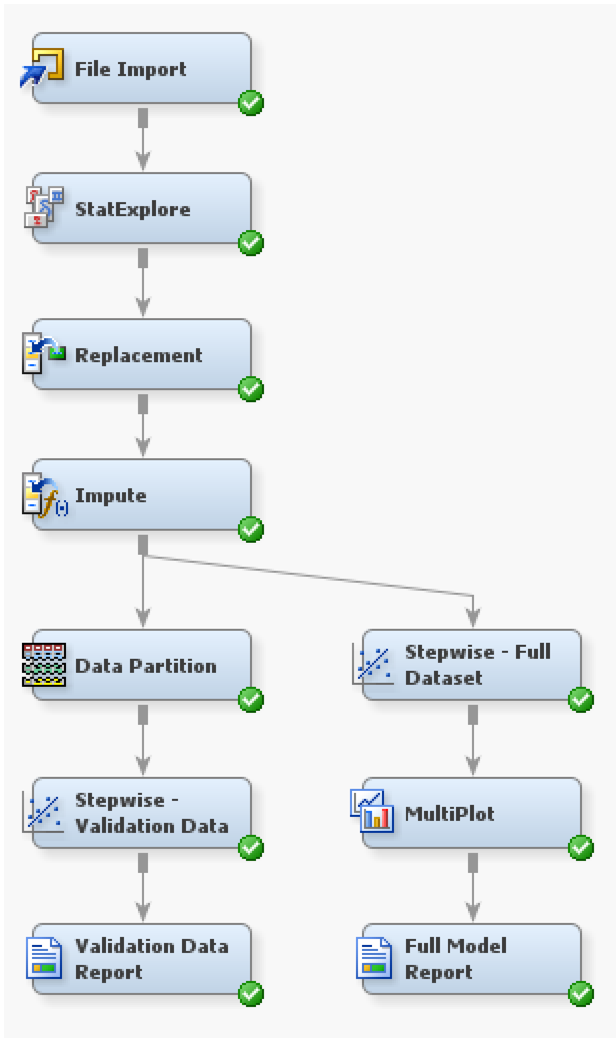

Figure 1 – Process Flow Diagram for Oil Production Model

After importing, the interval variables are summarized as follows in Table 1:

Ordered Inputs	Data Role	Variable	Median	Missing	Non Missing	Minimum	Maximum	Mean	Standard Deviation	Skewness
1	TRAIN	Log_Carbonate	1.428139	1	4751	-3.98184	3.777379	1.329109	0.863187	-0.84417
2	TRAIN	N_Stages	8	1	4751	2	14	7.888445	3.117991	-0.02086
3	TRAIN	Log_Cum_Production	12.67439	1	4751	8.798606	14.38292	12.6006	0.690231	-0.70733
4	TRAIN	Log_Proppant_LB	14.30122	1	4751	6.309918	17.98899	14.27667	0.771792	-1.91123
5	TRAIN	Log_Frac_Fluid_GL	14.86068	1	4751	7.824046	17.77034	14.73261	0.792053	-2.84318
6	TRAIN	Log_GrossPerforatedInterval	7.760041	1	4751	4.60517	8.794825	7.720973	0.400409	-1.20259
7	TRAIN	Log_UpperPerforation_xy	8.913954	1	4751	8.180321	9.380083	8.898754	0.126218	-0.65632
8	TRAIN	Log_TotalDepth	9.195227	1	4751	8.575462	9.63874	9.194772	0.118666	-0.11293
9	TRAIN	Log_LowerPerforation_xy	9.184407	1	4751	8.509161	9.618735	9.182296	0.117756	-0.18413
10	TRAIN	Y_Well	32.56777	1	4751	31.88254	33.40471	32.64008	0.295745	0.423351
11	TRAIN	X_Well	-97.4474	1	4751	-98.5066	-97.0224	-97.4724	0.233262	-0.78883

Table 1 – Interval Variable Statistics

No outliers are found per boundaries defined in the data dictionary; however, each attribute has exactly one missing value. The tree method is used to impute missing values; although for the small number it would be reasonable to simply drop records with missing values.

The resultant dataset was partitioned 70/30 into training and validation datasets. The model fit statistics for the training and validation models are as follows:

Label of Statistic	Train	Validation
Akaike's Information Criterion	-4249.27	.
Average Squared Error	0.28	0.28
Average Error Function	0.28	0.28
Degrees of Freedom for Error	3309.00	.
Model Degrees of Freedom	17.00	.
Total Degrees of Freedom	3326.00	.
Divisor for ASE	3326.00	1425.00
Error Function	917.55	404.30
Final Prediction Error	0.28	.
Maximum Absolute Error	3.52	3.03
Mean Square Error	0.28	0.28
Sum of Frequencies	3326.00	1425.00
Number of Estimate Weights	17.00	.
Root Average Sum of Squares	0.53	0.53
Root Final Prediction Error	0.53	.
Root Mean Squared Error	0.53	0.53
Schwarz's Bayesian Criterion	-4145.41	.
Sum of Squared Errors	917.55	404.30
Sum of Case Weights Times Freq	3326.00	1425.00

Table 2 – Model Fit Statistics

```
              Model Fit Statistics

R-Square         0.4144     Adj R-Sq         0.4116
AIC          -4249.2702     BIC          -4247.4003
SBC          -4145.4082     C(p)            46.7750
```

Table 3 – Fundamental Model Fit Statistics

The model effects table shows *Log_Lower_Perforation_xy* to be the attribute with the greatest absolute *t*-value, indicating that it is the strongest effect.

Effect Number	Variable	Level	Coefficient	T-value	P Value	Effect Number	Variable	Level	Coefficient	T-value	P Value
1	Intercept		-6.21423	-7.1448	0.00000	10	IMP_County	6	-0.25581	-3.6773	0.00024
2	IMP_Log_LowerPerforation_xy		1.66462	16.3470	0.00000	11	IMP_County	12	-0.18434	-1.9574	0.05039
3	IMP_County	1	-0.72239	-4.0206	0.00006	12	IMP_County	3	0.16988	3.1298	0.00176
4	IMP_County	13	0.59548	12.3273	0.00000	13	IMP_County	10	-0.16348	-1.4004	0.16149
5	IMP_County	2	0.45725	0.9326	0.35110	14	IMP_Log_Proppant_LB		0.14660	10.9108	0.00000
6	IMP_County	5	-0.35780	-4.6464	0.00000	15	IMP_County	7	-0.12809	-2.3596	0.01835
7	IMP_County	14	0.29601	5.2548	0.00000	16	IMP_Log_Frac_Fluid_GL		0.08127	6.2750	0.00000
8	IMP_County	8	-0.28486	-3.5795	0.00035	17	IMP_County	11	-0.02771	-0.5444	0.58621
9	IMP_County	4	0.26237	1.3781	0.16827	.			.	.	.

Table 4 – Validation Model Effects Table

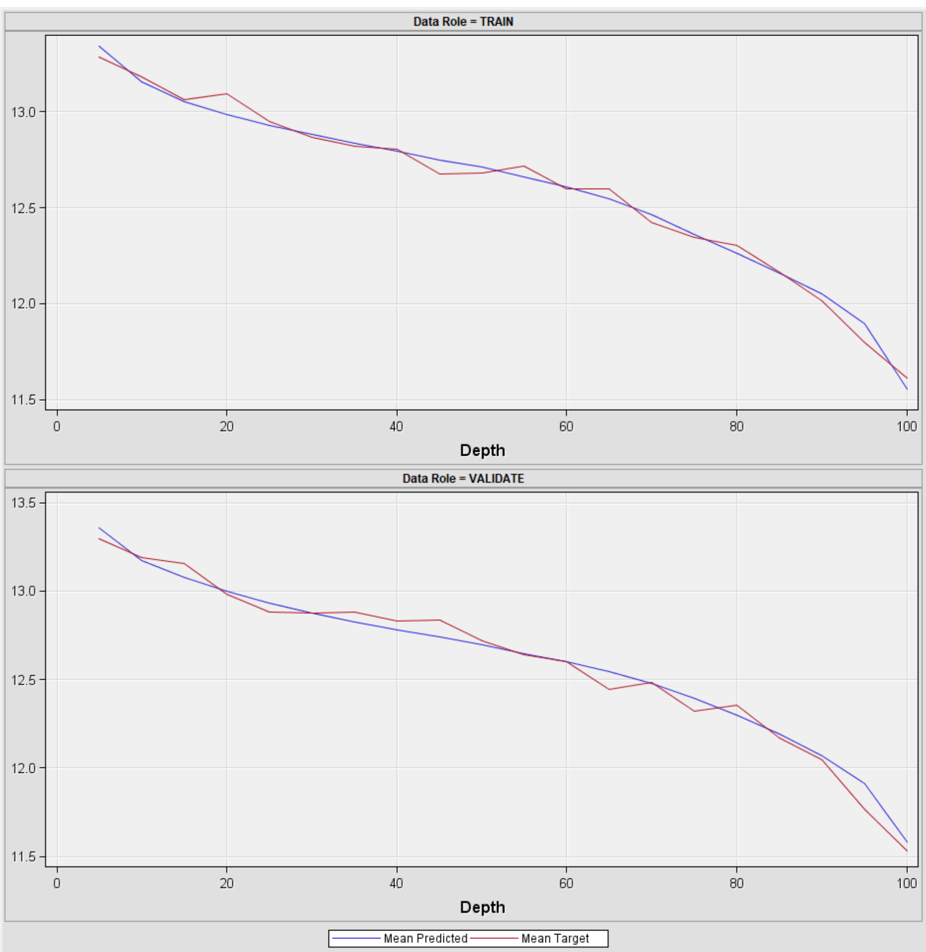

Figure 2 – Comparison of Mean Predicted vs. Mean Target, Log_Cum_Production

1.2 Assessment

While the R^2 is not particularly high, the predicted target appears to track the mean target reasonably well.

2 Python Assignment

2.1 General Results:

Of the 4752 observations and 12 attributes (prior to encoding) there were no outliers (as defined by the data dictionary) and only one observation of each of the 12 features had missing data.

Stepwise regression retained all attributes except for County9, using Log_Cum_Production as the target.

Log_LowerPerforation_xy is the attribute with the highest absolute value of the test statistic, $t = 17.180$. This suggests that this attribute is the strongest predictor of the target. Other strong predictors are *X_Well* (13.305), *County6* (-12.104), *Log_Proppant_LB* (11.959).

```
AICC:  7255.867607495069 AIC:  7255.570401737236 BIC:  7423.689272168096
************************************************************
                                   Target: Log_Cum_Production
                            OLS Regression Results
==============================================================================
Dep. Variable:                      y   R-squared:                       0.440
Model:                            OLS   Adj. R-squared:                  0.437
Method:                 Least Squares   F-statistic:                     154.8
Date:                Wed, 03 Jun 2020   Prob (F-statistic):               0.00
Time:                        22:44:45   Log-Likelihood:                -3601.8
```

Table 5 – Statistics for OLS Stepwise Model

	coef	std err	t	P>|t|	[0.025	0.975]
const	116.6109	9.358	12.462	0.000	98.266	134.956
Log_LowerPerforation_xy	1.6930	0.099	17.180	0.000	1.500	1.886
County13	0.1606	0.024	6.778	0.000	0.114	0.207
Log_Proppant_LB	0.1310	0.011	11.959	0.000	0.110	0.152
Y_Well	0.2392	0.044	5.491	0.000	0.154	0.325
X_Well	1.2009	0.090	13.305	0.000	1.024	1.378
Log_UpperPerforation_xy	-1.4815	0.157	-9.461	0.000	-1.788	-1.174
Log_Frac_Fluid_GL	0.0693	0.011	6.600	0.000	0.049	0.090
County3	-0.2896	0.038	-7.627	0.000	-0.364	-0.215
County8	-0.3264	0.070	-4.687	0.000	-0.463	-0.190
County1	-1.0569	0.166	-6.348	0.000	-1.383	-0.731
County6	-0.6049	0.050	-12.104	0.000	-0.703	-0.507
Operator15	0.3846	0.134	2.863	0.004	0.121	0.648
Log_Carbonate	0.0219	0.009	2.510	0.012	0.005	0.039
Operator7	0.0909	0.037	2.466	0.014	0.019	0.163
Operator5	-0.1069	0.055	-1.955	0.051	-0.214	0.000
Operator10	0.0937	0.049	1.912	0.056	-0.002	0.190
Operator27	0.1027	0.056	1.837	0.066	-0.007	0.212
County7	-0.3529	0.034	-10.373	0.000	-0.420	-0.286

Table 6 – Attributes and their statistics, OLS Stepwise Model

The large condition number (1.32E+05) indicates possible strong collinearity.

Plotting the observed *vs.* predicted values of the target variable shows these values to be clustered *mostly* symmetrically about a reference slope line of unity, (Figure 3) which may be indicative of a reasonable model fit. It is notable that the cluster seems to be rotated counter-clockwise off of the reference slope line, as a result of the model somewhat compressing the more extreme predictions of the target about the mean.

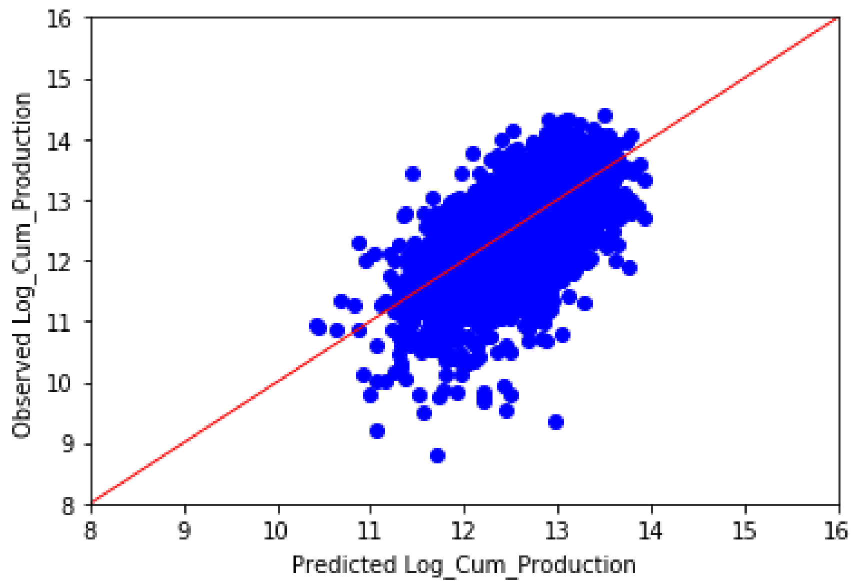

Figure 3 – Observed vs. Predicted Values of Log_Cum_Production

The means of the studentized residuals are zero, which is expected of a linear regression. As shown in Figure 4, A number of points have high values of Cook's Distance, indicating that they possess excessive leverage.

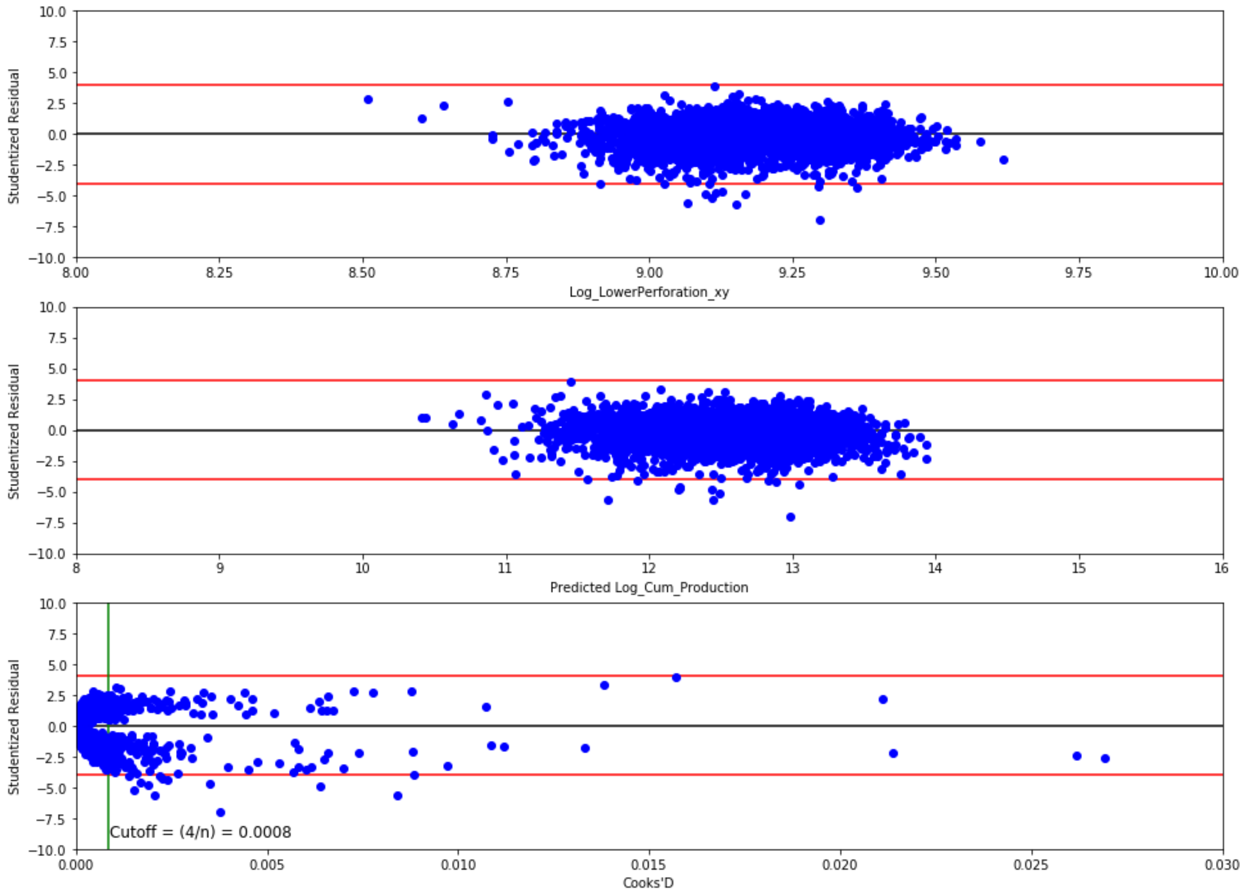

Figure 4 – Studentized Residuals

To evaluate fit of the model, the dataframe is subjected to a 70/30 training/validation split, to which an OLS model is applied using the same attributes as determined previously during stepwise feature selection.

```
*******************************************************
                              TRAINING MODEL
                              Target: Log_Cum_Production
                            OLS Regression Results
==============================================================================
Dep. Variable:                      y   R-squared:                       0.447
Model:                            OLS   Adj. R-squared:                  0.443
Method:                 Least Squares   F-statistic:                     111.1
Date:                Wed, 03 Jun 2020   Prob (F-statistic):               0.00
Time:                        22:44:45   Log-Likelihood:                -2471.3
No. Observations:                3325   AIC:                             4993.
Df Residuals:                    3300   BIC:                             5145.
Df Model:                          24
Covariance Type:            nonrobust
===========================================================================================
```

Table 7 – OLS Model Applied to Training Data

As can be seen in Table 7 – OLS Model Applied to Training DataTable 7, the value of R^2 changes only slightly from the original model. Comparison of statistics between the training and validation sets show similar values between the two for R^2 as well as Mean Absolute Error and ASE. Taken together, it would appear that the model is not overfitting the training data.

Comparison of the *Min/Mean/Max* values of the target indicates that the model is somewhat compressed about the mean, as was observed previously regarding Figure 3.

3 Conclusion

Fit of the OLS model appears to be reasonably adequate to be instructive and useful, although it tends to overestimate the lower values of the target and underestimate the higher values of the target. A number of observations with large values of Cook's Distance may be responsible for this model behavior, possibly indicating the presence of different data segments.

4 Appendix – Listing of Python Code

```
"""
Created 03 JUN 2020

@author: el-rainwater
"""
import pandas as pd
import numpy as np
from AdvancedAnalytics.ReplaceImputeEncode import ReplaceImputeEncode, DT
from AdvancedAnalytics.Regression import linreg, stepwise
import statsmodels.api as sm
import statsmodels.tools.eval_measures as em
from scipy.stats import norm
import matplotlib.pyplot as plt
from sklearn.model_selection import train_test_split


file = 'OilProductionHW2.xlsx'
filepath = 'C:/Users/rainwater-e/OneDrive - Texas A&M University/' \
    'Summer-2020/STAT 656 Applied Analytics/hw-02/'

df = pd.read_excel(filepath + file)
print(df.head(), df.shape)
print(df.dtypes)

attribute_map = {
    'Obs':[DT.ID, (1, np.inf)],
    'Log_Cum_Production':[DT.Interval, (8,15)],
    'Log_Proppant_LB':[DT.Interval, (6, 18)],
    'Log_Carbonate':[DT.Interval, (-4, 4)],
    'Log_Frac_Fluid_GL':[DT.Interval, (7, 18)],
    'Log_GrossPerforatedInterval':[DT.Interval, (4,9)],
    'Log_LowerPerforation_xy':[DT.Interval, (8,10)],
    'Log_UpperPerforation_xy':[DT.Interval, (8,10)],
    'Log_TotalDepth':[DT.Interval, (8,10)],
    'N_Stages':[DT.Interval, (2,14)],
    'X_Well':[DT.Interval, (-100, -95)],
    'Y_Well':[DT.Interval, (30,35)],
    'Operator':[DT.Nominal, list(range(1,29))],
    'County':[DT.Nominal, list(range(1,15))]
}

target = 'Log_Cum_Production'

# One-hot encode and impute missing values
rie = ReplaceImputeEncode(data_map=attribute_map, nominal_encoding='one-hot',
                          no_impute=[target], interval_scale=None, drop=True,
                          display=True)
df_encoded = rie.fit_transform(df).dropna() #drop rows with missing values


# Set up stepwise regression
sw = stepwise(df_encoded, target, reg='linear', method='stepwise',
              crit_in=0.1, crit_out=0.1, verbose=True)

selected = sw.fit_transform()

print('\nFinal selected attributes:\n', selected)
```

```python
y = df_encoded[target]
y = np.ravel(y) # Ravel it into a contiguous flattened array

X  = df_encoded[selected] # create a dataframe of the selected attributes
Xc = sm.add_constant(X)   # Add an intercept column to the regressors so that
                          # sm.OLS will work correctly
model   = sm.OLS(y,Xc)    # Use StatsModels.OLS
results = model.fit()

ll       = model.loglike(results.params) # Returns the log likelihood function
model_df = model.df_model + 2            # Corrects the DOF by adding for the
                                         # intercept and sigma
nobs     = y.shape[0]                    # Returns number of observations
aic      = em.aic(ll, nobs, model_df)
bic      = em.bic(ll, nobs, model_df)
aicc     = em.aicc(ll, nobs, model_df)

predicted  = results.fittedvalues
residual   = results.resid
influence  = results.get_influence()


# Gonna just copy this formatty stuff from Dr. J's example
# These are the correct values as reported in SAS
print("\n")
print("AICC: ", aicc, "AIC: ", aic, "BIC: ", bic)

print("\n************************************************************")
print("                                  Target: " + target)
print(results.summary())
print("\n************************************************************")


# Determine the attribute with the greatest absolute test statistic value
max_tvalue = results.tvalues[results.tvalues.keys()!='const'].abs().max()
max_attrib = results.tvalues[results.tvalues==max_tvalue].index[0]
print("\nStrongest Attribute: " + max_attrib)

# Set sigma intervals
n3 = 2.0*(1.0-norm.cdf(3.0)) * nobs
n4 = 2.0*(1.0-norm.cdf(4.0)) * nobs
n5 = 2.0*(1.0-norm.cdf(5.0)) * nobs
n6 = 2.0*(1.0-norm.cdf(6.0)) * nobs

print("\nExpected number of observations outside stated limits:")
print("-+ 3Sigma: ", int(round(n3)))
print("-+ 4Sigma: ", int(round(n4)))
print("-+ 5Sigma: ", int(round(n5)))

print("\n")
leverage  = influence.hat_matrix_diag
cooks_d   = influence.cooks_distance[0]

cutoffD = 4.0/nobs
print("Max Cooks D: {:8.4f}  Cutoff ( 4/n): {:8.5f}".
      format(cooks_d.max(), cutoffD))
cutoffH = 2.0 * model_df / nobs
print("Max H:       {:8.4f}  Cutoff (2p/n): {:8.5f}".
      format(leverage.max(), cutoffH))

std_residuals  = residual/np.sqrt(results.mse_resid)
stud_residuals = influence.resid_studentized_internal

print("\nResiduals beyond 4 sigma:")
outliers = np.nonzero(stud_residuals > 4)[0]
outliers = np.append(outliers, np.nonzero(stud_residuals<-4)[0])
print("Total Number of Outliers:", outliers.shape[0])
print("\nFirst Fifteen Residuals beyond 4 sigma:")
print("\n*****************************************************")
print(" Case      Observed      Predicted       Stud. Resid.")
```

```
cases = 0
for case in outliers:
        print("{:5d}{:13.2f}{:14.2f}{:17.2f}".
              format(case, y[case], predicted[case], stud_residuals[case]))
        cases += 1
        if cases==15: break
print("")

print("\n*****************************************************")
print("                             Min      Mean        Max")
print("Observed:                  {:10.4f} {:10.4f} {:10.4f}".
      format(y.min(),  y.mean(), y.max()))
print("Predicted:                 {:10.4f} {:10.4f} {:10.4f}".
      format(predicted.min(),  predicted.mean(), predicted.max()))
print("Residuals:                {:10.4f} {:10.4f} {:10.4f}".
      format(residual.min(),  residual.mean(), residual.max()))
print("Standardized Residuals:{:10.4f} {:10.4f} {:10.4f}".
      format(std_residuals.min(),std_residuals.mean(),std_residuals.max()))
print("Studentized  Residuals:{:10.4f} {:10.4f} {:10.4f}".
      format(stud_residuals.min(),stud_residuals.mean(),stud_residuals.max()))
print("Cooks'D:                   {:10.4f} {:10.4f} {:10.4f}".
      format(cooks_d.min(),  cooks_d.mean(), cooks_d.max()))
print("*****************************************************")
# Using MatPlot for residual and influence graphics
# Plot of Observed Values versus the Predicted Values
plt.figure()
plt.xlabel("Predicted " + target)
plt.ylabel("Observed " + target)
plt.plot(predicted, y, "bo")
plt.plot( [8,16], [8,16], "r", linewidth=1, markersize=9)
plt.axis([8, 16, 8, 16])
plt.show()

# Multiplot of 1.  Obs Number vs Studentized Residuals,
#              2.  Predicted Value vs. Studentized Residuals, and
#              3.  Cook's D vs. Studentized Residuals
plt.figure()
plt.subplots(figsize=(16,12))
plt.subplot(311)
plt.xlabel(max_attrib)
plt.ylabel("Studentized Residual")
plt.axis([8,10, -10.0, +10.0])
plt.axhline(0, color="k")
plt.axhline(4, color="r")
plt.axhline(-4, color="r")
plt.plot(df_encoded[max_attrib], stud_residuals, "bo")

plt.subplot(312)
plt.xlabel("Predicted " + target)
plt.ylabel("Studentized Residual")
plt.axis([8, 16, -10.0, +10.0])
#plt.axis([4, 14, -10.0, +10.0])
plt.axhline(0, color="k")
plt.axhline(4, color="r")
plt.axhline(-4, color="r")
plt.plot(predicted, stud_residuals, 'bo')

plt.subplot(313)
plt.xlabel("Cooks'D")
plt.ylabel("Studentized Residual")
plt.axis([0,0.03, -10.0, +10.0])
plt.axhline(0, color="k")
plt.axhline(4, color="r")
plt.axhline(-4, color="r")
plt.axvline(cutoffD, color="g")
plt.plot(cooks_d, stud_residuals, 'bo')
cutoffText = 'Cutoff = (4/n) = {0:.4f}'.format(cutoffD)
plt.text(cutoffD+0.00005, -9, cutoffText, fontsize=12)
plt.show()

# Split the model 70/30 for training/validation
```

```
X_train, X_validate, y_train, y_validate =  \
    train_test_split(Xc, y, test_size=0.3, random_state = 12345)

model    = sm.OLS(y_train, X_train)  # Using StatsModels for Linear Regression
results  = model.fit()
print("\n\n*******************************************************")
print("                              VALIDATION MODEL")
print("                              Target: " + target)
print(results.summary())
print("\n*******************************************************\n")

print("AdvancedAnalytics Display Split Metrics:")
linreg.display_split_metrics(results, X_train,y_train, X_validate,y_validate)
```

5 Appendix – Full list of Python output

```
runfile('C:/Users/rainwater-e/OneDrive - Texas A&M University/Summer-2020/STAT 656 Applied Analytics/hw-02/rainwater-stat656-hw02.py', wdir='C:/Users/rainwater-e/OneDrive - Texas A&M University/Summer-2020/STAT 656 Applied Analytics/hw-02')
   Obs  Log_Cum_Production  Log_Proppant_LB  ...    Y_Well  Operator  County
0    1           12.238153        13.925315  ...  32.63650       1.0    11.0
1    2           12.810446        13.191794  ...  32.81973       1.0    13.0
2    3           11.304855        14.188508  ...  32.72011       1.0    11.0
3    4           12.921434        13.548937  ...  32.77724       1.0    11.0
4    5           11.869739        14.707304  ...  32.88015       1.0    13.0

[5 rows x 14 columns] (4752, 14)
Obs                             int64
Log_Cum_Production            float64
Log_Proppant_LB               float64
Log_Carbonate                 float64
Log_Frac_Fluid_GL             float64
Log_GrossPerforatedInterval   float64
Log_LowerPerforation_xy       float64
Log_UpperPerforation_xy       float64
Log_TotalDepth                float64
N_Stages                      float64
X_Well                        float64
Y_Well                        float64
Operator                      float64
County                        float64
dtype: object

********** Data Preprocessing ***********
Features Dictionary Contains:
11 Interval,
0 Binary,
2 Nominal, and
1 Excluded Attribute(s).

Data contains 4752 observations & 14 columns.


Attribute Counts
................................... Missing  Outliers
Obs.........................         0         0
Log_Cum_Production...........         1         0
Log_Proppant_LB.............         1         0
Log_Carbonate...............         1         0
Log_Frac_Fluid_GL...........         1         0
Log_GrossPerforatedInterval..         1         0
Log_LowerPerforation_xy......         1         0
Log_UpperPerforation_xy......         1         0
Log_TotalDepth..............         1         0
N_Stages....................         1         0
X_Well......................         1         0
Y_Well......................         1         0
Operator....................         1         0
County......................         1         0
Add  Log_LowerPerforation_xy        with p-value 4.056e-295
```

```
Add  County13                   with p-value 5.99877e-92
Add  County9                    with p-value 1.23249e-61
Add  Log_Proppant_LB            with p-value 2.64986e-40
Add  Y_Well                     with p-value 4.22674e-34
Add  X_Well                     with p-value 1.13925e-21
Add  Log_UpperPerforation_xy    with p-value 7.2829e-28
Add  Log_Frac_Fluid_GL          with p-value 4.61247e-13
Add  County3                    with p-value 2.79897e-07
Add  County8                    with p-value 9.39085e-06
Add  County1                    with p-value 0.000374608
Add  County6                    with p-value 0.00154046
Add  Operator15                 with p-value 0.0061388
Add  Log_Carbonate              with p-value 0.00842697
Add  Operator7                  with p-value 0.0177631
Add  Operator5                  with p-value 0.0235112
Add  Operator10                 with p-value 0.0543365
Add  Operator27                 with p-value 0.0599095
Add  County7                    with p-value 0.060518
Add  N_Stages                   with p-value 0.0785451
Add  Operator1                  with p-value 0.0802613
Add  County5                    with p-value 0.0854933
Add  County11                   with p-value 0.0202283
Add  County12                   with p-value 0.00572919
Add  County4                    with p-value 0.0566588
Remove County9                    with p-value 0.160882

Final selected attributes:
 ['Log_LowerPerforation_xy', 'County13', 'Log_Proppant_LB', 'Y_Well', 'X_Well', 'Log_UpperPerforation_xy',
'Log_Frac_Fluid_GL', 'County3', 'County8', 'County1', 'County6', 'Operator15', 'Log_Carbonate', 'Operator7',
'Operator5', 'Operator10', 'Operator27', 'County7', 'N_Stages', 'Operator1', 'County5', 'County11',
'County12', 'County4']


AICC:  7255.867607495069 AIC:  7255.570401737236 BIC:  7423.689272168096

**********************************************************
                                  Target: Log_Cum_Production
                           OLS Regression Results
==============================================================================
Dep. Variable:                      y   R-squared:                       0.440
Model:                            OLS   Adj. R-squared:                  0.437
Method:                 Least Squares   F-statistic:                     154.8
Date:                Wed, 03 Jun 2020   Prob (F-statistic):               0.00
Time:                        22:44:45   Log-Likelihood:                -3601.8
No. Observations:                4751   AIC:                             7254.
Df Residuals:                    4726   BIC:                             7415.
Df Model:                          24
Covariance Type:            nonrobust
===========================================================================================
                              coef    std err          t      P>|t|      [0.025      0.975]
-------------------------------------------------------------------------------------------
const                     116.6109      9.358     12.462      0.000      98.266     134.956
Log_LowerPerforation_xy     1.6930      0.099     17.180      0.000       1.500       1.886
County13                    0.1606      0.024      6.778      0.000       0.114       0.207
Log_Proppant_LB             0.1310      0.011     11.959      0.000       0.110       0.152
Y_Well                      0.2392      0.044      5.491      0.000       0.154       0.325
X_Well                      1.2009      0.090     13.305      0.000       1.024       1.378
Log_UpperPerforation_xy    -1.4815      0.157     -9.461      0.000      -1.788      -1.174
Log_Frac_Fluid_GL           0.0693      0.011      6.600      0.000       0.049       0.090
County3                    -0.2896      0.038     -7.627      0.000      -0.364      -0.215
County8                    -0.3264      0.070     -4.687      0.000      -0.463      -0.190
County1                    -1.0569      0.166     -6.348      0.000      -1.383      -0.731
County6                    -0.6049      0.050    -12.104      0.000      -0.703      -0.507
Operator15                  0.3846      0.134      2.863      0.004       0.121       0.648
Log_Carbonate               0.0219      0.009      2.510      0.012       0.005       0.039
Operator7                   0.0909      0.037      2.466      0.014       0.019       0.163
Operator5                  -0.1069      0.055     -1.955      0.051      -0.214       0.000
Operator10                  0.0937      0.049      1.912      0.056      -0.002       0.190
Operator27                  0.1027      0.056      1.837      0.066      -0.007       0.212
County7                    -0.3529      0.034    -10.373      0.000      -0.420      -0.286
N_Stages                    0.0043      0.002      1.765      0.078      -0.000       0.009
```

```
Operator1                 -0.1733      0.098     -1.762      0.078      -0.366       0.020
County5                   -0.3950      0.072     -5.466      0.000      -0.537      -0.253
County11                  -0.2106      0.029     -7.229      0.000      -0.268      -0.153
County12                  -0.3974      0.081     -4.908      0.000      -0.556      -0.239
County4                   -0.4001      0.158     -2.537      0.011      -0.709      -0.091
==============================================================================
Omnibus:                      652.466   Durbin-Watson:                   1.905
Prob(Omnibus):                  0.000   Jarque-Bera (JB):             1446.021
Skew:                          -0.815   Prob(JB):                         0.00
Kurtosis:                       5.157   Cond. No.                     1.32e+05
==============================================================================

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 1.32e+05. This might indicate that there are
strong multicollinearity or other numerical problems.

***********************************************************

Strongest Attribute: Log_LowerPerforation_xy

Expected number of observations outside stated limits:
-+ 3Sigma:  13
-+ 4Sigma:  0
-+ 5Sigma:  0


Max Cooks D:   0.0269  Cutoff ( 4/n):  0.00084
Max H:         0.1071  Cutoff (2p/n):  0.01095

Residuals beyond 4 sigma:
Total Number of Outliers: 13

First Fifteen Residuals beyond 4 sigma:

*******************************************************
 Case      Observed       Predicted       Stud. Resid.
   92         9.51          12.41             -4.01
 1048         9.54          12.50             -5.63
 1693         9.82          12.84             -4.63
 2454        10.69          12.53             -4.25
 2977        10.73          12.63             -4.06
 3107         9.37          12.62             -6.98
 3498         9.80          12.41             -5.19
 3520         9.69          13.22             -4.87
 4049         9.82          12.27             -4.06
 4076         8.80          13.01             -5.65
 4337         9.93          12.88             -4.84
 4698         9.78          12.86             -4.72
 4711        10.78          12.30             -4.38


*******************************************************
                              Min       Mean        Max
Observed:                   8.7986    12.6006    14.3829
Predicted:                 10.4162    12.6006    13.9363
Residuals:                 -3.6092    -0.0000    2.0023
Standardized Residuals:    -6.9704    -0.0000     3.8670
Studentized  Residuals:    -6.9772     0.0000     3.9162
Cooks'D:                    0.0000     0.0002     0.0269
*******************************************************
<Figure size 432x288 with 0 Axes>


*******************************************************
                              TRAINING MODEL
                              Target: Log_Cum_Production
                            OLS Regression Results
==============================================================================
Dep. Variable:                      y   R-squared:                       0.447
Model:                            OLS   Adj. R-squared:                  0.443
```

```
Method:                 Least Squares   F-statistic:                     111.1
Date:                Wed, 03 Jun 2020   Prob (F-statistic):               0.00
Time:                        22:44:45   Log-Likelihood:                -2471.3
No. Observations:                3325   AIC:                             4993.
Df Residuals:                    3300   BIC:                             5145.
Df Model:                          24
Covariance Type:            nonrobust
=============================================================================================
                              coef    std err          t      P>|t|      [0.025      0.975]
---------------------------------------------------------------------------------------------
const                     107.3843     10.823      9.921      0.000      86.163     128.606
Log_LowerPerforation_xy     1.5212      0.115     13.240      0.000       1.296       1.746
County13                    0.1701      0.028      6.087      0.000       0.115       0.225
Log_Proppant_LB             0.1478      0.013     11.426      0.000       0.122       0.173
Y_Well                      0.2146      0.051      4.202      0.000       0.114       0.315
X_Well                      1.1109      0.104     10.632      0.000       0.906       1.316
Log_UpperPerforation_xy    -1.1702      0.182     -6.422      0.000      -1.527      -0.813
Log_Frac_Fluid_GL           0.0558      0.012      4.802      0.000       0.033       0.079
County3                    -0.2793      0.045     -6.230      0.000      -0.367      -0.191
County8                    -0.3925      0.081     -4.844      0.000      -0.551      -0.234
County1                    -0.9365      0.184     -5.084      0.000      -1.298      -0.575
County6                    -0.6195      0.059    -10.562      0.000      -0.734      -0.504
Operator15                  0.3916      0.148      2.639      0.008       0.101       0.683
Log_Carbonate               0.0302      0.010      2.908      0.004       0.010       0.051
Operator7                   0.1256      0.043      2.896      0.004       0.041       0.211
Operator5                  -0.0939      0.066     -1.417      0.157      -0.224       0.036
Operator10                  0.1186      0.058      2.044      0.041       0.005       0.232
Operator27                  0.1199      0.067      1.783      0.075      -0.012       0.252
County7                    -0.3506      0.040     -8.680      0.000      -0.430      -0.271
N_Stages                    0.0047      0.003      1.668      0.096      -0.001       0.010
Operator1                  -0.1262      0.115     -1.099      0.272      -0.351       0.099
County5                    -0.3633      0.088     -4.143      0.000      -0.535      -0.191
County11                   -0.2179      0.034     -6.393      0.000      -0.285      -0.151
County12                   -0.3273      0.090     -3.654      0.000      -0.503      -0.152
County4                    -0.3461      0.210     -1.647      0.100      -0.758       0.066
==============================================================================
Omnibus:                      325.925   Durbin-Watson:                   1.938
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              536.452
Skew:                          -0.701   Prob(JB):                    3.24e-117
Kurtosis:                       4.380   Cond. No.                     1.29e+05
==============================================================================

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 1.29e+05. This might indicate that there are
strong multicollinearity or other numerical problems.

*******************************************************

AdvancedAnalytics Display Split Metrics:


Model Metrics..........        Training     Validation
Observations...........            3325           1426
Coefficients...........              26             26
DF Error...............            3299           1400
R-Squared..............          0.4469         0.4203
Adj. R-Squared.........          0.4427         0.4100
Mean Absolute Error....          0.3911         0.4014
Median Absolute Error..          0.3167         0.3166
Avg Squared Error......          0.2589         0.2873
Square Root ASE........          0.5088         0.5360
Log Likelihood.........      -2471.2664     -1134.1419
AIC         .........         4996.5328      2322.2838
AICc        .........         4996.9914      2323.3654
BIC         .........         5161.4818      2464.3748

*******************************************************
                          Min        Mean        Max
Observed:              8.7986     12.6006     14.3829
Predicted:            10.4162     12.6006     13.9363
```

```
Residuals:                  -3.6092    -0.0000     2.0023
Studentized Residuals:      -6.9772     0.0000     3.9162
Cooks'D:                     0.0000     0.0002     0.0269
*****************************************************
```

```
         Top 10 Log_Cum_Production
 Obs.   Observed    Predicted
 2741    14.38        13.37
 3362    14.33        11.89
  447    14.33        13.01
  407    14.31        12.29
   85    14.29        12.76
 1065    14.25        12.78
   52    14.18        12.69
 2745    14.16        12.88
 4307    14.14        12.80
 4067    14.14        11.99
```