**DSDP 2020: Homework Regression**

Please enter your responses below. Submissions are due prior to the start of the next lecture.

1. Enter your CAI: LPZQ

Required to answer

2. Regression analysis can only be used for prediction.

*TRUE OR FALSE*

3. Residuals are the difference between the observed values of y and the fitted values of y.

*TRUE OR FALSE*

Using the Facebook\_Post\_Prediction\_Truncated.xlsx dataset, create a multivariate linear regression model using the post comment counts, post length, how many times the post has been shared, and if the post was made on Saturday as compared to the rest of the week to predict how many likes the Facebook post will receive in the first 24 hours.

1. What is the overall model p-value? 1.19e-151
2. Is the model statistically significant at the p<0.05 level? YES OR NO
3. Are there any variables in the model that are not statistically significant with the outcome after adjusting for the other variables? If so, which ones?

PostMonday, PostTuesday, PostWednesday

1. Write a sentence that interprets the beta coefficient for the number of times a post has been shared.

We estimate that there is a 0.95% increase in Like for every additional post on FB adjusting for post comment counts, post length, and if the post was made on Saturday

1. Write a sentence that interprets the beta coefficient for the post being made on a Saturday.

We estimate that there is a 5.9785 likes increase for every additional post on Saturday adjusting for post comment counts, post length, and how many times the post has been shared

1. Given that a post has 10 comments in the first 24 hours, was 148 characters long, was shared by 3 people and was made on a Friday, how many likes can we expect that post to have in the first 24 hours?

10.288 likes

1. Run the diagnostic tests we discussed in class on the above multivariate model. Explain your findings.

VIF < 10 : no Multicollinearity

Durbin-Watson = 1.7156273756937015, no Autocorrelation

1. Test to see if the statistical assumptions for a linear model have been met. Explain your findings.

Got NaN on ANOVA

1. Perform stepwise regression with the above model to determine the optimal combination of variables to predict the number of Facebook post likes in the first 24 hours. What model is the best fit?

Model Selected

NumberLikes24 ~ CommentCount24 + PostShareCount + PostSunday + PostSaturday + PostThursday + 1

R2= 0.1851637683607994

1. Provide the code you used to run the analysis above.

#Stepwise Regression', 'PostShareCount','PostSaturday','Pos

DataSubset = pd.DataFrame(FB, columns=['NumberLikes24', 'CommentCount24','PostLength', 'PostShareCount','PostSaturday','PostSunday','PostThursday'])

def forward\_selected(data, response):

"""Linear model designed by forward selection.

Parameters:

-----------

data : pandas DataFrame with all possible predictors and response

response: string, name of response column in data

Returns:

--------

model: an "optimal" fitted statsmodels linear model

with an intercept

selected by forward selection

evaluated by adjusted R-squared

"""

remaining = set(data.columns)

remaining.remove(response)

selected = []

current\_score, best\_new\_score = 0.0, 0.0

while remaining and current\_score == best\_new\_score:

scores\_with\_candidates = []

for candidate in remaining:

formula = "{} ~ {} + 1".format(response,

' + '.join(selected + [candidate]))

score = smf.ols(formula, data).fit().rsquared\_adj

scores\_with\_candidates.append((score, candidate))

scores\_with\_candidates.sort()

best\_new\_score, best\_candidate = scores\_with\_candidates.pop()

if current\_score < best\_new\_score:

remaining.remove(best\_candidate)

selected.append(best\_candidate)

current\_score = best\_new\_score

formula = "{} ~ {} + 1".format(response,

' + '.join(selected))

model = smf.ols(formula, data).fit()

return model

model = forward\_selected(DataSubset, 'NumberLikes24')

print('Model Selected')

print(model.model.formula)

print('R2=', model.rsquared\_adj)