ISE-529 Predictive Analytics

Mid-Term Examination – July 25, 2022

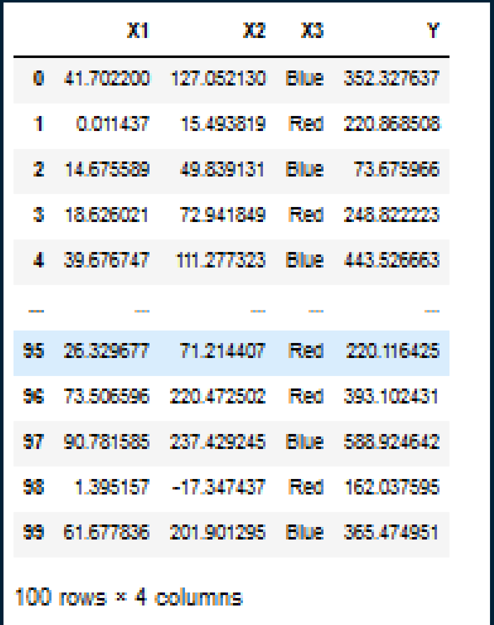
Instructions

- You are to complete the exam by typing your answers into this PowerPoint as indicated.
- You will have 90 minutes to complete the exam and submit it to GradeScope (in the same manner as done for homework assignments). Late submissions will be penalized.
- The exam is open-book / open-notes. You may consult any resource except another person.
- Good luck!

Problem 1

Linear Model Analysis

For this problem we will be working with the following dataset:

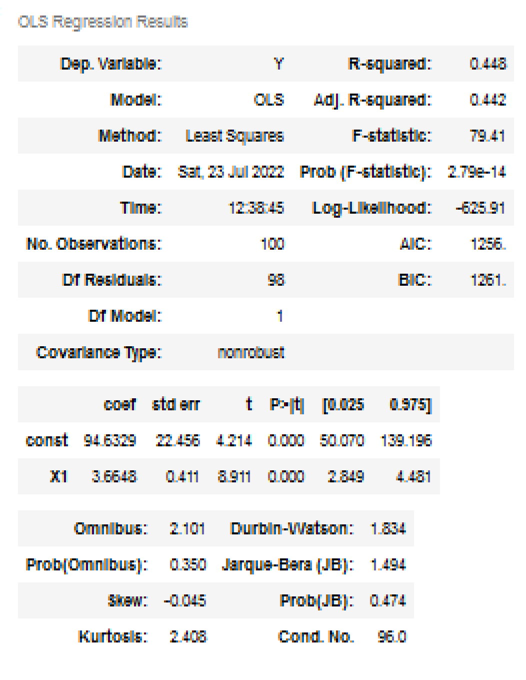


	X1	X2	X3	Y
0	41.702200	127.052130	Blue	352.327637
1	0.011437	15.493819	Red	220.868508
2	14.675589	49.839131	Blue	73.675966
3	18.626021	72.941849	Red	248.822223
4	39.676747	111.277323	Blue	443.526663
...	...	...	...	...
95	26.329677	71.214407	Red	220.116425
96	73.506596	220.472502	Red	393.102431
97	90.781585	237.429245	Blue	588.924642
98	1.395157	-17.347437	Red	162.037595
99	61.677836	201.901295	Blue	365.474951

100 rows × 4 columns

Problem 1

First, we create three models using X1, X2, and the combination of X1 & X2 to predict Y:

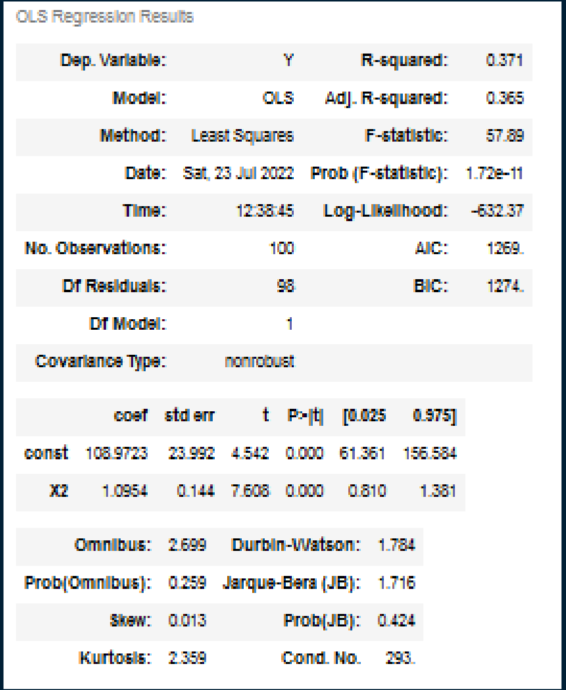


OLS Regression Results

Dep. Variable:	Y	R-squared:	0.448
Model:	OLS	Adj. R-squared:	0.442
Method:	Least Squares	F-statistic:	79.41
Date:	Sat, 23 Jul 2022	Prob (F-statistic):	2.79e-14
Time:	12:38:45	Log-Likelihood:	-625.91
No. Observations:	100	AIC:	1256.
Df Residuals:	98	BIC:	1261.
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	94.6329	22.456	4.214	0.000	50.070	139.196
X1	3.6648	0.411	8.911	0.000	2.849	4.481

Omnibus:	2.101	Durbin-Watson:	1.834
Prob(Omnibus):	0.350	Jarque-Bera (JB):	1.494
Skew:	-0.045	Prob(JB):	0.474
Kurtosis:	2.408	Cond. No.	96.0

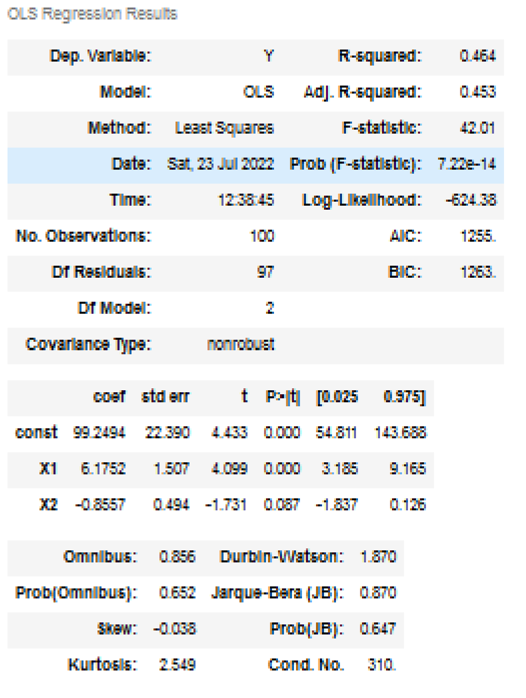


OLS Regression Results

Dep. Variable:	Y	R-squared:	0.371
Model:	OLS	Adj. R-squared:	0.365
Method:	Least Squares	F-statistic:	57.89
Date:	Sat, 23 Jul 2022	Prob (F-statistic):	1.72e-11
Time:	12:38:45	Log-Likelihood:	-632.37
No. Observations:	100	AIC:	1269.
Df Residuals:	98	BIC:	1274.
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	108.9723	23.992	4.542	0.000	61.361	156.584
X2	1.0954	0.144	7.608	0.000	0.810	1.381

Omnibus:	2.699	Durbin-Watson:	1.784
Prob(Omnibus):	0.259	Jarque-Bera (JB):	1.716
Skew:	0.013	Prob(JB):	0.424
Kurtosis:	2.359	Cond. No.	293.



OLS Regression Results

Dep. Variable:	Y	R-squared:	0.464
Model:	OLS	Adj. R-squared:	0.453
Method:	Least Squares	F-statistic:	42.01
Date:	Sat, 23 Jul 2022	Prob (F-statistic):	7.22e-14
Time:	12:38:45	Log-Likelihood:	-624.38
No. Observations:	100	AIC:	1255.
Df Residuals:	97	BIC:	1263.
Df Model:	2		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	99.2494	22.390	4.433	0.000	54.811	143.688
X1	6.1752	1.507	4.099	0.000	3.185	9.165
X2	-0.8557	0.494	-1.731	0.087	-1.837	0.126

Omnibus:	0.856	Durbin-Watson:	1.870
Prob(Omnibus):	0.652	Jarque-Bera (JB):	0.870
Skew:	-0.038	Prob(JB):	0.647
Kurtosis:	2.549	Cond. No.	310.

Problem 1

1A) For the two simple (single-predictor) models, are the predictors X1 & X2 significant?

- Yes

1B) For the multiple regression model, which predictors are significant?

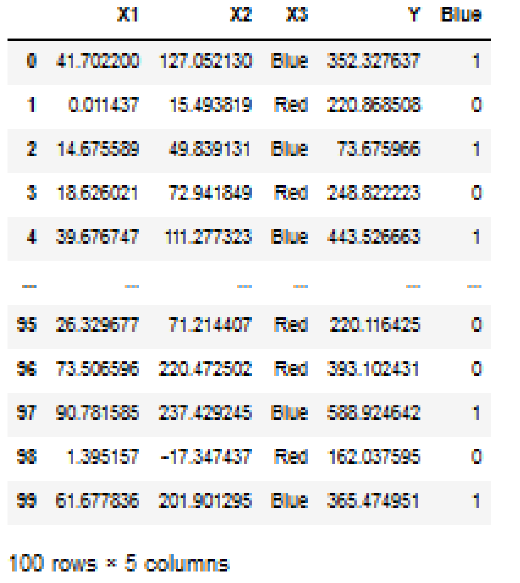
- Only X1

1C) How do you interpret what is going on here?

- It is likely that X1 and X2 are correlated and that X1 is the better predictor. Once the information from X1 is incorporated into the model, X2 is no longer significant

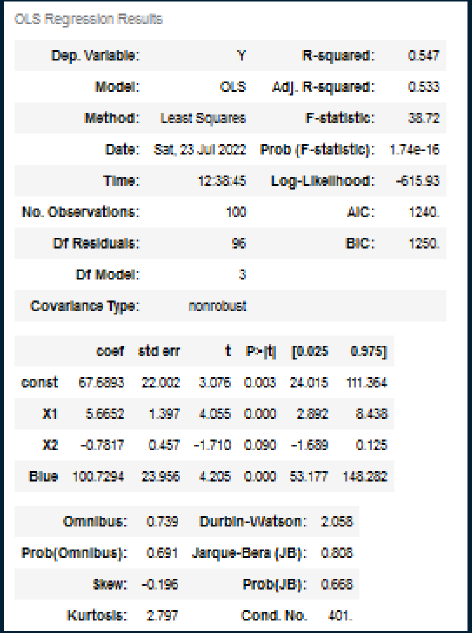
Problem 1

Now we incorporate the categorical variable into the model by creating a dummy variable “Blue” and incorporate it into the model as shown:



	X1	X2	X3	Y	Blue
0	41.702200	127.052130	Blue	352.327637	1
1	0.011437	15.493819	Red	220.868508	0
2	14.675589	49.839131	Blue	73.675966	1
3	18.626021	72.941849	Red	248.822223	0
4	39.676747	111.277323	Blue	443.526663	1
...	...	...	...	...	...
95	26.329677	71.214407	Red	220.116425	0
96	73.506596	220.472502	Red	393.102431	0
97	90.781585	237.429245	Blue	588.924642	1
98	1.395157	-17.347437	Red	162.037595	0
99	61.677836	201.901295	Blue	365.474951	1

100 rows × 5 columns



OLS Regression Results

Dep. Variable:	Y	R-squared:	0.547
Model:	OLS	Adj. R-squared:	0.533
Method:	Least Squares	F-statistic:	38.72
Date:	Sat, 23 Jul 2022	Prob (F-statistic):	1.74e-16
Time:	12:38:45	Log-Likelihood:	-615.93
No. Observations:	100	AIC:	1240.
Df Residuals:	96	BIC:	1250.
Df Model:	3		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	67.6893	22.002	3.076	0.003	24.015	111.364
X1	5.6652	1.397	4.055	0.000	2.892	8.438
X2	-0.7817	0.457	-1.710	0.090	-1.689	0.125
Blue	100.7294	23.956	4.205	0.000	53.177	148.282

Omnibus:	0.739	Durbin-Watson:	2.058
Prob(Omnibus):	0.691	Jarque-Bera (JB):	0.808
Skew:	-0.196	Prob(JB):	0.668
Kurtosis:	2.797	Cond. No.	401.

Problem 1

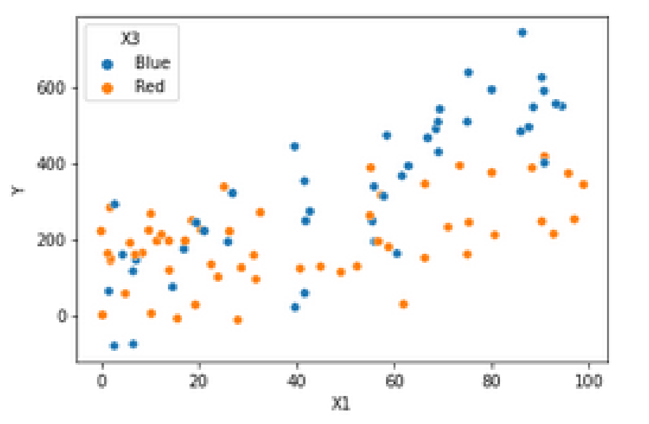
1D) Does adding this categorical variable to the model improve it’s overall performance? Why or why not?

- Yes, we see that the model R^2 has improved from 0.464 to 0.547

Problem 1

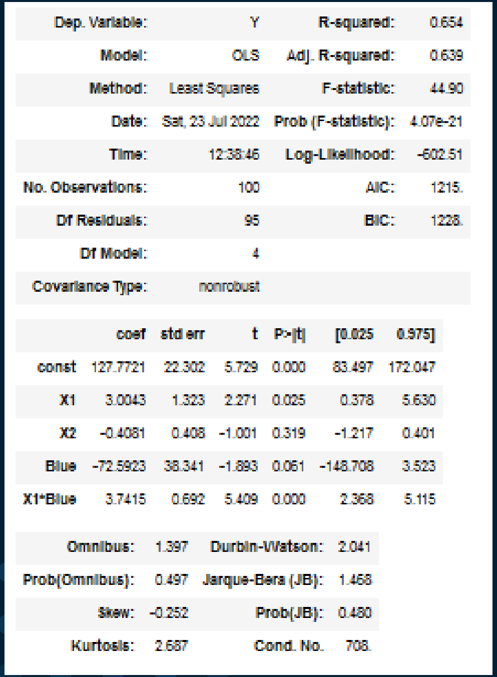
1E) Looking at this color-coded scatterplot of X1 vs Y, do you see any indication of an interaction effect between X1 and X3? Why or why not?

- Yes, it appears that the slop of the lines for the blue and red observations may be different



Problem 1

1E) Looking at these model results, do you see any indication of an interaction effect between X1 and X3? Why or why not?

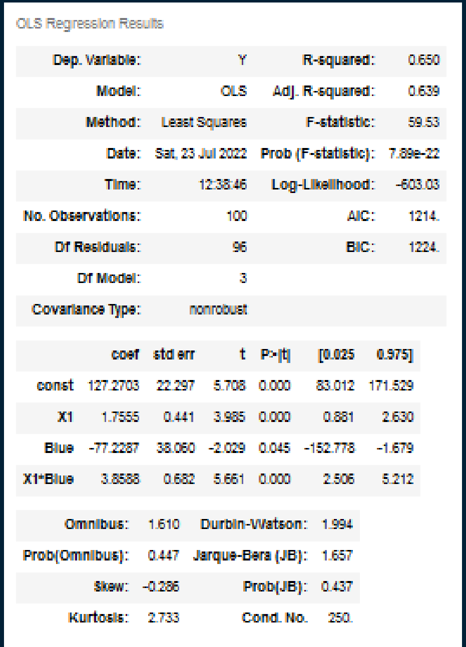


```
        Dep. Variable:               Y          R-squared:     0.654
                Model:             OLS      Adj. R-squared:    0.639
               Method:   Least Squares         F-statistic:    44.90
                 Date: Sat, 23 Jul 2022  Prob (F-statistic): 4.07e-21
                 Time:        12:38:46      Log-Likelihood:  -602.51
    No. Observations:             100                 AIC:     1215.
        Df Residuals:              95                 BIC:     1228.
            Df Model:               4
     Covariance Type:       nonrobust

              coef   std err       t   P>|t|    [0.025    0.975]
   const  127.7721    22.302   5.729   0.000    83.497   172.047
      X1    3.0043     1.323   2.271   0.025     0.378     5.630
      X2   -0.4081     0.408  -1.001   0.319    -1.217     0.401
    Blue  -72.5923    38.341  -1.893   0.061  -148.708     3.523
 X1*Blue    3.7415     0.692   5.409   0.000     2.368     5.115

       Omnibus:   1.397     Durbin-Watson:  2.041
 Prob(Omnibus):   0.497  Jarque-Bera (JB):  1.468
          Skew:  -0.252          Prob(JB):  0.480
      Kurtosis:   2.687          Cond. No.   708.
```

- Yes, the p-value for the interaction term is less than 0.05

Problem 1

After completing your modeling analysis, you decide to use the model shown below:



OLS Regression Results

Dep. Variable:	Y	R-squared:	0.650
Model:	OLS	Adj. R-squared:	0.639
Method:	Least Squares	F-statistic:	59.53
Date:	Sat, 23 Jul 2022	Prob (F-statistic):	7.89e-22
Time:	12:38:46	Log-Likelihood:	-603.03
No. Observations:	100	AIC:	1214.
Df Residuals:	96	BIC:	1224.
Df Model:	3		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	127.2703	22.297	5.708	0.000	83.012	171.529
X1	1.7555	0.441	3.985	0.000	0.881	2.630
Blue	-77.2287	38.060	-2.029	0.045	-152.778	-1.679
X1*Blue	3.8588	0.682	5.661	0.000	2.506	5.212

Omnibus:	1.610	Durbin-Watson:	1.994
Prob(Omnibus):	0.447	Jarque-Bera (JB):	1.657
Skew:	-0.286	Prob(JB):	0.437
Kurtosis:	2.733	Cond. No.	250.

Problem 1

1F) Write out the algebraic expression for this model (you do not need to include the error term):

- $Y = 127.3 + 1.76X1 \ - 77.2 * Blue + 3.86(X1 * Blue)$

1G) Write out the simplified algebraic expression for this model for the Blue observations

- Y = 127.3 + 1.76X1 – 77.2 + 3.3X1
- Y = 50.1 + 5.7X1

1H) Write out the simplified algebraic expression for this model for the Red observations

- Y = 127.3 + 1.76X1

Problem 2

2) We have developed a model to predict the sales (in thousands of dollars) at a new store our company may decide to open in a new city and we define and fit a model with five predictors:

- X_P: Population of the city (in thousands of people)
- X_I: Average income of the city (in thousands of dollars per adult)
- X_T: Type of store (1 for downtown store, 0 for a mall store)
- X_{PI}: Interaction between population and average income (in thousands)
- X_{IT}: Interaction between average income (in thousands) and store type

In the cities we are evaluating, the average income is generally less than $100,000 and the cities are in the size range of 0 – 500,000 people

After fitting this model using a linear regression, we get the following coefficients: $\hat{\beta}_0 = 10$, $\hat{\beta}_P = 20$, $\hat{\beta}_I = 50$, $\hat{\beta}_T = 350$, $\hat{\beta}_{PI} = 0.05$, $\hat{\beta}_{IT} =$ - 5

Problem 2

2a) Which answer is correct:

a) For a fixed value of population and average income, a downtown store would on average have greater sales than a mall store

b) For a fixed value of population and average income, a mall store would on average have greater sales than a downtown store

c) For a fixed value of population and average income, a downtown store would on average have more sales than a mall store provided that the average income is high enough

d) For a fixed value of population and average income, a mall store would on average have more sales than a downtown store provided that the average income is high enough

Response: D

Problem 2

2B) What is the predicted sales for a downtown store in a city with a population of 100,000 and an average income of $50,000?

- $4,860K

2C) Is this statement true or false and why: “Since the coefficient of the interaction term between population and average income is very small, there is very little evidence of an interaction effect:

- False. The scale of the interaction variable is much larger than the scales of the other predictors.

Problem 2

2D) Which predictor has the larger impact on sales, income or city population? Explain your answer

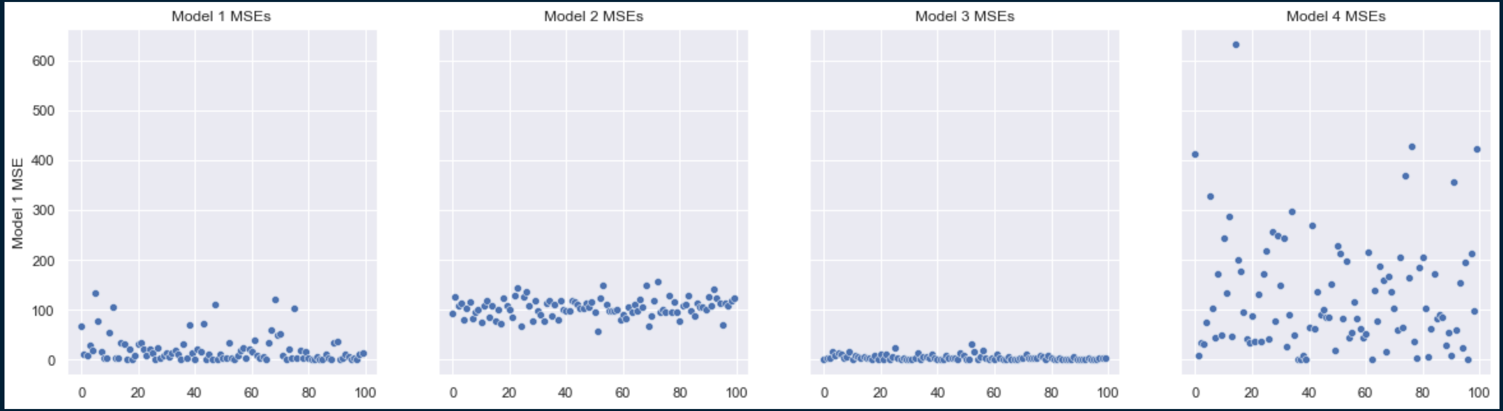
- The population variable has a scale that is 5 times greater than the income variable. Thus, after scaling, the effect of population is approximately twice as large as income.

Problem 3

You are assessing four candidate models (M1 through M4). You try training the models ten different times with different population samples and then assessing those models against test partitions by calculating their mean squared errors (MSE). The results of those tests are summarized on the following page.

Complete the figure on the bottom of the following page with one model for each of the four boxes.

Problem 3



	Low Variance	High Variance
Low Bias	M3	M1
High Bias	M2	M4

Problem 4

4A) Explain in your own words how k-fold cross-validation is implemented

- The dataset is divided into K equal-sized parts
- One at a time, one of the parts is held out for validation and the other K-1 are used for training
- The overall assessment is the weighted average of each of the individual assessments

Problem 4

4B) Provide one advantage and one disadvantage of k-fold cross validation relative to:

- The validation set approach?
  - Advantage: K-fold CV uses more of the available data for training (with decreased bias)
  - Disadvantage: K-fold CV is more computationally expensive
- Leave-Out-One-Cross-Validation?
  - Advantage:
    - K-fold CV is less computationally expnsive
    - K-fold CV has lower model variance (because each training iteration with LOOCV uses almost the same dataset
  - Disadvantage: K-fold CV has greater bias due to using less data for training

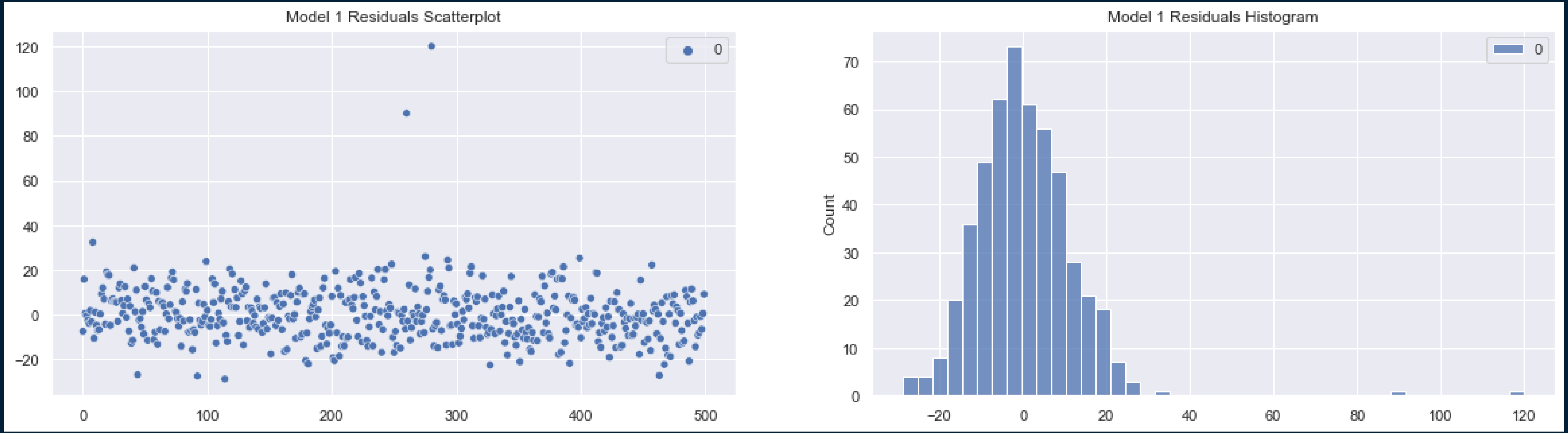
Problem 5

Residuals Analysis

The following pages present a residuals diagram and a residuals histogram for each of six different models. For each model, identify the apparent problem(s) with the model and provide one technique that you might use to remediate (correct) the problem.

5A – Model 1

Residuals Analysis

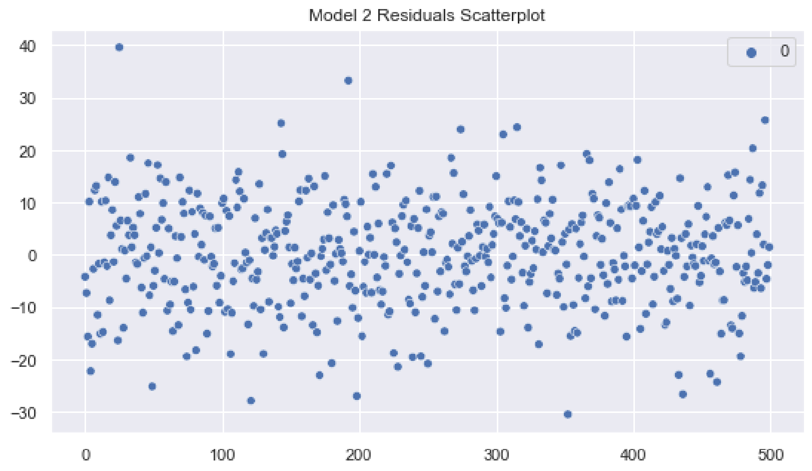


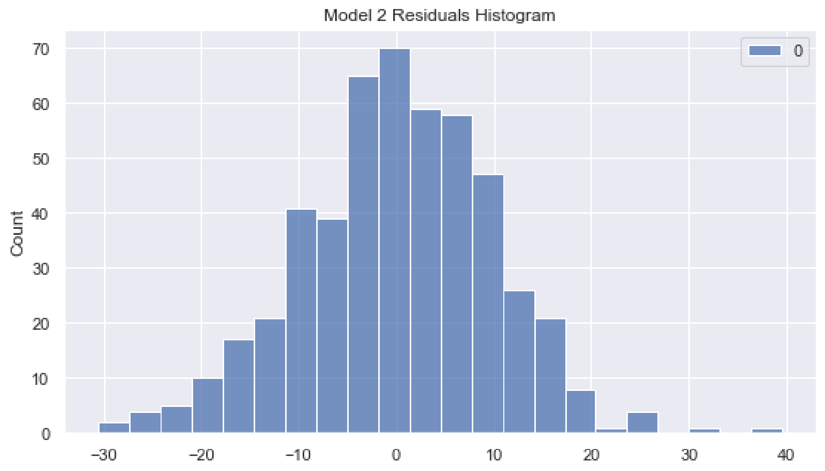
Model Issue: Model has outliers

Possible remediation: Remove outliers or otherwise deal with them

5A – Model 2

Residuals Analysis



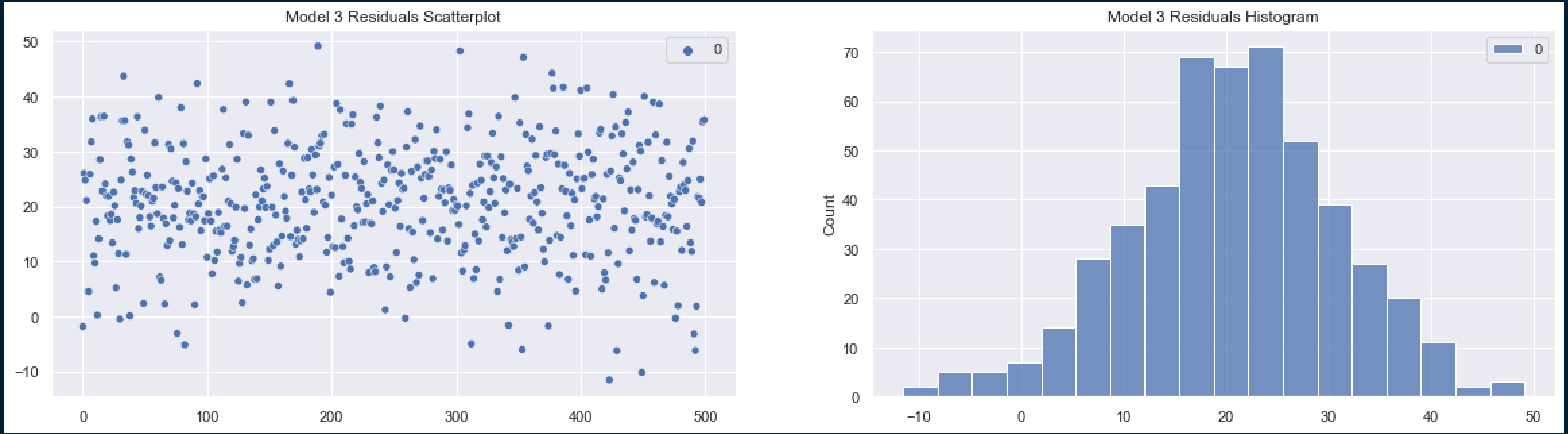


Model Issue: No apparent issues

Possible remediation: N/A

5A – Model 3

Residuals Analysis

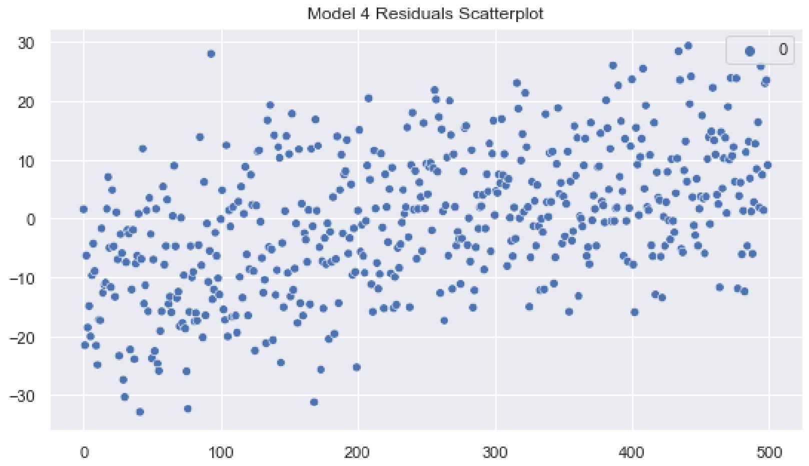


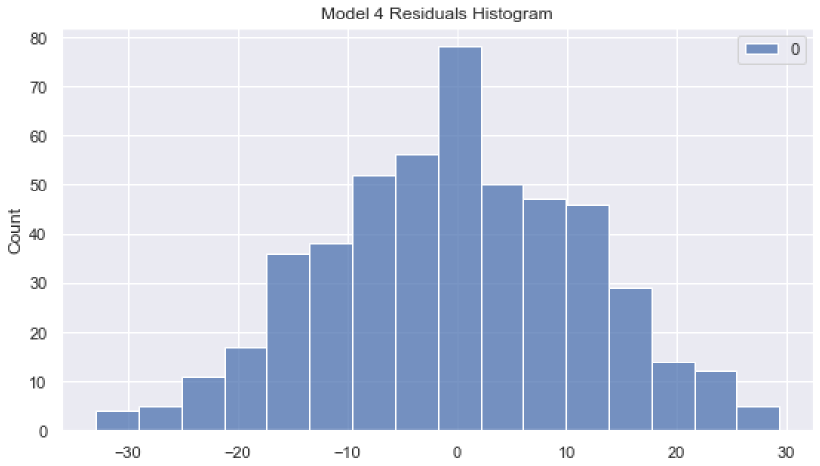
Model Issue: Model has a positive bias term. The residuals appear to be centered at around 20 instead of 0

Possible remediation: Include missing variable or perform transformation

5A – Model 4

Residuals Analysis



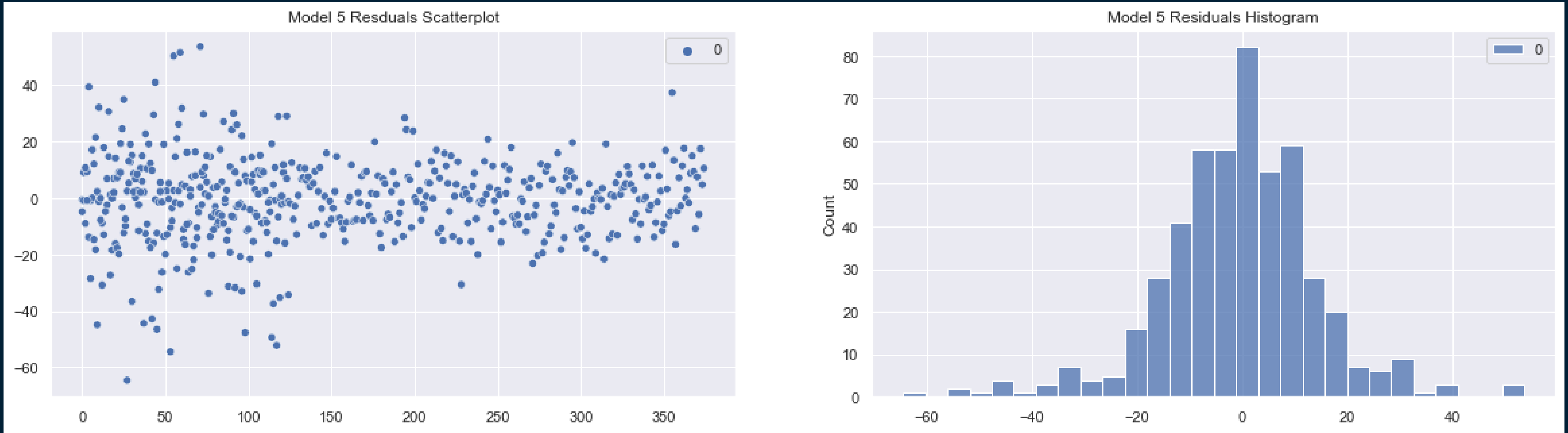


Model Issue: Residuals are not independent (pattern to the data)

Possible remediation: Include missing variable or perform transformation

5A – Model 5

Residuals Analysis

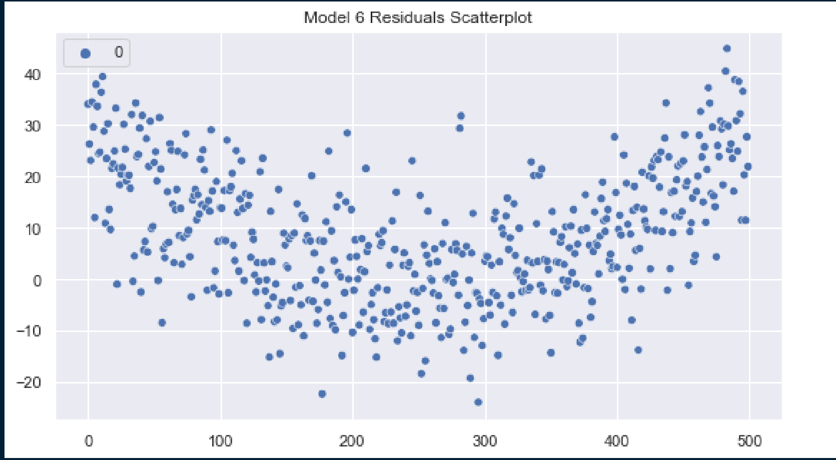


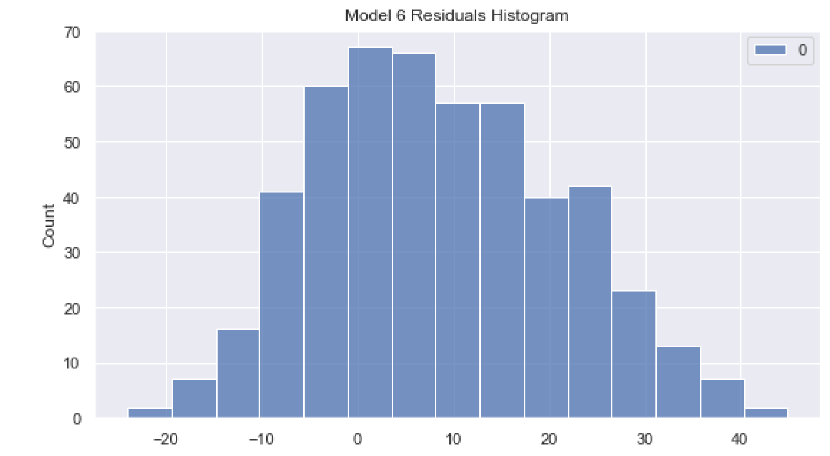
Model Issue: Heteroscadicity

Possible remediation: Perform transformation

5A – Model 6

Residuals Analysis


Model 6 Residuals Scatterplot


Model 6 Residuals Histogram

Model Issue: Residuals are not independent (pattern to the data)

Possible remediation: Include missing variable or perform transformation