Classifiers:
Logistic Regression

Affairs dataset

- 1969 study by *Psychology Today*
  - N = 601

- Original dataset 9 variables
  - Affair (number of times in the past year)
  - Gender
  - Age
  - Yrs Married
  - Children (binary)
  - Religiousness
  - Education
  - Occupation
  - Marriage Satisfaction

affairs
numeric. How often engaged in extramarital sexual intercourse during the past year? 0 = none, 1 = once, 2 = twice, 3 = 3 times, 7 = 4–10 times, 12 = monthly, 12 = weekly, 12 = daily.
age
numeric variable coding age in years: 17.5 = under 20, 22 = 20–24, 27 = 25–29, 32 = 30–34, 37 = 35–39, 42 = 40–44, 47 = 45–49, 52 = 50–54, 57 = 55 or over.
yearsmarried
numeric variable coding number of years married: 0.125 = 3 months or less, 0.417 = 4–6 months, 0.75 = 6 months–1 year, 1.5 = 1–2 years, 4 = 3–5 years, 7 = 6–8 years, 10 = 9–11 years, 15 = 12 or more years.
religiousness
numeric variable coding religiousness: 1 = anti, 2 = not at all, 3 = slightly, 4 = somewhat, 5 = very.
rating
numeric variable coding self rating of marriage: 1 = very unhappy, 2 = somewhat unhappy, 3 = average, 4 = happier than average, 5 = very happy.

What are we trying to do with this dataset?

Looking at the Outcome

- How often have you engaged in extramarital sexual intercourse during the past year?
  - 0 = None
  - 1 = Once
  - 2 = Twice
  - 3 = Three times
  - 7 = 4-10 times
  - 12 = Monthly/Weekly/Daily
- Continuous
  - Right?

How often have you engaged in extramarital sexual intercourse during the past year?

Freq of affair	Count of affairs	Percent of Total
0	451	75.04
1	34	5.66
2	17	2.83
3	19	3.16
7	42	6.99
12	38	6.32
Grand Total	**601**	

Looks like we will likely violate that homoscedacity assumption

Assumptions

Linear Regression

- Pre-model
  - Linearity
  - Independence of obs
  - Homoscedacity
  - No Multicollinearity
  - No “extreme cases”
- Post-model
  - Normality of residual

Logistic Regression

- Pre-model
  - Outcome is binary
  - Independence of obs
  - Large sample
  - No Multicollinearity
  - No “extreme cases”
- Post-model
  - Linearity between predictors and logit of the outcome

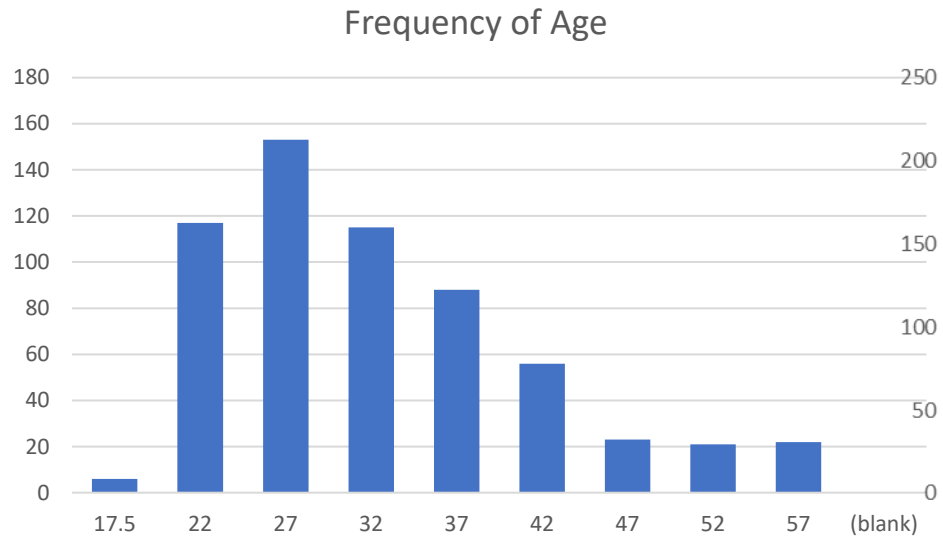
Logistic Regression Assumptions

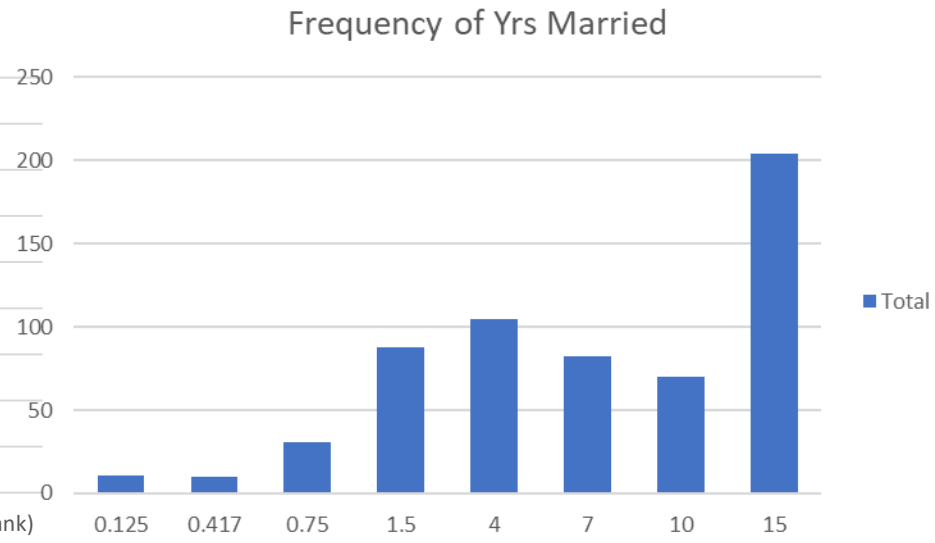
- Pre-model
  - Outcome is binary
  - Independence of obs
  - Large sample
  - No Multicollinearity
  - No “extreme cases”
- Post-model
  - Linearity between predictors and logit of the outcome

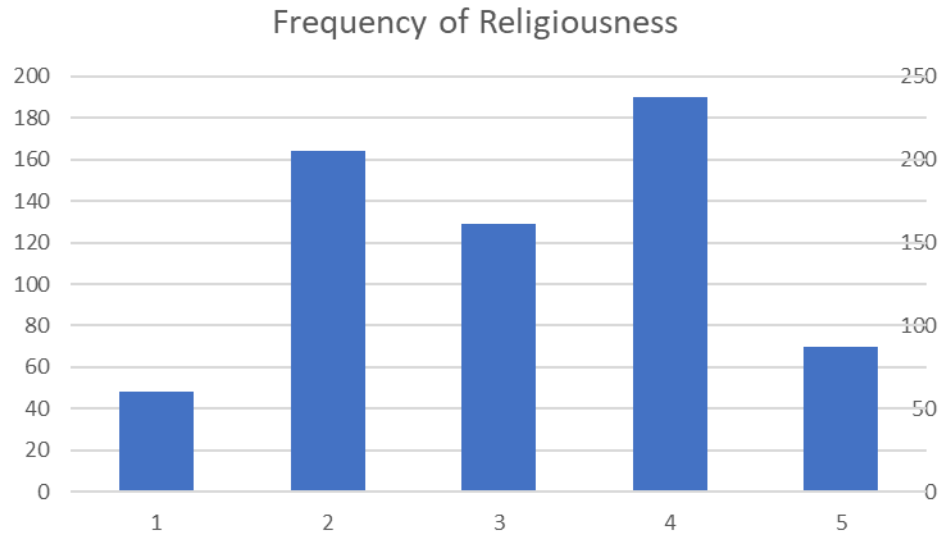
Multicollinearity?

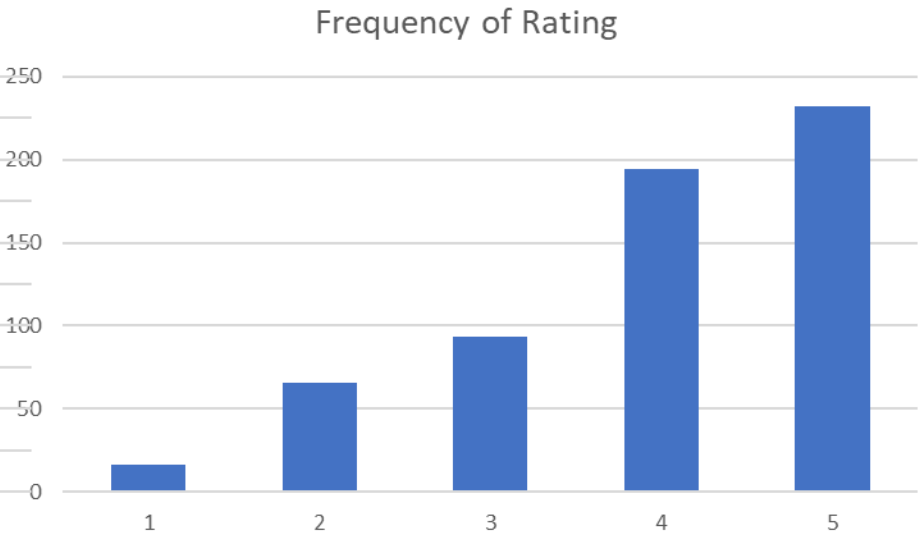
	Age	Yrs Married	Religiousness	Rating
Age	1.00			
Yrs Married	0.78	1.00		
Religiousness	0.19	0.22	1.00	
Rating	-0.19	-0.24	0.02	1.00

We have a flag here.


Frequency of Age
180
160
140
120
100
80
60
40
20
0
17.5
22
27
32
37
42
47
52
57
(blank)


Frequency of Yrs Married
250
200
150
100
50
0
0.125
0.417
0.75
1.5
4
7
10
15
Total


Frequency of Religiousness
200
180
160
140
120
100
80
60
40
20
0
1
2
3
4
5


Frequency of Rating
250
200
150
100
50
0
1
2
3
4
5

Seems okay

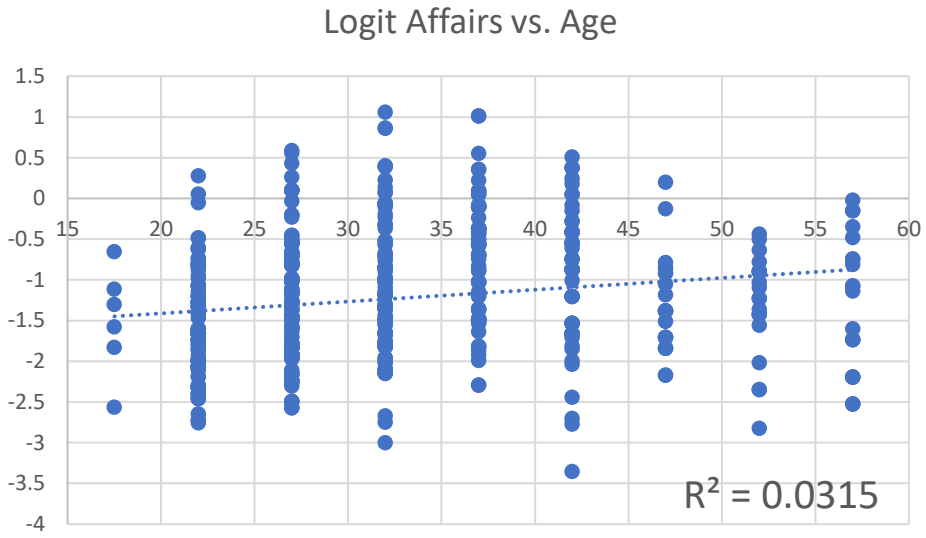
Logistic Regression Assumptions

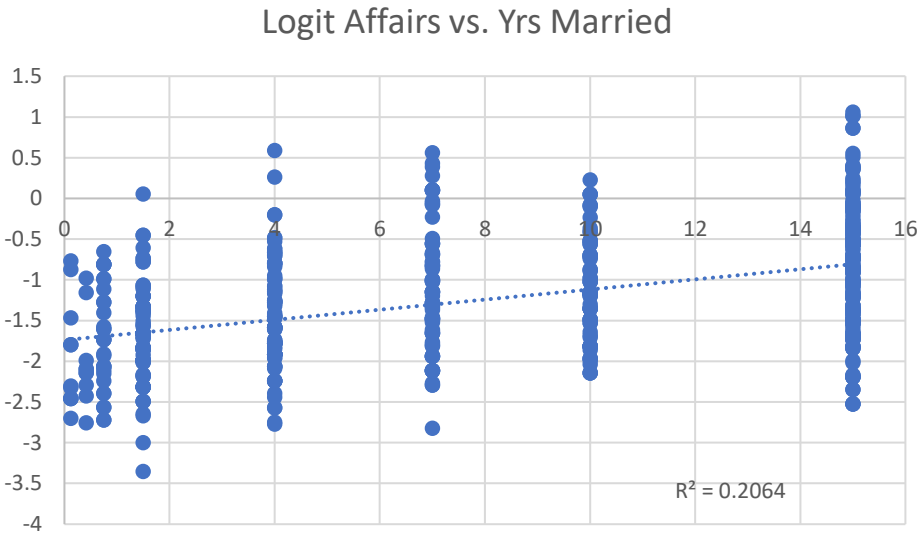
- Pre-model
  - Outcome is binary
  - Independence of obs
  - Large sample
  - No Multicollinearity
  - No “extreme cases”
- Post-model
  - Linearity between predictors and logit of the outcome

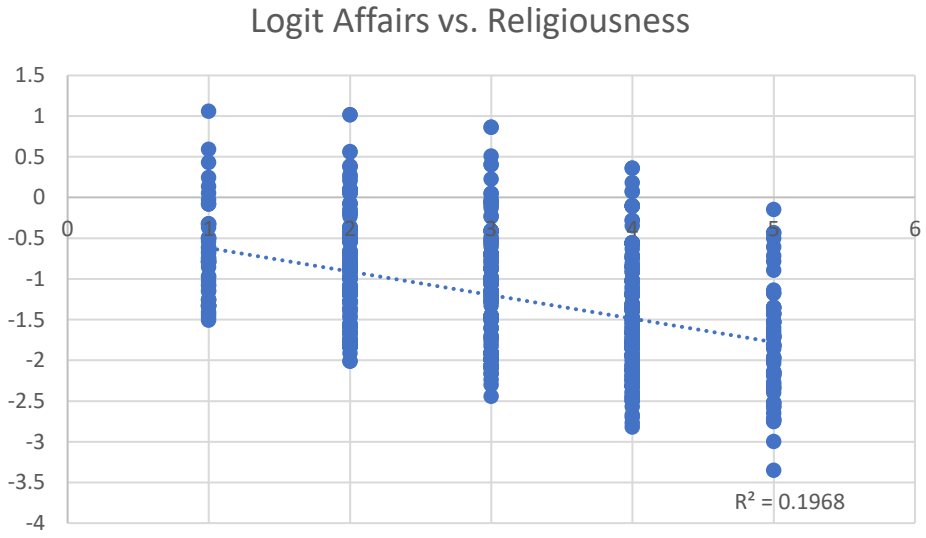
Let’s run the model and see where we’re at.

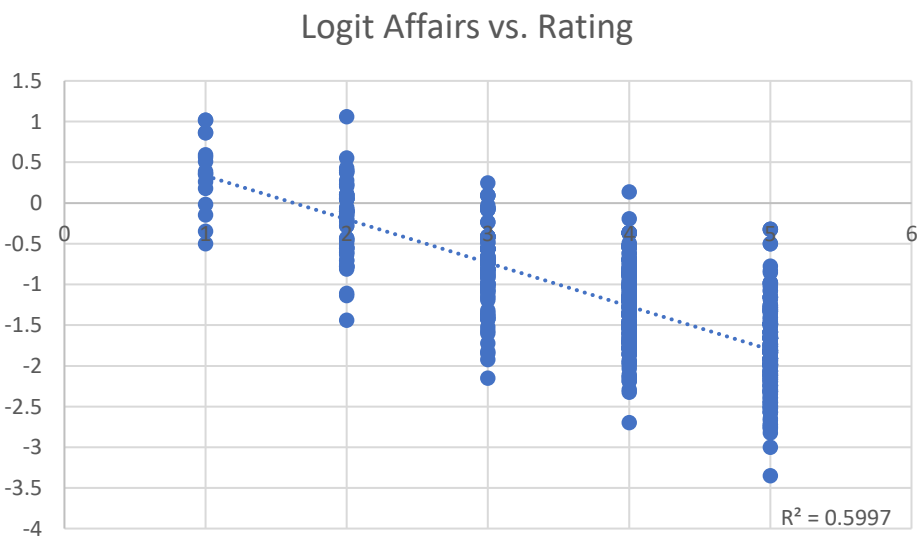
Logistic Regression Assumptions

- Pre-model
  - Outcome is binary
  - Independence of obs
  - Large sample
  - No Multicollinearity
  - No “extreme cases”
- Post-model
  - Linearity between predictors and logit of the outcome


Logit Affairs vs. Age
1.5
1
0.5
0
-0.5
-1
-1.5
-2
-2.5
-3
-3.5
-4
15
20
25
30
35
40
45
50
55
60
R² = 0.0315


Logit Affairs vs. Yrs Married
1.5
1
0.5
0
-0.5
-1
-1.5
-2
-2.5
-3
-3.5
-4
0
2
4
6
8
10
12
14
16
R² = 0.2064


Logit Affairs vs. Religiousness
1.5
1
0.5
0
-0.5
-1
-1.5
-2
-2.5
-3
-3.5
-4
0
1
2
3
4
5
6
R² = 0.1968


Logit Affairs vs. Rating
1.5
1
0.5
0
-0.5
-1
-1.5
-2
-2.5
-3
-3.5
-4
0
1
2
3
4
5
6
R² = 0.5997