

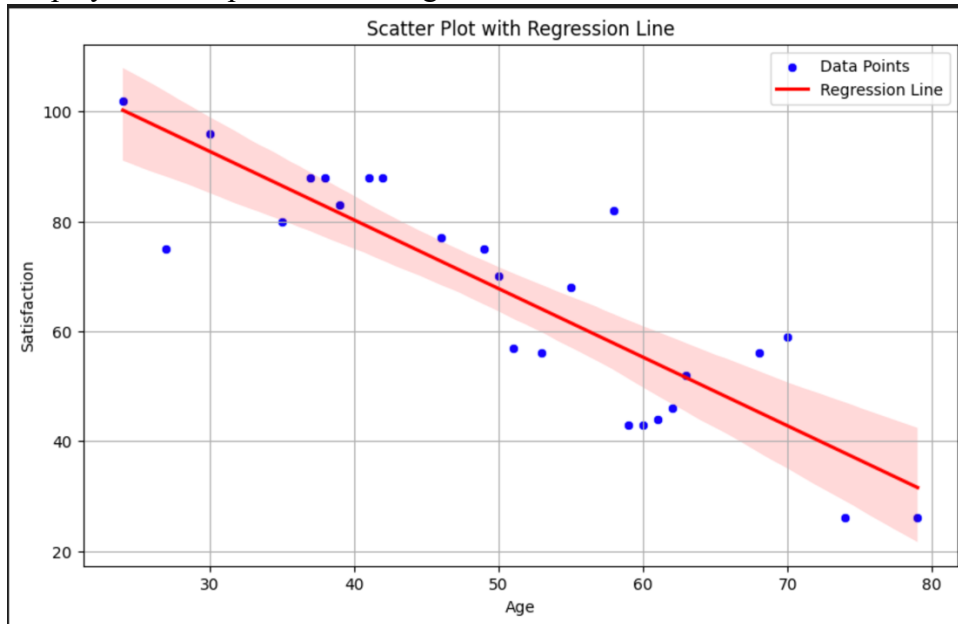
**ALL OF MY CODE IS AT THIS LINK, answers in this document:**

[https://github.com/bdupey/STAT-](https://github.com/bdupey/STAT-341A_Appplied_Regression_Analysis_Modeling/blob/main/hw_2.ipynb)

[341A\\_Appplied\\_Regression\\_Analysis\\_Modeling/blob/main/hw\\_2.ipynb](https://github.com/bdupey/STAT-341A_Appplied_Regression_Analysis_Modeling/blob/main/hw_2.ipynb)

**1) Section 2.7 – use this data to answer these questions: Use the variable Severity vs Satisfaction**

**a) Display a scatter plot with the regression line**



**b) Computer the regression line**

OLS Regression Results						
=====						
Dep. Variable:	Satisfaction		R-squared:	0.758		
Model:	OLS		Adj. R-squared:	0.748		
Method:	Least Squares		F-statistic:	72.09		
Date:	Tue, 19 Sep 2023		Prob (F-statistic):	1.52e-08		
Time:	09:41:56		Log-Likelihood:	-93.623		
No. Observations:	25		AIC:	191.2		
Df Residuals:	23		BIC:	193.7		
Df Model:	1					
Covariance Type:	nonrobust					
=====						
	coef	std err	t	P> t	[0.025	0.975]
-----						
const	130.2209	7.778	16.743	0.000	114.132	146.310
Age	-1.2490	0.147	-8.491	0.000	-1.553	-0.945
=====						
Omnibus:	0.156		Durbin-Watson:	1.712		
Prob(Omnibus):	0.925		Jarque-Bera (JB):	0.110		
Skew:	0.122		Prob(JB):	0.947		
Kurtosis:	2.787		Cond. No.	193.		
=====						

- c) Compute the ANOVA table

	sum_sq	df	F	PR(>F)
C(Surgical_Medical)	28.449455	1.0	0.289042	0.596767
Age	3545.147588	1.0	36.018177	0.000007
Severity	624.103110	1.0	6.340796	0.020428
Anxiety	70.989563	1.0	0.721243	0.405788
Residual	1968.532513	20.0	NaN	NaN

- d) Do a hypothesis test for the existence of an indirect linear relationship. Show all the steps

```
Indirect Effect: -1.2490
Direct Effect: -1.0462
Total Effect: -2.2952
Sobel Test Statistic: -4.8951
Sobel p-value: 0.0000
```

- e) Is the value of the y-intercept meaningful in the context of this problem? Justify!

The y-intercept, in this context, doesn't provide meaningful information because setting the Age variable to zero is often not meaningful or realistic.

- f) Compute and interpret the meaning of the coefficient of variation in the context of this problem.

```
The coefficient of variation for Age is 29.13%.
Interpretation: Moderate variability
```

Continued...

- g) Compute the 98% confidence interval for the slope  $\beta_1$ . Interpret its meaning. Is the slope of your regression equation inside this interval? Justify!

98% Confidence Interval for the Slope of Age:

Lower Bound: -1.6168

Upper Bound: -0.8813

We are 98 percent confident that the effect of 'Age' on 'Satisfaction' falls within this interval.

The slope coefficient from the regression equation is -1.2490. This means that for each one-unit increase in 'Age,' the 'Satisfaction' score decreases by approximately 1.2490 units on average.

The slope of the regression equation (-1.2490) is inside the 98% confidence interval (-1.6168 to -0.8813), confirming that the relationship between 'Age' and 'Satisfaction' is statistically significant and negative.

- h) Compute the 98% confidence interval for the average y value given  $x_0 = 45$ . Interpret its meaning.

98% Confidence Interval for Average 'Satisfaction' at Age 40:

Lower Bound: 54.6295

Upper Bound: 105.8896

We can conclude that the average 'Satisfaction' score for individuals aged 40 is likely to be within this range.

Continued....

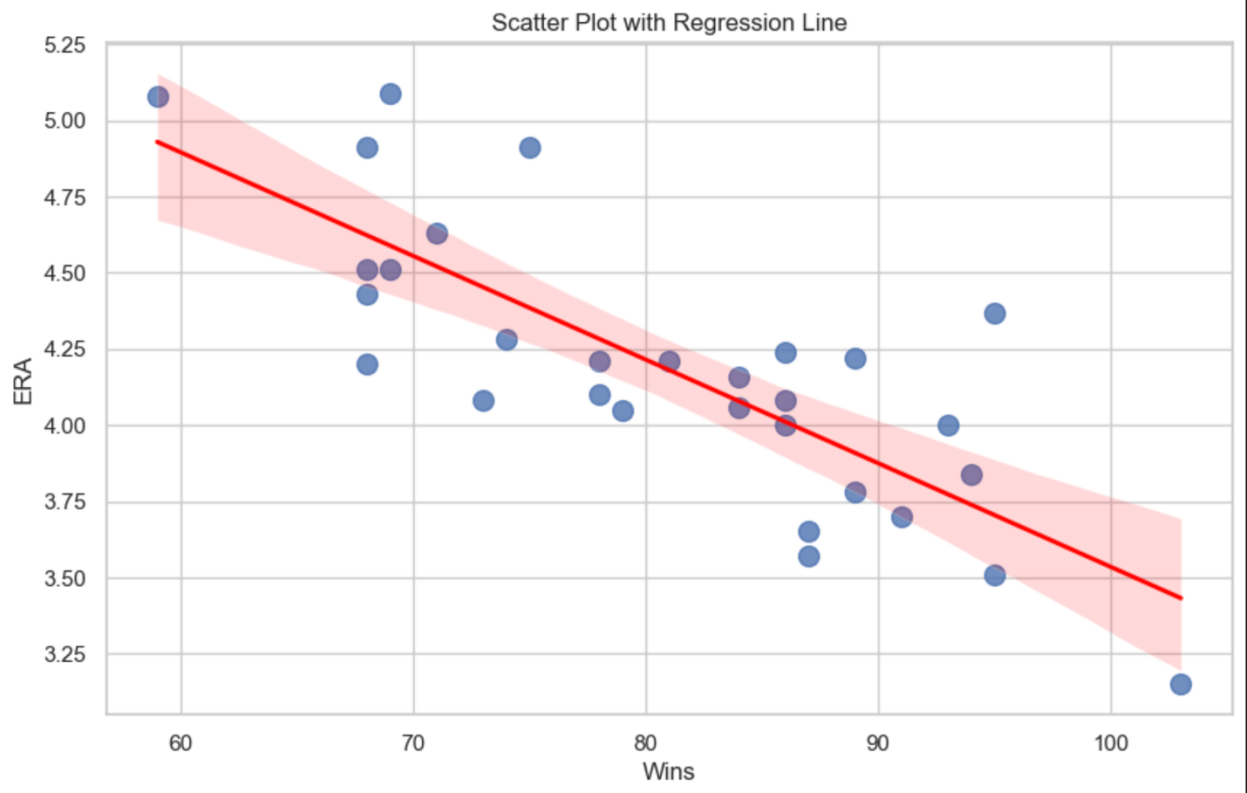
- i) Compute the 99% prediction interval for the y value given  $x_0 = 45$ . Interpret its meaning.

99% Prediction Interval for 'Satisfaction' at Age 40:  
Lower Bound: 51.5495  
Upper Bound: 108.9696

if you were to encounter an individual aged 40,  
you can be reasonably confident that their  
'Satisfaction' score will be somewhere within  
the range of 51.5495 to 108.9696

- 2) Section 2.8 – use this data to answer these questions: Use the variables Era vs Wins

- a) Display a scatter plot with the regression line



- b) Computer the regression line

Slope:  $-0.034031266095070734$   
Intercept:  $6.937462760424556$   
Correlation Coefficient (r-value):  $-0.7961123255322091$

- c) Compute the ANOVA table

	sum_sq	df	F	PR(>F)
Wins	3.822630	1.0	48.459872	1.437582e-07
Residual	2.208707	28.0	NaN	NaN

- d) Do a hypothesis test for the existence of an indirect linear relationship. Show all the steps.

OLS Regression Results						
=====						
Dep. Variable:	DV	R-squared:	0.634			
Model:	OLS	Adj. R-squared:	0.621			
Method:	Least Squares	F-statistic:	48.46			
Date:	Tue, 19 Sep 2023	Prob (F-statistic):	1.44e-07			
Time:	10:14:43	Log-Likelihood:	-3.4363			
No. Observations:	30	AIC:	10.87			
Df Residuals:	28	BIC:	13.67			
Df Model:	1					
Covariance Type:	nonrobust					
=====						
	coef	std err	t	P> t	[0.025	0.975]
-----						
Intercept	6.9375	0.399	17.396	0.000	6.121	7.754
IV	-0.0340	0.005	-6.961	0.000	-0.044	-0.024
=====						
Omnibus:	2.319	Durbin-Watson:	2.158			
Prob(Omnibus):	0.314	Jarque-Bera (JB):	1.819			
Skew:	0.597	Prob(JB):	0.403			
Kurtosis:	2.833	Cond. No.	635.			
=====						
Step 8: Mediator ('Batter Age') does not have a significant effect.						
There may not be a significant indirect linear relationship.						

- e) Is the value of the y-intercept meaningful in the context of this problem? Justify!  
 For each additional win, 'ERA' decreases by approximately 0.0340. This coefficient provides a more interpretable understanding of the relationship between 'Wins' and 'ERA'

- f) Compute and interpret the meaning of the coefficient of variation in the context of this problem.

The coefficient of variation (CV) for 'ERA' is 10.90%.  
In the context of ERA, a CV of 10.90% indicates Moderate variation.

- g) Compute the 97% confidence interval for the slope  $\beta_1$ . Interpret its meaning. Is the slope of your regression equation inside this interval? Justify!

97% Confidence Interval for the Slope ( $\beta_1$ ): (-0.0452, -0.0229)

With 97% confidence, we can say that there is a statistically significant negative linear relationship between the number of wins and earned run average.

- h) Compute the 97% confidence interval for the average y value given  $x_0 = 3.8$ . Interpret its meaning.

97% Confidence Interval for Average Y at  $x_0=3.8$ : (6.1660, 7.4503)

Setting 'Wins' to 3.8, you can be 97% confident that the average 'ERA' falls between 6.1660 and 7.4503 units

- i) Compute the 97% prediction interval for the y value given  $x_0 = 3.8$ . Interpret its meaning.

97% Prediction Interval for Average Y at  $x_0=3.8$ : (5.9384, 7.6778)

If a baseball team achieves 3.8 wins, we can be 97% confident that the average 'ERA' for teams with this level of performance will fall within the range of 5.9384 to 7.6778 units.