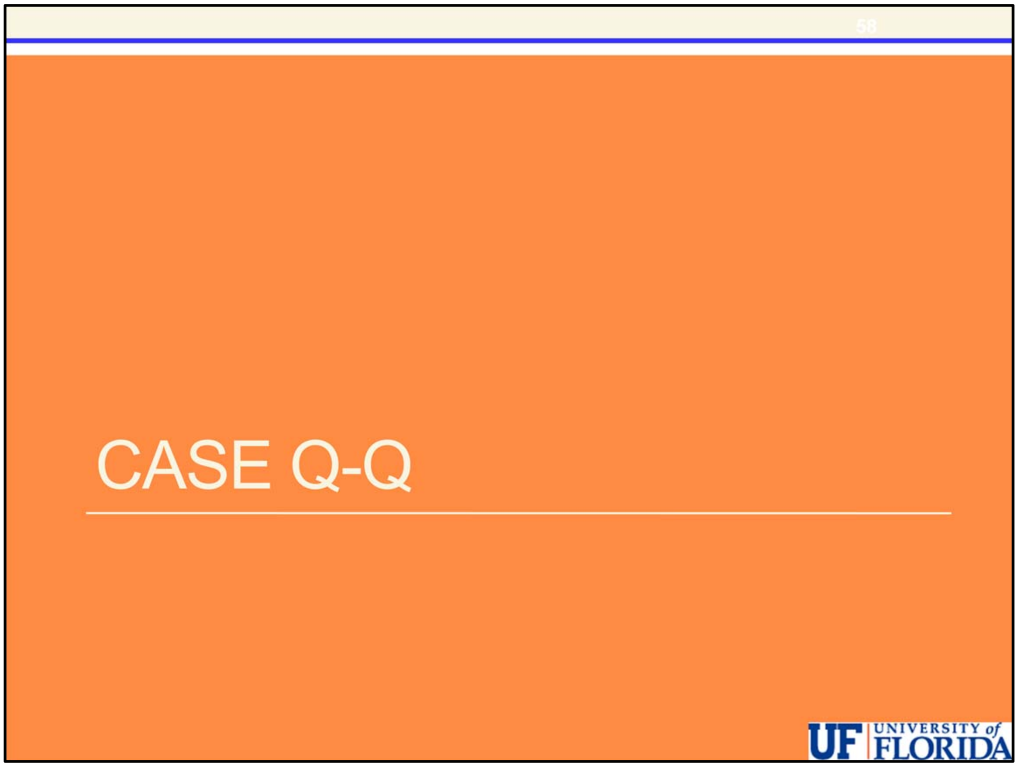
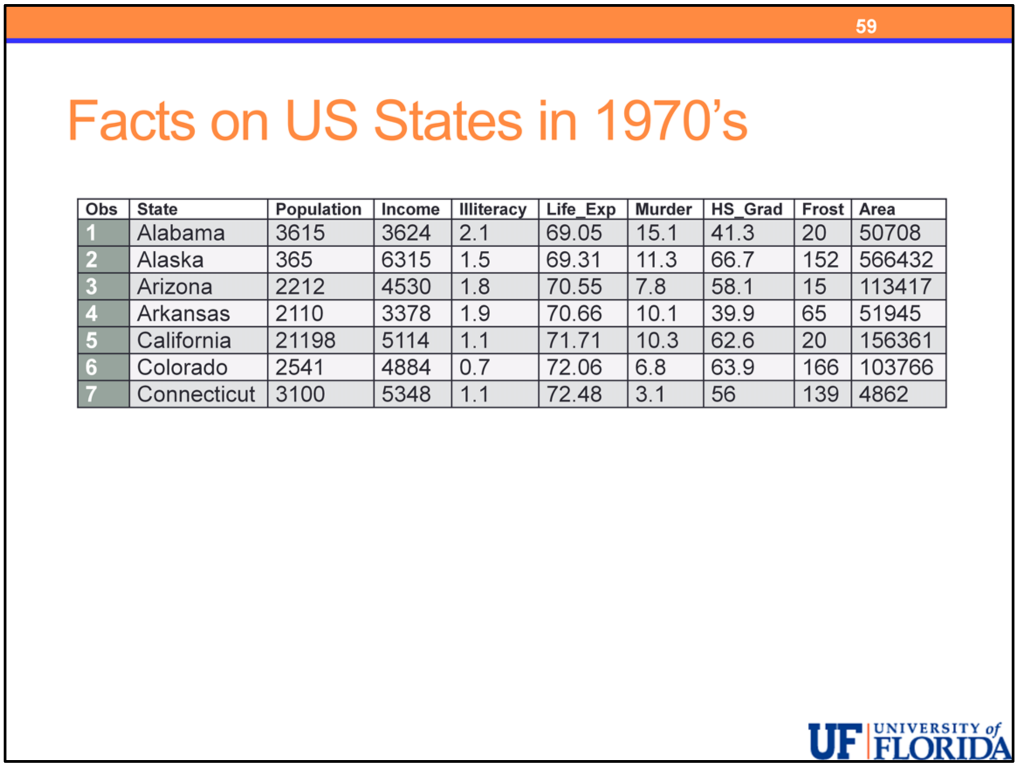


CASE Q-Q

Dataset information:

http://stat.ethz.ch/R-manual/R-devel/library/datasets/html/state.html

For case Q-Q, we will use a dataset containing information about U.S. states during the 1970’s.



Facts on US States in 1970's

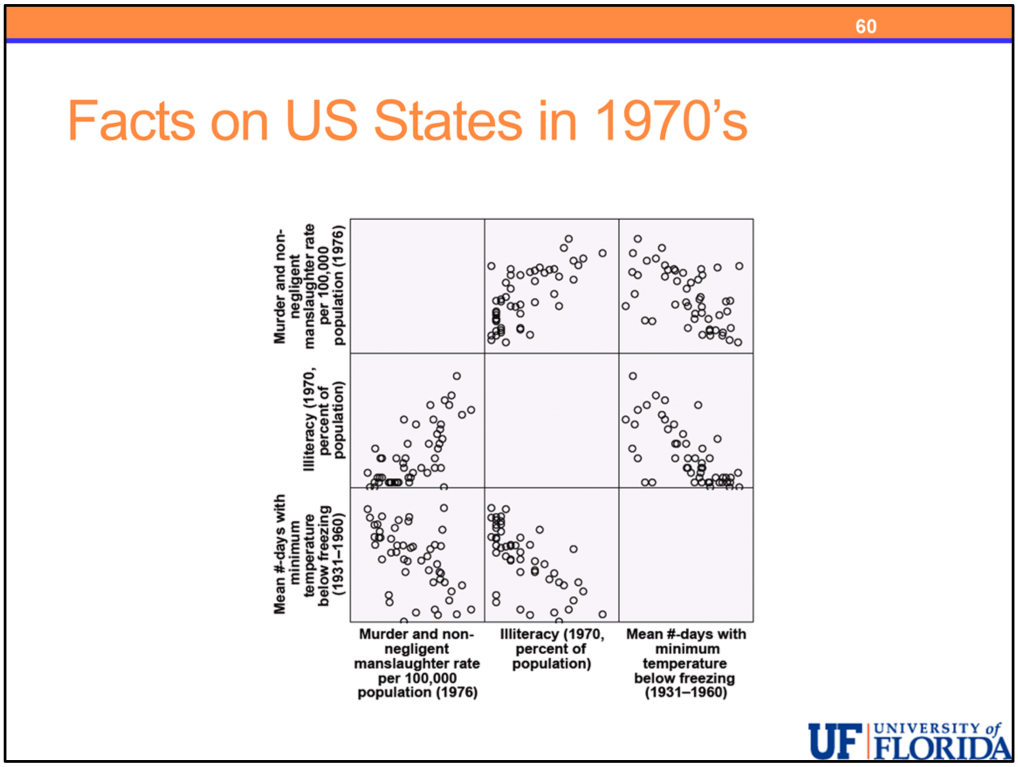
Obs	State	Population	Income	Illiteracy	Life_Exp	Murder	HS_Grad	Frost	Area
1	Alabama	3615	3624	2.1	69.05	15.1	41.3	20	50708
2	Alaska	365	6315	1.5	69.31	11.3	66.7	152	566432
3	Arizona	2212	4530	1.8	70.55	7.8	58.1	15	113417
4	Arkansas	2110	3378	1.9	70.66	10.1	39.9	65	51945
5	California	21198	5114	1.1	71.71	10.3	62.6	20	156361
6	Colorado	2541	4884	0.7	72.06	6.8	63.9	166	103766
7	Connecticut	3100	5348	1.1	72.48	3.1	56	139	4862

A few lines of the data are shown here.

The variables are:
- state name
- Population
- per-capita income
- illiteracy rate
- life expectancy
- Murder and non-negligent manslaughter rate per 100,000 population
- Percent high school graduates
- Mean number of days with the minimum temperature below freezing in capital or large city

And the
- Land area in square miles.

In particular we will investigate the associations between murder, frost, and illiteracy.

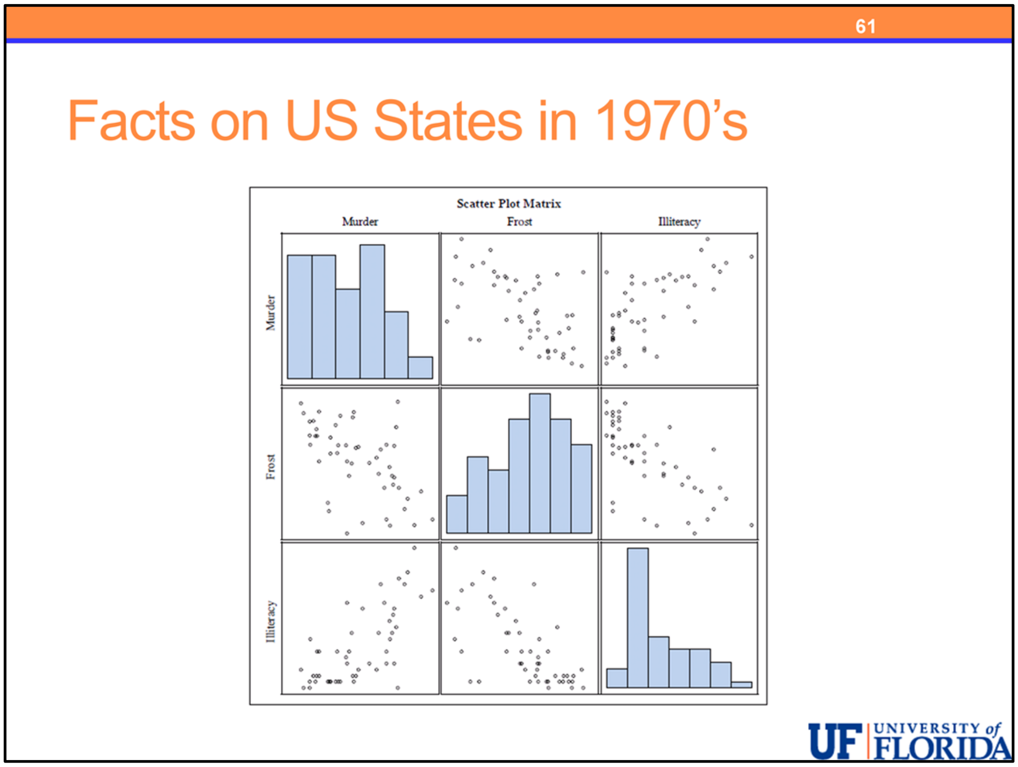


60

Facts on US States in 1970's

UF UNIVERSITY of FLORIDA

This is a scatterplot matrix from SPSS showing the scatterplots of all possible pairings between the variables murder, frost, and illiteracy.

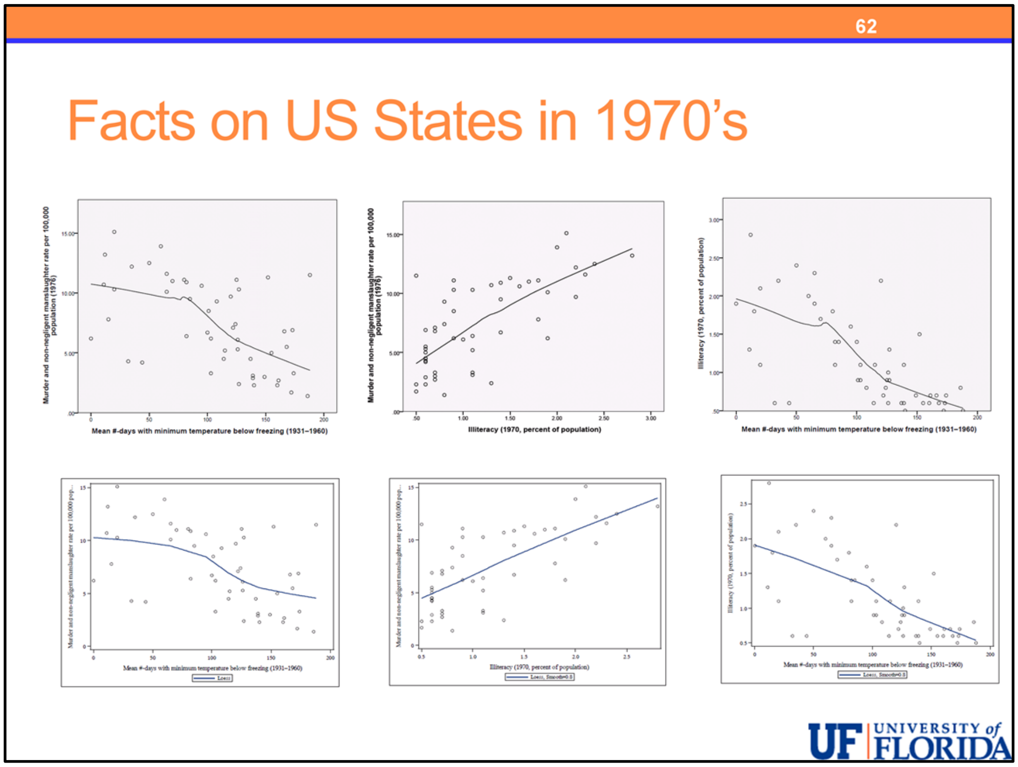


61

Facts on US States in 1970's

And a similar scatterplot matrix from SAS.

None of these scatterplots show any clear non-linear trends although there may be some outliers.



62

Facts on US States in 1970's

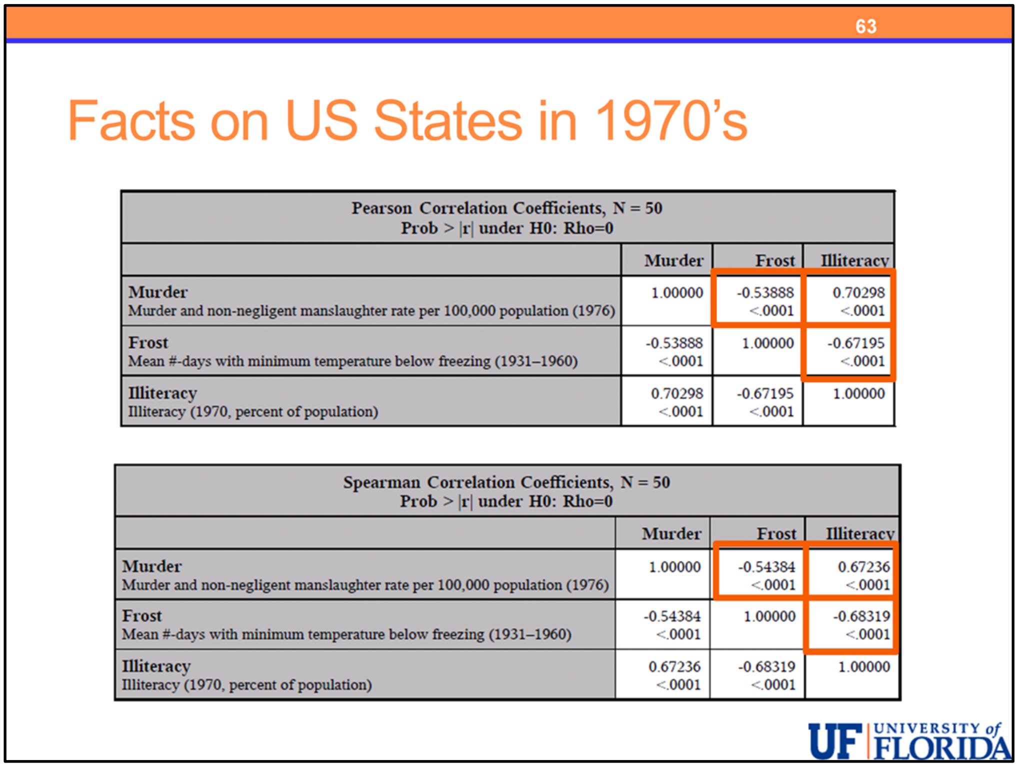
Here we have individual scatterplots with LOESS curves for
Murder vs. frost (on the left)
Murder vs. illiteracy (center)
Illiteracy vs. frost (right)

Of the three plots, the murder vs. illiteracy scatterplot in the center shows the most linear trend followed by illiteracy vs. frost (on the right) and finally murder vs. frost (on the left).

Although the plot for murder vs. frost (on the left) may be truly non-linear, we will investigate all three of these relationships further using correlation and regression.

From these plots we would expect a negative correlation between murder and frost (on the left) and between illiteracy and frost (on the right)

And a positive correlation between murder and illiteracy.



63

Facts on US States in 1970's

Pearson Correlation Coefficients, N = 50 Prob > |r| under H0: Rho=0	Murder	Frost	Illiteracy
Murder Murder and non-negligent manslaughter rate per 100,000 population (1976)	1.00000	-0.53888 <.0001	0.70298 <.0001
Frost Mean #-days with minimum temperature below freezing (1931–1960)	-0.53888 <.0001	1.00000	-0.67195 <.0001
Illiteracy Illiteracy (1970, percent of population)	0.70298 <.0001	-0.67195 <.0001	1.00000

Spearman Correlation Coefficients, N = 50 Prob > |r| under H0: Rho=0	Murder	Frost	Illiteracy
Murder Murder and non-negligent manslaughter rate per 100,000 population (1976)	1.00000	-0.54384 <.0001	0.67236 <.0001
Frost Mean #-days with minimum temperature below freezing (1931–1960)	-0.54384 <.0001	1.00000	-0.68319 <.0001
Illiteracy Illiteracy (1970, percent of population)	0.67236 <.0001	-0.68319 <.0001	1.00000

UF UNIVERSITY of FLORIDA

First we have the SAS output for both Pearson's and Spearman's correlation between all combinations.

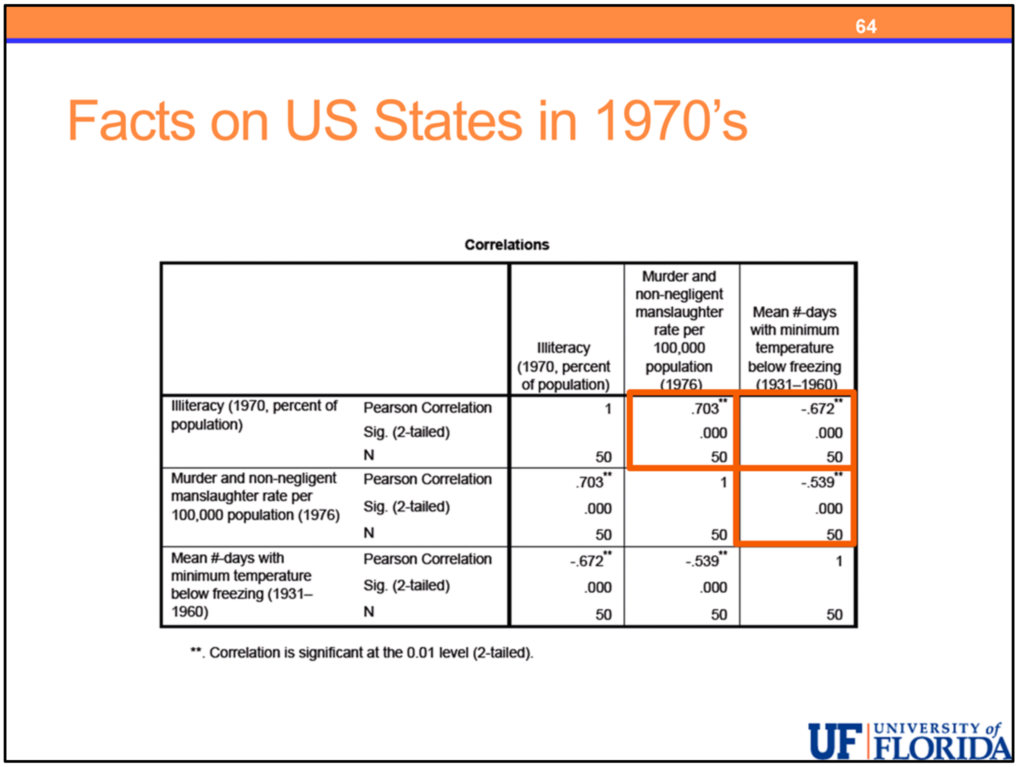
The results are all highly statistically significant.

For Murder vs. Frost, Pearson's correlation is -0.539 and Spearman's is -0.544. Both indicating a moderately strong negative linear association between murder and frost. As the mean number of days below freezing increases, the murder rate tends to decrease.

For Murder vs. Illiteracy, Pearson's correlation is 0.703 and Spearman's is 0.672. Both indicating a somewhat strong positive linear association between murder and illiteracy. As the illiteracy rate increases, the murder rate tends to increase.

For Frost vs. Illiteracy, Pearson's correlation is -0.672 and Spearman's is -0.683. Both indicating a somewhat strong negative linear association between frost and illiteracy. As the mean number of days below freezing increases, the illiteracy rate tends to decrease.

These values confirm what we found in the previous scatterplots.



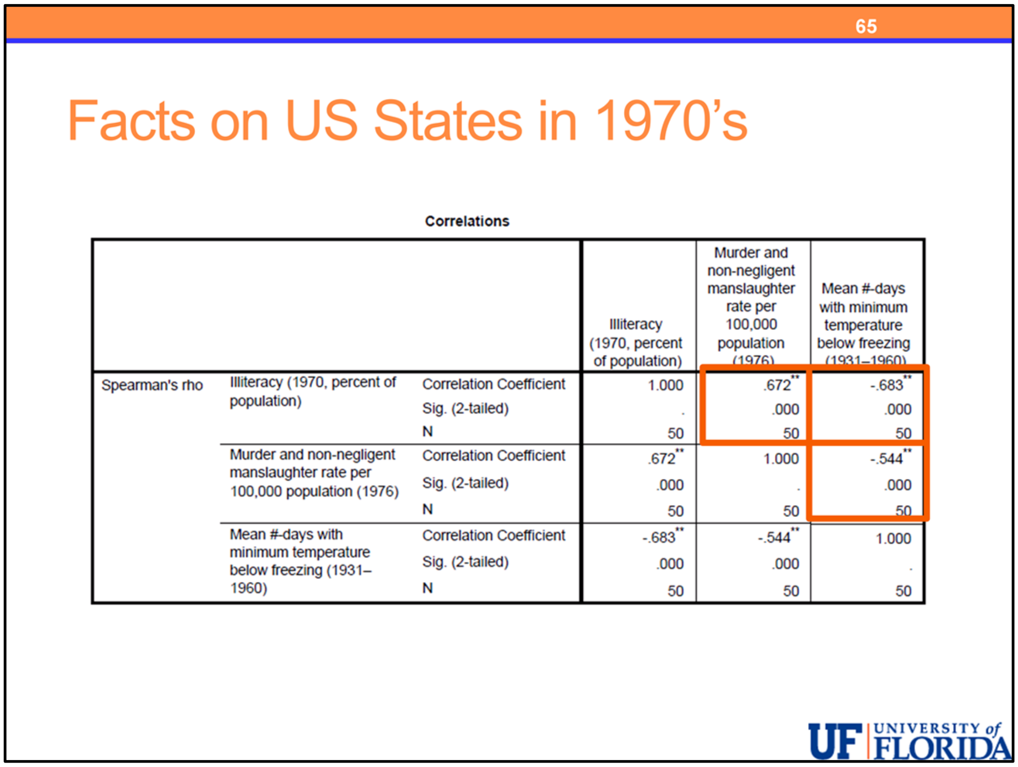
Facts on US States in 1970’s

Correlations

		Illiteracy (1970, percent of population)	Murder and non-negligent manslaughter rate per 100,000 population (1976)	Mean #-days with minimum temperature below freezing (1931–1960)
Illiteracy (1970, percent of population)	Pearson Correlation	1	.703**	-.672**
	Sig. (2-tailed)		.000	.000
	N	50	50	50
Murder and non-negligent manslaughter rate per 100,000 population (1976)	Pearson Correlation	.703**	1	-.539**
	Sig. (2-tailed)	.000		.000
	N	50	50	50
Mean #-days with minimum temperature below freezing (1931–1960)	Pearson Correlation	-.672**	-.539**	1
	Sig. (2-tailed)	.000	.000	
	N	50	50	50

**. Correlation is significant at the 0.01 level (2-tailed).

We find the same results for Pearson’s correlation in SPSS.

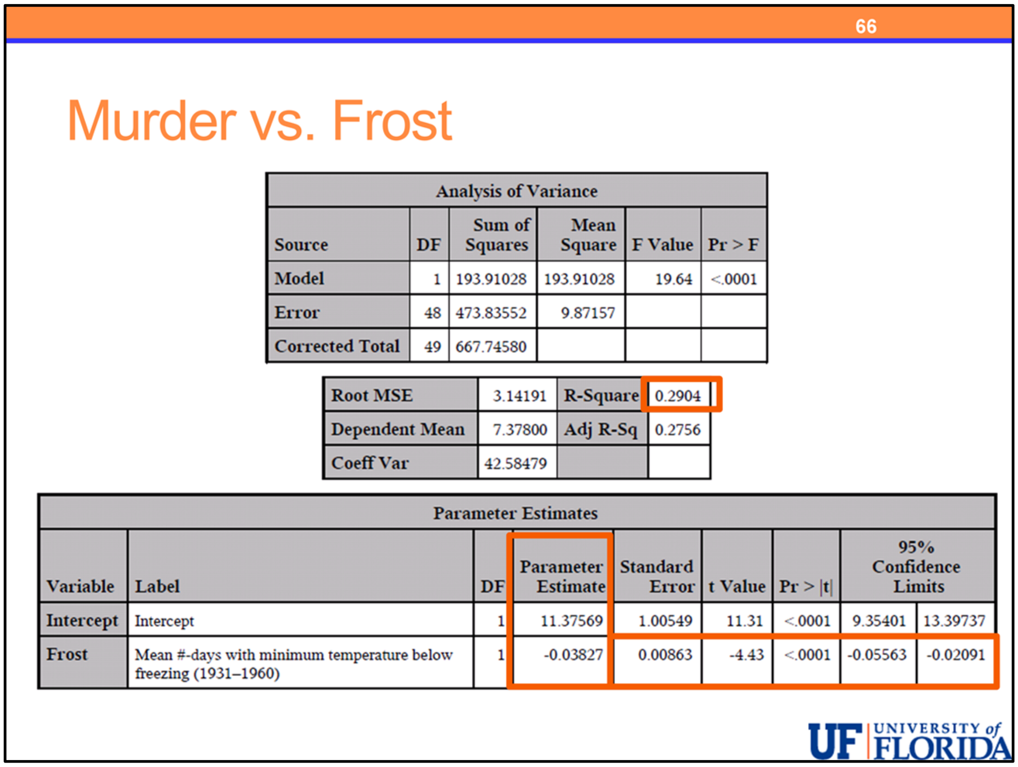


Facts on US States in 1970's

Correlations

			Illiteracy (1970, percent of population)	Murder and non-negligent manslaughter rate per 100,000 population (1976)	Mean #-days with minimum temperature below freezing (1931–1960)
Spearman's rho	Illiteracy (1970, percent of population)	Correlation Coefficient	1.000	.672**	-.683**
		Sig. (2-tailed)	.	.000	.000
		N	50	50	50
	Murder and non-negligent manslaughter rate per 100,000 population (1976)	Correlation Coefficient	.672**	1.000	-.544**
		Sig. (2-tailed)	.000	.	.000
		N	50	50	50
	Mean #-days with minimum temperature below freezing (1931–1960)	Correlation Coefficient	-.683**	-.544**	1.000
		Sig. (2-tailed)	.000	.000	.
		N	50	50	50

And for Spearman's correlation. The only difference is in the order the variables are presented.



Murder vs. Frost

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	193.91028	193.91028	19.64	<.0001
Error	48	473.83552	9.87157		
Corrected Total	49	667.74580			

Root MSE	3.14191	R-Square	0.2904
Dependent Mean	7.37800	Adj R-Sq	0.2756
Coeff Var	42.58479		

Parameter Estimates								
Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	95% Confidence Limits	
Intercept	Intercept	1	11.37569	1.00549	11.31	<.0001	9.35401	13.39737
Frost	Mean #-days with minimum temperature below freezing (1931–1960)	1	-0.03827	0.00863	-4.43	<.0001	-0.05563	-0.02091

Now we can continue with simple linear regression.

Values of particular interest are outlined.

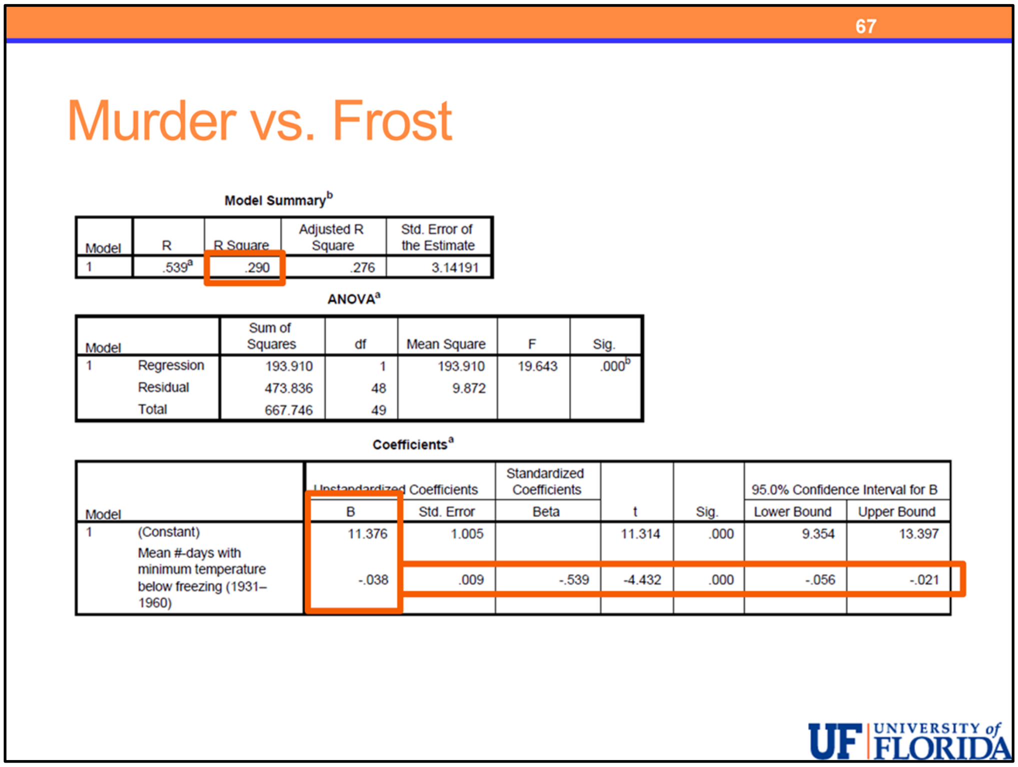
We have an R-squared of 0.2904 indicating that 29% of the variation in murder rate can be explained by the mean number of days below freezing.

The slope is statistically significant with a p-value <0.0001.

The linear regression equation is: Predicted Murder Rate = 11.38 – 0.038(Frost).

The 95% confidence interval for the slope is -0.056 to -0.021.

We can interpret the slope and it's confidence interval by saying: For each 1 day increase in the *mean number of days with minimum temperature below freezing*, the **average** murder rate decreases by 0.038. The 95% confidence interval suggests this decrease could be as little as 0.021 to as much as 0.056.



Murder vs. Frost

Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.539[a]	.290	.276	3.14191

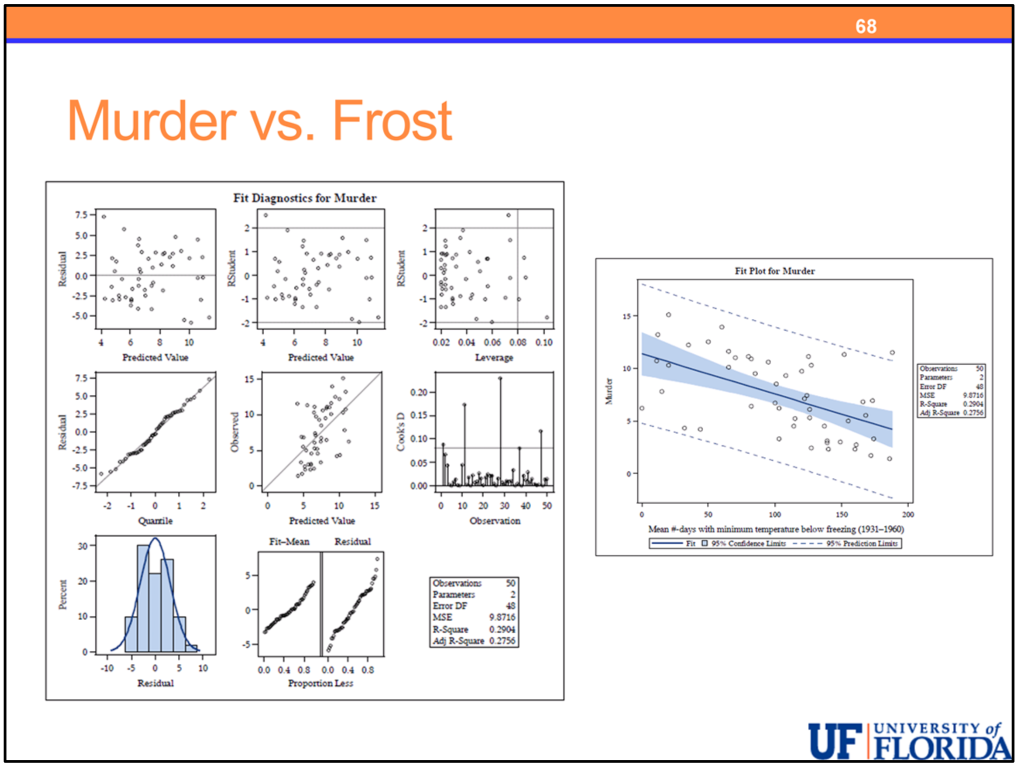
ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	193.910	1	193.910	19.643	.000[b]
	Residual	473.836	48	9.872		
	Total	667.746	49			

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients			95.0% Confidence Interval for B	
		B	Std. Error	Beta	t	Sig.	Lower Bound	Upper Bound
1	(Constant)	11.376	1.005		11.314	.000	9.354	13.397
	Mean #-days with minimum temperature below freezing (1931–1960)	-.038	.009	-.539	-4.432	.000	-.056	-.021

The results from SPSS are exactly the same except for differences in rounding.



Murder vs. Frost

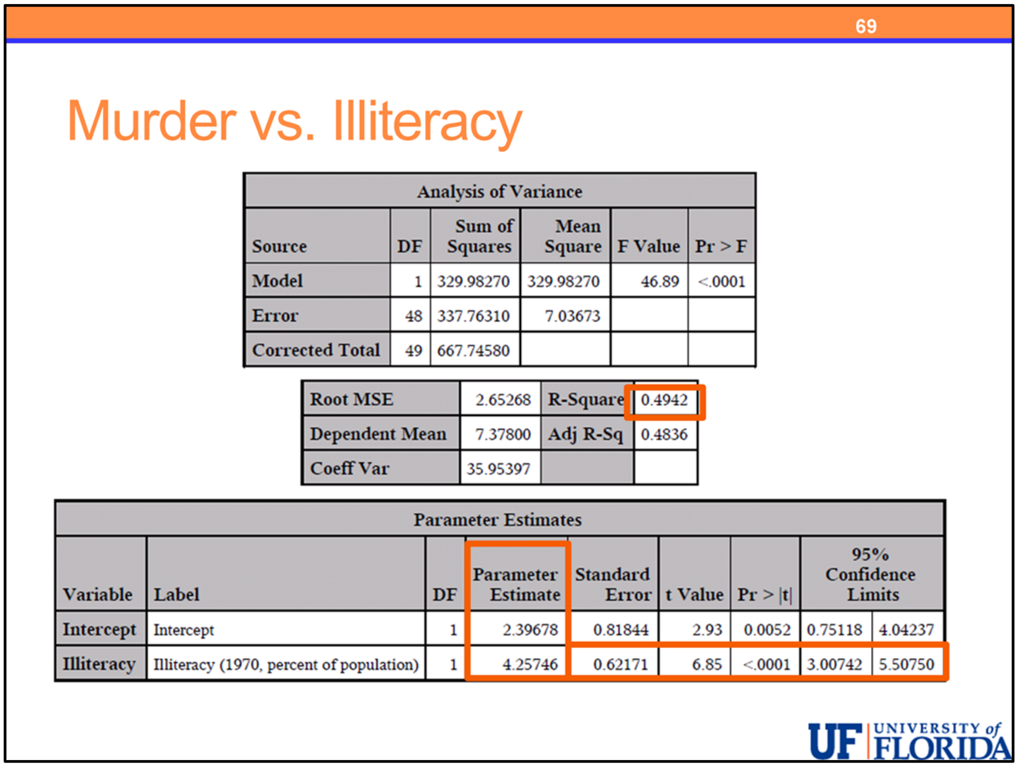
In SAS we obtain the following diagnostic plots and a fit plot by default when conducting a regression analysis.

We need to verify that the relationship is reasonably linear, which we have here.

We need to check that the residuals are approximately normally distributed. Looking at the QQ-plot and histogram of the residuals, the normality assumption seems completely reasonable.

We need to check the assumption of constant variance. From the plot of the residuals by the predicted values, there is no clear violation of this assumption. The points are relatively evenly distributed with similar spread around the horizontal line at zero over the range of predicted values. We could also look at the scatterplot of the data to see that the constant variance assumption is reasonable.

We haven’t learned about all of the graphs displayed here by SAS but if you go on to a regression course you will learn more about some of these plots and measures.



69

Murder vs. Illiteracy

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	329.98270	329.98270	46.89	<.0001
Error	48	337.76310	7.03673		
Corrected Total	49	667.74580			

Root MSE	2.65268	R-Square	0.4942
Dependent Mean	7.37800	Adj R-Sq	0.4836
Coeff Var	35.95397		

Parameter Estimates

Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	95% Confidence Limits	
Intercept	Intercept	1	2.39678	0.81844	2.93	0.0052	0.75118	4.04237
Illiteracy	Illiteracy (1970, percent of population)	1	4.25746	0.62171	6.85	<.0001	3.00742	5.50750

UF UNIVERSITY of FLORIDA

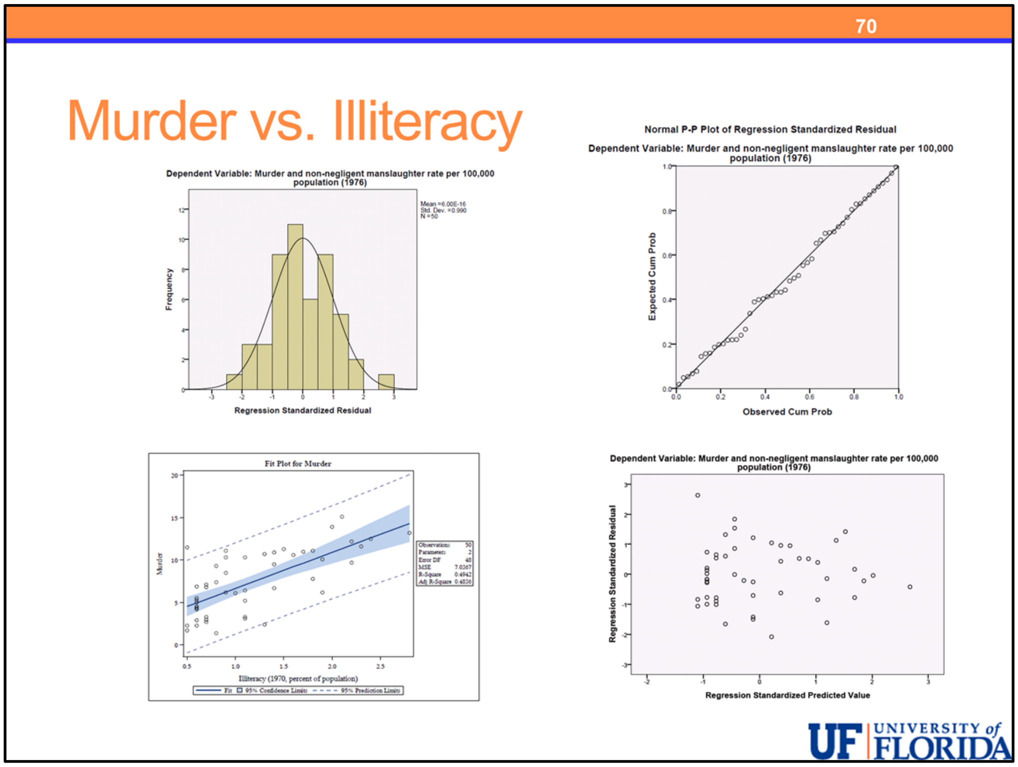
For murder vs. illiteracy, we have an R-squared of 0.4942 indicating that 49% of the variation in murder rate can be explained by illiteracy.

The slope is statistically significant with a p-value <0.0001.

The linear regression equation is: Predicted Murder Rate = 2.40 + 4.26(Illiteracy).

The 95% confidence interval for the slope is 3.007 to 5.508.

We can interpret the slope and it's confidence interval by saying: For each 1 percentage point increase in the *illiteracy rate*, the **average** murder rate increases by 4.26. The 95% confidence interval suggests this increase could be as little as 3.007 to as much as 5.508.



70

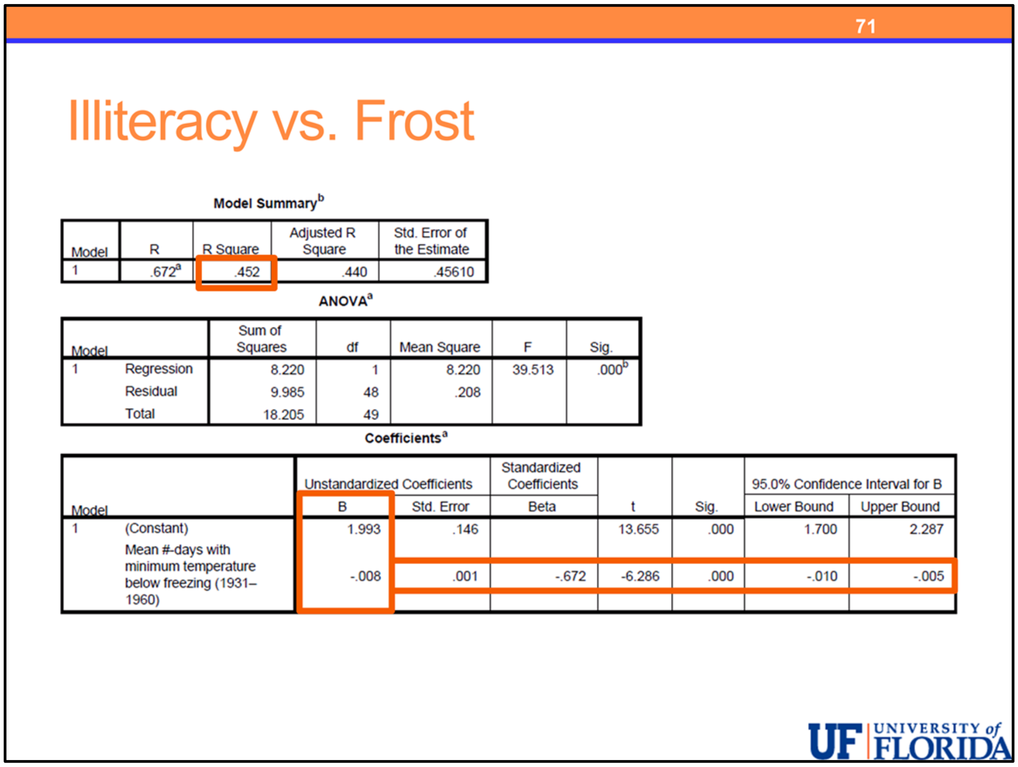
Murder vs. Illiteracy

Here we use the SPSS versions of the needed graphs to validate assumptions.

Linearity is reasonable from the scatterplot.

The histogram and normal probability plot indicate normality is reasonable. In regression, SPSS gives a PP-plot instead of a QQ-plot but these graphs are identical in what we expect to see and how they are interpreted and can be used interchangeably.

Finally the plot of the residuals by the predicted values shows no major issues although there does seem to be a slight decrease in the spread as the predicted value increases, this could be driven by two odd points in the scatterplot – one high value on the left side which is unusually far from the line and one on the right side corresponding to the largest x-value as if we ignore those two points, what remains seems to better satisfy the constant variance assumption.



Illiteracy vs. Frost

Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.672[a]	.452	.440	.45610

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	8.220	1	8.220	39.513	.000[b]
	Residual	9.985	48	.208		
	Total	18.205	49			

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95.0% Confidence Interval for B	
		B	Std. Error	Beta			Lower Bound	Upper Bound
1	(Constant)	1.993	.146		13.655	.000	1.700	2.287
	Mean #-days with minimum temperature below freezing (1931–1960)	-.008	.001	-.672	-6.286	.000	-.010	-.005

UF UNIVERSITY of FLORIDA

Finally for illiteracy vs. frost, we have an R-squared of 0.452 indicating that 45% of the variation in illiteracy can be explained by frost.

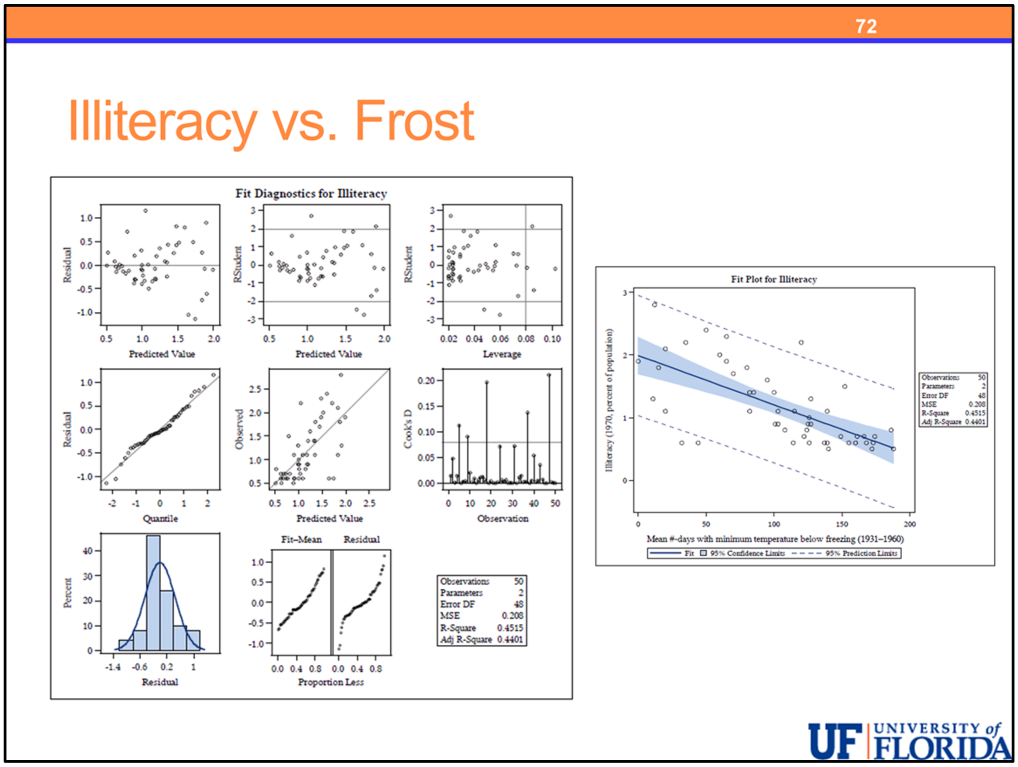
The slope is statistically significant with a p-value reported as 0.000.

The linear regression equation is: Predicted Percent Illiteracy = 1.993 – 0.008(Frost).

The 95% confidence interval for the slope is -0.010 to -0.005.

We can interpret the slope and it’s confidence interval by saying: For each 1 day increase in the *mean number of days with minimum temperature below freezing*, the **average** illiteracy percentage decreases by 0.008. The 95% confidence interval suggests this decrease could be as little as 0.005 to as much as 0.01.

Illiteracy vs. Frost

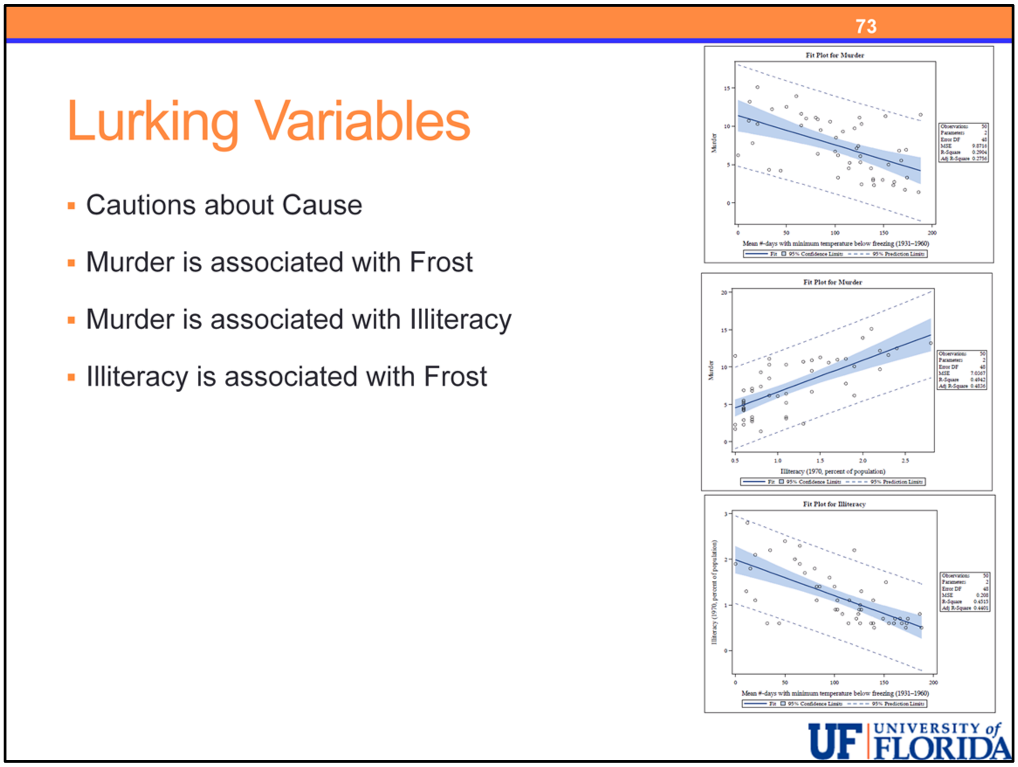


Observations	50
Parameters	2
Error DF	48
MSE	0.208
R-Square	0.4515
Adj R-Square	0.4401

In this case, linearity is reasonably.

The residuals are reasonably normally distributed based upon the QQ-plot and histogram of the residuals.

However, in this case, there does seem to be a strange pattern in the residual vs. predicted values plot and the original scatterplot. The residuals vs. predicted values shows an increasing spread as the predicted value increases. The scatterplot shows a similar trend in that as the variable Frost increases, the variation around the regression line seems to be decreasing. Thus there is some concern about the validity of the constant variance assumption.



Lurking Variables

- Cautions about Cause
- Murder is associated with Frost
- Murder is associated with Illiteracy
- Illiteracy is associated with Frost

UF UNIVERSITY of FLORIDA

Although we found associations in each of these three regression models, we must be careful about concluding the relationship is causal.

The first relationship found as the mean number of days with minimum temperature below freezing increases, the murder rate decreases but we CANNOT say that more days below freezing CAUSES the murder rate to decrease.

In the second relationship we see that as the illiteracy percentage increases, the murder rate also increases but again we CANNOT say that higher illiteracy percentage CAUSES the murder rate to increase.

The fact that illiteracy and frost are also related in the third relationship shows that when considering the relationship between murder and frost, we must be aware that illiteracy is also related to both frost and murder and thus illiteracy is a potential lurking variable in this relationship between murder and frost.

In general, unless you have performed a randomized controlled experiment, you should always be cautious about claiming a direct causal link between the explanatory and response variables in any analysis!