

Simple Linear Regression

$Y = B_0 + B_1X$	$Y = mx + b$
Y = dependent variable	Y = dependent variable
B_0 = Constant	m = slope
B_1 = Slope (a.k.a regression coefficient)	b = intercept
X = Value of the IV	x = the IV

ASSUMPTIONS:

1. Linearity
2. Independence
3. Homoscedasticity
4. Normality

A meteorologist is preparing his weather forecast for the morning news. He has two pieces of information that he hopes will help him make his prediction. He has data available to him reflecting the average temperature and the amount of precipitation for the last 62 days. He wants to know if the amount of precipitation can significantly predict what the temperature will be for the morning news.

Null & Alternative Hypotheses:

H_0 : $B_1 = 0$

H_a : $B_1 \neq 0$

***If there is a significant linear relationship between the independent variable (X) and the dependent variable (Y), the slope (B_1) will *not* equal zero.

Independent/Predictor/Explanatory Variable: precipitation

Dependent Variable: temperature

SPSS

Analyze → Regression → Linear
Dependent → temperature
Independent → precipitation

Statistics

Estimates

Model Fit

Descriptives

Residuals

Durbin Watson

Casewise Diagnostics

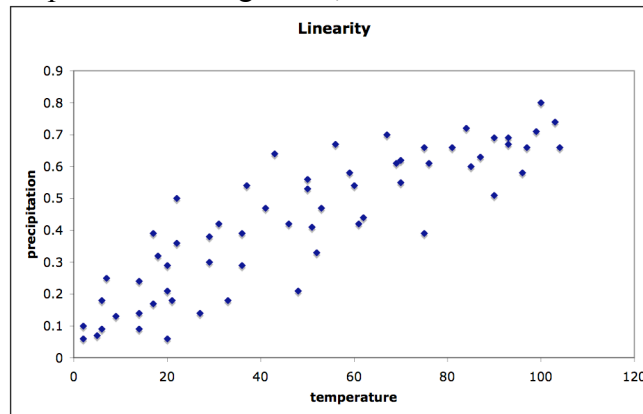
Ok.

SYNTAX.

```
REGRESSION
/DESCRIPTIVES MEAN STDDEV CORR SIG N
/MISSING LISTWISE
/STATISTICS COEFF OUTS R ANOVA
/CRITERIA=PIN(.05) POUT(.10)
/NOORIGIN
/DEPENDENT temp
/METHOD=ENTER precipitation
/RESIDUALS DURBIN
/CASEWISE PLOT(ZRESID) ALL .
```

Checking Assumptions

Linearity: Plot IV + DV. If these values together are not in a straight line, your data is not linear. This data represents a straight line, so we did not violate this assumption.



Independence: Durbin-Watson coefficient. Independence is violated if this statistic is outside the range of 1.5-2.5.

Model Summary(b)

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Durbin-Watson
1	.871(a)	.759	.755	15.451	1.543

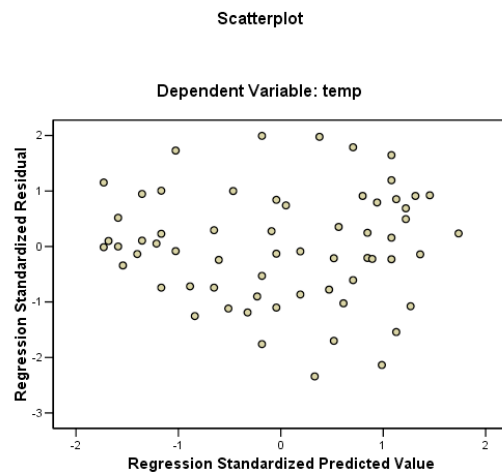
a Predictors: (Constant), precipitation
b Dependent Variable: temp

Based on the information we obtained from the model summary, we did not violate the assumption of independence ($d = 1.543$).

Homoscedasticity: The residuals should be relatively close to each other or equidistant. There are tests that can be used to test for homoscedasticity. Each test available makes assumptions about the shape of the errors. If you think your residuals are heteroscedastic, run one of these tests.

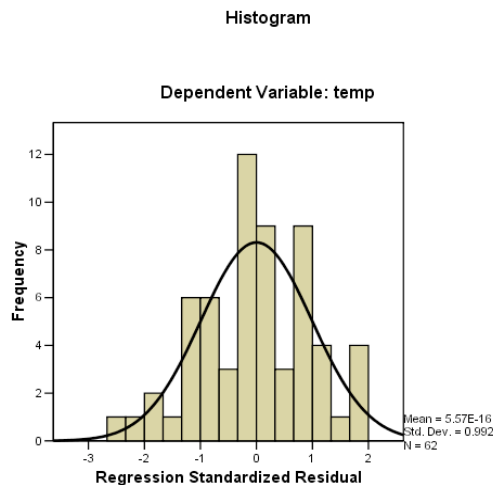
1. **Goldfeld-Quandt test:** funnel/fan shape
2. **Glejser test:** bow-tie shape
3. **Breusch-Pagan-Godfrey test:** for large samples

4. **White's test:** no prior knowledge of the heteroscedastic form, less powerful than the others



In this example, our residuals are relatively close to each other. This assumption is presumably not violated. If this assumption is violated, the model fit will be over-estimated, but the estimates can be improved using weighting procedures on the data (e.g., Box-Cox normality plot can help find the appropriate data transformation).

Normality: We want to check to see if our residuals are normally distributed. We should expect the data to show roughly a normal distribution. This data is not completely normal, it has a slight negative skew, but it is really close to a normal distribution. So, based on a visual assessment, we have not violated this assumption.



Just in case, we can test the normality of the Standardized Residuals using the Shapiro-Wilk's test (see below). We can get the Standardized Residuals from the SPSS Regression output.

Casewise Diagnostics(a)

Case Number	Std. Residual	temp	Predicted Value	Residual
-------------	---------------	------	-----------------	----------

1	-1.189	22	40.37	-18.367
2	.000	6	6.00	-.003
3	.853	93	79.82	13.179
4	.741	62	50.55	11.452
5	-.141	84	86.18	-2.185
6	-.718	14	25.09	-11.094
7	1.000	52	36.55	15.452
8	-.206	69	72.18	-3.185
9	1.647	104	78.55	25.451
10	.235	100	96.37	3.633
11	-.865	41	54.37	-13.367
12	.912	85	70.91	14.088
13	1.977	90	59.46	30.542
14	.947	27	12.37	14.634
15	-1.118	18	35.28	-17.276
16	1.730	48	21.28	26.724
17	-1.701	37	63.28	-26.276
18	-1.077	67	83.64	-16.639
19	-1.542	56	79.82	-23.821
20	-1.100	31	48.00	-17.003
21	-1.759	17	44.18	-27.185
22	-1.253	7	26.37	-19.366
23	-.012	2	2.18	-.184
24	-.088	53	54.37	-1.367
25	.353	70	64.55	5.452
26	.100	5	3.46	1.543
27	.494	90	82.37	7.633
28	-.130	46	48.00	-2.003
29	-.530	36	44.18	-8.185
30	.518	14	6.00	7.997
31	-.212	60	63.28	-3.276
32	.924	103	88.73	14.270
33	-2.136	43	76.00	-33.003
34	-2.342	22	58.18	-36.185
35	1.994	75	44.18	30.815
36	-.241	29	32.73	-3.730
37	.247	76	72.18	3.815
38	-.742	20	31.46	-11.457
39	-.900	29	42.91	-13.912
40	-.777	50	62.00	-12.003
41	-.606	59	68.37	-9.367
42	-.224	70	73.46	-3.458
43	.159	81	78.55	2.451
44	.688	93	82.37	10.633
45	.912	99	84.91	14.088
46	.106	14	12.37	1.634
47	.276	51	46.73	4.270

48	-.230	75	78.55	-3.549
49	-.742	6	17.46	-11.457
50	-.083	20	21.28	-1.276
51	.294	36	31.46	4.543
52	-1.024	50	65.82	-15.821
53	-.136	9	11.09	-2.094
54	-.341	2	7.28	-5.275
55	.229	21	17.46	3.543
56	.053	17	16.18	.815
57	.794	87	74.73	12.270
58	1.194	97	78.55	18.451
59	1.006	33	17.46	15.543
60	1.153	20	2.18	17.816
61	1.788	96	68.37	27.633
62	.841	61	48.00	12.997

a Dependent Variable: temp

Analyze → Descriptive statistics → Explore.

Click on Plots

Click on Options

Click on Normality Plots with tests.

Tests of Normality

	Kolmogorov-Smirnov(a)			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
Std. Residual	.081	62	.200(*)	.987	62	.756

* This is a lower bound of the true significance.

a Lilliefors Significance Correction

We the standardized residuals using Shapiro-Wilk's W test ($W(62) = .987, p = .756$). Since the Shapiro-Wilk's W test was not statistically significant, we did not violate the normality assumption.

Output & Interpretation

Now we can interpretation of the data!!! Woohoo, we made it!

Descriptive Statistics

	Mean	Std. Deviation	N
temp	49.19	31.205	62
precipitation	.4294	.21358	62

On average, over the past 62 days the temperature was 49.19°F ($\sigma = 31.21$), with 42.9% ($\sigma = 21.4\%$) precipitation.

Correlations

		temp	precipitation
Pearson Correlation	temp	1.000	.871
	precipitation	.871	1.000
Sig. (1-tailed)	temp	.	.000
	precipitation	.000	.
N	temp	62	62
	precipitation	62	62

There was a positive correlation between temperature and precipitation ($r = .871$, $n = 62$, $p < .001$).

Variables Entered/Removed(b)

Model	Variables Entered	Variables Removed	Method
1	precipitation (a)	.	Enter

a All requested variables entered.

b Dependent Variable: temp

Make sure you check this. This will tell you if you entered the correct predictor and dependent variable.

Model Summary(b)

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Durbin-Watson
1	.871(a)	.759	.755	15.451	1.543

a Predictors: (Constant), precipitation

b Dependent Variable: temp

We have already seen this table (Durbin-Watson statistic), when we were evaluating independence. This table also has important information about the overall fit of the model. First we look at R. R is the correlation. Look back at the correlations table. It is the same number. So, we have already addressed that there is a positive correlation.

R^2 : a.k.a. Coefficient of Determination; This value represents how much variability in the data can be accounted for by the predictor. 75.9% of the variability in temperature can be accounted for by precipitation. This is important! This is a good R^2 value, but it still means that 20.5% of the variability in temperature cannot be explained by precipitation. This means there are other factors influencing the temperature.

Adjusted R^2 : this will be important when there are multiple predictors.

Std. Error of the Estimate: the standard deviation of the residuals. In a good model, the SEE will be markedly less than the standard deviation of the dependent variable.

DV = time; standard deviation = 31.21

SEE = 15.451. The SEE is much smaller than the DV σ . This is good.

ANOVA(b)

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	45073.595	1	45073.595	188.802	.000(a)
	Residual	14324.082	60	238.735		
	Total	59397.677	61			

a Predictors: (Constant), precipitation

b Dependent Variable: temp

The ANOVA table tests the significance of the Model. This suggests that the slope of the regression line is significantly different than zero. This suggests that the model including precipitation significantly predicts temperature [$F(1,60) = 188.802, p < .001$], but it does not tell us by how much.

Coefficients(a)

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	-5.452	4.435		-1.229	.224
	precipitation	127.273	9.263	.871	13.741	.000

a Dependent Variable: temp

Unstandardized Coefficients: The coefficient table tells you how much precipitation influences temperature. For every unit increase in temperature, precipitation will increase by 127.273 units.

Beta: This is what we would get if we standardized all of the data. This value is also the same as the correlation coefficient and you can interpret it as its magnitude. In this example, the value is 0.871, or a difference of 0.871 standard deviations in precipitation per change of one standard deviation in temperature. That's a really strong relationship! This value will not be the same as the correlation coefficient when you have multiple predictors.

Write up

A meteorologist is interested in predicting today's weather based on precipitation. The meteorologist finds that precipitation and temperature are positively correlated ($r = .871, n = 62, p < .001$). He now knows that precipitation is a significant predictor of temperature [$F(1,60) = 188.802, p < .001$] and that 79.5% of the variance in temperature can be explained by precipitation. More specifically, for every unit increase in temperature, precipitation will increase by 127.273 units ($t = 13.741, p < .001$).

Now the meteorologist knows that precipitation is a good predictor of temperature. He knows what the precipitation is today (0.67%). Now he must predict the

average temperature for the day. We can use the linear regression equation to figure this out.

$$Y = B_0 + B_1X$$

$$\text{temperature} = -5.452 + 127.273(0.67)$$

$$\text{temperature} = -5.452 + 85.273$$

$$\text{temperature} = 79.821$$

The meteorologist would predict that the temperature for today will be 79.821° F.

What if the precipitation was 0.2?

$$\text{temperature} = -5.452 + 127.273(0.2)$$

$$\text{temperature} = -5.452 + 25.455$$

$$\text{temperature} = 20.453.$$

So, if the precipitation today is 0.2, the temperature should be 20.453° F.