Regression Analysis

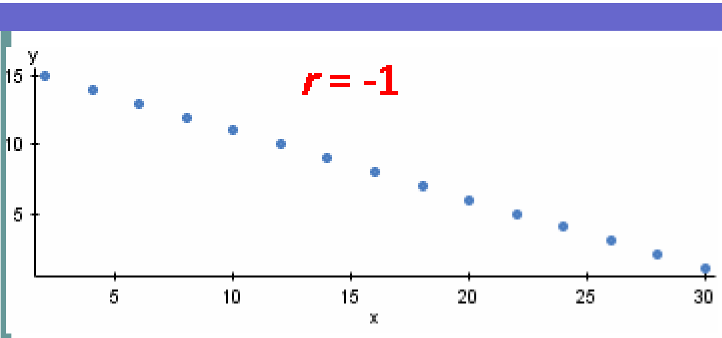
- Scatter Plots
- Correlation (r)
- r^2
- Hypothesis Test for ρ
- Correlation Vs. Causation

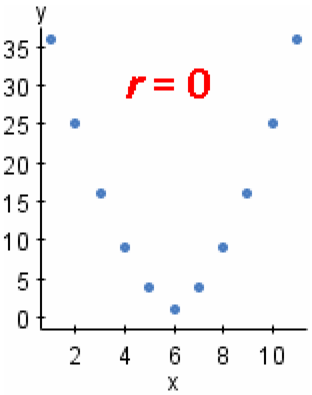
Correlation

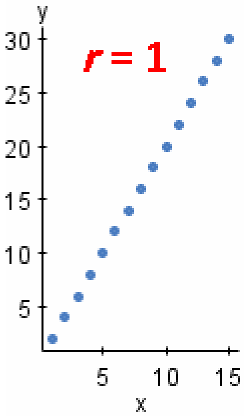
Correlation is a number that describes how close to a line the data lies. $-1 \leq r \leq 1$

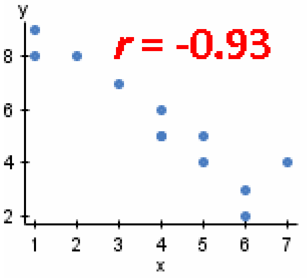
- If $r = -1$, the data is perfectly on negatively sloped line.
- If $r = 1$, the data is perfectly on a positively sloped line.
- If $r = 0$, then there is no line that is even close to describing the data.

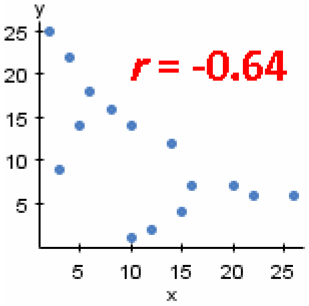
Examples of *r*

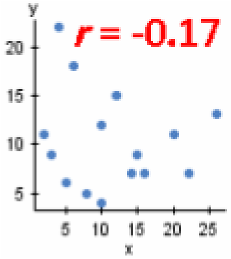


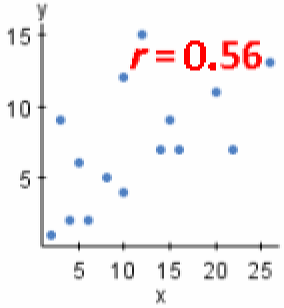


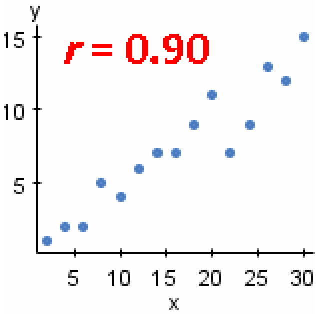












Scatter Plots and *r*

Sketch scatter plot that have the following correlations:

A. $r = 0.98$

B. $r = -0.02$

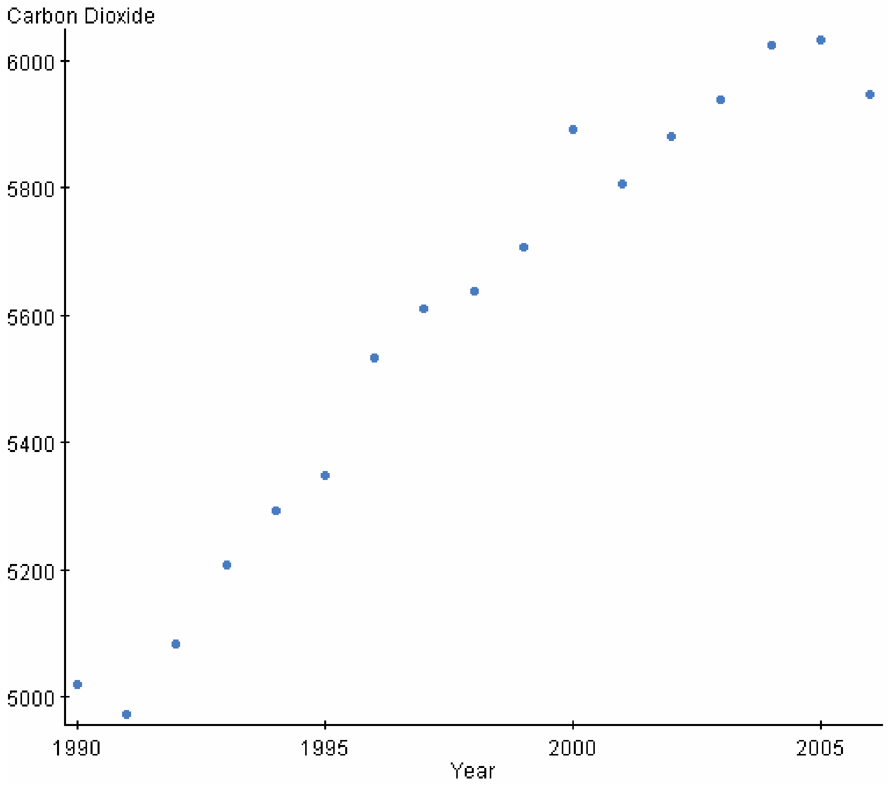
C. $r = 0.72$

D. $r = -0.23$

E. $r = -0.69$

Year vs. CO_2 Emissions

The Scatter Plot below shows the relationship between CO_2 emissions and the year. Discuss the correlation.



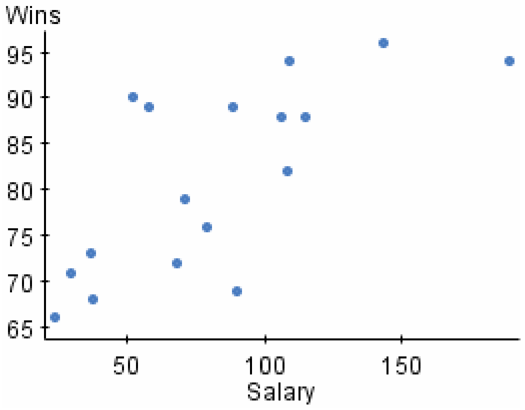
Baseball Wins vs. Salary

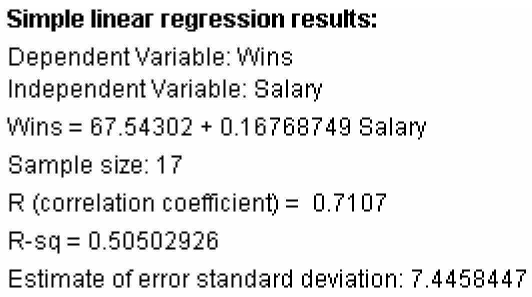
The table below gives the wins vs. salary in millions of major league baseball teams.

Salary	Wins
143	96
109	94
189	94
52	90
89	89
58	89
115	88
106	88
108	82
71	79
79	76
37	73
68	72
30	71
90	69
38	68
24	66

Baseball Wins vs. Salary: r^2

Some of the variation in the dependent variable can be explained by the independent variable while some of the variation in the dependent variable cannot be explained by the independent variable. r^2 is the proportion of variation in the dependent variable that can be explained by the independent variable.





```
Simple linear regression results:
Dependent Variable: Wins
Independent Variable: Salary
Wins = 67.54302 + 0.16768749 Salary
Sample size: 17
R (correlation coefficient) =  0.7107
R-sq = 0.50502926
Estimate of error standard deviation: 7.4458447
```

Hypothesis Test for ρ

- H_0: $\rho = 0$
- H_1: $\rho \neq 0$

Requirements:

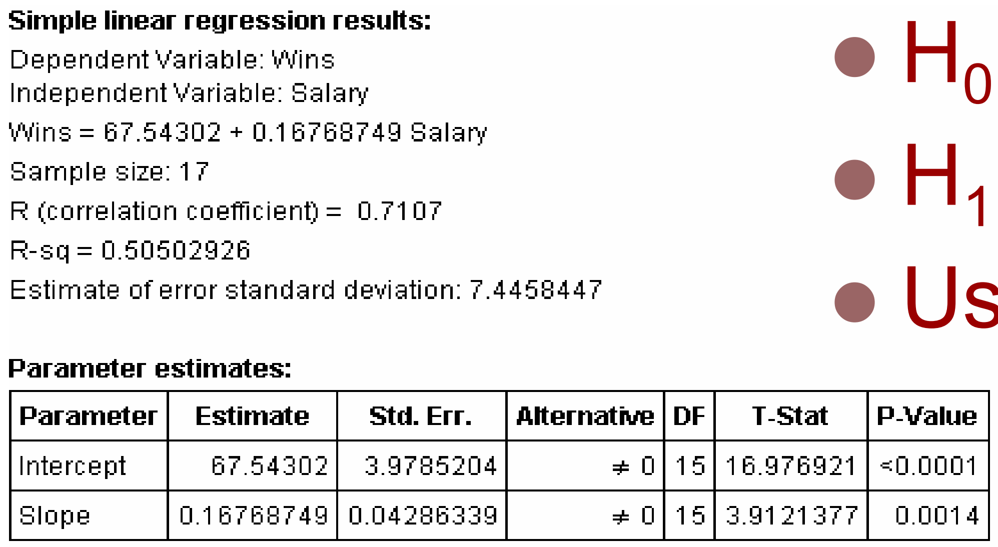
1. The population values of y for every individual value of x must follow approximately a normal distribution.
2. The pair (x,y) were gathered using simple random sampling.

TI 83/84: STAT → LinRegTTest

Conclusions: If P-Value < α, then there is statistically significant evidence to reject the null hypothesis and conclude that there is a linear correlation between x and y.

If P-Value > α, then fail to reject the null hypothesis and state that there is insufficient evidence to make a conclusion about there being a linear correlation between x and y.

Baseball Wins vs. Salary: r^2



Simple linear regression results:

Dependent Variable: Wins

Independent Variable: Salary

Wins = 67.54302 + 0.16768749 Salary

Sample size: 17

R (correlation coefficient) = 0.7107

R-sq = 0.50502926

Estimate of error standard deviation: 7.4458447

Parameter estimates:

Parameter	Estimate	Std. Err.	Alternative	DF	T-Stat	P-Value
Intercept	67.54302	3.9785204	≠ 0	15	16.976921	<0.0001
Slope	0.16768749	0.04286339	≠ 0	15	3.9121377	0.0014

- H_0: $\rho = 0$
- H_1: $\rho \neq 0$
- Use $\alpha = 0.05$

Correlation Does not Imply Causation

Correct:

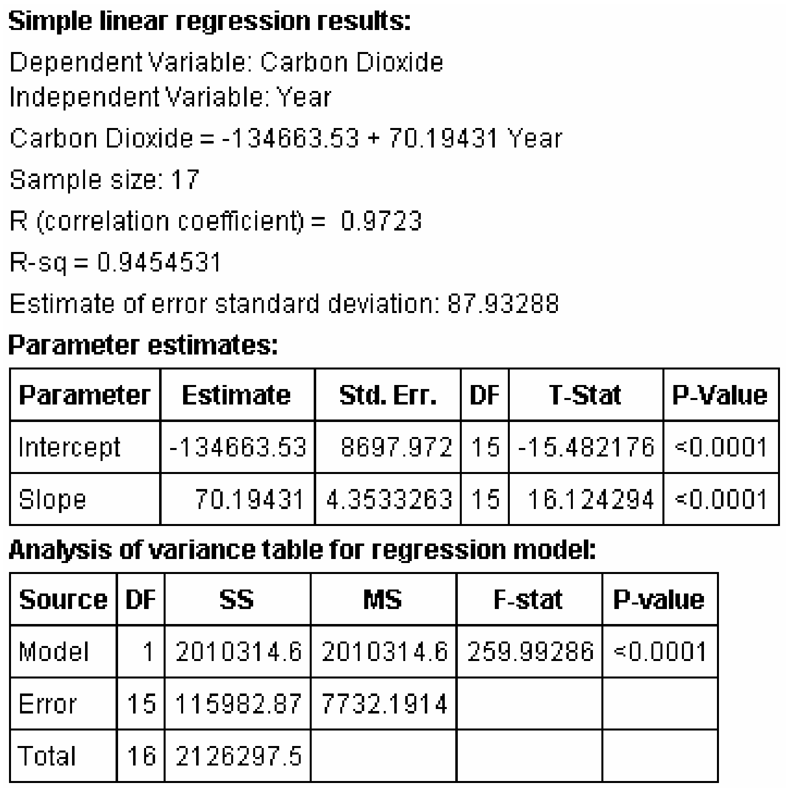
- There is a linear relationship between team baseball salaries and total wins.
- As the team baseball salaries increase the total wins tends to also increase.

Wrong:

- Increasing a team’s salary will make the team win more games.
- A salary increase will result in more wins for the team.
- If you want to win more games pay your players more money.

Year vs. CO$_2$ Emissions

The StatCrunch readout shows the regression analysis for the year vs. CO$_2$ emissions. Interpret r and r^2 and conduct the hypothesis test.



Simple linear regression results:

Dependent Variable: Carbon Dioxide
Independent Variable: Year

Carbon Dioxide = -134663.53 + 70.19431 Year

Sample size: 17

R (correlation coefficient) = 0.9723

R-sq = 0.9454531

Estimate of error standard deviation: 87.93288

Parameter estimates:

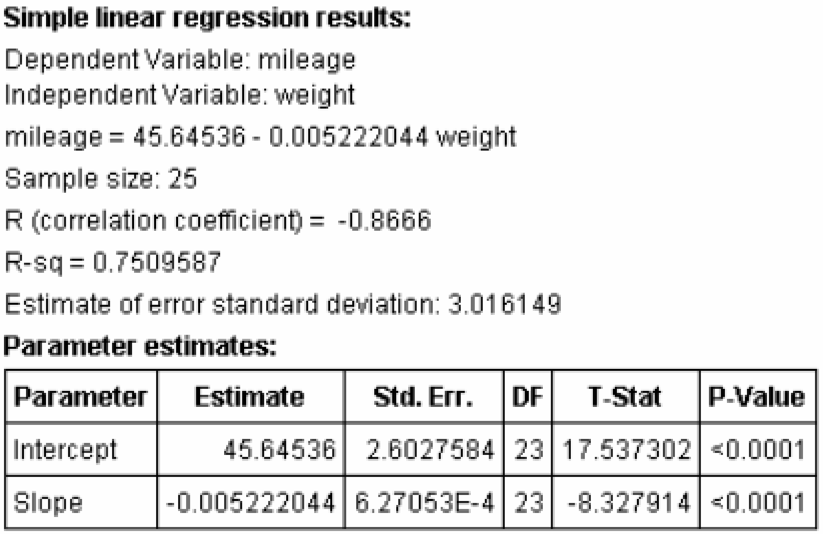
Parameter	Estimate	Std. Err.	DF	T-Stat	P-Value
Intercept	-134663.53	8697.972	15	-15.482176	<0.0001
Slope	70.19431	4.3533263	15	16.124294	<0.0001

Analysis of variance table for regression model:

Source	DF	SS	MS	F-stat	P-value
Model	1	2010314.6	2010314.6	259.99286	<0.0001
Error	15	115982.87	7732.1914		
Total	16	2126297.5			

Car Weight vs. Mileage

The StatCrunch readout shows the regression analysis for the weight of a car vs. gas mileage. Interpret r and r^2 and conduct the hypothesis test.



Simple linear regression results:

Dependent Variable: mileage
Independent Variable: weight

mileage = 45.64536 - 0.005222044 weight

Sample size: 25

R (correlation coefficient) = -0.8666

R-sq = 0.7509587

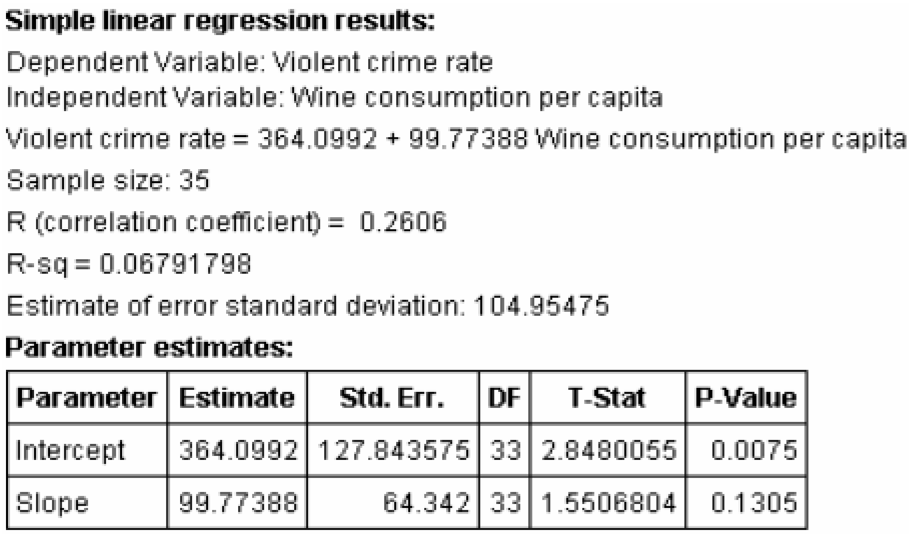
Estimate of error standard deviation: 3.016149

Parameter estimates:

Parameter	Estimate	Std. Err.	DF	T-Stat	P-Value
Intercept	45.64536	2.6027584	23	17.537302	<0.0001
Slope	-0.005222044	6.27053E-4	23	-8.327914	<0.0001

Wine Consumption vs. Crime

The StatCrunch readout shows the regression analysis for wine consumption per capita in cities and the city’s violent crime rate. Interpret r and r^2 and conduct the hypothesis test.



Simple linear regression results:

Dependent Variable: Violent crime rate

Independent Variable: Wine consumption per capita

Violent crime rate = 364.0992 + 99.77388 Wine consumption per capita

Sample size: 35

R (correlation coefficient) = 0.2606

R-sq = 0.06791798

Estimate of error standard deviation: 104.95475

Parameter estimates:

Parameter	Estimate	Std. Err.	DF	T-Stat	P-Value
Intercept	364.0992	127.843575	33	2.8480055	0.0075
Slope	99.77388	64.342	33	1.5506804	0.1305