

Chapter 3 Association: Contingency, Correlation, and Regression

When studying the association between two variables, we usually want to distinguish between:

- **explanatory** (or predictor) variable
- **response** variable

3.1 The Association between Two Categorical Variables

Contingency Tables:

- Both explanatory and response variables are categorical.
- Display counts (frequencies) on the table.
- Compute percentages to determine association.

Conditional Proportions: Find percentages by dividing each cell count by the total number of observations in their group (as defined by the explanatory variable).

Example: Do male college students follow their school's teams more closely than females? The following data was collected in class on a Monday morning, after a particularly exciting and important basketball game. The question asked was: Did you watch the game on TV last night?

	Whole Game	Part of the Game	None of It	Total
Male	10	12	4	26
Female	21	24	30	75
Total	31	36	34	101

a) What is the explanatory variable? **Gender (M/F)**

b) What is the response variable? **Game watching (whole, part, none)**

c) Find the conditional proportions of each gender that watched all, part, or none of the game.

	Whole Game	Part of the Game	None of It	Total
Male	$10/26 = 0.385$	$12/26 = 0.462$	$4/26 = 0.154$	1.00
Female	$21/75 = 0.28$	$24/75 = 0.32$	$30/75 = 0.40$	1.00

d) Is it fair to say that males were more likely to watch the game than females?

YES – even though the **number** of females who watched the whole game or part of it was larger than the number of men, the **percentages** were higher for males. Warning – this is only true for students who came to class that day, since some may have stayed up late watching the game and celebrating, and missed the 9:30 am class.

3.2 The Association between Two Quantitative Variables

Scatterplots

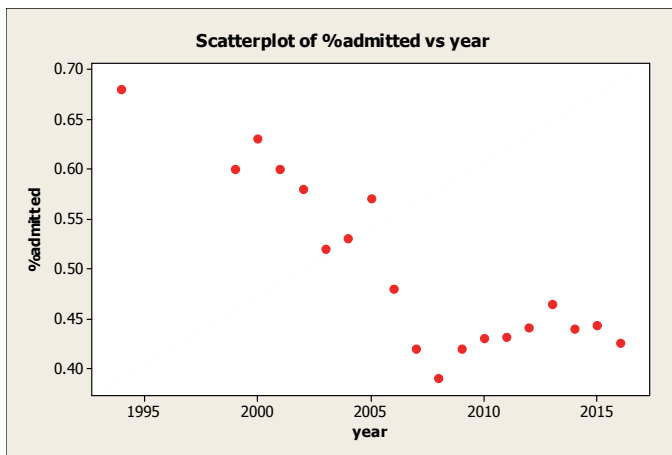
- Plot of Y vs. X, two quantitative variables, measured on the same individual
- X = explanatory variable, Y = response variable

Interpreting Scatterplots

- Direction – positive or negative?
- Linear trend? How strong?
- Any outliers?

Examples: Interpret the following scatterplots.

1. X= year and Y= percentage of Freshmen applicants admitted into UF, as reported on the UF website every year by the Office of Institutional Planning and Research.

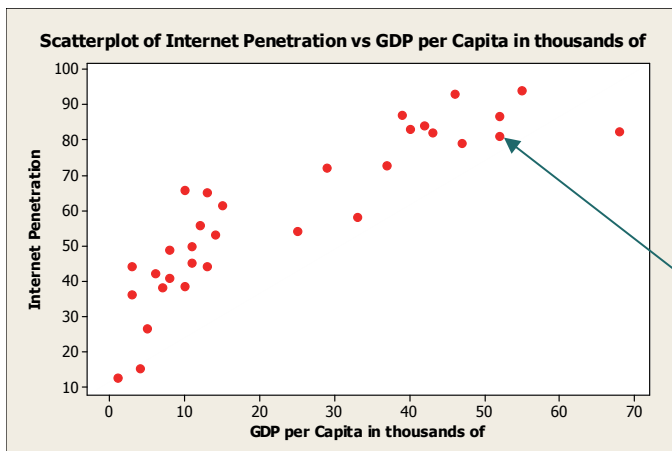


Fairly strong linear trend

Negative association – as the year increases, the percentage of Freshmen admitted decreases, on average.

But the fast decrease from the 1900's 60% to around 2008's low of under 40% seems to have stabilized around the mid 40% lately.

2. Y= percentage of population with Internet access in a country
X= GDP = country's Gross Domestic Product per capita (in thousands of US dollars)
Find the US: GDP=52.0 and 81.03% of adults have Internet access (Data from AFK)



Positive association – the higher GDP (richer country), the higher rate of Internet access, on average

Fairly strong linear association - no major outliers

US

Example: Determine which variable should be explanatory (x) and response (y), and sketch a scatterplot of the relationship you would expect between the father's height and his adult sons' heights.

x= fathers ht

y= sons ht

Scatterplot - positive, fairly strong

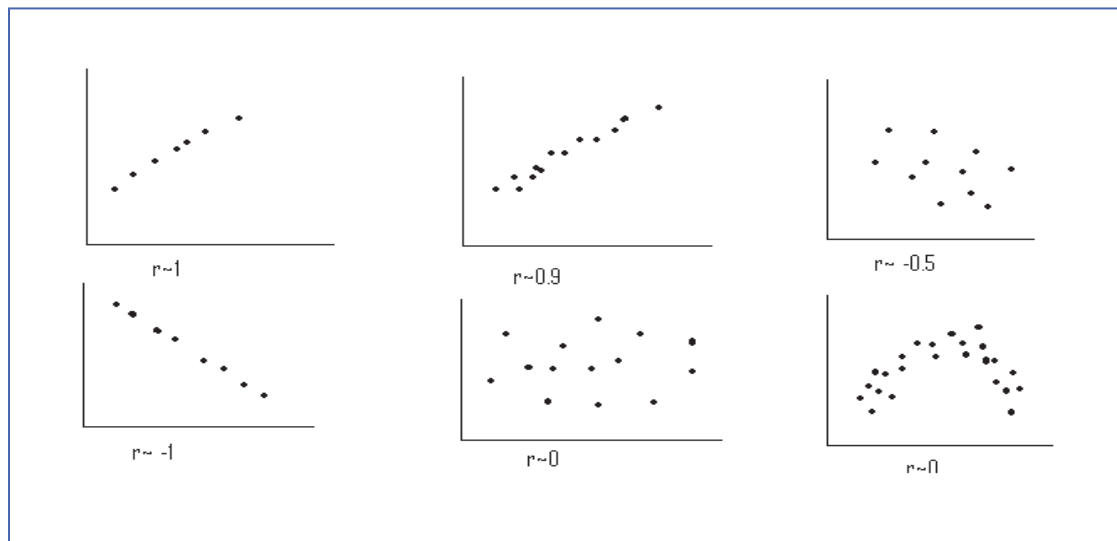
Example: Consider the relationship between father's height and mother's height. Would that pair of variables have a stronger or weaker linear relationship than the variables on the example above?

Much weaker

Correlation

- The correlation summarizes the **direction** and **strength** of the straight line relationship between x and y.
- The two variables have the same correlation regardless of which one is called the explanatory or the response variable.
- We will use the symbol **r** to represent the correlation coefficient.
- r is always between -1 and +1 (no units).
- **Interpretation:** positive/negative, strong/weak.
- Outliers can have a strong effect on r.

Some examples:



Formula: Correlation Coefficient

Note: The formula for the correlation coefficient is mostly for illustration purposes. For homework problems, you can use Minitab, or your calculator (if it does 2-variable statistics).

$$r = \frac{1}{n-1} \sum \left(\frac{x - \bar{x}}{s_x} \right) \left(\frac{y - \bar{y}}{s_y} \right)$$

3.3 Predicting the Outcome of a Variable

Review of Straight Lines:

- All points are exactly on the line
- Line extends forever in both directions
- Equation of a Line:
 $y=mx+b$ $m=\text{slope}$ $b=\text{y-intercept}$

In Statistics – Regression Line:

- Points are scattered
- Want the equation of the line that best fits through the middle of the points
- Use it to predict the response variable y for a particular value of x .
- We call the predicted values \hat{y} .
- Regression Equation: $\hat{y} = a + bx$

Interpreting the Regression Line:

- **b is the slope** (rise/run)
The slope represents the average (or predicted) change in y for a one-unit change in x .
- **a is the y-intercept** (the point where the regression line crosses the y axis)
The y -intercept corresponds to the predicted value of y when $x=0$ and it is necessary to complete the equation. However, we only interpret it if $x=0$ makes sense and it is close to the values of x observed.

Example: Suppose the regression equation to predict y = weight (in pounds) from x = height (in inches) for female college students is $\hat{y} = -200 + 5x$.

a) Interpret the equation.

Slope: 5

As height increases by 1 inch, weight increases, on average, by five pounds.

Intercept: -200

Do not interpret. $X=0$ inches tall for a college woman does not make sense. Trying to use the equation to predict her weight gives us the ridiculous answer of negative 200 pounds.

b) Predict the weight of a female college student whose height is 5'5" (65").

Answer: $\hat{y} = -200 + 5(65) = 125$

According to this regression equation, the average (or predicted) weight for 65" tall women is 125 pounds.

c) Roughly sketch the relationship between height and weight for college-aged women. Do you expect all 65" tall women to weight the same?

There will be variability in the weights for women of the same height.

Expect the graph to show a positive, fairly strong linear relationship between $x = \text{ht}$ and $y = \text{wt}$.

Using Data to Find the Slope and y-Intercept of a Regression Line

There are many lines that, visually, seem to fit a scatterplot well. So how do we find the "best" line to describe a data set? The least squares regression method finds the line that minimizes the prediction errors.

Residuals

- Residuals are the prediction errors for each observation.
- Graphically, they are the vertical distance from the point to the line.
- Residuals = difference between the observed and predicted values of y .
- Residuals = $y - \hat{y}$

Least Squares Regression Method

- The least squares regression line goes through the middle of the points, in the sense that the distances to the points above and below the line cancel each other. That means the sum of the residuals will be zero.
- The least squares regression line also minimizes the sum of squared residuals, or prediction errors. The formulas for slope and intercept are discussed below – they are not hard to derive, but it requires knowledge of calculus.
- Because it is the vertical distances that are minimized, it is important to determine which variable is x and which one is y.
- The least squares regression line passes through the point (\bar{x}, \bar{y}) .

Least Squares Regression Formulas:

$\hat{y} = a + bx$ where b is the SLOPE (unlike high school where $y=mx+b$), a is y-int

$b = r sy/sx$ (change in y over change in x times the correlation)

$a = \bar{y} - b\bar{x}$ (uses the slope and the fact LSR always goes through (\bar{x}, \bar{y}))

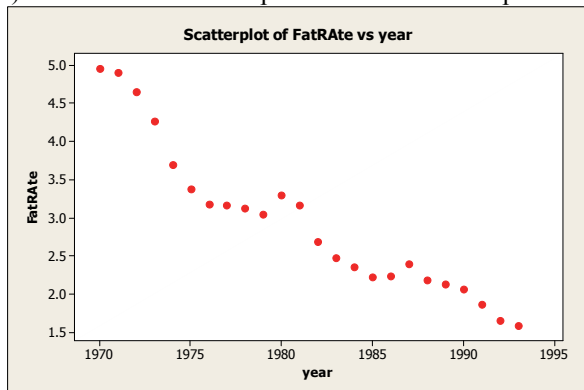
R²: Coefficient of Determination

- $R^2 = (r)^2 = (\text{correlation})^2$
- Easier to interpret than the correlation coefficient.
- It is interpreted as the percent of variability in y explained by the linear regression on x.

Example: After more than 20 years with a 55 mph speed limit, the state of New York raised the speed limit to 65 mph on some of its major highways, effective August 1, 1995. Will highway fatalities increase with the speed limit? Some people say yes, pointing to the reduction in deaths when the 55 mph limit was implemented in 1974. Others say no, pointing to better roads and car equipment (tires, anti-lock brakes, air bags) that help save lives. We will use data on highway fatalities and fatality rates (number of deaths per 100 million vehicle miles traveled) from 1970 to 1993 to do a regression analysis. (NHTSA/DOT)

a) Why is it better to use fatality rate instead of fatalities?
It adjusts for population increase and more cars on the road.

b) Describe the most important features of the plot of fatality rate vs. year that appears below.



Negative, strong correlation between year and death rate.

Not quite linear, there are “blips” in the data.

c) Here are some other laws passed during the period being studied. Do they seem to have affected highway fatalities?
They seem to roughly correspond to the “blips”.

- 1974 Speed limit reduced to 55 mph
- 1981 Anti DWI law, late November
- 1982 19 yr old drinking age on Dec 1
- 1985 Mandatory seat belt law on Jan 1
- 1985 21 yr old drinking age on Dec 1

- d) Interpret $R^2 = 90.9\%$

90.9% of the variability in death rate is explained by year.

- e) Compute the correlation coefficient, r , and interpret.

$$r = \sqrt{0.909} = -0.953 \quad \text{Negative, strong correlation}$$

Important: write R^2 as a decimal AND look at the plot to decide if correlation should be positive or negative.

- f) Use the following information to find the regression equation:

	year	death rate
mean	81.5	2.939
stdev	7.07	0.987

$$\hat{y} = 13.78 - 0.133x$$

- g) Plot the regression line on the scatterplot above. Comment on the fit of the regression line.

Find two points on the graph. Be careful with the coding of year.

$$x=70 \quad \hat{y} = 4.47$$

$$x=90 \quad \hat{y} = 1.81$$

Line should go through the middle of the points.

- h) Interpret the slope and the intercept (if appropriate).

Slope = -0.133 Each year, the death rate on highways decreases by .133 on average

Intercept = 13.78 $x=0$ corresponds to the year 1900. We should NOT interpret – this date is too far from those observed, and there were no highways back then (and there were very few cars, they were not mass produced yet).

- i) What fatality rate does the line predict for 1996? Do you trust this prediction? Explain.

$$x=96 \quad \hat{y} = 1.012$$

Although this date is not too far from the data collected (1993), the speed limit had just changed in 1995, so we expect a blip on the graph.

- j) The fatality rate for 1996 was 1.34. Find the residual for this point and interpret.

$$\text{residual} = \text{observed } y - \text{predicted } y = y - \hat{y} = 1.34 - 1.012 = 0.328$$

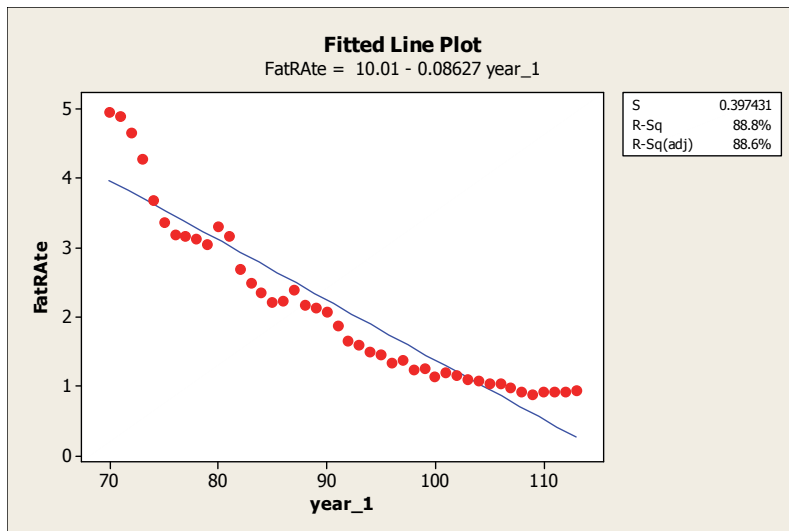
Positive, so the death rate was 0.328 higher than expected.

- k) What fatality rate does the line predict for 2006? Do you trust this prediction? Explain.

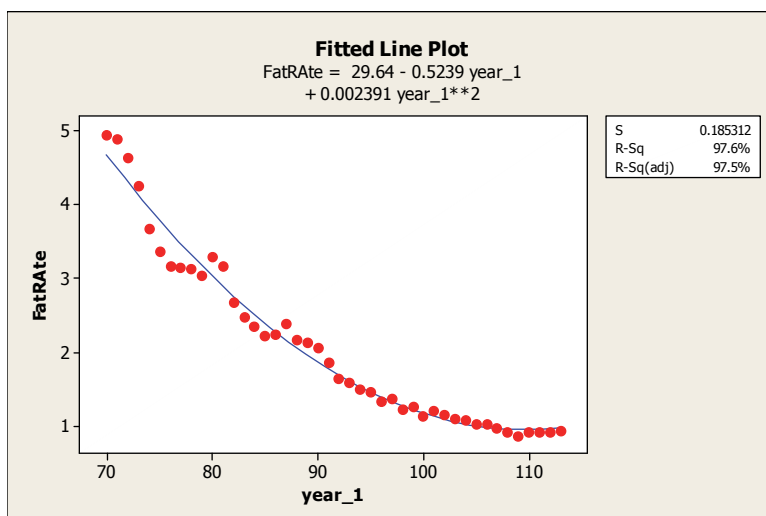
$$x=106 \quad \hat{y} = -0.316$$

Do not trust it; negative death rate makes no sense.

The graph below shows the results of the regression analysis that incorporates data for more recent years.



- l) How does this regression line compare to the old one?
 Slope is less steep, R^2 a bit lower
- m) Will the fatality rate ever reach zero in practice? Explain.
 Not as long as there are cars and highways.
- n) What fatality rate does this regression line predict for 2006?
 0.866
 Predictions will reach zero soon.
- o) What fatality rate does this regression line predict for 2020?
 -.34
 Negative death rate, does not make any sense
- p) What could be done to improve our predictions?
 Perhaps a curve that flattens out and never reaches zero.
- q) The plot below shows the results of the quadratic regression. Comment on the fit of this regression curve.



Much better fit and R^2

3.4 Cautions in Analyzing Associations

Extrapolation

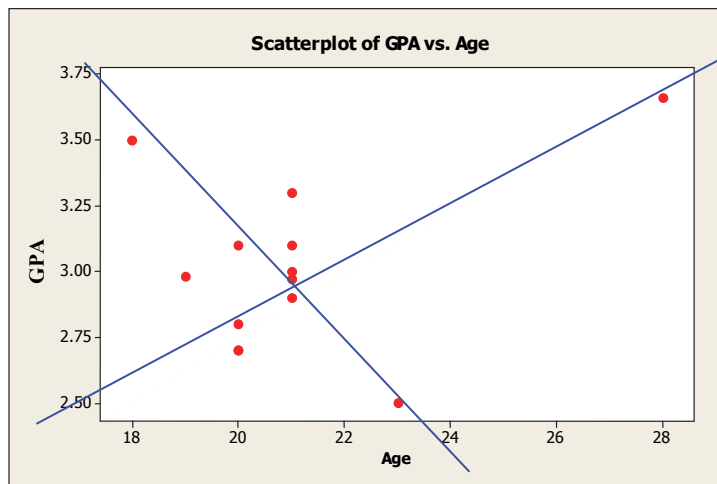
Predictions made using the regression equation can only be trusted for values of X within the observed range. Predicting outside this range is called **extrapolation**, and you might get ridiculous predictions.

Example: predicting death rate in NY for the year 2020, or even interpreting the intercept in many cases.

Influential Outliers

Outliers in regression are those points that are way away from the trend of the other observations. Influential outliers are only points that have an x value far away from the rest, and that fall far from the trend that the rest of the data follow. These points tend to pull the line towards them, and deleting them can have a huge effect on the regression line, sometimes even changing the sign of the slope.

Example: The following plot has the college GPA and age of several students. Sketch what you expect the least squares regression equation for this data will be, and then do the same thing after removing the 28-year old student. Comment on the difference between the two lines.



With outlier – positive slope (older students get higher GPA's).

Without outlier – negative slope (older students get lower GPA's).

The person who is 28 yr-old is much older than the rest. His observation does NOT follow same trend as the rest. He is an Influential Outlier.

Do not trust either regression line.

What to do if your data has outliers?

- Check the data and correct any typos.
- If there are still unusual observations, try to find out more about them. Do they belong in the data set? What makes them different? If they do not belong in the data set, you should delete the point before proceeding with the regression analysis.
- If the point is valid, conduct the regression analysis with and without that point. If their results are similar, you may use them. If they are very different, you should collect more data to find out the true relationship between x and y.

Correlation (or Association) Does Not Imply Causation

In many studies, the goal is to prove that changes in x cause changes in y. However, even strong association does not imply causation. It is very hard to prove causation, since there are usually other **lurking variables** that can affect the relationship between x and y.

Examples:

1. Newspaper report: Decaf drinkers have higher blood pressure levels than regular coffee or non-coffee drinkers. Is decaf bad for your health?
 - explanatory variable: type of coffee people choose to drink
 - response variable: blood pressure levels
 - graphical summary: side-by-side boxplots (categorical x, quantitative y)
 - potential lurking variables: blood pressure levels before they chose their drink
 - conclusion: Cannot prove decaf is bad for you. Perhaps decaf drinkers already had high blood pressure and their doctor advised them to switch from regular to decaf.
2. Very high positive correlation is found between the size of the head of school-aged children and their reading skills. Are big-headed kids smarter?
 - explanatory variable: size of head
 - response variable: test of reading skills
 - graphical summary: scatterplot (quantitative x, quantitative y)
 - potential lurking variables: child's age
 - conclusion: Cannot prove big headed kids are smarter (or read better). School-aged children range from kindergarteners who are just learning to read, to at least 5th graders who should be a lot better readers.
3. Thousands of studies over the years have found that smokers have a much higher incidence of lung cancer than non-smokers. Does smoking cause cancer?
 - explanatory variable: smoking (yes/no)
 - response variable: cancer (yes/no)
 - graphical summary: bar charts or contingency table (categorical x, categorical y)
 - potential lurking variables: age, gender, weight, eating, drinking, and exercise habits, genetics, second-hand smoke, exposure to other toxic substances, etc
 - conclusion: Only after thousands of studies, all over the world, taking all of these and other lurking variables into account, have found consistently that the incidence of cancer is higher for smokers than non-smokers have we come to accept that smoking causes cancer.

Simpson's Paradox

The direction of an association between two categorical variables can be reversed if we include a third variable and re-analyze the data. This is known as Simpson's Paradox.

Example: A study in the United Kingdom asked women if they smoked or not, and twenty years later determined if they were alive or not. Data appears on the table below. (AFK)

Smoker	Dead	Alive	Total
Yes	139	443	582
No	230	502	732

- a) Compute the percentage of smokers and non-smokers who died. What does the data suggest about the effects of smoking? Is this surprising?
- 24% of smokers and 31% of non-smokers died.
Looks like smoking is good for you, NOT what we expected.

- b) A potential lurking variable in this study is the age of the women at the beginning of the study. This information was also available, and has been included in the table below. Compute the percentage of smokers and non-smokers who died for each age group. What does the data suggest about the effects of smoking now?

	18-34 yrs old		35-54 yrs old		55-64 yrs old		65+ yrs old	
Smoker	Dead	Alive	Dead	Alive	Dead	Alive	Dead	Alive
Yes	5	174	41	198	51	64	42	7
No	6	213	19	180	40	81	165	28

Smoker	18-34 yrs old	35-54 yrs old	55-64 yrs old	65+ yrs old
Yes	$5/179 = 2.8\%$	17.20%	44.30%	85.70%
No	$6/219 = 2.7\%$	9.50%	33.10%	85.50%

For each age group, the percentage of people who died is higher for smokers than non-smokers. Now smoking seems bad for you. Simpson's Paradox occurred because older people were more likely to die, but they were also more likely to be non-smokers, so grouping women across all age groups obscured the effect of smoking, that is smoking and age were **confounded**.

- c) For each age group, compute the difference in proportion of smokers and non-smokers who died. Then compute the ratio of the two proportions (called the **Relative Risk**). Which comparison, difference or ratio, best illustrates the effect of smoking on these women?

	18-34 yrs old	35-54 yrs old	55-64 yrs old	65+ yrs old
Difference	$2.8\% - 2.7\% = 0.1\%$	7.7%	11.2%	0.2%
RR	$2.8\%/2.7\% = 1.04$	1.81	1.34	1.002

For the youngest and oldest groups, the difference in proportion dead for smokers and non-smokers is very small. The relative risk for both these age groups is close to one, which means that the chance of dying for smokers is almost the same as for non-smokers.

The effects of smoking are seen in the two middle groups. For women who started the study 35-54 yrs old, the proportion that had died after 20 years in the smoking group was 7.7% higher than in the non-smoking group. But the RELATIVE RISK says that, for this age group, the probability of dying for smokers is 1.81 times the probability of dying for non-smokers (almost twice as high).

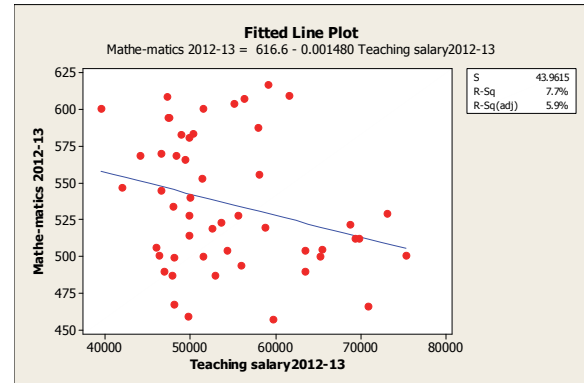
For women who started the study 55-64 yrs old, the proportion that had died after 20 years in the smoking group was 11.2% higher than in the non-smoking group. But the RELATIVE RISK says that, for this age group, the probability of dying for smokers is 1.34 times the probability of dying for non-smokers.

The RELATIVE RISK is a better way to compare these probabilities of dying.

Example: Teacher salaries vary dramatically from state to state. What is their relationship to student learning? Data from the National Center for Educational Statistics for the 2012-2013 school year was used for the following analyses.

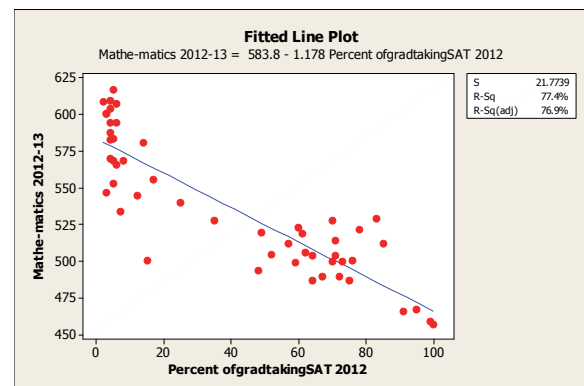
- a) The first plot represents the regression of average SAT Math score vs average teacher salary for each state. Describe the relationship between the two variables.

Weak, negative correlation – it seems like higher teacher salary is weakly associated with lower average SAT scores!



- b) The SAT is not the only college entrance test. In some states, the ACT is more popular and only students considering going to college out of state would take the SAT. The second plot represents the regression of average SAT Math score vs percentage of students who took the test, for each state. Describe the relationship between the two variables.

Fairly strong, negative correlation – states where most students take the SAT have lower average SAT scores



- c) The third plot represents the regression of average SAT Math score vs average teacher salary for each state, but now the states have been split up into those states where the percentage of students taking the SAT is Low, Medium or High. Describe the relationship between the three variables. Explain how this is an example of Simpson's Paradox

For states where a LOW or HIGH percentage of students take the test, average SAT scores are positively correlated with teacher salary. For states where a MED percentage of students take the test, average SAT scores seem very weakly or negatively correlated with teacher salary.

At least for states with LOW and HIGH percentage taking the SAT, the relationship between average SAT scores and teacher salary was reversed when we included the third (lurking) variable.

