Multiple Regression

Cautions About Simple Linear Regression

Correlation and regression are powerful tools for describing relationship between two variables, but be aware of their limitations

- Correlation and regression describe only linear relations
- Correlation and least-squares regression line are not resistant to outliers
- Predictions outside the range of observed data are often inaccurate
- Relationship between two variables often influenced by lurking variables not included in our model

Momentous sprint at the 2156 Olympics?

Women sprinters are closing the gap on men and may one day overtake them.

The 2004 Olympic women's 100-metre sprint champion, Yuliya Nesterenko, is assured of fame and fortune. But we show here that — if current trends continue — it is the winner of the event in the 2156 Olympics whose name will be etched in sporting history forever, because this may be the first occasion on which the race is won in a faster time than the men's event.

The Athens Olympic Games could be viewed as another giant experiment in human athletic achievement. Are women narrowing the gap with men, or falling further behind? Some argue that the gains made by women in running events between the 1930s and the 1980s are decreasing as the women's achievements plateau[1]. Others contend that there is no evidence that athletes, male or female, are reaching the limits of their potential[1,2].

In a limited test, we plot the winning times of the men's and women's Olympic finals over the past 100 years (ref. 3; for data set, see sup-

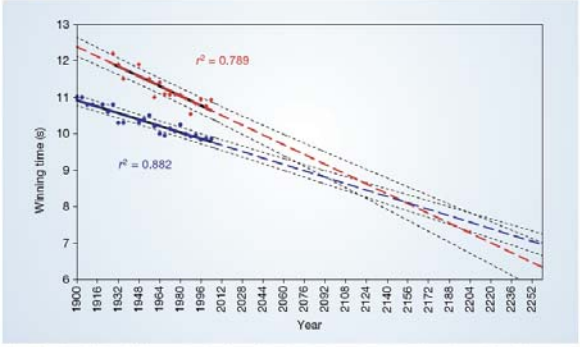
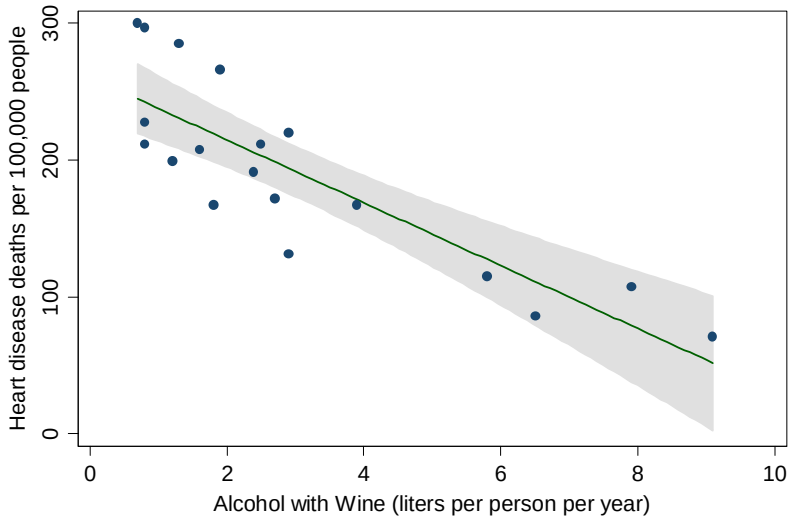


Figure 1 The winning Olympic 100-metre sprint times for men (blue points) and women (red points), with superimposed best-fit linear regression lines (solid black lines) and coefficients of determination. The regression lines are extrapolated (broken blue and red lines for men and women, respectively) and 95% confidence intervals (dotted black lines) based on the available points are superimposed. The projections intersect just before the 2156 Olympics, when the winning women's 100-metre sprint time of 8.079 s will be faster than the men's at 8.098 s.

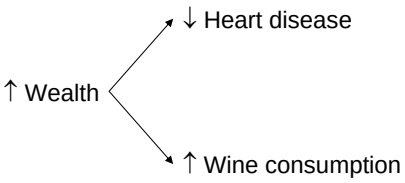
Least-Squares Regression of Heart Disease and Wine Consumption



Does this regression provide strong evidence that increased wine consumption lowers the risk of heart disease?

no

Lurking variables



Ecological fallacy

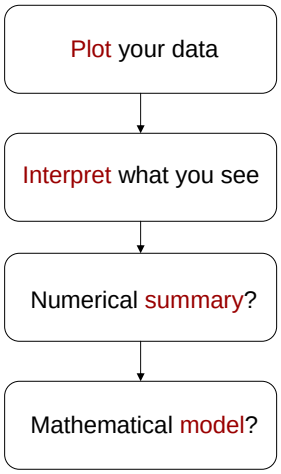
We can't make inferences about what individuals do, based on aggregate data

Are individuals who drink more wine suffering less heart disease?

General Principles of Data Analysis

Plot your data	To understand the data, always start with a series of graphs
Interpret what you see	Look for overall pattern and deviations from that pattern
Numerical summary?	Choose an appropriate measure to describe the pattern and deviation
Mathematical model?	If the pattern is regular, summarize the data in a compact mathematical model

Analysis of Two Quantitative Variables



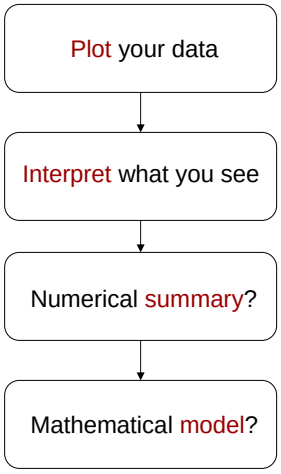
For two quantitative variables, use a scatterplot

Describe the direction, form, and strength of the relationship

If pattern is roughly linear, summarize with correlation, means, and standard deviations

Regression gives a compact model of overall pattern, if relationship is roughly linear

Analysis of Three or More Quantitative Variables

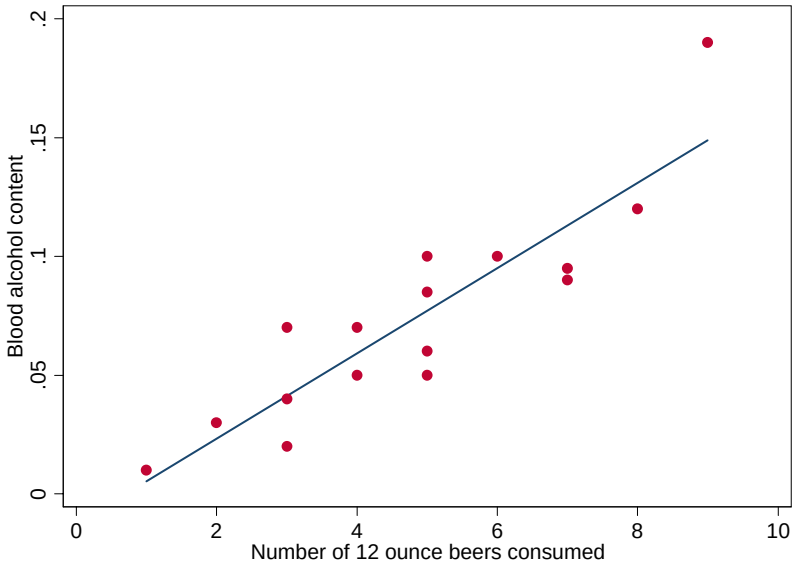


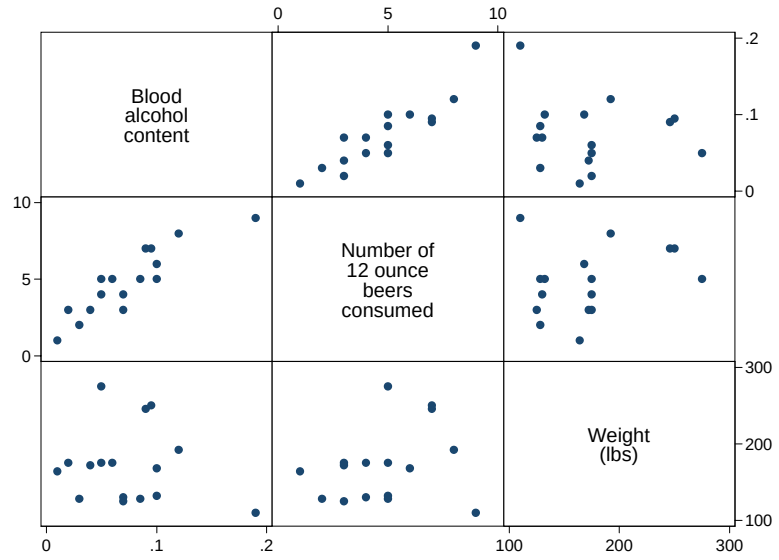
To examine relationships among all possible pairs use a scatterplot matrix

Describe the direction, form, and strength of the relationships

If pattern is roughly linear, summarize with correlations, means, and standard deviations

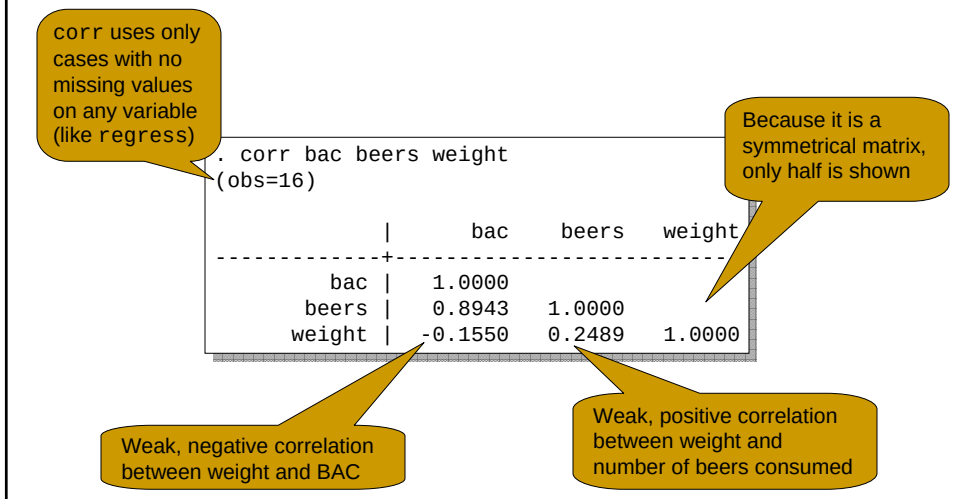
Multiple regression gives a compact model of relationship between response variable and a set of predictors



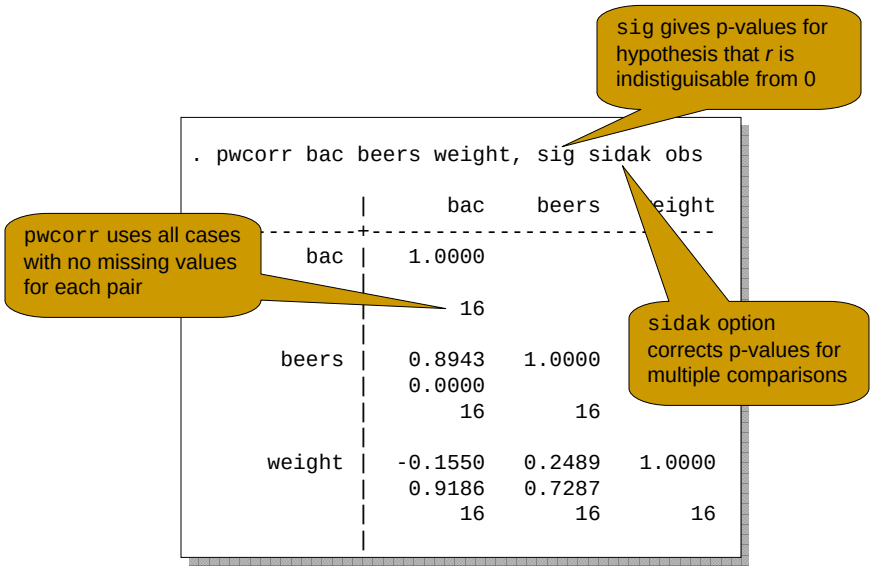


In Stata, obtain this graph with `graph matrix bac beers weight`

Correlation Matrix in Stata



Correlation Matrix in Stata



Multiple Regression in Stata

Overall *F*-test of model

```
. regress bac beers weight

      Source |       SS       df       MS              Number of obs =      16
-------------+------------------------------           F(  2,    13) =  128.33
       Model |  .027816116     2  .013908058           Prob > F      =  0.0000
    Residual |  .001408883    13  .000108376           R-squared     =  0.9518
-------------+------------------------------           Adj R-squared =  0.9444
       Total |    .029225    15  .001948333           Root MSE      =  .01041

------------------------------------------------------------------------------
         bac |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       beers |   .0199757   .0012629    15.82   0.000     .0172474     .022704
      weight |  -.0003628   .0000567    -6.40   0.000    -.0004853   -.0002404
       _cons |   .0398634   .0104333     3.82   0.002     .0173236    .0624031
------------------------------------------------------------------------------
```

R^2

slope, b_1

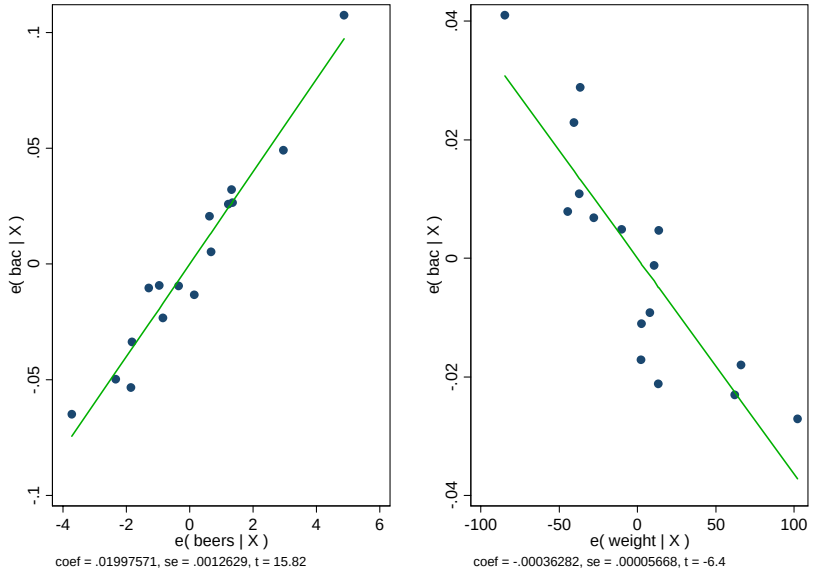
slope, b_2

y-intercept, a

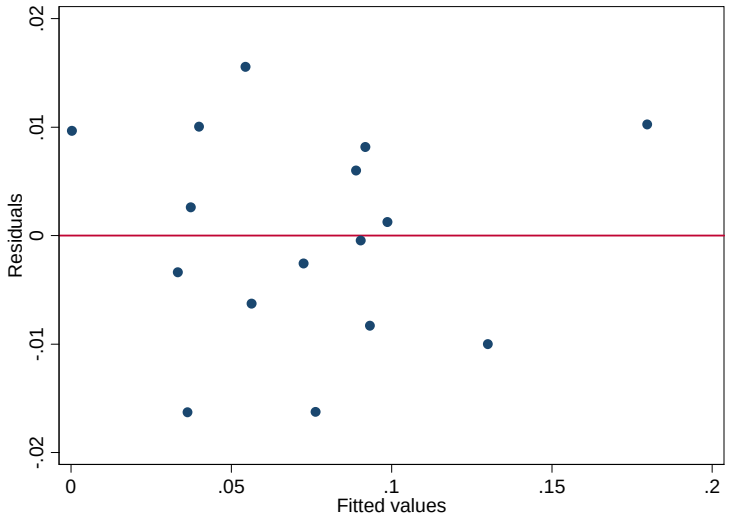
$$\hat{y} = a + b_1x_1 + b_2x_2$$

Estimated BAC = .0398 + (.0200)(Beers consumed) – (.0003)(Weight)

In Stata, obtain added-variable plots with `avplots`

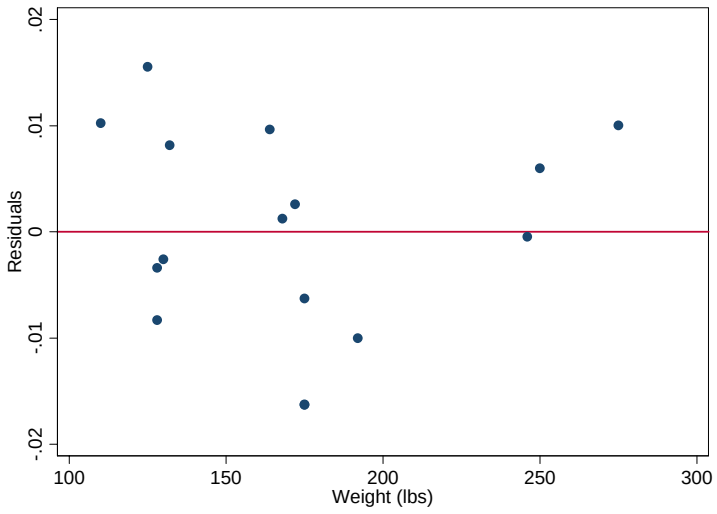


Residuals-versus-Fitted Plot



In Stata, obtain this plot after `regress` with `rvfplot, yline(0)`

Residuals-versus-Predictor Plot



In Stata, obtain this plot after `regress` with `rvpplot weight, yline(0)`