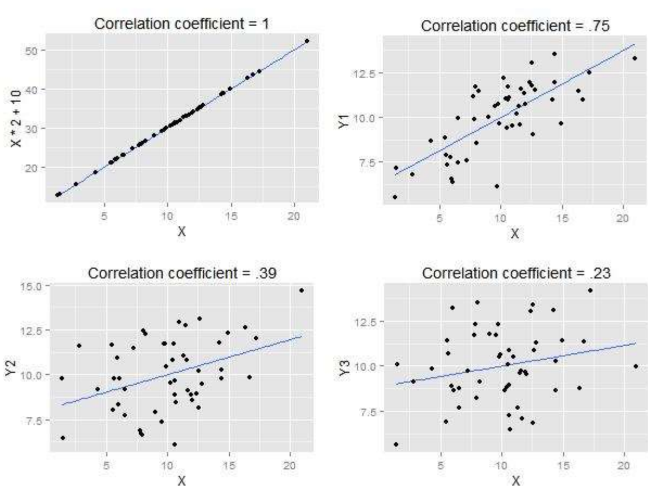
Regression Analysis

Regression

- Correlation implies an invisible line
- If we want to see that line, we use regression

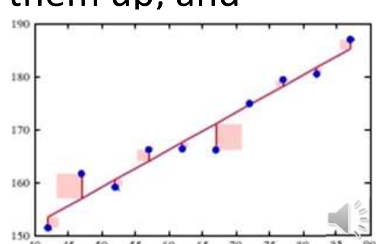


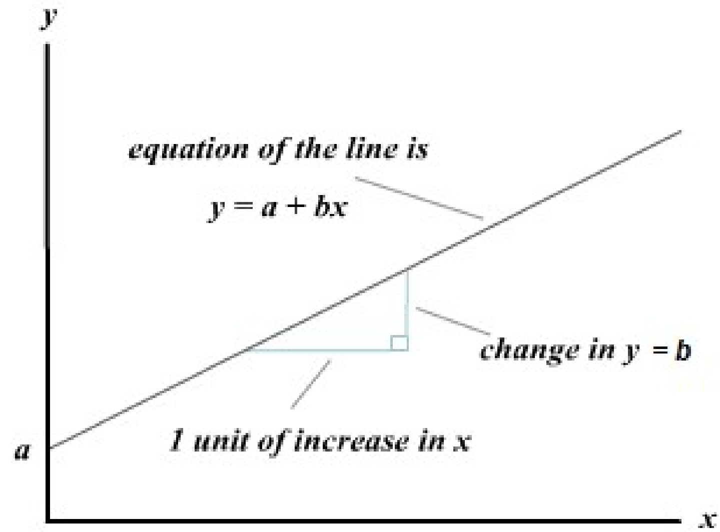
Dependent and Independent Variable

- Two variables → bivariate regression
- For correlation, X and Y are equal partners
- For regression, X is used to predict Y
- X = independent variable, Y = dependent variable
- We say: “We regressed Y on X”
- Both X and Y are interval/ratio

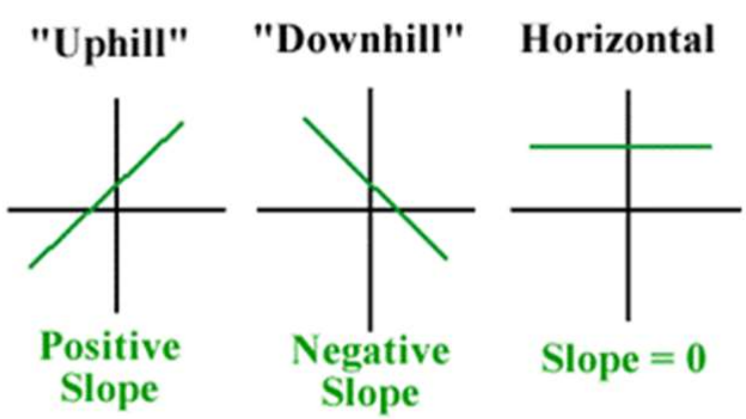
Finding the Best Fit

- Bivariate regression line is like a two-dimensional mean - it runs though the middle (equal spread on each side)
- Line of best fit -- minimizing the distances between all data points and the line
- “Least squares” regression – square all the distances from the line, add them up, and minimize that value





- a = intercept → value of Y when X is 0
- b = slope → if X increases by one unit, Y increases by b units



Formulas for Regression Slope and Intercept

$$b = \frac{n\left(\sum XY\right) - \left(\sum X\right)\left(\sum Y\right)}{n\left(\sum X^2\right) - \left(\sum X\right)^2}$$

$$a = \frac{\left(\sum Y\right) - b\left(\sum X\right)}{n}$$

Example of Calculation

	X	*Y*	XY	X^2
	5	2	10	25
	3	4	12	9
	7	1	7	49
	2	6	12	4
	4	5	20	16
	6	2	12	36
	4	3	12	16
	2	7	14	4
	8	1	8	64
	1	6	6	1
Σ	**42**	**37**	**113**	**224**

Example of Calculation

	X	Y	XY	X^2
Σ	42	37	113	224

$$b=\frac{10*113 \quad\quad *37}{10*224- \quad\quad *42}=-0.89$$

$$b=\frac{n\left(\sum XY\right)-\left(\sum X\right)\left(\sum Y\right)}{n\left(\sum X^2\right)-\left(\sum X\right)^2}$$

$$a=\frac{37-(-0.89)*42}{10}=7.4$$

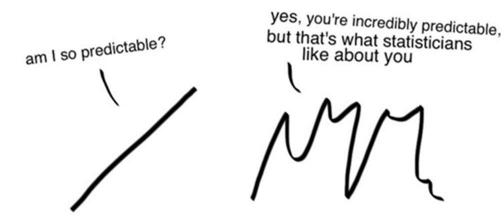
$$a=\frac{\left(\sum Y\right)-b\left(\sum X\right)}{n}$$

Writing an Equation

- Y’ = a + b*X
- Y’ = (Y prime) = predicted rather than actual value of Y
- Example: intercept = 7.4, slope = -0.89
- We write: Y’ = 7.4 - 0.89*X
- What about the actual Y?
- To get it, we need to include an error term, e:
- Y = a + b*X + e

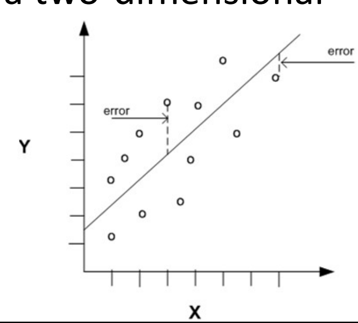
Prediction using Regression

- We can use a regression line to predict one variable based on another
- E.g., predict college GPA from one's SAT
- But regression is often used without prediction
- To predict, for a given value of X, we use the regression equation to calculate the predicted value of Y



Error of Estimate (Error of Prediction)

- Error of estimate = distance between an actual value and regression line (i.e., difference between prediction and reality)
- Standard error of estimate – all differences are squared, then we calculate their average and take its square root (like a two-dimensional standard deviation)

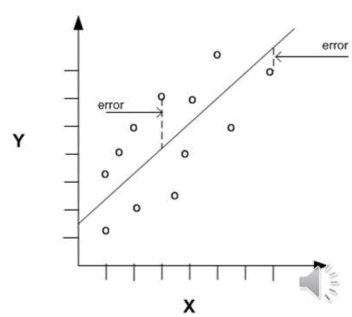


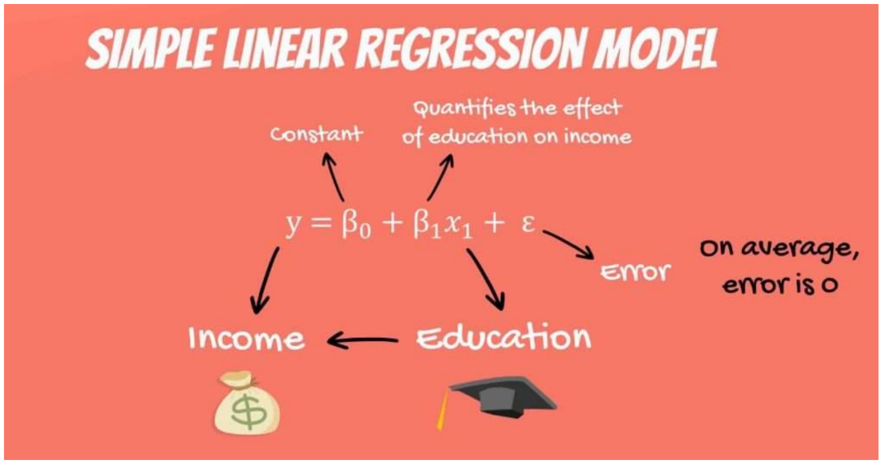
Example: Predicted and Actual Values

	X	Y	$Y' = 7.4 - 0.89*X$	error = Y – Y'	$error^2 = (Y - Y')^2$
	5	2	2.95	-0.95	0.9025
	3	4	4.73	-0.73	0.5329
	7	1	1.17	-0.17	0.0289
	2	6	5.62	0.38	0.1444
	4	5	3.84	1.16	1.3456
	6	2	2.06	-0.06	0.0036
	4	3	3.84	-0.84	0.7056
	2	7	5.62	1.38	1.9044
	8	1	0.28	0.72	0.5184
	1	6	6.51	-0.51	0.2601
Σ				0	6.3464

Standard Error of Estimate (Root MSE)

- Standard error of estimate – all differences are squared, then we calculate their average and take its square root (like a two-dimensional standard deviation)
- Divide by df=n-k where k is number of coefficients in regression
- Here it's 2 – a & b: df=10-2=8
- Sqrt(6.3464/8)=0.89





Testing Hypotheses About Regression Coefficients

- Hypothesis test determines if independent variable *X* has an effect on the dependent variable *Y* in the population
- In bivariate regression, two coefficients: constant (intercept) a and slope b
- For effect of X on Y → focus on the slope b
- If slope is significantly different from zero → there is a linear relationship in the population

Hypotheses Testing for Regression: Step by Step

1. State hypotheses:

- H0: β = 0
- H1: β > 0 — directional
- H1: β < 0 — directional
- H1: β ≠ 0 — non-directional

2. Select alpha: 0.05, 0.01, .001, .10
3. Test statistic: Student's t
4. $t = b/SE_b$
5. Use the table to find critical value:
Table B2 (df=n-2, alpha, one-tailed vs. two-tailed)
6. Compare computed value and critical value
7. State your decision about H0
8. Make a substantive conclusion

$$SE_b = \frac{\sqrt{\sum_{i=1}^{n}\left(y_i - y_i'\right)^2 / n-2}}{\sqrt{\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2}}$$

Example: Commuting and Stress

Do employed Americans who spend more time commuting by car have higher levels of stress? In a nationally representative sample of 75 employed individuals, we regress stress on time spent commuting by car and get: slope b = 0.247, standard error = .133.

Can we conclude that commuting time increases stress?

Example: Step by Step

1. State hypotheses:
- H0: β = 0
- H1: β > 0

2. Select alpha: 0.05
3. Test statistic: Student's t
4. $t = b/s_b$ = .247/.133 = 1.86
5. Use the table to find critical value: Table B2 (df=n-2 = 75-2=73, alpha = .05, one-tailed) → 1.666
6. Compare computed & critical value: 1.86 > 1.666
7. State your decision: We reject H0 in favor of H1.
8. Conclusion: Based on the sample of 75 employed individuals, we are 95% sure that the time spent commuting by car is associated with increased levels of stress among employed Americans (this relationship is statistically significant at .05 level)

Regression in Stata: Problem

- Problem: We would like to determine whether, for the U.S. population, one's level of education affects the age when that person has their first child.
- H0: One's level of education has no effect on the age when that person has her or his first child.
- H1: One's level of education affects the age when that person has her or his first child. [non-directional → two-tailed]
- H0: β = 0
- H1: β ≠ 0

Regression in Stata

reg agekdbrn educ

Coefficient of determination: Shows the percent of variance in age at first birth that is explained by education.

```
      Source |       SS       df       MS              Number of obs =    1429
-------------+------------------------------           F(  1,  1427) =  211.89
       Model |  5816.50574     1  5816.50574           Prob > F      =  0.0000
    Residual |  39171.7504  1427  27.4504207           R-squared     =  0.1293
-------------+------------------------------           Adj R-squared =  0.1287
       Total |  44988.2561  1428   31.504381           Root MSE      =  5.2393

------------------------------------------------------------------------------
    agekdbrn |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        educ |   .6272258   .0430891    14.56   0.000     .5427011    .7117506
       _cons |   15.80311   .5854251    26.99   0.000     14.65472    16.95149
------------------------------------------------------------------------------
```

ANOVA table

Regression intercept (a). Shows the predicted age at first birth for someone with 0 years of education.

Regression slope (b). Shows that as education increases by one year, age at first birth increases by .63 of a year (or if using between .54 and .71 of a year).

Conclusion

- b = .627, t = 14.56, p < .001 (two-tailed) → reject null
- The positive effect of education on one's age at first childbirth is statistically significant at 0.001 level → we are 99.9% confident that in the population, higher levels of education are linked to higher age at first childbirth
- In addition, we can say that we are 95% sure that one year increase in education is associated with between .54 and .71 of a year increase in one's age at first childbirth

Confidence Intervals in Regression in Stata

- The output shows a 95% confidence interval by default
- Use level option to change:

```
. reg agekdbrn educ, level(99.9)
      Source |       SS           df       MS      Number of obs   =     1,429
-------------+----------------------------------   F(1, 1427)      =    211.89
       Model |  5816.50574         1  5816.50574   Prob > F        =    0.0000
    Residual |  39171.7504     1,427  27.4504207   R-squared       =    0.1293
-------------+----------------------------------   Adj R-squared   =    0.1287
       Total |  44988.2561     1,428   31.504381   Root MSE        =    5.2393

------------------------------------------------------------------------------
    agekdbrn |      Coef.   Std. Err.      t    P>|t|     [99.9% Conf. Interval]
-------------+----------------------------------------------------------------
        educ |   .6272258   .0430891    14.56   0.000     .4851457     .769306
       _cons |   15.80311   .5854251    26.99   0.000     13.87275    17.73346
------------------------------------------------------------------------------
```

- We are 99.9% sure that one year increase in education is associated with a between one half and three quarters of a year increase in the age at first childbirth. Probability(.48<beta<.77)=.999

Two-tailed vs One-tailed Test for Regression in Stata

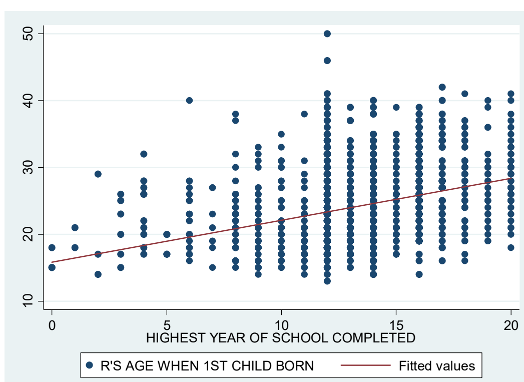
- The output shows a two-tailed test for regression coefficients
- But what if our research hypothesis is directional?
- If you want one-tailed test, just divide p-value (the value in P>|t|) by 2!

How to Decide Which Variable Is Your Dependent One?

- Make a statement both ways, e.g.:
  - The length of one's commute affects the level of stress they have. Commute → Stress
  - The level of stress one has affects the length of one's commute. Stress → Commute
- Select the one that makes more sense
- Independent variable → Dependent variable
- In Stata:
  - reg dep indep
  - scatter dep indep
  - lowess dep indep

Scatterplot with a Regression Line

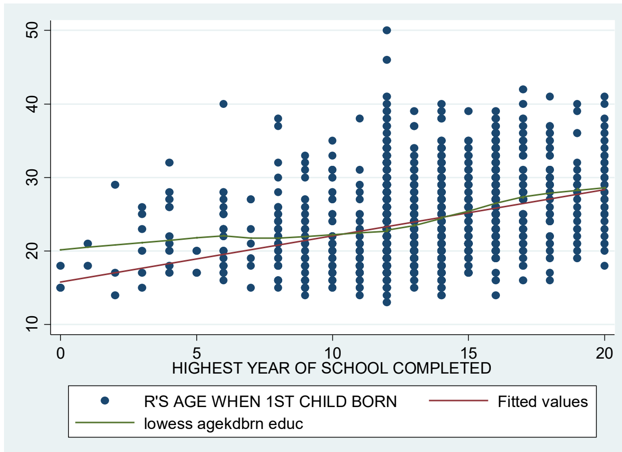
```
graph twoway (scatter agekdbrn educ) (lfit agekdbrn educ)
```



- Independent variable – horizontal axis (X)
- Dependent variable – vertical axis (Y)

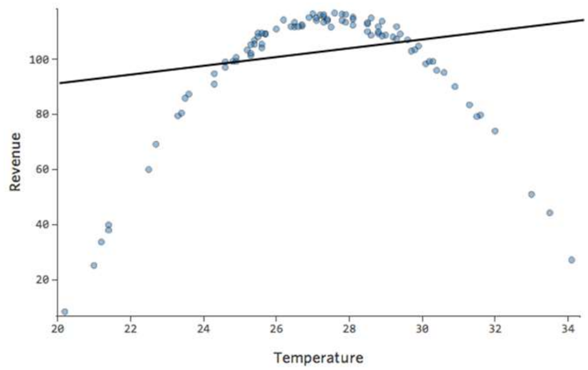
Regression and Lowess Combo

```
graph twoway (scatter agekdbrn educ) (lfit agekdbrn
educ) (lowess agekdbrn educ)
```



Common Problems with Regression

Problem 1: Underlying relationship is not linear



Problem 2: Highly correlated predictors (can't distinguish unique contributions)

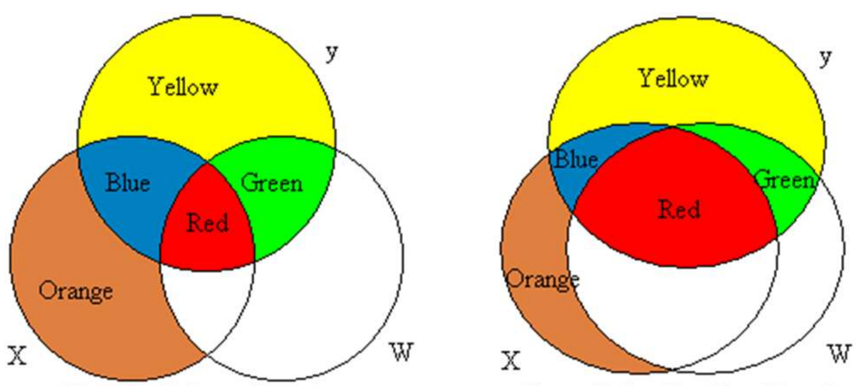
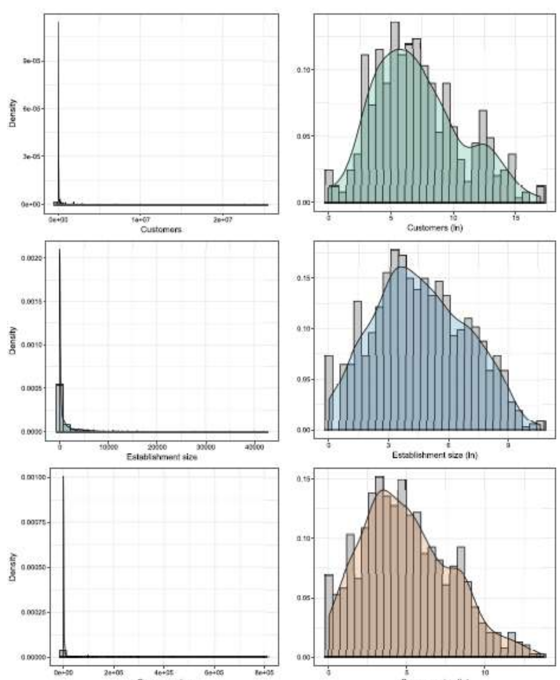


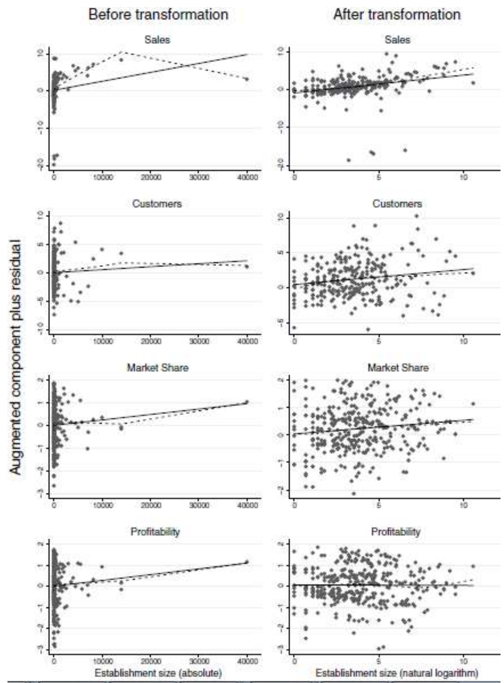
Figure 3a Modest collinearity

Figure 3b Considerable collinearity

Problem 3: Variables Are Skewed or Have Extreme Outliers

- American Sociological Review: Does Diversity Pay?: Race, Gender, and the Business Case for Diversity
- https://www.asanet.org/wp-content/uploads/savvy/images/journals/docs/pdf/asr/Apr09ASRFeature.pdf
- Response/reanalysis:
- http://journals.sagepub.com/doi/pdf/10.1177/0003122417714422





Original Article Abstract

- The value-in-diversity perspective argues that a diverse workforce, relative to a homogeneous one, is generally beneficial for business, including but not limited to corporate profits and earnings. This is in contrast to other accounts that view diversity as either nonconsequential to business success or actually detrimental by creating conflict, undermining cohesion, and thus decreasing productivity. Using data from the 1996 to 1997 National Organizations Survey, a national sample of for-profit business organizations, this article tests eight hypotheses derived from the value-in-diversity thesis. The results support seven of these hypotheses: racial diversity is associated with increased sales revenue, more customers, greater market share, and greater relative profits. Gender diversity is associated with increased sales revenue, more customers, and greater relative profits. I discuss the implications of these findings relative to alternative views of diversity in the workplace.

Does Diversity Pay? A Replication of Herring (2009)

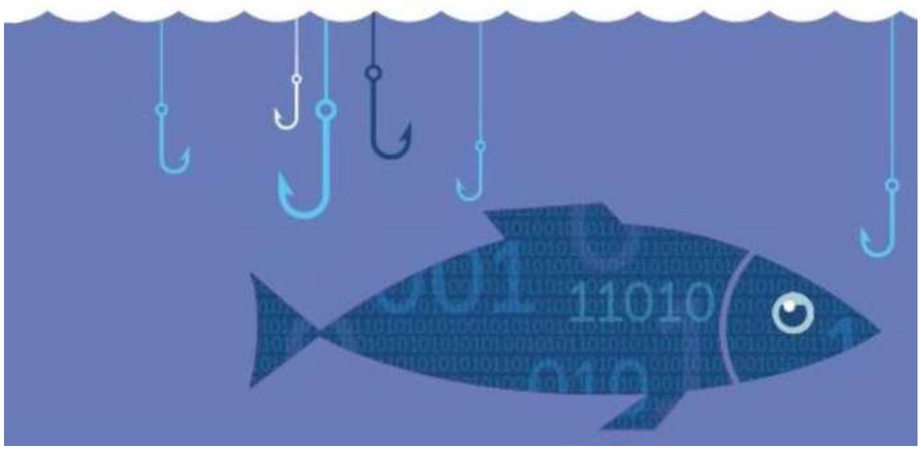
- In an influential article published in the *American Sociological Review* in 2009, Herring finds that diverse workforces are beneficial for business. His analysis supports seven out of eight hypotheses on the positive effects of gender and racial diversity on sales revenue, number of customers, perceived relative market share, and perceived relative profitability. This comment points out that Herring's analysis contains two errors. First, missing codes on the outcome variables are treated as substantive codes. Second, two control variables—company size and establishment size—are highly skewed, and this skew obscures their positive associations with the predictor and outcome variables. We replicate Herring's analysis correcting for both errors. The findings support only one of the original eight hypotheses, suggesting that diversity is nonconsequential, rather than beneficial, to business success.

Problem 4: Causation Issues

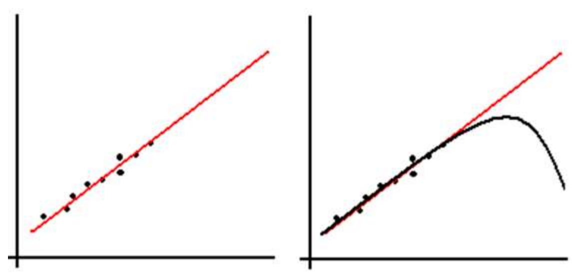
- Regression cannot prove causation (especially with cross-sectional data)
- We risk overlooking reverse causality (Y→X)
- Omitted variable bias ("third variable" explanations, etc.)

Excellent health statistics - smokers are less likely to die of age related illnesses.'

Problem 5: "Fishing" for effects (p-hacking and inflated alpha)

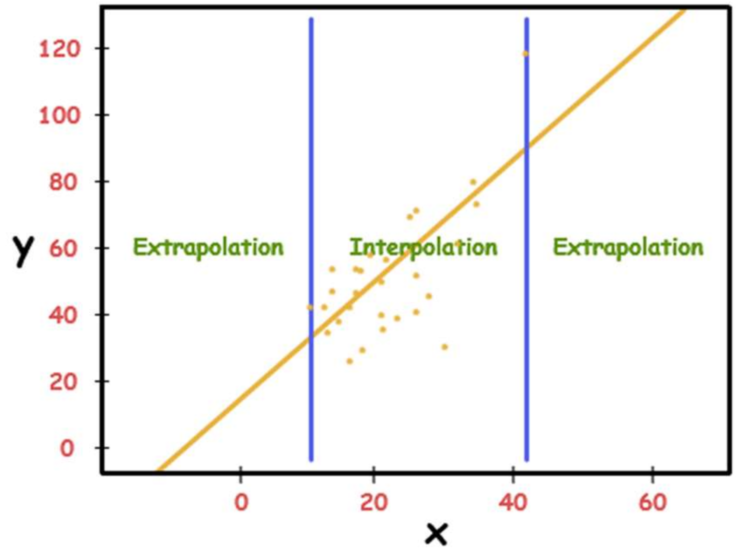


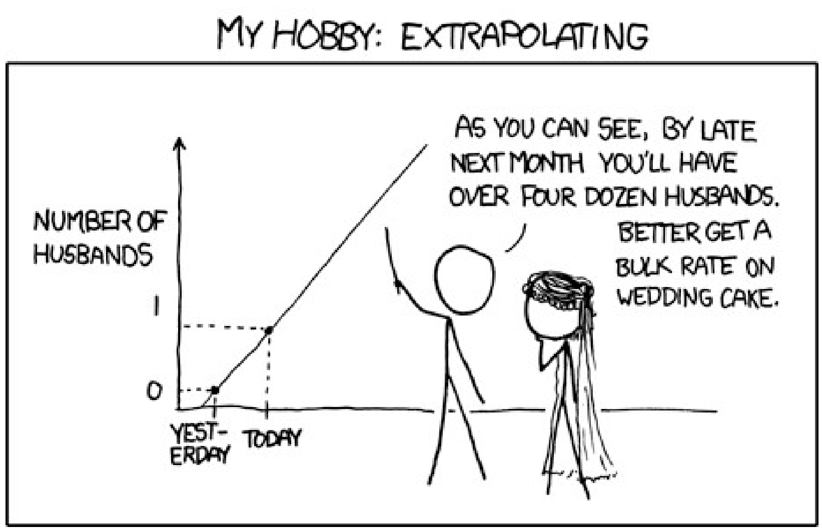
Problem 6: Extrapolation



- Red line = prediction based on regression
- Black line = the actual trend

Interpolation vs Extrapolation





If she loves you more each and every day,
by linear regression she hated you before you met.