Regression

Collin Cademartori

1/30/2020

Regression to The Mean

Regression to The Mean

- Regression to the mean is a common statistical artifact
- An example of phenomena explainable by random variation
- Responsible for many mistakes in published research

Do tall parents have shorter children

- Galton (1886) recorded the heights (in inches) of 205 parents and their 928 adult children.
- On average, men 8 percent taller than women so adjusted womens heights to be comparable.
- Galton compared average height of a parent to average height of each child.
- He noticed tall parents tended to have shorted children. Declared children appeared to “regress towards mediocrity”.
- At first posited evolutionary mechanism causing tendency to reduced variation around mean.
- Eventually figured out this was just a random effect

Sir Francis Galton (1903)

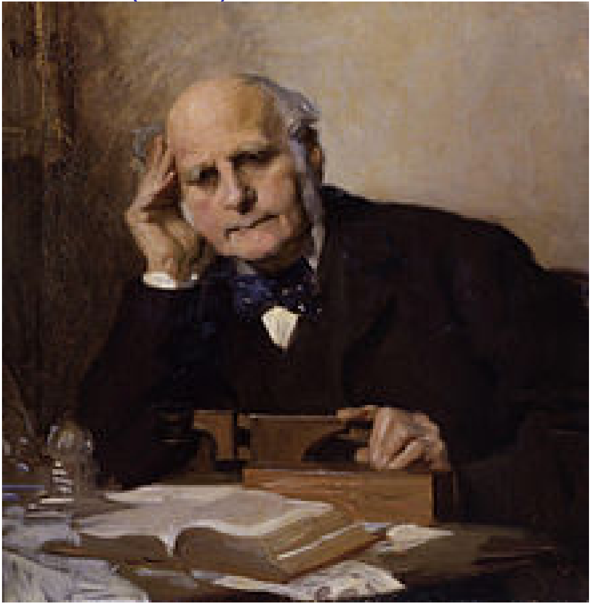


Figure 1: Galton, first cousin of Charles Darwin

Regression towards mediocrity in heriditary stature (1886)

TABLE I.

NUMBER OF ADULT CHILDREN OF VARIOUS STATURES BORN OF 205 MID-PARENTS OF VARIOUS STATURES.
(All Female heights have been multiplied by 1·08).

Heights of the Mid-parents in inches.	Heights of the Adult Children.														Total Number of		Medians.
	Below	62·2	63·2	64·2	65·2	66·2	67·2	68·2	69·2	70·2	71·2	72·2	73·2	Above	Adult Children.	Mid-parents.	
Above ..	..	..	..	..	..	..	..	..	..	..	..	1	3	..	4	5	..
72·5	..	..	..	..	..	..	..	1	2	1	2	7	2	4	19	6	72·2
71·5	..	..	..	..	1	3	4	3	5	10	4	9	2	2	43	11	69·9
70·5	1	..	1	..	1	1	3	12	18	14	7	4	3	3	68	22	69·5
69·5	..	..	1	16	4	17	27	20	33	25	20	11	4	5	183	41	68·9
68·5	1	..	7	11	16	25	31	34	48	21	18	4	3	..	219	49	68·2
67·5	..	3	5	14	15	36	38	28	38	19	11	4	..	..	211	33	67·6
66·5	..	3	3	5	2	17	17	14	13	4	..	..	..	..	78	20	67·2
65·5	1	..	9	5	7	11	11	7	7	5	2	1	..	..	66	12	66·7
64·5	1	1	4	4	1	5	5	..	2	..	..	..	..	..	23	5	65·8
Below ..	1	..	2	4	1	2	2	1	1	..	..	..	..	..	14	1	..
Totals ..	5	7	32	59	48	117	138	120	167	99	64	41	17	14	928	205	..
Medians ..	..	..	66·3	67·8	67·9	67·7	67·9	68·3	68·5	69·0	69·0	70·0	..	..	..	..	..

NOTE.—In calculating the Medians, the entries have been taken as referring to the middle of the squares in which they stand. The reason why the headings run 62·2, 63·2, &c., instead of 62·5, 63·5, &c., is that the observations are unequally distributed between 62 and 63, 63 and 64, &c., there being a strong bias in favour of integral inches. After careful consideration, I concluded that the headings, as adopted, best satisfied the conditions. This inequality was not apparent in the case of the Mid-parents.

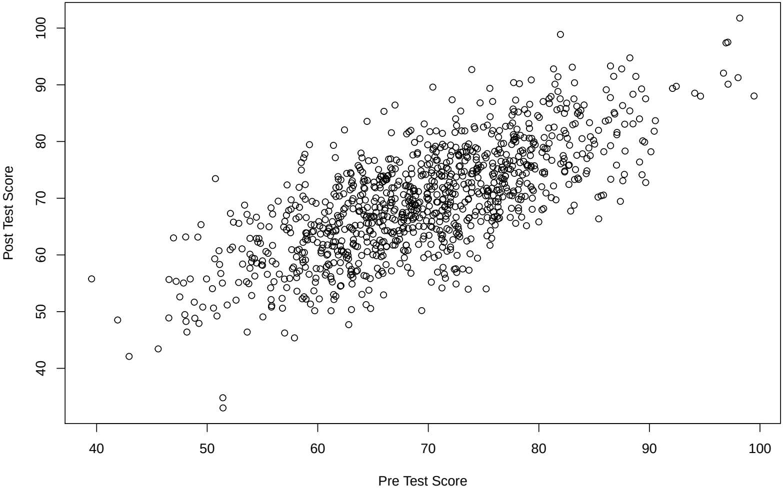
A Testing Problem

- Imagine students take a test on calculus
- Then they are instructed to study for three more hours
- Then they take an equivalent test
- We want to use the before and after scores to judge the effectiveness of the extra studying
- The best students might not benefit much, but we really want to target the worst students
- So we can look at the change in the scores of the students with lowest scores on the first test
- Suppose we see their scores all increased. Is this evidence that the studying helped?

Simulating No Effect

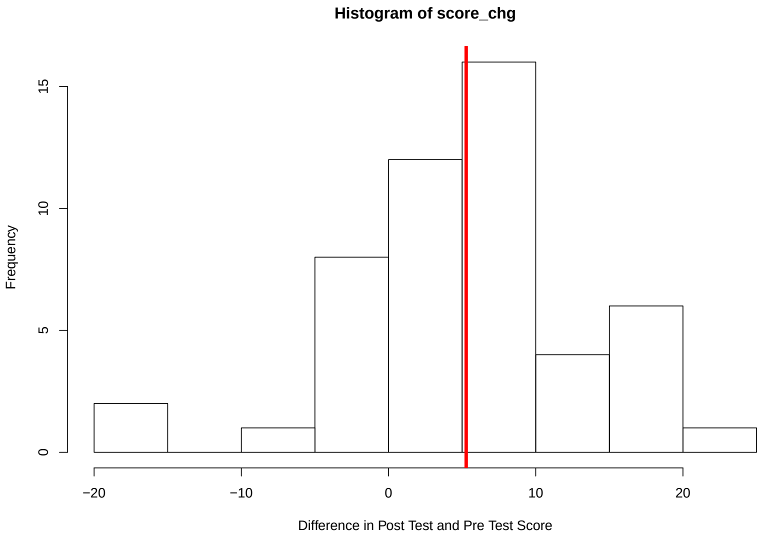
- Suppose there was no effect from the additional studying
- We can simulate this
- Suppose 500 initial test scores are distributed like $N(70, 8)$
- Suppose the post-studying test scores are distributed like initial scores plus some random noise ϵ
- Specifically suppose $\epsilon \sim N(0, 5)$
- Zero mean error corresponds to "no effect" assumption

Simulating No Effect

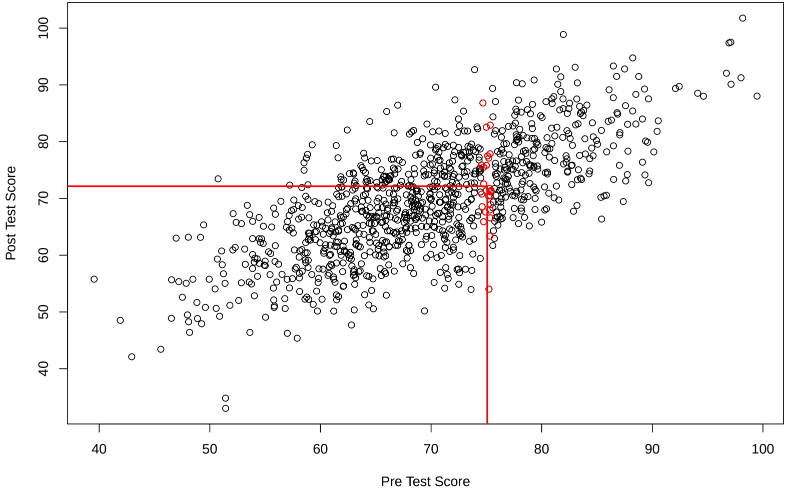


How Did The Lowest Scoring Students Perform?

- We can check how the lowest scoring students on the pre-test performed on the post-test

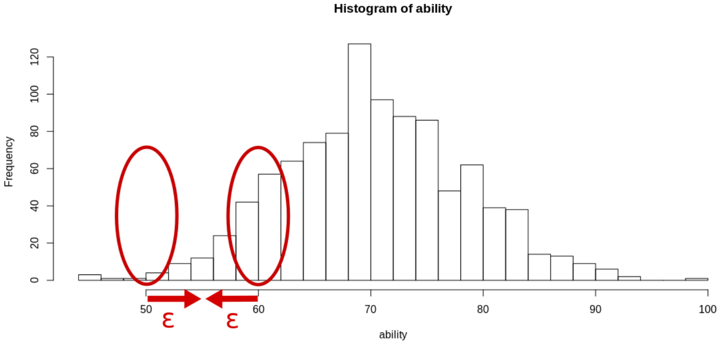


Not Just the Lowest Scores!



What Happened? Regresson to The Mean!

- We expect the students who did worst on the pre-test to improve on the post-test
- . . . even if the extra studying had no effect on their underlying ability!
- Why? Let's look at a picture



So, What Is Regression to The Mean?

We can summarize the lesson of regression to the mean in two ways:

- In many cases, when observing data that combine an underlying effect with noise, the more extreme a value we observe, the more probable it is that this value corresponds to a less extreme underlying effect and a more extreme noise value.
- When two variables are imperfectly correlated, more extreme values of one are associated, on average, with less extreme values of the other.
- Not a law of nature; doesn't *always* occur!
- The distribution of abilities could have a very long tail, for example.

Regression Models

Motivation: Imperfect Correlation

- In the last section we discussed a statistical artifact arising from imperfect correlation.
- We wanted to understand the effect of studying on ability.
- But we could not measure ability directly!
- We can think of the test score as consisting of true ability plus some error.
- In good cases, we can get at the quantity of interest directly (and with negligible measurement error).
- And ideally we get deterministic models like inverse square laws in physics.
- But often this is impossible. Our tools are either too crude, or we can't even get direct access in principle.
- More fundamentally, can't measure enough quantities to hope for deterministic relationships.
- Complex phenomena are highly multi-causal.

Statistical Models

- These problems motivate considering models of the form

$$y_i = f\left(x_i^1, \ldots, x_i^k\right) + \epsilon_i$$

- y_i are outcomes of interest.
- $x_i^1, \ldots, x_i^k$ are predictors or covariates.
- $f(\cdot)$ is a specified function that describes the relationship between y and the xs
- ϵ_i are error or noise terms representing variation in y unexplained by x
- The ys and xs are measured, but the ϵs are not.
- Want to infer the relationship f from the measured data.

Statistical Models

- The ys and xs are specified by the research question.
- How do we use this data to estimate f? in

$$y_i = f\left(x_i^1, \ldots, x_i^k\right) + \epsilon_i$$

- In principle, f could be anything!
- Without any resitriction on f, this is an infinite-dimensional inference problem!
- In general, the fewer assumption we make about f, the more data we need to infer it accurately.
- (This is a case of a more general phenomenon called the bias-variance tradeoff.)

Linear Regression Models

- As a starting point, we can assume that f is linear in its predictors.

$$y_i = \alpha + \beta_1 x_i^1 + \cdots + \beta_k x_i^k + \epsilon_i$$

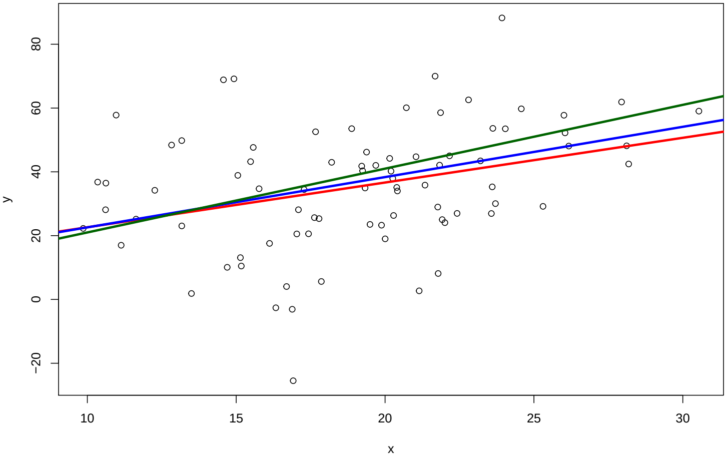
- This is a linear regression model.
- As we will see, the methods used for these models are easily extended to more general f.
- We will start with an even simpler form with one predictor.

$$y_i = \alpha + \beta x_i + \epsilon_i$$

- For example, for $1 \leq i \leq 50$, y_i could be Trump's share of the vote in state i and x_i the average share of the vote the polls predicted Trump would win.

Fitting a Linear Regression Model

- Now estimating f just requires estimating α and β.
- But this is still not obvious! Is there one right way to do this?
- First we can look at the question deterministically.
- We can just ask which line best fits the trend in the data.



The Best Fitting Line

- How should we think about what it means for line to fit the data well?
- For any line $y = a + bx$, we can use this line to predict y values

$$\hat{y}_i = a + bx_i$$

where the $\hat{\cdot}$ symbol is used to denote a predicted value.

- Then the best line might be the one which minimizes the sum of the distances between predictions and the true values:

$$\sum_{i=1}^{n} |y_i - \hat{y}_i|$$

- In the last graph, the blue line minimized this condition.

The Best Fitting Line

- Can we justify using $|y_i - \hat{y}_i|$ to measure the quality of a prediction?
- We could consider any discrepancy function $d(\hat{y}_i, y_i)$ that quantifies the quality of a prediction.
- For this to make sense, we would need $d(\hat{y}_i, y_i) \geq 0$ with equality only if $\hat{y}_i = y_i$.
- Furthermore, suppose that this discrepancy function is smooth (in this case, has two derivatives).
- Then Taylor's theorem tells us we can approximate $d(\hat{y}, y)$ as a function of $\hat{y}$ with a Taylor series centered at y (the true value):

$$\begin{aligned} d(\hat{y}, y) &\approx d(y, y) + d'(y, y)(\hat{y} - y) + 2d''(y, y)(\hat{y} - y)^2 \\ &= 2d''(y, y)(\hat{y} - y)^2 \end{aligned}$$

- Since the above guarantees that $d(y, y) = 0$ and $d'(y, y) = 0$ since y minimizes $d(\cdot, y)$.

The Best Fitting Line

- Now again since y is a minimum, we must have $d''(y, y) > 0$
- So minimizing $2d''(y, y)(\hat{y} - y)^2$ is equivalent to minimizing $(\hat{y} - y)^2$.
- So for a general smooth disprepancy function d, we can approximately minimize the sum $\sum_{i=1}^{n} d(\hat{y}_i, y_i)$ by minimizing

$$\sum_{i=1}^{n} (\hat{y}_i - y_i)^2$$

- In the above graph, the red line minimizes this condition.
- What about the green line? It's the true line from which the data were generated!

Introducing Probability

- The previous discussion was entirely deterministic.
- Phrased our problem as an optimization problem.
- What if we think of the points as being randomly scattered around the true line?
- This corresponds to thinking of errors ϵ_i as having some probability distribution.
- But we still have a choice: what distribution do we think ϵ_i has?

Error Distributions

- Often reasonable to think of errors as arising from a series of independent chance effects.
- For instance, a product moving down an assembly line.
- Each machine has some level of imprecision in its operation.
- Each question on a test is an imperfect measure of knowledge of a concept.
- A sum of many independent small chance effects have an approximately normal distirbution.
- By the central limit theorem!
- So we often take $\epsilon_i \overset{iid}{\sim} \mathrm{normal}(0, \sigma)$

From Distributions to Fitting Procedures

- If ϵ_i are normal, then we have $y_i \overset{iid}{\sim} \mathrm{normal}(\alpha + \beta x_i, \sigma)$
- Recall that the normal distribution has density

$$f(x \mid \mu, \sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(x-\mu)^2}$$

- The mean is $\mu = \alpha + \beta x_i$ and the standard deviation is σ.
- Since the y_i are independent, the density for $(y_1, \ldots, y_n)$ is the product of the densities:

$$\begin{aligned} f(y_1, \ldots, y_n \mid \alpha, \beta) &= \prod_{i=1}^{n} f(y_i \mid \alpha + \beta x_i, \sigma) \\ &= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(y_i - \alpha - \beta x_i)^2} \\ &= \left(\frac{1}{\sqrt{2\pi\sigma^2}} \right)^n e^{-\frac{1}{2\sigma^2} \sum_{i=1}^{n} (y_i - \alpha - \beta x_i)^2} \end{aligned}$$

Maximum Likelihood

- Now we think of our data points as having randomly generated y_i values.
- And we can derive a distribution for these values.
- The distribution depends on the parameters α and β.
- For each (α, β), we get a different value of $f(y_1, \ldots, y_n \mid \alpha, \beta)$.
- The larger this value, the more probable the data were.
- It is often reasonable to assume that the data we observed were as probable as possible.
- If that is true, then we should think that the true α and β maximize $f(y_1, \ldots, y_n \mid \alpha, \beta)$.
- This is called fitting the model by maximum likelihood.

Maximum Likelihood with Normal Errors

- What is the ML estimate of (α, β) with normal errors?
- Want to maximize

$$\left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n e^{-\frac{1}{2\sigma^2}\sum_{i=1}^n (y_i - \alpha - \beta x_i)^2}$$

- As a function of (α, β), $\left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n$ is a constant.
- So just want to maxmimize

$$e^{-\frac{1}{2\sigma^2}\sum_{i=1}^n (y_i - \alpha - \beta x_i)^2}$$

- This is equivalent to minimizing

$$\frac{1}{2\sigma^2}\sum_{i=1}^n (y_i - \alpha - \beta x_i)^2$$

- Using $\alpha + \beta x_i = \hat{y}_i$, this becomes

$$\sum_{i=1}^n (y_i - \hat{y}_i)^2$$

- So the ML estimate is the same as the least squares estimate!

Changing Error Distribution

- Now suppose the ϵ_i are not normal.
- Why? Sometimes central limit theorem doesn't hold.
- Points could be too scattered to follow normal distribution (outliers).
- What if we instead assume a Laplace distribution?
- Here $\epsilon_i \overset{iid}{\sim} \text{exponential}(\lambda)$.
- Where the Laplace distribution has density

$$f(y \mid \mu, \lambda) = \frac{1}{2\lambda} e^{-\frac{|y-\mu|}{\lambda}}$$

- Then the joint distribution of the y_i is

$$f(y_1, \ldots, y_n \mid \alpha, \beta) = \left(\frac{1}{2\lambda}\right)^n e^{-\frac{1}{\lambda}\sum_{i=1}^n |y_i - \alpha - \beta x_i|}$$

- Maximizing this is equivalent to minimizing

$$\sum_{i=1}^n |y_i - \hat{y}_i|$$

- So ML estimate is the minimum absolute deviation line!

So, What?

- What have we learned from this?
- We can treat our regression problem as a deterministic problem of finding the best fitting line.
- Then we need to choose a discrepancy measure $d(y, \hat{y})$ to define out optimization problem.
- Or we can treat regression as a probabilistic problem of finding the parameters from which our data was randomly generated.
- Then we need to choose an error distribution to define our estimation problem.
- These two views can give us the same solutions, but involve different ways of thinking.
- The former requires us to judge the severity of an error. This is a decision problem that is related to how we use the our predictions.
- The latter requires us to judge the underlying data-generating process. This requires us to use substantive knowledge about the world.

Beyond Linearity

- Linear regression makes a linearity assumption that appears restrictive
- Many common data display nonlinear association
- How do we capture this nonlinearity?
- Linear regression!
- First: regression with mulitple predictors.
- Recall from earlier, a linear model of the form

$$y_i = \alpha + \beta_1 x_i^{(1)} + \cdots + \beta_k x_i^{(k)} + \epsilon_i$$

  where $x_i^{(j)}$ is the value of the j^{th} predictor for the i^{th} data point.
- Now need to estimate α and teh β_j. This can be done as before.

Beyond Linearity

- Note that the function

$$\alpha + \beta_1 x_i^{(1)} + \cdots + \beta_k x_i^{(k)}$$

  is linear in the parameters α and β_j.
- This is true regardless of what $x^{(j)}$ are.
- So if we start with a single predictor $x_i^{(1)} = x_i \ldots$
- We can define $x_i^{(j)} = x_i^j$ for $j \geq 2$.
- The resulting function

$$y_i = \alpha + \beta_1 x_i + \beta_2 x_i^2 + \cdots + \beta_k x_i^k + \epsilon_i$$

  is a polynomial in x_i, but it is linear in the parameters!
- We can again fit such a model using least squares or maximum likelihood in exactly the same way as before!

Beyond Linearity

- In general we can take $x_i^{(j)} = b_j(x_i)$ for any "basis functions" b_i.
- In last slide, we took these to be powers of x.
- Could choose these to be trig functions with different periods.
- Or exponentials with different rates.
- So linear regression supports a wide variety of functional forms.
- However, the functional form must be specified in advance.
- What if we want to learn the functional form from the data?
- We will see one solution to this when we discuss high dimensional problems.