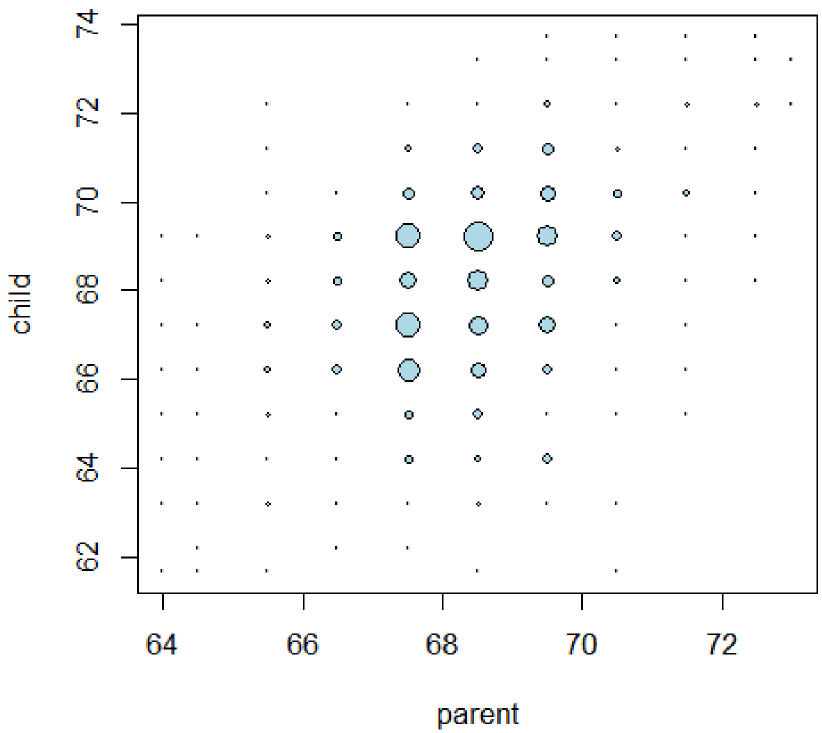
Least squares estimation of regression lines

Regression via least squares

Brian Caffo, Jeff Leek and Roger Peng
Johns Hopkins Bloomberg School of Public Health

General least squares for linear equations

Consider again the parent and child height data from Galton



Fitting the best line

- Let Y_i be the i^{th} child's height and X_i be the i^{th} (average over the pair of) parents' heights.
- Consider finding the best line
  - Child's Height = β_0 + Parent's Height β_1
- Use least squares

observed

$$\sum_{i=1}^{n}\{Y_i - (\beta_0 + \beta_1 X_i)\}^2$$

predicted

see Ng in
Machine Learning!

- How do we do it?

Let's solve this problem generally

- Let $\mu_i = \beta_0 + \beta_1 X_i$ and our estimates be $\hat{\mu}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i$.

mu_i^ und beta_0/1^ sind unsere Voraussagen.
mu_i resp beta_0/1 sind irgendwelche /anderen/ Voraussagen.

- We want to minimize

Add and extract my_i^ (wie beim Beweis, dass Min der Summe der squared Dist. = mean),
then expand the square:

$$\dagger \sum_{i=1}^{n} (Y_i - \mu_i)^2 = \sum_{i=1}^{n} (Y_i - \hat{\mu}_i)^2 + 2 \sum_{i=1}^{n} (Y_i - \hat{\mu}_i)(\hat{\mu}_i - \mu_i) + \sum_{i=1}^{n} (\hat{\mu}_i - \mu_i)^2$$

1.Summe:
Diff zw. Y_i und originalen Voraussage mu_i^

2. Summe:
Diff zw originalen Voraussage und neuen, geaenderten Voraussage

- Suppose that

$$\sum_{i=1}^{n} (Y_i - \hat{\mu}_i)(\hat{\mu}_i - \mu_i) = 0$$

then

$$\dagger = \sum_{i=1}^{n} (Y_i - \hat{\mu}_i)^2 + \sum_{i=1}^{n} (\hat{\mu}_i - \mu_i)^2 \geq \sum_{i=1}^{n} (Y_i - \hat{\mu}_i)^2$$

Diese Summe ist immer > 0

Also fuer /irgendein/ mu_i ist die Summe der quadrierten Distanzen groesser als wenn man das vorhergesagte mu nimmt! HX

Sofern die mittlere Summe == 0! Siehe unten..

=> gleicher Trick wie beim Beweis, dass Min der Summe der squared Dist. = mean)

Mean only regression

- So we know that if:

$$\sum_{i=1}^{n}(Y_i - \hat{\mu}_i)(\hat{\mu}_i - \mu_i) = 0$$

where $\mu_i = \beta_0 + \beta_1 X_i$ and $\hat{\mu}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i$ then the line

$$Y = \hat{\beta}_0 + \hat{\beta}_1 X$$

is the least squares line.

- Consider forcing $\beta_1 = 0$ and thus $\hat{\beta}_1 = 0$; that is, only considering horizontal lines
- The solution works out to be

$$\hat{\beta}_0 = \bar{Y}.$$

-> die beste Voraussage ist das mean, dann.

(haben wir schon frueher bewiesen, und wie die naechste Slide nochmal beweist)

Let's show it

$$\begin{aligned}\sum_{i=1}^{n}(Y_i-\hat{\mu}_i)(\hat{\mu}_i-\mu_i) &= \sum_{i=1}^{n}(Y_i-\hat{\beta}_0)(\hat{\beta}_0-\beta_0) \\ &= (\hat{\beta}_0-\beta_0)\sum_{i=1}^{n}(Y_i-\hat{\beta}_0)\end{aligned}$$

Thus, this will equal 0 if $\sum_{i=1}^{n}(Y_i-\hat{\beta}_0) = n\bar{Y} - n\hat{\beta}_0 = 0$

^ siehe frueher)

Thus $\hat{\beta}_0 = \bar{Y}$.

Regression through the origin

- Recall that if:

$$\sum_{i=1}^{n}(Y_i - \hat{\mu}_i)(\hat{\mu}_i - \mu_i) = 0$$

where $\mu_i = \beta_0 + \beta_1 X_i$ and $\hat{\mu}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i$ then the line

$$Y = \hat{\beta}_0 + \hat{\beta}_1 X$$

is the least squares line.

- Consider forcing $\beta_0 = 0$ and thus $\hat{\beta}_0 = 0$; that is, only considering lines through the origin
  not beta1, but beta0 ist jetzt = 0, also ^
- The solution works out to be

$$\hat{\beta}_1 = \frac{\sum_{i=1^n} Y_i X_i}{\sum_{i=1}^{n} X_i^2}.$$

$\dfrac{< Y, X >}{< X, X >}$ (inner product)

Let's show it

```
mu_i  = beta1 * x_i
mu_i^ = beta1^ * x_i
```

$$\sum_{i=1}^{n}(Y_i - \hat{\mu}_i)(\hat{\mu}_i - \mu_i) = \sum_{i=1}^{n}(Y_i - \hat{\beta}_1 X_i)(\hat{\beta}_1 X_i - \beta_1 X_i)$$

$$= (\hat{\beta}_1 - \beta_1) \sum_{i=1}^{n}(Y_i X_i - \hat{\beta}_1 X_i^2)$$

Thus, this will equal 0 if $\sum_{i=1}^{n}(Y_i X_i - \hat{\beta}_1 X_i^2) = \sum_{i=1}^{n} Y_i X_i - \hat{\beta}_1 \sum_{i=1}^{n} X_i^2 = 0$

Thus

$$\hat{\beta}_1 = \frac{\sum_{i=1^n} Y_i X_i}{\sum_{i=1}^{n} X_i^2}.$$

```
Bei /diesem/ beta1^ wird der erste Term
oben = 0, und /dann/ ist die Gerade

Y = beta0^ + beta1^ * X

die Gerade der kleinsten Quadrate!
```

Recapping what we know

- If we define $\mu_i = \beta_0$ then $\hat{\beta}_0 = \bar{Y}$.
  - If we only look at horizontal lines, the least squares estimate of the intercept of that line is the average of the outcomes.
- If we define $\mu_i = X_i \beta_1$ then $\hat{\beta}_1 = \frac{\sum_{i=1^n} Y_i X_i}{\sum_{i=1}^n X_i^2}$ =sum of squares of X_i
  - If we only look at lines through the origin, we get the estimated slope is the cross product of the X and Ys divided by the cross product of the Xs with themselves.
- What about when $\mu_i = \beta_0 + \beta_1 X_i$? That is, we don't want to restrict ourselves to horizontal lines or lines through the origin.

Let's figure it out

mu_i = beta0 + beta1 * x_i
mu_i^ = beta0^ + beta1^ * x_i

-> einsetzen:

$$\sum_{i=1}^{n}(Y_i - \hat{\mu}_i)(\hat{\mu}_i - \mu_i) = \sum_{i=1}^{n}(Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i)(\hat{\beta}_0 + \hat{\beta}_1 X_i - \beta_0 - \beta_1 X_i)$$

das soll 0 werden

$$= (\hat{\beta}_0 - \beta_0)\boxed{\sum_{i=1}^{n}(Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i)} + (\hat{\beta}_1 - \beta_1)\sum_{i=1}^{n}(Y_i - \boxed{\hat{\beta}_0} - \hat{\beta}_1 X_i)X_i$$

falls dieser Term 0 waere …

… und dieser Term auch 0 waere, dann haetten wir, was wir wollen!

Note that

Wir haben also 2 Gleichungen und zwei Unbekannte!

einsetzen

$$0 = \boxed{\sum_{i=1}^{n}(Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i)} = n\bar{Y} - n\hat{\beta}_0 - n\hat{\beta}_1\bar{X} \text{ implies that } \boxed{\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1\bar{X}}$$

einfach nach beta0^ aufloesen

Then

$$\sum_{i=1}^{n}(Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i)X_i = \sum_{i=1}^{n}(Y_i - \bar{Y} + \hat{\beta}_1\bar{X} - \hat{\beta}_1 X_i)X_i$$

das koennen wir jetzt 0 setzen

Continued

$$= \sum_{i=1}^{n} \{\underbrace{(Y_i - \bar{Y})}_{\text{centered Ys}} - \hat{\beta}_1 \underbrace{(X_i - \bar{X})}_{\text{centered Xs}}\} X_i$$

And thus

$$\sum_{i=1}^{n} (Y_i - \bar{Y}) X_i - \hat{\beta}_1 \sum_{i=1}^{n} (X_i - \bar{X}) X_i = 0.$$

nach beta1^ aufloesen:

So we arrive at

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} \{(Y_i - \bar{Y}) X_i}{\sum_{i=1}^{n} (X_i - \bar{X}) X_i} = \frac{\sum_{i=1}^{n} (Y_i - \bar{Y})(X_i - \bar{X})}{\sum_{i=1}^{n} (X_i - \bar{X})(X_i - \bar{X})} = \boxed{\mathrm{Cor}(Y, X) \frac{\mathrm{Sd}(Y)}{\mathrm{Sd}(X)}.}$$

And recall

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}.$$

Das stellt sicher, dass die Regressionslinie
immer durch den Punkt (X_, Y_) geht!

sum(y_i - y_) x_ = x_ sum(y_i - y_) =

Consequences

- The least squares model fit to the line $Y = \beta_0 + \beta_1 X$ through the data pairs (X_i, Y_i) with Y_i as the outcome obtains the line $\boxed{Y = \hat{\beta}_0 + \hat{\beta}_1 X}$ where

mit $$\boxed{\hat{\beta}_1 = Cor(Y, X) \frac{Sd(Y)}{Sd(X)}} \quad \boxed{\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}}$$

slope intercept

- $\hat{\beta}_1$ has the units of Y/X, $\hat{\beta}_0$ has the units of Y. weil Cor(Y,X) hat keine Einheit, also bleibt Einheit Y/Einheit X, was Sinn macht fuer eine Steigung
- The line passes through the point $(\bar{X}, \bar{Y})$
- The slope of the regression line with X as the outcome and Y as the predictor is $Cor(Y, X) Sd(X)/Sd(Y)$. Inverse Funktion
- The slope is the same one you would get if you centered the data, $(X_i - \bar{X}, Y_i - \bar{Y})$, and did regression through the origin.
- If you normalized the data, $\{\frac{X_i - \bar{X}}{Sd(X)}, \frac{Y_i - \bar{Y}}{Sd(Y)}\}$, the slope is $Cor(Y, X)$.

weil dann die SDs = 1 eins, also bleibt die Korrelation.

Revisiting Galton's data

Double check our calculations using R

```r
y <- galton$child
x <- galton$parent
beta1 <- cor(y, x) *  sd(y) / sd(x)
beta0 <- mean(y) - beta1 * mean(x)
rbind(c(beta0, beta1), coef(lm(y ~ x)))
```

```
     (Intercept)      x
[1,]       23.94 0.6463
[2,]       23.94 0.6463
```

Revisiting Galton's data

Reversing the outcome/predictor relationship

```
beta1 <- cor(y, x) *  sd(x) / sd(y)
beta0 <- mean(x) - beta1 * mean(y)
rbind(c(beta0, beta1), coef(lm(x ~ y)))    einfach umgekehrt
```

```
     (Intercept)      y
[1,]       46.14 0.3256
[2,]       46.14 0.3256
```

Revisiting Galton's data

Regression through the origin yields an equivalent slope if you center the data first

```
yc <- y - mean(y)
xc <- x - mean(x)
beta1 <- sum(yc * xc) / sum(xc ^ 2)
c(beta1, coef(lm(y ~ x))[2])
```

centering

```
            x
0.6463 0.6463
```

Revisiting Galton's data

Normalizing variables results in the slope being the correlation

```
yn <- (y - mean(y))/sd(y)
xn <- (x - mean(x))/sd(x)
c(cor(y, x), cor(yn, xn), coef(lm(yn ~ xn))[2])
```

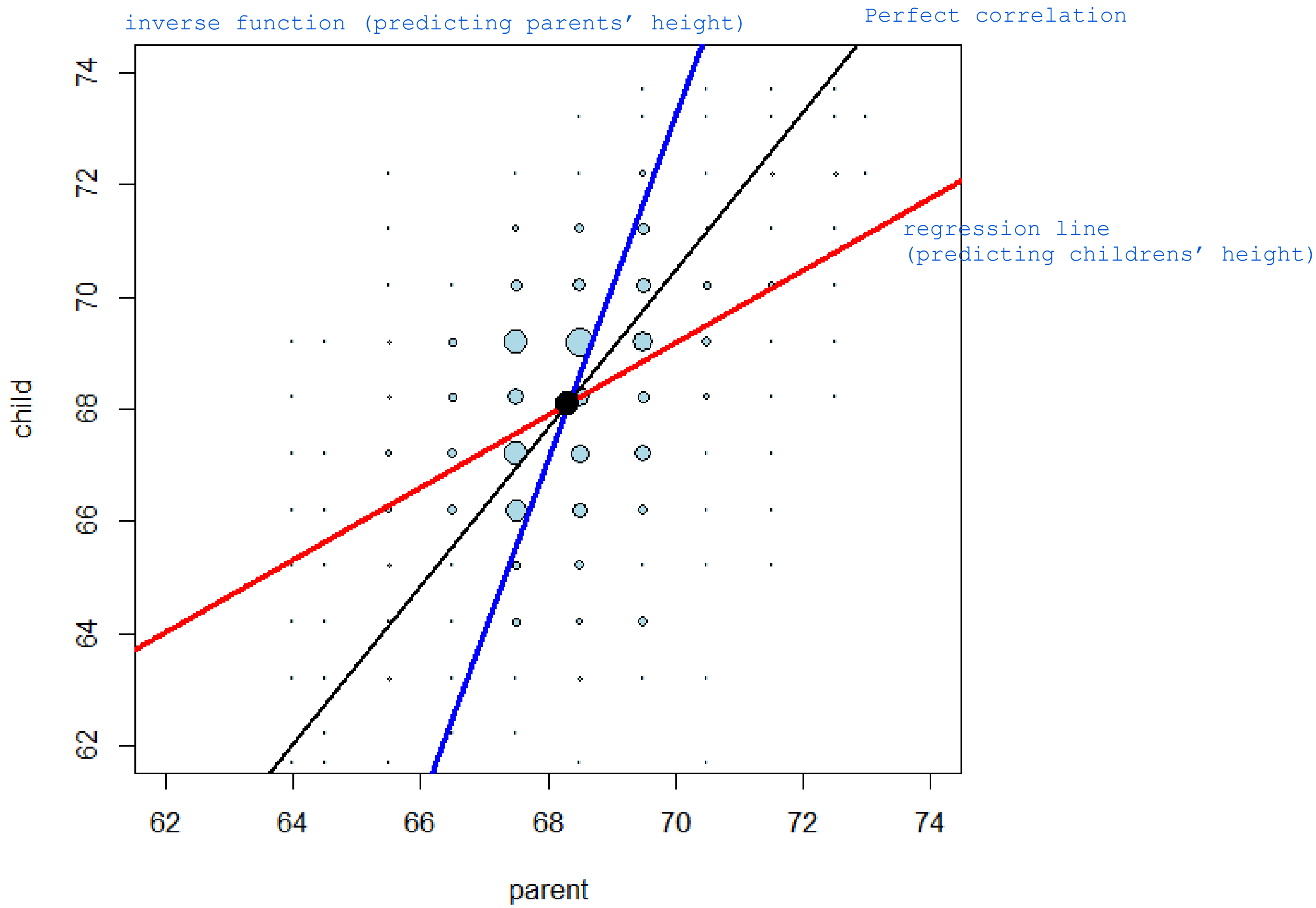
```
                  xn
0.4588 0.4588 0.4588
```

Plotting the fit

- Size of points are frequencies at that X, Y combination.
- For the red lie the child is outcome.
- For the blue, the parent is the outcome (accounting for the fact that the response is plotted on the horizontal axis).
- Black line assumes $\mathrm{Cor}(Y, X) = 1$ (slope is $\mathrm{Sd}(Y)/\mathrm{Sd}(x)$).
- Big black dot is $(\bar{X}, \bar{Y})$.

The code to add the lines

```
abline(mean(y) - mean(x) * cor(y, x) * sd(y) / sd(x),
  sd(y) / sd(x) * cor(y, x),
  lwd = 3, col = "red")
abline(mean(y) - mean(x) * sd(y) / sd(x) / cor(y, x),
  sd(y) cor(y, x) / sd(x),
  lwd = 3, col = "blue")
abline(mean(y) - mean(x) * sd(y) / sd(x),
  sd(y) / sd(x),
  lwd = 2)
points(mean(x), mean(y), cex = 2, pch = 19)
```



var(data)=var(estimate)+var(residuals)