1 Least Squares Fit

Setup: Suppose we observe pairs of data values x_i and y_i for $i = 1, \ldots, n$ individuals. We know also that a graph of y vs. x shows linear association

- Recall: A linear fit is of the form $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$, identifying a 'typical' y value called $\hat{y}$ to go with a given x value.

A) The least squares linear fit chooses $\hat{\beta}_0$ and $\hat{\beta}_1$ to minimize the Sum of Squared Errors

$$SSE = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$ for the observed x_i and y_i values.

We are asked to show that $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$

→ First: note that the SSE can be rewritten as...

$$SSE = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} (y_i - \hat{\beta}_0 + \hat{\beta}_1 x_i)^2$$

→ Take the partial derivative with respect to $\hat{\beta}_0$

$$\frac{\partial SSE}{\partial \hat{\beta}_0} = -2 \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0$$

→ Set equal to zero and solve $\sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0$, we get...

$$\sum y_i = n\hat{\beta}_0 + \hat{\beta}_1 \sum x_i$$

→ Divide by n and solve for $\hat{\beta}_0$

$$\frac{\sum y_i}{n} = \frac{n\hat{\beta}_0}{n} + \hat{\beta}_1 \frac{\sum x_i}{n}$$

$$\Rightarrow \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

A cont.) We are also asked to show that $\hat{\beta}_1 = r\frac{S_y}{S_x}$

→ Begin by going back to the SSE, substitute $\hat{\beta}_0$ in, and take the partial derivative with respect to $\hat{\beta}_1$

$$\frac{\partial SSE}{\partial \beta_1} = 2\sum_{i=1}^{n}\left[(y_i-\bar{y})+\hat{\beta}_1(\bar{x}-x_i)\right]\cdot(\bar{x}-x_i)$$

→ Set equal to zero and expand

$$\sum(y_i-\bar{y})(\bar{x}-x_i)+\hat{\beta}_1\sum(\bar{x}-x_i)^2=0$$

$$\Rightarrow -\sum(x_i-\bar{x})(y_i-\bar{y})+\hat{\beta}_1\sum(x_i-\bar{x})^2=0$$

→ Solve for $\hat{\beta}_1$,

$$\hat{\beta}_1 = \frac{\sum(x_i-\bar{x})(y_i-\bar{y})}{\sum(x_i-\bar{x})^2} = \frac{\cancel{(n-1)}\overbrace{S_xS_yr}^{Cov(X,Y)}}{\cancel{(n-1)}\underbrace{S_x^2}_{Var(X)}} = r\frac{S_y}{S_x}$$

Finally, we are asked to note the similarity to the conditional mean of the Bivariate Normal, so consider if we have

$$\begin{pmatrix}Y\\X\end{pmatrix}\sim N_2\left(\begin{pmatrix}\mu_y\\\mu_x\end{pmatrix},\begin{pmatrix}\sigma^2_y & \rho\sigma_x\sigma_y\\ \rho\sigma_x\sigma_y & \sigma^2_x\end{pmatrix}\right)$$

→ then the conditional mean is given as

$$E[Y|X=x]=\beta_0+\beta_1 x \quad \text{where} \quad \beta_1=\rho\frac{\sigma_y}{\sigma_x}, \quad \beta_0=\mu_y-\beta_1\mu_x$$

$$E[Y|X=x]=\mu_y+\rho\frac{\sigma_y}{\sigma_x}(x-\mu_x)$$

B) We are then asked to show the inequality (1)...

$$\sum_{i=1}^{n}(y_i-(\beta_0+\beta_1 x_i))^2=\sum_{i=1}^{n}(y_i-(\hat{\beta}_0+\hat{\beta}_1 x_i))^2+\sum_{i=1}^{n}((\hat{\beta}_0+\hat{\beta}_1 x_i-(\beta_0+\beta_1 x_i))^2$$

... which shows us a Pythagorean relationship where the Total Squared Error from μ = Squared Error from the Least Squares Estimate + the squared distance between μ and the LSE. We'll define some candidate fit, μ_i, and our least squares fit, $\hat{y}_i$:

DEF. $\mu_i=\beta_0+\beta_1 x_i$ $\quad \hat{y}_i=\hat{\beta}_0+\hat{\beta}_1 x_i$

(1) can be rewritten as (2)... $\sum_{i=1}^{n}(y_i-\mu_i)^2=\sum_{i=1}^{n}(y_i-\hat{y}_i)^2+\sum_{i=1}^{n}(\hat{y}_i-\mu_i)^2$

→ Expand LHS (expand)

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

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

(why is the cross-term = 0 ?)

→ This is because in Least Squares regression the residuals are orthogonal to a linear combination of the predictors. When this cancels,

$$\sum_{i=1}^{n}(y_i-\mu_i)^2=\sum_{i=1}^{n}(y_i-\hat{y}_i)^2+\sum_{i=1}^{n}(\hat{y}_i-\mu_i)^2$$ as desired.

Finally, we are asked to explain how this inequality shows, without calculus, that $\hat{\beta}_0$ and $\hat{\beta}_1$ are the least squares estimates for β_0 and β_1.

→ If we look at the equation, we are stuck with the first term (+), so we need to minimize the second term, potentially to 0, which occurs when $\beta_0=\hat{\beta}_0$ and $\beta_1=\hat{\beta}_1$ so that $\mu_i=\hat{\mu}_i$. Any other prediction choices will increase the SS

C) Suppose we instead require $\hat{\beta}_0$ and $\hat{\beta}_1$ to be chosen such that

① $\sum(y_i - \hat{y}_i) = 0$ and ② $\sum x_i(y_i - \hat{y}_i) = 0$

→ This forces the $\hat{y}_i$ fitted means to have the same average as the y_i's and for the errors $(y_i - \hat{y}_i)$ to be uncorrelated with the x_i's.

We want to show that these requirements lead to the same estimates as the least squares.

→ Recall fitted values, $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$, so the residuals are equal to $(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)$. If we plug into both conditions ...

- Condition 1: The fitted mean values must have the same avg. as the y_i's, in other words, the sum of our errors must be 0.

$$\sum(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0 = \sum y_i - n\hat{\beta}_0 - \hat{\beta}_1 \sum x_i$$

→ Solve for $\hat{\beta}_0$:

$$n\hat{\beta}_0 = \sum y_i - \hat{\beta}_1 \sum x_i \Rightarrow \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

Which is the same formula as from Least Squares.

- Condition 2: The errors are uncorrelated with x_i, in other words, the sum of the products $x_i(y_i - \hat{y}_i)$ must be zero. We'll show by going back and substitute our know estimate for $\hat{\beta}_0$ ...

distribute $\sum x_i$: $\sum x_i(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0 = \sum x_i y_i - \hat{\beta}_0 \sum x_i - \hat{\beta}_1 \sum x_i^2$

substitute: $\sum x_i y_i - (\bar{y} - \hat{\beta}_1 \bar{x}) \sum x_i - \hat{\beta}_1 \sum x_i^2 = 0$

rearrange: $\sum x_i y_i - \bar{y} \sum x_i - \hat{\beta}_1 \bar{x} \sum x_i - \hat{\beta}_1 \sum x_i^2 = 0$

C cont.) $\sum x_i y_i - \bar{y}\sum x_i - \hat{\beta}_1(\sum x_i^2 - \bar{x}\sum x_i) = 0$

(* because * $\bar{x}\sum x_i = n\bar{x}^2$)

$$\Rightarrow \hat{\beta}_1(\sum x_i^2 - \bar{x}\sum x_i) = \sum x_i y_i - \bar{y}\sum x_i$$

$$\Rightarrow \hat{\beta}_1(\sum x_i^2 - n\bar{x}^2) = \sum x_i y_i - n\bar{x}\bar{y}$$

$$\Rightarrow \hat{\beta}_1 = \frac{\sum(x_i y_i) - n\bar{x}\bar{y}}{\sum x_i^2 - n\bar{x}^2} = \frac{n\,Cov(X,Y)}{n\,Var(X)} = r\frac{S_y}{S_x}$$

Which is the same formula as from Least Squares.

Note: Neither criterion requires an assumption of independence or any specific distribution of errors.

D) We are asked to rearrange the least squares estimates to show the regression towards the mean effect.

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1\bar{x}, \quad \hat{\beta}_1 = r\frac{S_y}{S_x}$$

→ Substitute into our regression line equation,

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i = \bar{y} - \hat{\beta}_1\bar{x} + \hat{\beta}_1 x_i$$

$$\Rightarrow \quad \hat{y}_i - \bar{y} = \hat{\beta}_1(x_i - \bar{x}) = r\frac{S_y}{S_x}(x_i - x)$$

$$\frac{\hat{y}_i - \bar{y}}{S_y} = r\left(\frac{x_i - \bar{x}}{S_x}\right)$$

→ This is our "regression towards the mean formula". The form of the equation shows the regression toward the mean effect.

→ The predicted value of y, $\hat{y}$, measured in standard deviations from its mean, $\bar{y}$, is only a factor of r of how far x is from its mean (in SDs)

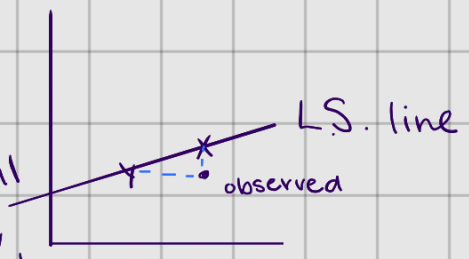
→ If $r<1$, the predicted $\hat{y}$ is closer to $\bar{y}$ than x is to $\bar{x}$

→ This means the most extreme values of x will result in less extreme predictions for y.

D cont) This effect is <u>not</u> reversible. The effect applies to both x and y, so if you tried to regress x on y, you'd get a different slope.

- Slope of y on x: $r\frac{S_y}{S_x}$
- Slope of x on y: $r\frac{S_x}{S_y}$

→ Which are not reciprocals unless $S_x = S_y$, so the regression line of y on x is not the inverse of the regression line of x on y, it is about prediction, not symmetry, so the squared errors are minimized in one direction only.



→ To illustrate, we're given a hypothetical example where we have husband, y, and wife, x, pairs, where the mean # of years of education $\bar{x} = \bar{y} = 12$, $S_x = S_y = 3$, and correlation $r = .5$

→ Say we know the wife is 6 years above average, that we know $X = 18$, then the predicted husband's education level is

$$\hat{y} = 12 + 0.5\left(\frac{3}{3}\right)(18-12) = 12+3 = 15 \text{ years}$$

only 3 years above average. In this simple example, that makes sense. We wouldn't necessarily expect both couples to be far above the mean.

BUT if we try to predict x from y, we see that $y = 15$ leads to $\hat{x} = 12 + 0.5\left(\frac{3}{3}\right)(15-12) = 12 + .5(3) = 13.5$, which is now only 1.5 years above average. This is the regressing to the mean we've been talking about, the equation isn't going to predict a more unusual result than the one we give it.

E) Finally, we have an example about casino betting on NFL games. The casino wants the

x_i - betting spread y_i - true spread (home team pts - visiting team pts)

betting spread to be close to the true spread, with a least squares fit of y_i on x_i which ideally has $\hat{\beta}_0 = 0$ and $\hat{\beta}_1 = 1$.

We suppose the casino can achieve a correlation of about $r = 0.33$ with the actual spreads (which have a standard deviation of about 15 points)

- We are asked: what should be the standard deviation of the betting spreads?

So, the casino's Least Squares regression line is $\hat{y} = \beta_0 + \beta_1 x$ but ideally it is $\hat{y} = 0 + 1x = x$

We know r and S_y, so we want to find the S_x that make $\hat{\beta}_1 = 1$. Using the slope formula

$$\hat{\beta}_1 = r \cdot \frac{S_y}{S_x} = .33 \frac{15}{S_x} = 1$$

$$\Rightarrow S_x = 0.33(15) = 4.95$$

- So, the casino should make sure the standard deviation of the betting spreads is about 4.95 in order for the regression slope of actual spreads on betting spreads to be near 1.
- This means that the betting spreads set by the casino should have much less variability than actual game spreads.
- Phil gave me data and code for an R-plot for this ↗