Gov 51: Least Squares Estimation for Linear Regression

Matthew Blackwell

Harvard University

Linear regression model

- Model for the line of best fit:

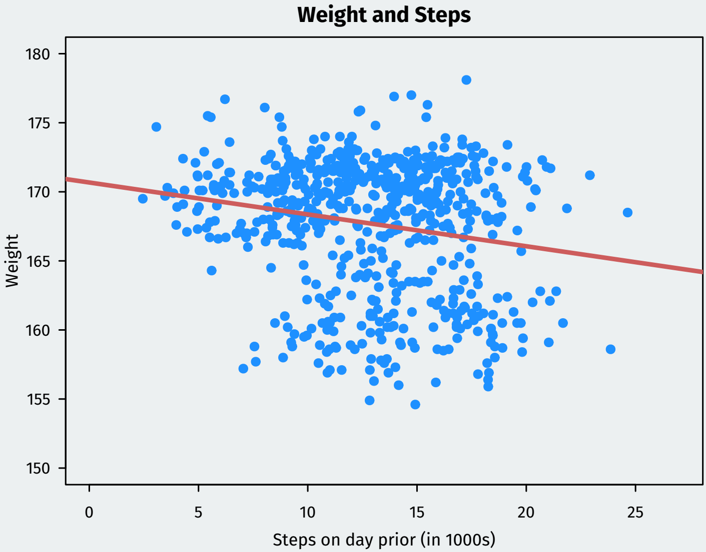
$$Y_i = \underbrace{\alpha}_{\text{intercept}} + \underbrace{\beta}_{\text{slope}} \cdot X_i + \underbrace{\epsilon_i}_{\text{error term}}$$

- **Coefficients/parameters** (α, β): true unknown intercept/slope of the line of best fit.
- **Chance error** ϵ_i: accounts for the fact that the line doesn't perfectly fit the data.
  - Each observation allowed to be off the regression line.
  - Chance errors are 0 on average.

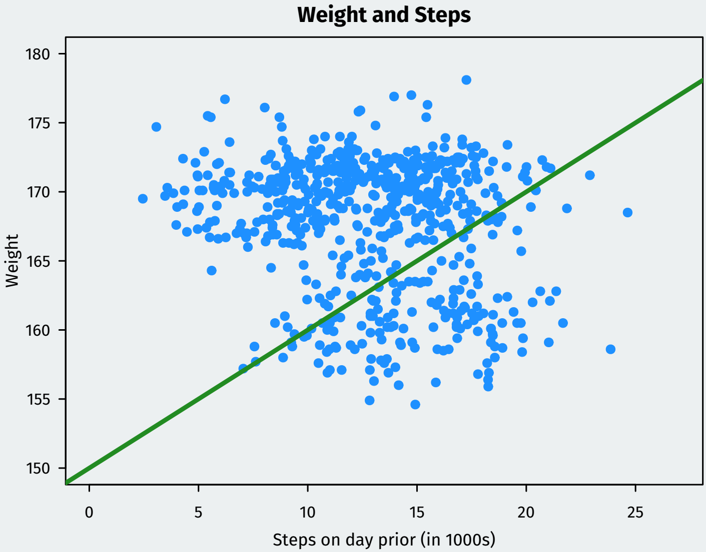
Estimated coefficients

- Parameters: α, β
  - Unknown features of the **data-generating process**.
  - Chance error makes these impossible to observe directly.
- Estimates: $\widehat{\alpha}, \widehat{\beta}$
  - An **estimate** is our best guess about some parameter.
- **Regression line**: $\widehat{Y} = \widehat{\alpha} + \widehat{\beta} \cdot x$
  - Average value of Y when X is equal to x.
  - Represents the best guess or **predicted value** of the outcome at x.

Line of best fit



Why not this line?



Least squares

- How do we figure out the best line to draw?
  - **Fitted/predicted value** for each observation: $\widehat{Y}_i = \hat{\alpha} + \hat{\beta} X_i$
  - **Residual/prediction error**: $\hat{\epsilon}_i = Y_i - \widehat{Y}$
- Get these estimates by the **least squares method**.
- Minimize the **sum of the squared residuals** (SSR):

$$\text{SSR} = \sum_{i=1}^{n} \hat{\epsilon}_i^2 = \sum_{i=1}^{n} (Y_i - \hat{\alpha} - \hat{\beta} X_i)^2$$

- Finds the line that minimizes the magnitude of the prediction errors!

Linear regression in R

- R will calculate least squares line for a data set using `lm()`.
  - Syntax: `lm(y ~ x, data = mydata)`
  - `y` is the name of the dependent variance
  - `x` is the name of the independent variable
  - `mydata` is the data.frame where they live

```r
fit <- lm(weight ~ steps.lag, data = health)
fit
```

```
##
## Call:
## lm(formula = weight ~ steps.lag, data = health)
##
## Coefficients:
## (Intercept)    steps.lag
##     170.675       -0.231
```

Coefficients and fitted values

- Use `coef()` to extract estimated coefficients:

```
coef(fit)
```

```
## (Intercept)   steps.lag
##     170.675      -0.231
```

- R can show you each of the fitted values as well:

```
head(fitted(fit))
```

```
##   2   3   4   5   6   7
## 167 166 166 168 166 169
```

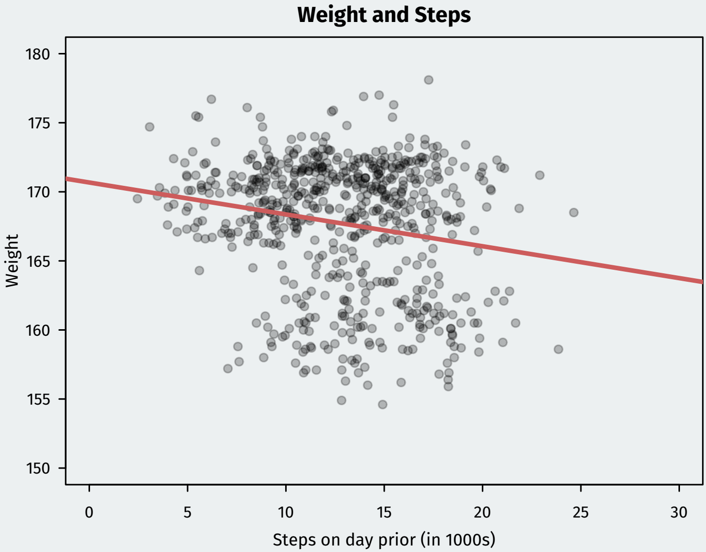
Properties of least squares

- Least squares line always goes through $(\overline{X}, \overline{Y})$.
- Estimated slope is related to correlation:

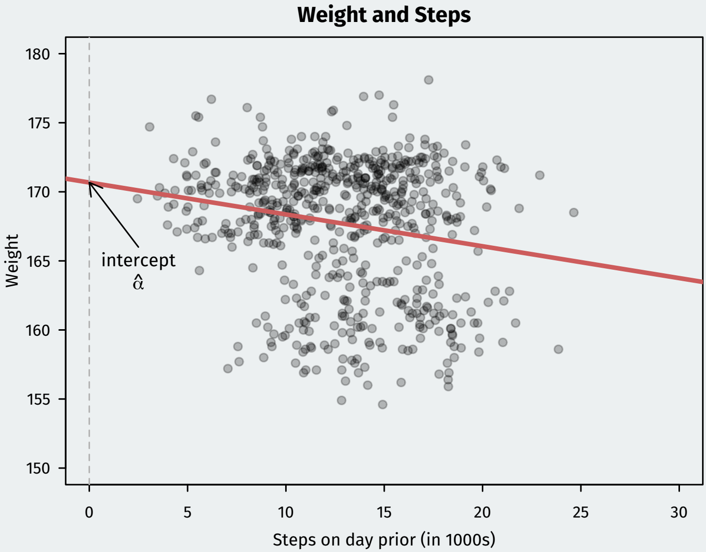
$$\hat{\beta} = (\text{correlation of } X \text{ and } Y) \times \frac{\text{SD of } Y}{\text{SD of } X}$$

- Mean of residuals is always 0.

Visual components of least squares



Visual components of least squares



Visual components of least squares

Visual components of least squares

Visual components of least squares

Visual components of least squares