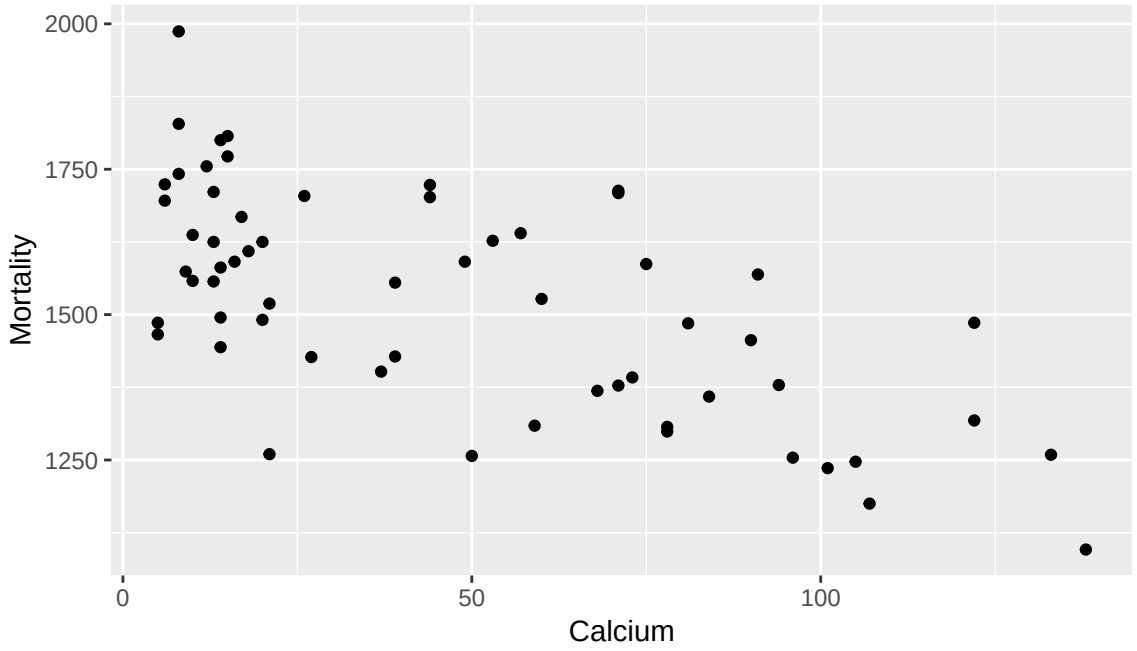
Linear Regression - First Example

Mortality and Hard Water

- Scientists believe that hard water (water with high concentrations of calcium and magnesium) is beneficial for health (Sengupta, P. (2013). IJPM, 4(8), 866-875.)
- We have recordings of the mortality rate (deaths per 100,000 population) and concentration of calcium in drinking water (parts per million) in 61 large towns in England and Wales

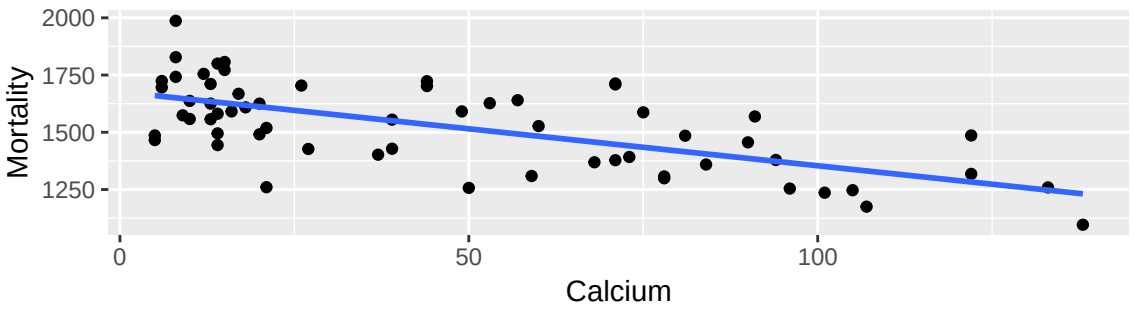
```
mortality_water <- read_csv("https://mhc-stat140-2017.github.io/data/sdm4/Har

ggplot(data = mortality_water, mapping = aes(x = Calcium, y = Mortality)) +
  geom_point()
```



- **Explanatory** (~~independent~~) variable goes on the x (horizontal) axis
- **Response** (~~dependent~~) variable on the y (vertical) axis
- **Notation**:
  - x_i: value of explanatory variable (Calcium) for observational unit number i
  - y_i: value of response variable (Mortality) for observational unit number i
- **Big idea**: Summarize the relationship between these variables with a line.

Add a line to the plot

```
ggplot(data = mortality_water, mapping = aes(x = Calcium, y = Mortality)) +
  geom_point() +
  geom_smooth(method = "lm", se = FALSE)
```



Estimate intercept and slope of line

```
linear_fit <- lm(Mortality ~ Calcium, data = mortality_water)
```

General form of `lm` command:

```
lm(response_variable ~ explanatory_variable, data = data_frame)
```

View summary of linear model fit

```
summary(linear_fit)
```

```
##
## Call:
## lm(formula = Mortality ~ Calcium, data = mortality_water)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -348.61 -114.52   -7.09  111.52  336.45
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 1676.3556    29.2981  57.217  < 2e-16 ***
## Calcium       -3.2261     0.4847  -6.656 1.03e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 143 on 59 degrees of freedom
## Multiple R-squared:  0.4288, Adjusted R-squared:  0.4191
## F-statistic:  44.3 on 1 and 59 DF,  p-value: 1.033e-08
```

Foundations

1. What is the estimated equation for the regression line?

2. What is the estimated slope and its interpretation?

3. What is the estimated intercept and its interpretation?

4. One of the towns in our sample had a measured Calcium concentration of 71. What is the predicted value for the mortality rate in that town?

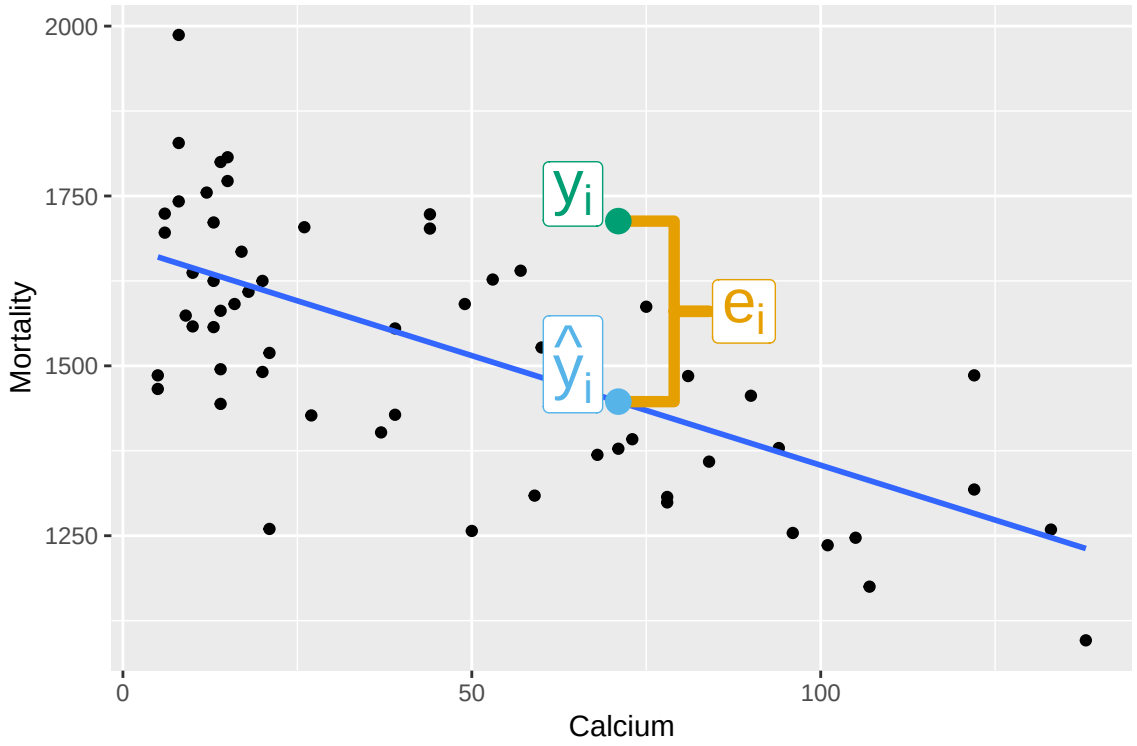
```
predict_data <- data.frame(
  Calcium = 71
)
predict(linear_fit, newdata = predict_data)
```

```
##        1
## 1447.303
```

5. Based on this analysis, does increasing the concentration of Calcium in drinking water cause a reduction in the mortality rate?

Residuals

- Residual = *Observed* - *Predicted*
- $e_i = y_i$ - $\hat{y}_i$ (e stands for error)



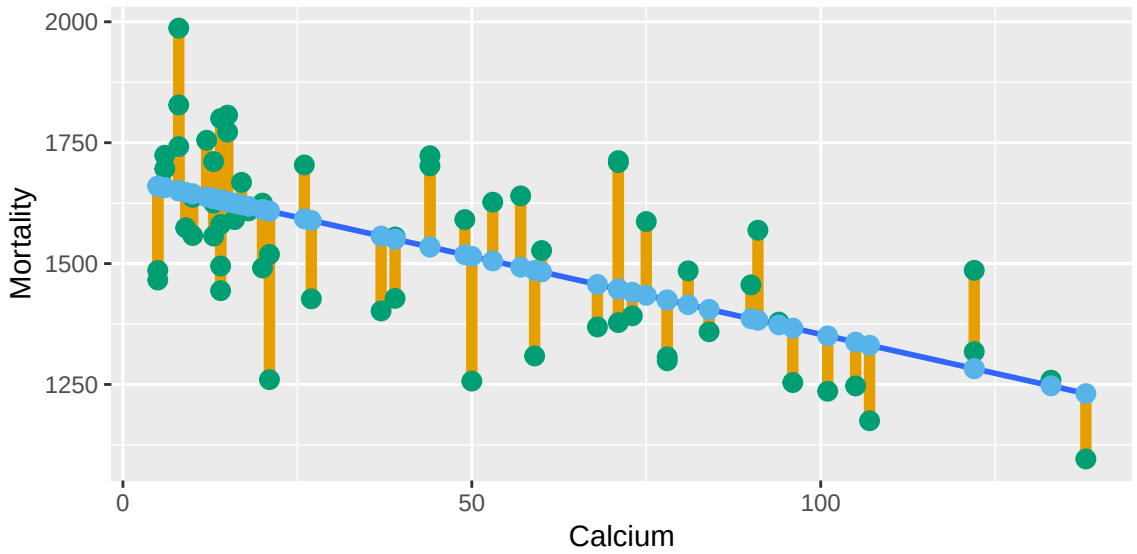
6. For the town we considered in part 4 above, with an observed Calcium measurement of 71 ppm, the observed Mortality rate was 1713 deaths per 100,000 population. What was the residual for that town?

7. Another town had an observed Calcium measurement of 50 ppm, and an observed Mortality rate of 1257 deaths per 100,000 population. What was the residual for that town?

```
predict(linear_fit, newdata = data.frame(Calcium = 50))
```

```
##        1
## 1515.051
```

Finding the Line of Best Fit



- The "best" line has the smallest residuals
- Pick $\hat{\beta}_0, \hat{\beta}_1$ to minimize sum of squared errors/residuals: $SSE = \sum_{i=1}^{n} e_i^2$

$$\text{Slope: } \hat{\beta}_1 = r\frac{s_y}{s_x} \qquad \text{Intercept: } \hat{\beta}_0 = \bar{y} - b_1\bar{x}$$