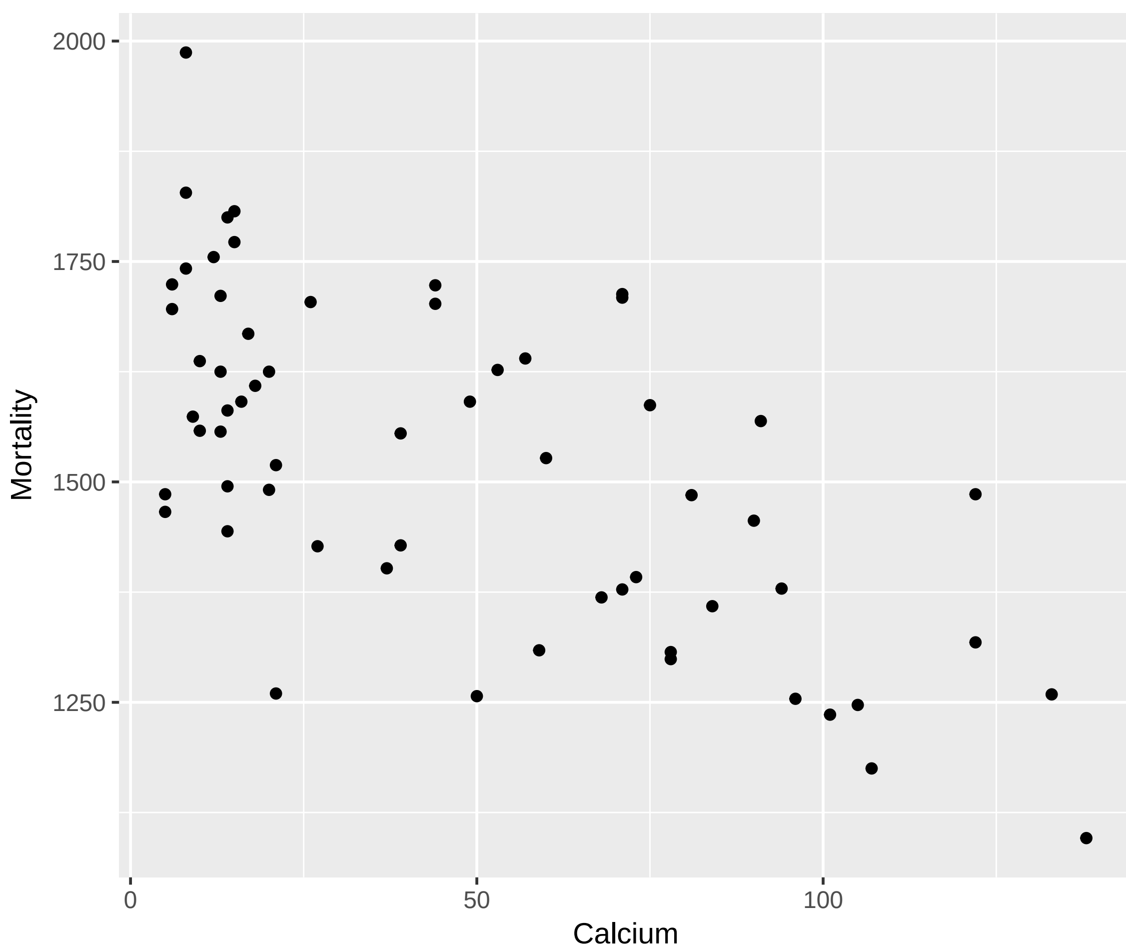
Linear Regression - First Example

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/Hard_water_Derby.csv")

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

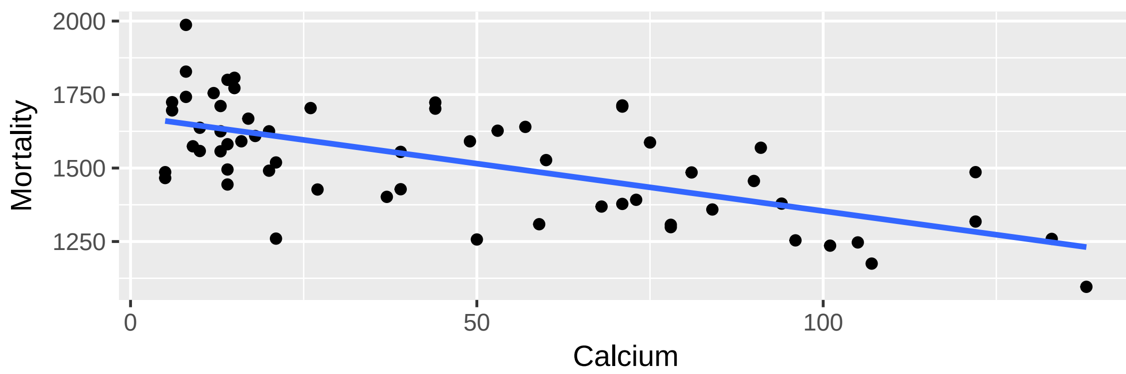


- **Explanatory** (~~independent~~) variable goes on the x axis, **Response** (~~dependent~~) variable on the y axis
- **Notation**:
  - x_i: value of explanatory variable (Calcium) for observational unit number i
  - y_i: value of response variable (Mortality) for observational unit number i
- **Big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

R Commands (first look - more to come)

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 intercept and its interpretation?

2. What is the estimated slope and its interpretation?

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

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?

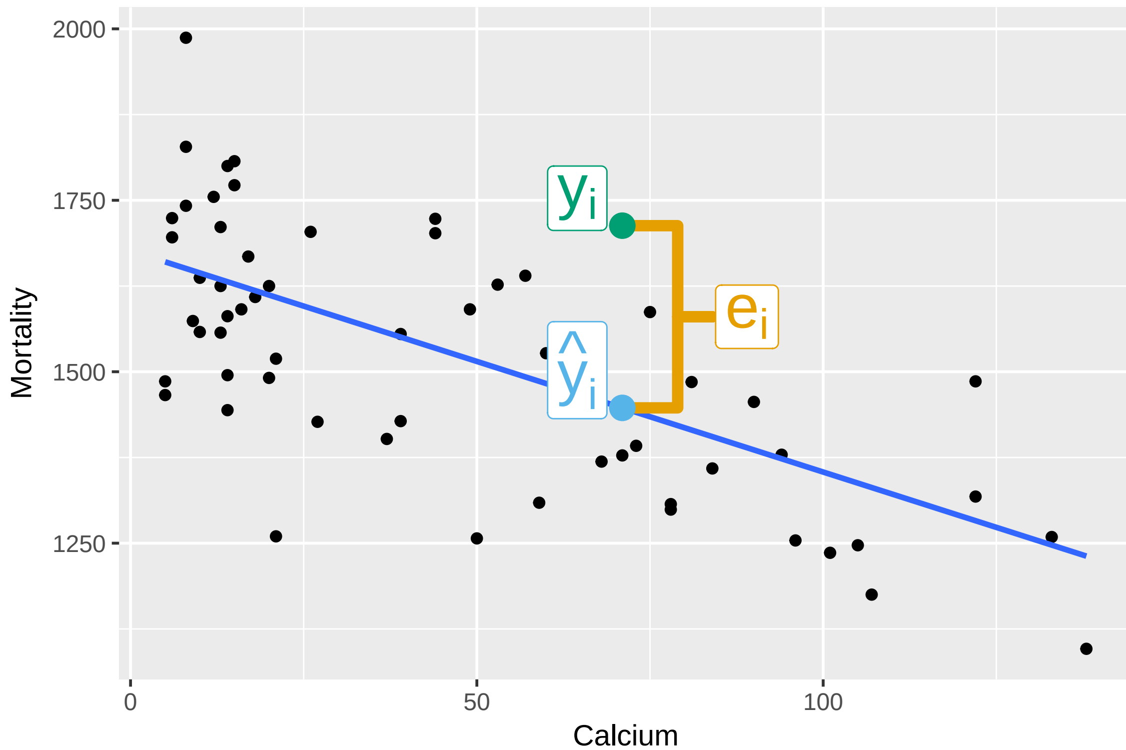
```
predict_data <- data.frame(
  Calcium = 71
)
predict(linear_fit, newdata = predict_data)
```

```
##        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*
- $e_i = y_i - \widehat{y}_i$ (e stands for error)



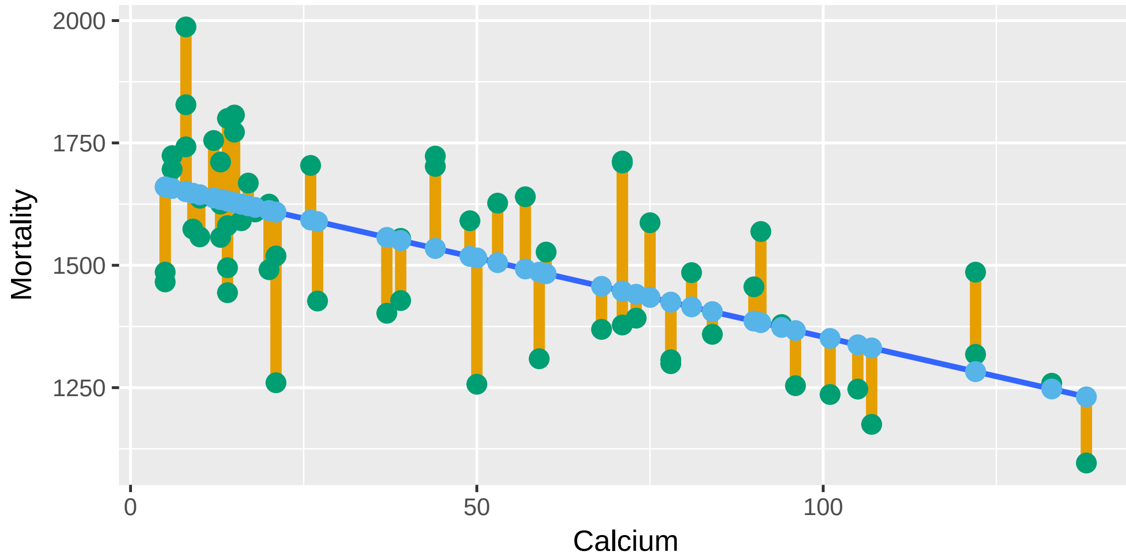
6. For the town we considered in part 4 above, with an observed Calcium measurement of 71 ppm, the observed Mortality rate was 1713 deaths per 100,000 population. What was the residual for that town?

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?

```r
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 b_0, b_1 to minimize sum of squared errors/residuals: $SSE = \sum_{i=1}^{n} e_i^2$

$$b_1 = r\frac{s_y}{s_x} \qquad\qquad b_0 = \bar{y} - b_1\bar{x}$$

How useful is the line?

Rephrased: How close to the line do the points fall?

Answer 1: R^2

Recall: Correlation (r) is a measure of the strength of a linear relationship.

R^2 (read as "R squared") is the square of the correlation.

"Proportion of variability in the response variable accounted for by the linear model." (See office hours or intermediate statistics for more.)

8. What is the R^2 value for this model fit?

```
mortality_water %>%
  select(Calcium, Mortality) %>%
  cor()
```

```
##               Calcium  Mortality
## Calcium    1.0000000 -0.6548486
## Mortality -0.6548486  1.0000000
```

```
(-0.6548486)^2
```

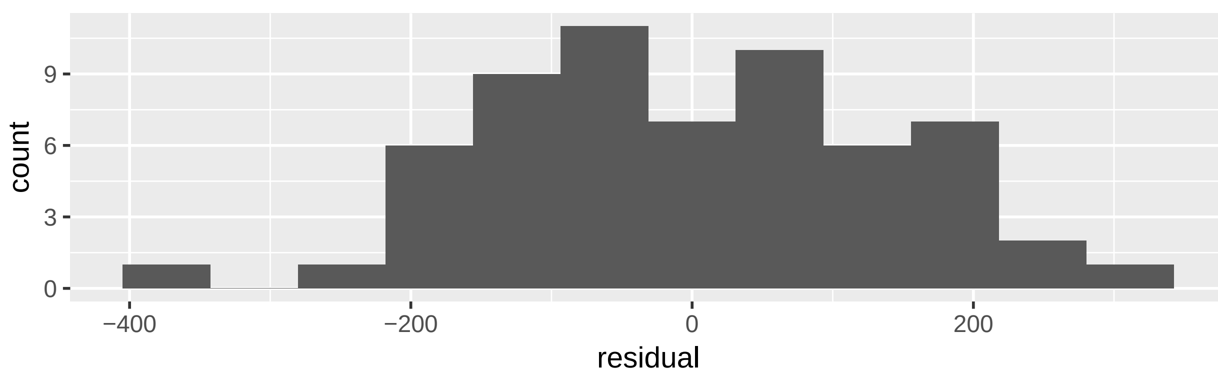
```
## [1] 0.4288267
```

Answer 2: How big do the residuals tend to be?

```r
mortality_water <- mortality_water %>% mutate(
  predicted = predict(linear_fit),
  residual = residuals(linear_fit)
)
head(mortality_water)
```

```
## # A tibble: 6 x 5
##   Mortality Calcium Derby predicted   residual
##       <dbl>   <int> <chr>     <dbl>      <dbl>
## 1      1702      44 South  1534.408  167.59243
## 2      1309      59 South  1486.016 -177.01620
## 3      1259     133 South  1247.285   11.71458
## 4      1427      27 North  1589.251 -162.25113
## 5      1724       6 North  1656.999   67.00095
## 6      1175     107 South  1331.164 -156.16380
```

```r
ggplot(data = mortality_water, mapping = aes(x = residual)) + geom_histogram(bins = 12)
```



If the residuals are all close to 0, the observed values are all close to the predicted values (i.e., the points are close to the line).

9. What is the residual standard deviation?

10. Assuming that the residuals follow an approximately normal distribution, what is the interpretation of the residual standard deviation based on the "95" part of the 68-95-99.7 rule?