Week 3: Basic regression

3. Compute a linear model in R

Stat 140 - 04

Mount Holyoke College

Dr. Shan Shan

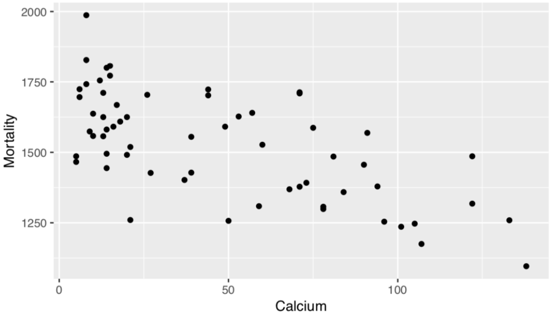
Slides posted at *http://sshanshans.github.io/stat140*

1. Main ideas
    1. Finding the least square line in R
    2. How useful is the model?

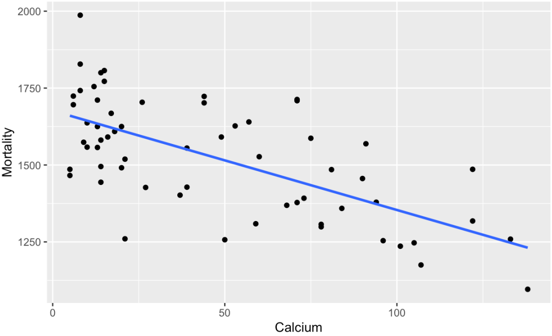
2. Summary

Scientists believe that water with high concentrations of calcium and magnesium is beneficial for health.

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.



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



geom_smooth(method = "lm", se = FALSE)

General form of ‘lm‘ command:

```
lm(y_variable ~ x_variable, data = data_frame)
```

Use this to estimate the intercept and the slope of line in the Mortality/Health data.

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

```
Coefficients:
(Intercept)      Calcium
   1676.356       -3.226
```

From the coefficients, we can write the regression line in the Mortality/Health data as

$$\widehat{\text{Mortality}} = 1676 - 3 \text{ Calcium}$$

Abstractly,

$$\hat{y} = 1676 - 3x$$

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?

$$\widehat{\text{Mortality}} = 1676 - 3 \text{ Calcium}$$

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?

$$\widehat{\text{Mortality}} = 1676 - 3 \text{ Calcium}$$

By hand:

$$\widehat{\text{Mortality}} = 1676 - 3 \times 71 = 1463$$

The predicted value for the mortality rate in that town is 1463 deaths per 100,000 population.

By R:

The general form of ‘predict‘ command:

predict(linear_model, newdata = data_frame)

Code:

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

Output:

```
       1
1447.303
```

By R:

The general form of ‘predict‘ command:

predict(linear_model, newdata = data_frame)

Code:

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

Output:

```
       1
1447.303
```

The outputs by hand and by R are different because of rounding errors.

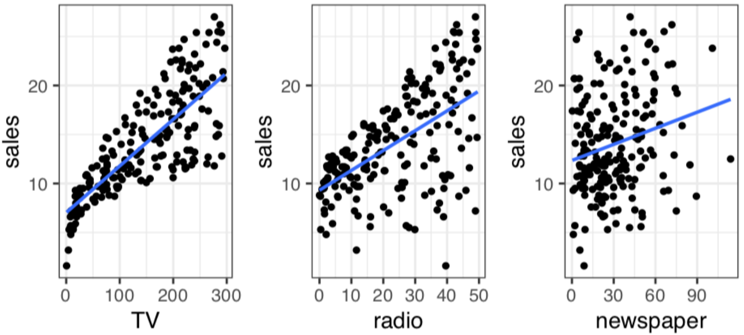
We have a data set of 200 markets, and we are interested in the relationship between sales and advertising budget. We look at the following variables.

- **sales** is a measure of sales volume in thousands of units
- **TV** is TV advertising budget
- **radio** is radio advertising budget
- **newspaper** is newspaper advertising budget

Poll question

Which of these models would you prefer to use for predicting sales?

- (a) TV
- (b) radio
- (c) newspaper

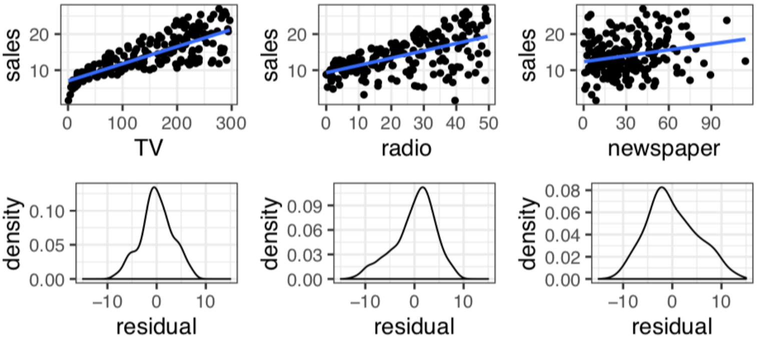


Being as specific and concrete as possible, write down a rule for selecting your preferred model

1. based only on **visual characteristics** of the plot.
2. based only on **a quantitative summary** of the data. You can describe how you would calculate your numeric summary of the data in a general sense; if you'd like you can write down a formula.

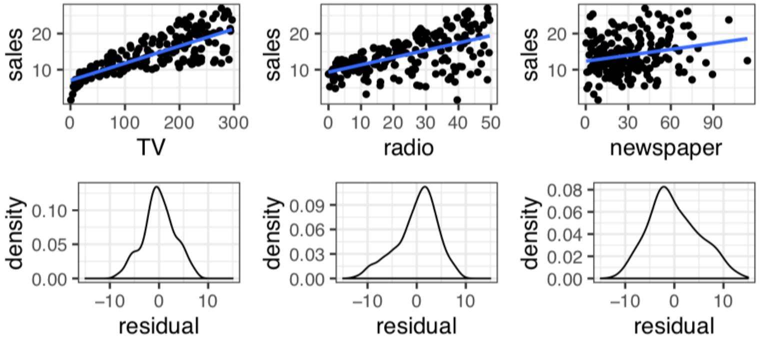
Residuals:

- $e_i = y_i - \hat{y}_i$ (vertical distance between point and line)
- Smaller residuals mean the predictions were better.
- The key is to measure the spread of residuals.

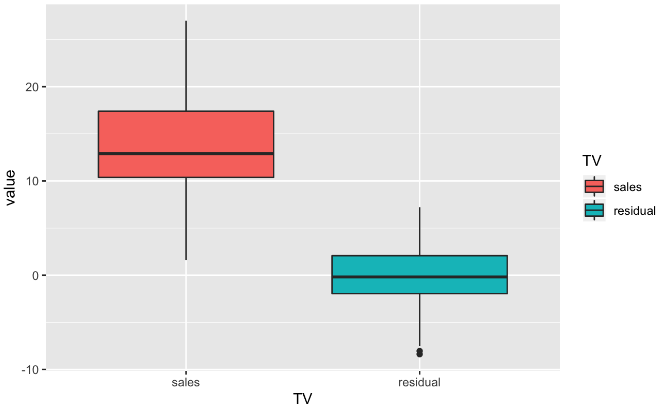


Measure spread of residuals with the standard deviation. We call this the **residual standard error**, s_{RES}.

- TV: 3.26
- radio: 4.28
- newspaper: 5.09



The variability in the residuals describes how much variation remains after using the model



Let's compute the reduction in variation.

$$\frac{s^2_{\text{sales}} - s^2_{\text{RES}}}{s^2_{\text{sales}}} = 0.61$$

This number describes the amount of variation in the y-variable that is explained by the least squares line.

An value of 61% indicates that 61% of the variation in sales can be accounted for by the TV advertisement budget.

Let's compute the reduction in variation.

$$\frac{s^2_{\textsf{sales}} - s^2_{\textsf{RES}}}{s^2_{\textsf{sales}}} = 0.61$$

This number describes the amount of variation in the y-variable that is explained by the least squares line.

An value of 61% indicates that 61% of the variation in sales can be accounted for by the TV advertisement budget.

How do we get 61%?

Let's compute the reduction in variation.

$$\frac{s^2_{\textsf{sales}} - s^2_{\textsf{RES}}}{s^2_{\textsf{sales}}} = 0.61$$

This number describes the amount of variation in the y-variable that is explained by the least squares line.

An value of 61% indicates that 61% of the variation in sales can be accounted for by the TV advertisement budget.

How do we get 61%?

Statisticians found this value is R^2, the **square of Correlation**.

Square of the correlation coefficient R: between 0 and 1, closer to 1 is better.

R^2 describes the amount of variation in the y-variable that is explained by the least squares line.

- TV: 0.61
- radio: 0.33
- newspaper: 0.05

meaning, 61% of the variation in sales can be accounted for by the TV advertisement budget; 33% of the variation in sales can be accounted for by the radio advertisement budget; 5% of the variation in sales can be accounted for by the newspaper advertisement budget.

```
##
## 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
```

b0 intercept

b1 slope

Useful later

R squared

1. Finding the least square line in R
2. How useful is the model?

Interpretation is the most important thing in this module.

- slope
- intercept
- residual
- R^2