# 7: Linear Regression Models with a Single Predictor (Part 1)

# Packages Needed To Recreate Code on Slides

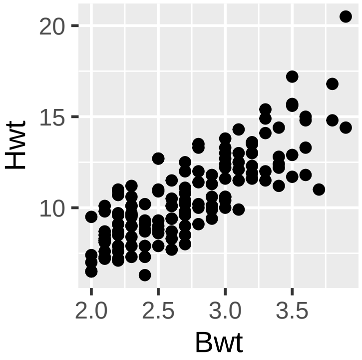
```r
library(MASS)       # for data set
library(tidyverse) # for ggplot functions/plotting
```

*Warning:* It is not expected that you understand all the R code in this presentation right now. You will go over more R code in SDS 100 in the coming weeks. However, you are welcome to try to make these plots on your own.

# Linear Regression with a Single Predictor

In this class we will focus on linear regression models, where we seek to model a relationship between two numerical variables using a straight line.

```
ggplot(cats, aes(x = Bwt, y = Hwt))+
  geom_point()
```



# Linear Regression with a Single Predictor

What do we mean by saying *"linear regression model with a single predictor"*

- **predict**: indicate in advance
  - *$x$ can help us indicate what $y$ will be.*
- **regress**: to tend to approach or revert to a value/relation
  - *$x$ & $y$ values approach a common relationship.*
- **linear**: $y = b_0 + b_1 x$
  - *$x$ & $y$ relationship can roughly be described by a straight line.*
- **model**: an informative representation of an object, person or system.
  - *Educated guess for $b_0$ & $b_1$ describing the relationsip of $x$ & $y$.*

# Linear Regression with a Single Predictor

Linear regression models can be used for:

1) prediction

*If I have a new data value $x^*$, can I guess what it's corresponding value for $y$ would be?*

2) evaluate whether there is a linear relationship between two numerical variables.

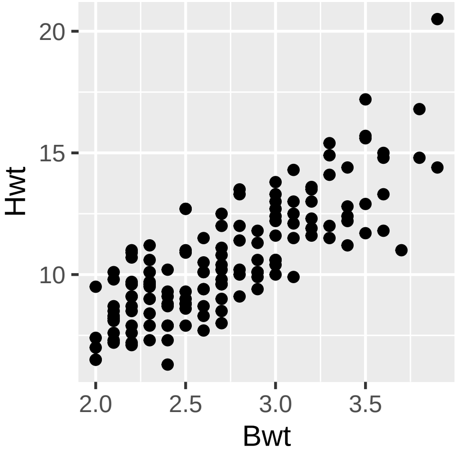
*Does a linear relationship exist between $x$ and $y$?*

# Linear Regression with a Single Predictor

- Larger heart weights indicate a higher risk of heart attacks/disease in cats; however, heart weight is hard to measure.
- Want to see if there is a relationship between heart weight (`Hwt`) and body weight (`Bwt`) for domestic cats.
- If so, we will have a better idea of which cats are at risk for heart attacks/disease.

# Real Sample Data for Domestic Cats

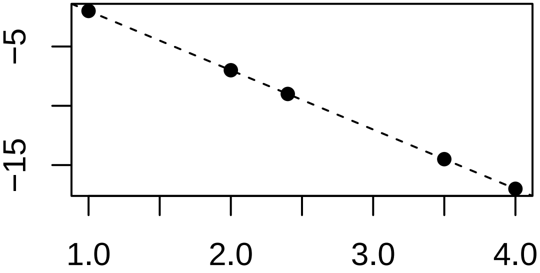
```r
ggplot(cats, aes(x = Bwt, y = Hwt))+
  geom_point()
```



# Lines in Mathematics

$$y = b_0 + b_1 x$$

- A linear regression line in mathematics usually takes the above form.
- In a typical math class this is a perfect relationship!
- For example: Let $y = 3 - 5x$ and consider points $x = 1, 2, 2.4, 4, 3.5$



# Fitting a line to data

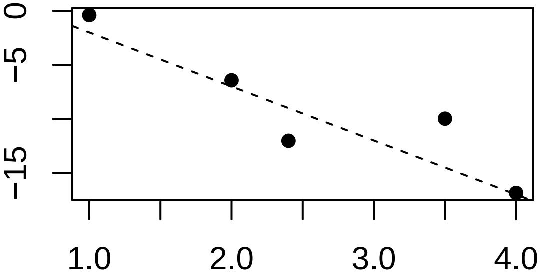
- Instead of the usual math equation we add an *error* term

$$y = b_0 + b_1 x + e$$

- $b_0$: intercept
- $b_1$: slope
- $x$: **predictor** variable
- $y$: **response** variable
- $e$: error (source of wiggliness around the line)

# Fitting a line to data

When we add the random error.

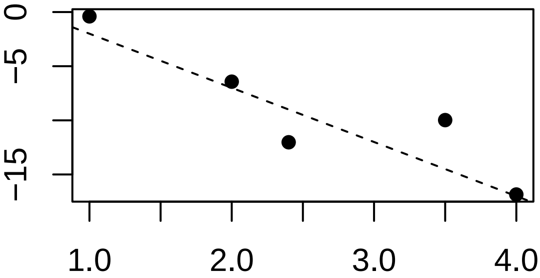



We want to find a good $b_0$ and $b_1$, but now it is not as straight forward because the points do not perfectly fall on the line.

If the dashed line was *not* known, how would we find it?

# Fitting a line to data

When we add the random error.




We want to find a good $b_0$ and $b_1$, but now it is not as straight forward because the points do not perfectly fall on the line.

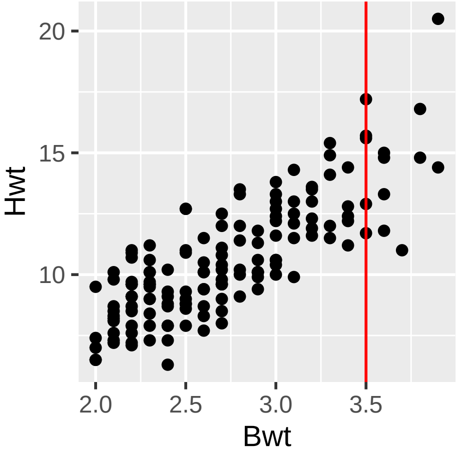
If the dashed line was *not* known, how would we find it?

*Coming soon!*

# Predicting New Values

The cat below weighs 3.5 kg, but we do NOT know the heart weight.

Can we guess this cat's heart weight?



# Predicting New Values

Suppose our regression line is

$$\texttt{Hwt} = -0.3567 + 4.0341\ \texttt{Bwt}$$

We can use this line to predict a new value. We denote predicted values with a hat.

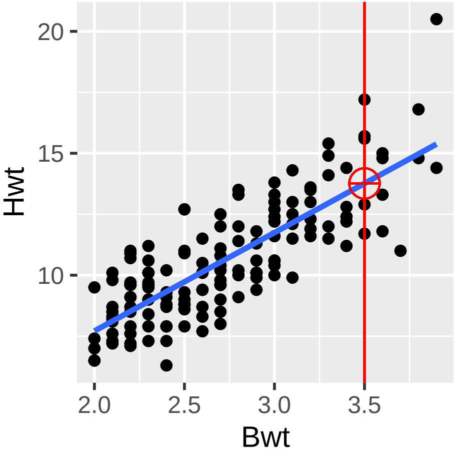
$$\hat{y} = b_0 + b_1 x$$

or for this specific situation we can write

$$\widehat{\texttt{Hwt}} = -0.3567 + 4.0341\ \texttt{Bwt}$$

# Predicting New Values

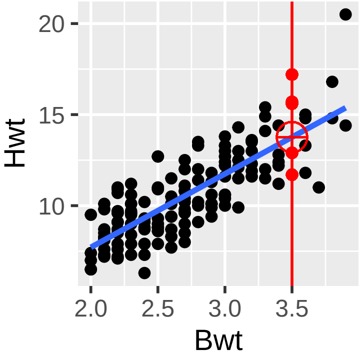
New value is $\widehat{\texttt{Hwt}} = 13.8$ (rounded)




# What does this predicted value mean?

- Note: five cats in the data set that had body weight 3.5. We can of the predicted value as the mean heart weight for cats with a body weight of 3.5.
- That is, if we believe this is the true linear relationship, *the equation predicts that cats with a body weight of 3.5 kg will have an mean heart weight of 13.8 g*.

| Subject | Bwt | Hwt |
|---|---|---|
| 67 | 3.50 | 17.20 |
| 106 | 3.50 | 15.70 |
| 112 | 3.50 | 15.60 |
| 81 | 3.50 | 12.90 |
| 125 | 3.50 | 11.70 |



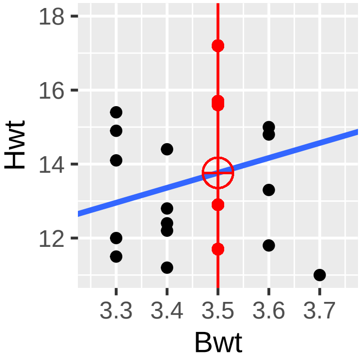
# What does this predicted value mean?

- **predict**: indicate in advance
  - *Bwt can help us indicate what Hwt will be.*
- **regress**: to tend to approach or revert to a value/relation
  - *For any value of Bwt, there is a mean Hwt.*
- **linear**: $y = b_0 + b_1 x$
  - *The relationship between Hwt and Bwt loosly resemble a line, so the means probably do too.*
- **model**: an informative representation of an object, person or system.
  - *We do not know the values for $b_0$ and $b_1$, we have to guess.*

# Residuals

- None of the cats in the data set with `Bwt = 3.5` had $\widehat{\texttt{Hwt}} = 13.8$.
- The difference between the values in the data set and their corresponding fitted value is called the **residual**.

| Subject | Bwt | Hwt | residual |
|---:|---:|---:|---:|
| 138 | 3.50 | 17.20 | 3.40 |
| 110 | 3.50 | 15.70 | 1.90 |
| 115 | 3.50 | 15.60 | 1.80 |
| 23 | 3.50 | 12.90 | -0.90 |
| 96 | 3.50 | 11.70 | -2.10 |



# Residuals

More formula (and general):

$$e_i = y_i - \hat{y}_i$$

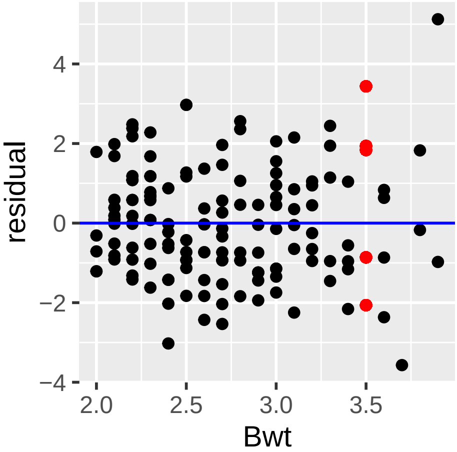
For this situation:

$$e_i = \texttt{Hwt}_i - \widehat{\texttt{Hwt}}_i$$

- We can calculate the residual for *every* observation in the data set. This is also called the **error** (the wiggiliness from the line).
- A residual can also be described as the difference between observed ($y_i$) and fitted ($\hat{y}_i$) values.

# Residual Plot

- Residuals are helpful in evaluating how well a linear model fits a data set.
- The residual is on the y-axis, and the predictor variable is still on the x-axis.



# Residual Plot

What are we looking for in a residual plot?

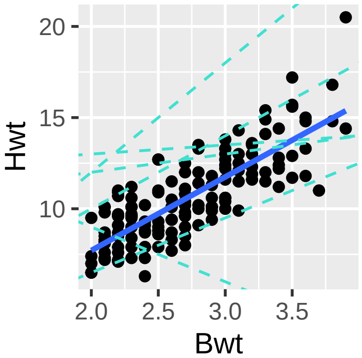
- We want to see a cloud of points with no pattern.
- Recall single numerical variable plots: dot plot, bar plot, histogram.
- Imagine one of these plots for each value on the x axis. Would these plots look the same?

# Connection to Correlation

- Correlation and linear regression models with a single predictor are VERY RELATED!
- Correlation is an indicator of how well the *best linear model* would represent the data.

# How do we find the best line?

- We want the line that makes *all* of the residuals as small as possible.
- There are infinitely many possible lines, we can find the *best* line with mathematics!



# How do we find the best line?

Ideally, we might want a line that minimizes the distance betweeen the observed and fitted values,

$$\sum_{i=1}^{n} |e_i| = |e_1| + |e_2| + ... + |e_n|$$

this is a great goal! However, in practice squared distance is more practical

$$\sum_{i=1}^{n} e_i^2 = e_1^2 + e_2^2 + ... + e_n^2$$

# How do we find the best line?

Why squared residuals:

1) A residual twice as large as another residual is more than twice as bad.
2) Easier to work with then absolute values.
3) Better statistical properties (this metric comes up more naturally in theorems).
4) The most supported technique in current statistical software.

# How do we find the best line?

Thus we want to find $b_0$ and $b_1$ that minimize

$$\sum_{i=1}^{n} e_i^2 = e_1^2 + e_2^2 + ... + e_n^2$$

Equivalently, we want to find $b_0$ and $b_1$ that minimize

$$\sum_{i=1}^{n} (y_i - b_0 - b_1 x_i)^2$$

# Coming Up!

# Other Variables

It is possible that other factors could influence these variables. For example, what about Sex?