# *Lab2 Tutorial: Linear Regression*

## *BIOL 414/514*

## *Quantitative Analysis of Vertebrate Populations*

*Lab adapted from Linear Regression I by Dr. Nicholas Nagel at the University of Tennessee*

First let's load the R packages that we will use:

```
library(dplyr)
library(ggplot2)
```

### *Regression with the sparrow data*

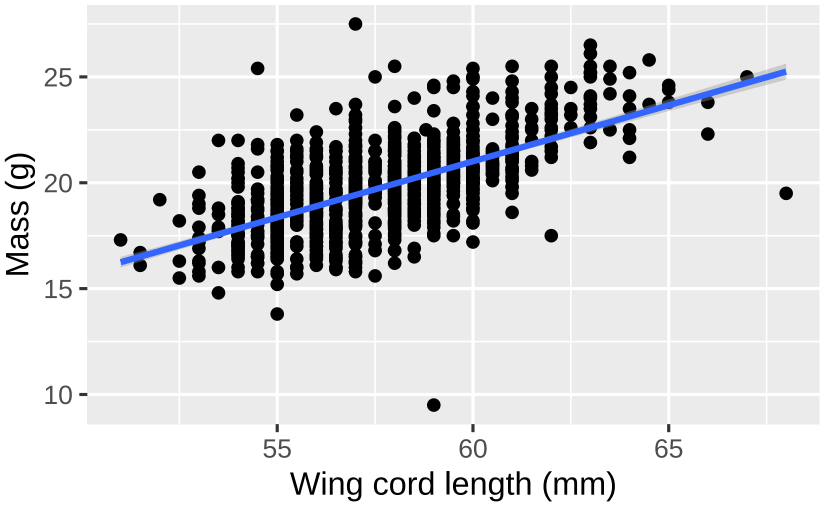
Let's use the sparrow data again because we are familiar with it.

```
sparrows <- read.table("Data/Sparrows.txt",
                       header = TRUE,
                       stringsAsFactors = FALSE)
str(sparrows)
```

```
## 'data.frame':    1281 obs. of  12 variables:
##  $ Speciescode: int  1 1 1 1 1 1 1 1 1 1 ...
##  $ Sex        : int  0 0 0 0 0 0 5 0 0 5 ...
##  $ SexNew     : int  1 1 1 1 1 1 6 1 1 6 ...
##  $ wingcrd    : num  59 55 53.5 55 52.5 57.5 53 55 55 55.5 ...
##  $ flatwing   : num  60 56 54.5 56 53.5 59 54 56 56 56.5 ...
##  $ tarsus     : num  22.3 19.7 20.8 20.3 20.8 21.5 20.6 21.5 20.8 20.5 ...
##  $ head       : num  31.2 30.4 30.6 30.3 30.3 30.8 32.5 31.2 31.6 31.4 ...
##  $ culmen     : num  12.3 12.1 12.8 11.9 12.6 12 13.5 12.3 13.2 13.2 ...
##  $ nalospi    : num  13 8.3 8.6 8.7 8.8 8.1 10.7 8.7 9.1 10.5 ...
##  $ wt         : num  9.5 13.8 14.8 15.2 15.5 15.6 15.6 15.7 15.7 15.7 ...
##  $ observer   : int  2 8 7 3 3 2 3 5 2 6 ...
##  $ Age        : int  2 2 2 1 1 2 1 1 1 1 ...
```

Let's pick a length measurement, wingcrd, and estimate the relationship with mass (unfortunately named wt for weight). Last lab we plotted it and included a regression line using the ggplot() function in the ggplot2 package:

```
ggplot(sparrows, aes(x = wingcrd, y = wt)) +
  geom_point() +
  labs(x='Wing cord length (mm)', y='Mass (g)') +
  geom_smooth(method = "lm")
```



*The linear model function and interpretation*

Now, let's fit the regression using the function `lm`. To use `lm`, we need to specify the y and x variables. R has a type of "short hand" for linear models that looks like this : `y ~ x`. R will interpret that $y = b_0 + b_1 x$.

```
my_lm <- lm(wt ~ wingcrd, sparrows)
```

Right now, R has finished the regression, and we just need to know how to print it out.

```
summary(my_lm)
```

```
##
## Call:
## lm(formula = wt ~ wingcrd, data = sparrows)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -10.9821  -0.8766   0.0017   0.7828   8.0774
##
## Coefficients:
##              Estimate Std. Error t value
## (Intercept) -10.77121    1.05992  -10.16
## wingcrd       0.52972    0.01835   28.86
##             Pr(>|t|)
## (Intercept)   <2e-16 ***
## wingcrd       <2e-16 ***
## ---
## Signif. codes:
```

```
##   0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1  ' ' 1
##
## Residual standard error: 1.506 on 1279 degrees of freedom
## Multiple R-squared:  0.3945, Adjusted R-squared:  0.394
## F-statistic: 833.1 on 1 and 1279 DF,  p-value: < 2.2e-16
```

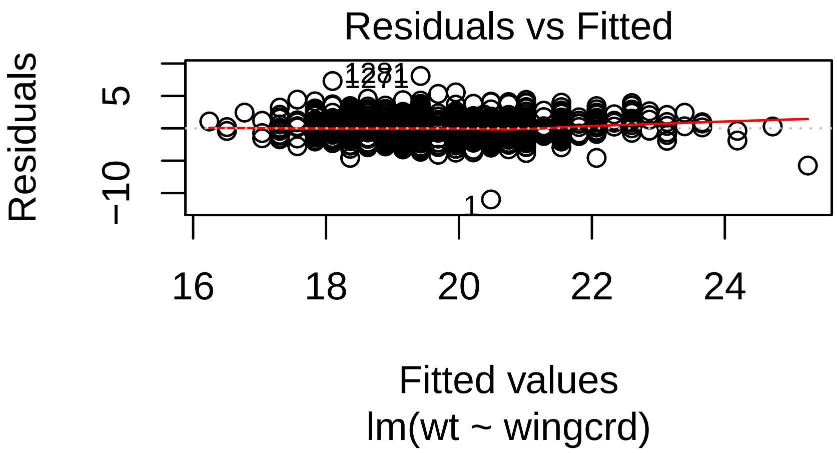
That's it. Pretty easy, right?

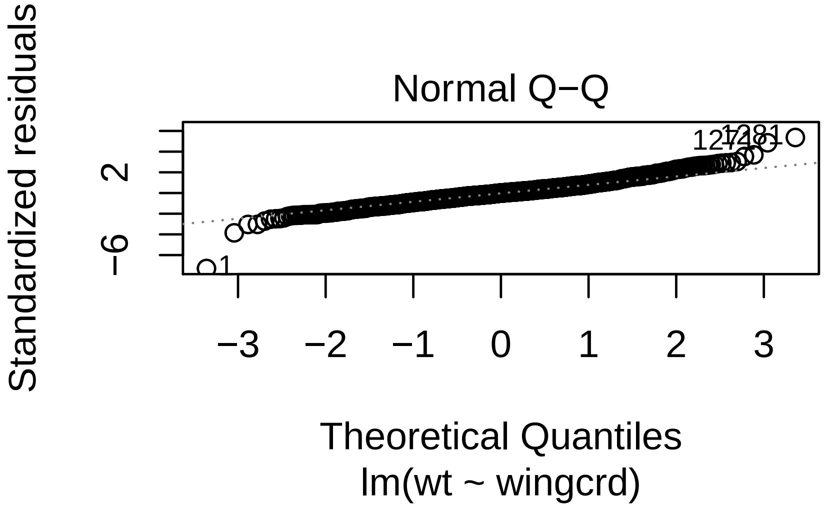
What does the line predict for the mass for a bird with a wing cord of 70 mm? It predicts $-10.77121 + 0.52972 * 70$ or 26.3 g. This might be reasonable.

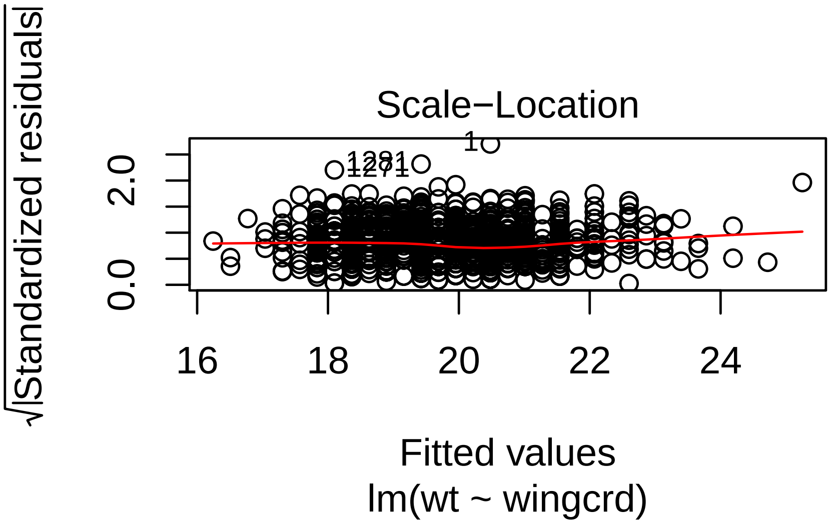
What does the line predict for the mass for a bird with a wing cord of 200 mm? It predicts $-10.77121 + 0.52972 * 200$ or 95.1 g. This would be a big bird and is totally unrealistic for a sparrow. Be careful when making predictions. In general, don't make predictions that aren't supported by your data.

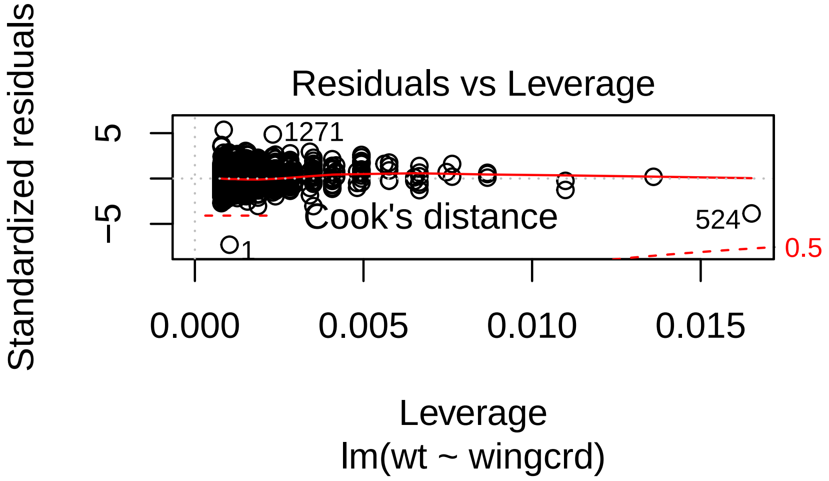
We always want to check the to see if our model and data matched the assumptions of the linear model (refer back to your lecture notes). We can do this by visualizing the residuals. It's easy with the plot function on your fitted model object.

```
plot(my_lm)
```




Normal Q−Q
Standardized residuals
−6
2
1271
1281
1
−3
−2
−1
0
1
2
3
Theoretical Quantiles
lm(wt ~ wingcrd)

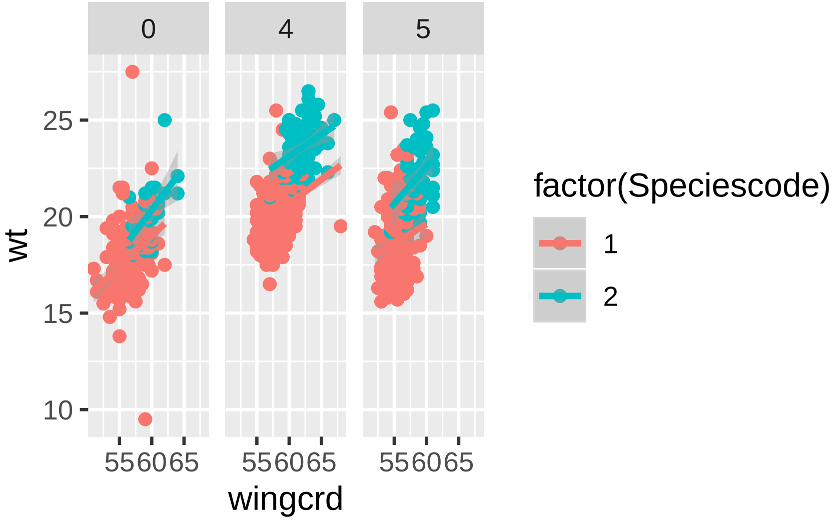

Scale−Location
√|Standardized residuals|
0.0
2.0
1
1281
1271
16
18
20
22
24
Fitted values
lm(wt ~ wingcrd)


Residuals vs Leverage
Standardized residuals
5
−5
1271
524
1
Cook's distance
0.5
0.000
0.005
0.010
0.015
Leverage
lm(wt ~ wingcrd)

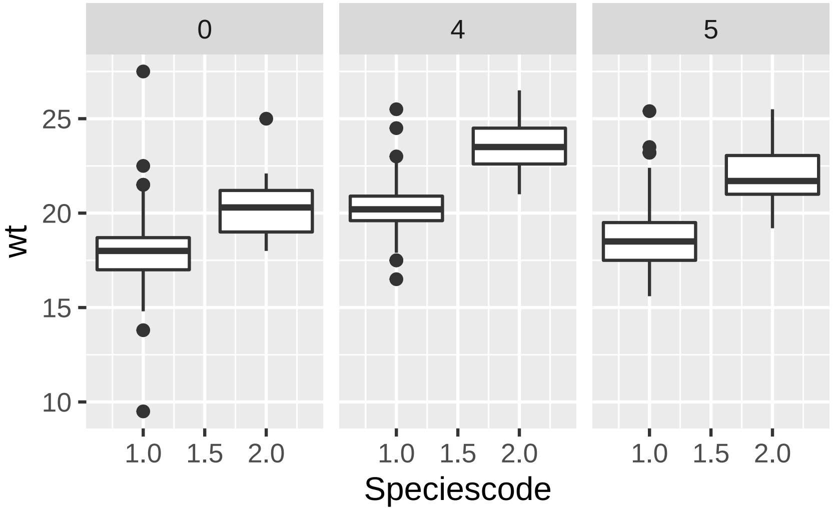
*Calculate average response for each Species.*

Now let's look at the average response by spevies and sex.

```
ggplot(data=sparrows, aes(wingcrd, wt,
                          colour = factor(Speciescode))) +
  geom_point() +
  geom_smooth(method = "lm") +
  facet_wrap(~Sex)
```



```
ggplot(data=sparrows, aes(Speciescode, wt,
                          group = Speciescode)) +
  geom_boxplot() +
  facet_wrap(~Sex)
```



To calculate species averages, we will need to (1) `group_by` species, and then (2) summarize by `mean`. I've been doing this for awhile, so

I know that mean won't like the NAs (missing values) in the data. So we will have to tell mean that it can remove NAs if it needs to.

```
species_data <- sparrows %>%
  group_by(Speciescode) %>%
  summarise(wingcrd = mean(wingcrd, na.rm=TRUE),
            wt = mean(wt, na.rm=TRUE)) %>%
  ungroup()

species_data
```

| Speciescode | wingcrd | wt |
|---:|---|---:|
| 1 | 57.32921 | 19.43989 |
| 2 | 60.62027 | 22.54324 |

Now one independent variable to model a dependent variable is unlikely to be ecologically realistic. What if we want to have more than independent variable? Maybe we want to include species as a predictor of mass, in addition to the predictor wing cord length.

Don't be a dummy, use dummy variables for categorical (group) predictor variables!

Remember that sex is a categorical variable so we have to use dummy variables. We could set up our own but R will do it for us if the data are stored as character or factor. Unfortunately, in this case Speciescode is stored as an integer. This means that R would interpret species 1 to be twice as large as species 2 and species 3 would be three times as larger. Therefore, we should use the mutate function in dplyr to convert it or create a new variable that is a character string.

```
sparrows <- sparrows %>%
  mutate(species = as.character(Speciescode))
str(sparrows)
```

```
## 'data.frame':    1281 obs. of  13 variables:
##  $ Speciescode: int  1 1 1 1 1 1 1 1 1 1 ...
##  $ Sex        : int  0 0 0 0 0 0 5 0 0 5 ...
##  $ SexNew     : int  1 1 1 1 1 1 6 1 1 6 ...
##  $ wingcrd    : num  59 55 53.5 55 52.5 57.5 53 55 55 55.5 ...
##  $ flatwing   : num  60 56 54.5 56 53.5 59 54 56 56 56.5 ...
##  $ tarsus     : num  22.3 19.7 20.8 20.3 20.8 21.5 20.6 21.5 20.8 20.5 ...
##  $ head       : num  31.2 30.4 30.6 30.3 30.3 30.8 32.5 31.2 31.6 31.4 ...
##  $ culmen     : num  12.3 12.1 12.8 11.9 12.6 12 13.5 12.3 13.2 13.2 ...
##  $ nalospi    : num  13 8.3 8.6 8.7 8.8 8.1 10.7 8.7 9.1 10.5 ...
##  $ wt         : num  9.5 13.8 14.8 15.2 15.5 15.6 15.6 15.7 15.7 15.7 ...
##  $ observer   : int  2 8 7 3 3 2 3 5 2 6 ...
##  $ Age        : int  2 2 2 1 1 2 1 1 1 1 ...
##  $ species    : chr  "1" "1" "1" "1" ...
```

Now you can see we have a new column called `species` that stores the different species as characters (`chr`). Now we are ready to run our regression.

```
lm2 <- lm(wt ~ wingcrd + species, sparrows)
summary(lm2)
```

```
##
## Call:
## lm(formula = wt ~ wingcrd + species, data = sparrows)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -10.6408  -0.8602   0.0592   0.7982   8.1982
##
## Coefficients:
##             Estimate Std. Error t value
## (Intercept) -4.60888    1.12102  -4.111
## wingcrd      0.41949    0.01954  21.468
## species2     1.72280    0.14014  12.293
##             Pr(>|t|)
## (Intercept) 4.19e-05 ***
## wingcrd      < 2e-16 ***
## species2     < 2e-16 ***
## ---
## Signif. codes:
##   0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1  ' ' 1
##
## Residual standard error: 1.425 on 1278 degrees of freedom
## Multiple R-squared:  0.4585, Adjusted R-squared:  0.4576
## F-statistic:   541 on 2 and 1278 DF,  p-value: < 2.2e-16
```

You can see that species has a significant affect on mass. Species two is generally 1.7 g heavier than species one and a bird is expected to get 0.42 g heavier for each 1 cm increase in wing cord.

## *Export Data*

Now imagine you want to save some data or results you worked up in R. To save a set of R objects for later use in R, you can use the save function.

```
if(!dir.exists("Output")) dir.create("Output")
```

```
save(lm2, species_data, file = "Output/lab2_tutorial.RData")
```

For the above code to work, you must already have a folder called "Output" to save the results to. Otherwise you may get an error like: `Error in gzfile(file, "wb") : cannot open the connection`.

If you have a table you want to be able to open in a spreadsheet like MS Excel you can use the `write.csv` function.

```
write.csv(species_data, file = "Output/species_data.csv",
          row.names = FALSE)
```

### *More Linear Regression Background and Details*

#### *Linear model example*

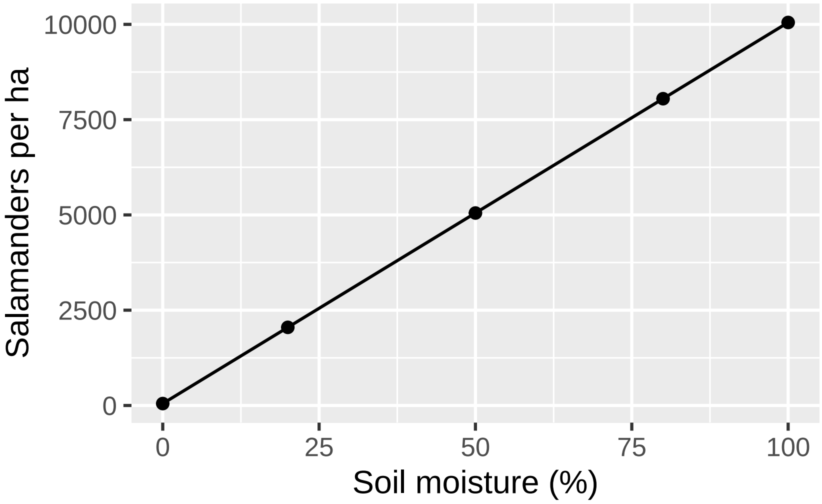
$$y_i = \beta_0 + \beta_1 soil_i + \epsilon_i$$

$$\epsilon_i \sim \mathcal{N}(\mu, \sigma)$$

The data, $y_i$, don't have to be normally distributed but the error (residuals) do.

Let's see how this works and the difference between a mathematical model and a statistical model.

Make a plot of soil moisture (x-axis) vs. salamander density per hectare ($y_i$ on the y-axis). Use soil values of 0, 20, 50, 80, and 100. Assume that the intercept, $\beta_0$, equals 50 and the slope (effect of soil moisture on salamander density, $\beta_1$) is 100.



How many salamanders per hectare do you expect when the soil moisture is 0? What about 50? 100?

*Do you think that if you went out that every hectare you measured there would be the exact number of salamanders predicted?*

Now assume that there is variation (error) associated with that relationship and that in the real world even the actual density of salamanders will vary around that line.

Before we can decide how much variation, we need to review the normal distribution.

*Normal (Gaussian) distribution*

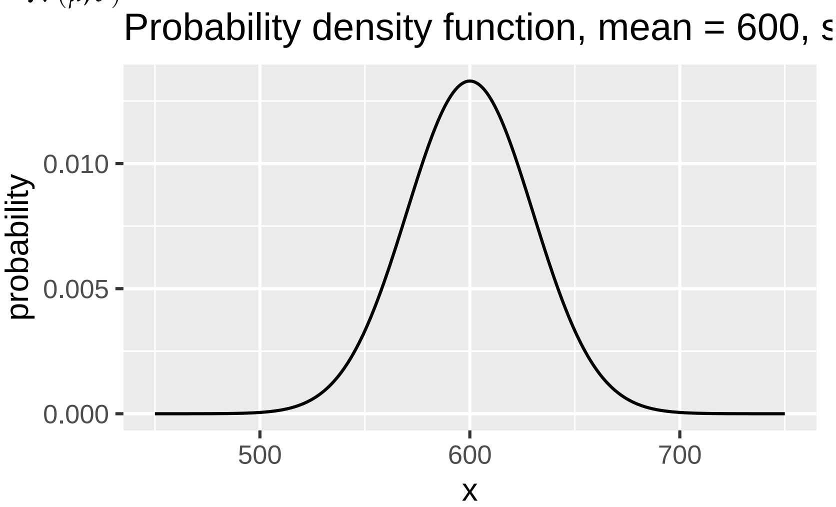
$\epsilon_i \sim \mathcal{N}(\mu, \sigma)$
$\sim$ : distributed as
$\mu$ : mean
$\sigma$ : standard deviation
$\sigma^2$ : variance
Can see normal distribution expressed in terms of the variance or standard deviation

*Probability density function*

$$F(x|\mu,\sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2 \big/ 2\sigma^2}$$

*Shorthand*

$\epsilon_i \sim \mathcal{N}(\mu, \sigma)$



*Now lets add some variation*

*Putting it together - adding noise*

$y_i = \beta_0 + \beta_1 X + \epsilon_i$
Deterministic:

- Intercept: $\beta_0 = 1$
- Slope: $\beta_1 = 2$
- Sample from -10 to 10 m at 1 m intervals $[X = (-10, -9, ...9, 10)]$

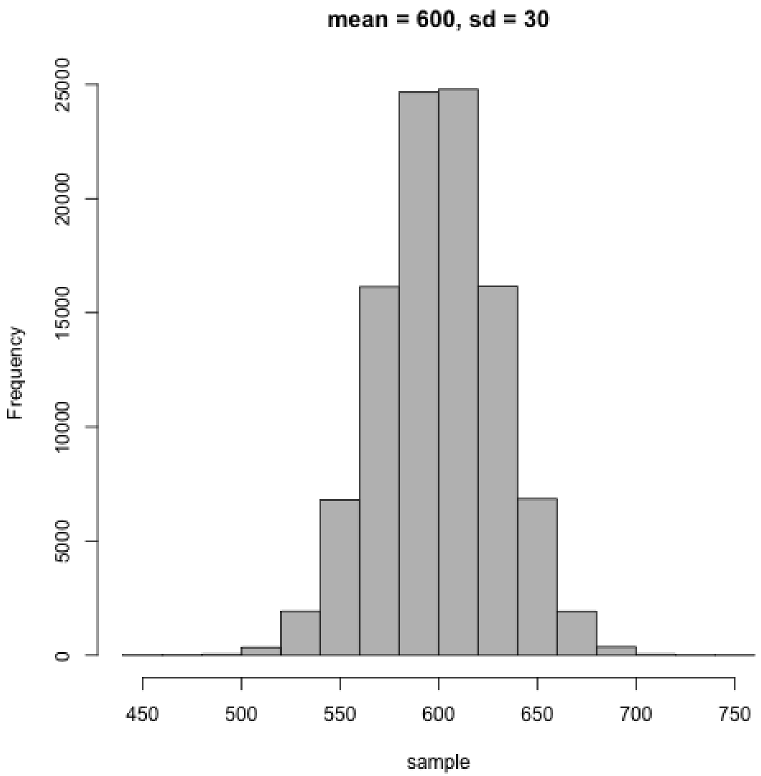


Figure 1: Normal Distribution Histogram

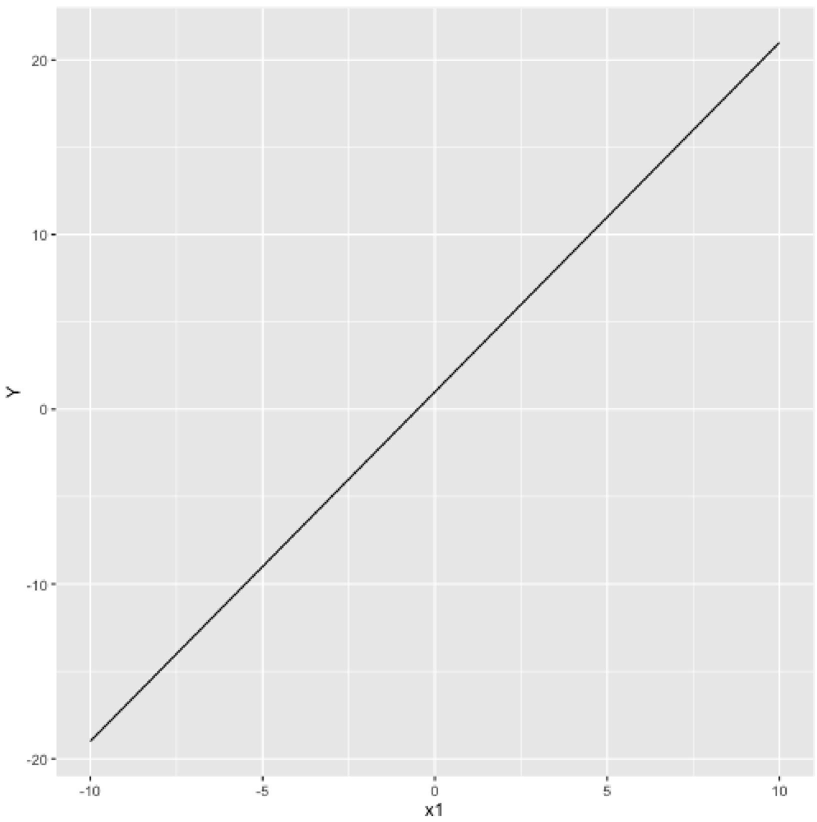


Figure 2: Mathematical Model

$y_i = \beta_0 + \beta_1 X + \epsilon_i$
Stochastic:

- $\epsilon_i \sim \mathcal{N}(\mu, \sigma)$
- $\mu = 0$
- $\sigma = 4$

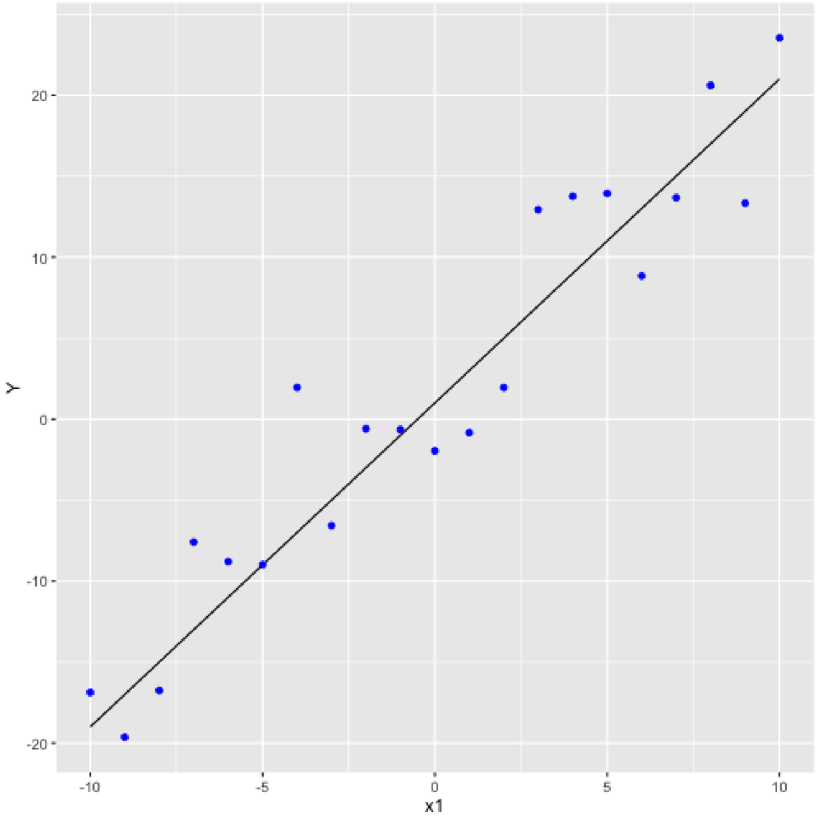


Figure 3: Statistical Model

*Adding covariates (more independent variables)*

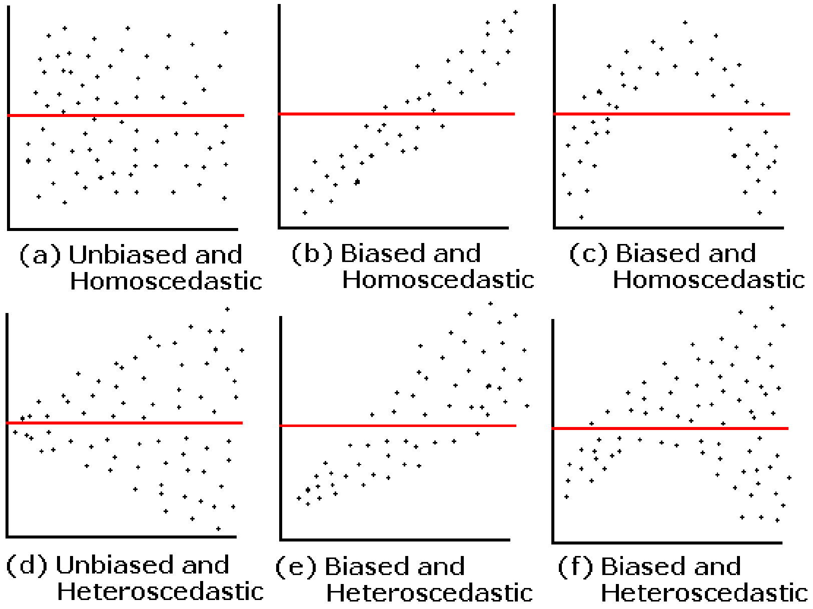
$y_i = \beta_0 + \beta_1 soil_i + \beta_2 litter_i + \epsilon_i$
$\epsilon_i \sim \mathcal{N}(\mu, \sigma)$
Plug in values to try it.

*Model assumptions*

**Know the assumptions of a linear model for the exam!!!**

- Linearity
- Homogeneity of variance (Homoscedasticity)
- Normally distributed error
- Minimal multicollinearity (if multiple $X$)
- Independence of observations (no autocorrelation)

### *Visual Checks*



(a) Unbiased and Homoscedastic

(b) Biased and Homoscedastic

(c) Biased and Homoscedastic

(d) Unbiased and Heteroscedastic

(e) Biased and Heteroscedastic

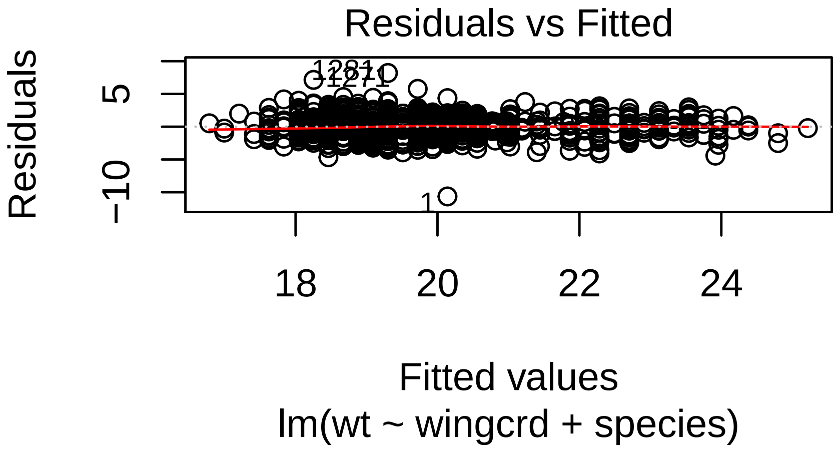
(f) Biased and Heteroscedastic

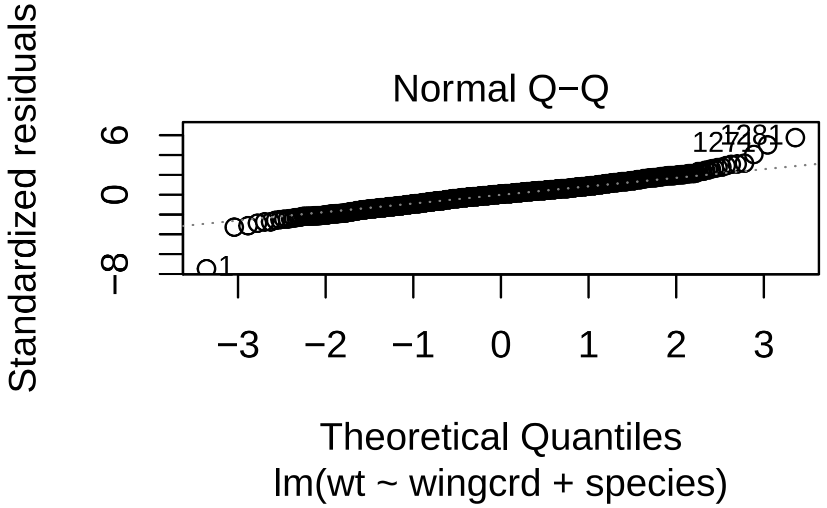
Figure 4: Error distributions

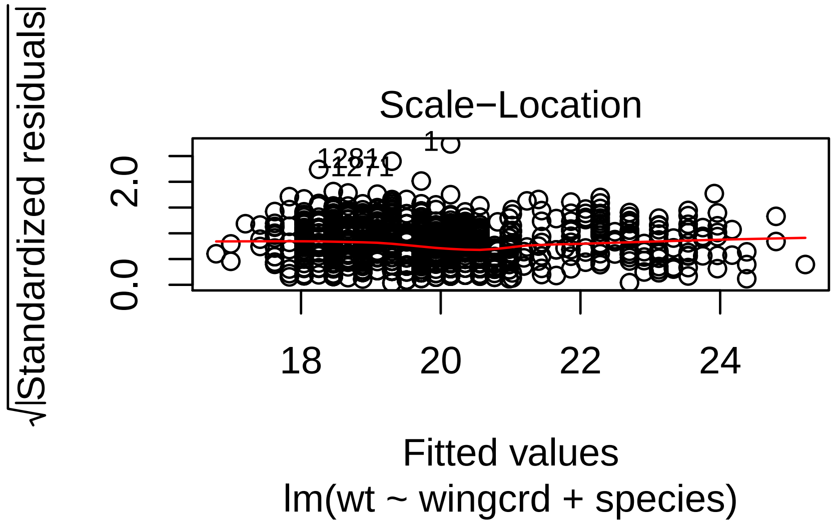
### *Checking Model Assumptions in R*

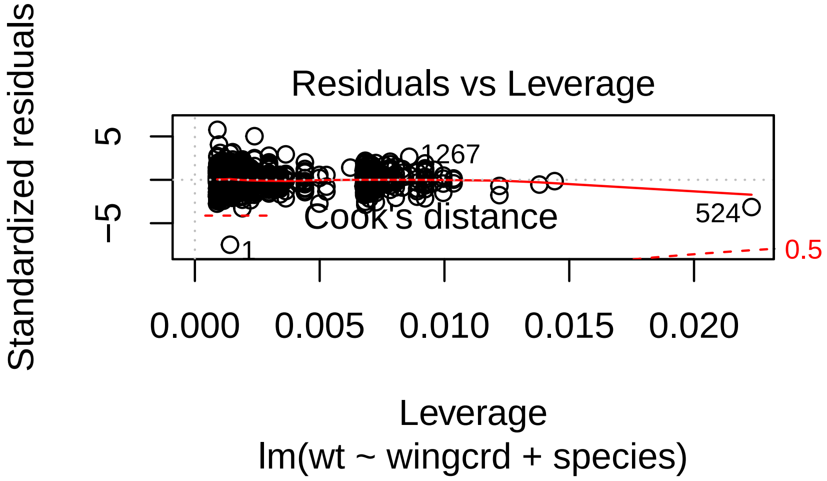
Now if you remember, we ran a regression of wing cord and sex on mass (`wt`) for the sparrows. Let's check if the data and model matched the assumptions for a linear regression. The key are the residual plots. R has the handy feature of having special plotting options for many special objects such as those from regression models. Just type `plot()` and insert your model results object.

```
plot(lm2)
```









This is great, the fitted vs. residuals plot doesn't show any significant heterogeneity or bias or other patterning. Of less importance but still good to look at, the qq-plot doesn't show major deviations from normality. I would be very happy to use these data with this model. I trust the results of this model and would draw biological inference from it.

## *Commands learned in this lab*

Base R

- `$`
- `mean( , na.rm=TRUE)`
- `lm()`
- `save()`
- `write.csv()`
- `as.character()`

tidyverse (dplyr)

- `%>%`
- `group_by()`
- `summarise()`
- `mean()`

ggplot

- `geom_smooth(, method='lm')`

## *Rprojects and Working Directories*