# Simple Linear Models

## Lecture 04.1: General Linear Models

Lauren Sullivan

Module: Linear, Non-linear, and Mixed Effects Models

# Readings

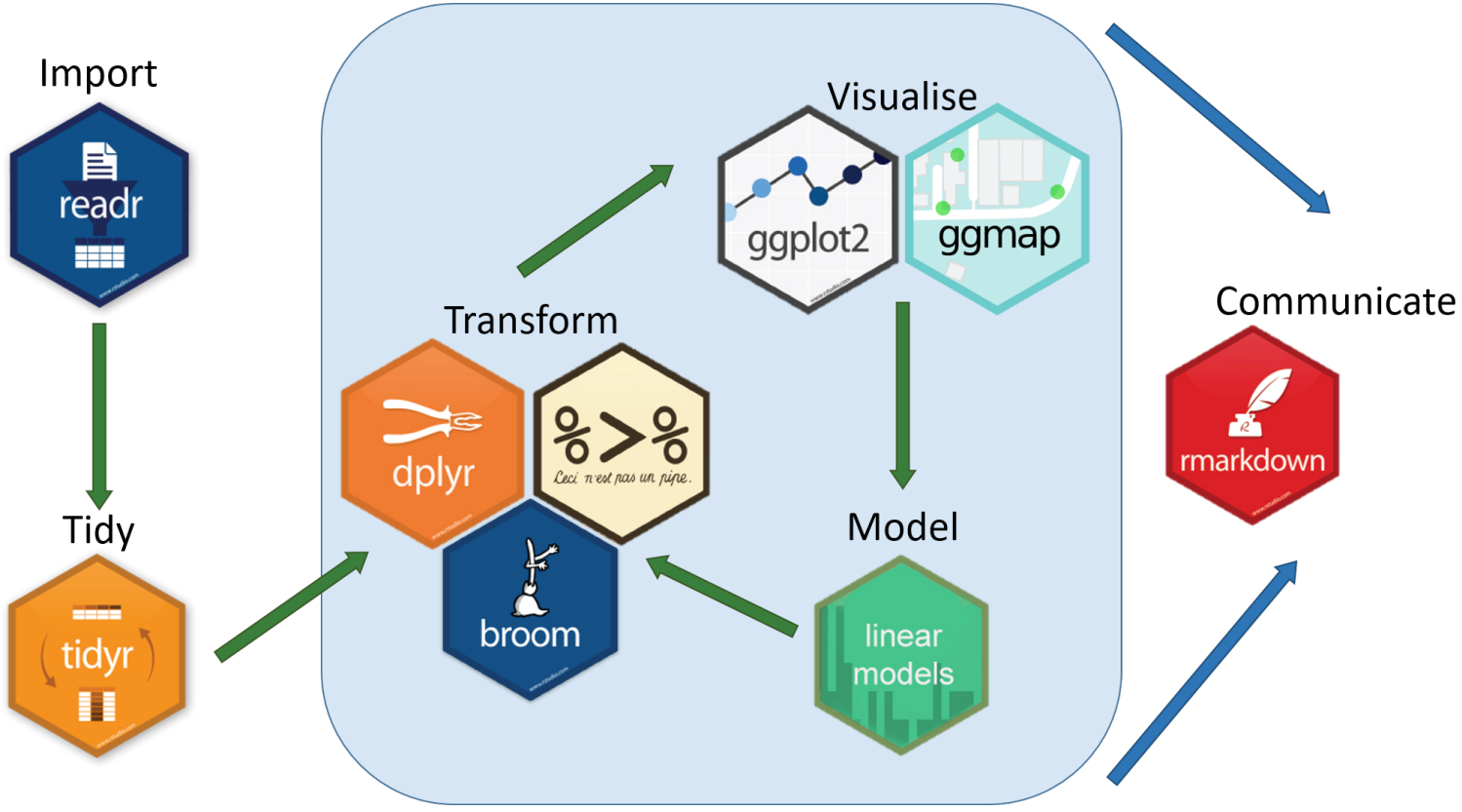
**Required for class:**

- NA

**Optional:**

- Crawley, M. *Statistics: An Introduction Using R*
- Bolker, B. *Ecological Models and Data in R - Ebook version*

# On to statistical analysis.



# Data

A dataset looking at salmon residence time in streams and how that varies with sex, age, year and preciptiation.

# Data

This dataset has multiple types of X variables (both categorical and continuous) and a single Y variable so we can look at different types of linear models.

```
## # A tibble: 68 x 5
##    ResidenceTime Sex   Age    Year Precip
##            <dbl> <chr> <chr> <dbl>  <dbl>
##  1           1   f     a      2001  16.1
##  2           2.5 m     a      2000   9.10
##  3           3   f     a      2001  21.6
##  4           3   m     a      2001  17.9
##  5           3   f     a      2002  17.7
##  6           3   f     a      2002   8.10
##  7           3   f     j      2002   4.48
##  8           3   f     j      2002   9.74
##  9           3   f     j      2002   7.48
## 10           3   f     j      2002  17.7
## # ... with 58 more rows
```

# General Linear Models

*General Linear Models* refer to linear regression models that have a continuous dependent variable (Y) and a single or series of independent variables (X's) that can be either continuous or categorical. *They all assume normal distributions.* These models can have specific names depending on their type, but are all linear models.

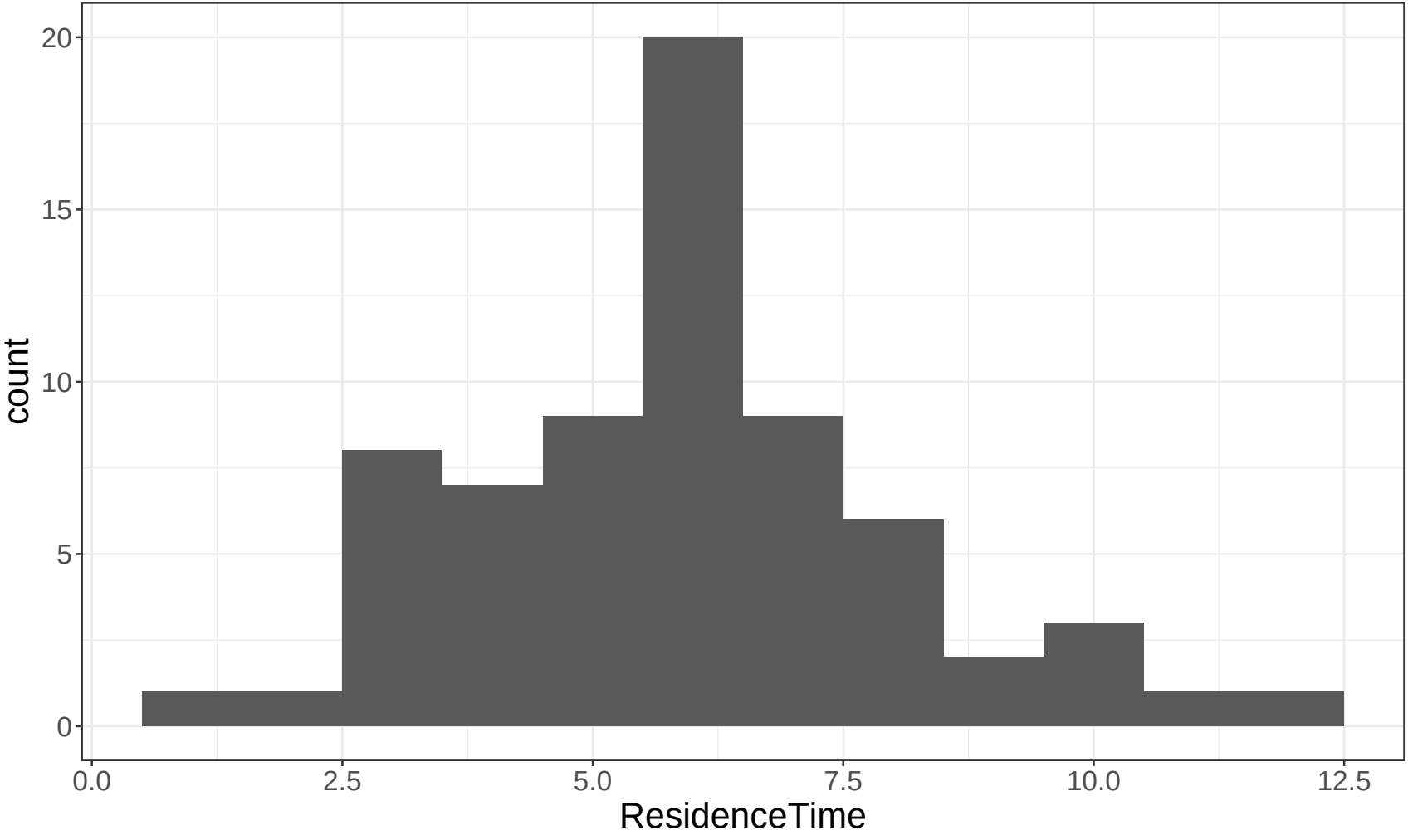
- **Regression** - continuous Y, 1 continuous X. `lm()`
- **ANOVA** - continuous Y, categorical X. `aov()`
- **Multiple Linear Regression** - continuous Y, multiple continuous X. `lm()`
- **ANCOVA** - continuous Y, at least 1 continuous X and at least 1 categorical X. `lm()`

# General Linear Models - Assumptions

1. Relationships are (all) linear
   - *For regressions only*
2. (Multivariate) Normal distributions of error variance $\epsilon$
3. Equal variance (aka - Homoscedasticity)
   - ANOVA's are pretty robust to this
4. Independence of observed samples

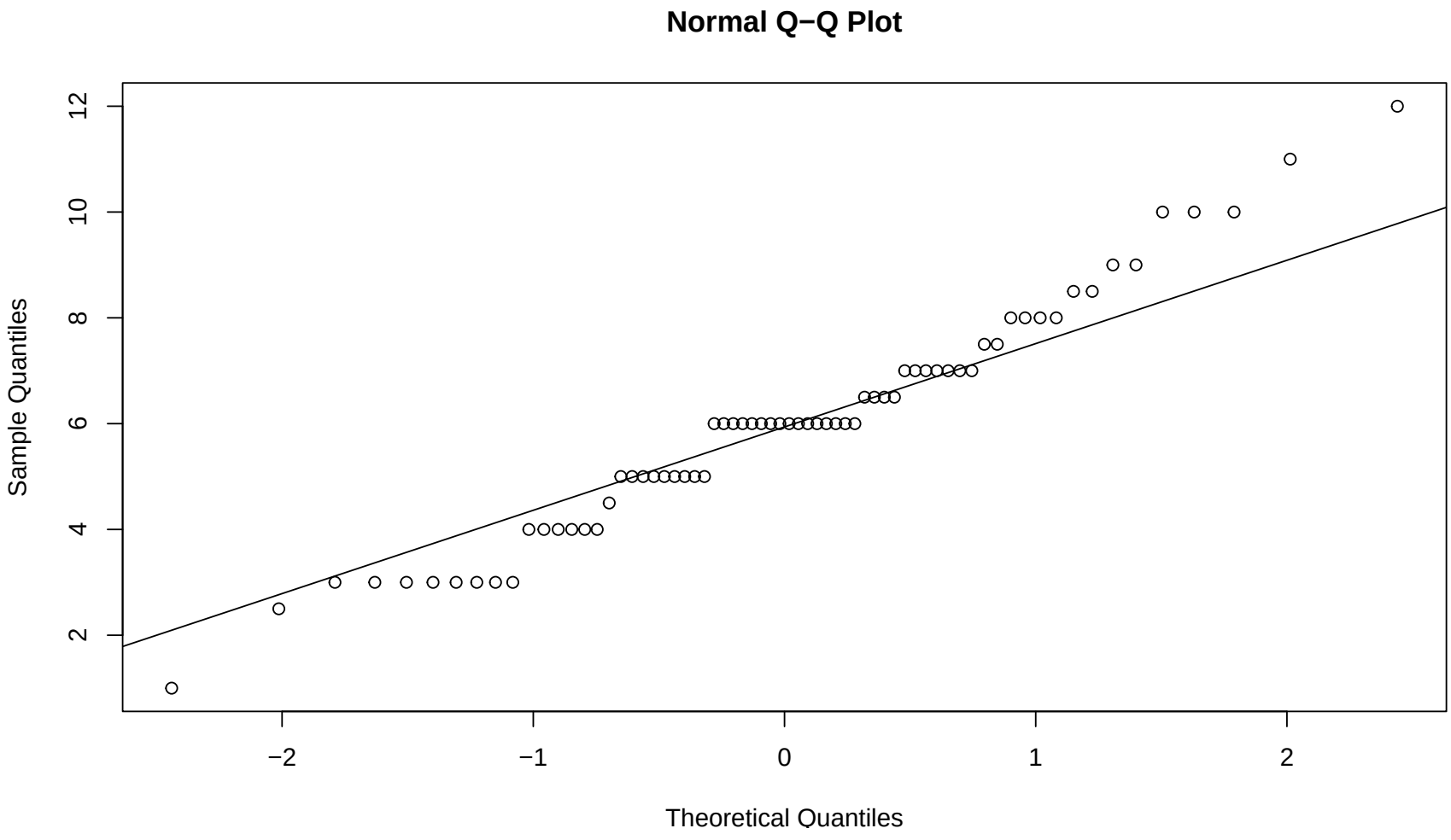
# Normality Assumption

How are the data distributed?



# Normality Assumption

Try a **normal Q-Q plot**



# Normality Assumption

Try a **Shapiro-Wilk Test**

- $H_0$: Data are not different from a normal distribution
- $H_a$: Data are different from a normal distribution

```
shapiro.test(salmon$ResidenceTime)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  salmon$ResidenceTime
## W = 0.97122, p-value = 0.1169
```

# Regression - Experimental Design

**Dependent variable (Y) is continuous, independent variable (X) is continuous.**

$Y_i = \beta_0 + \beta_i X_i + \epsilon$ ($\beta_0$ is the intercept, $\beta_i$ is the slope coefficient and $\epsilon$ is the error)

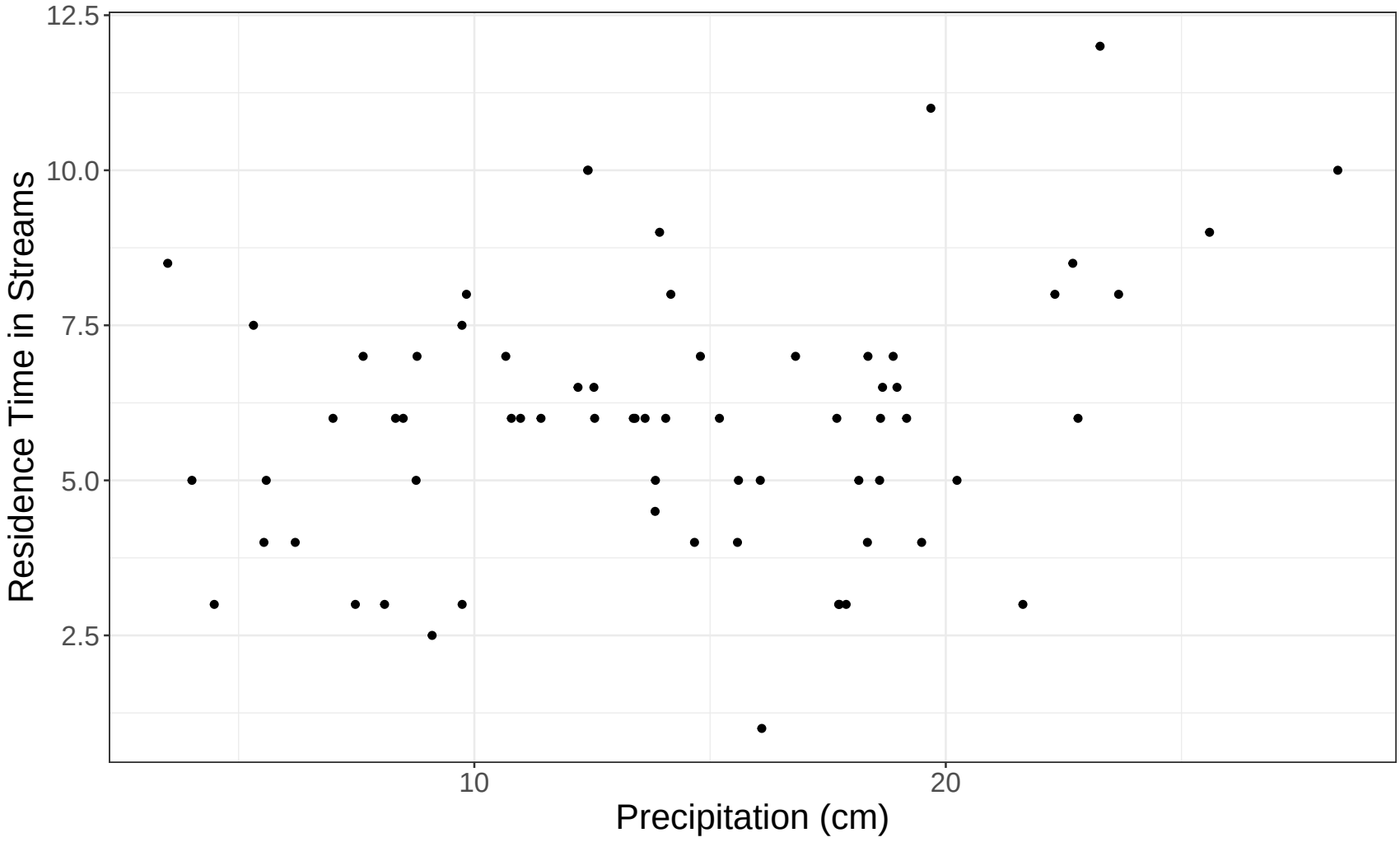
$H_0$: no relationship between X and Y

Some Questions...

- How does elevation (X) alter a plant's seed production (Y)?
- How does temperature (X) alter a lizard's metabolic rate (Y)?
- How does the year (X) influence the average global temperature (Y)?

# Regression

Does residence time within a stream depend on the amount of precipitation?



# Regression

Does residence time within a stream depend on the amount of precipitation?

```
test_reg <- lm(ResidenceTime ~ Precip, data = salmon)
summary(test_reg)
```

```
##
## Call:
## lm(formula = ResidenceTime ~ Precip, data = salmon)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.1676 -1.3786  0.0781  1.1288  5.1342
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  4.60153    0.69182   6.651 6.72e-09 ***
## Precip       0.09729    0.04522   2.152   0.0351 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.103 on 66 degrees of freedom
## Multiple R-squared:  0.06555,    Adjusted R-squared:  0.0514
## F-statistic:  4.63 on 1 and 66 DF,  p-value: 0.03508
```

# Regression

$$Y_i = \beta_0 + \beta_i X_i + \epsilon$$

$$\mathrm{Y} = 4.6015 + 0.0973\mathrm{X} + \epsilon$$

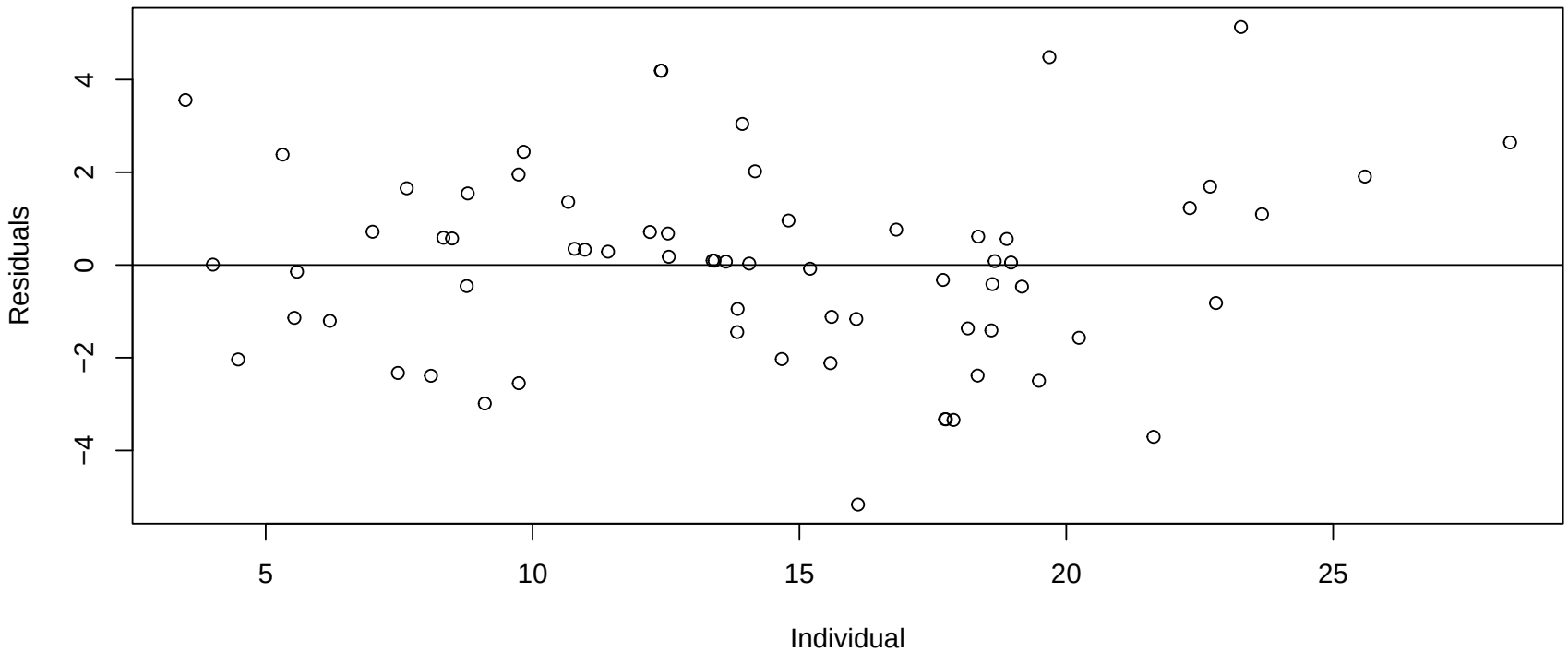
```
##
## Call:
## lm(formula = ResidenceTime ~ Precip, data = salmon)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.1676 -1.3786  0.0781  1.1288  5.1342
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  4.60153    0.69182   6.651 6.72e-09 ***
## Precip       0.09729    0.04522   2.152   0.0351 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.103 on 66 degrees of freedom
## Multiple R-squared:  0.06555,    Adjusted R-squared:  0.0514
## F-statistic:  4.63 on 1 and 66 DF,  p-value: 0.03508
```

# Residuals

Residual plots show you the difference between your observed data ($y$), and the expected, or fitted value ($\hat{y}$).

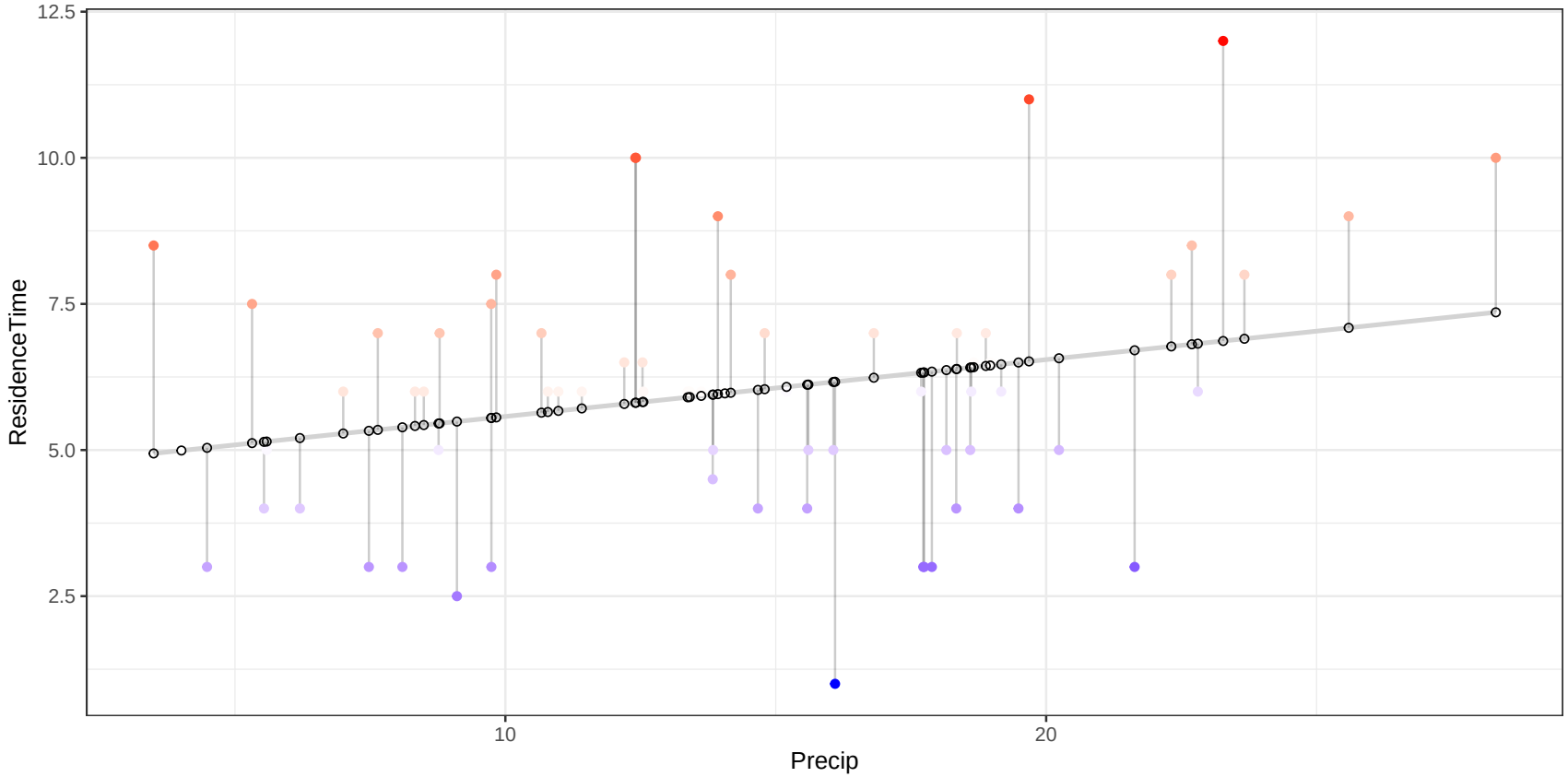
$$Residual = y - \hat{y}$$

```
test_reg_resid <- resid(test_reg)
```



# Residuals

$$Residual = y - \hat{y}$$



# ANOVA - Experimental Design

**Dependent variable (Y) is continuous, independent variable (X) is categorical.**

$Y_{ij} = \mu + \alpha_i + \epsilon_{ij}$ ($\mu$ is the grand mean, $\alpha_i$ is the $i^{th}$ group mean and $\epsilon_{ij}$ is the error)

$H_0$: no difference among groups

Some Questions...

- How does a diet treatment (X) alter an animal's growth rate (Y)?
- How do nutrient additions (X) alter plant species diversity (Y)?
- How does sex of an organism (X) alter it's feeding behavior (Y)?
- How does plant family (X) alter a plant's SLA (Y)?

# ANOVA vs t-test

For 2 groups of equal size, ANOVA's and t-tests give you the same result.

```
x <- c("m", "m", "m", "m", "m", "f", "f", "f", "f", "f")
y <- c(5,4,4,3,3,7,5,7,6,6)

t.test(y ~ x)
```

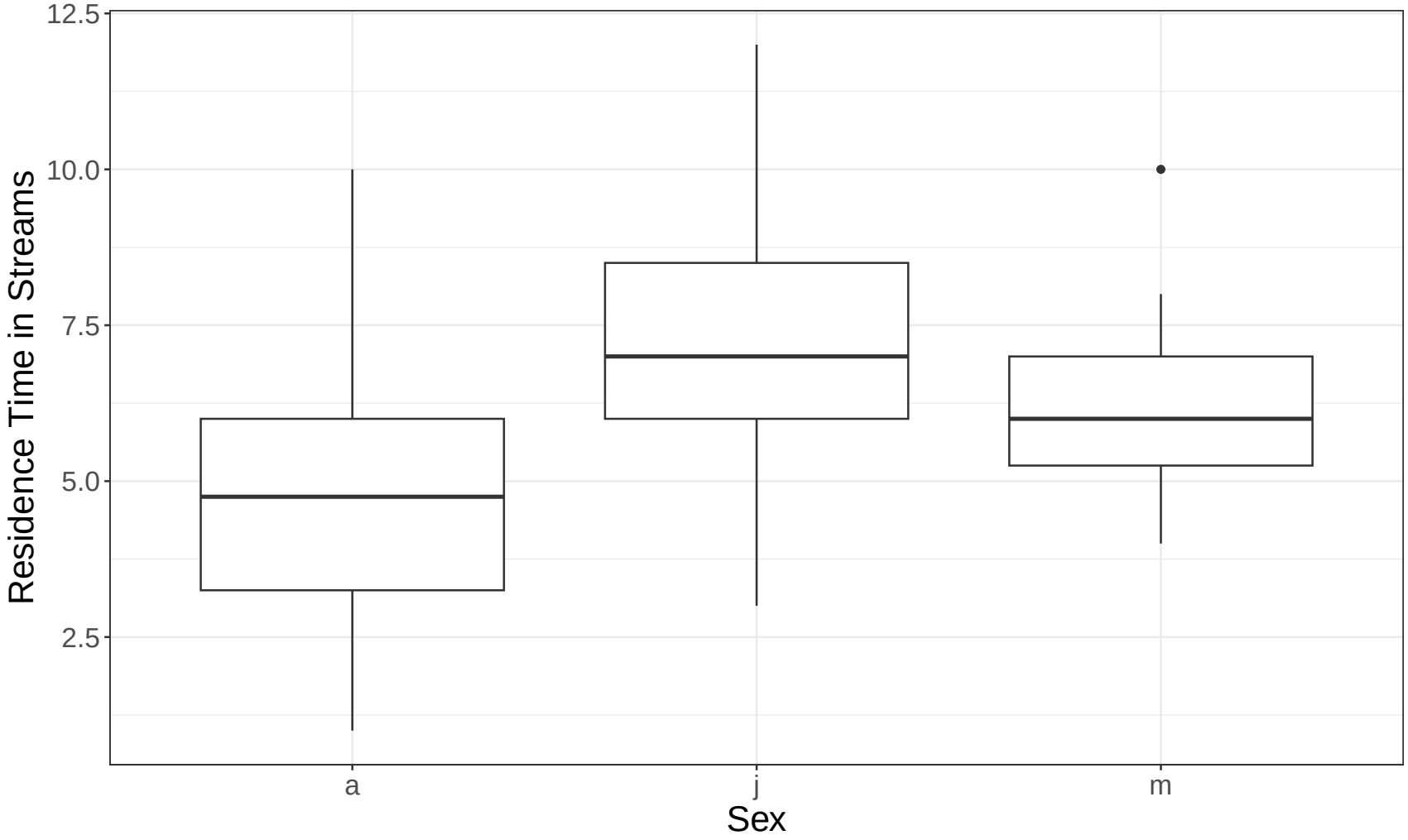
```
##
##  Welch Two Sample t-test
##
## data:  y by x
## t = 4.5356, df = 8, p-value = 0.00191
## alternative hypothesis: true difference in means between group f and group m is not equal to 0
## 95 percent confidence interval:
##  1.179777 3.620223
## sample estimates:
## mean in group f mean in group m
##             6.2             3.8
```

```
anova(lm(y ~ x))
```

```
## Analysis of Variance Table
##
## Response: y
##           Df Sum Sq Mean Sq F value  Pr(>F)
## x          1   14.4    14.4  20.571 0.00191 **
## Residuals  8    5.6     0.7
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

# ANOVA

Is residence time within a stream a function of the age of the fish?



# ANOVA

Is residence time within a stream a function of the age of the fish?

```
test_aov <- aov(ResidenceTime ~ Age, data = salmon)
summary(test_aov)
```

```
##              Df Sum Sq Mean Sq F value   Pr(>F)
## Age           2  61.39  30.695   7.946 0.000818 ***
## Residuals    65 251.09   3.863
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

# ANOVA - Post Hoc Test

But this only tells you that age significantly predicts residence time. Which groups are different from each other?

```
TukeyHSD(test_aov)
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = ResidenceTime ~ Age, data = salmon)
##
## $Age
##           diff        lwr       upr     p adj
## j-a  2.2765152  0.8850487 3.6679816 0.0006159
## m-a  1.5454545  0.1240606 2.9668485 0.0299676
## m-j -0.7310606 -2.1225271 0.6604059 0.4226850
```

# Multiple Linear Regression - Experimental Design

**Dependent variable (Y) is continuous, multiple independent variables (X) that are all continuous.**

$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + ... + \epsilon$

- $\beta_i$ are **partial regression coefficients** - the effect of $X_i$ while holding all other $X$ constant

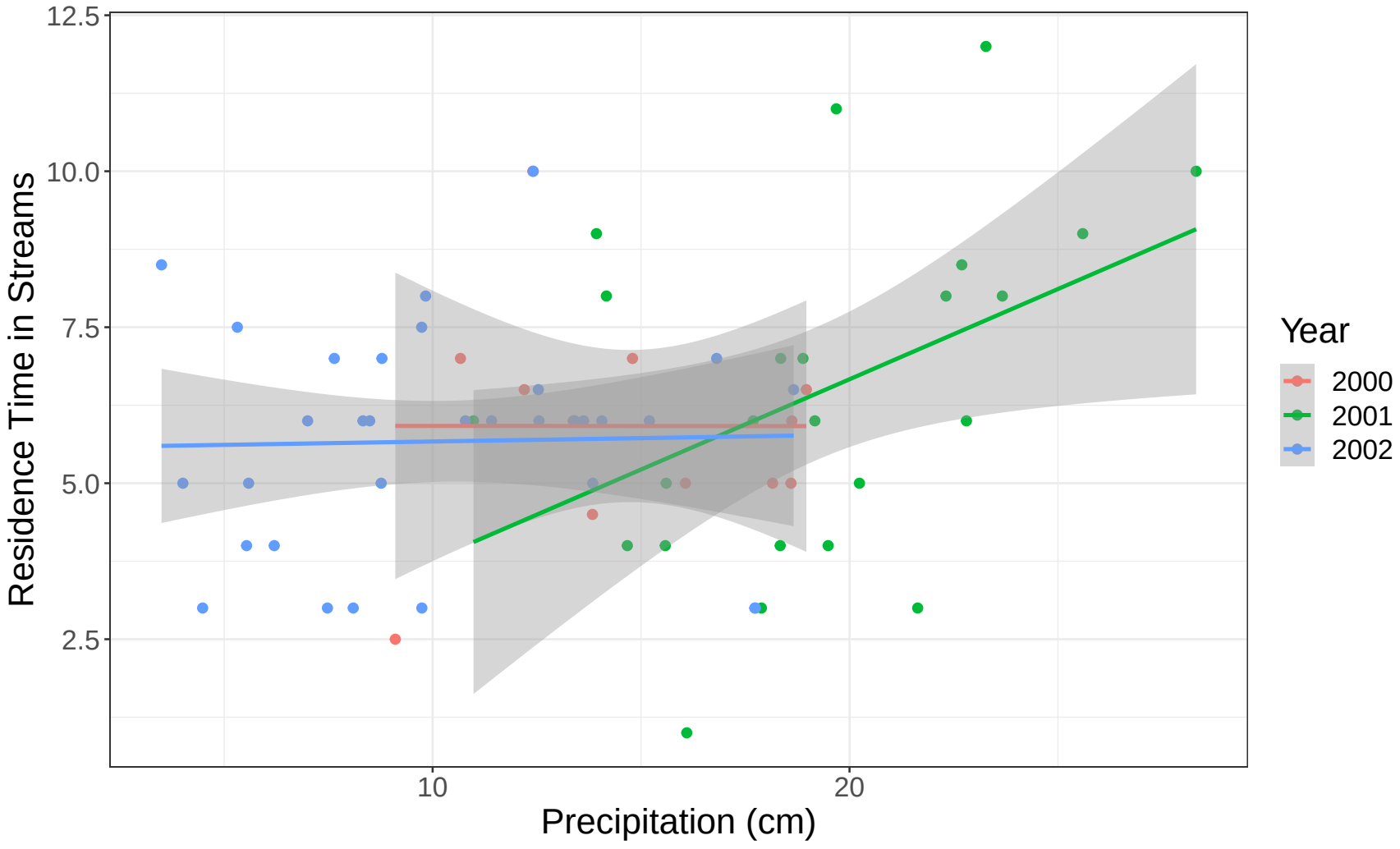
$H_0$: no difference among slopes

Some Questions...

- How does elevation (X1) and temperature (X2) alter a plant's seed production (Y)?
- How does temperature (X1) and humidity (X2) alter a lizard's metabolic rate (Y)?
- How does the year (X1) and atmospheric CO2 level (X2) influence the average global temperature (Y)?

# Multiple Linear Regression

Does residence time within a stream depend on the amount of precipitation and the year?



# Multiple Linear Regression

Does residence time within a stream depend on the amount of precipitation and the year?

```
test_mreg <- lm(ResidenceTime ~ Year + Precip, data = salmon)
summary(test_mreg)
```

```
##
## Call:
## lm(formula = ResidenceTime ~ Year + Precip, data = salmon)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.1459 -1.2522  0.0064  1.1341  5.1031
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -236.96285  773.21091  -0.306   0.7602
## Year           0.12065    0.38619   0.312   0.7557
## Precip         0.10465    0.05125   2.042   0.0452 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.118 on 65 degrees of freedom
## Multiple R-squared:  0.06695,    Adjusted R-squared:  0.03825
## F-statistic: 2.332 on 2 and 65 DF,  p-value: 0.1052
```

# ANCOVA - Experimental Design

**Dependent variable (Y) is continuous, multiple independent variables (X), where at least one is continuous and one is categorical.**

$$Y_{ij} = \mu + \alpha_i + \beta_{within}(X_{ij} - \overline{X_i}) + \epsilon_{ij}$$

$H_0$: no difference among slopes, no difference among groups.

- First compares slopes, then compares groups while holding effects of covariates constant.

Some Questions...

- How does elevation (X1) and nutrient addition (X2) alter a plant's seed production (Y)?
- How does temperature (X1) and the sex of an individual (X2) alter lizard's metabolic rate (Y)?
- How does the year (X1) and atmospheric CO2 level (X2) and habitat type (X3) influence plant biomass (Y)?

# ANCOVA

Does residence time within a stream depend on the amount of precipitation, the year, and the sex of the fish?

```
test_mreg <- lm(ResidenceTime ~ Precip + Year + Sex , data = salmon)
summary(test_mreg)
```

```
##
## Call:
## lm(formula = ResidenceTime ~ Precip + Year + Sex, data = salmon)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.2393 -0.8995 -0.1261  0.8873  4.0080
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -317.30460  589.84225  -0.538  0.59248
## Precip         0.13975    0.03942   3.545  0.00074 ***
## Year           0.15936    0.29460   0.541  0.59042
## Sexm           3.15523    0.45666   6.909 2.68e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.615 on 64 degrees of freedom
## Multiple R-squared:  0.4656, Adjusted R-squared:  0.4405
## F-statistic: 18.59 on 3 and 64 DF,  p-value: 8.78e-09
```

# ANCOVA

Does residence time within a stream depend on the amount of precipitation, the year, and the sex of the fish?