Linear Models and ANOVA

What we will discuss today

- Simple Linear Regression
- Multiple Linear Regression
- ANOVA

After the workshop, you will be able to

- Write code in R for regression and anova
- Interpret results and output from R or other statistical software
- Have a better understanding of how to work with data similar to examples we will discuss

Functions in R we will use today

- lm( )
- summary( )
- anova( )
- aov( )

Preliminaries

What is the goal of my analysis?

- ❖ To analyze the relationship between two variables
  - How does sea ice changes with year?
- ❖ To analyze the effect of several variables on an outcome
  - What are some factors related to blood viscosity?
- ❖ To determine if a treatment is effective
  - Do improvement scores differ by types of treatment?
- ❖ To determine whether two or more factors have an effect on an outcome
  - How do source of supplements and dose levels affect tooth growth?

What are the variables?

- ❖ By their purposes
  - Dependent Variable, or Response Variable (y)
  - Independent Variable, or Feature, or Predictor, or Covariate (x)
- ❖ By their nature
  - Numerical or Continuous Variable
    - ➢ When values of this variable are continuous and have meaning numerically
  - Categorical or Nominal Variable
    - ➢ When this variable is qualitative instead of quantitative
  - Ordinal Variable
    - ➢ When this variable is qualitative with an order

Before Starting Any Analysis

- Take a look at the dataset
  - Is it in the correct format for the purpose of the analysis?
    - Rearrange variables if necessary
  - Are there missing values?
    - Remove an observation with missing values
    - Imputation
  - Are there outliers?
    - Consider removing them
  - What are the demographics?
    - Will the results be generalizable?
    - Could sampling methods have been improved?

Ronald Aylmer Fisher

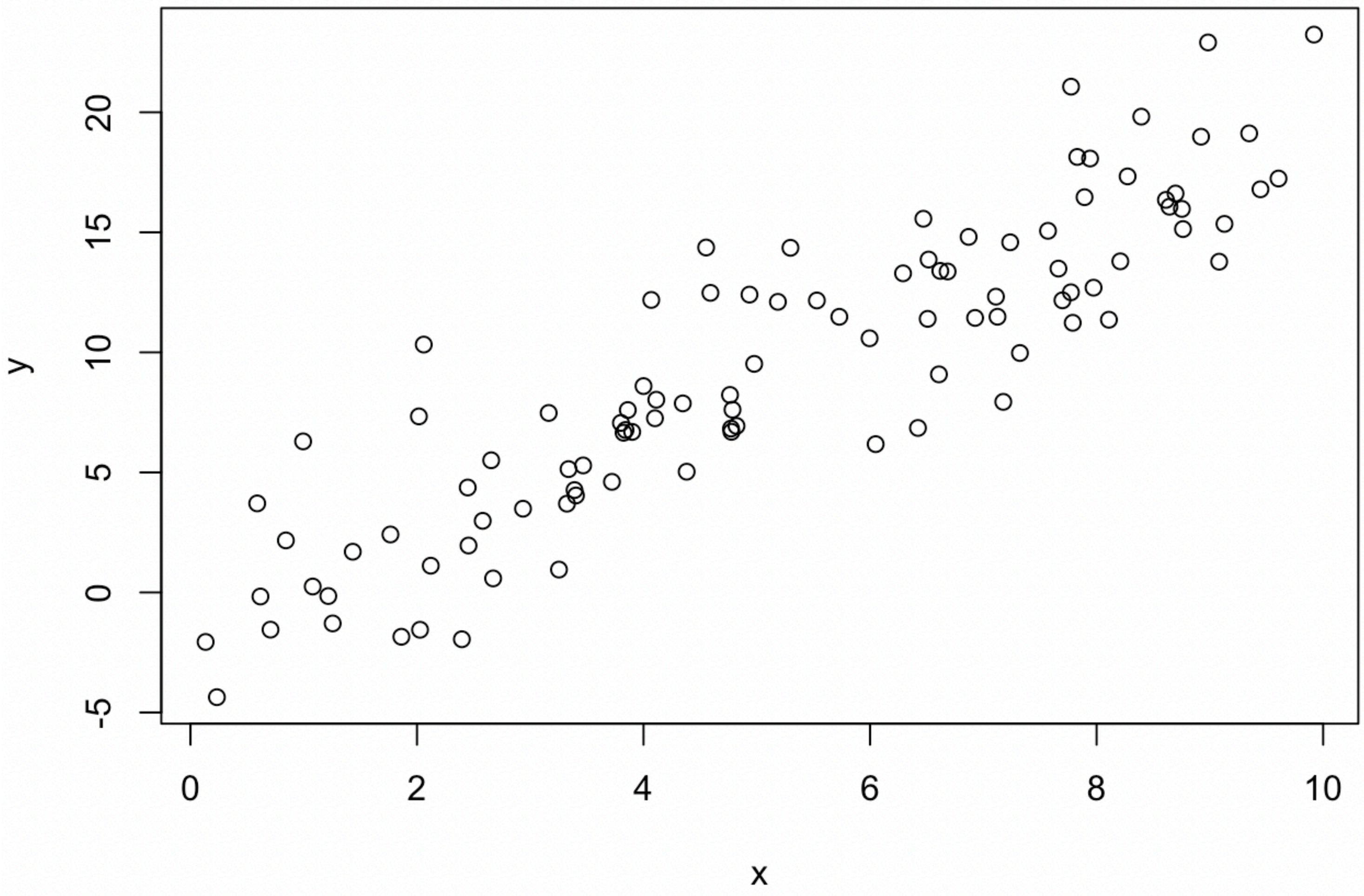
(17 February 1890 – 29 July 1962)

- Analysis of Variance (ANOVA)
- F-distribution

"To call in the statistician after the experiment is done may be no more than asking him to perform a post-mortem examination: he may be able to say what the experiment died of."

Simple Linear Regression

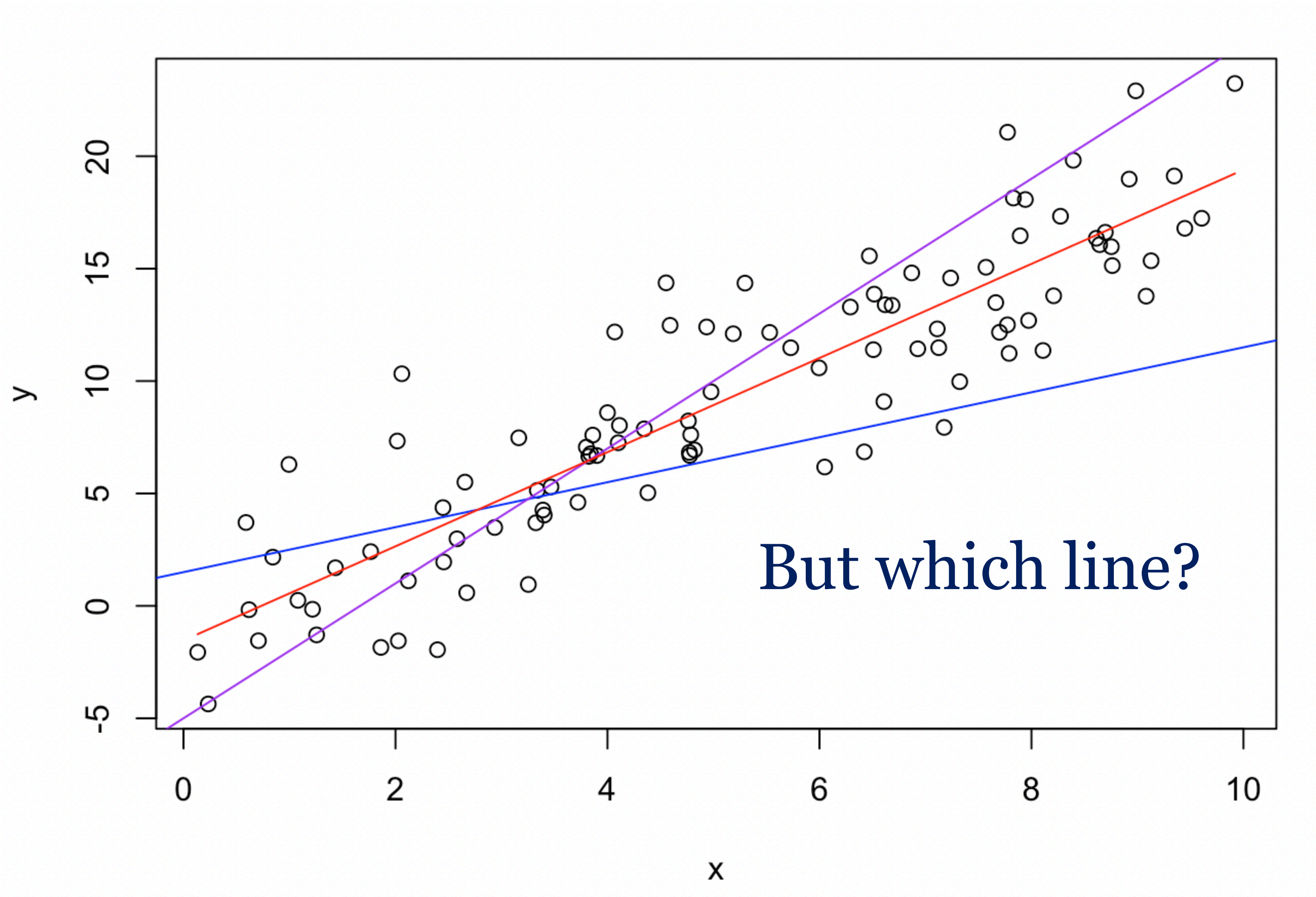
A Simulated Example



Intuitions

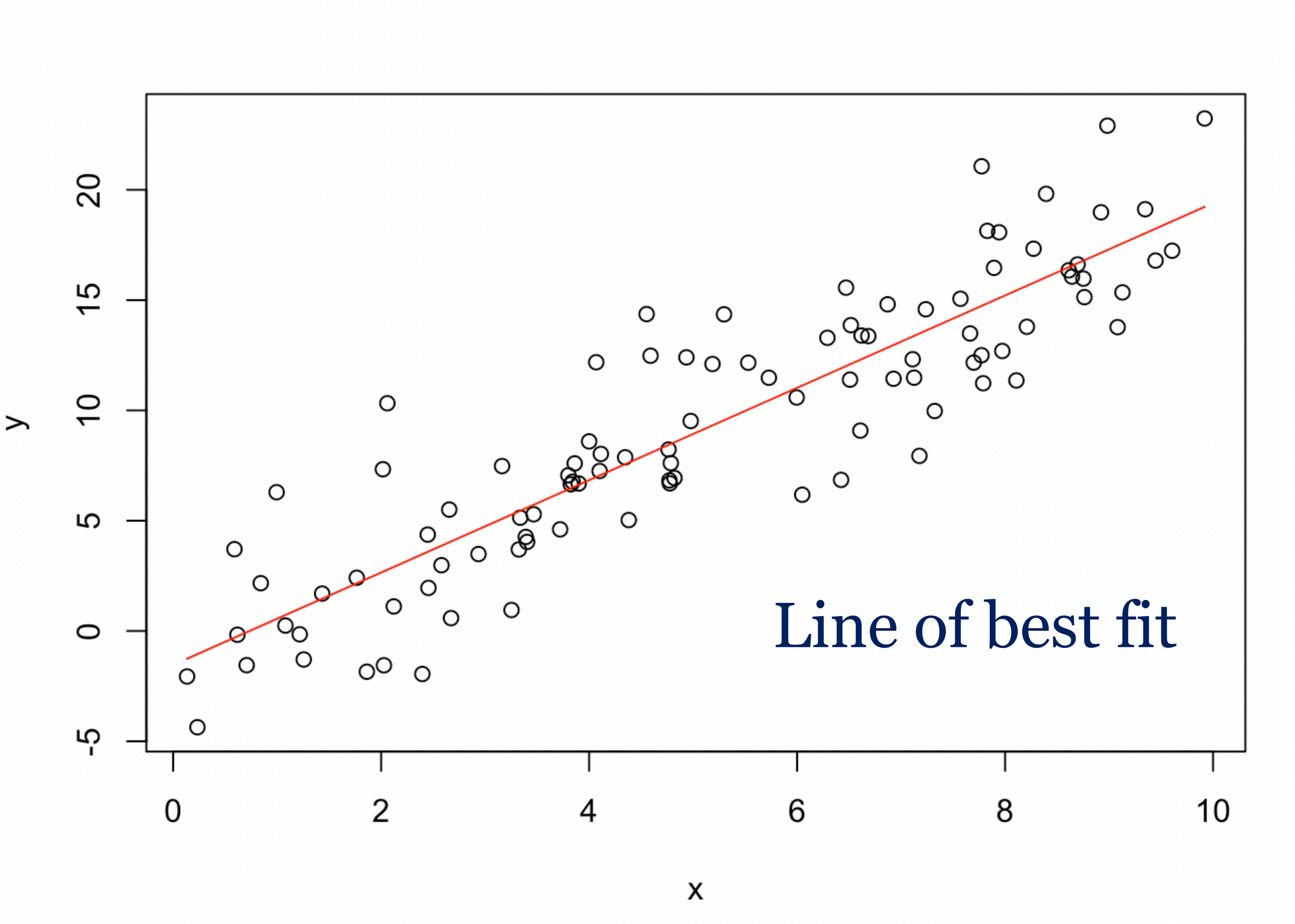
- There seems to be a strong linear trend
- The correlation between the two variables appears to be positive
- It would be plausible to model the data with a straight line

A Simulated Example



But which line?

A Simulated Example

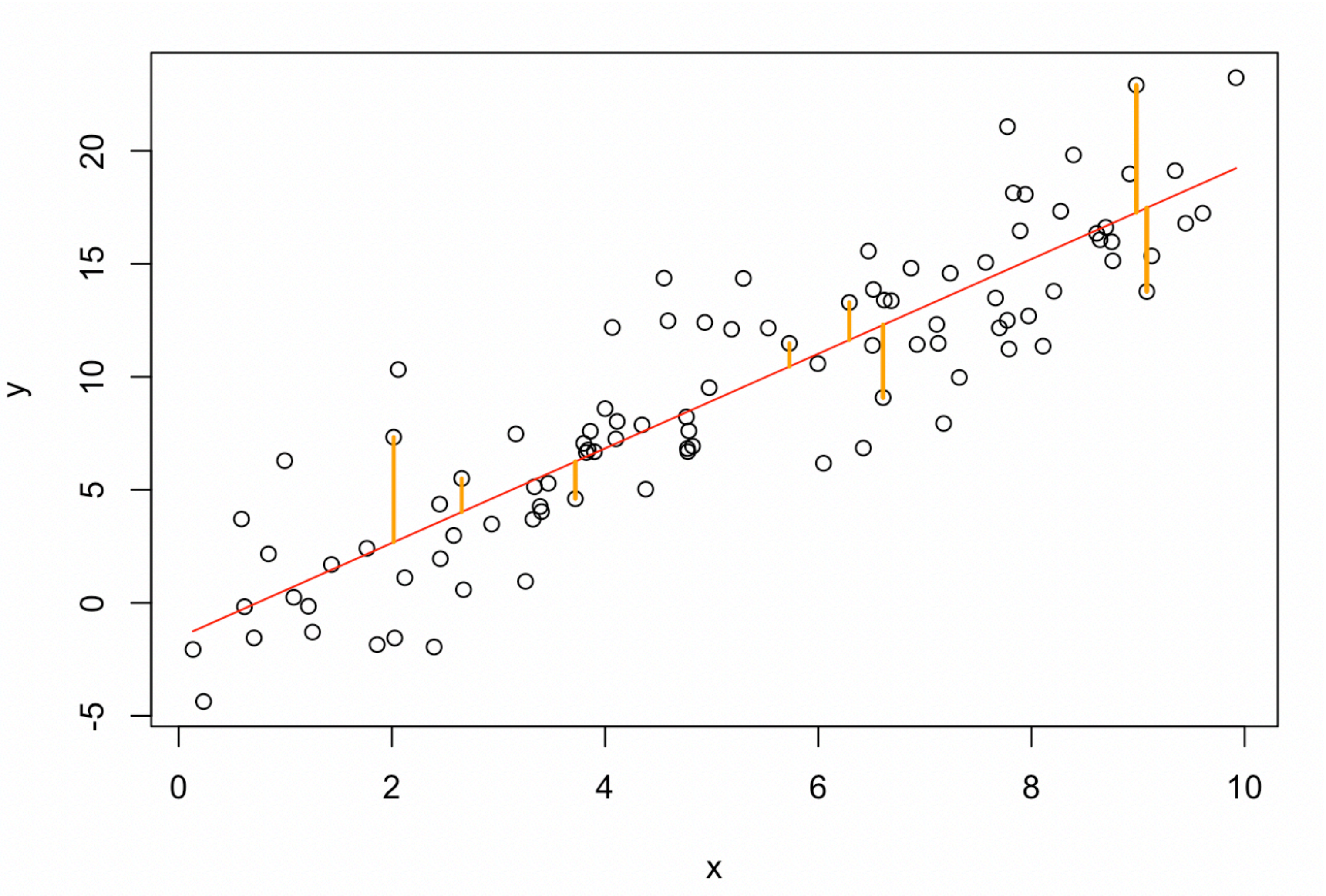


Line of best fit

Intuitions

- This line best captures the relationship between the two variables
- Data points seem to scatter evenly and randomly around the line
- How to quantify best fit?

A Simulated Example



McGill initiative in Computational Medicine

Formulating the Model

$$y = \alpha + \beta x + \epsilon$$

α: intercept parameter

β: slope parameter

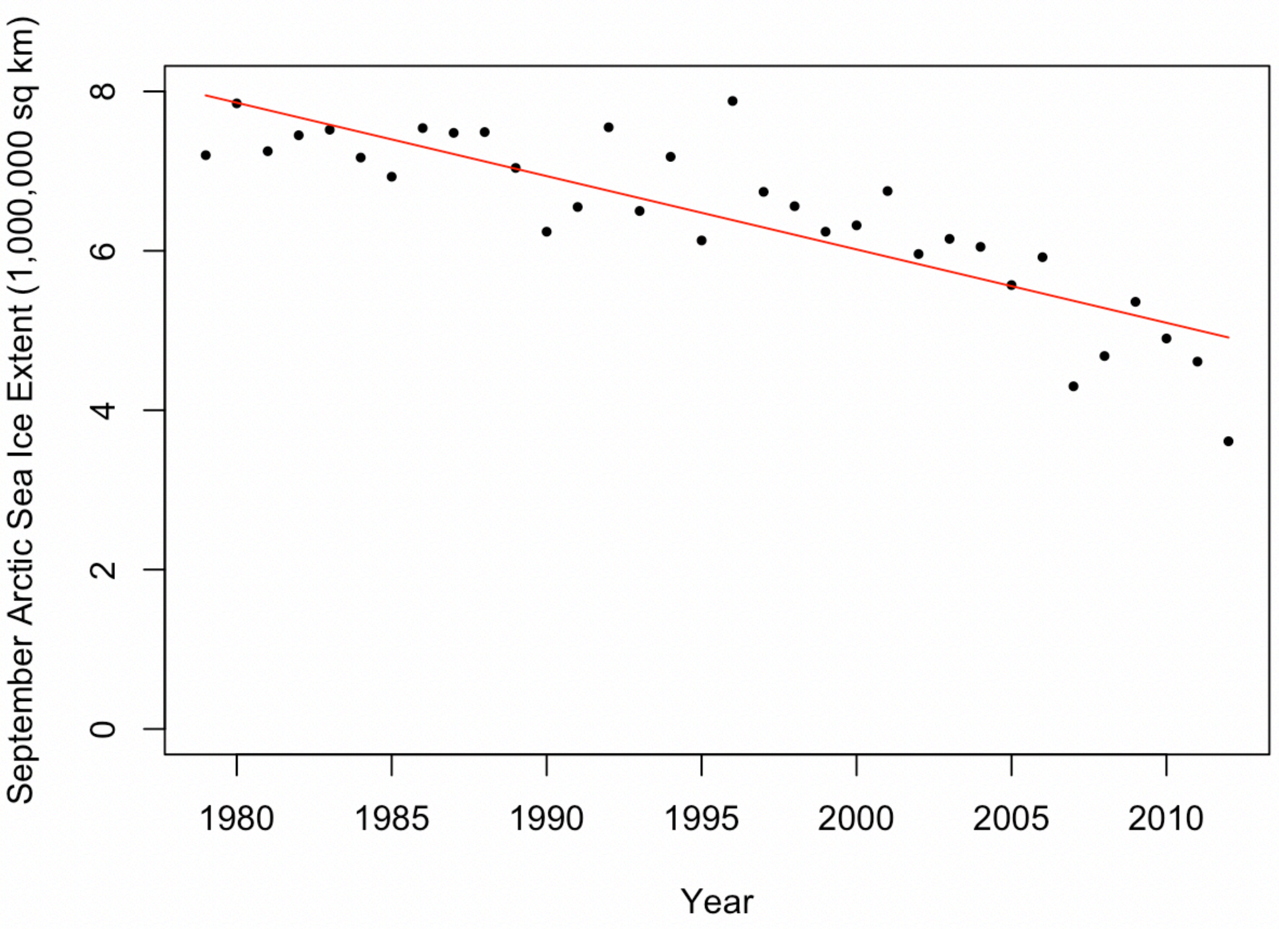
ϵ: error, independent random variable normally distributed with zero mean and variance σ^2

$$Minimize \quad SS_{Error} = \sum_{i=1}^{n} (y_i - (\alpha + \beta x_i))^2$$

Example: SeaIce Data

Year: years 1979-2012, independent variable

SeaIce : September Arctic Sea Ice Extent (1,000,000 sq km) measured in each year, dependent variable



Example: SeaIce Data

$$y = 190.12418 - 0.09205x$$

Every year there is an estimated decrease of 0.09205 (1,000,000 sq km) of sea ice.

t-values and their p-values, indicating if the parameters are significantly $\neq 0$ in the model.

```
> model.seaice <- lm(SeaIce~Year)
> summary(model.seaice)

Call:
lm(formula = SeaIce ~ Year)

Residuals:
    Min          1Q      Median        3Q        Max
-1.30259    -0.34064    0.01161     0.36576   1.49456

Coefficients:
              Estimate      Std. Error     t value    Pr(>|t|)
(Intercept)  190.12418 α    20.00964        9.502     7.80e-11 ***
Year          -0.09205 β     0.01003       -9.180     1.76e-10 ***
---
Signif. codes:  0 '***'  0.001 '**'  0.01 '*'  0.05 '.'  0.1 ' '  1

Residual standard error: 0.5736 on 32 degrees of freedom
Multiple R-squared:  0.7248,  Adjusted R-squared:  0.7162
F-statistic: 84.28 on 1 and 32 DF,  p-value: 1.76e-10
```

Example: SeaIce Data

```
> model.seaice <- lm(SeaIce~Year)
> summary(model.seaice)

Call:
lm(formula = SeaIce ~ Year)

Residuals:
    Min         1Q       Median       3Q       Max
-1.30259    -0.34064    0.01161     0.36576   1.49456

Coefficients:
               Estimate      Std. Error     t value     Pr(>|t|)
(Intercept)   190.12418      20.00964        9.502       7.80e-11 ***
Year           -0.09205       0.01003       -9.180       1.76e-10 ***
---
Signif. codes:  0 '***'   0.001 '**'   0.01 '*'   0.05 '.'   0.1 ' '   1

Residual standard error: 0.5736 on 32 degrees of freedom
Multiple R-squared:  0.7248, Adjusted R-squared:  0.7162
F-statistic: 84.28 on 1 and 32 DF,  p-value: 1.76e-10
```

R^2, always between 0 and 1, is a measure of the global adequacy of x as a predictor of y, indicating the portion of variability explained by the model;
72.48% of variation is explained here.

Is the model adequate?

Like R^2, but penalized for the number of parameters included in the model.

Example: SeaIce Data

The ANOVA F-test is a global test of the regression model: it tests whether the covariate is an influential variable associated with a systematic change in the response.

```
> model.seaice <- lm(SeaIce~Year)
> summary(model.seaice)

Call:
lm(formula = SeaIce ~ Year)

Residuals:
    Min        1Q      Median      3Q       Max
-1.30259  -0.34064   0.01161   0.36576  1.49456

Coefficients:
             Estimate    Std. Error   t value   Pr(>|t|)
(Intercept) 190.12418    20.00964      9.502    7.80e-11 ***
Year         -0.09205     0.01003     -9.180    1.76e-10 ***
---
Signif. codes:  0 ‘***’  0.001 ‘**’  0.01 ‘*’  0.05 ‘.’  0.1 ‘ ’  1

Residual standard error: 0.5736 on 32 degrees of freedom
Multiple R-squared:  0.7248, Adjusted R-squared:  0.7162
F-statistic: 84.28 on 1 and 32 DF,  p-value: 1.76e-10
```

Example: SeaIce Data

```
> anova(model.seaice)
Analysis of Variance Table

Response: SeaIce
            Df     Sum Sq     Mean Sq      F value      Pr(>F)
Year         1     27.731     27.731       84.278       1.76e-10 ***
Residuals   32     10.529      0.329
---
Signif. codes:  0 ‘***’  0.001 ‘**’  0.01 ‘*’  0.05 ‘.’  0.1 ‘ ’  1
```

Year is an influential variable associated with the change of sea ice.

$$SS_{Total} = SS_{Regression} + SS_{Error}$$

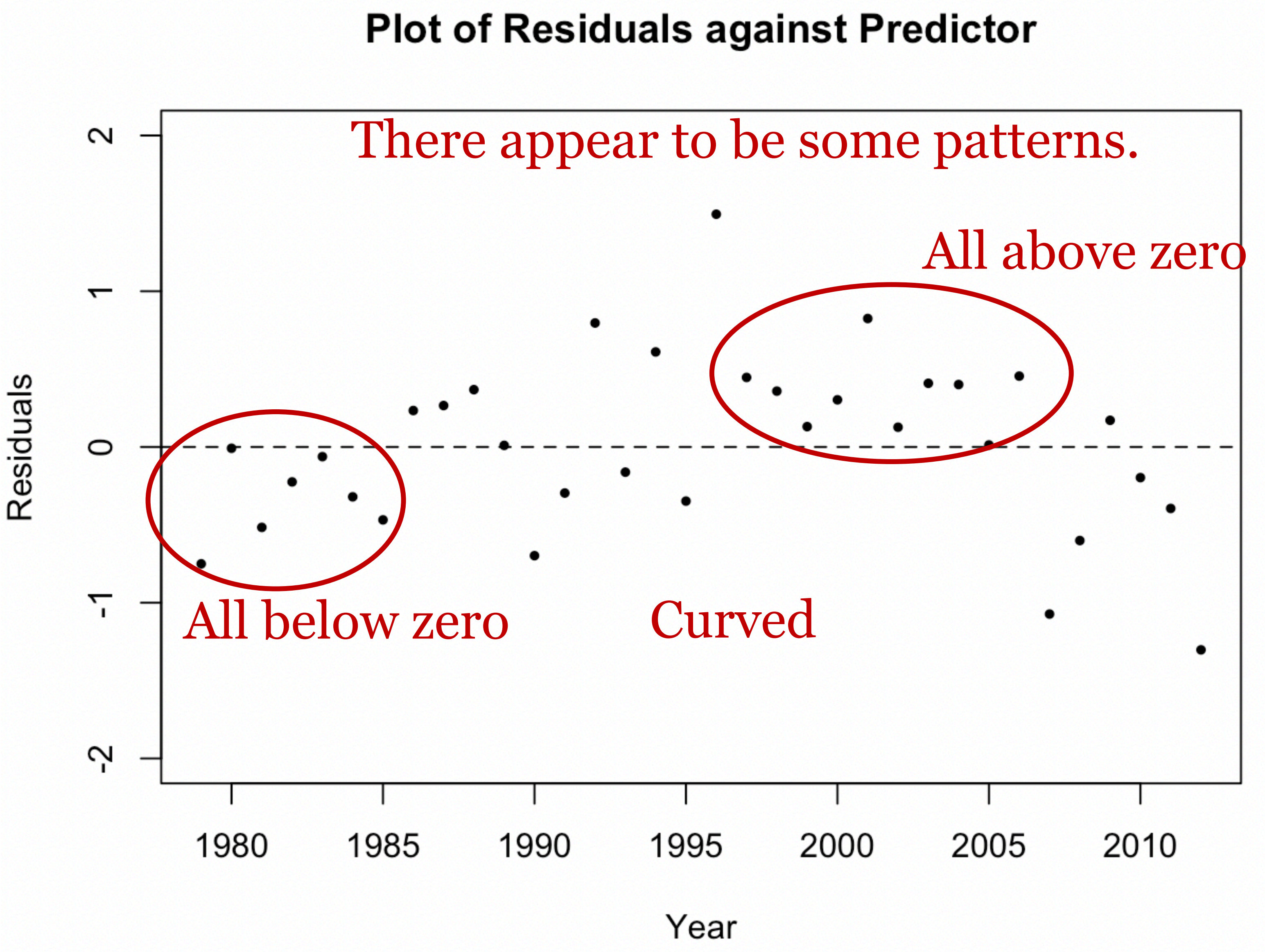
$$F = \frac{MS_{Regression}}{MS_{Error}}$$

❖ The F-test and t-test are equivalent in simple linear regression only.

Assessing Model Adequacy

- R^2 is a measure of global adequacy of the model.
- ANOVA F-test is used to determine if the covariate is associated with the change in the response.
- Residual plots are used to assess "local" model adequacy.
- If the model assumptions are correct, then the residual plots should not exhibit systematic patterns in either mean-level or variability and should appear "pattern-less".
- The residuals should form a horizontal "band" around zero, with equal variability around zero everywhere.

Example: SeaIce Data



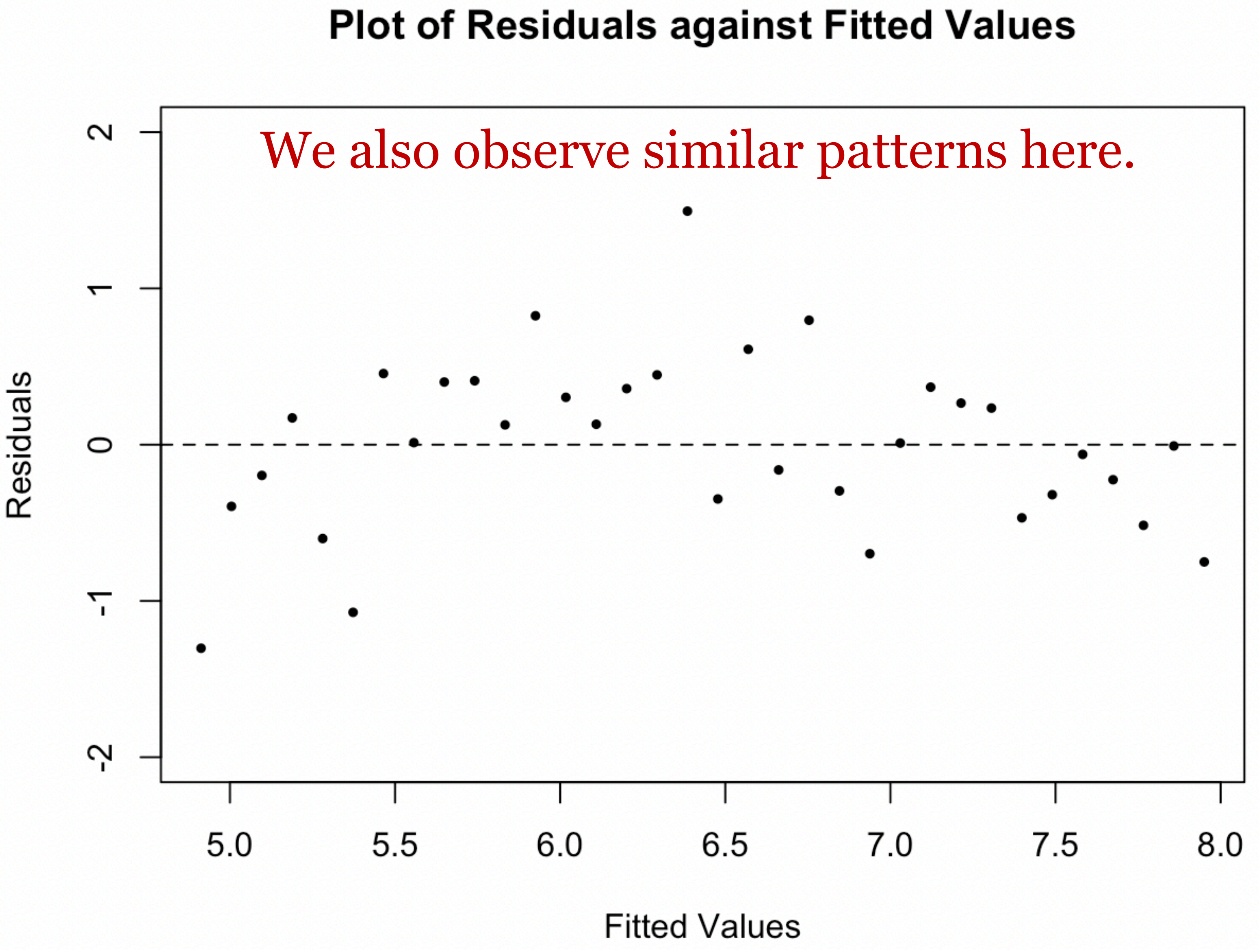
There appear to be some patterns.

All above zero

All below zero

Curved

Example: SeaIce Data



We also observe similar patterns here.

Example: SeaIce Data

```
> Year.transformed <- (Year-1979)^2
> model.seaice.quadratic <- lm(SeaIce~Year.transformed)
> summary(model.seaice.quadratic)

Call:
lm(formula = SeaIce ~ Year.transformed)

Residuals:
     Min       1Q   Median       3Q      Max
-0.94366 -0.27455  0.06663  0.29295  1.22126

Coefficients:
                  Estimate   Std. Error   t value   Pr(>|t|)
(Intercept)       7.4849184  0.1199383     62.41    < 2e-16 ***
Year.transformed -0.0028587  0.0002408    -11.87   2.92e-13 ***
---
Signif. codes:  0 '***'  0.001 '**'  0.01 '*'  0.05 '.'  0.1 ' '  1

Residual standard error: 0.4704 on 32 degrees of freedom
Multiple R-squared:  0.8149, Adjusted R-squared:  0.8091
F-statistic: 140.9 on 1 and 32 DF,  p-value: 2.921e-13
```

$$y = 7.48492 - 0.00286(x - 1979)^2$$

R^2 has improved.

81.49% explained.

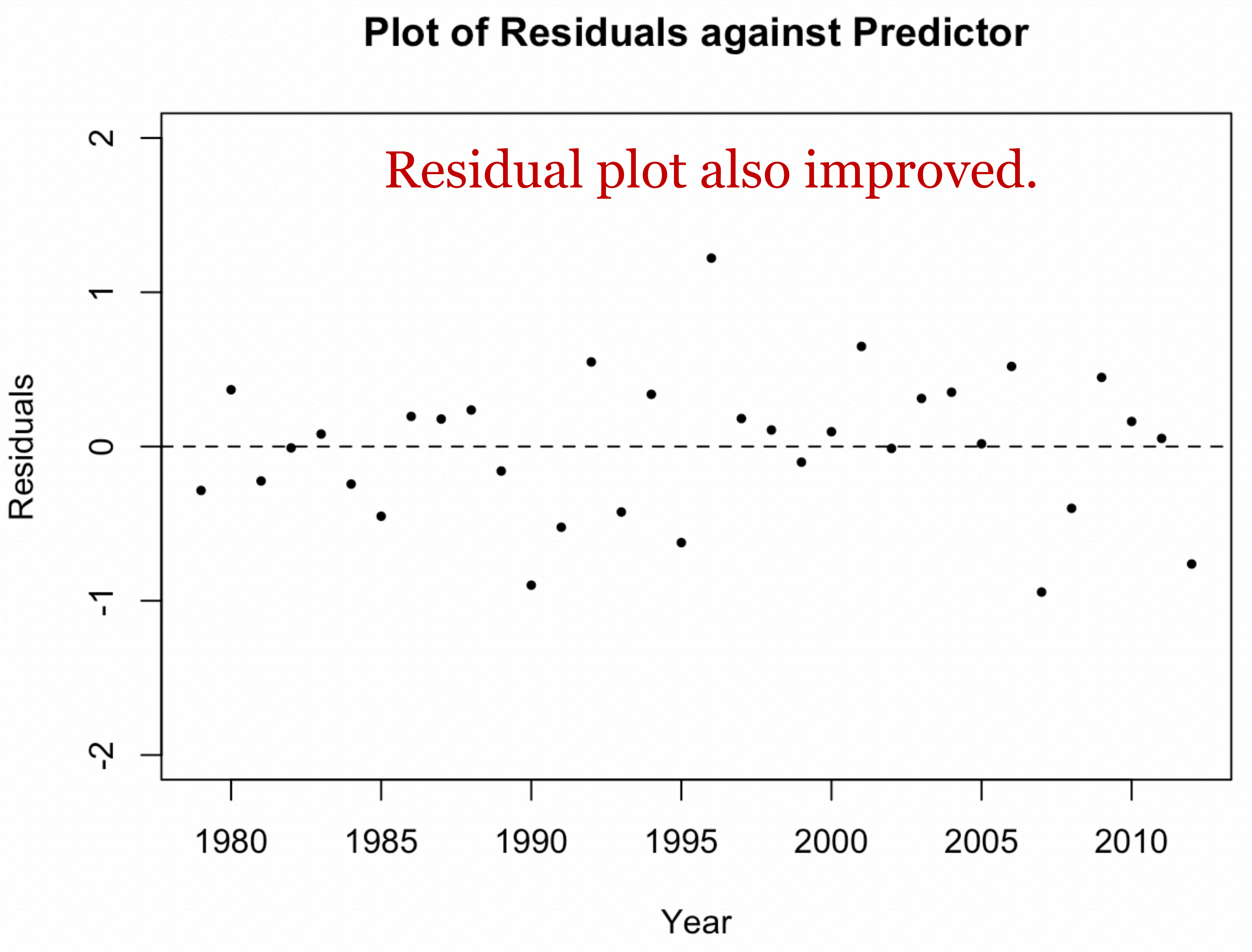
Example: SeaIce Data

```
> anova(model.seaice.quadratic)
Analysis of Variance Table

Response: SeaIce
                  Df      Sum Sq     Mean Sq     F value     Pr(>F)
Year.transformed   1     31.1785     31.1785      140.89     2.921e-13 ***
Residuals         32      7.0814      0.2213
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

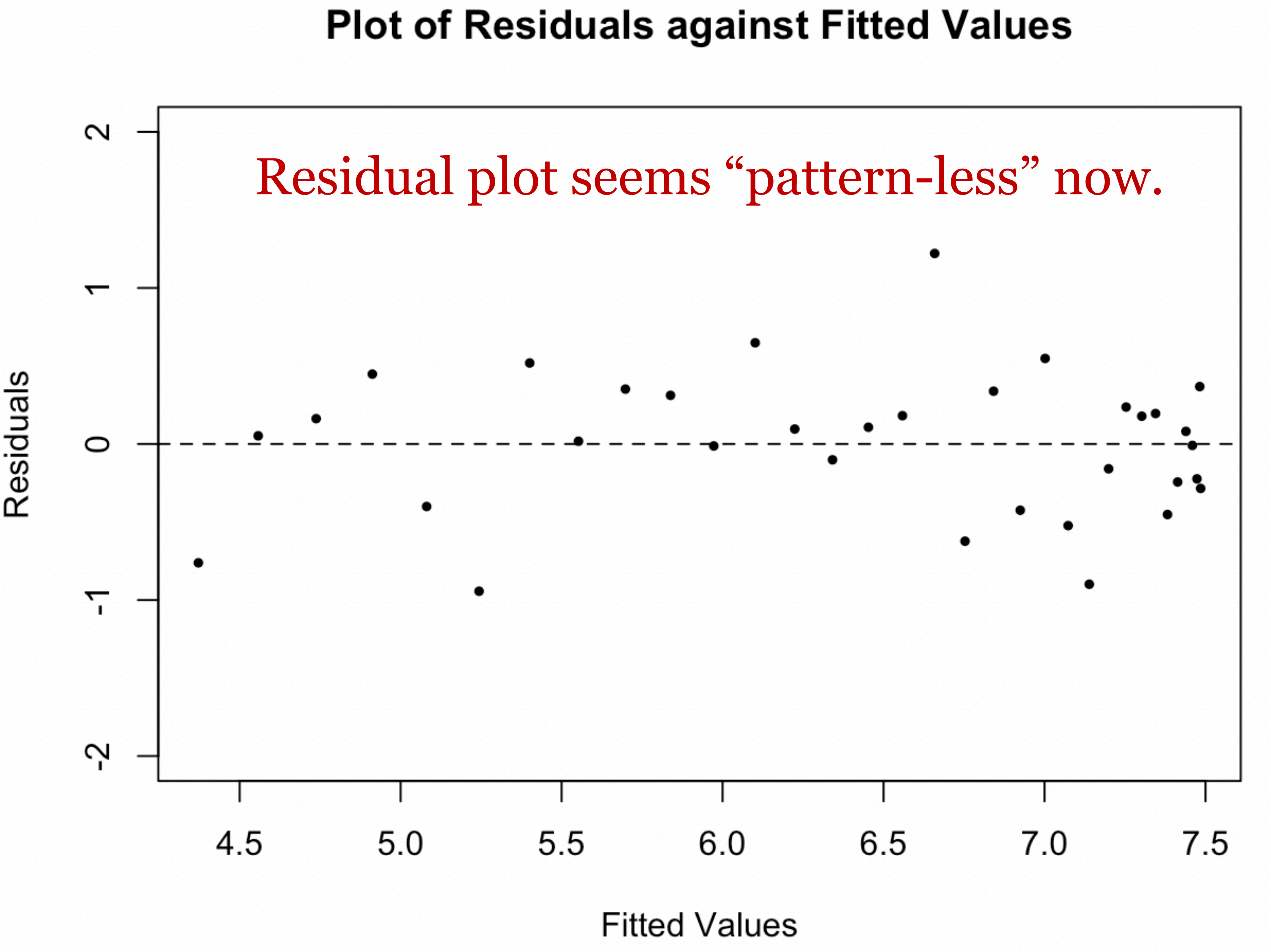
Year as a transformed variable is still an influential variable associated with the change of sea ice.

Example: SeaIce Data



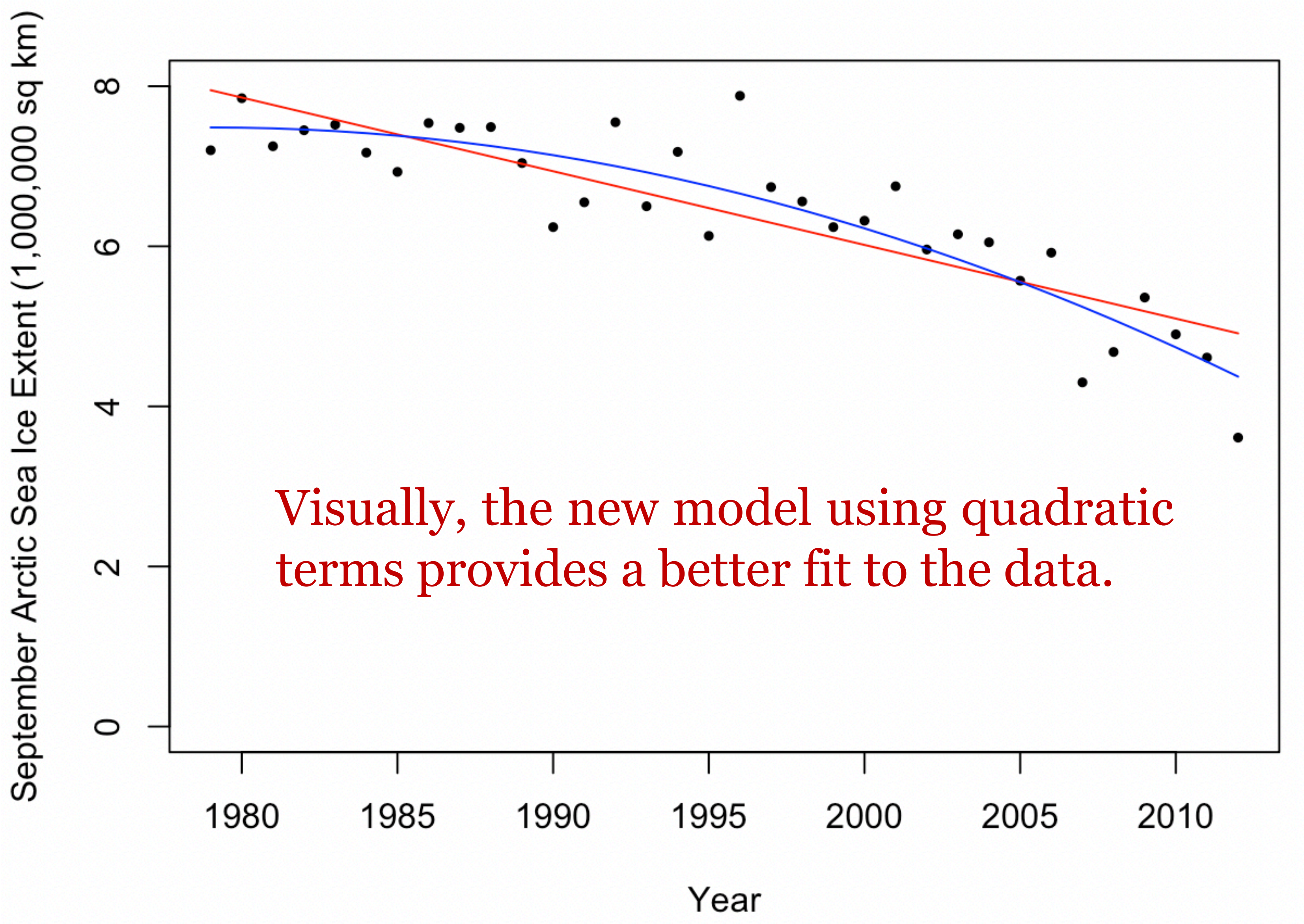
Residual plot also improved.

Example: SeaIce Data



Residual plot seems “pattern-less” now.

Example: SeaIce Data



Visually, the new model using quadratic terms provides a better fit to the data.

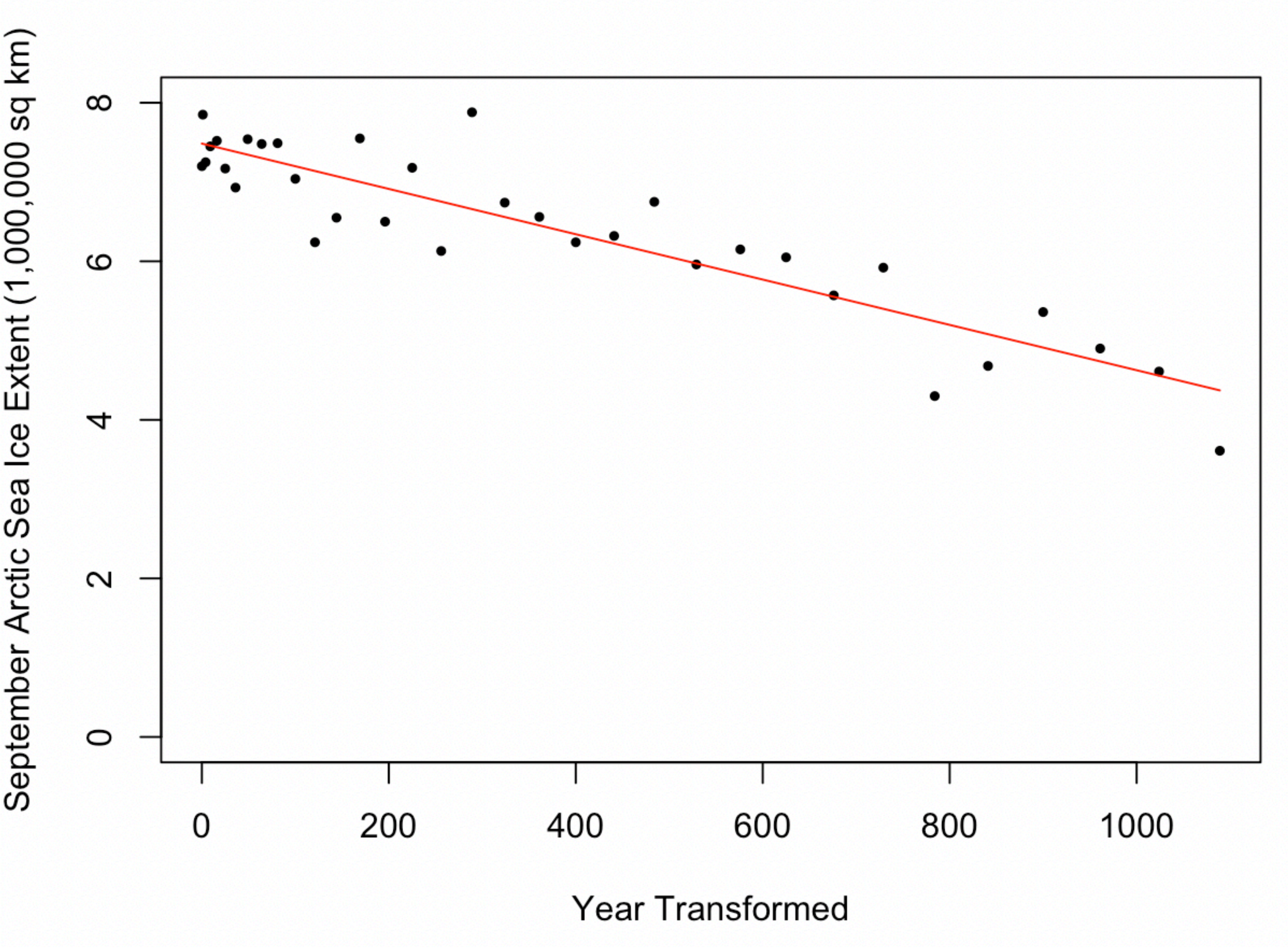
Comments on Linearity

- In the original space, the quadratic curve is the preferred model for the data.

- In the transformed space, the model is still the line of best fit.

- Linearity is with respect to the model parameters, rather than the original independent variables.

Example: SeaIce Data



Multiple Linear Regression

Formulating the Model

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_{p-1} x_{p-1} + \epsilon$$

β_j: parameter associated with the j^{th} covariate x_j

$j = 1, 2, \cdots, p - 1$

p: the number of parameters in the model

$p = 2$ in simple linear regression

ϵ: error, independent random variable normally distributed with zero mean and variance σ^2

Introducing Matrix Form

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_{p-1} x_{i,p-1} + \epsilon_i$$

$$i = 1, 2, \cdots, n$$

$$\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1,p-1} \\ 1 & x_{21} & x_{22} & \cdots & x_{2,p-1} \\ \vdots & & \ddots & & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{n,p-1} \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_{p-1} \end{pmatrix} + \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \vdots \\ \epsilon_n \end{pmatrix}$$

$$\boldsymbol{y} = X\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

$$Minimize \quad SS_{Error} = (\boldsymbol{y} - X\boldsymbol{\beta})^T (\boldsymbol{y} - X\boldsymbol{\beta})$$

Example: Blood Viscosity Data

Model Blood Viscosity as a function of Packed Cell Volume (PCV), Plasma Fibrinogen, and Plasma Protein.

Unit	Viscosity	PCV	Plasma.Fib.	Plasma.Pro.
1	3.71	40.00	344	6.27
2	3.78	40.00	330	4.86
3	3.85	42.50	280	5.09
4	3.88	42.00	418	6.79
5	3.98	45.00	774	6.40
6	4.03	42.00	388	5.48
7	4.05	42.50	336	6.27
8	4.14	47.00	431	6.89
9	4.14	46.75	276	5.18
10	4.20	48.00	422	5.73

Reference: Begg, C. B. and Hearns, J. B. (1966) Components of Blood Viscosity. The relative contributions of haematocrit, plasma fibrinogen and other proteins, Clinical Science, 31, 87-92.

Example: Blood Viscosity Data

Model1 with three **main effects**: PCV, Plasma Fib., Plasma Pro.

$$y = -1.3782 + 0.1168x_1 + 0.0004x_2 + 0.0400x_3$$

```
> model1 <- lm(visc~pcv+fib+pro)
> summary(model1)

Call:
lm(formula = visc ~ pcv + fib + pro)

Coefficients:
               Estimate  Std. Error  t value   Pr(>|t|)
(Intercept)  -1.3782383   0.8966650   -1.537   0.136
pcv           0.1168232   0.0136089    8.584   2.5e-09 ***
fib           0.0004019   0.0003505    1.147   0.261
pro           0.0400364   0.0971527    0.412   0.683
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.3037 on 28 degrees of freedom
Multiple R-squared:  0.7839  Adjusted R-squared:  0.7607
F-statistic: 33.86 on 3 and 28 DF,  p-value: 1.876e-09
```

Only the parameter for PCV seems to be significantly different from 0.

78.39% of variability is explained here.

The model accounts for a significant amount of variability in the data.

Example: Blood Viscosity Data

Model2 with two **main effects**: PCV, Plasma Fib.

$$y = -1.1030 + 0.1160x_1 + 0.0004x_2$$

```
> model2 <- lm(visc~pcv+fib)
> summary(model2)

Call:
lm(formula = visc ~ pcv + fib)

Coefficients:
               Estimate    Std. Error    t value    Pr(>|t|)
(Intercept)   -1.1030187   0.5896907     -1.871     0.0715 .
pcv            0.1159823   0.0132611      8.746     1.26e-09 ***
fib            0.0004042   0.0003454      1.170     0.2514
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.2993 on 29 degrees of freedom
Multiple R-squared:  0.7826, Adjusted R-squared:  0.7676
F-statistic: 52.19 on 2 and 29 DF,  p-value: 2.458e-10
```

Only the parameter for PCV seems to be significantly different from 0.

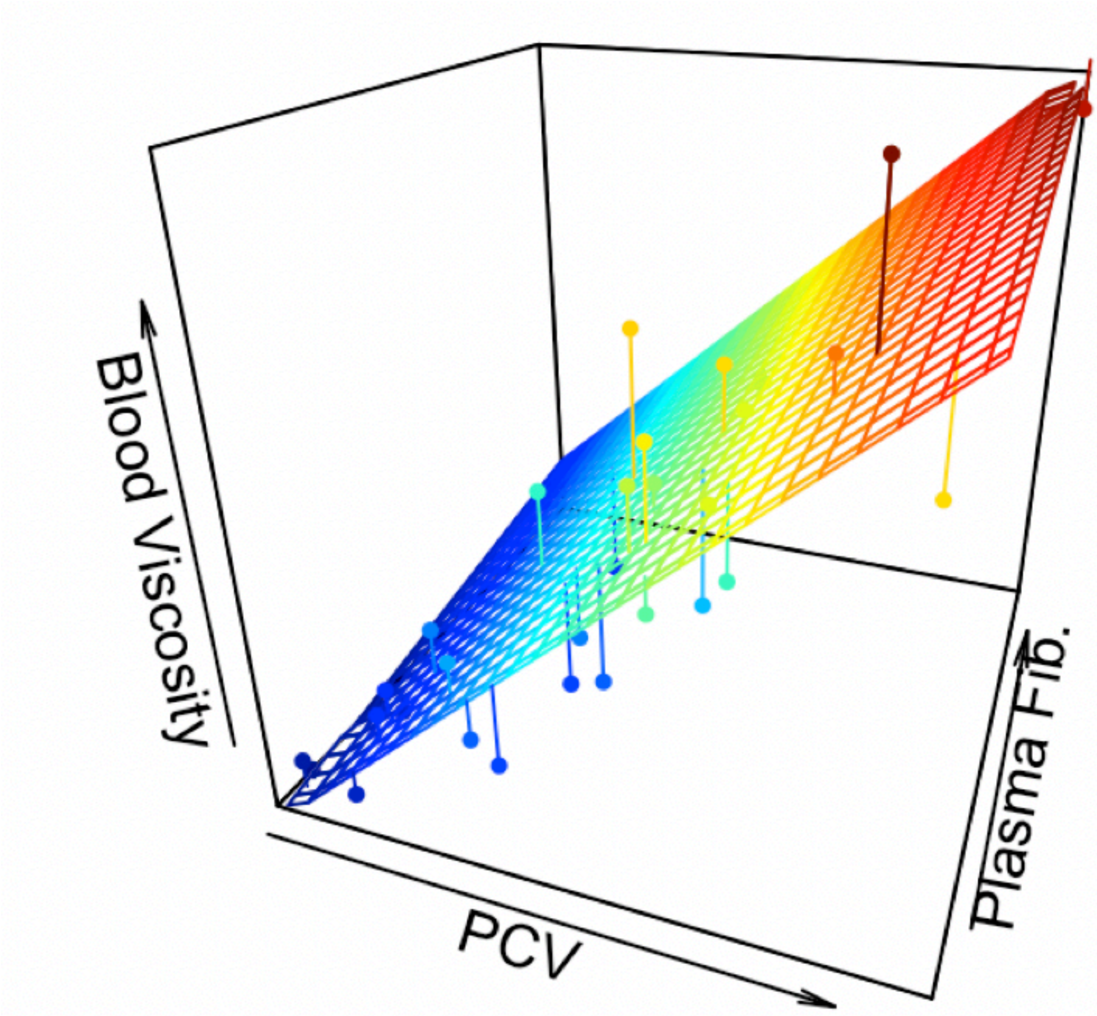
78.26% of variability is explained here.

Adjusted R^2 improved from 0.7607 to 0.7676, since model2 is simpler than model1.

The model accounts for a significant amount of variability in the data.

Example: Blood Viscosity Data

Visualizing Model 2 with PCV and Plasma Fib. as main effects.



Plane of best fit

In higher dimensions, we obtain the hyperplane of best fit.

Example: Blood Viscosity Data

Model3 with only one **main effect**: Plasma Fib.

$$y = 3.8798 + 0.0017x_2$$

```
> model3 <- lm(visc~fib)
> summary(model3)

Call:
lm(formula = visc ~ fib)

Coefficients:
              Estimate    Std. Error    t value    Pr(>|t|)
(Intercept)  3.8708133    0.2924499     13.236     4.64e-14 ***
fib          0.0016595    0.0005892      2.816     0.0085 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.5613 on 30 degrees of freedom
Multiple R-squared:  0.2091, Adjusted R-squared:  0.1828
F-statistic: 7.932 on 1 and 30 DF,  p-value: 0.008504
```

Plasma Fib. is now a significant term in the model.

Only 20.91% of variability is explained here.

Plasma Fib. is significantly associated with the change in blood viscosity.

Example: Blood Viscosity Data

Model4 with only one **main effect**: PCV.

$$y = -1.2234 + 0.1224x_1$$

```
> model4 <- lm(visc~pcv)
> summary(model4)

Call:
lm(formula = visc ~ pcv)

Coefficients:
              Estimate   Std. Error   t value    Pr(>|t|)
(Intercept)   -1.22336    0.58422     -2.094     0.0448 *
pcv            0.12243    0.01214     10.088    3.73e-11 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.3012 on 30 degrees of freedom
Multiple R-squared:  0.7723, Adjusted R-squared:  0.7647
F-statistic: 101.8 on 1 and 30 DF,  p-value: 3.731e-11
```

PCV is a significant term in the model.

77.23% of variability is explained here.

Adjusted R^2 is still higher than 0.7607, since model3 is much simpler than model1.

PCV is significantly associated with the change in blood viscosity.

Model Selection

- We want to select a simple model that explains the variability in the data well and is also easy to interpret.

- We look for a balance among model adequacy, simplicity and interpretability.

- We also want to avoid overfitting, so that: the model is not too sensitive towards new data points; it has better prediction power and can be generalized more easily.

Example: Blood Viscosity Data

```
> anova(model3,model1)
Analysis of Variance Table

Model 1: visc ~ fib
Model 2: visc ~ pcv + fib + pro
     Res.Df    RSS    Df   Sum of Sq      F      Pr(>F)
1      30    9.4513
2      28    2.5825   2      6.8688    37.237  1.293e-08 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Evidence against model3, indicating that model3 is not an adequate simplification of model1.

```
> anova(model4,model1)
Analysis of Variance Table

Model 1: visc ~ pcv
Model 2: visc ~ pcv + fib + pro
     Res.Df    RSS    Df   Sum of Sq      F      Pr(>F)
1      30    2.7208
2      28    2.5825   2      0.13835    0.75     0.4816
```

No evidence against model4, indicating that model4 is an acceptable simplification of model1.

Example: Blood Viscosity Data

```
> anova(model3,model2,model1)
Analysis of Variance Table

Model 1: visc ~ fib
Model 2: visc ~ pcv + fib
Model 3: visc ~ pcv + fib + pro
     Res.Df    RSS    Df   Sum of Sq       F        Pr(>F)
1      30     9.4513
2      29     2.5982   1      6.8532    74.3037   2.293e-09 ***
3      28     2.5825   1      0.0157     0.1698   0.6834
```

No evidence against model2 when compared with model1, but model3 is rejected when compared with model2.

```
> anova(model4,model2,model1)
Analysis of Variance Table

Model 1: visc ~ pcv
Model 2: visc ~ pcv + fib
Model 3: visc ~ pcv + fib + pro
     Res.Df    RSS    Df   Sum of Sq       F       Pr(>F)
1      30    2.7208
2      29    2.5982   1     0.122691    1.3302   0.2585
3      28    2.5825   1     0.015663    0.1698   0.6834
```

Models must be nested.

No evidence against model simplification at each step.

Analysis of Variance (ANOVA)

Example: Honey Cough Data

- **One-way ANOVA** analyzing the effect of types of treatment (Honey, DM, or no treatment) on children’s nocturnal cough.

	Honey	DM	Control
1	12	4	5
2	11	6	8
3	15	9	6
4	11	4	1
5	10	7	0
6	13	7	8
7	10	7	12
8	4	9	8
9	15	12	7

→

ImproveScore	Treatment
12	H
4	DM
5	C
11	H
6	DM
8	C
15	H
9	DM
6	C

Reference: Paul, I. M., Beiler, J., McMonagle, A., Shaffer, M. L., Duda, L., & Berlin, C. M. (2007). Effect of honey, dextromethorphan, and no treatment on nocturnal cough and sleep quality for coughing children and their parents. Archives of pediatrics & adolescent medicine, 161(12), 1140-1146.

Example: Honey Cough Data

```
> anova.cough <- aov(ImproveScore ~ Treatment)
> summary(anova.cough)
            Df    Sum Sq   Mean Sq   F value   Pr(>F)
Treatment    2    318.5    159.2     17.51     2.9e-07 ***
Residuals  102    927.7      9.1
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Improvement differs significantly by types of treatment.

However, we cannot identify which treatment types are different from the others.

- We have a **completely randomized design** (CRD).

$$SS_{Total} = SS_{Treatment} + SS_{Error}$$

$$F = \frac{MS_{Treatment}}{MS_{Error}}$$

Example: Honey Cough Data

```
> anova.cough <- aov(ImproveScore ~ Treatment)
> summary(anova.cough)
              Df    Sum Sq    Mean Sq   F value   Pr(>F)
Treatment      2     318.5     159.2     17.51    2.9e-07 ***
Residuals    102     927.7       9.1
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

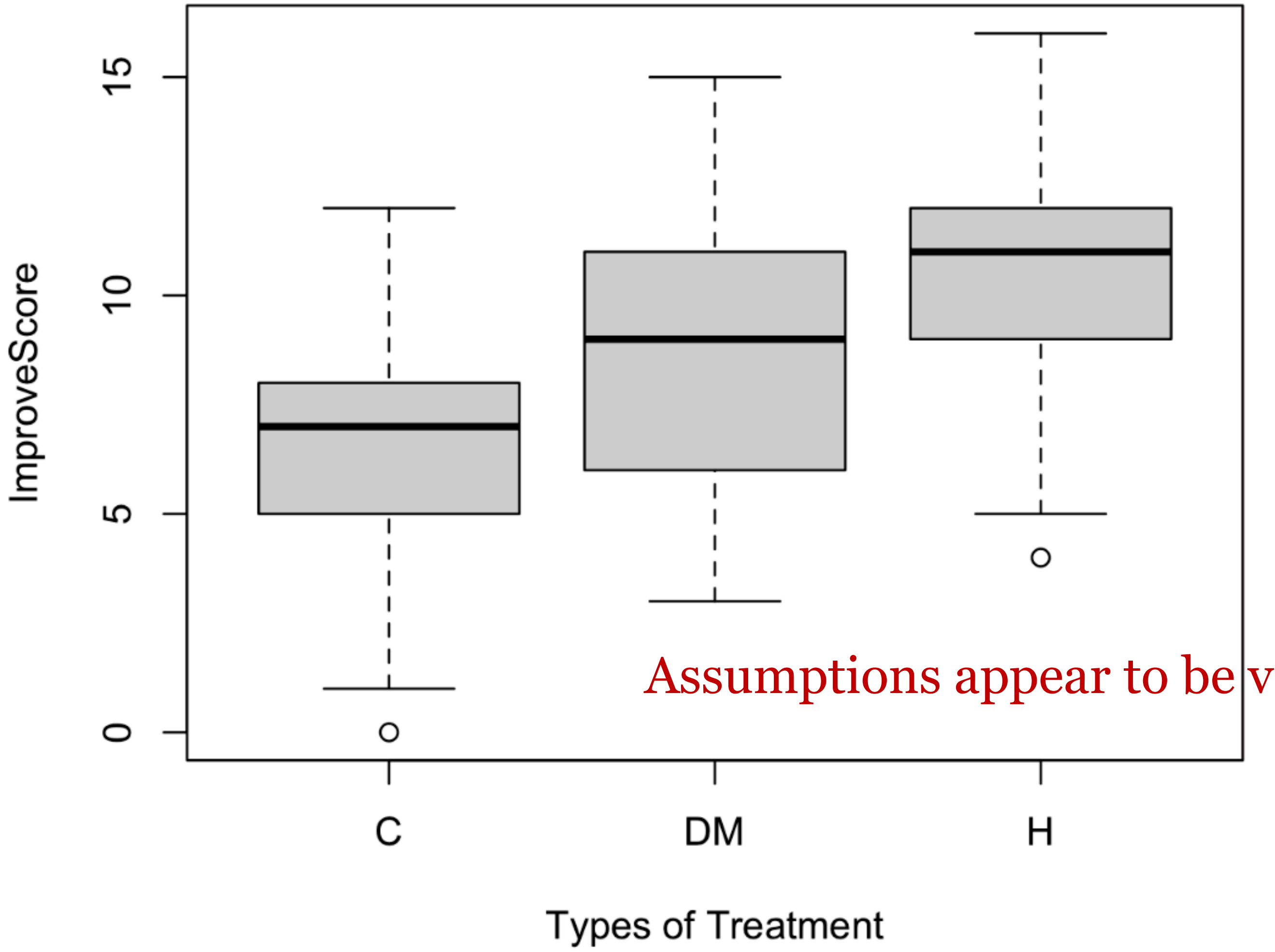
ANOVA for a simple linear regression model is equivalent to a one-way ANOVA for CRD.

```
> model.cough <- lm(ImproveScore ~ Treatment)
> anova(model.cough)
Analysis of Variance Table

Response: ImproveScore
              Df    Sum Sq    Mean Sq   F value   Pr(>F)
Treatment      2    318.51    159.255     17.51   2.902e-07 ***
Residuals    102    927.72      9.095
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Example: Honey Cough Data

- We could use boxplots to visually check the assumptions of normality and equal variance.



Assumptions appear to be valid here.

Example: Honey Cough Data

- Levene’s Test is used to check the assumption of equal variance.

```
> leveneTest(ImproveScore ~ Treatment, ="mean")
Levene's Test for Homogeneity of Variance (center = "mean")
          Df    F value    Pr(>F)
group     2     0.9218     0.4011
        102

> leveneTest(ImproveScore ~ Treatment, center="median")
Levene's Test for Homogeneity of Variance (center = "median")
          Df    F value    Pr(>F)
group     2     0.9169     0.403
        102
```

Assumption of equal variance is met here.

Example: Honey Cough Data

- Post Hoc Test using Tukey HSD can be used to determine pairwise differences.

```
> TukeyHSD(anova.cough)
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = ImproveScore ~ Treatment)

$Treatment
             diff         lwr        upr      p adj
DM-C     1.819820   0.1023625   3.537277  0.0351562
H-C      4.200772   2.5094509   5.892094  0.0000001
H-DM     2.380952   0.6405157   4.121389  0.0043728
```

Each pair has a significant difference in the mean improvement score.

Example: Honey Cough Data

```
> summary(model.cough)

Call:
lm(formula = ImproveScore ~ Treatment)

Residuals:
    Min      1Q  Median      3Q     Max 
-6.7143 -1.7143  0.4865  1.6667  6.6667 

Coefficients:
                 Estimate     Std. Error     t value        Pr(>|t|)
(Intercept)       6.5135        0.4958        13.137       < 2e-16 ***
TreatmentDM       1.8198        0.7221         2.520        0.0133 *
TreatmentH        4.2008        0.7111         5.907       4.62e-08 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.016 on 102 degrees of freedom
Multiple R-squared:  0.2556,  Adjusted R-squared:  0.241
F-statistic: 17.51 on 2 and 102 DF,  p-value: 2.902e-07
```

A simple linear regression model with a factor predictor as the only main effect is modelling the group means.

Differences between treatment groups and the control group are the estimated model parameters.

Example: Honey Cough Data

- Pairwise t test with Bonferroni correction can be used to check significance of pairwise differences as post hoc test.

```
> pairwise.t.test(ImproveScore, Treatment, p.adj = "bonf")

            Pairwise comparisons using t tests with pooled SD

data:  ImproveScore and Treatment
        C         DM
DM  0.0398      -
H    1.4e-07  0.0046

P value adjustment method: bonferroni
```

Each pair has a significant difference in the mean improvement score.

Example: Dosage Data

- ANOVA for **Randomized Block Design** (RBD) analyzing the effect of dosage level on improvement of headaches.

Participant_ID	Control	20mg	60mg
1	-20.88	-15.75	-8.62
2	-4.76	0.11	6.20
3	-0.46	5.64	10.42
4	10.78	16.98	20.05
5	-10.47	-6.03	-1.29

ID	Dosage	Response
1	Control	-20.88
1	20mg	-15.57
1	60mg	-8.62
2	Control	-4.76
2	20mg	0.11
2	60mg	6.20

Each block has 3 units, assigned to each of the three dosage levels.

Blocking factor

Be careful that in general, ID is often not a useful variable. We only use participant ID as a blocking factor when each participant is subjected to the same treatments.

Why is it important to account for blocking effect?

Example: Dosage Data

```
> anova.noID <- aov(Response~Dosage)
> summary(anova.noID)
            Df    Sum Sq    Mean Sq    F value    Pr(>F)
Dosage       2     276.2      138.1      1.012     0.393
Residuals   12    1637.8      136.5
```

Simply treating the data as a CRD and using a one-way ANOVA reveals no significant effect of dosage levels.

After controlling for the blocking effect, we see that response in fact differs by dosage levels.

```
> anova.dosage <- aov(Response ~ Dosage + ID)
> summary(anova.dosage)
            Df    Sum Sq    Mean Sq    F value    Pr(>F)
Dosage       2     276.2      138.1      178.8    2.29e-07 ***
ID           4    1631.6      407.9      528.2    1.01e-09 ***
Residuals    8   6.2    0.8
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Notice that ID is indeed a significant blocking factor.

Example: Dosage Data

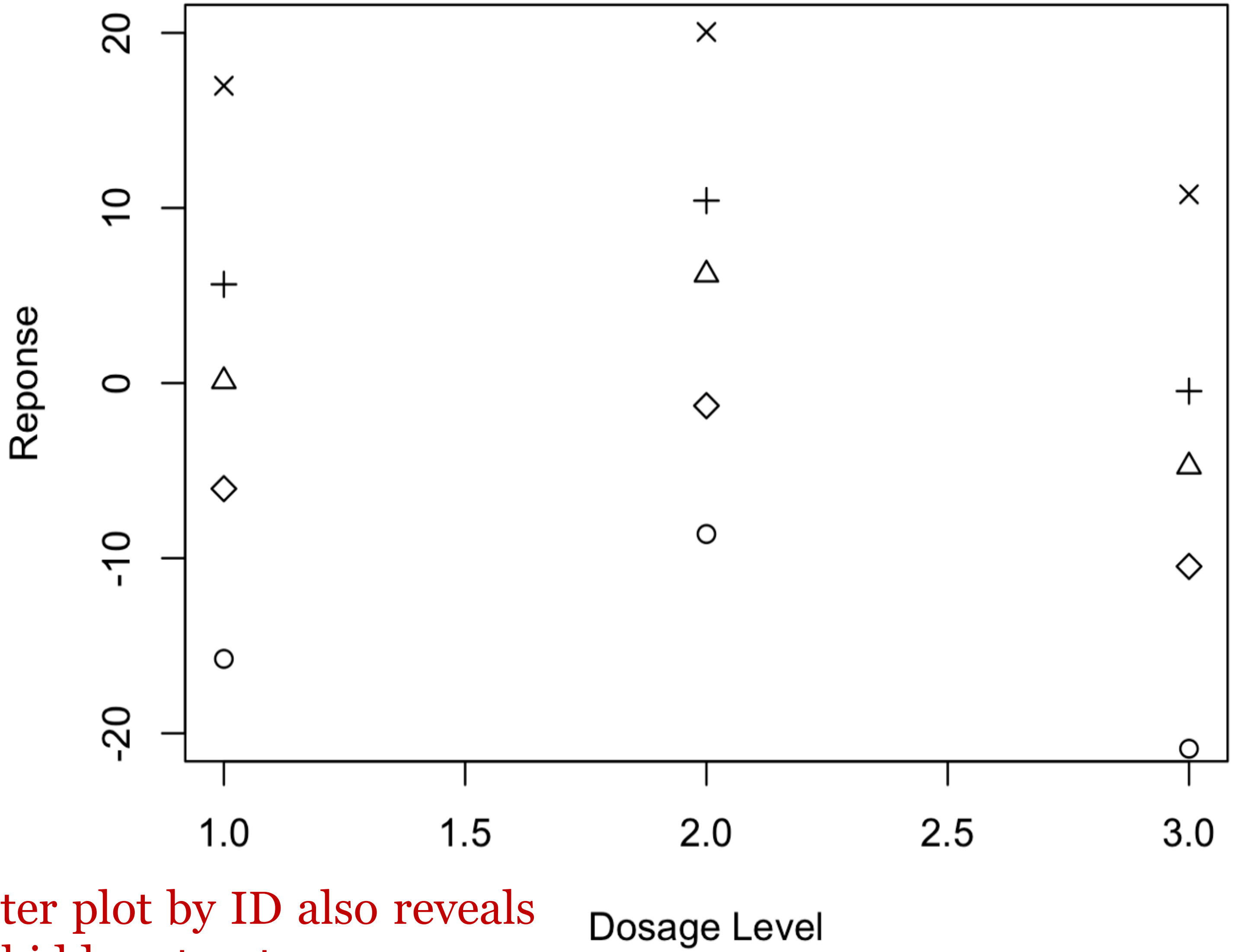
- In RBD, we account for blocking effect when decomposing the total variability.

$$SS_{Total} = SS_{Treatment} + SS_{Block} + SS_{Error}$$

$$F = \frac{MS_{Treatment}}{MS_{Error}}$$

- If we did not control for the blocking effect, SS_{Block} would have been included into SS_{Error}, and we would have missed this structure in the data.

Example: Dosage Data



Scatter plot by ID also reveals the hidden structure.

Example: Dosage Data

```
> TukeyHSD(anova.dosage)
  Tukey multiple comparisons of means
    95% family-wise confidence level
Fit: aov(formula = Response ~ Dosage + ID)
$Dosage
                diff        lwr        upr     p adj
60mg-20mg      5.162   3.573826   6.750174  3.84e-05
Control-20mg  -5.348  -6.936174  -3.759826  2.96e-05
Control-60mg -10.510 -12.098174  -8.921826  1.00e-07
$ID
          diff        lwr        upr     p adj
2-1  15.600000  13.121088  18.078912 0.0000001
3-1  20.283333  17.804421  22.762246 0.0000000
4-1  31.020000  28.541088  33.498912 0.0000000
5-1   9.153333   6.674421  11.632246 0.0000096
3-2   4.683333   2.204421   7.162246 0.0012339
4-2  15.420000  12.941088  17.898912 0.0000001
5-2  -6.446667  -8.925579  -3.967754 0.0001307
4-3  10.736667   8.257754  13.215579 0.0000029
5-3 -11.130000 -13.608912  -8.651088 0.0000022
5-4 -21.866667 -24.345579 -19.387754 0.0000000
```

Each pair of dosage levels has significant differences in response.

Each pair of participants also shows significant differences, which corresponds to the significant blocking effect.

Example: ToothGrowth Data

- Two- or Multi-Way ANOVA for **Factorial Design** (FD) analyzing the effects of two or more **factors** (sources of supplements and dose levels), and their **interaction** on lengths of teeth.

len	supp	dose
17.3	VC	1.0
4.2	VC	0.5
20.0	OJ	1.0
23.6	VC	2.0
23.3	OJ	1.0
5.2	VC	0.5

supp has two levels: OJ, VC
dose has three levels: 0.5, 1.0, 2.0

- There is no difference between FD and RBD in terms of the calculations or methods of analysis. However, in RBD, one of the factors is known beforehand or strongly believed to have a significant effect on the response, whereas in FD, the effects of the factors are unknow before the analysis.

Example: ToothGrowth Data

- In FD, we decompose total variability into variability caused by main effects and by interaction of main effects.

$$SS_{Total} = SS_{FactorA} + SS_{FactorB} + SS_{AB} + SS_{Error}$$

$$F = \frac{MS_{FactorA}}{MS_{Error}} \qquad F = \frac{MS_{FactorB}}{MS_{Error}} \qquad F = \frac{MS_{AB}}{MS_{Error}}$$

```
> anova.tooth <- aov(len ~ supp*as.factor(dose), data = ToothGrowth)
> summary(anova.tooth)
                     Df   Sum Sq   Mean Sq   F value   Pr(>F)
supp                  1    205.4     205.4    15.572   0.000231 ***
as.factor(dose)       2   2426.4    1213.2    92.000    < 2e-16 ***
supp:as.factor(dose)  2    108.3      54.2     4.107   0.021860 *
Residuals            54    712.1      13.2
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

The interaction between supp and dose has a significant effect on tooth growth.

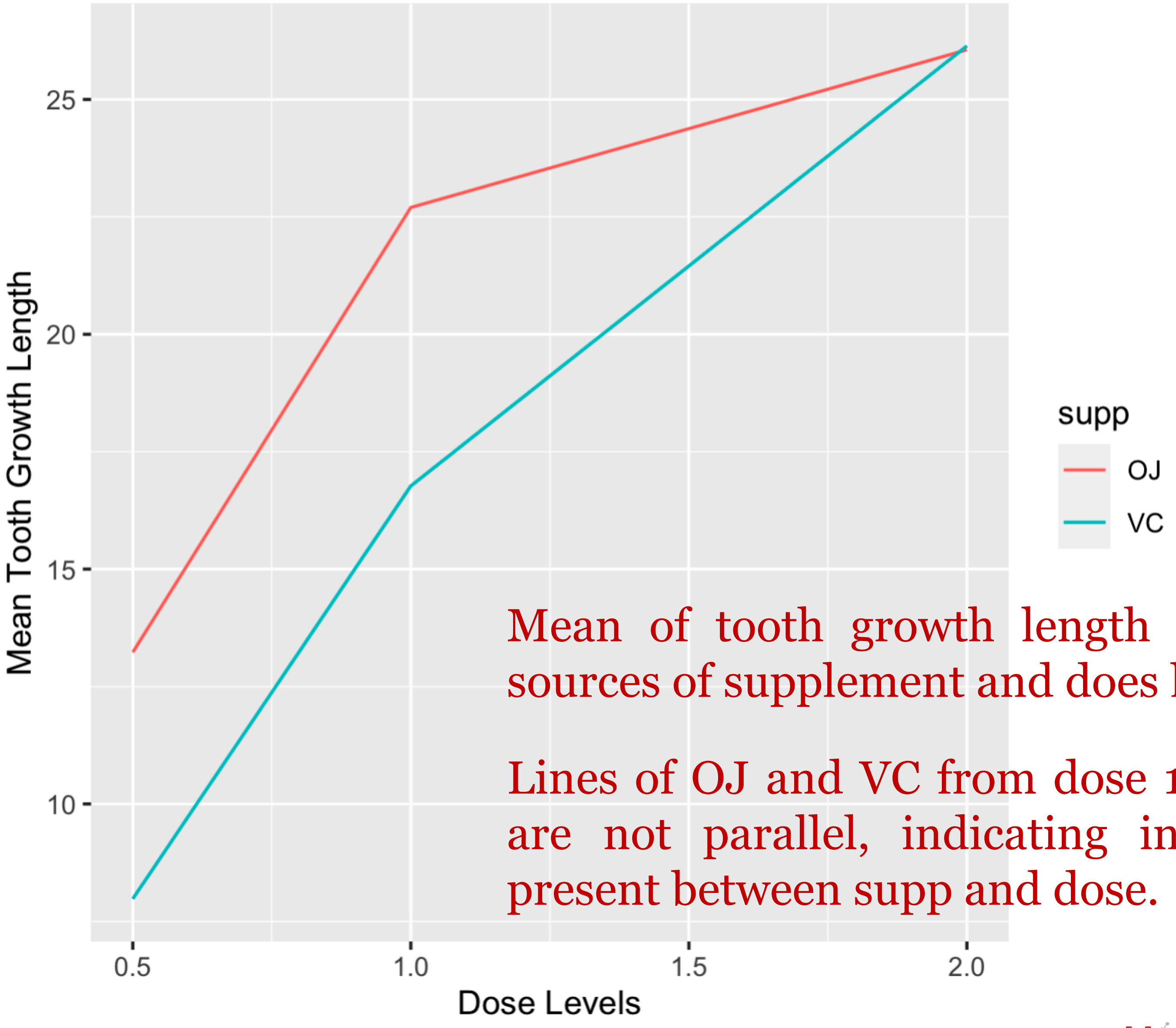
Example: ToothGrowth Data

```
> leveneTest(len ~supp*as.factor(dose), data = ToothGrowth)
Levene's Test for Homogeneity of Variance (center = median)
        Df     F value     Pr(>F)
group    5     1.7086      0.1484
        54
```

Assumption of equal variance is met.

```
> leveneTest(len ~supp*as.factor(dose), data = ToothGrowth, center = "mean")
Levene's Test for Homogeneity of Variance (center = "mean")
        Df     F value     Pr(>F)
group    5      1.9401     0.1027
        54
```

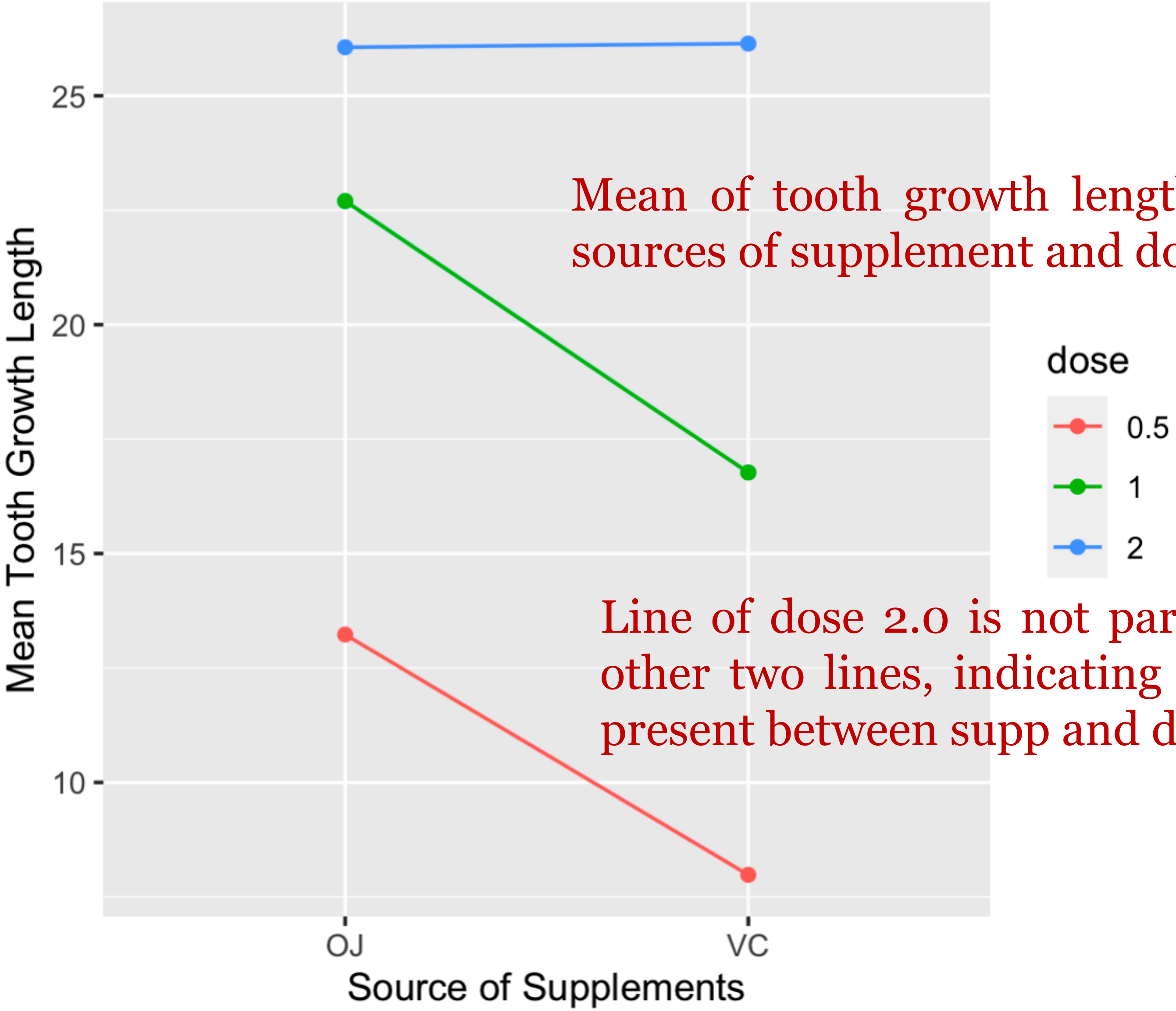
Example: ToothGrowth Data



Mean of tooth growth length differ by sources of supplement and does levels.

Lines of OJ and VC from dose 1.0 to 2.0 are not parallel, indicating interaction present between supp and dose.

Example: ToothGrowth Data



Mean of tooth growth length differ by sources of supplement and does levels.

Line of dose 2.0 is not parallel to the other two lines, indicating interaction present between supp and dose.

McGill initiative in Computational Medicine

A Final Encouraging Quote

All models are wrong,
but some are useful!

George Edward Pelham Box
(18 October 1919 – 28 March 2013)

If you have more questions

xiaonan.da@mail.mcgill.ca