SUMMER 2025

TEXAS
The University of Texas at Austin

INTRODUCTION TO STATISTICAL MODELING

Center for Biomedical Research Support

LAYLA GUYOT
Assistant Professor of Instruction, Ph.D.
Department of Statistics and Data Sciences
The University of Texas at Austin

Access materials

laylaguyot

Overview | Repositories 8 | Projects

Layla Guyot
laylaguyot

Statistics and Data Science enthusiast: teacher and researcher in education, focusing on bridging the gap between academia and industry.

https://github.com/laylaguyot/CBRS_Intro_Statistical_Modeling

README

Introduction to Statistical Modeling

Materials

Download the slides and corresponding .Rmd files each day

Access RStudio

Go to https://gsafcomp02.ccbb.utexas.edu and select RStudio

Tentative Schedule

Day 1
Exploring Data

- Study design and variables
- Descriptive statistics and visualizations
- Introduction to hypothesis testing

Day 2
Making Inferences

- Probability, random variables, and common probability distributions
- Sampling distributions and Central Limit Theorem
- Confidence intervals, t-tests, ANOVA, and Chi-square tests

Day 3
Linear Regression

- Simple Linear Regression
- Multiple Regression with different types of predictors
- Model assumptions, evaluation, and comparisons

Day 4
Logistic Regression

- Odds
- Logistic Regression
- Model evaluation with ROC curves or confusion matrix

Day 5
Model Building

- Underfitting, overfitting, and cross-validation
- Regularization with Lasso and Ridge
- Missing data

Summary of all tests:

Name of test	Variables involved	Hypotheses	Test statistic	df	Assumptions*	Effect size
One-sample t-test	numeric response	H_0: $\mu = \mu_0$ H_A: $\mu \neq \mu_0$	$t = \frac{\bar{X} - \mu_0}{SE}$	$n - 1$	✓ normal	$\bar{X} \pm t^* \cdot SE$
Independent t-test	numeric response binary predictor	H_0: $\mu_1 = \mu_2$ H_A: $\mu_1 \neq \mu_2$	$t = \frac{\bar{X}_1 - \bar{X}_2}{SE}$	$n_1 + n_2 - 1$	✓ normal ✓ equal variance	$\bar{X}_1 - \bar{X}_2 \pm t^* \cdot SE$
ANOVA	numeric response categorical predictor	H_0: $\mu_1 = \mu_2 = \cdots$ H_A: $not\ all\ equal$	$F = \frac{MS_{group}}{MS_{error}}$	$df_{group} = k - 1$ $df_{error} = n - k$	✓ normal ✓ equal variance	Post Hoc Model fit R^2
Chi2 Goodness-of-Fit	categorical response	H_0: $distrib\ is\ \ldots$ H_A: $distrib\ is\ not\ \ldots$	$\chi^2 = \sum \frac{(obs\ - exp)^2}{exp}$	$\#\ cat\ - 1$	✓ sample size	percentages
Chi2 Test of Independence	categorical response categorical predictor	H_0: $independent$ H_A: $not\ independent$	$\chi^2 = \sum \frac{(obs\ - exp)^2}{exp}$	$(\#\ cat_1\ - 1)$ $\cdot (\#\ cat_2\ - 1)$	✓ sample size	percentages

***Random sample and Independent observations are common assumptions to all these tests**

Checking in

https://www.menti.com/al187iyoshb7

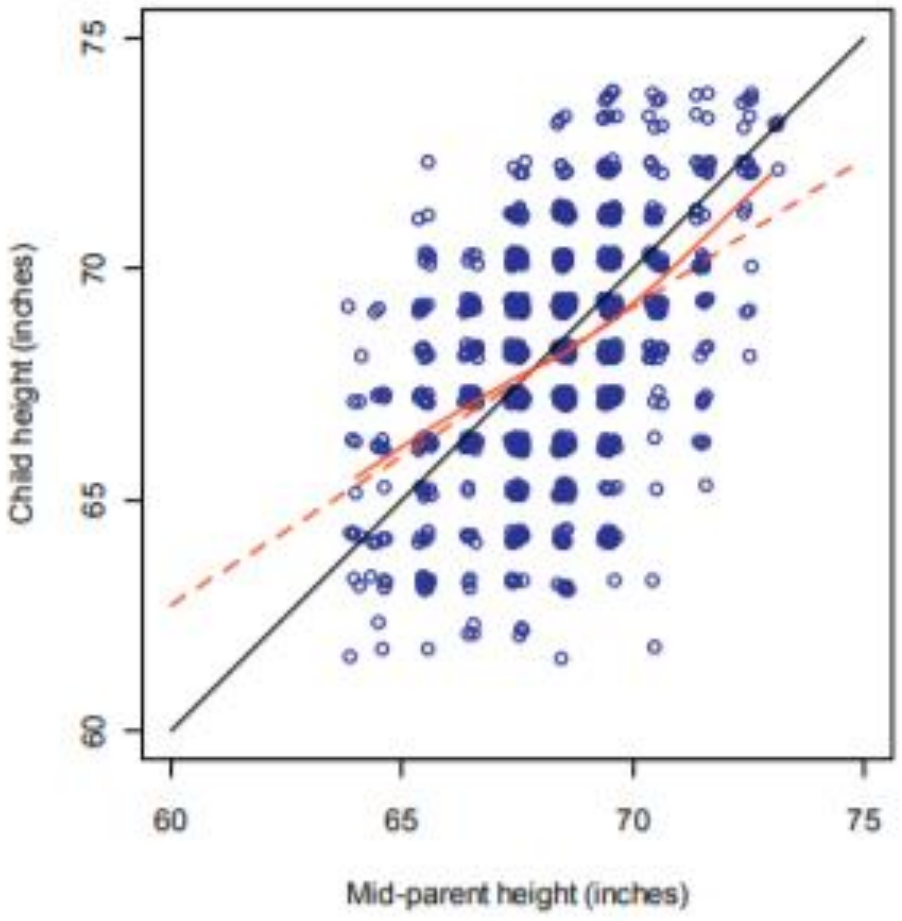
Mentimeter

What is something you have learned (or re-discovered) during this workshop so far?

Linear Regression

Where does the term "regression" come from?

- Coined by Francis Galton in the late 1800s.
- Galton observed that tall parents tended to have shorter children (closer to the population average) and vice versa. He called this pattern "regression toward mediocrity“, what we now call regression to the mean.
- Galton and Pearson used family height data to describe and model these trends.



⚠ While foundational to modern statistics for the least squares linear model, Galton and Pearson’s work is inseparable from their **promotion of eugenics**.

Simple Linear Regression

- Population model:

error/residual

$$Y = \beta_0 + \beta_1 X_1 + \varepsilon$$

- Estimated regression function:

$$\hat{Y} = b_0 + b_1 X_1$$

- Residuals:

$$e_i = Y_i - \hat{Y}_i$$

What should we test about to check for a "significant" linear relationship?

Simple Linear Regression

How to determine the line to model our data?

Least Squares Regression Line

$$\hat{Y} = b_0 + b_1 X \quad \textbf{with} \quad b_1 = r\frac{s_Y}{s_X}$$

$$b_0 = \bar{Y} - b_1\bar{X}$$

Least-Squares Regression Demo

USING R AND RSTUDIO

Simple Linear Regression

Besides random sample, there are 4 assumptions for the SLR model:

Linearity: the mean response is a linear function of X_i

Independent observations: the errors, ϵ_i, are independent

Normality of residuals: the errors, ϵ_i, are normally distributed

Equal variances: the errors, ϵ_i, have equal variances (σ^2)

Simple Linear Regression

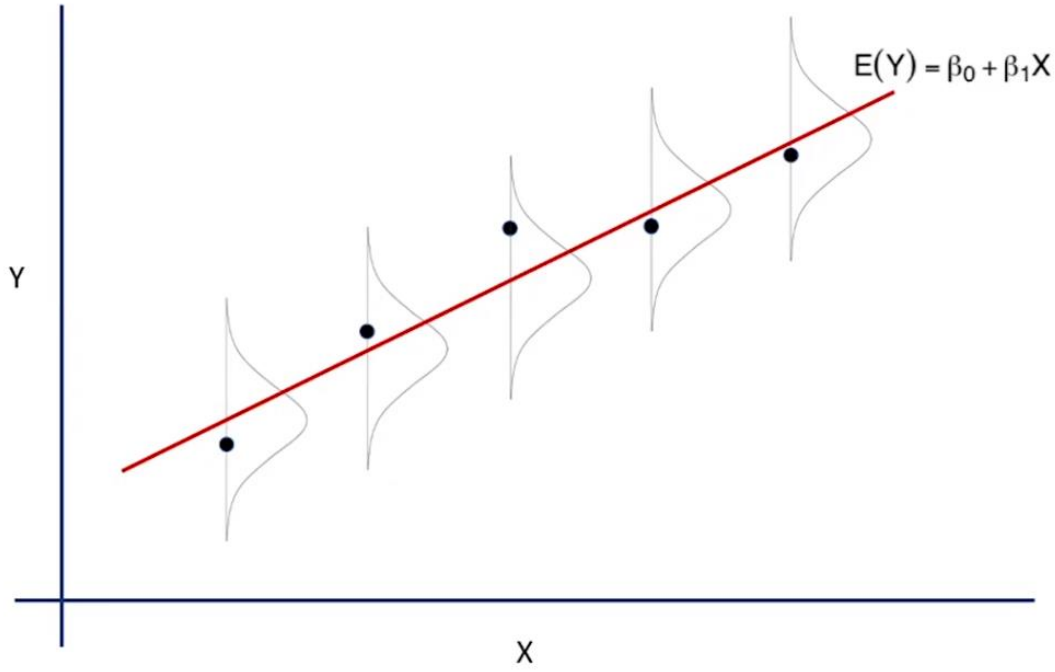
Besides random sample, there are 4 assumptions for the SLR model:

Linearity

Independent

Normality

Equal variance



Simple Linear Regression

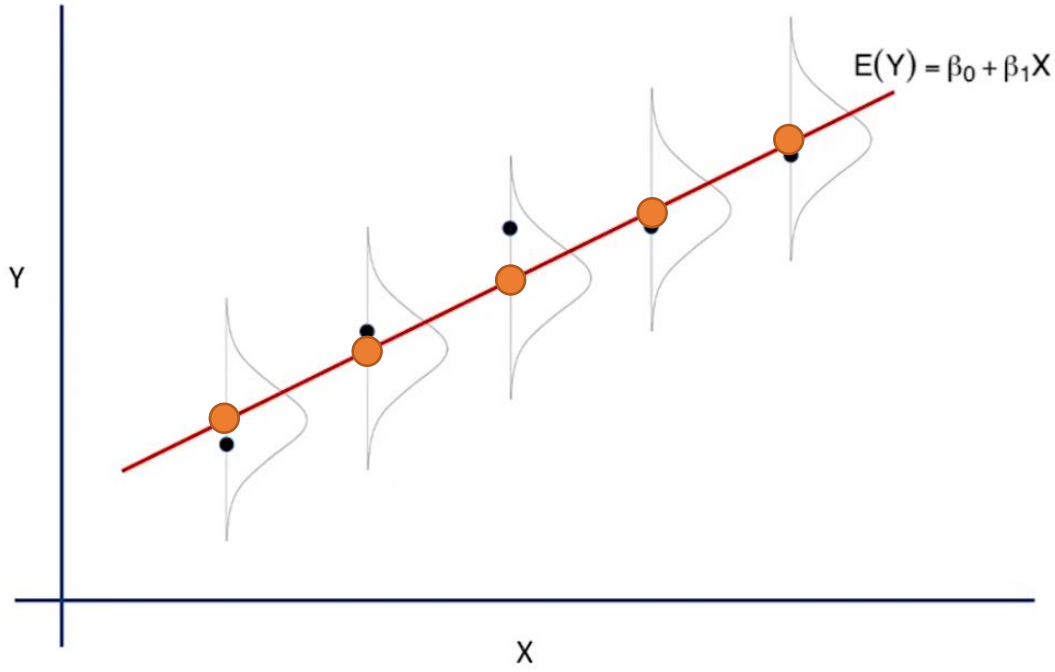
Besides random sample, there are 4 assumptions for the SLR model:

→ **L**inearity

Independent

Normality

Equal variance



Simple Linear Regression

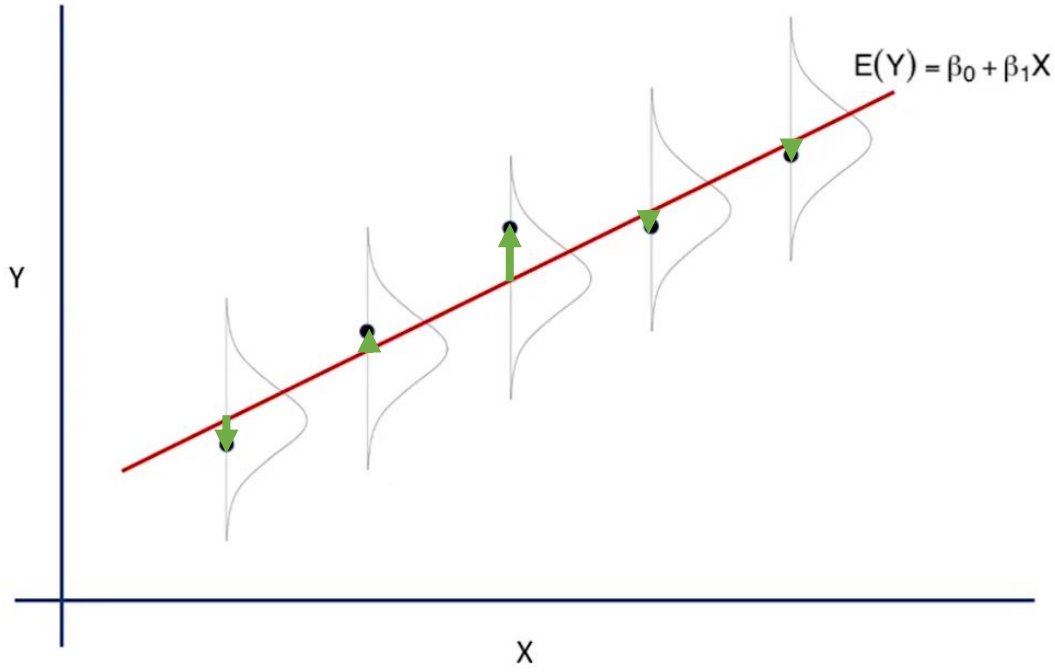
Besides random sample, there are 4 assumptions for the SLR model:

Linearity

→ **I**ndependent

Normality

Equal variance



Simple Linear Regression

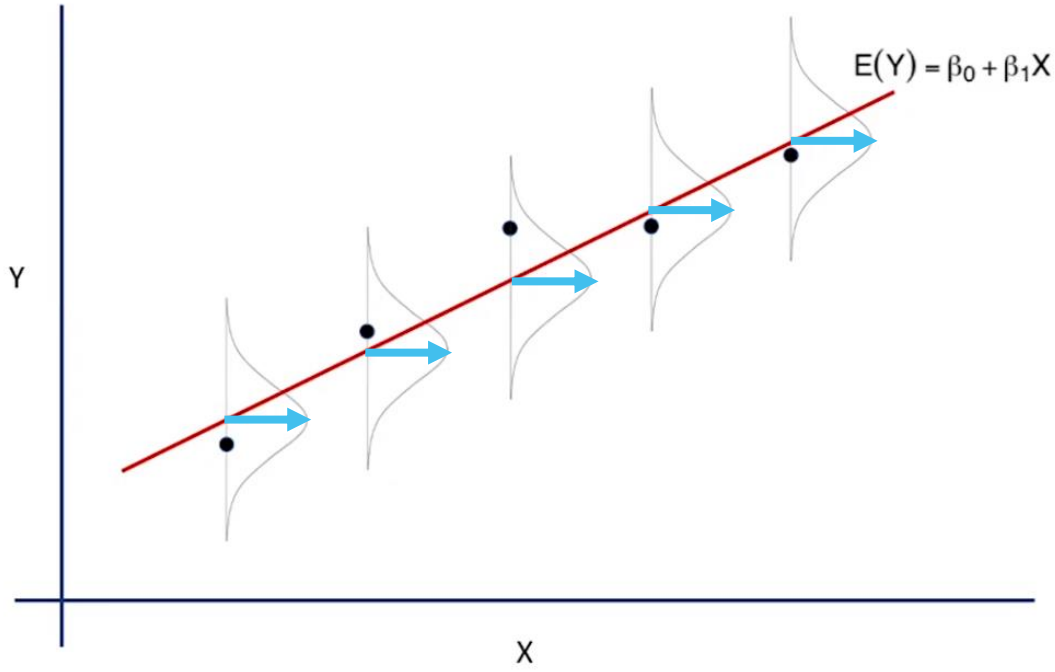
Besides random sample, there are 4 assumptions for the SLR model:

Linearity

Independent

→ **N**ormality

Equal variance



Simple Linear Regression

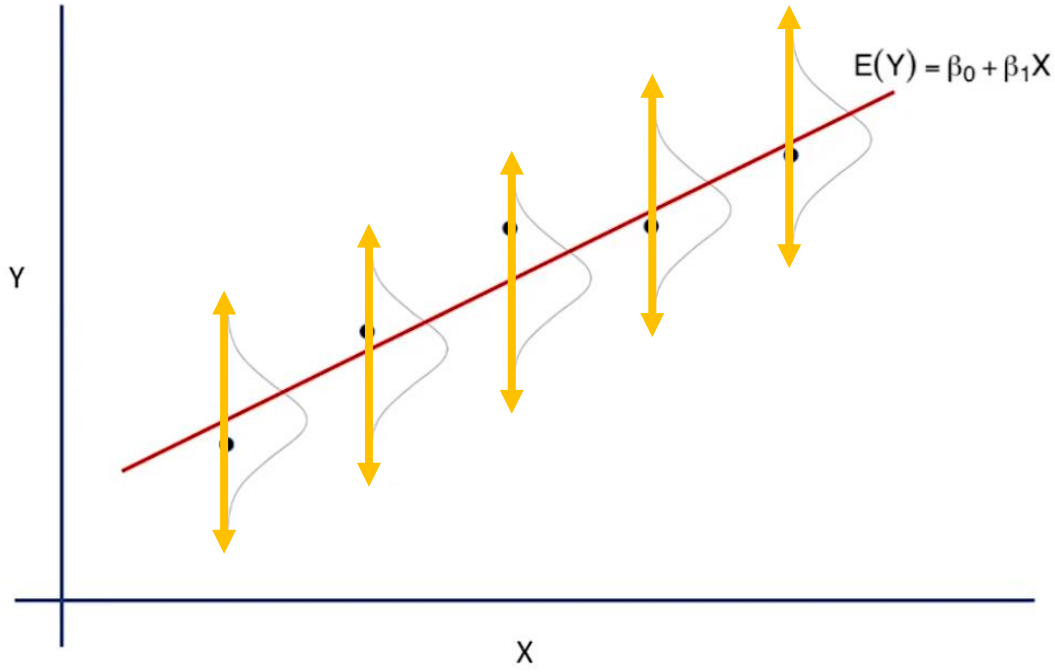
Besides random sample, there are 4 assumptions for the SLR model:

Linearity

Independent

Normality

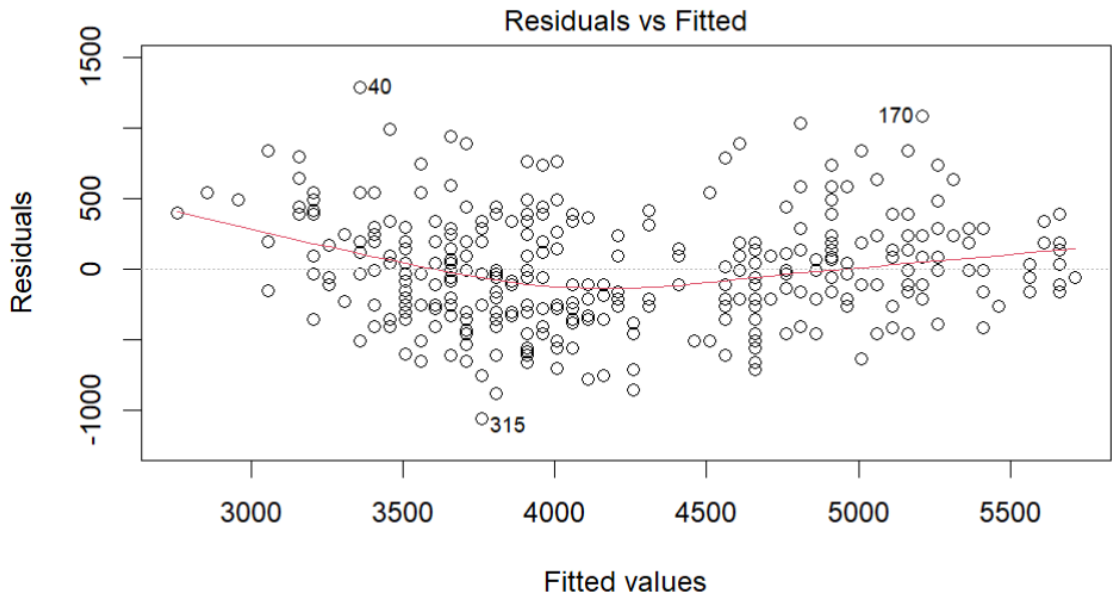
Equal variance



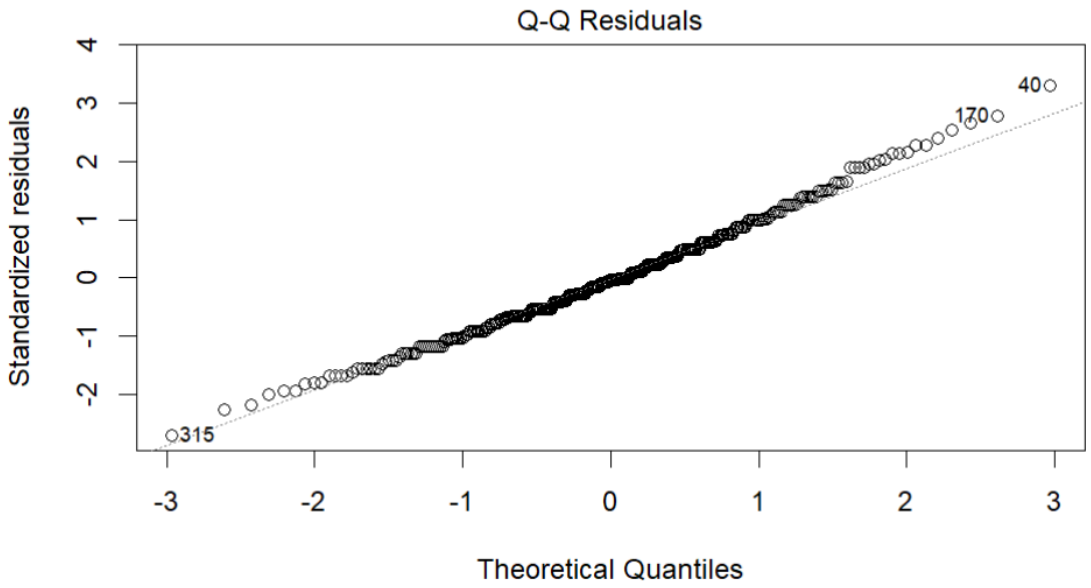
Simple Linear Regression

Checking for assumptions:

Residual plot



QQ-plot



Linearity, Equal variance

Normality

Simple Linear Regression

Checking for assumptions:

Residual plot

✓Nonlinearity

Look for clear nonlinear patterns in the residuals plot

✓ Unequal error variance

Look for unequal distances from y=0

✓ Outliers

Look for extreme values in the residuals plot

QQ-plot

✓Not normally

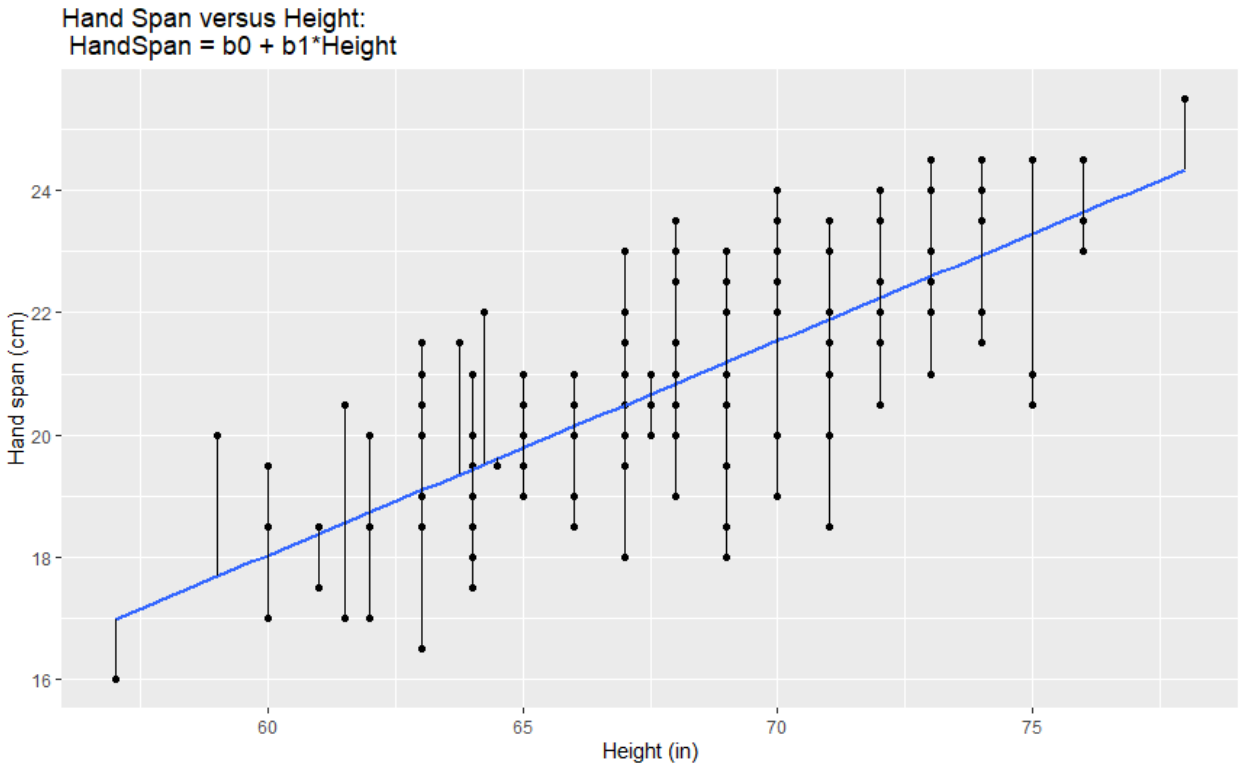
Look for deviation from the normality:

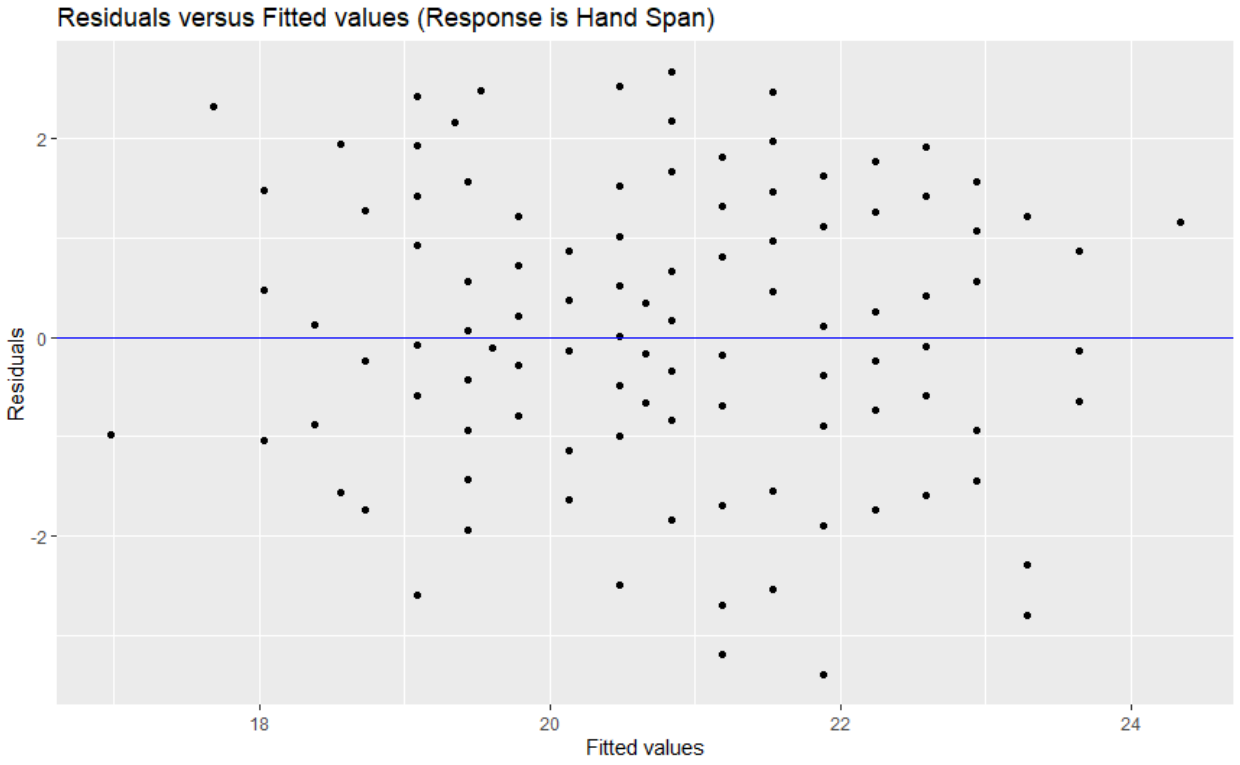
We compare the observed sample percentiles to the theoretical percentiles of the normal distribution. The normal probability plot should be approximately linear.

Simple Linear Regression

Checking for assumptions:

Residual plot

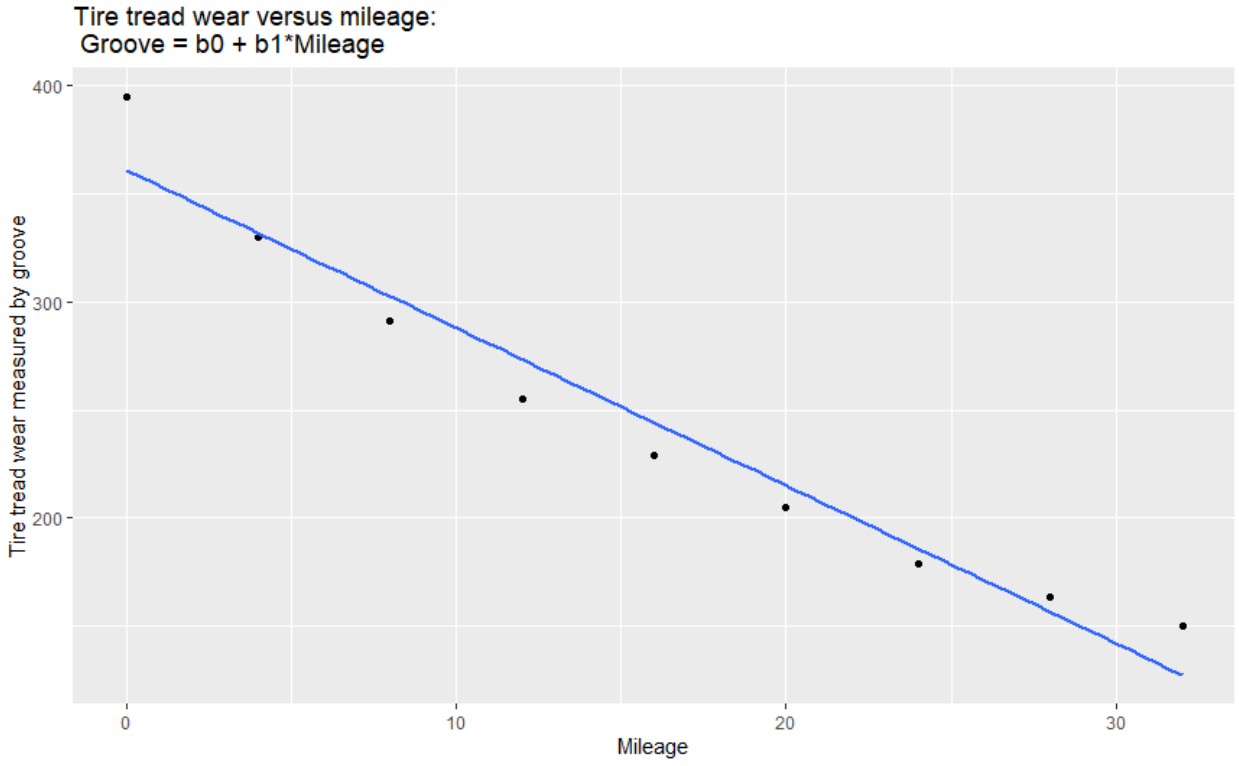


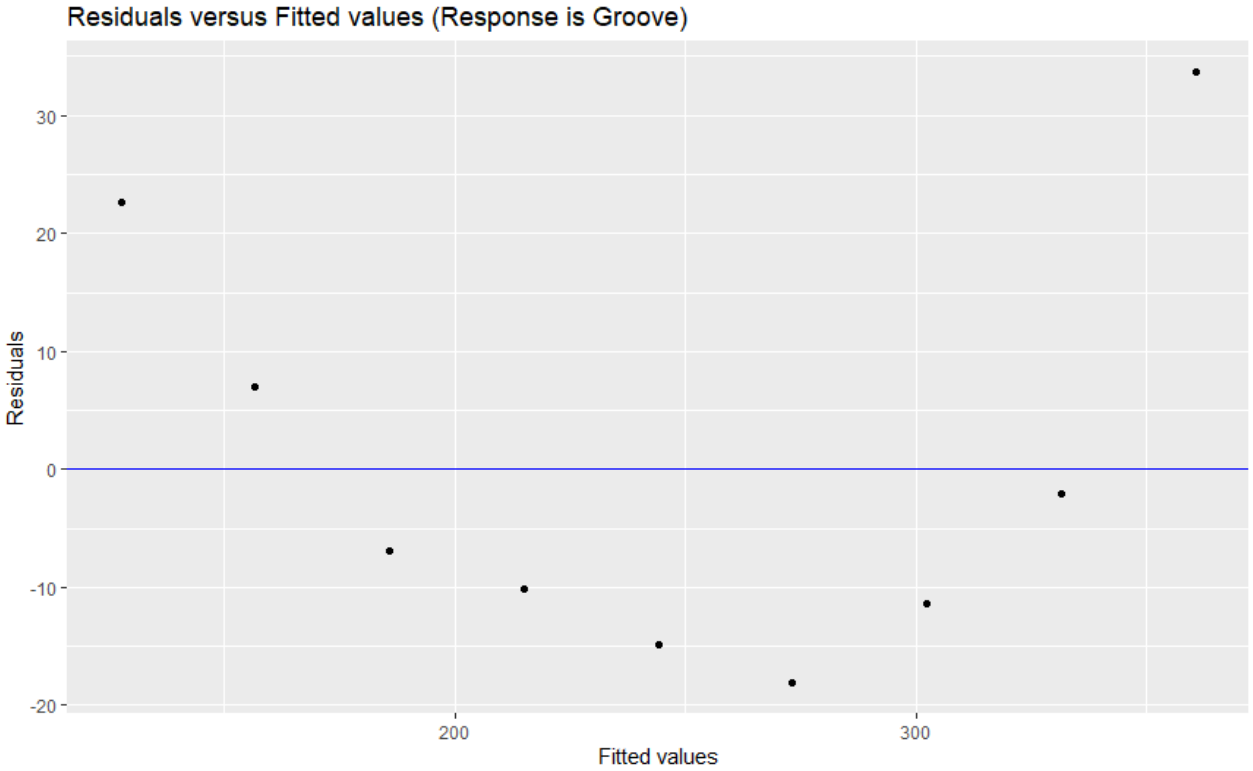


Simple Linear Regression

Checking for assumptions:

Residual plot

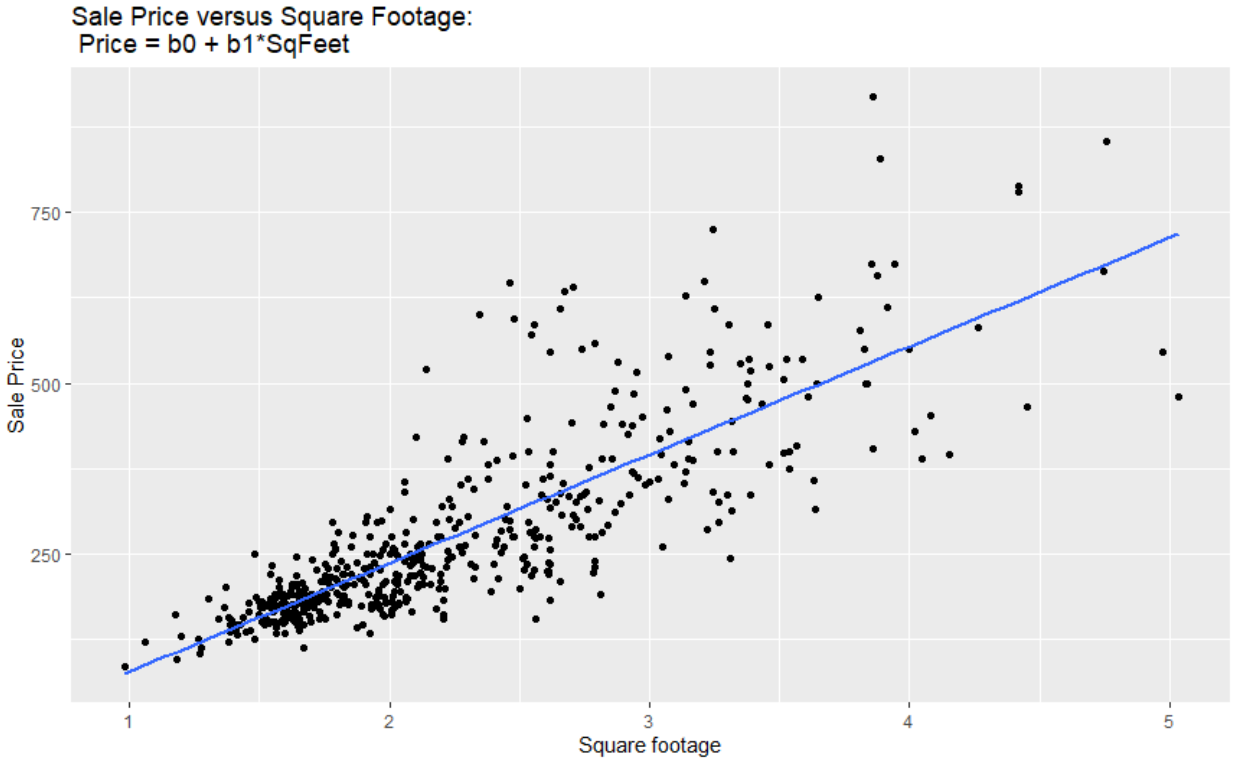


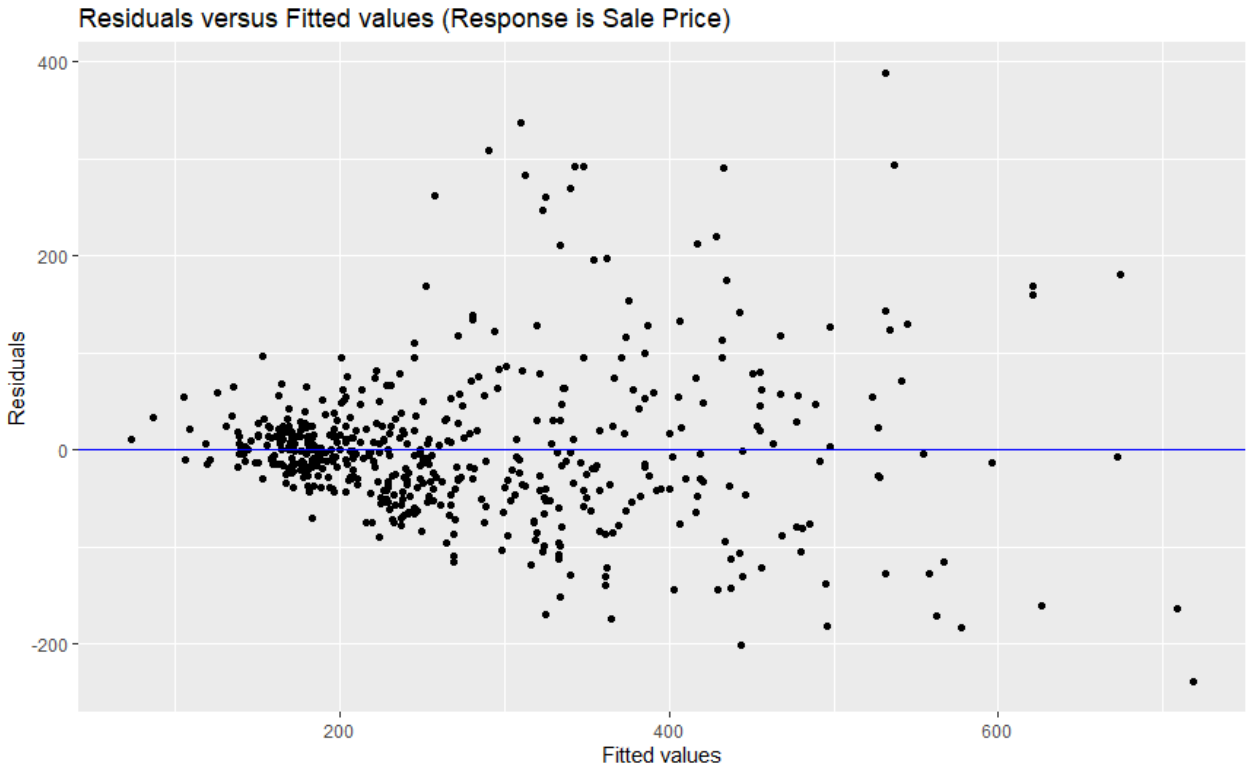


Simple Linear Regression

Checking for assumptions:

Residual plot

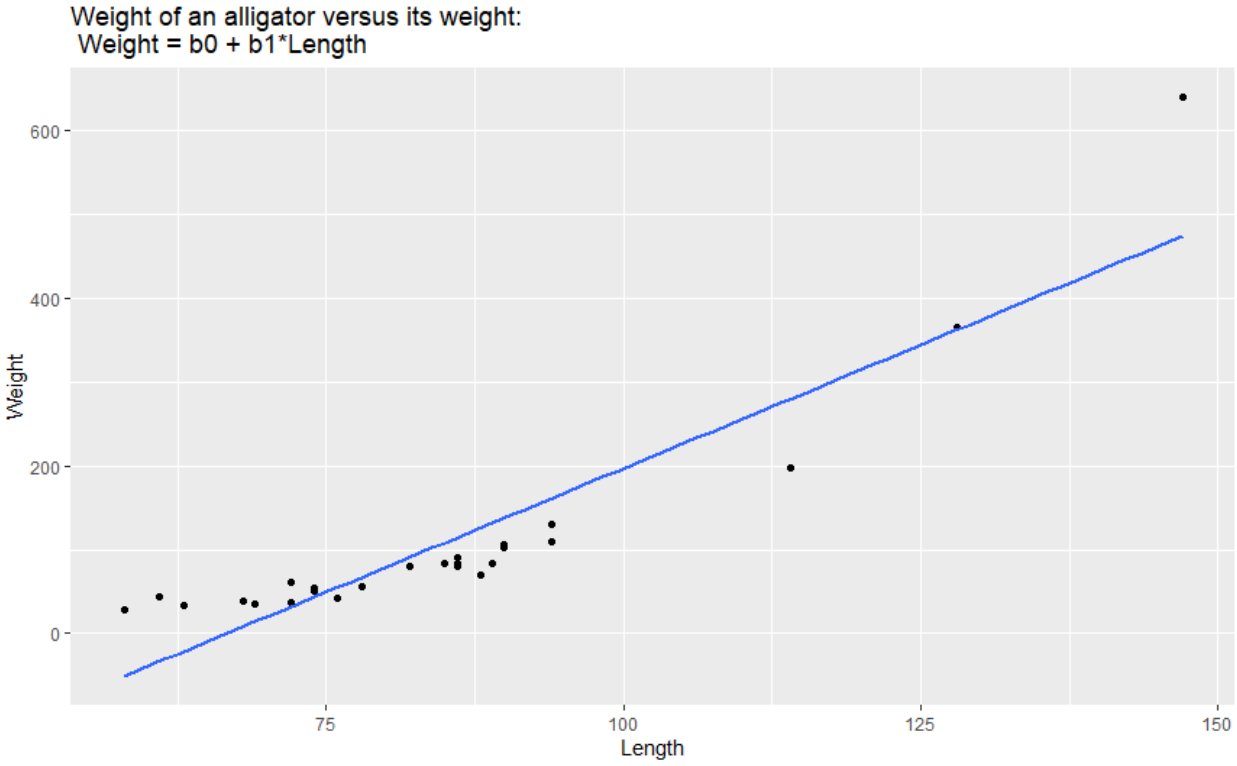




Simple Linear Regression

Checking for assumptions:

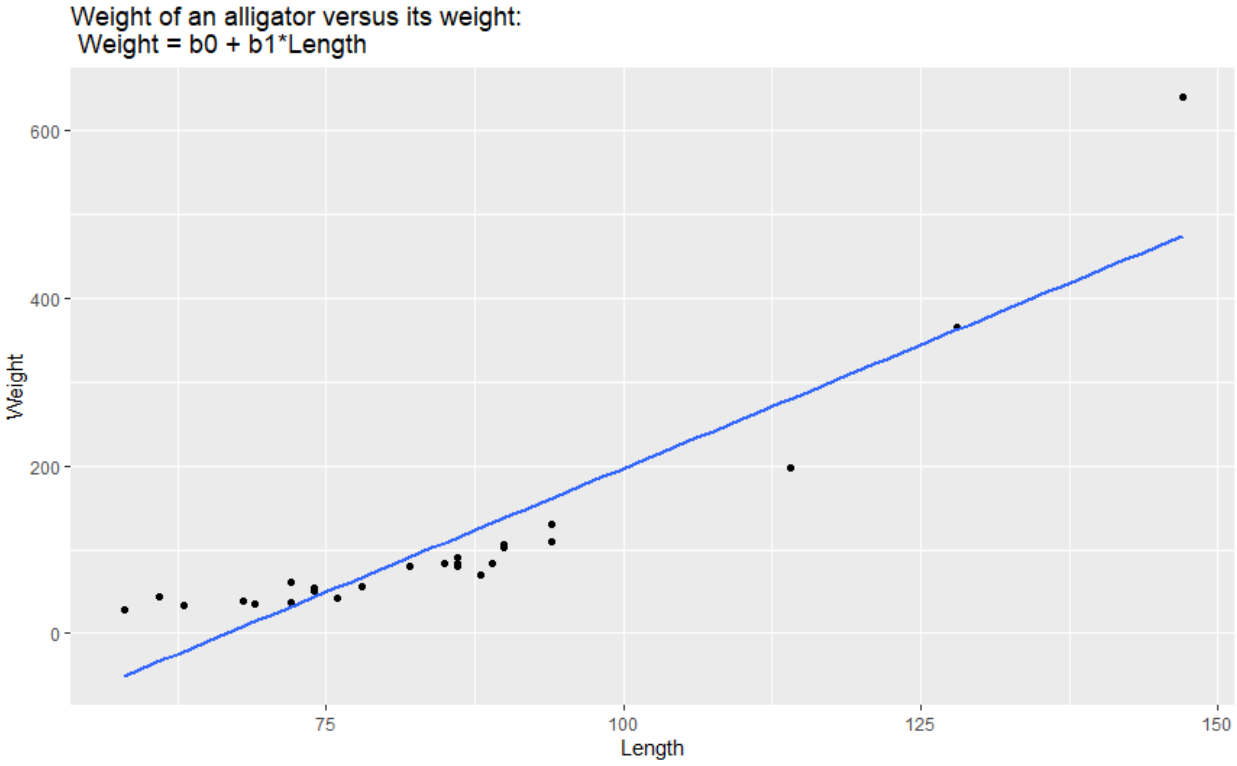
Residual plot



Simple Linear Regression

Checking for assumptions:

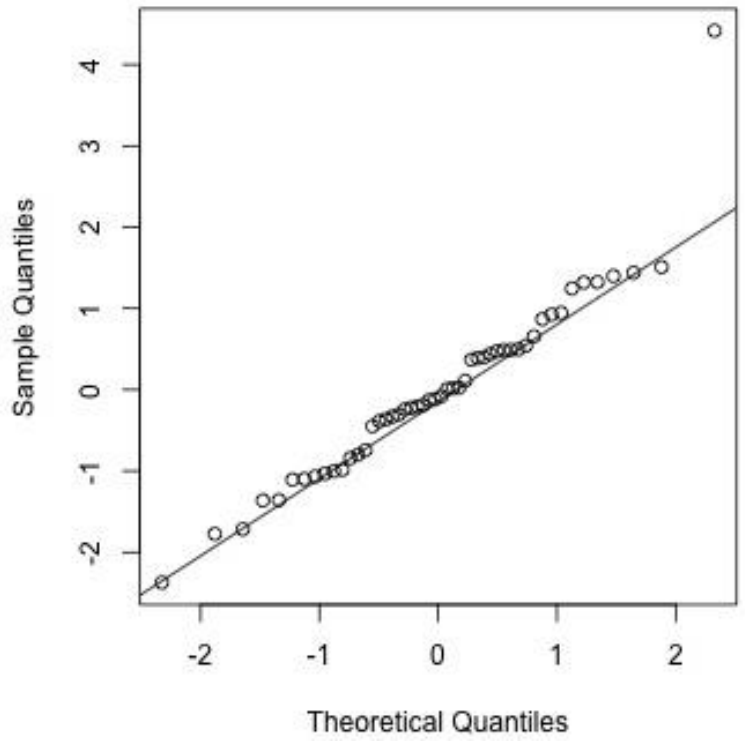
Residual plot



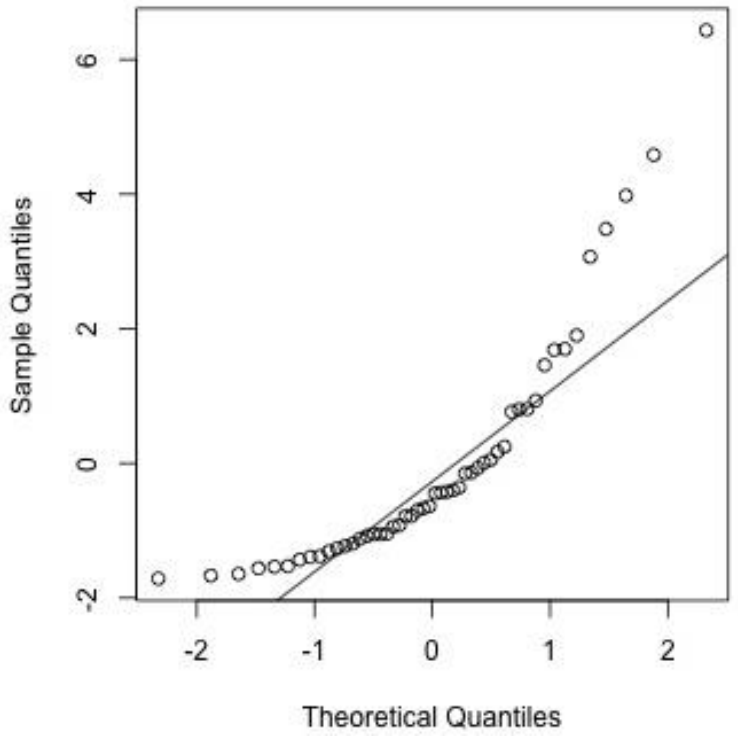
Simple Linear Regression

Checking for assumptions:

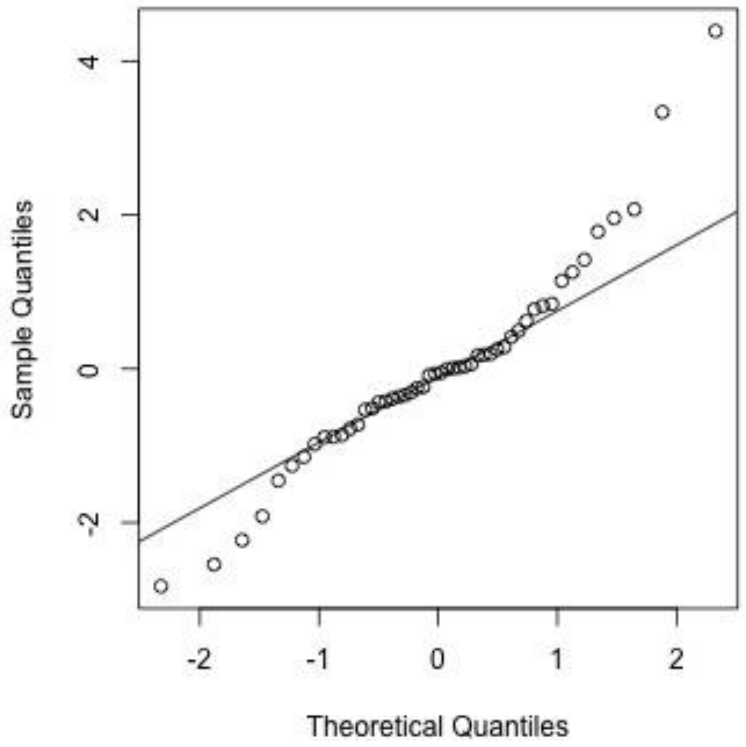
QQ-plot 1



QQ-plot 2



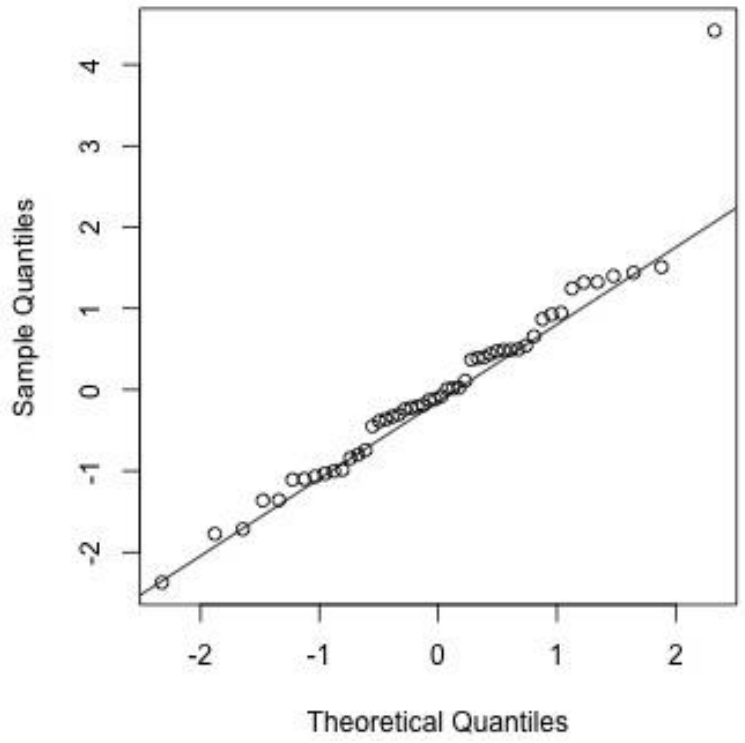
QQ-plot 3



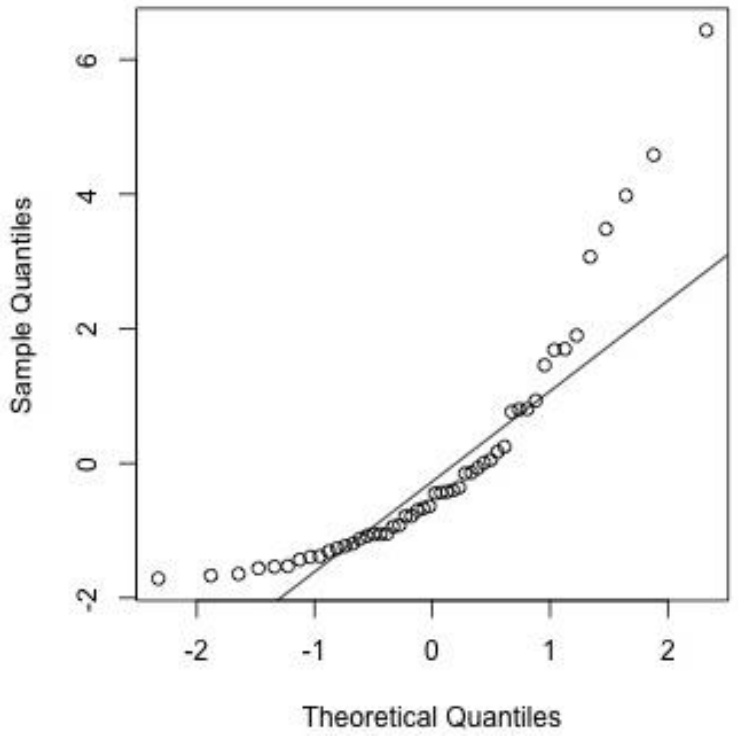
Simple Linear Regression

Checking for assumptions:

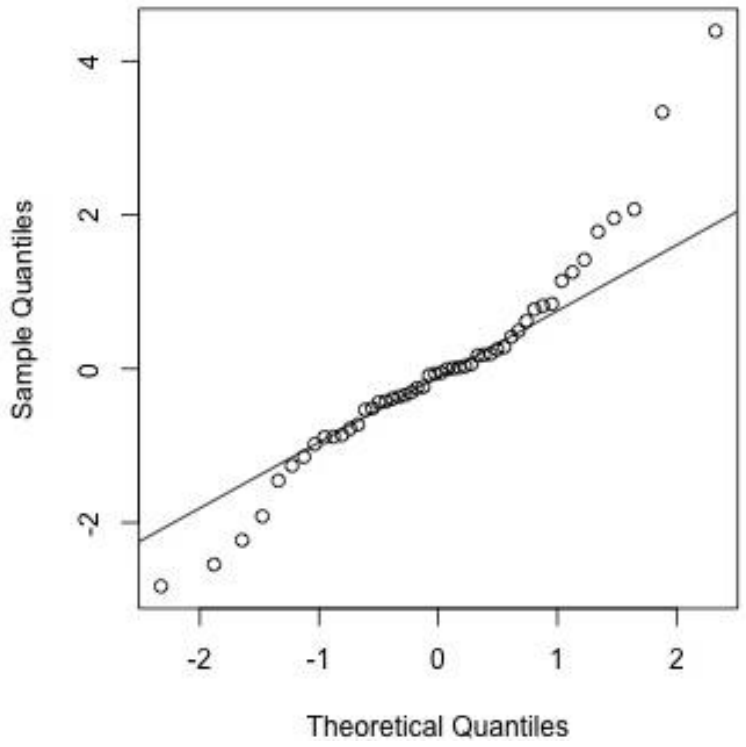
QQ-plot 1



QQ-plot 2



QQ-plot 3



USING R AND RSTUDIO

Simple Linear Regression

Comparing population slope to 0

1. State your hypotheses

$$H_0: \beta_1 = 0$$
$$H_A: \beta_1 \neq 0$$

Simple Linear Regression

Comparing population slope to 0

1. State your hypotheses

2. Calculate the test statistic $\boldsymbol{t}$ (based on sample data)

$$\boldsymbol{t} = \frac{\boldsymbol{b_1}}{\boldsymbol{SE_{b_1}}}$$

We will explain this standard error later

Simple Linear Regression

Comparing population slope to 0

1. State your hypotheses
2. Calculate the test statistic $\boldsymbol{t}$ (based on sample data)
3. Compare test statistic to null distribution (calculate ***p*-value**)
4. Make a conclusion in context, reporting the appropriate statistics (t, df, p-value).

$$\boldsymbol{df = n - 2}$$

Simple Linear Regression

When reporting results of a significant test, we should report a measure of the effect size with a **confidence interval** of the **population slope:**

$$\boldsymbol{b_1 \pm t^*_{df} \cdot SE_{b_1}}$$

We will explain this standard error later

Simple Linear Regression

Inference for predicted values

<u>Example</u>:

1. What is the mean hand span, μ, of **<u>all</u>** US adults?

   → Confidence interval for the mean response

2. What is the hand span, Y, of an **<u>individual</u>** US adult?

   → Prediction interval for a predicted response

Simple Linear Regression

To evaluate the performance of a linear regression model, we can consider:

- R^2: measures the percentage of variation in the response variable that can be explained by the predictor variable.
  The higher the R^2, the better the model is.

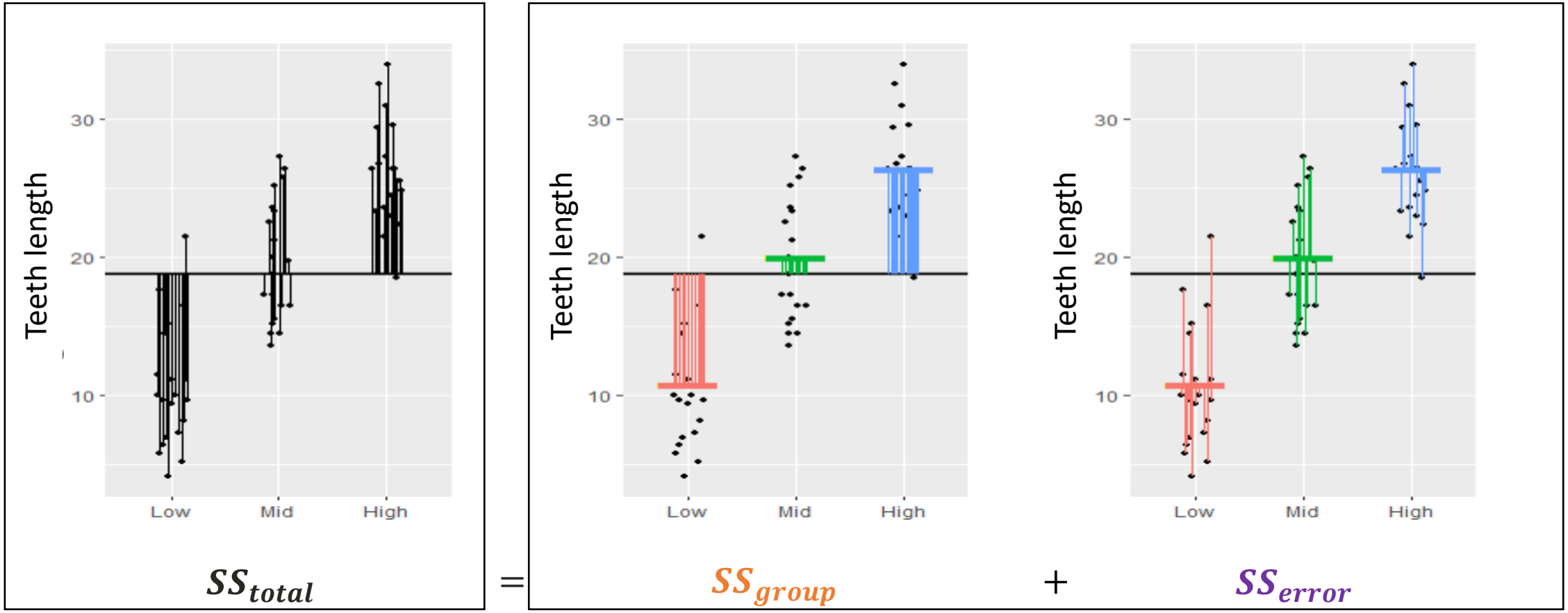
- Mean Squared Error (MSE): measures the average squared residuals.
  The lower the MSE, the better the model is.

- Root Mean Squared Error (RMSE): measures how far apart the predicted values are from the observed values in a dataset, on average.
  The lower the RMSE, the better the model is.

USING R AND RSTUDIO

Connection with ANOVA

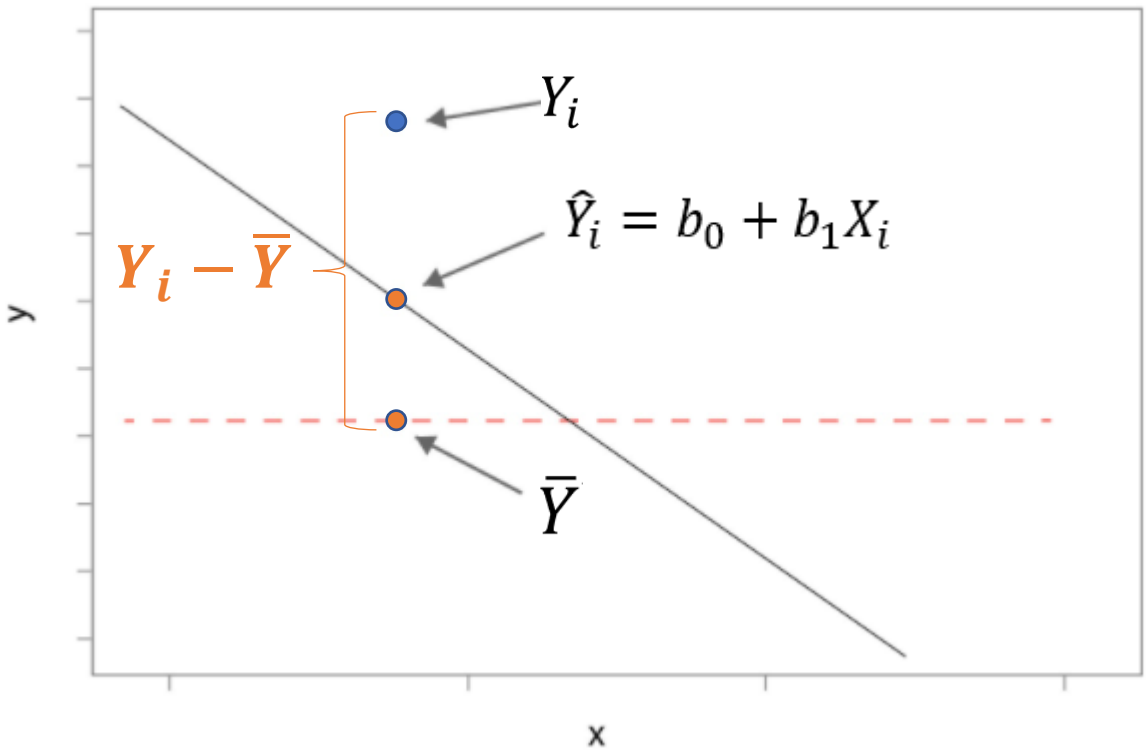
Decomposing variation in ANOVA



$$\sum (X_{ik} - \bar{X})^2 = \sum n_k(\bar{X}_k - \bar{X})^2 + \sum\sum (X_{ik} - \bar{X}_k)^2$$

Connection with ANOVA

Decomposing variation in Simple Linear Regression

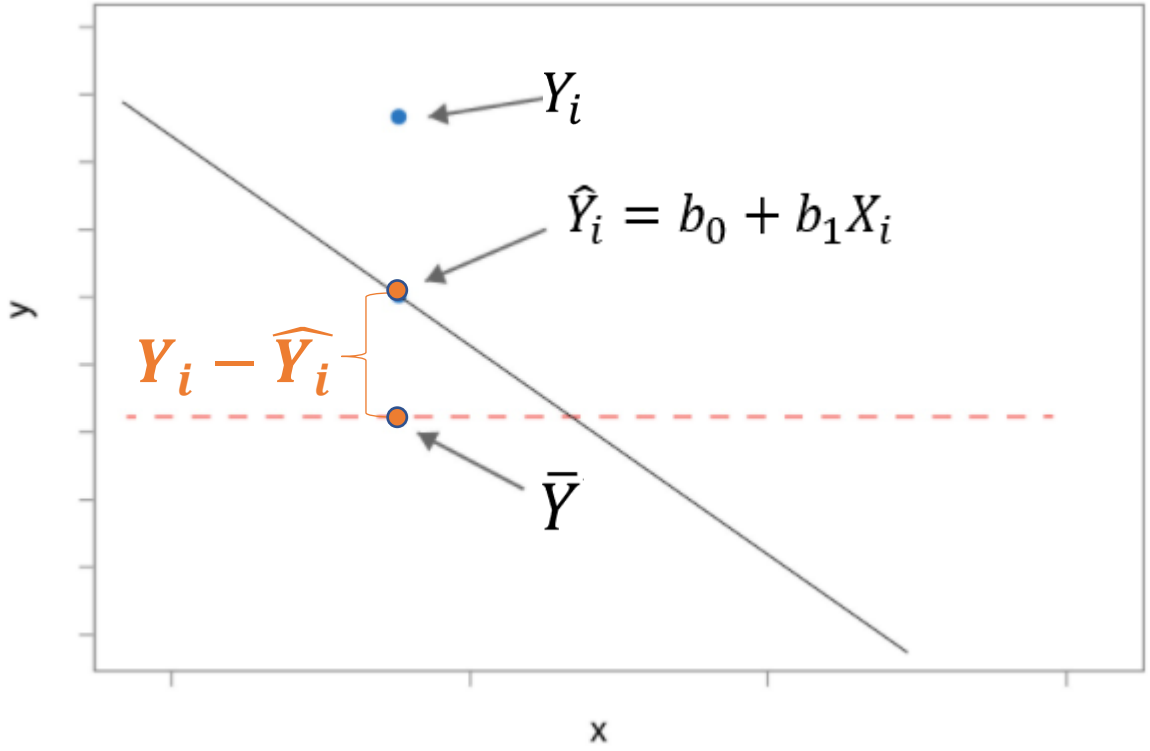


$$SS_{total}$$

$$\sum (Y_i - \bar{Y})^2$$

Connection with ANOVA

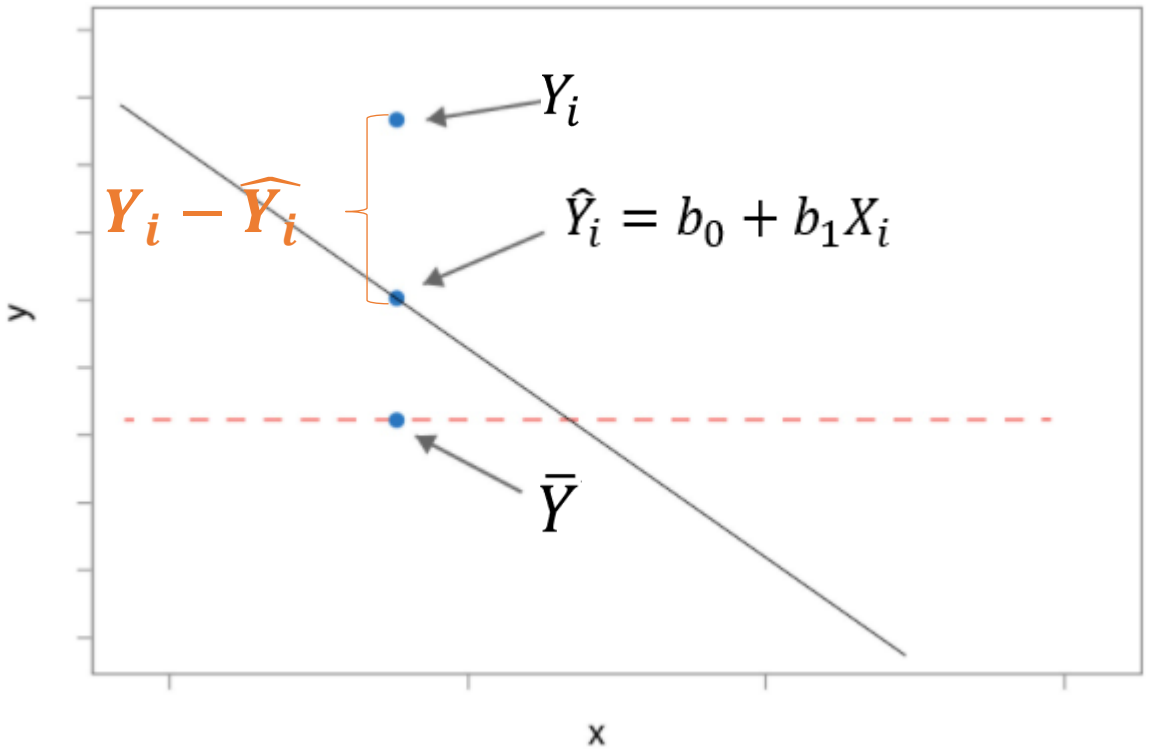
Decomposing variation in Simple Linear Regression



$$\overset{SS_{total}}{\sum (Y_i - \bar{Y})^2} = \overset{SS_{\text{regression}}}{\sum \left(\widehat{Y_i} - \bar{Y}\right)^2} + \overset{SS_{error}}{\sum \left(Y_i - \widehat{Y_i}\right)^2}$$

Connection with ANOVA

Decomposing variation in Simple Linear Regression



$$\overset{SS_{total}}{\sum (Y_i - \bar{Y})^2} = \overset{SS_{\text{regression}}}{\sum \left(\widehat{Y_i} - \bar{Y}\right)^2} + \overset{SS_{error}}{\sum \left(Y_i - \widehat{Y_i}\right)^2}$$

Connection with ANOVA

Inference with ANOVA

$$H_0: \beta_1 = 0$$

$$H_A: \beta_1 \neq 0$$

$$F = \frac{MS_{regression}}{MS_{error}} \quad \text{with} \quad MS_{regression} = \frac{SS_{regression}}{1}$$

$$MS_{error} = \frac{SS_{error}}{n-2}$$

USING R AND RSTUDIO

BREAK TIME

BACK AT ...

Multiple Linear Regression

- Population model:

$$\boldsymbol{Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \varepsilon}$$

- Estimated regression function:

$$\boldsymbol{\widehat{Y} = b_0 + b_1 X_1 + b_2 X_2 + \cdots + b_p X_p}$$

- Residuals:

$$\boldsymbol{e_i = Y_i - \widehat{Y}_i}$$

Multiple Linear Regression

Example: On a sample of 38 college students, the following variables were collected:

Y : Performance IQ scores (PIQ) of the revised Wechsler Adult Intelligence Scale

X_1: Brain size index based on MRI scans

X_2: Height (in inches)

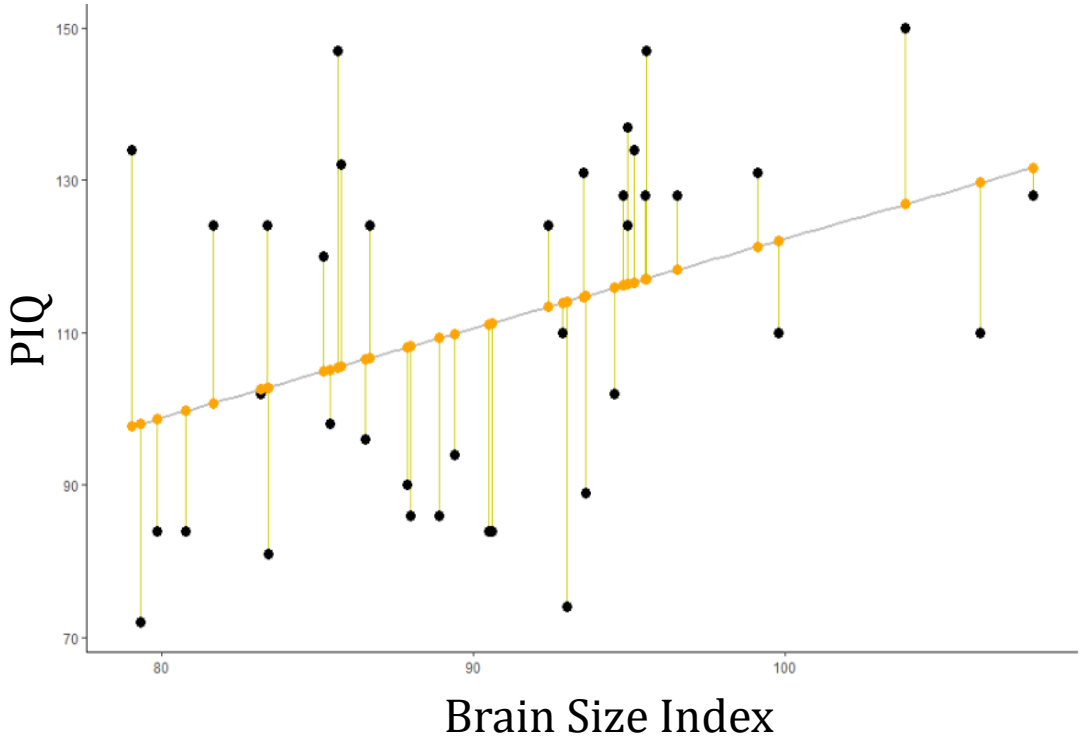
X_3: Weight (in pounds)

Can a student's intelligence be predicted by brain size and height?

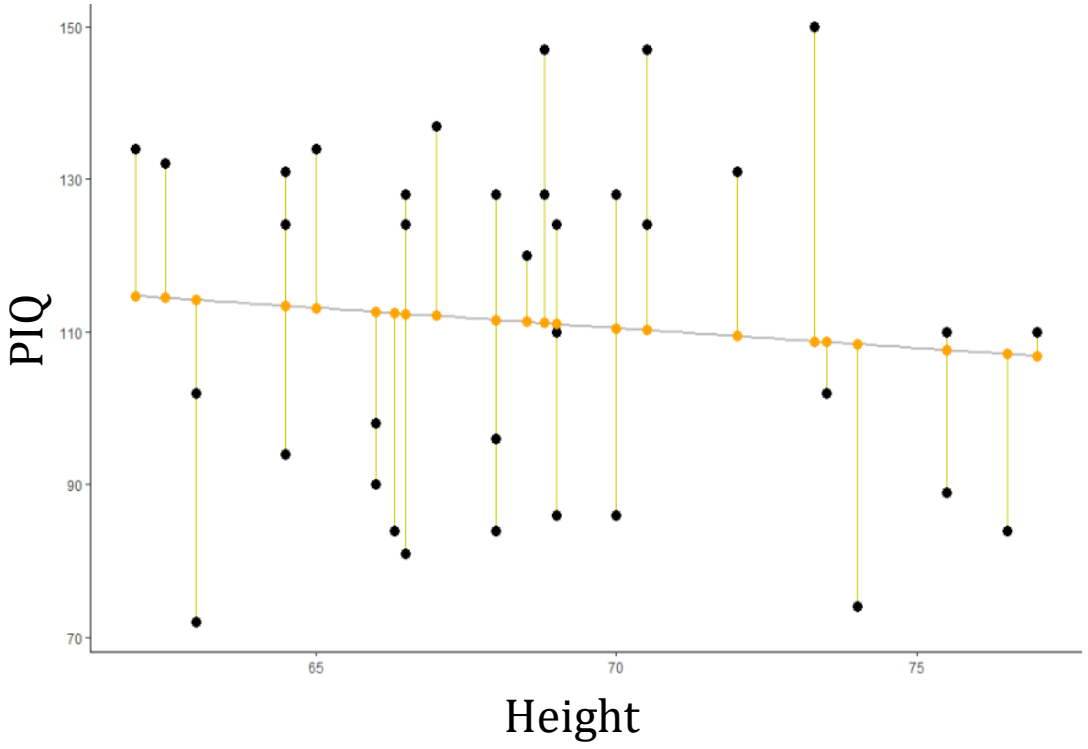
Multiple Linear Regression

Can a person's intelligence be predicted by brain size? by height?

$$\widehat{PIQ} = b_0 + b_1 * Brain$$



$$\widehat{PIQ} = b_0 + b_1 * Height$$

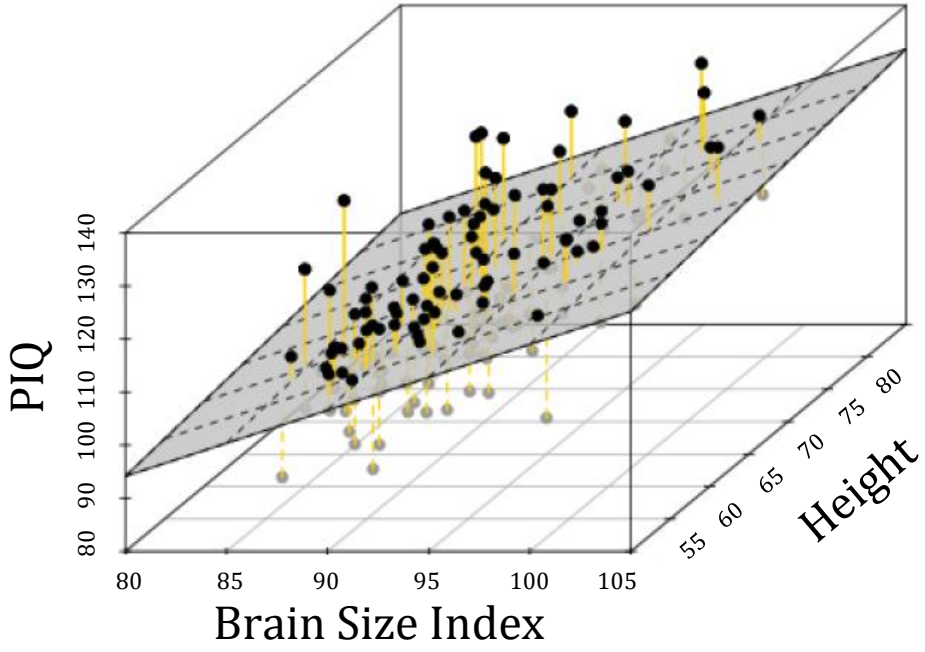


Multiple Linear Regression

Can a person's intelligence be predicted by brain size and height?

$$\widehat{PIQ} = b_0 + b_1 * Brain + b_2 * Height$$

For 2 predictors (brain size index and height), the equation technically yields a plane:



Multiple Linear Regression

Can a person's intelligence be predicted by brain size and height?

$$\widehat{PIQ} = b_0 + b_1 * Brain + b_2 * Height$$

Interpret the values of the slopes

As brain size index increases by 1, PIQ increases by 2.1 on average, <u>while holding height constant</u>.

As height increases by 1 inch, PIQ decreases by 2.7 on average, <u>while holding brain size index constant</u>.

```
> my_model <- lm(PIQ ~ Brain + Height, data = iqsize)
> summary(my_model)

Call:
lm(formula = PIQ ~ Brain + Height, data = iqsize)

Residuals:
    Min      1Q  Median      3Q     Max 
-32.750 -12.090  -3.841  14.174  51.690 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 111.2757    55.8673   1.992 0.054243 .  
Brain         2.0606     0.5466   3.770 0.000604 ***
Height       -2.7299     0.9932  -2.749 0.009399 ** 
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 19.51 on 35 degrees of freedom
Multiple R-squared:  0.2949,	Adjusted R-squared:  0.2546 
F-statistic: 7.321 on 2 and 35 DF,  p-value: 0.002208
```

Multiple Linear Regression

Comparing population slope to 0

for each slope

1. State your hypotheses

$$\boxed{\begin{aligned} H_0&: \beta_i = 0 \\ H_A&: \beta_i \neq 0 \end{aligned}}$$

2. Calculate the test statistic $\boldsymbol{t = \dfrac{b_i}{SE_{b_i}}}$

3. Compare test statistic to null distribution (calculate ***p*-value**)

3. Make a conclusion in context, reporting the appropriate statistics (t, df, p-value).

$$\boldsymbol{df = n - p - 1}$$

Multiple Linear Regression

Can a person's intelligence be predicted by brain size and height?

$$\widehat{PIQ} = b_0 + b_1 * Brain + b_2 * Height$$

Interpret the values of the slopes

Brain size index is a significant predictor of PIQ <u>while holding height constant</u> (t = 3.77, df = 35, p = 0.0006).

Height is a significant predictor of PIQ <u>while holding brain size index constant</u> (t = −2.75, df = 35, p = 0.009).

```
> my_model <- lm(PIQ ~ Brain + Height, data = iqsize)
> summary(my_model)

Call:
lm(formula = PIQ ~ Brain + Height, data = iqsize)

Residuals:
    Min      1Q  Median      3Q     Max 
-32.750 -12.090  -3.841  14.174  51.690 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 111.2757    55.8673   1.992 0.054243 .  
Brain         2.0606     0.5466   3.770 0.000604 ***
Height       -2.7299     0.9932  -2.749 0.009399 ** 
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 19.51 on 35 degrees of freedom
Multiple R-squared:  0.2949,	Adjusted R-squared:  0.2546 
F-statistic: 7.321 on 2 and 35 DF,  p-value: 0.002208
```

Multiple Linear Regression

Besides random sample, there are 4 assumptions for the SLR model:

Linearity: the mean response is a linear function of X_i

Independent observations: the errors, ϵ_i, are independent

Normality of residuals: the errors, ϵ_i, are normally distributed

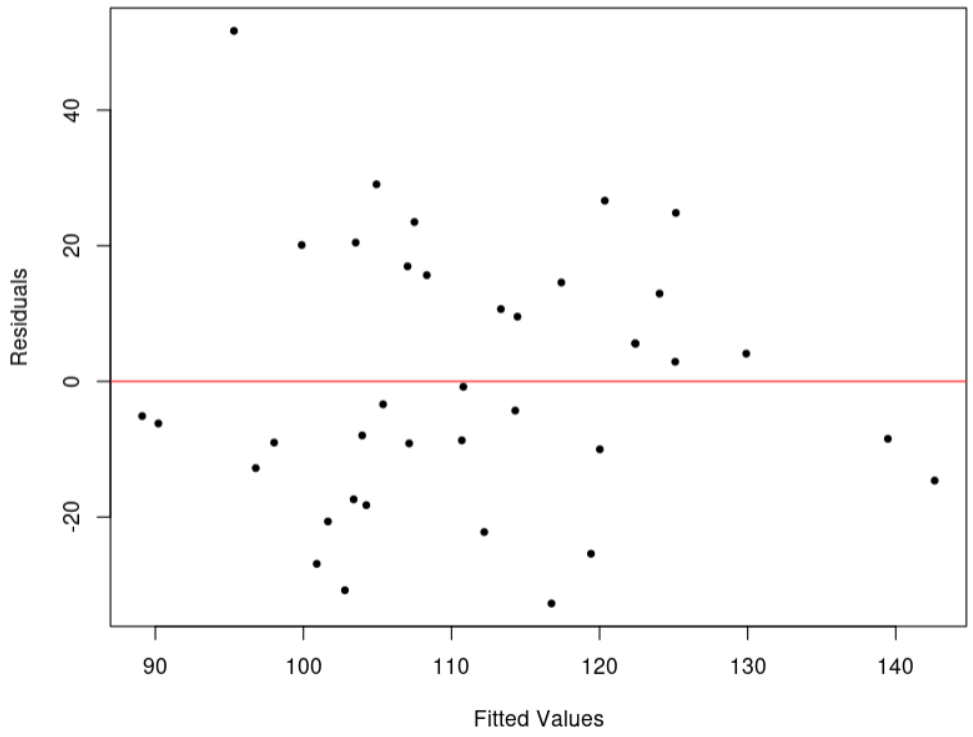
Equal variances: the errors, ϵ_i, have equal variances (σ^2)

Same assumptions than for the Simple Linear Regression model!

Multiple Linear Regression

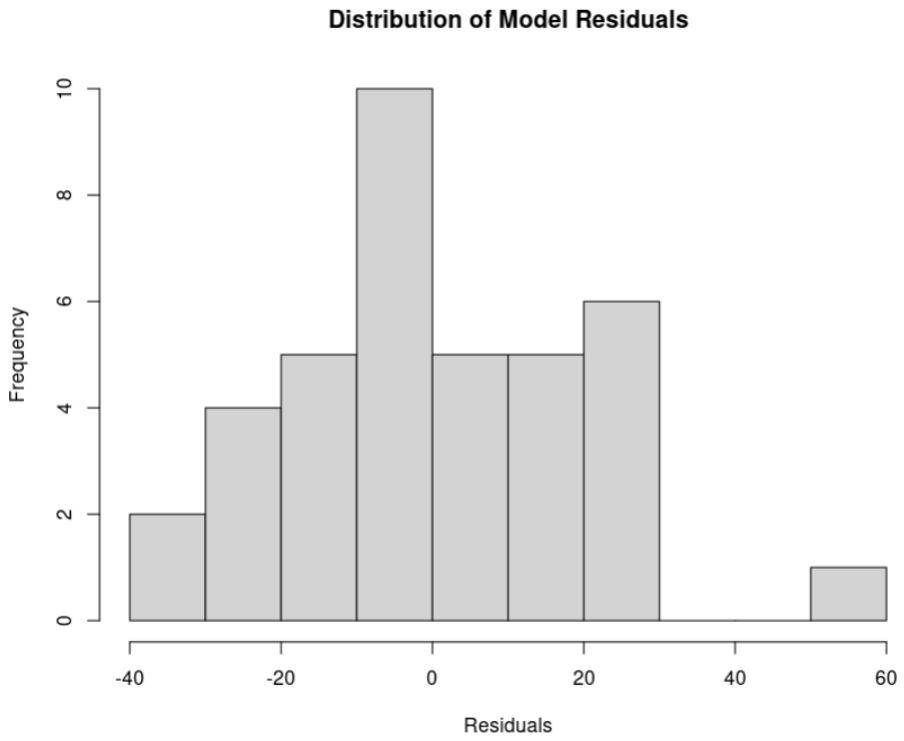
Are the assumptions met?

Residual plot



- Linearity
- Equal variance

Histogram of the residuals



- Normality

```
Shapiro-Wilk normality test

data:  my_model$residuals
W = 0.976, p-value = 0.5764
```

Multiple Linear Regression

Can a person's intelligence be predicted by brain size and height?

$$\widehat{PIQ} = b_0 + b_1 * Brain + b_2 * Height$$

Anything else we should check?

```
> my_model <- lm(PIQ ~ Brain + Height, data = iqsize)
> summary(my_model)

Call:
lm(formula = PIQ ~ Brain + Height, data = iqsize)

Residuals:
    Min      1Q  Median      3Q     Max 
-32.750 -12.090  -3.841  14.174  51.690 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 111.2757    55.8673   1.992 0.054243 .  
Brain         2.0606     0.5466   3.770 0.000604 ***
Height       -2.7299     0.9932  -2.749 0.009399 ** 
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 19.51 on 35 degrees of freedom
Multiple R-squared:  0.2949,	Adjusted R-squared:  0.2546 
F-statistic: 7.321 on 2 and 35 DF,  p-value: 0.002208
```

Multiple Linear Regression

Can a person's intelligence be predicted by brain size and height?

$$\widehat{PIQ} = b_0 + b_1 * Brain + b_2 * Height$$

```
> my_model <- lm(PIQ ~ Brain + Height, data = iqsize)
> summary(my_model)

Call:
lm(formula = PIQ ~ Brain + Height, data = iqsize)

Residuals:
    Min      1Q  Median      3Q     Max 
-32.750 -12.090  -3.841  14.174  51.690 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 111.2757    55.8673   1.992 0.054243 .  
Brain         2.0606     0.5466   3.770 0.000604 ***
Height       -2.7299     0.9932  -2.749 0.009399 ** 
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 19.51 on 35 degrees of freedom
Multiple R-squared:  0.2949,	Adjusted R-squared:  0.2546 
F-statistic: 7.321 on 2 and 35 DF,  p-value: 0.002208
```

What could be a potential issue if we had many predictors? What would happen to R^2?

Multiple Linear Regression

Can a person's intelligence be predicted by brain size and height?

Model with Brain size, Height

```
> my_model <- lm(PIQ ~ Brain + Height, data = iqsize)
> summary(my_model)

Call:
lm(formula = PIQ ~ Brain + Height, data = iqsize)

Residuals:
    Min      1Q  Median      3Q     Max 
-32.750 -12.090  -3.841  14.174  51.690 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 111.2757    55.8673   1.992 0.054243 .  
Brain         2.0606     0.5466   3.770 0.000604 ***
Height       -2.7299     0.9932  -2.749 0.009399 ** 
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 19.51 on 35 degrees of freedom
Multiple R-squared:  0.2949,	Adjusted R-squared:  0.2546 
F-statistic: 7.321 on 2 and 35 DF,  p-value: 0.002208
```

Which model is better?

Model with Brain size, Height, Weight

```
> my_model <- lm(PIQ ~ Brain + Height + Weight, data = iqsize)
> summary(my_model)

Call:
lm(formula = PIQ ~ Brain + Height + Weight, data = iqsize)

Residuals:
   Min     1Q Median     3Q    Max 
-32.74 -12.09  -3.84  14.17  51.69 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept)  1.114e+02  6.297e+01   1.768 0.085979 .  
Brain        2.060e+00  5.634e-01   3.657 0.000856 ***
Height      -2.732e+00  1.229e+00  -2.222 0.033034 *  
Weight       5.599e-04  1.971e-01   0.003 0.997750    
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 19.79 on 34 degrees of freedom
Multiple R-squared:  0.2949,	Adjusted R-squared:  0.2327 
F-statistic: 4.741 on 3 and 34 DF,  p-value: 0.007215
```

USING R AND RSTUDIO

Multiple Linear Regression

On a random sample of 32 births, Daniel (1999) collected these three variables:

Y : Birth weight of baby (Wgt) in grams

X_1: Length of gestation (Gest) in weeks

X_2: Smoking status of mother (Smoke = 1 if yes or Smoke = 0 if no)

Does smoking during pregnancy affect birth weight, while accounting for the length of gestation?

Multiple Linear Regression

Does smoking during pregnancy affect birth weight, while accounting for the length of gestation?

$$\widehat{Wgt} = -2389.573 + 143.1\,Gest - 244.544\,Smoke$$

Interpret the values of the slopes

As the length of gestation increases by 1 week, birth weight increases by 143.1 grams on average, while holding smoking status constant.

Birth weight decreases by 244.544 grams on average for babies born from smoking mothers compared to nonsmoking, while holding length of gestation constant.

```
> reg <- lm(Wgt~Gest+Smoke,birthsmokers)
> summary(reg)

Call:
lm(formula = Wgt ~ Gest + Smoke, data = birthsmokers)

Residuals:
     Min       1Q   Median       3Q      Max 
-223.693  -92.063   -9.365   79.663  197.507 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) -2389.573    349.206  -6.843 1.63e-07 ***
Gest          143.100      9.128  15.677 1.07e-15 ***
Smoke        -244.544     41.982  -5.825 2.58e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 115.5 on 29 degrees of freedom
Multiple R-squared:  0.8964,	Adjusted R-squared:  0.8892 
F-statistic: 125.4 on 2 and 29 DF,  p-value: 5.289e-15
```

Multiple Linear Regression

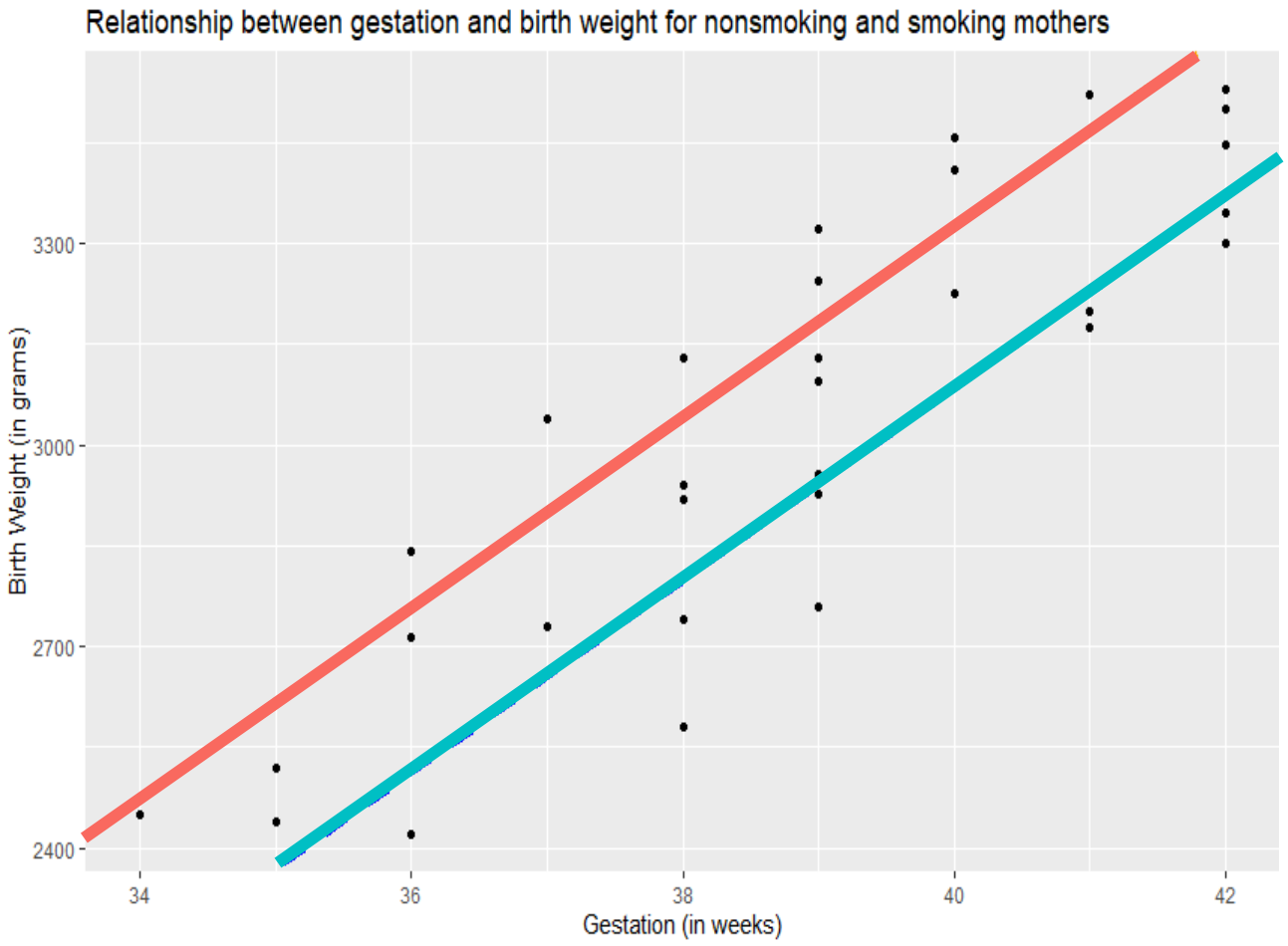
Does smoking during pregnancy affect birth weight, while accounting for the length of gestation?

$$\widehat{Wgt} = -2389.573 + 143.1\,Gest - 244.544\,Smoke$$

Interpret the values of the slopes

As the length of gestation increases by 1 week, birth weight increases by 143.1 grams on average, while holding smoking status constant.

Birth weight decreases by 244.544 grams on average for babies born from smoking mothers compared to nonsmoking, while holding length of gestation constant.



Multiple Linear Regression

Interactions

- Does the effect of one predictor on the response depends on the levels of another predictor?
- Model with interaction effect for two predictors:

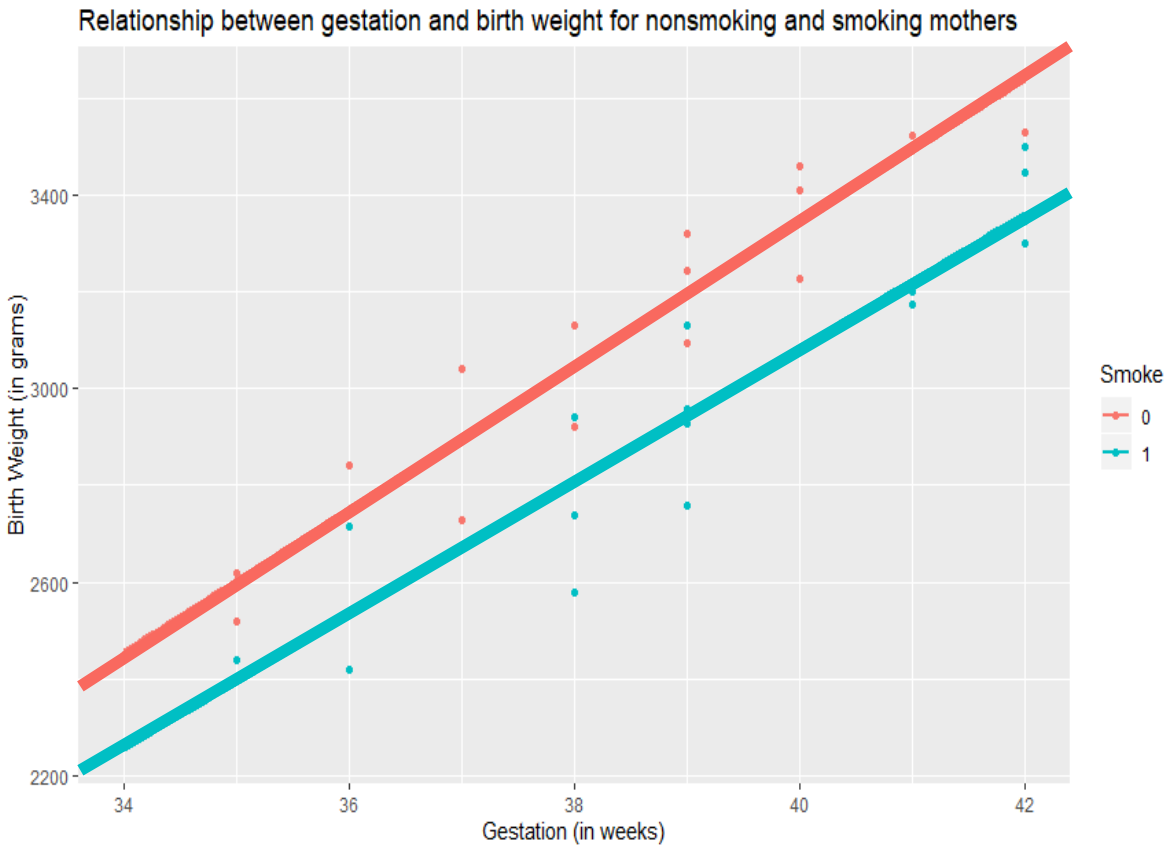
$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_{1*2}(X_1 * X_2) + \varepsilon$$

Main effects: effects between each predictor and the response

Interaction effect: a predictor might affect the response differently, depending on another predictor

Multiple Linear Regression

Interactions



Does the effect of the gestation length on birth weight depend on smoking status?

OR

Does the effect of smoking on birth weight depend on the length of gestation?

Multiple Linear Regression

Is there an interaction between the length of gestation and smoking status to predict birth weight?

$$\widehat{Wgt} = -2546.1 + 147.2\,Gest + 71.6\,Smoke - 8.2\,Gest * Smoke$$

Interpret the slopes of the main effects.

As the length of gestation increases by 1 week, birth weight increases by 147.2 grams for <u>nonsmoking</u> mothers.

Birth weight is 71.6 grams more on average for babies born from smoking mothers compared to nonsmoking mothers for a length gestation of 0.

Does not make sense to interpret

```
> reg <- lm(Wgt~Gest*Smoke, data = birthsmokers)
> summary(reg)

Call:
lm(formula = Wgt ~ Gest * Smoke, data = birthsmokers)

Residuals:
     Min       1Q   Median       3Q      Max 
-228.528  -89.560    0.273   83.629  184.529 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) -2546.138    501.067  -5.081 2.22e-05 ***
Gest          147.207     13.120  11.220 7.15e-12 ***
Smoke          71.574    716.950   0.100    0.921    
Gest:Smoke     -8.178     18.515  -0.442    0.662    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 117.2 on 28 degrees of freedom
Multiple R-squared:  0.8971,	Adjusted R-squared:  0.8861 
F-statistic: 81.37 on 3 and 28 DF,  p-value: 6.144e-14
```

Multiple Linear Regression

Interactions

- Centering numeric predictors to better interpret slopes:

$$\boldsymbol{Variable\_c = Variable - mean(Variable)}$$

How do we interpret $\boldsymbol{Variable\_c}$ = 0?

Multiple Linear Regression

Is there an interaction between the length of gestation and smoking status to predict birth weight?

$$\widehat{Wgt} = 31144.3 + 147.2\,Gest - 244.6\,Smoke - 8.2\,Gest * Smoke$$

Interpret the slopes of the main effects.

As the length of gestation increases by 1 week, birth weight increases by 147.2 grams for <u>nonsmoking</u> mothers.

Birth weight is 244.6 grams less on average for babies born from smoking mothers compared to nonsmoking mothers for an average gestation length.

```
> Gest_c <- birthsmokers$Gest-mean(birthsmokers$Gest)
> reg <- lm(Wgt~Gest_c*Smoke, data = birthsmokers)
> summary(reg)

Call:
lm(formula = Wgt ~ Gest_c * Smoke, data = birthsmokers)

Residuals:
     Min       1Q   Median       3Q      Max 
-228.528  -89.560    0.273   83.629  184.529 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept)   3144.329     30.110 104.429  < 2e-16 ***
Gest_c         147.207     13.120  11.220 7.15e-12 ***
Smoke         -244.563     42.577  -5.744 3.65e-06 ***
Gest_c:Smoke    -8.178     18.515  -0.442    0.662    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 117.2 on 28 degrees of freedom
Multiple R-squared:  0.8971,	Adjusted R-squared:  0.8861 
F-statistic: 81.37 on 3 and 28 DF,  p-value: 6.144e-14
```

Multiple Linear Regression

Is there an interaction between the length of gestation and smoking status to predict birth weight?

$$\widehat{Wgt} = 31144.3 + 147.2\,Gest - 244.6\,Smoke - 8.2\,Gest * Smoke$$

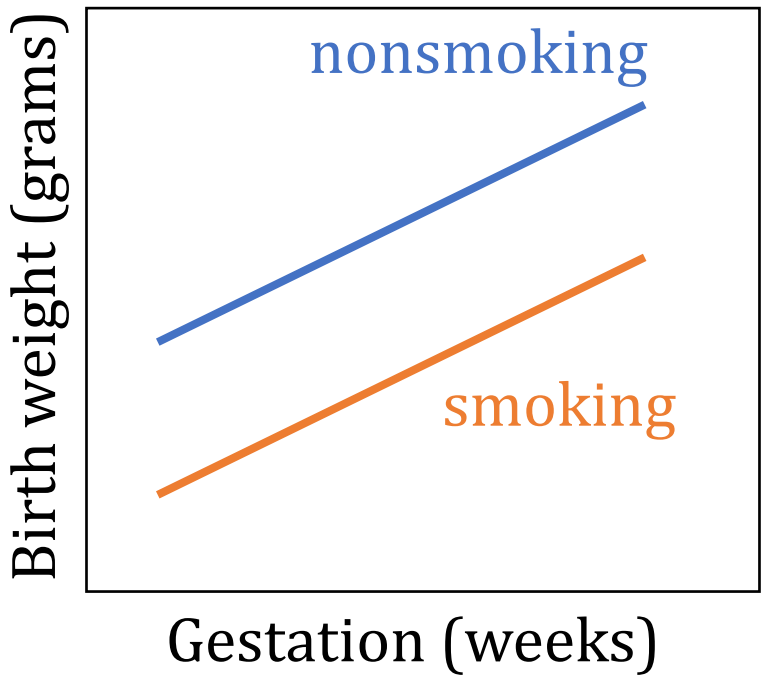
Interpret the slope for interaction effect.

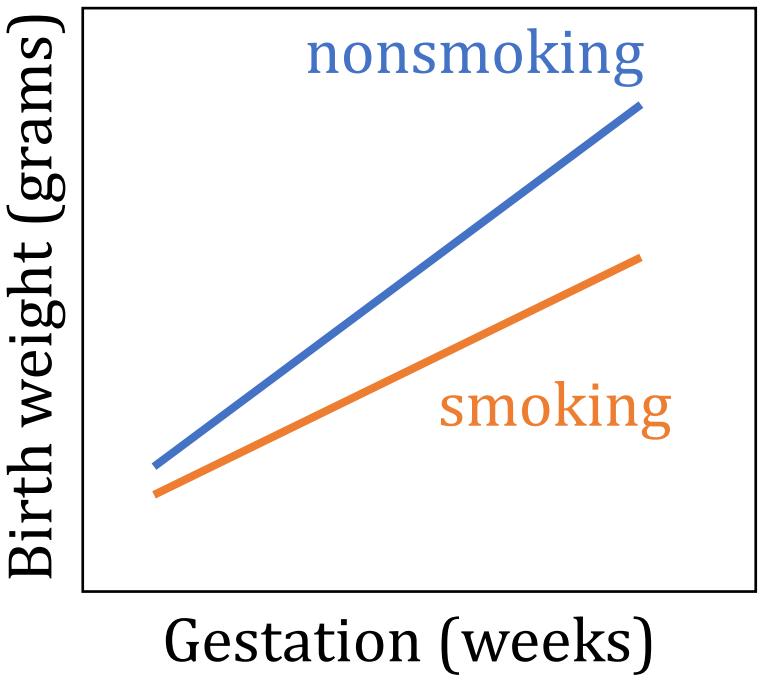
The effect of the length of gestation on the birth weight decreases by 8.178 grams on average for smoking mothers compared to nonsmoking mothers.

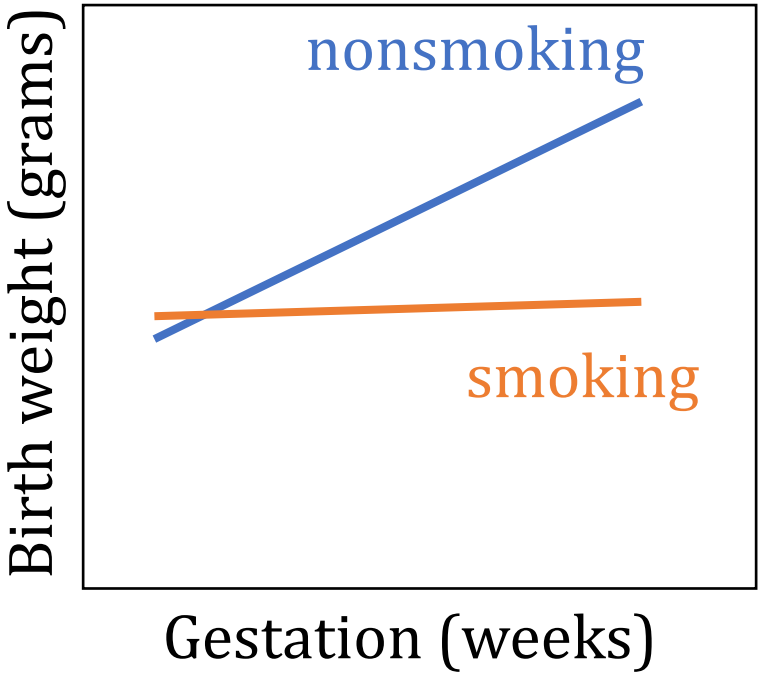
```
> Gest_c <- birthsmokers$Gest-mean(birthsmokers$Gest)
> reg <- lm(Wgt~Gest_c*Smoke, data = birthsmokers)
> summary(reg)

Call:
lm(formula = Wgt ~ Gest_c * Smoke, data = birthsmokers)

Residuals:
     Min       1Q   Median       3Q      Max 
-228.528  -89.560    0.273   83.629  184.529 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept)   3144.329     30.110 104.429  < 2e-16 ***
Gest_c         147.207     13.120  11.220 7.15e-12 ***
Smoke         -244.563     42.577  -5.744 3.65e-06 ***
Gest_c:Smoke    -8.178     18.515  -0.442    0.662    
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 117.2 on 28 degrees of freedom
Multiple R-squared:  0.8971,	Adjusted R-squared:  0.8861 
F-statistic: 81.37 on 3 and 28 DF,  p-value: 6.144e-14
```

Multiple Linear Regression

In which of these graphs do you see an interaction between gestation and smoking status?







USING R AND RSTUDIO

Multiple Linear Regression

A case study: Predicting course ratings

Hamermesh & Parker (2004) Beauty in the classroom: instructors' pulchritude and putative pedagogical productivity. *Economics of Education Review*.

Collected data at the University of Texas at Austin :

- Evaluations at the end of the semester for the academic years 2000-2002: 463 courses.
- Instructors at all levels, with pictures on their department's websites: 94 professors.
- Beauty judgements were made by 6 students who had not attended the classes and were not aware of the course evaluations: combined into a beauty score (out of 10).

What are the sampling units?
What is the population?

Multiple Linear Regression

Preliminary Analysis: Descriptive statistics

Table 1 page 3

Descriptive statistics, courses, instructors and evaluations

Variable	All
Course evaluation	4.022 (0.525)
Instructor evaluation	4.217 (0.540)
Number of students	55.18 (75.07)
Percent evaluating	74.43
Female	0.359
Minority	0.099
Non-native English	0.037
Tenure track	0.851
Lower division	0.339
One credit	0.029
Number of courses	463
Number of faculty	94

Note: Means with standard deviations in parentheses. All statistics except for those describing the number of students, the percent evaluating the instructor and the lower–upper division distinction are weighted by the number of students completing the course evaluation forms.

Look at the list of variables:
Which one can be considered as the response?
Which ones as potential predictors?

Multiple Linear Regression

Preliminary Analysis: Descriptive statistics

Table 1 page 3

Descriptive statistics, courses, instructors and evaluations

Variable	All
Course evaluation	4.022 (0.525)
Instructor evaluation	4.217 (0.540)
Number of students	55.18 (75.07)
Percent evaluating	74.43
Female	0.359
Minority	0.099
Non-native English	0.037
Tenure track	0.851
Lower division	0.339
One credit	0.029
Number of courses	463
Number of faculty	94

Note: Means with standard deviations in parentheses. All statistics except for those describing the number of students, the percent evaluating the instructor and the lower–upper division distinction are weighted by the number of students completing the course evaluation forms.

Look at the descriptive statistics. Interpret the highlighted values.

Multiple Linear Regression

Preliminary Analysis: Comparisons

Table 2 page 3

Beauty evaluations, individual and composite

	Average	Standard deviation	Standardized	
			Minimum	Maximum
Individual ratings:				
Male, upper division—1	4.43	2.18	−1.57	2.10
Male, upper division—2	4.87	1.65	−2.34	2.50
Female, upper division—1	5.18	2.05	−2.03	1.84
Female, upper division—2	5.39	2.10	−2.10	2.20
Male, lower division	3.53	1.70	−1.49	2.04
Female, lower division	4.14	1.88	−1.67	2.05
Composite standardized rating	0	0.83	−1.54	1.88

Compare the highlighted values.

Multiple Linear Regression

Preliminary Analysis: Comparisons

Table 2 page 3

Beauty evaluations, individual and composite

	Average	Standard deviation	Standardized	
			Minimum	Maximum
Individual ratings:				
Male, upper division—1	4.43	2.18	−1.57	2.10
Male, upper division—2	4.87	1.65	−2.34	2.50
Female, upper division—1	5.18	2.05	−2.03	1.84
Female, upper division—2	5.39	2.10	−2.10	2.20
Male, lower division	3.53	1.70	−1.49	2.04
Female, lower division	4.14	1.88	−1.67	2.05
Composite standardized rating				
	0	0.83	−1.54	1.88

If we wanted to compare the highlighted values with a test, what test shall we conduct?

Multiple Linear Regression

Preliminary Analysis: Comparisons

Table 2 page 3

Beauty evaluations, individual and composite

	Average	Standard deviation	Standardized	
			Minimum	Maximum
Individual ratings:				
Male, upper division—1	4.43	2.18	−1.57	2.10
Male, upper division—2	4.87	1.65	−2.34	2.50
Female, upper division—1	5.18	2.05	−2.03	1.84
Female, upper division—2	5.39	2.10	−2.10	2.20
Male, lower division	3.53	1.70	−1.49	2.04
Female, lower division	4.14	1.88	−1.67	2.05
Composite standardized rating	0	0.83	−1.54	1.88

What did the researcher do when they "standardized" the beauty scores?

Multiple Linear Regression

Predicting course ratings: Multiple Regression

Table 3 page 4

Weighted least-squares estimates of the determinants of class ratings

Variable	All
Composite standardized beauty	0.275 (0.059)
Female	−0.239 (0.085)
Minority	−0.249 (0.112)
Non-native English	−0.253 (0.134)
Tenure track	−0.136 (0.094)
Lower division	−0.046 (0.111)
One-credit course	0.687 (0.166)
R^2	.279
N courses	463
N faculty	94

Note: Robust standard errors in parentheses here and in Table 4.

Look at the estimates for each predictor. Which predictors have a positive effect on the class ratings (controlling for other variables)?

On page 4, the authors wrote: "The striking fact from the estimates [...] is the statistical significance of the composite standardized beauty measure".

Does Table 3 show which variables are significant?

Multiple Linear Regression

Predicting course ratings

Based on the RStudio output, which predictors have a significant effect on class ratings?

```
Coefficients:
                    Estimate Std. Error t value             Pr(>|t|)
(Intercept)          4.22314    0.06512  64.856 < 0.0000000000000002 ***
beauty_standardized  0.27481    0.02759   9.959 < 0.0000000000000002 ***
genderfemale        -0.23899    0.04586  -5.212          0.000000284 ***
minorityyes         -0.24894    0.07999  -3.112              0.00197 **
nativeno            -0.25271    0.11985  -2.109              0.03553 *
tenureyes           -0.13592    0.06247  -2.176              0.03007 *
divisionlower       -0.04589    0.04380  -1.048              0.29523
creditssingle        0.68651    0.13685   5.016          0.000000757 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.719 on 455 degrees of freedom
Multiple R-squared:  0.2788,    Adjusted R-squared:  0.2677
F-statistic: 25.13 on 7 and 455 DF,  p-value: < 0.0000000000000022
```

Multiple Linear Regression

Predicting course ratings

The authors acknowledged some issues that came up during their study:

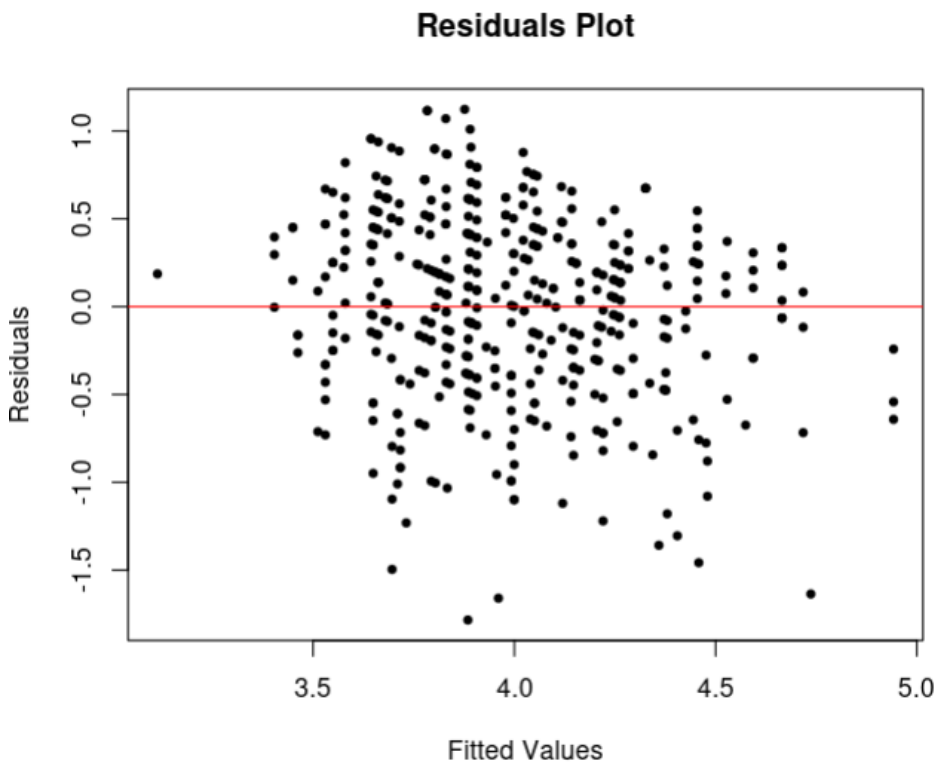
- "the most serious potential problem may result from a type of sample selectivity"
- "our measure of beauty may merely be a proxy"
- "whether higher instructional ratings mean that the faculty member is a better teacher"

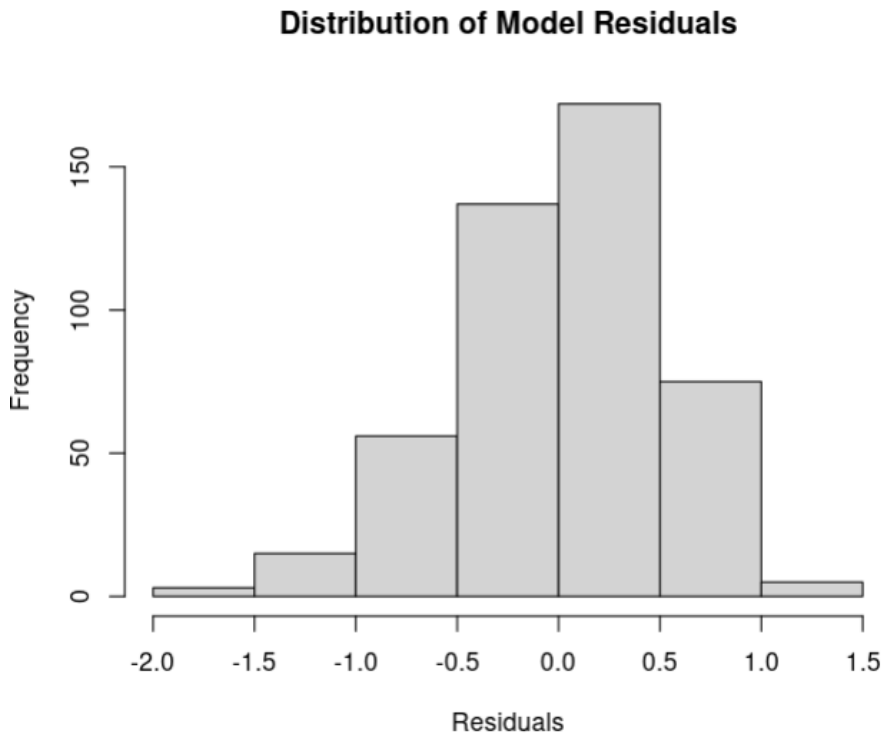
Why is important for the authors to acknowledge these issues?

What other potential issues should they have also addressed?

Multiple Linear Regression

Predicting course ratings





Check the assumptions

Multiple Linear Regression

Predicting course ratings

Table 3 page 4

Weighted least-squares estimates of the determinants of class ratings

Variable	All
Composite standardized beauty	0.275 (0.059)
Female	−0.239 (0.085)
Minority	−0.249 (0.112)
Non-native English	−0.253 (0.134)
Tenure track	−0.136 (0.094)
Lower division	−0.046 (0.111)
One-credit course	0.687 (0.166)
R^2	.279
N courses	463
N faculty	94

Note: Robust standard errors in parentheses here and in Table 4.

What's the model fit?
What does it mean?

Next

Day 4
Logistic Regression

- Odds
- Logistic Regression
- Model evaluation with ROC curves or confusion matrix

Any questions? comments?