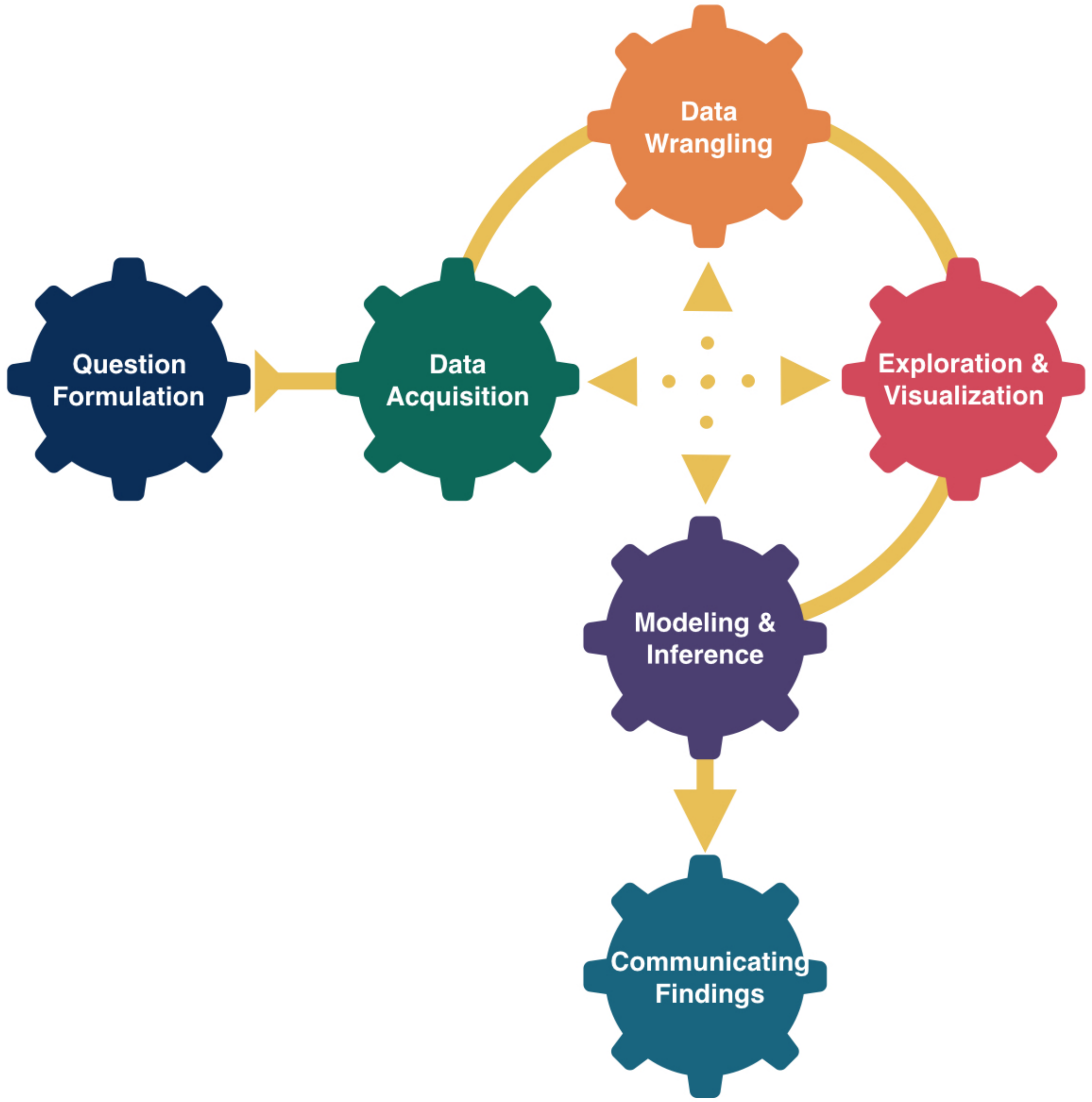


Inference for Linear Regression

Kelly McConville

Stat 100

Week 13 | Fall 2023

Announcements

- Lecture Quizzes
  - Last one this week.
  - Plus Extra Credit Lecture Quiz: Due Tues, Dec 5th at 5pm
- Last section this week!
  - Receive the last p-set.
- The material from next Monday's lecture may appear on the final and so we have included relevant practice problems on the review sheet.

Goals for Today

- Recap multiple linear regression
- Check assumptions for linear regression inference
- Hypothesis testing for linear regression
- Estimation and prediction inference for linear regression

Games and Prizes!

Lunch and Treats!

ggparty

Th Dec 7 12-2 pm

Location: SC 316

If you are able to attend, please RSVP: bit.ly/ggpartyf23

What does statistical inference (estimation and hypothesis testing) look like when I have more than 0 or 1 explanatory variables?

One route: Multiple Linear Regression!

Multiple Linear Regression

Linear regression is a flexible class of models that allow for:

- Both quantitative and categorical explanatory variables.
- Multiple explanatory variables.
- Curved relationships between the response variable and the explanatory variable.
- BUT the response variable is quantitative.

In this week's p-set you will explore the importance of controlling for key explanatory variables when making inferences about relationships.

Multiple Linear Regression

Form of the Model:

$$y = \beta_o + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p + \epsilon$$

Fitted Model: Using the Method of Least Squares,

$$\hat{y} = \hat{\beta}_o + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \cdots + \hat{\beta}_p x_p$$

Typical Inferential Questions – Hypothesis Testing

Should x_2 be in the model that already contains x_1 and x_3? Also often asked as “Controlling for x_1 and x_3, is there evidence that x_2 has a relationship with y?”

$$y = \beta_o + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \epsilon$$

In other words, should $\beta_2 = 0$?

Typical Inferential Questions – Estimation

After controlling for the other explanatory variables, what is the range of plausible values for β_3 (which summarizes the relationship between y and x_3)?

$$y = \beta_o + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \epsilon$$

Typical Inferential Questions – Prediction

While $\hat{y}$ is a point estimate for y, can we also get an interval estimate for y? In other words, can we get a range of plausible predictions for y?

$$y = \beta_o + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \epsilon$$

To answer these questions, we need to add some assumptions to our linear regression model.

Multiple Linear Regression

Form of the Model:

$$y = \beta_o + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p + \epsilon$$

Additional Assumptions:

$$\epsilon \overset{\text{ind}}{\sim} N(\mu = 0, \sigma = \sigma_\epsilon)$$

σ_ϵ = typical deviations from the model

Let's unpack these assumptions!

Assumptions – Independence

For ease of visualization, let's assume a simple linear regression model:

$$y = \beta_o + \beta_1 x_1 + \epsilon \quad \text{where} \quad \epsilon \overset{\text{ind}}{\sim} N\left(0, \sigma_\epsilon\right)$$

Assumption: The cases are independent of each other.

Question: How do we check this assumption?

Consider how the data were collected.

Assumptions – Normality

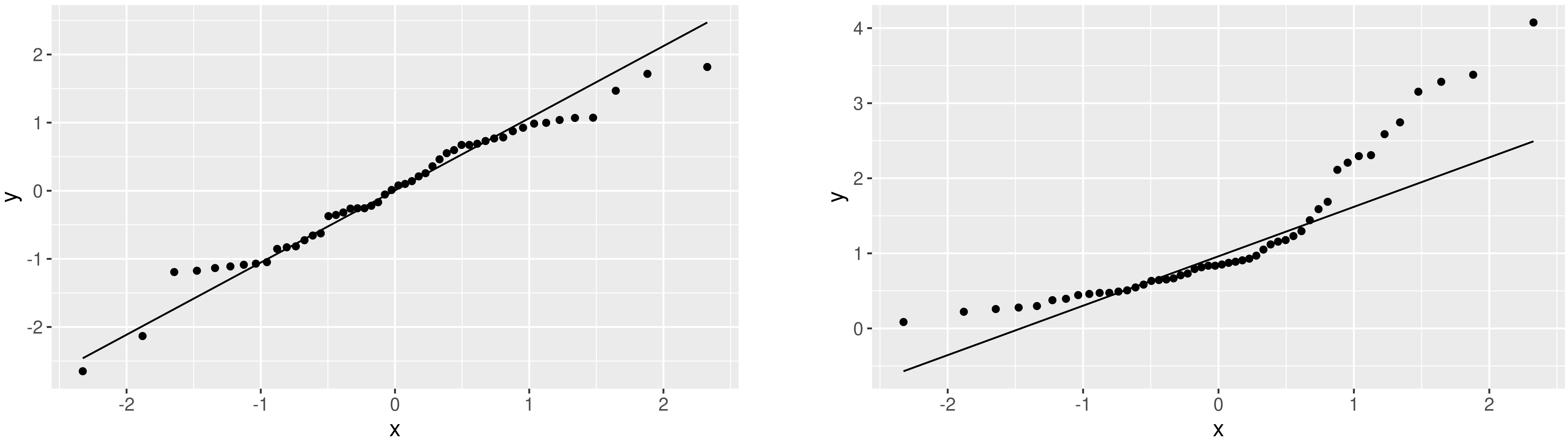
$$y = \beta_o + \beta_1 x_1 + \epsilon \quad \text{where} \quad \epsilon \overset{\text{ind}}{\sim} N(0, \sigma_\epsilon)$$

Assumption: The errors are normally distributed.

Question: How do we check this assumption?

Recall the residual: $e = y - \hat{y}$

QQ-plot: Plot the residuals against the quantiles of a normal distribution!



Assumptions – Mean of Errors

$$y = \beta_o + \beta_1 x_1 + \epsilon \quad \text{where} \quad \epsilon \overset{\text{ind}}{\sim} N\left(0, \sigma_\epsilon\right)$$

Assumption: The points will, on average, fall on the line.

Question: How do we check this assumption?

If you use the Method of Least Squares, then you don't have to check.

It will be true by construction:

$$\sum e = 0$$

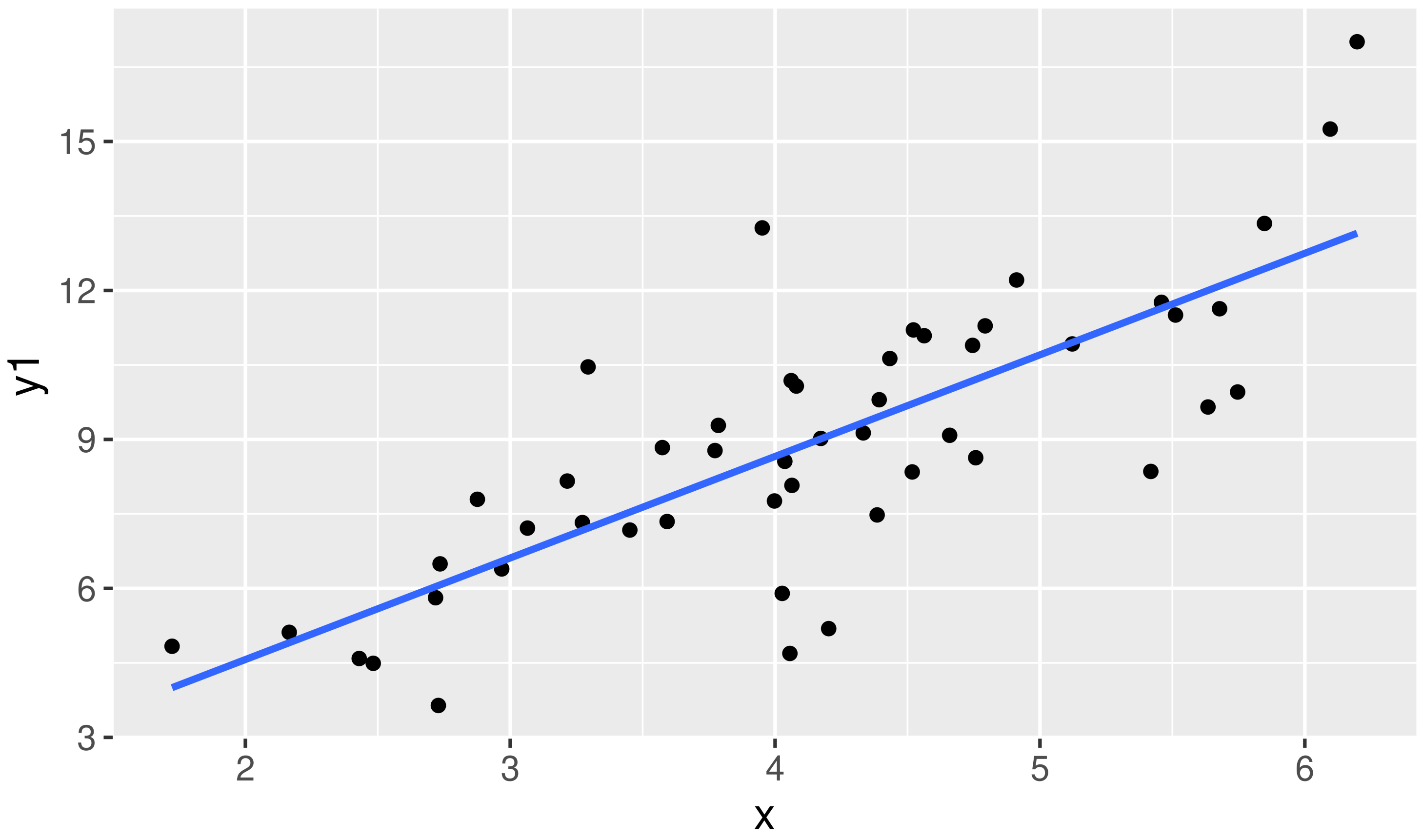
Assumptions – Constant Variance

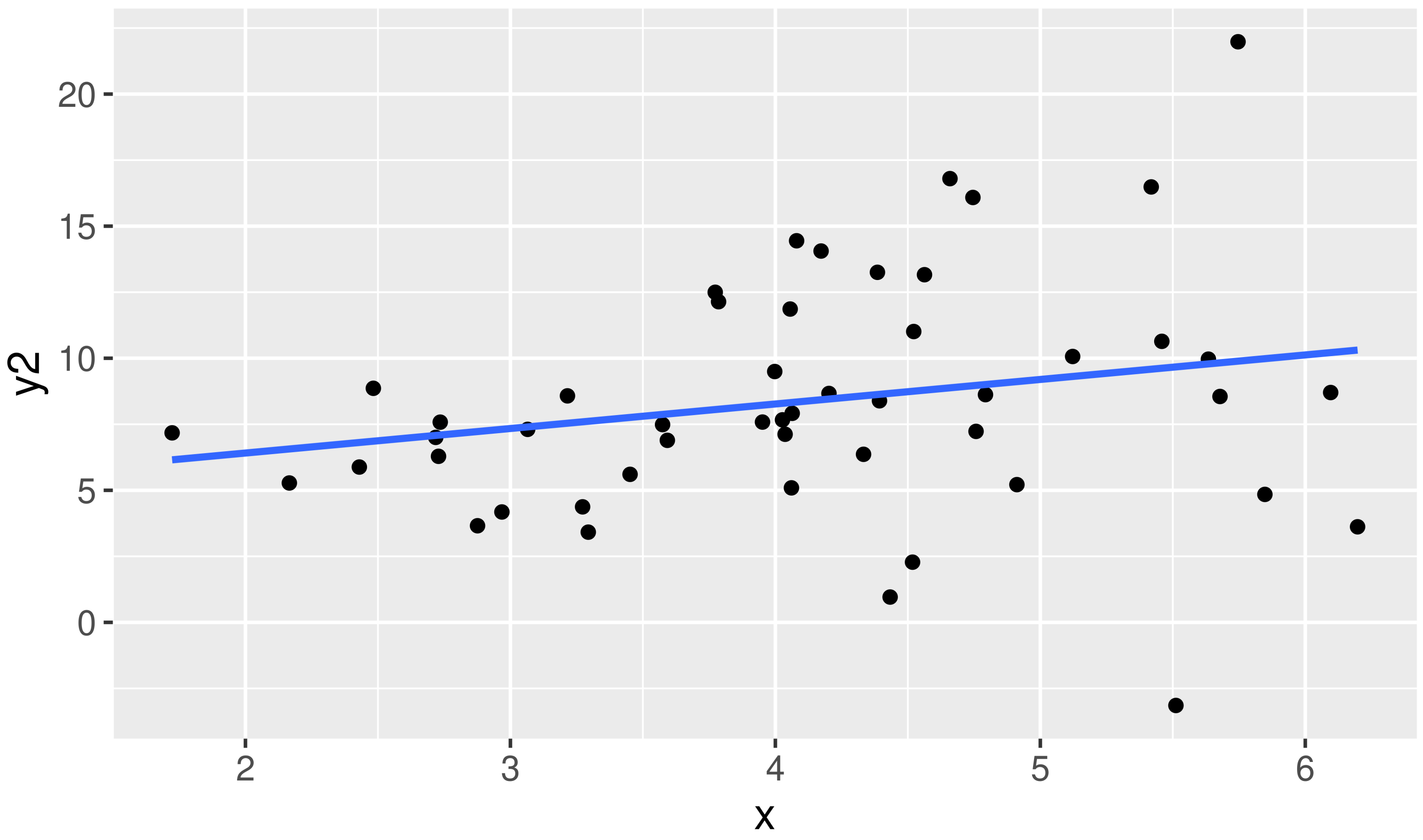
$$y = \beta_o + \beta_1 x_1 + \epsilon \quad \text{where} \quad \epsilon \overset{\text{ind}}{\sim} N\left(0, \sigma_\epsilon\right)$$

Assumption: The variability in the errors is constant.

Question: How do we check this assumption?

One option: Scatterplot


y1
15
12
9
6
3
2
3
4
5
6
x


y2
20
15
10
5
0
2
3
4
5
6
x

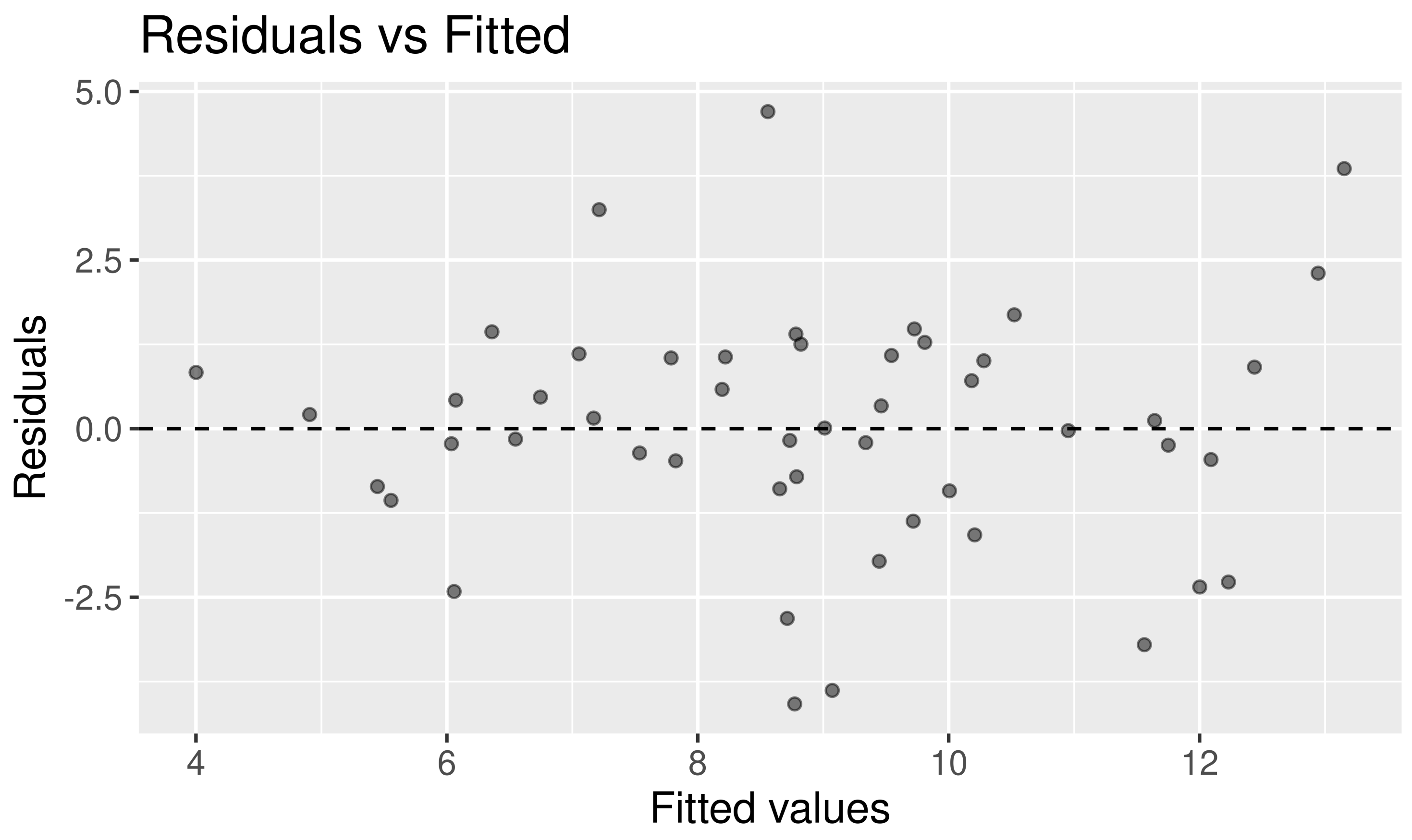
Assumptions – Constant Variance

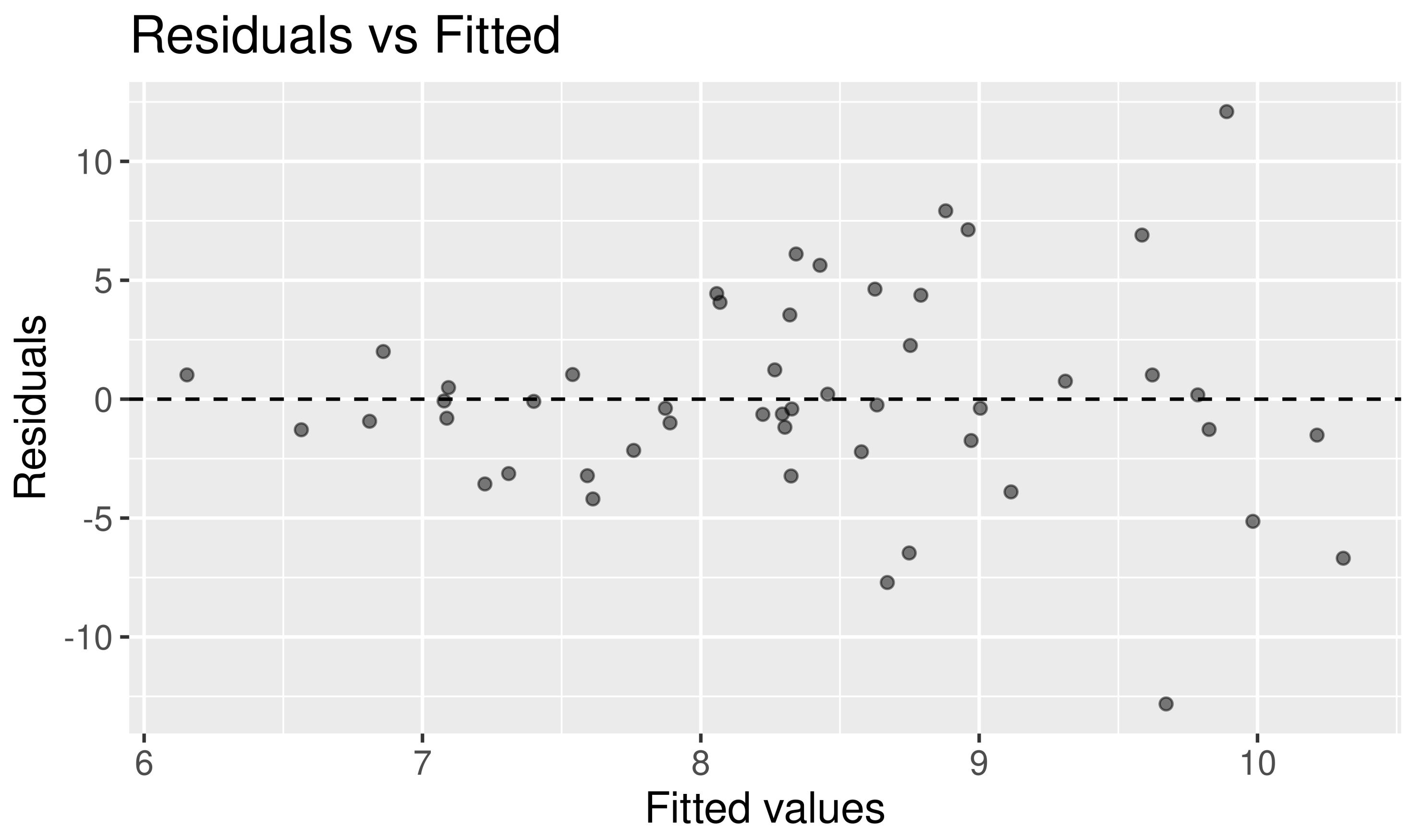
$$y = \beta_o + \beta_1 x_1 + \epsilon \quad \text{where} \quad \epsilon \overset{\text{ind}}{\sim} N(0, \sigma_\epsilon)$$

Assumption: The variability in the errors is constant.

Question: How do we check this assumption?

Better option (especially when have more than 1 explanatory variable): Residual Plot


Residuals vs Fitted
Residuals
Fitted values
5.0
2.5
0.0
-2.5
4
6
8
10
12


Residuals vs Fitted
Residuals
Fitted values
10
5
0
-5
-10
6
7
8
9
10

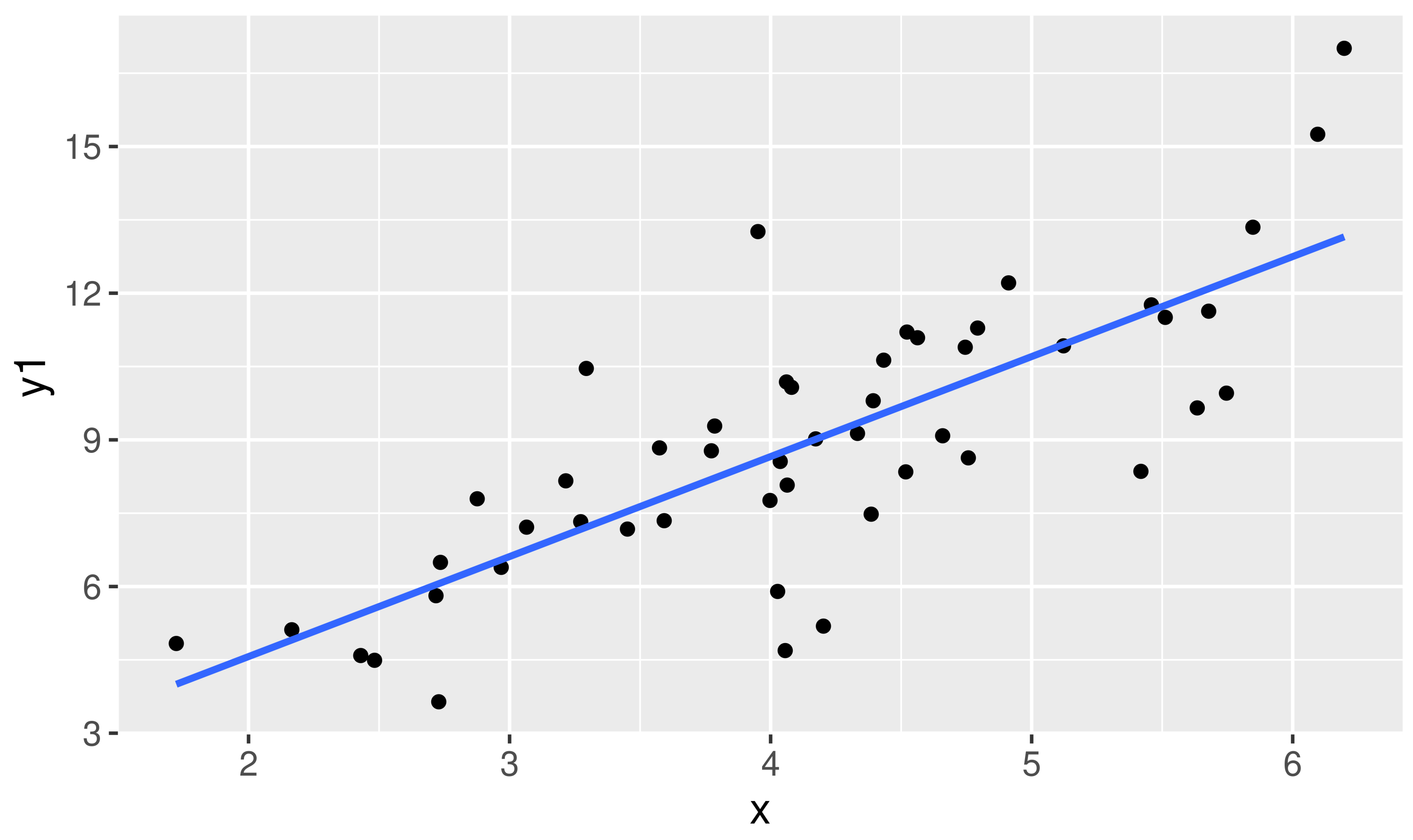
Assumptions – Model Form

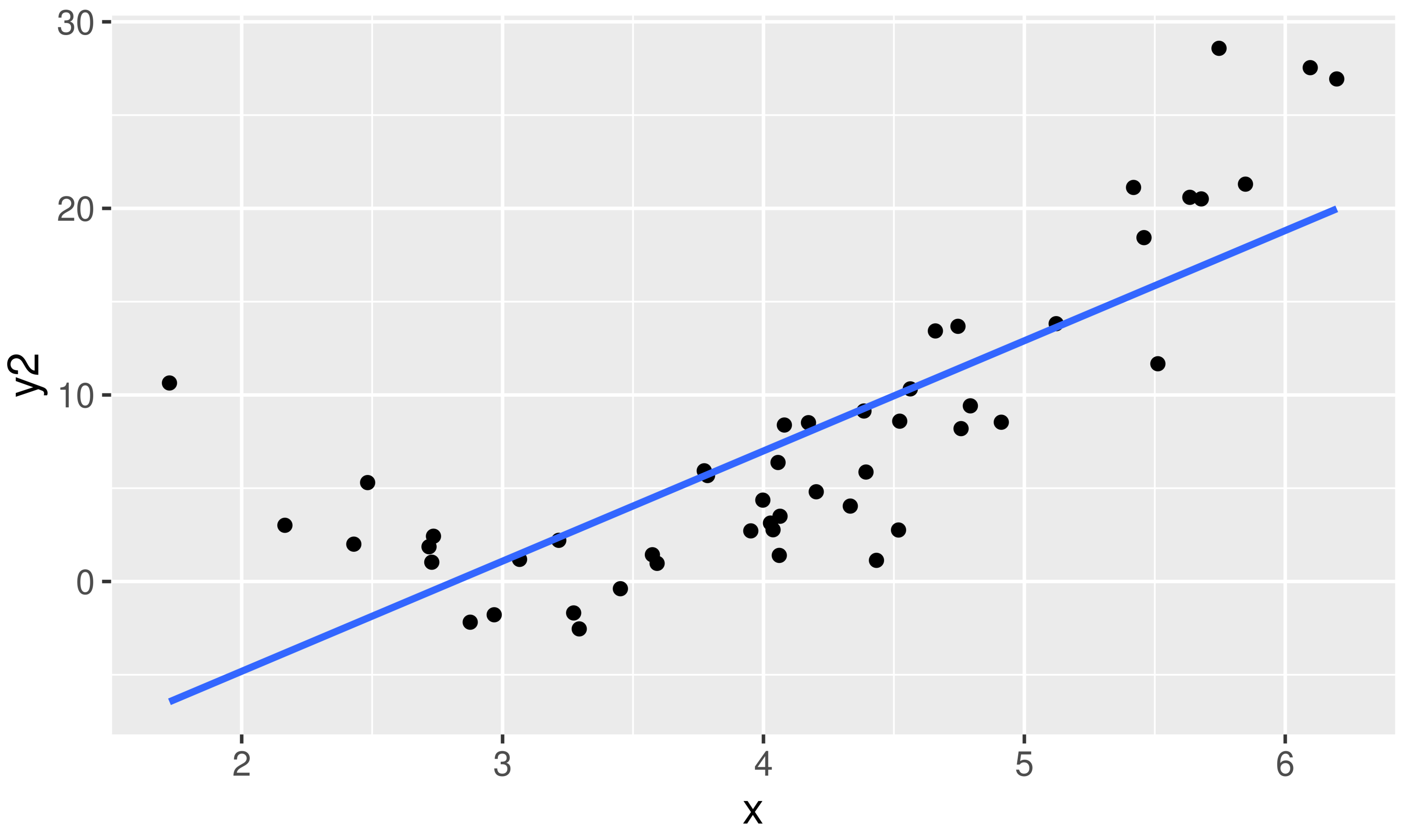
$$y = \beta_o + \beta_1 x_1 + \epsilon \quad \text{where} \quad \epsilon \overset{\text{ind}}{\sim} N(0, \sigma_\epsilon)$$

Assumption: The model form is appropriate.

Question: How do we check this assumption?

One option: Scatterplot(s)





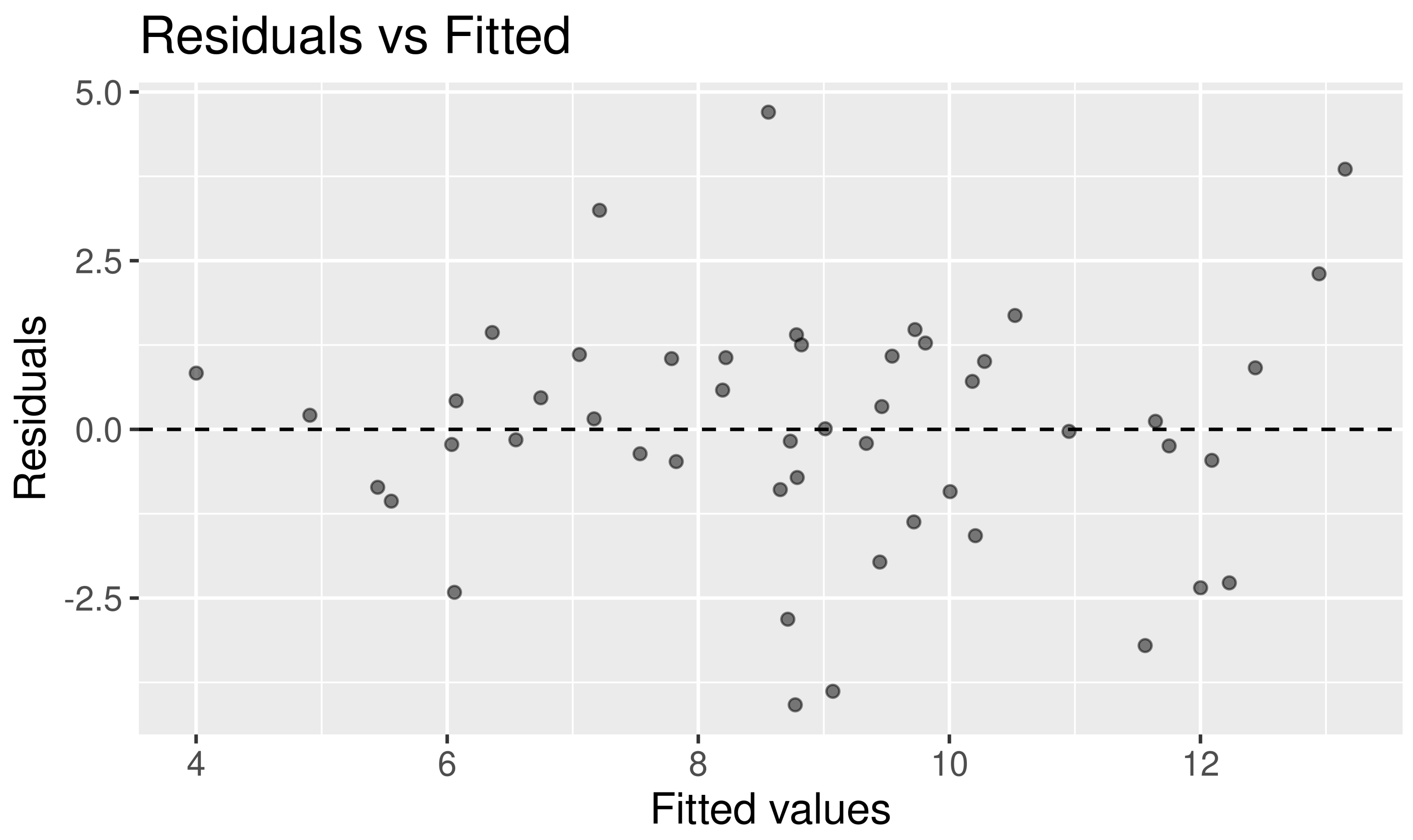
Assumptions – Model Form

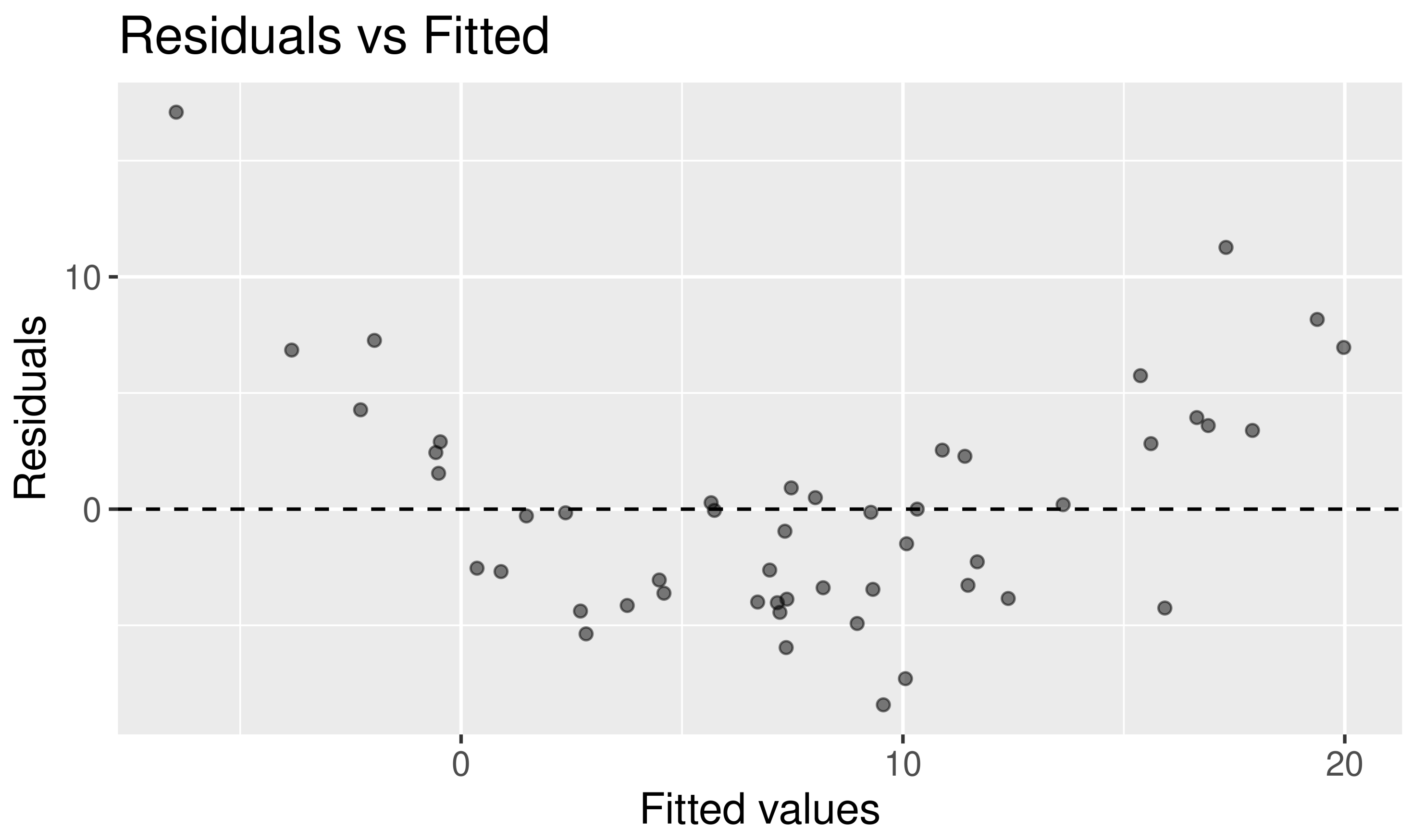
$$y = \beta_o + \beta_1 x_1 + \epsilon \quad \text{where} \quad \epsilon \overset{\text{ind}}{\sim} N(0, \sigma_\epsilon)$$

Assumption: The model form is appropriate.

Question: How do we check this assumption?

Better option (especially when have more than 1 explanatory variable): **Residual Plot**





Assumption Checking

Question: What if the assumptions aren't all satisfied?

- Try transforming the data and building the model again.
- Use a modeling technique beyond linear regression.

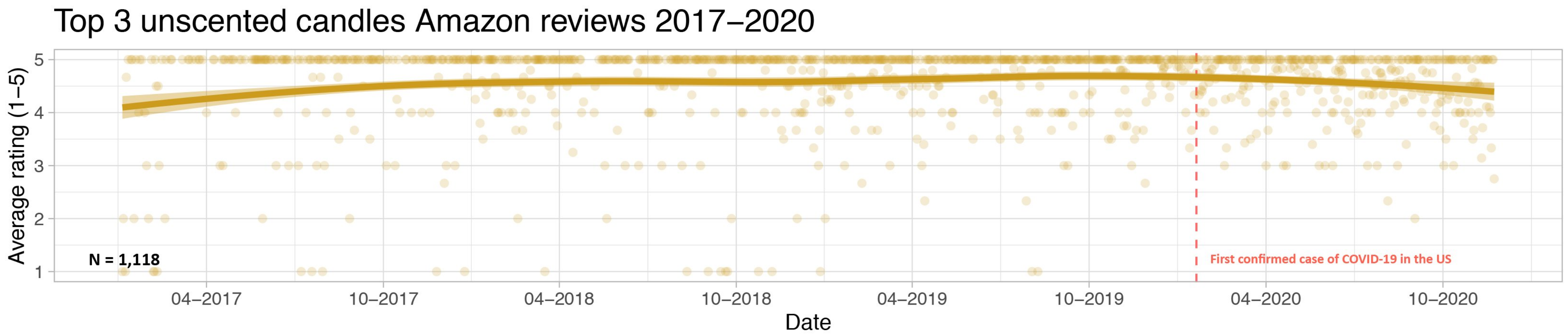
Question: What if the assumptions are all (roughly) satisfied?

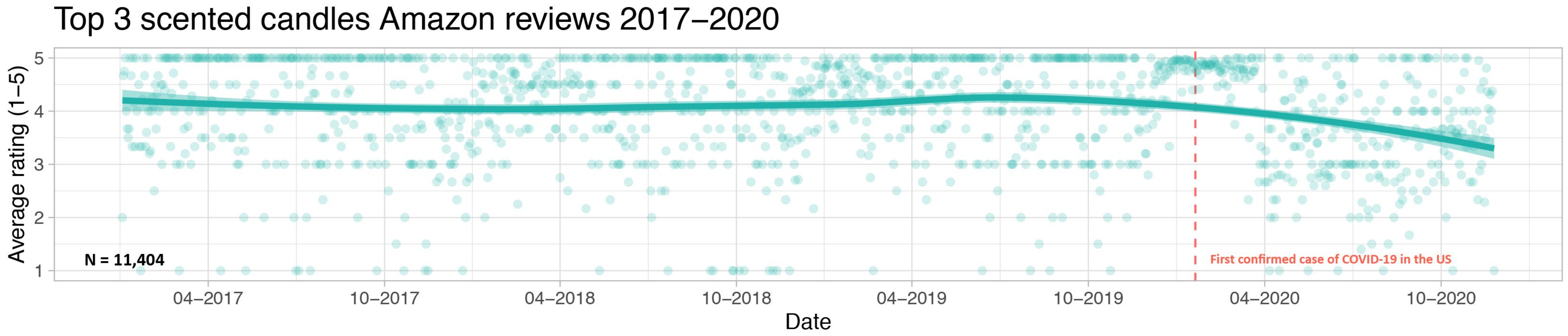
- Can now start answering your inference questions!

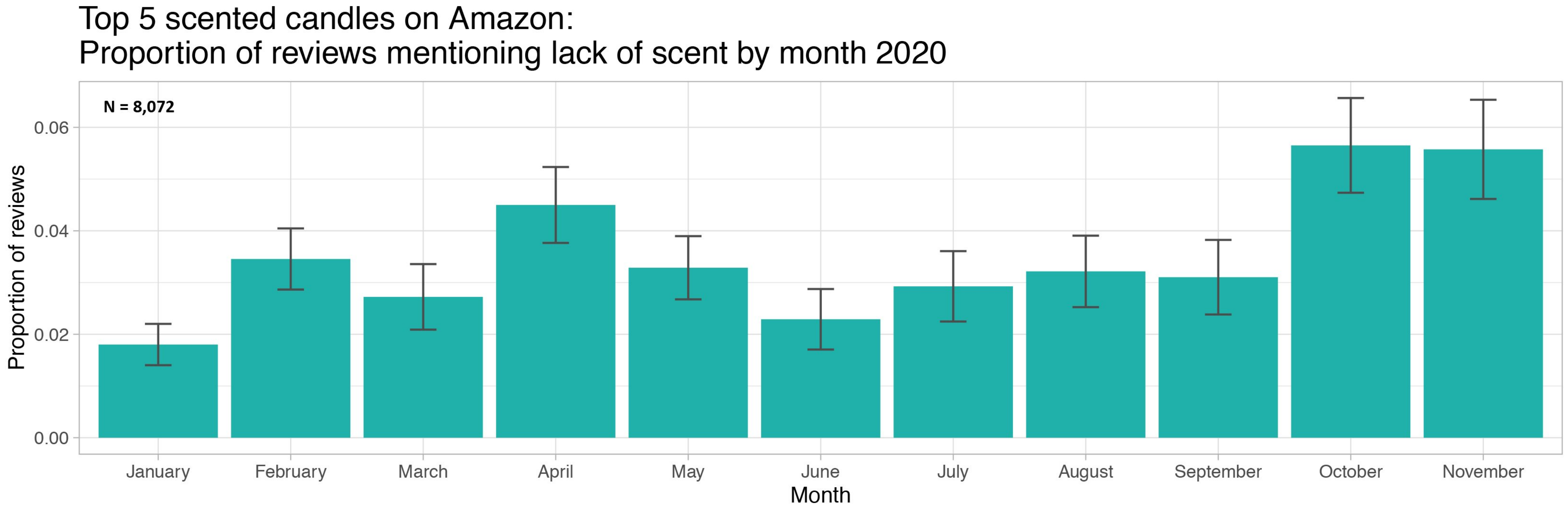
Let’s now look at an example and learn how to create qq-plots and residual plots in R.

Example: COVID and Candle Ratings

Kate Petrova created a dataset that made the rounds on Twitter:


Top 3 unscented candles Amazon reviews 2017–2020
Average rating (1–5)
N = 1,118
First confirmed case of COVID-19 in the US
04–2017
10–2017
04–2018
10–2018
04–2019
10–2019
04–2020
10–2020
Date


Top 3 scented candles Amazon reviews 2017–2020
Average rating (1–5)
N = 11,404
First confirmed case of COVID-19 in the US
04–2017
10–2017
04–2018
10–2018
04–2019
10–2019
04–2020
10–2020
Date


Top 5 scented candles on Amazon:
Proportion of reviews mentioning lack of scent by month 2020
N = 8,072
Proportion of reviews
0.00
0.02
0.04
0.06
January
February
March
April
May
June
July
August
September
October
November
Month

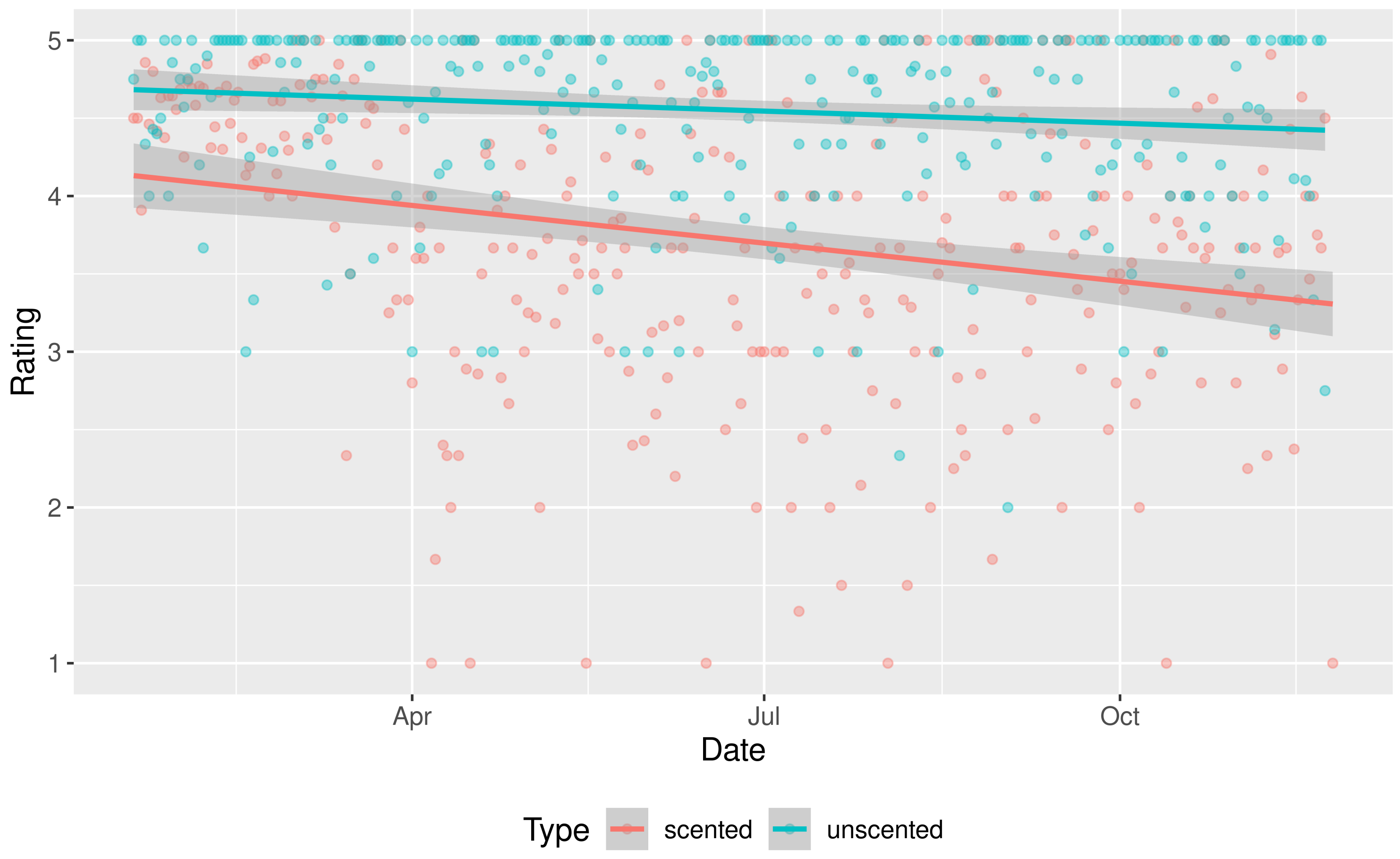
COVID and Candle Ratings

She posted all her data and code to GitHub and I did some light wrangling so that we could answer the question:

> Do we have evidence that early in the pandemic the association between time and Amazon rating varies by whether or not a candle is scented and in particular, that scented candles have a steeper decline in ratings over time?

In other words, do we have evidence that we should allow the slopes to vary?

```
ggplot(data = all,
       mapping = aes(x = Date,
                     y = Rating,
                     color = Type)) +
  geom_point(alpha = 0.4) +
  geom_smooth(method = lm) +
  theme(legend.position = "bottom")
```



COVID and Candle Ratings

Checking assumptions:

Assumption: The cases are independent of each other.

Question: What needs to be true about the candles sampled?

Assumption Checking in R

The R package we will use to check model assumptions is called `gglm` and was written by one of my former Reed students, Grayson White.

```
library(gglm)
```

First need to fit the model:

```
glimpse(all)
```

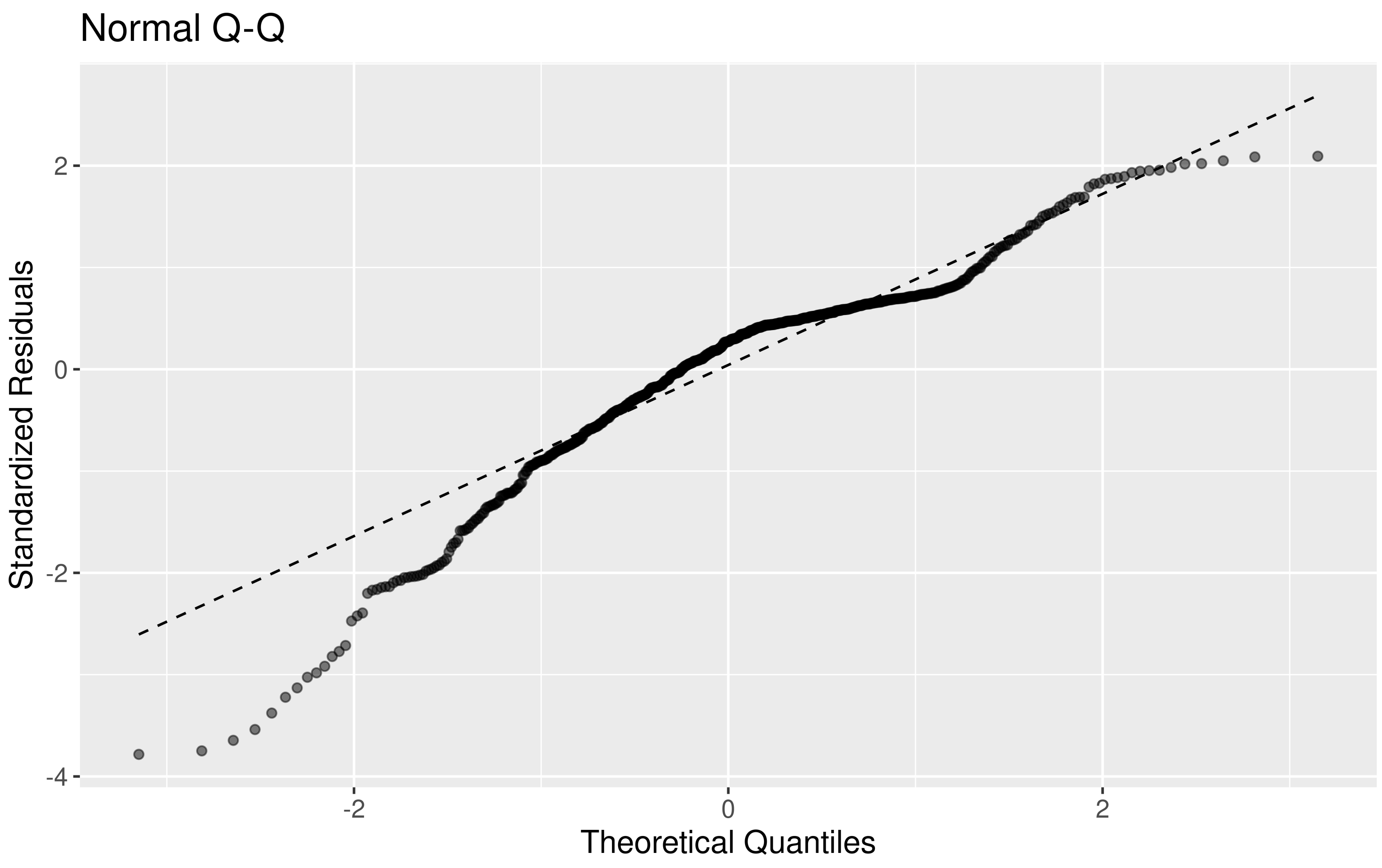
```
Rows: 612
Columns: 3
$ Date   <date> 2020-01-20, 2020-01-21, 2020-01-22, 2020-01-23, 2020-01-24, 20…
$ Rating <dbl> 4.500000, 4.500000, 3.909091, 4.857143, 4.461538, 4.800000, 4.4…
$ Type   <chr> "scented", "scented", "scented", "scented", "scented", "scented…
```

```
mod <- lm(Rating ~ Date * Type, data = all)
```

qq-plot

Assumption: The errors are normally distributed.

```
# Normal qq plot
ggplot(data = mod) +
  stat_normal_qq()
```

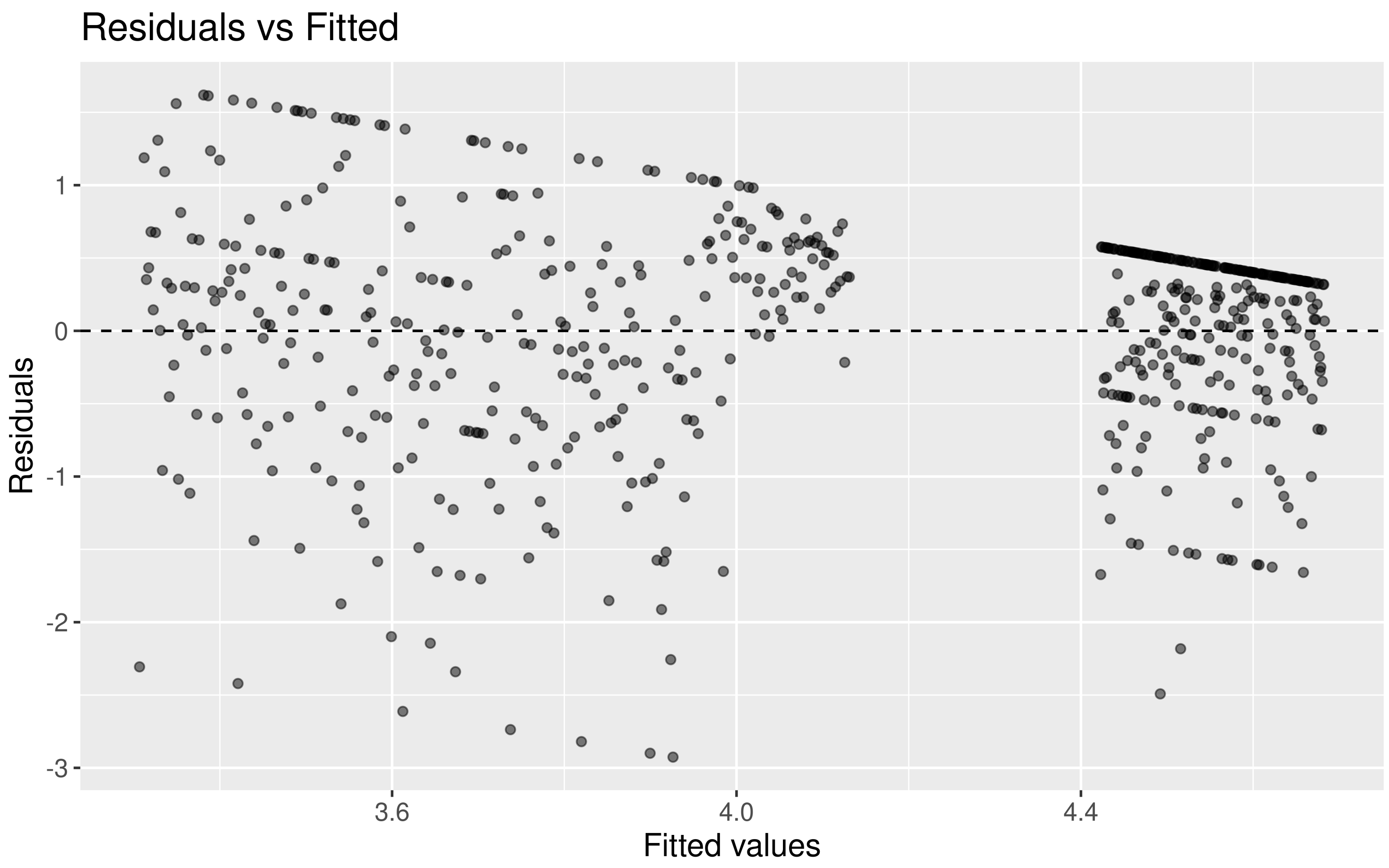


Residual Plot

Assumption: The variability in the errors is constant.

Assumption: The model form is appropriate.

```
# Residual plot
ggplot(data = mod) +
  stat_fitted_resid()
```



Hypothesis Testing

Question: What tests is `get_regression_table()` conducting?

For the moment, let's focus on the equal slopes model.

```r
mod <- lm(Rating ~ Date + Type, data = all)
get_regression_table(mod)
```

```
# A tibble: 3 × 7
  term             estimate std_error statistic p_value lower_ci upper_ci
  <chr>               <dbl>     <dbl>     <dbl>   <dbl>    <dbl>    <dbl>
1 intercept          36.2       6.50       5.58       0   23.5     49.0
2 Date               -0.002     0         -5.00       0   -0.002   -0.001
3 Type: unscented     0.831     0.063     13.2        0    0.707    0.955
```

In General:

$$H_o : \beta_j = 0 \quad \text{assuming all other predictors are in the model}$$

$$H_a : \beta_j \neq 0 \quad \text{assuming all other predictors are in the model}$$

Hypothesis Testing

Question: What tests is `get_regression_table()` conducting?

```r
mod <- lm(Rating ~ Date + Type, data = all)
get_regression_table(mod)
```

```
# A tibble: 3 × 7
  term            estimate std_error statistic p_value lower_ci upper_ci
  <chr>              <dbl>     <dbl>     <dbl>   <dbl>    <dbl>    <dbl>
1 intercept         36.2       6.50       5.58       0   23.5     49.0
2 Date              -0.002     0         -5.00       0   -0.002   -0.001
3 Type: unscented    0.831     0.063     13.2        0    0.707    0.955
```

For our Example:

Row 2:

$$H_o : \beta_1 = 0 \quad \text{given Type is already in the model}$$

$$H_a : \beta_1 \neq 0 \quad \text{given Type is already in the model}$$

Hypothesis Testing

Question: What tests is `get_regression_table()` conducting?

```r
mod <- lm(Rating ~ Date + Type, data = all)
get_regression_table(mod)
```

```
# A tibble: 3 × 7
  term            estimate std_error statistic p_value lower_ci upper_ci
  <chr>              <dbl>     <dbl>     <dbl>   <dbl>    <dbl>    <dbl>
1 intercept         36.2       6.50       5.58       0   23.5     49.0
2 Date              -0.002     0         -5.00       0   -0.002   -0.001
3 Type: unscented    0.831     0.063     13.2        0    0.707    0.955
```

For our Example:

Row 3:

$$H_o : \beta_2 = 0 \quad \text{given Date is already in the model}$$

$$H_a : \beta_2 \neq 0 \quad \text{given Date is already in the model}$$

Hypothesis Testing

Question: What tests is `get_regression_table()` conducting?

In General:

$$H_o : \beta_j = 0 \quad \text{assuming all other predictors are in the model}$$

$$H_a : \beta_j \neq 0 \quad \text{assuming all other predictors are in the model}$$

Test Statistic: Let p = number of explanatory variables.

$$t = \frac{\hat{\beta}_j - 0}{SE(\hat{\beta}_j)} \sim t(df = n - p)$$

when H_o is true and the model assumptions are met.

Our Example

```r
get_regression_table(mod)
```

```
# A tibble: 3 × 7
  term            estimate std_error statistic p_value lower_ci upper_ci
  <chr>              <dbl>     <dbl>     <dbl>   <dbl>    <dbl>    <dbl>
1 intercept         36.2       6.50       5.58       0   23.5     49.0
2 Date              -0.002     0         -5.00       0   -0.002   -0.001
3 Type: unscented    0.831     0.063     13.2        0    0.707    0.955
```

Row 3:

$$H_o : \beta_2 = 0 \quad \text{given Date is already in the model}$$

$$H_a : \beta_2 \neq 0 \quad \text{given Date is already in the model}$$

Test Statistic:

$$t = \frac{\hat{\beta}_2 - 0}{SE(\hat{\beta}_2)} = \frac{0.831 - 0}{0.063} = 13.2$$

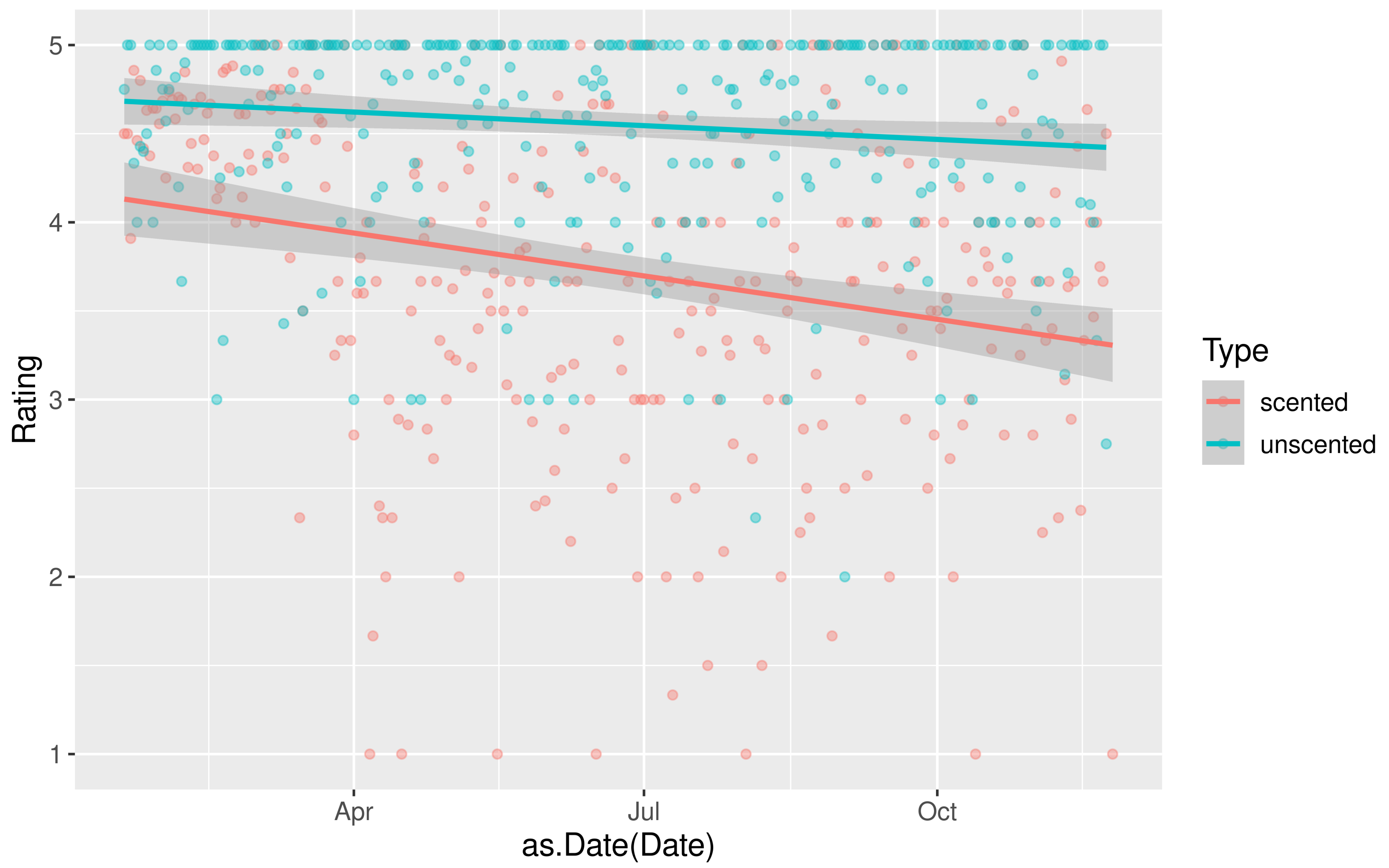
with p-value $= P(t \leq -13.2) + P(t \geq 13.2) \approx 0.$

There is evidence that including whether or not the candle is scented adds useful information to the linear regression model for Amazon ratings that already controls for date.

Example

Do we have evidence that early in the pandemic the association between time and Amazon rating varies by whether or not a candle is scented and in particular, that scented candles have a steeper decline in ratings over time?

```r
ggplot(data = all, mapping = aes(x = as.Date(Date),
                                 y = Rating,
                                 color = Type)) +
  geom_point(alpha = 0.4) +
  geom_smooth(method = lm)
```



Example

Do we have evidence that early in the pandemic the association between time and Amazon rating varies by whether or not a candle is scented and in particular, that scented candles have a steeper decline in ratings over time?

```r
mod <- lm(Rating ~ Date * Type, data = all)
get_regression_table(mod)
```

```
# A tibble: 4 × 7
  term                 estimate std_error statistic p_value lower_ci upper_ci
  <chr>                   <dbl>     <dbl>     <dbl>   <dbl>    <dbl>    <dbl>
1 intercept              52.7       9.09       5.80   0         34.9     70.6
2 Date                   -0.003     0         -5.40   0         -0.004   -0.002
3 Type: unscented       -32.6      12.9       -2.52   0.012   -58.0      -7.24
4 Date:Typeunscented      0.002     0.001      2.59   0.01      0         0.003
```

One More Example – Prices of Houses in Saratoga Springs, NY

Does whether or not a house has central air conditioning relate to its price for houses in Saratoga Springs?

```r
library(mosaicData)
mod1 <- lm(price ~ centralAir, data = SaratogaHouses)
get_regression_table(mod1)
```

```
# A tibble: 2 × 7
  term            estimate std_error statistic p_value lower_ci upper_ci
  <chr>              <dbl>     <dbl>     <dbl>   <dbl>    <dbl>    <dbl>
1 intercept        254904.     3685.      69.2       0  247676.  262132.
2 centralAir: No  -67882.     4634.     -14.6       0  -76971.  -58794.
```

Potential confounding variables?

One More Example – Prices of Houses in Saratoga Springs, NY

- Want to control for many explanatory variables
  - Notice that you generally don't include interaction terms for the control variables.

```
get_regression_table(mod1)
```

```
# A tibble: 2 × 7
  term          estimate std_error statistic p_value lower_ci upper_ci
  <chr>            <dbl>     <dbl>     <dbl>   <dbl>    <dbl>    <dbl>
1 intercept      254904.     3685.      69.2       0  247676.  262132.
2 centralAir: No -67882.     4634.     -14.6       0  -76971.  -58794.
```

```
mod2 <- lm(price ~ livingArea + age + bathrooms + centralAir, data = SaratogaHouses)
get_regression_table(mod2)
```

```
# A tibble: 5 × 7
  term          estimate std_error statistic p_value lower_ci upper_ci
  <chr>            <dbl>     <dbl>     <dbl>   <dbl>    <dbl>    <dbl>
1 intercept       26749.    7127.       3.75   0       12770.   40728.
2 livingArea         91.7      3.80    24.1    0          84.2     99.1
3 age               -15.7     61.0     -0.257  0.797    -135.     104.
4 bathrooms       20968.    3802.       5.52   0       13511.   28426.
5 centralAir: No -23819.    3648.      -6.53   0      -30974.  -16665.
```

Now let’s shift our focus to estimation and prediction!

Estimation

Typical Inferential Question:

After controlling for the other explanatory variables, what is the range of plausible values for β_j (which summarizes the relationship between y and x_j)?

Confidence Interval Formula:

$$\text{statistic} \pm ME$$

$$\hat{\beta}_j \pm t^* SE(\hat{\beta}_j)$$

```
get_regression_table(mod2)
```

```
# A tibble: 5 × 7
  term            estimate std_error statistic p_value lower_ci upper_ci
  <chr>              <dbl>     <dbl>     <dbl>   <dbl>    <dbl>    <dbl>
1 intercept         26749.    7127.       3.75   0       12770.   40728.
2 livingArea           91.7      3.80    24.1    0          84.2     99.1
3 age                 -15.7     61.0     -0.257  0.797    -135.     104.
4 bathrooms         20968.    3802.       5.52   0       13511.   28426.
5 centralAir: No -23819.    3648.      -6.53   0      -30974.  -16665.
```

Prediction

Typical Inferential Question:

While $\hat{y}$ is a point estimate for y, can we also get an interval estimate for y? In other words, can we get a range of plausible predictions for y?

Two Types of Predictions

Confidence Interval for the Mean Response

→ Defined at given values of the explanatory variables

→ Estimates the average response

→ Centered at $\hat{y}$

→ Smaller SE

Prediction Interval for an Individual Response

→ Defined at given values of the explanatory variables

→ Predicts the response of a single, new observation

→ Centered at $\hat{y}$

→ Larger SE

CI for mean response at a given level of X:

We want to construct a 95% CI for the average price of Saratoga Houses (in 2006!) where the houses meet the following conditions: 1500 square feet, 20 years old, 2 bathrooms, and have central air.

```
house_of_interest <- data.frame(livingArea = 1500, age = 20,
                                bathrooms = 2, centralAir = "Yes")
predict(mod2, house_of_interest, interval = "confidence", level = 0.95)
```

```
       fit      lwr      upr
1 205876.7 199919.1 211834.3
```

- Interpretation: We are 95% confident that the average price of 20 year old, 1500 square feet Saratoga houses with central air and 2 bathrooms is between $199,919 and $211834.

PI for a new Y at a given level of X:

Say we want to construct a 95% PI for the price of an individual house that meets the following conditions: 1500 square feet, 20 years old, 2 bathrooms, and have central air.

Notice: Predicting for a new observation not the mean!

```r
predict(mod2, house_of_interest, interval = "prediction", level = 0.95)
```

```
       fit      lwr      upr
1 205876.7 73884.51 337868.9
```

- Interpretation: For a 20 year old, 1500 square feet Saratoga house with central air and 2 bathrooms, we predict, with 95% confidence, that the price will be between $73,885 and $337,869.

Next Time: Comparing Models and Chi-Squared Tests!

Reminders:

- Lecture Quizzes
  - Last one this week.
  - Plus Extra Credit Lecture Quiz: Due Tues, Dec 5th at 5pm
- Last section this week!
  - Receive the last p-set.
- The material from next Monday's lecture will be on the final and so we will include relevant practice problems on the review sheet.