Chapter 23: Inference for regression

Corinne Riddell

November 16, 2020

### Regression review

Reminder of what we’ve done in regression in Part I of the course:

* Graph the data: scatter plot of the relationship between X and Y
  + Does the relationship look linear? If so, what is the correlation coefficient, ?
  + If not, can we transform X, Y, or both to have a linear relationship on the transformed scale?
* Calculate the line of best fit using lm()
* Using glance() and tidy() from the library broom to summarize the linear model findings
* Interpret the slope () and intercept () parameters
* Interpret the value

### What are the regression “statistics?”

* I changed something about the way I presented the notation on the last slide vs. earlier in the course to give you a clue.
* , , , and are all statistics based on the sample we chose. That is, if we chose a different SRS and re-plotted the data and re-ran the regression, their values would also change.
* When we are specifically interested in the **effect** of some explanatory variable on , then our main interest is often in the underlying parameter , the slope coefficient for .
* For now, we interpret as an **association** rather than a causal effect because we have not learned in this class how to build causal models.
* Today we revisit the output from regression models and apply the inference techniques from Part III of the course to regression.

### Assumptions that require checking for regression inference

* The way we state the assumptions is different from the text book
* Focus on the four assumptions stated on the next slide, **not** the textbook’s version

### Assumptions that require checking for regression inference

1. The relationship between and is linear in the population

### Assumptions that require checking for regression inference

1. varies Normally around the line of best fit. Said differently, the **residuals** vary Normally around the line of best fit.

**Residuals**: Residuals refer to the vertical distance between the line of best fit and the observed value. You can draw a residual for every point on the scatter plot between its value and the line of best fit.

The assumption is that the lengths of these residuals is Normally distributed.

### Assumptions that require checking for regression inference

1. Observations are independent.

Often we can’t check this using a plot, it is based on what we know about the study design.

### Assumptions that require checking for regression inference

1. The standard deviation of the responses is the same for all values of

What does this mean? Consider a line of best fit where the underlying values vary between 1 and 10. This assumption means that the spread of the responses (the values) where is similar to the spread of the responses when .

### Assumptions that require checking for regression inference

1. The relationship between and is linear in the population
2. varies Normally around the line of best fit. That is, the **residuals** vary Normally around the line of best fit.
3. Observations are independent.
4. The standard deviation of the responses is the same for all values of

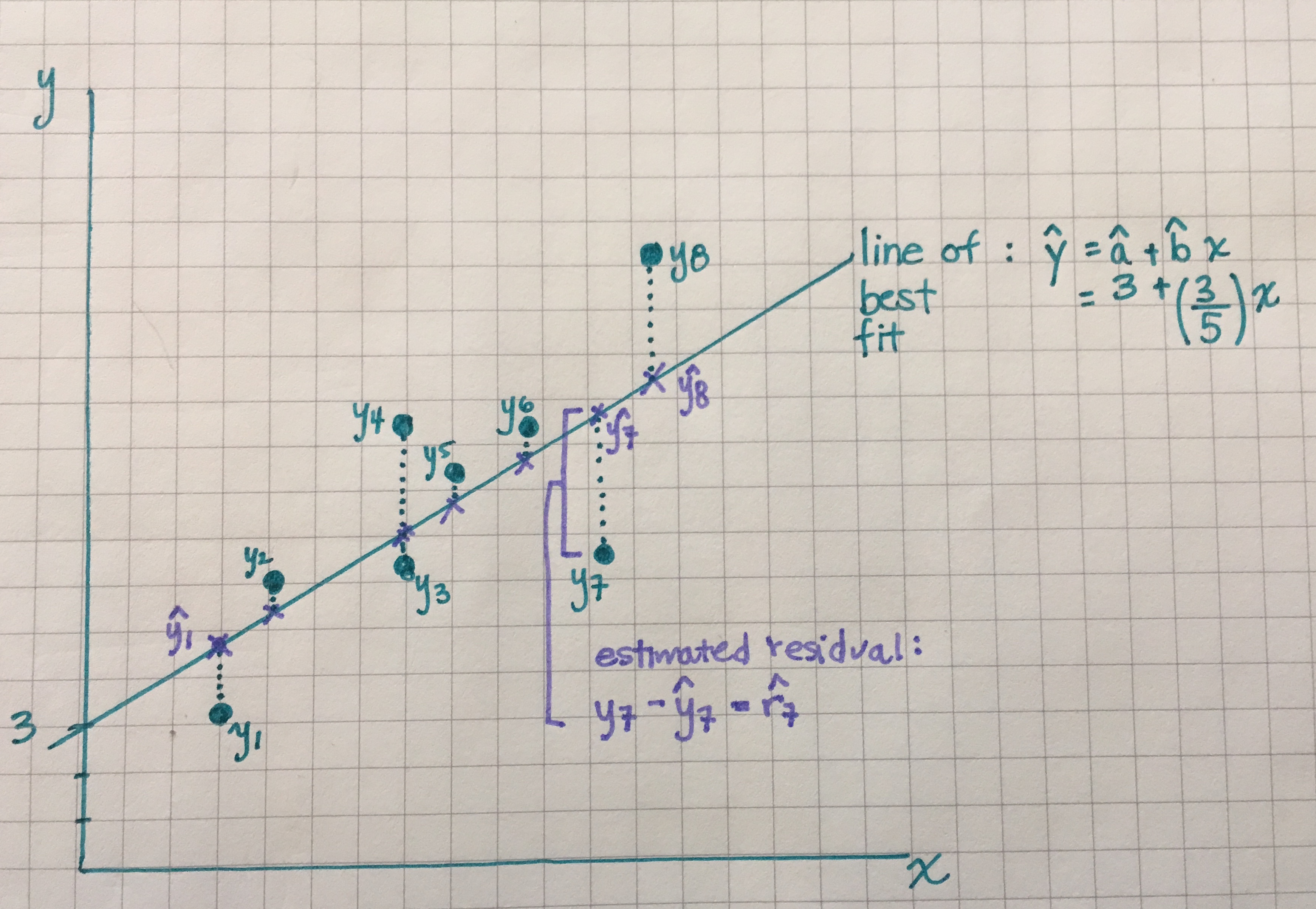
Except for assumption #3, these assumptions can be investigated by examining the **estimated residuals**

We also use these plots to keep an eye out for **outliers**, which can sometimes have a large effect on and

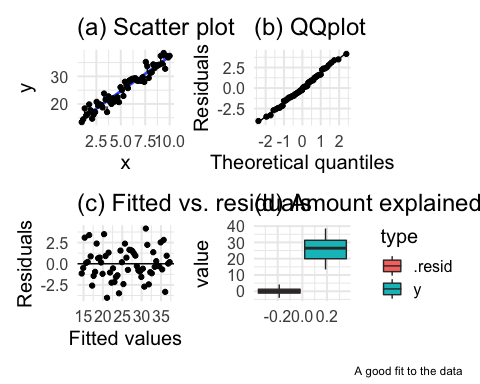
### Terminology needed to understand the assumptions

* **Observed value**:
* **Fitted (or predicted) value**:
* **Estimated residual**:

### Terminology needed to understand the assumptions, visualized



### Example 1: Investigating the assumptions



### Some information about each of the four plots

Plot (a) shows a fitted regression line and the data. The estimated residuals are shown by the dashed lines. We want to see that the residuals are sometimes positive and sometimes negative with no trend in their location

Plot (b) shows a QQ plot of the residuals (to check if they’re Normally distributed)

Plot (c) shows a plot of the fitted values vs. the residuals. We want this to look like a random scatter. If their is a pattern then an assumption has been violated. We will shown examples of this.

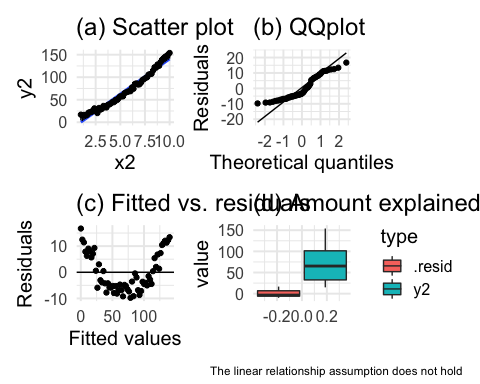
Plot (d) shows a boxplot of the distribution of y vs. the distribution of the residuals. If x does a good job describing y, then the box plot for the residuals will be much shorter because the model fit is good

### Example 1: Investigating the assumptions

* Plot (a): The residuals are sometimes positive and sometimes negative and their magnitude varies randomly as x increases
* Plot (b): The residuals appear to be Normally distributed
* Plot (c): A random scatter - good
* Plot (d): The model fits the data well because the variation in the residuals is much smaller than the variation in the y variable to begin with.

### Example 2: Investigating the assumptions

## `geom\_smooth()` using formula 'y ~ x'

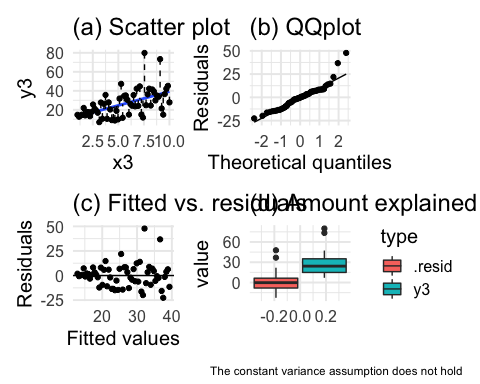


### Example 2: Investigating the assumptions

* Plot (a): While the residuals are small there is a pattern: they start positive, then turn negative and become positive again (as x increases).
* Plot (b): The QQ plot does not support Normality because it is much different from a line
* Plot (c): There is a trend in the residuals vs. fitted. This accentuates the pattern observed in plot (a)
* Plots (a)-(c) all provide evidence against the assumption that a linear fit is the most appropriate one. Because the fit is actually curved, this relationship would require a term in the model, i.e.,
* Plot (d): However, even though the linearity assumption is violated, the linear model still explains a lot of the variation so it still offers insight into explaining y, even if it isn’t the best model

### Example 3: Investigating the assumptions

## `geom\_smooth()` using formula 'y ~ x'



### Example 3: Investigating the assumptions

* Plot (a): This might look okay at first glance, but notice that the magnitude of the residuals is very small for x-values < 2.5, and then it increases
* Plot (b): Also shows some issues in the upper tail
* Plot (c): There is a definite pattern in this plot known as “fanning out”. “Fanning out” is describing the sideways triangle you would see if you were to draw an outline around the set of points. Here, we see that as the fitted value increases, the residuals become further from 0. Fanning out happens when the constant variance assumptions does not hold.

### A note on these diagnostic plots

* If you chose a different sample, the diagnostic plots would change
* Be careful not to over interpret them
* Our goal is to learn about the population, but we only have our one sample

### A note on these diagnostic plots

* Regression procedures are not too sensitive to lack of Normality
* Outliers are important though because they have the potential to have a large effect on the intercept and/or slope terms.

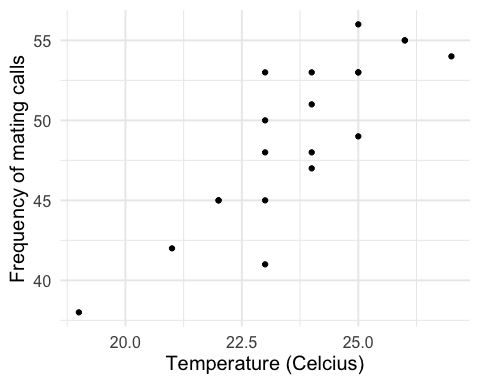
### Example from the text book

Read in the data on frog mating call frequency and temperature:

library(tibble)  
  
frog\_data <- tibble(id = 1:20,  
 temp = c(19, 21, 22, 22, 23, 23, 23, 23, 23,   
 24, 24, 24, 24,  
 25, 25, 25, 25, 26, 26, 27),  
 freq = c(38, 42, 45, 45, 41, 45, 48, 50, 53, 51, 48, 53, 47,  
 53, 49, 56, 53, 55, 55, 54))

### Scatter plot

ggplot(frog\_data, aes(x = temp, y = freq)) +  
 geom\_point() +  
 theme\_minimal(base\_size = 15) +  
 labs(x = "Temperature (Celcius)", y = "Frequency of mating calls")



* Does the relationship look linear?
* Is the relationship positive or negative?

### Run the linear model

Here are the functions we learnt in Part I of the course:

frog\_lm <- lm(formula = freq ~ temp, data = frog\_data)  
tidy(frog\_lm)

## # A tibble: 2 x 5  
## term estimate std.error statistic p.value  
## <chr> <dbl> <dbl> <dbl> <dbl>  
## 1 (Intercept) -6.19 8.24 -0.751 0.462   
## 2 temp 2.33 0.347 6.72 0.00000266

glance(frog\_lm)

## # A tibble: 1 x 12  
## r.squared adj.r.squared sigma statistic p.value df logLik AIC BIC  
## <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>  
## 1 0.715 0.699 2.82 45.2 2.66e-6 1 -48.1 102. 105.  
## # … with 3 more variables: deviance <dbl>, df.residual <int>, nobs <int>

### Check the model diagnostics

frog\_lm\_augment <- augment(frog\_lm)  
  
frog\_lm\_augment %>% select(freq, temp, .fitted, .resid) %>% head()

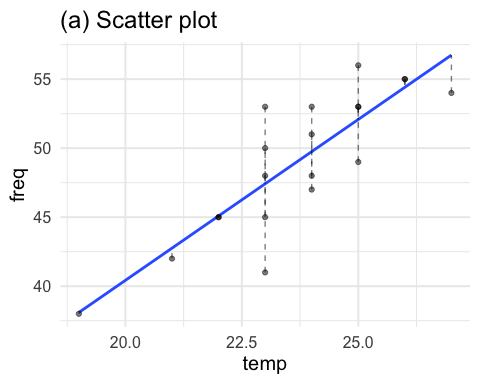
## # A tibble: 6 x 4  
## freq temp .fitted .resid  
## <dbl> <dbl> <dbl> <dbl>  
## 1 38 19 38.1 -0.0952  
## 2 42 21 42.8 -0.757   
## 3 45 22 45.1 -0.0876  
## 4 45 22 45.1 -0.0876  
## 5 41 23 47.4 -6.42   
## 6 45 23 47.4 -2.42

* augment() is another broom function. It augments the original data frame with the residual (.resid) and fitted (.fitted) values, among other values that we won’t talk about now.
* Make sure to know the augment command!

### Check the model diagnostics

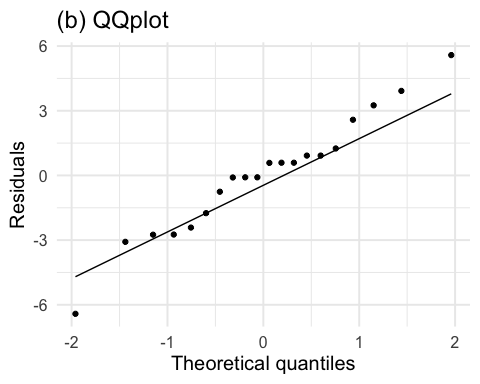
ggplot(frog\_lm\_augment, aes(temp, freq)) +  
 geom\_smooth(method = "lm", se = F) +  
 geom\_point(alpha = 0.5) +  
 geom\_segment(aes(xend = temp, yend = .fitted), lty = 2, alpha = 0.5) +  
 theme\_minimal(base\_size = 15) +  
 labs(title = "(a) Scatter plot")

## `geom\_smooth()` using formula 'y ~ x'



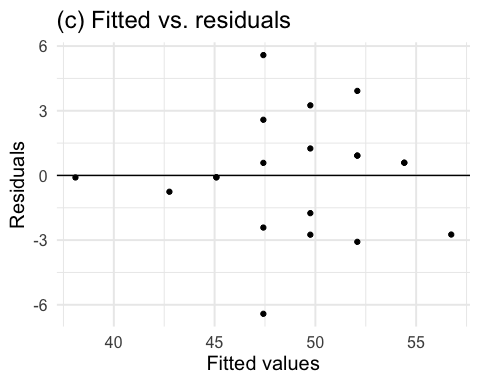
### Check the model diagnostics

# QQ plot  
 ggplot(frog\_lm\_augment, aes(sample = .resid)) +   
 geom\_qq() +   
 geom\_qq\_line() +  
 theme\_minimal(base\_size = 15) +  
 labs(y = "Residuals", x = "Theoretical quantiles", title = "(b) QQplot")



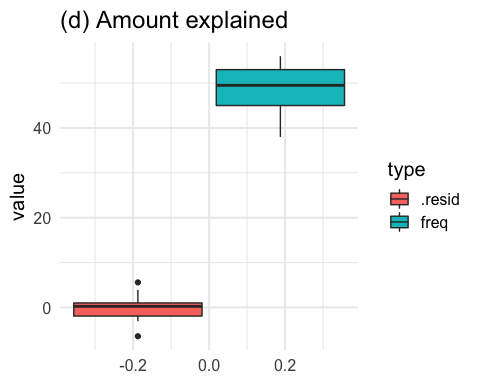
### Check the model diagnostics

## Fitted vs. residuals  
 ggplot(frog\_lm\_augment, aes(y = .resid, x = .fitted)) +   
 geom\_point() +   
 theme\_minimal(base\_size = 15) +  
 geom\_hline(aes(yintercept = 0)) +  
 labs(y = "Residuals", x = "Fitted values", title = "(c) Fitted vs. residuals")



### Check the model diagnostics

## Amount explained  
frog\_gather <- frog\_lm\_augment %>% select(freq, .resid) %>%  
 gather(key = "type", value = "value", freq, .resid)  
  
ggplot(frog\_gather, aes(y = value)) +   
 geom\_boxplot(aes(fill = type)) +  
 theme\_minimal(base\_size = 15) +  
 labs(title = "(d) Amount explained")



### Example for you to work on

The breaking strength of steel bolts is measured by subjecting a bolt to increasing (lateral) force and determining the force at which the bolt breaks. This force is called the breaking strength; it depends on the diameter of the bolt and the material the bolt is composed of. There is variability in breaking strengths: Two bolts of the same dimension and aterial will generally break at different forces. Understanding the distribution of breaking strengths is very important in construction and other areas.

The data below show the breaking strengths of six steel bolts at each of five different bolt diameters.

diameter <- c(rep(0.10, 6), rep(0.20, 6), rep(0.3, 6), rep(0.4, 6), rep(0.5, 6))  
breaking\_strength <- c(1.62, 1.73, 1.70, 1.66, 1.74, 1.72,   
 1.71, 1.78, 1.79, 1.86, 1.70, 1.84,  
 1.86, 1.86, 1.90, 1.95, 1.96, 2.00,  
 2.14, 2.07, 2.11, 2.18, 2.17, 2.07,   
 2.45, 2.42, 2.33, 2.36, 2.38, 2.31)  
  
bolt\_data <- tibble(diameter, breaking\_strength)

### Example for you to work on

1. Which variable is the response and which variable is the explanatory?
2. Fit a linear model to these data and add the residuals and fitted values to a data frame alongside the original data.
3. Make a plot of the residuals vs. the fitted values. Comment on what you see. Is an assumption violated?
4. Add a quadratic term to the dataset. You can do this using mutate(). Then re-run the linear model using this template: lm(formula = y ~ x + x\_squared, dat = your\_data)
5. Re-make the residual vs. fitted plot after this model and comment on the difference.

### Example’s solutions

1. Which variable is the response and which variable is the explanatory?

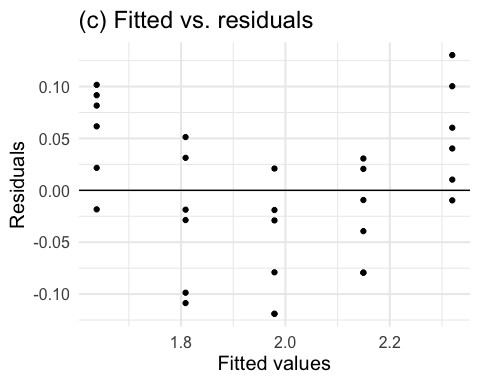
Answer: The breaking strength is the response variable and the diameter is the explanatory variable. This is because the question says that breaking strength depends on the diameter of the bolt.

1. Fit a linear model to these data and add the residuals and fitted values to a data frame alongside the original data.

bolt\_model <- lm(breaking\_strength ~ diameter, bolt\_data)  
augmented\_bolt\_data <- augment(bolt\_model)

1. Make a plot of the residuals vs. the fitted values. Comment on what you see. Is an assumption violated?

ggplot(data = augmented\_bolt\_data, aes(x = .fitted, y = .resid)) + geom\_point() +   
 theme\_minimal(base\_size = 15) +  
 geom\_hline(aes(yintercept = 0)) +  
 labs(y = "Residuals", x = "Fitted values", title = "(c) Fitted vs. residuals")



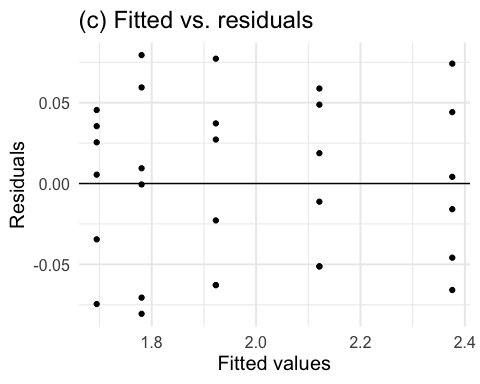
There is a curved pattern in this plot – the residuals are generally positive, then generally negative, followed by positive again. This indicates that the linearity assumption is violated.

1. Add a quadratic term to the dataset. You can do this using mutate(). Then re-run the linear model using this template: lm(formula = y ~ x + x\_squared, dat = your\_data)

#add a squared term to the dataset  
bolt\_data <- bolt\_data %>% mutate(diameter2 = diameter^2)  
  
bolt\_model\_2 <- lm(formula = breaking\_strength ~ diameter + diameter2, dat = bolt\_data)  
  
augment\_bolt\_data2 <- augment(bolt\_model\_2)

1. Re-make the residual vs. fitted plot after this model and comment on the difference.

ggplot(data = augment\_bolt\_data2, aes(x = .fitted, y = .resid)) + geom\_point() +   
 theme\_minimal(base\_size = 15) +  
 geom\_hline(aes(yintercept = 0)) +  
 labs(y = "Residuals", x = "Fitted values", title = "(c) Fitted vs. residuals")



The pattern from the previous plot cannot be seen in this plot, showing that the model including the quadratic term is a better fit to these data.