

Statistics is Everywhere

Recap of part 1 (chapters
3,4, lectures 4,5,6)

Regression and assumptions
needed for inference

testing and CI with linear
regression models

Hypothesis testing for
regression

Confidence intervals for
regression coefficient

Inference for prediction

Inference for linear regressions

Statistics is Everywhere

Recap of part 1 (chapters
3,4, lectures 4,5,6)

Regression and assumptions
needed for inference

testing and CI with linear
regression models

Hypothesis testing for
regression

Confidence intervals for
regression coefficient

Inference for prediction

Statistics is Everywhere

Statistics is Everywhere

Inference for linear
regressions

Statistics is Everywhere

Recap of part 1 (chapters
3,4, lectures 4,5,6)

Regression and assumptions
needed for inference

testing and CI with linear
regression models

Hypothesis testing for
regression

Confidence intervals for
regression coefficient

Inference for prediction

Concussions



Inference for linear regressions

Statistics is Everywhere

Recap of part 1 (chapters 3,4, lectures 4,5,6)

Regression and assumptions needed for inference

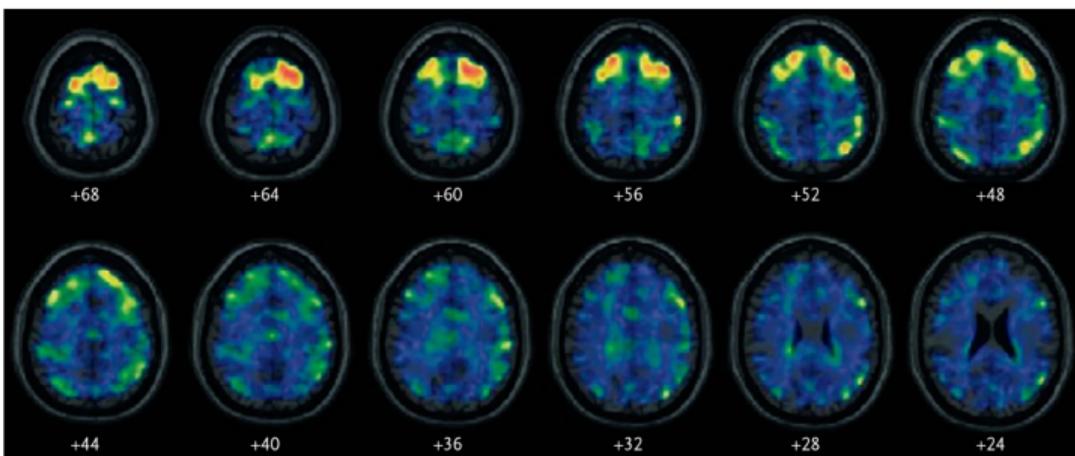
testing and CI with linear regression models

Hypothesis testing for regression

Confidence intervals for regression coefficient

Inference for prediction

Abnormal Levels of a Protein Linked to C.T.E. Found in N.F.L Players' Brains, Study Shows



NYtimes article

Statistics is Everywhere

Recap of part 1 (chapters 3,4, lectures 4,5,6)

Regression and assumptions
needed for inference

testing and CI with linear
regression models

Hypothesis testing for
regression

Confidence intervals for
regression coefficient

Inference for prediction

Tau Positron-Emission Tomography in Former National Football League Players

Robert A. Stern, Ph.D., Charles H. Adler, M.D., Ph.D., Kewei Chen, Ph.D., Michael Navitsky, M.S., Ji Luo, M.S., David W. Dodick, M.D., Michael L. Alosco, Ph.D., Yorghos Tripodis, Ph.D., Dhruvan D. Goradia, Ph.D., Brett Martin, M.S., Diego Mastrolia, Ph.D., Nathan G. Fritts, B.A., *et al.*

The authors of the study and outside experts stressed that such tau imaging is far from a diagnostic test for C.T.E., which is likely years away and could include other markers, from blood and spinal fluid.

The results of the study, published in The New England Journal of Medicine on Wednesday, are considered preliminary, but constitute a first step toward developing a clinical test to determine the presence of C.T.E. in living players, as well as early signs and potential risk.

Statistics is Everywhere

Recap of part 1 (chapters 3,4, lectures 4,5,6)

Regression and assumptions needed for inference

testing and CI with linear regression models

Hypothesis testing for regression

Confidence intervals for regression coefficient

Inference for prediction

STATISTICAL ANALYSIS Between-group comparisons of age, years of education, and MMSE scores were analyzed with [Mann–Whitney U tests](#). Group differences in race were analyzed with the use of chi-square tests. For between-group comparisons of amyloid-beta plaque burden, chi-square tests were used to compare the proportion of participants with a positive florbetapir PET, and [t-tests](#) were used to compare the mean cortical:cerebellar florbetapir standard uptake value ratio (SUVR, the ratio of radioactivity in a cerebral region to that in the cerebellum as a reference) between the groups.

Statistics is Everywhere

Recap of part 1 (chapters 3,4, lectures 4,5,6)

Regression and assumptions needed for inference

testing and CI with linear regression models

Hypothesis testing for regression

Confidence intervals for regression coefficient

Inference for prediction

From the article

Statistics is Everywhere

Recap of part 1 (chapters 3,4, lectures 4,5,6)

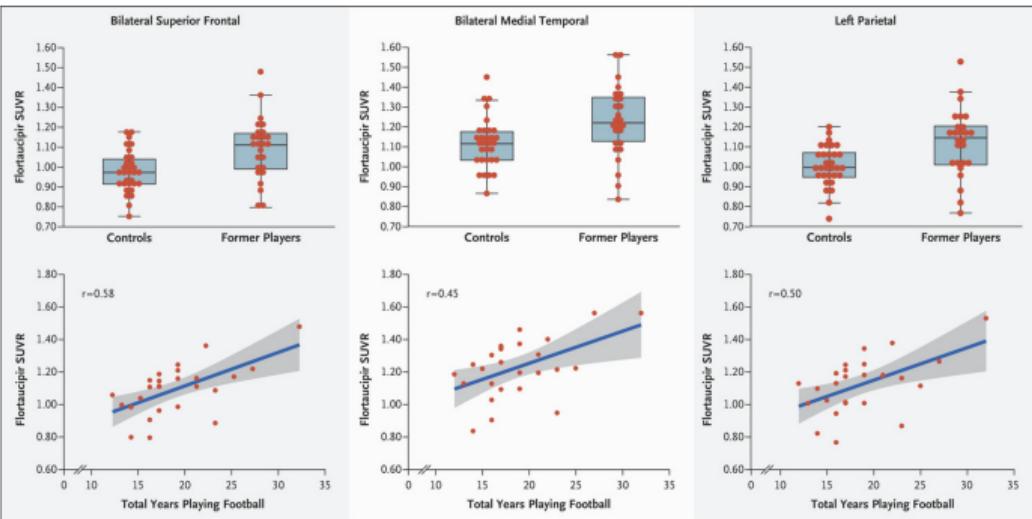
Regression and assumptions
needed for inference

testing and CI with linear
regression models

Hypothesis testing for
regression

Confidence intervals for
regression coefficient

Inference for prediction



So far in part 2:

- ▶ continuous outcomes by categories (ie continuous outcome, categorical predictor)

Next up:

- ▶ continuous outcomes with continuous predictors
- ▶ a brief touch on multiple predictor variables with one continuous outcome

Statistics is Everywhere

Recap of part 1 (chapters 3,4, lectures 4,5,6)

Regression and assumptions needed for inference

testing and CI with linear regression models

Hypothesis testing for regression

Confidence intervals for regression coefficient

Inference for prediction

Statistics is Everywhere

Recap of part 1 (chapters
3,4, lectures 4,5,6)

Regression and assumptions
needed for inference

testing and CI with linear
regression models

Hypothesis testing for
regression

Confidence intervals for
regression coefficient

Inference for prediction

Recap of part 1 (chapters 3,4, lectures 4,5,6)

Reminder of what we've done with continuous vs continuous variables 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, \hat{r} ?
 - ▶ If not, can we transform X, Y, or both to have a linear relationship on the transformed scale?
- ▶ Fit the line of best fit using `lm()`
- ▶ Using `glance()` and `tidy()` from the library `broom` to summarize the linear model findings
- ▶ Interpret the slope (\hat{b}) and intercept (\hat{a}) parameters
- ▶ Interpret the \hat{r}^2 value

Statistics is Everywhere

Recap of part 1 (chapters
3,4, lectures 4,5,6)Regression and assumptions
needed for inferencetesting and CI with linear
regression modelsHypothesis testing for
regressionConfidence intervals for
regression coefficient

Inference for prediction

Recap: Visualizing continuous-continuous relationships

Inference for linear regressions

- ▶ Scatterplots are a good way to visualize a relationship between two continuous variables
- ▶ When we look at a scatterplot we want to evaluate:
 - ▶ The overall Pattern of the dots
 - ▶ Any notable exceptions to the pattern
 - ▶ Direction (positive or negative)
 - ▶ Form (straight line or curved)
 - ▶ Strength (how closely the points follow a line)
 - ▶ Are there any obvious outliers

Statistics is Everywhere

Recap of part 1 (chapters 3,4, lectures 4,5,6)

Regression and assumptions needed for inference

testing and CI with linear regression models

Hypothesis testing for regression

Confidence intervals for regression coefficient

Inference for prediction

Scatterplot Syntax in R

Inference for linear
regressions

```
name of plot <- ggplot(data = dataset, aes(x = xvariable, y = yvariable)) +  
geom_point(na.rm=TRUE) + theme_minimal(base_size = 15)+  
labs(x = " xlabel", y = " ylabel", title = " Title")
```

Statistics is Everywhere

Recap of part 1 (chapters
3,4, lectures 4,5,6)

Regression and assumptions
needed for inference

testing and CI with linear
regression models

Hypothesis testing for
regression

Confidence intervals for
regression coefficient

Inference for prediction

Remember the Manatees?

Manatee data set from your textbook:

```
mana_data <- read_csv("Ch03_Manatee-deaths.csv")  
head(mana_data)
```

```
## # A tibble: 6 x 3  
##   year powerboats deaths  
##   <dbl>     <dbl>   <dbl>  
## 1 1977       447     13  
## 2 1987       645     39  
## 3 1997       755     54  
## 4 2007      1027     73  
## 5 1978       460     21  
## 6 1988       675     43
```

```
mana_scatter <- ggplot(data = mana_data, aes(x = powerboats, y = deaths)) +  
  geom_point() + theme_minimal(base_size = 15)
```

Statistics is Everywhere

Recap of part 1 (chapters
3,4, lectures 4,5,6)

Regression and assumptions
needed for inference

testing and CI with linear
regression models

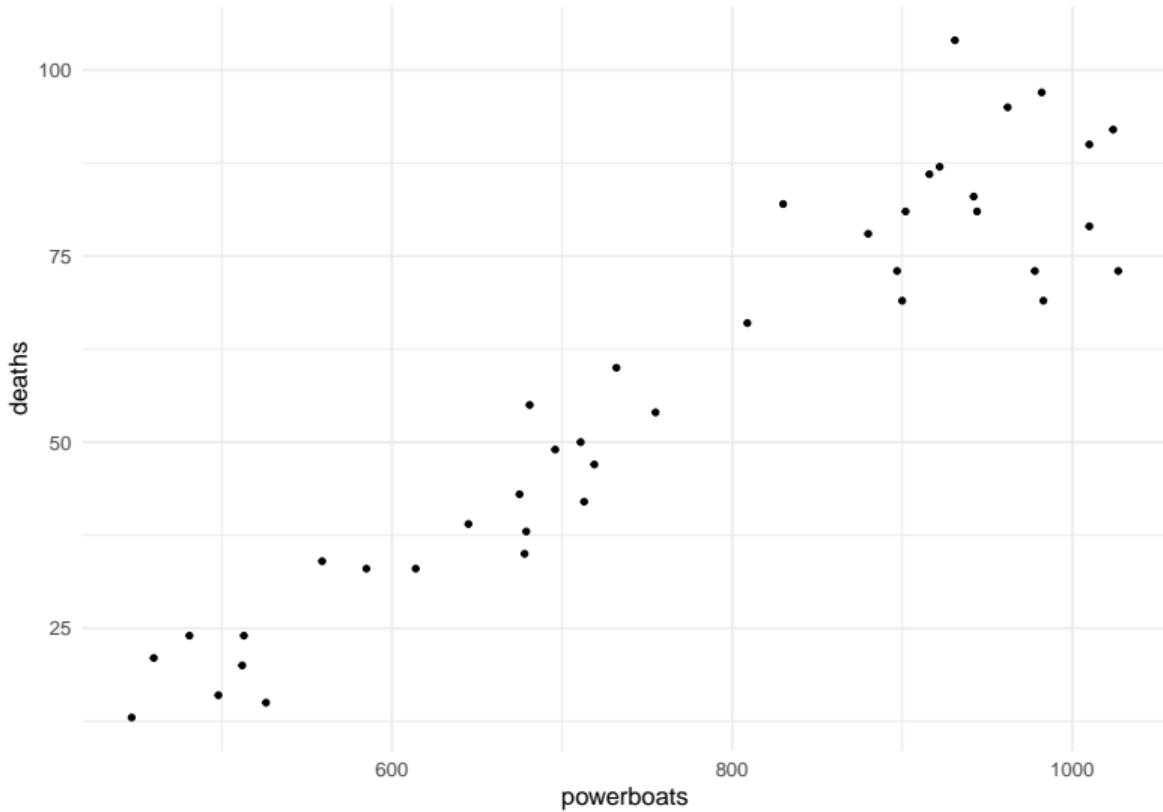
Hypothesis testing for
regression

Confidence intervals for
regression coefficient

Inference for prediction

Remember the Manatees?

Inference for linear regressions



Statistics is Everywhere

Recap of part 1 (chapters 3,4, lectures 4,5,6)

Regression and assumptions needed for inference

testing and CI with linear regression models

Hypothesis testing for regression

Confidence intervals for regression coefficient

Inference for prediction

Recap: Pearson's

Inference for linear regressions

Statistics is Everywhere

Recap of part 1 (chapters 3,4, lectures 4,5,6)

Regression and assumptions needed for inference

testing and CI with linear regression models

Hypothesis testing for regression

Confidence intervals for regression coefficient

Inference for prediction

- ▶ Pearson's correlation coefficient measures linear association between two continuous variables
- ▶ It characterizes the extent to which the points cluster around a straight line
- ▶ the correlation coefficient can take on any value between -1 to 1 (inclusive)
 - ▶ -1: A perfect, negative linear association
 - ▶ 1: A perfect, positive linear association
 - ▶ 0: No linear association
- ▶ usually we use ρ when referring to the correlation in a population and r when referring to the correlation observed in a sample

Recap: Pearson's

Inference for linear regressions

```
mana_cor <- mana_data %>%
  summarize(corr_mana = cor(powerboats, deaths))
mana_cor
```

```
## # A tibble: 1 x 1
##   corr_mana
##       <dbl>
## 1     0.945
```

Statistics is Everywhere

Recap of part 1 (chapters 3,4, lectures 4,5,6)

Regression and assumptions needed for inference

testing and CI with linear regression models

Hypothesis testing for regression

Confidence intervals for regression coefficient

Inference for prediction

lm() of manatee deaths and powerboat purchases

Statistics is Everywhere

Recap of part 1 (chapters 3,4, lectures 4,5,6)

Regression and assumptions needed for inference

testing and CI with linear regression models

Hypothesis testing for regression

Confidence intervals for regression coefficient

Inference for prediction

Calculate the line of best fit:

```
mana_lm <- lm(deaths ~ powerboats, mana_data)
# we use the package broom to look at the output of the linear model
tidy(mana_lm)
```

```
## # A tibble: 2 x 5
##   term      estimate std.error statistic p.value
##   <chr>     <dbl>     <dbl>     <dbl>    <dbl>
## 1 (Intercept) -46.8      6.03     -7.75 2.43e- 9
## 2 powerboats    0.136     0.00764    17.8  5.21e-20
```

Interpreting the intercept and slope

```
## # A tibble: 2 x 5
##   term      estimate std.error statistic p.value
##   <chr>     <dbl>     <dbl>     <dbl>     <dbl>
## 1 (Intercept) -46.8      6.03    -7.75 2.43e- 9
## 2 powerboats     0.136    0.00764    17.8  5.21e-20
```

- ▶ Intercept: The predicted number of deaths if there were no powerboats.
- ▶ Slope: A one unit change in the number of powerboats registered (X 1,000) is associated with an increase of manatee deaths of 0.1358. That is, an increase in the number of powerboats registered by 1,000 is association with 0.1358 more manatee deaths.

Statistics is Everywhere

Recap of part 1 (chapters 3,4, lectures 4,5,6)

Regression and assumptions needed for inference

testing and CI with linear regression models

Hypothesis testing for regression

Confidence intervals for regression coefficient

Inference for prediction

Getting the R-squared from your model

When we run a linear model, the r-squared is also calculated. Here is how to see the r-squared for the manatee data:

```
glance(mana_lm)
```

```
## # A tibble: 1 x 12
##   r.squared adj.r.squared sigma statistic p.value    df logLik
##       <dbl>          <dbl>  <dbl>     <dbl>    <dbl>    <dbl>  <dbl>
## 1     0.893         0.890  8.82      316. 5.21e-20     1  -143.
## # ... with 3 more variables: deviance <dbl>, df.residual <int>,
## #   nobs <int>
```

Focus on:

- ▶ Column called `r.squared` values only.
- ▶ Interpretation of r-squared: The fraction of the variation in the values of y that is explained by the line of best fit.

Statistics is Everywhere

Recap of part 1 (chapters 3,4, lectures 4,5,6)

Regression and assumptions needed for inference

testing and CI with linear regression models

Hypothesis testing for regression

Confidence intervals for regression coefficient

Inference for prediction

AIC BI

<dbl> <dbl>

292. 297

nobs <int>

Correlation vs R Squared

```
mana_cor <- mana_data %>%  
  summarize(corr_mana = cor(powerboats, deaths))  
mana_cor
```

```
## # A tibble: 1 x 1  
##   corr_mana  
##       <dbl>  
## 1     0.945
```

```
glance(mana_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     0.893          0.890  8.82     316.  5.21e-20     1  -143.  292.  297.  
## # ... with 3 more variables: deviance <dbl>, df.residual <int>, nobs <int>
```

Statistics is Everywhere

Recap of part 1 (chapters
3,4, lectures 4,5,6)Regression and assumptions
needed for inferencetesting and CI with linear
regression modelsHypothesis testing for
regressionConfidence intervals for
regression coefficient

Inference for prediction

Statistics is Everywhere

Recap of part 1 (chapters
3,4, lectures 4,5,6)

**Regression and assumptions
needed for inference**

testing and CI with linear
regression models

Hypothesis testing for
regression

Confidence intervals for
regression coefficient

Inference for prediction

Regression and assumptions needed for inference

What are the regression “statistics?”

When we are estimating values from a sample, we often put a “hat” on them.

- ▶ \hat{e} , \hat{r}^2 , \hat{a} , and \hat{b} are all statistics based on the sample we chose. That is, if we chose a different SRS and re-plotted the data and re-run the regression, we would expect their values to change somewhat.
- ▶ When we are specifically interested in the **effect** of some explanatory variable x on y , then our main interest is often in the underlying parameter b , the slope coefficient for x .
- ▶ For now, we interpret b 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.

Statistics is Everywhere

Recap of part 1 (chapters
3,4, lectures 4,5,6)Regression and assumptions
needed for inferencetesting and CI with linear
regression modelsHypothesis testing for
regressionConfidence intervals for
regression coefficient

Inference for prediction

Assumptions that require checking for regression inference

Inference for linear regressions

Statistics is Everywhere

Recap of part 1 (chapters 3,4, lectures 4,5,6)

Regression and assumptions needed for inference

testing and CI with linear regression models

Hypothesis testing for regression

Confidence intervals for regression coefficient

Inference for prediction

- ▶ 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 x and y is linear in the population
2. y varies Normally about the line of best fit. That is, the residuals vary Normally around the line of best fit.
3. Observations are independent. Often we can't check this using a plot, it is based on what we know about the study design.
4. The standard deviation of the responses is the same for all values of x

Except for #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 larger effect on \hat{a} and \hat{b}

Statistics is Everywhere
Recap of part 1 (chapters 3,4, lectures 4,5,6)
Regression and assumptions needed for inference
testing and CI with linear regression models
Hypothesis testing for regression
Confidence intervals for regression coefficient
Inference for prediction

Terminology needed to understand the assumptions

Observed value: y

Fitted value: $\hat{y} = \hat{a} + \hat{b}x$

Estimated residuals:

\hat{e} = observed value - fitted value

$\hat{e} = y - (\hat{a} + \hat{b}x)$

Statistics is Everywhere

Recap of part 1 (chapters
3,4, lectures 4,5,6)

Regression and assumptions
needed for inference

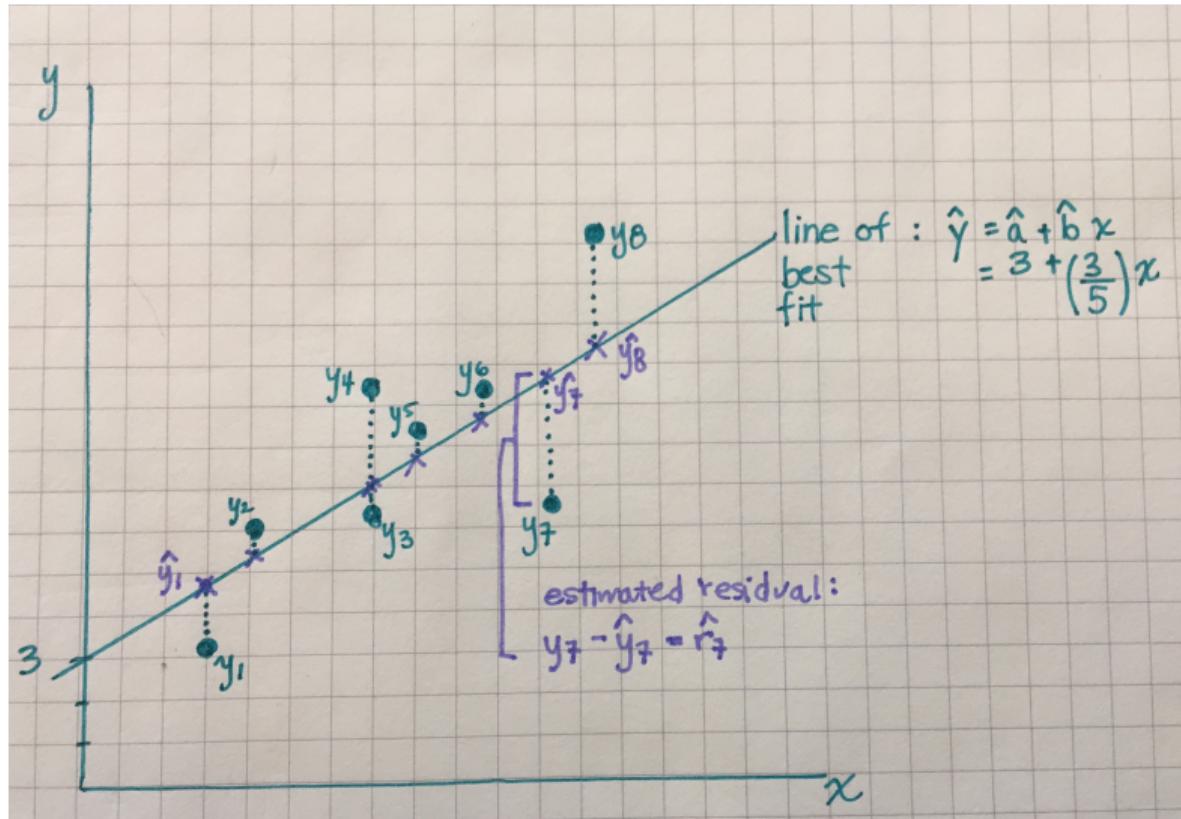
testing and CI with linear
regression models

Hypothesis testing for
regression

Confidence intervals for
regression coefficient

Inference for prediction

Terminology needed to understand the assumptions, visualized



Statistics is Everywhere

Recap of part 1 (chapters 3,4, lectures 4,5,6)

Regression and assumptions needed for inference

testing and CI with linear regression models

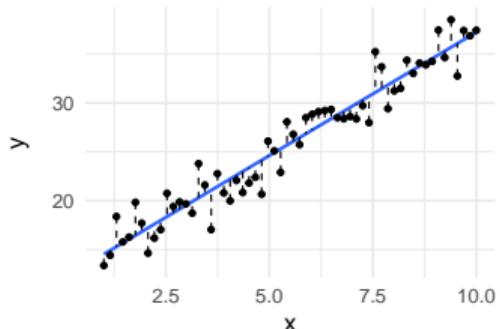
Hypothesis testing for regression

Confidence intervals for regression coefficient

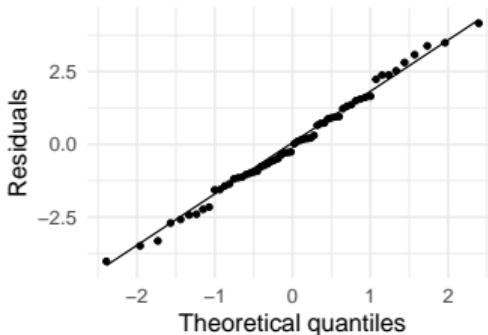
Inference for prediction

Example 1: Investigating the assumptions

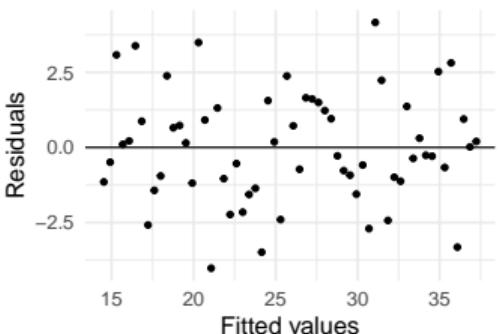
(a) Scatter plot



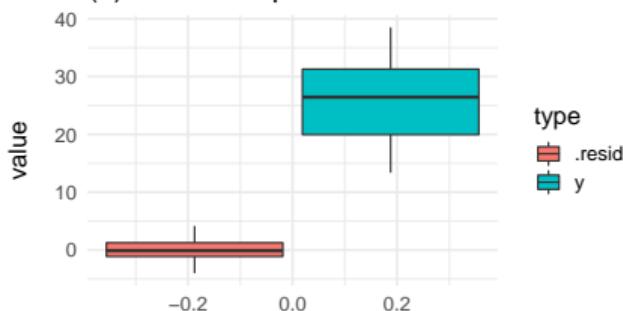
(b) QQplot



(c) Fitted vs. residuals



(d) Amount explained



A good fit to the data

Statistics is Everywhere

Recap of part 1 (chapters 3,4, lectures 4,5,6)

Regression and assumptions needed for inference

testing and CI with linear regression models

Hypothesis testing for regression

Confidence intervals for regression coefficient

Inference for prediction

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 there is a pattern then an assumption has been violated. We will show 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

Statistics is Everywhere

Recap of part 1 (chapters
3,4, lectures 4,5,6)Regression and assumptions
needed for inferencetesting and CI with linear
regression modelsHypothesis testing for
regressionConfidence intervals for
regression coefficient

Inference for prediction

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.

Statistics is Everywhere

Recap of part 1 (chapters
3,4, lectures 4,5,6)

Regression and assumptions
needed for inference

testing and CI with linear
regression models

Hypothesis testing for
regression

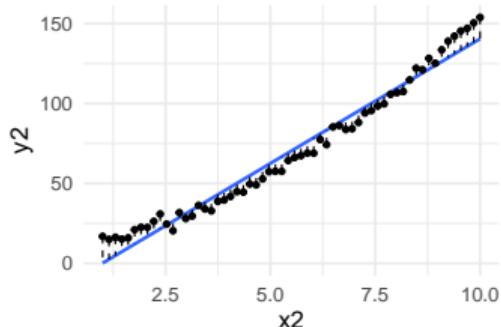
Confidence intervals for
regression coefficient

Inference for prediction

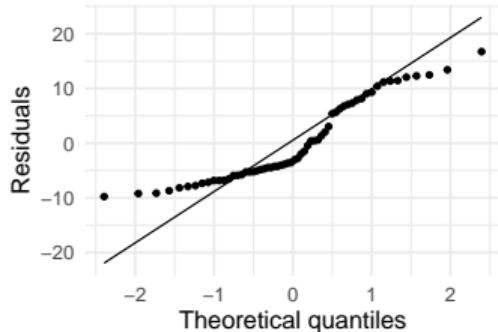
Example 2: Investigating the assumptions

```
## `geom_smooth()` using formula 'y ~ x'
```

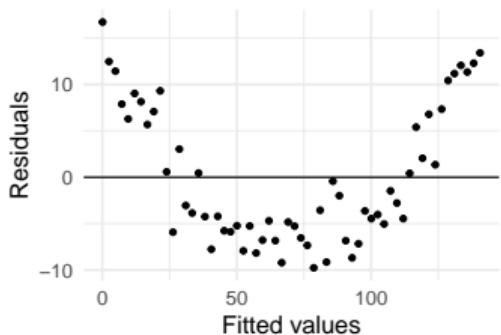
(a) Scatter plot



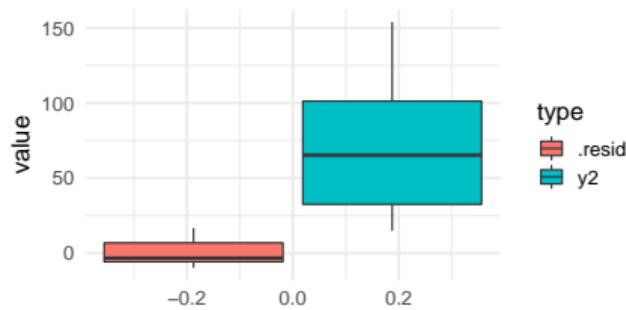
(b) QQplot



(c) Fitted vs. residuals



(d) Amount explained



The linear systematic assumption does not hold

Statistics is Everywhere

Recap of part 1 (chapters 3,4, lectures 4,5,6)

Regression and assumptions needed for inference

testing and CI with linear regression models

Hypothesis testing for regression

Confidence intervals for regression coefficient

Inference for prediction

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 x^2 term in the model, i.e., $\hat{y} = \hat{a} + \hat{b}x + \hat{c}x^2$
- ▶ 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

Statistics is Everywhere

Recap of part 1 (chapters 3,4, lectures 4,5,6)

Regression and assumptions needed for inference

testing and CI with linear regression models

Hypothesis testing for regression

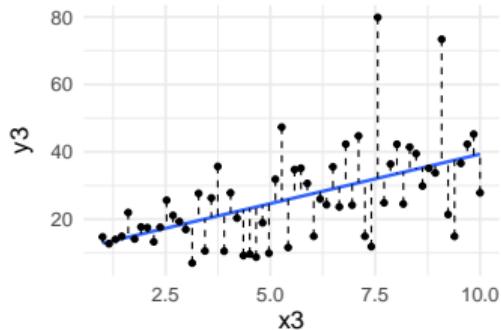
Confidence intervals for regression coefficient

Inference for prediction

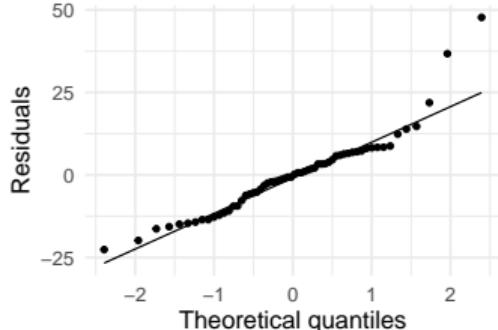
Example 3: Investigating the assumptions

```
## `geom_smooth()` using formula 'y ~ x'
```

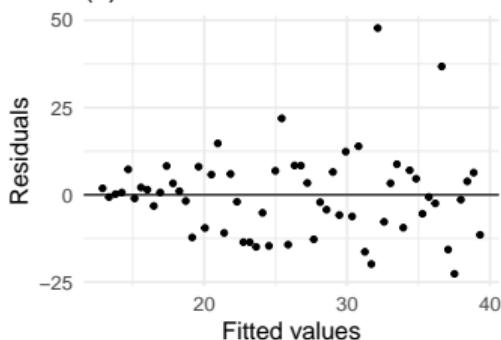
(a) Scatter plot



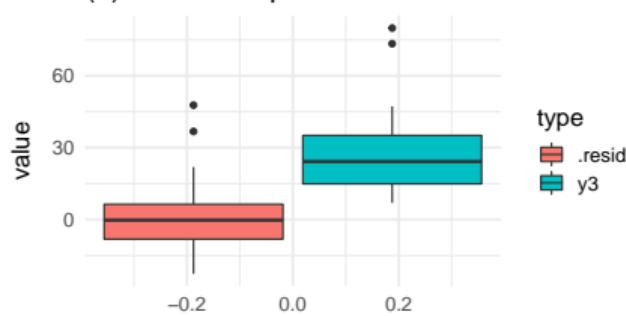
(b) QQplot



(c) Fitted vs. residuals



(d) Amount explained



Statistics is Everywhere

Recap of part 1 (chapters 3,4, lectures 4,5,6)

Regression and assumptions needed for inference

testing and CI with linear regression models

Hypothesis testing for regression

Confidence intervals for regression coefficient

Inference for prediction

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”. Here, we see that as the fitted value increases, the residuals become further from 0.

Statistics is Everywhere

Recap of part 1 (chapters
3,4, lectures 4,5,6)

Regression and assumptions
needed for inference

testing and CI with linear
regression models

Hypothesis testing for
regression

Confidence intervals for
regression coefficient

Inference for prediction

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

Statistics is Everywhere

Recap of part 1 (chapters
3,4, lectures 4,5,6)

Regression and assumptions
needed for inference

testing and CI with linear
regression models

Hypothesis testing for
regression

Confidence intervals for
regression coefficient

Inference for prediction

A note on these diagnostic plots

Statistics is Everywhere

Recap of part 1 (chapters
3,4, lectures 4,5,6)

Regression and assumptions
needed for inference

testing and CI with linear
regression models

Hypothesis testing for
regression

Confidence intervals for
regression coefficient

Inference for prediction

- ▶ 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))
```

Statistics is Everywhere

Recap of part 1 (chapters
3,4, lectures 4,5,6)

Regression and assumptions
needed for inference

testing and CI with linear
regression models

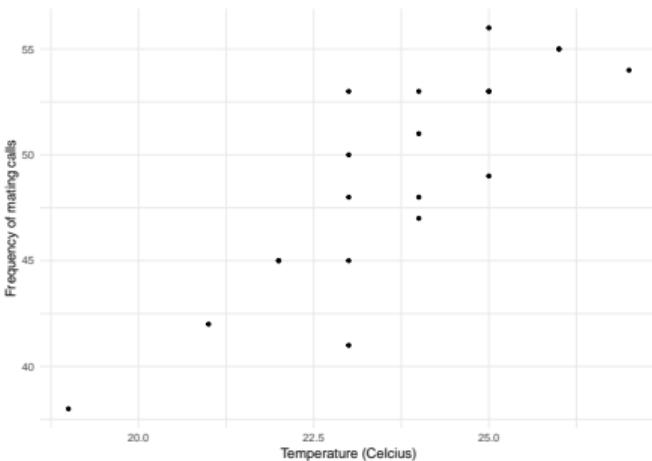
Hypothesis testing for
regression

Confidence intervals for
regression coefficient

Inference for prediction

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?

Statistics is Everywhere

Recap of part 1 (chapters 3,4, lectures 4,5,6)

Regression and assumptions needed for inference

testing and CI with linear regression models

Hypothesis testing for regression

Confidence intervals for regression coefficient

Inference for prediction

Run the linear model

```
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
##   <dbl>        <dbl> <dbl>     <dbl>    <dbl> <dbl> <dbl> <dbl> <dbl>
## 1 0.715        0.699  2.82      45.2  0.00000266     1 -48.1  102.  1
## # ... with 3 more variables: deviance <dbl>, df.residual <int>, nobs <int>
```

Statistics is Everywhere

Recap of part 1 (chapters 3,4, lectures 4,5,6)

Regression and assumptions needed for inference

testing and CI with linear regression models

Hypothesis testing for regression

Confidence intervals for regression coefficient

Inference for prediction

Check the model diagnostics

```
frog_lm_augment <- augment(frog_lm)
```

- ▶ `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!

Statistics is Everywhere

Recap of part 1 (chapters
3,4, lectures 4,5,6)

Regression and assumptions
needed for inference

testing and CI with linear
regression models

Hypothesis testing for
regression

Confidence intervals for
regression coefficient

Inference for prediction

Check the model diagnostics

Inference for linear regressions

Statistics is Everywhere

Recap of part 1 (chapters 3,4, lectures 4,5,6)

Regression and assumptions needed for inference

testing and CI with linear regression models

Hypothesis testing for regression

Confidence intervals for regression coefficient

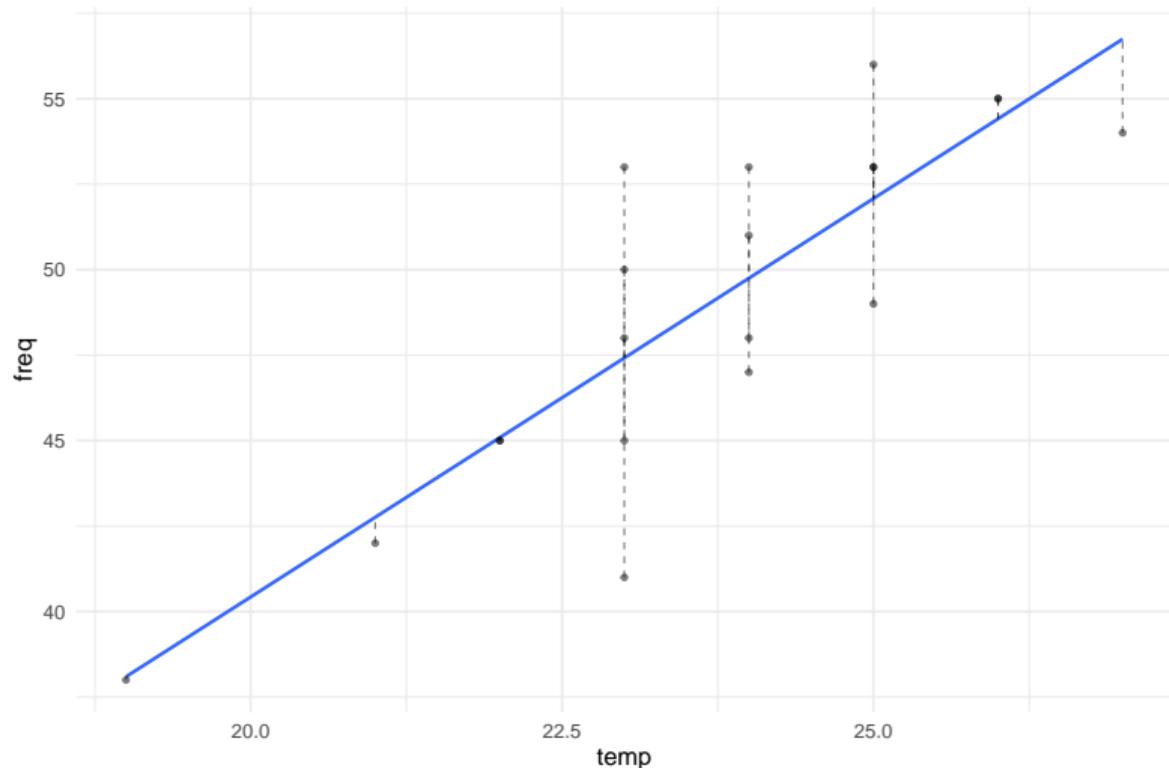
Inference for prediction

```
frog_resid<- 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")
```

Check the model diagnostics: residuals

```
## `geom_smooth()` using formula 'y ~ x'
```

(a) Scatter plot



Statistics is Everywhere

Recap of part 1 (chapters 3,4, lectures 4,5,6)

Regression and assumptions needed for inference

testing and CI with linear regression models

Hypothesis testing for regression

Confidence intervals for regression coefficient

Inference for prediction

Check the model diagnostics

Inference for linear
regressions

Statistics is Everywhere

Recap of part 1 (chapters
3,4, lectures 4,5,6)

Regression and assumptions
needed for inference

testing and CI with linear
regression models

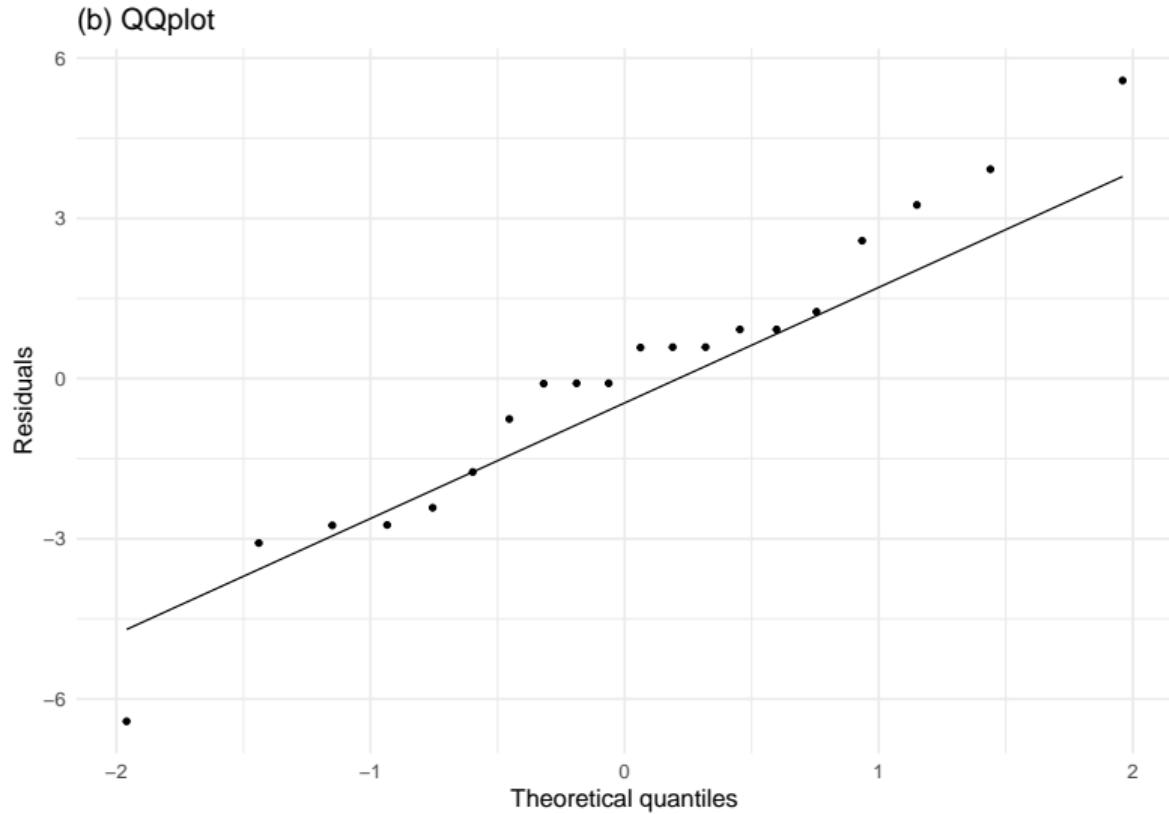
Hypothesis testing for
regression

Confidence intervals for
regression coefficient

Inference for prediction

```
# QQ plot
frog_qq<-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



Statistics is Everywhere

Recap of part 1 (chapters
3,4, lectures 4,5,6)

Regression and assumptions
needed for inference

testing and CI with linear
regression models

Hypothesis testing for
regression

Confidence intervals for
regression coefficient

Inference for prediction

Check the model diagnostics

Inference for linear regressions

```
## Fitted vs. residuals
frog_fitresid<-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. residual")
```

Statistics is Everywhere

Recap of part 1 (chapters 3,4, lectures 4,5,6)

Regression and assumptions needed for inference

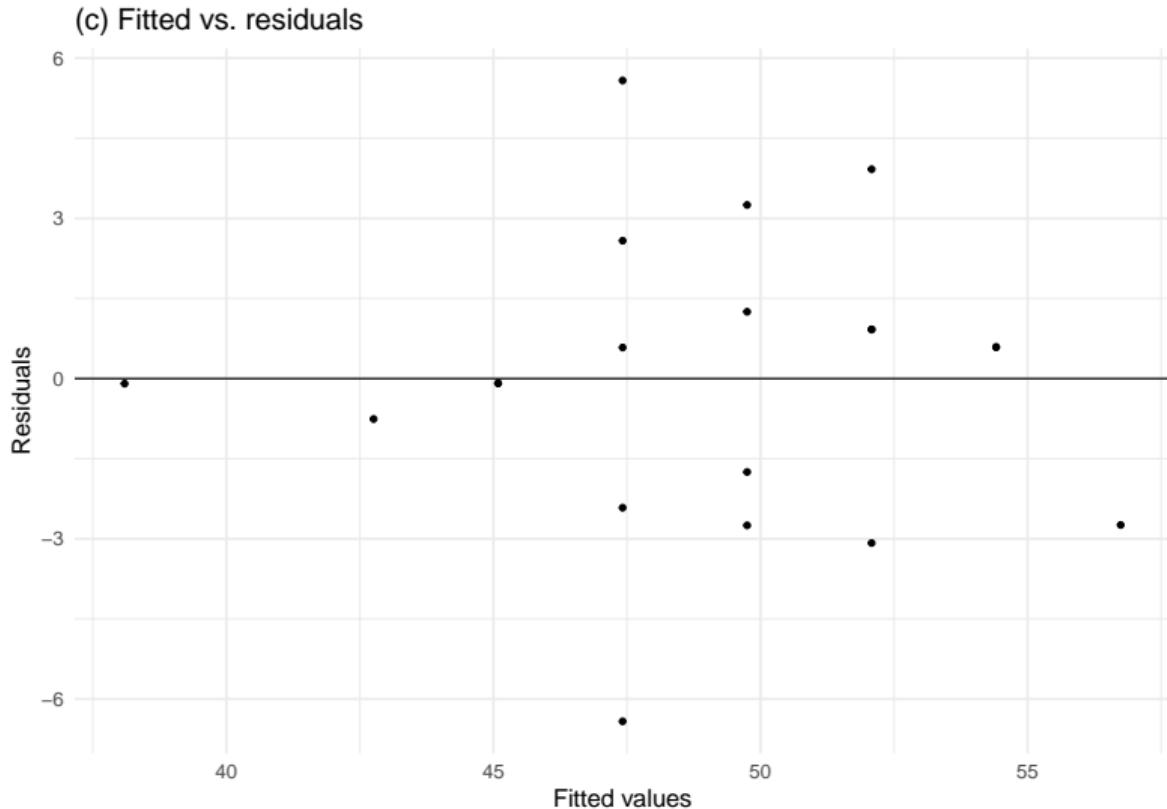
testing and CI with linear regression models

Hypothesis testing for regression

Confidence intervals for regression coefficient

Inference for prediction

Check the model diagnostics



Statistics is Everywhere

Recap of part 1 (chapters
3,4, lectures 4,5,6)Regression and assumptions
needed for inferencetesting and CI with linear
regression modelsHypothesis testing for
regressionConfidence intervals for
regression coefficient

Inference for prediction

Check the model diagnostics

```
## Amount explained
frog_gather <- frog_lm_augment %>% select(freq, .resid) %>%
  gather(key = "type", value = "value", freq, .resid)

frog_expl<-ggplot(frog_gather, aes(y = value)) +
  geom_boxplot(aes(fill = type)) +
  theme_minimal(base_size = 15) +
  labs(title = "(d) Amount explained")
```

Statistics is Everywhere

Recap of part 1 (chapters
3,4, lectures 4,5,6)

Regression and assumptions
needed for inference

testing and CI with linear
regression models

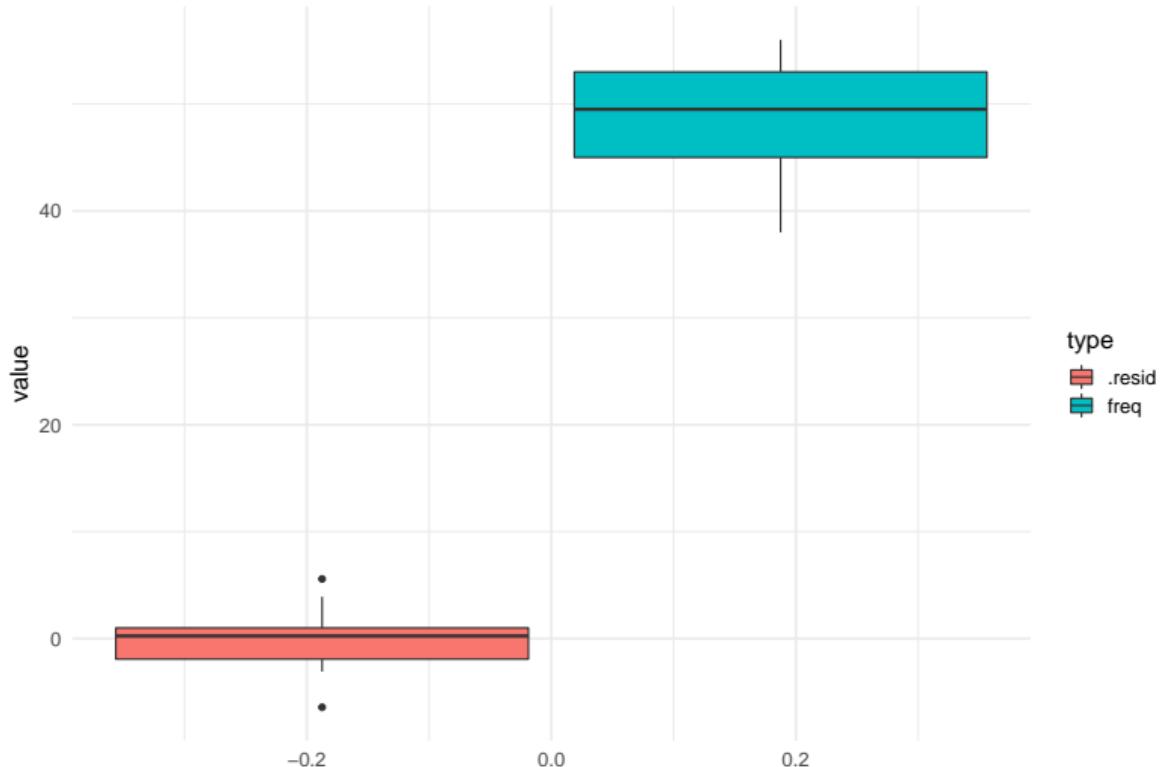
Hypothesis testing for
regression

Confidence intervals for
regression coefficient

Inference for prediction

Check the model diagnostics

(d) Amount explained



Statistics is Everywhere

Recap of part 1 (chapters
3,4, lectures 4,5,6)Regression and assumptions
needed for inferencetesting and CI with linear
regression modelsHypothesis testing for
regressionConfidence intervals for
regression coefficient

Inference for prediction

Statistics is Everywhere

Recap of part 1 (chapters
3,4, lectures 4,5,6)

Regression and assumptions
needed for inference

**testing and CI with linear
regression models**

Hypothesis testing for
regression

Confidence intervals for
regression coefficient

Inference for prediction

testing and CI with linear regression models

Part III hypothesis testing in regression models

Inference for linear regressions

- ▶ How can we account for the imprecision of only having a sample?
- ▶ Confidence interval for the slope
- ▶ Hypothesis testing for the slope

Statistics is Everywhere

Recap of part 1 (chapters 3,4, lectures 4,5,6)

Regression and assumptions needed for inference

testing and CI with linear regression models

Hypothesis testing for regression

Confidence intervals for regression coefficient

Inference for prediction

Regression standard error

Inference for linear regressions

Statistics is Everywhere

Recap of part 1 (chapters 3,4, lectures 4,5,6)

Regression and assumptions needed for inference

testing and CI with linear regression models

Hypothesis testing for regression

Confidence intervals for regression coefficient

Inference for prediction

All of our hypothesis tests are looking for the “signal” in the “noise”.

The “noise” here is the regression standard error, the variability around our regression line.

Regression standard error

- ▶ A good-fitting model will have a low regression standard error because \hat{y} will be close to y .
- ▶ Look at s after running a linear model to assess the model's fit to the data.
- ▶ s is on the same scale as y (i.e., they have the same units).
- ▶ `glance(your_lm)` prints s , denoted by sigma

$$s = \sqrt{\frac{1}{n-2} \sum_{i=1}^n \hat{e}^2}$$

$$s = \sqrt{\frac{1}{n-2} \sum_{i=1}^n (y - \hat{y})^2}$$

Statistics is Everywhere

Recap of part 1 (chapters
3,4, lectures 4,5,6)Regression and assumptions
needed for inferencetesting and CI with linear
regression modelsHypothesis testing for
regressionConfidence intervals for
regression coefficient

Inference for prediction

glance() to view the regression standard error in our frog data

```
glance(frog_lm)
```

```
## # A tibble: 1 x 12
##   r.squared adj.r.squared sigma statistic  p.value    df logLik AIC
##       <dbl>          <dbl>  <dbl>      <dbl>     <dbl>  <dbl>  <dbl> <dbl>
## 1     0.715          0.699  2.82      45.2 0.00000266     1 -48.1  102.
## # ... with 3 more variables: deviance <dbl>, df.residual <int>, nobs <int>
```

- ▶ $\sigma = 2.82$. This is the regression standard error.
- ▶ It can be interpreted as an average measure of the absolute distance between the observed and predicted response variable across all observations (y and \hat{y}).

Statistics is Everywhere

Recap of part 1 (chapters
3,4, lectures 4,5,6)Regression and assumptions
needed for inferencetesting and CI with linear
regression modelsHypothesis testing for
regressionConfidence intervals for
Regression coefficient

Inference for prediction

Another way to contextualize the regression standard error

You can compute a five number summary on the residuals using the augmented data frame:

```
frog_lm_augment %>% summarise(min_resid = min(.resid),  
                                q25_resid = quantile(.resid, 0.25),  
                                mean_resid = mean(.resid),  
                                q75_resid = quantile(.resid, 0.75),  
                                max_resid = max(.resid))
```

```
## # A tibble: 1 x 5  
##   min_resid  q25_resid  mean_resid  q75_resid  max_resid  
##     <dbl>      <dbl>      <dbl>      <dbl>      <dbl>  
## 1    -6.42     -1.92   -2.63e-14      1.00      5.58
```

Statistics is Everywhere
Recap of part 1 (chapters 3,4, lectures 4,5,6)
Regression and assumptions needed for inference
testing and CI with linear regression models
Hypothesis testing for regression
Confidence intervals for regression coefficient
Inference for prediction

Another way to contextualize the regression standard error

- ▶ The smallest residual is -6.42 and the largest is 5.58.
- ▶ The IQR for the residuals goes from -1.92 to 1.00.
- ▶ The mean residual is very close to 0.
- ▶ The residual standard error (2.82) is capturing the standard deviation of this distribution of residuals.

Statistics is Everywhere

Recap of part 1 (chapters
3,4, lectures 4,5,6)

Regression and assumptions
needed for inference

testing and CI with linear
regression models

Hypothesis testing for
regression

Confidence intervals for
regression coefficient

Inference for prediction

Statistics is Everywhere

Recap of part 1 (chapters
3,4, lectures 4,5,6)

Regression and assumptions
needed for inference

testing and CI with linear
regression models

**Hypothesis testing for
regression**

Confidence intervals for
regression coefficient

Inference for prediction

Hypothesis testing for regression

Hypothesis testing for regression

Inference for linear
regressions

Statistics is Everywhere

Recap of part 1 (chapters
3,4, lectures 4,5,6)

Regression and assumptions
needed for inference

testing and CI with linear
regression models

**Hypothesis testing for
regression**

Confidence intervals for
regression coefficient

Inference for prediction

What are the null and alternative hypotheses?

Hypothesis testing for regression

$H_0 : b = 0$ (i.e., There is no association between temperature and the frequency of mating calls)

$H_a : b \neq 0$ (i.e., There is an association between temperature and the frequency of mating calls)

side note: your book has a section on “Testing lack of correlation” please ignore this section

Statistics is Everywhere

Recap of part 1 (chapters 3,4, lectures 4,5,6)

Regression and assumptions needed for inference

testing and CI with linear regression models

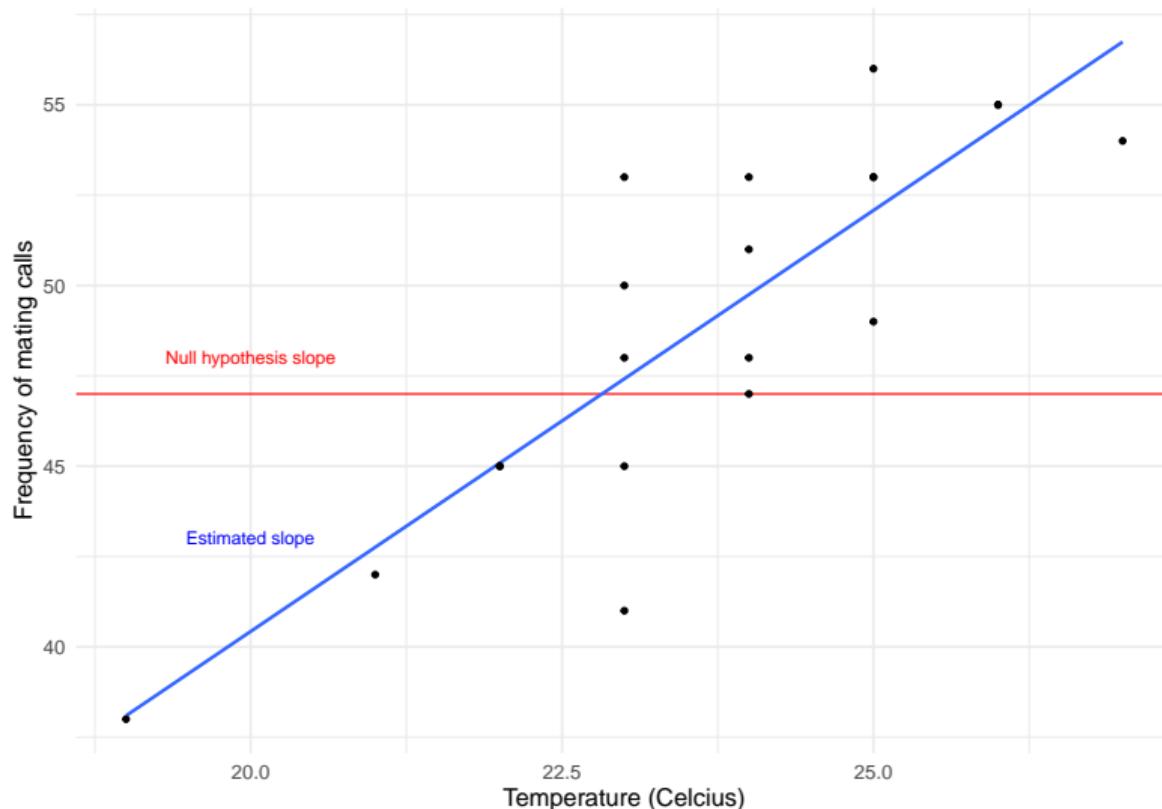
Hypothesis testing for regression

Confidence intervals for regression coefficient

Inference for prediction

Frog data showing the estimates slope vs. null hypothesis slope

```
## `geom_smooth()` using formula 'y ~ x'
```



Statistics is Everywhere

Recap of part 1 (chapters 3,4, lectures 4,5,6)

Regression and assumptions needed for inference

testing and CI with linear regression models

Hypothesis testing for regression

Confidence intervals for regression coefficient

Inference for prediction

Hypothesis testing for regression

- The regression standard error is used as part of the test statistic for the slope coefficient

To test the null hypothesis, the t-test statistic is:

$$t = \frac{\hat{b}}{SE_b}$$

where $SE_b = \frac{s}{\sqrt{\sum(x-\bar{x})^2}}$ and $s = \sqrt{\frac{1}{n-2} \sum_{i=1}^n (y - \hat{y})^2}$

We will use R to compute the test statistic, SE_b and s . Be sure you know where all of these values come from and which functions we use to run a linear model and print these values.

Statistics is Everywhere

Recap of part 1 (chapters 3,4, lectures 4,5,6)

Regression and assumptions needed for inference
testing and CI with linear regression models

Hypothesis testing for regression

Confidence intervals for regression coefficient

Inference for prediction

Two-sided hypothesis testing for regression using `tidy()`

```
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
```

Focus on the row of data for `temp`:

- ▶ `estimate` is the estimated slope coefficient \hat{b} : 2.33
- ▶ `std.error` is the standard error, $SE_b = 0.347$
- ▶ `statistic` is the t-test statistic: $\frac{\hat{b}}{SE_b} = 2.330816/0.3467893 = 6.72$
- ▶ The test has $n - 2$ degrees of freedom, where n is the number of observations in the data frame.
- ▶ `p-value` is the p-value corresponding to the test

Statistics is Everywhere
Recap of part 1 (chapters 3,4, lectures 4,5,6)
Regression and assumptions needed for inference
testing and CI with linear regression models
Hypothesis testing for regression
Confidence intervals for regression coefficient
Inference for prediction

p value for the slope

Statistics is Everywhere

Recap of part 1 (chapters
3,4, lectures 4,5,6)

Regression and assumptions
needed for inference

testing and CI with linear
regression models

Hypothesis testing for
regression

Confidence intervals for
regression coefficient

Inference for prediction

Remember we can check this in R using our `pt()` function

- ▶ statistic is the t-test statistic: $\frac{\hat{b}}{SE_b} = 2.330816/0.3467893 = 6.72$
- ▶ The test has $n - 2$ degrees of freedom, where n is the number of observations (in our frog data $n=20$)

```
pt(q = 6.7211302, df = 18, lower.tail = F)*2
```

```
## [1] 2.663401e-06
```

Statistics is Everywhere

Recap of part 1 (chapters
3,4, lectures 4,5,6)

Regression and assumptions
needed for inference

testing and CI with linear
regression models

Hypothesis testing for
regression

Confidence intervals for
regression coefficient

Inference for prediction

Confidence intervals for regression coefficient

Confidence intervals for the regression coefficient

We can also use the output from `tidy(your_lm)` to create a 95% confidence interval for the slope coefficient.

estimate \pm margin of error

$$\hat{b} \pm t^* SE_b$$

Where t^* is the critical value for the t distribution with $n - 2$ degrees of freedom with area C (e.g., 95%) between $-t^*$ and t^* .

Statistics is Everywhere

Recap of part 1 (chapters
3,4, lectures 4,5,6)Regression and assumptions
needed for inferencetesting and CI with linear
regression modelsHypothesis testing for
regressionConfidence intervals for
regression coefficient

Inference for prediction

Confidence intervals for the regression coefficient

First, find the critical value t^* , such that 95% of the area is between t^* and $-t^*$:
notice the p value I am entering - why is this not .95?

```
t_star<-qt(p = 0.975, df = 18)
```

```
t_star
```

```
## [1] 2.100922
```

95% CI:

$2.330816 \pm t^*0.3467893$ or $2.330816 \pm 2.100922 \times 0.3467893$

95% CI: 1.60 to 3.06

Interpretation: The estimate for the slope coefficient is 2.33 (95% CI: 1.60-3.06). We found this interval using a method that gives an interval that captures the true population slope parameter (b) 95% of the time.

Statistics is Everywhere

Recap of part 1 (chapters 3,4, lectures 4,5,6)

Regression and assumptions needed for inference

testing and CI with linear regression models

Hypothesis testing for regression

Confidence intervals for regression coefficient

Inference for prediction

Statistics is Everywhere

Recap of part 1 (chapters
3,4, lectures 4,5,6)

Regression and assumptions
needed for inference

testing and CI with linear
regression models

Hypothesis testing for
regression

Confidence intervals for
regression coefficient

Inference for prediction

Inference for prediction

Inference for prediction

Inference for linear regressions

Statistics is Everywhere

Recap of part 1 (chapters 3,4, lectures 4,5,6)

Regression and assumptions needed for inference

testing and CI with linear regression models

Hypothesis testing for regression

Confidence intervals for regression coefficient

Inference for prediction

- ▶ So far we've learned only about inference for the regression coefficient
- ▶ But what if you wanted to use the model to make a prediction?
- ▶ We already know how to predict the **average** number of mating calls corresponding to a specific x value, say of 21 degrees Celsius:

$$\hat{y} = -6.190332 + 2.330816x$$

$$\hat{y} = -6.190332 + 2.330816(21) = 42.8$$

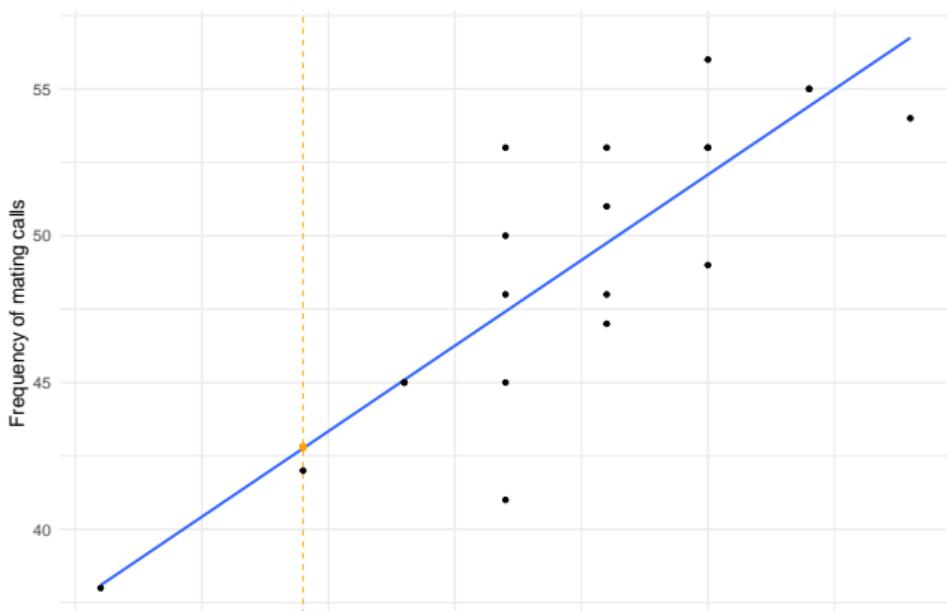
We expect 42.8 mating calls, so 43 mating calls (rounding because the outcome is a discrete variable) when the temperature is 21 degrees Celsius.

Inference for prediction

How do we make a confidence interval for this prediction?

- ▶ It depends on whether you want to make a CI for the **average response** or for an **individual's response**

```
## `geom_smooth()` using formula 'y ~ x'
```



Statistics is Everywhere

Recap of part 1 (chapters 3,4, lectures 4,5,6)

Regression and assumptions needed for inference

testing and CI with linear regression models

Hypothesis testing for regression

Confidence intervals for regression coefficient

Inference for prediction

Inference for prediction

Inference for linear regressions

Statistics is Everywhere

Recap of part 1 (chapters 3,4, lectures 4,5,6)

Regression and assumptions needed for inference

testing and CI with linear regression models

Hypothesis testing for regression

Confidence intervals for regression coefficient

Inference for prediction

If you want to make inference for the mean response μ_y when x takes the value x^* ($x^*=21$ in our example):

$$\hat{y} \pm t * SE_{\hat{\mu}}, \text{ where } SE_{\hat{\mu}} = s \sqrt{\frac{1}{n} + \frac{(x^* - \bar{x})^2}{\sum(x - \bar{x})^2}}$$

If you want to make inference for a single observation y when x takes the value x^* ($x^*=21$ in our example):

$$\hat{y} \pm t * SE_{\hat{y}}, \text{ where } SE_{\hat{y}} = s \sqrt{1 + \frac{1}{n} + \frac{(x^* - \bar{x})^2}{\sum(x - \bar{x})^2}}$$

Corresponding R code for prediction and confidence interval:

```
# specify the value of the explanatory variable for which you want
newdata = data.frame(temp = 21)

# use `predict()` to make prediction and confidence intervals
prediction_interval <- predict(frog_lm, newdata, interval = "predict")
prediction_interval

##          fit      lwr      upr
## 1 42.7568 36.37187 49.14173

confidence_interval <- predict(frog_lm, newdata, interval = "confidence")
confidence_interval

##          fit      lwr      upr
## 1 42.7568 40.38472 45.12887
```

Statistics is Everywhere
Notes of p. 10 (Chap. 3,4, lectures 4,5,6)

Regression and assumptions
needed for inference

testing and CI with linear
regression models

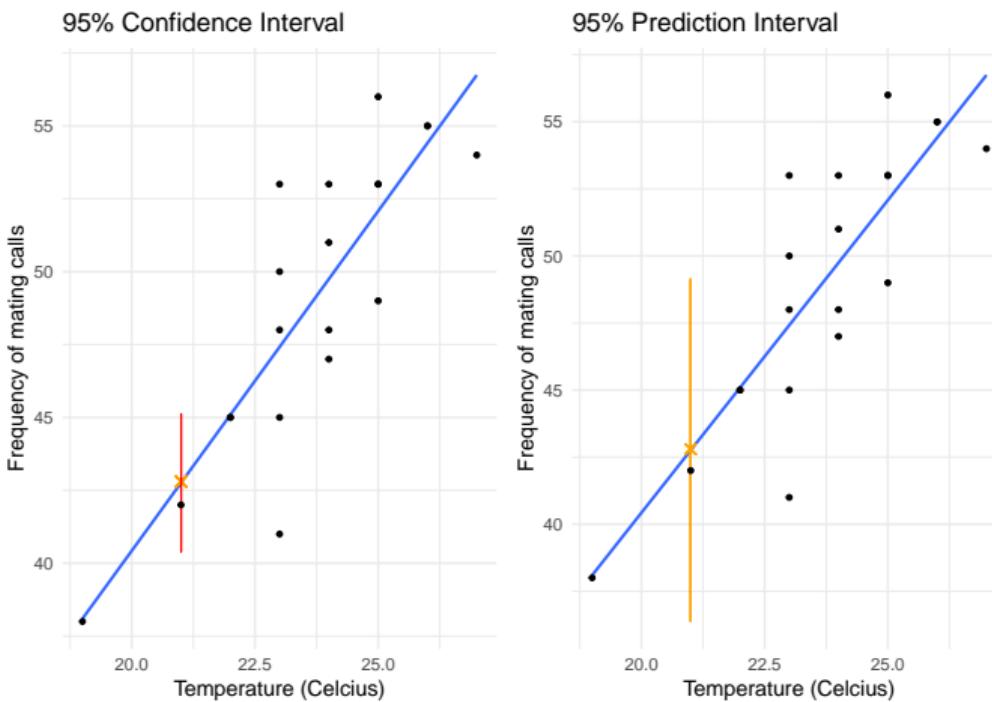
Hypothesis testing for
regression

Confidence intervals for
regression coefficient

Inference for prediction

Inference for prediction, visualized

```
## `geom_smooth()` using formula 'y ~ x'  
## `geom_smooth()` using formula 'y ~ x'
```



Statistics is Everywhere
Recap of part 1 (chapters 3,4, lectures 4,5,6)
Regression and assumptions needed for inference
testing and CI with linear regression models
Hypothesis testing for regression
Confidence intervals for regression coefficient
Inference for prediction

Recap on notation

Term	Population	Sample
Intercept	a or α	\hat{a}
Slope	b or β	\hat{b}
Residual	e	\hat{e}

Note: Although many sources will use r to indicate residuals, we will try to be consistent and use e , because we use r and r^2 to represent the correlation coefficient and r-squared respectively and this is confusing.

Statistics is Everywhere

Recap of part 1 (chapters 3,4, lectures 4,5,6)

Regression and assumptions needed for inference

testing and CI with linear regression models

Hypothesis testing for regression

Confidence intervals for regression coefficient

Inference for prediction

Recap: Use `lm()` + broom functions to look at your linear model

- ▶ `tidy(your_lm)`: Presents the output of the linear model in a tidy way
- ▶ `glance(your_lm)`: Takes a quick (one line) look at the fit statistics.
- ▶ `augment(your_lm)`: Creates an augmented data frame that contains a column for the fitted y-values (\hat{y}) and the residuals ($\hat{e} = y - \hat{y}$) among other columns

Know these functions, what they do, and how to use them.

Statistics is Everywhere

Recap of part 1 (chapters 3,4, lectures 4,5,6)

Regression and assumptions needed for inference

testing and CI with linear regression models

Hypothesis testing for regression

Confidence intervals for regression coefficient

Inference for prediction

Parting humor

Inference for linear regressions

Statistics is Everywhere

Recap of part 1 (chapters 3,4, lectures 4,5,6)

Regression and assumptions needed for inference

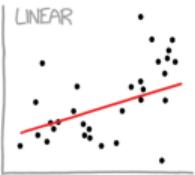
testing and CI with linear regression models

Hypothesis testing for regression

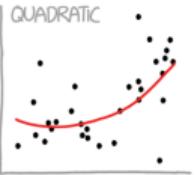
Confidence intervals for regression coefficient

Inference for prediction

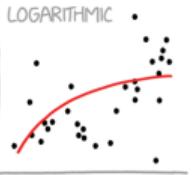
CURVE-FITTING METHODS AND THE MESSAGES THEY SEND



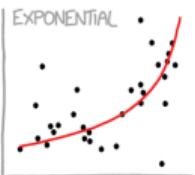
"HEY, I DID A
REGRESSION."



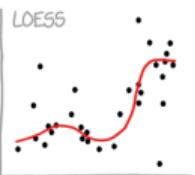
"I WANTED A CURVED
LINE, SO I MADE ONE
WITH MATH."



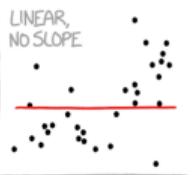
"LOOK, IT'S
TAPERING OFF!"



"LOOK, IT'S GROWING
UNCONTROLLABLY!"



"I'M SOPHISTICATED, NOT
LIKE THOSE BUMBLING
POLYNOMIAL PEOPLE."



"I'M MAKING A
SCATTER PLOT BUT
I DON'T WANT TO."