Continous-continous and regressions

Roadmap

So for 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

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

Reminder of what we've done with continous vs continous 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

Recap: Visualizing continous-continous relationships

- 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

Scatterplot Syntax in R

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”)

Remember the Manatees?

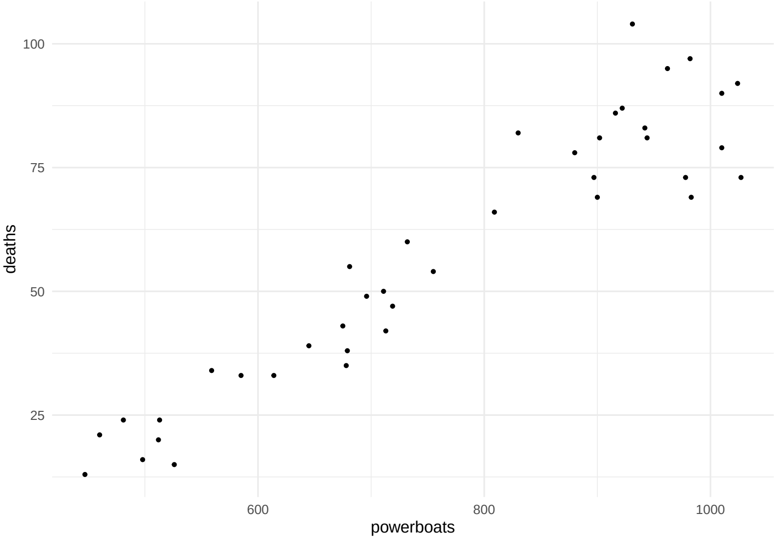
Manatee data set from your textbook:

```
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)
```

Remember the Manatees?



Recap: Pearson's

- 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

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

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

lm() of manatee deaths and powerboat purchases

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.

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   AIC   BI
##       <dbl>         <dbl> <dbl>     <dbl>    <dbl> <dbl>  <dbl> <dbl> <dbl
## 1     0.893         0.890  8.82      316. 5.21e-20     1  -143.  292.  297
## # i 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.

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

Regression and assuptions 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.

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 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 $\hat{b}$

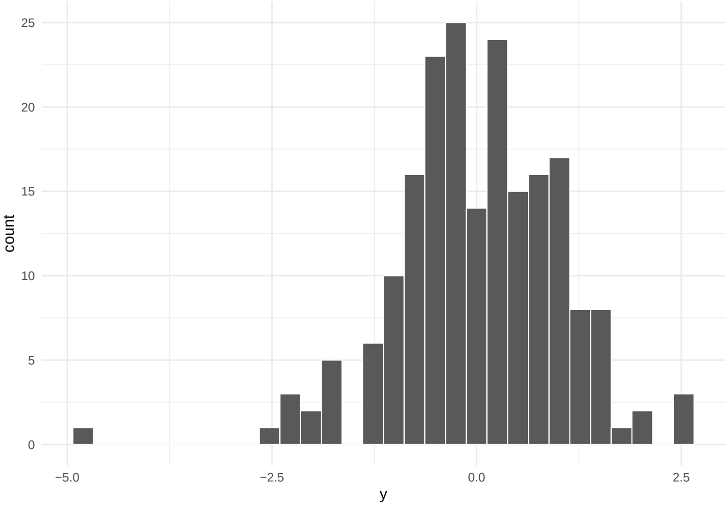
The Normal quantile plot (a.k.a the Q-Q plot)

The Normal quantile plot (a.k.a the Q-Q plot)

- The purpose of making a Q-Q plot is to examine the Normality of a distribution of a variable.
- If you want to know whether variable is Normally distributed you could examine its histogram to see if it is unimodal and symmetric. However, it is still sometimes hard to say if it is truly Normal. To do so you can use a Q-Q plot.

Are these data Normally distributed?

- The data is unimodal and symmetric, but is its distribution Normal?



Making a QQ plot step by step

1. First, arrange the variable in ascending order. Calculate the percentile for each measurement. For example if you had ten measurements in ascending order, the first measurement is at the 10th percentile because 10% of the data is at or below its value. The second measurement is at the 20% because 20% of the data is at or below its value, and so forth.
2. Then, for each of the percentiles you calculated, use that percentile to calculate the corresponding x-value of the Normal distribution that occurs at that percentile. For example, at $x = -1$ at the 16th percentile of the N(0, 1) distribution.
3. Make a scatter plot of the calculated x-values on the x-axis and the original variable values on the y-axis.
4. The closer the data is to a straight line, the more closely it approximates a Normal distribution.

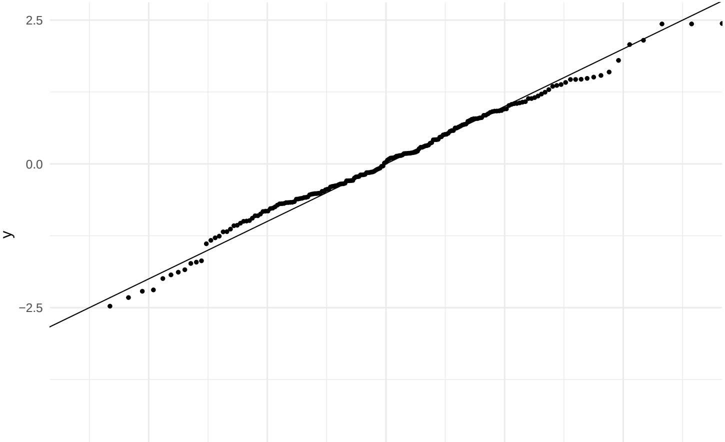
Making a QQ plot step by step

```
#1. calculate the percentile :
example_data <- example_data %>%  arrange(y) %>%
  mutate(quantile = row_number()/n())
# 2. then calculate the x-value at each percentile from the previous step
# 3. this x-value can be called a z-score because it is from the standard Nor
example_data <- example_data %>%
  mutate(z_score = qnorm(quantile, mean = 0, sd = 1))
head(example_data)
```

```
##             y quantile   z_score
## 1 -4.895129    0.005 -2.575829
## 2 -2.474167    0.010 -2.326348
## 3 -2.324430    0.015 -2.170090
## 4 -2.214979    0.020 -2.053749
## 5 -2.191381    0.025 -1.959964
## 6 -1.993573    0.030 -1.880794
```

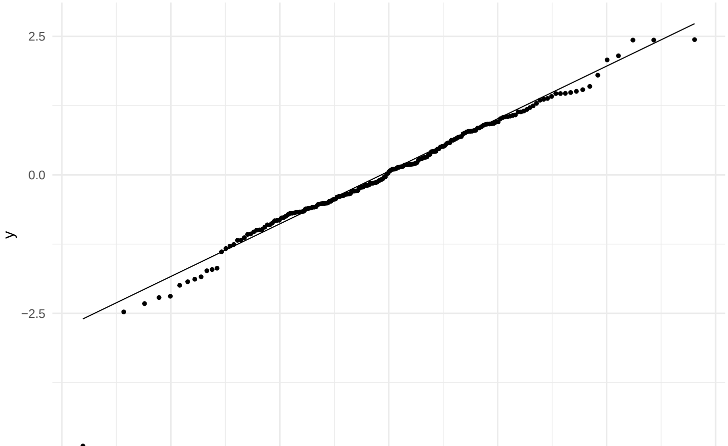
Look at the QQ plot for these data

- Notice that the data overlays the 45 degree line in the middle but not in the tails of the distribution. This sort of pattern shows that these data are “wider” (have larger standard deviation) that a Normally distributed variable.



Easy way to make a qqplot() where R does all the calculating for you

```
ggplot(example_data, aes(sample = y)) + stat_qq() + stat_qq_line() +
  theme_minimal(base_size = 15)
```



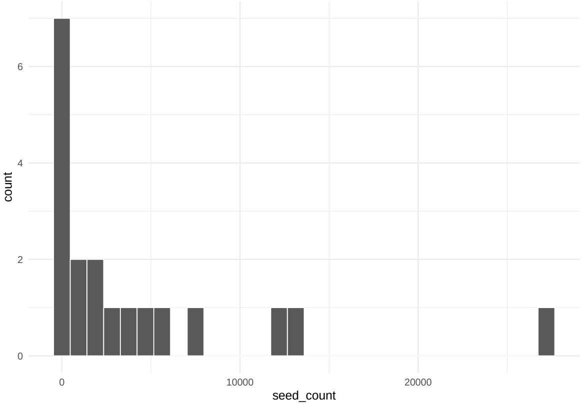
Another example: Seed Data

```r
library(readr)
seed_data <- read_csv("Ch04_seed-data")

head(seed_data)
```

```
## # A tibble: 6 x 3
##   species          seed_count seed_weight
##   <chr>                 <dbl>       <dbl>
## 1 Paper birch           27239         0.6
## 2 Yellow birch          12158         1.6
## 3 White spruce           7202         2
## 4 Engelman spruce        3671         3.3
## 5 Red spruce             5051         3.4
## 6 Tulip tree            13509         9.1
```

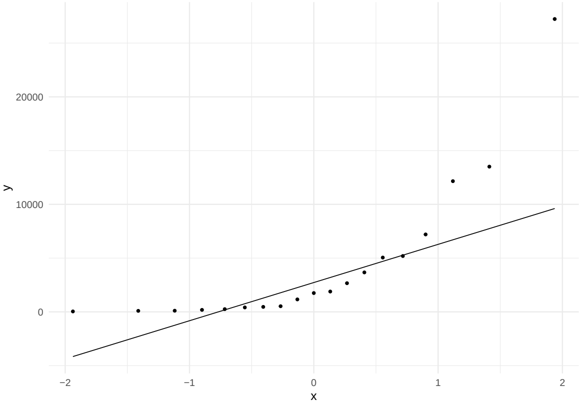
Another example

Check out its distribution. It definitely does not look normal:



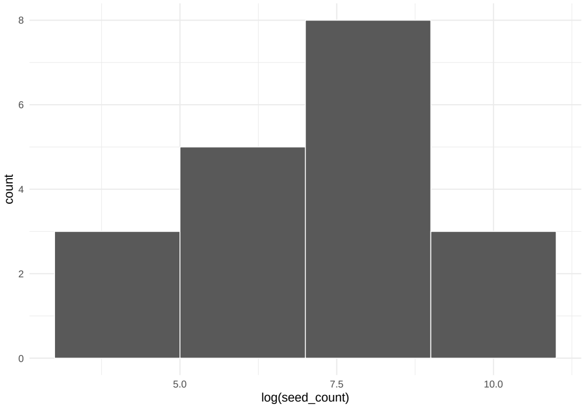
Another example

And look at its QQ plot. Does the data appear to follow a Normal distribution?



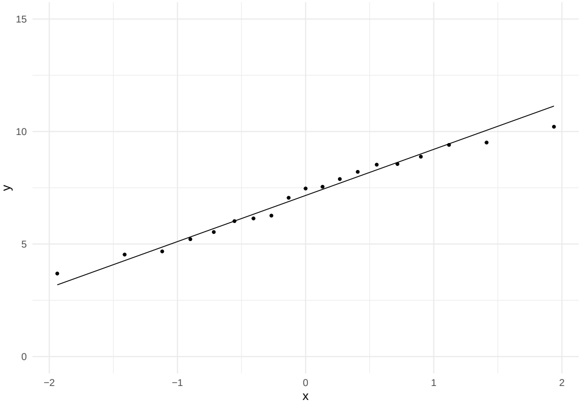
Another example (logged)

You might remember that we took the log of seed_count before we used it in regression. The log values look like this:



Another example (logged)

How does the QQ plot look for the logged variable?



QQ plot summary

- Review the QQ plots from the book on page 290-292 of B&M Edition 4
- Try and gain intuition about when a variable does not appear to fit a Normal distribution
  - Was the distribution skewed?
  - Was there an outlier?
- For each scenario how do these deviations from Normality affect the QQ plot?

Terminology used to investigate the assumptions

Observed value: y

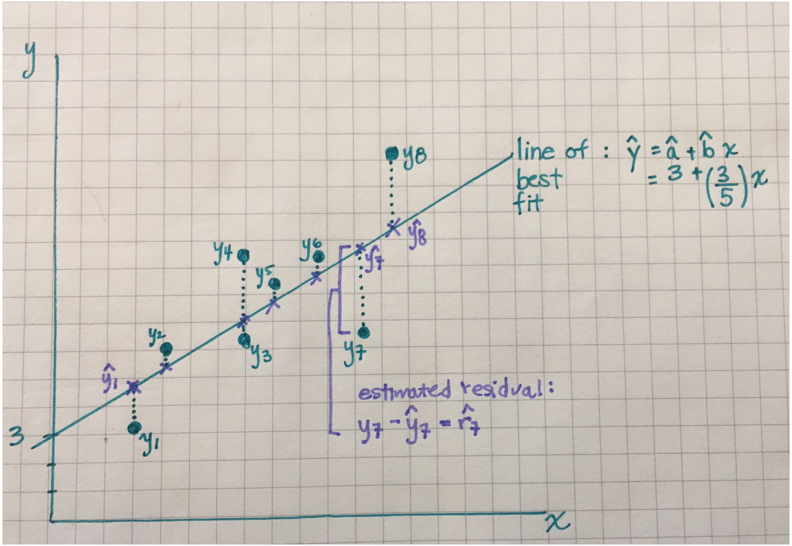
Fitted value: $\hat{y} = \hat{a} + \hat{b}x$

Estimated residuals:

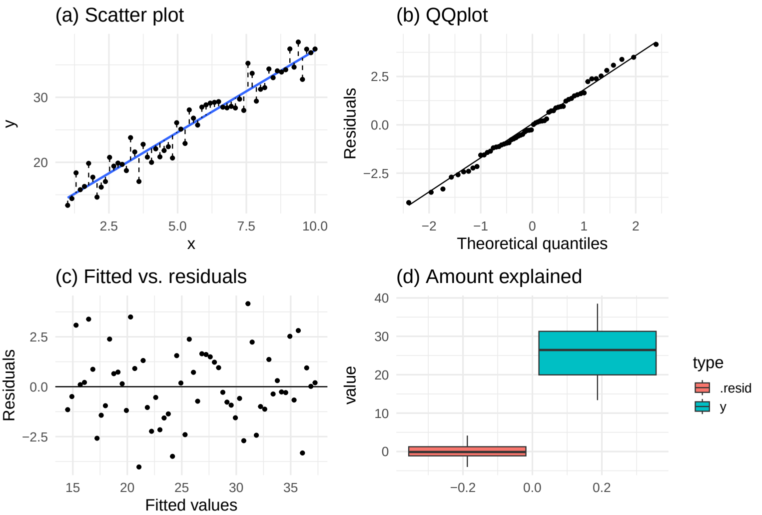
$\hat{e} =$ observed value - fitted value

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

Terminology used to investigate the assumptions, visualized



Example 1: Investigating the assumptions



A good fit to the data

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.

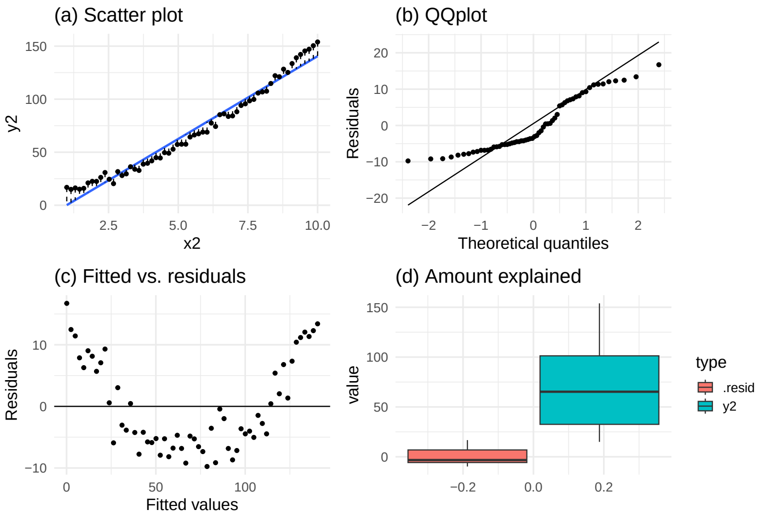
Plot (d) shows a boxplot of the distribution of y vs. the distribution of the residuals. If x does a good job describing y, then the box plot for the residuals will be much shorter because the model fit is good

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'
```



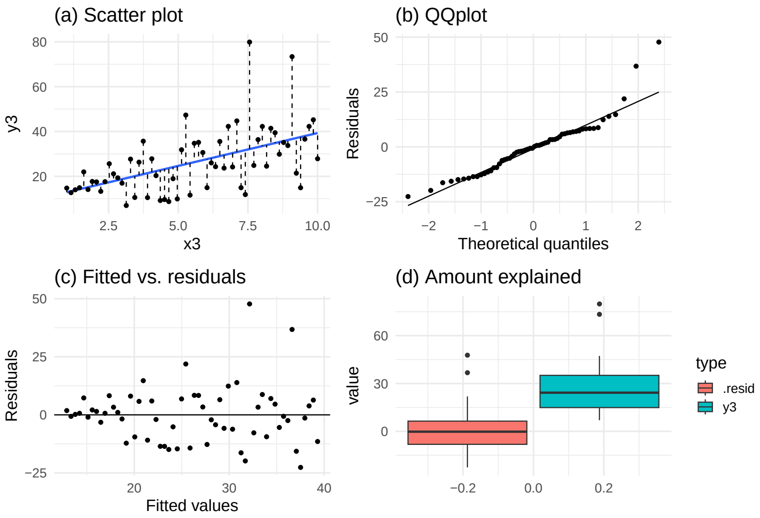
The linear systematic assumption does not hold

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

Example 3: Investigating the assumptions

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



The constant variance assumption does not hold

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.

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.

Hypothesis testing for regression

Hypothesis testing for regression

What are the null and alternative hypotheses?

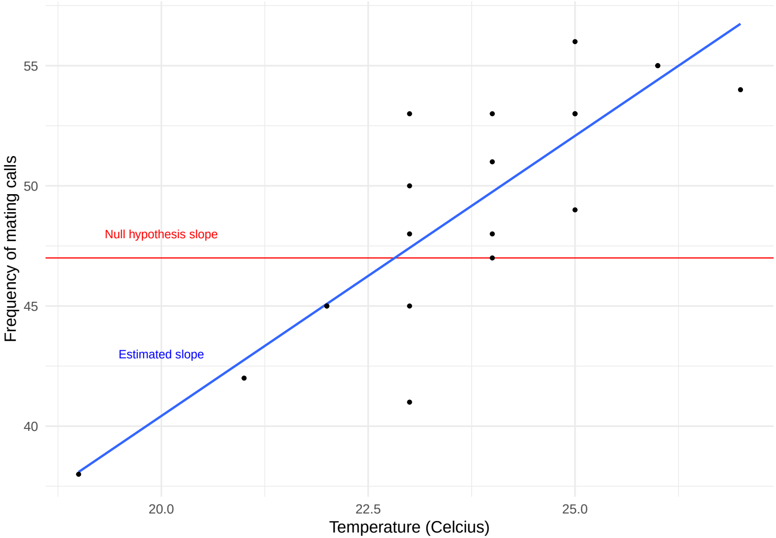
Hypothesis testing for regression

$H_0 : b = 0$ (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

Frog data showing the estimates slope vs. null hypothesis slope



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.

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

p value for the slope

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
```

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^*.

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.

Inference for prediction

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:

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

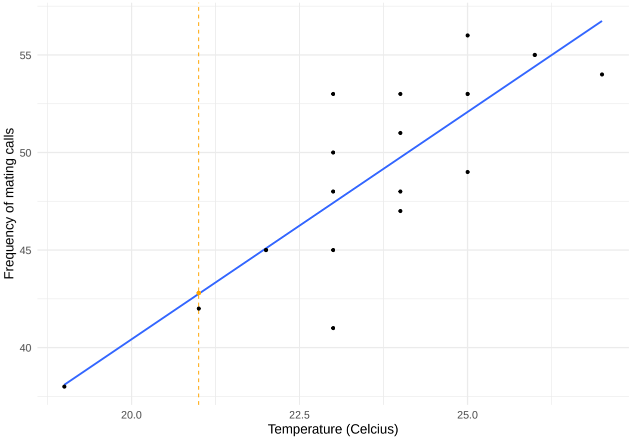
$$\hat{y} = -6.190332 + 2.330816(21) = 42.8$$

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



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:

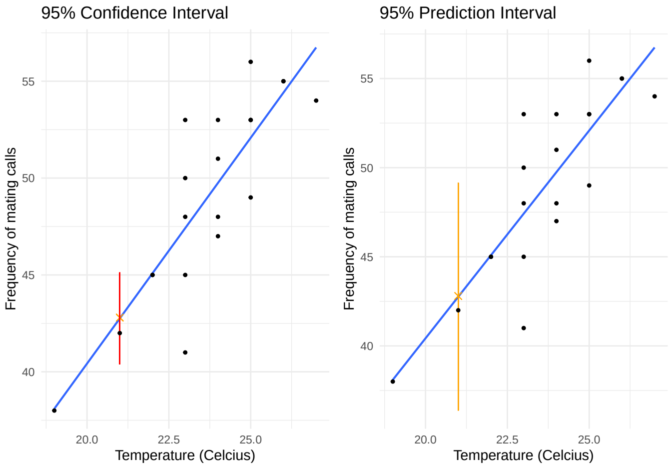
```
# specify the value of the explanatory variable for which you want the predic
newdata = data.frame(temp = 21)

# 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
```

Inference for prediction, visualized



- Why is the prediction interval *wider* than the confidence interval?

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.

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.

Parting humor