Stats Demo (papaja version), Data to Manuscript in R 2024

Abstract

This document demos some of the most commonly used methods for descriptive statistics and basic hypothesis testing in psychology research.

Keywords: keywords

Word count: X

Stats Demo (papaja version), Data to Manuscript in R 2024

Descriptive statistics

The quickest way to see summary statistics for a numeric variable is the `summary()` function:

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
   66.0   170.0   180.0   174.4   192.0   234.0 
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
  15.00   55.60   79.00   97.31   84.50 1358.00 
```

Or if you need to keep things a little more organized, create a summarized dataframe:

```
# A tibble: 2 x 7
  measure  mean median    sd   min   max range
  <chr>   <dbl>  <dbl> <dbl> <dbl> <dbl> <dbl>
1 height  174.     180  35.5    66   234   168
2 mass     97.3     79 169.     15  1358  1343
```

Distributions

Calculate measures of center. You can also calculate everything piece by piece.

“Measures of center” are different ways of talking about averages. Usually we think about “mean” as synonymous with “average”, so calling these measures of center instead can be more precise.

Calculate mean and median with `mean()` and `median()`. There is no built-in mode function, but if you need one you can either write your own function or use the `modeest` library.

Table 1

measure	variance	median1	median2	quartile	iqr1
height	1,262.85	180.00	180.00	170.00	22.00
height	1,262.85	180.00	180.00	192.00	22.00
mass	28,715.73	79.00	79.00	55.60	28.90
mass	28,715.73	79.00	79.00	84.50	28.90

- The mean height is 174.36 cm and median height is 180 cm.
- The mean mass is 97.31 kg and median mass is 79 kg.

Calculate measures of spread. Measures of spread describe the distribution of continuous data around the center. Calculate standard deviation with `sd()`. Calculate range by getting a list of the minimum and maximum with `range()` and then using `diff()` to find the difference between the two.

- The standard deviation of height is 35.54 cm and the range is 168 cm.
- The standard deviation of mass is 169.46 kg and the range is 1343 kg.

Other common measures of spread include variance, quantiles, and interquartile range (Table 1:

```
‘summarise()‘ has grouped output by ’measure’. You can override using the
‘.groups‘ argument.
```

Visualize center and spread. Distribution plots visualize center and spread, for example histograms (Figure 1), density plots (Figure 2), boxplots (Figure 3), and violin plots (Figure 4).

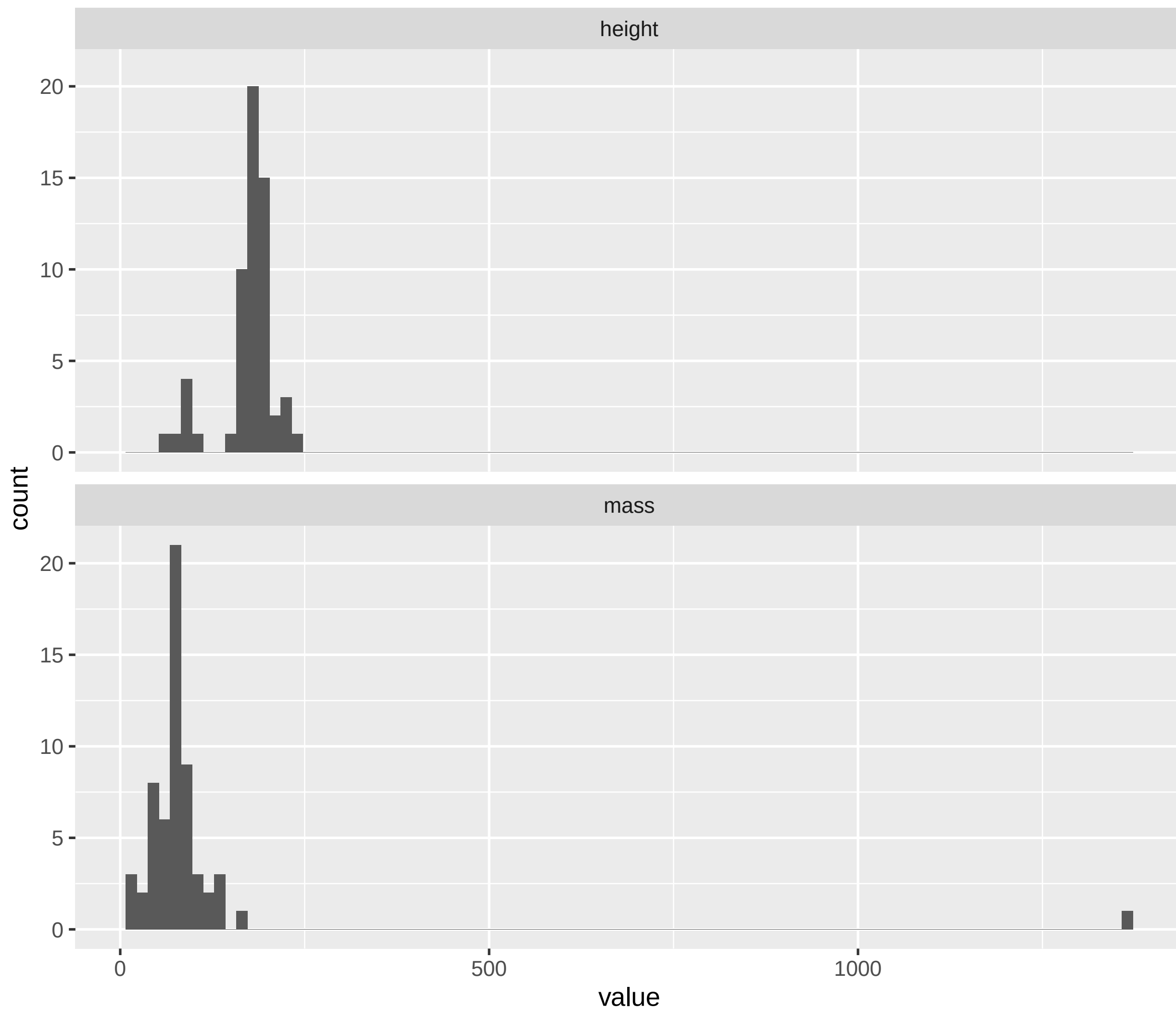


Figure 1. Histogram of height and mass distributions

Correlation

The `cor()` function creates a correlation matrix:

```
         height     mass
height 1.000000 0.130859
mass   0.130859 1.000000
```

```
           mass
height 0.130859
```

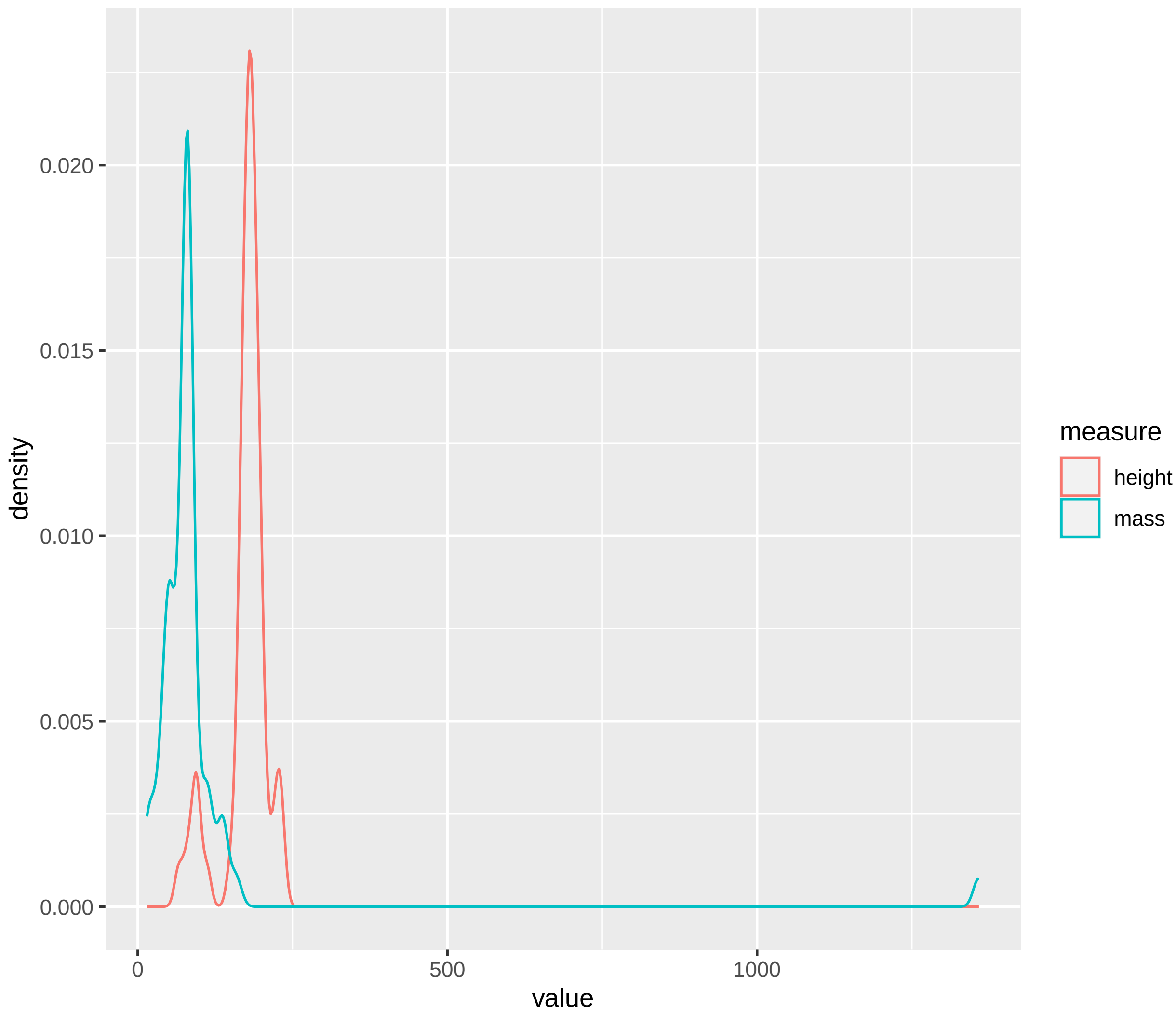


Figure 2. Density of height and mass distributions

The correlation of height and mass is 0.13.

If you intend to use a correlation as a (quasi)hypothesis test, you'll need the `corr.test()` function in the `psych` package to give you p-values,

```
Call:corr.test(x = sw.desc)
Correlation matrix
       height mass
height   1.00 0.13
```

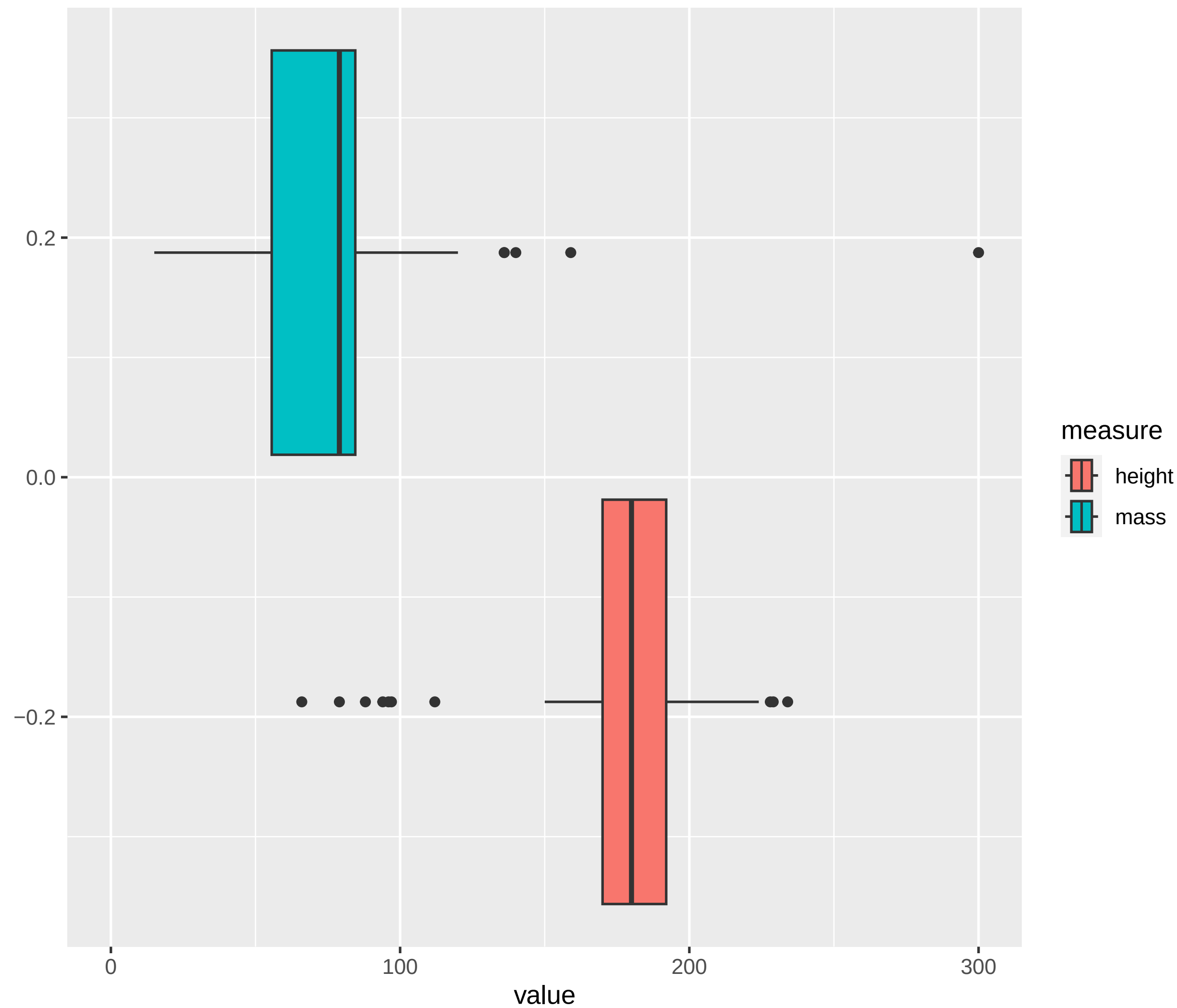


Figure 3. Boxplot of height and mass distributions

```
mass      0.13 1.00
Sample Size
[1] 59
Probability values (Entries above the diagonal are adjusted for multiple tests.)
        height mass
height   0.00 0.32
mass     0.32 0.00
```

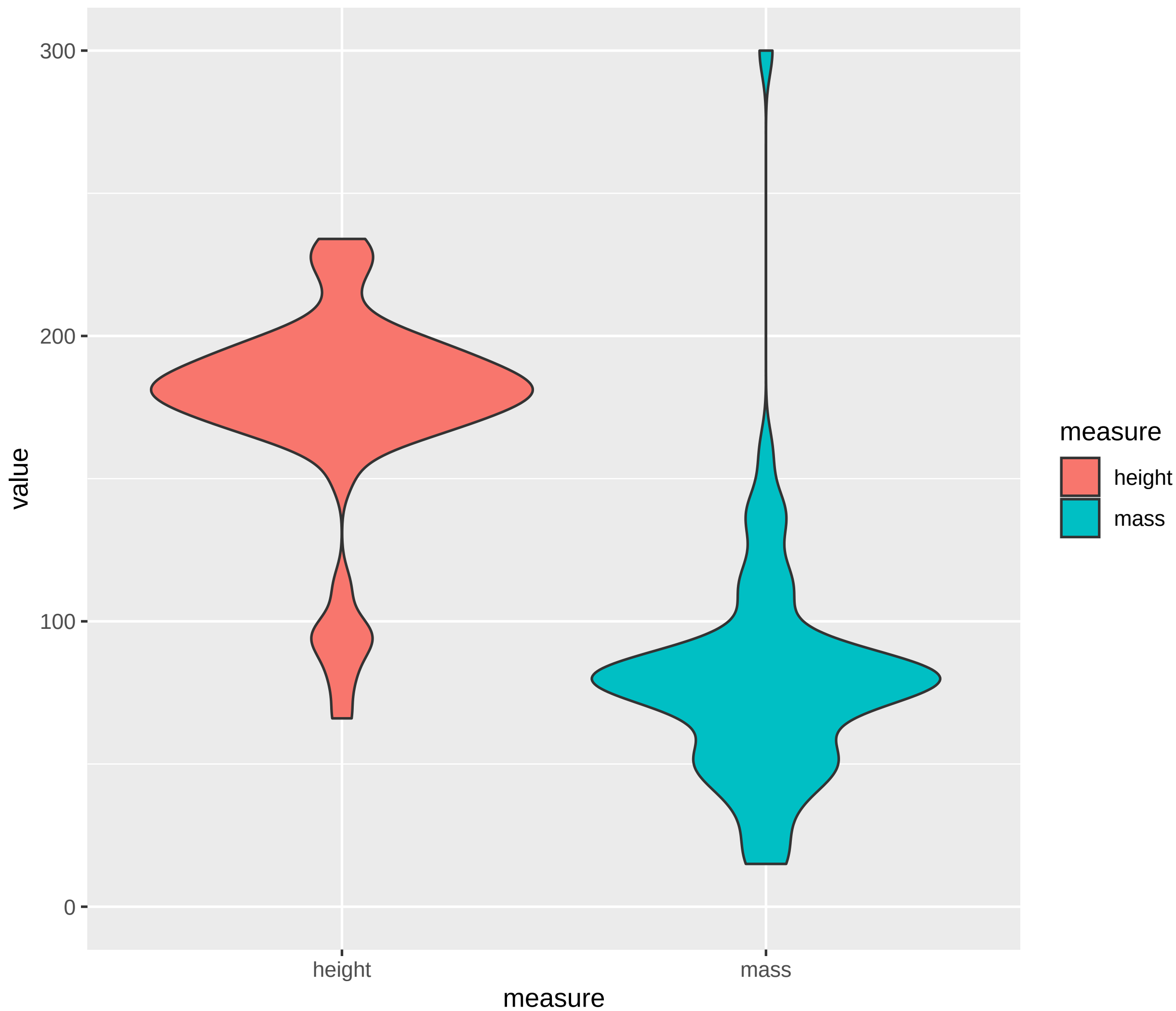


Figure 4. Violin plot of height and mass distributions

```
 To see confidence intervals of the correlations, print with the short=FALSE option

Call:corr.test(x = sw.desc[, 1], y = sw.desc[, 2])
Correlation matrix
       mass
height 0.13
Sample Size
[1] 59
These are the unadjusted probability values.
```

```
 The probability values  adjusted for multiple tests are in the p.adj object.
       mass
height 0.32

 To see confidence intervals of the correlations, print with the short=FALSE option
```

`View()` the object to see what the output contains and then extract elements like p-value:

```
          height      mass
height 0.0000000 0.3232031
mass   0.3232031 0.0000000
```

```
[1] 0.3232031
```

```
            mass
height 0.3232031
```

The default corr method is pearson, but you can change this. For example, Spearman rank correlation is useful for small samples:

- Pearson $\rho = 0.13$ ($p = 0.32$)
- Spearman ranked $\rho = 0.72$ ($p = 0$)

Visualize correlation. Visualizing correlation is functionally the same as visualizing linear regression (though to truly visualize correlation you'd need to normalize the axes). Figure 5 combines a scatter plot with a regression line (using `geom_smooth()`):

```
‘geom_smooth()‘ using formula = ’y ~ x’
```

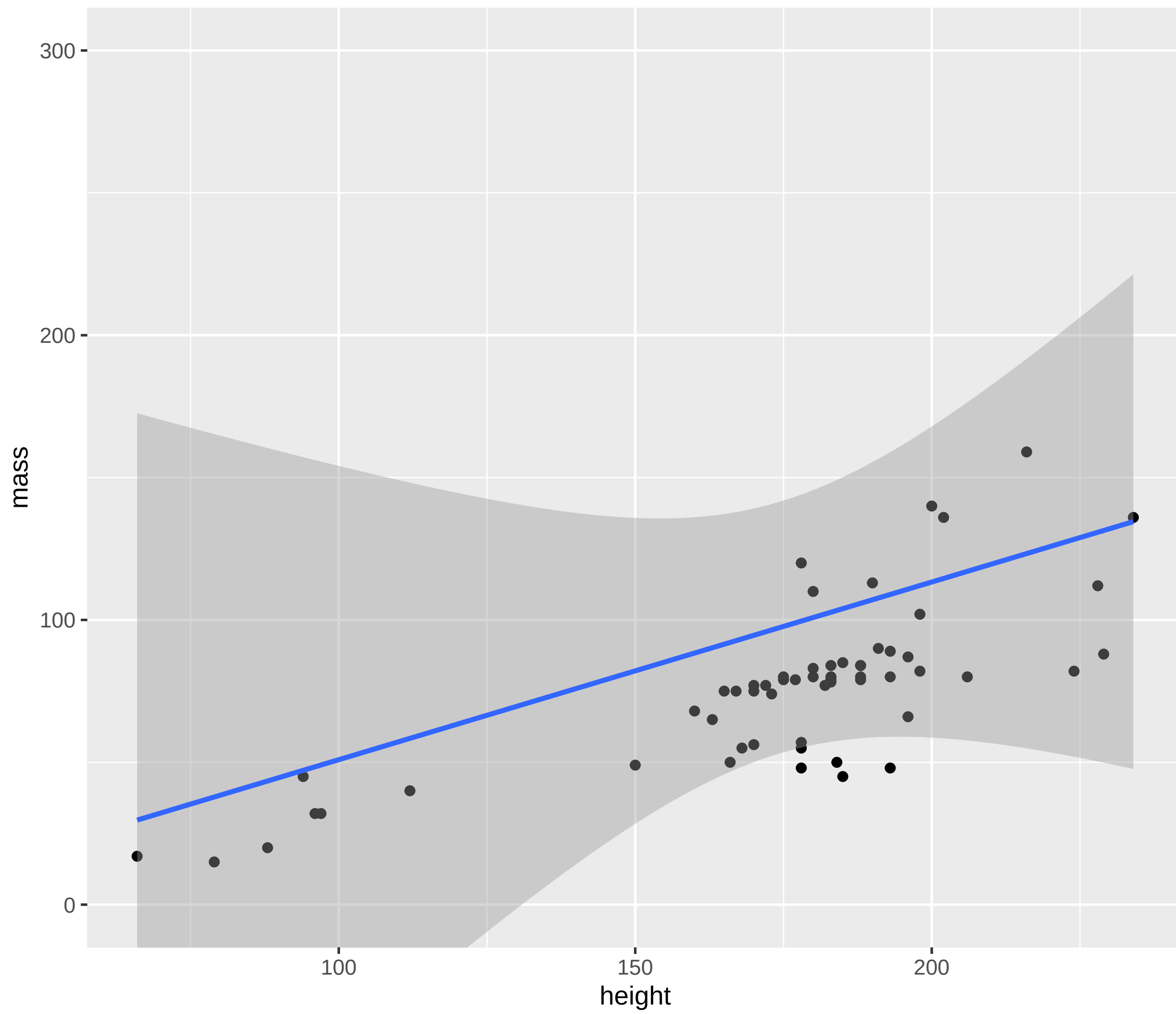
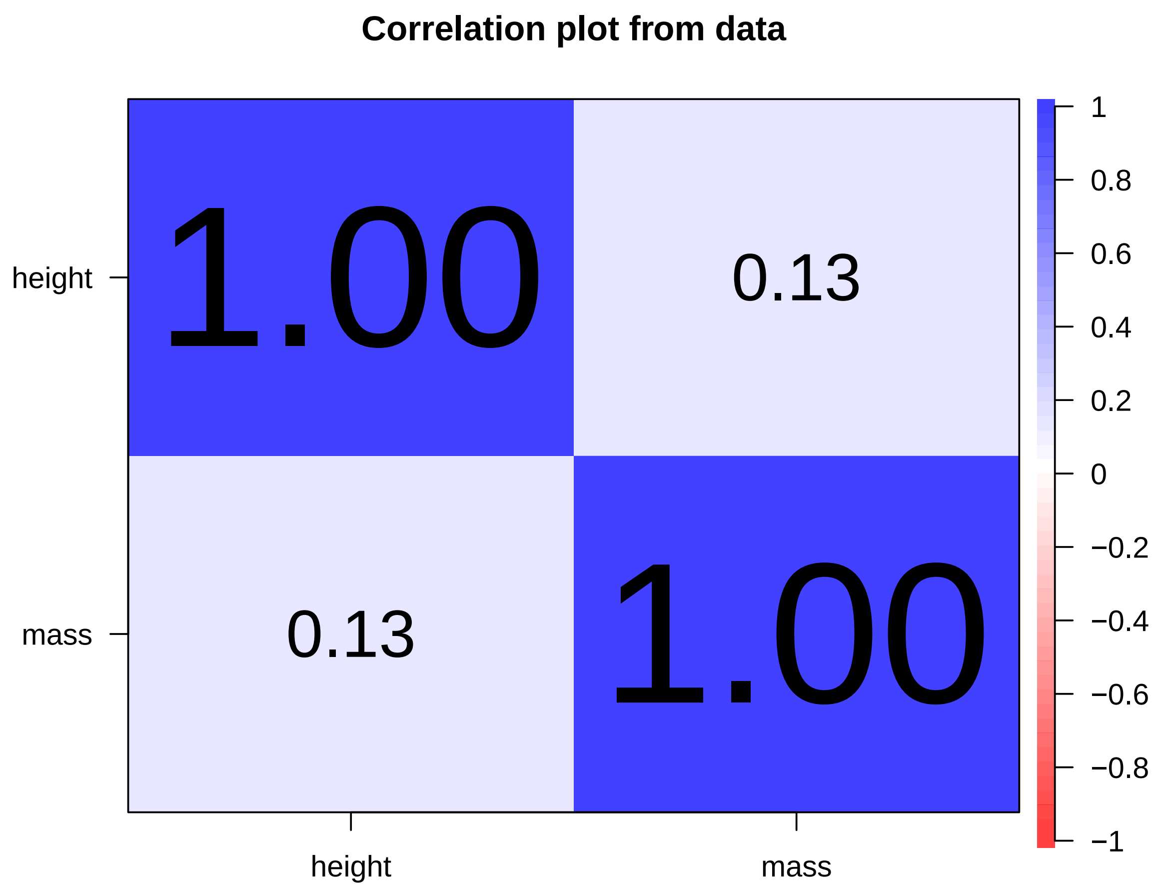
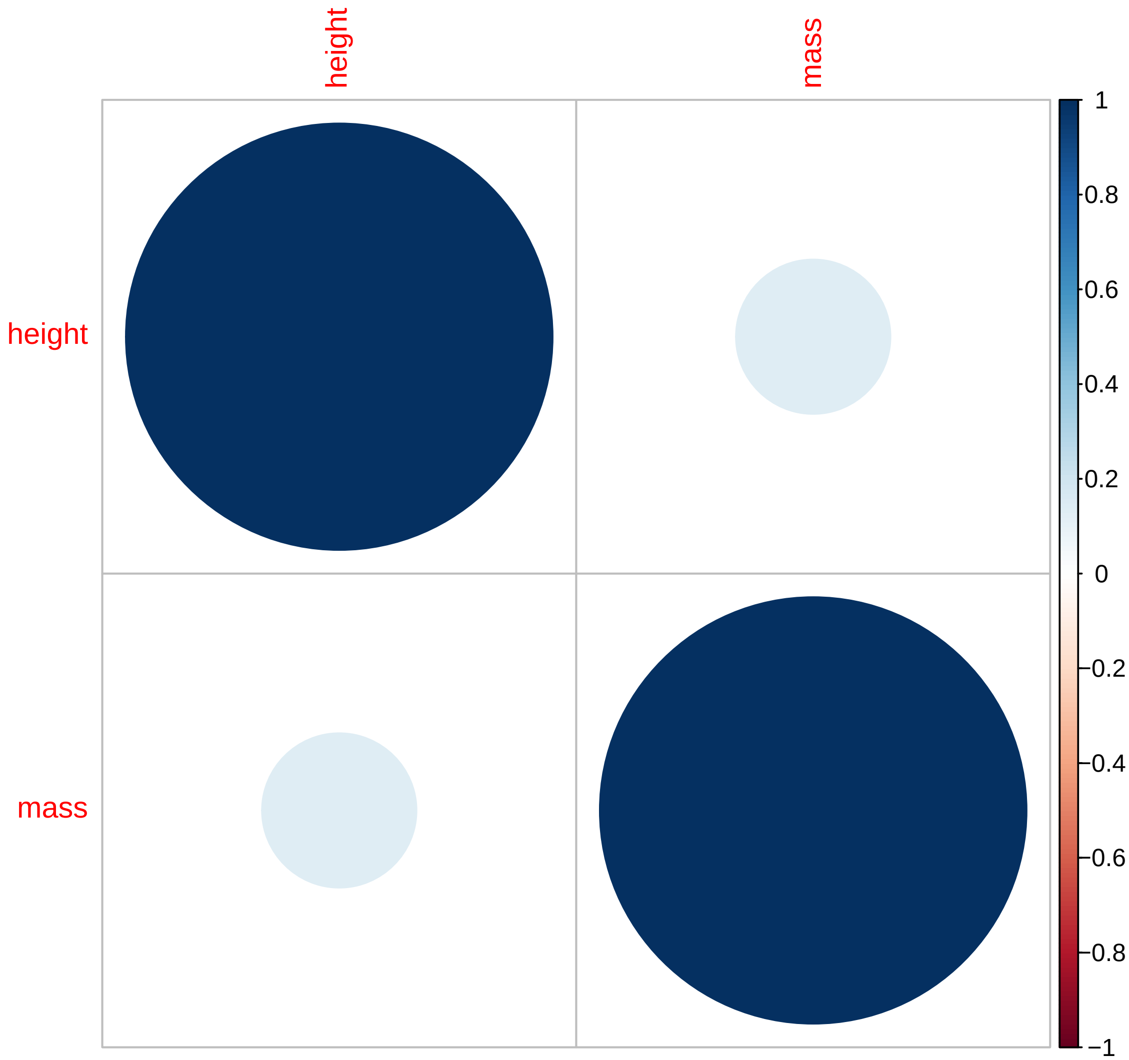


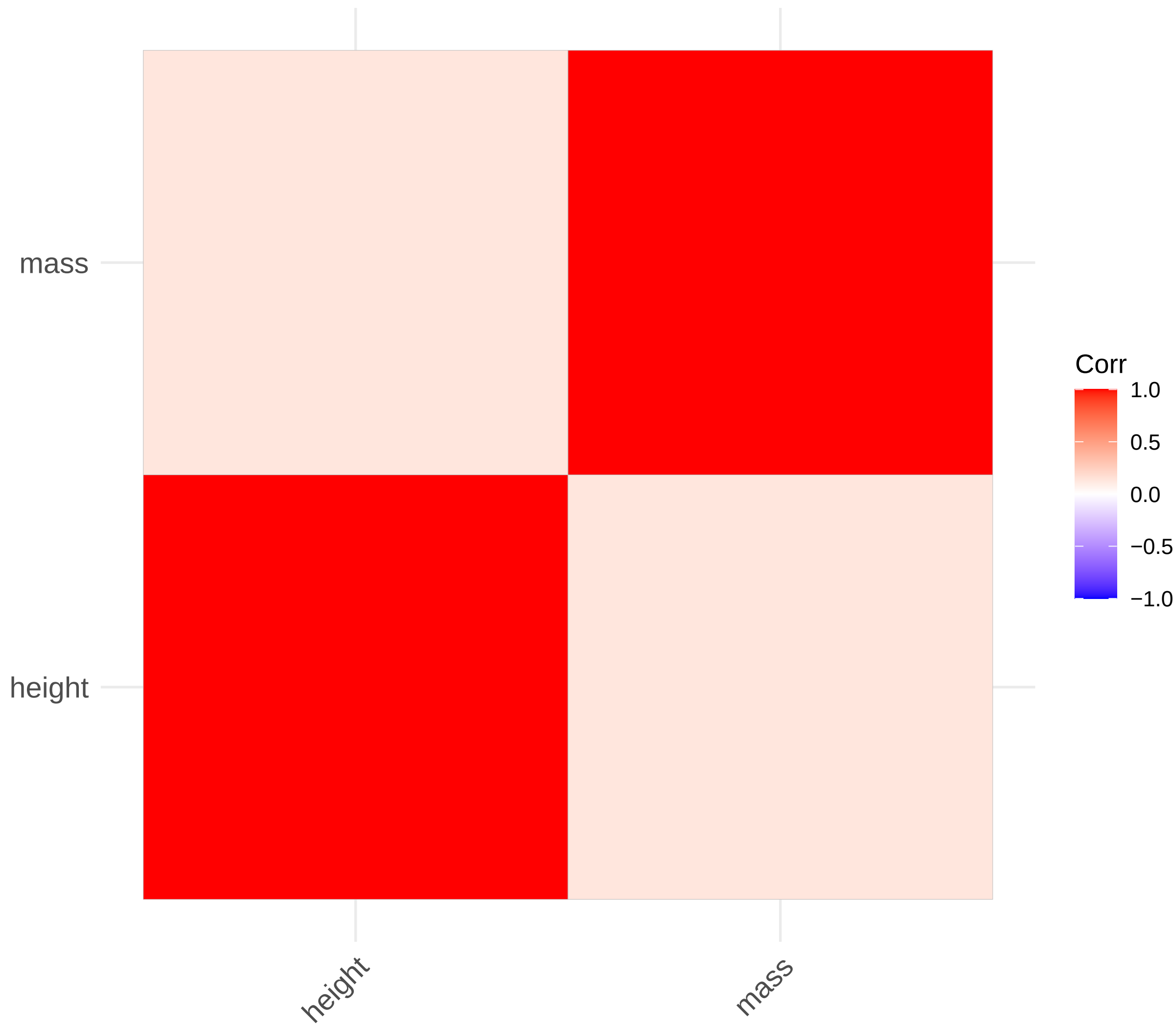
Figure 5. Plot of correlation/regression between height and mass.

You can also visualize the correlation matrix with functions from other packages, including `corPlot()` from the `psych` package (Figure **??**), `corrplot()` from the `corrplot` package (Figure **??**), and `ggcorrplot()` from the `ggcorrplot` package (Figure @(fig:ggcorrplot)).


Correlation plot from data
height
mass
1.00
0.13
0.13
1.00
height
mass
1
0.8
0.6
0.4
0.2
0
−0.2
−0.4
−0.6
−0.8
−1

102





None of these are thrilling with just two variables, but they can be very useful when you're using a correlation matrix across many variables. Figure

Hypothesis Testing

Hypothesis testing is anything you might usually think of as "results." Essentially: *do these data suggest some kind of non-random pattern?* The best hypothesis test to use will depend on a few factors, most significantly the data type of the independent (predictor) and dependent (outcome) variables.

This flowchart is a quick-and-dirty, imperfect cheatsheet:

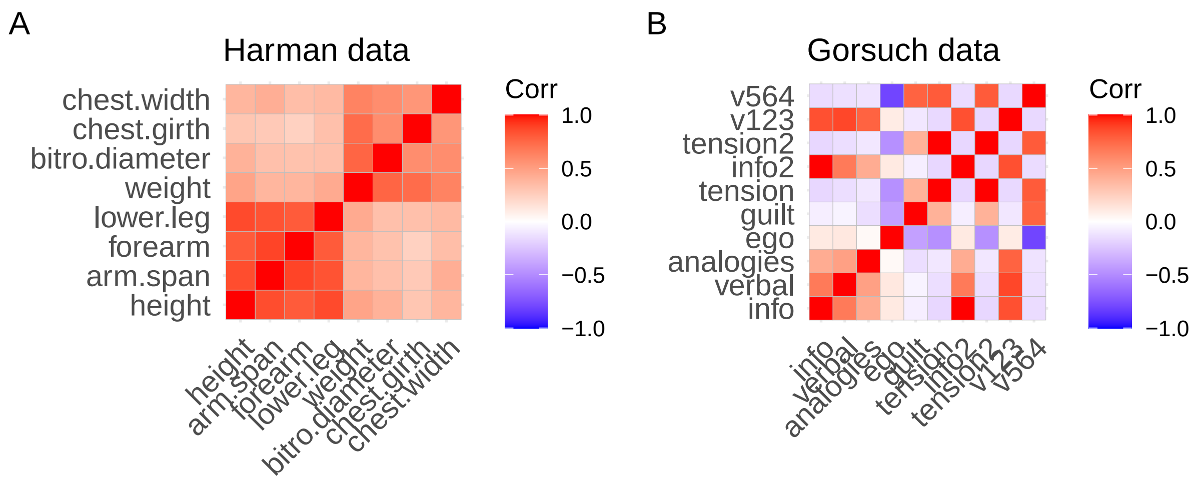
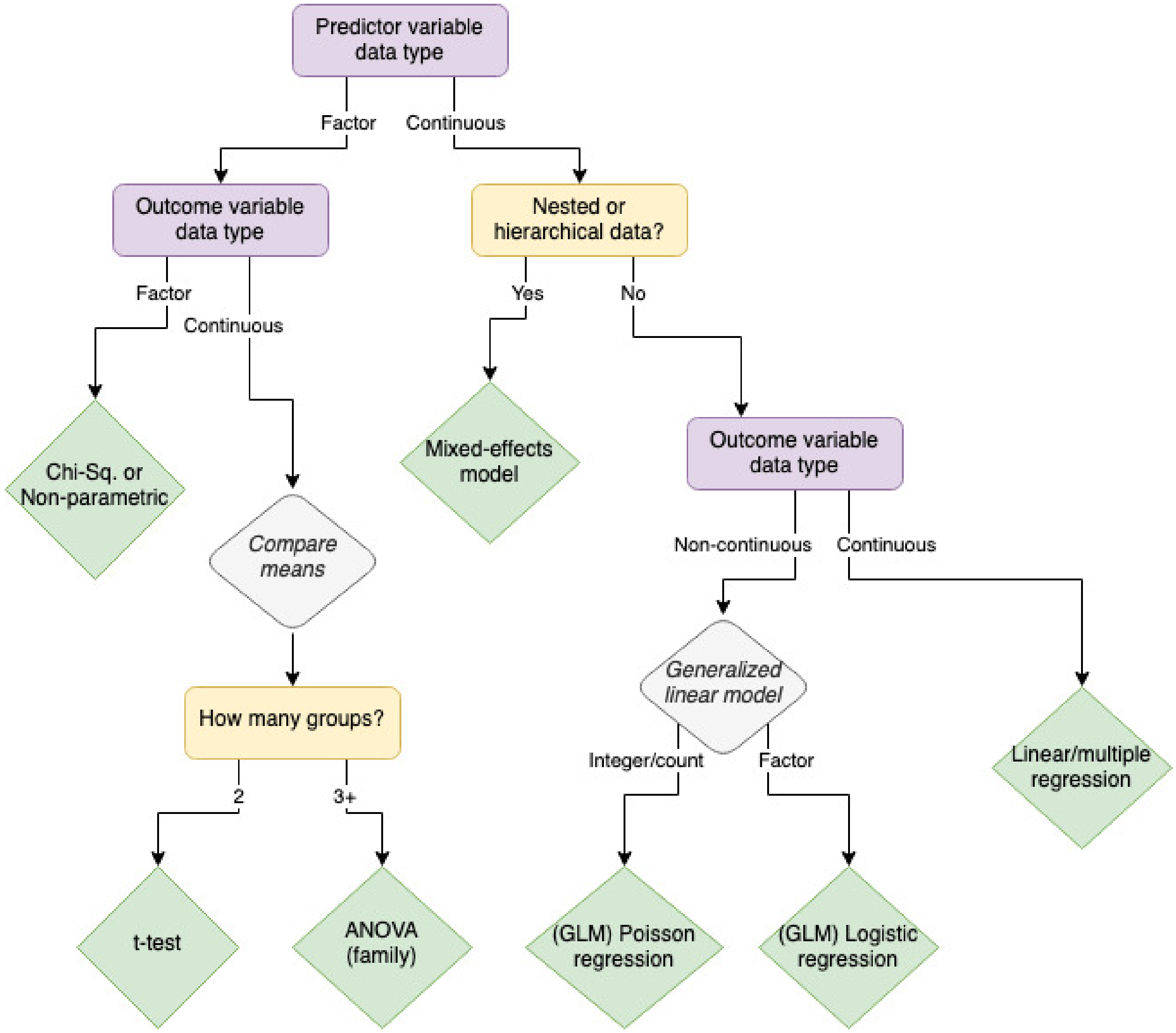


Figure 6. Correlation matrices for complex data, like the Harman (A) and Gorsuch (B) datasets in the psych package.



Categorical Predictors

t-tests.

1-sample. A 1-sample t-test tells you the likelihood that the "true" mean of a value is not equal to 0 (or another reasonable, specific alternative).

For example, a 1-sample t-test on the `mass` variable should return significant results, rejecting the null hypothesis that the true mean is 0. Since mass is necessarily a positive value, it is impossible that the true mean would be 0.

```

	One Sample t-test

data:  sw.desc$mass
t = 4.4109, df = 58, p-value = 4.525e-05
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
  53.15109 141.47264
sample estimates:
mean of x
 97.31186
```

We can alternatively specify the mean that the null hypothesis should assume. Let's assume that the `starwars` dataset contains the mass of literally every character in Star Wars. In that case, the true population mean mass for Star Wars characters is 97.31 kg. We can specify the null/true mean with the `mu` (/*mu*) argument.

Is the mean of this sample different than the true mean?

```

	One Sample t-test

data:  sw.desc$mass
t = 0, df = 58, p-value = 1
alternative hypothesis: true mean is not equal to 97.31186
95 percent confidence interval:
  53.15109 141.47264
sample estimates:
mean of x
```

```
 97.31186
```

Obviously not, since the "sample" is what the true mean was calculated on. But if we consider the full dataset the true, full population, we can compare a sample to that population.

```

    Welch Two Sample t-test

data:  . and sw.desc$mass
t = -3.2264, df = 81.373, p-value = 0.001807
alternative hypothesis: true difference in means is not equal to 97.31186
95 percent confidence interval:
 -32.10151  66.62445
sample estimates:
mean of x mean of y
114.57333  97.31186
```

In nearly all cases (depending on your random seed) this will result in rejecting the null hypothesis. Essentially this is showing that there are some values (that of one mister The-Hutt, with a mass of 1358) of mass that is such an outlier it makes the mean of the full sample not actually representative of the "average". (This is a case where median might be a better measure measure of center than mean.) If we get rid of the extreme outlier and use that as the "true" mean, things might look different.

Now randomly sampling from the dataframe will usually *not* be significantly different from that mean. Sometimes it will be though, just because of random variation. Sometimes is will be *extremely* significantly different. Why?

```
    Welch Two Sample t-test

data:  . and sw.desc$mass
t = -0.89444, df = 100.33, p-value = 0.3732
alternative hypothesis: true difference in means is not equal to 75.57586
95 percent confidence interval:
 -23.1651 112.9494
sample estimates:
mean of x mean of y
142.20400  97.31186
```

2-sample. While a 1-sample t-test compares a sample mean against a static value (like 0), a 2-sample t-test compares two sample means against each other. The null hypothesis of a 2-sample t-test is that the true means of the group are not different.

Is the mass of male characters different from female characters?

```
    Welch Two Sample t-test

data:  filter(sw.desc2, sex == "male")$mass and filter(sw.desc2, sex == "female")$mass
t = 4.8612, df = 43.298, p-value = 1.571e-05
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 14.94106 36.11926
sample estimates:
mean of x mean of y
 80.21905  54.68889
```

For all characters, yes. We can reject the null hypothesis that the means are the same ($t = 4.86$, $p < 0$).

What if we just look at the humans?

```

	Welch Two Sample t-test

data:  filter(sw.desc2, sex == "male", species == "Human")$mass and filter(sw.desc2, sex
t = 2.8744, df = 2.7808, p-value = 0.06986
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -4.648771 63.417398
sample estimates:
mean of x mean of y
 85.71765  56.33333
```

For humans, no. We cannot reject the null hypothesis that the means are the same ($t = 2.87$, $p < 0.07$).

Outside the Star Wars Cinematic Universe, we know that the mean mass of male humans is higher than that of female humans. Rather than looking for *any* difference in means, we have a theoretical reason to look for a difference in one particular direction. If we set `alternative = "greater"`, the null hypothesis is that the true difference in means (mean-of-males - mean-of-females) is less than or equal to 0.

```

	Welch Two Sample t-test

data:  filter(sw.desc2, sex == "male", species == "Human")$mass and filter(sw.desc2, sex
```

```
t = 2.8744, df = 2.7808, p-value = 0.03493
alternative hypothesis: true difference in means is greater than 0
95 percent confidence interval:
 4.533443      Inf
sample estimates:
mean of x mean of y
 85.71765  56.33333
```

Now we do see a significant effect. We can reject the null hypothesis that the mean mass of females is greater than or equal to that of males are the same ($t = 2.87$, $p < 0.04$).

Important optional arguments for t-tests:

- True mean (μ): `mu`
  - In a 1-sample test, the null hypothesis will compare the mean to 0 by default. You can change this to the "true mean".
- Alt hypothesis: `alternative = c("two.sided", "less", "greater")`
  - By default this tests that the 1-var mean is not equal to 0 (or μ) or that the 2-vars means are not equal to each other. If you are specifically looking to demonstrate that the mean is greater than or less than 0 (or μ) or that one particular group's mean is greater than the others (e.g., you expect the control group to have poorer outcomes than the treatment/intervention group), set this to `less` or `greater`.
- Paired: `paired = FALSE`
  - If the observations are related in some way, you can use a paired t-test. For example if you want to compare growth between pre-test and post-test, you're more interested in the change for each individual rather than either mean test score *per se*.

- Confidence level: `conf.level = 0.95`
  - Set an alternative confidence interval when comparing means. This is rarely changed; 95% is almost always the expectation here.

ANOVA. Think of an Analysis of Variance (ANOVA) as an extension of the t-test. With a t-test you can compare the mean of 1 group to a static value or the means of 2 groups to each other. The basic functionality of ANOVA is to allow you compare three or more groups.

ANOVA is a whole family of analyses, but we'll focus on just 1-way ANOVA and 2-way ANOVA. One-way ANOVA is appropriate when there is one categorical independent variable with multiple levels, while two-way ANOVA is used when there are two categorical independent variables and their interaction effect needs to be examined.

1-Way ANOVA. **Example:** A psychologist wants to compare the effectiveness of three different stress reduction techniques (e.g., mindfulness meditation, progressive muscle relaxation, and deep breathing exercises) on reducing anxiety levels among participants.

One-way ANOVA can be used to test for significant differences in anxiety levels (dependent variable, continuous) across the three stress reduction techniques (independent variable, factor).

If the p-value from the ANOVA test is significant, post-hoc tests (e.g., Tukey's HSD) can be conducted to determine which techniques differ significantly from each other.

```
            Df Sum Sq Mean Sq F value Pr(>F)
sex3cat      2   5917    2958   2.541 0.0853 .
Residuals   78  90822    1164
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
6 observations deleted due to missingness
```

```
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = height ~ sex3cat, data = sw.hyp)

$sex3cat
                   diff       lwr       upr     p adj
male-female    7.551378 -16.76734 31.870094 0.7393848
other-female -18.471429 -52.22770 15.284846 0.3953726
other-male   -26.022807 -53.97481  1.929192 0.0733375
```

Here there is a trending but non-significant difference in height across the 3 sex categories $F() = 2.54$, $p = 0.08$.

Using Tukey post-hoc adjustment we can see this difference is primarily driven by the difference in height between those in the "male" and "other" category.

2-Way ANOVA. **Example:** A psychologist conducts a study to investigate the effects of both gender (male vs. female) and stress level (low vs. high) on performance in a cognitive task.

In this scenario, there are two independent variables: gender (with two levels: male and female) and stress level (with two levels: low and high). The dependent variable is performance in the cognitive task. Two-way ANOVA would be used to assess the main effects of gender and stress level, as well as their interaction effect on performance. The interaction effect indicates whether the effect of one independent variable depends on the level of the other independent variable.

```
            Df Sum Sq Mean Sq F value Pr(>F)
sex3cat      2   5917  2958.4   2.520 0.0873 .
```

```
hair4cat     3   2771   923.6   0.787 0.5051
Residuals   75  88052  1174.0
---
Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
6 observations deleted due to missingness

  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = height ~ sex3cat + hair4cat, data = sw.hyp)

$sex3cat
                   diff       lwr       upr     p adj
male-female    7.551378 -16.88666 31.989414 0.7412438
other-female -18.471429 -52.39333 15.450471 0.3985203
other-male   -26.022807 -54.11195  2.066339 0.0751112

$hair4cat
                   diff       lwr      upr     p adj
dark-blond    0.8780075 -46.57273 48.32875 0.9999583
light-blond -16.0775689 -76.47250 44.31736 0.8969627
none-blond    7.5086536 -39.94208 54.95939 0.9756109
light-dark  -16.9555764 -59.92405 26.01289 0.7284309
none-dark     6.6306461 -14.58997 27.85126 0.8443509
none-light   23.5862225 -19.38225 66.55469 0.4773422

            Df Sum Sq Mean Sq F value Pr(>F)
sex3cat      2   9114    4557   3.914 0.0243 *
```

```
gender        1   2167    2167   1.862 0.1766
Residuals    73  85000    1164
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
10 observations deleted due to missingness

  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = height ~ sex3cat + gender, data = sw.hyp)

$sex3cat
                   diff       lwr       upr     p adj
male-female    7.551378 -16.79951 31.902268 0.7394416
other-female -33.071429 -72.90621  6.763355 0.1228312
other-male   -40.622807 -75.66121 -5.584402 0.0190413

$gender
                       diff       lwr      upr    p adj
masculine-feminine 3.518817 -16.04968 23.08732 0.721092
```

Aside from being used as a hypothesis test itself, another important use for ANOVA is comparing model fit. For example, you create 3 possible regressions to test whether household income and/or proximity to grocery stores affects stress level using one variable, both variables, or both and an interaction effect. Passing these models to the `anova()` function can tell you which model best explains a predictive effect, so you can move forward just using that model.

We can use the `mtcars` dataset to show a simple example: Does horsepower and/or weight predict a car's fuel consumption?

```
Call:
lm(formula = mpg ~ hp, data = mtcars)

Residuals:
    Min      1Q  Median      3Q     Max 
-5.7121 -2.1122 -0.8854  1.5819  8.2360 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 30.09886    1.63392  18.421  < 2e-16 ***
hp          -0.06823    0.01012  -6.742 1.79e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.863 on 30 degrees of freedom
Multiple R-squared:  0.6024,    Adjusted R-squared:  0.5892 
F-statistic: 45.46 on 1 and 30 DF,  p-value: 1.788e-07
```

```
Call:
lm(formula = mpg ~ hp + wt, data = mtcars)

Residuals:
   Min     1Q Median     3Q    Max 
-3.941 -1.600 -0.182  1.050  5.854 
```

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 37.22727    1.59879  23.285  < 2e-16 ***
hp          -0.03177    0.00903  -3.519  0.00145 **
wt          -3.87783    0.63273  -6.129 1.12e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.593 on 29 degrees of freedom
Multiple R-squared:  0.8268,    Adjusted R-squared:  0.8148
F-statistic: 69.21 on 2 and 29 DF,  p-value: 9.109e-12


Call:
lm(formula = mpg ~ hp * wt, data = mtcars)

Residuals:
    Min      1Q  Median      3Q     Max
-3.0632 -1.6491 -0.7362  1.4211  4.5513

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 49.80842    3.60516  13.816 5.01e-14 ***
hp          -0.12010    0.02470  -4.863 4.04e-05 ***
wt          -8.21662    1.26971  -6.471 5.20e-07 ***
hp:wt        0.02785    0.00742   3.753 0.000811 ***
---
```

```
Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1

Residual standard error: 2.153 on 28 degrees of freedom
Multiple R-squared:  0.8848,    Adjusted R-squared:  0.8724
F-statistic: 71.66 on 3 and 28 DF,  p-value: 2.981e-13
```

All three models show a significant effect of the predictor variable(s). The question becomes which of these to use for the rest of the analyses and in the interpretation of our results. Comparing these models in an ANOVA tells us which model (if any) has a significantly better predictive fit.

```
Analysis of Variance Table

Model 1: mpg ~ hp
Model 2: mpg ~ hp + wt
Model 3: mpg ~ hp * wt
  Res.Df    RSS Df Sum of Sq      F    Pr(>F)
1     30 447.67
2     29 195.05  1   252.627 54.512 4.856e-08 ***
3     28 129.76  1    65.286 14.088 0.0008108 ***
---
Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
```

The p-values here indicate whether there is a significant difference in fit between one model and the model that came before it. Assuming significant difference, the best model fit is the one with the lowest residual sum of squares (RSS).

Note that depending on the type of models you're comparing, you might need to find the lowest value of something else. For example with mixed-effects models you'll (typically) look for the lowest BIC.

Chi-square. The Chi-Square Test is used to determine whether there is a significant association between categorical variables. You can think of it like a kind of "correlation" between categorical data.

Example: A psychologist conducting research on the effectiveness of different therapy interventions for treating phobias wants to examine whether there is a significant association between the type of therapy (exposure therapy, cognitive-behavioral therapy, or relaxation therapy) and the self-reported effectiveness (reduction of symptoms, increase in symptoms, or no change).

In this scenario, there are two categorical variables (therapy type & symptom change). The experimental design allows for a directional association: sensibly, therapy type is the predictor variable and symptom change is the outcome variable. The null hypothesis of a Chi-sq. test is that there is no association between the two variables. That is, a subject in any of the three therapy groups is equally likely to fall into any of the three outcome groups.

Like with correlation of continuous variables, directionality isn't required. Amount of time spent outside and amount of time spent with family may be positively correlated, but it's not clear which would cause the other (if either). There may be a significant association between favorite ice cream flavor and favorite candy flavor, but it's not clear that one of those is the independent predictor and the other the dependent outcome.

In the Star Wars dataset, we can use chi-sq. to look for associations between any of the factor variables we've already defined (or that existed in the original dataset). For example, is there a relationship between sex and hair color?

A contingency table shows the frequency of observations in each possible combination of factor levels:

```
        blond dark light none
```

```
  female      1    9     1    5
  male        3   24     4   29
  other       0    7     0    4
```

The chi-sq. test compares this contingency table to what we'd expect if the observations were evenly distributed *based on the number of observations per level* within each variable (i.e., not just dividing the total number of observations up evenly across all cells.

```

    Pearson's Chi-squared test

data:  sex_hair_table
X-squared = 3.9272, df = 6, p-value = 0.6865
```

These results are not significant. We can't reject the null hypothesis that there is any non-random relationship between sex and hair color.

What about the relationship between hair color and skin color?

```

             blond dark light none
  cool hue       0    3     1    9
  fair/light     4   24     2    4
  metallic       0    1     0    2
  other          0    1     2   14
  tan/dark       0    9     0    5
  warm hue       0    2     0    4
```

```
    Pearson's Chi-squared test

data:  skin_hair_table
X-squared = 38.94, df = 15, p-value = 0.0006543
```

In this case, $\chi^2 = 38.94$ ($p < .001$). We can reject the null hypothesis and claim that there is an association between hair and skin color in characters in the Star Wars Universe. We cannot make any claims about direction of the association.

Continuous Predictors

Linear Regression. Linear regression models the relationship between a continuous dependent variable and one or more (i.e., multiple regression) independent variables, at least one of which is also continuous. Linear modeling can also incorporate interaction effects between predictors.

Example: A psychologist is interested in understanding the relationship between hours of study per week and exam scores among college students. Using a linear regression to model the this relationship shows 1) whether there is a relationship, 2) whether that association is statistically significant, 3) the association's direction, and 4) the magnitude of the association.

Since both variables in this case are continuous, the psychologist could have used a correlation instead of a regression. One advantage of the regression is that the magnitude of the effect has more immediate application. Correlation is always normed to be between 0 and 1, so the magnitude of the correlation coefficient can be interpreted as a kind of percentage change.

With regression, the slope (magnitude) is not normed and applies directly to the variables. It can be interpreted as change-in-outcome per change-in-predictor, i.e. the expected change (probably increase?) in exam score for every additional hour of studying.

Without norming, the regression will also give an intercept, which tells you what the predicted value of y would be if $x = 0$ (i.e., what would we expect the exam score to be for someone who does not study at all?).

Another advantage of linear models is the opportunity to consider multiple predictor variables. Additional independent variables may be variables of interest (maybe both study hours and sleep hours affect exam scores) or one may be a control (maybe the effect of study hours differs based on students' pre-test scores).

```
         height     mass
height 1.000000 0.130859
mass   0.130859 1.000000


Call:
lm(formula = mass ~ height, data = sw.desc)

Residuals:
    Min      1Q  Median      3Q     Max
 -60.95  -29.51  -20.83  -17.65 1260.29

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -11.4868   111.3842  -0.103    0.918
height        0.6240     0.6262   0.997    0.323

Residual standard error: 169.5 on 57 degrees of freedom
Multiple R-squared:  0.01712,   Adjusted R-squared:  -0.0001194
F-statistic: 0.9931 on 1 and 57 DF,  p-value: 0.3232
```

The correlation between height and mass shows that there is some positive association: as height increases, mass increases too (though remember from the correlation matrix that this effect was not significant).

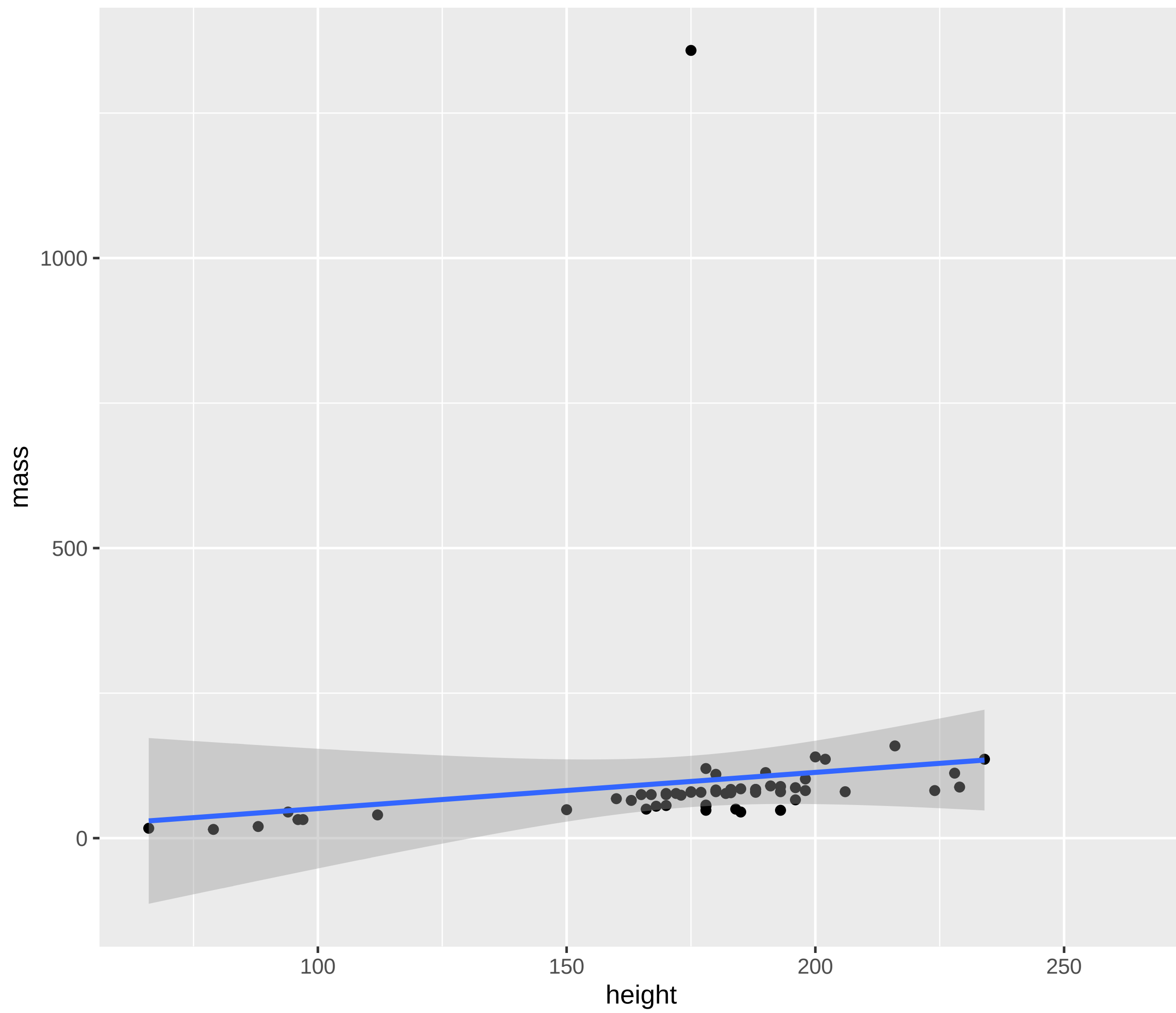
The linear model shows that for every additional unit of height (cm), mass increases by 0.62 units (kg), *but* this effect is not significant.

(BTW: To see that this very simple regression is doing basically the same thing as the correlation, compare the *p*-values of the `cor.test` and of the `height` slope!)

The lm also gives an intercept, which in this case is a fantastic example of when the intercept is simply not a useful thing to interpret: what should we expect the mass of a 0cm being to be? Apparently -11.49, which is nonsensical. First of all, mass cannot be negative, but that could potentially just be the result of a bad model fit. More clearly, a being cannot exist with literally 0 height. If a real observation's predictor measurement *can literally never be 0*, the intercept does not have a meaninful interpretation. It's still important for the model's overall functionality and fit, but only the slope will go into our interpretation of the results. **Be careful: the intercept will have its own significance value!** It's almost always the significance of the *slope* that matters, so don't get excited when you see `***` on the intercept line of the model output.

As discussed above, visualizing simple regression is the same as visualizing correlation: scatter plot and linear smoothing (Figure **??**):

```
‘geom_smooth()‘ using formula = ’y ~ x’
```



Multiple regression works exactly the same way. Add predictor or control variables to the right side of the formula. Connect them with an asterisk to look for an interaction effect.

Using the `mtcars` dataset, (how) do horsepower and displacement predict fuel efficiency (miles per gallon)?

```
Call:
lm(formula = mpg ~ hp, data = mtcars)
```

```
Residuals:
    Min      1Q  Median      3Q     Max 
-5.7121 -2.1122 -0.8854  1.5819  8.2360 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 30.09886    1.63392  18.421  < 2e-16 ***
hp          -0.06823    0.01012  -6.742 1.79e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.863 on 30 degrees of freedom
Multiple R-squared:  0.6024,	Adjusted R-squared:  0.5892 
F-statistic: 45.46 on 1 and 30 DF,  p-value: 1.788e-07


Call:
lm(formula = mpg ~ hp + disp, data = mtcars)

Residuals:
    Min      1Q  Median      3Q     Max 
-4.7945 -2.3036 -0.8246  1.8582  6.9363 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) 30.735904   1.331566  23.083  < 2e-16 ***
hp          -0.024840   0.013385  -1.856 0.073679 .  
disp        -0.030346   0.007405  -4.098 0.000306 ***
```

```
---
Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1

Residual standard error: 3.127 on 29 degrees of freedom
Multiple R-squared:  0.7482,    Adjusted R-squared:  0.7309
F-statistic: 43.09 on 2 and 29 DF,  p-value: 2.062e-09


Call:
lm(formula = mpg ~ hp + disp + hp * disp, data = mtcars)

Residuals:
    Min      1Q  Median      3Q     Max
-3.5153 -1.6315 -0.6346  0.9038  5.7030

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  3.967e+01  2.914e+00  13.614 7.18e-14 ***
hp          -9.789e-02  2.474e-02  -3.956 0.000473 ***
disp        -7.337e-02  1.439e-02  -5.100 2.11e-05 ***
hp:disp      2.900e-04  8.694e-05   3.336 0.002407 **
---
Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1

Residual standard error: 2.692 on 28 degrees of freedom
Multiple R-squared:  0.8198,    Adjusted R-squared:  0.8005
F-statistic: 42.48 on 3 and 28 DF,  p-value: 1.499e-10
```

In Model 1, which includes just one independent variable (`mpg ~ hp`), horsepower is negatively associated with fuel efficiency (β = -0.07, $p < .001$). That is, for every additional unit of horsepower we expect a reduction of 0.07 mpg.

Model 2 includes a second (continuous) predictor variable: displacement. In this regression, displacement (β = -0.03, $p < .001$) is a better predictor of mpg than horsepower, which in fact is no longer even significant (β = -0.03, $p < =$ .074).

Model 3 adds a potential interaction effect between horsepower and displacement. In this example, an interaction would mean that the strength of the effect on mpg of horsepower changes across changing levels of displacement. (As a simple psychology example, we might be interested in the interaction of age and sleep deprivation on exam scores. Sleep deprivation will probably lower exam scores for everyone, but it might lower them *a lot* for younger kids and just a little for older kids or vice versa). In this model, both horsepower and displacement have a significant effect on mpg, *and* there is a significant interaction effect ($\beta = 0$, p .002). The effect of displacement on mpg gets stronger as horsepower increases, above and beyond overall effects of displacement and horsepower.

If we have multiple reasonable models that give different results, which one should we use? We definitely don't want to create a bunch of models and pick the one that gives us the results we like the best. Instead, remember that ANOVA can compare model fit to help us make an informed and (relatively) impartial choice.

```
Analysis of Variance Table

Model 1: mpg ~ hp
Model 2: mpg ~ hp + disp
Model 3: mpg ~ hp + disp + hp * disp
  Res.Df    RSS Df Sum of Sq      F    Pr(>F)
1     30 447.67
```

```
2      29 283.49  1   164.181 22.662 5.339e-05 ***
3      28 202.86  1    80.635 11.130  0.002407 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

We look for the model that has the lowest residual sum of squares (RSS) *that is also* significantly improved from the next best model. In this case, Model 2 is a significantly better fit than Model 1 ($p < .001$), and Model 3 is a significantly better fit than Model 2 ($p = .002$). Moving forward, it makes sense to use Model 3 that includes the interaction between the two independent variables of interest.

Visualizing multiple regression. Visualizing relationships between more than two continuous variables gets very complicated very quickly. Although there are ways to plot a regression onto three axes (e.g., the `plot3D` package), it's not super easy to produce or interpret, and there's to way to create plot with more than 3 dimensions.

If you only have one continuous independent variable (the others are categorical or logical), you can use grouping strategies. Figure 7 demonstrates this approach to show the effects of horsepower, transmission type (`am`), and engine type (`vs`) on fuel efficiency using color grouping and faceting.

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

With multiple continuous predictors, you can use color, transparency, size, etc. to add another dimension without *literally* adding another dimension. Figure **??** again shows horsepower's primary effect on MPG, while color adds in information about displacement.

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

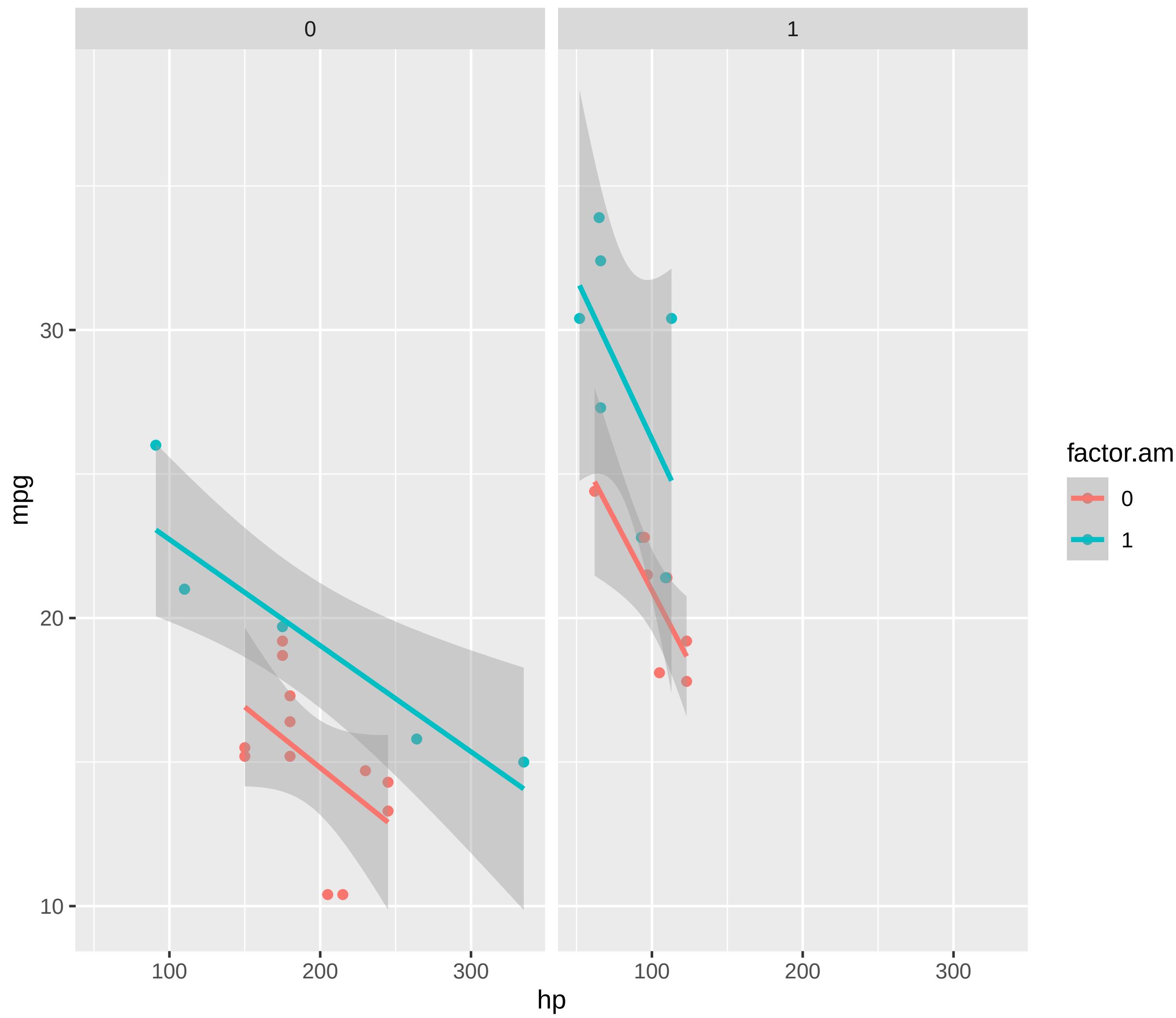
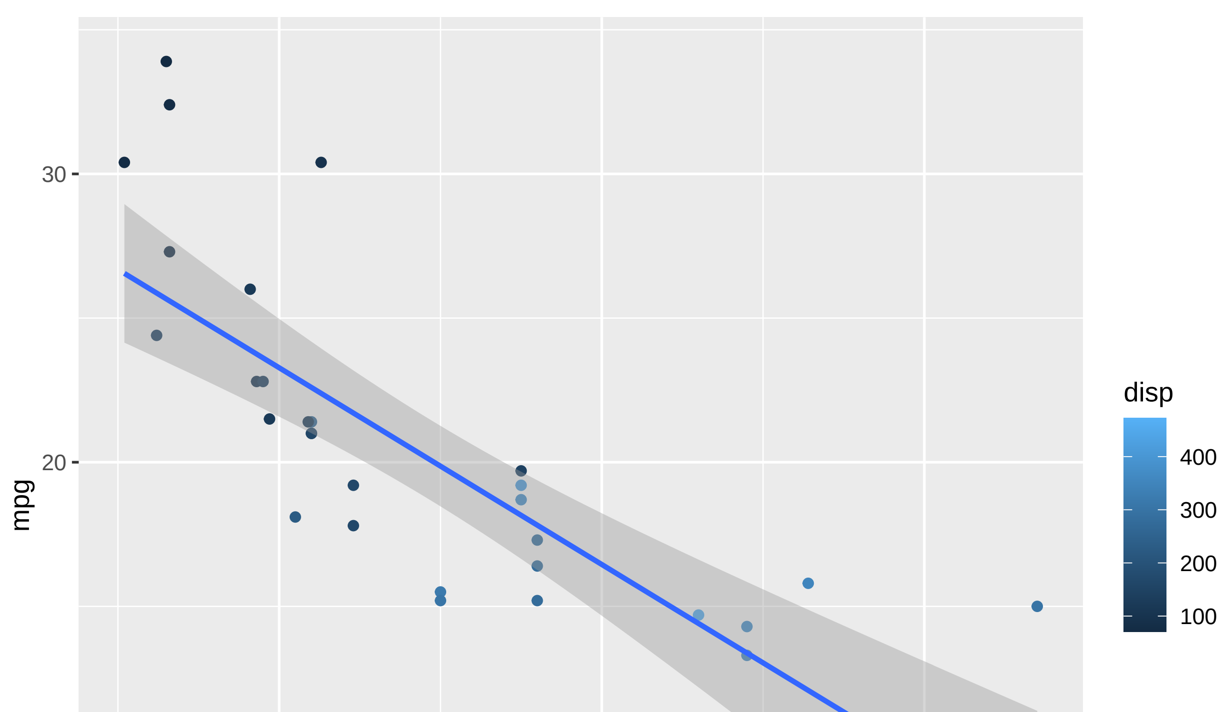
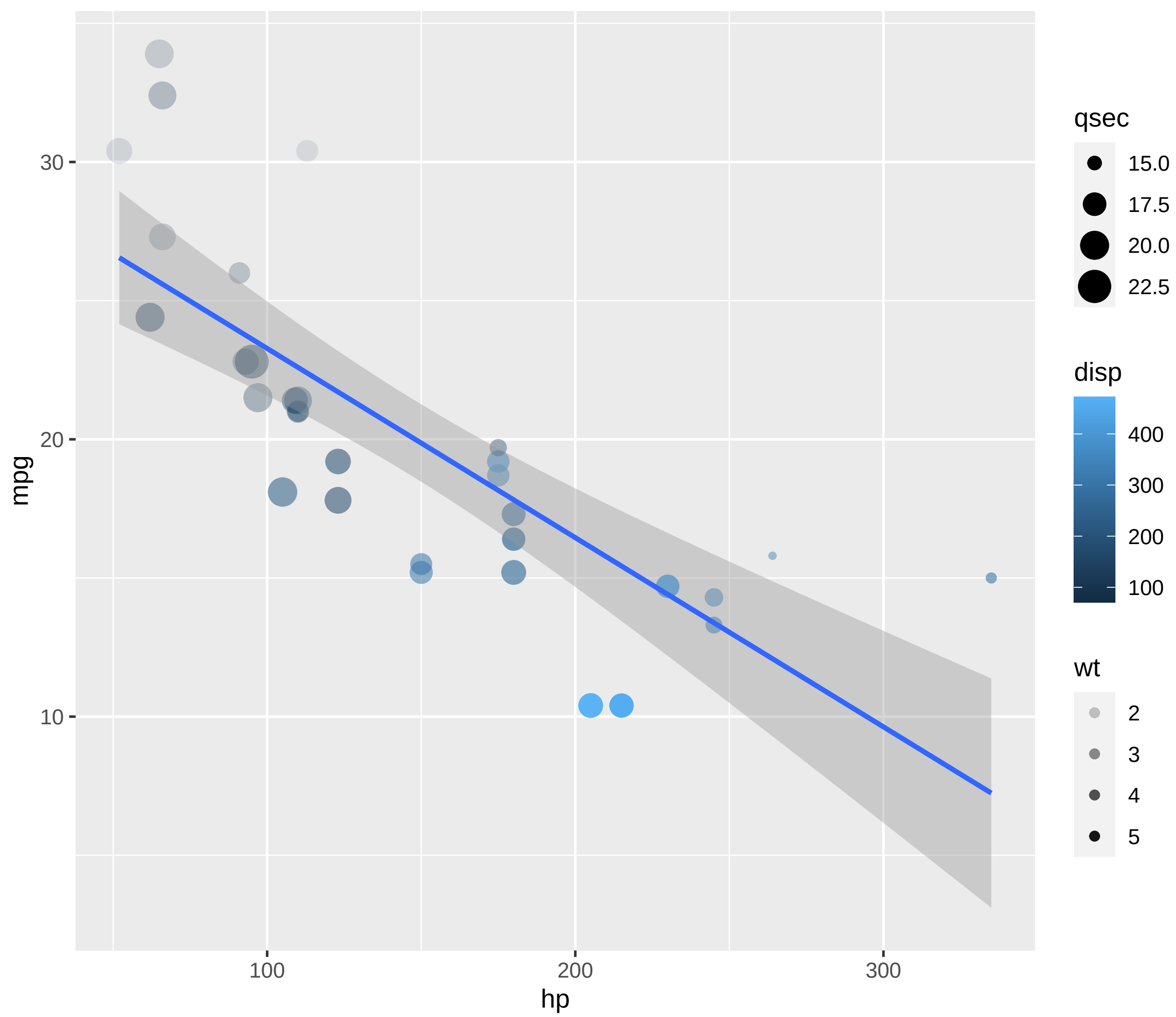


Figure 7. MPG by HP, across engine and transmission type



You can go crazy with even more continuous variables, but you probably shouldn't (Figure **??**).

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



Logistic Regression. Logistic regression is a type of generalized linear modeling (GLM) used when the dependent variable is binary (two categories). It models the probability of the occurrence of an event based on one or more independent variables. You can interpret a logistic regression as the change in probability that an outcome will occur given changes in your predictors.

Example: A psychologist is interested in identifying risk factors associated with the presence of anxiety disorders among college students, such as stress levels, sleep quality, and academic performance.

In this scenario, the outcome is not *how much* anxiety students experience (however you'd quantify that as a continuous variable), but simply the binary option of has-anxiety-disorder or doesn't-have-anxiety-disorder.

Note that this is a good example of where the direction between variables is not certain. In this model, we are treating the presence of an anxiety disorder as the outcome, which implies that the independent variables of stress, sleep, and academic performance are what lead to that diagnosis. While that may be what's happening, it's also reasonable to suspect that having an anxiety disorder is actually what leads to stress, sleep disturbance, and changes in academic performance. The logistic regression is still useful even if the cause-and-effect relationship is murky at best, so long as we are cautious and transparent when interpreting the results.

It is typical, but is not strictly necessary, that at least one predictor is continuous. If all predictors are categorical, it may be better to use something like a Chi-square test.

The `glm()` function in the `stats` package allows us to run logistic regressions (and other GLMs) with a syntax very similar to linear regression by specifying a distribution "family." For logistic regression, the "family" is "binomial." Here, rather than asking how much a change in horsepower will change MPG, we ask whether a change in horsepower changes the probability of a car being in the "High efficiency" category (defined as MPG above the median).

```
Call:
glm(formula = highMPG ~ hp, family = binomial, data = mt2)
```

```
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  7.62119    2.64469   2.882  0.00396 **
hp          -0.05901    0.02114  -2.791  0.00525 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 44.236  on 31  degrees of freedom
Residual deviance: 18.022  on 30  degrees of freedom
AIC: 22.022

Number of Fisher Scoring iterations: 7
```

Unsurprisingly (given what we saw with the linear models), higher horsepower makes it *less* likely that a car falls in the high efficiency category.

You can visualize logistic regression with point and smooth geoms just like "regular" (Gaussian) regressions. Specify the `glm` method and set the family to `binomial` with the syntax used here to produce Figure 8.

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

Notice that the y-axis goes from 0 to 1, and that all values fall either on y=0 or y=1. We can make that more interpretable by changing the y-axis labels (Figure 9).

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

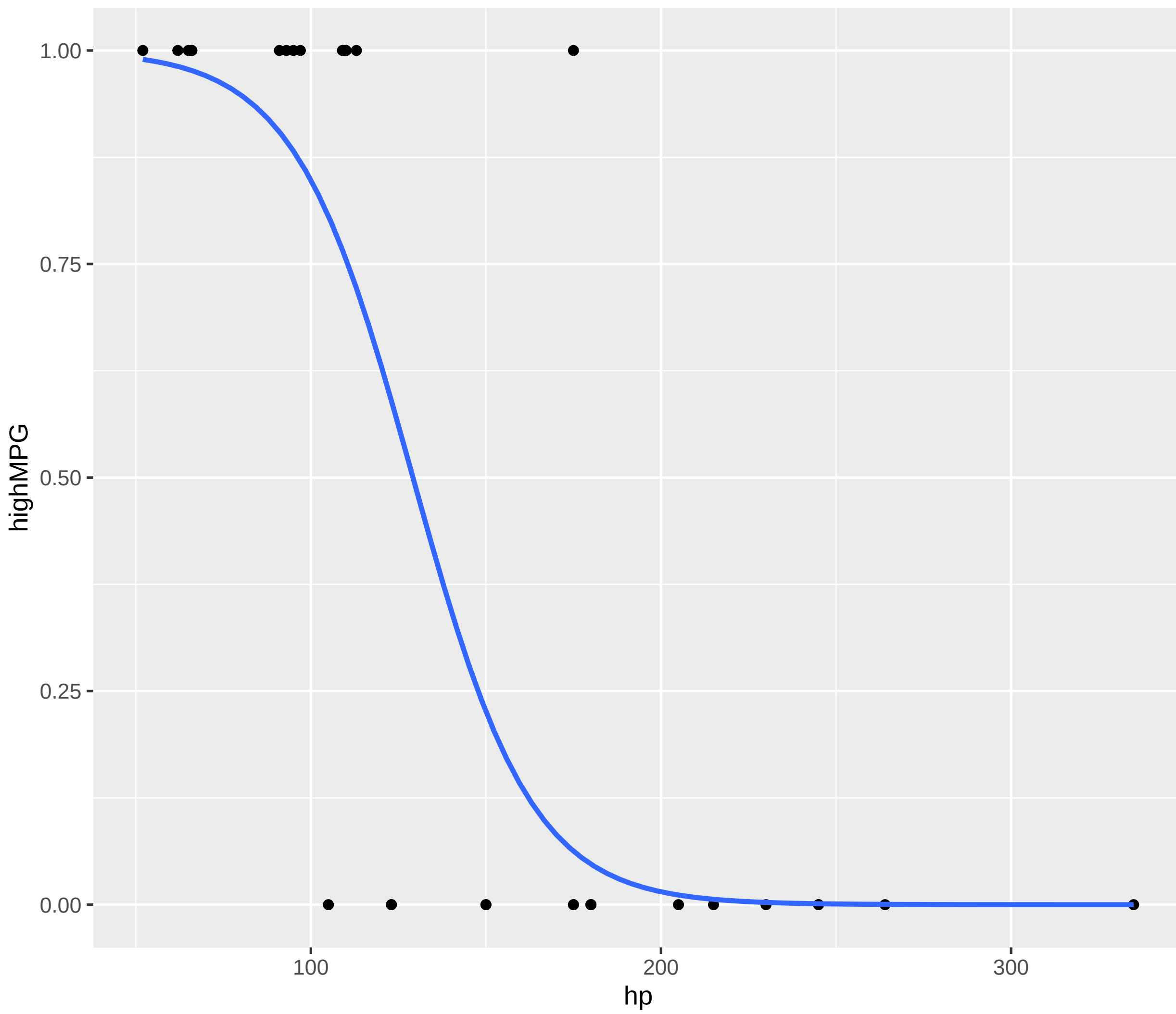


Figure 8. A logistic regression, a probabilistic relationship between horsepower and MPG

More...

- Poisson GLM (GLM with `family=poisson`)
  - used for count data
  - e.g., a psychologist wants to determine whether each of 3 intervention options decreases the number of times symptomatic behaviors are used in an observation period
- Generalized Linear Mixed-Effects Models (GLMM)

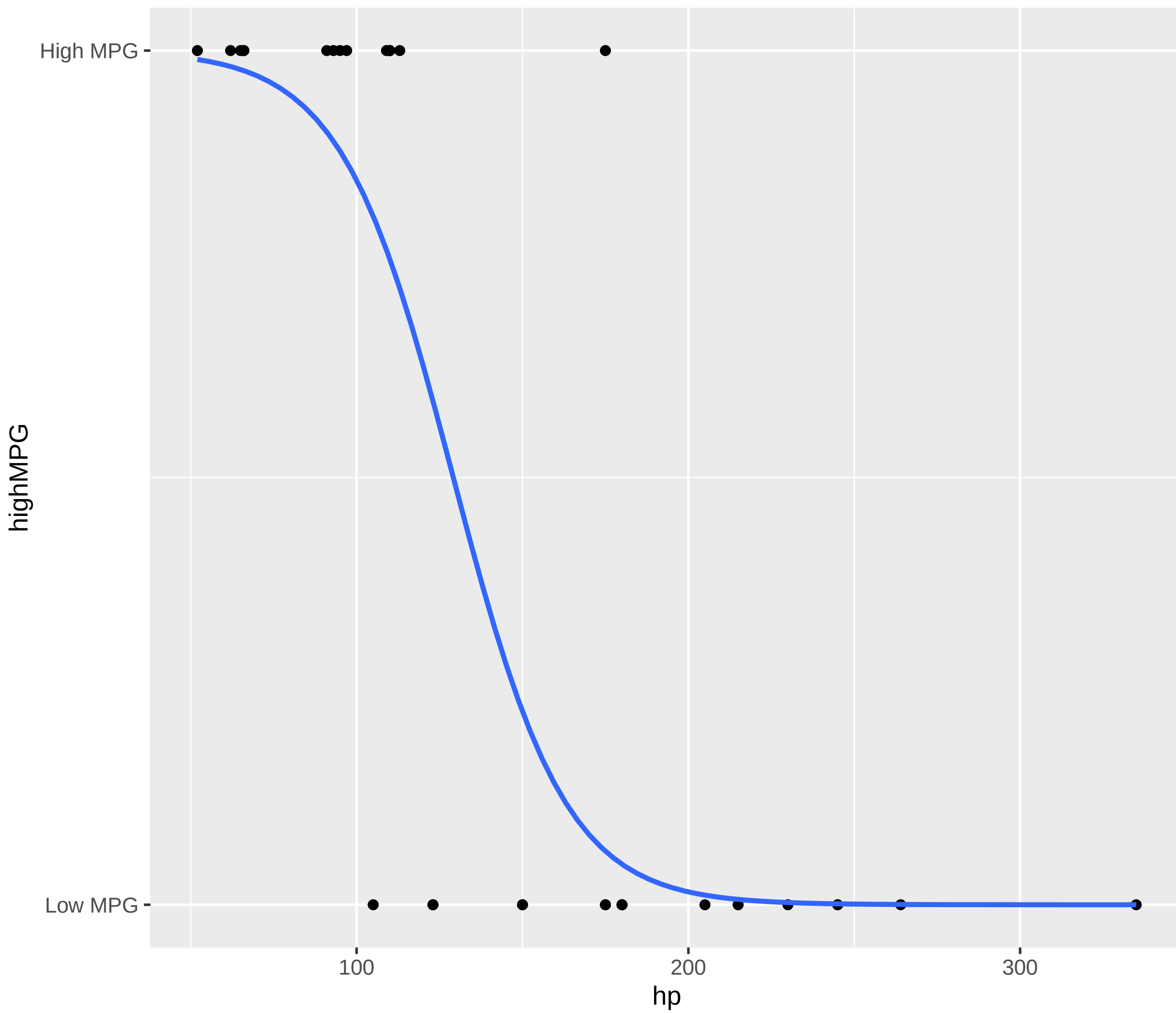


Figure 9. A logistic regression, a probabilistic relationship between horsepower and MPG with better axes

- used for nested or hierarchical data, where you need to account for random or spillover effects
- e.g., a psychologist want to determine the effectiveness of a teaching intervention. the intervention is administered at classroom level, but measured at the student level. the psychologist includes School ID as a random effect because they expect students will perform similarly to other students in their own school based on many factors unrelated to the intervention