Bivariate Inference

Overview

We've been visually exploring relationships between two variables by creating appropriate plots to assess how the distribution of a primary outcome (response/dependent) variable changes according to the level of a predictor (explanatory/independent/covariate) variable. We can learn a lot by conducting exploratory data analysis, and if description is the goal then this is where your work can stop.

However, if you want to make conclusions or **inference** about a relationship, then formal statistical analysis techniques are needed. We start here by formally testing if relationships or associations between *two* measures exist, then later will see how additional third variables can potentially disrupt or enhance any association that you may find.

Assignment Instructions

In this assignment you will practice **FIVE(5)** different types of bivariate analysis:

1. (Q~B) Quantitative Outcome ~ Binary Categorical Explanatory == Two-sample t-tests for a difference in means
2. (Q~C) Quantitative Outcome ~ Categorical Explanatory == ANOVA
3. (B~C) Binary Outcome ~ Categorical (or Binary) Explanatory == χ^2 test of Association.
4. (Q~Q) Quantitative Outcome ~ Quantitative Explanatory == Correlation analysis
5. (Q~Q) Quantitative Outcome ~ Quantitative Explanatory == Linear regression analysis

For each analysis you will do the following steps:

1. State which variable (including the variable name from your codebook) will be your explanatory variable and which will be your response variable.
    - Remember, you have some variables in your codebook that can act as both categorical and quantitative.
    - Decide which of those variables makes sense to "explain" the other. Don't just blindly pick a bunch of variables.
    - Think about the relationship among your variables, keeping in mind your original research questions. You may use gender as your categorical explanatory variable if you are struggling to find an explanatory and response relationship that makes sense.
2. Create an appropriate bivariate plot to visualize the relationship you are exploring. Summarize the relationship between the explanatory and outcome variables in short paragraph form.
3. Write the relationship you want to examine in the form of a research question.
    - State the null and alternative hypotheses as sentences.
4. Perform an appropriate statistical analysis using the full five step method as outlined in class and described below.
    - Define the parameters being tested. (ρ, p_1, μ_1, β_1 etc)
    - Translate the null and alternative hypotheses into H_0 and H_A with symbols.
    - State and verify assumptions of the test. Even if these assumptions are potentially violated, for the purposes of this assignment, acknowledge this limitation and continue with the prescribed analysis.
    - Conduct the analysis using your software program of choice. Make a decision whether or not to reject the null hypothesis. State your p-value.
5. Write a conclusion in context of the problem.

Submission Instructions:

- This is a team assignment, but NOT peer reviewed.
- Use this template to answer the questions.
  - Save the file as `team_name_bivariate.Rmd`
- After each completed question - upload your document (word or pdf) to the assignment folder in Google Drive
  - This assignment is *not* officially peer reviewed, but allows others who may be struggling to see what you have done.
  - Everyone is encouraged to leave comments and questions on your peers' submission.
  - You have access to a large learning community - contribute to it and use it!
- Submit the final version as a PDF to Blackboard Learn by the due date.

```
library(knitr)
opts_chunk$set(warning=FALSE, message=FALSE)
library(dplyr)
library(ggplot2)


load("C:/Box Sync/Data/AddHealth/addhealth_clean.Rdata")
```

(Q~B) Two sample T-Test for Independent Groups

We would like to know, is there convincing evidence that the average BMI differs between men and women?

Example

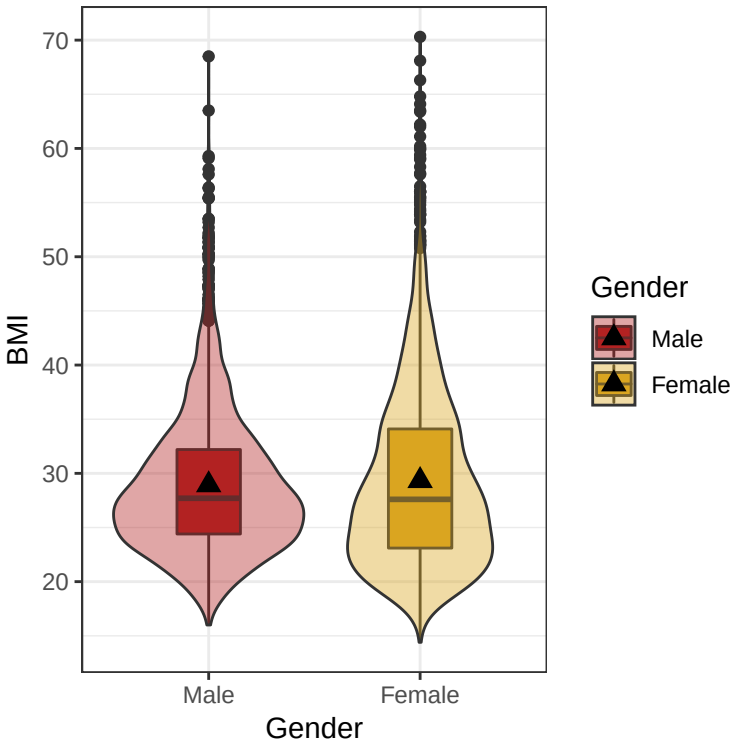
1. **Identify response and explanatory variables**

- Gender = binary explanatory variable (variable `female_c`)
- BMI = quantitative response variable (variable `BMI`)

2. **Visualize and summarise bivariate relationship**

```
plot.bmi.gender <- addhealth %>% select(female_c, BMI) %>% na.omit()

ggplot(plot.bmi.gender, aes(x=female_c, y=BMI, fill=female_c)) +
      geom_boxplot(width=.3) + geom_violin(alpha=.4) + theme_bw() +
      labs(x="Gender") +
      scale_fill_manual(values=c("firebrick", "goldenrod"), name="Gender") +
      stat_summary(fun.y="mean", geom="point", size=3, pch=17,
      position=position_dodge(width=0.75))
```




```
addhealth %>% group_by(female_c) %>%
              summarise(mean=mean(BMI, na.rm=TRUE),
                        var = var(BMI, na.rm=TRUE))
```

```
## # A tibble: 3 x 3
##   female_c  mean   var
##   <fct>    <dbl> <dbl>
## 1 Male      28.9  43.5
## 2 Female    29.3  66.7
## 3 <NA>      27.2  NA
```

Males have on average a BMI of 28.9, a little smaller than the average BMI of women at 29.3. Females have more variation in their weights, but the distributions both look normal, if slightly skewed right.

3. **Write the relationship you want to examine in the form of a research question.**

- Null Hypothesis: There is no difference in the average BMI between males and females.

- Alternate Hypothesis: There is a difference in the average BMI between males and females.

4. **Perform an appropriate statistical analysis.**

I. Let μ_1 be the average BMI for males, and μ_2 be the average BMI for females.

II. $H_0 : \mu_1 - \mu_2 = 0$
$H_A : \mu_1 - \mu_2 \neq 0$

III. We are comparing the means between two independent samples. A Two-Sample T-Test for a difference in means will be conducted. The assumptions that the groups are independent is upheld because each individual can only be one gender at a time. The difference in sample means $\bar{x}_1 - \bar{x}_2$ is normally distributed - this is a valid assumption due to the large sample size and that differences typically are normally distributed. The observations are independent, and the variances are roughly equal ($67/44 = 1.5$ is smaller than 2).

IV. Do not reject the null hypothesis: p-value = 0.06.

```
t.test(BMI ~ female_c, data=addhealth)
```

```
##
##  Welch Two Sample t-test
##
## data:  BMI by female_c
## t = -1.8696, df = 5025.5, p-value = 0.0616
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.79672844  0.01889772
## sample estimates:
##   mean in group Male mean in group Female
##             28.93414             29.32305
```

5. **Write a conclusion in context of the problem.**

Males, on average, have a slightly lower 0.4 (-0.9, 0.03) BMI compared to females. This is a non-significant difference with a p-value of 0.06.

(Q~C) Analysis of Variance

Analysis of variance assesses whether the means of two or more groups are statistically different from each other. This analysis is appropriate whenever you want to compare the means (quantitative variables) of groups (categorical variables). The null hypothesis is that there is no difference in the mean of the quantitative variable across groups (categorical variable), while the alternative is that there is a difference.

Post-Hoc Analysis

Run Post Hoc tests ("Tukeys HSD", or "Duncan"), if your ANOVA is significant.

- We only do post hoc tests if the following 2 requirements are met
  - You have a statistically significant difference
  - Your explanatory variable has more than 2 levels
- The overall ANOVA can be significant and NOT have any significant differences when you look at the post hoc results. The reason is that the two analyses ask two different questions.
  - The ANOVA is testing the overall pattern of the data and asking if as a whole the explanatory variable has a relationship (or lack thereof) with the response variable.
  - The post hoc is asking if one level of the explanatory variable is significantly different than another for the response variable. The post hoc is not as sensitive to differences as the ANOVA.
- Differences in group means can be non-significant at the post hoc level, but significant at the ANOVA level.

Example

1. **Identify response and explanatory variables**

- Species of flower = categorical explanatory variable (variable `Species`)
- Length of the Petal = quantitative response variable (variable `Petal.Length`)

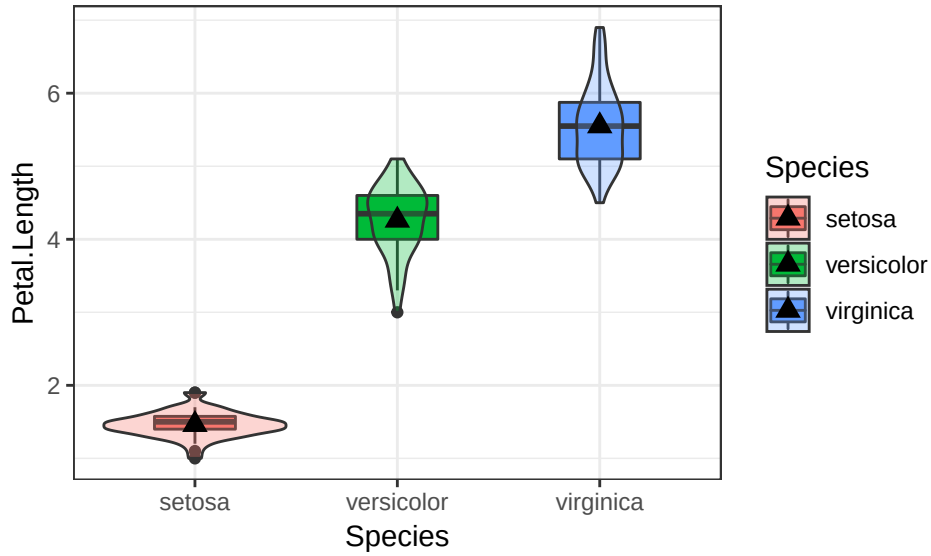
2. **Visualize and summarise bivariate relationship**

```
iris %>% group_by(Species) %>%
         summarise(mean=mean(Petal.Length),
```

```
                  sd=sd(Petal.Length),
                  n=n()) %>%
        kable()
```

Species	mean	sd	n
setosa	1.462	0.1736640	50
versicolor	4.260	0.4699110	50
virginica	5.552	0.5518947	50

```
ggplot(iris, aes(x=Species, y=Petal.Length, fill=Species)) +
  geom_boxplot(width=.4) + geom_violin(alpha=.3) +
  stat_summary(fun.y="mean", geom="point", size=3, pch=17,
    position=position_dodge(width=0.75)) + theme_bw()
```



There is clear difference in the average Petal length across iris Species. *Setosa* has an average petal length of 1.5 (sd 0.17), *Veriscolor* has an average petal length of 4.3 (sd 0.50), and *Virginica* has the largest average petal length of 5.6 (sd 0.55).

3. **Write the relationship you want to examine in the form of a research question.**

Is there a relationship between the Petal length of an iris flower and the species of flower?

- Null Hypothesis: There is no relationship between Petal length and Species.
- Alternate Hypothesis: There is a relationship between petal length and Species.

4. **Perform an appropriate statistical analysis.**

I. Let μ_1 be the true mean petal length for *Setosa* Let μ_2 be the true mean petal length for *Versicolor* Let μ_3 be the true mean petal length for *Virginica*

II. $H_0 : \mu_1 = \mu_2 = \mu_3$
H_A : At least one group mean is different.

III. I will conduct an analysis of variance using ANOVA. The distribution of petal length looks approximately normal within each species group. We can assume that the group means are normally distributed due to the sample size within each group $n = 50$ being large enough for the

CLT to hold. The assumption of equal variances may be violated here; the sd of *setosa* is less than half that of the other two species.

IV. Reject the null, the p-value is <.0001.

```
summary(aov(Petal.Length ~ Species, data=iris))
```

```
##              Df Sum Sq Mean Sq F value Pr(>F)
## Species       2  437.1  218.55    1180 <2e-16 ***
## Residuals   147   27.2    0.19
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Since the ANOVA was significant, I need to conduct a post-hoc test to identify which pairs are different.

```
TukeyHSD(aov(Petal.Length ~ Species, data=iris))
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = Petal.Length ~ Species, data = iris)
##
## $Species
##                       diff     lwr     upr p adj
## versicolor-setosa    2.798 2.59422 3.00178     0
## virginica-setosa     4.090 3.88622 4.29378     0
## virginica-versicolor 1.292 1.08822 1.49578     0
```

All pairs are significantly different from each other. The p-value for all post-hoc tests are all less than .0001.

The code below is just an example of how to make a better looking table

```
kable(TukeyHSD(aov(Petal.Length ~ Species, data=iris))$Species, digits=3)
```

	diff	lwr	upr	p adj
versicolor-setosa	2.798	2.594	3.002	0
virginica-setosa	4.090	3.886	4.294	0
virginica-versicolor	1.292	1.088	1.496	0

5. **Write a conclusion in context of the problem.**

There is sufficient evidence to conclude that the average petal length of an iris flower is associated with the species of the Iris (p<.0001). Specifically the length of the petal for species *Virginica* is 4.1 (95% CI 3.9, 4.3) cm longer than *Setosa*, and 1.3 (95%CI 1.1, 1.5) cm longer than *Versicolor*. *Versicolor* is also significantly longer than *Setosa* (2.8, 95% CI 2.6, 3.0) cm. All pairwise comparisons were significant at the .0001 level.

(B ~ C) Chi-Square Test of Association

A chi square (χ^2) test of association (independence) compares frequencies of one categorical variable for different values of a second categorical variable. The null hypothesis is that the relative proportions (values) of one variable are "independent" of the second variable; in other words, the proportions of one variable are the same for different values of the second variable. The alternate hypothesis is that the relative proportions of one variable are associated with the second variable

Although it is possible to run large chi square tables (i.e., 5 x 5, 4 x 6, etc.), the test is most easily interpretable when your response variable has only 2 levels. Therefore, if your chosen response variable has more than 2 levels, you must collapse it down to two levels like you did with the bivariate graphing assignment

Post hoc Analysis

The `RVAideMemoire` package provides the `fisher.multcomp` function to conduct post-hoc pairwise comparisons after a significant chi-squared test of association. See the RVAideMemoire Package reference for more information on how to use this function. There are many methods to control for multiple comparisons, the default is to use the *fdr* or "false discovery rate".

Example

1. **Identify response and explanatory variables**

- Gender = binary explanatory variable (variable `female`)
- General Health = categorical response variable (variable `genhealth`)

2. **Visualize and summarise bivariate relationship**

```
crosstab.gender.genhealth <- table(addhealth$female_c, addhealth$genhealth)
kable(crosstab.gender.genhealth)
```

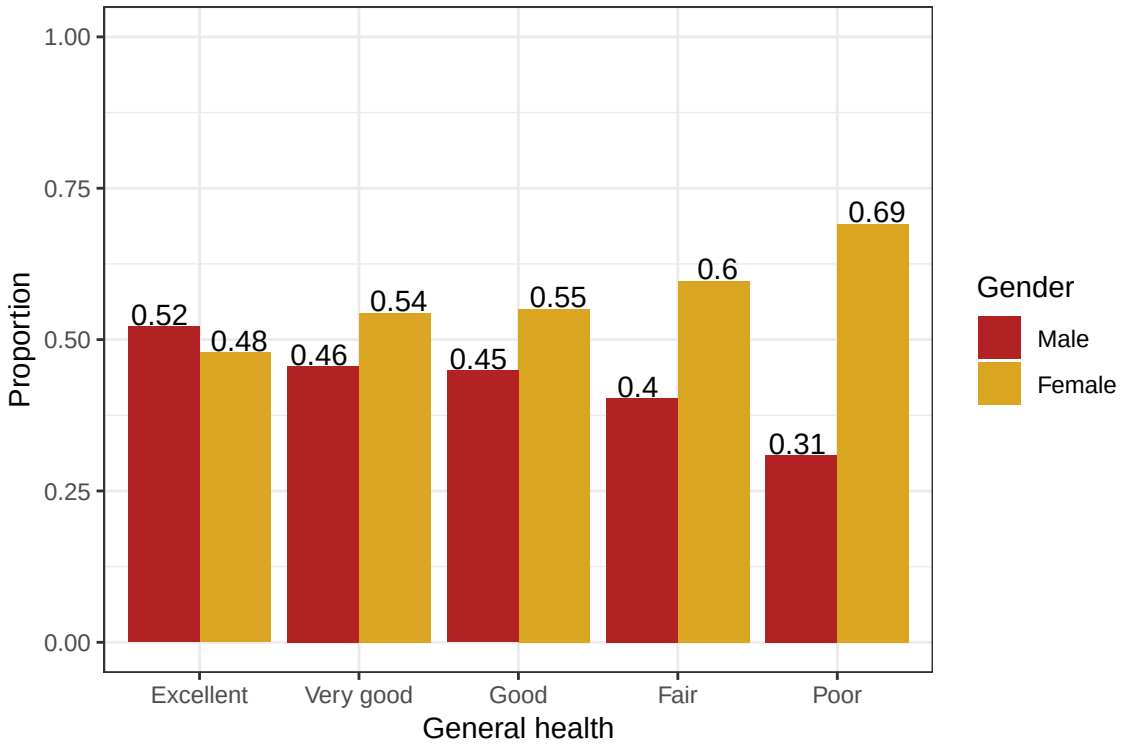
	Excellent	Very good	Good	Fair	Poor
Male	510	895	756	175	17
Female	468	1068	927	259	38

```
proptab.gender.genhealth <- prop.table(crosstab.gender.genhealth, margin=2)
kable(proptab.gender.genhealth, digits=3)
```

	Excellent	Very good	Good	Fair	Poor
Male	0.521	0.456	0.449	0.403	0.309
Female	0.479	0.544	0.551	0.597	0.691

```
plot.crostab <- data.frame(proptab.gender.genhealth)

ggplot(plot.crostab, aes(x=Var2, y=Freq, fill=Var1)) +
        geom_col(position = position_dodge()) + theme_bw() +
        geom_text(aes(y=Freq+.02, label=round(Freq,2)), position = position_dodge(width=1)) +
        labs(x="General health", y="Proportion") +
         scale_fill_manual(name="Gender", values=c("firebrick", "goldenrod")) +
        scale_y_continuous(limits=c(0,1))
```

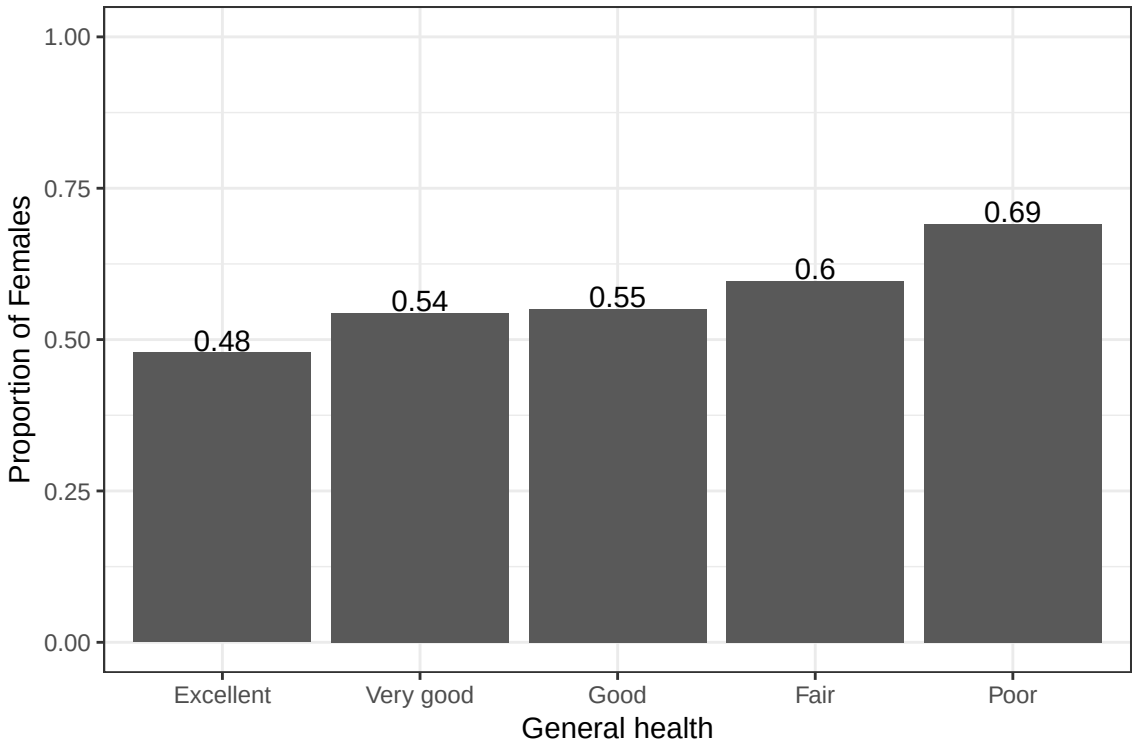


The distribution of gender appears to be different across the levels of general health categories. Females make up 48% of those reporting excellent health, and 69% of those reporting poor health

If your categorical variable is binary as in this example, what we're really interested in is the proportion of "yes" responses (here it's female). This is the group that we identify specifically in our hypothesis generation. To simplify the graph then, you can plot only the % of the group that you're interested in. Again, this only holds for binary categorical variables since within each categorical grouping variable (along the x axis) the proportions of the outcome variable will sum to 1. Note that I have to better specify what the proportion is along the Y axis now (i.e. *proportion of females* instead of simply *proportion*).

```
plot.crostab.female <- filter(plot.crostab, Var1=="Female")

ggplot(plot.crostab.female, aes(x=Var2, y=Freq)) +
        geom_col(position = position_dodge()) + theme_bw() +
        geom_text(aes(y=Freq+.02, label=round(Freq,2)), position = position_dodge(width=1)) +
        labs(x="General health", y="Proportion of Females") +
        scale_y_continuous(limits=c(0,1))
```



3. **Write the relationship you want to examine in the form of a research question.**

Is the gender distribution equal across all levels of general health?

- Null Hypothesis: The proportion of females in each general health category is the same.
- Alternate Hypothesis: At least one proportion is different.

4. **Perform an appropriate statistical analysis.**

   I. Let p_i be the true proportion of females in general health category i.

   II. $H_0 : p_1 = p_2 = p_3$
   H_A : At least one proportion is different.

   III. I will conduct a Chi-squared test of association. There is at least 5 observations in each combination of gender and general health.

   IV. Reject the null, the p-value is $<.0001$.

```
chisq.test(addhealth$genhealth, addhealth$female_c)
```

```
##
##  Pearson's Chi-squared test
##
## data:  addhealth$genhealth and addhealth$female_c
## X-squared = 26.471, df = 4, p-value = 2.543e-05
```

Since the test was significant, I will conduct a post-hoc test to identify which pairs are different.

```
RVAideMemoire::fisher.multcomp(crosstab.gender.genhealth, p.method="fdr")
```

```
##
##         Pairwise comparisons using Fisher's exact test for count data
##
## data:  crosstab.gender.genhealth
```

```
##
##             Excellent:Very good Excellent:Good Excellent:Fair
## Male:Female            0.002884       0.001696       0.000416
##             Excellent:Poor Very good:Good Very good:Fair Very good:Poor
## Male:Female       0.005538         0.6889        0.07518        0.07518
##             Good:Fair Good:Poor Fair:Poor
## Male:Female    0.1159   0.07518    0.2115
##
## P value adjustment method: fdr
```

The p-values for all tests comparing Excellent to other groups are all significant at $< .01$, Very good vs fair, and vs poor, and good vs poor are all marginally significant with p-values around 0.07.

5. **Write a conclusion in context of the problem.**

We can conclude that there is an association between gender and perceived general health (X2 = 26.4, df=4, p<.0001). The proportion of females in the Excellent health group (48%) is significantly lower than any other group 54% for Very Good to 69% for Poor). Furthermore there may be indiciation that there is a higher proportion of females in the Very good health group (54%) compared to the Fair (60%) and poor (69%) groups (p-values = .075).

(Q~Q) Correlation analysis

A correlation coefficient assesses the degree of linear relationship between two variables. It ranges from +1 to -1. A correlation of +1 means that there is a perfect, positive, linear relationship between the two variables. A correlation of -1 means there is a perfect, negative linear relationship between the two variables. In both cases, knowing the value of one variable, you can perfectly predict the value of the second.

Explain your results for each correlation in terms of strength (actual number value) and direction (using the sign and scatter plot) within the context of the relationship you are examining. Then discuss the r-squared value. Please put each explanation after the appropriate results.

Writing correlation values (replace highlighted words with numbers):

- `$r` (`N`) = `correlation coefficient`, p = `p-value`

Below are rough estimates for interpreting strengths of correlations (to be used in write up):

If r is ...

- +.70 and up Very strong positive relationship
- +.40 to +.69 Strong positive relationship
- +.30 to +.39 Moderate positive relationship
- +.20 to +.29 Weak positive relationship
- +.01 to +.19 No or negligible relationship
- -.01 to -.19 No or negligible relationship
- -.20 to -.29 Weak negative relationship
- -.30 to -.39 Moderate negative relationship
- -.40 to -.69 Strong negative relationship
- -.70 or below Very strong negative relationship

Correlation as a measure of model fit. When we square r (i.e. R^2), it tells us what proportion of the variability in one variable that is described by variation in the second variable.

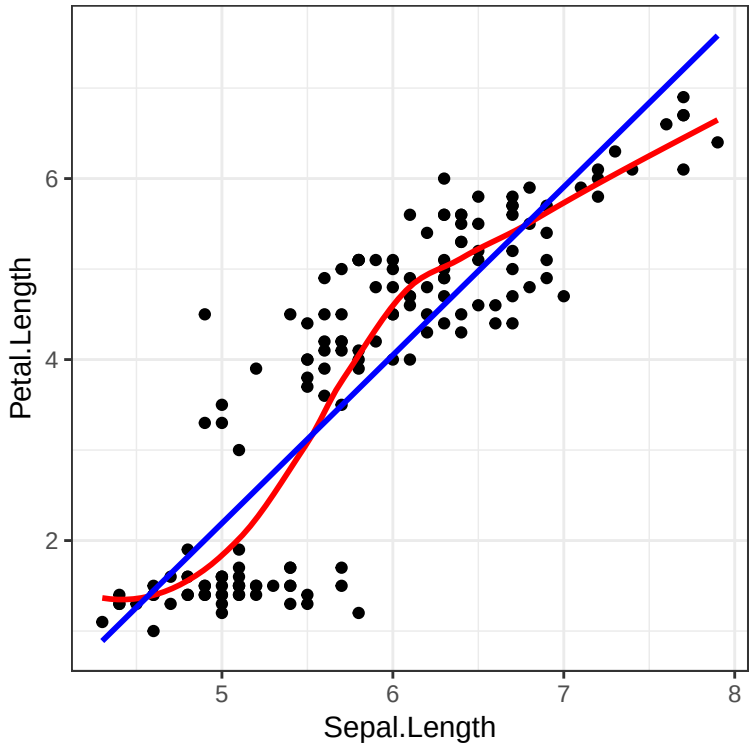
Example

1. **Identify response and explanatory variables**

- Sepal Length = quantitative explanatory variable (variable `Sepal.Length`)
- Petal Length = categorical response variable (variable `Petal.Length`)

2. **Visualize and summarise bivariate relationship**

```
ggplot(iris, aes(x=Sepal.Length, y=Petal.Length)) + geom_point() +
      geom_smooth(se=FALSE, col="red") +
      geom_smooth(method="lm", col="blue", se=FALSE) +
      theme_bw()
```




```
cor(iris$Sepal.Length, iris$Petal.Length)
```

```
## [1] 0.8717538
```

There is a strong, positive, linear relationship between the sepal length of the flower and the petal length (r=0.87).

3. **Write the relationship you want to examine in the form of a research question.**

- Null Hypothesis: There is no correlation between length of sepal and petal.
- Alternate Hypothesis: Sepal and petal lengths are correlated

4. **Perform an appropriate statistical analysis.**

I. Let ρ be the true correlation between sepal and petal length.

II. $H_0 : \rho = 0$
$H_A : \rho \neq 0$

III. Both variables are quantitative, a correlation analysis will be conducted.

```r
cor.test(iris$Petal.Length, iris$Sepal.Length)
```

```
##
##  Pearson's product-moment correlation
##
## data:  iris$Petal.Length and iris$Sepal.Length
## t = 21.646, df = 148, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.8270363 0.9055080
## sample estimates:
##       cor
## 0.8717538
```

IV. Reject the null hypothesis: p-value for ρ is <.0001.

5. **Write a conclusion in context of the problem.**

There was a statistically significant and very strong correlation between the sepal length of an iris and the petal length, r (148) = 0.87, p < .0001. The significant positive correlation shows that as the sepal length increases so does the petal length. These results suggest that 76% (95% CI: 68.9-82.8) of the variance in petal length can be explained by the length of the sepal.

(Q~Q) Linear Regression Analysis

Linear regression models can be used to evaluate whether there is a linear relationship between two numerical variables. When only one explanatory variable x is considered, this is termed simple linear regression (SLR). The basic idea of SLR is to use data to fit a straight line that relates the response Y to the predictor X. The mathematical relationship between x and y is written as

$$y_i \sim \beta_0 + \beta_1 x_i + \epsilon_i \qquad \epsilon_i \sim \mathcal{N}(0, \sigma^2)$$

Using a sample of data we can calculate point estimates b_0 and b_1 to estimate the population parameter values β_0 and β_1 respectively. Calculating these estimates uses a procedure called Least Squares Regression.

A linear regression analysis tests the hypothesis that there is no linear relationship between x and y ($\beta_1 = 0$) versus there is a linear relationship ($\beta_1 \neq 0$)

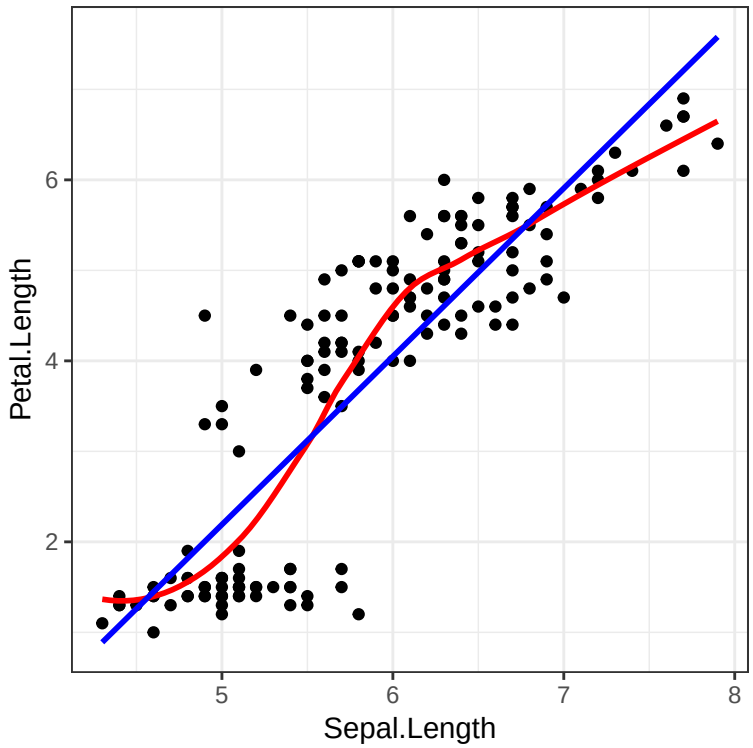
Example

1. **Identify response and explanatory variables**

- Sepal Length = quantitative explanatory variable (variable `Sepal.Length`)
- Petal Length = categorical response variable (variable `Petal.Length`)

2. **Visualize and summarise bivariate relationship**

```r
ggplot(iris, aes(x=Sepal.Length, y=Petal.Length)) + geom_point() +
      geom_smooth(se=FALSE, col="red") +
      geom_smooth(method="lm", col="blue", se=FALSE) +
      theme_bw()
```




```
cor(iris$Sepal.Length, iris$Petal.Length)
```

```
## [1] 0.8717538
```

There is a strong, positive, linear relationship between the sepal length of the flower and the petal length (r=0.87).

3. **Write the relationship you want to examine in the form of a research question.**

Does the length of the flower's sepal linearly correlate with the length of the flower's petal?

- Null Hypothesis: There is no linear relationship between length of sepal and petal.
- Alternate Hypothesis: Sepal and petal lengths are linearly related.

4. **Perform an appropriate statistical analysis.**

I. Let β_1 be the true measure of linear association between sepal and petal length.

II. $H_0 : \beta_1 = 0$
$H_A : \beta_1 \neq 0$

III. Both variables are quantitative, a linear regression analysis will be conducted. The first assumption that the relationship is linear is verified using the scatterplot in part 2. Further assumptions are that the residuals are normally distributed, centered around zero and have constant variance. All assumptions are checked after the model has been fit.

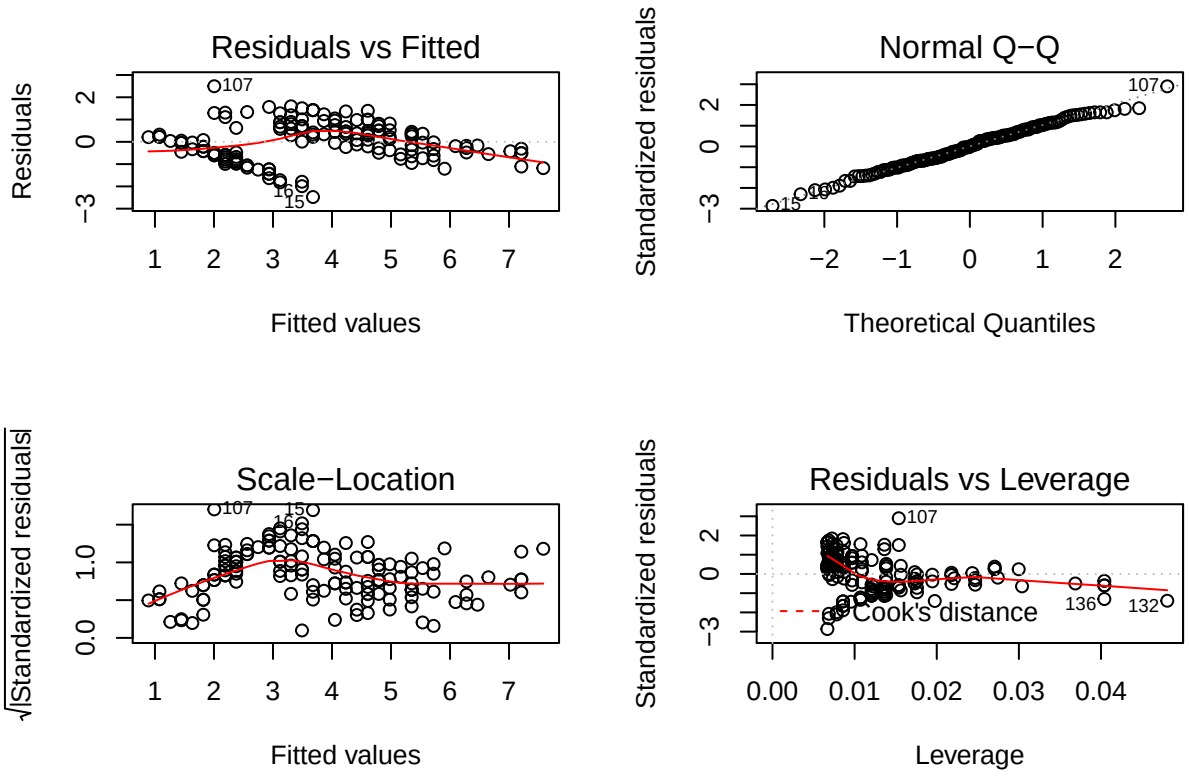
IV. Reject the null hypothesis: p-value for β_1 is $<.0001$.

```
iris.linear.model <- lm(Petal.Length ~ Sepal.Length, data=iris)
summary(iris.linear.model)
```

```
##
## Call:
## lm(formula = Petal.Length ~ Sepal.Length, data = iris)
##
```

```
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.47747 -0.59072 -0.00668  0.60484  2.49512
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -7.10144    0.50666  -14.02   <2e-16 ***
## Sepal.Length  1.85843    0.08586   21.65   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.8678 on 148 degrees of freedom
## Multiple R-squared:   0.76,  Adjusted R-squared:  0.7583
## F-statistic: 468.6 on 1 and 148 DF,  p-value: < 2.2e-16
```

```
par(mfrow=c(2,2))
plot(iris.linear.model)
```



The model assumptions appear to be upheld. The normal Q-Q plot indicate that the residuals are normally distributed, the residuals vs fitted plot indicate that the residuals are centered around zero. There might be a slight violation of the assumption of homoscedascitiy (constant variance), the variance of the residuals appears to decrease as the fitted values increase.

```
confint(iris.linear.model)
```

```
##                 2.5 %    97.5 %
## (Intercept)  -8.102670 -6.100217
## Sepal.Length  1.688772  2.028094
```

5. **Write a conclusion in context of the problem.**

The length of a petal and sepal on an iris flower are linearly correlated. For every 1 cm longer the sepal of the flower is, the petal length is increased by 1.85cm (95% CI 1.69, 2.03, $p<.0001$).