Regression Analysis

Lecture 9

1 Outline

- Definition of Terms
- Simple Linear Regression
- Multiple Linear Regression
- Diagnostics
- Correlation

2 Definition of Terms

2.1 Prediction vs. Explanation

> **Note**
>
> In a prediction problem, the investigators do not necessarily need to know the rule used to predict the outcome. Attention is focused on accuracy.
> In an explanation problem, the investigators will often assume a functional form for the dependent variable, similar to the linear model in Lecture 8.

> **Important**
>
> Prediction problems do not need a functional form, hence functional relations between independent and dependent variables will not be identified.

2.2 Correlation vs. Regression

Regression Analysis

Regression analysis is helpful in assessing specific forms of the relationship between variables, and the ultimate objective when this method of analysis is employed usually is to predict or estimate the value of one variable corresponding to a given value of another variable.

Correlation Analysis

Correlation analysis is concerned with measuring the strength of the relationship between variables. It cannot predict or estimate unlike regression.

3 Simple Linear Regression

3.1 Linear Regression

We are often interested in how one or more predictor/independent variables are associated with an outcome/response/dependent variable. In linear regression, we assume this association is defined by a linear function.

> **!** Important
>
> Linear regression models are used when the independent and dependent variables are both **continuous**.

> **!** Important
>
> It is important to keep in mind that our linear assumption is rarely met in practice. We may never know the true model, but information obtained from the linear model can still yield useful results, especially with exploratory data analysis.

3.2 Simple Linear Regression Model

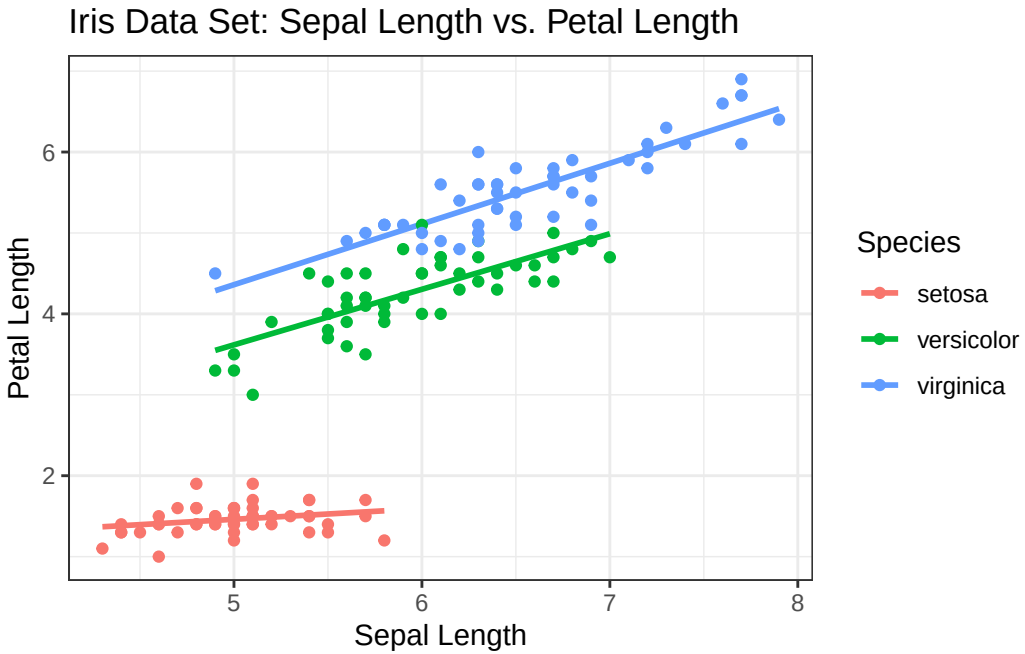
Recall Lecture 8: Linear regression models are a part of general linear models. In linear regression, a linear relationship is assumed between the independent variable X and the dependent variable Y. For the i^{th} observation (X_i,Y_i)

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$

> **!** Important
>
> The error term ε_i are independent and identically distributed Gaussian errors (mean 0 and variance σ^2). Recall that the error term contributes solely to the variance and not the mean of Y.

3.3 Example

```
-- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
v dplyr     1.1.4     v readr     2.1.5
v forcats   1.0.0     v stringr   1.5.1
v ggplot2   4.0.0     v tibble    3.3.0
v lubridate 1.9.4     v tidyr     1.3.1
v purrr     1.1.0
-- Conflicts ------------------------------------------ tidyverse_conflicts() --
x dplyr::filter() masks stats::filter()
x dplyr::lag()    masks stats::lag()
i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```



3.4 Method of Least Squares

To define the linear regression model, we need to estimate the following terms: the intercept, β_0; the slope β_1, and the variance σ^2. The intercept and the slope can be estimated using the **method of least squares**.

> **Note**
>
> The method of least squares can be visualized using the following link.

3.5 Estimation

Suppose we have an estimate of the intercept $\hat{\beta}_0$ and the slope $\hat{\beta}_1$. The estimated value of the dependent variable $\hat{Y}_i$ for a value of the independent variable X_i can be expressed as:

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i$$

> **!** Important
>
> The slope estimate $\hat{\beta}_1$ is the expected change in the dependent variable for a 1-unit increase in X_i. Generally, for a d unit increase in X_i, the dependent variable is expected to change by $d\hat{\beta}_1$.

3.6 Assumptions of Linear Regression

- The means of the subpopulations of Y all lie on the same straight line, also known as the linearity assumption.
- The values of Y are statistically independent.
- For each value of X there is a subpopulation of Y values. For the usual inferential procedures of estimation and hypothesis testing to be valid, these subpopulations must be normally distributed.
- The variances of the subpopulations of Y are all equal to σ^2.
- The independent variable can be measured without error or with negligible error.

> **!** Important
>
> The first four assumptions are also known as the LINE assumptions: **Linearity, Independence, Normality, Equal Variances**

3.7 Hypothesis Testing: Linear Regression

In examining the association between X and Y, the important parameter for inference is the slope β_1.

3.7.1 Slope and Association

> **!** Important
>
> When $\beta_1 = 0$, the best fit line resembles a horizontal line, indicating a lack of association between X and Y.
> If $\beta_1 > 0$, the slope of the best fit line is positive, indicating a positive direct association between X and Y.
> If $\beta_1 < 0$, the slope of the best fit line is negative, indicating a negative direct association between X and Y.
> Simply put, $\beta_1 \neq 0$ indicates that there is an association between X and Y.

3.7.2 Statistical Hypotheses

The null hypothesis of no association can be expressed as $H_0 : \beta_1 = 0$. The alternatives can be expressed as one-sided or two-sided.

- One-sided: $\beta_1 > 0$ or $\beta_1 < 0$
- Two-sided: $\beta_1 \neq 0$

3.8 Test Statistic: t

Suppose the estimated value of β_1 is $\hat{\beta}_1$. The test statistic when testing the hypotheses for β_1 can be expressed as:

$$t = \frac{\hat{\beta}_1}{\sqrt{S^2/S_{XX}^2}}$$

> **Note**
>
> S^2 is given by the following equation:
>
> $$S^2 = \frac{\sum_i (Y_i - \hat{Y}_i)^2}{(n-2)}$$
>
> The term $(Y_i - \hat{Y}_i)$ is also called the *residual* of the model. On the other hand, S_{XX}^2 is defined as the sum of squares of the independent variable.
>
> $$S_{XX}^2 = \sum_i (X_i - \bar{X})^2$$

3.9 Test Statistic: F

An analysis of variance (ANOVA) table can also be constructed from the regression model. This framework separates the source of variation accounted by the model and the error.

Source of Variation	Sum of Squares (SS)	Degrees of Freedom (DF)	Mean Squares (MS)	Variance Ratio (F)
Regression	SSR	1	MSR = SSR	MSR/MSE
Error	SSE	n-2	MSE = SSE/(n-2)	
Total	SST	n-1		

The test statistic $F = MSR/MSE$ follows an F-distribution with degrees of freedom $(df_1, df_2) = (1, n-2)$.

> **Note**
>
> This tests the following hypotheses: $H_0 : \beta_1 = 0$ and $H_a : \beta_1 \neq 0$.

3.10 Goodness of Fit: R^2

We define the coefficient of determination, also known as R^2, as the following ratio:

$$R^2 = SSR/(SSR + SSE)$$

> **!** Important
>
> R^2 is non-negative and cannot exceed 1. The R^2 measures the proportion of the variability explained by the predictor in the model. The higher the value of R^2, the closer the fit of the data to the model.
> $R^2 = 1$ indicates a perfect fit, but is almost always not observed in real-life situations.

3.11 `R` Implementation: `lm()`

The function `lm()` can perform calculations for linear regression. The function `summary()` is used with `lm()` to provide estimates for the intercept, slope, and residual variance. The `summary()` output also includes the results of the tests involving the slopes (both t and F).

Sample code should follow how we used `lm()` in Lecture 8:

```
mod1 <- lm(DepVar~IndepVar,data=df)
summary(mod1)
```

> **Tip**
>
> The `summary()` function also provides the R^2 value of the model.

3.12 `R` implementation: Visualization

The `ggplot2` package is primarily used for data visualization. This package is included in the `tidyverse` package.

> **Note**
>
> The `geom_point()` is used to plot the observed data points, while `geom_smooth(method,formula)` is used to overlay the best fit line on the observed data points.

3.12.1 Sample code

```
ggplot(data=df,aes(x=IndepVar,y=DepVar)) +
geom_point() +
geom_smooth(method="lm", formula=y~x)
```

3.13 Framework for Analysis

Here are the recommended steps in performing regression analysis:

- Visualize the data using `ggplot2()`. Is there a discernible linear trend?
- Obtain the equation of the best fit line using `lm()` and `summary()`.
- Evaluate the equation to obtain a measure of the strength of association between the independent and dependent variables (hypothesis tests).
- Use `geom_smooth()` as a litmus test of model fit.

3.14 Example

The data set `trees` (loaded in `R`) includes the measurements of the diameter (inches), height (ft), and volume (cubic ft) of timber in 31 felled black cherry trees. The diameter, labeled as Girth, is measured at 4'6" above the ground.

3.14.1 Question

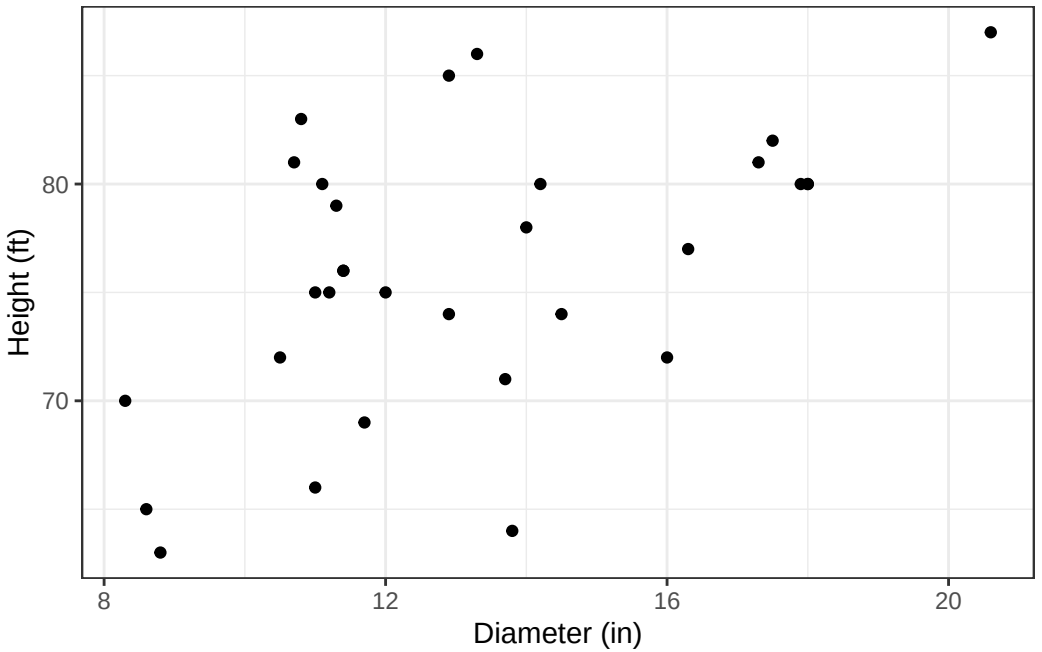
Test the hypothesis that there is an association between the diameter and height of the cherry trees. Use simple linear regression to estimate the best fit line that estimates the height of the cherry trees based on its diameter. What is the R^2 value?

3.14.2 Plot before Fit

```
library(tidyverse)
glimpse(trees)
```

```
Rows: 31
Columns: 3
$ Girth  <dbl> 8.3, 8.6, 8.8, 10.5, 10.7, 10.8, 11.0, 11.0, 11.1, 11.2, 11.3, ~
$ Height <dbl> 70, 65, 63, 72, 81, 83, 66, 75, 80, 75, 79, 76, 76, 69, 75, 74,~
$ Volume <dbl> 10.3, 10.3, 10.2, 16.4, 18.8, 19.7, 15.6, 18.2, 22.6, 19.9, 24.~
```

```
ggplot(data=trees,aes(x=Girth,y=Height)) +
  geom_point()  +
  theme_bw() +
  labs(x="Diameter (in)", y="Height (ft)")
```


Height (ft)
80
70
8
12
16
20
Diameter (in)

3.14.3 Fit

```r
mod1 <- lm(Height~Girth,data=trees)
summary(mod1)
```

```
Call:
lm(formula = Height ~ Girth, data = trees)

Residuals:
     Min       1Q   Median       3Q      Max
-12.5816  -2.7686   0.3163   2.4728   9.9456

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  62.0313     4.3833  14.152 1.49e-14 ***
Girth         1.0544     0.3222   3.272  0.00276 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 5.538 on 29 degrees of freedom
Multiple R-squared:  0.2697,    Adjusted R-squared:  0.2445
F-statistic: 10.71 on 1 and 29 DF,  p-value: 0.002758
```

The estimated equation for the tree height is $\widehat{height} = 1.0544 * Diameter + 62.0313$.

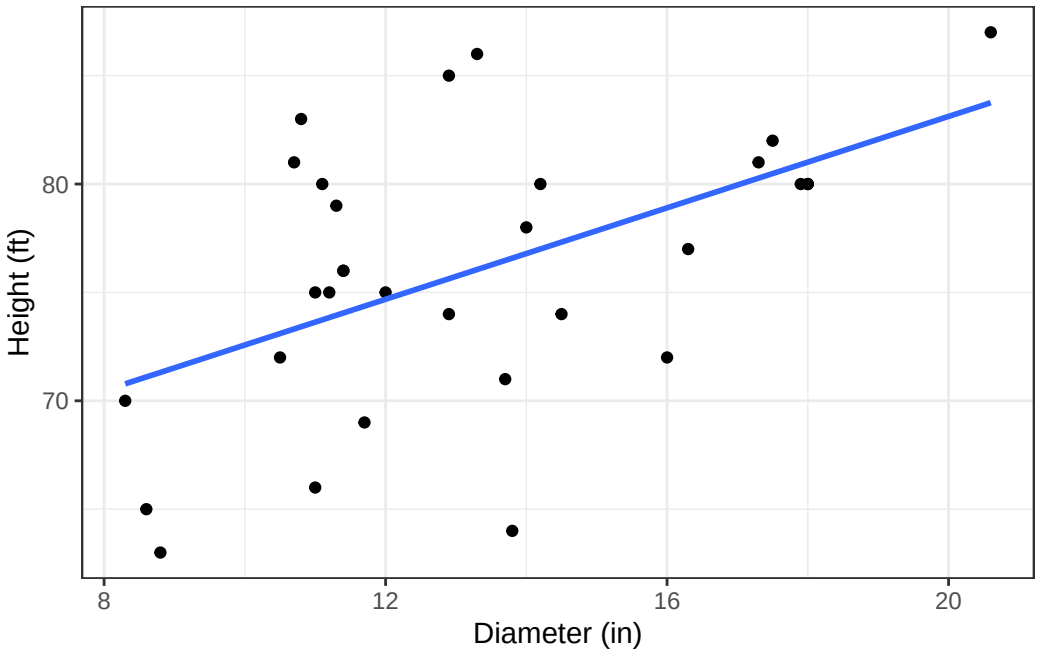
The p-value of the coefficient is 0.0028, and the test statistic t is 3.27. There is sufficient evidence of an association between diameter and height. The value of R^2 is 0.27.

3.14.4 Plot after Fit

```
library(tidyverse)
glimpse(trees)
```

```
Rows: 31
Columns: 3
$ Girth  <dbl> 8.3, 8.6, 8.8, 10.5, 10.7, 10.8, 11.0, 11.0, 11.1, 11.2, 11.3, ~
$ Height <dbl> 70, 65, 63, 72, 81, 83, 66, 75, 80, 75, 79, 76, 76, 69, 75, 74,~
$ Volume <dbl> 10.3, 10.3, 10.2, 16.4, 18.8, 19.7, 15.6, 18.2, 22.6, 19.9, 24.~
```

```
ggplot(data=trees,aes(x=Girth,y=Height)) +
  geom_point() +
  geom_smooth(method="lm", formula=y~x,se=F) +
  theme_bw() +
  labs(x="Diameter (in)", y="Height (ft)")
```


Height (ft)
80
70
8
12
16
20
Diameter (in)

3.15 `R` implementation: Estimation/Prediction

Recall that we can estimate the value of the dependent variable using our best fit line. Recall that for a value of the independent variable X, the estimated value of the dependent variable $\hat{Y}$ is:

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 * X$$

In `R`, we can use the `predict` function with `lm()` to calculate predicted values for one or more values of the independent variable. It can also output confidence and prediction intervals

3.15.1 Extrapolation

> **Warning**
>
> Estimation/Prediction using linear regression works best if the considered value of the independent variable is within the range of the data used to fit the regression line (interpolation). Estimation outside this range is referred to as extrapolation, which is not a recommended practice.

3.15.2 Intercept

The intercept is the estimated value of the dependent variable when the independent variable is equal to 0. Interpreting the intercept can be considered as extrapolation if 0 is not within the range of the independent variable. It might also be trivial to interpret the intercept if the case of $X = 0$ is trivial.

3.15.3 Confidence vs. Prediction Intervals

Confidence intervals are made for parameters, whereas an interval for random variables are called **prediction intervals**. Prediction interval are generally wider than confidence intervals.

> **!** Important
>
> For example, if we want to determine an interval estimate for the mean height of a tree with girth X, then we use a confidence interval. If we want to estimate the height of a particular tree with girth X, then we use a prediction interval

3.15.4 Sample Code

```
mod1 <- lm(DepVar~IndepVar,data=df)
newdata <- data.frame(IndepVar=c(X)) # substitute X with the value of the independent variable you want to estimate the value

newdata$prediction <- predict(mod1,newdata,interval = "confidence",level = 0.95)
```

```r
newdata$prediction <- predict(mod1,newdata,interval = "prediction",level = 0.95)
newdata
```

3.16 Example

The data set `trees` (loaded in `R`) includes the measurements of the diameter, height, and volume of timber in 31 felled black cherry trees. The diameter, labeled as Girth, is measured at 4'6" above the ground.

3.16.1 Question

- Calculate the estimated height of the black cherry trees with the following diameters: 10 inches, 15 inches, 20 inches with the corresponding prediction intervals.
- Is it advisable to interpret the intercept? Explain.
- Is it advisable to use this regression line to estimate the height of a black cherry tree with a 25-inch diameter? Explain.

3.16.2 Answer

```
mod1 <- lm(Height~Girth,data=trees)
newdata <- data.frame(Girth=c(10,15,20))
newdata$prediction <- predict(mod1,newdata=newdata,interval = "prediction",level = 0.95)
newdata
```

```
  Girth prediction.fit prediction.lwr prediction.upr
1    10       72.57500       60.86890       84.28110
2    15       77.84685       66.28041       89.41328
3    20       83.11869       70.77983       95.45755
```

It is not advisable to interpret the intercept because it is trivial to examine the case when the tree diameter is equal to zero. It is also outside the range of diameters considered in the study. Specifically, the range shown below does not include zero.

```
range(trees$Girth)
```

```
[1]  8.3 20.6
```

Similarly, X=25 is not in the range of diameters considered in the study and would lead to extrapolation if used in estimation. Hence, estimating the height of a tree with a 25-inch diameter using our best fit line is not recommended.

3.17 Exercise

Consider the liver steatosis data set `Liver_Steatosis.csv`. This dataset contains information on the 443 of 451 patients who had bariatric surgery at the Cleveland Clinic between 2005 and 2009 and underwent livery biopsy.

```
liver <- read.csv("Liver_Steatosis.csv")
```

3.17.1 Question

Suppose we want to investigate whether we can use an individual's weight in kg (variable weight) to estimate their LDL levels (variable LDL).

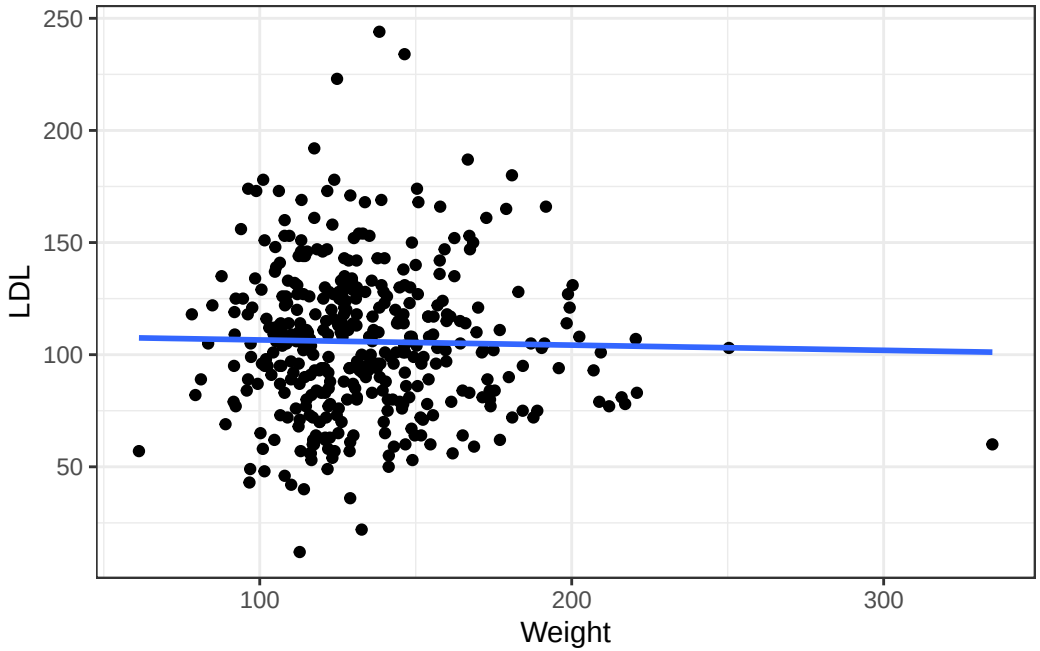
- Estimate the best fit line using regression methods.
- Test for an association between weight and LDL values.
- Predict the LDL value for an individual weighing 120 kg. Provide a 95% prediction interval.
- Provide the value of R^2.
- Create a scatter plot with the regression line overlayed using `ggplot`.

3.17.2 Answer

```
ggplot(data=liver,aes(x=Weight,y=LDL))+
  geom_point() +
  geom_smooth(method="lm", formula=y~x,se=F) +
  theme_bw()
```

```
Warning: Removed 48 rows containing non-finite outside the scale range
(`stat_smooth()`).
```

```
Warning: Removed 48 rows containing missing values or values outside the scale range
(`geom_point()`).
```



There appears to be a weak association between LDL and weight.

```r
mod1 <- lm(LDL~Weight,data=liver)
summary(mod1)
```

```
Call:
lm(formula = LDL ~ Weight, data = liver)

Residuals:
    Min      1Q  Median      3Q     Max
-94.282 -22.991  -0.458  19.363 138.310

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 108.89842    7.68701  14.167   <2e-16 ***
Weight       -0.02319    0.05618  -0.413     0.68
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 33 on 393 degrees of freedom
  (48 observations deleted due to missingness)
Multiple R-squared:  0.0004335, Adjusted R-squared:  -0.00211
F-statistic: 0.1704 on 1 and 393 DF,  p-value: 0.68
```

The best fit line is $\hat{LDL} = 108.90 - 0.02 * Weight$. The p-value of the coefficient is 0.68, and the test statistic t is -0.41. There is no evidence of an association between weight and LDL levels. The R^2 value is 4×10^{-4}. Note that **there are 48 observations deleted due to missingness**. These variables were excluded from the analysis.

```r
newdata <- data.frame(Weight=c(120))
newdata$prediction <- predict(mod1,newdata=newdata,interval = "prediction",level = 0.95)
newdata
```

```
  Weight prediction.fit prediction.lwr prediction.upr
1    120      106.11543       41.13065      171.10020
```

4 Multiple Linear Regression

4.1 Multiple Linear Regression

Often, we want to consider the effect of multiple predictors on the dependent variable. The simple linear regression can be extended to any number of predictor variables. We define X_{ij} as the value of predictor j on unit i.

The multiple linear regression model extended to p predictors is then:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + ... + \beta_p X_{ip} + \varepsilon_i$$

The estimated best fit line can be written as:

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_{1i} + \hat{\beta}_2 X_{2i} + ... + \hat{\beta}_p X_{ip}$$

> **!** Important
>
> The slope estimate $\hat{\beta}_k$ is the expected change in the dependent variable for a 1-unit increase in $X_k i$ assuming all other variables remain constant. Generally, for a d unit increase in $X_k i$, the dependent variable is expected to change by $d\hat{\beta}_1$.

4.2 Hypothesis Tests: Significance of Regression

The testing of significance of regression involves the null hypothesis $H_0 : \beta_1 = \beta_2 = ... = \beta_p = 0$, assuming that there is a total of p predictors. The alternative hypothesis that at least one coefficient is non-zero. The test of significance of regression is used to test if there are any predictors with a non-zero coefficient. This test uses the F-statistic similar to what was discussed for simple linear regression.

> **Note**
>
> `R` provides this information in the `summary()` output of `lm()` with the F-statistic and the corresponding p-value. The corresponding degrees of freedom are (p,N-p-1).

4.3 Hypothesis Tests: Individual Coefficients

When the null hypothesis of the significance of regression is rejected, we can perform hypothesis tests for each coefficient estimated.

> **Note**
>
> `R` provides this information in the `summary()` output of `lm()` with the t-statistic and the corresponding p-value.

4.4 Discrete Variables as Predictors

It is possible that we have a categorical independent variable that we want to consider for the regression model.

4.4.1 Dichotomous Variables

For a *dichotomous* variable, we can introduce a variable such that a "Yes" condition is assigned to `1` and `0` otherwise.

4.4.2 Multi-level

For a categorical predictor variable with k levels, we can introduce a $k-1$ **dichotomous dummy variables** to describe each level of the categorical predictor variable.

> **Note**
>
> For example, blood type (A, B, AB, O) has four levels. This means we need three dummy variables (X_1, X_2, X_3) to account for blood type. We define these dummy variables as:
>
> $$X_1 = \begin{cases} 1, BloodType = A \\ 0, otherwise \end{cases}$$
>
> $$X_2 = \begin{cases} 1, BloodType = B \\ 0, otherwise \end{cases}$$
>
> $$X_3 = \begin{cases} 1, BloodType = AB \\ 0, otherwise \end{cases}$$
>
> For an individual with blood type A, the dummy variable values are $(X_1, X_2, X_3) = (1, 0, 0)$. Similarly, B corresponds to $(X_1, X_2, X_3) = (0, 1, 0)$. For blood type O, $(X_1, X_2, X_3) = (0, 0, 0)$.

4.4.3 `R` Implementation

`lm()` creates dummy variables for character/factor variables. If the variable as a numeric variable, we need to use `as.factor()` for `R` to recognize its categorical nature.

4.5 Adjusted R^2

When we add more predictor variables to a model, the R^2 can only go up or stay the same. It is tempting to add more predictor variables to increase R^2 even if these variables have little to no effect.

> Tip
>
> The adjusted R^2 value, also known as R^2_{adj}, is used to penalize models with more parameters. One form of the adjusted R^2 is given by:
>
> $$R^2_{adj} = 1 - \frac{(N-1)}{(N-p-1)}(1-R^2)$$
>
> where N is the total number of samples. This is also provided in the `summary()` output of `lm()`.

4.6 Model Checking

Recall: The residual is defined as the difference between the observed and predicted responses, specifically $r_i = y_i - \hat{y}_i$. The residuals can be used to test the normality, equal variance, and independence assumptions. What is of interest to us are the **residual plots**.

The following tabs explain the plots shown when plotting an `lm()` object using `plot()`.

```
mod1 <- lm(y~x1+x2,data=df)
plot(mod1)
```

4.6.1 Residual v. Fitted

This plot shows the fitted values of the predictor data to the residuals. If the results can be confined in a horizontal band, there are no obvious model defects.

4.6.2 Q-Q Residuals

The normal probability plot of residuals, also known as the QQ plot, plots the theoretical quantiles of the normal distribution to the residuals. A good plot shows the points coinciding with the dashed line, defined as the expected behavior of the residuals if they follow a normal distribution.

> **!** Important
>
> Small deviations from this line is not a major source of concern. Visual inspection of the QQ plot is more recommended than formal statistical tests for normality, as these residuals are not independent. The non-independence of the residuals breaks the independence assumptions of most normality tests.

4.6.3 Scale-Location

The Scale-Location plot is a plot of the fitted values against the square root of the standardized residuals. Ideally, the red line should be generally horizontal with no upward trend, sharp angles, or slope.

4.6.4 Residuals v. Leverage

The residuals vs. leverage plot provides information on highly influential points in the model. Highly influential data points might lead to a skewed model. If the points are within the dashed lines, then there are no highly influential points in the model.

4.7 Example

Consider the sleep health data set in `SleepHealthData.csv`.

```
sleep <- read.csv("SleepHealthData.csv")
```

4.7.1 Question

Suppose researchers are interested in estimating sleep duration based on heart rate, age, and gender.

- Estimate the best fit line using multiple regression
- Comment on the results of the test of significance of regression. Is there evidence of non-zero coefficients?
- Comment on the results of the hypothesis tests for each coefficient. Which predictors do we have evidence of an association with sleep duration?
- Provide the R^2 and the adjusted R^2 for this model.
- Comment on the plots used for model checking. Are there major model defects?

4.7.2 Answer

```
mod1 <- lm(sleep_duration~gender + age + heart_rate,data=sleep)
summary(mod1)
```

```
Call:
lm(formula = sleep_duration ~ gender + age + heart_rate, data = sleep)

Residuals:
    Min      1Q  Median      3Q     Max
-1.0447 -0.6814  0.1215  0.5053  2.0566

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 12.084352   0.650015  18.591  < 2e-16 ***
genderMale   0.304680   0.083770   3.637 0.000315 ***
age          0.032200   0.004845   6.646 1.08e-10 ***
heart_rate  -0.092133   0.008356 -11.027  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.6465 on 370 degrees of freedom
Multiple R-squared:  0.345, Adjusted R-squared:  0.3397
F-statistic: 64.96 on 3 and 370 DF,  p-value: < 2.2e-16
```

The best fit line to estimate sleep duration is $\hat{y} = 12.08 + 0.305 * X_{male} + 0.03 * Age - 0.09 * HeartRate$, where $X_{male} = 1$ if the individual is male and 0 otherwise.

The overall test of significance of regression yielded an F-statistic of F(3,370) = 64.96 corresponding to a p-value <2.2e-16, indicating there is sufficient evidence that there is at least one non-zero coefficient in the assumed model.

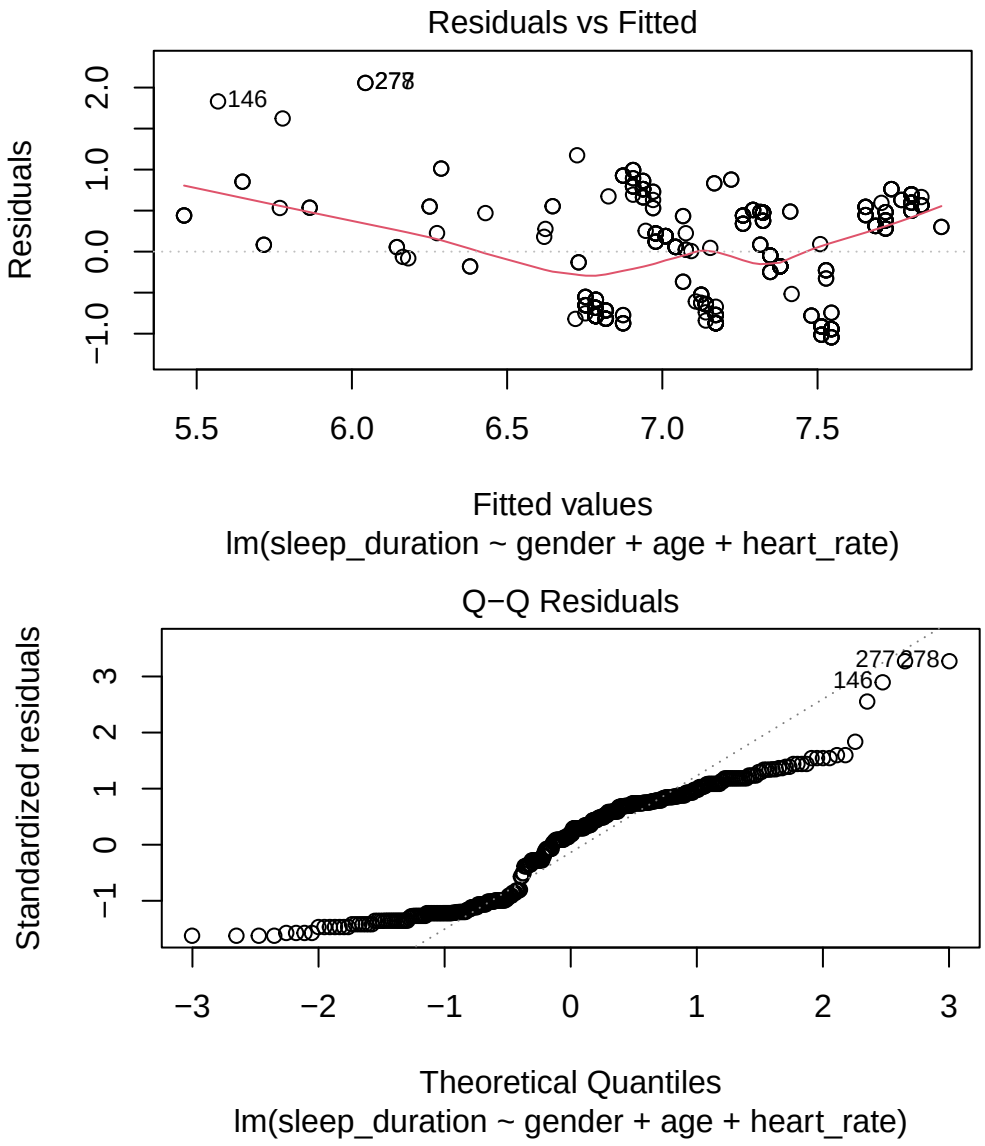
The individual hypothesis tests show that there is evidence of associations between sleep duration and gender ($\beta = 0.30$, p=0.0003), age (β=0.03, p=1.08e-10), and heart rate (β=-0.09,p<2.2e-16).

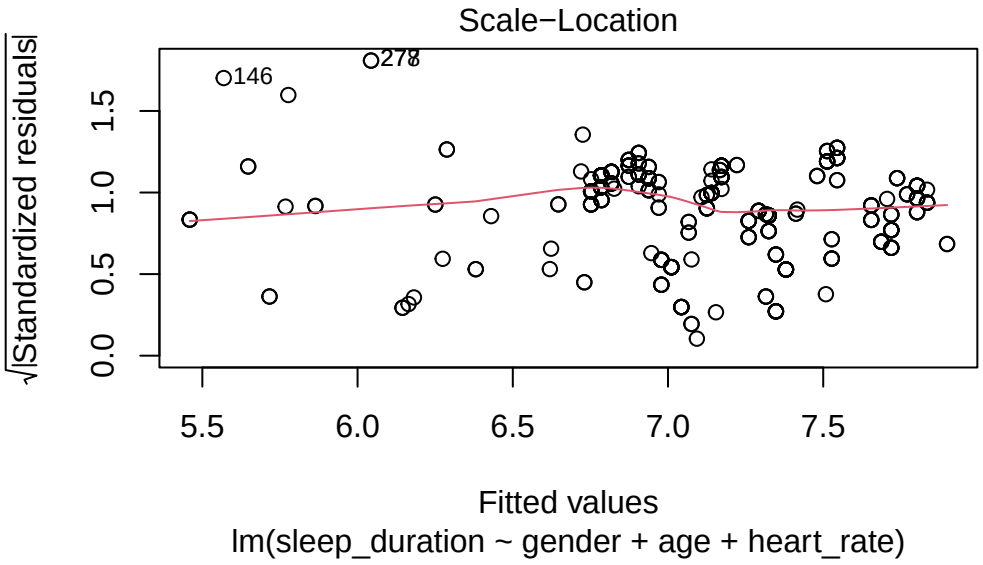
The corresponding R^2 and adjusted R^2 values are 0.345 and 0.3397.

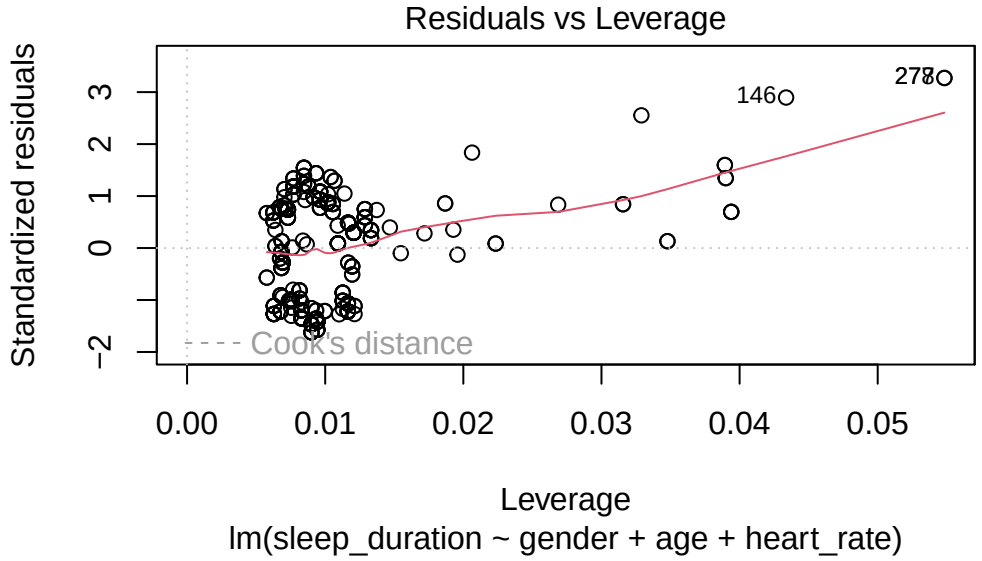
4.7.3 Plots

The plots show major defects on the normality assumption of the residuals, which we must keep in mind when we are interpreting our results. We can also see that the residuals vs. fitted values cannot be enclosed in two horizontal bands, implying evidence of unequal variances.

```
plot(mod1)
```


Residuals vs Fitted
Residuals
146
278
277
2.0
1.0
0.0
−1.0
5.5
6.0
6.5
7.0
7.5
Fitted values
lm(sleep_duration ~ gender + age + heart_rate)
Q−Q Residuals
Standardized residuals
277
278
146
3
2
1
0
−1
−3
−2
−1
0
1
2
3
Theoretical Quantiles
lm(sleep_duration ~ gender + age + heart_rate)


Scale−Location
√|Standardized residuals|
146
278
0.0
0.5
1.0
1.5
5.5
6.0
6.5
7.0
7.5
Fitted values
lm(sleep_duration ~ gender + age + heart_rate)


Residuals vs Leverage
Standardized residuals
146
278
Cook's distance
−2
0
1
2
3
0.00
0.01
0.02
0.03
0.04
0.05
Leverage
lm(sleep_duration ~ gender + age + heart_rate)

4.8 Exercise

Consider the liver steatosis data set `Liver_Steatosis.csv`. This dataset contains information on the 443 of 451 patients who had bariatric surgery at the Cleveland Clinic between 2005 and 2009 and underwent livery biopsy.

```
liver <- read.csv("Liver_Steatosis.csv")
```

4.8.1 Question

Suppose researchers are interested in estimating a person's total cholesterol levels (variable CHOL) based on their BMI (variable BMI), age (variable Age), history of metabolic syndrome (variable MET_Syndrome, equal to 1 means patient has history), and plasma triglycerides (variable TG).

- Estimate the best fit line using multiple regression.
- Comment on the results of the test of significance of regression. Is there evidence of non-zero coefficients?
- Comment on the results of the hypothesis tests for each coefficient. Which predictors do we have evidence of an association with total cholesterol levels?
- Provide the R^2 and the adjusted R^2 for this model.
- Comment on the plots used for model checking. Are there major model defects?

4.8.2 Answer

```
mod1 <- lm(CHOL~BMI+ Age + MET_Syndrome+TG,data=liver)
summary(mod1)
```

```
Call:
lm(formula = CHOL ~ BMI + Age + MET_Syndrome + TG, data = liver)

Residuals:
     Min       1Q   Median       3Q      Max
-142.490  -25.349   -1.504   22.242  134.882

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  152.18618   12.46599  12.208   <2e-16 ***
BMI            0.09060    0.17538   0.517   0.6057
Age            0.19541    0.17158   1.139   0.2554
MET_Syndrome  -9.98380    4.36471  -2.287   0.0227 *
TG             0.15812    0.01478  10.699   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 37.59 on 406 degrees of freedom
  (32 observations deleted due to missingness)
Multiple R-squared:  0.2233,    Adjusted R-squared:  0.2157
F-statistic: 29.19 on 4 and 406 DF,  p-value: < 2.2e-16
```

The best fit line to estimate total cholesterol levels is $\hat{y} = 152.19 + 0.091 * BMI + 0.195 * Age - 9.98 * X_{MetSyndrome} + 0.158 * TG$. where $X_{MetSyndrome} = 1$ when patient has history of metabolic syndrome.

The overall test of significance of regression yielded an F-statistic of F(4,406) = 29.19 corresponding to a p-value <2.2e-16, indicating there is sufficient evidence that there is at least one non-zero coefficient in the assumed model.

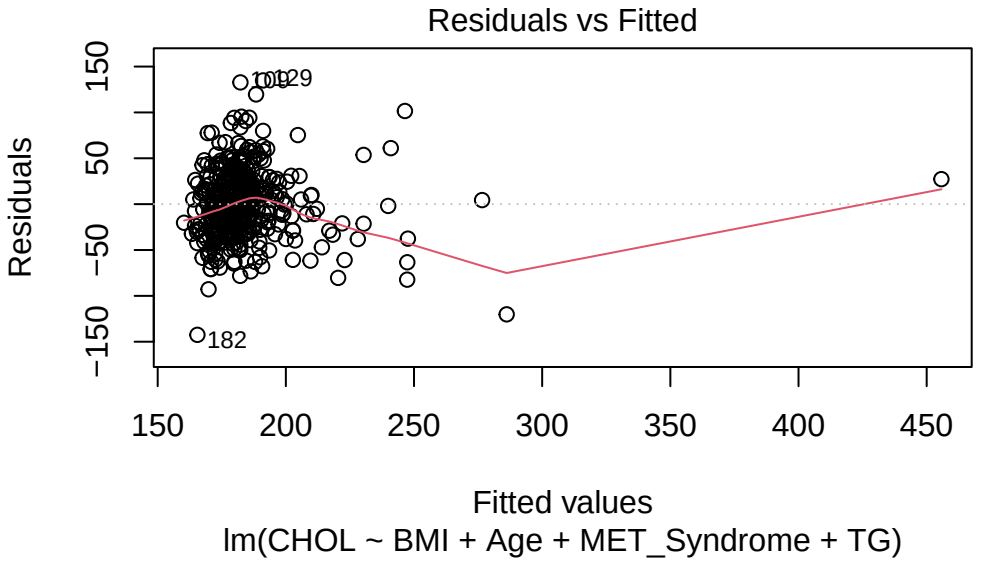
The individual hypothesis tests show that there is evidence of associations between total cholesterol and Metabolic Syndrome history (β = -9.98, p=0.0003) and triglycerides (β=0.158,p<2.2e-16). There is no evidence of association between total cholesterol and BMI (β = 0.09, p=0.61) and Age (β = 0.195, p=0.25).

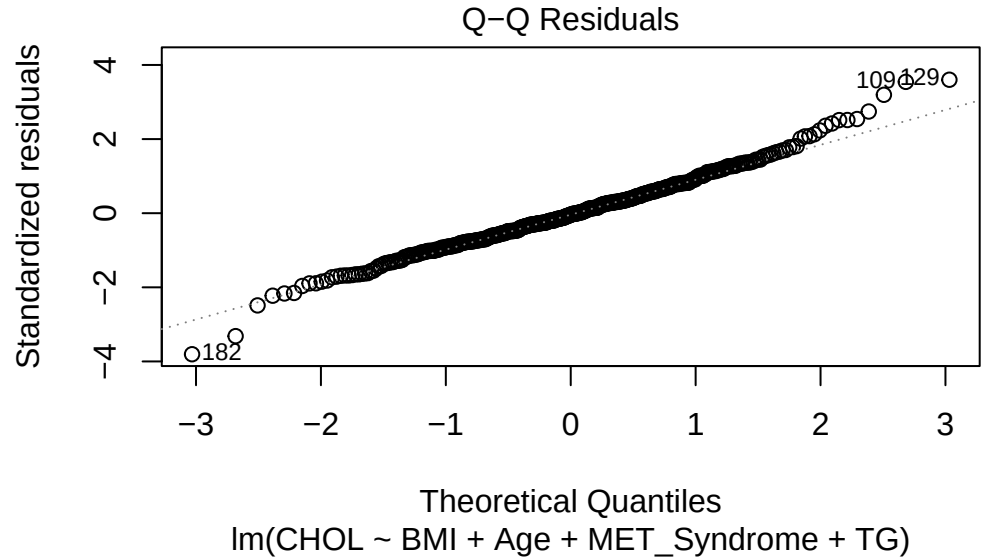
The corresponding R^2 and adjusted R^2 values are 0.2233 and 0.2157.

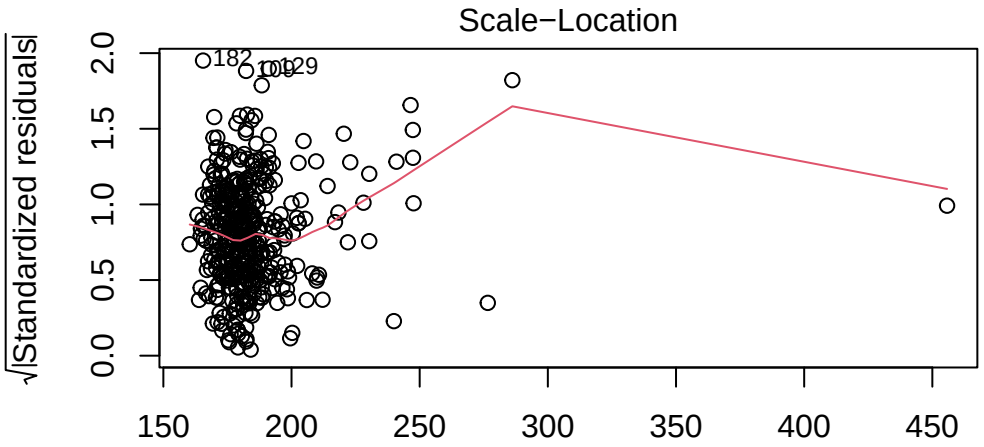
4.8.3 Plots

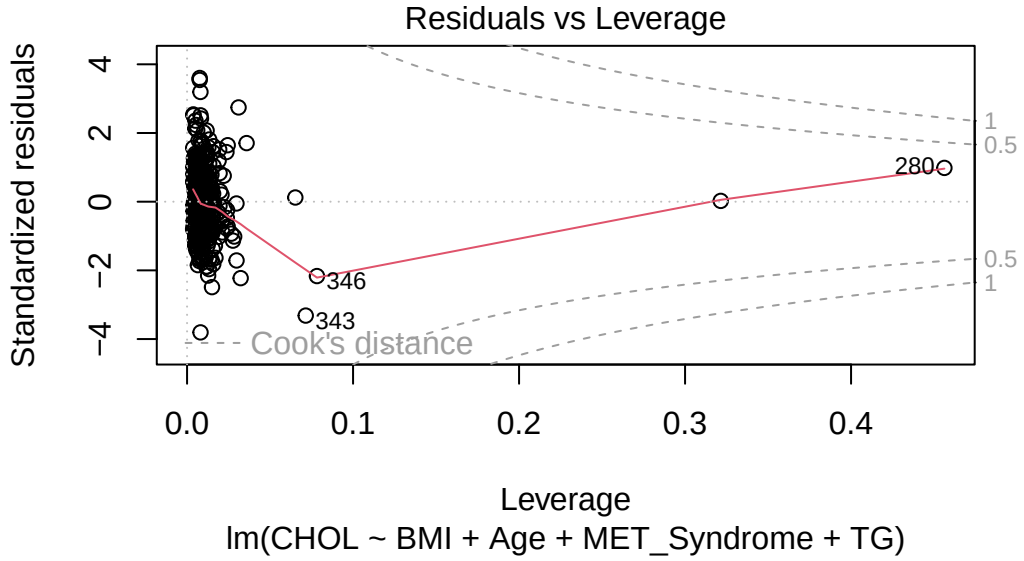
A sharp curve in the residuals vs. fitted plot could be indicative of unequal variances. Normality assumption appears to hold well.

```r
plot(mod1)
```


Residuals vs Fitted
Residuals
150
50
−50
−150
1029
182
150
200
250
300
350
400
450
Fitted values
lm(CHOL ~ BMI + Age + MET_Syndrome + TG)


Q−Q Residuals
Standardized residuals
4
2
0
−2
−4
1094
129
182
−3
−2
−1
0
1
2
3
Theoretical Quantiles
lm(CHOL ~ BMI + Age + MET_Syndrome + TG)


Scale−Location
√|Standardized residuals|
0.0
0.5
1.0
1.5
2.0
150
200
250
300
350
400
450
Fitted values
lm(CHOL ~ BMI + Age + MET_Syndrome + TG)


Residuals vs Leverage
Standardized residuals
4
2
0
−2
−4
0.0
0.1
0.2
0.3
0.4
1
0.5
0.5
1
280
346
343
Cook's distance
Leverage
lm(CHOL ~ BMI + Age + MET_Syndrome + TG)

5 Correlation

5.1 Correlation

In regression analysis, we assumed that the predictor variables were constant and the only random variable was the response or outcome. Often, we encounter scenarios where both response and predictor variables are random. This is when we use a *correlation model*.

> **Note**
>
> In the previous exercise, cholesterol and triglycerides can be treated as random if they are measured at the same time.

> **Important**
>
> Although correlation analysis cannot be carried out meaningfully under the classic regression model, regression analysis can be carried out under the correlation model.

5.2 Correlation Coefficient

The population correlation coefficient ρ measures the strength of linear relationship between the predictor X and Y. On the other hand, The sample correlation coefficient, r, describes the linear relationship between the sample observations on two variables in the same way that ρ describes the relationship in a population.

> **!** Important
>
> ρ and r can take on values between -1 and 1. Negative values of the correlation coefficient indicate a negative/inverse association, while positive values of the correlation coefficient indicate a positive/direct association.
> A value of the correlation coefficient equal to 1 indicates perfect positive linear association, i.e. all data points fit in a line with a positive slope.
> A value of the correlation coefficient equal to -1 indicates perfect negative linear association, i.e. all data points fit in a line with a negative slope.
> A value of the correlation coefficient equal to 0 indicates no association.

5.3 Types of Correlation Coefficients

There are two main types of correlation coefficients: Pearson and Spearman correlation coefficients.

> **Note**
>
> Pearson correlation coefficients measures the strength of **linear relationship**. This is typically used in most studies.
> Spearman correlation coefficients measures the strength of monotonic relationship using non-parametric methods. It is used to measure association that might not be linear, or in the presence of outliers.

We will focus on Pearson correlation coefficients in this chapter.

5.4 A Stern Warning

Always remember: **CORRELATION DOES NOT MEAN CAUSATION**. No matter how high the correlation coefficient values are, this is not evidence that X causes Y or vice versa.

Examples of spurious correlation can be found here

5.5 Hypothesis Test: Correlation Coefficient ρ

We can perform a test on the correlation coefficient ρ to determine if we have evidence of linear association between the two variables. We use the null hypothesis of no association $H_0 : \rho = 0$. The alternative hypotheses can be one-sided ($\rho < 0$ or $\rho > 0$), or two-sided ($\rho \neq 0$).

The test statistic can be calculated from the sample correlation coefficient r.

$$t = r\sqrt{\frac{(n-2)}{(1-r^2)}}$$

This test statistic follows a t-distribution with $n-2$ degrees of freedom, where n is the total sample size.

5.6 R implementation

The function `cor.test(x,y)` can be used to estimate the sample correlation coefficient r, as well as perform the hypothesis test of no association and estimate confidence interval for ρ.

```
cor.test(df$x,df$y,conf.level=0.95)

# OR

cor.test(~x+y, data=df,conf.level=0.95)
```

> **!** Important
>
> Note that the formula is slightly different from `lm()` and `aov()`.

5.7 Example

Consider the sleep health data `SleepHealthData.csv`.

```
sleep <- read.csv("SleepHealthData.csv")
```

5.7.1 Question

Test for an association between stress level and sleep duration using correlation analysis.

5.7.2 Answer

```
cortest <- cor.test(sleep$sleep_duration,sleep$stress_level,conf.level=0.95)

cortest
```

```
    Pearson's product-moment correlation

data:  sleep$sleep_duration and sleep$stress_level
t = -26.739, df = 372, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 -0.8430912 -0.7732074
sample estimates:
      cor
-0.811023
```

```
cor.test(~sleep_duration+stress_level,data=sleep,conf.level=0.95)
```

```
    Pearson's product-moment correlation

data:  sleep_duration and stress_level
t = -26.739, df = 372, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 -0.8430912 -0.7732074
sample estimates:
      cor
-0.811023
```

The estimated correlation coefficient between sleep duration and stress level is -0.811 with a 95% confidence interval of -0.843, -0.773.

The hypothesis test shows that the test statistic is -26.739 with 372 degrees of freedom. This corresponds to a p-value <2.2e-16, indicating we have sufficient evidence of a correlation between sleep duration and stress level.

5.8 Exercise

Consider the liver steatosis data set `Liver_Steatosis.csv`. This dataset contains information on the 443 of 451 patients who had bariatric surgery at the Cleveland Clinic between 2005 and 2009 and underwent livery biopsy.

```
liver <- read.csv("Liver_Steatosis.csv")
```

5.8.1 Question

Test for an association between BMI (variable BMI) and cholesterol (variable CHOL) using correlation analysis.

5.8.2 Answer

```
cortest <- cor.test(liver$BMI,liver$CHOL,conf.level=0.95)

cortest
```

```
    Pearson's product-moment correlation

data:  liver$BMI and liver$CHOL
t = -0.13162, df = 409, p-value = 0.8954
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 -0.10317221  0.09027825
sample estimates:
        cor
-0.00650787
```

The estimated correlation coefficient between BMI and cholesterol is -0.007 with a 95% confidence interval of -0.103, 0.09.

The hypothesis test shows that the test statistic is -0.132 with 409 degrees of freedom. This corresponds to a p-value of 0.8953525, indicating we have insufficient evidence of a correlation between BMI and cholesterol.