Lessons this week will cover Chapter 15 in the textbook.

Lesson 1: Covariance and Correlation

Key Takeaways

By the end of this lesson you should be able to:

- Solve and interpret the covariance between two random variables, X and Y
- Solve and interpret the correlation between two random variables, X and Y.

Introduction to Simple Linear Regression

Regression analysis is used when investigators are interested in exploring a possible relationship between two **quantitative (numeric)** random variables. These variables are typically categorised as being a:

- **Response (Dependent) Variable (Y)**: Outcome of interest
- **Explanatory (Independent) Variable(s)** $(\boldsymbol{X_1}, \boldsymbol{X_2}, \dots)$: Variable(s) to help *predict* the response.

Before moving into Simple Linear regression, we take a step back and review simpler ways of measuring ***linear*** relationships between two quantitative random variables:

- **Covariance**
- **Correlation**
- **Slope**

Covariance

Notation: The Covariance is denoted by Cov(X,Y) = S_{xy}

Purpose: Covariance is more useful from a statisticians perspective because we use S_{xy} to calculate the correlation.

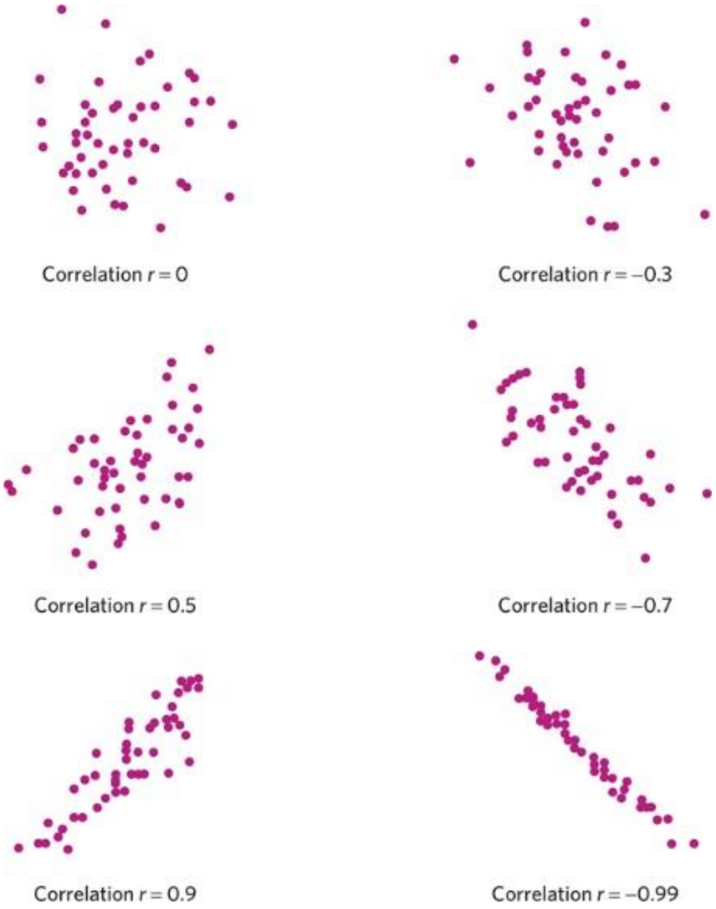
Interpretation: No Magnitude, Just Direction. With covariance the magnitude (size of the value) is NOT important. It can have values from minus infinity to positive infinity and the size of the number is meaningless. Only the direction of the relationship can be determined and is based on the sign:

- A positive covariance suggests a positive association (as X increases, Y increases)
- A negative covariance suggests a negative association (as X increases, Y decreases or vice-versa)

Formula: $\boldsymbol{Cov(X,Y) = s_{xy} = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{n-1} = \frac{(\sum_{i=1}^{n} x_i y_i) - n\bar{x}\bar{y}}{n-1}}$. The terms in the sum are positive if x_i and y_i are larger than their means together or smaller than their means together (as X increases, Y increases). The terms are negative when x_i is smaller than its mean but y_i is larger than its mean, or vice versa (as X increases, Y decreases or vice-versa).

The Pearson Correlation Coefficient

What if we also want to understand the strength of the relationship?
Examples:



Correlation $r = 0$

Correlation $r = -0.3$

Correlation $r = 0.5$

Correlation $r = -0.7$

Correlation $r = 0.9$

Correlation $r = -0.99$

Figure 3: Variety of correlations

What Do You Notice?

In Figure 3 the correlation coefficient appears to change in

1. Magnitude (Size) and,
2. Sign (positive or negative)

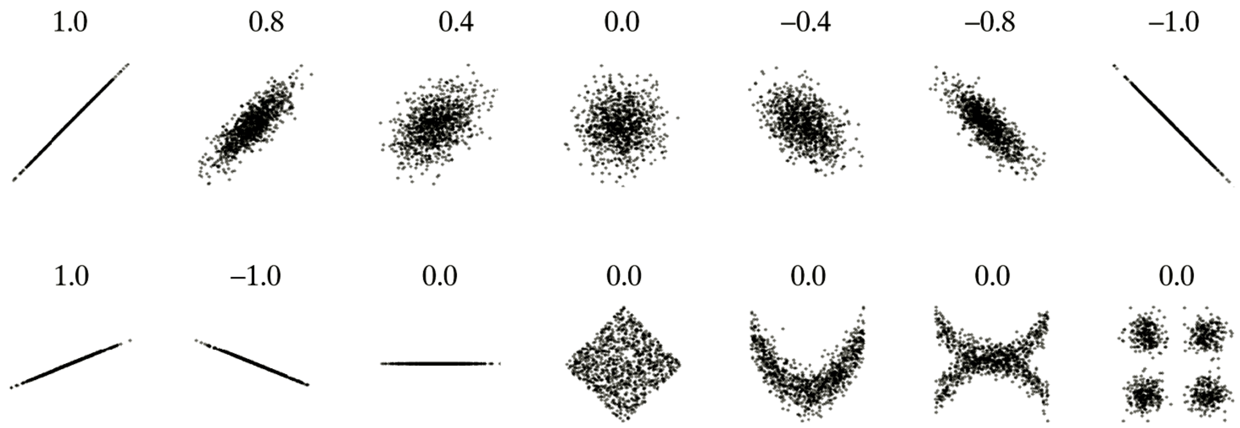
Depending on how X and Y are related.

Interpretation: The correlation coefficient (r) is a number between **-1 and 1**. There are 2 important features about this number.

1. Magnitude – the size of the number
   The closer |r| is to 1, the stronger the linear relationship. Graphically this is indicated by a tightness in the data about a line. If |r|=1 we call it a perfectly linear relationship. The closer |r| is to zero, the less linear the graph.
2. Direction – the sign of the number
   The sign of r indicates the direction.
   - A positive r indicates that the points have a positive slope.

- A negative r indicates that the points have a negative slope.

Note: The Pearson correlation coefficient is checking for a linear relationship only, i.e. can we fit a straight line to the data. It is not accurate when other types relationships, say quadratic, exist. This is why it is always important to plot your data. This way you can see the relationship and make proper sense of the correlation value obtained. Other examples can be found below:



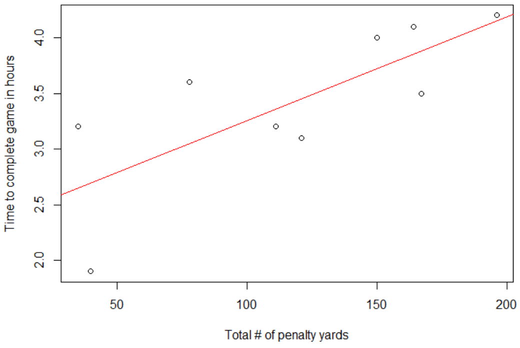
Formula: $\boldsymbol{\rho} = \boldsymbol{r_{xy}} = \frac{\boldsymbol{s_{xy}}}{\boldsymbol{s_x s_y}} = \frac{\sum(\boldsymbol{X_i}-\bar{\boldsymbol{X}})(\boldsymbol{Y_i}-\bar{\boldsymbol{Y}})}{\sqrt{\sum(\boldsymbol{X_i}-\bar{\boldsymbol{X}})^2\sum(\boldsymbol{Y_i}-\bar{\boldsymbol{Y}})^2}}$, where:

- s_{xy} is the covariance between X and Y
- s_x is the standard deviation of X
- s_y is the standard deviation of Y

Example 1

Many factors affect the length of a professional football game, for example the number of running plays versus the number of passing plays. A study was conducted to determine the relationship between the total number of penalty yards (*x*) and the time required to complete a game (*y, in hours).* The table below provides the data.

Total # of penalty yards (X)	196	164	167	35	111	78	150	121	40
Time to complete game (Y)	4.2	4.1	3.5	3.2	3.2	3.6	4.0	3.1	1.9



1. Solve and interpret the covariance
2. Solve and interpret the correlation coefficient

Soln:

1. Using the data we can solve for $\bar{x} = 118$ and $\bar{y} = 3.422$

$$Cov(X,Y) = \frac{(\sum_{i=1}^{9} x_i y_i) - 9\bar{x}\bar{y}}{9-1} = \frac{[(196 \times 4.2) + \cdots + (40 \times 1.9)] - [9 \times 118 \times 3.422]}{9-1}$$
$$= 30.6$$

This value suggests a **positive association** between total # of penalty yards (X) and the time to complete the game (Y).

2. To solve for the correlation coefficient we will also need the standard deviations where:

$$s_X = \sqrt{\frac{\left(\sum_{i=1}^{9} x_i^2\right) - 9\bar{x}^2}{9-1}} = \sqrt{\frac{[196^2 + \cdots + 40^2] - (9 \times 118^2)}{9-1}} = 57.2887$$

Similarly we can solve for $s_Y = 0.7032$, hence $r = \frac{Cov(X,Y)}{s_X s_Y} = \frac{30.6}{57.2887 \times 0.7032} = 0.7596$

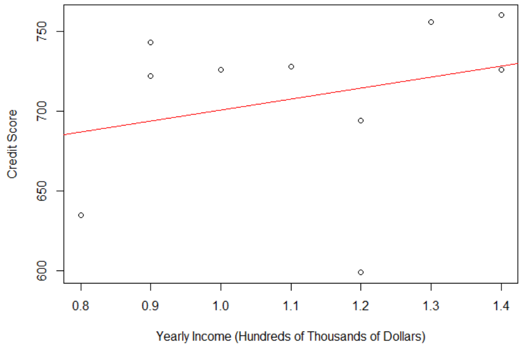
This value suggests that we have a **strong, positive**, linear association between total # of penalty yards (X) and the time to complete the game (Y).

Example 2

A new graduate is thinking ahead and wanting to better understand what goes in to a good credit score. They manage to collect a random sample of annual income (in 100's of thousands) and Credit score values for 10 people. Help this new graduate check for a possible association by

1. Solving for and interpreting the covariance
2. Solving for and interpreting the correlation coefficient

Income (X)	1.3	1.1	0.8	1.2	1.4	0.9	0.9	1.4	1.2	1.0
Credit Score (Y)	756	728	635	599	760	722	743	726	694	726



Soln:

1. Using the data we can solve for $\bar{x} = 1.12$ and $\bar{y} = 708.9$

$$Cov(X,Y) = \frac{\left(\sum_{i=1}^{10} x_i y_i\right) - 10\bar{x}\bar{y}}{10-1} = \frac{[(1.3 \times 756) + \cdots + (1.0 \times 726)] - [10 \times 1.12 \times 708.9]}{10-1}$$
$$= 3.15778$$

This value suggests a **positive** association between annual income (X) and credit score(Y).

2. To solve for the correlation coefficient, we will also need the standard deviations where:

$$s_X = \sqrt{\frac{\left(\sum_{i=1}^{10} x_i^2\right) - 10\bar{x}^2}{10-1}} = \sqrt{\frac{[1.3^2 + \cdots + 1.0^2] - (10 \times 1.12^2)}{10-1}} = 0.214994$$

Similarly, we can solve for $s_Y = 52.5726$, hence $r = \frac{Cov(X,Y)}{s_X s_Y} = \frac{3.15778}{0.214994 \times 52.5726} = 0.27938.$

This value suggests that we have a **fairly weak, positive**, linear association between annual income (X) and credit score(Y).

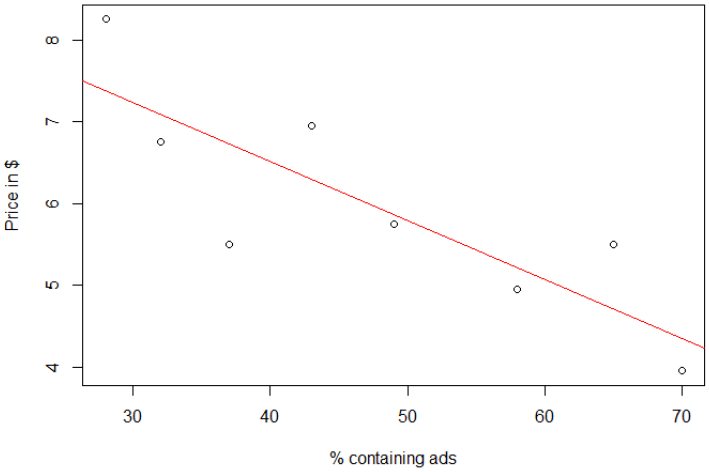
Reminders:

- Correlation measures the strength of the **linear** relationship between 2 **quantitative** variables.
- Like the mean and the standard deviation, the correlation is **not robust** to outliers.

You Try 1

While browsing through the magazine rack at a bookstore, a statistician decides to examine the relationship between the price of a magazine and the percentage of the magazine space that contains advertisements.

% containing ads (X)	37	43	58	49	70	28	65	32
Price in $ (Y)	5.50	6.95	4.95	5.75	3.95	8.25	5.50	6.75



Solve for and interpret the covariance. Solve for and interpret the correlation coefficient. Do these values agree with what's shown on the graph? Feel free to use R to calculate the covariance and correlation coefficient using cov() and cor(). If using another language eg Numpy in Python to calculate these values, please make sure the functions use n-1 and not n in the denominator!

Causation

Often we want to use data collected from an study to asses whether or not there is evidence that X ***affects*** Y, i.e. Do changes in X ***cause*** changes in Y. This is called Causation.

e.g. Smoking Causes Lung Cance; Lack of sleep causes poor grades.

Correlation is NOT Causation

Correlation only tells us that as X increases, Y increases or as X increases, Y decreases, i.e. it defines a trend. It does **NOT** imply that changes in X ***induce*** changes in Y. Correlation ONLY allows us to make conclusions of **Association.**

Practice

Chapter 15: 1, 2, 3, 17, 18, 29

Lesson 2: Slope and Line of Best Fit

Key Takeaways

By the end of this lesson, you should be able to:

- Solve for the line of best fit.
- Make predictions using your line of best fit

Line of best fit - Slope

So far, we have spoken about determining the direction and strength of the linear association between two quantitative variables. What about the actual size of the association, i.e effect of X on Y?

To quantify the actual size of the association we want to solve for the **<u>Slope.</u>**

When a scatter plot shows a linear relationship between two quantitative random variables we can try and summarize the overall pattern by drawing a **<u>line of best fit</u>**.

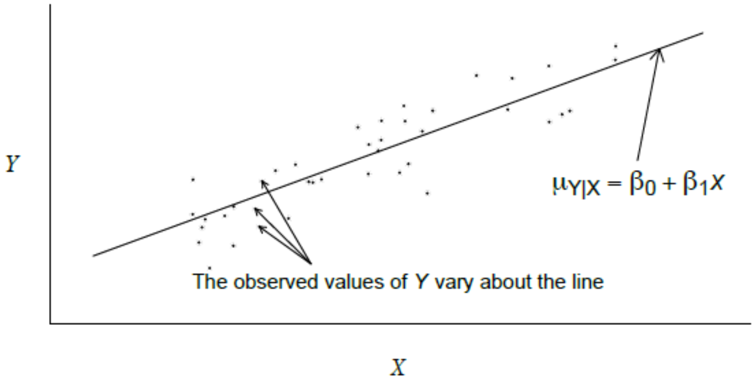
In simple linear regression there is only one explanatory variable and the notation used is as follows:

- Response Variable, Y
- Explanatory Variable, X

We are interested in estimating the line of best fit, where a straight-line relating *y* (response variable) to *x* (explanatory variable) has an equation of the form: $\boldsymbol{\mu_{Y|X} = \beta_0 + \beta_1 x,}$ where:

- $\boldsymbol{\mu_{Y|X} = E(Y|X)}$ and represents the true mean value of Y for a given value of X.
- $\boldsymbol{\mu_{Y|X}}$ is typically referred to as the **deterministic part** of the model and captures the **known variation.**

Visualising the line of best fit:



More conventionally we write the model as: $\boldsymbol{Y = \beta_0 + \beta_1 X + \epsilon}$, which can be interpreted as **Total Variation = known variation + unknown variation**

- Y is the response variable (observed) – This represents the total variation
- X is the explanatory variable (observed)
- β_0 is the Y-intercept (unknown parameter and needs to be estimated)
- β_1 is the slope (unknown parameter and needs to be estimated)

  - $\boldsymbol{\beta_0 + \beta_1 X}$ represents the known variation
- ϵ is a random error term (residual), and represents the unknown variation

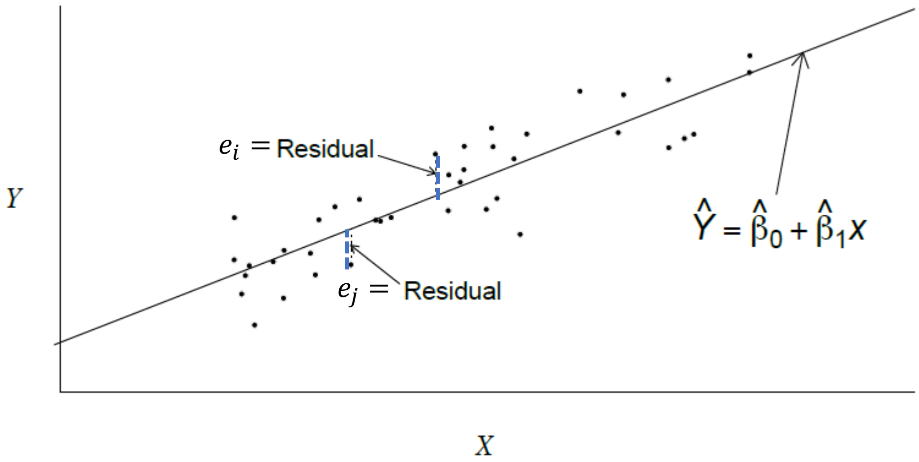
Once the unknown parameters are estimated using the data collected, we can use our **estimated regression line** to make predictions on the response variable: $\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X$, where

- $\hat{Y}$ represents the **predicted value of Y** for a given value of X, strictly speaking is captures $\hat{\mu}_{Y|X}$
- $\hat{\beta}_0$ and $\hat{\beta}_1$ represent the statistics that estimate β_0 and β_1 respectively.
- Interpreting the parameter estimates $\hat{\beta}_0$ and $\hat{\beta}_1$:

$\boldsymbol{\widehat{\beta}_0}$ is the estimated average response when **X=0** (may not be of interest depending on whether X=0 has meaning or not), and $\boldsymbol{\widehat{\beta}_1}$ is the estimated change in the average response for a **one unit increase** in X.

The Least Squares Regression Line

The formula's we use fit what is called the **Least Squares Regression line.** It estimates $\boldsymbol{\beta_0}$ and $\boldsymbol{\beta_1}$ by trying to ***minimize*** the sum of the squares of the vertical distances between each of the **observed Y** and **predicted Y ($\boldsymbol{\widehat{Y}}$)** values. Each one of these distances represent a **residual term: $\boldsymbol{e_i = Y_i - \widehat{Y}_i}$.** Residual = Observed Y- Predicted Y.



This residual term is calculated for every unit in the study and a ***sum of the squared residuals*** *(error sum of squares)* is calculated. Using this and some additional mathematics we find that the parameter estimates to minimise our error sum of squares and hence produce the line of best fit is given by: $\boldsymbol{\widehat{\beta}_0 = \overline{Y} - \widehat{\beta}_1 \overline{X}}$, and $\boldsymbol{\widehat{\beta}_1 = \frac{SP_{XY}}{SS_{XX}} = \frac{Cov(X,Y)}{Var(X)} = r \times \frac{s_y}{s_x}}$.

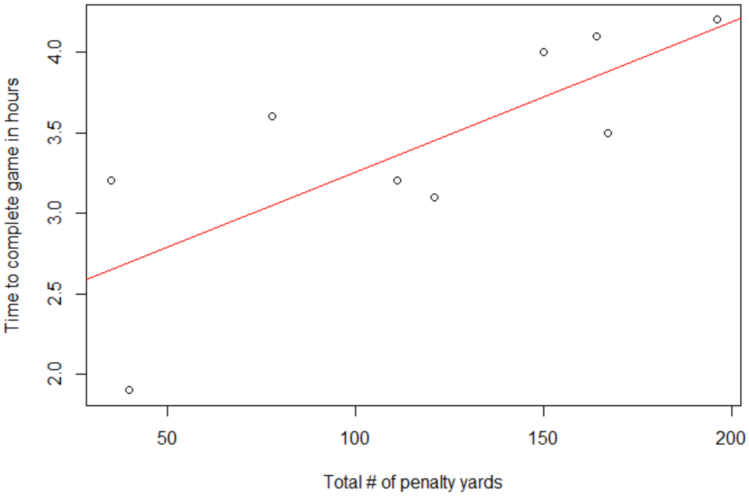
Where:

- $SS_{XX} = \sum_i (X_i - \bar{X})^2 = (n-1)s_x^2$
- $SS_{YY} = \sum_i (Y_i - \bar{Y})^2 = (n-1)s_y^2$
- $SP_{XY} = \sum_i (X_i - \bar{X})(Y_i - \bar{Y}) = (n-1)Cov(X,Y)$

Example 1

Many factors affect the length of a professional football game, for example the number of running plays versus the number of passing plays. A study was conducted to determine the relationship between the total number of penalty yards (*x*) and the time required to complete a game (*y, in hours).* The table below provides the data.

Total # of penalty yards (X)	196	164	167	35	111	78	150	121	40
Time to complete game (Y)	4.2	4.1	3.5	3.2	3.2	3.6	4.0	3.1	1.9



1. Solve for the line of best fit.
2. Interpret your intercept and slope.
3. Use your model to predict the time to complete a game if the number of penalty yards is 180. Do you think this is a good estimate?

Soln:

1. Previously we found that $r = 0.75961, \bar{x} = 118, s_X = 57.28874, \bar{y} = 3.422, s_Y = 0.70317$. Hence we have $\hat{\beta}_1 = r \times \frac{s_Y}{s_X} = 0.75961 \times \left(\frac{0.70317}{57.28874}\right) = 0.0093$ and $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1\bar{x} = 3.422 - (0.0093 \times 118) = 2.322$ Thus, the line of best fit is given by $\hat{y} = 2.322 + 0.0093x$
2. $\hat{\beta}_0 = 2.322$ which implies that on average we will complete the game in 2.322 hours when the total # of penalty yards is 0.

   $\hat{\beta}_1 = 0.0093$ which implies that on average the time to complete the game increases by 0.0093 for every 1 unit increase in the total # of penalty yards.

   R output to check:

```
➢ model<-lm(time~py)
➢ > summary(model)
➢ Call: lm(formula = time ~ py)
➢ Residuals:
➢ Min 1Q Median 3Q Max
➢ -0.79498 -0.35019 0.05054 0.27942 0.55164
➢ Coefficients:
➢             Estimate Std. Error t value Pr(>|t|)
➢ (Intercept) 2.322039  0.391554   5.93   0.000581 ***
➢  py         0.009324  0.003017   3.09   0.017563 *
➢ ---
➢ Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

➢ Residual standard error: 0.4889 on 7 degrees of freedom
➢ Multiple R-squared: 0.577, Adjusted R-squared: 0.5166
➢ F-statistic: 9.549 on 1 and 7 DF, p-value: 0.01756
```

3. Our prediction is $\hat{y} = 2.322 + (0.0093 \times 180) = 3.996$. This estimate is likely not very accurate for two reasons:
    a. The dataset used is quite small leading to a model that is likely not very accurate.
    b. The X-value being used is far larger than the maximum value observed in the dataset, i.e. we are trying to **extrapolate**. This is likely reducing the accuracy of the prediction even more.

Extrapolation

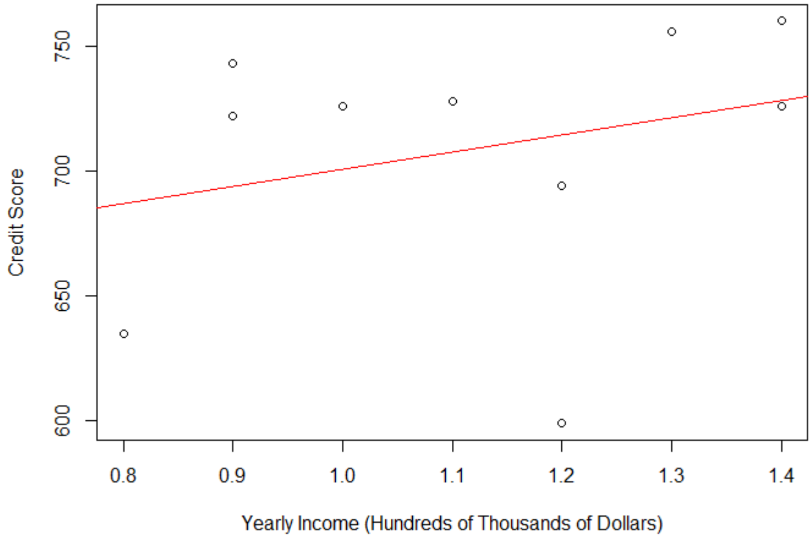
Using the line of best fit for prediction or estimation outside of the range of the observed values of the explanatory variable (X) is known as **extrapolation** and it should be avoided. The line of best fit comes about from the observed values and so does the best job when looking within that range. Once we move outside that range it can result in very misleading estimates.

Example 2

A new graduate is thinking ahead and wanting to better understand what goes in to a good credit score. They manage to collect a random sample of annual income (in 100's of thousands) and Credit score values for 10 people. Help this new graduate check for a possible association by:

1. Solve for the line of best fit.
2. Interpreting your intercept and slope.

Income (X)	1.3	1.1	0.8	1.2	1.4	0.9	0.9	1.4	1.2	1.0
Credit Score (Y)	756	728	635	599	760	722	743	726	694	726



Soln:

1. Previously we found that $Cov(X,Y) = 3.15778, r = 0.27928, \bar{x} = 1.12, s_X = 0.215, \bar{y} = 708.9, s_Y = 52.57$. Hence, we have $\hat{\beta}_1 = \frac{Cov(X,Y)}{Var(X)} = \frac{3.15778}{0.215^2} = 68.32$ and $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} = 708.9 - (68.32 \times 1.12) = 632.38$. Line of best fit is given by $\hat{y} = 632.38 + 68.32x$
2. $\hat{\beta}_0 = 632.38$ which implies that for a person with \$0 income the average credit score is 632.28. $\hat{\beta}_1 = 68.32$ which implies that on average the credit score increases by 68.32 for every 1 unit increase in annual income.

R output to check:

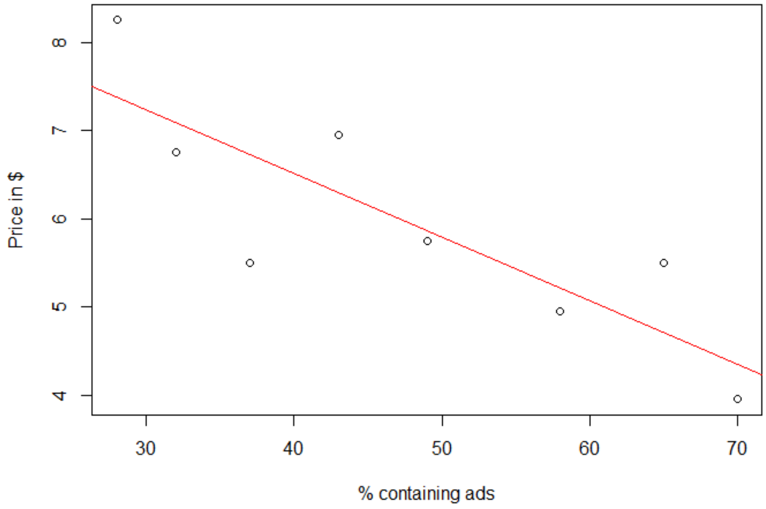
```
model<-lm(creditScore~income) #fitting a simple linear model

  ➢ summary(model) #provide output of model
  ➢ Call: lm(formula = creditScore ~ income)
  ➢ Residuals:
  ➢ Min 1Q Median 3Q Max
  ➢ -115.36 -15.78 22.88 31.01 49.13
  ➢ Coefficients:
  ➢            Estimate Std. Error t value Pr(>|t|)
  ➢ (Intercept) 632.38     94.50    6.692  0.000154 ***
  ➢ income      68.32      83.01    0.823  0.434364
  ➢ ---
  ➢ Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
  ➢  Residual standard error: 53.54 on 8 degrees of freedom
  ➢ Multiple R-squared: 0.07805, Adjusted R-squared: -0.03719
  ➢ F-statistic: 0.6773 on 1 and 8 DF, p-value: 0.4344
```

You Try 1

While browsing through the magazine rack at a bookstore, a statistician decides to examine the relationship between the price of a magazine and the percentage of the magazine space that contains advertisements.

% containing ads (X)	37	43	58	49	70	28	65	32
Price in $ (Y)	5.50	6.95	4.95	5.75	3.95	8.25	5.50	6.75

1. Solve for the line of best fit.
2. Interpret your intercept and slope.



Practice

Chapter 15: 4, 5, 7, 8, 9a)-c)

Lesson 3: Coefficient of Determination, Outliers and Influential points

Key takeaways

By the end of this lesson you should be able to:

- Solve for and interpret the coefficient of determination
- Identify outliers and influential points

Coefficient of Determination

When a linear relationship exists between two variables some of the variation observed in the response variable (Y) is accounted for by the explanatory variable (X) of interest. The rest of the variation is attributed to ***unknown factors*** that were not considered.

Question of interest: "How do we determine just how much of the variation in y is accounted for by X?"

Under simple linear regression, the strength of the relationship between Y and X gives us an idea of how much of the variation in Y is explained by the explanatory variable X. Recall that this strength in relationship is captured by the correlation coefficient, $\boldsymbol{r}$. After some manipulation we come to find that: **$\boldsymbol{r^2}$ captures the fraction of the variation in $\boldsymbol{Y}$ that is explained by the line of best fit. $\boldsymbol{r^2}$** is called the **Coefficient of Determination**.

Example 1

Interpret the $\boldsymbol{r^2}$ values from our previous examples:

- Time to complete football game, we found that $r = 0.76$
- Credit Score example, we found that $r = 0.28$

Soln:

Time to complete football game we have that $r^2 = (0.76)^2 = 0.5776$. In words this is to say that the model explains ~57.76% of the variation in Y. (Note: This is fair value suggesting that on its own the total # of penalty yards does ok at predicting the length of the game but there is room to improve the model.)

Credit Score example we have that $r^2 = (0.28)^2 = 0.0784$. In words this is to say that the model explains ~7.84% of the variation in Y. (Note: This is extremely low again suggesting that income, on its own, is not a good predictor of Credit Score.)

In the R outputs above, this value is the `Multiple R-squared` value. They are slightly different because we rounded our values.

You Try 1

While browsing through the magazine rack at a bookstore, a statistician decides to examine the relationship between the price of a magazine and the percentage of the magazine space that contains advertisements.

% containing ads (X)	37	43	58	49	70	28	65	32
Price in $ (Y)	5.50	6.95	4.95	5.75	3.95	8.25	5.50	6.75

In "You Try 1" you should have solved for $r = -0.84$. Use this value to solve for the coefficient of determination and interpret the value.

Outliers and Influential points

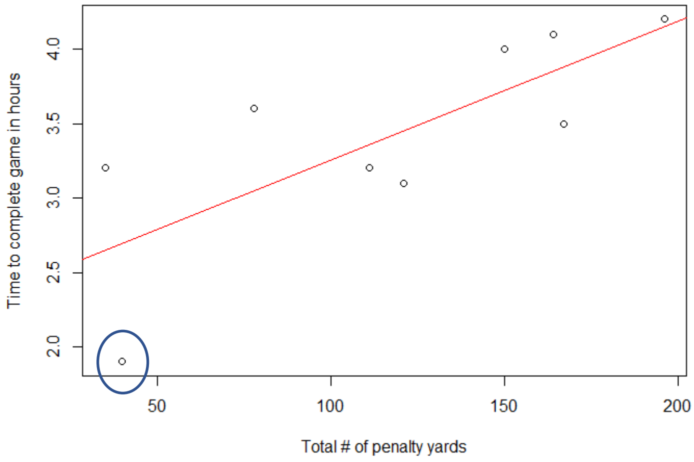
An **influential point** is an observation that has a large influence on the statistical calculations being done. Both the correlation coefficient and parameter estimates in the regression line are typically affected by influential points. Under linear regression such a point is identified as one where if removing it from the data it would cause our line of best fit to ***change markedly***.

Typically, **outliers** in either the X or Y direction are influential points. If such points exist, the investigator should make an effort to see if this due to an error that occurred while capturing the data or if there is some other factor surrounding the unit from which this point was collected e.t.c. Under certain scenarios the investigator may choose to remove such points from the analysis.

Example 2

Many factors affect the length of a professional football game, for example the number of running plays versus the number of passing plays. A study was conducted to determine the relationship between the total number of penalty yards (*x*) and the time required to complete a game (*y, in hours).* The table below provides the data.

Total # of penalty yards (X)	196	164	167	35	111	78	150	121	40
Time to complete game (Y)	4.2	4.1	3.5	3.2	3.2	3.6	4.0	3.1	1.9



In Example 3 we solved for the line of best fit defining it as: $\hat{y} = 2.322 + 0.0093x$

But notice that the one (circled) point seems a little odd compared to the rest. The time to complete the game seems a lot lower for its corresponding X co-ordinate. Investigate whether this point is influential by solving for the line of best fit when that point is removed.

Soln:

This odd point corresponds to the last data point in the table above, i.e co-ordinates (40, 1.9). If we remove it from the dataset and re-calculate our line of best fit we obtain:

Hence we have: $\hat{\beta}_1 = r \times \frac{s_Y}{s_X} = 0.68739 \times \left(\frac{0.43895}{52.660}\right) = 0.00573,$
and $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} = 3.6125 - (0.00573 \times 127.75) = 2.88.$

Then, the line of best fit is given by $\hat{y} = 2.88 + 0.00573x$.

X	Y	
196	4.2	
164	4.1	
167	3.5	
35	3.2	
111	3.2	
78	3.6	
150	4	
121	3.1	
127.75	3.6125	**Mean**
52.65996039	0.438951673	**Std. Dev**
0.687396217		**Correlation**

Notice that in removing this “suspicious” point the correlation coefficient got a little weaker $(0.75961\ vs.\ 0.68729)$. The variability in Y decreased and the slope estimate changed a lot $(0.0093\ to\ 0.00573)$. Hence this point appears to be influential and needs to be further investigated.

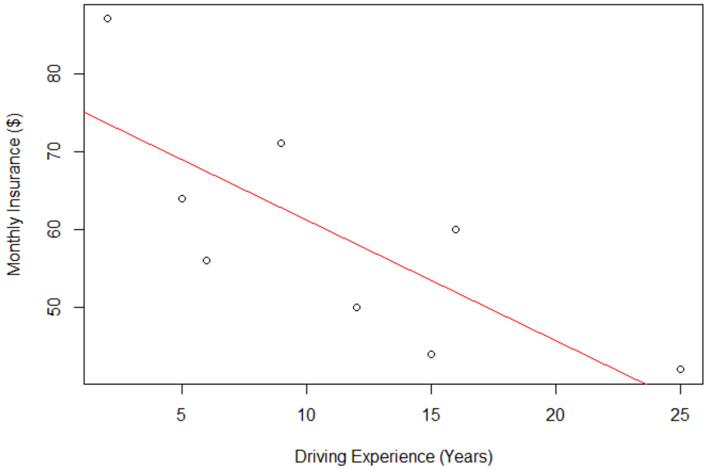
Example 3

A random sample of eight drivers selected from a small town insured with a company and having similar minimum required auto insurance policies was selected. The interest is in investigating the relationship between Driving experience (years) and Monthly auto insurance premium ($).

The data is presented in the table:

	Driving Experience (Years), X	**Monthly Auto Insurance Premium ($), Y**
	5	64
	2	87
	12	50
	9	71
	15	44
	6	56
	25	42
	16	60
Mean	11.25	59.25
Std. Dev	7.40	14.92

1. Use the data to solve for the line of best fit.
2. Use your model to predict the monthly auto insurance premium for a driver with 12 years of driving experience.
3. Solve for the estimated residual for a driver with 12 years of driving experience who pays $50 of premium monthly. What does this value suggest?



Soln:

1. We can use our data to find $r = -0.76793$. This suggests a strong, negative linear association.
   Hence we have $\hat{\beta}_1 = r \times \frac{s_Y}{s_X} = -0.76793 \times \left(\frac{14.92}{59.25}\right) = -1.548$
   And $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} = 59.25 - (-1.548 \times 11.25) = 76.67$
   The line of best fit is given by $\hat{y} = 76.67 - 1.548x$
2. Our predicted response is given by $\hat{y} = 76.67 - (1.548 \times 12) = \$58.09$
3. Our estimated residual is given by, $e_i = y_i - \hat{y}_i = 50 - 58.094 = -8.094$. Since the residual is negative it suggests that the line overfit the true premium the driver is paying.

Statistical Inference in Simple Linear Regression

To make valid statistical inference procedures in regression we must make a few assumptions:

To start it is assumed that observations are ***independent*** of each other, i.e., when collecting information from our units and/or objects it is assumed that each response is unrelated (not dependent) on any other unit or objects response. Mathematically, this assumption is stating that the Y_is are independent.

The residuals are assumed to be random variables that:

- ***Have a mean of 0***
- ***Have a constant variance (σ^2)***
- ***Are independent of each other***
- ***Are normally distributed***

Summarising these: $\boldsymbol{e_i \sim N(0, \sigma^2)}$

Before concluding that a model is adequate these assumptions must be tested for and shown to be satisfied. Various plots as well as statistics can be looked at to determine whether these assumptions have been satisfied or violated. We will consider these plots once we have a firmer understanding of random variables and the Normal distribution.

Practice

Chapter 15: 24 (a), d)-e)), 25(a), b), d)), 26a), 27, 30, 31(a), b)), 37, 40, 41b)

R component

Will be covered in Tutorial. We will cover simple linear regression – the lm() function and how to interpret output. How to interpret R output for linear regression WILL be tested.

Fun fact: Linear regression is one of the coolest parts of statistics! In fact, even a lot of what Machine Learning practitioners do ends up being a regression model of some kind. I currently work at the SSO at UW, and quite a lot of my analysis is done with linear regression, so even something covered in an introductory course can have loads of real life applications and be useful!