Chapter 1

Two-variable data analysis

In chapter 1, we leaned how to describe the distribution of one random variable both numerically and graphically. Most of the the real life problems are about more than one variables and their interactions. In this chapter we will learn how to model and describe the relationships between two variables.

1.1 The relationship between two categorical variables

By reading table **??**, we may suspect that the is a relation between categorical variable *BASKETALL* and *GENDER*. Maybe boys are more likely to play basketball than girls. How can we describe the relation between *BASKETALL* and *GENDER*?

- **Two-way table**

  Table 1.1 is a two-way table describing the relationship between *BASKETALL* and *GENDER*.

		BASKETBALL		Total
		Yes	No	
GENDER	Male	21	6	27
	Female	5	20	25
Total		26	26	52

Table 1.1: Two-way table of Gender × Basketball

- **Conditional distribution**

  The **conditional distribution** of a categorical variable is defined as the distribution of this variable while the value of the other variable is fixed.

BASKETBALL	Percentage of Male	Percentage of Female
Yes	$\frac{21}{26} \approx 81\%$	$\frac{5}{26} \approx 19\%$
No	$\frac{6}{26} \approx 23\%$	$\frac{20}{26} \approx 77\%$

Table 1.2: Conditional distribution

  Table 1.2 gives the conditional distribution of *GENDER* conditioned on different different values of *BASKETBALL*. For example, the conditional distribution of *GENDER* among those who play basketball is that: about 81% of them are boys, 19% are girls.

  If a student plays basketball, this students is more likely to be a boy than a girl. Clearly, there is some association between *GENDER* and *BASKETBALL*.

- **Association**

  If the conditional distributions of a variable are different while conditioned on different values of the other variable, we say there is an **association** between those variables. Otherwise, they are **independent(with no association)**.

  The conditional distributions of *GENDER* are different conditioned on different values(Yes, No) of *BASKETBALL*. Therefore, there is an association between *GENDER* and *BASKETBALL*.

- **Marginal distribution**

  If we consider the distribution of *GENDER* regardless of the *BASKETBALL*, we just look at the data at the right margin of tabel 1.1.

  $$\text{Percentage of girls: } \frac{25}{52} = 48\%, \qquad \text{Percentage of boys: } \frac{27}{52} = 52\%$$

  Similarly, we can find out the distribution of *BASKETBALL* regardless of the *GENDER* by looking at the data at the bottom margin of tabel 1.1.

  All the distributions are **marginal distribuions**, for we only consider the data at the margin of the two-way table.

Is it true that if two variables have no association, the marginal distributions and the conditional distributions are the same?

- **Side-by-side bar graph** is a graph to show the relation between two categorical variables.

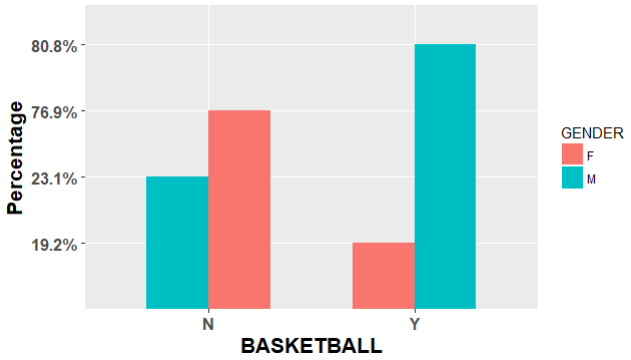


Figure 1.1: Side-by-side bar graph of *GENDER* and *BASKETBALL*

- **Stacked bar graph** can play the same role as side-by-side bar graph.

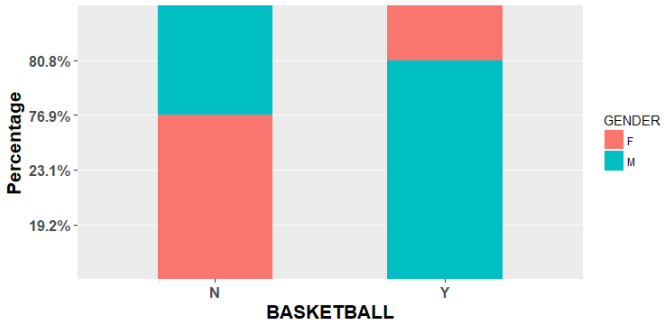


Figure 1.2: Stacked bar graph

A Titanic disaster

In 1912 the luxury liner *Titanic*, on its first voyage across the Atlantic, struck an iceberg and sank. Some passengers got off the ship in lifeboats, but many died. The two-way table below gives information about adult passengers who lived and who died, by class of travel.

	Class of Travel		
Survival status	**First class**	**Second class**	**Third class**
Lived	197	94	151
Died	122	167	476

Here's another table that displays data on survival status by gender and class of travel.

	Class of Travel					
	First class		**Second class**		**Third class**	
Survival status	**Female**	**Male**	**Female**	**Male**	**Female**	**Male**
Lived	140	57	80	14	76	75
Died	4	118	13	154	89	387

The movie *Titanic*, starring Leonardo DiCaprio and Kate Winslet, suggested the following:

- First-class passengers received special treatment in boarding the lifeboats, while some other passengers were prevented from doing so(especially third-class passengers).
- Women and children boarded the lifeboats first, followed by the men.

1 What do the data tell us about these two suggestions? Give appropriate graphical and numerical evidence to support your answer.

2 How does gender affect the relationship between class of travel and survival status? Explain.

1.2 The relationship between two quantitative variables

Table 1.3: Small data set

Name	Calculus	Physics
James	75	74
Sam	87	83
Crystal	88	92
Evelyne	84	82
Phoebe	89	86
Vince	76	85
Mike	71	76
Lucy	81	89
Kitty	86	91
Owen	88	82
Angela	96	93
Christina	87	81
Jamie	84	80
Meggie	80	82
Kevin	97	99
Tom	90	86
John	81	88
Jason	72	80

- **Scatter plot**

  For each student in table 1.3, there is a two dimensional coordinates (*Calculus*,*Physics*). For example, the coordinates of 'James' is $(75, 74)$. If we draw all those coordinates out in a coordinates system, the graph is called the **scatter plot**, as shown in figure 1.3.

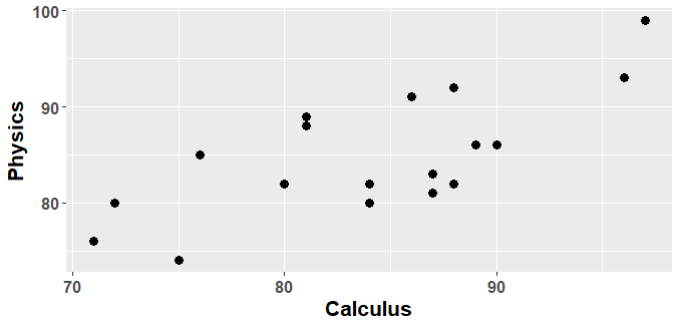


Figure 1.3: Scatter plot of *FINAL* **vs** *MID*

– **Explanatory variable, Response variable**

If we want to use *MID* to explain *FINAL*, then *MID* is called **explanatory variable** and *FINAL* is called **response variable**. The explanatory variable goes to the x-axis and the response variable goes to the y-axis in the scatter plot.

– **Direction**

The direction of a scatter plot gives a general trend, and can be described by terms **positive association** or **negative association**. If one variable increases while the other increases, those two variables have a positive association. If one variable increases while the other decreases, they have a negative association. *FINAL* and *MID* have a positive association according to figure 1.3.

– **Form**

We describe the form of a scatter plot by terms **linear** or **curved**. If the points in the scatter plot forms a linear pattern, the scatter plot has a linear form, otherwise a curved form. The scatter plot in figure 1.3 has a linear form.

– **Strength**

Strength describes how strong is the form. Terms used to describe the strength are: **strong, moderate, weak**. The scatter plot in figure 1.3 has a strong linear relation.

– **Outliers**

Outliers are points don't follow the pattern of the majority either in the **x** direction or in the **y** direction. We modify figure 1.3 a little bit as shown by figure 1.4. Point **A** and **B** are outliers in the **x** direction and the **y** direction respectively.

Describe the scatter plot in figure 1.3.

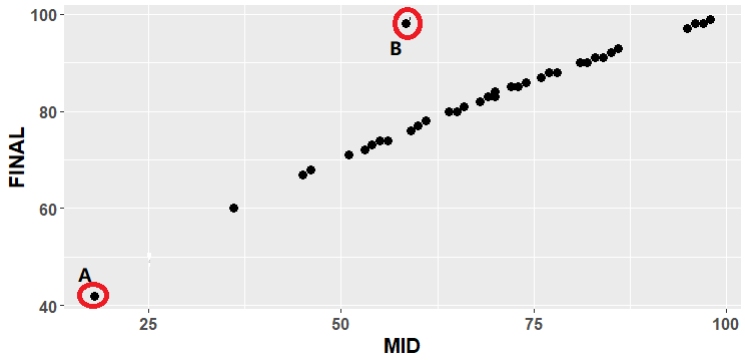


Figure 1.4: Outliers of a scatter plot

- **Correlation r**

  Correlation r measures the strength and the direction of the linear relationship. Suppose $\{(x_1, y_1), (x_2, y_2), \cdots, (x_n, y_n\}$ are the data points, then

$$\begin{aligned} r &= \frac{1}{n-1}[(\frac{x_1-\bar{x}}{s_x}\frac{y_1-\bar{y}}{s_y}) + (\frac{x_2-\bar{x}}{s_x}\frac{y_2-\bar{y}}{s_y}) + \cdots + (\frac{x_n-\bar{x}}{s_x}\frac{y_n-\bar{y}}{s_y})] \\ &= \frac{1}{n-1}\sum_{i=1}^{n}(\frac{x_i-\bar{x}}{s_x})(\frac{y_i-\bar{y}}{s_y}) \\ &= \frac{1}{n-1}\sum_{i=1}^{n} z_{x_i} z_{y_i} \end{aligned}$$

  - **$\mathbf{r} \in [-1, 1]$ $\mathbf{r > 0}$ means there is a positive association.** $r < 0$ **means there is a negative association.**

  - **Correlation r is the measurement of the strength of the linear relationship.**

    If there is a linear relationship, $|r|$ is close to 1, means the linear relationship is strong. " $|r|$ is close to 1" itself can not guarantee there is a linear relation. There may be a curved relation with $|r|$ close to 1.

    $r = 0$ means there is no linear relation, but there may be a curved relationship. If $r = \pm 1$, the scatter plot is strictly linear.

    Correlation r only works on condition that there is a linear relationship. How can we tell there is a linear relationship? We will learn latter.

  - **Correlation r doesn't imply causation**

  - **Correlation r only describes the relationship between two quantitative variables and has no unit.**

Which of the following statements about correlation r is true?

(1) Perfect correlation, that is, when the points lie exactly on a straight line, results in $r = 0$.

(2) If x increase, y increases, $r < 0$.

(3) Correlation is not affected by which variable is called x and which is called y.

(4) Correlation r is not affected by extreme values.

(5) If the unit of the explanatory variable changes, r changes.

(6) r has the same unit as the response variable.

1.3 Least-squares regression

Least-squares regression is a simple model of the relationship between two quantitative variables. It is simple, but it embodies the basic idea of modelling, which plays an important role in machine learning as well.

- **Regression lines**

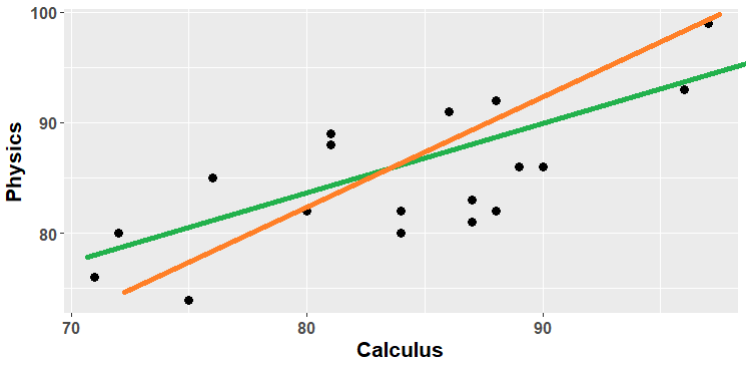


Figure 1.5: Linear regression lines

A regression line is a line drawn to model the scatter plot. There may be more than one regression lines as shown in figure 1.5. We need to set up a criteria and find the best one according to this criteria.

- **Criteria**
  - **Residuals**

Suppose the equation for a regression line is

$$\hat{\mathbf{y}} = \mathbf{a} + \mathbf{bx}.$$

$\hat{\mathbf{y}}$ is the predicted value of *Physics* and $\mathbf{x}$ is the value of *Calculus*. For each value of *Calculus*, we can find out a predicted value of *Physics* by the equation of the regression line. Our prediction $\hat{\mathbf{y}}$ may be different from the observation $\mathbf{y}$. The difference is called **residual**.

$$\textbf{residual} = \mathbf{y} - \hat{\mathbf{y}}$$

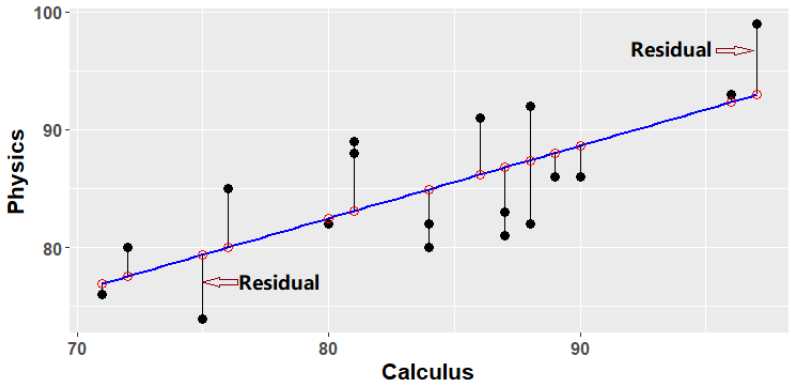


Figure 1.6: Visualizing residuals

A visualization of the residuals is shown in figure 1.6, with black dots observations and red circles predicted value, and the vertical lines connecting the red circles and the black dots represents residuals.

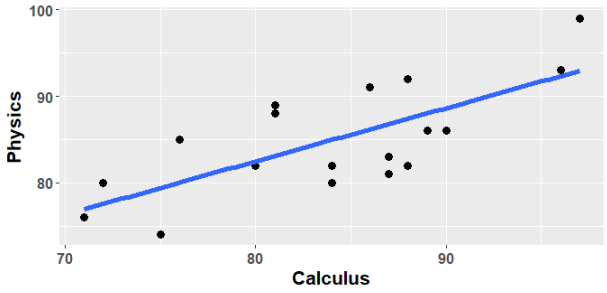
– **The criteria**

We want to find a line which minimizes the sum of the squares of the residuals:

$$\Sigma(\mathbf{y} - \hat{\mathbf{y}}^2).$$

The line minimizes $\Sigma(\mathbf{y} - \hat{\mathbf{y}}^2)$ is called **the least-squares regression line**

- **Least-squares regression line**




```
##                          Summary
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)      33.2662    12.4828   2.665 0.016944 *
##  Calculus         0.6152     0.1481   4.155 0.000746 ***
## ---
##
## Residual standard error: 4.491 on 16 degrees of freedom
## Multiple R-squared:  0.519,Adjusted R-squared:  0.4889
```

Figure 1.7: Least-squares regression line

Figure 1.7 gives a graph and a summary of the least-squares regression line. Lets read the summary.

– **The equation $\hat{\mathbf{y}} = \mathbf{a} + \mathbf{bx}$**

Find the predicted value of *Physics* of James and its residual. Interpret the residual.

In the summary, **a** is the estimated coefficient of the `Intercept`: `33.2622`, and **b** is the estimated coefficient of the `Calculus`: `0.6152`. Thus the equation of the least squares regression line is

$$\widehat{\textbf{Physics}} = 33.262 + 0.6152 \times \textbf{Calculus}$$

The **intercept a** is interpreted as: when the scores of *Calculus* is 0, the predicted value of *Physics* is 33.3 points. The **slope b** is interpreted as: when the scores of the *Calculus* increases by 1 point, the scores of the *Physics* will increase by 0.62 points.

Some formulas:

$$\bar{y} = a + b\bar{x}, \quad b = r\frac{s_y}{s_x}, \quad \Sigma\textbf{residuals} = \Sigma(y - \hat{y}) = 0.$$

– **Standard deviation of the residuals: S**

$$\begin{aligned} S &= \Sigma(\textbf{residuals} - \textbf{mean of residuals})^2/(n-2) \\ &= \Sigma(y - \hat{y})^2/(n-2) \end{aligned}$$

S gives the typical error of the predictions.

– **Coefficient of determination: $\mathbf{r^2}$**

If we know nothing about the relationship between *Calculus* and *Physics*. For a given value of *Calculus*, a reasonable prediction of the value of *Physics* is its mean, since it is the center of *Physics*. In this case, the variation of the predictions from the observations is given by

$$\frac{\Sigma(y - \bar{y})^2}{n}.$$

If the relationship between *Calculus* and *Physics* is given by the least-squares regression line. For a given value **x** of *Calculus*, the predicted value $\hat{y}$ will be given by $\hat{y} = a + bx$. In this case, the variation of the predictions from the observations is given by

$$\frac{\Sigma(y - \hat{y})^2}{n} = \frac{\Sigma\text{residual}^2}{n}.$$

Interpret the standard deviation of the residuals **s** and the coefficient of determination $\mathbf{r^2}$ in figure 1.7.

The percentage of the variation of the response variable explained by the least-squares regression line is

$$1 - \frac{\Sigma\text{residual}^2}{n} \Big/ \frac{\Sigma(y - \bar{y})^2}{n} = 1 - \frac{\Sigma\text{residual}^2}{\Sigma(y - \bar{y})^2} = \mathbf{r^2}$$

This value is **the coefficient of determination**

$\mathbf{r^2}$ is interpreted as the percentage of the the variation of the response variable explained by the least-squares regression line relating the response variable to the explanatory variable.

– **Residual plot**

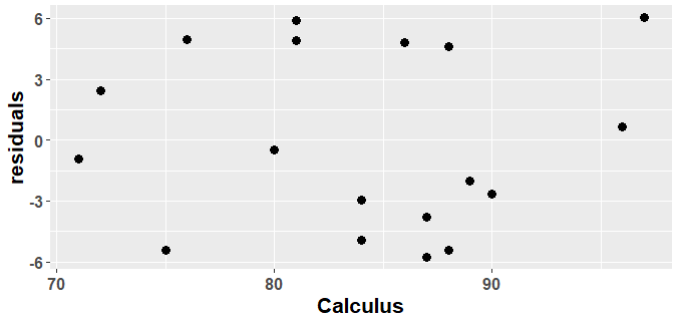


Figure 1.8: Residual plot

The residual plot in figure 1.8 is a scatter plot of residuals against the explanatory variable *Calculus*. Sometimes a residual plot may be generated by plotting the residuals against the response variable.

The residual plot gives us information about whether there is any "left-over" pattern after the response variable is explained by the model. If there is no left-over pattern in the residual plot and the residuals are randomly scattered around 0, our model is appropriate. In figure 1.8, there is no left-over pattern and the residuals is randomly scattered around 0. Thus the linear model is appropriate to describe the relationship between *Physics* and *Calculus*.

Residual plot gives us information about whether a model is appropriate.

– **Extrapolation and influential points**

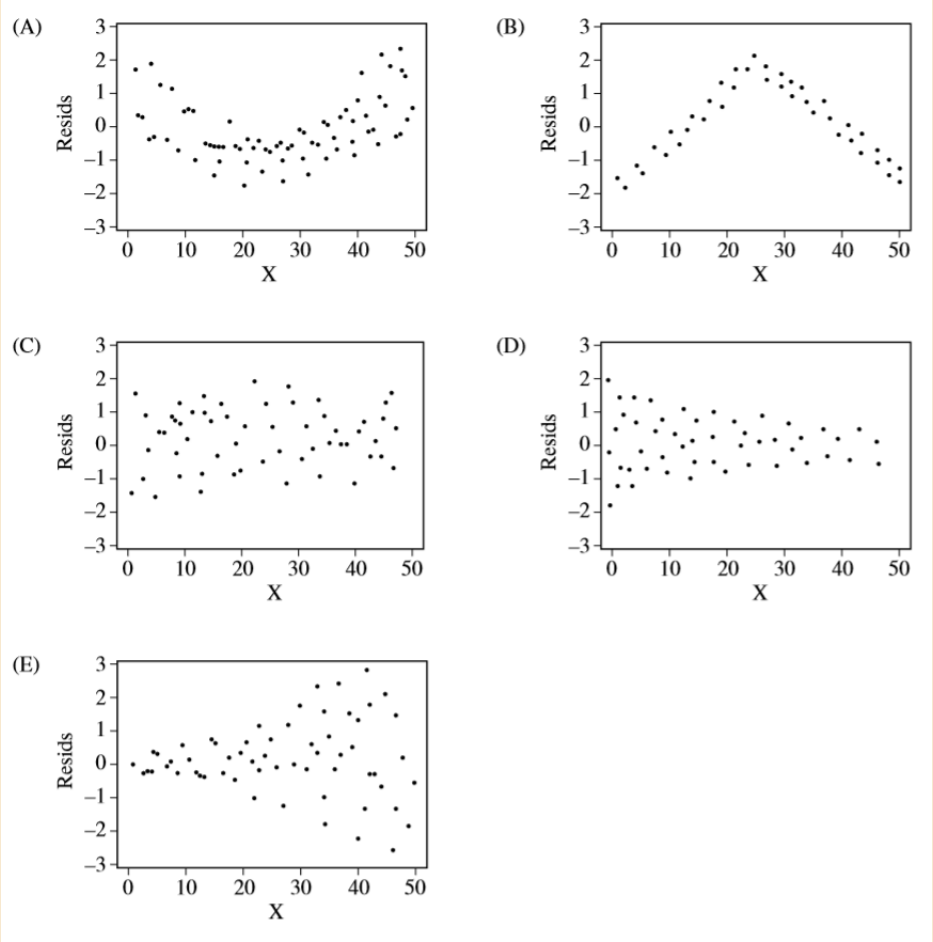
Extrapolation is the use of the regression line to predict the response values based on the explanatory values far out of the range from which we obtain the regression line. For the linear regression in figure 1.7, if a give *Calculus* score is 0, the predicted value for *Physics* is 33.3.

We have to watch out for those extrapolations. Most of the time they are not accurate.

An **influential point** is the point that can markedly influence the result of the least-squares regression line if it is removed.

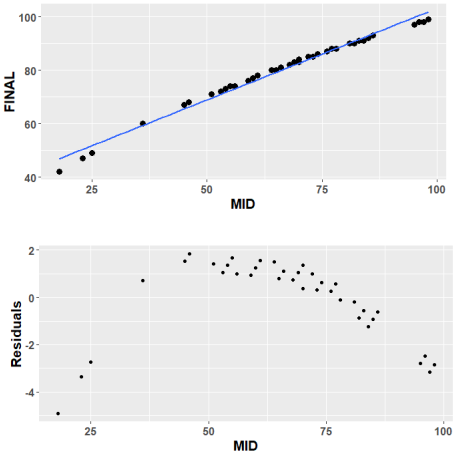
Outliers in the x direction are more influential.

Exercise:
The residual plots from five different least squares regression lines are shown below. Which of the plots provides the strongest evidence that its regression line is an appropriate model for the data and is consistent with the assumptions required for inference for regression.



- **Something more about least-squares regression**

A linear model may be appropriate, but weak, with a low correlation. And, alternatively, a linear model may not be the best model(as evidenced by the residual plot), but it still might be a very good model with high r^2.




```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  34.56800    0.83986   41.16   <2e-16 ***
ExampleData$MID  0.68655    0.01211   56.72   <2e-16 ***
---

Residual standard error: 1.622 on 50 degrees of freedom
Multiple R-squared:  0.9847,    Adjusted R-squared:  0.9844
```

Figure 1.9: Linear regression of *FINAL* to *MID* of data in table **??**.

Figure 1.9 gives the output of **R packages** for the linear regression of *FINAL* to *MID* of data in table **??**. Clearly the residual plot is not "randomly scattered around 0" and there is a left-over pattern. Linear model is not appropriate. However, this linear model is very good with $r^2 = 0.985$, which means **98.5% of the variation of the the *FINAL* is explained by our model relating *FINAL* to *MID*.**

In figure 1.9, **even though a lot of the variance of the response variable is explained by the least-squares regression line, there is a left-over pattern in the residual plot, and the response variable can be further explained by some other models**. While for the model in figure 1.7, **there is no left-over pattern in the residual plot, and the response variable can not be further explained.**

It is important to read the output of different of different softwares in AP exam.

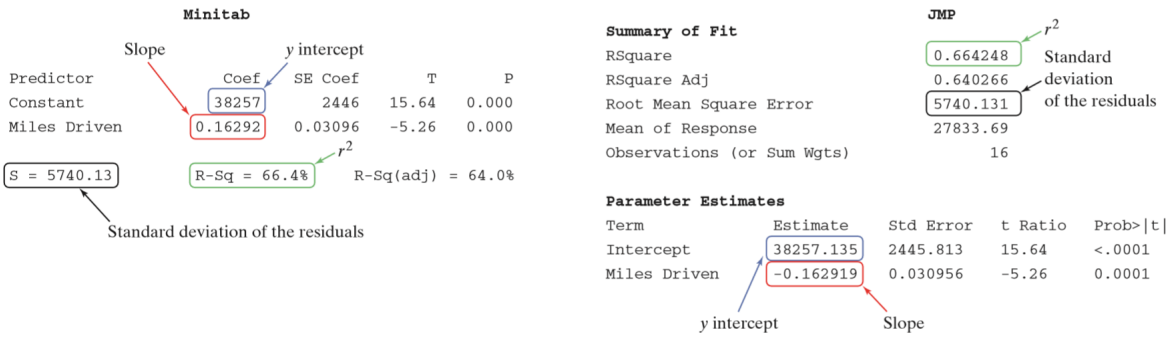
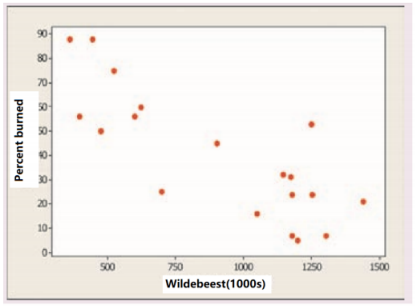
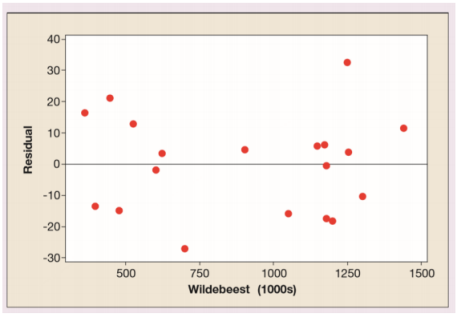


Figure 1.10: Output formats

Serengeti National Park

Long-term records from the Serengeti National Park in Tanzania show interesting ecological relationships. When wildebeest are more abundant, they graze the grass more heavily, so there are fewer fires and more trees grow. Lions feed more successfully when there are more trees, so the lion population increases. Researchers collected data on one part of this cycle, wildebeest abundance (in thousands of animals) and the percent of the grass area burned in the same year. The results of a least-squares regression on the data are shown here.





Predictor	Coef	SE Coef	T	P
Constant	92.29	10.06	9.17	0.000
Wildebeest (1000s)	−0.05762	0.01035	−5.56	0.000

S = 15.9880 R-Sq = 64.6% R-Sq(adj) = 62.5%

(a) Give the equation of the least-squares regression line. Be sure to define any variables you use.

(b) Explain what the slope of the regression line means in this setting.

(c) Find the correlation. Interpret this value in context.

(d) Is a linear model appropriate for describing the relationship between wildebeest abundance and percent of grass area burned? Support your answer with appropriate evidence.

1.4 Transform to achieve linearity

In many situations, the relationship between two quantitative variables can not be described by linear models, but can be described by linear models after proper transformation. Here we will learn how to transform the **exponential models** and the **power models** into linear models.

- **Exponential model**

  The exponential model is given by

  $$\mathbf{y = ab^x}, \quad \mathbf{a} \text{ and } \mathbf{b} \text{ are constant.}$$

  Take natural log of both sides,

  $$\ln \mathbf{y} = \ln(\mathbf{ab^x}) = \ln \mathbf{a} + (\ln \mathbf{b})\ \mathbf{x}.$$

  The relationship between $\ln y$ and x can be described by a linear model, with intercept $\ln a$ and slope $\ln b$.

- **Power model**

  The power model is given by

  $$\mathbf{y = ax^b}, \quad \mathbf{a} \text{ and } \mathbf{b} \text{ are constant.}$$

  Take natural log of both sides,

  $$\ln \mathbf{y} = \ln(\mathbf{ax^b}) = \ln \mathbf{a} + \mathbf{b}\ \ln \mathbf{x}.$$

  The relationship between $\ln y$ and $\ln x$ can be described by a linear model, with intercept $\ln \mathbf{a}$ and slope $\mathbf{b}$.

Sometimes, we don't know whether it is a power model or an exponential model. We can draw two scatter plots to show the relationship of $\ln \mathbf{y} \sim \ln \mathbf{x}$ and $\ln \mathbf{y} \sim \mathbf{x}$, to see which one is more linear.

Exercise:
Here is the data for the nine planets in figure 1.11.

Planet	Distance from sun (astronomical units)	Period of revolution (Earth years)
Mercury	0.387	0.241
Venus	0.723	0.615
Earth	1.000	1.000
Mars	1.524	1.881
Jupiter	5.203	11.862
Saturn	9.539	29.456
Uranus	19.191	84.070
Neptune	30.061	164.810
Pluto	39.529	248.530

Figure 1.11: Data for the nine planets (*Adapted from 'The practice of statistics'*)

(1) Is power model or exponential model more appropriate to for the relationship of **Period** $\sim$ **Distance**?

(2) If a new planet called Eris is discovered with distance 102.15 AU, what is the period?