STAT2401: Analysis of Experiments

Darfiana Nur

Dept. of Math. & Stat.

May 14, 2024

Aims of Lecture Week 11

- AIM 1 Extensions of the Linear Model:
  - Removing additive terms;
  - Interaction
- AIM 2 ANCOVA: Parallel regression lines
- AIM 3 Non-linear Relationships

AIM 1 Extensions of the Linear Model

- The standard linear regression model provides interpretable results and works quite well on many real world problems.
- However, it makes several highly restrictive assumptions that are often violated in practice.
- Two of the most important assumptions state that the relationship between the predictors and response are *additive* and *linear*.
- The *additive assumption* means that the effect of changes in a predictor X_j on the response Y is *independent* of the values of the other predictors.

AIM 1 Extensions of the Linear Model

- The standard linear regression model provides interpretable results and works quite well on many real world problems.
- However, it makes several highly restrictive assumptions that are often violated in practice.
- Two of the most important assumptions state that the relationship between the predictors and response are *additive* and *linear*.
- The *additive assumption* means that the effect of changes in a predictor X_j on the response Y is *independent* of the values of the other predictors.
- The *linear assumption* states that the change in the response Y due to a one – unit change in X_j is constant, regardless of the value of X_j.
- We examine a number of sophisticated methods that relax these two assumptions. Here, we briefly examine some common classical approaches for extending the linear model.

Removing the Additive Assumption

- In our previous analysis of the `Advertising` data, we concluded that both `TV` and `radio` seem to be associated with `sales`.
- The linear models that formed the basis for this conclusion assumed that the effect on `sales` of increasing one `advertising` medium is independent of the amount spent on the other media.

Removing the Additive Assumption

- For example, the linear model

$$\texttt{sales} = \beta_0 + \beta_1 \times \texttt{TV} + \beta_2 \times \texttt{radio} + \beta_3 \times \texttt{newspaper} + \epsilon$$

  states that the average effect on sales of a one-unit increase in `TV` is always β_1, regardless of the amount spent on `radio`.
- However, this simple model may be incorrect. Suppose that spending money on `radio` advertising actually increases the effectiveness of `TV` advertising, so that the slope term for `TV` should increase as `radio` increases.
- In this situation, given a fixed budget of $100,000, spending half on `radio` and half on `TV` may increase `sales` more than allocating the entire amount to either `TV` or to `radio`.

Removing the Additive Assumption

- In marketing, this is known as a *synergy* effect, and in statistics it is referred to as an *interaction* effect.
- The truth may suggest that such an effect may be present in the `advertising` data.
- Notice that when levels of either `TV` or `radio` are low, then the true `sales` are lower than predicted by the linear model.
- But when advertising is split between the two media, then the model tends to underestimate `sales`.

Removing the Additive Assumption

- Consider the standard linear regression model with two variables,

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \epsilon$$

  According to this model, if we increase X_1 by one unit, then Y will increase by an average of β_1 units.
- Notice that the presence of X_2 does not alter this statement – that is, regardless of the value of X_2, a one-unit increase in X_1 will lead to a β_1 – unit increase in Y.
- One way of extending this model to allow for *interaction* effects is to include a third predictor, called an *interaction* term, which is constructed by computing the product of X_1 and X_2. This results in the model

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2 + \epsilon$$

Removing the Additive Assumption

- How does inclusion of this **interaction** term relax the additive assumption? Notice that the model can be rewritten as

$$
\begin{aligned}
Y &= \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2 + \epsilon \\
&= \beta_0 + (\beta_1 + \beta_3 X_2) X_1 + \beta_2 X_2 + \epsilon \\
&= \beta_0 + \widetilde{\beta}_1 X_1 + \beta_2 X_2 + \epsilon
\end{aligned}
$$

  where $\widetilde{\beta}_1 = \beta_1 + \beta_3 X_2$. Since $\widetilde{\beta}_1$ changes with X_2, the effect of X_1 on Y is no longer constant: adjusting X_2 will change the impact of X_1 on Y.

Removing the Additive Assumption

- For example, suppose that we are interested in studying the productivity of a factory.
- We wish to predict the number of `units` produced on the basis of the number of production `lines` and the total number of `workers`.
- It seems likely that the effect of increasing the number of production `lines` will depend on the number of `workers`, since if no `workers` are available to operate the `lines`, then increasing the number of `lines` will not increase production.
- This suggests that it would be appropriate to include an *interaction* term between `lines` and `workers` in a linear model to predict units.

- Suppose that when we fit the model, we obtain

$$
\begin{aligned}
\widehat{\texttt{units}} &= \widehat{\beta}_0 + \widehat{\beta}_1 \times \texttt{lines} + \widehat{\beta}_2 \times \texttt{workers} + \widehat{\beta}_3 \times \texttt{lines} \times \texttt{workers} \\
&= \widehat{\beta}_0 + (\widehat{\beta}_1 + \widehat{\beta}_3 \times \texttt{workers}) \times \texttt{lines} + \widehat{\beta}_2 \times \texttt{workers}
\end{aligned}
$$

In other words, adding an additional line will increase the number of `units` produced by $\widehat{\beta}_1 + \widehat{\beta}_3 \times \texttt{workers}$. Say $\widehat{\beta}_1$, $\widehat{\beta}_2$ and $\widehat{\beta}_3$ are all positive. Hence the more `workers` we have, the stronger will be the effect of `lines`.

Removing the Additive Assumption

- We now return to the Advertising example.
- A linear model that uses radio, TV, and an *interaction* between the two to predict sales takes the

$$\begin{aligned} \texttt{sales} &= \beta_0 + \beta_1 \times \texttt{TV} + \beta_2 \times \texttt{radio} + \beta_3 \times \texttt{TV} \times \texttt{radio} + \epsilon \\ &= \beta_0 + (\beta_1 + \beta_3 \times \texttt{radio}) \times \texttt{TV} + \beta_2 \times \texttt{radio} + \epsilon \end{aligned}$$

We can interpret β_3 as the increase in the effectiveness of TV advertising for a one unit increase in radio advertising (or vice-versa).

Removing the Additive Assumption

- The result from fitting the model are given by

```
> summary(lm(sales~TV+radio+TV*radio,data=Advertising))
```

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 6.750e+00  2.479e-01  27.233   <2e-16 ***
TV          1.910e-02  1.504e-03  12.699   <2e-16 ***
radio       2.886e-02  8.905e-03   3.241   0.0014 **
TV:radio    1.086e-03  5.242e-05  20.727   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

- The fitted model is given by

$$\begin{aligned}\widehat{\texttt{sales}} &= 6.7502 + 0.0191 \times \texttt{TV} + 0.0289 \times \texttt{radio} + 0.0011 \times \texttt{TV} \times \texttt{radio} \\ &= 6.7502 + (0.0191 + 0.0011 \times \texttt{radio}) \times \texttt{TV} + 0.0289 \times \texttt{radio}\end{aligned}$$

Removing the Additive Assumption

- The results strongly suggest that the model that includes the the *interaction* term is superior to the model that contains only *main effects*.
- The p-value for the *interaction* term, `TV` $\times$ `radio`, is extremely low, indicating that there is strong evidence for $H_1 : \beta_3 \neq 0$.

```
> summary(lm(sales~TV+radio+TV*radio,data=Advertising))
```

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 6.750e+00  2.479e-01  27.233   <2e-16 ***
TV          1.910e-02  1.504e-03  12.699   <2e-16 ***
radio       2.886e-02  8.905e-03   3.241   0.0014 **
TV:radio    1.086e-03  5.242e-05  20.727   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

- In other words, it is clear that the true relationship is not additive.

Removing the Additive Assumption

- The adjusted R^2 for the model is 96.73%,

```
> summary(lm(sales~TV+radio+TV*radio,data=Advertising))
```

```
Residual standard error: 0.9435 on 196 degrees of freedom
Multiple R-squared:  0.9678,Adjusted R-squared:  0.9673
F-statistic:  1963 on 3 and 196 DF,  p-value: < 2.2e-16
```

  compared to only 89.62% for the model that predicts `sales` using `TV` and `radio` without an *interaction* term:

```
> summary(lm(sales~TV+radio,data=Advertising))
```

```
Residual standard error: 1.681 on 197 degrees of freedom
Multiple R-squared:  0.8972,Adjusted R-squared:  0.8962
F-statistic: 859.6 on 2 and 197 DF,  p-value: < 2.2e-16
```

- This means that $(96.73 - 89.62)/(100 - 89.62) = 68.50\%$ of the variability in sales that remains after fitting the *additive* model has been explained by the *interaction* term.

- Recall the fitted model

$$
\begin{aligned}
\widehat{\texttt{sales}} &= 6.7502 + 0.0191 \times \texttt{TV} + 0.0289 \times \texttt{radio} + 0.0011 \times \texttt{TV} \times \texttt{radio} \\
&= 6.7502 + (0.0191 + 0.0011 \times \texttt{radio}) \times \texttt{TV} + 0.0289 \times \texttt{radio} \\
&= 6.7502 + 0.0191\texttt{TV} + (0.0289 + 0.0011 \times \texttt{TV}) \times \texttt{Radio}
\end{aligned}
$$

- The coefficient estimates suggest that an increase in TV advertising of $\$1,000$ is associated with increased sales of

$$(0.0191 + 0.0011 \times \texttt{radio}) \times 1000 = 19.1 + 1.1 \times \texttt{radio} \quad \text{units}$$

  and an increase in radio advertising of $\$1,000$ will be associated with an increase in sales of

$$(0.0289 + 0.0011 \times \texttt{TV}) \times 1000 = 28.9 + 1.1 \times \texttt{TV} \quad \text{units}$$

Removing the Additive Assumption

- In this example, the *p*-values associated with `TV`, `radio`, and the *interaction* term all are statistically significant:

```
> summary(lm(sales~TV+radio+TV*radio,data=Advertising))
```

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 6.750e+00  2.479e-01  27.233   <2e-16 ***
TV          1.910e-02  1.504e-03  12.699   <2e-16 ***
radio       2.886e-02  8.905e-03   3.241   0.0014 **
TV:radio    1.086e-03  5.242e-05  20.727   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

  so it is obvious that all three variables should be included in the model.

- However, it is sometimes the case that an *interaction* term has a very small *p*-value, but the associated *main effects* (in this case, `TV` and `radio`) do not.

Removing the Additive Assumption

- The hierarchical principle states that if we include an *interaction* in a model, we should also include the *main effects*, even if the p-values associated with their coefficients are not significant.
- In other words, if the *interaction* between X_1 and X_2 seems important, then we should include both X_1 and X_2 in the model even if their coefficient estimates have large p-values.
- The rationale for this principle is that if $X_1 \times X_2$ is related to the response, then whether or not the coefficients of X_1 or X_2 are exactly zero is of little interest.
- Also $X_1 \times X_2$ is typically correlated with X_1 and X_2, and so leaving them out tends to alter the meaning of the interaction.

Removing the Additive Assumption

- We have considered an interaction between `TV` and `radio`, both of which are quantitative variables.
- However, the concept of *interactions* applies just as well to qualitative variables, or to a combination of quantitative and qualitative variables.
- In fact, an *interaction* between a qualitative variable and a quantitative variable has a particularly nice interpretation.

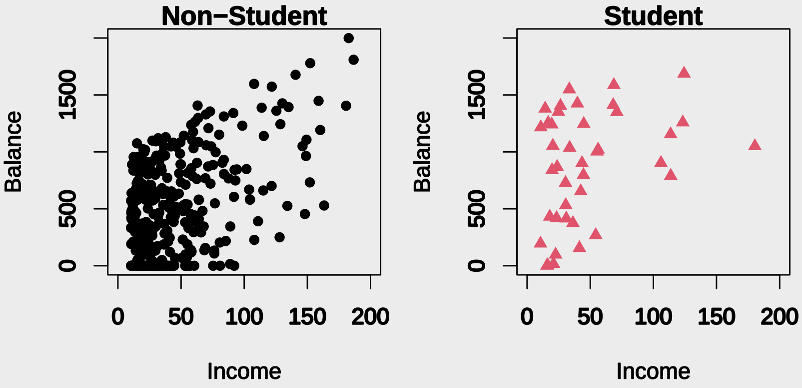
Removing the Additive Assumption

- Consider the `Credit` data set, and suppose that we wish to predict balance using the `Income` (quantitative) and `Student` (qualitative) variables.

```
> Credit = read.csv("Credit.csv",header=TRUE)
> str(Credit)
'data.frame': 400 obs. of  11 variables:
 $ Income   : num  14.9 106 104.6 148.9 55.9 ...
 $ Limit    : int  3606 6645 7075 9504 4897 8047 3388 7114 3300 6819 ...
 $ Rating   : int  283 483 514 681 357 569 259 512 266 491 ...
 $ Cards    : int  2 3 4 3 2 4 2 2 5 3 ...
 $ Age      : int  34 82 71 36 68 77 37 87 66 41 ...
 $ Education: int  11 15 11 11 16 10 12 9 13 19 ...
 $ Gender   : Factor w/ 2 levels "Female","Male": 2 1 2 1 2 2 1 2 1 1 ...
 $ Student  : Factor w/ 2 levels "No","Yes": 1 2 1 1 1 1 1 1 1 2 ...
 $ Married  : Factor w/ 2 levels "No","Yes": 2 2 1 1 2 1 1 1 1 2 ...
 $ Ethnicity: Factor w/ 3 levels "African American",..: 3 2 2 2 3 3 1 2 3 1 ..
 $ Balance  : int  333 903 580 964 331 1151 203 872 279 1350 ...
```

- Alternatively, we have the figures

```
> par(mfrow=c(1,2))
> with(subset(Credit,Student=="No"),plot(Balance~Income,xlab="Income",
+       ylab="Balance",col=1,pch=16,xlim=c(0,200),ylim=c(0,2000),
+       main="Non-Student"))
> with(subset(Credit,Student=="Yes"),plot(Balance~Income,xlab="Income",
+       ylab="Balance",col=2,pch=17,xlim=c(0,200),ylim=c(0,2000),
+       main="Student"))
```



Removing the Additive Assumption

- The model takes the form

$$\begin{aligned}
\texttt{Balance}_i &= \beta_0 + \beta_1 \times \texttt{Income}_i + \beta_2 \times \texttt{Student}_i + \epsilon_i \\
&= \beta_0 + \beta_1 \times \texttt{Income}_i \\
&\quad + \begin{cases} \beta_2 & \text{if } i\text{th person is a } \texttt{Student} \\ 0 & \text{if } i\text{th person is NOT a } \texttt{Student} \end{cases} + \epsilon_i \\
&= \begin{cases} (\beta_0 + \beta_2) + \beta_1 \times \texttt{Income}_i + \epsilon_i \\ \qquad\qquad \text{if } i\text{th person is a } \texttt{Student} \\ \beta_0 + \beta_1 \times \texttt{Income}_i + \epsilon_i \\ \qquad\qquad \text{if } i\text{th person is NOT a } \texttt{Student} \end{cases}
\end{aligned}$$

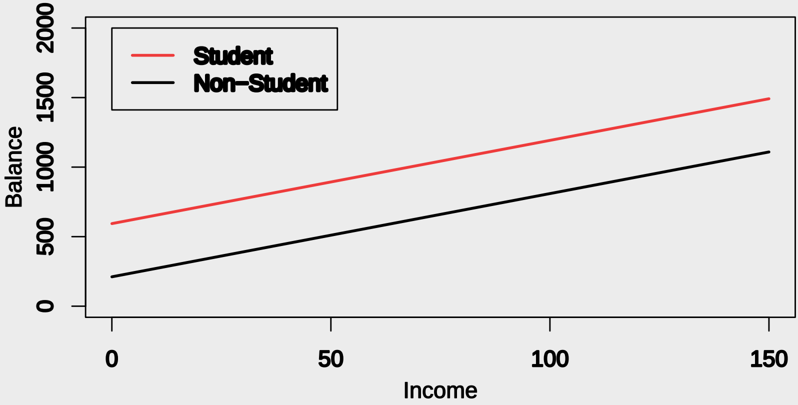
Here

$$\texttt{Student}_i = \begin{cases} 1 & \text{if } i\text{th person is a } \texttt{Student} \\ 0 & \text{if } i\text{th person is NOT a } \texttt{Student} \end{cases}$$

Removing the Additive Assumption

- Notice that this amounts to fitting two parallel lines to the data, one for `Student`s and one for non-`Student`s.
- The lines for `Student`s and non-`Student`s have different intercepts, $\beta_0 + \beta_2$ versus β_0, but the same slope, β_1.

- This is illustrated here:



- The fact that the lines are parallel means that the average effect on balance of a one-unit increase in `Income` does not depend on whether or not the individual is a `Student`.

Removing the Additive Assumption

- This represents a potentially serious limitation of the model, since in fact a change in `Income` may have a very different effect on the credit card `Balance` of a `Student` versus a non-`Student`.
- This limitation can be addressed by adding an interaction variable, created by multiplying `Income` with the dummy variable for `Student`.

Removing the Additive Assumption

- Our model now becomes

$$
\begin{aligned}
\texttt{Balance}_i &= \beta_0 + \beta_1 \times \texttt{Income}_i + \beta_2 \times \texttt{Student}_i \\
&\quad + \beta_3 \times \texttt{Income}_i \times \texttt{Student}_i + \epsilon_i \\
&= \beta_0 + \beta_1 \times \texttt{Income}_i \\
&\quad + \begin{cases} \beta_2 + \beta_3 \times \texttt{Income}_i & \text{if } i\text{th person is a } \texttt{Student} \\ 0 & \text{if } i\text{th person is NOT a } \texttt{Student} \end{cases} + \epsilon_i \\
&= \begin{cases} (\beta_0 + \beta_2) + (\beta_1 + \beta_3) \times \texttt{Income}_i + \epsilon_i & \text{if } i\text{th person is a } \texttt{Student} \\ \beta_0 + \beta_1 \times \texttt{Income}_i + \epsilon_i & \text{if } i\text{th person is NOT a } \texttt{Student} \end{cases}
\end{aligned}
$$

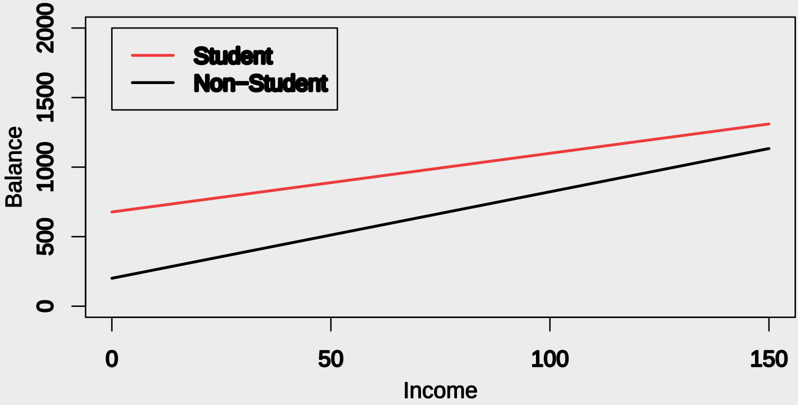
Here

$$
\texttt{Student}_i = \begin{cases} 1 & \text{if } i\text{th person is a } \texttt{Student} \\ 0 & \text{if } i\text{th person is NOT a } \texttt{Student} \end{cases}
$$

Removing the Additive Assumption

- Once again, we have two different regression lines for the `students`s and the non-`Students`s.
- But now those regression lines have different intercepts, $\beta_0 + \beta_2$ versus β_0, as well as different slopes, $\beta_1 + \beta_3$ versus β_1.
- This allows for the possibility that changes in income may affect the credit card balances of `Students`s and non-`Students`s differently.

- This is illustrated here:



- This figure shows the estimated relationships between `Income` and `Balance` for `Student`s and non-`Student`s in the model.

Removing the Additive Assumption

- We note that the slope for `Student`s is lower than the slope for non-`Student`s. This suggests that increases in `Income` are associated with smaller increases in credit card `Balance` among `Student`s as compared to non-`Student`s.

Analysis of Covariance (ANCOVA)

- Consider the situation in which we want to model a response variable, Y based on a continuous predictor, X and a dummy variable, Z. Suppose that the effect of X on Y is linear. This situation is the simplest version of what is commonly referred as Analysis of Covariance, since the predictors include both quantitative variables (i.e., X) and qualitative variables (i.e., Z). The model is then say

$$Y = \beta_0 + \beta_1 X + \beta_2 Z + \beta_3 ZX + \epsilon$$

  There are 4 possible models based on Z

  - **Coincident Regression Lines**/**Simple Regression Models**: The simplest model in the given situation is one in which the dummy variable Z has no effect on Y (both $\beta_2 = \beta_3 = 0$), that is, and the regression line is exactly the same for both values of the dummy variable. That is

$$Y = \beta_0 + \beta_1 X + \epsilon$$

    and the regression line is exactly the same for both values of the dummy variable.

Analysis of Covariance (ANCOVA)

- **Parallel Regression Lines/Parallel Regression Models**: Another model to consider for this situation is one in which the dummy variable produces only an additive change in Y ($\beta_2 \neq 0$ & $\beta_3 = 0$),

$$Y = \begin{cases} \beta_0 + \beta_1 X + \epsilon & Z = 0 \\ (\beta_0 + \beta_2) + \beta_1 X + \epsilon & Z = 1 \end{cases}$$

  In this case, the regression coeffcient β_2 measures the additive change in Y due to the dummy variable.

- **Regression Lines with equal intercepts but different slopes**: A third model to consider for this situation is one in which the dummy variable only changes the size of the effect of X on Y ($\beta_2 = 0$ & $\beta_3 \neq 0$), that is,

$$Y = \begin{cases} \beta_0 + \beta_1 X + \epsilon & Z = 0 \\ \beta_0 + (\beta_1 + \beta_3) X + \epsilon & Z = 1 \end{cases}$$

  This case is rarely considered since it is not that interesting to focus on equal intercepts (or intercept alone)

Analysis of Covariance (ANCOVA)

- - **Unrelated Regression Lines/Separate Regression Models**: The most general model is appropriate when the dummy variable produces an additive change in Y and also changes the size of the effect of X on Y ($\beta_2 \neq 0$ & $\beta_3 \neq 0$). In this case the appropriate model is,

$$Y = \begin{cases} \beta_0 + \beta_1 X + \epsilon & Z = 0 \\ (\beta_0 + \beta_2) + (\beta_1 + \beta_3) X + \epsilon & Z = 1 \end{cases}$$

- In Summary, we have the models under the following conditions

	$Z = 0$	$Z = 1$
$\beta_2 = 0$ & $\beta_3 = 0$	$Y = \beta_0 + \beta_1 X + \epsilon$	$Y = \beta_0 + \beta_1 X + \epsilon$
$\beta_2 \neq 0$ & $\beta_3 = 0$	$Y = \beta_0 + \beta_1 X + \epsilon$	$Y = (\beta_0 + \beta_2) + \beta_1 X + \epsilon$
$\beta_2 = 0$ & $\beta_3 \neq 0$	$Y = \beta_0 + \beta_1 X + \epsilon$	$Y = \beta_0 + (\beta_1 + \beta_3) X + \epsilon$
$\beta_2 \neq 0$ & $\beta_3 \neq 0$	$Y = \beta_0 + \beta_1 X + \epsilon$	$Y = (\beta_0 + \beta_2) + (\beta_1 + \beta_3) X + \epsilon$

- This then turns to be a model selection problem that we select the best models among them or choose variables from (X, Z, ZX).

Example: Amount spent on travel

- This example is based on a problem in a text on business statistics. The background to the example is as follows:

  Small travel agency has retained your services to help them better understand two important customer segments. The first segment, which we will denote by `A`, consists of those customers who have purchased an adventure tour in the last twelve months. The second segment, which we will denote by `C`, consists of those customers who have purchased a cultural tour in the last twelve months. Data are available on 925 customers (i.e. on 466 customers from segment `A` and 459 customers from segment `C`). Note that the two segments are completely separate in the sense that there are no customers who are in both segments. Interest centres on *identifying any differences between the two segments in terms of the amount of money spent in the last twelve months*. In addition, data are also available on the age of each customer, since age is thought to have an effect on the amount spent.

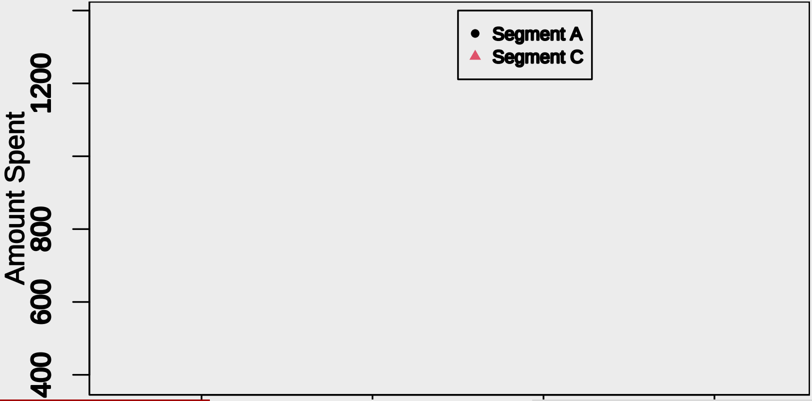
Example: Amount spent on travel

- ```
  > travel = read.table("travel.txt",header=TRUE)
  > str(travel)
  'data.frame': 925 obs. of  4 variables:
   $ Amount : int  997 997 951 649 1265 1059 837 924 852 963 ...
   $ Age    : int  44 43 41 59 25 38 46 42 48 39 ...
   $ Segment: chr  "A" "A" "A" "A" ...
   $ C      : int  0 0 0 0 0 0 0 0 0 0 ...
  > travel[1:10,]
     Amount Age Segment C
  1     997  44       A 0
  2     997  43       A 0
  3     951  41       A 0
  4     649  59       A 0
  5    1265  25       A 0
  6    1059  38       A 0
  7     837  46       A 0
  8     924  42       A 0
  9     852  48       A 0
  10    963  39       A 0
  ```
  ```

Example: Amount spent on travel

- ```
  > travel[907:925,]
      Amount Age Segment C
  907    918  42       C 1
  908    944  47       C 1
  909    770  40       C 1
  910    835  41       C 1
  911   1300  61       C 1
  912    548  25       C 1
  913    726  36       C 1
  914    760  39       C 1
  915   1150  55       C 1
  916   1117  59       C 1
  917    535  30       C 1
  918    985  52       C 1
  919    547  26       C 1
  920    954  51       C 1
  921   1110  59       C 1
  922    907  44       C 1
  923   1111  57       C 1
  924    883  43       C 1
  925   1038  53       C 1
  ```
  ```

Example: Amount spent on travel

```
> with(travel,plot(Amount~Age,xlab="Age",ylab="Amount Spent",
+                   col=as.numeric(Segment),pch=as.numeric(Segment)+15))
> ## read str(travel) for as.numeric(Segment) or check as.numeric(Segment)
> legend(45,1400,c("Segment A","Segment C"),col=c(1,2),pch=c(16,17),cex=0.7)
```

```
Warning in FUN(X[[i]], ...):  NAs introduced by coercion
Warning in FUN(X[[i]], ...):  NAs introduced by coercion
```



Example: Amount spent on travel

- The dummy variable for `segment` changes the size of the effect of `Age` on `Amount` Spent.
- We shall also allow for the dummy variable for `Segment` to produce an additive change in `Amount` Spent.
- The appropriate model is what we referred to above as Unrelated regression lines/Separate Regression Models.

$$\begin{aligned} \texttt{Amount} &= \beta_0 + \beta_1 \texttt{Age} + \beta_2 \texttt{C} + \beta_3 \texttt{CAge} + \epsilon \\ &= \begin{cases} \texttt{Amount} = \beta_0 + \beta_1 \texttt{Age} + \epsilon & \texttt{C} = 0 \\ \texttt{Amount} = (\beta_0 + \beta_2) + (\beta_1 + \beta_3)\texttt{Age} + \epsilon & \texttt{C} = 1 \end{cases} \end{aligned}$$

where `C` is a dummy variable that

$$\texttt{C} = \begin{cases} 1 & \text{if the customer is in Segment } \texttt{C} \\ 0 & \text{if the customer is in Segment } \texttt{A} \end{cases}$$

Example: Amount spent on travel

```
> M3 = lm(Amount~Age+C+Age:C,data=travel)
> summary(M3)

Call:
lm(formula = Amount ~ Age + C + Age:C, data = travel)

Residuals:
     Min       1Q   Median       3Q      Max 
-143.298  -30.541   -0.034   31.108  130.743 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept)  1814.5445     8.6011   211.0   <2e-16 ***
Age           -20.3175     0.1878  -108.2   <2e-16 ***
C           -1821.2337    12.5736  -144.8   <2e-16 ***
Age:C          40.4461     0.2724   148.5   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 47.63 on 921 degrees of freedom
Multiple R-squared:  0.9601,Adjusted R-squared:  0.9599 
F-statistic:  7379 on 3 and 921 DF,  p-value: < 2.2e-16
```

Example: Amount spent on travel

- Notice that all the regression coefficients are highly statistically significant.
- Thus, the ovarall fitted model is

$$\widehat{\texttt{Amount}} = 1814.5445 - 20.3175\texttt{Age} - 1821.2337\texttt{C} + 40.4461\texttt{CAge}$$

- For customers in segment $\texttt{A}$, (i.e., $\texttt{C} = 0$) our model predicts,

$$\widehat{\texttt{Amount}} = 1814.5445 - 20.3175\texttt{Age}$$

while for customers in segment $\texttt{C}$ (i.e., $\texttt{C} = 1$) our model predicts

$$\begin{aligned}\widehat{\texttt{Amount}} &= (1814.5445 - 1821.2337) + (-20.3175 + 40.4461)\,\texttt{Age} \\ &= -6.6892 + 20.1286\texttt{Age}\end{aligned}$$

Example: Amount spent on travel

- We shall compare this model with Coincident Regression Line/Simple Regression Model and Parallel Regression Line/Parallel Regression Model to see which the models we should prefer.

```
> M3 = lm(Amount~Age+C+Age:C,data=travel)  ## Separate Regression Model
> M2 = lm(Amount~Age+C,data=travel)  ## Parallel Regression Model
> M1 = lm(Amount~Age,data=travel)    ## Simple Regression Model
```

Example: Amount spent on travel

- partial output of `summary(M1)` and `summary(M2)`

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 957.9103    31.3056  30.599   <2e-16 ***
Age          -1.1140     0.6784  -1.642    0.101

Residual standard error: 237.7 on 923 degrees of freedom
Multiple R-squared:  0.002913,Adjusted R-squared:  0.001833
F-statistic: 2.697 on 1 and 923 DF,  p-value: 0.1009
```

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 963.4254    32.0143  30.094   <2e-16 ***
Age          -1.0939     0.6789  -1.611    0.107
C           -12.9291    15.6455  -0.826    0.409

Residual standard error: 237.8 on 922 degrees of freedom
Multiple R-squared:  0.003651,Adjusted R-squared:  0.00149
F-statistic: 1.689 on 2 and 922 DF,  p-value: 0.1852
```

The coefficients (in both models) are not statistically signifiance!

Example: Amount spent on travel

- Compare M3 and M2

```
> anova(M2,M3)
Analysis of Variance Table

Model 1: Amount ~ Age + C
Model 2: Amount ~ Age + C + Age:C
  Res.Df      RSS Df Sum of Sq     F    Pr(>F)
1    922 52120341
2    921  2089377  1  50030964 22054 < 2.2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Clearly, we prefer M3

Analysis of Covariance

Example: Amount spent on travel

- Compare M2 and M1

```
> anova(M1,M2)
Analysis of Variance Table

Model 1: Amount ~ Age
Model 2: Amount ~ Age + C
  Res.Df      RSS Df Sum of Sq      F Pr(>F)
1    923 52158945
2    922 52120341  1     38604 0.6829 0.4088
```

M3 is the best!

Note that the p-value with high precision can be extracted from

```
> anova(M1,M2)$'Pr(>F)'[2]
[1] 0.4088046
```

Example: Amount spent on travel

- double check: Compare M3 and M1

```
> anova(M1,M3)
Analysis of Variance Table

Model 1: Amount ~ Age
Model 2: Amount ~ Age + C + Age:C
  Res.Df      RSS Df Sum of Sq     F    Pr(>F)
1    923 52158945
2    921  2089377  2  50069568 11035 < 2.2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

AIM 3 Non-linear Relationships

- The linear regression model assumes a linear relationship between the response and predictors.
- But in some cases, the true relationship between the response and the predictors may be nonlinear.
- Here we present a very simple way to directly extend the linear model to accommodate non-linear relationships, using *polynomial regression*.

Non-linear Relationships

- Consider the `Auto` data set, in which the `mpg` (gas mileage in miles per gallon) versus `horsepower` is shown for a number of cars.

```
> Auto = read.csv("Auto.csv",header=TRUE,na.strings="?")
> str(Auto)
'data.frame': 397 obs. of  9 variables:
 $ mpg         : num  18 15 18 16 17 15 14 14 14 15 ...
 $ cylinders   : int  8 8 8 8 8 8 8 8 8 8 ...
 $ displacement: num  307 350 318 304 302 429 454 440 455 390 ...
 $ horsepower  : int  130 165 150 150 140 198 220 215 225 190 ...
 $ weight      : int  3504 3693 3436 3433 3449 4341 4354 4312 4425 3850 ...
 $ acceleration: num  12 11.5 11 12 10.5 10 9 8.5 10 8.5 ...
 $ year        : int  70 70 70 70 70 70 70 70 70 70 ...
 $ origin      : int  1 1 1 1 1 1 1 1 1 1 ...
 $ name        : chr  "chevrolet chevelle malibu" "buick skylark 320" "plymout
```

- The data has 397 observations, or rows, and 9 variables, or columns.

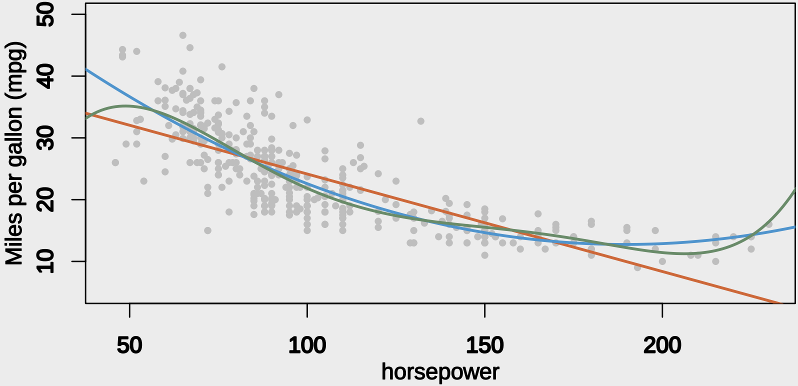
Non-linear Relationships

- There are various ways to deal with the missing data. In this case, only five of the rows contain missing observations, and so we choose to use the `na.omit()` function to simply remove these rows.

```
> Auto = na.omit(Auto)
> str(Auto)
'data.frame': 392 obs. of  9 variables:
 $ mpg         : num  18 15 18 16 17 15 14 14 14 15 ...
 $ cylinders   : int  8 8 8 8 8 8 8 8 8 8 ...
 $ displacement: num  307 350 318 304 302 429 454 440 455 390 ...
 $ horsepower  : int  130 165 150 150 140 198 220 215 225 190 ...
 $ weight      : int  3504 3693 3436 3433 3449 4341 4354 4312 4425 3850 ...
 $ acceleration: num  12 11.5 11 12 10.5 10 9 8.5 10 8.5 ...
 $ year        : int  70 70 70 70 70 70 70 70 70 70 ...
 $ origin      : int  1 1 1 1 1 1 1 1 1 1 ...
 $ name        : chr  "chevrolet chevelle malibu" "buick skylark 320" "plymout
 - attr(*, "na.action")= 'omit' Named int [1:5] 33 127 331 337 355
  ..- attr(*, "names")= chr [1:5] "33" "127" "331" "337" ...
```

Non-linear Relationships

- For a number of cars, `mpg` and `horsepower` are shown. The linear regression fit is shown in orange. The linear regression fit for a model that includes `horsepower`^{2} is shown as a blue curve. The linear regression fit for a model that includes all polynomials of `horsepower` up to fifth-degree is shown in green.



Non-linear Relationships

- The orange line represents the linear regression fit. There is a pronounced relationship between `mpg` and `horsepower`,

```
> summary(lm(mpg~horsepower,data=Auto))

Call:
lm(formula = mpg ~ horsepower, data = Auto)

Residuals:
     Min       1Q   Median       3Q      Max
-13.5710  -3.2592  -0.3435   2.7630  16.9240

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 39.935861   0.717499   55.66   <2e-16 ***
horsepower  -0.157845   0.006446  -24.49   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.906 on 390 degrees of freedom
Multiple R-squared:  0.6059,Adjusted R-squared:  0.6049
F-statistic: 599.7 on 1 and 390 DF,  p-value: < 2.2e-16
```

but it seems clear that this relationship is in fact non-linear: the data

Non-linear Relationships

- A simple approach for incorporating non-linear associations in a linear model is to include transformed versions of the predictors in the model.
- For example, the points seem to have a quadratic shape, suggesting that a quadratic model of the form

$$\texttt{mpg} = \beta_0 + \beta_1 \times \texttt{horsepower} + \beta_2 \times \texttt{horsepower}^2 + \epsilon$$

  may provide a better fit.
- This model involves predicting `mpg` using a non-linear function of `horsepower`. But it is still a linear model!
- That is, it is simply a multiple linear regression model with $X_1 = \texttt{horsepower}$ and $X_2 = \texttt{horsepower}^2$.
- So we can use standard linear regression software to estimate β_0, β_1, and β_2 in order to produce a non-linear fit.

Non-linear Relationships

- The blue curve shows the resulting quadratic fit to the data. The quadratic fit appears to be substantially better than the fit obtained when just the linear term is included.

```
> summary(lm(mpg~horsepower+I(horsepower^2),data=Auto))

Call:
lm(formula = mpg ~ horsepower + I(horsepower^2), data = Auto)

Residuals:
     Min       1Q   Median       3Q      Max
-14.7135  -2.5943  -0.0859   2.2868  15.8961

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)     56.9000997  1.8004268   31.60   <2e-16 ***
horsepower      -0.4661896  0.0311246  -14.98   <2e-16 ***
I(horsepower^2)  0.0012305  0.0001221   10.08   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.374 on 389 degrees of freedom
Multiple R-squared:  0.6876,Adjusted R-squared:  0.686
F-statistic:   428 on 2 and 389 DF,  p-value: < 2.2e-16
```

Non-linear Relationships

- The R^2 of the quadratic fit is 0.6860, compared to 0.6059 for the linear fit, and the p-value for the quadratic term is highly significant.

Non-linear Relationships

- If including `horsepower`^{2} led to such a big improvement in themodel, why not include `horsepower`^{3}, `horsepower`^{4}, or even `horsepower`^{5}?

Non-linear Relationships

- The green curve displays the fit that results from including all polynomials up to fifth degree in the model:

```
> summary(lm(mpg~horsepower+I(horsepower^2)+I(horsepower^3)
+           +I(horsepower^4)+I(horsepower^5),data=Auto))

Call:
lm(formula = mpg ~ horsepower + I(horsepower^2) + I(horsepower^3) +
    I(horsepower^4) + I(horsepower^5), data = Auto)

Residuals:
     Min       1Q   Median       3Q      Max
-15.4326  -2.5285  -0.2925   2.1750  15.9730

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)     -3.223e+01  2.857e+01  -1.128  0.26003
horsepower       3.700e+00  1.303e+00   2.840  0.00475 **
I(horsepower^2) -7.142e-02  2.253e-02  -3.170  0.00164 **
I(horsepower^3)  5.931e-04  1.850e-04   3.206  0.00146 **
I(horsepower^4) -2.281e-06  7.243e-07  -3.150  0.00176 **
I(horsepower^5)  3.330e-09  1.085e-09   3.068  0.00231 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Non-linear Relationships

- The resulting fit seems unnecessarily wiggly – that is, it is unclear that including the additional terms really has led to a better fit to the data.

Non-linear Relationships

- The approach that we have just described for extending the linear model to accommodate non-linear relationships is known as *polynomial regression*, since we have included polynomial functions of the predictors in the regression model.