# MAST7866: Multiple Linear Regression

Professor Jian Zhang

## Using more explanatory variables

We have looked at modelling the relationship between an explanatory variable and a response variable. The model allows us to make predictions.

More explanatory variables can lead to better predictions if they provide additional information about the response.

## Example: Factors affecting the price of food in New York

The data concerns the price of food in high-end Italian restaurants in New York. The scores for food, décor and service are taken from www.zagat.com.

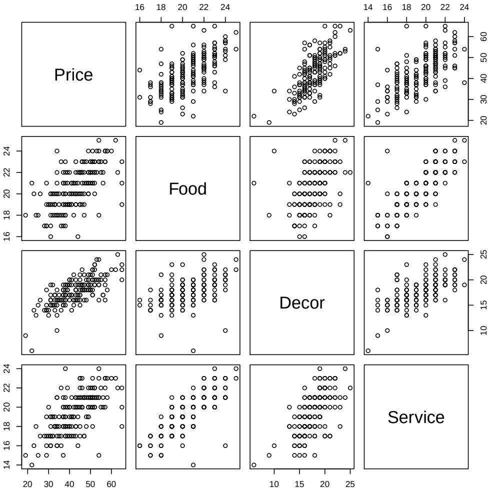
The first three restaurants in the data set are

| Case | Restaurant | Price | Food | Decor | Service |
|---|---|---|---|---|---|
| 1 | Daniella Ristorante | 43 | 22 | 18 | 20 |
| 2 | Tello's Ristorante | 32 | 20 | 19 | 19 |
| 3 | Biricchino | 34 | 21 | 13 | 18 |

where

- `Price` is the price (in $US) of dinner (including one drink & a tip).
- `Food` is the customer rating of the food (out of 30).
- `Decor` is the customer rating of the decor (out of 30).
- `Service` is the customer rating of the service (out of 30).

# Example: Factors affecting the price of food in New York



## Questions you might want to ask

You could imagine working with a new restaurant to set competitive prices for their food.

Some questions that you might want to ask are:

- How does the quality of food, décor and service affect the price of food in these restaurants?
- Could a restaurant charge a higher price for its food if the level of service increases?
- Which variables are useful for predicting the price of food in New York restaurants?

## Multiple Linear Regression model

We extend the simple linear regression model by assuming that the effects of each variable is linear

$$y_i = \alpha + \beta_1 x_{i,1} + \cdots + \beta_K x_{i,K} + e_i$$

where

- $y_i$ is the response variable.
- $x_{i,j}$ is the value of the $j$-th explanatory variable for the $i$-th subject.
- $e_i$ is the error with $\mathrm{E}(e_i) = 0$ and $\mathrm{Var}(e_i) = \sigma^2$.
- $\beta_j$ is called the effect of the $j$-th variable.

This implies that the mean of $y_i$ is $\alpha + \beta_1\, x_{i,1} + \cdots + \beta_K\, x_{i,K}$ and the variance of $y_i$ is $\sigma^2$.

## Questions you might want to ask – some answers

$$\text{Price} = \alpha + \beta_1\,\text{Food} + \beta_2\,\text{Decor} + \beta_3\,\text{Service}$$

- How does the quality of food, décor and service affect the price of food in these restaurants?
  Answer:
  On average, the price increases by $\beta_1$ if the Food score increases by 1 and the other scores stay the same.

  Similarly, on average, the price increases by $\beta_2$ if the Décor score increases by 1 (and the other scores are the same) and the price increases by $\beta_3$ if the Service score increases by 1 (and the other scores are the same).

## Questions you might want to ask – some answers

$$\text{Price} = \alpha + \beta_1\,\text{Food} + \beta_2\,\text{Decor} + \beta_3\,\text{Service}$$

- Could a restaurant charge a higher price for its food if the level of service increases?
  Answer: Yes, if $\beta_3 > 0$.
- Which variables are useful for predicting the price of food in New York restaurants?
  Answer: Food is useful if $\beta_1 \neq 0$, Decor is useful if $\beta_2 \neq 0$ and Service is useful is $\beta_3 \neq 0$.

## Example: New York food prices

```
> fit_nyc <- lm(Price ~ Food + Decor + Service, data=nyc)
> fit_nyc

Call:
lm(formula = Price ~ Food + Decor + Service, data = nyc)

Coefficients:
(Intercept)         Food        Decor      Service
    -24.641        1.556        1.847        0.135
```

## Interpretation

$$\text{Price} = -24.641 + 1.556 \times \text{Food} + 1.847 \times \text{Decor} + 0.135 \times \text{Service}$$

- The effect of Food on Price is 1.556, *i.e. a one unit increase in the food score leads a 1.556 unit increase in price on average*.
- The effect of Decor on Price is 1.847, *i.e. a one unit increase in the decor score leads a 1.847 unit increase in price on average*.
- The effect of Service on Price is 0.135, *i.e. a one unit increase in the service score leads a 0.135 unit increase in price on average*.

## Estimating $\sigma^2$

Similarly to simple linear regression, we can define the fitted values by

$$\hat{y}_i = \hat{\alpha} + \hat{\beta}_1 x_{i,1} + \cdots + \hat{\beta}_K x_{i,K}$$

and residuals by

$$\begin{aligned} r_i &= y_i - \hat{y}_i \\ &= y_i - \hat{\alpha} - \hat{\beta}_1 x_{i,1} - \cdots - \hat{\beta}_K x_{i,K} \,. \end{aligned}$$

The residual sum of squares is $\mathrm{RSS} = r_1^2 + r_2^2 + \cdots + r_n^2$.

We can estimate $\sigma^2$ by

$$\hat{\sigma}^2 = \frac{\mathrm{RSS}}{n - p}$$

where $p = K + 1$.

## Example: New York food prices

```
Call:
lm(formula = Price ~ Food + Decor + Service, data = nyc)

Residuals:
     Min       1Q   Median       3Q      Max
-14.8440  -3.7039  -0.1525   3.6218  19.0576

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -24.6409     4.7536  -5.184 6.33e-07 ***
Food          1.5556     0.3731   4.170 4.93e-05 ***
Decor         1.8473     0.2176   8.491 1.17e-14 ***
Service       0.1350     0.3957   0.341    0.733
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 5.803 on 164 degrees of freedom
Multiple R-squared:  0.617,      Adjusted R-squared:   0.61
F-statistic: 88.06 on 3 and 164 DF,  p-value: < 2.2e-16
```

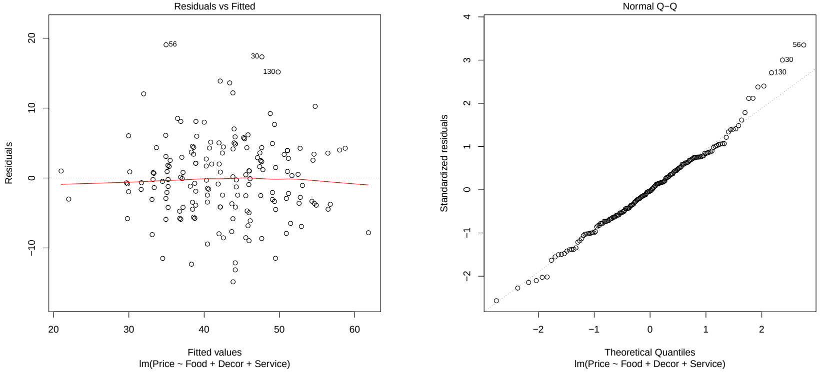
## Example: New York food prices

```
Call:
lm(formula = Price ~ Food + Decor + Service, data = nyc)

Residuals:
     Min       1Q   Median       3Q      Max 
-14.8440  -3.7039  -0.1525   3.6218  19.0576 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) -24.6409     4.7536  -5.184 6.33e-07 ***
Food          1.5556     0.3731   4.170 4.93e-05 ***
Decor         1.8473     0.2176   8.491 1.17e-14 ***
Service       0.1350     0.3957   0.341    0.733    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 5.803 on 164 degrees of freedom
Multiple R-squared:  0.617,	Adjusted R-squared:   0.61 
F-statistic: 88.06 on 3 and 164 DF,  p-value: < 2.2e-16
```

# Checking the model

Again we can check the suitability of the model using the residuals.

```
plot(fit_nyc, which = c(1, 2))
```



## Prediction

Often we want to predict values of the response for a new value of the explanatory variables $x_{0,1}, x_{0,2}, \ldots, x_{0,K}$.

The predicted value of the response from the fitted regression model is

$$\alpha + \beta_1 x_{0,1} + \beta_2 x_{0,2} + \cdots + \beta_K x_{0,K}.$$

Again, we can construct a confidence interval and a prediction interval.

1. confidence interval – the long-term average of many response values for the values $x_{0,1}, x_{0,2}, \ldots, x_{0,K}$ of the explanatory variables.
2. prediction interval – one particular response value for the value $x_{0,1}, x_{0,2}, \ldots, x_{0,K}$ of the explanatory variables.

## Example: New York food prices

Suppose that I plan to open a new restaurant in New York, *Casa Mia*. I think that the score will be

| Food | Decor | Service |
|---|---|---|
| 20 | 15 | 20 |

How much could I charge?
A reasonable range of prices is given by the prediction interval which is

```
> predict(fit_nyc, newdata = Data.frame(Food = 20, Decor = 15,
                                         Service = 20),
          interval = "prediction", level = 0.95)
      fit      lwr      upr
1 36.88147 25.27703 48.48591
```

Between \$25.28 and \$48.49.

## Example: New York food prices

What is the average price charged by restaurants with these scores?

The confidence interval gives a range for the average price which is

```
> predict(fit_nyc, newdata, interval = "confidence", level=0.95)
       fit      lwr      upr
1 36.88147 35.05131 38.71163
```

The 95% confidence interval for the average price is \$35.05 to \$38.71.

## Measuring the importance of variables

To build good models, we need to understand whether some (or all) of the variables are related to the response.

This can be addressed using $t$-values and analysis of variance.

# Example: New York food prices

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -24.6409     4.7536  -5.184 6.33e-07 ***
Food          1.5556     0.3731   4.170 4.93e-05 ***
Decor         1.8473     0.2176   8.491 1.17e-14 ***
Service       0.1350     0.3957   0.341    0.733
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Standard errors are available for the effect of each variable and measure how close an estimate is to the corresponding true parameter value.

## Testing the effect of a variable using a $t$-test

To test whether the effect of the $j$-th variable is different from zero (assuming that the effect of all other variables is different to zero), we can use a $t$-test.

The null hypothesis is
$H_0 : \beta_j = 0$ (and $\beta_1 \neq 0, \ldots, \beta_{j-1} \neq 0, \beta_{j+1} \neq 0, \ldots, \beta_K \neq 0$).

The alternative hypothesis is
$H_A : \beta_j \neq 0$ (and $\beta_1 \neq 0, \ldots, \beta_{j-1} \neq 0, \beta_{j+1} \neq 0, \ldots, \beta_K \neq 0$).

[Usually, we don't write the parts in brackets.]

## Testing the effect of a variable using a $t$-test

The null hypothesis can be tested using the following $t$-test statistic (for the $j$-th variable)

$$\frac{\text{Estimate}}{\text{Standard Error}}$$

Under the null hypothesis, this follows a $t$-distribution with the residual degrees of freedom which is $n - K - 1$.

## Example: New York food prices

| | Estimate | Std. Error | t value | Pr(>\|t\|) | |
|---|---|---|---|---|---|
| (Intercept) | -24.6409 | 4.7536 | -5.184 | 6.33e-07 | *** |
| Food | 1.5556 | 0.3731 | $\boxed{4.170}$ | $\boxed{4.93e-05}$ | *** |
| Decor | 1.8473 | 0.2176 | 8.491 | 1.17e-14 | *** |
| Service | 0.1350 | 0.3957 | 0.341 | 0.733 | |

To test whether the effect of food is zero if the effect of decor and service are not zero, we can use the *t*-statistic which is 4.17.

The *p*-value is 4.93e-05 which is very small and so the null hypothesis can be rejected, there is very strong evidence that Food has an effect on Price if Decor and Service have an effect.

## Example: New York food prices

```
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -24.6409     4.7536  -5.184 6.33e-07 ***
Food          1.5556     0.3731   4.170 4.93e-05 ***
Decor         1.8473     0.2176   8.491 1.17e-14 ***
Service       0.1350     0.3957   0.341    0.733
```

To test whether the effect of service is zero if the effect of food and decor are not zero, we can use the *t*-statistic which is 0.341.

The p-value is 0.733 which is large and so the null hypothesis cannot be rejected, Service does not have an effect on Price if Food and Decor are not zero.

# Confidence interval for a regression effect

The $100\gamma\%$ confidence interval for the effect of the $j$-th variable, $\beta_j$, has the form

$$\left(\hat{\beta}_j - \text{s.e.}(\hat{\beta}_j) \times \text{t-point}\left(\frac{\gamma}{2}\right), \hat{\beta}_j + \text{s.e.}(\hat{\beta}_j) \times \text{t-point}\left(\frac{\gamma}{2}\right)\right)$$

To calculate the 95% confidence interval for the effect of service in the New York food prices example, we type

```
> confint(nyc_food, 'Service', level=0.95)
             2.5 %    97.5 %
Service -0.6461839 0.9162753
```

The 95% confidence interval is $(-0.647, 0.917)$.

## Example: New York food prices

```
Call:
lm(formula = Price ~ Food + Decor + Service, data = nyc)

Residuals:
     Min       1Q   Median       3Q      Max
-14.8440  -3.7039  -0.1525   3.6218  19.0576

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -24.6409     4.7536  -5.184 6.33e-07 ***
Food          1.5556     0.3731   4.170 4.93e-05 ***
Decor         1.8473     0.2176   8.491 1.17e-14 ***
Service       0.1350     0.3957   0.341    0.733
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 5.803 on 164 degrees of freedom
Multiple R-squared:  0.617,       Adjusted R-squared:   0.61
F-statistic: 88.06 on 3 and 164 DF,  p-value: < 2.2e-16
```

This $F$-statistic can be used to test the following hypotheses:

The null hypothesis is
$H_0 : \beta_1 = 0, \beta_2 = 0, \ldots, \beta_{K-1} = 0, \beta_K = 0.$

The alternative hypothesis is
$H_A$ : at least one $\beta_j \neq 0$.

*i.e.*, no effect vs some effect.

## Questions you might want to ask – some answers

How does the quality of food, décor and service affect the price of food in these restaurants?

Answer: The 95% confidence intervals are

| Variable | 95% CI | Effect |
|---|---|---|
| Food | (0.819, 2.292) | Price increases with food quality |
| Décor | (1.418, 2.278) | Price increases with décor quality |
| Service | $(-0.647, 0.917)$ | Service quality has no effect on price |

## Questions you might want to ask – some answers

- Could a restaurant charge a higher price for its food if the level of service increases?

  Answer: A $t$-test of the null hypothesis that the effect of service is equal is not significant. Therefore, there is no evidence that the level of service affects the price and so no evidence that a higher price could be charged if the level of service increases.

## Questions you might want to ask – some answers

- Which variables are useful for predicting the price of food in New York restaurants?

  Answer: We find evidence that Food and Décor are useful for predicting the price of food but no evidence that service effects the food of prices. This is supported by the result of a t-test for the effect of Service after including the effect of Food and Décor.

## Writing a report

Some points to bear in mind:

1. Consider the audience for your reports. Often reports should be readable by non-technical staff with a clear introduction and conclusions.
2. Your report should allow other analysts to reproduce your work and understand the decisions that you made (which variables are important, which transformations are used).
3. All graphs and tables should be carefully chosen (with a caption) and discussed in the text with clear references.

## Writing a report - structure

The report should include

1. An introduction – explaining the problem that you want to address, the data (including how it was collected/sourced), and any specific questions or aims that will be addressed during the analysis.
2. The analysis – The analysis will usually include: a basic description of the data (using graphs or tables), a clear description of the models fitted and discussion of any choices made.
3. A conclusion – explaining the main points from your analysis (for example, which variables are important, which variables have positive effects and which have negative effects), any limitations of your analysis and answers to any specific questions raised in the introduction.

## Building a model

- Which variables to include?
    1. Too few variables – some effects are excluded leading to poor predictions.
    2. Too many variables – effects are estimated with large standard errors leading to poor predictions
- Should some variables be transformed for the linear regression model?

## Building a model

1. Draw some graphs of the data.
   - Do these show any relationships between the variables (particularly, the explanatory variables and the response variable)?
2. Fit an initial model (often the model with all variables or the "current" model).
3. Does this model fit the data? Check residuals.
4. Are all variables important? Can some variables be removed?
   - Use $t$-value for effects and check the effect on the fit of the model (both $R^2$ and residuals).