Multiple Linear Regression

Multiple Linear Regression (MLR)

- We extend the Simple Linear Regression (SLR) model to accommodate multiple predictors.
- The resulting model is known as Multiple Linear Regression (MLR), which is given by

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \epsilon,$$

  where $X_1, X_2, \ldots, X_p$ are p predictors.
- We interpret β_j as the average effect on Y of a one unit increase in X_j, holding all other predictors fixed.
- In the advertising example, the MLR model becomes

$$sales = \beta_0 + \beta_1 \times TV + \beta_2 \times radio + \beta_3 \times newspaper + \epsilon.$$

Estimating the Regression Coefficients

- The parameters are estimated using the same least squares approach that we saw in the context of simple linear regression.
- We choose $\beta_0, \beta_1, \ldots, \beta_p$ to minimize the sum of squared residuals

$$RSS = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

$$= \sum_{i=1}^{n} \left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \hat{\beta}_2 x_{i2} - \cdots - \hat{\beta}_p x_{ip}\right)^2.$$

Regression Summary Table

	Coefficients	Std. error	t-statistic	p-value
Intercept	2.939	0.312	9.422	<0.0001
TV	0.046	0.001	32.809	<0.0001
Radio	0.189	0.009	21.893	<0.0001
Newspaper	-0.001	0.006	-0.177	0.860

Interpretation

- For a given amount of TV and newspaper advertising, spending an additional \$1,000 on radio advertising leads to an increase in sales by approximately 189 units.
- The MLR coefficient estimates for TV and radio are pretty similar to the SLR coefficient estimates.
- However, while the newspaper regression coefficient estimate in the SLR model was significantly non-zero, the coefficient estimate for newspaper in the MLR model is close to zero, with corresponding p-value around 0.86.

Interpretation

- In SLR, the slope term represents the average effect of a $\$1,000$ increase in newspaper advertising, ignoring other predictors such as TV and radio.
- In contrast, in the multiple regression setting, the coefficient for newspaper represents the average effect of increasing newspaper spending by $\$1,000$ while holding TV and radio fixed.

Correlation Matrix

	TV	radio	newspaper	sales
TV	1.000	0.055	0.057	0.782
radio		1.000	0.354	0.576
Newspaper			1.000	0.228
sales				1.000

Interpretation

- The correlation between radio and newspaper is 0.35.
- This suggests that there is a tendency to spend more on newspaper advertising in markets where more is spent on radio advertising.
- Suppose that the MLR model is correct and newspaper advertising has no direct impact on sales, but radio advertising does increase sales.
- Then in markets where we spend more on radio our sales will tend to be higher, and as our correlation matrix shows, we also spend more on newspaper advertising in those same markets.

Interpretation

- In a SLR model that only checks sales versus newspaper, higher values of newspaper tend to be linked with higher values of sales, even though newspaper advertising does not affect sales.
- So newspaper gets “credit” for the effect of radio on sales.

Some Important Questions

1. Is at least one of the predictors $X_1, \ldots, X_p$ useful in predicting the response?
2. How well does the model fit the data?
3. Given a set of predictor values, what response value should we predict, and how accurate is our prediction?

Is at least one of the predictors $X_1, \ldots, X_p$ useful in predicting the response?

- We test the null hypothesis

$$H_0\colon \beta_1 = \beta_2 = \cdots = \beta_p = 0$$

vs the alternative

$$H_a\colon at\ least\ one\ \beta_j\ is\ non-zero.$$

- We use the following F-statistic to test the above hypothesis:

$$F = \frac{(TSS - RSS)/p}{RSS/(n-p-1)},$$

which, under H_0, follows a F-distribution with d.f. p and $n-p-1$.

- The observed p-value is $P\left(F_{p.\,n-p-1} > F_{obs}\right)$. Reject H_0 if p-value is small.
- For the *Advertising* data set, the observed p-value is very low. Thus we reject H_0.

Model Fit

- The quality of a linear regression fit is assessed using following quantities:
  - R^2 statistic
  - Adjusted R^2 statistic
  - Residual Standard Error (RSE)

R^2 and Adjusted-R^2 Statistic

- In MLR, R^2 equals $Cor(Y, \hat{Y})^2$, i.e., the square of the correlation between the response and the fitted linear model.
- R^2 close to 1 indicates that the model explains the large portion of the variance in the response variable.
- However, R^2 always increases with the addition of every new variable.
- This is remedied using

$$\text{Adjusted} - R^2 = 1 - \frac{RSS/(n-p-1)}{TSS/(n-1)}.$$

- A model with more variables can have lower Adjusted-R^2.

Residual Standard Error (RSE)

- The RSE is defined as

$$RSE = \sqrt{\frac{1}{n-p-1} RSS}.$$

- A model with more variables can have higher RSE if the decrease in RSS is small relative to the increase in the number of variables (p).

Model Fit for *Advertising* Data Set

Model	**Predictors**	R^2	Adjusted $-R^2$	RSE
1	TV	0.61	0.61	3.26
2	Radio	0.33	0.33	4.28
3	Newspaper	0.05	0.05	5.09
4	TV & Radio	0.90	0.90	1.68
5	TV & Newspaper	0.65	0.64	3.12
6	Radio & Newspaper	0.33	0.33	4.28
7	TV, Radio & Newspaer	0.90	0.90	1.69

Prediction

- Given the coefficient estimates, the predicted response is given by
$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \cdots + \hat{\beta}_p x_p.$$
- In order to assess the uncertainty associated with the predicted response, consider the following two cases:
  - *How should we quantify the uncertainty associated with the average sales over a number of cities, given that $100,000 is spent on TV advertising and $20,000 is spent on radio advertising in each city?*
  - *How should we quantify the uncertainty associated with the sales of a particular city, given that $100,000 is spent on TV advertising and $20,000 is spent on radio advertising in that city?*

Prediction

- Use a *confidence interval* in the first case.
- Given that \$100,000 is spent on TV advertising and \$20,000 is spent on radio advertising in each city, the 95% confidence interval is [11258, 11563].
- Thus 95% of the intervals of this form will contain the true value of the average sales.
- *To elaborate, if we collect a large number of data sets (perhaps hypothetical) like the Advertising data sets and we construct a confidence interval for the average sales in each case (given 100,000 on TV advertising and \$20,000 on radio advertising), then 95% of these intervals will contain the true value of the average sales.*

Prediction

- Use a *prediction interval* in the second case.
- Given that \$100,000 is spent on TV advertising and \$20,000 is spent on radio advertising in a particular city, the 95% prediction interval is [9544, 13278].
- Thus 95% of the intervals of this form will contain the true value of the sales of that city.
- It is obvious that the prediction interval is substantially wider than the confidence interval.

Extension of the Linear Model

- Two important assumptions in the MLR models considered so far are
  - ✓ The relationship between the predictors and the response are additive.
  - ✓ The relationship between the predictors and the response are linear.
- The additive assumption means that the effect of changes in a particular predictor, say X_j, on the response Y is independent of the values of the other predictors.
- The linear assumption means 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.

Removing the Additive Assumption

- For *Advertising* data set, we saw that both TV and Radio are associated with the Sales.
- The linear model considered previously says 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 assumption may be incorrect!
- Suppose that spending money on Radio advertising increases the effectiveness of TV advertising.
- This clearly suggests that the slope term for TV should increase with Radio.

Removing the Additive Assumption

- Given a fixed budget of $100,000, spending half on Radio and half on TV may result in higher Sales as compared to allocating the entire amount to either TV or Radio.
- In Marketing, this is known as a synergy effect and in statistics, it is referred to as an interaction effect.

Modelling Interaction

- Now we model the synergy (or interaction) effect among the advertising media.
- This is done by adding an additional interaction term in the regression model.
- This results a non-additive model, which is given by

$$sales = \beta_0 + \beta_1 \times TV + \beta_2 \times Radio + \beta_3 \times (TV \times Radio) + \epsilon.$$

- To check whether there is synergy among different media, we check the following hypothesis:

$$H_0: \beta_3 = 0 \text{ vs } H_0: \beta_3 \neq 0.$$

Modelling Interaction

- Note that the interaction model, given in the previous slide, can be rewritten as

$$sales = \beta_0 + (\beta_1 + \beta_3 \times Radio) \times TV + \beta_2 \times Radio + \epsilon.$$

- 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).
- The next table provides the summary of the regression analysis for this model.

Regression Table Summary

	Coefficients	Std. Error	t-statistic	p-value
Intercept	6.750	2.479×10^{-1}	27.233	<0.0001
TV	0.019	1.504×10^{-3}	12.699	<0.0001
radio	0.029	8.905×10^{-3}	3.241	0.0014
$\mathrm{TV} \times \mathrm{radio}$	0.001	5.242×10^{-5}	20.727	<0.0001

Model Fit Summary

Quantity	Value
R^2	0.9678
$\text{Adjusted} - R^2$	0.9673
RSE	0.9435

Interpretation

- The p-value for the interaction term, $TV \times Radio$, is extremely low.
- This clearly suggests that the true relationship is not additive.
- Looking at the coefficient estimates, we observe that an increase in TV advertising of \$1,000 is associated with increased sales of
$$\left(\hat{\beta}_1 + \hat{\beta}_3 \times Radio\right) \times 1{,}000 = 19 + 1.1 \times Radio$$
units.
- Similarly, an increase in Radio advertising of \$1,000 is associated with increased sales of $29 + 1.1 \times TV$ units.

Interpretation

- We also observe that the R^2 for this interaction model is 96.8%.
- Note that R^2 for the additive model with TV and Radio as predictors was 89.7%.
- This means that $(96.8 - 89.7)/(100 - 89.7) = 69\%$ of the variability in Sales that remains after fitting the additive model has been explained by the interaction term.

Non-linear Relationship

- The additive MLR model with TV and Radio as predictors assumed a linear relationship between sales and the predictors.
- In some cases, the true relationship between the response and the predictors may be non-linear.
- Here we extend the additive MLR model to incorporate the non-linear relationships in a linear model by including transformed predictors in the model.

Non-linear Relationship

- Here we consider the following model:

$$sales = \beta_0 + \beta_1 \times TV + \beta_2 \times Radio + \beta_3 \times TV^2 + \beta_4 \times (TV \times Radio) + \epsilon.$$

- Note that the above model involves the non-linear function of TV.
- *But this model is still a linear model.*

Regression Table Summary

	Coefficients	Std. Error	t-statistic	p-value
Intercept	5.1371	1.927×10^{-1}	26.663	<0.0001
TV	0.0509	2.232×10^{-3}	22.810	<0.0001
Radio	0.0351	5.901×10^{-3}	5.959	<0.0001
TV^2	-0.0001	6.893×10^{-6}	-15.920	<0.0001
$\mathrm{TV} \times \mathrm{Radio}$	0.0011	3.466×10^{-5}	31.061	<0.0001

Model Fit Summary

Quantity	Value
R^2	0.986
$\text{Adjusted} - R^2$	0.986
RSE	0.624

Interpretation

- The p-value for the TV^2 term is extremely low.
- This clearly suggests that there is a non-linear relationship between Sales and TV.
- Also note that R^2 for this model is 98.6%.
- This means that $(98.6 - 96.8)/(100 - 96.8) = 56.25\%$ of the variability in Sales that remains after fitting the non-additive interaction model has been explained by the inclusion of the TV^2 term in the model.

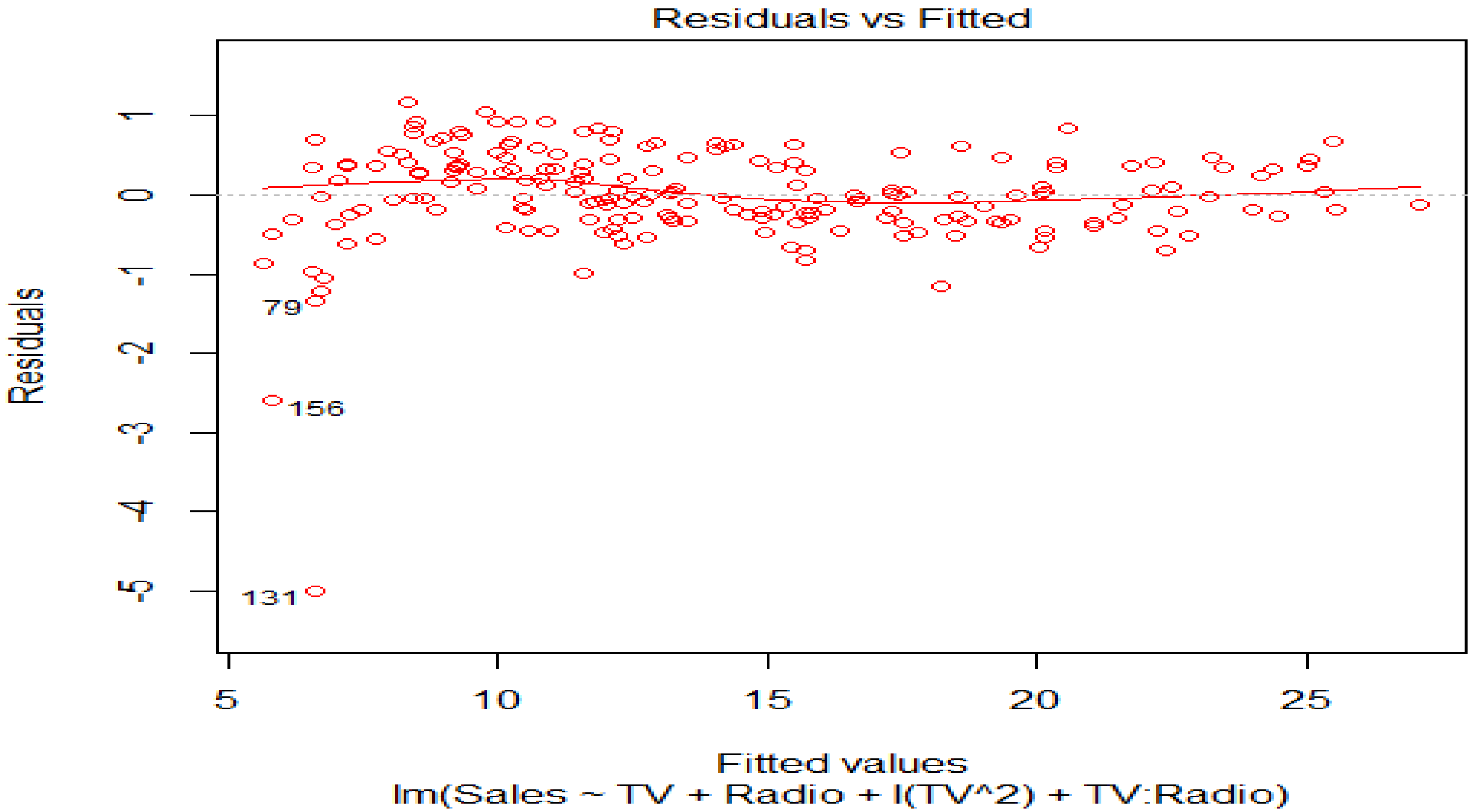
Some Potential Problems

- Non-linearity of the response-predictor relationship
- Non-constant variance of error terms
- Collinearity
- Non-Normality

Non-linearity of the Response-Predictor Relationship

- The additive MLR model considered at the beginning assumes a straight-linear relationship between the predictors and the response.
- If the true relationship is far from linear, then all the conclusions are suspect.
- Also the prediction accuracy of the model can be significantly reduced.
- We can use Residual vs fitted plot to detect non-linear relationship.
- If the non-linear relationship is present, a simple approach can be to use non-linear transformations such as $\log X$, $\sqrt{X}$, or X^2.

Residual vs Fitted Plot



Non-constant variance of error terms

- An important assumption in the MLR model is that the error terms have constant variance.
- Unfortunately the variance of the error terms are often non-constant.
- One can identify the non-constant variance or heteroscedasticity from the presence of a funnel shape in the residual vs fitted plot.
- One way to resolve this issue is to transform the response Y such as $\log Y$ or $\sqrt{Y}$.

Collinearity

- Collinearity refers to the situation in which two or more predictors are closely related to each other.
- A simple way to detect collinearity is to look at the correlation matrix of the predictors.
- Unfortunately, the collinearity problems may not be always detected by looking at the correlation matrix .
- Sometimes collinearity exists among three or more variables even if no pair of variables has a particularly high correlation.
- This situation is called *multicollinearity*.

Collinearity

- *A better way to assess the multicollinearity is to compute the variance inflation factor (VIF).*
- The VIF is defined as

$$VIF\left(\hat{\beta}_j\right) = \frac{1}{1 - R^2_{X_j|X_{-j}}},$$

where $R^2_{X_j|X_{-j}}$ is the R^2 from a regression of X_j onto all of the other predictors.

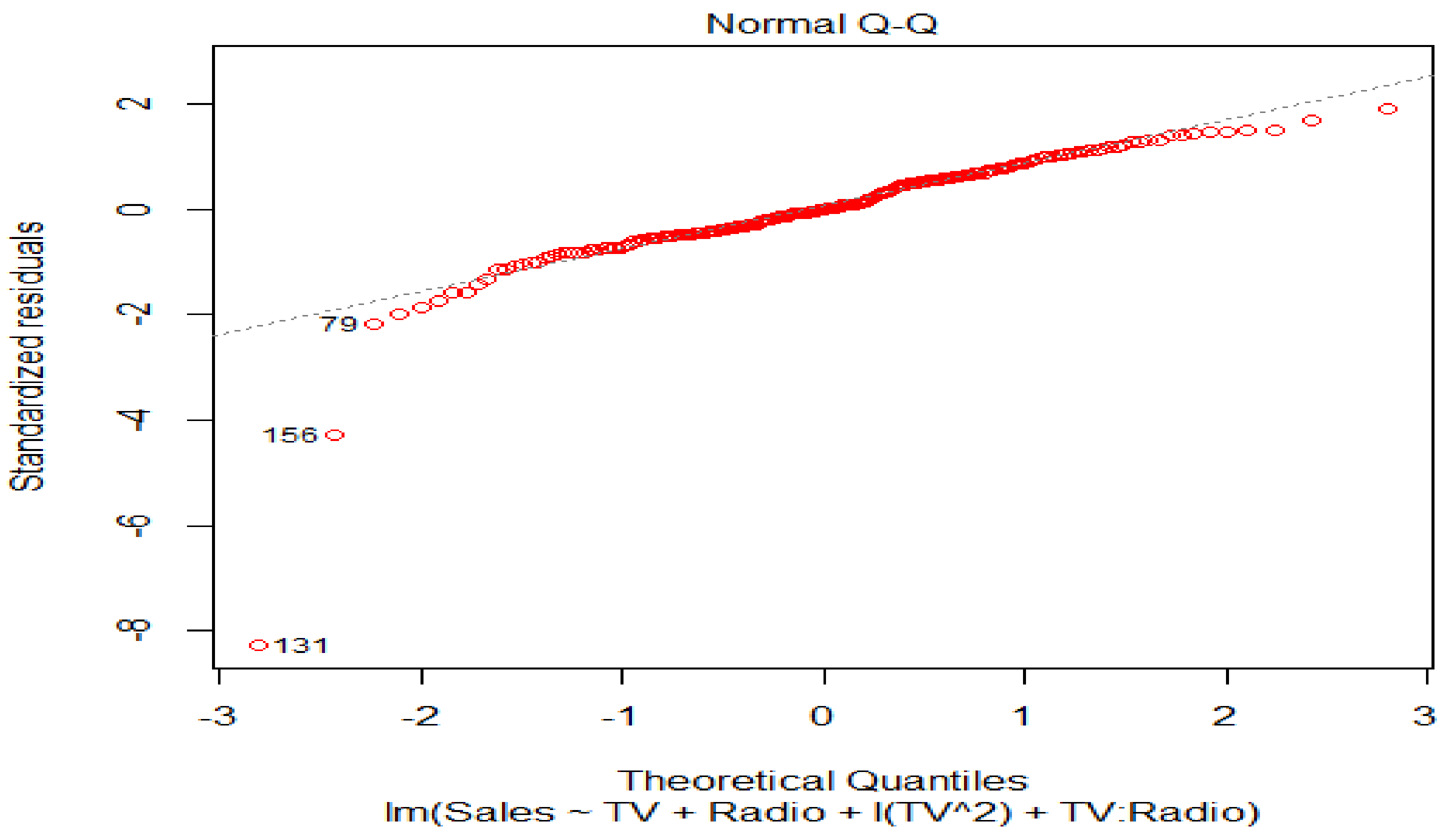
- As a rule of thumb, a VIF value that exceeds 5 or 10 indicates a problematic amount of collinearity.

Collinearity

TV	Radio
1.003	1.003

VIFs are low for Advertising data set, indicating that the collinearity is not significant in Advertising data set.

Non-Normality



Model Fit Summary for *Advertising* Data Set

Model	Predictors	R^2	$\text{Adjusted} - R^2$	RSE
1	TV	0.61	0.61	3.26
2	Radio	0.33	0.33	4.28
3	Newspaper	0.05	0.05	5.09
4	TV & Radio	0.90	0.90	1.68
5	TV & Newspaper	0.65	0.64	3.12
6	Radio & Newspaper	0.33	0.33	4.28
7	TV, Radio & Newspaer	0.90	0.90	1.69
8	$\text{TV, Radio, TV} \times \text{Radio}$	0.97	0.97	0.94
9	$\text{TV, Radio, TV}^2\text{, TV} \times \text{Radio}$	0.99	0.99	0.62

Marketing Plan

- *Is there a relationship between advertising sales and budget?*
  - We fitted a regression model with sales as response variable and TV, radio, and newspaper as predictors.
  - We use the F-statistic to determine whether or not we should reject H_0.
  - The observed p-value is very low, indicating that there is a *relationship between advertising sales and budget.*

Marketing Plan

- *How strong is the relationship?*
  - *The observed value for RSE is 620 units, while the mean value for the response is 14,022 (Model 9).*
  - *This indicates a percentage error of roughly 4.42%.*
  - *The observed value for R^2 statistic is around 99%.*
  - *This indicates that the predictors explain almost 99% of the variance in $sales$.*
  - *This indicates that there is a strong relationship between advertising sales and budget.*

Marketing Plan

- *Which media contribute to sales?*
    - ➢ *From Model 7, we observe that the p-values for TV and radio are very low, but the p-value for newspaper is not.*
    - ➢ *This suggests that only TV and radio are related to sales.*

Marketing Plan

- How large is the effect of each medium on sales?
  - ➢ We observed that both TV and radio are related to sales, but newspaper is not statistically significant.
  - ➢ To assess the individual effect of each medium on sales, we have performed three separate linear regressions.
  - ➢ We observed a very strong association between TV and sales, and between radio and sales.
  - ➢ However, the association between newspaper and sales is a mild one.

Marketing Plan

- How accurately can we predict future sales?
  - The response can be predicted using the formula $\widehat{sales} = \hat{\beta}_0 + \hat{\beta}_1 \times TV + \hat{\beta}_2 \times Radio + \hat{\beta}_3 \times (TV \times Radio).$
  - The accuracy associated with this estimate depends on whether we are interested in individual sales for a particular city or the average sales of a number of cities.
  - In the former case, we use a prediction interval, and in the latter case, we use a confidence interval.
  - Prediction intervals are always wider than the confidence interval.

Marketing Plan

- *Is the relationship linear?*
  - ➢ *Residual plots can be used in order to identify non-linearity.*
  - ➢ *If the relationships are linear, then the residual plots should display no pattern.*
  - ➢ *In the case of Advertising data, the non-linear relationship can be identified in a residual plot.*
  - ➢ *This non-linear relationship is accommodated here with the inclusion of interaction term and the transformed predictors (Model 9).*

Marketing Plan

- *Is there synergy among the advertising media?*
  - For the Advertising data set, we have included an interaction term in the MLR model (Model 8) to accommodate the non-additive relationship.
  - The corresponding low p-value confirms the presence of such relationship.
  - Inclusion of an interaction term in the MLR model increases R^2 from 90% to 97%.
  - This clearly indicates the synergy between two advertising media- *TV and radio*.