

# LINEAR REGRESSION

Esam Mahdi

Data Analytics (ECMP 5005)  
School of Mathematics and Statistics  
Master of Engineering - Engineering Practice  
Carleton University

September 22, 2023

By the end of this chapter, you should be able to do the following:

- ① Use simple and multiple linear regression models to estimate the mean response, make predictions, and interpret the results.
- ② Perform analysis of variance and statistically evaluate the utility of the estimated linear model.
- ③ Test hypotheses and construct confidence intervals on the regression coefficients.
- ④ Use the appropriate interaction, reduced, and full models.
- ⑤ Explore diagnostic methods to check the validity assumptions of linear regression models:
  - Homogeneity variance (homoscedasticity): the variance of the errors is constant.
  - Linearity: relationships between predictors and response variables are linear.
  - Independence: the errors are not correlated.
  - Normality: the errors are normally distributed.
  - Model specification: include only significant and relevant variables and exclude insignificant variables.
- ⑥ Use R with some real-life applications.

# SIMPLE LINEAR REGRESSION

- $X$  is the *feature*, or *input*, or, *independent*, or *predictor* variable - can be quantitative (numeric) or qualitative (category) variable.
- $Y$  is the *target*, or *dependent*, or *response* variable - must be quantitative variable.
- *Simple linear regression* assumes that  $Y$  is linearly depends on  $X$  so that

$$Y = f(X) + \varepsilon = \beta_0 + \beta_1 X + \varepsilon, \quad (1)$$

where  $\beta_0$  and  $\beta_1$  are two unknown constants that represent the *intercept* and *slope*, also known as *coefficients* or *parameters*, and  $\varepsilon$  captures measurement errors and other discrepancies.

- It is assumed that  $\varepsilon_i \stackrel{\sim}{\text{iid}} \mathcal{N}(0, \sigma^2)$  which is also independent from  $x_i$  for all  $i$ .
- *iid* stands for *independent and identically distributed* and  $\mathcal{N}$  for *normal distribution*. Thus,  $\varepsilon_i$  are independent and identically normally distributed with mean zero and variance  $\sigma^2$ .
- Note that  $f(x) = E(Y|X = x)$  means *expected value (average)* of  $Y$  given  $X = x$  which is called the *regression function*.

# ESTIMATION OF THE PARAMETERS BY LEAST SQUARES

- From the sample values  $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$  of  $(X, Y)$ , we estimate  $\beta_0$  by  $\hat{\beta}_0$  and  $\beta_1$  by  $\hat{\beta}_1$ ;
- The prediction for  $Y$  based on the  $i$ th value of  $X$  is given by

$$\hat{y}_i = E(Y|X = x_i) = \hat{\beta}_0 + \hat{\beta}_1 x_i \quad (2)$$

- Let  $e_i = \hat{\varepsilon}_i = y_i - \hat{y}_i$  denotes the  $i$ th *residual* and define the *residual sum of squares (RSS)* as

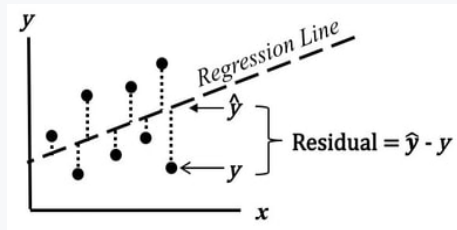
$$\text{RSS} = e_1^2 + e_2^2 + \dots + e_n^2 = \sum_{i=1}^n (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2;$$

- The *least squares method* chooses  $\hat{\beta}_0$  and  $\hat{\beta}_1$  to minimize the RSS as follows:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2},$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x},$$

where  $\bar{y} \equiv \frac{1}{n} \sum_{i=1}^n y_i$  and  $\bar{x} \equiv \frac{1}{n} \sum_{i=1}^n x_i$  are the sample means.



Note that the  $\sum_{i=1}^n e_i = 0$ , so that  $\bar{e} = 0$ .

- 1 **Linearity:** There exists a linear relationship between the predictor and the response variables.
  - Note that the **linear model** is a function that is linear in the parameters  $\beta_j$ , so the polynomial model, for example  $Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \varepsilon$  is quadratic as a function of  $X$  but linear in the coefficients  $\beta_0, \beta_1$ , and  $\beta_2$ .
  - $Y = \frac{\beta_0 X}{\beta_1 + X(1 + \varepsilon)}$  and  $Y = \beta_0 e^{\beta_1 X} \varepsilon$  are two examples of **nonlinear models**.
- 2 **Independency:**
  - The error terms are independent of each other.
  - The error terms are independent from the independent variables.
- 3 **Normality:** The error terms are normally distributed with mean equal to zero [i.e.,  $E(\varepsilon) = 0$ ] and common variance equal to  $\sigma^2$  [i.e.,  $\text{Var}(\varepsilon) = \sigma^2$ ].
- 4 **Homoscedasticity:** The variance of error terms  $\sigma^2$  are similar across the values of the independent variables.

Note that assumptions (2-4) means that  $\varepsilon_i \stackrel{iid}{\sim} \mathcal{N}(0, \sigma^2)$ , for  $i = 1, \dots, n$ .

- The standard error of an estimator reflects how it varies under repeated sampling. We have

$$\text{SE}(\hat{\beta}_1) = \sqrt{\frac{\sigma^2}{\sum_{i=1}^n (x_i - \bar{x})^2}}, \quad \text{SE}(\hat{\beta}_0) = \sqrt{\sigma^2 \left[ \frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^n (x_i - \bar{x})^2} \right]},$$

where  $\sigma^2 = \text{Var}(\varepsilon)$ , which can be estimated by  $\hat{\sigma}^2 = s^2 = \text{RSS}/(n - 2)$ .

- These standard errors can be used to compute *confidence intervals*. A 95% confidence interval is defined as a range of values such that with 95% probability, the range will contain the true unknown value of the parameter. It has the form

$$\hat{\beta}_i \pm 2 \cdot \text{SE}(\hat{\beta}_i), \quad i = 0, 1.$$

That is, there is approximately a 95% chance that the interval

$$\left[ \hat{\beta}_i - 2 \cdot \text{SE}(\hat{\beta}_i), \hat{\beta}_i + 2 \cdot \text{SE}(\hat{\beta}_i) \right]$$

will contain the true value of  $\beta_i$ , for  $i = 0, 1$ .

- Standard errors can also be used to perform *hypothesis* tests on the coefficients. The most common hypothesis test involves testing the *null hypothesis* of

$H_0$  : There is no relationship between  $X$  and  $Y$ ,

versus the *alternative hypothesis*

$H_A$  : There is some relationship between  $X$  and  $Y$ .

- Alternatively, we are testing

$$H_0 : \beta_1 = 0$$

versus

$$H_A : \beta_1 \neq 0,$$

since if  $\beta_1 = 0$  then the model reduces to  $Y = \beta_0 + \varepsilon$ , and  $X$  is not associated with  $Y$ .

- To test the null hypothesis, we compute a *t-statistic*, given by

$$t = \frac{\hat{\beta}_1 - 0}{SE(\hat{\beta}_1)} \sim t_{df=n-2}.$$

- Using statistical software, one can reject the null hypothesis, at a 5% level of significance, if the probability value (*p-value*) is less than or equal to 0.05.

# MULTIPLE LINEAR REGRESSION MODEL

The *multiple linear regression model* relating the *dependent variable (response)*  $Y$  to  $p$  *independent variables (predictors)*  $x_1, x_2, \dots, x_p$  is

$$Y = \underbrace{\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_p X_p}_{\mu_{Y|X_1, X_2, \dots, X_p}} + \varepsilon, \quad (3)$$

where

- The parameters  $\beta_j$ ,  $j = 0, 1, \dots, p$ , are unknown constants, called the *regression coefficients* ( $\beta_0$  denotes the intercept).
- $\varepsilon$  is an *error term* that describes the effects on  $Y$  of all factors other than the independent variables  $X_1, X_2, \dots, X_p$  (all unexplained variations in  $Y$ ).
- $X_1, X_2, \dots, X_p$  are independent predictor variables, measured without error, which may represent higher-order terms for quantitative predictors (e.g.,  $X_2 = X_1^2$ ) or terms for qualitative (categorical) predictors.
- Let  $x_{ij}$  denote the  $i$ th observation of variable  $X_j$ , and  $y_i$  denote the  $i$ th observation of variable  $Y$ , where  $i = 1, \dots, n$  and  $j = 1, \dots, p$ , then the *estimation/prediction equation* is

$$E(Y|X_1 = x_1, \dots, X_p = x_p) = \hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \hat{\beta}_2 x_{i2} + \dots + \hat{\beta}_p x_{ip} \quad (4)$$



# LEAST SQUARES ESTIMATION IN MATRIX NOTATION

The multiple regression model in (3) can be written in matrix notation as

$$\mathbf{Y} = \mathbf{X}\beta + \varepsilon, \quad \varepsilon \sim \mathcal{N}(\mathbf{0}, \sigma^2 I_{n \times n}), \quad (5)$$

or

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1p} \\ 1 & x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{np} \end{bmatrix} \times \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

The estimated values for  $\beta$  will be called  $\hat{\beta}$  and the best fitted model is

$$\hat{\mathbf{Y}} = \mathbf{X}\hat{\beta}, \quad \text{where } \mathbf{X} \text{ is called } \textit{design matrix}.$$

Mathematically, the vector  $\hat{\beta} = [\hat{\beta}_0, \hat{\beta}_1, \dots, \hat{\beta}_p]'$  is obtained by minimizing the *residual sum of squares*

$$\text{RSS} = \sum_{i=1}^n (y_i - \hat{y}_i)^2 = \sum_{i=1}^n e_i^2 = \mathbf{e}'\mathbf{e} = [e_1, e_2, \dots, e_n] \begin{bmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{bmatrix},$$

where  $\mathbf{e} = \mathbf{Y} - \hat{\mathbf{Y}} = \mathbf{Y} - \mathbf{X}\hat{\beta}$  are the *residuals* (estimate the unobserved errors  $\varepsilon = (\varepsilon_1, \dots, \varepsilon_n)'$ ).

The least squares estimates of  $\beta$  are given by

$$\hat{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$$

# SOME IMPORTANT QUESTIONS

- ① *Is at least one of the predictors  $X_1, X_2, \dots, X_p$  useful in predicting the response?*
- ② *Do all the predictors help to explain  $Y$ , or is only a subset of the predictors useful?*
- ③ *How well does the model fit our data? (Does our data violate the linear regression assumptions?)*
- ④ *Given a set of predictor values, what response value should we predict, and how accurate is our prediction?*

# THE ANALYSIS OF VARIANCE (ANOVA)

To answer these questions, we can use the *analysis of variance (ANOVA)* based on the following estimators:

- Total sum of squares (*Total variation*) is given by  $TSS = \sum_{i=1}^n (y_i - \bar{y})^2$ .
- Sum of squares for regression (*Explained variation*) is given by  $SS_{reg} = \sum_{i=1}^n (\hat{y}_i - \bar{y})^2$ .
- Sum of squared errors (*Unexplained variation*) is given by  $RSS = \sum_{i=1}^n (y_i - \hat{y}_i)^2$ .
- Total variation is the sum of explained and unexplained variation. That is

$$TSS = SS_{reg} + RSS.$$

- *Mean squares errors (MSE)* is the point estimate of the variance of the error term  $\sigma^2$ :

$$s^2 = MSE = \frac{RSS}{n - (p + 1)}.$$

- *Residual standard error (RSM)* is the point estimate of the standard deviation of the error term  $\sigma$ :

$$s = RSM = \sqrt{MSE} = \sqrt{\frac{RSS}{n - p - 1}}.$$

# ASSESSING THE OVERALL ACCURACY OF THE MODEL

For the first question (to test whether the model is adequate for predicting  $Y$ ) we use the global  $F$ -test based on the **analysis of variance (ANOVA)** to test the following hypotheses:

$$H_0 : \beta_1 = \beta_2 = \dots = \beta_p = 0 \text{ (Overall model is not useful for predicting } Y \text{)}$$

$$H_A : \text{at least one of } \beta_1, \beta_2, \dots, \beta_p \neq 0, \text{ (At least one model term is useful for predicting } Y \text{)}$$


The test statistic is

$$F = \frac{\text{Explained variation}/p}{\text{Unexplained variation}/(n - p - 1)} = \frac{(TSS - RSS)/p}{RSS/(n - p - 1)}$$

Reject  $H_0$  at level of significance  $\alpha$  (usually 0.05) in favor of  $H_A$  if:

- $F \geq F_\alpha$  or p-value  $\leq \alpha$ , where  $F_\alpha$  is based on  $p$  numerator and  $n - (p + 1)$  denominator degrees of freedom (df).

**TABLE:** The analysis of variance (ANOVA)

Source	SS	df	MS	F	Rejection region
Regression	$SS_{reg}$	$p$	$MSR = \frac{SS_{reg}}{p}$	$\frac{MSR}{MSE} = \frac{SS_{reg}/p}{RSS/(n-p-1)}$	
Residuals	$RSS$	$n - p - 1$	$MSE = \frac{RSS}{n-p-1}$		
Total	$TSS$	$n - 1$			

## DEFINITION

The **multiple coefficient of determination**,  $R^2$ , is the ratio of explained variation to total variation. That is,

$$R^2 = \frac{SS_{reg}}{TSS} = 1 - \frac{RSS}{TSS}, \quad 0 \leq R^2 \leq 1.$$

- $R^2$  measures the proportion of the total variation in the response variable that can be explained by using the independent variables in the model. For example,  $R^2 = 0.6$  means that 60% of the variability in  $Y$  can be explained by the regression model.
- The higher the proportion of variation that is explained by the model, the better your predictions will be.
- $R^2 = 0$  implies a complete lack of fit of the model to the data, and  $R^2 = 1$  implies a perfect fit, with the model passing through every data point.
- Adding an independent variable to multiple regression will raise  $R^2$  (even if this variable does not relate to  $Y$ ). Thus, in practice always use **the adjusted value of  $R^2$**  that is given by

$$R_{adj}^2 = 1 - \frac{RSS/(n-p-1)}{TSS/(n-1)}.$$

- Note that  $R_{adj}^2 \leq R^2$  and, for poor-fitting models  $R^2$  may be negative.

# INFERENCES ABOUT THE INDIVIDUAL $\beta$ PARAMETERS

## ONE-TAILED TESTS

## TWO-TAILED TEST

$$H_0 : \beta_i = 0 \quad H_0 : \beta_i = 0 \quad H_0 : \beta_i = 0$$

$$H_A : \beta_i < 0 \quad H_A : \beta_i > 0 \quad H_A : \beta_i \neq 0$$

$$\text{Test statistic:} \quad T = \frac{\hat{\beta}_i}{s(\hat{\beta}_i)}$$

$$\text{Rejection region:} \quad T < -t_\alpha \quad T > t_\alpha \quad |T| > t_{\alpha/2}$$

where  $t_\alpha$  and  $t_{\alpha/2}$  are based on  $n - (p + 1)$  degrees of freedom and

$n$  = Number of observations,

$p + 1$  = Number of  $\beta$  parameters in the model.

Note: Most statistical software programs report two-tailed  $p$ -values on their output.

To find the appropriate  $p$ -value for a one-tailed test, make the following adjustment to  $\mathcal{P}$  = two-tailed  $p$ -value:

$$\text{For } H_A : \beta_i > 0, \quad p\text{-value} = \begin{cases} \mathcal{P}/2 & \text{if } T > 0 \\ 1 - \mathcal{P}/2 & \text{if } T < 0 \end{cases}$$

$$\text{For } H_A : \beta_i < 0, \quad p\text{-value} = \begin{cases} 1 - \mathcal{P}/2 & \text{if } T > 0 \\ \mathcal{P}/2 & \text{if } T < 0 \end{cases}$$

A  $(1 - \alpha)100\%$  **confidence interval** for a regression coefficient  $\beta_i, i = 0, 1, \dots, p$  is given by

$$\hat{\beta}_i \pm t_{\alpha/2}(n - p - 1)SE(\hat{\beta}_i) \quad (6)$$

## TESTING WHETHER A SPECIFIED SUBSET OF THE PREDICTORS HAVE REGRESSION COEFFICIENTS EQUAL TO ZERO

Suppose that we are interested in testing

$$H_0 : \beta_1 = \beta_2 = \dots = \beta_k = 0, \text{ where } k < p$$

$$\text{i.e., } Y = \beta_0 + \beta_{k+1}x_{k+1} + \dots + \beta_p x_p + \varepsilon \text{ (reduced model)}$$

against  $H_A : H_0$  is not true

$$\text{i.e., } Y = \beta_0 + \beta_1 x_1 + \dots + \beta_k x_k + \beta_{k+1} x_{k+1} + \dots + \beta_p x_p + \varepsilon \text{ (full model)}.$$

This can be achieved using an F-test. Let  $RSS(\text{Full})$  be the residual sum of squares under the full model (i.e., the model which includes all the predictors, i.e.,  $H_A$ ) and  $RSS(\text{Reduced})$  be the residual sum of squares under the reduced model (i.e., the model which includes only the predictors thought to be non-zero, i.e.,  $H_0$ ). Then the F-statistic is given by

$$\begin{aligned} F &= \frac{(RSS(\text{reduced}) - RSS(\text{full})) / (df_{\text{reduced}} - df_{\text{full}})}{RSS(\text{full}) / df_{\text{full}}} \\ &= \frac{(RSS(\text{reduced}) - RSS(\text{full})) / k}{RSS(\text{full}) / (n - p - 1)} \end{aligned}$$

since the reduced model has  $p + 1 - k$  predictors and

$$[n - (p + 1 - k)] - [n - (p + 1)] = k.$$

This is called a *partial F-test*.

## MENU PRICING IN A NEW ITALIAN RESTAURANT IN NEW YORK CITY

Consider the dataset that is available in the file `nyc.csv` which can be found from the website of the book *"A Modern Approach to Regression with R"* written by Simon Sheather

[https://gattoweb.uky.edu/sheather/book/data\\_sets.php](https://gattoweb.uky.edu/sheather/book/data_sets.php). The data are in the form of the average of customer views on

- $Y = \text{Price}$  = the price (in \$US) of dinner (including one drink & a tip).
  - $X_1 = \text{Food}$  = customer rating of the food (out of 30).
  - $X_2 = \text{Décor}$  = customer rating of the decor (out of 30).
  - $X_3 = \text{Service}$  = customer rating of the service (out of 30).
  - $D = \text{East}$  = dummy variable = 1 (or 0) if the restaurant is east (or west) of Fifth Avenue.
- 1 Develop a regression model that directly predicts the price of dinner (in dollars) using a subset or all of the 4 potential predictor variables listed above.
  - 2 Determine which of the predictor variables Food, Décor and Service has the largest estimated effect on Price? Is this effect also the most statistically significant?
  - 3 If the aim is to choose the location of the restaurant so that the price achieved for dinner is maximized, should the new restaurant be on the east or west of Fifth Avenue?
  - 4 Does it seem possible to achieve a price premium for “setting a new standard for high-quality service in Manhattan” for Italian restaurants?



# R-CODE: MENU PRICING IN A NEW ITALIAN RESTAURANT IN NEW YORK CITY

We use the built in R function `lm()` to estimate the linear model  $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 D + \varepsilon$

```
nyc <- read.csv("https://gattoweb.uky.edu/sheather/book/docs/datasets/nyc.csv", header=TRUE)
attach(nyc)
m1 <- lm(Price ~ Food + Decor + Service + East)
summary(m1)

##
## Call:
## lm(formula = Price ~ Food + Decor + Service + East)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -14.0465  -3.8837   0.0373   3.3942  17.7491
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -24.023800   4.708359  -5.102 9.24e-07 ***
## Food         1.538120   0.368951   4.169 4.96e-05 ***
## Decor        1.910087   0.217005   8.802 1.87e-15 ***
## Service     -0.002727   0.396232  -0.007  0.9945
## East         2.068050   0.946739   2.184  0.0304 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.738 on 163 degrees of freedom
## Multiple R-squared:  0.6279, Adjusted R-squared:  0.6187
## F-statistic: 68.76 on 4 and 163 DF,  p-value: < 2.2e-16
```

## MENU PRICING IN A NEW ITALIAN RESTAURANT IN NEW YORK CITY

- 1 The initial regression model including all predictors is

$$\text{Price} = -24.02 + 1.54\text{Food} + 1.91\text{Décor} - 0.003\text{Service} + 2.07\text{East}$$

- 2 Décor has the largest effect on Price since its regression coefficient is largest. Also it is the most statistically significant since its  $p$ -value is the smallest of the three.  
*Be careful! in general we can't compare the regression coefficients of the variable, but in this example we can! Why?*
- 3 In order that the price achieved for dinner is maximized, the new restaurant should be on the east of Fifth Avenue since the coefficient of the dummy variable is statistically significantly larger than 0.
- 4 It does not seem possible to achieve a price premium for "setting a new standard for high quality service in Manhattan" for Italian restaurants since the regression coefficient of Service is not statistically significantly greater than zero.

## R-CODE: MENU PRICING AFTER REMOVING THE VARIABLE SERVICE

```
m2 <- update(m1, ~.-Service)
summary(m2)
```

	Estimate	Std. Error	t value	Pr(> t )
<b>Intercept</b>	-24.0269	4.6727	-5.14	< 0.0001
<b>Food</b>	1.5363	0.2632	5.84	< 0.0001
<b>Decor</b>	1.9094	0.1900	10.05	< 0.0001
<b>East</b>	2.0670	0.9318	2.22	0.0279

The final regression model is

$$\text{Price} = -24.03 + 1.54 \text{ Food} + 1.91 \text{ Décor} + 2.07 \text{ East}$$

Comparing the last two sets of output from R, we see that the regression coefficients for the variables in both models are very similar. This **does not** always occur.

# A POLYNOMIAL MODEL WITH A QUANTITATIVE PREDICTOR

A response  $Y$  is related to a single independent variable  $X$ , but not in a linear manner. The *polynomial model of order  $p$*  is:

$$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \cdots + \beta_p X^p + \varepsilon.$$

Here, we consider the second-order model ( $p = 2$ ). This model is called the *quadratic model* which is given by

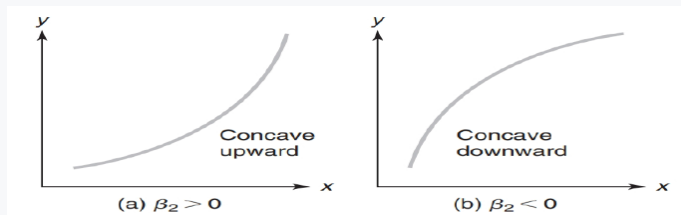
$$E(Y|X) = \beta_0 + \beta_1 X + \beta_2 X^2,$$

where

$\beta_0$  is the  $Y$ -intercept of the curve

$\beta_1$  is a shift parameter

$\beta_2$  is the rate of curvature



## AUTO DATA IN ISLR2 PACKAGE: POLYNOMIAL REGRESSION

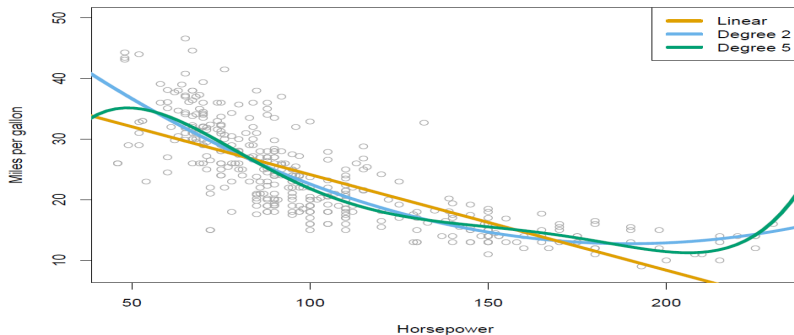


FIGURE: polynomial regression on Auto data

The figure suggests that

$$\text{mpg} = \beta_0 + \beta_1 \times \text{horsepower} + \beta_2 \times \text{horsepower}^2 + \varepsilon$$

might be suitable model to predict the miles per gallon based on a quadratic form of the engine horsepower.

## AUTO DATA: POLYNOMIAL REGRESSION

```
require("ISLR2")  
poly_fit <- lm(mpg ~ horsepower + I(horsepower^2), data = Auto)  
summary(poly_fit)
```

Results:

	Estimate	Std. Error	t value	Pr(> t )
Intercept	56.9001	1.8004	31.60	< 0.0001
horsepower	-0.4662	0.0311	-14.98	< 0.0001
horsepower <sup>2</sup>	0.0012	0.0001	10.08	< 0.0001

# INTERACTIONS MODELS

- The *first-order model* (in polynomial model) includes only terms for quantitative variables that are not functions of other independent variables (*no interaction*).
- The *Interaction model* is considered when the effect of an independent variable on a dependent variable changes, depending on the value(s) of one or more other independent variables. The interaction effect is represented as a *cross-product terms* =  $\times$ .

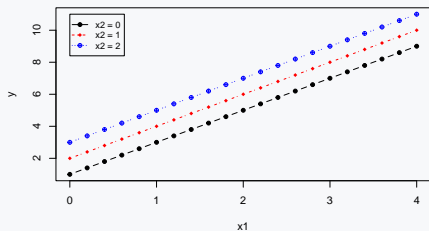


FIGURE: *Non interaction model* graph of  $E(Y) = 1 + 2X_1 + X_2$ , for  $X_2 = 0, 1, 2$  where  $X_1$  is independent from  $X_2$ .

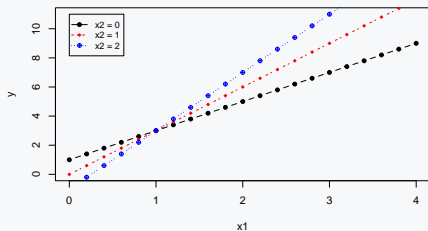


FIGURE: *Interaction model* graph of  $E(Y) = 1 + 2X_1 - X_2 + X_1 \times X_2$ , for  $X_2 = 0, 1, 2$  where  $X_1$  and  $X_2$  are interacted.

## AN INTERACTION MODEL RELATING $E(Y)$ TO AT LEAST ONE QUANTITATIVE INDEPENDENT VARIABLE

The mean value  $E(Y|X_1, X_2)$  of a response  $Y$  is related to two quantitative independent variables  $X_1$  and  $X_2$  by the interaction model

$$E(Y) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2.$$

where

$(\beta_1 + \beta_3 X_2)$  represents the change in  $E(Y)$  for every 1-unit increase in  $X_1$ , holding  $X_2$  fixed,

$(\beta_2 + \beta_3 X_1)$  represents the change in  $E(Y)$  for every 1-unit increase in  $X_2$ , holding  $X_1$  fixed

### Note:

- The interaction between  $X_1$  and  $X_2$  is called a *two-way interaction*, because it is the interaction between two independent variables.
- Higher-order interactions are possible (e.g., *three-way interaction*). For example, consider the three quantitative independent variables  $X_1$ ,  $X_2$ , and  $X_3$ , then  $E(Y)$  is related to  $X_1$ ,  $X_2$ , and  $X_3$  by the interaction model

$$E(Y|X_1, X_2, X_3) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_1 X_2 + \beta_5 X_1 X_3 + \beta_6 X_2 X_3 + \beta_7 X_1 X_2 X_3.$$



The linear models is commonly referred as *Analysis of Covariance* when we model a response variable,  $Y$  based on *quantitative* and *qualitative variables dummy* predictor variables. There are 4 cases:

## CASE 1: COINCIDENT REGRESSION LINES:

The simplest model in which the dummy variable has no effect on  $Y$ , that is,

$$Y = \beta_0 + \beta_1 X + 0D + \varepsilon \quad (\text{equivalently } Y = \beta_0 + \beta_1 X + \varepsilon)$$

and the regression line is exactly the same for both values of the dummy variable ( $D$ ).

## CASE 2: PARALLEL REGRESSION LINES:

When dummy variable produces an additive change in  $Y$ , that is,

$$Y = \beta_0 + \beta_1 X + \beta_2 D + \varepsilon = \begin{cases} \beta_0 + \beta_1 X + \varepsilon & \text{when } D = 0 \\ \beta_0 + \beta_1 X + \beta_2 + \varepsilon & \text{when } D = 1 \end{cases}$$

In this case, the regression coefficient  $\beta_2$  measures the additive change in  $Y$  due to the dummy variable.

**CASE 3: 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  $Y$ , that is,

$$Y = \beta_0 + \beta_1 X + \beta_2 X \times D + \varepsilon = \begin{cases} \beta_0 + \beta_1 X + \varepsilon & \text{when } D = 0 \\ \beta_0 + (\beta_1 + \beta_2)X + \varepsilon & \text{when } D = 1 \end{cases}$$

**CASE 4: UNRELATED REGRESSION LINES (INTERACTION MODEL):**

When quantitative and dummy variables are interacted, that is,

$$Y = \beta_0 + \beta_1 X + \beta_2 D + \beta_3 X \times D + \varepsilon = \begin{cases} \beta_0 + \beta_1 X + \varepsilon & \text{when } D = 0 \\ \beta_0 + \beta_2 + (\beta_1 + \beta_3)X + \varepsilon & \text{when } D = 1 \end{cases}$$

In this case, the regression coefficient  $\beta_2$  measures the additive change in  $Y$  due to the dummy variable, while the regression coefficient  $\beta_3$  measures the change in the size of the effect of  $X$  on  $Y$  due to the dummy variable.

## ADVERTISING DATA: INTERACTIONS MODEL

Consider the advertising data in the package **ISLR2** that has the following variables:

- **TV**: Advertising budgets on TV.
- **radio**: Advertising budgets on radio.
- **newspaper**: Advertising budgets on newspaper.
- **sales**: Response variable.

Suppose we need to build a linear regression to predict the sales based on the advertising on TV and radio. Suppose also 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 case, we need to add an interaction term to our model by multiplying the effects of both TV and radio.

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

## ADVERTISING DATA: MODELLING INTERACTIONS

```
Advertising <- read.csv("C:/1/Advertising.csv")
attach(Advertising)
interact_fit1 <- lm(sales ~ TV + radio + TV:radio)
# Alternatively, use "*" instead of ":" as follows
# interact_fit1 <- lm(sales~TV*radio) #no need to include "TV+radio"

summary(interact_fit1)
```

Results:

	Estimate	Std. Error	t value	Pr(> t )
<b>Intercept</b>	6.7502	0.2479	27.23	< 0.0001
<b>TV</b>	0.0191	0.0015	12.70	< 0.0001
<b>radio</b>	0.0289	0.0089	3.24	0.0014
<b>TV × radio</b>	0.0011	0.0001	20.73	< 0.0001

## ADVERTISING DATA: INTERPRETATION

- The results in the previous table suggests that interactions are important.
- The p-value for the interaction term **TV**  $\times$  **radio** is extremely low, indicating that there is strong evidence for  $H_A : \beta_3 \neq 0$ .
- The  $R^2$  for the interaction model is 96.8%, compared to only 89.7% for the model that predicts sales using **TV** and **radio** without an interaction term (check this as an exercise!).
- 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.
- The coefficient estimates in the table suggest that an increase in **TV** advertising of \$1,000 is associated with increased **sales** of

$$(\hat{\beta}_1 + \hat{\beta}_3 \times \text{radio}) \times 1000 = 19.1 + 1.1 \times \text{radio units.}$$

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

$$(\hat{\beta}_2 + \hat{\beta}_3 \times \text{TV}) \times 1000 = 28.9 + 1.1 \times \text{TV units.}$$

## STYLIZED EXAMPLE: AMOUNT SPENT ON TRAVEL

A 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. The data are given on the web site

<https://gatonweb.uky.edu/sheather/book/docs/datasets/travel.txt> in the file **travel.txt**.

Clearly from Figure 31 (see next slide) that the dummy variable for segment changes the size of the effect of Age,  $X$  on Amount Spent,  $Y$ . We shall also allow for the dummy variable for Segment to produce an additive change in  $Y$ . In this case the appropriate model is what we referred to above as Unrelated regression lines

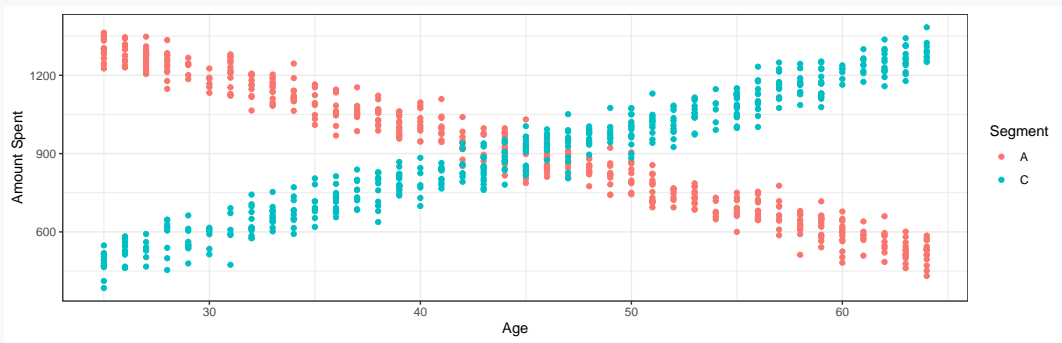
$$Y = \beta_0 + \beta_1 X + \beta_2 C + \beta_3 X \times C + \varepsilon = \begin{cases} \beta_0 + \beta_1 X + \varepsilon & \text{when } C = 0 \\ \beta_0 + \beta_2 + (\beta_1 + \beta_3)X + \varepsilon & \text{when } C = 1 \end{cases}$$

where  $Y$  = amount spent;  $X$  = Age; and  $C$  is a dummy variable which is 1 when the customer is from Segment C and 0 otherwise (i.e., if the customer is in Segment A).

```

travel <- read.table("https://gatonweb.uky.edu/sheather/book/docs/datasets/travel")
# install.packages("tidyverse")
library("ggplot2")
p1 <- ggplot(travel)
p1 + geom_point(aes(x = Age, y = Amount, color = Segment, group = C)) +
  theme_bw() + labs(y = "Amount Spent", x = "Age")

```



**FIGURE:** A scatter plot of Amount Spent versus Age for segments *A* and *C*.

## R-CODE FOR AMOUNT SPENT VERSUS AGE FOR SEGMENTS EXAMPLE

```
attach(travel)
mfull <- lm(Amount ~ Age + C + C:Age) # or use "*" to include the interaction term
summary(mfull)

##
## Call:
## lm(formula = Amount ~ Age + C + C:Age)
##
## 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
```



## R-OUTPUT

Notice that all the regression coefficients are highly statistically significant. Thus, we shall use as a final model (full model) The fitted model is

$$\text{Amount Spent} = \$1814.54 - \$20.32 \times \text{Age} - 1821.2337 \times C + 40.4461 \text{ Age} \times C$$

For customers in segment  $A$  (i.e.,  $C = 0$ ) the model predicts

$$\text{Amount Spent} = \$1814.54 - \$20.32 \times \text{Age}$$

while for customers in segment  $C$  (i.e.,  $C = 1$ ) it predicts

$$\begin{aligned} \text{Amount Spent} &= \$1814.54 - \$20.32 \times \text{Age} - 1821.2337 \times 1 + 40.4461 \text{ Age} \times 1 \\ &= -\$6.69 + \$20.13 \times \text{Age} \end{aligned}$$

Thus, in segment  $A$  (i.e., those customers who have purchased an adventure tour) the amount spent decreases with Age while in Segment  $C$  (i.e., those customers who have purchased a cultural tour) the amount spent increases with Age.

## R-CODE - AMOUNT SPENT VERSUS AGE FOR SEGMENTS EXAMPLE (CONT.)

Imagine that we are interested in an overall test of

$$H_0 : \beta_2 = \beta_3 = 0$$

i.e.,  $Y = \beta_0 + \beta_1x + \varepsilon$  (reduced model: coincident regression lines:) against

$$H_A : H_0 \text{ is not true}$$

i.e.,  $Y = \beta_0 + \beta_1x + \beta_2C + \beta_3C \times x + \varepsilon$  (full model: unrelated lines).

The fit under the **reduced model** is

```
mreduced <- lm(Amount ~ Age)
summary(mreduced)
```

	Estimate	Std. Error	t value	Pr(> t )
Intercept	957.9103	31.3056	30.60	< 0.0001
Age	-1.1140	0.6784	-1.64	0.1009

## R-CODE - AMOUNT SPENT VERSUS AGE FOR SEGMENTS EXAMPLE (CONT.)

Then the F-statistic is given by

$$F = \frac{(RSS(\text{reduced}) - RSS(\text{full})) / (df(\text{reduced}) - df(\text{full}))}{RSS(\text{full}) / df_{\text{full}}}$$
$$= \frac{(52158945 - 2089377) / (923 - 921)}{2089377 / 921} = 1103$$

and the output from R associated with this F-statistic is

```
anova(mreduced, mfull)

## Analysis of Variance Table
##
## Model 1: Amount ~ Age
## Model 2: Amount ~ Age + C + C:Age
##   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
```

There is very strong evidence against the reduced model in favour of the full model. Thus, we prefer the unrelated regression lines model to the coincident lines model.

## MENU PRICING IN A NEW ITALIAN RESTAURANT IN MANHATTAN (CONT.)

Earlier we obtained the final regression model (reduced model) for predicting the menu pricing in a new Italian restaurant in New York City as

$$\text{Price} = -24.03 + 1.54 \text{ Food} + 1.91 \text{ Décor} + 2.07 \text{ East}$$

We wonder whether the restaurants on the east side of Fifth Avenue are very different from those on the west side with service and Décor thought to be more important on the east of Fifth Avenue. Thus, to investigate whether the effect of the predictors depends on the dummy variable East, we consider the extended model (full model):

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 D + \beta_5 \times X_1 \times D + \beta_6 \times X_2 \times D + \beta_7 \times X_3 \times D + \varepsilon,$$

where  $Y = \text{Price}$ ,  $X_1 = \text{Food}$ ,  $X_2 = \text{Décor}$ ,  $X_3 = \text{Service}$ , and  $D = \text{East} = \text{dummy variable}$ .

We test the hypothesis

$$H_0 : \beta_3 = \beta_5 = \beta_6 = \beta_7 = 0$$

i.e.,  $Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_4 \times D + \varepsilon$  (reduced model) against

$$H_A : H_0 \text{ is not true}$$

i.e.,  $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 D + \beta_5 \times X_1 \times D + \beta_6 \times X_2 \times D + \beta_7 \times X_3 \times D + \varepsilon$  (full model).  
Regression output from R showing the test procedures appears is given in the next slide.

## R-CODE FOR ANOVA TEST - ITALIAN RESTAURANTS IN MANHATTAN

```
# We will run the code and interpret the results in lecture class
nyc<-read.csv("https://gattonweb.uky.edu/sheather/book/docs/datasets/nyc.csv", header=T)
attach(nyc)
mfull <- lm(Price~Food+Decor+Service+East+Food:East+Decor:East+Service:East)
summary(mfull)
mreduced <- lm(Price~Food+Decor+East)
summary(mreduced)
anova(mreduced, mfull)
detach(nyc)
```

The F-statistic for comparing the reduced and full models based on ANOVA is given by

$$F = \frac{(RSS(\text{reduced}) - RSS(\text{full})) / (df(\text{reduced}) - df(\text{full}))}{RSS(\text{full}) / df_{\text{full}}} \approx 1.11$$

The  $p$ -value of ANOVA test equals 0.36. Thus, we can't adopt the full model and we conclude that the reduced final model

$$\text{Price} = -24.03 + 1.54 \text{ Food} + 1.91 \text{ Décor} + 2.07 \text{ East}$$

is a good to be adopted.

After fitting a multiple regression model, we check the validity of the model assumptions by:

- 1 Examining the plots of the *standardized residuals* and/or *fitted values*.
- 2 Determining which (if any) of the observation points are *unusual and influential*, that are substantially different from all other observations:
  - which (if any) of the response variable  $Y$  values are unusual (*outliers*).
  - which (if any) of the predictor values that have an influential large effect (*leverage points*) on the estimated regression model.
- 3 Checking whether the errors have constant variance; i.e., *homogeneous (homoscedasticity)* or not; i.e., *heterogeneous (heteroskedasticity)*.
- 4 Checking the extent of *collinearity* among the predictor variables using *variance inflation factors (VIF)*.
- 5 For time series data, check for *autocorrelation* over time.

The model is *valid model* if the conditional mean  $Y$  given  $\mathbf{X} = \{X_1, X_2, \dots, X_p\}$  is a *linear function of  $\mathbf{X}$  and the conditional variance of  $Y$  given  $\mathbf{X}$  is constant vector*. In other words,

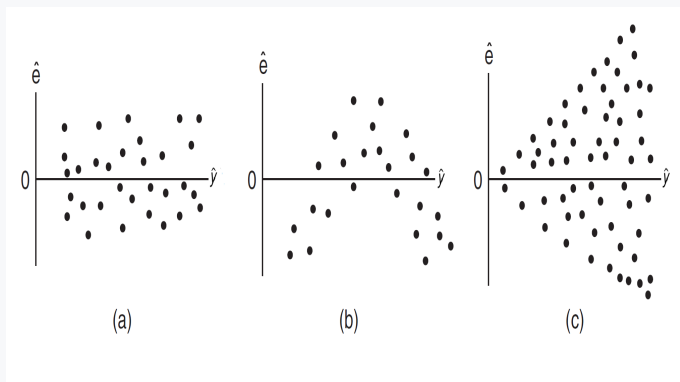
$$E(Y | \mathbf{X} = \mathbf{x}) = \beta_0 + \beta_1 x_1 + \dots + \beta_p x_p \quad \text{and} \quad \text{Var}(Y | \mathbf{X} = \mathbf{x}) = \sigma^2.$$

When a valid model has been fit, a plot of standardized residuals against any predictor or any linear combination of the predictors will have the following features:

- A random scatter of points around the horizontal axis, since the mean function of the error term is zero when a correct model has been fit
- Constant variability as we look along the horizontal axis.

*Thus, any pattern in a plot of standardized residuals is indicative that an invalid model has been fit to the data.*

- Furthermore, when the model is valid, then the plot of  $Y$  against  $\hat{Y}$  should produce points scattered around a straight line.



**FIGURE:** Residuals plotted against linear-model fitted values that reflect (a) model adequacy, (b) quadratic and not linear relationship, and (c) nonconstant variance.



## MENU PRICING IN A NEW ITALIAN RESTAURANT IN NEW YORK CITY

```
library("tidyverse")
library("gridExtra") #Arrange multiple grobs
library("GGally")    #Same as "pairs" function of base R
nyc=read.csv("https://gatonweb.uky.edu/sheather/book/docs/datasets/nyc.csv",header=T)
attach(nyc)
ggpairs(nyc, columns = 4:6)
```

The plot from this code is shown in the next slide. We will see that the model is misspecified as the predictors seem to be related linearly at least approximately.

The model is misspecified when the following two conditions hold:

$$E(Y | \mathbf{X} = \mathbf{x}) = g(\beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p)$$

and

$$E(X_i | X_j) \approx \alpha_0 + \alpha_1 X_j$$

where  $g$  is not the identity function (i.e.,  $g(x) = x$ ).

Note that the model is valid when  $g(x) = x$  is the identity function. In this case, the plot of  $Y$  against  $\hat{Y}$  produces points scattered closely around a straight line.

## SCATTER PLOT MATRIX OF THE CONTINUOUS PREDICTOR VARIABLES

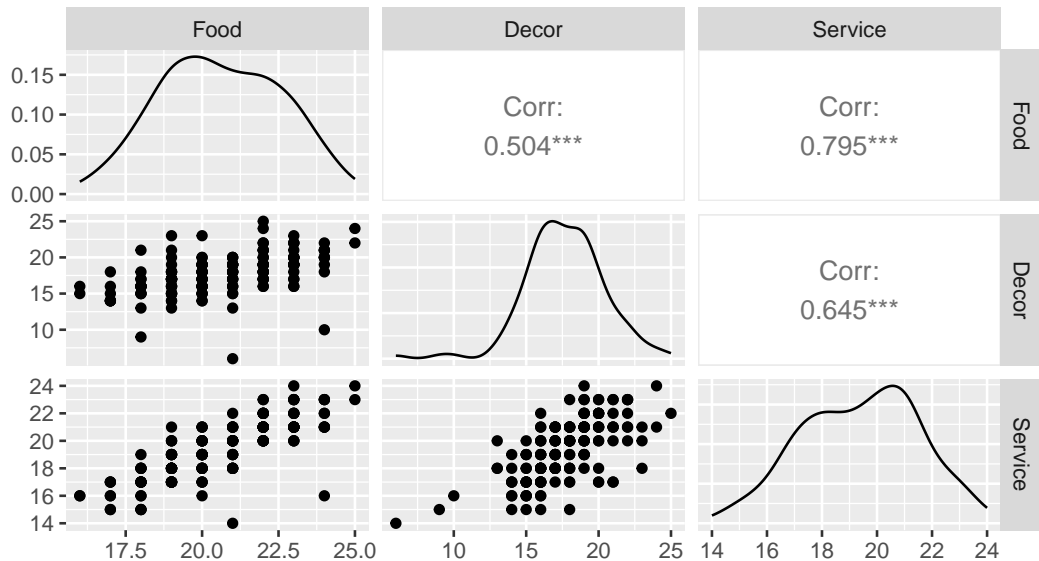


FIGURE: Scatter plot matrix of the three numerical predictor variables.

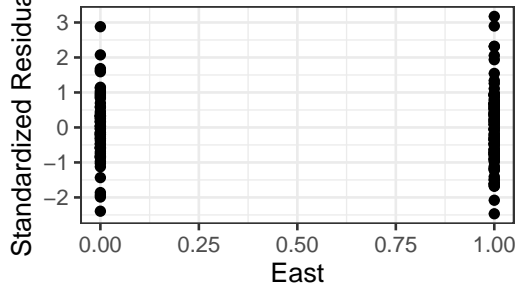
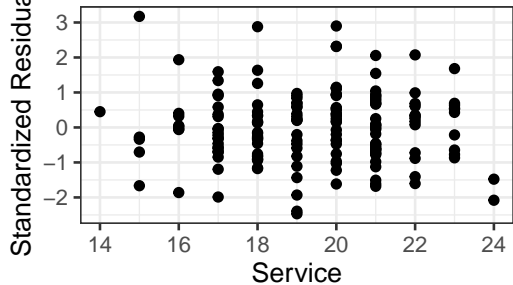
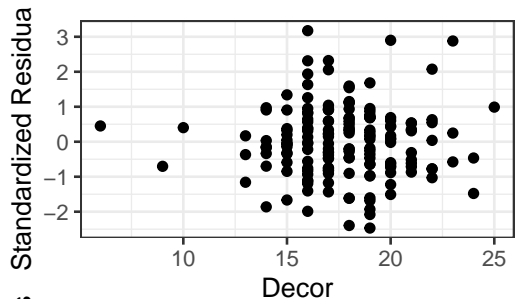
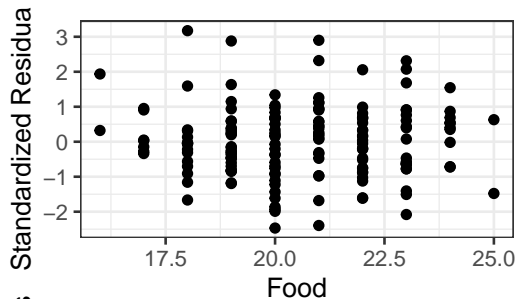
## PLOTS OF STANDARDIZED RESIDUALS AGAINST EACH PREDICTOR VARIABLE (CHECKING FOR LINEARITY)

To check linearity residuals, we plot the standardized residuals against each predictor. If any of these plots show systematic shapes, then the linear model is not appropriate.

```
m1 <- lm(Price ~ Food + Decor + Service + East)
res.std1 <- rstandard(m1)
nyc <- nyc %>% mutate(res.std = res.std1)
p1 <- ggplot(nyc) + geom_point(aes(Food, res.std)) + theme_bw() +
ylab("Standardized Residuals")
p2 <- ggplot(nyc) + geom_point(aes(Decor, res.std)) + theme_bw() +
ylab("Standardized Residuals")
p3 <- ggplot(nyc) + geom_point(aes(Service, res.std)) + theme_bw() +
ylab("Standardized Residuals")
p4 <- ggplot(nyc) + geom_point(aes(East, res.std)) + theme_bw() +
ylab("Standardized Residuals")
layout <- rbind(c(1, 2), c(3, 4))
grid.arrange(grobs=list(p1, p2, p3, p4), ncol=2, layout_matrix=layout)
```

The plots from this code are shown in the next slide.

## PLOTS OF STANDARDIZED RESIDUALS AGAINST EACH PREDICTOR VARIABLE (CHECKING FOR INDEPENDENCY BETWEEN RESIDUALS AND PREDICTORS)



## PLOTS OF FITTED VALUES AGAINST THE MENU PRICING VARIABLE (CHECKING THAT THE REGRESSION IS A LINEAR MODEL)

```
ggplot(m1, aes(.fitted, Price)) + geom_point() + geom_abline() + xlab("Fitted Values") + theme_bw()
```

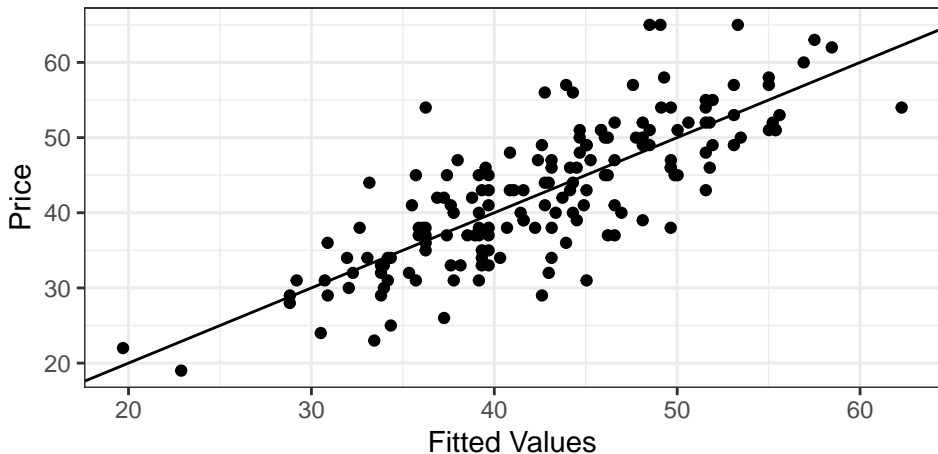
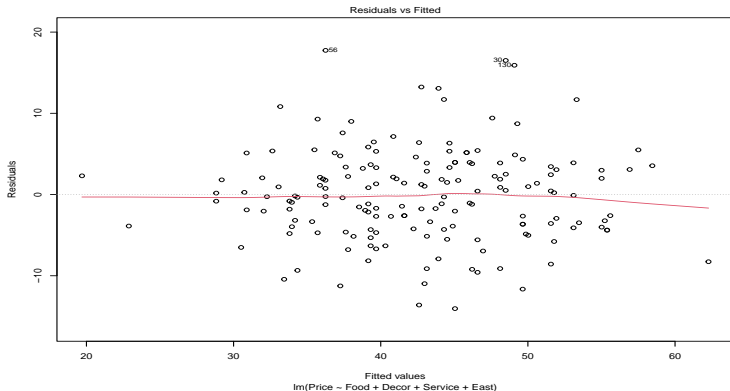


FIGURE: A plot of Price against fitted values.

## CHECK FOR NONLINEARITY USING PLOT FROM BASE R

We can, also, check whether the linearity is valid or not by examining the first plot obtained from the following code `plot(m1)`

```
plot(m1, which = 1)
```

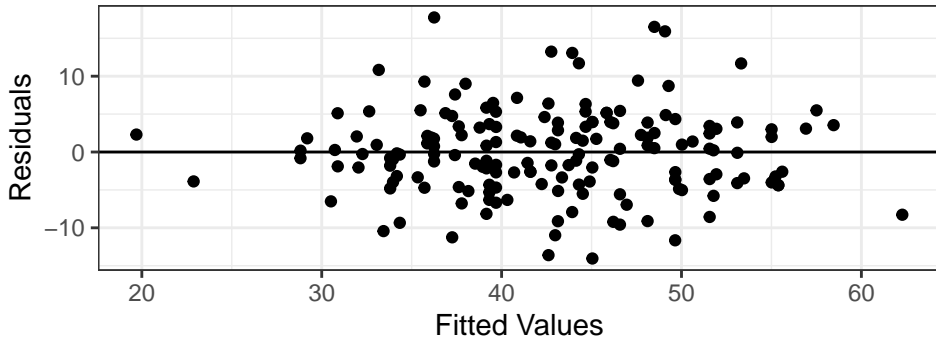


**FIGURE:** A plot of residuals against fitted values to check for any nonlinearity pattern.

## PLOTS OF STANDARDIZED RESIDUALS AGAINST FITTED VALUES FOR MENU PRICING DATA (CHECK THE CONSTANT VARIANCE OF ERRORS)

We can check whether the errors (residuals) have constant variance (*homoscedasticity*) or not (*heteroscedastic*) by examining the scatterplot of standardized residuals against fitted values. If the assumption is valid, then there should be no pattern in the plot.

```
ggplot(m1, aes(.fitted, .resid)) + geom_point() + geom_hline(yintercept = 0) +  
  labs(x = "Fitted Values", y = "Residuals") + theme_bw()
```

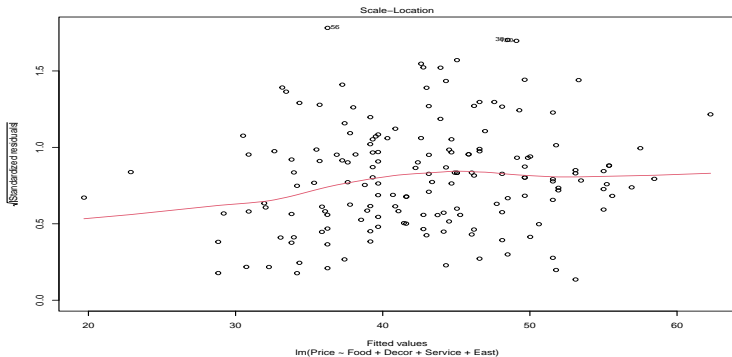


**FIGURE:** A plot of standardized residuals against fitted values.

## CHECK FOR HETEROSKEDASTICITY USING PLOT FROM BASE R

We can, also, check whether the variance of the errors is constant or not by examining the third plot obtained from the following code `plot(m1)`

```
plot(m1, which = 3)
```



**FIGURE:** A plot of square root of absolute standardized residuals against fitted values to check for heterogeneous variance of errors.



## FORMAL TEST STATISTIC TO DETECT HETEROSCEDASTICITY

The function `ncvTest()` from the R package **car** is the score test that can be used to test for non-constant error variance. The null hypothesis

$H_0$  : The errors have constant variance

is tested against the alternative hypothesis

$H_A$  : The errors have non constant variances

```
car::ncvTest(m1)

## Non-constant Variance Score Test
## Variance formula: ~ fitted.values
## Chisquare = 0.8396175, Df = 1, p = 0.35951
```

The large p-value ( $p = 0.35951 > \alpha = 0.05$ ) suggests that the variance is constant.

Recall that the linear regression model in a matrix notation is  $Y = X\beta + \varepsilon$ , and the fitted values are

$$\hat{Y} = X\hat{\beta}, \quad \text{where} \quad \hat{\beta} = (X'X)^{-1}X'Y,$$

Thus, the predicted values of  $Y$  can be calculated as follows:

$$\hat{Y} = X(X'X)^{-1}X'Y = HY, \quad \text{where} \quad H = X(X'X)^{-1}X'$$

The  $(n \times n)$  idempotent matrix  $H$  is commonly called the *hat matrix*. Let  $h_{ij}$  denote the  $(i, j)$  th element of  $H$ , then

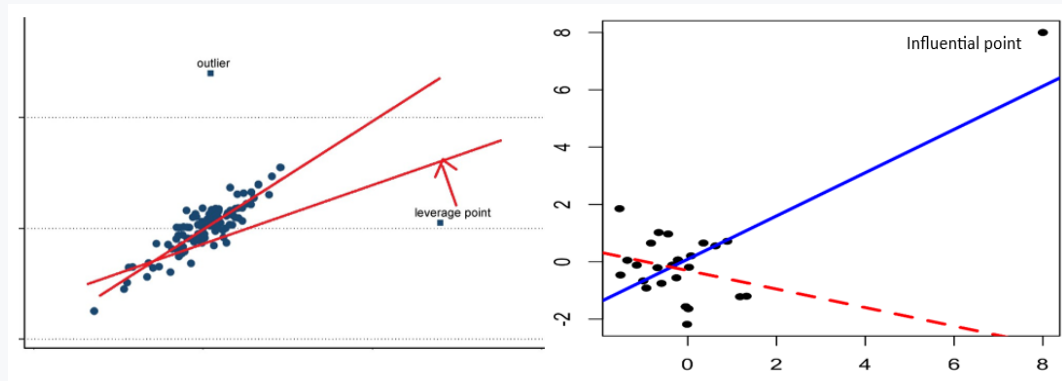
$$\hat{y}_i = h_{ii}y_i + \sum_{j \neq i} h_{ij}y_j$$

where  $h_{ii}$  denotes the  $i$  th diagonal element of  $H$ , which measures the extent to which the fitted regression model  $\hat{y}_i$  is attracted by the given data point,  $y_i$ .

Note that the hat matrix is an *idempotent* because it satisfies  $HH' = H^2 = H$ .

# OUTLIER AND INFLUENTIAL LEVERAGE POINTS

A value whose absence would significantly change the regression equation is termed an *influential observation*.



**FIGURE:** The effect of leverage and outlier points on the fitted model. The estimated fitted line with and without leverage influential point.

# IDENTIFY THE POTENTIAL INFLUENTIAL LEVERAGE POINTS

## HAT-VALUES

A common measure of leverage is the *hat-value* =  $h_{ii}$ ; as a rule of thumb,  $h_{ii} > 2(p+1)/n$  indicate a *high-leverage (influential) data value*.

With multiple explanatory variables and values  $\mathbf{x}_i$  for observation  $i$  with mean  $\bar{\mathbf{x}}$  (as row vectors), let  $\tilde{\mathbf{X}}$  denote the model matrix using centered variables. Then, the *leverage* of the  $i$ th observation is

$$h_{ii} = \frac{1}{n} + (\mathbf{x}_i - \bar{\mathbf{x}})(\tilde{\mathbf{X}}' \tilde{\mathbf{X}})^{-1}(\mathbf{x}_i - \bar{\mathbf{x}})', \quad \text{where } 0 < h_{ii} < 1$$

Notice how the leverage increases as the  $\mathbf{x}_i$  gets farther from  $\bar{\mathbf{x}}$ .

## COOK'S DISTANCE

Another metric to measure the leverage points is *Cook's distance*, as a rule of thumb, *values above*  $4/(n-p-1)$  *indicate a high-influential points*.

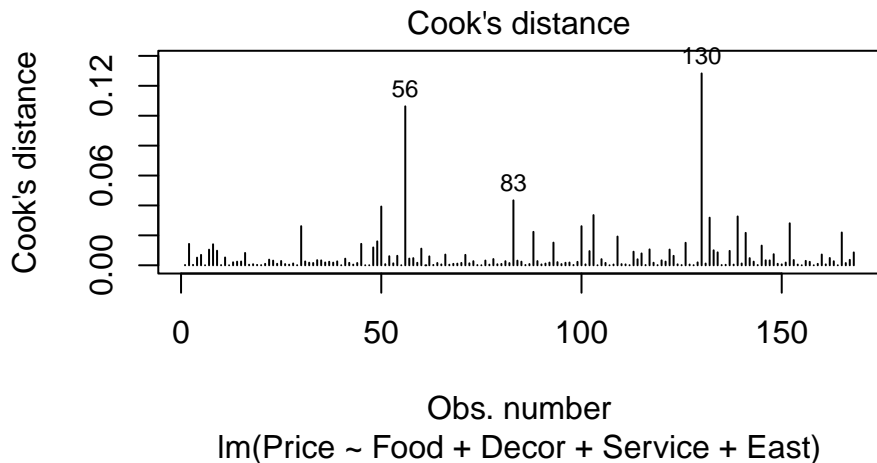
Cook's distance ( $D_i$ ) of observation  $i$ ,  $i = 1, \dots, n$  is defined as the sum of all the changes in the regression model when observation  $i$  is removed from it.

$$D_i = \frac{\sum_{j=1}^n (\hat{y}_j - \hat{y}_{(-i)j})^2}{(p+1)s^2} = \frac{\hat{e}_i^2}{(p+1)s^2} \left[ \frac{h_{ii}}{(1-h_{ii})^2} \right] = \frac{t_i^2}{p+1} \left[ \frac{h_{ii}}{(1-h_{ii})} \right] \sim F(p+1, n-p-1)$$

where  $\hat{y}_{(-i)j}$  is the  $j$ th fitted response value obtained when excluding  $i$ ,  $t_i$  denotes the studentized residual, and  $s^2 = \text{MSE}$ .

## COOK'S DISTANCES TO SPOT FOR INFLUENCE POINTS USING **base R**

```
plot(m1, which = 4, cook.levels = 1) ## try the code: cooks.distance(m1)
```



## REGRESSION LEVERAGE PLOTS USING THE PACKAGE `car`

```
car::leveragePlots(m1)
```

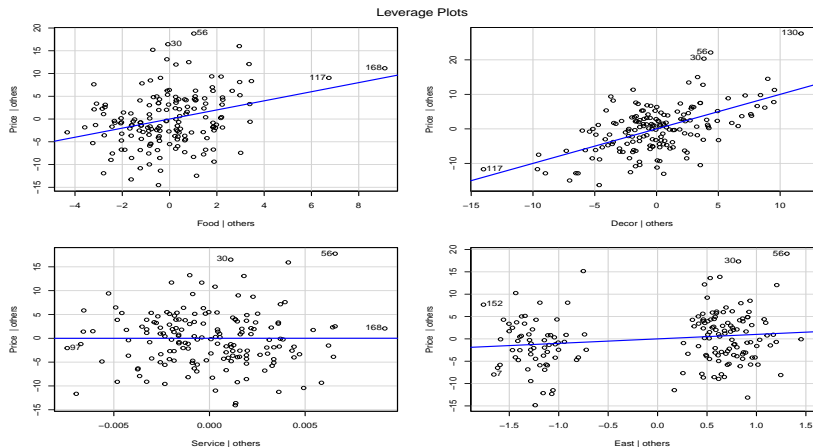
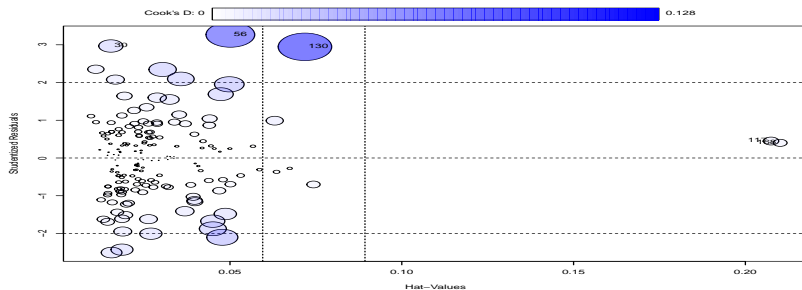


FIGURE: Regression leverage plots for price menu in Italian restaurants data set

## REGRESSION INFLUENTIAL LEVERAGE POINTS USING THE PACKAGE `car`

```
car::influencePlot(m1)
```

```
##          StudRes          Hat          CookD
## 30  2.9679503 0.01532064 0.026157895
## 56  3.2666518 0.05010858 0.106277650
## 117 0.4493433 0.20746530 0.010622954
## 130 2.9463084 0.07181092 0.128275446
## 168 0.4012884 0.21011533 0.008611493
```



**FIGURE:** Regression influential plots for price menu in Italian restaurants data set

## IDENTIFY THE POTENTIAL OUTLIERS IN THE MENU PRICING DATA

One way to identify the outliers is to check the plot of the fitted values versus residuals resulted from the `lm()` function.

```
plot(m1, which = 1)
```

Another way is to plot the standardized residual and find out which data point is outside of  $(-2, 2)$ . The resulted plots of the code are shown in the next slide

```
plot(res.std1, ylab="Standardized Residual", ylim=c(-4, 4), cex=0.2)
#Add horizontal lines 2 and -2 to identify potential outliers
abline(h =c(-2, 0, 2), lty = 2, col = 2)
#Find out which data point is outside of 2 standard deviation cut-off
index <- which(res.std1 > 2 | res.std1 < -2)
#Add price value next to points that have extreme value
text(index, res.std1[index], labels=Price[index])
```



## IDENTIFY THE POTENTIAL OUTLIERS IN THE MENU PRICING DATA

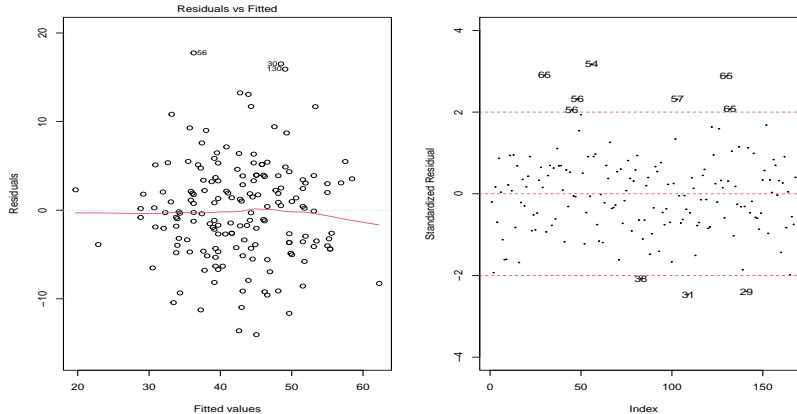


FIGURE: Potential outliers for menu pricing in a new Italian restaurant at Manhattan.

## USE THE R PACKAGE **car** TO TEST FOR OUTLIERS

We can use the function `outlierTest()` build in R package **car** to test the hypotheses

$H_0$  : Data has no outliers

$H_A$  : Data has at least one outlier

```
car::outlierTest(m1)

## No Studentized residuals with Bonferroni p < 0.05
## Largest |rstudent|:
##      rstudent unadjusted p-value Bonferroni p
## 56 3.266652      0.0013284      0.22318
```

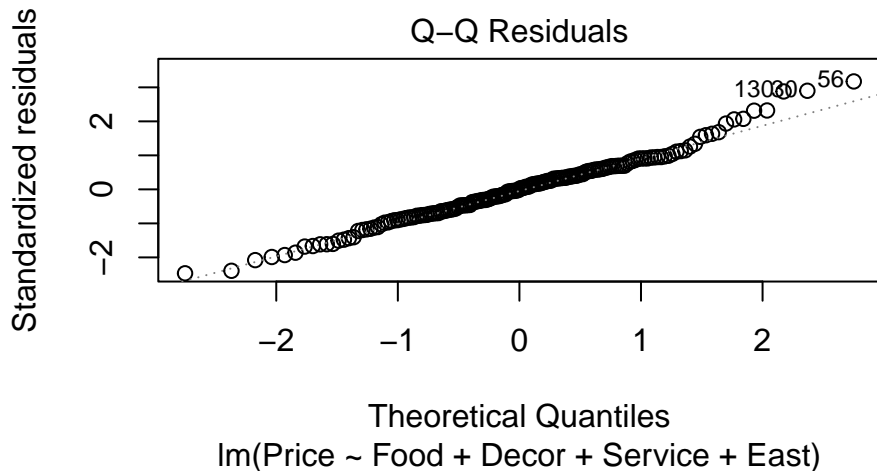
The resulted p-value from the `outlierTest()` function suggests that we have some outliers in the menu pricing. We can use the following code to extract which restaurants have the highest extreme prices:

```
Restaurant[index]

## [1] "Harry Cipriani"      "Bravo Gianni"          "Il Valletto Due Mila"
## [4] "Nello"                 "Rughetta"              "Rao's"
## [7] "Casa Mia"              "Rainbow Grill"         "San Domenico"
## [10] "Trattoria Del Sogno"
```

## CHECK FOR NORMALITY AND DETECT FOR OUTLIERS: QQ PLOT

```
plot(m1, which = 2) # Try the code car::qqPlot(m1,main="Normal Q-Q")
```



# MULTICOLLINEARITY AND VARIANCE INFLATION FACTORS

The multicollinearity can lead to a serious problem as the matrix  $\mathbf{X}\mathbf{X}'$  might be singular (inverse does not exist). This singularity might inflate the variance of the regression coefficients.

## INFORMAL DIAGNOSTICS FOR EXISTENCE OF COLLINEARITY PROBLEM

- Strong correlation between the independent variables.
- Adding/removing one predictor implies a large change in the estimated regression coefficients.
- The standard errors of the regression coefficients are large.
- The estimated regression coefficients do not make sense (wrong sign  $+/-$  for what we expect to get).
- The overall  $F$ -test is highly significant indicating that the overall model is good, but none of the  $t$ -tests on the regression coefficients for the predictor variables is significant.
- The value of the coefficient of determination  $R^2$  has very small change when an independent variable is added or removed.
- The inverse of the matrix  $\mathbf{X}\mathbf{X}'$  might not exist (singular) and its determinant has a value close to zero.

One solution to alleviate the multicollinearity problem is to increase the sample size. Another solution is to merge some highly dependent variables together. For example, the score of math ( $X_1$ ) and the score of statistics ( $X_2$ ) might be combined into one variable, say the score of mathematical statistics ( $X_{12} = aX_1 + bX_2$ ) where  $a$  and  $b$  are constants.

# VARIANCE INFLATION FACTOR (VIF)

Consider the bivariate regression model

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$$

and let  $r_{12}$  denote the correlation between  $x_1$  and  $x_2$  and  $S_{x_j}$  denote the standard deviation of  $x_j$ . Then it can be shown that

$$\text{Var}(\hat{\beta}_j) = \frac{1}{1 - r_{12}^2} \times \frac{\sigma^2}{(n - 1)S_{x_j}^2} \quad j = 1, 2$$

Notice how the variance of  $\hat{\beta}_j$  gets larger as the absolute value of  $r_{12}$  increases. Thus, correlation amongst the predictors increases the variance of the estimated regression coefficients.

For example, when  $r_{12}^2 = 0.99$  the variance of  $\hat{\beta}_j$  is  $\frac{1}{1 - r_{12}^2} = \frac{1}{1 - 0.99^2} = 50.25$  times larger than it would be if  $r_{12}^2 = 0$ .

## DEFINITION

The term  $\frac{1}{1 - r_{12}^2}$  is called a variance inflation factor (VIF).

# VARIANCE INFLATION FACTOR (VIF)

Recall the general multiple regression model

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_p x_p + \varepsilon.$$

Let  $R_j^2$  denote the value of  $R^2$  obtained from the regression of  $x_j$  on the other  $x'$ 's (i.e., the amount of variability explained by this regression). Then it can be shown that

$$\text{Var}(\hat{\beta}_j) = \frac{1}{1 - R_j^2} \times \frac{\sigma^2}{(n - 1)S_{x_j}^2} \quad j = 1, \dots, p$$

## DEFINITION

The term  $1 / (1 - R_j^2)$  is called the  $j$ th *variance inflation factor (VIF)*.

As a rule of thumb

- **VIF** < 5 is an indicative of **no** multicollinearity.
- $5 \leq$  **VIF** < 10 is an indicative of **minor** multicollinearity.
- **VIF**  $\geq$  10 is an indicative of **serious** multicollinearity.

## VARIANCE INFLATION FACTOR (VIF) FOR THE MENU PRICING DATA SET

Revisit the example of modeling the menu pricing in a new Italian restaurant in New York City, where the final model is the reduced model that we obtained by fitting the following code.

```
m1 <- lm(Price~Food+Decor+Service+East)
```

The variance inflation factor (VIF) for this data set are as follows:

```
library("car")  
vif(m1)
```

```
##      Food      Decor  Service      East  
## 2.714261 1.744851 3.558735 1.064985
```

We noticed that all variance inflation factors less than 5 and so no multicollinearity is detected.

There are three situations for transformations

- ① Only the response variable needs to be transformed.
- ② Only the predictor variables needs to be transformed.
- ③ Both the response and predictor variables need to be transformed.

There two general methods for transforming

- ① Inverse response plots.
- ② Box-Cox procedure.



## TRANSFORMING ONLY THE RESPONSE VARIABLE $Y$

Suppose that the true regression model between  $Y$  and  $X_1, X_2, \dots, X_p$  is given by

$$Y = g(\beta_0 + \beta_1 x_1 + \dots + \beta_p x_p + \varepsilon)$$

where  $g$  is a function that is generally unknown. This model can be turned into a multiple linear regression model by transforming  $Y$  by  $g^{-1}$ , the inverse of  $g$ , since,

$$g^{-1}(Y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_p x_p + \varepsilon$$

For example, suppose that

$$Y = \exp(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_p x_p + \varepsilon)$$

then,

$$g(Y) = \exp(\mathbf{X}) \text{ and so } g^{-1}(Y) = \log(\mathbf{X}).$$

We next look at methods for estimating  $g^{-1}$ .

## INVERSE RESPONSE PLOT

The transformed function  $g^{-1}$  can be estimated from the scatter plot of  $Y$  (on the horizontal axis) and the  $\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \dots + \hat{\beta}_p x_p$  (on the vertical axis).

## BOX-COX TRANSFORMATION

The Box-Cox procedure aims to find a transformation that makes the transformed response variable close to normally distributed. The simple case for Box-Cox transformation is given by

$$g_\lambda(y) = \begin{cases} \frac{y^\lambda - 1}{\lambda} & \lambda \neq 0 \\ \log(y) & \lambda = 0 \end{cases}$$

The objective is to use the data to choose a value of the parameter  $\lambda$  that maximizes the normality of the residuals  $(g_\lambda(Y) - \mathbf{X}\beta)$ .

## MODELLING DEFECTIVE RATES

This example is taken from Simon Sheather book (See Ch6) which was adapted from Siegel (1997, pp. 509-510). Data can be found on the web site

[https://gatonweb.uky.edu/sheather/book/data\\_sets.php](https://gatonweb.uky.edu/sheather/book/data_sets.php) in the file **defects.txt**.

According to Siegel:

*Everybody seems to disagree about just why so many parts have to be fixed or thrown away after they are produced. Some say that it's the standard deviation of the temperature of the production process, which needs to be minimised. Others claim it is clearly the density of the product, and that the problems would disappear if the density is increased. Then there is Ole, who has been warning everyone forever to take care not to push the equipment beyond its limits. This problem would be easiest to fix, simply by slowing down the production rate; however, this would increase some costs. The data has the average number of defects per 1,000 parts produced (denoted by Defective) along with values of the other variables described above for 30 independent production runs.*

The variables are

- $Y = \text{Defective.}$
- $X_1 = \text{Temperature.}$
- $X_2 = \text{Density.}$
- $X_3 = \text{Rate.}$

See the R-markdown Example 2 posted in Brightspace.

You need to

- Read and understand the examples and code in Section 3.6 Lab: Linear Regression from the textbook *"An Introduction to Statistical Learning: With Applications in R"*
- Solve question 2 in Exercise 3.7.
- Solve question 3 in Exercise 3.7.
- Solve question 4 in Exercise 3.7.
- Solve question 8 in Exercise 3.7.
- Solve question 9 in Exercise 3.7.
- Solve questions 13, 14, 15 in Exercise 3.7.