Lecture 9:
Multiple Linear Regression
STAT 432, Spring 2021

Multiple Linear Regression (MLR)

Suppose y is a response variable, and $x_1, \cdots, x_p$ are p explanatory variables. Then the multiple linear regression model can be written as

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p + \epsilon$$

where $\epsilon \sim N(0, \sigma^2)$ is the random error term.

Multiple Linear Regression (MLR)

Suppose we have a collection $i = 1, \cdots, n$ observations. Then the multiple linear regression model for case i is written as

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} \cdots + \beta_p x_{ip} + \epsilon_i$$

where $\epsilon_i \sim N(0, \sigma^2)$ independently.

Given estimates $\hat{\beta}_0, \hat{\beta}_1, \cdots, \hat{\beta}_p$ of the unknown regression parameters $\beta_0, \beta_1, \cdots, \beta_p$:

- The i^{th} fitted (or predicted) value:

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \hat{\beta}_2 x_{i2} + \cdots + \hat{\beta}_p x_{ip}$$

- The i^{th} residual:

$$\hat{e}_i = y_i - \hat{y}_i = y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \hat{\beta}_2 x_{i2} - \cdots - \hat{\beta}_p x_{ip}$$

Least Squares Estimation

The parameter estimates $\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2, \cdots, \hat{\beta}_p$ are found by minimizing the sum of squared residuals:

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

- Using some calculus, a closed form solution for the parameter estimates can be derived and expressed using matrix notation.
- In practice, for a specific data set, we can use the `lm()` function in R to compute the least squares estimates of the parameters.

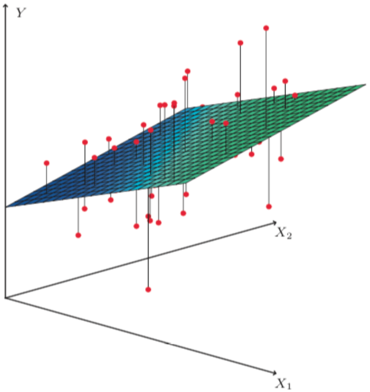


FIGURE 3.4. *In a three-dimensional setting, with two predictors and one response, the least squares regression line becomes a plane. The plane is chosen to minimize the sum of the squared vertical distances between each observation (shown in red) and the plane.*

From Chapter 3, p. 73, of *An Introduction to Statistical Learning*.

Estimating σ^2

Estimate of σ^2:

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

Residual standard error:

$$\hat{\sigma} = \sqrt{\frac{\mathrm{RSS}}{n-p-1}}$$

Hypothesis Test for a Single Predictor

Test whether parameter β_j is zero.

$H_0 : \beta_j = 0$
$H_A : \beta_j \neq 0$

Test statistic:

$$t_j = \frac{\hat{\beta}_j}{se(\hat{\beta}_j)}; \quad \text{df} = n - p - 1$$

- $se(\hat{\beta}_j)$ is the standard error of $\hat{\beta}_j$
- n is the number of observations
- p is the number of predictor variables
- degrees of freedom (df) $=$
  sample size - number of parameters estimated $= n - p - 1$
  (since, when including the intercept, there are $p + 1$ parameters)

Confidence Interval for a Single Predictor

A $1-\alpha$ confidence interval for β_j:

$$\hat{\beta}_j \pm t_{\alpha/2;n-p-1} se(\hat{\beta}_j)$$

The R function `confint()` can be used to calculate confidence intervals for the parameters.

Coefficient of Determination (R^2)

The coefficient of determination R^2 has the same definition for simple and multiple linear regression.

$$R^2 = \frac{\text{RegSS}}{\text{TSS}} = 1 - \frac{\text{RSS}}{\text{TSS}}$$

- $\text{TSS} = \sum_{i=1}^{n}(y_i - \bar{y})^2$ is the total sum of squares
  - total variability in the response variable
- $\text{RegSS} = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2$ is the regression sum of squares
  - variability in the response explained by the model
- $\text{RSS} = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$ is the residual sum of squares
  - unexplained variability

Coefficient of Determination (R^2)

$$R^2 = \frac{\text{RegSS}}{\text{TSS}} = 1 - \frac{\text{RSS}}{\text{TSS}}$$

- R^2 can be interpreted as the proportion of variability in the response y that is explained by the regression model.
- $0 \leq R^2 \leq 1$, where values closer to 1 indicate a better linear fit to the data.
- **Problem with R^2 in MLR**: Adding predictor variables to the regression model will always increases R^2 (or, equivalently decrease RSS). Even if the predictor variable is irrelevant (noise) the R^2 will increase slightly. This is not ideal since simpler models are preferred to more complicated models.

Occam's razor, or the law of parsimony, is a problem solving principle that states that simpler solutions are preferred to more complex ones.[1]

"Everything should be kept as simple as possible, but not simpler"
–Albert Einstein

[1] https://en.wikipedia.org/wiki/Occam's_razor

Adjusted Coefficient of Determination (R^2_{adj})

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

- R^2_{adj} penalizes for adding variables to regressional model.
- The idea is that the R^2_{adj} should decrease when adding irrelevant predictor variables into a model.
- When comparing models with different numbers of predictors one should use R^2_{adj} and not R^2.

MLR Example: Menu Pricing Data Set

Data set from surveys of customers of 168 Italian restaurants in New York City.[2]

The variables are:

- y = Price = the price (in \$US) of dinner (including 1 drink and tip)
- x_1 = Food = customer rating of the food (out of 30)
- x_2 = Decor = customer rating of the decor (out of 30)
- x_3 = Service = customer rating of the service (out of 30)
- x_4 = East = dummy variable, 1 (0) if the restaurant is east (west) of Fifth Avenue

[2] Zagat Survey 2001: New York City Restaurants

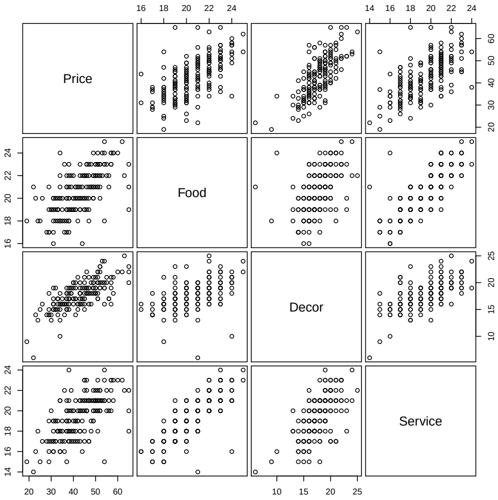
MLR Example: Menu Pricing Data Set

```
> nyc <- read.csv("https://ericwfox.github.io/data/nyc.csv")

> head(nyc)
          Restaurant Price Food Decor Service East
1 Daniella Ristorante    43   22    18      20    0
2  Tello's Ristorante    32   20    19      19    0
3          Biricchino    34   21    13      18    0
4             Bottino    41   20    20      17    0
5          Da Umberto    54   24    19      21    0
6           Le Madri    52   22    22      21    0
```

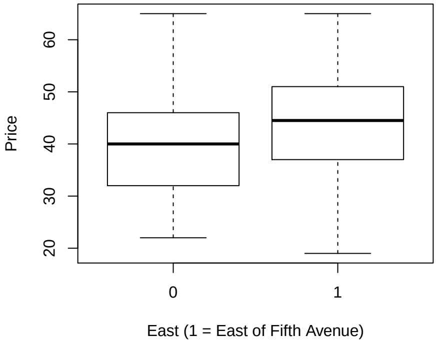
MLR Example: Menu Pricing Data Set

```
> pairs(Price ~ Food + Decor + Service, data=nyc)
```



MLR Example: Menu Pricing Data Set

```
> boxplot(Price ~ East, data= nyc,
          ylab="Price", xlab="East (1 = East of Fifth Avenue)")
```



MLR Example

```
> lm1 <- lm(Price ~ Food + Decor + Service + East, data=nyc)
> summary(lm1)

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
```

MLR Example

Since Service is not significant we remove it from the model.

```
> lm2 <- lm(Price ~ Food + Decor + East, data=nyc)
> summary(lm2)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -24.0269     4.6727  -5.142 7.67e-07 ***
Food          1.5363     0.2632   5.838 2.76e-08 ***
Decor         1.9094     0.1900  10.049  < 2e-16 ***
East          2.0670     0.9318   2.218   0.0279 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 5.72 on 164 degrees of freedom
Multiple R-squared:  0.6279,Adjusted R-squared:  0.6211
F-statistic: 92.24 on 3 and 164 DF,  p-value: < 2.2e-16
```

```
> s1 <- summary(lm1)
> s2 <- summary(lm2)

> s1$r.squared
[1] 0.6278809
> s2$r.squared
[1] 0.6278808

> s1$adj.r.squared
[1] 0.6187492
> s2$adj.r.squared
[1] 0.6210738

#---------------------------------
> confint(lm2)
                 2.5 %     97.5 %
(Intercept) -33.253364 -14.800395
Food          1.016695   2.055996
Decor         1.534181   2.284565
East          0.227114   3.906912
```

MLR Example

The final regression model is:

$$\widehat{\texttt{Price}} = -24.03 + 1.54\,\texttt{Food} + 1.91\,\texttt{Decor} + 2.07\,\texttt{East}$$

For example, we can use the model to predict `Price` when `Food=20`, `Decor=16` and `East=1`:

$$\widehat{\texttt{Price}} = -24.03 + 1.54(20) + 1.91(16) + 2.07(1) = 39.4$$

We can also use R to make this prediction and to calculate a 95% prediction interval.

```
> new_x <- data.frame(Food = 20, Decor = 16, East = 1)
> predict(lm2, newdata = new_x, interval="prediction")
       fit      lwr      upr
1 39.31701 27.95384 50.68019
```

MLR Example

The final regression model is:

$$\widehat{\texttt{Price}} = -24.03 + 1.54\,\texttt{Food} + 1.91\,\texttt{Decor} + 2.07\,\texttt{East}$$

- `Decor` has the largest effect on `Price` since its regression coefficient is largest. Note that `Food`, `Decor`, and `Service` are on the same 0 to 30 scale, so it is meaningful to make the comparison.

- If a goal is to maximize `Price` for a new restaurant, it should be located east of Fifth Avenue (i.e., $\texttt{East} = 1$).

Interpreting Regression Coefficients

Suppose we fit a multiple linear regression model

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \cdots + \hat{\beta}_p x_p,$$

where x_j is the j^{th} predictor and $\hat{\beta}_j$ the estimated coefficient for the variable.

How do we interpret $\hat{\beta}_j$?

The usual interpretation is as follows: an increase in x_j by 1, *with all other predictors in the model held fixed*, is associated with a change of $\hat{\beta}_j$ in the predicted response, $\hat{y}$.

Interpreting Regression Coefficients

Going back to the example, the final regression model is

$$\widehat{\texttt{Price}} = -24.03 + 1.54\texttt{Food} + 1.91\texttt{Decor} + 2.07\texttt{East}$$

- Interpret the coefficient for `Decor`: a one unit increase in the customer rating of decor, with the other predictors (`Food` and `East`) held fixed, is associated with an increase in `Price` by \$1.91.
- Interpret the coefficient for the dummy variable `East`: the price of dinner at a restaurant east of Fifth Avenue will cost \$2.07 more, on average, than a restaurant west of Fifth Avenue, when all other predictors (`Food` and `Decor`) are held fixed.

Interpreting Regression Coefficients

Some problems when interpreting regression coefficients:

- The interpretation of β_j as the average change in y per unit change in x_j, *with all other predictors held fixed*, assumes predictors can be changed without affecting other predictors.
- Interpretation becomes hazardous when there are correlations amongst predictors. When x_j changes, then values for other predictors also change.
- The magnitude and sign of a coefficient can change when including (or removing) another predictor from the model.
- For observational data we can only make claims about associations, not causation.

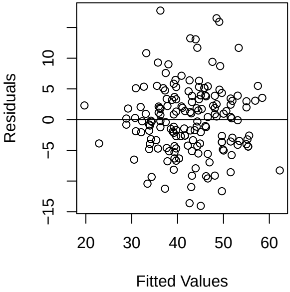
Conditions for MLR

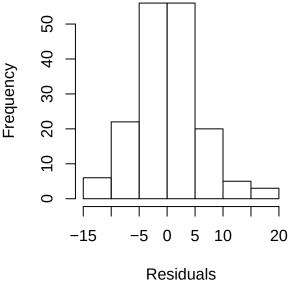
1. **Linearity**. There is a linear relationship between the the response variable and predictor variables. More precisely, the response variable can be modeled as a linear combination of predictor variables.
2. **Constant Variability**. The variability of the residuals should be roughly constant.
3. **Normality**. The residuals should be approximately normally distributed with mean 0.
4. **Independence**. The values of the response variable are independent of each other.

Regression diagnostics are numerical and graphical techniques to check the conditions. Residuals plots will be our primary diagnostic for checking the conditions.

Diagnostics

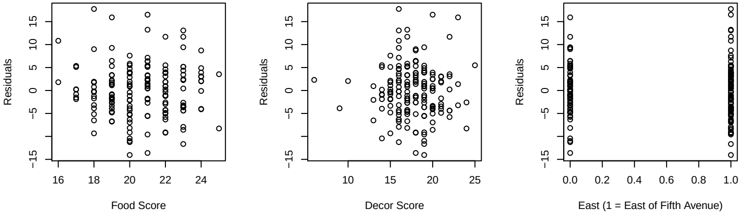
Going back to menu pricing example, below is a plot of the residuals versus fitted values, and a histogram of the residuals. Based on these plots the conditions of linearity, constant variance, and normality are reasonably satisfied.





Diagnostics

We can also look at a plot of the residuals versus each predictor variable. Again, these plots show that the conditions for MLR are reasonably satisfied. For instance, there appears to be constant variability in the residuals in each plot, and no indications of nonlinearity.



Diagnostics

Here is the code for the diagnostic plots in the previous slides:

```
lm2 <- lm(Price ~ Food + Decor + East, data=nyc)

# residuals versus fitted values
plot(predict(lm2), resid(lm2), xlab="Fitted Values", ylab="Residuals")
abline(h=0)

# histogram of residuals
hist(resid(lm2), xlab="Residuals", main='')

# residuals versus each predictor
plot(nyc$Food, resid(lm2), xlab="Food Score", ylab="Residuals")
plot(nyc$Decor, resid(lm2), xlab="Decor Score", ylab="Residuals")
plot(nyc$East, resid(lm2), xlab="East (1 = East of Fifth Avenue)", ylab="Residuals")
```