STA302 METHODS OF DATA ANALYSIS 1

MODULE 3: ASSUMPTIONS OF LINEAR REGRESSION MODELS

PROF. KATHERINE DAIGNAULT

MODULE 3 OUTLINE

1. Introduction to Linear Regression Assumptions
2. Verifying Assumptions using Residual Plots
3. Additional Conditions for Multiple Linear Models

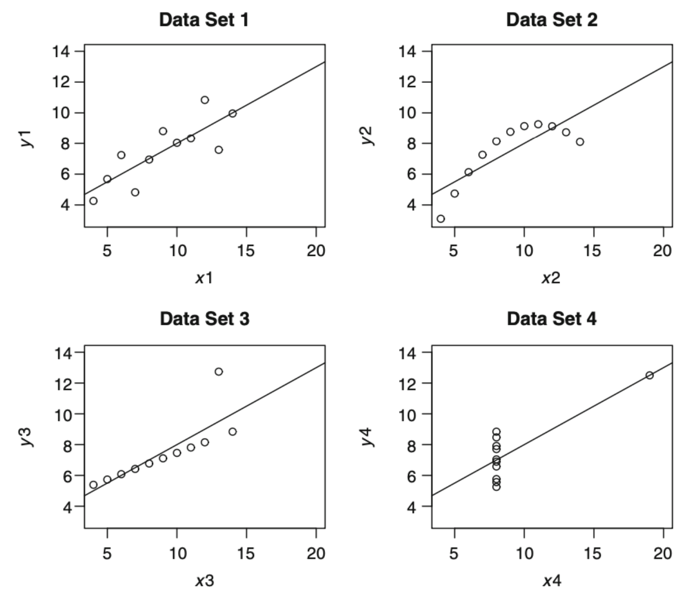


Figure 3.1 Plots of Anscombe's four data sets

From Sheather 's A Modern Approach to Regression with R, pg 46

```
> lm(y1 ~ x1, data=anscombe)

Call:
lm(formula = y1 ~ x1, data = anscombe)

Coefficients:
(Intercept)           x1
     3.0001       0.5001

> lm(y2 ~ x2, data=anscombe)

Call:
lm(formula = y2 ~ x2, data = anscombe)

Coefficients:
(Intercept)           x2
      3.001        0.500
```

```
> lm(y3 ~ x3, data=anscombe)

Call:
lm(formula = y3 ~ x3, data = anscombe)

Coefficients:
(Intercept)           x3
     3.0025       0.4997

> lm(y4 ~ x4, data=anscombe)

Call:
lm(formula = y4 ~ x4, data = anscombe)

Coefficients:
(Intercept)           x4
     3.0017       0.4999
```

ANSCOMBE'S DATASETS

```
> anscombe <- read.table("anscombe.txt", header=T)
> anscombe
   case x1 x2 x3 x4    y1   y2    y3    y4
1     1 10 10 10  8  8.04 9.14  7.46  6.58
2     2  8  8  8  8  6.95 8.14  6.77  5.76
3     3 13 13 13  8  7.58 8.74 12.74  7.71
4     4  9  9  9  8  8.81 8.77  7.11  8.84
5     5 11 11 11  8  8.33 9.26  7.81  8.47
6     6 14 14 14  8  9.96 8.10  8.84  7.04
7     7  6  6  6  8  7.24 6.13  6.08  5.25
8     8  4  4  4 19  4.26 3.10  5.39 12.50
9     9 12 12 12  8 10.84 9.13  8.15  5.56
10   10  7  7  7  8  4.82 7.26  6.42  7.91
11   11  5  5  5  8  5.68 4.74  5.73  6.89
```

LINEAR REGRESSION ASSUMPTIONS

1. Linearity of the Relationship (also known as Mean Zero Errors) assumption

$$E(\boldsymbol{\varepsilon}\,|\boldsymbol{X}) = \boldsymbol{0} \ \text{ or } \ E(\boldsymbol{Y}|\boldsymbol{X}) = \boldsymbol{X\beta} \ \text{ or } \ \boldsymbol{Y} = \boldsymbol{X\beta} + \boldsymbol{\varepsilon}$$

2. Uncorrelated Errors (sometimes referred to as Independence) assumption

$$Cov(\varepsilon_i, \varepsilon_j) = 0 \ \text{ or } \ Cov(y_i, y_j) = 0$$

3. Constant Error Variance (also known as Homoskedasticity) assumption

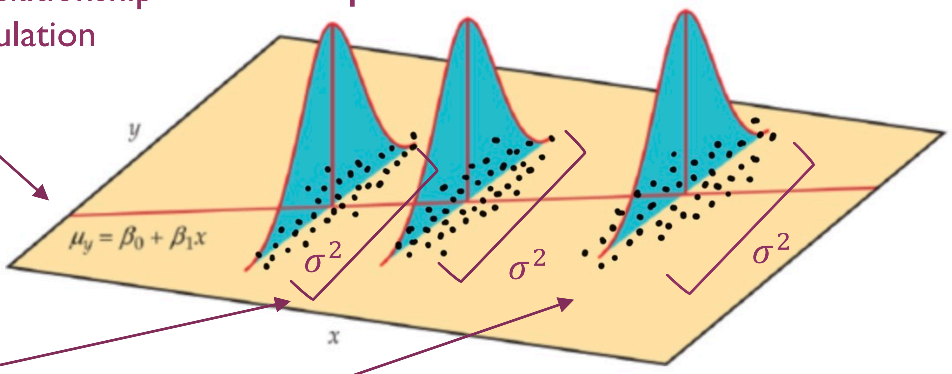
$$Var(\boldsymbol{\varepsilon}|\boldsymbol{X}) = \sigma^2\boldsymbol{I} \ \text{ or } \ Var(\varepsilon_i|X) = Var(y_i|X) = \sigma^2$$

4. Normal Errors assumption

$$\boldsymbol{\varepsilon}|\boldsymbol{X} \sim N_n(\boldsymbol{0}, \sigma^2\boldsymbol{I}) \ \text{or} \ \boldsymbol{Y}|\boldsymbol{X} \sim N_n(\boldsymbol{X\beta}, \sigma^2\boldsymbol{I}) \ \text{ or } \ \varepsilon_i \sim N(0, \sigma^2)$$

True relationship in population

Population



Normally distributed responses/errors

MORE ON LINEAR REGRESSION ASSUMPTIONS

Assumptions relate to properties of the population, not the sample

- When fitting a model, we implicitly make these assumptions EVERY time

Uncorrelated Errors Assumption

- Each data point in population must not be related or connected to any other data point
  - i.e. knowing information about one does not give any information about another
- *Examples of violation: stock price data, weather data, measurements on the same person, etc.*

Linearity/Mean Zero Error Assumption

- Implies two things about the population relationship:
  - The true relationship is linear in the coefficients
  - The true relationship is exactly $\boldsymbol{Y} = \boldsymbol{X\beta} + \boldsymbol{\varepsilon}$ with
    - no predictors omitted from $\boldsymbol{X}$ that should be present,
    - no predictors included in $\boldsymbol{X}$ that should not be present, and
    - no predictors in $\boldsymbol{X}$ that are in the wrong functional form
- *Examples of violation: omitting a predictor known to influence response; fitting a linear model when the truth is $y_i = \log(\beta_0 + \beta_1 x_i + \varepsilon_i)$; including x when it should be x^2; etc.*

MORE ON LINEAR REGRESSION ASSUMPTIONS

Constant Variance Assumption

- Each conditional distribution must have the same spread
  - So only difference between each one is that the mean changed by a specific amount

Normality Assumption

- Each conditional distribution must have the same shape
- Harder to verify in small samples
- Not needed to estimate coefficients by least squares

Importance of Assumptions

- Linearity ensures we estimate coefficients unbiasedly
- Uncorrelated errors ensure correct precision of estimates
- Constant variance ensures we obtain reasonable estimates of variability for all conditional means
- Normality allows us to utilize properties of Normal random variables for inferential purposes (e.g. computing confidence intervals)
- Nothing stops us from fitting an invalid model with violated assumptions

MODULE 3 OUTLINE

1. Introduction to Linear Regression Assumptions
2. Verifying Assumptions using Residual Plots
3. Additional Conditions for Multiple Linear Models

CHECK ASSUMPTIONS USING RESIDUALS

Population Error Assumptions: $\boldsymbol{\varepsilon}|\boldsymbol{X} \sim N_n(\boldsymbol{0}, \sigma^2 \boldsymbol{I})$

- Residuals are sample analogues of errors
  - $\varepsilon_i = y_i - E(y_i|x_i)$ whereas $\hat{e}_i = y_i - \hat{E}(y_i|x_i)$
  - Assuming sample was collected appropriately from population
- Residuals capture noise leftover after estimating the trend between $\boldsymbol{Y}$ and $\boldsymbol{X}$.
  - If estimated coefficient close to the truth (unbiased) then $\hat{\boldsymbol{e}} = \boldsymbol{Y} - \widehat{\boldsymbol{Y}} = \boldsymbol{X}(\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}}) + \boldsymbol{\varepsilon} \approx \boldsymbol{\varepsilon}$
- If a violation of the error assumptions occurs, we should be able to see it in the residuals
- E.g. the true relationship we want to estimate is

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i1}^2 + \varepsilon_i$$

- But model fit in sample data was different

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{i1}$$

- The residuals would pick up the linearity issue since

$$\hat{e}_i = y_i - \hat{y}_i = (\beta_0 - \hat{\beta}_0) + (\beta_1 - \hat{\beta}_1)x_{i1} + \beta_2 x_{i1}^2 + \varepsilon_i \approx \boxed{\beta_2 x_{i1}^2 + \varepsilon_i}$$

LOOK FOR PATTERNS IN RESIDUAL PLOTS

- **Residuals versus each predictor (scatterplot)** → For linearity, uncorrelated errors, and constant variance
- **Residuals versus fitted values (scatterplot)** → For linearity, uncorrelated errors, and constant variance
- **Normal Quantile-Quantile (QQ) plot** → For Normality

Random bands of residuals indicates no violations

- **Linearity**: Any systematic pattern, but especially curves or other functions of predictors
- **Uncorrelated Errors**: Large clusters of many points or a pattern across time or some other sequencing information
- **Constant Variance**: Any systematic pattern, but especially a fanning pattern with increasing or decreasing spread
- **Normality**: Stark deviations/ curving/wiggling from the diagonal line (minimal deviations ok)

HOW TO MAKE RESIDUAL PLOTS

Extracting Components from Model

1. Fit the model to your data:

```
> model1 <- lm(y1 ~ x1, data = anscombe)
```

2. Extract the fitted/predicted values ($\hat{y}_i$):

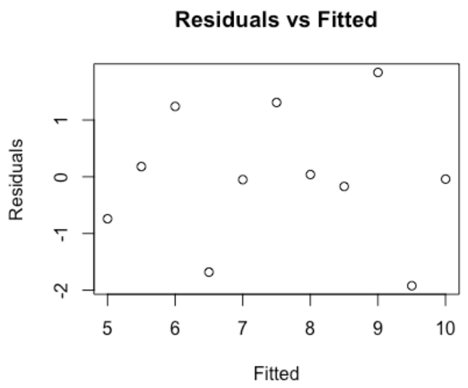
```
> y_hat <- fitted(model1)
> y_hat
        1         2         3         4         5 
 8.001000  7.000818  9.501273  7.500909  8.501091 
        6         7         8         9        10 
10.001364  6.000636  5.000455  9.001182  6.500727 
       11 
 5.500545 
```

3. Extract the residuals from the model ($\hat{e}_i$):

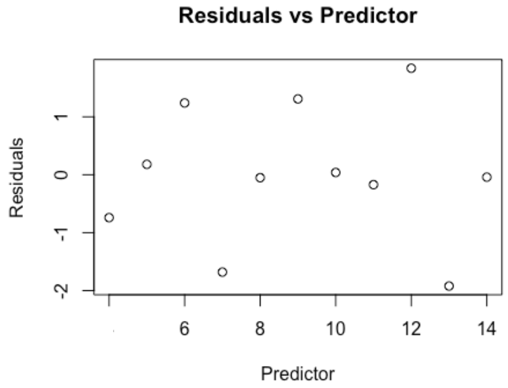
```
> e_hat <- resid(model1)
> e_hat
          1           2           3           4 
 0.03900000 -0.05081818 -1.92127273  1.30909091 
          5           6           7           8 
-0.17109091 -0.04136364  1.23936364 -0.74045455 
          9          10          11 
 1.83881818 -1.68072727  0.17945455 
```

Creating Residual Plots

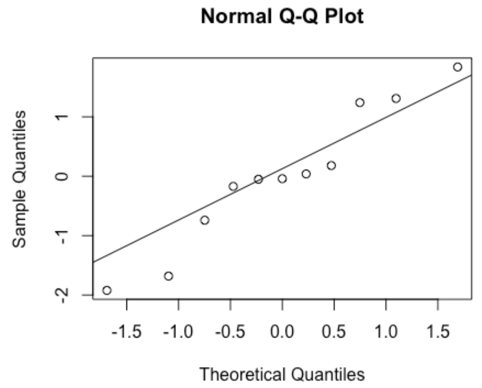
```
> plot(x = y_hat, y = e_hat, main = "Residuals vs Fitted", ylab = "Residuals", xlab = "Fitted")
```




```
> plot(x = anscombe$x1, y = e_hat, main = "Residuals vs Predictor", ylab = "Residuals", xlab = "Predictor")
```




```
> qqnorm(e_hat)
> qqline(e_hat)
```



EXAMPLE OF NO DISTINCT PATTERNS

```
> nyc <- read.csv("nyc.csv", header=T)
> head(nyc)
  Case           Restaurant Price Food Decor Service East
1    1 Daniella Ristorante    43   22    18      20    0
2    2  Tello's Ristorante    32   20    19      19    0
3    3          Biricchino    34   21    13      18    0
4    4             Bottino    41   20    20      17    0
5    5          Da Umberto    54   24    19      21    0
6    6            Le Madri    52   22    22      21    0
```

Response (Y)

Indicator variable
1 = East location
0 = West location

Identifiers

Numerical predictors

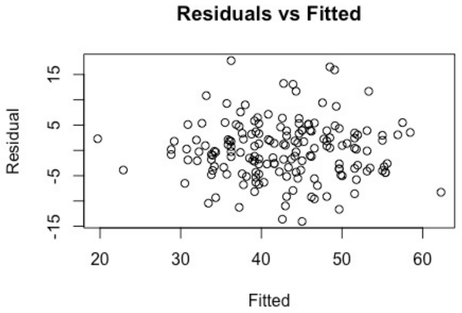
```
> model <- lm(Price ~ Food + Decor + Service + East, data=nyc)
> model

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

Coefficients:
(Intercept)         Food        Decor      Service         East
 -24.023800     1.538120     1.910087    -0.002727     2.068050
```

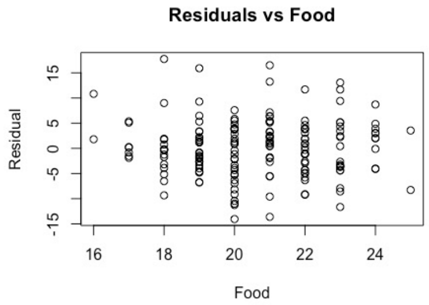
When East = 1, add this to intercept

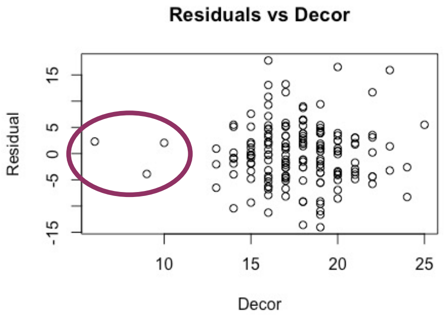
"For a restaurant with a fixed Décor and Service rating and location, the mean Price increases by $1.54 for a one-rating increase in Food"

```
> e_hat <- resid(model)
> y_hat <- fitted(model)
> plot(e_hat ~ y_hat, main="Residuals vs Fitted", xlab="Fitted", ylab="Residual")
```

```
> plot(e_hat ~ nyc$Food, main="Residuals vs Food", xlab="Food", ylab="Residual")
> plot(e_hat ~ nyc$Decor, main="Residuals vs Decor", xlab="Decor", ylab="Residual")
```



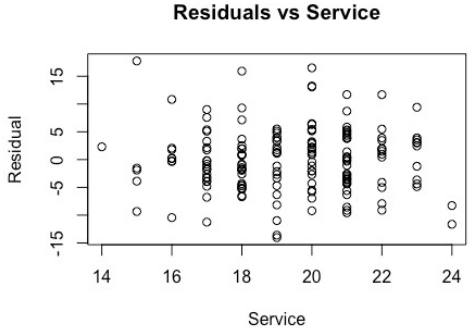


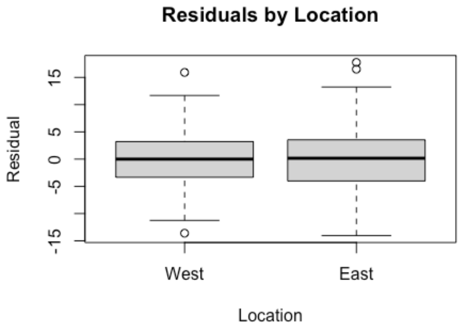


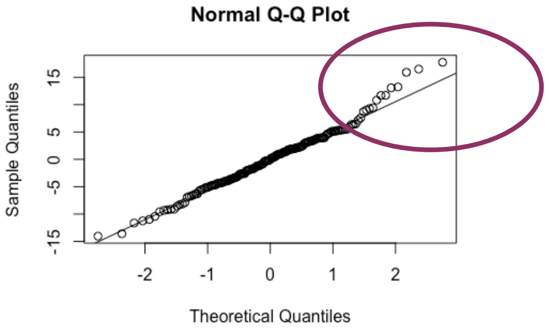

```
> plot(e_hat ~ nyc$Service, main="Residuals vs Service", xlab="Service", ylab="Residual")
```

```
> boxplot(e_hat ~ nyc$East, main="Residuals by Location", xlab="Location", ylab="Residual", names=c("West", "East"))
```

```
> qqnorm(e_hat)
> qqline(e_hat)
```

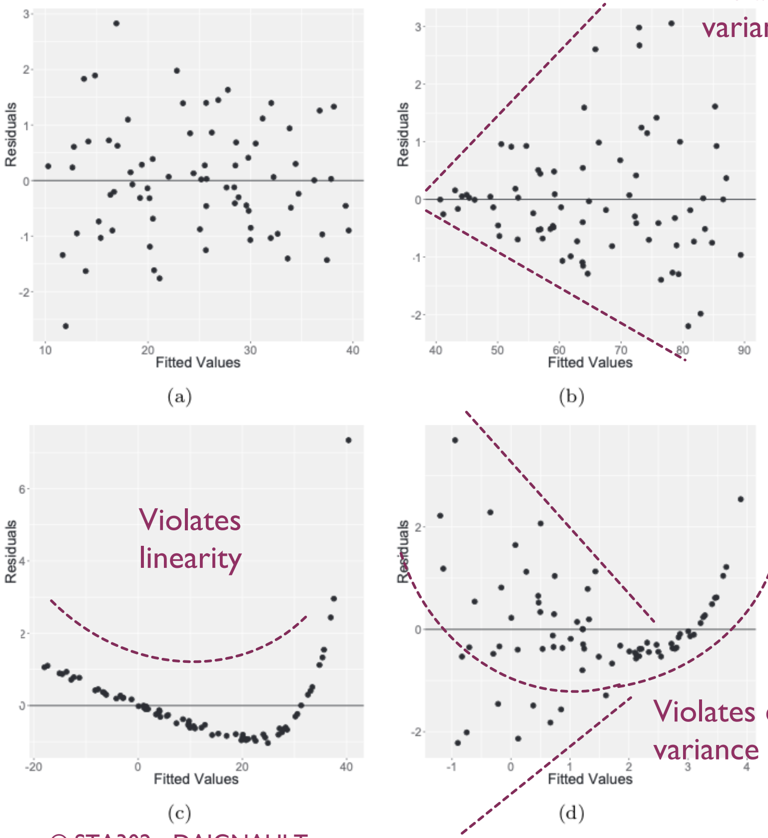






EXAMPLES OF DISTINCT PATTERNS

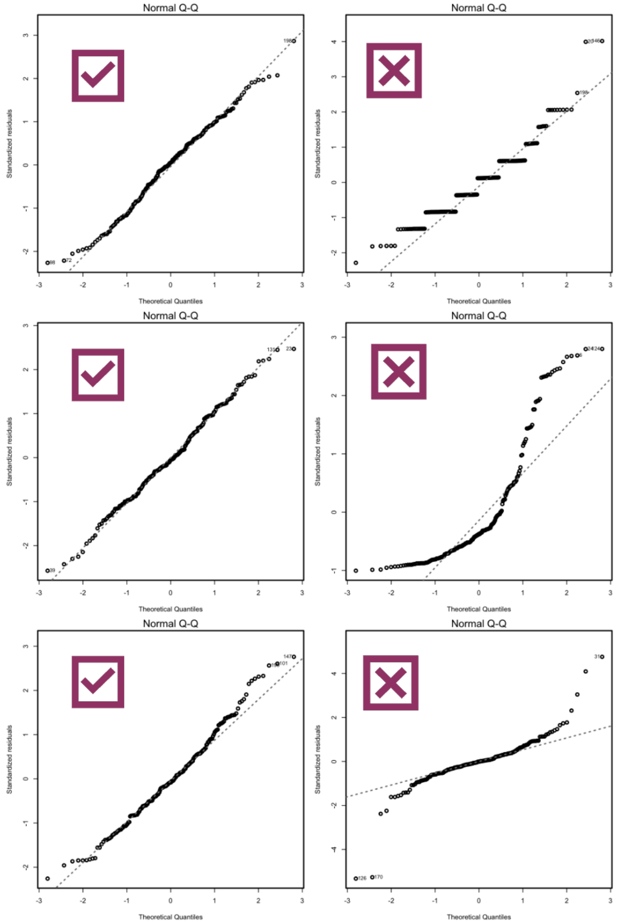
From Young's Handbook of Regression Methods, pg 55

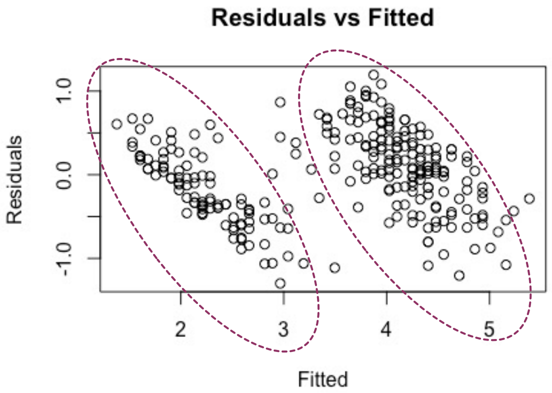


Violates constant variance

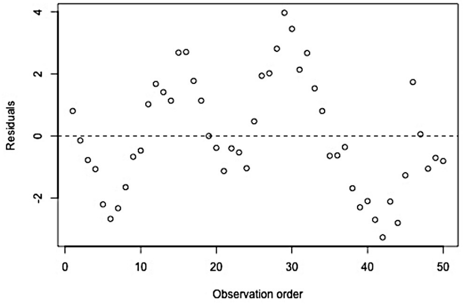
Violates linearity

Violates constant variance and linearity





Violates uncorrelated errors



https://online.stat.psu.edu/stat501/book/export/html/915

García-Portugués, E. (2023). *Notes for Predictive Modeling*. Version 5.9.12.
ISBN 978-84-09-29679-8. Available at https://bookdown.org/egarpor/PM-UC3M/.

EXPLORE AND UNDERSTAND THE DATA

- Assumptions are formally checked using residual plots but knowing the data can also help
- Always conduct an exploratory data analysis before fitting a model
  - Skew in response variable → probably will have an issue with Normality or Linearity
  - Skews in predictor variables → may see an issue with Linearity
  - Think about underlying characteristic → may help deciding how to include predictor
- Existing literature informs your knowledge about true relationship

- Thinking about the data collection and population important too
  - *"Data was collected by voluntary response survey…"*
    - Means likely an issue with linearity assumption
  - *"Each variable was measured every day for a month…"*
    - Means likely an issue with uncorrelated errors
  - *"Neighbourhoods were randomly sampled and all subjects from selected neighbourhoods were included in the study on income inequality…"*
    - Means likely an issue with uncorrelated errors

MODULE 3 OUTLINE

1. Introduction to Linear Regression Assumptions
2. Verifying Assumptions using Residual Plots
3. Additional Conditions for Multiple Linear Models

ADDITIONAL CONDITIONS FOR MLR

- Recall: the interpretation of coefficients involved holding other predictors fixed
  - MLR estimates relationship using predictors jointly
- Certain relationships 1) between $\boldsymbol{Y}$ and $\boldsymbol{X}$, and 2) between predictors must be identified
- In either case, presence causes residual plots to become unreliable
  - Plots can be used to say that the model is not valid
  - **Patterns in plots <u>cannot</u> be used to identify a specific violation and can give misleading conclusions**
- Two conditions must be checked in MLR before using residual plots

1. Conditional mean response condition: the mean responses are a single function of a linear combination involving $\boldsymbol{\beta}$

$$E(Y_i \mid \boldsymbol{X} = \boldsymbol{x_i}) = g(\beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip})$$

   - $E(Y|X) = log(\beta_0 + \beta_1 x_i)$ satisfies this condition
   - $E(Y|X) = {}^{\beta_1 x_{i1}}/_{\beta_2 x_{i2}} = {}^{g_1(x_1)}/_{g_2(x_2)}$ violates it

2. Conditional mean predictor condition: the mean of each predictor is related to each other predictor in no more complicated way than linearly

$$E\left(X_i \middle| X_j\right) = \alpha_0 + \alpha_1 X_j$$

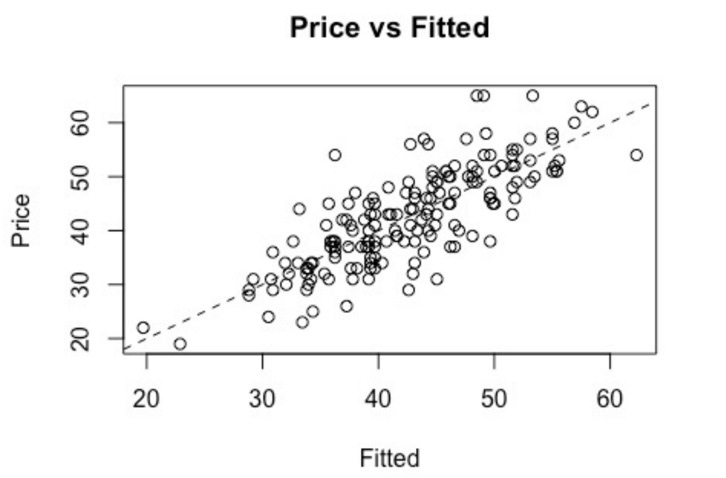
   - Linear or no relationship satisfy condition; anything else violates

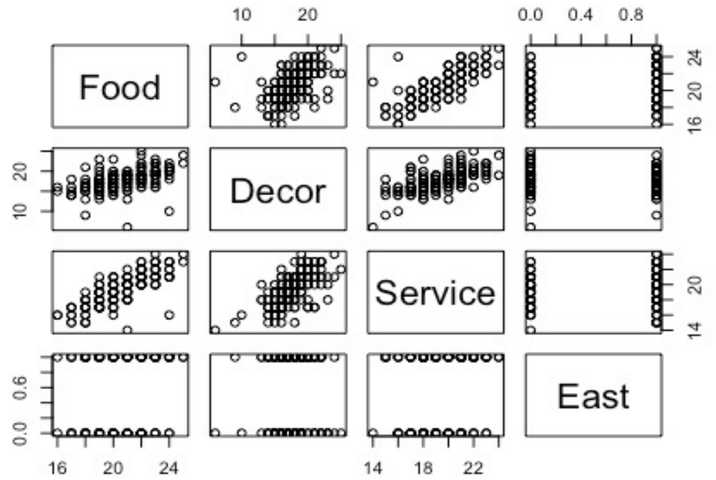
Response versus Fitted Values

```
> plot(x = y_hat, y = nyc$Price, main="Price vs Fitted",
xlab="Fitted", ylab="Price")
> abline(a = 0, b = 1, lty=2)
```

Pairwise Scatterplots

```
> pairs(nyc[,4:7])
```





HOW ARE CONDITIONS CHECKED

1. Conditional mean response

Scatterplot of Response versus Fitted values

Look for random diagonal scatter or an easily identifiable non-linear trend

2. Conditional mean predictors

All pairwise scatterplots of predictors

Look for lack of curves or other non-linear patterns

CONDITION 1 DOES NOT HOLD

- Population (i.e., true relationships):
  - Condition 2 holds:

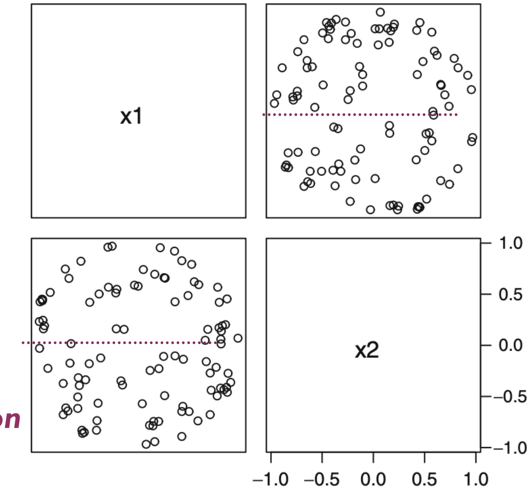
$$E(X_i|X_j) = \alpha_0 + \alpha_1 X_j$$

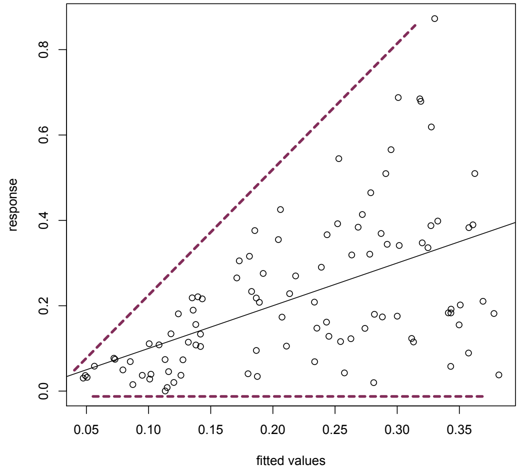
No non-linear pattern

  - Condition 1 fails:

$$\boxed{E(Y|X) = \frac{|x_1|}{2 + (1.5 + x_2)^2} = \frac{g(x_1)}{g(x_2)}}$$

Linear model is not appropriate for this situation





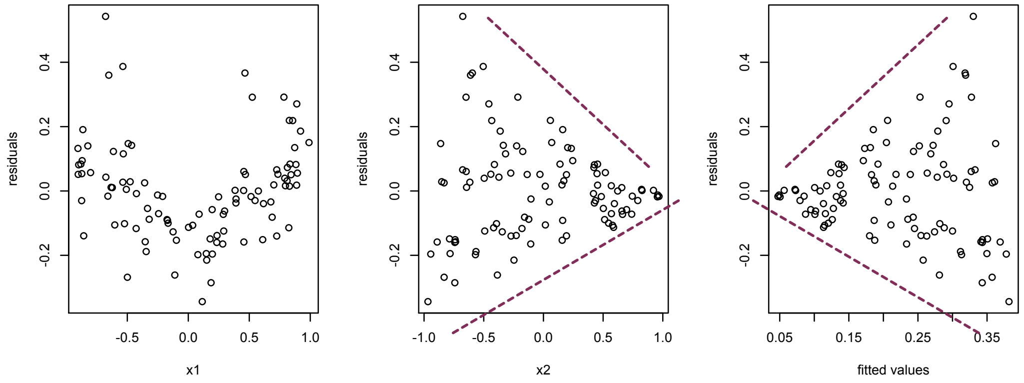
Not a single identifiable pattern/function and not random scatter

- Data are simulated from above so that constant variance is not violated

- Model fit in the sample:

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

- Can look at plots for conditions and residual plots for assumptions

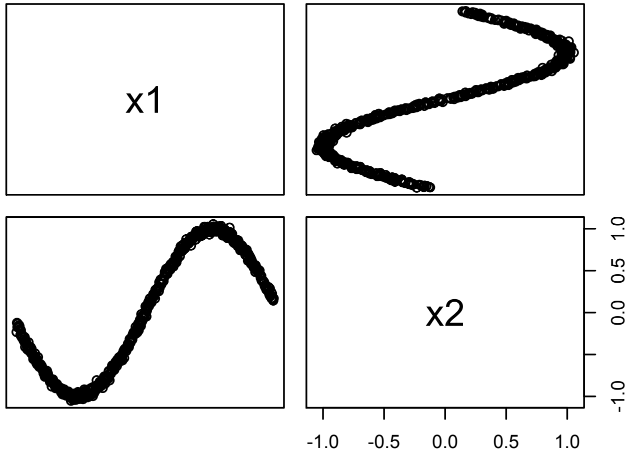


Pattern implies non-constant variance, yet data generated to ensure assumptions hold

CONDITION 2 DOES NOT HOLD

- Population (i.e., true relationship):
  - Condition 1 hold:
    $$Y = x_1 + 3x_2^2 + \varepsilon$$
  - Condition 2 fails:
    $$E(x_2|x_1) = \sin(x_1)$$

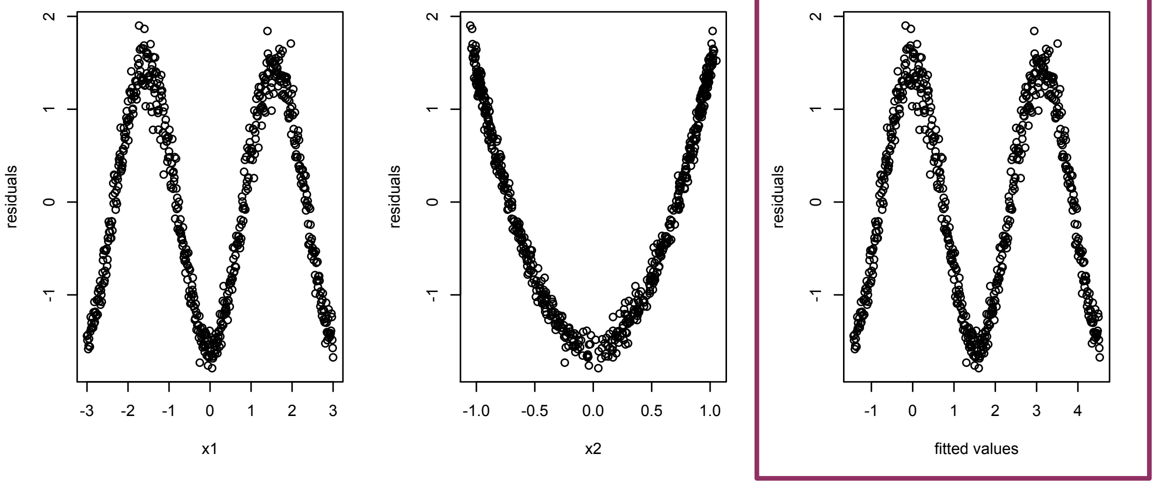
Periodic pattern visible, a non-linear relationship



- Data are simulated from above
- Model fit in the sample:
  $$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + e$$
- We should see that linearity is violated because we are missing a squared term

Quadratic curve in residual vs x_2 plot highlights the missing square term from true relationship

Without checking conditions, would imply also missing a sinusoidal/periodic term as well → Due to failure of condition 2



MODULE TAKE-AWAYS

1. What are the assumptions of linear regression and what do they mean?
2. What tools are used to assess whether each assumption holds?
3. What do we look for to know whether each assumption holds?
4. Why is it important to check the additional conditions in multiple linear models?
5. What other aspects of data, data collection or population characteristics tell us assumption may not hold?