Introduction to

Statistical Data Analysis IV

JULY 2011

Afsaneh Yazdani

Inferences about More than Population Central Values

Analysis of Variance Method:

ANOVA (AOV) is a test for comparing more than ‘2’ populations means, which is developed under following conditions:

Inferences about More than Population Central Values

Analysis of Variance Method:

ANOVA (AOV) is a test for comparing more than '2' populations means, which is developed under following conditions:

- Each of the populations has a normal distribution.
- The variances of the populations are equal.
- Measurements are independent random samples from their respective populations.

Inferences about More than Population Central Values

Analysis of Variance Method:

	Population '1'	…	Population 't'
Sample Values	y_{11}	⋯	y_{t1}
	y_{21}	⋯	y_{t2}
	$\vdots$	⋯	$\vdots$
	y_{1n_1}	⋯	y_{tn_t}
Mean	$\bar{y}_{1.}$	⋯	$\bar{y}_{t.}$

$\rightarrow \bar{y}_{..}$
overall mean

Inferences about More than Population Central Values

Analysis of Variance Method:

Let s_T^2 be the sample variance of the $n_T = \sum_{i=1}^{T} n_i$ measurements y_{ij} (variability of the whole measurements about the overall mean)

$$s_T^2 = \frac{\sum_{i=1}^{t} \sum_{j=1}^{n_t} \left(y_{ij} - \bar{y}_{..}\right)^2}{n_T - 1}$$

Inferences about More than Population Central Values

Analysis of Variance Method:

Let s_T^2 be the sample variance of the $n_T = \sum_{i=1}^{T} n_i$ measurements y_{ij} (variability of the whole measurements about the overall mean)

Total sum of squares

$$s_T^2 = \frac{\sum_{i=1}^{t}\sum_{j=1}^{n_t}\left(y_{ij} - \bar{y}_{..}\right)^2}{n_T - 1}$$

Inferences about More than Population Central Values

Analysis of Variance Method:

$$\sum_{i=1}^{t}\sum_{j=1}^{n_t}(y_{ij}-\bar{y}_{..})^2 = \underbrace{\sum_{i=1}^{t}\sum_{j=1}^{n_t}(y_{ij}-\bar{y}_{i.})^2}_{\text{Within-Sample Sum of squares}} + \underbrace{\sum_{i=1}^{t} n_i(\bar{y}_{i.}-\bar{y}_{..})^2}_{\text{Between-Sample Sum of squares}}$$

$$s_B^2 = \frac{SSB}{t-1}, \quad s_W^2 = \frac{SSW}{n_t - t}$$

Inferences about More than Population Central Values

Analysis of Variance Method:

Source	Sum of Squares	Degrees of Freedom	Mean Square	F Test
Between Samples	SSB	$t-1$	$s_B^2 = SSB/(t-1)$	s_B^2/s_W^2
Within Samples	SSW	$n_T - T$	$s_W^2 = SSW/(n_T - t)$	
Total	TSS	$n_T - 1$		

Inferences about More than Population Central Values

Analysis of Variance Method:

Test Statistic: $F = \frac{s_B^2}{s_w^2} \sim F(t-1, n_T - t)$

$H_0: \mu_1 = \cdots = \mu_t$

H_a: at least one of the 't' population means differ from the rest

- Reject H_0 if 'F' exceeds $F_{\alpha,(t-1),(n_T-t)}$

Inferences about More than Population Central Values

Checking on AOV Conditions:

- Equality of the population variances
  - Using Hartely's of BFL Test (Brown-Forsythe-Levene)
  - When the sample sizes are nearly equal ,
    this assumption is less critical

Inferences about More than Population Central Values

Checking on AOV Conditions:

- Equality of the population variances
  - Using Hartely's of BFL Test (Brown-Forsythe-Levene)
  - When the sample sizes are nearly equal ,
    this assumption is less critical

Using a transformation to stabilize the variance

Inferences about More than Population Central Values

Checking on AOV Conditions:

- Equality of the population variances
  - Using Hartely's of BFL Test
  - When the sample sizes are nearly equal , this assumption is less critical
- Normality
  - Using graphs and normality tests

Inferences about More than Population Central Values

Checking on AOV Conditions:

- Equality of the population variances
  - Using Hartely's of BFL Test
  - When the sample sizes are nearly equal , this assumption is less critical

- Normality
  - Using graphs and normality tests

Using Kruskal-Wallis Test

Inferences about More than Population Central Values

Checking on AOV Conditions:

- Equality of the population variances
  - Using Hartely's of BFL Test
  - When the sample sizes are nearly equal , this assumption is less critical
- Normality
  - Using graphs and normality tests
- Independence
  - Careful review of how measurements has been gathered

Inferences about More than Population Central Values

Multiple Comparisons:

If $H_0: \mu_1 = \cdots = \mu_t$ is rejected, we want to know which means differ from each other.

Inferences about More than Population Central Values

Multiple Comparisons:

If $H_0: \mu_1 = \cdots = \mu_t$ is rejected, we want to know which means differ from each other.

→ **Multiple-Comparison Procedures**

Multiple Comparisons Procedures

- Fisher's Least Significant Difference (LSD)
- Tukey's W
- Student–Newman–Keuls

Procedures for Pairwise Comparisons of 't' Population Means

Inferences about More than Population Central Values

Multiple Comparisons Procedures

- Fisher's Least Significant Difference (LSD)
- Tukey's W
- Student–Newman–Keuls

Tukey is more conservative than LSD and SNK

Inferences about More than Population Central Values

Multiple Comparisons Procedures

- Fisher's Least Significant Difference (LSD)
- Tukey's W
- Student–Newman–Keuls

Tukey's limitation is that, it must be based on equal size 'n' from each population

Inferences about More than Population Central Values

Multiple Comparisons Procedures

- Fisher's Least Significant Difference (LSD)
- Tukey's W (Tukey-Kramer W^*)
- Student–Newman–Keuls

Tukey's limitation
is that, it must be based
on equal size 'n' from
each population

Multiple Comparisons Procedures

- Scheffe's Method

More general procedure that can be used to make all possible comparisons among the '*t*' population means.

Inferences about More than Population Central Values

Multiple Comparisons Procedures

- Scheffe's Method

- More conservative Procedure
- Can also be used for constructing a simultaneous confidence interval for all possible (not necessarily pairwise) contrasts using the 't' population means.

Inferences about More than Population Central Values

Multiple Comparisons Procedures

- Fisher's Least Significant Difference (LSD)
- Tukey's W (Tukey-Kramer W^*)
- Student–Newman–Keuls
- Scheffe's Method

- Kruskal–Wallis Nonparametric Procedure

Categorical Data

Categorical Data

We sometimes encounter situations in which levels of a variable of interest are identified by:

- Name
- Rank
- Number of observations occurred at each level of variable, …

Categorical Data

We sometimes encounter situations in which levels of a variable of interest are identified by:

- Name
- Rank
- Number of observations occurred at each level of variable, …

Categorical or Count Data

Categorical Data

Inferences about a Population Proportion ‘π’

In binomial experiment, the probability distribution of ‘y’ (number of success in ‘n’ identical trials) is:

$$P(y) = \frac{n!}{y!\,(n-y)!}\pi^{y}(1-\pi)^{n-y}$$

Inferences about a Population Proportion ‘π’

In binomial experiment, the probability distribution of ‘y’ (number of success in ‘n’ identical trials) is:

$$P(y) = \frac{n!}{y!\,(n-y)!}\pi^{y}(1-\pi)^{n-y}$$

Probability of Success

Categorical Data

Inferences about a Population Proportion 'π'

In binomial experiment, the probability distribution of 'y' (number of success in 'n' identical trials) is:

$$P(y) = \frac{n!}{y!\,(n-y)!}\pi^{y}(1-\pi)^{n-y}$$

$$\mu_{\hat{\pi}} = \pi \quad , \quad \sigma_{\hat{\pi}} = \sqrt{\frac{\pi(1-\pi)}{n}}$$

Confidence Interval for 'π' with Confidence Coefficient of (1-α)

$$\left(\hat{\pi} - z_{\frac{\alpha}{2}}\,\hat{\sigma}_{\hat{\pi}}\ ,\ \hat{\pi} + z_{\frac{\alpha}{2}}\,\hat{\sigma}_{\hat{\pi}}\right)$$

Where

$$\hat{\pi} = \frac{y}{n} \text{ and } \hat{\sigma}_{\hat{\pi}} = \sqrt{\frac{\hat{\pi}(1-\hat{\pi})}{n}}$$

Categorical Data

Sample Size Required for a $100(1-\alpha)\%$ Confidence Interval for 'π' of the form $\hat{\pi} \pm E$

$$\mathbf{n} = \frac{z_{\frac{\alpha}{2}}^{2}\,\pi(1-\pi)}{E^{2}}$$

Categorical Data

Statistical Test for ‘π’

(Under H_0, $\sigma_{\hat{\pi}} = \sqrt{\pi_0(1-\pi_0)/n}$, and ‘n’ must satisfy both $n\pi_0 \geq 5$ and $n(1-\pi_0) \geq 5$)

Test Statistic: $z = \dfrac{\hat{\pi}-\pi_0}{\sigma_{\hat{\pi}}} \sim \mathrm{N}(0,1)$

$\begin{cases} \mathrm{H_0}: \pi \leq \pi_0 \\ \mathrm{H_a}: \pi > \pi_0 \end{cases}$	• Reject $\mathrm{H_0}$ if $z > -z_\alpha$
$\begin{cases} \mathrm{H_0}: \pi \geq \pi_0 \\ \mathrm{H_a}: \pi < \pi_0 \end{cases}$	• Reject $\mathrm{H_0}$ if $z < -z_\alpha$
$\begin{cases} \mathrm{H_0}: \pi = \pi_0 \\ \mathrm{H_a}: \pi \neq \pi_0 \end{cases}$	• Reject $\mathrm{H_0}$ if $\lvert z\rvert > -z_{\frac{\alpha}{2}}$

Categorical Data

Inferences about two populations proportions

	Population 1	Population 2
Population Proportion	π_1	π_2
Sample Size	n_1	n_1
Number of Success	y_1	y_2
Sample Proportion	$\hat{\pi}_1 = \frac{y_1}{n_1}$	$\hat{\pi}_2 = \frac{y_2}{n_2}$

Confidence Interval for '$\pi_1 - \pi_2$' with Confidence Coefficient of $(1-\alpha)$

$$(\hat{\pi}_1 - \hat{\pi}_2) \pm z_{\frac{\alpha}{2}}\, \hat{\sigma}_{\hat{\pi}_1 - \hat{\pi}_2}$$

Where

$$\hat{\sigma}_{\hat{\pi}_1 - \hat{\pi}_2} = \sqrt{\frac{\hat{\pi}_1(1-\hat{\pi}_1)}{n_1} + \frac{\hat{\pi}_2(1-\hat{\pi}_2)}{n_2}}$$

Categorical Data

Statistical Test for '$\pi_1 - \pi_2$'

(Under H_0, $\hat{\sigma}_{\hat{\pi}_1 - \hat{\pi}_2} = \sqrt{\hat{\pi}_1(1-\hat{\pi}_1)/n_1 + \hat{\pi}_2(1-\hat{\pi}_2)/n_2}$,

'n_1' and 'n_2' must satisfy both $n\pi_0 \geq 5$ and $n(1-\pi_0) \geq 5$)

Test Statistic: $z = \dfrac{\hat{\pi}_1 - \hat{\pi}_2}{\hat{\sigma}_{\hat{\pi}_1 - \hat{\pi}_2}} \sim \mathrm{N}(0, 1)$

$\begin{cases} \mathrm{H}_0: \pi_1 \leq \pi_2 \\ \mathrm{H}_a: \pi_1 > \pi_2 \end{cases}$	• Reject H_0 if $z > -z_\alpha$
$\begin{cases} \mathrm{H}_0: \pi_1 \geq \pi_2 \\ \mathrm{H}_a: \pi_1 < \pi_2 \end{cases}$	• Reject H_0 if $z < -z_\alpha$
$\begin{cases} \mathrm{H}_0: \pi_1 = \pi_2 \\ \mathrm{H}_a: \pi_1 \neq \pi_2 \end{cases}$	• Reject H_0 if $\lvert z \rvert > -z_{\frac{\alpha}{2}}$

Categorical Data

Inferences about 'k' proportions

(Chi-square Goodness-of-Fit Test, where $E_i = n\pi_{i0}$)

Test Statistic: $$\chi^2 = \sum \frac{(n_i - E_i)^2}{E_i} \sim \chi^2_{\alpha}(k-1)$$

H_0: $\pi_i = \pi_{i0}$ for categories $i = 1, \ldots, k$, π_{i0} are specified probabilities or proportions

H_a: At least one of the cell probabilities differs from the hypothesized values

Categorical Data

Contingency Tables (Cross Tabulations)

		Variable 2			
		Level 1	…	Level c	
Variable 1	Level 1	n_{11}	$\cdots$	n_{1c}	$n_{1.}$
	$\vdots$	$\vdots$		$\vdots$	
	Level r	n_{r1}	$\cdots$	n_{rc}	$n_{r.}$
		$n_{.1}$	$\cdots$	$n_{.c}$	$n_{..}$

Categorical Data

Contingency Tables (Cross Tabulations)

		Variable 2			
		Level 1	…	Level c	
Variable 1	Level 1	n_{11}	$\cdots$	n_{1c}	$n_{1.}$
	$\vdots$	$\vdots$		$\vdots$	
	Level r	n_{r1}	$\cdots$	n_{rc}	$n_{r.}$
		$n_{.1}$	$\cdots$	$n_{.c}$	$n_{..}$

Dependence of variables means that one variable has some value for predicting the other

Categorical Data

Test of Independence

Test Statistic:

$$\chi^2 = \sum_{i=1}^{r}\sum_{j=1}^{c}\frac{(n_{ij}-\hat{E}_{ij})^2}{\hat{E}_{ij}} \sim\chi_{\alpha}^{2}[(r-1)(c-1)]$$

H_0: The row and column variables are independent

H_a: The row and column variables are dependent (associated)

Categorical Data

Test of Independence

Test Statistic:

$$\chi^2 = \sum_{i=1}^{r} \sum_{j=1} \frac{(n}{}$$

H_0: The row and col

H_a: The row and column
(associated)

There is an alternative statistic, called the **likelihood ratio statistic** that is often shown in computer outputs.

Categorical Data

Measuring Strength of Relation

- Kendall's Tau Correlation Coefficient
- Contingency Coefficient
- Spearman's Ranked Correlation Coefficient
- Phi's Coefficient
- Cramer's V

Linear Regression

Linear Regression

Modeling of the relationship between a response variable and a set of explanatory variables.

Regression Analysis

Linear Regression

A regression model provides the user with a **functional relationship** between the response variable and explanatory variables that allows the user to:

❶ Determine which of the explanatory variables have an effect on the response.

❷ Explore what happens to the response variable for specified changes in the explanatory variables.

Linear Regression

Uses of Regression Models:

- Provides a description of data set (which of the explanatory variables affect the response variable)
- Provides estimates of the response variable for values of the explanatory not observed in the study, or expensive to measure
- Prediction

Uses of Regression Models:

The accuracy of the estimates and prediction depends on:

- How well the final model fits the observed data
- Stability of the conditions during which observed data were collected, over the prediction period

- Prediction

Linear Regression

Prediction Versus Explanation

Future Value | **Current or past values**

Explanation is easier than Prediction

Both of them use the connection between explanatory (independent) and response (dependent)

Linear Regression

Simple Regression

There is a single independent variable and the equation for predicting a dependent variable 'y' is a linear function of a given independent variable 'x'.

$$\underbrace{\beta_0}_{\text{Intercept}} \quad \underbrace{\beta_1}_{\text{Slope}} \quad \underbrace{\varepsilon_i}_{\text{Random Error Term}}$$

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$

Intercept

Slope

Random Error Term

Linear Regression

Simple Regression

There is a single independent variable and the equation for predicting a dependent variable ‘y’ is a linear function of a given independent variable ‘x’.

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$

Intercept

Slope

The slope of the equation does not change as ‘x’ changes

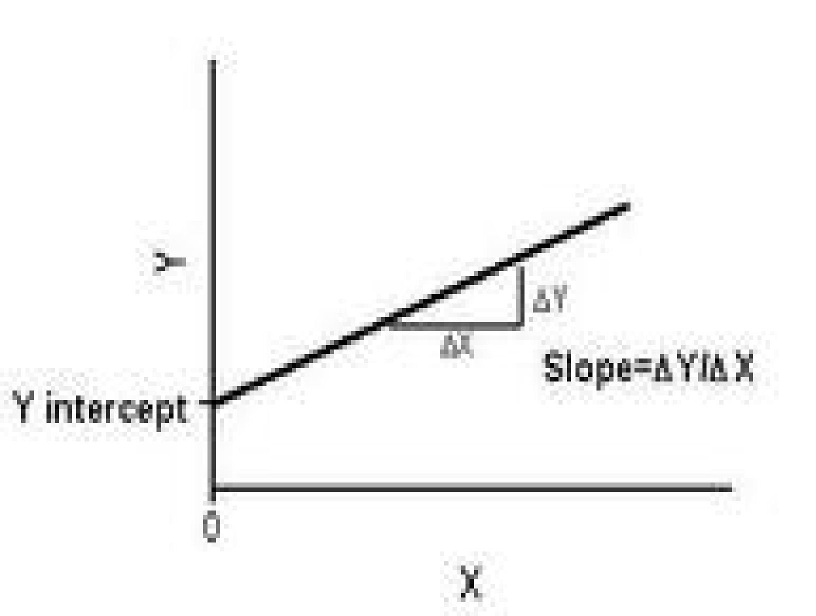
Linear Regression

Simple Regression

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$

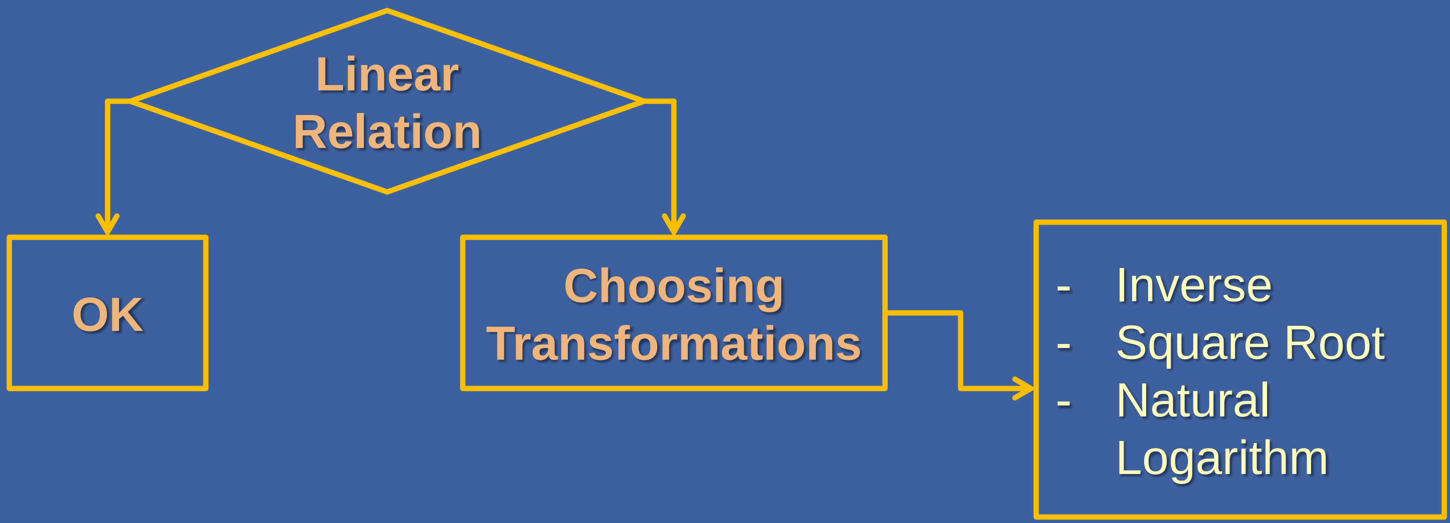
Intercept

Slope



Checking for Linearity

Checking by looking at a scatterplot of data



Linear Regression

Regression Modeling Steps:

1- Specify model and estimate unknown parameters

2- Evaluate model

3- Use model for prediction and estimation

Linear Regression – Specifying Model

Estimating Model Parameters

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$

Least-square estimates for slope and intercept:

$$\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}} \quad , \quad \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

$$S_{xy} = \sum_i (x_i - \bar{x})(y_i - \bar{y}) \text{ and } S_{xx} = \sum_i (x_i - \bar{x})^2$$

Linear Regression – Specifying Model

Estimating Model Parameters

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$

The estimate of the regression slope can potentially be greatly affected by "high leverage points".

Linear Regression – Specifying Model

Estimating Model Parameters

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$

The estimate of the regression slope can potentially be greatly affected by "high leverage points".

Points that have
very high or very low values
of independent variables

Linear Regression – Specifying Model

Estimating Model Parameters

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$

The estimate of the regression slope can potentially be greatly affected by "high leverage points".

High leverage point whose 'y' value is outlier,
is "High Influence Point"

Linear Regression – Specifying Model

Estimating Model Parameters

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$

The estimate of the regression slope can potentially be greatly affected by "high leverage points".

A 'y' which is outlier can not much affect the slope, if it is not a "high influence point"

Linear Regression – Evaluating Model

Inferences about Regression Parameters

Test Statistic: $t = \frac{\hat{\beta}_1 - 0}{s_\varepsilon / \sqrt{S_{xx}}} \sim t_\alpha(n-2)$

$\begin{cases} H_0: \beta_1 \leq 0 \\ H_a: \beta_1 > 0 \end{cases}$	• Reject H_0 if $t > t_\alpha$
$\begin{cases} H_0: \beta_1 \geq 0 \\ H_a: \beta_1 < 0 \end{cases}$	• Reject H_0 if $t < -t_\alpha$
$\begin{cases} H_0: \beta_1 = 0 \\ H_a: \beta_1 \neq 0 \end{cases}$	• Reject H_0 if $\lvert t \rvert > t_{\frac{\alpha}{2}}$

Linear Regression – Evaluating Model

Inferences about Regression Parameters

Test Statistic: $$\mathrm{F} = \frac{MS(Regresstion)}{MS(Residual)} \sim F_{\alpha}\,(1, n-2)$$

$\begin{cases} \mathrm{H_0}: \beta_1 \leq 0 \\ \mathrm{H_a}: \beta_1 > 0 \end{cases}$	• Reject $\mathrm{H_0}$ if $\mathrm{t} > t_{\alpha}$
$\begin{cases} \mathrm{H_0}: \beta_1 \geq 0 \\ \mathrm{H_a}: \beta_1 < 0 \end{cases}$	• Reject $\mathrm{H_0}$ if $t < -t_{\alpha}$
$\begin{cases} \mathrm{H_0}: \beta_1 = 0 \\ \mathrm{H_a}: \beta_1 \neq 0 \end{cases}$	• Reject $\mathrm{H_0}$ if $\lvert\mathrm{t}\rvert > t_{\frac{\alpha}{2}}$

Linear Regression – Evaluating Model

Inferences about Regression Parameters

Test Statistic:

$$t = \frac{\hat{\beta}_0 - 0}{s_\varepsilon / \sqrt{\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}}} \sim t_\alpha(n-2)$$

$\begin{cases} H_0: \beta_0 \le 0 \\ H_a: \beta_0 > 0 \end{cases}$	• Reject H_0 if $t > t_\alpha$
$\begin{cases} H_0: \beta_0 \ge 0 \\ H_a: \beta_0 < 0 \end{cases}$	• Reject H_0 if $t < -t_\alpha$
$\begin{cases} H_0: \beta_0 = 0 \\ H_a: \beta_0 \ne 0 \end{cases}$	• Reject H_0 if $\lvert t \rvert > t_{\frac{\alpha}{2}}$

Linear Regression – Evaluating Model

Examining the model using 'R^2':

$$R^2 = \frac{Explained\,Variation}{Total\,Variation} = \frac{SS_{Model}}{SS_{Total}}$$

$$R^2_{Adj} = 1 - \frac{MS_{Residual}}{MS_{Total}}$$

Linear Regression – Evaluating Model

Examining the model using 'R^2':

$$R^2 = \frac{Explained\,Variation}{Total\,Variation} = \frac{SS_{Model}}{SS_{Total}}$$

$$R^2_{Adj} = 1 - \frac{MS_{Residual}}{MS_{Total}}$$

Value closer to '1'
the model explains the variation more

Linear Regression - Evaluation

Assumptions of Regression Analysis

- The relation is linear, so that the errors all have expected value zero ($E(\varepsilon_i) = 0$; for all 'i')
- The errors are independent of each other.
- The errors are all normally distributed
- The errors all have the same variance ($Var(\varepsilon_i) = \sigma^2$)

$$\varepsilon_i \sim N\,(0\,, \sigma^2)$$

Linear Regression – Evaluating Model

Checking Regression Assumptions

1- Linearity

- Draw Residual Plot Versus $\hat{y}_i = b_0 + b_1 x_i$ to check for existence of non-linearity pattern

- Using F-Test, where $F^* = \dfrac{MS_{Lack}}{MS_{Pure\ Experimental}}$, ($H_0$: A linear regression is appropriate)

Linear Regression – Evaluating Model

Checking Regression Assumptions

2- Independency of residuals

- Draw Residual Plot Versus Observation Number

- Using Durbin Watson

Linear Regression – Evaluating Model

Checking Regression Assumptions

2- Independency of residuals

- Draw Residual Plot Versus Observation Number

- Using Durbin Watson

values of *d* less than approximately
1.5 (or greater than approximately 2.5) lead
one to suspect positive (or negative)
serial correlation.

Linear Regression – Evaluating Model

Checking Regression Assumptions

3- Normality of Residuals

- Draw Q-Q Plot, or Box-Plot of Residuals

- Using Normality Tests (such as Kolmogrov-Smirnof, Shapiro Wilk, …)

Linear Regression – Evaluating Model

Checking Regression Assumptions

4- Homogeneous Residuals' Variance

- Draw Residuals Versus x_i

- Divide the observations into two groups, then test the equality of variance of the groups

Linear Regression – Estimation and Prediction

Confidence interval for $\mathrm{E}(y_{n+1})$

$$\hat{y}_{n+1} \pm t_{\frac{\alpha}{2}} S_{\varepsilon} \sqrt{\frac{1}{n} + \frac{(x_{n+1} - \bar{x})^2}{S_{xx}}}$$

It is easier to estimate an average value E(y) than predict an individual ‘y’ value.

Linear Regression – Estimation and Prediction

Prediction interval for y_{n+1}

$$\hat{y}_{n+1} \pm t_{\frac{\alpha}{2}} S_{\varepsilon} \sqrt{1 + \frac{1}{n} + \frac{(x_{n+1} - \overline{x})^2}{S_{xx}}}$$

Linear Regression

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$

Multivariate Regression
When there are more than one response variables

Multiple Regression
When there are more than one explanatory variables