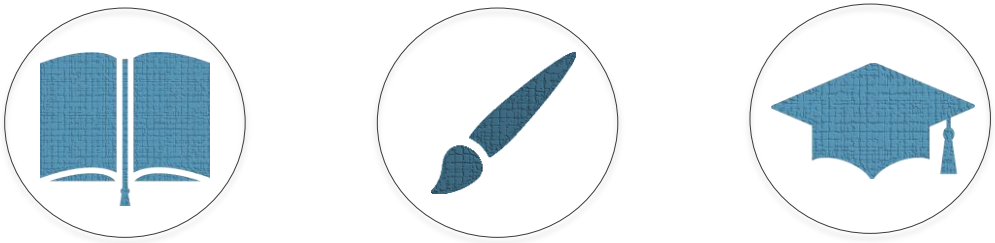


Data Science Concepts

Lesson02–Regression Concepts

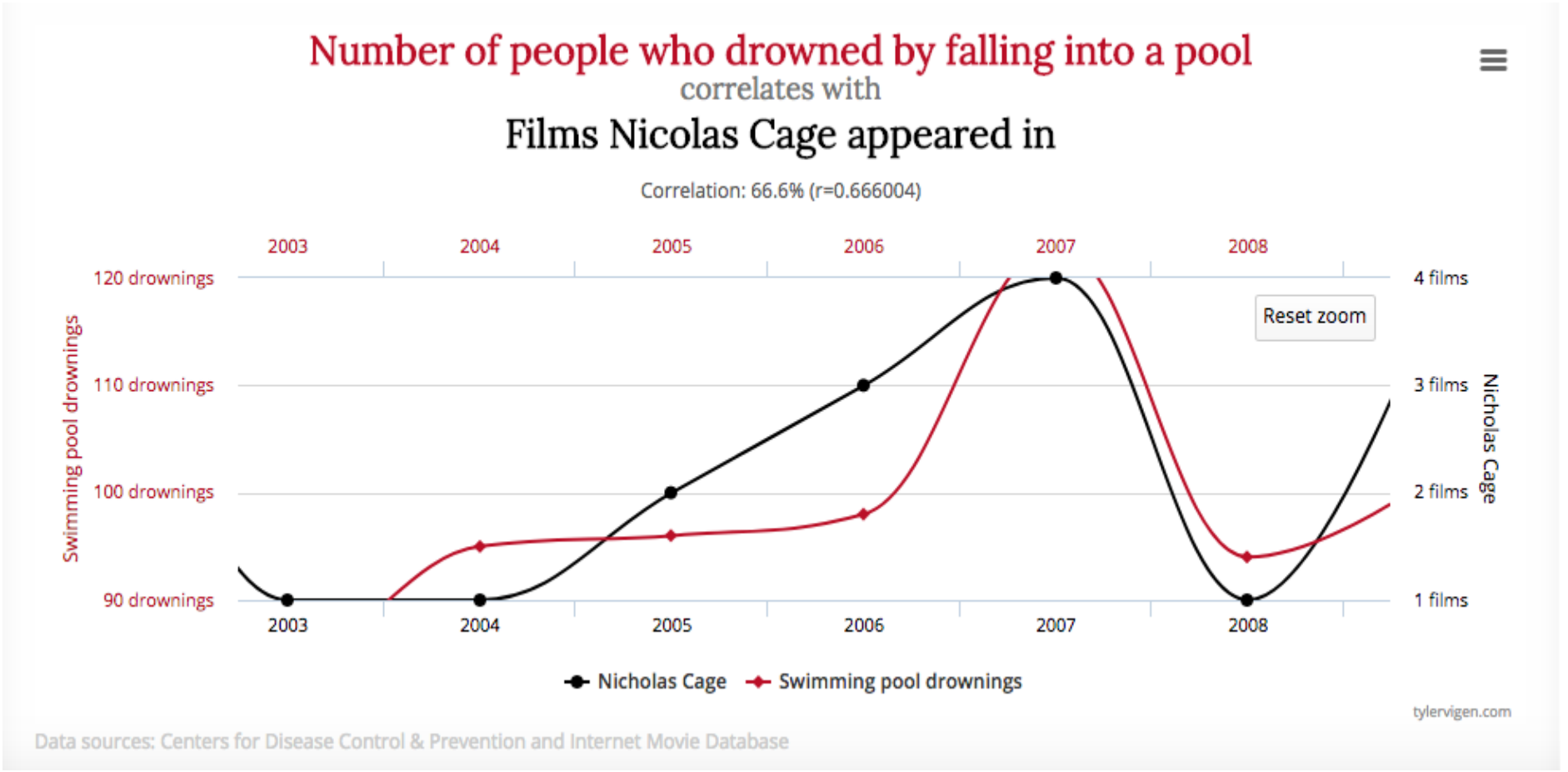
Objective

After completing this lesson you will be able to:

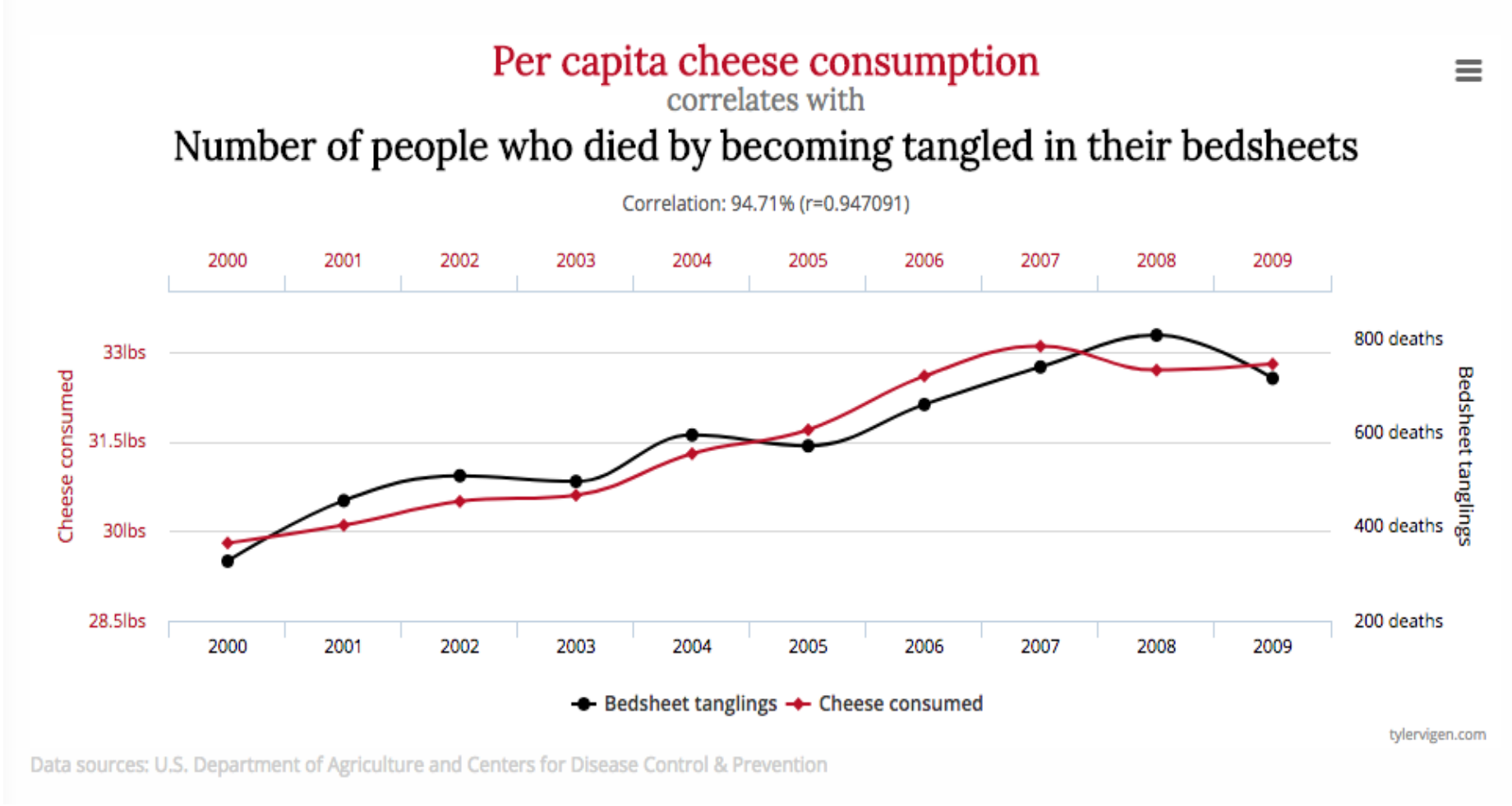
- Explain Regression analysis
- Describe the assumptions of linear regression
- Explain the need of transformations of data
- Understand the representation of qualitative variables in linear regression

Married men earn more money

Which is a dependent variable and which one an independent variable?



Source : http://www.tylervigen.com/spurious-correlations



Source : http://www.tylervigen.com/spurious-correlations

CRIME A BLOG ABOUT MURDER, THEFT, AND OTHER WICKEDNESS. | JULY 9 2013 2:59 PM

When Ice Cream Sales Rise, So Do Homicides. Coincidence, or Will Your Next Cone Murder You?

By Justin Peters

*Crime is **Slate**'s crime blog. Like us on **Facebook**, and follow us on Twitter **@slatecrime**.*

JUSTIN PETERS

97% Correlation between these two. But correlation is not causation!!!

"The England tour has been an eye-opener for everyone. From whatever information we have gathered, it's been seen that even if players wanted to focus on their cricket, their wives were being a big distraction.

- Indian official, after test series loss to England 2014

The essence of regression analysis is to be use the information available about surrounding (independent variables) to better predict an outcome (dependent variable).

Regression–Building the Concept

	Weekly family income X (Rs.)									
X	**800**	**1000**	**1200**	**1400**	**1600**	**1800**	**2000**	**2200**	**2400**	**2600**
Weekly expenditure (Rs.) Y	550	650	790	800	1020	1100	1200	1350	1370	1500
	600	700	840	930	1070	1150	1360	1370	1450	1520
	650	740	900	950	1100	1200	1400	1400	1550	1750
	700	800	940	1030	1160	1300	1450	1520	1650	1780
	750	850	980	1080	1180	1350	-	1570	1750	1800
	-	880	-	1130	1250	1400	-	1600	1890	1850
	-		-	1150	-	-	-	1620	-	1910
Total	3250	4620	4450	7070	6780	7500	6850	10430	9660	12110
E(Y|X)	650	770	890	1010	1130	1250	1370	1490	1610	1730

- The unconditional mean i.e. E(Y) = 72720/60 = 1212.
- The essence of regression analysis is to be use the knowledge of income level to better predict the weekly expenditure.

Regression–Population Regression Function

E(Y|X) is called the population regression function and tells how the mean response of Y varies with X.

The first assumption of PRF is a linear function of X:

$$E(Y|X_i) = \beta_1 + \beta_2 * X_i$$

- β_1 is the estimated average value of Y when the value of X is zero. More often than not it does not have a physical interpretation
- β_2 is the estimated change in the average value of Y as a result of a one-unit change in X.

Linearity for regression assumes linearity in beta values and not in X variables. Example of non linear form: Y = β0 + 1/ (β1 + β2X1) + X2β3 + ε.

Regression–Sample Regression Function

Generally the information available will be a randomly selected sample of Y values for fixed X values.

Y (Exp)	X (Inc)
700	800
650	1000
900	1200
950	1400
1100	1600
1150	1800
1200	2000
1400	2200
1550	2400
1500	2600

Y (Exp)	X (Inc)
550	800
880	1000
900	1200
800	1400
1180	1600
1200	1800
1450	2000
1350	2200
1450	2400
1750	2600

Sample regression function (SRF) takes the form:

$$\widehat{Y_i} = \widehat{\beta_1} + \widehat{\beta_2} * \widehat{X_i}$$

where

- $\widehat{Y_i}$ = estimator of $E(Y|X_i)$
- $\widehat{\beta_1}$ = estimator of β_1
- $\widehat{\beta_2}$ = estimator of β_2

Regression–Sample Regression Function

Method of ordinary least squared (OLS) is used to choose SRF in such a way that $\sum \hat{u}_i^2 = \sum (Y_i - \widehat{Y_i})^2$ is minimized.

The equation obtained

$$\widehat{Y_i} = \widehat{\beta_1} + \widehat{\beta_2} * \hat{X}_i$$

will have following properties:

- The sum of the squared residuals is a minimum.
- The sum of the residuals from the least squares regression line is 0.
- The simple regression line always passes through the sample mean of the Y and X variable.

Objective is to not only estimate $\widehat{\beta_1}$ and $\widehat{\beta_2}$ but also ensure it is close as possible to the true β_1 and β_2?

Regression Model–Validity and Usefulness

Will the model work on the population data? Is the model generalizable and useful?

Is the model valid?

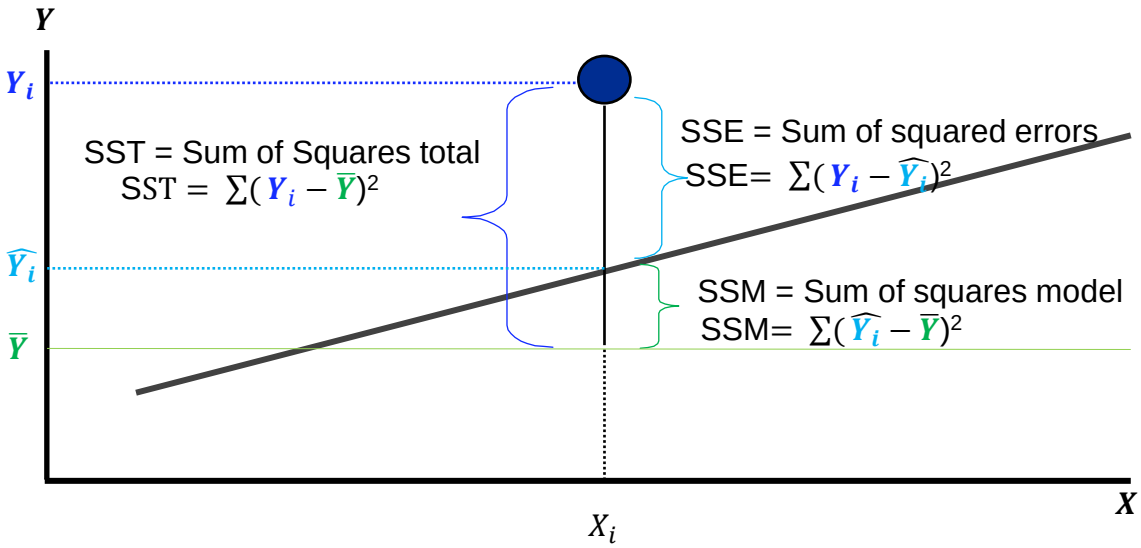
- Use of co-efficient of determination to check the goodness of fit of regression.
- Precision of OLS estimates and t-tests to validate the beta coefficients are significant
- Analysis of Variance (ANOVA) and F test to check the overall fitness of the regression model.
- Residual analysis to check the model adequacies and Multicollinearity

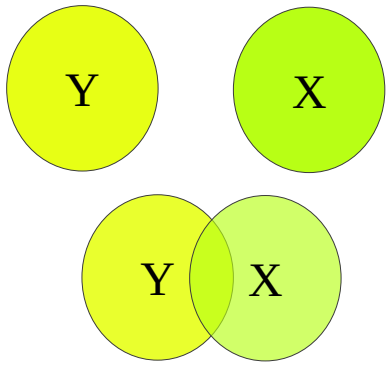
Is the model useful?

- Is the confidence interval estimating the average value of Y for a given value of X?
- Is the prediction interval estimating the individual Y for a given value of X?
- Is the prediction inline with the natural belief?

This is the gist of the assumptions of Classical Linear Regression Model (CLRM). More on the assumptions at: http://www.ats.ucla.edu/stat/stata/webbooks/reg/chapter2/statareg2.htm

Is the model valid–Goodness of fit test





Venn diagram representation

- Coefficient of determination is a measure of the extent to which the variation in Y is explained by X

- The r - squared (coefficient of determination) tells how well the sample regression line fits the data.

$$R^2 = \frac{SSM}{SST} \; where \; 0 \leq R^2 \leq 1$$

Closer the value of R^2 towards 1 more is the variation in Y explained by X

Is the model valid–Goodness of fit test

Regression model for the Expense (Y) and Income (X):

Y (Exp)	X (Inc)
700	800
650	1000
900	1200
950	1400
1100	1600
1150	1800
1200	2000
1400	2200
1550	2400
1500	2600

Y (Weekly Expense) = 244.5 + 0.509* X(Weekly Income)

Regression Statistics	
Multiple R	0.980847369
R Square	0.96206156
Adjusted R Square	0.957319256
Standard Error	64.93003227
Observations	10

Is the Model Valid–Precision of OLS Estimates

OLS estimates $(\widehat{\beta_1}, \widehat{\beta_2})$ are a function of sample data. If sample changes estimates will change. How to get the reliability of the estimate then?

- Precision or reliability of an estimate i.e. $\widehat{\beta_1}\ and\ \widehat{\beta_2}$, is measured by the standard error of the $\widehat{\beta_1}\ and\ \widehat{\beta_2}$

$$\text{se}(\widehat{\beta_2}) = \widehat{\sigma} / \sqrt{\sum {x_i}^2}$$

where $\hat{\sigma} = \sqrt{\frac{\sum {u_i}^2}{n-2}}$; $\hat{\sigma}$ is the measure of standard deviation of y values about the estimated regression line. Also called standard error of the regression.

Standard error of beta coefficients goes on to decide the range in which the population beta coefficients may fall (in repeated sampling). Smaller the SE better is the range.

Is the Model Valid–Precision of OLS Estimates

How to know that the sample beta coefficient ($\widehat{\beta_2}$) is a true estimate of the population beta coefficient (β_2)?

- t-Test on the beta co-efficient with the following hypothesis:

$$\boldsymbol{H_0}: \beta_2 = 0; \boldsymbol{H_a}: \beta_2 \# 0$$

Decision rule: Reject H_0 if |t| $> t_{\alpha/2,df}$ where $|t| = (\widehat{\beta_2} - 0)/\text{s.e}(\widehat{\beta_2})$

P-value corresponding to the t-stats helps make decision. P < .05, reject the null hypothesis.

Is the Model Valid–Precision of OLS Estimates

Regression model for the Expense (Y) and Income (X)

ε

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	244.5454545	64.13817299	3.812791091	0.005142172	96.64256241	392.4483467
Weekly_income (X)	0.509090909	0.035742806	14.24317115	5.75275E-07	0.42666785	0.591513968

Y (Exp)	X (Inc)
700	800
650	1000
900	1200
950	1400
1100	1600
1150	1800
1200	2000
1400	2200
1550	2400
1500	2600

- The confidence interval range (.4266, .5915) suggests that if we do repeated sampling, then in 95 out of 100 cases the above interval will contain the true beta.
- The larger the standard error, greater is the uncertainty of estimating true beta.

Calculating the confidence interval (in Excel):

Lower 95% = 0.50909 - T.INV.2T(0.05,8)*0.03574; upper 95% = 0.50909 - T.INV.2T(0.05,8)*0.03574

How is t – test applicable

$$\widehat{Y_i} = \widehat{\beta_1} + \widehat{\beta_2} * \widehat{X_i} + \in$$

If $\in$ is normally distributed, then from the properties of normal distribution, $\widehat{Y_i}$ is also normally distributed. Also,

$$\widehat{\beta_2} = \frac{\sum_{i=1}^{n}(X_i - \bar{X}) * (Y_i - \bar{Y})}{\sum_{i=1}^{n}(X_i - \bar{X})^2}$$

Which can be shown equivalent to $\widehat{\beta_2} = \sum_{i=1}^{n} k * Y_i$

Since $\widehat{Y_i}$ is normally distributed, $\widehat{\beta_2}$ will follow normal distribution and hence t-test is used as hypothesis testing.

Is the Model Valid–ANNOVA and F Statistics

How to know that the sample coefficients which is an estimate of population coefficients are not simultaneously equal to zero?

Null hypothesis: The error of the intercept-only model and fitted model are equal.

Alternative hypothesis: The error of the fitted model is significantly reduced compared intercept only model.

F test compares a model with no predictors to the model that you specify. A regression model that contains no predictors is also known as an intercept-only model.

Intuition behind F–test

Assume a model with

$$(cost\ of\ treatment \sim age\ +\ past\ medical\ history\ +\ family\ income)$$

If a categorical variable like past medical history has “diabetic”, “hypertension” and “glaucoma” as three level.

$$C.O.T = \beta_0 + \beta_1 * age + \beta_2 * family\ income + \beta_3 * medical\ history(hypertension) + \beta_4 * medical\ history(glaucoma)$$

The t-tests are partial co-efficient tests and does not alone or together, directly test the overall significance of the categorical predictor from which they are derived.

Intuition behind F–test

A F-test is the appropriate test to use when the simultaneous test of the statistical significance of a group of variables is needed.

A F-test requires fitting two regression models.

- A full model—a model that includes all the variables currently of interest
- A reduced model—a model that includes all the variables currently of interest except those whose statistical significance is to be tested. In this case, the reduced model is a null model i.e. without any predictor.

Intuition behind F–test

$$F = \frac{\frac{SSE(reduced\ model) - SSE(full\ model)}{(\Delta \neq regressors(full - reduced)}}{\frac{SSE(full\ model)}{df\ error\ (full\ model)}}$$

F has an F-distribution with

- numerator degrees of freedom = $\Delta \neq regressors$
- denominator degrees of freedom = the degrees of freedom for error in the full model.

Calculation for F - test

Full Model:

$$C.O.T = \beta_0 + \beta_1 * age + \beta_2 * family\ income + \beta_3 * medical\ history(hypertension) + \beta_4 * medical\ history(glaucoma)$$

SSE = 2000, df(error) = 40

Reduced Model:

$$C.O.T = \beta_0$$

SSE = 7505

$$F = \frac{\frac{7505 - 2000}{(4 - 0)}}{\frac{2000}{40}} = 27.52 \sim F_{2,40} = 1.3 * 10^{-8}$$

Calculation for F - test

Since p-value is less that 0.05, reject the null hypothesis.

Implies the age, income and cost of treatment has a significant impact on the mean cost of treatment.

F-test and its relation to R-squared

While R-squared provides an estimate of the strength of the relationship between your model and the response variable, it does not provide a formal hypothesis test for this relationship.

The overall F-test determines whether this relationship is statistically significant. If the P value for the overall F-test is less than the significance level, you can conclude that the R-squared value is significantly different from zero.

F-test is significant but individual t – test is not!!!

The F-test adds up the explanatory power of each predictor and collectively the total explanatory power is statistically significant.

However, each individual predictor doesn't explain enough by itself to be statistically significant.

Is the Model Valid–ANNOVA and F Statistics

Source of Variability	DF	Sum of Squares	Mean Sum of Squares
Regression(Explained)	k	SSM	MSM=SSM/DFM
Error(Unexplained)	n-k-1	SSE	MSE=SSE/DFE
Total	n-1	SST=SSM+SSE	SST/DFT

F-test is always a single tailed test while testing the hypothesis that the coefficients are simultaneously equal to zero. F statistics is given by:

$$F = \frac{MSM}{MSE} = \frac{SSM/k}{SSE/n-k-1}$$

Computed F value is compared with the critical F value from the F table. Or obtain the p-value. P < .05, reject the null hypothesis that all the beta values are simultaneously equal to zero. Valid for regression models with more than one X variables.

Is the Model Valid–ANNOVA and F Statistics

Regression model for the Expense (Y) and Income (X) .

ε

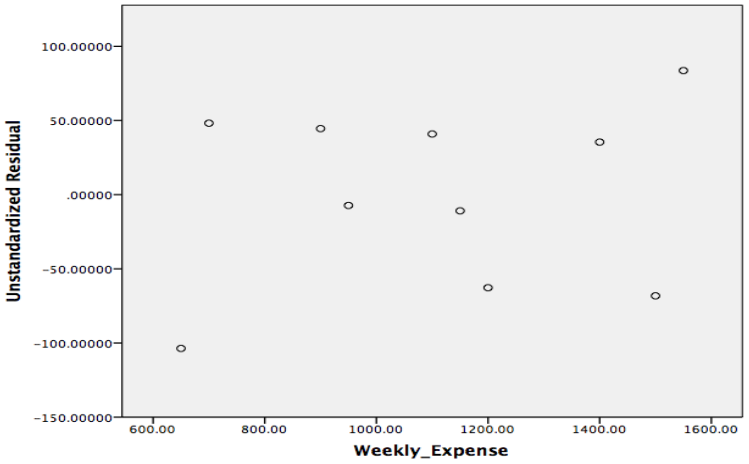
	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	855272.7273	855272.7273	202.8679245	5.75275E-07
Residual	8	33727.27273	4215.909091		
Total	9	889000			

Y (Exp)	X (Inc)
700	800
650	1000
900	1200
950	1400
1100	1600
1150	1800
1200	2000
1400	2200
1550	2400
1500	2600

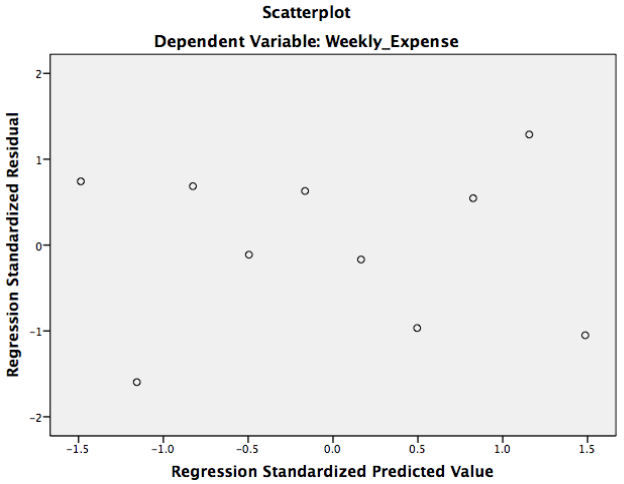
- Significance F (p-value) is less that 0.05 which signifies that the beta coefficients are not simultaneously equal to zero.

Is the Model Valid: Residual Analysis

No correlation between error term and X variable

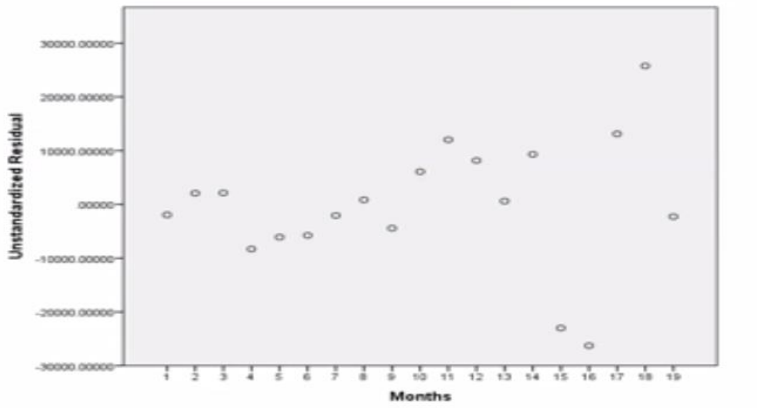


Heteroscedasticity problem (non constant variance for the error term)

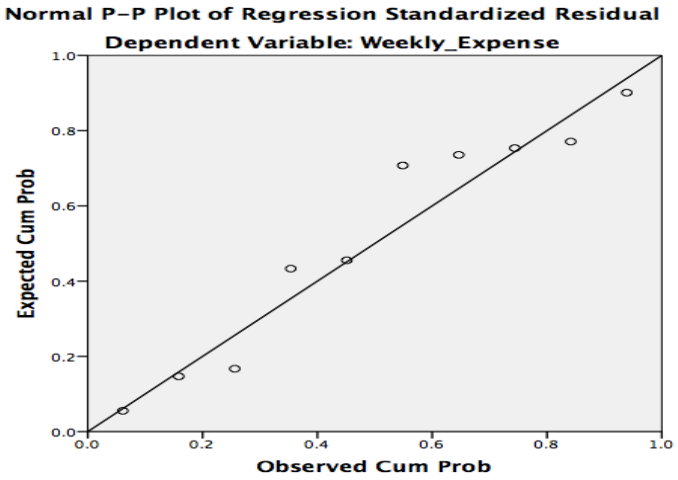


- White test
- Parks test

Autocorrelation in case of time series data (Plot of residual vs time)



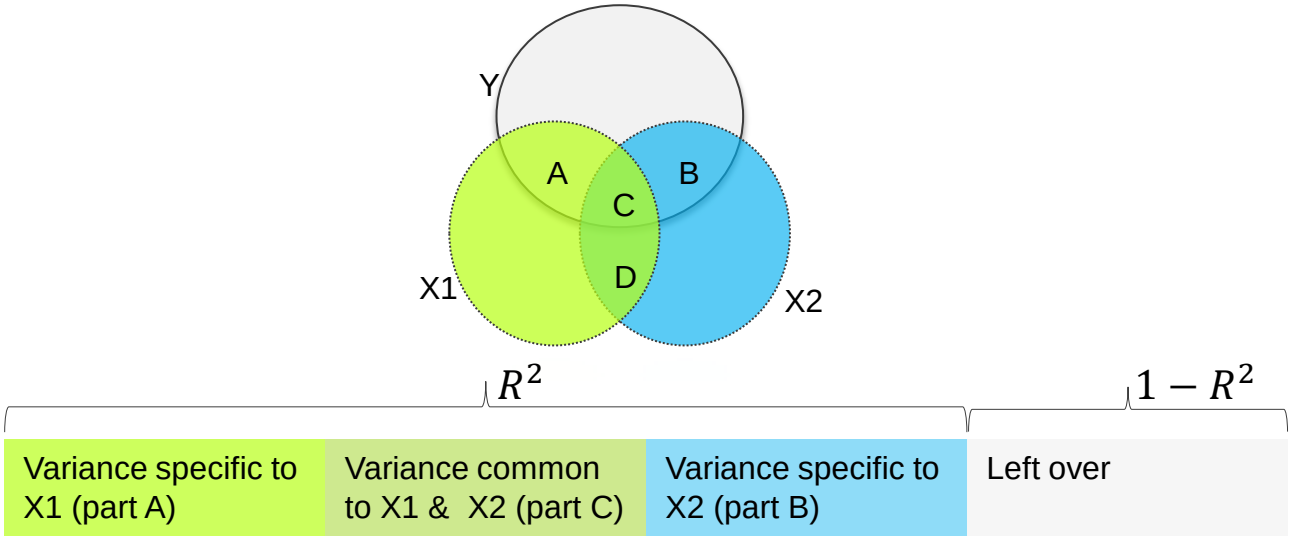
Normality of error term



- AD test

Is the Model Valid–Multicollinearity

The independent variables when correlated with each other leads to multicollinearity issue.



- Consequence of multicollinearity:
  - This correlation leads to larger variances and covariance's in the OLS estimators.
  - This can lead to wider CI of beta estimates and nullify the t-statistic for statistical significance.

Multicollinearity: If t-test concludes that the coefficients are not statistically different from zero but the F-test is significant and the coefficient of determination (R^2) is high. VIF (variance inflating factor) is a measure of MC. VIF>10 implies high degree of multicollinearity.

Is the model useful–Confidence and prediction interval

- Is the model depicting the general belief?
- Is the CI and PI encompassing the real world scenario?

Confidence Interval

- Confidence interval provides the interval estimate of the expected value of Y given X
- Used for interpolation of data within the range. $(1-\alpha)*100\%$ CI for $E(Y|X)$ is:

$$\hat{Y}_i \pm t_{\frac{\alpha}{2},n-2} * S_e * \sqrt{\frac{1}{n} + \frac{(X_i - \bar{X})^2}{SS_X}}$$

Where

$$S_e = standard\ error\ of\ regression$$

$$SS_X = \sum_{1}^{n} (X_i - \bar{X})^2$$

$$\hat{Y}_i = E(Y|X) = \widehat{\beta_1} + \widehat{\beta_2} * \hat{X}_i$$

Prediction Interval

- Prediction interval provides the interval estimate for Y given X
- Used for interpolation of data within the range. $(1-\alpha)*100\%$ PI for Y is:

$$\hat{Y}_i \pm t_{\frac{\alpha}{2},n-2} * S_e * \sqrt{1 + \frac{1}{n} + \frac{(X_i - \bar{X})^2}{SS_X}}$$

Where

$$S_e = standard\ error\ of\ regression$$

$$SS_X = \sum_{1}^{n} (X_i - \bar{X})^2$$

$$\hat{Y}_i = E(Y|X) = \widehat{\beta_1} + \widehat{\beta_2} * \hat{X}_i$$

Transformations on Data

Transformations on the dataset may be needed to use linear regression techniques more effectively and achieve:

- Normality in error term.
- Homoscedasticity of variance.
- Normality of regression equation.
- Better strength of relationship between response and explanatory variables.

Relationship between σ^2 and $E(Y)$	Transformation (Y')
σ^2 is constant	$Y' = Y$ (no transformation)
$\sigma^2 \alpha\, E(Y)$	$Y' = \sqrt{Y}$
$\sigma^2 \alpha\, E(Y)^2$	$Y' = \ln(Y)$
$\sigma^2 \alpha\, E(Y)^3$	$Y' = 1/\sqrt{Y}$
$\sigma^2 \alpha\, E(Y)^4$	$Y' = 1/Y$

Representing Qualitative Factors

Representing Qualitative factors in a regression equation:

- By using 'dummy variables‘. variables that take values of either 1 or 0, depending whether it is true or false.

Martial Status (MS)	MS_Married	MS_Single	MS_Divorced
Married	1	0	0
Single	0	1	0
Divorced	0	0	0

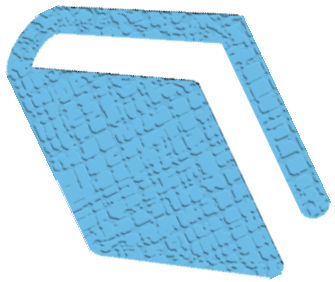
- If there are **n** factors, they can be represented by **n-1** dummy coded variables. This is derived from the concept of degrees of freedom.

More on: http://www.ats.ucla.edu/stat/mult_pkg/faq/general/dummy.htm. Package ‘dummy’ can be used in R for Dummy Variable Coding.

Regression in R Using an Example

Summary

Summary of the topics covered in this lesson:



- The intent of performing regression analysis is to predict the outcome of an event outside the sample dataset.
- Assumptions of linear regression needs to be satisfied to bring in better generalizability of the model.
- Usefulness of the model is understood by interpreting the signs of the coefficients and whether it is inline with the natural belief.

QUIZ TIME

Quiz Question 1

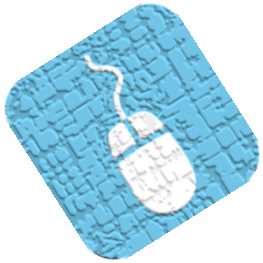
Quiz 1	Which library in R can be used for dummy variable coding? *Select all that apply.*

a. dummy

b. *dummies*

c. *dummys*

d. *dumb*



Quiz Question 1

Quiz 1 Which library in R can be used for dummy variable coding?
Select all that apply.

a. dummy

b. *dummies*

c. *dummys*

d. *dumb*

Correct answer is: ***a & b***

dummys and dumb are not defined packages in R.

End of Lesson02–Regression Concepts