Regression Analysis

Terms in this set (61)

The Population Regression Model	Y = B0 +B1X + E Y= dependent, X= independent, B0= y-int, B1= slope, E= error. B0 & B1 are unknown pop. parameters, we estimate these by taking a sample from the data.
The estimated sample regression line	AKA prediction equation. yhat = b0 + b1x yhat= sample regression line, x= independent variable, b0 (a)= y-int of sample line, b1 (b)= slope of sample line. b0 & b1 are estimates of pop. parameters which makes them sample statistics.
OLS Regression	AKA ordinary least squares regression. When we conduct regression analysis, we select the line that minimizes

Regression Analysis

To calculate the estimates of the coefficients that minimize the differences between the data points and the line:	b1 = cov(x,y)/sx^2 b0 = ybar - b1*xbar
The regression equation that estimates the equation of the first order linear model:	yhat = b0 +b1x
Coefficient of Determination (RSquared)	Standard error of the estimate is an important RELATED measure. Used when we want to measure the strength of the linear relationship. Measures the proportion of the variation in y that is explained by the variation in x. 1-(SSE/SST) or (SST-SSE)/SST or (SSR/SST). Absvalue(r) = SQRT(R^2) sign depends on slope coefficient for SIMPLE regression. Can be between 0-1. 1= perfect match between line and data points; 0= no linear relationship between x&y. The % of variation of y that is explained by the model. Adding independent variables increases RSquared.
F-Test for Overall Validity of the Model	Multiple regression version of this test asks: Is there atleast one independent variable linearly related to the dependent variable?. H0: B1=B2=...Bk=0. If atleast one B is not equal to 0 (H1), the model is valid. Simple regression F-test duplicates t-test for slope (F=t^2). Simple regression: H0: B1=0; H1: B1 =/0. If H0 is rejected for H1, the model is

Regression Analysis

T-test for Slope (b1)	When no linear relationship exists between two variables, regression line should be horizontal. Testing for non-zero slope. using b1 (estimate of the slope). H0: B0 = 0 (variable is not helpful). Test statistic: (b1-B1)/sb1 sb1 = se/SQRT((n-1)sx^2). CI: b1 +- t alpha/2*sb1. D.F. = n-2.
T-test for slope (r)	When no linear relationship exists between two variables, regression line should be horizontal. Testing for non-zero slope. using r (correlation coefficient: used to measure strength of association. goes from -1 to 1). H0: p = 0 (variable not helpful) Test statistic: r*SQRT((n-2)/(1-r^2) r (correlation coef.) = Cov/(Sx*Sy) d.f. = n-2
Model assessment:	RSquared, F-Test for overall validity, & model assessment.
Sum of squares total	SST = SSR + SSE. Total variation in dependent variable (y). Used to find sample variance of y. SST/(n-1) = sy^2
Sum of squares regression	SSR. The regression model. The variation explained by the regression line. =R^2*SST

Regression Analysis

Sum of squares for error	SSE. The error. The unexplained variation. =(n-k-1)*stderror^2
Standard Error of the Estimate	Measure related to how data points fit the line. The standard deviation of the data points around the regression line. The term we use for standard deviation of a sampling distribution. Used in conjunction with empirical rule. Located at top of data output. =SQRT(SSE/(n-2)) or =SQRT(MSE) for simple =SQRT(SSE/(n-k-1)) for multiple
What does the standard error of the estimate measure?	How spread out the data are around the regression line
F test statistic	MSR/MSE; MSR = SSR/k, MSE = SSE/(n-k-1)
Analysis of Variance (F-test)	F = MSR/MSE. MSR = SSR/k; k: # of independent variables. MSE = SSE/(n-k-1). Large F results from large SSR, meaning much of the variation in y is explained by regression model, H0 should be rejected, model is valid. Always upper tailed test

Regression Analysis

Regression Model Assumptions	3 requirements involving distribution of E which must be satisfied for model results to be valid: Probability distribution of E is normal with mean of 0 (normality), the standard deviation of E is a fixed amount for all values of x (homoskedaticity), the set of errors associated with different values of y are all independent (independence or non-A/C). Other assumptions that when violated can threaten usefulness: no unnecessary outliers, no serious multicollinerarity (independent variables are very closely related; only in multiple regression). Ideally: model has normally distributed y with mean = B0 + B1X1 and constant standard deviation
Residual Analysis	The i th residual is defined as ei = yi - y^i. Start by graphing residuals (create histogram of residuals [assumption 1], place residuals on vertical axis and graph against predicted values of sample y [assumption 2], and if a time series against time [assumption 3].
Normality of errors	Violation of normality can happen anywhere. Use excel to obtain standardized residual histogram. Examine histogram and look for bell shaped diagram with mean close to 0. To fix problem, transform y

Regression Analysis

Constant Variance of errors	AKA homoscedasticity. Violation can happen anytime. Use standard residual plot vs predicted y plot. The residuals, our estimates of the errors, seem to have approximately equal spread around the regression line. When requirement is violated we have heteroscedasticity, the spread of residuals varies at different points along regression line. To fix, transform y
Independence of errors	AKA non-autocorrelation. When the data is a time series, error terms are frequently related thereby violating independence. Use standard residual vs time plot to detect or DW test. When consecutive error terms measured across time are correlated, the errors are autocorrelated. Examining residuals over time, no pattern should be observed if errors are independent. +A/C means residuals are close, -A/C means residuals are far. only a problem in time series regression. to fix, add time variable
Time series	a series with data collected over time
No unnecessary outliers	When outlier is observed these possibilities need to be investigated: There was an error in recording the value, The point does not belong in the sample, or The observation is valid. Unnecessary outlier can occur anytime, found from XY scatter plot. It is customary to suspect an observation is an outlier if its Istandard residuall > 2. the cause determines the response. If the first or second situation, delete it, and re-estimate model. If it appears to be a valid observation, keep it in, but recognize the impact it is having. To fix, from outlier

Regression Analysis

Outlier	observation that is unusually small or large. Identified on scatter diagram
Multicollinearity	linear association between independent variables used in a regression. can be detected using correlation matrix.
Which of the following statements is a possible description of heteroscedasticity?	The absolute size of the residuals is related to the dependent variable. i.e. The spread of the data points gets wider as you move along the regression line.
Which regression problem would we typically be looking for by examining a histogram of the residuals?	Non-normality
Consider the following statements about outliers: 1. An outlier is an invalid observation and should be removed from the data set 2. An outlier may or may not be an invalid observation, and should be investigated 3. An outlier is likely to have a large effect on the regression coefficients Which of the statements is correct?	2 & 3

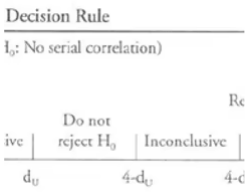
Regression Analysis

In simple linear regression, which of the following defines the residual associated with the first observation.	e1 = y1 - b0 - b1x
Applications of Regression Equation	Estimating the value of y, establish prediction and confidence intervals, and interpreting slope of the regression.
Prediction interval vs confidence interval	Prediction interval for a particular value of y. Confidence interval for the expected value of y. Prediction interval is wider because of added 1

Regression Analysis

Which of the following statements about simple linear regression is false? a. A t-statistic close to zero for the slope coefficient implies that the x-variable is not linearly related to the y-variable b. The value of the slope coefficient estimates how much the y-variable changes for a unit change in the x-variable c. The sum of squared errors is the sum of squared differences between the actual y-values and the predicted y- values d. The standard error of the estimate measures how spread out the data points are around the regression line e. The correlation coefficient between the two variables is equal to the square root of R2	e. The correlation coefficient between the two variables is equal to the square root of R2
Simple vs Multiple Regression	Simple: 1 variable, straight line Multiple: k variables, plain
Remedying non-normality and heteroscedasticity	Transformation of y variable. The transformations can improve the linear relationship between the dependent variable and the independent

Regression Analysis

List of transformations	y' = ln y (for y > 0) Use when the se increases with y, or Use when the error distribution is positively skewed y' = y^2 Use when the s2ε is proportional to E(y), or Use when the error distribution is negatively skewed y' = y^1/2 (for y > 0) Use when the s2ε is proportional to E(y) y' = 1/y Use when s2ε increases significantly when y increases beyond some value.
Durbin-Watson Test	This test detects first order auto-correlation between consecutive residuals in a time series. If autocorrelation exists the error terms are not independent. d can be approximated using: 2(1-r). range of d is (0,4). H0: No A/C H1: A/C Two tail test. Autocorrelation occurs over time, so a time dependent variable can correct the problem. 

Regression Analysis

Positive first order autocorrelation	Occurs when consecutive residuals tend to be similar. Then, the value of d is small (less than 2)
Negative first order autocorrelation	Occurs when consecutive residuals tend to markedly differ. Then, the value of d is large (greater than 2).
What would be the most appropriate solution to the problem of autocorrelated error terms?	Add a time related variable to the model
Durbin-Watson formula	sigma(ei-e-1)^2/sigma(ei)^2
Serious multicollinearity	Only found in multiple regression, detected using correlation matrix. Exists when independent variables included in the same regression, are linearly related to one another. Multicollinearity nearly always exists, We consider it serious if the absolute value of the correlation coefficient between any pair of independent variables exceeds .80. Multicolinearity inflates the standard errors of the slope estimates, this brings t-stats close to 0 and insignificant, making the coefficients unable to be interpreted as slopes.
Correcting for Multicollinearity	Get rid of one of the variables (the one with highest absolute value in correlation matrix, then remove based on p-value for every subsequent removal) that is a duplicate, and re-estimate.

Regression Analysis

Consider the following statements about the error term in linear regression. 1. The standard deviation of ε is one 2. The standard deviation of ε is constant for all values of the independent variables 3. The mean of ε is zero 4. The errors are independent as measured across time Which of the above statements need to be true for the ordinary least squares regression procedure to be valid?	2, 3, 4

Regression Analysis

The Modeling Process	Develop a model that has a sound basis (assign a priori expectations for signs). Gather data for the variables in the model. Draw the scatter diagram to determine whether a linear model (or other forms) appears to be appropriate. Estimate the model coefficients and statistics using statistical computer software. (combine below) Assess the model fit and usefulness using the model statistics (3 step process in simple). combine^Diagnose violations of required conditions. Try to remedy problems when identified Assess the model fit and usefulness using the model statistics. Use finished model to: predict values of dependent variable, provide interval estimates, provide insight to impact of each independent variable
Model assessment for multiple regression	1. R2 (Coefficient of Determination) - Adjusted R2 - Standard error of the estimate 2. F-Test for overall validity of the model 3. T-test for slope - using b (estimate of the slope) - Partial F-test to verify elimination of some independent variables

Regression Analysis

"Adjusted" Coefficient of Determination	1-[SSE/(n-k-1)]/[SST/(n-1)]. Penalizes you a small amount for each additional independent variable you add, new variable must contribute significantly to explaining SST b4 AdjR2 goes up. The "adjustment" is adjusting after degrees of freedom or number of independent variables. Uncertain which way AdjR^2 will go when adding variables; if new variable has high correlation it will increase and vise versa. Can be negative; will never be greater than R^2
The following is a list of some of the steps in the modelling process. Please put them in order: Assess the model fit Develop a model with a sound basis Estimate the model coefficients	2, 3, 1
Sample variance	SST/(n-1)

Regression Analysis

You have run a model with 9 independent variables and then through the modelling process decide to exclude 4 of the original independent variables. What can you say will happen to the final values of R2 and adjusted R2 compared to their initial values?

a. R2 and adjusted R2 will not decrease
b. R2 and adjusted R2 will not increase
c. R2 will not increase, however you are unsure what will to happen to adjusted R2
d. R2 will not decrease, however you are unsure what will to happen to adjusted R2
e. Adjusted R2 will not decrease, however you are unsure what will to happen to R2

R2 will not increase, however you are unsure what will to happen to adjusted R2

Regression Analysis

The Partial F-test.	First, consider your individual t-test results (which variables should you keep? are there variables that officials should be eliminated but have a close enough p-value? are any variables that are believed strongly to be in the model despite t-tests? H0: B1 = B2 = ... = Bi = 0 (full model) H1: At least one Bi is not = 0 (reduced) Reject H0 if ((SSRf - SSRr)/kd)/MSEf > F alpha,kd,n-k-1f Bi refer to the variables which were eliminated; SSRf is full equation, SSRr is reduced equation; MSEf is full equation, Kd is # variables deleted If there is a large difference in numerator, some of the variables eliminated have significant power in which case you use reduced model. Always one side upper tail test. If only one variable is dropped, the slope t-test statistic = partial F statistic same for p-value.
In multiple regression, if we reject the F-test for the overall validity of the model, we can conclude: a. All of the independent variables' slope coefficients are different than zero b. None of the independent variables' slope coefficients are different than zero c. At least one independent variable's slope coefficient is different than zero	c

Regression Analysis

If our F statistic for the overall validity of the model is a very large number but our model contradicts our a priori expectations:

i. Eliminating variables and conducting a partial F-test may help confirm our expectations
ii. Our model is not valid
iii. Our a priori expectations may have been incorrect
iv. Since the F statistic is significant, we would never need to adjust our model

iii

Regression Analysis

When conducting a partial F-test, if the SSR of the reduced model is significantly lower than the SSR of the full model: a. The eliminated variables had a significant amount of explanatory power in the overall model b. The eliminated variables had very little explanatory power in the overall model c. We would reject the null hypothesis of the partial F-test and accept the reduced model d. We do not have sufficient information	a

Regression Analysis

Curvilinear Relationships	In regression, it is assumed that the relationship between the dependent and each independent variable is linear. You have to specifically introduce the possibility of a curvilinear relationship for it to be considered. A commonly used functional form used to create curvilinear relationships in multiple regression analysis is the polynomial model. A polynomial model of order p takes the form: $y = \beta 0 + \beta 1x + \beta 2x2 + ...+\beta pxp + \varepsilon$ (most common are quadratic and cubic) When introducing a polynomial model, you will need to decide what order of model is appropriate. Do this by adding additional independent variables each of which is a higher order versions of the original independent variable (x1,x12,x13,x14). Estimate the model by including the lowest order term first, and working your way higher. The final polynomial model will be the one which has the highest order polynomial term still showing a significant t-test, while the next higher polynomial term does not. This requires testing of multiple models of different powers before selecting the most appropriate model. In multiple regression it is common to have independent variables that have a curvilinear relationship with the dependent variable while also having other independent variables that have a linear relationship with the dependent variable in the same model. We want to distinguish between the number of independent variables and the number of predictor variables.

Regression Analysis

Interaction term	y = b0 + b1x1 + b2x2 + b3x22 + b4x1x2 + e The x1x2 is the interaction term. When interaction terms are present, neither the slope nor the intercept between any of the x variables included in the interaction term and the y variable are constant.
Qualitative Independent Variables	Including qualitative variables in a regression analysis model is done via indicator variables. An indicator variable (I) can assume one out of two values, "zero" or "one" I = 1 means yes/is I = 0 means no/is not Create m-1 (m=levels) indicator variables. We create one less dummy variable than the categories we have. Coefficients act as comparison to left out dummy variable
Suppose you have a database on profitability of intercity bus routes. Which of the following would be represented with a dummy variable? The percentage of women travelling on a particular route The average income of the cities Whether or not the average income of the cities is above $30,000	3

Regression Analysis

You want to create a multiple regression for factors that influence GDP. You decide that geographical regions may influence a country's GDP (for reasons such as trade agreements, natural resources available, etc.). You would like to compare each of the continents GDP, excluding Antarctica, to Northern America's GDP. How many dummy variables would you need to do this? a. 5, not including Northern America b. 5, not including Antarctica c. 6, not including Northern America d. 6, not including Antarctica	a
What is true about the values of R^2 and AdjR^2	R^2 can take values [0,1], Adj can take values less than 0
Multiple R	(in simple regression) absvalue of r = absvalue of SQRT(R^2).