Regression - STAT110 Otago

Students also viewed

TAT110 STUDY TYPES

2 terms

Laura_Siegert57

Preview

R CODES STAT110

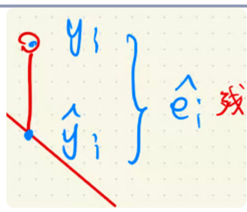
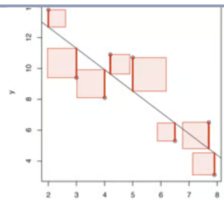
18 terms

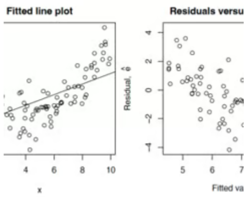
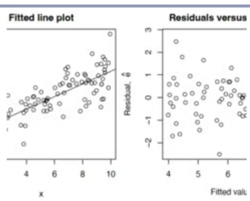
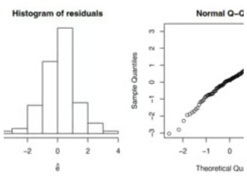
Laura_Siegert57

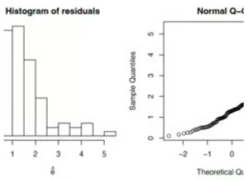
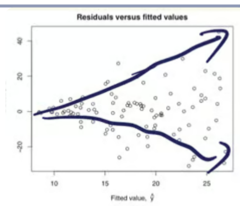
Preview

Terms in this set (60)

types of regression	**linear (continuous data), logistic (categorical data),** cox (categorical data in a survival analysis)
X	*Explanatory variable (X), also known as covariate, predictor, or independent variable.*
Y	*Outcome variable (Y), also known as response or dependent variable*
Simple linear regression (SLR)	*looks at a relationship between two continuous variables where the relationship between the two variables is* ***approximately a straight line***

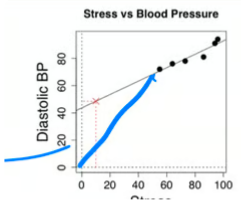
SLR equation	***Y = β0 + β1x + e*** *implies that the mean response is related to x by* ***μY = β0 + β1x.*** ***Y*** *is the numerical outcome variable (continuous or approximately so)* ***x*** *is the explanatory variable* ***β0*** *is the intercept or constant (where the line crosses the y axis)* ***β1*** *is the slope of the line* ***e*** *(often denoted* ***ϵ****) is the random error or residual term*
SLR equation for estimating	$= \hat{\beta}_0 +$
residual ('estimated error') term	*ˆei = yi - ˆyi* 
How to find regression line	*The line of best fit minimises the sum of the squares of the residuals.* 
equation for how to find regression line	$= \sum_{i=1}^{n} (y_i$
How to calculate beta1 and beta0	$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$ $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$

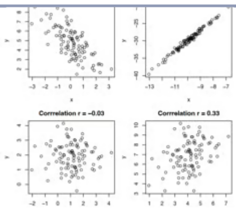
example for how to calculate regression (*Stress and Blood Pressure*)	1. get n, n = 6 2. find the *explanatory and outcome* 3. calculate beta 1 and beta 0 4. get the regression equation 5. using R for SLR
Assumptions for Simple Linear Regression(LINE)	***Linearity:*** *The relationship between the mean response µY and x is described by a straight line.* ***Independence*** *The responses Y1,Y2, . . . ,Yn are statistically independent.* ***Normality*** *The error terms e1, e2, . . . , en come from a normal distribution.* ***Equal variance*** *The errors terms all have the same variance, σ2 ('homoscedastic').*
What diagram is used for checking linearity	***residual*** $= y_i - \hat{y}_i$ $= y_i - (\hat{\beta}_0 + \hat{\beta}_1$ $= y_i - \hat{\beta}_0 - \hat{\beta}_1 x$
failure of linearity assumption	
linearity assumption holds	
Checking independence assumption	*May get insight by thinking about the study design.* (Ask yourself questions)
plot for *Checking the normality assumption*	Q - Q plot
pass of *normality assumption*	

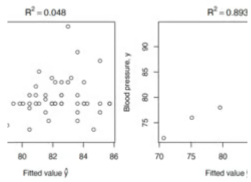
fail of *normality assumption*	
Checking equal variance assumption (homoscedasticity)	pass if the residual plot is not like this 
what is the impact if *Fail of the linearity assumption*	***critical*** *If that assumption fails,* ***all conclusions drawn from the model will be invalid.***
what is the impact if *Fail of independence or equal variance assumptions*	***remain valid*** *However, estimates can be inefficient* *Follows that* ***fitted regression line is useable.*** *Any* ***test results or confidence intervals based on the regression model will be invalid.***
what is the impact if *Fail of inormality assumption*	*typically* ***least important.*** *Effects validity of confidence intervals and test results when the sample size* ***n is small.***
what to do with outliers	*the first thing to do is check that the data are correctly recorded* *If data cannot be corrected, try refitting regression with outliers removed, but still investigate cause of outliers - may be very important.*
estimate of error variance	$s_e^2 = \frac{1}{n-2}\sum_{i=1}^{n}$ $_{=1}\ \hat{e}_i^2$ is the *residu*
(estimated) standard error	$\frac{s_e}{\sqrt{\sum_{i=1}^{n}(x_i}}$
degree of freedom for SLR's CI	v = n - 2 because there're two parameters

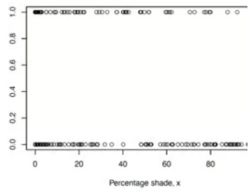
what is the multiplier for SLR's CI	$\frac{\text{stimate} -}{\text{std. err}}$
what is the SE for SLR's CI	$\sqrt{\sum_{i=1}^{n}}$

Using R to find SLR's CI	`(model1)`
	`2.5 %`
	`:) 24.8300345 6`
	0.2557407

β1 = 0 indicates what	*that the response is not (linearly) related to the predictor.* so *the estimated slope will (almost) always be non-zero: ˆ β1≠ 0.*
Steps to *test to assess strength of evidence in the data for β1 != 0*	1. *Setting up the hypotheses* *H0 : β1 = 0, HA : β1≠ 0.* 2. Calculating *The test statistic (picture)* 3. *Computing the p-value* 4. draw conclusion with rejecting or not H0 $t = \frac{\hat{\beta}_1}{s_{\hat{\beta}_1}}.$
when predicting the data	ignore e_0
why not recommend to extrapolate when predicting data	plot may not be linear 
prediction error	*The prediction error is* ***analogous to a standard error, but takes account of both sources of uncertainty.*** *For prediction at x0, the prediction error is:* $\sqrt{1 + \frac{1}{n} + \frac{}{\sum}}$

prediction interval formula	$-\frac{\alpha}{2}, n-2)$
correlation coefficient (r)	*summarises the strength of a linear relationship between variables.* *It is a measure of linear association between variables* *It describes both the strength and direction of the relationship.* ***r*** ∈ **[-1, 1]** *A **positive value** of r means that **Y and X increase together.*** *A **negative value** of r means that as **X increases, Y decreases** (and vice-versa).* *The **strength of the linear relationship increases** as r tends **towards 1 or -1.*** ***r = 0 corresponds to no linear relationship** between the variables.* $\frac{{}_{=1}(x_i - \bar{x})}{x_i - \bar{x})^2 \sum}$
scatterplots for r	
re-write for r	***s_x and s_y** are **sample standard deviations** for x and y variables.* ***s_xy** is the **sample covariance** between x and y.* $= \frac{s_{xy}}{s_x s_y}$

Correlation coefficient *versus* ***regression models***	*The* ***correlation coefficient is a summary of the data.*** *Unlike linear regression, the* ***correlation coefficient does not specify a model for the data****, and cannot (for example) be used for prediction.* *The* ***correlation coefficient is symmetric*** *in the variables. That is, correlation between x and y is the same as correlation between y and x.* *In* ***regression****, the variables are* ***not handled symmetrically****.* ***Regression*** *models look at* ***variation in Y for fixed values of x****.*
coefficient of determination (R^2)	*R^2, is a measure of how well a regression model describes the data.* *R^2 is the* ***squared correlation between the observed and predicted responses*** ***R^2*** ∈ **[0, 1]**
meaning for the value of R^2	*A high value of R2 (close to 1) indicates a regression model that describes the data very well.* *Conversely, a low value of R2 (close to 0) indicates a regression that describes the data poorly.* 
what *describes the* ***overall variation in the response variable****?*	*total sum of squares* $= \sum_{i=1}^{n} (y_i -$
what *describes the* ***total variation of the data points*** *about the regression line?*	residual sum of squares *RSS can be thought of as variation not explained by the regression model* $= \sum_{i=1}^{n} (y_i -$

what describes *as the* ***amount of variation in the response*** *that is explained by the regression model?*	*explained sum of squares* ESS = TSS - RSS
Equation of R^2	$\frac{ESS}{TSS} = 1 -$
Correlation does not equal causation	e.g., just because there's more ice cream in the summer and more drowning in the summer doesn't mean there's a link between ice cream and drowning.
logistic regression	*outcome variable is binary* 
equation for logistic regression	***Y is the binary outcome*** *variable, Y = 1 or Y = 0 for each observation.* ***p is the probability*** *that specified category will occur; i.e. p = Pr(Y = 1).* *x is the explanatory variable.* *Parameters β0, β1 are the regression coefficients.* *β0 is intercept and β1 slope 'on the logit scale'* *In formula,* ***log is the natural logarithm (log to base e).*** $g\left(\frac{p}{1-p}\right)$
	$\frac{p}{1-p})$
Which technique we use when we estimating the regression coefficients?	*maximum likelihood estimation*

what will *increasing x by one unit results in?*	*a multiplicative change of e^β1 to the odds*
formula for *logistic curve for the probability p*	$\frac{\exp(\beta_0 +}{+ \exp(\beta_0 -}$
Testing in logistic regression	*1. Define the hypotheses: H0 : β1 = 0 and HA : β1 != 0.* *2. The test statistic is:* $z = \frac{\hat{\beta}_1}{s_{\hat{\beta}_1}}$ *where s_ ˆβ1 is the standard error of ˆ β1.* *The further away our test statistic, z, is from zero, the greater the evidence against H0.* 3. get the *corresponding p-value* 4. reject/not reject H_0 5. conclusion
Multiple regression model	$x_1 + \cdots$
mean value de la Multiple regression model	$\beta_1 x_1 +$
Applications of multiple regression	*1 Adjusting for the effect of confounding variables.* *2 Establishing which variables are important in explaining the values of* *the response variable.* *3 Predicting values of the response variable.* *4 Describing the strength of the association between the response variable and the explanatory variables.*

least squares estimates	$_{=1}(y_i\ -$
RSS for ^e_i	*to estimate the error variance σ^2_e* $= \sum_i^r$
usual estimate	$= \dfrac{RSS}{n - k -}$