roduction To Regression

follows ↓ mean ↓ ↙ st. deviation

$$Z \sim N(0, 1)$$

value ↓ mean

2 score $\quad \frac{x - \mu}{\sigma}$ ← st. deviation

26/4/13

$$\sum_{i=1}^{n} (x_i - \bar{x})^2$$

$$\Rightarrow \sum_{i=1}^{n} (x_n^2 + \bar{x}^2 - 2x_i\bar{x})$$

$$\Rightarrow \sum_{i=1}^{n} \left(x_n^2 + n\bar{x}^2 - 2\sum_{i=1}^{n} x_i\bar{x}\right)$$

$$\Rightarrow \sum_{i=1}^{n} x_i^2 + n\bar{x}^2 - 2\bar{x}\sum_{i=1}^{n} x_i$$

$$\Rightarrow \sum_{i=1}^{n} x_i^2 + n\bar{x}^2 - 2\bar{x}(n\bar{x}) \quad \leftarrow \quad \bar{x} = \frac{\sum_{i=1}^{n} x_i}{2}$$

Variance always ADDED

IF $x \sim N(0, 2^2)$ and $y \sim N(3, 1^2)$

1. $\frac{1}{4}\text{var}(x+y) = \frac{1}{4}(\text{var}(x) + \text{var}(y))$

→ (constant get Squared

2. $\text{Var}(4x+3y) = 16\,\text{var}(x) + 9\,\text{var}(y)$ → Variance always ADDED

3. $\text{Var}(x-y) = \text{Var}(x) + \text{Var}(y)$

$$\mathrm{Var}(\overline{X}) = \frac{\sigma^2}{n}$$

$$E\left(\sum_{i=1}^{n} \frac{1}{n}\mu\right)$$

$$\frac{\mu}{n}$$

$$E(\bar{x}) = E\left(\frac{1}{n}\sum_{i=1}^{N} x\right)$$

$$= \frac{1}{n} E\left(\sum_{i=1}^{N} x_i\right)$$

$$= \frac{1}{n} \sum_{i=1}^{N} E(x)$$

$$= \frac{1}{n} \sum_{i=1}^{n} \mu$$

$$= \frac{1}{\mu} (n\mu)$$

$$= \mu$$

$$\mathrm{Var}(\bar{x}) = \mathrm{Var}(\bar{x})$$

$$\mathrm{Var}(\bar{x}) = \mathrm{Var}\left(\frac{1}{n}\sum_{i=1}^{n} x_i\right)$$

$$= \left(\frac{1}{n}\right)^2 \mathrm{Var}\left(\sum_{i=1}^{n} x_i\right)$$

$$\frac{1}{n^2} \sum_{i=1}^{n} \mathrm{Var} x_i$$

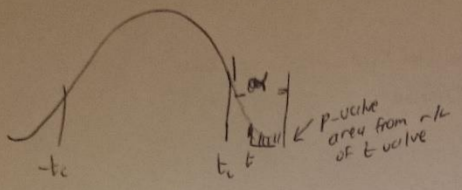
$$\frac{\sum_{i=1}^{N} \sigma^2}{n^2}$$

$$= \frac{1}{n^2} (n\sigma^2)$$

$$= \frac{\sigma^2}{n}$$

t-test.

$\alpha = 0.05$



$t = t(n-1)$

t_c t-critical

$$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} > t_c$$

(data, data, in t-table)

reject.

p value < 0.05
$\Rightarrow$ reject.

$< -t_c$

Accept

$$-t_c < t < t_c$$

$$\Rightarrow -t_c < \frac{\bar{x} - \mu_0}{s/\sqrt{n}} < t_c$$

$$\Rightarrow -t_c \frac{s}{\sqrt{n}} < \bar{x} - \mu_0 < t_c \frac{s}{\sqrt{n}}$$

$$\Rightarrow \bar{x} - t_c \frac{s}{\sqrt{n}} < \mu_0 < \bar{x} + t_c \frac{s}{\sqrt{n}}$$

CI =

$$\left(\bar{x} - t_c \frac{s}{\sqrt{n}},\ \bar{x} + t_c \frac{s}{\sqrt{n}}\right)$$

Regression

$$R = \frac{Cov(X,Y)}{\sqrt{var(x)}\sqrt{var(y)}}$$

$$= \frac{Cov(X,Y)}{Sd(x)\ Sd(y)}$$

$$= \frac{\frac{1}{n-1}\sum (x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\frac{1}{n-1}\sum (x_i-\bar{x})^2}\sqrt{\frac{1}{n-1}\sum (y_i-\bar{y})^2}}$$

$$= \frac{\sum (x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum (x_i-\bar{x})^2 \sum (y_i-\bar{y})^2}}$$

$$-1 \leq r \leq 1$$

$r > 0$: positive correlation
$r < 0$: negative
$r = 0$: no correlation

No linear Relationship

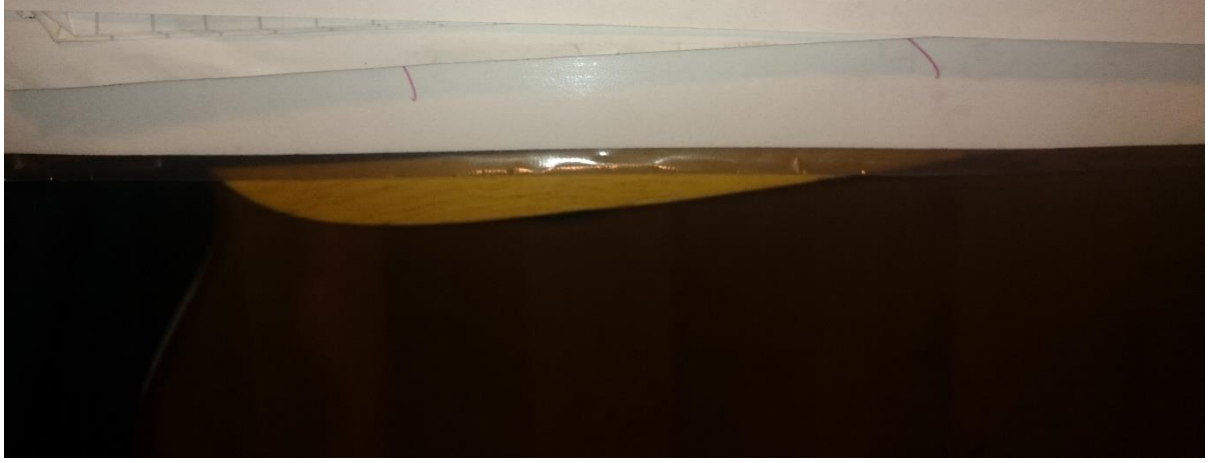
3

Covariance: $Cov(X,Y) = E[(X - E(x))(Y - E(y))]$

Sample: $Cov(x,y) = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})$

If values of x lie above $\bar{x}$ indicates values of y lie above $\bar{y}$"

then we have positive products. Similar for negative ones



9/10/13 Regression

Correlation

$$R = \frac{Cov(x,y)}{Sd(x)\,Sd(y)}$$

1. Change in X cause change in Y
2. " Y " X
3. Change in some 3rd variable Z, could change in X and Y
4. The relationship is a coincidence

Correlation does not imply causality

Simple Linear Regression

Regression is a statistical tool that utilises the relationship between two or more variables

Relationship

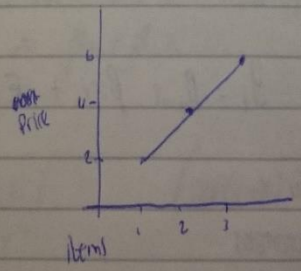
1. Functional
2. Statistical

Functional relationship is deterministic and can be exactly expressed by a formula

$$Y = f(x)$$

eg. $Y = 2x$

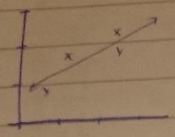
Price of 1 item = €2.



2

9/10/12

Statistical relationship is not perfect like a functional one



FR
$Y = F(x)$

SR
$E(Y) = f(x)$

Regression model is formal means of expressing a statistical relationship

A statistical relation can be summed up as:

1. The dependant variable Y varies with respect to independent variables in some systematic way.

2. Observations (Y_i) are scattered around the curve of relationship

These 2 are represented in a regression model:

1. In the population of observations associated with the sampled proces, there is a probability distribution of Y for every value of x

2. The means of Y varies in some systematic (a line) relationship with respect to x.

$(x_1, y_1) \dots (x_n, y_n)$

Simple linear equation. $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$

$y_i \rightarrow i^{th}$ data pair of Y

$x_i \rightarrow$ " fixed value for X

$\beta_0, \beta_1 \rightarrow$ unknown parameters

$\varepsilon_i \rightarrow$ random error term

10/13 Regression

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

i. Y_i → i^{th} datapoint
ii. β_0, β_1 → unknown parameter (need to be estimated)
iii. X_i are constants
iv. ε_i is a random error term $E(\varepsilon_i) = 0$
$Var(\varepsilon_i) = \sigma^2$ for all i
$Cov(\varepsilon_i, \varepsilon_j) = 0$ for all $i \neq j$.

Features of the model

1. The observed value of Y of the i^{th} datapoint is the sum of 2 components

$Y_i = (\beta_0 + \beta_1 X_i) + (\varepsilon_i)$
Systematic Chance (random term)

Linear in parameters $Y_i = \beta_0^2 + \underbrace{e^{\beta_1}}_{\log \beta_2} + \varepsilon_i$

2. Since $E(\varepsilon_i) = 0$ ($E(\varepsilon_i) = 0$)
$E(Y_i) = E(\beta_0 + \beta_1 X_i + \varepsilon_i) = E(\beta_0 + \beta_1 X_i) + E(\varepsilon_i)$
$= \beta_0 + \beta_1 X_i$

3. $Var(\varepsilon_i) = \sigma^2$
$Var(Y_i) = Var(\beta_0 + \beta_1 X_i)$ $(Var(c+x) = Var(x)$ case constant$)$
$= Var(\varepsilon_i) = \sigma^2$

4. Important part for regression is $E(Y) = \beta_0 + \beta_1 X$

5. Y_i and Y_j are uncorrelated for $i \neq j$. because $Cov(\varepsilon_i, \varepsilon_j) = 0$
$\Rightarrow Cov(Y_i, Y_j) = 0$

$Cov(x,y)=0 \not\Rightarrow$ x and Y are independent
Y_i and Y_j are assumed to be independent

One more assumption which is not necessary $\Rightarrow Y \sim N(0,\sigma^2)$

Estimate β_0 and β_1 only required.

i. $E(Y_i)=0$ $\not\Rightarrow Y_i \sim N(0,\sigma^2)$
ii. $Var(x_i)=\sigma^2$
iii. $Cov(Y_i, Y_j)=0$ $\not\Rightarrow Y_i$ and Y_j are independent

Exam

* ASSUMPTIONS LEADING TO LINEAR REGRESSION

1. Y's are assumed to be independent of each other.
2. Y's are normally distributed with mean 0.
3. All Y's have the same variance
4. It is assumed that $E(Y_i)$'s can be joined by a straight line. $E(Y_i)=\beta_0+\beta_1 x_i$

Example: Price Of Diamond Dataset

$Y_i^{Price} = -259.63 + 3721.02 x_i^{Carat\ value} + \varepsilon_i$

Price = β_0 β_1 carat

x_i = 0.23 carats. actual value $$591

$E(Y_i) = -259.63 + 3721.02(0.23)$
$= 596.20$
$\varepsilon_i = -1.2$ (difference - error - residual)

Meaning of β_0 and β_1

β_0 = y intercept
β_1 = slope of line - change in y with respect to one unit change in x.

2/10/13 Introduction To Regression

Regression Model

$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i, \quad i = 1, \dots n$

Assumptions

1. X_i is the i^{th} value of the predictor variable and is constant
2. ε_i is the error value and is a random variable
   - (i) $E(\varepsilon_i) = 0$
   - (ii) $Var(\varepsilon_i) = \sigma^2$ } for all i
   - (iii) $Cov(\varepsilon_i, \varepsilon_j) = 0$ for $i \neq j$
3. The mean of Y_i can be joined by a straight line

   $E(Y_i) = \beta_0 + \beta_1 X_i \rightarrow \beta_0, \beta_1$ are parameter

   (i) β_1 is the slope of the regression line. It indicates the change in mean of Y for unit change in X

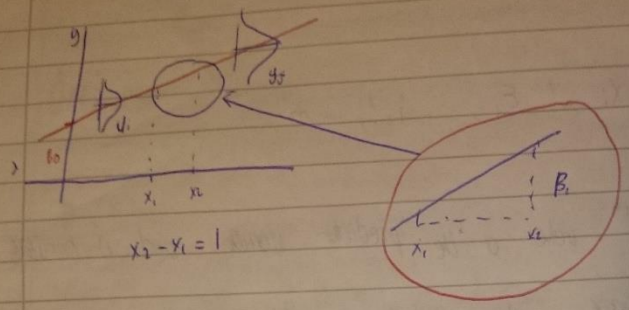
   (ii) β_0 is the slope
   Physically Speaking, β_0 is the mean value of Y when $X = 0$ and may not always have any meaning

Simplified Assumptions

1. X_i is the ...
2. The observation Y_i (or equivalently ε_i) are independent of each other.
3. For any i, Y_i is normally distributed
4. For any i, Y_i's have same standard deviation of
5. The mean of Y_i can be.

2

Regression



$x_2 - x_1 = 1$

Diamonds Dataset

$Y_i = -259.63 + 3721.02X_i + \varepsilon_i$

Price: $Y_i = 595$ carat: $X_i = 0.23$

$595 = -259.63 + 3721.02(0.23) + \varepsilon_i$ ← error

$\varepsilon_i = -1.20$

$E(\varepsilon_i) = 0$

Estimation of parameter β_0 and β_1

For the price of diamonds example, it was provided. How did we get:

$\beta_0 = -259.63$, $\beta_1 = 3721.01$

If X followed a Normal distribution with unknown mean μ and σ^2, μ was estimated by $\bar{x}$.

$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$

Procedure

One of the outcomes of estimating β_0 and β_1 is that they should result in a "line of best fit".

⇒ one which reflects the data well

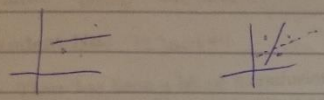
Best in what sense?

16/10/13 3. Regression

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

$\Rightarrow Y_i - (\beta_0 + \beta_1 X_i)$ for all i

We consider that line to be best for which the sum of squared error is minimised ($\sum_{i=1}^{n} \varepsilon_i^2$)



Method of least Squares

The method of LS gives estimates (estimate) b_0 ($\hat{\beta}_0$ (β0 hat)) for β_0 (parameter)
b_1 (β_1) for β_1

In principle $Q(\beta_0, \beta_1) = \sum_{i=1}^{n} \varepsilon_i^2$

$= \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 X_i)^2$

Taking partial derivatives

$$\frac{\partial Q}{\partial \beta_0} = -2 \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 X_i)$$

$$\frac{\partial Q}{\partial \beta_1} = -2 \sum_{i=1}^{n} X_i (Y_i - \beta_0 - \beta_1 X_i)$$

Then solve the Normal Equations (equate partial derivate to 0)
Solve for β_0, β_1

We use b_0 and b_1 to denote the sample

$$b_1 = \frac{\sum (x_i y_i) - \frac{\sum x_i \sum y_i}{n}}{\sum x_i^2 - \frac{(\sum x_i)^2}{n}}$$

23/10/13 Regression

Estimation of β_0 and β_1

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

All the y_i's and ε_i's have the same variance σ^2

What if σ^2 is unknown?

Estimate σ^2

When we had a single population

$X_i \sim N(\mu, \sigma^2)$

we estimate μ by $\bar{x}$

Estimate σ^2 by $\hat{\sigma}^2 = S^2 = \frac{\sum (x_i - \bar{x})^2}{n-1}$

$(n-1)$ is known as degree of freedom

Estimating σ^2 in SLR

$Y \sim N(\ ,\ )$

$E[Y_i] = \beta_0 + \beta_1 X_i$

$Var[Y_i] = \sigma^2$

$\Rightarrow Y_i \sim N((\beta_0 + \beta_1 X_i, \sigma^2))$

Firstly estimate β_0 and β_1 we call them b_0 and b_1

We also know that a fitted $\hat{y}_i$ is defined as

$\hat{y}_i = b_0 + b_1 X_i$

Define $SSE = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$ SSE = error sum of squared (residual sum of squares)

$= \sum_{i=1}^{n} (y_i - (b_0 + b_1 X_i))^2$

$= \sum_{i=1}^{n} e_i^2$ where e_i is residual $(y_i - \hat{y}_i)$

4.

$$b_1 = \frac{\Sigma (x_i - \bar{x})(y_i - \bar{y})}{\Sigma (x_i - \bar{x})^2}$$

$$b_0 = \frac{1}{n} (\Sigma y_i - b_1 \Sigma x_i)$$

$$= \bar{y} - b_1 \bar{x}$$

Some notation

$$S_{xx} = \Sigma (x_i - \bar{x})^2$$

$$S_{xy} = \Sigma (x_i - \bar{x})(y_i - \bar{y})$$

$\hat{y}_i$ → fitted value for the i^{th} observation x_i
y_i → observed value

$x_i = 0.17$ $\quad \hat{y}_i = -259.63 + 3721.02 * 0.17$
$= 372.95$

Interpetation
The mean if many diamond of carats $x_i = 0.17$ were sold under similar conditions as this data set, the average price of the would be 372.95 $

Of ccurse the price of a single diamond with carat = 0.17 would not be 372.95 $.

This is obvious, since there is chane variation in the pricing structure

Residuals
The i^{th} residual is the difference between the fitted value and the observed value of y.
$e_i = y_i - \hat{y}_i$

The residuall are the vertical deviations of y_i from the fitted regression line

Differnce between e_i and ε_i
$e_i = y_i - \hat{y}_i$ $\quad \varepsilon_i = y_i - (\beta_0 + \beta_1 x_i)$
↓ ↓
① Known ① Unknown quantity (theory → hypothetical.
② Distance between y_i and $\hat{y}_i$ ② Vertical deviation between y_i and "unknown" regression

17/10/13 Regression

Estimating σ^2 in SLR

Var $(\varepsilon_i) = \sigma^2$ } unknown parameter, → need to estimate
Var $(Y_i) = \sigma^2$ }

$Y \sim (\mu, \sigma^2)$ both μ and σ^2 unknown
μ is estimated by $\bar{y}$

$s^2 = \frac{\sum (Y_i - \bar{Y})^2}{n-1}$ ← Degrees of Freedom

- In estimating μ, we have lost 1 degree of freedom.
- = number of parameters that you have estimated

SLR simple linear regression

$Y_i \sim N(\beta_0 + \beta_1 X_i, \sigma^2)$

Since we have to estimate β_0 and β_1 before we can go on to estimate σ^2, we loose 2 degrees of freedom

Define $Y_i - \hat{Y}_i = e_i$

The appropriate sum of squares is: SSE (error sum of squares)

$SSE = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2$

$= \sum (Y_i - b_0 - b_1 X_i)^2$

The estimate of σ^2 also known as MSE is

$$\hat{\sigma}^2 = MSE = \frac{SSE}{n-2} = \frac{\Sigma(Y_i - \hat{Y}_i)^2}{n-2}$$

It can be shown that

$$E(MSE) = \sigma^2$$

For the diamonds example:

$\hat{Y}_i = -259.36 + 3721.03 X_i$ regression equation

SSE =

Y_i	X_i
355	0.17
⋮	⋮
316	0.15

$$SSE = [355 - (-259.36 + 3721.03 \times 0.17)]^2 + \dots + [316 - (-259.36 + 3721.03 \times 0.15)]^2$$

$$= 46636$$

$$MSE = \frac{SSE}{n-2} = \frac{46636}{48-2} = 1013.82$$

Inferences about b_0 and b_1

Fundamentally interested in distributions, t-tests and confidence intervals.

Recall: If $Y_i \sim N(\mu, \sigma^2)$

Then $\hat{\mu}$ → estimate of μ is $\bar{Y}$

remember $\bar{Y} \sim N(\mu, \frac{\sigma^2}{n})$

So we look at sampling distributions of b_0 and b_1

Inference of b_1

point estimate for b_1 is $b_1 = \frac{\Sigma(X_i - \bar{X})(Y_i - \bar{Y})}{\Sigma(X_i - \bar{X})^2}$

23/10/13 Regression

ST2002

It can be shown $E(b_1) = \beta$

$$Var[b_1] = \frac{\sigma^2}{\Sigma(x_i - \bar{x})^2} = \frac{\sigma^2}{S_{xx}}$$

If σ^2 is known, $b_1 \sim N(\beta_1, \frac{\sigma^2}{S_{xx}})$

However, if σ^2 is unknown, we need to estimate $Var[b]$ and that is $\hat{Var}[b_1] = s^2_{b_1} = \frac{MSE}{S_{xx}}$

Standard error S.E is the square root of $\hat{Var}[b_1]$

$s.e[b_1] = \sqrt{\frac{MSE}{S_{xx}}}$ packages give out this

For the diamond example:

$s.e(b_1) = \sqrt{\frac{10.3383}{0.1510}} = 81.78$

Hypothesis testing for β_1 / b_1

$Y_i = \beta_0 + \beta_1 X_i$ β_1 has a physical meaning

Refer to change in Y for unit change in X.

Experts think change in Y unit change in X should be 7, whereas when we estimate β_1 from data it is 5

Test: $H_0 : \beta_1 = 7$ vs H_1 $\beta_1 \neq 7$ < 7 or > 7

Golden rule for t test

$$t_{calc} = \frac{(\text{estimator}) - (\text{value from } H_0)}{S.E.(\text{estimator})}$$

Confidence Interval $(\text{estimator}) \pm t_{critical} \, se(\text{estimator})$

↑

one you get from table

2 Regression

Diamonds data set.

$H_0 : \beta_1 = 0 \qquad H_1 : \beta_1 \neq 0$

$$t_{calc} = \frac{b_1}{se.(b_1)} = \frac{372.02}{81.79}$$

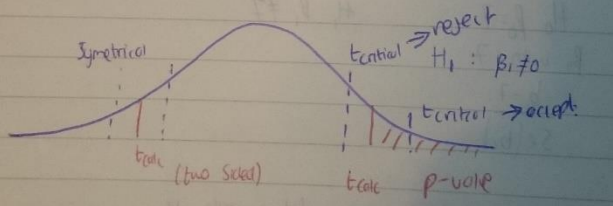
From the t-table $t_{0.975,\ 46} = 2.011$

Since $|t_{calc}| > t_{critical}$ we reject null hypothesis

Interpretation: we can reject the null hypothesis that the rate of change of price of diamonds with respect to carats is equal to zero.

The test also provides p-values

P-value: Given the null hypothesis is true, the p-value means the probability of observing a more extreme value that t_{calc}. (in either direction)



area to right of t_{calc}

area to right of t_{calc} is α

If p value $< \alpha \Rightarrow t_{critical} < t_{calc} \Rightarrow$ reject

24/10/13 Regression

Inference on β_1

$$Se(b_1) = \sqrt{\frac{MSE}{\sum (x_i - \bar{x})^2}} \quad \text{Sxx} = \sqrt{\frac{MSE}{S_{xx}}}$$

estimate

$$\frac{b_1 - \beta_1}{Se(b_1)} \sim t \text{ distribution with } (n-2) \text{ d.f.}$$

Test. parameter

H_0 $\beta_1 = 0$ vs H_1 $\beta_1 \neq 0$

$$t_{calc} \quad \frac{b_1 - \beta_1}{s.e.(b_1)} = \frac{b_1}{s.e.(b_1)} \quad (\beta_1 = 0 \text{ under } H_0)$$

~~$b_1 = \beta$~~

example H_0 $\beta_0 = 7$ H_1 $\beta_1 \neq 7$

Replace β_0 with 7.

$$t_{calc} = \frac{b_1 - 7}{Se(b_1)}$$

For a significance level α the rule is:

- If (t_{calc}) is $\leq t_{critical}$, then accept H_0.
- Reject otherwise
- $t_{critical} \Rightarrow$ value you get from t-tables α 0.05 d.f. $= n-2$

- Most often 5%

$$t_{critical} = t_{[1-\frac{\alpha}{2}, n-2]}$$

area d.f.

4 REGRESSION

We are very interested in
$H_0: \beta_1 = 0$ vs $H_1: \beta_1 \neq 0$

$$[Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i]$$
$\beta_1 = 0$ no relationship

Note that $\beta_1 = 0$
$\Rightarrow$ no association between X and Y

So, under the null hypothesis, the sampling distribution of
$\frac{b_1 - \beta_1}{SE(b_1)} \sim t$ distribution with $(n-2)$ d.f.

The $(n-2)$ d.f. is borrowed from $S.E(b_1)$

For our example

p-value $< 2 \times 10^{-16}$

$\alpha = 0.05$

$p << \alpha \Rightarrow$ reject H_0

Confidence Intervals for β_1:

estimator $\pm$ $t_{critical}$ $\times$ S.e(estimator)

The 95% CI for β_1 is

$b_1 \pm t_{0.975, n-2} \times S.e.(b_1)$

$t_{0.975, 40}$ $se(b_1)$

Example: $3721.03 \pm 2.011 \times 81.79$

$= 3721.03 \pm 164.48$

95% CI $= (3556.54, 3885.5)$

This is an example of Interval estimation

Interpretation:

Firstly a 95% CI does not mean that there is a 95% chance that the true β_1 lies in the interval

CI just means that if we were to compute intervals in this way from similar data sets, then 95% of those intervals will contain the true slope β_1

Correspondence between t-test and CI:

Both should give the same inference.

$\Rightarrow$ t-test rejects H_0 then CI should not contain 0.

(For H_0: $\beta_1 = 0$) if [(H_0: $\beta_1 = 7$) should not contain 7]

2/13 Regression

Testing of hypothesis and C.I

$H_0: \mu = \mu^*$ vs $H_1: \mu \neq \mu^*$

One approach → t-tables or p-value

Other is to compute C.I. for μ
Check if μ^* falls within the interval or not

Inference for β_1

CI: $b_1 \pm t_{crit} \, se(b_1)$

$(b_1 - t_{crit} \, s.e(b_1) \; ; \; b_1 + t_{crit} \, se[b_1])$

If $H_0: \beta_1 = 7$ (say) vs $H_1: \beta_1 \neq 7$

If 7 falls within CI we accept H_0 Reject O/W.

Inference about the intercept β_0

Mostly we are interested when $X=0$, makes sense

point estimate of β_0 $b_0 = \bar{y} - b_1 \bar{x}$

It can be shown $E(b_0) = \beta_0$

$Var(b_0) = \sigma^2 \left[\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right]$

where $S_{xx} = \sum (x_i - \bar{x})^2$

The sampling distribution of b_0 is

$b_0 \sim N\left(\beta_0, \sigma^2 \left[\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right]\right)$

$s.e[b_0]$ (square root of estimated variance) $= \sqrt{MSE\left[\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right]}$

↑ estimated value

Regression

From the diamond data set

$$se[b_0] = \sqrt{MSE\left[\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right]}$$

$$= \sqrt{1.1332\left[\frac{1}{48} + \frac{0.2042^2}{0.1510}\right]}$$

$$= 17.32$$

The sampling distribution of $\frac{b_0 - \beta_0}{s.e(b_0)} \sim$ t distribution with $(n-2)$ d.f

Hypothesis testing:

$H_0: \beta_0 = 0$ vs $H_1: \beta_0 \neq 0$

Test Statistic $t_{calc} = \frac{b_0}{s.e(b_0)}$

For α level of significance

* If $|t_{calc}| \leq t_{critical}$ accept H_0.

Reject otherwise

$t_{calc} = \frac{-254.63}{17.32} = -14.94$ [intercept is negative] Flaw.

For $\alpha = 0.05$, $t_{critical} = t_{0.95, 46} = 2.011$

Since $|t_{calc}| > t_{critical}$ we reject the null hypothesis.

The true intercept is not equal to zero, in light of the data at hand

R gives us p-value of 2×10^{-16} which is < 0.05.

CI. For β_0

$100(1-\alpha)\%$ CI for β_0:

$b_0 \pm t_{critical} \cdot s.e.[b_0]$

If you consider the example

$-254.63 \pm 2.011 \times (17.32)$

$\Rightarrow -254.63 \pm 34.38 \Rightarrow (-289.46, -220.48)$

13 Regression

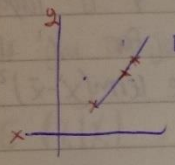
(-294.46, -224.8)

Since this does not contain 0, we can get the same conclusion as before: Reject H_0

Interpretation of C.I.: very important

- 95% of time when calculate CI, it will contain true parameter.

Inference on $E[Y]$ for $X = X_{some\ value}$



One of the major goals in regression is to make inference about the mean of the distribution of Y at X.

We want to construct point and interval estimates for

$E(Y|X = x_{some\ value})$

Let x' be some value of X for which we want to estimate the mean response

$E(Y|x) = \beta_0 + \beta_1 x'$

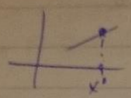
The point estimate, say y' of $E[Y|x']$

$\hat{y}' = b_0 + b_1 x'$

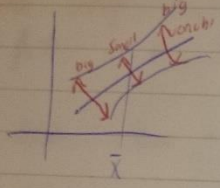
$$E(\hat{y}') = E[b_0 + b_1 x']$$
$$= E[b_0] + E[b_1 x']$$
$$= E[b_0] + E[b_1] x' \quad (x' \text{ is a constant})$$
$$= \beta_0 + \beta_1 x' \rightarrow \text{true value}$$

4

$$\text{Var}(\hat{y}') = \sigma^2\left[\frac{1}{n} + \frac{(x'-\bar{x})^2}{S_{xx}}\right]$$



In term of a dataset, $x' = 0.33$



Note that variablity of $\hat{y}'$ is largely affected by how much far x' is from $\bar{x}$, through the term $(x'-\bar{x})^2$

Hence further away causes greater variability. Than values closer to $\bar{x}$

(Talking about variability of a point on the regression line)

$$\hat{y}' \sim N\left((\beta_0+\beta_1 x'), \sigma^2\left[\frac{1}{n} + \frac{(x'-\bar{x})^2}{S_{xx}}\right]\right)$$

$$s.e(\hat{y}') = \sqrt{MSE\left[\frac{1}{n} + \frac{(x'-\bar{x})^2}{S_{xx}}\right]}$$

Finally, sampling distribution of $\frac{\hat{y}' - (\beta_0 + \beta_1 x')}{s.e(\hat{y}')}$ is t distribution with $n-2$ degree freedom

HYPOTHESIS TESTING

$H_0: \hat{y}' = y'$ vs $H_1: \hat{y}' \neq y'$

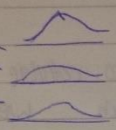
$$t_{calc} = \frac{\hat{y}' - y'}{se[\hat{y}']}$$

CI: $\hat{y}' \pm t_{critical}\, s.e(\hat{y}')$

10/8. Regression. ①

Assignment Q3c write down the values of SSE, MSE, and RSE

$RSE = \sqrt{MSE}$ residual standard error

Y	X constant	
y_1	x_1	$y_1 \rightarrow$  different means by
y_2	x_2	ally's Same variance σ^2 (unknown)
y_n	y_n	

If you can join their means by a straight line, then do linear regression

$E(Y_i|X_i) = \beta_0 + \beta_1 X_i$

$\Rightarrow Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$

$= E(Y_i|X_i) + \epsilon_i$

Estimate this line estimate β_0 and β_1

estimated by b_0 and b_1

$b_1 = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2}$ sampling distribution of this one

$b_0 = \bar{y} - b_1 \bar{x}$ also ok

Estimate of a point on a line $\hat{y}_i = b_0 + b_1 x_i$

$E(y|x')$

For some value $x = x'$, estimate of $E(y|x') = \hat{y}'$

$\hat{y}' = b_0 + b_1 x'$

$s.e[\hat{y}'] = \sqrt{mse\left(\frac{1}{n} + \frac{(x' - \bar{x})^2}{\sum (x_i - \bar{x})^2}\right)}$

Diamond dataset

$\bar{x} = 0.2012$

1. $x' = 0.2$ · plug in:

$\hat{y}' = -2596 + 3721.03 * 0.2 = 484.57$

fitted value here refers to the average price of diamonds with weight 0.2 carats.
if you collected data on every 0.2 carat diamond this would be the average of all 0.2 carat diamonds.

So, if one was to calculate the average of price of all the diamonds which weigh 0.2 carat. Then it would be 484.57 $.

$s.e.(\hat{y}') = \sqrt{MSE\left[\frac{1}{n} + \frac{(x'-\bar{x})^2}{\Sigma(x_i-\bar{x})^2}\right]}$

$= \sqrt{101.383\left[\frac{1}{48} + \frac{(0.2-0.2042)^2}{0.1516}\right]}$

$= 4.609$ closer to mean

CI: $\hat{y}' \pm t_{0.975,46}\, se(\hat{y}') = (475.307,\ 493.85)$

$x' = 0.33$ $\quad s.e.\,\hat{y}' = \sqrt{101.383\left[\frac{1}{48} + \frac{(0.33-0.2042)^2}{0.1516}\right]}$

$\hat{y}' = 968.31$ $\quad = 11.267$ away from mean

CI: $\hat{y}' \pm t_{0.975,46} * se(\hat{y}') = (945.646,\ 990.9739)$.

diamond.lm $ coefficients → gives intercept B_0 and B_1

b0 = diamond.lm $ coefficients [1] ← First position

RSE = summary.diamond.lm $ sigma

getting mean = mean(carats)
mean(price)

making vector name = c (1, 2, 3, 4)

ST2002: Introduction to regression

Computer laboratory 5

Name: DAVID WEITBRECHT

Student No.: 12300644

A way for setting one's path in R.

Regression

13/11/13 C.I. / t-test / Sampling Distribution

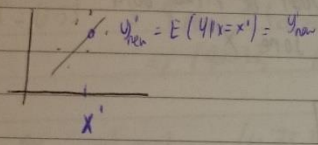
of $E(Y|X=x') = y'$

Prediction of a new observation

We consider prediction of a new observation y corresponding to some x.

Let us denote a new X as x' (x takes the value x'), and we call the new observation as y'_{new}

y'_{new} is a single observation corresponding to x'



$$Z \sim N(\mu, \sigma^2)$$
$$E[Z] = \bar{Z} \sim N(\mu, \frac{\sigma^2}{n})$$

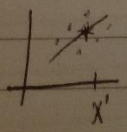
Big distinction between $E(Y|X=x') = y'$ in the previous section and our new response y'_{new}.

First case refers to mean of distribution of Y for some $X=x'$.

In the later case, we predict an individual outcome drawn from the distribution of Y for $X=x'$.

$\Rightarrow$ Greater variability for the 2nd case.

Point estimate: We predict y'_{new} by its estimate $\hat{y}'_{new}$, given by $\hat{y}'_{new} = b_0 + b_1 x'$ - point on regression line.



Note that the point estimate is same as before

Properties

$E(\hat{y}'_{new}) = y'_{new}$

$Var(\hat{y}'_{new}) = \sigma^2\left[1 + \frac{1}{n} + \frac{(x'-\bar{x})^2}{\sum(x_i-\bar{x})^2}\right]$

For $E(y \mid x = x') = y'$

$Var(y') = \sigma^2\left[\frac{1}{n} + \frac{(x'-\bar{x})^2}{\sum(x_i-\bar{x})^2}\right]$

Note that, $var(\hat{y}'_{new})$ has 2 components,

1. The variance of the sampling distribution of fitted value $\hat{y}'$
2. The variance of distribution of y at some x.

$Var(\hat{y}'_{new}) = \sigma^2 + \sigma^2\left[\frac{1}{n} + \frac{(x'-\bar{x})^2}{\sum(x_i-\bar{x})^2}\right]$

$= var(\hat{y}') + var(y')$

$Var(\hat{y}'_{new}) > Var(\hat{y}')$

Estimate of variance is:

$\hat{Var}(\hat{y}'_{new}) = s^2(\hat{y}'_{new})$

$= MSE\left[1 + \frac{1}{n} \frac{(x'-\bar{x})^2}{\sum(x_i-\bar{x})^2}\right]$

$S.e(\hat{y}'_{new}) = \sqrt{mse\left[1 + \frac{1}{n} + \frac{(x'-\bar{x})^2}{\sum(x_i-\bar{x})^2}\right]}$

$x' = \bar{x} \Rightarrow Var(\hat{y}'_{new}) = MSE\left[1 + \frac{1}{n}\right]$

Intervals are now called prediction intervals

$\hat{y}'_{new} \pm t_{critical}\, S.e(\hat{y}'_{new})$

$\hat{y}'_{new} \pm t_{critical} \sqrt{MSE\left[1 + \frac{1}{n} + \frac{(x'-\bar{x})^2}{\sum(x_i-\bar{x})^2}\right]}$

lm(price ~ carats)
lm (formula, dataset

diamond.dataframe → 2 columns price and carats

price = diamond.dataframe $ price;
carats = diamond.dataframe $ carats
lm (price ~ carats)
could of used lm (Price ~ carats, diamond.dataframe)

coef = diamond.lm $ coefficients
b_0 = coef[1]
b_1 = coef[2]

Significance code

*	0.05
**	0.01
***	0.001
.	0.1

explain value

Confidence Interval for β_0 and β_1 ⇒ confint (lm object)
confint (lm object, level=0.95)
= confint (price, 0.95)

new.X = data.frame (X = c(x_1, x_2, x_3))

20/11/13 2 Regr 2

1 iii The plot suggests a strong linear relationship between size of company and delay time. A strong high correlation coefficient of -0.84 supports this.

There is a causal relationship between the 2 variables, linear regression can be used

b Since Size = x and delay = y
The "true" equation is given by
$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i \quad i = 1, \dots n$
where β_0 and β_1 are parameters and ε_i are the errors

The fitted regression line is:
$\hat{Y}_i = 25.47 - 0.18 X_i$.

Assumptions
i. The predictor variable X is assumed to be constant.
ii. The observations Y_i are independent.
iii. For a given x_i, the corresponding Y_i is normally distributed.
iv. All Y_i's have the same variance σ^2.
v. The means of Y_i given X_i can be joined by a straight line
$E(Y_i | X_i) = \beta_0 + \beta_1 X_i$
β_1 = Slope parameter. Signifys the change in Y for unit change in X
β_0 = Intercept. Value of Y for $X = 0$

c For β_0
The hypothesis tested is: $H_0 : \beta_0 = 0 \quad H_1 : \beta_0 \neq 0$
The calculated t value is $t_{cal} = 11.687$.
From table $t_{0.975,18} = 2.306$
$\rightarrow t_{(1-\frac{\alpha}{2}, n-2)}$

Since $|t| > t_{0.975,18}$ we reject null hypothesis

3. Regression

Example

Carats = 0.2

Instead of estimating interval for the true mean price of diamond that weigh 0.2 carats, we are interested in estimating the interval within which the true price of a single diamond of weigh 0.2 lie

For $x'=0.2$ $\hat{y}'_{new} \pm t_{0.975,46}\ se(\hat{y}'_{new})$

$\Rightarrow (419.92, 549.34)$

For $x'=0.33$ $= \hat{y}'_{new} \pm t_{0.975,46}\ se(\hat{y}'_{new})$

$\Rightarrow (900.32, 1036.30)$

CI for $x'=0.2$ (475.31, 494.85)

CI for $x'=0.33$ = (945.65, 990.98)

Prediction interval → P.I

Regression R

Vector

(2,5,7)
c(2,5,7)

matrix

1 4 7
2 5 8
3 6 9

Data frame

A	B
1	4
2	5
3	6

how to have name

A = c(1,2,3)

A = data.frame(A = c(1,2,3)

function to read data from file.

read.table (.txt, csv etc)

read.table (filename, header, sep)

header = true

excel

	Price	Card
row1	:	:
row2	:	:

data.read = read.table (sonefile, header = True, sep = "I")

If you say false it will assign names on its own

read.csv (filname, header, sep, col.names)

col.name) = c("A", "B")

csv = comma, seperated values

Regression

20/11/13 Assignment 1 Answers

1a ii) correlation coeffn $= \dfrac{\frac{1}{n-1}\sum (x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\frac{1}{n-1}\sum (x_i-\bar{x})^2}\sqrt{\frac{1}{n-1}\sum (y_i-\bar{y})^2}}$

$= \dfrac{A}{\sqrt{B}\sqrt{C}}$

$A = \frac{1}{n-1}\sum (x_i-\bar{x})(y_i-\bar{y})$

$= \frac{1}{n-1}\sum (x_iy_i - x_i\bar{y} - y_i\bar{x} + \bar{x}\bar{y})$

$= \frac{1}{n-1}\sum [x_iy_i - \sum x_i\bar{y} - \sum y_i\bar{x} + \sum \bar{x}\bar{y}$

$= \frac{1}{n-1}\sum x_iy_i - n\bar{x}\bar{y} - n\bar{y}\bar{x} + n\bar{x}\bar{y}$

$= \frac{1}{n-1}\sum x_iy_i - n\bar{x}\bar{y}$

$= \frac{1}{n-1}\sum x_iy_i - n(\frac{1}{n}\sum x_i)(\frac{1}{n}\sum y_i)$

$= \frac{1}{n-1}\sum x_iy_i - \frac{\sum x_i \sum y_i}{n}$

$B = \frac{1}{n-1}\sum (x_i-\bar{x})^2$

$= \frac{1}{n-1}\sum (x_i^2 + \bar{x}^2 - 2x_i\bar{x})$

$= \frac{1}{n-1}\sum x_i^2 + n\bar{x}^2 - 2\bar{x}\sum x_i$

$= \frac{1}{n-1}\sum x_i^2 + n\bar{x}^2 - 2\bar{x}(n\bar{x})$

$= \frac{1}{n-1}\sum x_i^2 - n\bar{x}^2$

$= \frac{1}{n-1}\left(\sum x_i^2 - n(\frac{1}{n}\sum x_i)^2\right)$

$= \frac{1}{n-1}\left(\sum x_i^2 - \frac{(\sum x_i)^2}{n}\right)$

$C = \frac{1}{n-1}\left(\sum y_i^2 - \frac{(\sum y_i)^2}{n}\right)$

From data: $n=10$ $\sum x_i = 435$ $\sum x_i^2 = 27771$

$\sum y_i = 173$ $\sum y_i^2 = 3481$ $\sum x_iy_i = 5990$

Plug into formula $= -0.8429$

2.

E. CI for mean response size = 40

$x' = 40 \quad \hat{y}' = b_0 + b_1 x'$

$= 18.14144$

$s.e.(\hat{y}') = \sqrt{MSE\left[\frac{1}{n} \frac{(x'-\bar{x})^2}{\sum (x_i - \bar{x})^2}\right]}$

$= 1.23896$

CI $\quad \hat{y}' \pm t_{0.975,8}\, se(\hat{y}') = (15.2844,\ 20.9985)$

~~t_{8} 1.24426~~

2A. $\sum x_i = 27 \quad \sum x_i y_i = 180$

$\sum x_i^2 = 330 \quad \sum y_i = 210$

$n = 16 \quad SSE = 17.50$

Regression model 1) plot : $Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$

or

$E(Y_i | X_i) = \beta_0 + \beta_1 X_i$

$\hat{\beta_1} = b_1 = \dfrac{\sum x_i y_i - \frac{\sum x_i \sum y_i}{n}}{\sum x_i^2 - \frac{(\sum x_i)^2}{n}} = 4.6605$

$\hat{\beta_0} = b_0 = \bar{y} - b_1 \bar{x}$

$= \frac{1}{n}(\sum y_i - b_1 \sum x_i) = 5.26$

b. Hypothesis $\quad H_0 : \beta_1 = 6 \quad$ vs $\quad H_1 : \beta_1 \neq 6$

$t = \dfrac{b_1 - 6}{s.e.(b_1)}$

$s.e.(b_1) = \sqrt{MSE\left(\frac{1}{n} + \frac{\bar{x}^2}{\sum (x_i - \bar{x})^2}\right)}$

$\sqrt{\dfrac{MSE}{\sum (x_i - \bar{x})^2}}$

1/11/13 3 Regression

$S.e(b_1) = 0.6624$

$t = \frac{b_1 - 6}{S.e(b_1)} = \frac{6.66 - 6}{0.6624} = -2.62$

$t_{0.975,14} = 2.145$ $|t| < t_{0.975,14}$ accept null hypothesis.

accept H_0 that slope or parameter is equal to 6

SOLUTION:

```
box.office.dataframe = read.csv("boxoffice.csv", header = TRUE)
box.office.lm = lm(currentweek ~ LastWeek, boxoffice.dataframe)
coef(box.office.lm)
Summary.box.office = summary.lm(box.office.lm)
summary.box.office$coefficients

RSE = summary.box.office$sigma
MSE = RSE^2

confint(box.office.lm)
fitted(box.office.lm)[1]
fitted(box.office.lm)[length(box.office.dataframe$currentweek)]
new.boxoffice = dataframe(lastweek = 400000)
predict.lm(box.office.lm, newdata = new.boxoffice, se.fit=T, interval=c("confidance"))
predict.lm(box.office.lm, newdata=new.boxoffice, se.fit=T, interval=c("prediction"))
```

7/11/13 Regression.

Problems in Assignment

Q1 E $H_0: \beta_0 = 0$
$H_1: \beta_1 = 0$ β_0 and β_1 cannot be in same hypotheses!

Should be $H_0: \beta_0 = 0$ $H_0: \beta_1 = 0$
$H_1: \beta_0 \neq 0$ $H_1: \beta_1 \neq 0$

β_0, β_1 are theoritical values of parameter

b_0, b_1 are estimate of that parameter

$se(\beta_1)$ wrong
$s.e.(b_0)$ or $se(b_1)$ correct.

2C. IF $x_i = 0$ Yield $= \beta_0$
According to the question the yield will be 0 tonnes TEST
H_0 $\beta_0 = 0$ vs H_1 $\beta_0 \neq 0$

$$se(b_0) = \sqrt{MSE\left[\frac{1}{n} + \frac{\bar{x}^2}{\Sigma(x_i - \bar{x})^2}\right]}$$

$$= \sqrt{\frac{SSE}{n-2}\left[\frac{1}{n} + \frac{\bar{x}^2}{\Sigma(x_i - \bar{x})^2}\right]}$$

$$= 3.01$$

$$t = \frac{b_0}{se(b_0)} = \frac{5.26}{3.01} = 1.74$$

$t_{0.975, 14} = 2.145$

$\Rightarrow |t| < t_{0.975, 14}$ Do not reject the null hypothesis

The analyst was right in pointing out that allocating 0 hectares could the yield to be 0 tonnes

20/11/13 3 Regression

Thus, we can conclude that the true intercept is not equal to 0, given the data.

In other words the mean delay time for a company of size 0 is not equal to 0; as seen from the test

For β_1

$t = -4.432$

$t_{0.975,8} = 2.306$

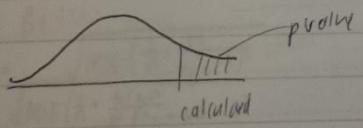
$\Rightarrow |t| > t_{0.975,8}$ we reject the null hypothesis that the slope is 0.

In other words, the change in ~~xxxxx~~ delay time for unit change in size is not equal 0.

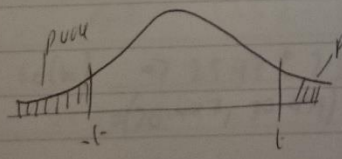
P-value:

Given that the null hypothesis is true, p-value measures the probability of observing a more extreme value than that of the calculated one

One Sided test:



Two Sided test:



If is the area in tails, more extreme than the calculated test statistic

If p-value is less than α, we reject the null hypothesis

$|t| > t_{critical} \equiv p < \alpha$ In this case, p-value for both β_1 and β_0 are less than $\alpha = 0.05$, hence we reject null hypothesis

1/13. Regression Assignment Solution

$|t| > t_{critical}$ $\Leftrightarrow$ related to p-value (<0.05)

If t = reject p will also reject

1.d. Estimate of $\sigma^2 = MSE = \frac{SSE}{n-2}$

$SSE = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} (y_i - (b_0 + b_1 x_1))^2$

y_i	$\hat{y}_i = 25.47 \pm 0.8x$	$y_i - \hat{y}_i$
24	→ *	
14	→ *	
⋮	⋮	

$= 121.14071$

$MSE = \frac{121.14071}{(18-2)} = 15.14$

R output:

$RSE = 3.891$ $RSE^2 = MSE = (3.891)^2 = 15.139 = 15.14$

CI For β_0:

$s.e.(\beta_0) = \sqrt{MSE\left(\frac{1}{n} + \frac{\bar{x}^2}{\sum (x_i - \bar{x})^2}\right)}$

$= \sqrt{MSE\left(\frac{1}{n} + \frac{\frac{1}{n}(x_i)^2}{\sum (x_i - \frac{1}{n}\sum x_i)^2}\right)}$

$= 2.1795$

$b_0 \pm t_{critical} \, se(b_0) \Rightarrow 25.47 \pm 2.306(2.1795)$

$= (20.447, 30.499)$

C.I for β_1:

$s.e.(b_1) = \sqrt{\frac{MSE}{\sum (x_i - \bar{x})^2}} = S_{xx}$

$= 0.04136$

$b_1 \pm t_{critical} \, s.e.(b_1) = (-0.274, -0.0879)$

Anova - Analysis of Variance

Total Variation = Systematic Variation + Chance Variation.

Aim to max Systematic and min chance variation.

Given any random variable $y_1 \dots y_n$ the variability is denoted by variance.

$$Var[y] = \frac{1}{n-1} \Sigma [y_i - \bar{y}]^2$$

In other words, a measure of variability is $\Sigma (y_i - \bar{y})^2$
→ refers to total variation.

In regression context, this is known as Total Sum of Squared or SSTO

$$= \Sigma [y_i - \bar{y}]^2 = \text{Corrected Sum of Squares}$$

Note: Uncorrected Sum of Squared = Σy_i^2

SSTO = 0

⇒ All y_i's are equal to $\bar{y}$ (all same number eg 5)

The greater the value of SSTO, the greater the variability in the data.

Aim:

Basic Aim is to partition total variability (SSTO) into 2 components: ① Variability due to model
② Variability caused due to error or chance

① Variability explained by regression equation
This is used to find out if our model was good enough in explaining variation in the data

② We have done this already using Error Sum of Squared SSE

$$SSE = \Sigma [y_i - \hat{y}_i]^2$$

4/11/13 Regression

The value quantified the variability of data around the fitted line
↳ Refer to chance variation

The rest of the variability is defined as Regression Sum of Squares or SSR
$SSR = \Sigma (\hat{y}_i - \bar{y})^2$

The greater the SSR, the better is the regression eqn in accounting for variability in the data

SSTO = SSR + SSE
Total Variation = Systematic variation + chance variation

* If SSE = 0, means all the data lies perfectly on the regression line

$SSE = \Sigma(y_i - \hat{y}_i)^2$ equals 0 when $y_i = \hat{y}_i$

Partitioning the total Sum of Squares
$y_i - \bar{y} = \underbrace{\hat{y}_i - \bar{y}}_{\text{due to regression}} + \underbrace{y_i - \hat{y}}_{\text{chance}}$

Squaring and Summing them on both sides gives
$\Sigma(y_i - \bar{y})^2 = \Sigma(\hat{y}_i - \bar{y})^2 + \Sigma(y_i - \hat{y}_i)^2$
The intersection / cross-product then is equal to 0.

To prove this use the following properties:
(1) The regression line can be written as $\hat{y}_i = \bar{y} + b_1(x_i - \bar{x})$
Proof: $\hat{y}_i = b_0 + b_1 x_i$
$b_0 = \bar{y} - b_1 \bar{x}$
$\Rightarrow \hat{y}_i = \bar{y} - b_1 \bar{x} + b_1 x_i \Rightarrow = \bar{y} + b_1(x_i - \bar{x})$

4

2 The Sum of observed value y_i is equal to that of fitted values:

$$\Sigma y_i = \Sigma \hat{y}_i$$

3 If $e_i = y_i - \hat{y}_i$

$$\Sigma x_i e_i = 0$$

4 $\Sigma \hat{y}_i e_i = 0$

Regression

New dat = data.frame (cInSize = c (100, 200, 300))

		systematic		chance
SSTo	=	SSR	+	SSE
total	=	due to regress	+	error sum of squa

$\Sigma (y_i - \bar{y})^2 = \Sigma (\hat{y}_i - \bar{y})^2 + \Sigma (y_i - \hat{y}_i)^2$

better model is SSE is smaller

the right hand side used to judge new model
Also used to compare with other models

Degrees of Freedom

$\Sigma (y_i - \bar{y})^2$ has (n-1) df SSTo only 1 for estimating $\bar{y}$.

$\Sigma (y_i - \hat{y}_i)^2$ has (n-2) df SSE

df [SSTo] = df [SSR] + df [SSE]
$\Rightarrow$ df [SSR] = 1.
Partitioning of degrees of freedom

Mean Sum of Squares

This is defined as the ratio of sum of squares and the corresponding d.f.

$MSE = \frac{SSE}{n-2}$

Mean squared error.

$MSR = \frac{SSR}{1}$
Mean Squared regression

Source of Variation	SS	df	Mean Square MS	F
Due to regression	$SSR = \Sigma(\hat{y} - \bar{y})^2$	1	$MSR = \frac{SSR}{1}$	$F = \frac{MSR}{MSE}$
Error	$SSE = \Sigma(y - \hat{y})^2$	n-2	$MSE = \frac{SSE}{n-2}$	
Total	$SST_o = \Sigma(y_i - \bar{y})^2$	n-1	←	

The F value is used to test the following hypothesis:

$H_0 : \beta_1 = 0$ vs $H_1 : \beta_1 \neq 0$

Tells whether it is meaningful to do a regression model

If $F_{calc} \leq F_{(1-\alpha, df1, df2)}$ then accept H_0

$F > $ reject H_0

df2 \ df1	1	2	
1	"	"	$F_{(0.05, 1, 3.)}$
2	"	"	"
3	"	"	"

Diamond Dataset Anova table

	SS	DF	MS	F
Regression	2144218	1	214428	211499
Error	46636	46	1014	
Total	2145232	47	—	

$F(0.95, 1, 46) = 7.95$.

Since $F(211499) \gg F(0.95, 1, 46)$ we reject H_0

R anova.lm (data, lm) ← lm object

$H_0 : \beta_1 = 0$ vs $H_1 : \beta_1 \neq 0$ t-test had been used

Relationship between t-test and f-test

2

Residual Analysis

Basic idea is to check if assumption regarding SLR is followed or not

Assumptions:

1. Means of Y_i, given X_i can be joined by a straight line
2. Mean of Y_i given X_i = $\beta_0 + \beta_1 X_i$
2. Y_i's are independent.
3. Y_i's have a constant variance
4. Y_i's rise from a normal distribution

Equivalent:

1. Mean of ε_i equal to zero
2. ε_i's are independent
3. Variance of ε_i's is constant
4. ε_i follows a normal distribution

$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$ > imply $\varepsilon_i = Y_i - E[Y_i]$

$E(Y_i) = \beta_0 + \beta_1 X_i$

or $\hat{Y}_i = b_0 + b_1 X_i$ } $e_i = Y_i - \hat{Y}_i$ — residual

Residual Analysis

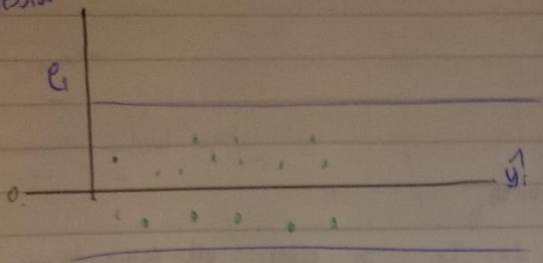
Deal with residuals e_i

1. Check for mean error term equal to 0 and error terms having constant variance.

- Plot residual against fitted value e_i vs $\hat{Y}_i$
- Look for pattern
- No pattern implies ⇒ Randomly distributed around 0 means assumption hold

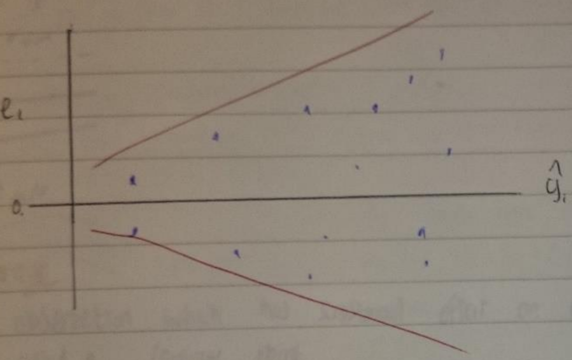
3 Regression

Good



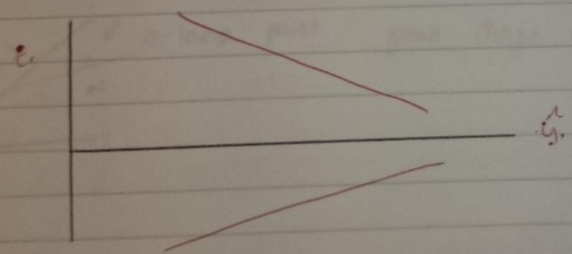
lies within band

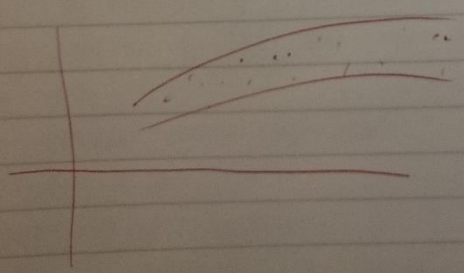
Bad



Funnel shaped band width

Constant variance violated





i) violated ...

ii) Constant varian...

3 Regression

Relationship between t-test and F-test

Can be shown that $f_{calc} = (t_{calc})^2$

For diamond example: $t_{calc} = t_{calc} = (45.44)^2 = 2114.99 = F$.

So for some level of α we can use either the t-test or the F-test for $H_0: \beta_1 = 0$ vs $H_1: \beta_1 \neq 0$

t-test is more versatile, since it allows us to test for one sided alternative

Mean $H_0: \beta_1 = 0$ $H_1: \beta_1 > 0$ or < 0 using t-test.

In hypothesis β_1 or β_0

In t-test use b_0 or b_1

13 Regression Anaa

Sources of variation	SS	df	Mean Square	F
Regression	SSR	1	MSR	$= \frac{MSR}{MSE}$
Error	SSE	n-2	MES	
total	SSTo	n-1	—	—

F is used to test $H_0: \beta_1 = 0$ vs $H: \beta_1 \neq 0$

Compare F with F critical $= F_{(1-\alpha, 1, n-2)}$ (df. SSR, df. SSE)

R^2 Statistic

$R^2 = \frac{\text{Sum of Squares due to Regression}}{\text{Total Sum of Squares}} = \frac{SSR}{SST_o} = 1 - \frac{SSE}{SST_o}$

R^2 measures the proportion of total variation around the mean as defined by the regression equation.

R^2 lies between 0 and 1.

R^2 is the correlation coefficient between y and ŷ.

In SLR, $R^2 = r^2$, where r is the correlation of x and y

Summary.lm() → in output multiple R-Sq is R^2.

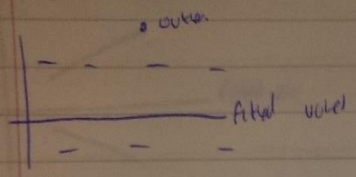
For a good fit, R^2 is close to 1.

For diamond dataset, $R^2 = 0.9783$

For bad fit, $R^2 < 0.7$ no set in stone

2/11 Regress

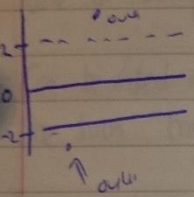
residual



Usually points outside the range of -2, +2 are considered outliers

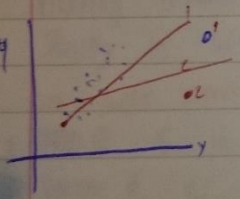
(not a proper rule more ga)

residuals



Leverage

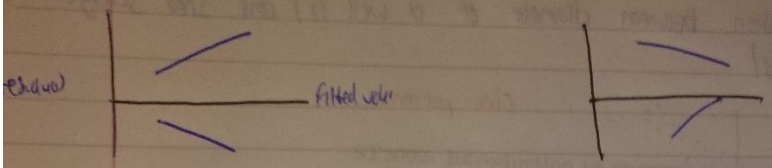
-An observation which has substantial effect on the simple linear regression fit is called a leverage point.



← leverage point

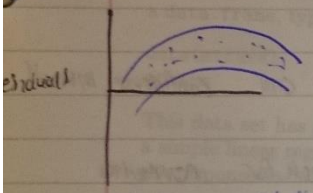
points change the whole data!!!

2/13 Regression



This problem mean variance is not constant.
Name of the problem is "heteroscedasticity" ⇒ $Var(\epsilon_i) \neq \sigma^2$ for some i

Sometimes: Can be fixed by transformation of Y and X
One such transformation is square root transform

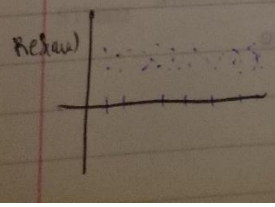


A non linear model needs to be fit
Eg $Y_i = \beta_0 + \beta_1 X_i + \beta_2 X^2 + \epsilon_i$
or higher " = " $+\beta_3 X_i^3 + \epsilon_i$

Plotting residuals vs fitted is similar to plotting residuals vs X
Strictly in SLR.

We were checking: in lab
i mean of ϵ_i is 0
ii Variance of ϵ_i is constant

Assumption:
- All ϵ_is are independent of each other
- It is only possible this to tell when there is some sort of time order in the data
⇒ plot residuals against time order
No pattern ⇒ ok
pattern ⇒ not independent

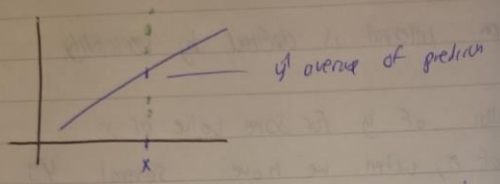


2

Alternative

Average r.v. Y_i = mean + chance variation

Hence, variability for a single Y_i is greater than variability of mean



CI of mean in some sense refers to how much the point on the regression line can vary

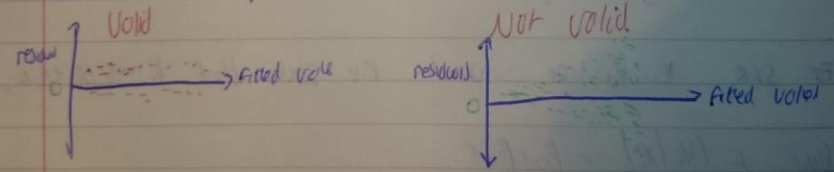
PI of a single point refers to in some sense variability of any of the green dots

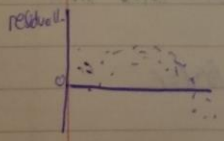
2010/2011 Q1

Q 1b

i. constant.

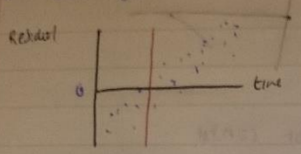
Constant variance for ε_i for all i.





2

Fit a regression between diameter of a well (x) and shear strength of the well (y)



Clear pattern

The observations are not independent

There is correlation between successive observations (points) (auto-correlation)

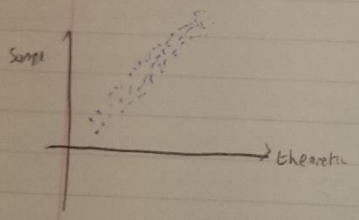
Assumption:

The error ε_i are normally distributed.

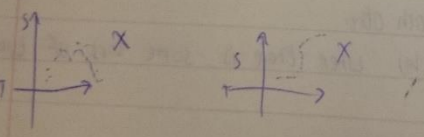
-Normal distribution necessary for computing t-tests and confidence intervals

-Plot Q-Q Plot or normal probability plot: is used to check for normality

Here, the ordered residuals are plotted against ordered expected normal scores



IF points lie on a straight line, then normality assumption holds



Outliers

-Outliers are extreme observations

-Detection of outliers is better done with Standardized residuals.

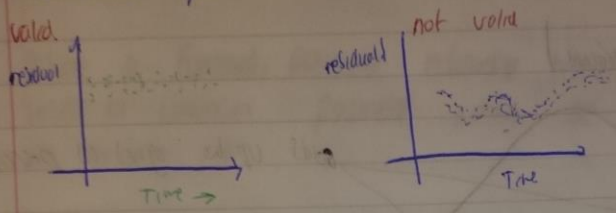
-Standardized residuals → $\frac{e_i - \bar{e}}{\sqrt{MSE}} = \frac{e_i}{\sqrt{MSE}}$ since $\bar{e} = 0$

-Do a residual plot residual vs fitted values

3 Regression

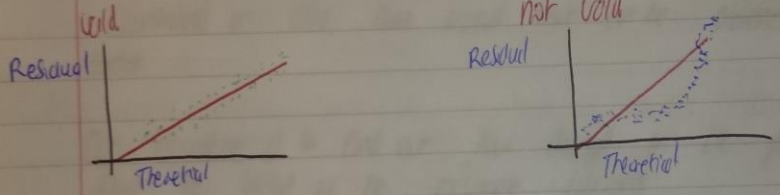
(ii) independence of ε_i

Time order plot



(iii) Normality of ε_i

Q-Q plot



c i $H_0: \beta_0 = 0$ vs $H_1: \beta_0 \neq 0$

Compar $t_{calculated}$ with $t_{critical}$

$t_{critical} = t_{0.975,8} = 2.306$

$|t| = 1.69 \gg t_{0.975,8}$ reject H_0, thus we reject that time taken to adopt the innovation for a company of size = 0.

ii. $H_0: \beta_1 = 0$ vs $H_1: \beta_1 \neq 0$

$t_{0.975,8} = 2.306$

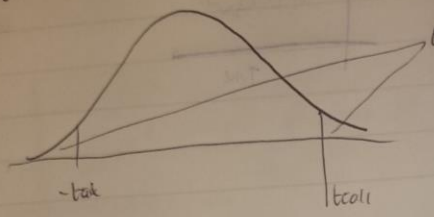
$|t| = 4.43 \gg t_{critical}$ reject H_0.

reject H that there is zero change in time taken to adopt innovation for 1 unit increase in size

4

c. P-value is the probability of getting a t value more extreme than the one we got in the test.

$P(|t| > t_{calculated})$



Regression

Confidence and Prediction intervals for prediction

Interval = point estimate ± $t_{critical}$ * S.E.(estimate)
variation or variability

The width or range of an interval is defined by variability.

(CI) for mean prediction of y for some value of x
For a certain value of x, when we have several y's then we construct the mean (the average) of those y's and finally get the standard error of average of y

This way we get CI of average of y for a certain value of x.

$Y_1 \dots Y_n$ are normal r.v. such that $Y_i \sim N(\mu_i, \sigma^2)$

Then $\bar{y} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$
$\Rightarrow$ variability of $\bar{y}$ < variability of y_i for all i.

For SLR, $Y_1 \dots Y_n$ are normal r.v. such that $Y_i \sim N(\beta_0 + \beta_1 x_i, \sigma^2)$

Now $E(Y_i | x_i) = \beta_0 + \beta_1 x_i$
Average of y for $X = x_i$
$\Rightarrow$ average of y for $x = x_i$, has less variability than Y_i for $X = x_i$.

In SLR, the P.I. is constructed for a single Y_i for $X = x_i$.

Hence PI is wider than CI for prediction