Regression Analysis 7th EDITION

WILLIAM MENDENHALL TERRY SINCICH

Regression Modelling Procedure:

1. Hypothesise the form of the model for $E(y)$
2. Collect the sample data
3. Use the sample data to estimate unknown parameters in the model
4. Specify the probability distribution of the random error term, and estimate any unknown parameters of this distribution
5. Statistically check the usefulness of the model
6. When satisfied that the model is useful, use it for prediction, estimation etc

Types of regression Data

1. Observational - values of x's are uncontrolled
2. Experimental - values of x's are controlled via a designed experiment

The random error component ϵ represents all unexplained variations in y's caused by important but omitted variables or by unexplainable random phenomena

SSE = Sum of Squared Errors = $\sum (y_i - \hat{y})^2$

~~SSR = Sum of Squares of residuals -~~

4 Assumptions of linear model

1. Mean of the probability distribution of ϵ is 0. That is, the average of the errors over an infinitely long series of experiments is 0 for each setting of independent variable x. This assumption implies that the mean value of y, $E(y)$, for a given x is $E(y) = \beta_0 + \beta_1 x_i$
2. The variance of the probability distribution of ϵ is constant for all settings of independent variable x. For straight line; the variance of ϵ is equal

2.
to a constant say σ^2 for all values of x.

3 The probability dist of ϵ is normal

4 The errors associated with any two different observations are independent. That is, error associated with one value of y has no effect on the error associated with other y values

Estimation of σ^2 and σ for Straight line First order Model

$$s^2 = \frac{SSE}{\text{D.F. for error}} = \frac{SSE}{n-2} \qquad s = \sqrt{s^2}$$

Where $SSE = \sum(y_i - \hat{y})^2$

$$= SS_{yy} - \hat{b}_1 SS_{xy}$$

$$SS_{yy} = \sum(y_i - \bar{y})^2 = \sum y_i^2 - n(\bar{y})^2$$

We refer to s as the estimated standard error of the regression model

Correlation Coefficient.

The Pearson product moment coefficient of correlation r is a measure of the strength of the linear relationship between two variables x and y :

$$r = \frac{SS_{xy}}{\sqrt{SS_{xx}SS_{yy}}}$$

Interpretation of the Coefficient of Determination, r^2

About $100(r^2)\%$ of the sample variation in y (measured by the total sum of squares of deviation $\sum(y_i - \bar{y}_i)^2$ about their mean $\bar{y}$) can be explained by (or attributed to) using x to predict y in the straight line model

3

Sampling Error for the estimator of the Mean of y and the Predictor of an Individual y for $x = x_p$

1. The Standard deviation of the sampling distribution of the estimator $\hat{y}$ of the mean value of y at a particular value of x, say x_p is:

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

$$\sigma_{\hat{y}} = \sigma \sqrt{\frac{1}{n} + \frac{(x_p - \bar{x})^2}{\Sigma(x_i - \bar{x})^2}}$$

where σ is the standard deviation of the random error ε.
we refer to $\sigma_{\hat{y}}$ as standard error of $\hat{y}$.

2. The Standard deviation of the prediction error for the predictor $\hat{y}$ of an individual y-value for $x = x_p$ is:

$$\sigma_{(y-\hat{y})} = \sigma \sqrt{1 + \frac{1}{n} + \frac{(x_p - \bar{x})^2}{\Sigma(x_i - \bar{x})^2}}$$

where σ is the Standard deviation of the random error ε. we refer to $\sigma_{(y-\hat{y})}$ as the standard error of prediction

The true value of σ will rarely be known. We estimate σ by s and calculate the estimation and prediction interval as shown

A $100(1-\alpha)\%$ Confidence Interval for the Mean Value of y for $x = x_p$.

$$\hat{y} \pm t_{(n-2,\ 1-\frac{1}{2}\alpha)} \text{(estimated standard deviation of } \hat{y})$$

or

$$\hat{y} \pm t_{(n-2,\ 1-\frac{1}{2}\alpha)} \cdot s \cdot \sqrt{\frac{1}{n} + \frac{(x_p - \bar{x})^2}{\Sigma(x_i - \bar{x})^2}}$$

Where $t_{(n-2,\ 1-\frac{1}{2}\alpha)}$ has $(n-2)$ d.f.

4

A $100(1-\alpha)\%$ Prediction Interval for an Individual y for $x = x_p$

$\hat{y} \pm t_{(n-2, 1-\frac{1}{2}\alpha)}$ [estimated Standard deviation of $(y - \hat{y})$]

or

$$\hat{y} \pm t_{(n-2, 1-\frac{1}{2}\alpha)} \cdot s \cdot \sqrt{1 + \frac{1}{n} + \frac{(x_p - \bar{x})^2}{\sum (x_i - \bar{x})^2}}$$

1^{st} uses model for estimating the mean value of y $E(y)$ for a Specific value of x.

- Estimating mean result of a very large number of experiments at given x value

Example, we want to estimate the mean sales revenue for All months during which $x=4$ is spent on advertising

- 2^{nd} model entails predicting a particular y value for a given x. That is if we decide to spend $x=4$ a month, we want to predict the firm's sale revenue for that month
- Predicting outcome of single experiment given x-value

CI is narrow why?

- The error in estimating the mean value of y, $E[y]$, for a given x say x_p, is the distance between the least squares line and the true line of means, $E(y) = \beta_0 + \beta_1 x$
- Error shown in Fig 4 $[\hat{y} - E(y)]$

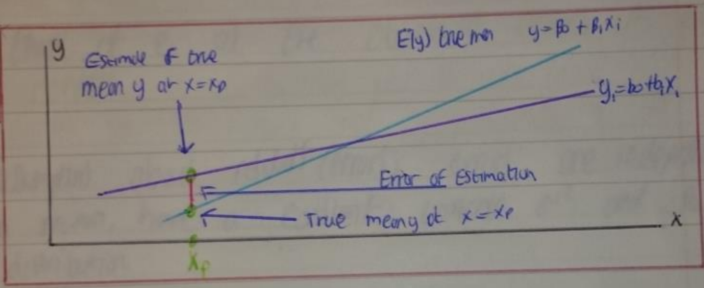
- In contrast, the error $(y_p - \hat{y})$ in predicting some future value of y is the sum of two errors - the error of estimating the mean of y, $E[y]$ plus the random error that is a component of the value of y to be predicted

- Consequently the error of predicting a particular value of y will always be larger than the error of estimating the mean value of y for a particular value of x.

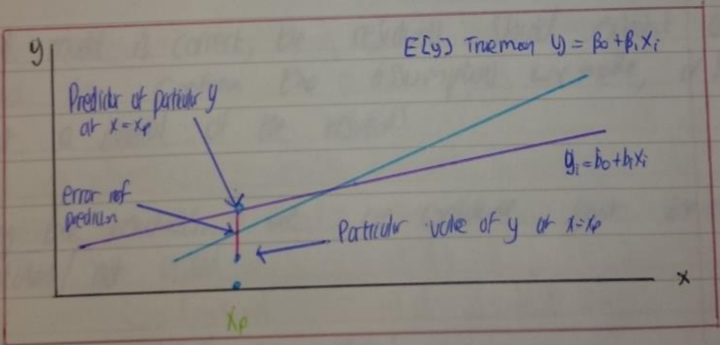
The further x (x_p) lies from $\bar{x}$, the larger will be the error of estimation and prediction.

- A graph showing both the confidence limits for $E(y)$ and the prediction limits for y over the entire range of the advertising x values is shown below:
- You see that CI is always narrower than PI, and they are both narrowest at the mean $\bar{x}$.

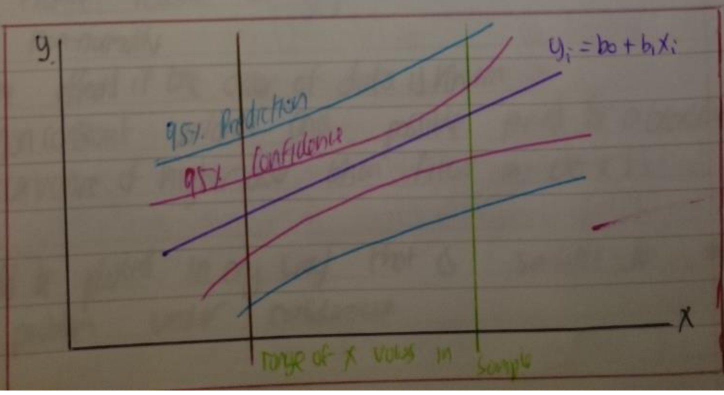
Error of Estimating the mean value of Y for a Given Value of X.



Error of Predicting a future single value of y for a Given value of X.



Comparison of Widths of 95% Confidence and Prediction Intervals.



Residual Analysis

The residuals are defined as the n differences $e_i = Y_i - \hat{Y}_i$, $i = 1, 2 \dots n$ where Y_i is an observation and $\hat{Y}_i$ is the corresponding fitted value obtained by use of the fitted regression eqⁿ

We can see from their definition that the residuals e_i are the differences between what is actually observed and what is predicted by the regression equation, - that is the amount that the regression equation had not been able to explain

We can thus think of e_i as the observed errors if the model is correct.

We make assumptions about residuals (errors); errors are independent, have zero mean, have a constant variance σ^2 and follow a normal distribution

If our fitted model is correct, the residuals should exhibit tendencies that tend to confirm the assumptions we made, or should not exhibit a denial of the residuals

After examining the residuals, we can conclude that the assumptions are violated / not violated

The ways to examine residuals are graphic:

1. To check for non-normality
2. Check for time effects if the order of data is known
2. Check for non constant variance and possible need for a transformation of Y
3. Check for curvature of higher order than fitted in the X's

5. Residuals should be plotted in any way that is sensible for the particular problem under consideration

Basic plot should always be done and will often pick up deficiencies present in many sets of residuals

Non-Normality Check on Residuals.

We usually assume that $\varepsilon_i \sim N(0, \sigma^2)$ and that all errors are independent of one another.

Then estimated, the residuals cannot be independent.

The estimation of parameters (p of them; say $p=2$ for straight line) means that the n residuals carry only $(n-p)$ df.

The p normal eqn' are restrictions on the e_i essentially. Unless p is large compared with n, this typically has little effect on our non normality check.

We first note that:

For any model with a β_0 (intercept) term in it, the least squared residuals must in theory, add to zero.

This is seen from the first normal eq^n obtained by differentiating the error sum of squares with respect to β_0. If the model fitted is $E[y] = \beta_0 + \beta_1 x_1 + \dots \beta_k x_k$, the eq^n can be written

$$-2\,\Sigma(y_i - b_0 - b_1 x_{1i} - \dots b_k x_{ki}) = 0$$

This reduces to $\Sigma(y_i - \hat{y}_i) = 0$

Thus $\Sigma e_i = 0$

Because the least squared fitting procedure guarantees this, there is no need to check that the mean $\bar{e} = \frac{\Sigma e_i}{n}$ is zero, we have made it so!

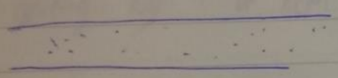
09 03

Checks for time effects, non constant variance, need for transformation and curvature

We plot the residuals e_i vertically against, in turn:

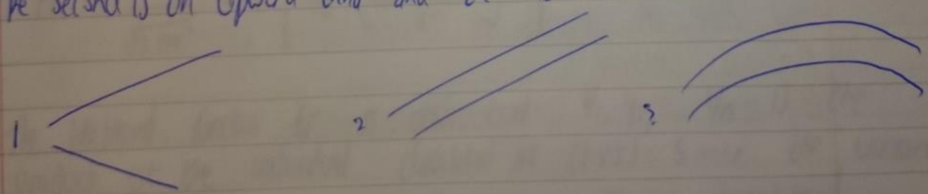
1. The time order of the data if known
2. The corresponding fitted values $\hat{Y}_i$ using the fitted model
3. The corresponding X_i values if there is only one predictor variable; or in general, each set of X_{ji}, where $J = 1, 2 \ldots k$ represent the X's in regression

In all these cases a satisfactory plot is one that shows a (more or less) horizontal band of points giving the impression:



There are many possible unsatisfactory plots. The funnel displays the band of residuals widening to the right showing non constant variance

The second is an upward trend and the third is curvature



Why do we plot the residuals $e_i = Y_i - \hat{Y}_i$ against $\hat{Y}_i$ and not against Y_i, for the usual linear model?

Because the e's and the Y's are usually correlated but the e's and the $\hat{Y}$'s are not.

Think of plots of the e_i as ordinate against (i) the Y_i

and (ii) against $\hat{Y}_i$ and find the slope of a least squares lines through the points

For (i) it will be $1-R^2$, for (ii) it will be 0.

This means that, unless $R^2=1$, there will always be a slope of $1-R^2$ in the e_i versus Y_i plot even if there is nothing wrong.

However a slope in the e_i versus $\hat{Y}_i$ plot indicates that something is wrong

Method of estimating the parameters of a statistical model.

Assuming x's are normally distributed with unknown mean and variance, the mean and variance can be estimated with MLE while only knowing x's of some sample or overall population.

Maximises the "agreement" of the selected model with the observed data, for discrete random variables it maximises the probability of the observed data under the resulting distribution

Maximise by getting log of both sides, makes computation easier.

In general, the density of an observation Y_i for the normal error regression model is as follows, utilizing the fact that $E[Y_i] = \bar{y} = \beta_0 + \beta_1 X_i$ and $\sigma[Y_i] = \sigma^2$

$$f_i = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{1}{2}\left(\frac{Y_i - \beta_0 - \beta_1 X_i}{\sigma}\right)^2\right] \quad 1.25$$

The likelihood function for n observations $Y_1, Y_2, \dots Y_n$ is the product of the individual densities in (1.25). Since the variance σ^2 of the error terms is usually unknown, the likelihood function is a function of three parameters: $\beta_0, \beta_1, \sigma^2$:

$$L = \prod_{i=1}^{n} \frac{1}{(2\pi\sigma^2)^{1/2}} \exp\left[-\frac{1}{2\sigma^2}(Y_i - \beta_0 - \beta_1 X_i)^2\right]$$

$$= \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(Y_i - \beta_0 - \beta_1 X_i)^2\right] \quad 1.26$$

2

The values of β_0, β_1 and σ^2 that maximise the likelihood function are the maximum likelihood estimates and are denoted by $\hat{\beta}_0$, $\hat{\beta}_1$, $\hat{\sigma}^2$. They are as follows:

Parameter	Maximum Likelihood estimator
β_0	$\hat{\beta}_0 = b_0$ Same as SLS
β_1	$\hat{\beta}_1 = b_1$ Same as SLS
σ^2	$\hat{\sigma}^2 = \frac{\sum(Y_i - \hat{Y})^2}{n}$

Maximum likelihood estimators of β_0 and β_1 are the same estimators as those provided by method of least squares

The MLE $\hat{\sigma}^2$ is biased, and ordinarily the unbiased MSE as given is used

Note that the unbiased estimator MSE or s^2 differs slightly from MLE $\hat{\sigma}^2$, especially if n is not small:

$$s^2 = MSE = \frac{n}{n-2}\hat{\sigma}^2$$

Comments

Since MLE $\hat{\beta}_0$ and $\hat{\beta}_1$ are the same as least squares estimators, b_0 and b_1 they have the properties of all least square estimators:

1. They are unbiased
2. They have minimum variance among all unbiased linear estimators
They are consistent
They are sufficient.
They are variance unbiased; they have minimum variance in the class of all unbiased estimators (linear or otherwise)

We find the values of β_0, β_1 and σ^2 that maximise the likelihood function L in (1.26) by taking partial derivatives of L with respect to β_0, β_1 and σ^2 equating them to 0.

We work with $\log_e L$ rather than L, because both L and $\log_e L$ are maximised for the same values of β_0, β_1 and σ^2?

$$\text{Log } L = \text{Log}\left[\frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left[\frac{-1}{2\sigma^2}\Sigma(y_i-\beta_0-\beta_1x_i)^2\right]\right] \qquad \text{Log}(AB) = \text{Log } A + \text{Log } B$$

$$= \text{Log}\left(\frac{1}{(2\pi\sigma^2)^{n/2}}\right) + \log\left(\exp\left[\frac{-1}{2\sigma^2}\Sigma(y_i-\beta_0-\beta_1x_i)^2\right]\right) \qquad \log(\exp) = 1$$

$$\log\frac{A}{B} = \log A - \log B$$

$$\text{Log}(1) - \log(2\pi\sigma^2)^{n/2} + \left(\frac{-1}{2\sigma^2}\Sigma(y_i-\beta_0-\beta_1x_i)^2\right) \qquad \log a^n = n\log a$$

$$-\frac{n}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\Sigma(y_i-\beta_0-\beta_1x_i)^2 \qquad \log AB = \log A + \log B$$

$$= -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log(\sigma^2) - \frac{1}{2\sigma^2}\Sigma(y_i-\beta_0-\beta_1x_i)^2$$

Partial differentiation of the log of the likelihood function is much easier:

$$\frac{d(\log_e L)}{d\beta_0} = \frac{-1}{2\sigma^2}\Sigma(y_i-\beta_0-\beta_1x_i)(2)(-1) = 0$$

$$\frac{+2}{2\sigma^2}\Sigma(y_i-\beta_0-\beta_1x_i) = 0$$

$$\Sigma(y_i-\beta_0-\beta_1x_i) = 0 \qquad \text{Same as least squares}$$

$$\frac{d(\log_e L)}{d\beta_1} = \frac{-1}{2\sigma^2}\Sigma(y_i-\beta_0-\beta_1x_i)(2)(-x_i) = 0$$

$$= \frac{2}{2\sigma^2}\Sigma(y_i-\beta_0-\beta_1x_i)(x_i) = 0$$

$$\Sigma x_i(y_i-\beta_0-\beta_1x_i) = 0 \qquad \text{Same as least squares}$$

$$\frac{d(\log_e L)}{d\sigma^2} \quad \frac{-n}{2\sigma^2} + \frac{1}{2\sigma^4}\Sigma(y_i-\beta_0-\beta_1x_i)^2 = 0$$

$$\frac{+1}{2\sigma^4}\Sigma(y_i-\beta_0-\beta_1x_i)^2 = \frac{n}{2\sigma^2}$$

$$+\Sigma(y_i-\beta_0-\beta_1x_i)^2 = n\sigma^2$$

$$\sigma^2 = \frac{\Sigma(y_i-\beta_0-\beta_1x_i)^2}{n}$$

1

Method of Maximum likelihood

$y_1 \dots y_n \sim N(\mu, \sigma^2)$ independent

Joint distribution of $y_1 \dots y_n$

$p(y_1 \dots y_n \mid \mu, \sigma^2) =$

product

$$\prod_{i=1} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(y_i-\mu)^2}$$

Likelihood function $L(\mu, \sigma^2 \mid y_1 \dots y_n) = \prod_{i=n} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(y_i-\mu)^2}$

product becomes sum

$$\text{Log}\left[L(\mu, \sigma^2 \mid y_1 \dots y_n)\right] = \sum_{i=1} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[\frac{-1}{2\sigma^2}(y_i-\mu)^2\right]$$

Some maths....

$$-\frac{n}{2}\log(2\pi) - \frac{n}{2}\log(\sigma^2) - \frac{1}{2\sigma^2}\sum(y_i-\mu)^2$$ log and cancel out

$$\frac{-1}{\sqrt{2\pi\sigma^2}} = \log(2\pi)$$

maximising $L(\mu, \sigma^2 \mid y_1 \dots y_n)$ is equivalent to minimising $\text{Log}(L(\mu, \sigma^2 \mid y_1 \dots y_n))$

① trying to compute $\frac{d}{d\mu} L(\mu, \sigma^2 \mid y_1 \dots y_n) = \frac{+2}{2\sigma^2} \sum(y_i-\mu) = 0$

② and $\frac{d}{d\sigma^2} L(\mu, \sigma^2 \mid y_1 \dots y_n) = \frac{-n}{2\sigma^2} \frac{-1}{2\sigma^4} \sum(y_i-\mu)^2 = 0$

If you solve $\mu = \bar{y}$

$$\sigma^2 = \frac{\sum(y_i-\bar{y})^2}{n}$$

2

Only One Difference

Here $y_i \sim N(\mu, \sigma^2)$ — 2 unknowns

In SLR $y_i \sim N(\beta_0 + \beta_1 x_i, \sigma^2)$ ← 3 unknowns

estimate β_0, β_1 and σ^2 using maximum likelihood

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

$\sum (y_i - \hat{y}_i)(\hat{y}_i - \bar{y})$ ← break up

$\sum_i \hat{y}_i (y_i - \hat{y}_i) - \sum_i \bar{y} (y_i - \hat{y}_i)$

$\sum \hat{y}_i (y_i - \hat{y}_i) - \bar{y} \sum (y_i - \hat{y}_i)$ $\quad$ $\hat{y}_i = b_0 + b_1 x_i$

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

using (1). $\sum (\bar{y} + b_1(x_i - \bar{x}))(y_i - \hat{y}_i) \underbrace{- \bar{y} \sum (y_i - \hat{y}_i)}_{\text{cancel 2}}$ $\quad$ $\hat{y}_i = \bar{y} - b_1\bar{x} + b_1 x_i$ (1).

$\sum \underline{b_1 (x_i - \bar{x})}(y_i - \hat{y}_i)$

$= b_1 \sum (x_i - \bar{x})(y_i - \hat{y}_i)$

$= b_1 \sum (x_i)(y_i - \hat{y}_i) - \bar{x} \sum (y_i - \hat{y}_i)$ ← 0

$\sum \hat{y}_i = \sum (\bar{y} + b_1(x_i - \bar{x}))$

$= \sum \bar{y} + b_1 \underbrace{\sum (x_i - \bar{x})}_{=0}$

$= n\bar{y}$

$= \sum y_i$

$\sum x_i (y_i - \hat{y})$

$= \sum x_i \left[y_i - (\bar{y} + b_1(x_i - \bar{x})) \right]$

$= \sum x_i \left[(y_i - \bar{y}) - b_1 (x_i - \bar{x}) \right]$

$\sum x_i (y_i - \bar{y}) - b_1 \sum (x_i - \bar{x})$

rule $\quad = \sum (x_i - \bar{x})(y_i - \bar{y})$

$\sum (x_i - \bar{x})(y_i - \bar{y}) - b_1 \sum (x_i - \bar{x})^2$

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

cancel and $= 0$.

$$f(y_i) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{y_i - \beta_1 - \beta_2 x_i}{\sigma}\right)^2}$$

$$\text{Probability} = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{y_1 - \beta_1 - \beta_2 x_1}{\sigma}\right)^2} \times \ldots \times \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{y_n - \beta_1 - \beta_2 x_n}{\sigma}\right)^2}$$

$$L(\beta_1, \beta_2, \sigma \mid y_1 \ldots y_n) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{y_1 - \beta_1 - \beta_2 x}{\sigma}\right)^2} \times \ldots \times \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{y_1 - \beta_1 - \beta_2 x}{\sigma}\right)^2}$$

$$\log L = \log(\quad \text{''} \quad \text{''} \quad)$$

Sum $\log L = \log\left(\frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{y_1 - \beta_1 - \beta_2 x}{\sigma}\right)^2}\right) + \ldots \log\left(\frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{y_n - \beta_1 - \beta_2 x_n}{\sigma}\right)^2}\right)$

$$= n\log\left(\frac{1}{\sigma\sqrt{2\pi}}\right) - \frac{1}{2}\left(\frac{y_1 - \beta_1 - \beta_2 x_1}{\sigma}\right)^2 - \ldots \frac{1}{2}\left(\frac{y_n - \beta_1 - \beta_2 x_n}{\sigma}\right)^2$$

$$= n\log\left(\frac{1}{\sigma\sqrt{2\pi}}\right) - \frac{\sigma^{-2}}{2} Z \quad \text{where } Z = \sum (y_i - \beta_1 - \beta_2 x_i)^2$$

Z is $\sum (y_i - \hat{y}_i)^2$

$= \sum e_i^2$ where $e_i = y_i - b_1 - b_2 x_i$

$$n\log\left(\frac{1}{\sigma}\right) + n\log\left(\frac{1}{\sqrt{2\pi}}\right) - \frac{\sigma^{-2}}{2} Z$$

$$= -n\log\sigma + n\log\left(\frac{1}{\sqrt{2\pi}}\right) - \frac{\sigma^{-2}}{2} Z$$

$$\frac{d \log L}{d\sigma} = -\frac{n}{\sigma} + \sigma^{-3} Z = \sigma^{-3}(Z - n\sigma^2) = 0.$$

$$-n\sigma^2 = -Z$$

$$n\sigma^2 = Z$$

$$\sigma^2 = \frac{\sum e_i^2}{N}$$

This is biased finite samples. To obtain unbiased estimates, we should divide by $n-k$, where k is number of parameters in this case 2, bias disappears with larger n.

$$-n\log\sigma + n\log\left(\frac{1}{\sqrt{2\pi}}\right) - \frac{1}{2\sigma}\sum(y_i - b_0 - b_1x_i)^2$$

$$\frac{d\log L}{db_0} = \frac{+2\sum(y_i - b_0 - b_1x_i)}{2\sigma} = 0 \qquad y_i - b_0 - b_1x_i = 0$$

$$b_0 = y_i - b_1x_i$$

$$\frac{d\log L}{db_1} = \frac{-1}{2\sigma}(-x_i)\ \sum(y_i - b_0 - b_1x_i)$$

$$\frac{-1}{2\sigma}2\sum(y_i - b_0 - b_1x_i) - \quad = \frac{1}{\sigma}\sum(y_i - b_0 - b_1x_i) = 0$$

$$\sum e_i = 0$$

$$\frac{\pm 1}{2\sigma}\sum(y_i - b_0 - b_1x_i)x_i = 0$$

$$\sum e_ix_i = 0.$$

1

$r = \dfrac{S_{xy}}{S_x S_y} \quad \dfrac{\Sigma(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\Sigma(x_i-\bar{x})^2}\sqrt{\Sigma(y_i-\bar{y})^2}}$

$r = \dfrac{S_{xy}}{\sqrt{S_{xx}}\sqrt{S_{yy}}}$

Measures relationship ~~not~~ slope

Assumptions

1. x_i is the i^{th} value of the predictor variable, which is a known constant for all i.
2. The observations y_i (or ε_i) are independent.
3. At any given x_i, y_i (or ε_i) is normally distributed
4. The observations y_i (or ε_i) have constant deviation σ^2.
5. The means of y_i can be joined by a straight line given as $E(y_i) = \beta_0 + \beta_1 x_i$ where β_0 and β_1 are unknown parameters

Such that: β_1 is the slope of the regression line and indicates the change in the mean of y for unit increase in x

β_0 is the intercept of the regression model. β_0 gives the mean distribution of y at $x=0$.

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

$S_{yy} = \Sigma(y_i - \bar{y})^2 = \Sigma y_i^2 - \dfrac{(\Sigma y_i)^2}{n}$

$b_1 = \dfrac{S_{xy}}{S_{xx}} \qquad r = b_1 \times \dfrac{S_x}{S_y}$

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

$S_{xy} = \Sigma(y_i - \bar{y})^2(x_i - \bar{x})^2 = \Sigma x_i y_i - \dfrac{\Sigma x_i \Sigma y_i}{n}$

Model states that the estimated mean value of y (response) at a given x is $\hat{y}$, also known as mean response

It does NOT predict an individual y of a given x

$\hat{y}_i = b_0 + b_1 x_i$ for $i = 1, \dots n$

$\hat{y}_i$ known as the fitted value for i^{th} observation
y_i is the observed value of y

We have to differentiate between the model error term $\epsilon_i = y_i - E(y_i)$ which is an unknown quantity; whereas the residual $e_i = y_i - \hat{y}_i$ is known since both the terms in the right hand side is known

For our model $y_i \sim N(\beta_0 + \beta_1 x_i, \sigma^2)$
$\beta_0 + \beta_1 x_i$ is mean μ.

Usually $s^2(\sigma^2) = \dfrac{\sum_1^n (y_i - \bar{y})^2}{n-1}$

but we estimate β_0 and β_1 and hence lose 2 d.f.
Remember $y_i - \hat{y}_i = e_i$ (the residual)

Sum of Squares $SSE = \sum_1^n (y_i - \hat{y}_i)^2$
$= \sum_1^n (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2$
$= \sum_1^n e_i^2$
known as Error Sum of Squares or Residual Sum of Squares.

Thus the estimate of σ^2 also denoted by MSE - mean squared error.
$\hat{\sigma}^2 = MSE = \dfrac{SSE}{n-2} = \dfrac{\sum (y_i - \hat{y}_i)^2}{n-2}$

$\hat{\sigma}^2$ measures the variation of the data around the regression line

It can be shown $E(MSE) = \sigma^2$ true stan variance

3

Recall the sampling distribution of $\bar{x}$ is:

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

Sampling distribution of b_1

$E(b_1) = \beta_1$

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

It can be shown b_1 follows normal distribution, thus sampling dist. of b_1 is:

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

However, if σ^2 is unknown, it would indicate that $Var(b_1)$ is also unknown.
Hence we need to estimate the variance of slope, turns out to be:

$\hat{Var}(b_1) = S^2_{b_1} = \frac{MSE}{S_{xx}}$

Also note that standard error (se) is square root of estimated variance:

$s.e(b_1) = \sqrt{\frac{MSE}{S_{xx}}}$

Hypothesis testing for β_1

t-test: $t_{calc} = \frac{(estimator) - (value\ from\ H_0)}{se(estimator)}$

To construct a confidence interval: estimator $\pm$ $t_{critical}$ se(estimator)

Normally testing to investigate whether any relationship exists between Y and X, we would test whether the slope of line = 0

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

Slope coef of 0 implies that there is no association between X and Y
When $\beta_1 = 0$ we have $E(y) = \beta_0$

If $|t_{calc}| \leq t_{critical}$ accept H_0 $H_0: \beta_1 = 0$ ✓

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

2

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

or

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

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

$$se(b_1) = \sqrt{\frac{MSE}{S_{xx}}}$$

ANOVA.

Total Sum of Squares (Corrected Sum of Squares)

$$SSTO = \Sigma(y_i - \bar{y})^2$$

Note: total uncorrected sum of squares $= \Sigma y_i^2$

If SSTO=0, all the responses are equal to some single value.
The greater the SSTO, the bigger the variation between the responses

Partition SSTO into
- Variability due to model (variability explained by regression eqn).
- Variability caused due to error/chance (unexplained or residual variability)

We already calculated Error Sum of Squares SSE the variability of the data around the fitted line

$$SSE = \Sigma(y_i - \hat{y})^2$$

If SSE = 0, all the data lies on the fitted line. Larger the SSE, greater is the variation of the observations around the fitted regression line

The other part of variability is Regression sum of squares SSR

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

SSR can be considered as a measure of the variability of the data that is associated with the r. line

Big distinction between $y' = E(y|x')$ in the previous section and prediction of a new response y_{new}.

In first case we refer to the mean of the distribution of y, for a particular value of x, whereas in the latter, we predict an individual outcome drawn from the distribution of y for a given value of x. In latter case we have to account for greater variability.

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

$$E[\hat{y}_{new}] = y'_{new}$$

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

Note that the variance of a new y has two components:

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

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

The estimate of the variance of the predicted value is:

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

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

The $100(1-\alpha)\%$ Prediction Intervals are given by:

$$\hat{y}'_{new} \pm t_{critical}\, se(\hat{y}'_{new})$$

$$\hat{y}'_{new} \pm t_{critical} \sqrt{MSE \; 1 + \frac{1}{n} + \frac{(x'-\bar{x})^2}{S_{xx}}}$$

Prediction interval is interested in estimating the interval within which the true price of a single diamond of weight $\_\_\_$

8

The larger the SSR in relation to SSTO the greater is the effect of the regression in accounting for the total variation in the observation. Provides us with a measure of how good a job is being done in fitting the straight line

Partitioning the total sum of squares

$y_i - \bar{y} \quad = \quad \hat{y}_i - \bar{y} \quad + \quad y_i - \hat{y}_i$

Total deviation — Dev of fitted r line around mean — Dev around fitted regression line

Mean Squares

A Sum of Squares divided by its degrees of freedom is defined as mean square (MS.) Mean Squared error or MSE is

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

Regression mean square or MSR = $MSR = \frac{SSR}{1}$

Prove

SST = SSE + SSR

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

$\Sigma(y_i-\bar{y})^2$ add $\hat{y}$ and $-\hat{y}$

$\Sigma\left((y_i-\hat{y}) + (\hat{y}-\bar{y})\right)^2$

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

need $2\Sigma(y_i-\hat{y})(\hat{y}-\bar{y})$ to dissapear

$\Sigma\hat{y}(y_i-\hat{y}) - \Sigma\bar{y}(y_i-\hat{y})$

$\Sigma\hat{y}_i(y_i-\hat{y}) - \bar{y}\Sigma(y_i-\hat{y})$

$\Sigma(\bar{y}+b_1(x_i-\bar{x}))(y_i-\hat{y}_i) - \bar{y}\Sigma(y_i-\hat{y}_i$

$\Sigma b_1(x_i-\bar{x})(y_i-\hat{y}_i)$

$b_1\Sigma(x_i-\bar{x})(y_i-\hat{y})$

$b_1\Sigma x_i(y_i-\hat{y}) - \Sigma\bar{x}(y_i-\hat{y})$

$\Sigma\bar{x}(y_i-y_i)$

$=0$

Aside

$\hat{y}_i = b_0 + b_1x_i$

~~$b_0 = \bar{y}_i - b_1\bar{x}$~~

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

$\Rightarrow \hat{y}_i = \bar{y} - b_1\bar{x} + b_1x_i$ (1)

$\Sigma\hat{y}_i = \Sigma(\bar{y}+b_1(x_i-\bar{x}))$

$= \Sigma\bar{y} + b_1\underbrace{\Sigma(x_i-\bar{x})}_{=0}$

$= n\bar{y}$

$= \Sigma y_i$

$b_1\Sigma x_i(y_i-\hat{y}_i)$

$\Sigma x_i[y_i - (\bar{y} + b_1(x_i(x_i-\bar{x}))]$

$\Sigma x_i[(y_i-\bar{y}) - b_1(x_i-\bar{x})]$

$\Sigma x_i(y_i-\bar{y}) - b_1\Sigma x_i-\bar{x})$

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

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

$\Sigma(x-\bar{x})(y_i-\bar{y}) - \dfrac{\Sigma(x_i-\bar{x})(y_i-\bar{y})\cancel{\Sigma(x_i-\bar{x})^2}}{\cancel{\Sigma(x_i-\bar{x})^2}}$

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

cancel $=0$

P value measures probability of observing a more extreme t-value (in either direction) than the one calculated

In other words, given that the null hypothesis is true, the p-value is the area in the tails further from the observed t-statistic (or negative of it)

If pvalue < α reject H_0

The t 95% CI ⇒ we expect long run of true slope to be between (..., ..) If interval does not contain 0, reject $H_0 : \beta_1 = 0$

Interpretation: CI means that if we are to calculate intervals in this way, from data sets sampled under similar conditions as the one we have, then 95% of those calculated intervals will contain the true value of the slope

This implies that we are 95% confident that the calculated interval contains the true slope

Inference about intercept β_0:
The point estimator for β_0 is b_0 and is given by

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

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

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

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

Standard error of intercept coeff is now:

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

When calculating intervals for a specific x' gives us $\hat{y}'$:

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

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

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

Values further away causes greater variability than values closer to the mean of predictor variable x. Thus the $s.e(\hat{y}_i)$ is

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

Hypothesis testing for $\hat{y}'$

The t-test corresponding to the mean response for a particular level of the predictor is the following:

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

Test statistic is: $t_{calc} = \frac{\hat{y}_i - y'}{se(\hat{y}')}$

Confidence Interval for $\hat{y}'$

$$\hat{y}' \pm t_{critical}\, se(\hat{y}')$$

Prediction of a new observation

The prediction of a new observation y corresponding to a given x.

Denote predictor variable as x' and observation for the response as y'_{new}.

Thus y'_{new} is a single value of the variable y, corresponding to $x = x'$.