Linear Regression

$Y \approx \beta_0 + \beta_1 X$

Regressing Y on (onto) X.

β_0 - intercept
β_1 - slope
} coefficients or parameters

find β_0, β_1 such that the mean square is lowest

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

RSS (residual sum of squares) $= e_1^2 + e_2^2 + \cdots + e_n^2$

$= (y_1 - \hat{\beta}_0 - \hat{\beta}_1 x_1)^2 + (y_2 - \hat{\beta}_0 - \hat{\beta}_1 x_2)^2 + \cdots (y_n - \hat{\beta}_0 - \hat{\beta}_1 x_n)^2$

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2}$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}.$$

Assessing the accuracy

Sample mean will not be same as population

but the average of different population means will be near to the sample mean.

Same applies to the unknown coefficients β_0 & β_1

average amount the estimate $\hat{\mu}$ differs from actual μ

$$\text{Var}(\hat{\mu}) = SE(\hat{\mu})^2 = \frac{\sigma^2}{n}$$

σ - std deviation

n - no of observations

$$SE(\hat{\beta}_0)^2 = \sigma^2 \left[\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \right]$$

$$SE(\hat{\beta}_1)^2 = \frac{\sigma^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2}$$

where $\sigma^2 = \text{Var}(\epsilon)$

when x_i s are more spread out, $SE(\hat{\beta}_1)$ is smaller
⇒ more leverage to estimate slope
$SE(\hat{\beta}_0) \approx SE(\hat{\mu})$ if $\bar{x}$ were zero.
($\hat{\beta}_0$ would be equal to $\bar{y}$)

estimate of σ is Residual Standard Error

$$RSE = \sqrt{\frac{RSS}{(n-2)}}$$

Standard errors can be used to compute confidence intervals.
with 95% confidence we can say the true values of β_0 & β_1 lie in

$$\hat{\beta}_0 \pm 2 \cdot SE(\hat{\beta}_0) \quad \& \quad \hat{\beta}_1 \pm 2 \cdot SE(\hat{\beta}_1) \quad \text{respectively}$$

Standard errors can also be used for hypothesis testing

H_0 : no relationship btw X & Y i.e $\beta_1 = 0$

H_A : there's a relationship btw X & Y. ie $\beta_1 \neq 0$

to test this we need to check if β_1 is far from zero.
this depends on the accuracy of β_1. ($SE(\beta_1)$)
[remember from t statistic that, larger the t value,
stronger the evidence against H_0
larger the sample size - greater than 30 - the bell curve becomes
normal distribution]

$$t = \frac{\hat{\beta}_1 - 0}{SE(\hat{\beta}_1)}$$

Also reject the H_0 if p value is small enough

if the coefficients of β_0 & β_1 are large with their SE, then
t statistic will also be large.
Hence $\beta_0 \neq 0$ & $\beta_1 \neq 0$ ⇒ H_0 is false

Assessing the accuracy of model

assess the model via 2 related quantities - RSE & R^2 statistic

~~RSE = the irreducible ϵ's std estimation~~

RSE is the estimation of std deviation of ϵ

$$RSE = \sqrt{\frac{1}{n-2} RSS} = \sqrt{\frac{1}{n-2} \sum_{i=1}^{n} (y_i - \hat{y})^2}$$

RSE measures the lack of fit

smaller the value → better fit

in the example of ads on TV & sales, RSE = 3.26

meaning the predicted value would be 3,260 units off on avg.

whether it is acceptable or not depends on the context.

mean value of sales over all markets = 14,000 units.

so % error = $\frac{3260}{14000}$ = 23%

RSE is measured in the units of Y. so its not clear as to what constitutes a good RSE

$$R^2 = \frac{TSS - RSS}{TSS} = 1 - \frac{RSS}{TSS} \quad \in [0,1]$$

$TSS = \sum (y_i - \bar{y})^2$ total sum of squares

TSS - total variance in the response Y before regression

hence TSS - RSS measures amount of variability in the response that is explained by regression

R^2 measures the proportion of variability in Y that's explained by X.

R^2 near 1 explains the variability is due to regression

R^2 near 0 could be because linear regression is wrong, or error variance σ^2 is high or both.

(it is as good as the mean. of that knowing x doesn't help predict y)

in certain problems in Physics where the data comes with a small residual error, R^2 will be close to 1.

but in biology, marketing etc residual errors due to unmeasured data is large. in this case R^2 may go below 0.1.

in simple linear regression $R^2 = r^2$

$r = Cor(X, Y)$ instead of R^2 in order to assess the fit of linear model

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

Multiple linear regression

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p + \epsilon$$

p - predictors

X_j - j^{th} predictor

β_j - association b/w variable & response

running independent simple linear regressions vs multiple linear regression

↓

predicts β while holding other variables fixed

gives less weightage for one of the predictors that's highly correlated with other variable.

1) Is there a relationship b/w responses and predictors?

$H_0 = \beta_1 = \beta_2 = \cdots = \beta_p = 0$

H_A = at least one β_j is non zero

F statistic, $f = \frac{(TSS - RSS)/p}{RSS/(n-p-1)}$

if the linear models assumptions are correct,

$$E(RSS/(n-p-1)) = \sigma^2$$

and if H_0 is true

$$E((TSS-RSS)/p) = \sigma^2$$

Hence when there is no relationship b/w responses & predictors, one would expect F statistic to be nearly 1.

If H_a is true, then $E((TSS-RSS)/p) > \sigma^2$, $F > 1$

When n is large, an F statistic just larger than 1 is enough
n is small, a larger F statistic is needed

Sometimes we want to test that a particular subset q coefficients are zero.

$$H_0 = \beta_{p-q+1} = \beta_{p-q+2} = \cdots = \beta_p = 0$$

in this case, we fit a model using all variables except last q.
Suppose R_0 is the residual sum of squares of that model, then

$$F = \frac{(RSS_0 - RSS)/q}{RSS/(n-p-1)}$$

the F statistic & p value will be same when we fit the model with individual variables. ($q=1$)

Given the individual p values, why check for overall F statistic?
because when the no. of predictors p is large, there will be atleast one predictor with small p value below 0.05 by chance.
Hence if we use individual t statistic and associated p value in order to decide if there is a relationship, we might be wrong.
However F statistic does not get affected by no. of predictors

Sometimes when $p > n$, we can't use F statistic or linear regressions

2) Deciding on important variables

p predictors. 2^p models.

we can choose the best one using Mallow's C_p, Akaike Information Criterion (AIC), Bayesian Information Criterion (BIC), adjusted R^2 etc

But 2^p is not feasible for large no. of p

efficient approaches -

a) Forward selection

begin with a null model. just intercept. no predictors.

then fit p linear models. find the least RSS. Add that predictor to the model. then fit p-1 models. Add the new variable to the model with least RSS.

Now model is of 2 predictors & intercept.

proceed till a stopping condition is met

a, b, c, d

M . M_a, M_b, M_c, M_d

least RSS → M_c

a, b, d

M_c . M_{ca}, M_{cb}, M_{cd}

least RSS → M_{cb}

M_{cb} .

...

b) Backward selection

start with all variables

remove predictor with largest value of RSS

now repeat for (p-1) predictors until stopping condition

it could be when all the predictors are below a threshold of p value

c) Mixed selection

start with null. do forward selection.

if at any point p value for a predictor is above the threshold, remove it. backward selection.

continue until stopping condition

adjusted $R^2 = 1 - \frac{RSS/(n-p-1)}{TSS/(n-1)}$

3) Model fit

R^2 & RSE are used to measure the model fit

R^2 near 1 explains a large portion of variance in the response variable

R^2 increases as more variables are added to the equation

if the addition of new variable doesn't change R^2 significantly, then its not suggested to add

RSE might increase with the addition of some variable.

$RSE = \sqrt{\frac{RSS}{n-p-1}}$

models with more variables can have higher RSE if the decrease in RSS is too small relative to the increase in p

plotting the graph helps further analysis:
and the synergy is indicated

4) Predictions

there is some model bias: assuming a linear model is an approximation of reality

Even if the model is correct, there is irreducible error: how much Y will vary from $\hat{Y}$?

prediction intervals are wider than confidence intervals because they incorporate both the error in estimation for $f(X)$, and the uncertainty as to how much an individual point will differ from the training sample (population regression plane)

ex: 95% confidence interval in TV sales - [10985, 11528]

95% prediction interval - [7930, 14580]

⇓

attributes to confidence intervals of the variability of city

Other considerations in the regression model

1) Qualitative predictors

choose some arbitrary values : 0, 1

then the model $y_i = \beta_0 + \beta_1 x_i + \epsilon_i = \begin{cases} \beta_0 + \beta_1 + \epsilon_i \\ \beta_0 + \epsilon_i \end{cases}$

$x_i = \begin{cases} 1 \\ 0 \end{cases}$

$\beta_0 \rightarrow$ avg of one category

$\beta_0 + \beta_1 \rightarrow$ avg of another category

$\beta_1 \rightarrow$ avg difference b/w the categories

instead of 0 & 1, we can choose $-1, +1$

the predictions will be same

but the interpretation of coefficients will be different

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i = \begin{cases} \beta_0 + \beta_1 + \epsilon_i \\ \beta_0 - \beta_1 + \epsilon_i \end{cases}$$

Now, $\beta_0 \rightarrow$ avg

$\beta_1 \rightarrow$ difference above & below the average

more than 2 levels

$\therefore y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \epsilon_i$

the level with no dummy variable is known as baseline

choice of ~~dummy~~ baseline variable is arbitrary. but the results would be same.

do a Ho test, find out F statistic test & p value. then see if any variables are insignificant

2) Extensions of linear model

the linear regression model makes several restrictive assumptions

2 major ones are additive and linear

a) removing the additive assumption

$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \epsilon$

assumes that for every unit change in X_1, it affects Y constantly, without considering X_2

this fails to address synergy effect

we can introduce a interaction term

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2 + \epsilon$$
$$= \beta_0 + (\beta_1 + \beta_3 X_2) X_1 + \beta_2 X_2 + \epsilon$$
$$= \beta_0 + \tilde{\beta}_1 X_1 + \beta_2 X_2 + \epsilon$$

Since $\tilde{\beta}_1$ is a function of X_2, the association b/w X_1 & Y is no longer a constant

it will also be clear that $\beta_3 \neq 0$ via the t test, F test & p values.

heirarchical principle states that if we include an interaction in a model, we should also include the main effects, even if the p values associated with their coefficients are not significant.

otherwise leaving them out alters the meaning of interaction

in a model if the intercepts are different, but have same slopes, then the avg effect is not dependent on the variable that's associated with slope coefficient (refers to the parallel slopes of student vs not student & balance)

b) non linear relationships

we can extend the linear model by including transformed version of predictors - polynomial regression

ex: $mpg = \beta_0 + \beta_1 \times horsepower + \beta_2 \times horsepower^2 + \epsilon$

in the residual plot (fitted values vs residuals), the residuals should be randomly around residual zero for different fitted values. a visible trend means the model is wrong.

Potential problems

i) non linearity of the data

if the data is non linear, then the linear regression is not a good idea

plot residual plot to identify non linearity

fitted values vs residual values. if any pattern is found, the linearity should be questioned

if any pattern is found / non linear associations exist, consider to use non linear transformations of predictors. such as $\log X$, $\sqrt{X}$, X^2...

ii) correlation of error terms

(i don't understand)

error term ϵ.

in linear regression with n samples, all the ϵ is assumed to be independent / random.

in case they're correlated, then we actually have less than n samples in the trained model.

so the confidence interval of 95% maybe actually less (60-70%.)

estimated SE < true SE

p value will be too small. falsely concluding that the variable is significant

error correlations are present in time series data

iii) non constant variance of error terms

assumption - $Var(\epsilon_i) = \sigma^2$ is constant

but the variances of errors may increase with the value of response.

funnel shape in residual plot.

possible solution - transform the response to $\log Y$, $\sqrt{Y}$ ...

such transformations result in greater shrinkage of larger responses.

leading to a reduction in heteroscedasticity

ex:

often the data behaves like $Y = f(X)\, x_n$

by logging $\Rightarrow$ $\log Y = \log(f(X)) + \log u$

and error $e' = \log Y - \log \hat{Y}$

weighted least squares give more influences to observations with smaller error variance

ex: avg of 2 points - noisy → less weightage
avg of 100 points - stable → more weightage

$Var(\epsilon_i) = \frac{\sigma^2}{n_i}$

$w_i \propto \frac{1}{\frac{\sigma^2}{n_i}} \propto n_i \quad \Rightarrow w_i = n_i$

iv) outliers

how big of an outlier to be ignored?

plot the studentized residuals - divide each ϵ_i by its estimated standard error.

if the studentized residuals are greater than 3 in absolute value, they are possible outliers.

outlier could be due to many reasons such as error in recording data we can remove that observation. but they might also be an important predictor criteria

v) high leverage points

some observations maybe far from the mean but have significant impact on the model. unlike outliers, they are necessary.

it may not be easily visualised.

leverage statistic $h_i = \frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2}$

with the increase in distance of x_i from $\bar{x}$, the leverage increases

h_i is always s/w $\frac{1}{n}$ & 1.

avg leverage for all observations = $\frac{p+1}{n}$

so if a given observation has leverage statistic greater than $\frac{p+1}{n}$, then we may suspect the point has high leverage

vi) collinearity

2 or more predictor variables are closely related to one another
it can be difficult to separate the individual effects of collinear variables in the response

even small data changes can alter the coefficients more
since the model gets confused as the result would be identical as the input parameters change in pairs (since correlated)

the standard error & p value increases for the collinear variables,
t-statistic reduces
means the H_0 can not be disproved ($\beta_j = 0$)

we may not be able to spot collinearity on correlation matrix.
multicollinearity - 3-4 variables are collineas together
(than just 2 in the matrix)

Variance inflation factor (VIF) =

$$VIF(\hat{\beta}_j) = \frac{1}{1 - R^2_{X_j | X_{-j}}}$$

R^2 is the regression of X onto all other predictors.
VIF - ratio of variance of $\hat{\beta}_j$ when fitting the full model divided by the variance of $\hat{\beta}_j$ if fit on its own

smallest possible value of VIF is 1.
VIF more than 5 or 10 → collinearity

$$VIF = \frac{Var(\hat{\beta}_j \mid \text{all predictors})}{Var(\hat{\beta}_j \mid X_j \text{ alone})}$$

estimating β_j alone → clean signal
β_j with correlated predictors → noisy signal

high R^2_j ⇒ X_j is redundant
⇒ more of the information is already known to the model

either drop one of the collinear variables
or combine them & use as a single variable

Marketing plan: Ads vs Sales example

1) Is there a relationship b/w sales & advertising budget?
→ fit a multiple regression model of sale onto TV, radio, newspaper.
test the $H_0 : \beta_{TV} = \beta_{radio} = \beta_{newspaper} = 0$
use F statistic & p value to determine the significance of coefficients

2) How strong is the relationship?
→ RSE estimates standard deviation of response from the regression eq.
ex: RSE = 1.69, mean = 14.022. % error = $\frac{1.69}{14.022}$ = 12%

R^2 measures the fraction of variability in the response that is explained by the predictors
$R^2 = 90\%$ → means 90% of the results are captured by the predictors. 10% due to randomness & unable to capture by the model

3) Which media are associated with sales?
→ multiple linear regression → find the p value associated with each predictor's t statistic
ones with lower p value (TV, radio) are related (to sales)

4) How large is the association b/w each medium & sales?
→ perform independent linear regression & compare the t statistic/p value
t stat = $\frac{\text{coeff}}{\text{standard error}}$
answers how many std deviation away from zero is this effect?

Confidence intervals: (0.043, 0.049) for TV
(0.172, 0.206) for radio
(−0.013, 0.011) for newspaper — includes zero, insignificant

is collinearity the reason for this confidence interval to be wide?
VIFs 1.005, 1.145, 1.145 → no collinearity

5) How accurately can we predict future sales?

individual response : $Y = f(X) + \epsilon$ - prediction interval

average response : $f(X)$ - confidence interval

R^2 & RSE tells the accuracy.

plot the graph for visualising the model

6) Is the relationship linear?

Residual plot [pattern - non linear
no pattern - linear

non linear → transform regression predictors → linear

7) Is there synergy among the advertising media?

Draw the model fit curve & see if there's synergy effect

use interaction term in the regression to accommodate non-additive relationship.

a small p value & significant increase in R^2 confirms the effect.

Comparison of linear regression with K-nearest neighbours

K-means regression

When the problem is non linear K-means performs better for lower numbers of p.

but as the p increases K-means performs bad.

curse of dimensionality

- given observation has no nearest neighbours
- as a general rule, parametric methods will tend to outperform non parametric approaches when there is a small no. of observations per predictor

$$\hat{f}(x_0) = \frac{1}{K} \sum_{x_i \in N_0} y_i$$

K training observations that are closest to x_0, represented by N_0

even when the dimensions are small, we might prefer regression, due to the interpretability.

classmate
Date
Page

Exercises

1) Describe the null hypothesis to which p values given in below table correspond. Explain what conclusions can be drawn based on the p values. Your explanation shall be phrased in terms of sales, TV, radio, newspaper, rather than in terms of coefficients of linear model.

	Coeff	Std Error	t-statistic	p-value
Intercept	2.939	0.3119	9.42	<0.0001
TV	0.046	0.0014	32.81	"
radio	0.189	0.0086	21.89	"
newspaper	−0.001	0.0059	−0.18	0.8599

→ H_0 for TV: in the presence of radio & newspaper ads, there is no relationship b/w TV & sales

H_0 for radio: in the presence of TV & newspaper ads, there is no relationship b/w radio & sales

H_0 for newspaper: in the presence of TV & radio ads, there is no relationship b/w newspaper & sales

H_0 for intercept: in the absence of TV, radio, newspaper, sales are zero.

$H_0 = \beta_0 = \beta_1 = \beta_2 = \beta_3 = 0$

$H_A = \beta_0 \neq \beta_1 \neq \beta_2 \neq \beta_3 \neq 0$ - there is some relationship b/w TV, radio, newspaper, sales are non zero in the absence of these 3.

p value for newspaper is large, hence we can't reject the H_0 for newspaper ∴ Newspaper & sales don't have a relationship.

2) Carefully explain the differences b/w KNN classifier & KNN regression models

→ both use the concept of finding k number of nearest neighbours. In case of KNN classifier, we choose the category that is most prevalent in the neighbourhood. If there's a tie, choose either or increase / decrease the k.

In case of KNN regression, we compute the average of the k observations in the neighbourhood and try to fit the model.

3) Suppose we have a dataset with 5 predictors & their fitted model coefficients as below for the starting salary (in 1000$) after graduation,

X_1 = GPA, $\hat{\beta}_1 = 20$

X_2 = IQ, $\hat{\beta}_2 = 0.07$

X_3 = Level (High school or college), $\hat{\beta}_3 = 35$ (1 for college, 0 for high school)

X_4 = GPA & IQ, $\hat{\beta}_4 = 0.01$

X_5 = GPA & Level, $\hat{\beta}_5 = -10$

$\hat{\beta}_0 = 50$

a) Which of the following are correct, why?

i) for a fixed value of IQ & GPA

i) High school graduates earn more on avg, than college students

ii) College graduates earn > high school

iii) high school graduates earn more on avg than college graduates provided that the GPA is high enough

iv) alta of iii)

→ Model,

$$\hat{salary} (Y) = \beta_0 + \beta_1 \times GPA + \beta_2 \times IQ + \beta_3 \times Level + \beta_4 \times (GPA \times IQ) + \beta_5 \times (GPA \times Level)$$

$$Y_{college} - Y_{High\ school} = \beta_3 + \beta_5 \times GPA \quad (Level = 1 \ - \ level = 0)$$

$$= 35 - 10\ GPA$$

If the GPA is more than 3.5, then high school graduates earn more than college graduates

iii) is correct

b) predict the salary of college graduate with IQ of 110 & GPA of 4.

→ Add to the formula: $Y = 50 + (20 \times 4) + (0.07 \times 110) + (35 \times 1) + 0.01 \times (4 \times 110) - 10 (4 \times 1)$

= 137.1

c) Since the coefficient for GPA, IQ interaction is small, the interaction effect is not significant / has very little effect. True or False? Justify

→ To test the hypothesis of $H_0: \beta_4 = 0$, we need to perform t-stat or F-stat test and find out p value. Without that we can't tell the interaction effect significance.
Further, the $\beta_4 \times IQ \times GPA$ will range from ~ 2 to 6. So that'll matter for the prediction.

4) A dataset with n=100 with sigle predictor & quantitative response. Then fit a linear regression model & cubic regression-

$$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \beta_3 X^3 + \epsilon.$$

a) Suppose the true relationship b/w X & Y is linear, $Y = \beta_0 + \beta_1 X + \epsilon$, which will be lower? RSS for linear or cubic regression? or is there not enough info? Justify

→ RSS of cubic will be less for training data. As cubic is more flexible, it will capture most of the observations.

b) for testing data?

→ RSS of linear regression will be lower. As the relationship b/w X & Y is linear. Cubic regression will have less bias, but more variance

c) If the relationship b/w X & Y is non linear, RSS of train data less for linear or cubic?

→ RSS of cubic will be less. As it will capture more non linear data. Cubic is more flexible.

d) for test data?
there's not enough info on how much is the non linearity.
if its slightly non linear, then RSS of linear would do better (lower) in the test data. otherwise RSS of cubic will be lower.

5) consider the model w/o intercept

$$\hat{y}_i = x_i \hat{\beta}$$

where $\hat{\beta} = \dfrac{\sum x_i y_i}{\sum x_i^2}$

show that we can write : $\hat{y}_i = \sum a_i y_i$

→ $\hat{\beta} = \dfrac{x_1 y_1 + x_2 y_2 + \cdots + x_n y_n}{x_1^2 + x_2^2 + \cdots + x_n^2}$

$$\hat{y}_i = x_i \frac{x_1 y_1 + x_2 y_2 + \cdots + x_n y_n}{x_1^2 + x_2^2 + \cdots + x_n^2} \quad \text{—①}$$

$$\hat{y}_i = \sum a_i y_i$$

$$= a_1 y_1 + a_2 y_2 + \cdots a_n y_n \quad \text{—②}$$

equating ① & ②

$$a_{i1} = x_i \times \frac{x_{i1}}{\sum_{j=1}^{n} x_j^2}$$

6) $$\hat{\beta}_1 = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2}$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

Prove that in simple linear regression, the least squares line always passes through the point $(\bar{x}, \bar{y})$

→ the least squares line is given by

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$$

from the given equation

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} \rightarrow \bar{y} = \hat{\beta}_0 + \hat{\beta}_1 \bar{x}$$

substituting

$$\hat{y} = (\bar{y} - \hat{\beta}_1 \bar{x}) + \hat{\beta}_1 x$$

we define the residuals $e_i = y_i - (\beta_0 + \beta_1 x_i)$;

least squares chooses β_0, β_1 to minimize

$$RSS = \sum_1^n e_i^2 = \sum_1^n (y_i - \beta_0 - \beta_1 x_i)^2$$

derivative wrt β_0,

$$\frac{\delta RSS}{\delta \beta_0} = -2 \sum y_i - \beta_0 - \beta_1 x_i$$

set to zero (minimum condition) $\quad \Sigma (y_i - \beta_0 - \beta_1 x_i) = 0$

rearrange

$$\Sigma y_i - n\beta_0 - \beta_1 \Sigma x_i = 0$$

% by n

$$\bar{y} - \beta_0 - \beta_1 \bar{x} = 0$$

Solve for β_0 : $\beta_0 = \bar{y} - \beta_1 \bar{x}$ → this is available in the simple linear regression itself

least square minimization : $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$

regression equation : $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$

evaluate at $x = \bar{x}$: $\hat{y} = (\bar{y} - \hat{\beta}_1 \bar{x}) + \hat{\beta}_1 \bar{x}$

$$\hat{y} = \bar{y}$$

Hence proved that the line passes through $(\bar{x}, \bar{y})$

7) Prove that $R^2 = r^2 \equiv Cor^2(X, Y)$ for simple linear regression of Y onto X.

for simplicity, assume $\bar{x} = \bar{y} = 0$

→ $$Cor(X, Y) = \frac{\Sigma (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\Sigma (x_i - \bar{x})^2} \sqrt{\Sigma (y_i - \bar{y})^2}}$$

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$$

classmate
Date
Page

$\hat{y}$ for $x = \bar{x} \Rightarrow \bar{y} = \beta_0 + \beta_1 \bar{x}$

$\beta_0 = \bar{y} - \beta_1 \bar{x}$

$= 0$

$\beta_1 = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2}$ (available info)

$= \frac{\sum x_i y_i}{\sum x_i^2}$

now, $R^2 = 1 - \frac{RSS}{TSS} = 1 - \frac{\sum (y_i - \hat{y})^2}{\sum (y_i - \bar{y})^2}$

$= 1 - \frac{\sum (y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i))^2}{\sum (y_i - \bar{y})^2}$ $(\hat{\beta}_0 = 0)$

$= 1 - \frac{\sum (y_i - \hat{\beta}_1 x_i)^2}{\sum y_i^2}$ $(\bar{y} = 0)$

$= 1 - \frac{\sum (y_i^2 + \beta_1 x_i^2 - 2 y_i \hat{\beta}_1 x_i)}{\sum y_i^2} = \frac{\sum y_i^2 - \sum (y_i^2 + \beta_1 x_i^2 - 2 y_i \hat{\beta}_1 x_i)}{\sum y_i^2}$

$= \frac{2 \hat{\beta}_1 \sum x_i y_i - \beta_1^2 \sum x^2}{\sum y_i^2}$

substitute for β_1

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

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