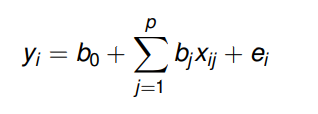
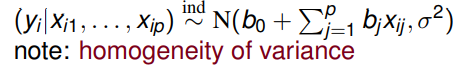
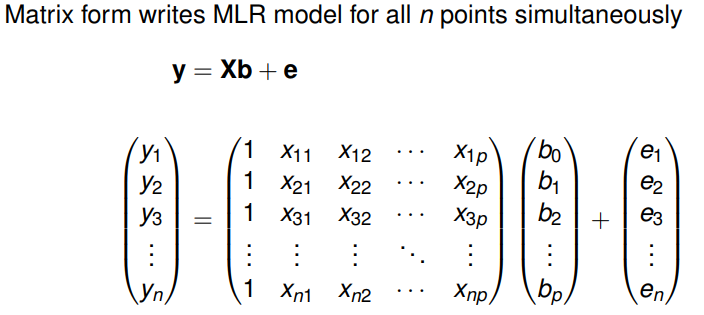
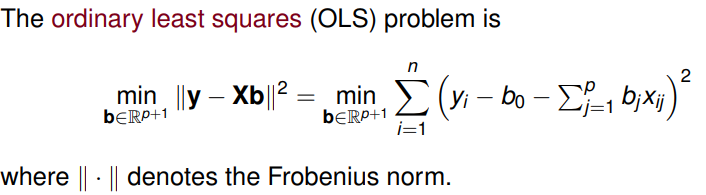
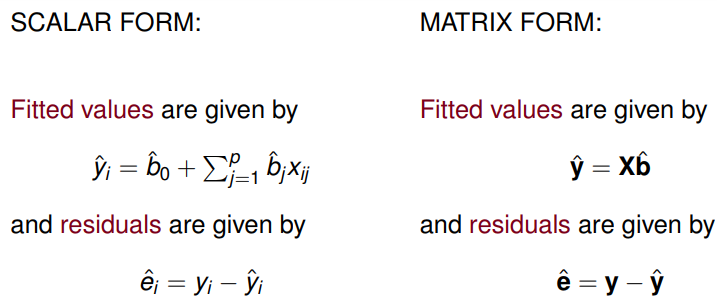
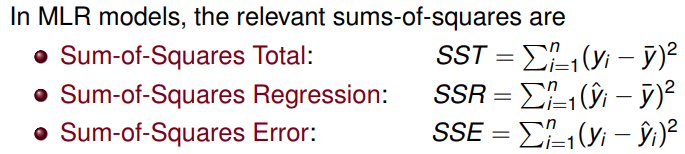
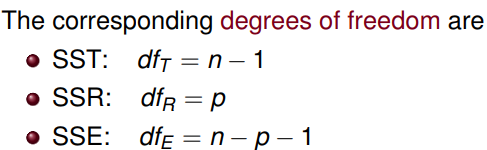
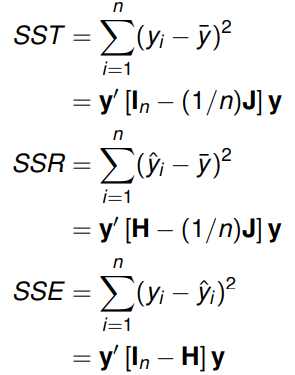
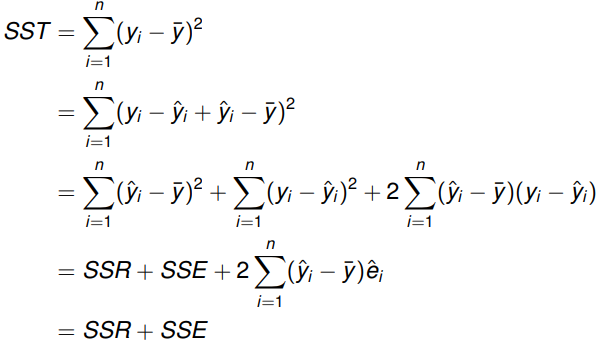
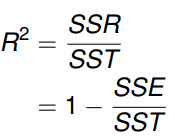
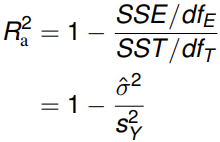
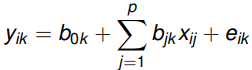
**Regression** is the statistical model that you use to predict a continuous outcome on the basis of one or more continuous predictor variables.

**ANOVA** is the statistical model that you use to predict a continuous outcome on the basis of one or more categorical predictor variables.

<https://ncss-wpengine.netdna-ssl.com/wp-content/themes/ncss/pdf/Procedures/NCSS/Discriminant_Analysis.pdf>

The main difference between these two techniques is that **regression analysis** deals with a continuous dependent variable, while **discriminant analysis** must have a discrete dependent variable.

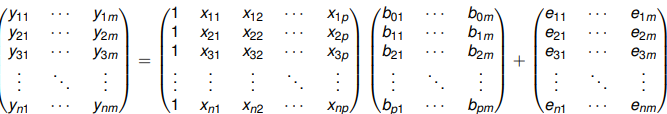
<https://www.stat-d.si/mz/mz1.1/pohar.pdf>  
  
Two of the most widely used statistical methods for analyzing categorical outcome variables are linear discriminant analysis and logistic regression. While both are appropriate for the development of linear classification models, linear discriminant analysis makes more assumptions about the underlying data. Hence, it is assumed that logistic regression is the more flexible and more robust method in case of violations of these assumptions.

**multiple vs multivariate linear regression**  
A **regression** analysis with one dependent variable **and** 8 independent variables is NOT a **multivariate regression**. It's a **multiple regression**. **Multivariate** analysis ALWAYS refers to the dependent variable. So when you're in SPSS, choose univariate GLM for this model, not **multivariate**.  
  
The multiple linear regression model has the form   
  
for i ∈ {1, . . . , n} where   
yi ∈ R is the real-valued response for the i-th observation   
b0 ∈ R is the regression intercept   
bj ∈ R is the j-th predictor’s regression slope   
xij ∈ R is the j-th predictor for the i-th observation   
ei iid∼ N(0, σ2 ) is a Gaussian error term  
The model is multiple because we have p > 1 predictors.   
If p = 1, we have a simple linear regression model  
The model is linear because yi is a linear function of the parameters (b0, b1, . . . , bp are the parameters). The model is a regression model because we are modeling a response variable (Y) as a function of predictor variables (X1, . . . , Xp).  
The fundamental assumptions of the MLR model are:   
1 Relationship between Xj and Y is linear (given other predictors)   
2 xij and yi are observed random variables (known constants)   
3 ei iid∼ N(0, σ2 ) is an unobserved random variable   
4 b0, b1, . . . , bp are unknown constants  
5   
Note: bj is expected increase in Y for 1-unit increase in Xj with all other predictor variables held constant  
The multiple linear regression model has the form  
where   
y = (y1, . . . , yn) 0 ∈ R^n is the n × 1 response vector   
X = [1n, x1, . . . , xp] ∈ R^n×(p+1) is the n × (p + 1) design matrix   
• 1n is an n × 1 vector of ones   
• xj = (x1j , . . . , xnj) 0 ∈ R^n is j-th predictor vector (n × 1)   
b = (b0, b1, . . . , bp) 0 ∈ R^p+1 is (p + 1) × 1 vector of coefficients   
e = (e1, . . . , en) 0 ∈ R^n is the n × 1 error vector  
  
  
  
  
The OLS solution has the form  
  
  
yˆ = Xbˆ = X(X 0X) −1X 0y = Hy  
where H = X(X 0X) −1X 0 is the hat matrix.   
H is a symmetric and idempotent matrix:   
HH = H H projects y onto the column space of X.  
  
  
  
   
  
  
The coefficient of multiple determination is defined as  
  
  
and gives the amount of variation in yi that is explained by the linear relationships with xi1...xip.  
When interpreting R 2 values, note that. . . 0 ≤ R 2 ≤ 1 Large R 2 values do not necessarily imply a good model  
  
Including more predictors in a MLR model can artificially inflate R 2 : Capitalizing on spurious effects present in noisy data Phenomenon of over-fitting the data  
  
The adjusted R 2 is a relative measure of fit:  
  
  
<http://users.stat.umn.edu/~helwig/notes/mvlr-Notes.pdf>  
  
The multivariate (multiple) linear regression model has the form  
  


for i ∈ {1, . . . , n} and k ∈ {1, . . . , m} where   
yik ∈ R is the k-th real-valued response for the i-th observation  
b0k ∈ R is the regression intercept for k-th response   
bjk ∈ R is the j-th predictor’s regression slope for k-th response   
xij ∈ R is the j-th predictor for the i-th observation   
(ei1, . . . , eim) iid∼ N(0m, Σ) is a multivariate Gaussian error vector

The model is multivariate because we have m > 1 response variables.  
The model is multiple because we have p > 1 predictors.   
If p = 1, we have a multivariate simple linear regression model  
The model is linear because yik is a linear function of the parameters (bjk are the parameters for j ∈ {1, . . . , p + 1} and k ∈ {1, . . . , m}).   
The model is a regression model because we are modeling response variables (Y1, . . . , Ym) as a function of predictor variables (X1, . . . , Xp).

The fundamental assumptions of the MLR model are:   
1 Relationship between Xj and Yk is linear (given other predictors)   
2 xij and yik are observed random variables (known constants)   
3 (ei1, . . . , eim) iid∼ N(0m, Σ) is an unobserved random vector   
4 bk = (b0k , b1k , . . . , bpk ) 0 for k ∈ {1, . . . , m} are unknown constants   
5 (yik |xi1, . . . , xip) ∼ N(b0k + Pp j=1 bjk xij, σkk ) for each k ∈ {1, . . . , m}   
note: homogeneity of variance for each response   
Note: bjk is expected increase in Yk for 1-unit increase in Xj with all other predictor variables held constant

The multivariate multiple linear regression model has the form  
Matrix form writes MLR model for all nm points simultaneously  


in the polynomial regression case, we try and find if there are higher-order relationships between X and Y, beyond the linear relationships. As a good thumb rule of model development, we explore higher order relationships to improve model fit if we have difficulty building linear models to describe the case.  
*Y*=*mX*+*C AND/ OR Y=m1X+m2X^2+C*  
In a multiple regression case, we’re interested in the impact of not only one, but many different factors, on the response variable. This is usually representative of real world problems more than the stock single-factor-vs-response model described above.   
*Y*=*m*1*X*1+*m*2*X*2+*C*  
Naturally, you can expect a more complex equation which has coefficients for squares and other higher powers of these multiple factors, in case of a multiple polynomial regression model.  
  
BOTH *Y=m1X1+m2X1^2* + *m*1*X*2+*m*2*X*2*^2*+*C*

<https://towardsdatascience.com/roc-curve-a-complete-introduction-2f2da2e0434c>

Support: (a+b)/total. Interpreted as: Fraction of transactions that contain both A and B.

Confidence: (a+b)/a. Interpreted as: How often items in B appear in transactions that contain A only.

Lift: Confidence / (b/total) Interpreted as: How much our confidence has increased that B will be purchased *given* that A was purchased.

https://www.statisticssolutions.com/correlation-pearson-kendall-spearman/

The most commonly used procedure used for regression analysis is called ordinary least

squares (OLS). The OLS procedure minimizes the sum of squared residuals.

The amount of variation to be explained by the regression is the \*\*total sum of squares (TSS)\*\*.

The \*\*explained sum of squares (ESS)\*\* represents the explained variation

The \*\*residual sum of squares (RSS)\*\* represents the unexplained variation

The ratio ESS / TSS = 1 - (RSS/TSS) is called the \*\*coefficient of determination and is denoted by ܴR^2\*\* and lies between 0 and 1.

ܴR^2 measures the percentage of variation of ܻY hataround ܻthe mean y that is explained by the regression

equation. The closer the observed points are to the estimated regression line, the better

the fit, the higher the ܴଶ.

The way we have defined ܴଶis problematic. The addition of any X variable, will never

decrease the ܴଶ. In fact, ܴଶis likely to increase. A different measure of goodness of fit is used, the \*\*adjusted R^2 (or R-bar squared)\*\*. This value has a maximum of 1 but a minimum that can be negative.

In addition to the ܴR^2, there is the \*\*simple correlation coefficient, r,\*\* which measures

strength and direction of a linear relationship between two variables. r lies between –1 and 1.

The Betas produced by OLS will differ when a different sample is used for the same model.

They are random variables with a sampling distribution.

mean > an estimator is unbiased if its expected value equals its true value, i.e. E(Betahat)==B

what we know so far is that the sampling distribution of the Bhats are normally

distributed with a mean of Beta

If Assumptions 1 – 6 are satisfied, OLS works well! Why? Because of the G-M Theorem.

This theorem states that when assumptions 1 – 6 are true, the OLS estimator of BsubJ is the

minimum variance estimator from among the set of all linear unbiased estimators of BsubJ

for all ݆ j = 0,1, … , ݇. That is, the OLS estimator is BLUE, the Best Linear Unbiased

Estimator.

Best: This just means that BhatsubJ has the smallest variance out of all linear unbiased

estimators of BsubJ This property is known as efficiency

OLS is the best procedure for estimating a linear regression model only under certain

assumptions.

The word classical refers to these assumptions that are required to hold.

Assumptions of the Classical Linear Regression Model:

1. The regression model is linear, correctly specified, and has an additive error term.

2. The error term has a zero population mean.

3. All explanatory variables are uncorrelated with the error term

4. Observations of the error term are uncorrelated with each other (no serial

correlation).

5. The error term has a constant variance (no heteroskedasticity).

6. No explanatory variable is a perfect linear function of any other explanatory

variables (no perfect multicollinearity).

7. The error term is normally distributed (not required).

Essentially, the rest of the course deals with what happens when one or more of these

assumptions do not hold and what we can do to remedy the situation.

If the variance of the distribution of the error term changes for each observation or range of observations then we have what’s called heteroskedasticity

Recall, we had the following six assumptions required for the Gauss-Markov Theorem:

1. The regression model is linear, correctly specified, and has an additive error term.

2. The error term has a zero population mean.

3. All explanatory variables are uncorrelated with the error term.

4. Observations of the error term are uncorrelated with each other (no serial

correlation).

5. The error term has a constant variance (no heteroskedasticity).

6. No explanatory variable is a perfect linear function of any other explanatory

variables (no perfect multicollinearity).

We have assumed these assumptions have been satisfied. For the rest of the course, we

will deal with violations of these assumptions

A specification error occurs when the model is misspecified in terms of the choice of

variables, functional form or error structure.

In choosing explanatory variables, two types of errors are likely:

1) Omitting a variable that belongs in the model

2) Including an irrelevant variable

Correcting for an Omitted Variable

Omitted variable bias is hard to detect:

1) invest time in thinking about the equation before you even look at the data

2) estimated coefficient has the wrong sign (and significant) or magnitude

Corrections:

1) Include the variable

2) Report the expected bias

Irrelevant Variables

What happens when a variable that does not belong in the model is included in the

equation?

Consequence:

The estimated values of all the other regression coefficients included in the model will

still be unbiased, their variance however will be higher so we can expect lower r^2 and

larger standard errors for our estimated coefficients.

This will happen unless:

1) the irrelevant variable is uncorrelated with every included variable

Choosing which variables belong in the model and which do not is difficult. Let

economic theory and careful judgment guide your choice of variables. Thinking about the

problem is the hard part and must be done before you estimate the model.

There are some formal specification criteria we will look at.

What you should not do:

1)Do not test various combination of variables until you find something that you like (Data mining – estimating a lot of specifications before “the” equation has been chosen)

2) Sequential specification search – sequentially dropping variables

What you should do:

Ramsey’s Regression Specification Error Test (RESET)

Most-used test next to ܴR^2 The RESET test is a general test that measures whether the

overall fit of a regression equation can be significantly improved by adding polynomials

in ܻY^hat. If so, then there is specification error.

Akaike’s Information Criterion and the Schwarz Criterion

These criteria allow you to compare alternative specifications by adjusting the residual

sum of squares (RSS) for the sample size and the number of explanatory variables:

The Linear Form

- This choice of equation is linear in the independent variables:

- A constant slope between ܺx1 and Y and ܺx2 and Y

- The elasticity of Y with respect to x1 and ܺx2 is not constant

The Double-Log Form

- This choice of equation is linear in the coefficients but not linear in the independent variables

- The slope between ܺx1 and Y and ܺx2 and Y is not constant:

- The elasticity of Y with respect to ܺx1 and ܺx2 is constant:

- All observations on X and Y are strictly positive

The Semi-Log Form

- Two popular choices of equation for this form

- In the first case, ܺx2 is linearly related to Y, while ܺx1 is nonlinearly related to Y.

- In the second case, the coefficients have a very useful interpretation – If ܺx1 increases by one unit, beta1 represents the percentage change in Y.

-- Why would this be a useful interpretation for a coefficient? Suppose Y is an employee’s

salary and X is years of experience. Each additional year of experience may be associated

with a given percentage increase in salary.

The Polynomial Form

- These types of forms are more general the linear form. The dependent variable is

expressed as a function of the independent variables, some of which have been raised to

powers greater than one

- The slope between ܺଵ and Y is not constant:

- Note: It is difficult to interpret individual regression coefficients.

- The elasticities are not constan

The Inverse Form

- This form expresses the dependent variable as a function of the reciprocal (inverse) of one or more of the independent variables

- The impact of a variable is expected to approach zero as it gets very very large

- The slope between ܺx1 and Y is not constant. Notice ܺx1 cannot be zero

Problems with Incorrect Functional Form

- Once again, economic theory should be used to establish the functional form of the relationship between the dependent variable and each independent variable.

- Avoid using goodness of fit over the sample to make your decision about functional form:

-- Cannot use ܴR^2 to compare models when Y has been transformed because the TSS will be different:

-- The “best” fitting functional form might not be accurate outside of your sample

The Consequences of Serial Correlation (seen mostly in time series data)

1. Pure serial correlation does not cause bias in the regression coefficient estimates.

2. Serial correlation causes OLS to no longer be a minimum variance estimator.

3. Serial correlation causes the estimated variances of the regression coefficients to be

biased, leading to unreliable hypothesis testing. The t-statistics will actually appear

to be more significant than they really are.

Testing for First-Order Serial Correlation

Plotting the residuals is always a good first step.

The most common formal test is the Durbin-Watson d test.

Remedies for Serial Correlation

1. Cochrane-Orcutt Procedure

Transform the equation with AR(1) error structure into one that is not autocorrelated.

2. Newey-West Standard Errors

Adjust the standard errors of the estimated regression coefficients but not the

estimates themselves since they are still unbiased.

Large-Sample Distributions

Two very useful properties when the sample size is large:

1) Law of large numbers (LLN)

o As the sample size n increases, the sample mean of a set of random variables

approaches its expected value

2) Central limit theorem (CLT)

o Let ܺଵ,…,ܺ௡ be a random sample from the same distribution with ܧሺܺ௜ሻ ൌ ߤ

and ܸܽݎሺܺ௜ሻ ൌ ߪଶ. Then the sampling distribution of the random variable

ܼ௡ ൌ ሺܺ

ത

െ ߤሻ/ሾߪ݊√/ሿ converges to the standard normal ܰሺ0, 1ሻ as n

converges to infinity.

Heteroskedasticity is the violation of Assumption 5 (the error term has a constant

variance).

Pure Heteroskedasticity

Tends to be seen in cross-sectional data more than time series data. Tends to be seen

when there is a lot of variation in the dependent variable.

Heteroskedasticity that is a function of the error term of a correctly specified regression

equation.

Assumption 5 is the assumption of homoskedasticity

The Consequences of Heteroskedasticity

1. Pure heteroskedasticity does not cause bias in the regression coefficient estimates.

2. Heteroskedasticity causes OLS to no longer be a minimum variance estimator.

3. Heteroskedasticity causes the estimated variances of the regression coefficients to be

biased, leading to unreliable hypothesis testing. The t-statistics will actually appear

to be more significant than they really are.

Testing for Heteroskedasticity

Plotting the residuals is always a good first step.

The Park Test

The White Tes

Remedies for Heteroskedasticity

As always, make sure there is no obvious specification error.

1. Weighted Least Squares

2. Heteroskedasticity-Corrected Standard Errors

Adjust the standard errors of the estimated regression coefficients but not the

estimates themselves since they are still unbiased. These standard errors are called

White Heteroskedasticity-Consistent Standard Errors.

3. Redefine the variables

Switching from a linear model to a double-log model might do it.

Type I and Type II errors

For any test procedure three outcomes are possible:

1) A correct decision

2) Rejecting H0 when it is true (Type I error)

3) Not rejecting H0 when it is false (Type II error)

The probability of committing a Type I error is denoted as Alpha and is also referred to as the significance level of the test.

The probability of committing a Type II error is denoted as Beta .The power of the test,

defined as Hnot ,is the probability of rejecting H0 when it is false.

any attempt to reduce the probability of a type I error automatically

increases the probability of a type II error.

In practice, for hypothesis testing, we choose a maximum value for the type I error that is

acceptable to us (say 5%) and then derive the decision rule for which the type II error is a

minimum.

For Normally distributed datasets we may use sigmoid

As this sampling distribution is known, Table B-1 in the text can be used to pin down

what “large” means in a statistical sense.

Look up the entry corresponding to n – 1 degrees of freedom and the given level of

significance ߙ and obtain the critical value.

Reject Hnot if the observed t is greater than the critical value, otherwise do not reject Hnot

P-values

We could have conducted our hypothesis test using a p-value. A p-value is the

probability of observing a value of a test statistic as large as we did when the null

hypothesis is true. It is also the largest probability of a Type I error when ܪ ଴is true.

http://www.sfu.ca/~dsignori/buec333/lecture%2016.pdf

Perfect (or Exact) Multicollinearity

If two or more independent variables have an exact linear relationship between them then

we have perfect multicollinearity.

Examples: including the same information twice (weight in pounds and weight in

kilograms), not using dummy variables correctly (falling into the dummy

variable trap), etc.

Consequence: OLS cannot generate estimates of regression coefficients (error message).

Why? OLS cannot estimate the marginal effect of ܺଵ on ܻ while holding ܺଶ constant

because ܺଶ moves exactly when ܺଵ moves!

Solution: Easy - Drop one of the variables!

The Consequences of Multicollinearity

1. Imperfect multicollinearity does not violate Assumption 6. Therefore the GaussMarkov Theorem tells us that the OLS estimators are BLUE.

So then why do we care about multicollinearity?

2. The variances and the standard errors of the regression coefficient estimates will

increase. This means lower t-statistics.

3. The overall fit of the regression equation will be largely unaffected by

multicollinearity. This also means that forecasting and prediction will be largely

unaffected.

4. Regression coefficients will be sensitive to specifications. Regression coefficients

can change substantially when variables are added or dropped.

The Detection of Multicollinearity

High Correlation Coefficients

Pairwise correlations among independent variables might be high (in absolute value).

Rule of thumb: If the correlation > 0.8 then severe multicollinearity may be present.

High ࡾ૛with low t-Statistic Values

Possible for individual regression coefficients to be insignificant but for the overall fit of

the equation to be high.

High Variance Inflation Factors (VIFs)

A VIF measures the extent to which multicollinearity has increased the variance of an

estimated coefficient. It looks at the extent to which an explanatory variable can be

explained by all the other explanatory variables in the equation.

Remedies for Multicollinearity

No single solution exists that will eliminate multicollinearity. Certain approaches may be

useful:

1. Do Nothing

Live with what you have.

2. Drop a Redundant Variable

If a variable is redundant, it should have never been included in the model in the

first place. So dropping it actually is just correcting for a specification error. Use

economic theory to guide your choice of which variable to drop.

3. Transform the Multicollinear Variables

Sometimes you can reduce multicollinearity by re-specifying the model, for

instance, create a combination of the multicollinear variables. As an example, rather

than including the variables GDP and population in the model, include

GDP/population (GDP per capita) instead.

4. Increase the Sample Size

Increasing the sample size improves the precision of an estimator and reduces the

adverse effects of multicollinearity. Usually adding data though is not feasible.

Even if you correct the model there still may be an imperfect multicollinearity

One-hot encoding converts it into n variables, while dummy encoding converts it into n-1 variables.

I hear that for one-hot encoding, intercept can lead to collinearity problem, which makes the model not sound. Someone call it "dummy variable trap".

one-hot-encoding directly induces perfect multicollinearity, we drop one of the columns from the encoded features. For e.g., we may choose to drop speed\_medium in this case, but the choice is completely arbitrary.

Multicollinearity

Multicollinearity occurs when two or more independent variables (a.k.a. features) in the dataset are correlated with each other. There are several methods using which we can measure the degree and direction of correlation for bivariate cases (more information on measures of correlation), while multicollinearity is generally measured using Variance Inflation Factor (more information on measures of multicollinearity). In a nutshell, multicollinearity is said to exist in a dataset when the independent variables are (nearly) linearly related to each other. Cases like as shown in Fig. 1. are called Perfect Multicollinearity. Likewise, we also have cases of Imperfect Multicollinearity, in which one or more highly linear relationships may be of our concern. These directly impact the linear regression analysis (refer to these lecture notes for more information on this)

he vectors that we use to encode the categorical columns are called ‘Dummy Variables’. We intended to solve the problem of using categorical variables, but got trapped by the problem of Multicollinearity. This is called the Dummy Variable Trap.

As mentioned earlier, this directly impacts the linear regression analysis because linear regression assumes non-existence of multicollinearity in the dataset. However, it also poses some other problems in Machine Learning tasks. Let us say, we train a logistic regression model on the dataset. We would expect our model to learn weights

we can express any one of the three independent variables in terms of the other two

(atleast) one of the features we are working with is redundant- that feature could be any one of the three, since equation-2 could be written with any one of them in the LHS. So, we are making our model learn an additional weight which is not really needed. This consumes computational power and time. This also gives an optimisation objective that might not be very reasonable and might also be difficult to work with. Too many independent variables may also lead to Curse of Dimensionality. If multicollinearity also comes alongwith that, things become worse.

We not only want our model to predict well, but we also want it to be interpretable. For e.g., Logistic Regression is expected to learn relatively higher values for weights corresponding to relatively more important features. More important features have a greater impact on the final prediction. But if features are correlated, then it becomes hard to judge which feature has more “say” in the final decision because their values are actually dependent on one another. This affects the values of the weights. In other words, the weights not only get decided based on how an independent variable correlates to the dependent variable, they also get influenced by how independent variables correlate with one another.