Unlock the Power of Linear regression

SCSRU Workshop

Statistical Consulting and Survey Research Unit
University of Waterloo

2025-02-18

1. Planning and conducting a study

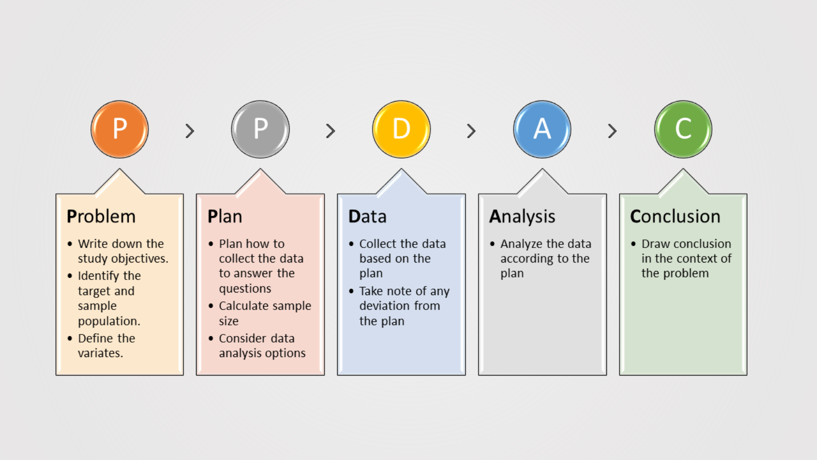


Figure 1: Consider PPDAC when planning and conducting a study

After collecting the data

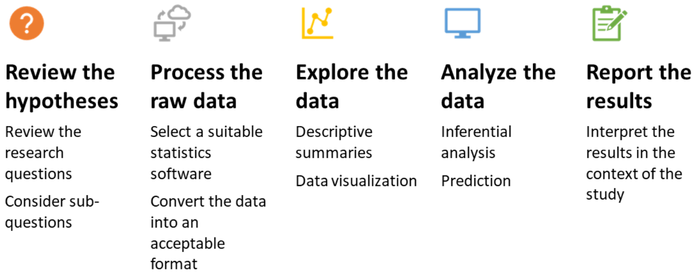


Figure 2: Recommended process

2.1 The R libraries

In this workshop, we will be using the following packages:

```r
# load the required packages
library(wooldridge)
library(corrplot)
library(lmtest)
library(MASS)
```

Please install the libraries if you have not done so. For example, to install `wooldridge`,

```r
install.packages("wooldridge")
```

2.2 The data

Throughout this workshop, we will be looking at the data set `econmath` from the R package *wooldridge*.

The data set was sent out to you a few days ago. Please set your working directory to where you saved the data set and load the data set into your R environment.

```
data("econmath") # load the data econmath
```

This data set contains information about students taking an economics class in college.

2.2 The data (Continued)

A data set is usually represented by a table of rows and columns.

- The rows represent individual observations;
- The column represents "features" or "factors" of the individual observations.
- Use the function `head()` to preview the first six rows of the data set.
- Use the function `View()` to see the whole data set.
- Use the function `summary()` for a brief summary of the data, including the minimum value, maximum value, the mean and median of each variable in the data set.

2.2 The data (Continued)

```r
head(econmath) # preview of the data set
```

```
##   age work study econhs colgpa hsgpa acteng actmth act mathscr male calculus
## 1  23   15  10.0      0 3.4909 3.355     24     26  27      10    1        1
## 2  23    0  22.5      1 2.1000 3.219     23     20  24       9    1        0
## 3  21   25  12.0      0 3.0851 3.306     21     24  21       8    1        1
## 4  22   30  40.0      0 2.6805 3.977     31     28  31      10    0        1
## 5  22   25  15.0      1 3.7454 3.890     28     31  32       8    1        1
## 6  22    0  30.0      0 3.0555 3.500     25     30  28      10    1        1
##   attexc attgood fathcoll mothcoll score
## 1      0       0        1        1 84.43
## 2      0       0        0        1 57.38
## 3      1       0        0        1 66.39
## 4      0       1        1        1 81.15
## 5      0       1        0        1 95.90
## 6      1       0        0        1 83.61
```

2.2 The data (Continued)

- The data set contains some missing data.
- Discard the data points with missing fields and gather them in a new data set

```
econ <- econmath[complete.cases(econmath), ]
```

Research goals

Question of Interest:

What factors are significantly associated with a student's score in a college economics course?

- Find how the variable `score`, i.e., the final score in an economics course measured as a percentage, can be "explained" by other variables.

3. Types of variables

In an regression problem, variables can be broadly categorized into two groups:

- **Dependent/Response/Outcome/Explained/Predicted Variable**: the variable that we want to study, usually denoted as y in linear regression models.
  - In our case, the dependent variable is `score`.
  - Linear regression is typically used to model *continuous* outcomes.
- **Independent/Control/Explanatory/Covariate/Predictor Variables**: variables which may influence the dependent variable, denoted as X in linear models.
  - These variables can be of different data types, *continuous* or *categorical*.
- What are **continuous** or **categorical** variables?

3.1 Continuous variables

- A continuous variable is a variable that can take any value over a continuous range.
- Usually, the variable will have a measurement unit, e.g., in our dataset:
  - `age` (years),
  - `work` (hours worked per week),
  - `study` (hours studying per week)
- In R, continuous data is usually defined as `num` or `int`.
- In our dataset, other continuous variables include:
  - `colgpa` (college GPA at the beginning of the semester), `hsgpa` (high school GPA), `acteng` (ACT English score), `actmth` (ACT math score), and `act` (ACT composite score).

3.2 Categorical variables

- Also known as *discrete* or *qualitative* variables.
- A categorical variable is a variable that can only take values over a finite set of values (or levels).
  - A university student's major.
  - A person's blood type.
  - The type of drinks at Starbucks.
  - A person's eye colour.
  - A person's level of agreement about a statement.
- We introduce three major types of categorical variables: **binary**, **nominal** and **ordinal** variable.

3.2.1 Binary variables

- **Binary variable**: a special categorical variables with only 2 levels. E.g., in our dataset,
  - `male` (=1 if male)
  - `econhs` (=1 if taken economics),
  - `calculus` (=1 if taken calculus),
  - `fathcoll` (=1 if father has BA),
  - `mothcoll` (=1 if mother has BA).

3.2.2 Nominal variables

- **Nominal variable**: a categorical variable with no specific order. Examples include:
  - A university student's major.
  - A person's blood type.
  - The type of drinks at Starbucks.
  - A person's eye color.

3.2.3 Ordinal variables

- **Ordinal variables**: a categorical variable with natural ordering. Examples include:
  - A person's eye color.
  - A person's level of agreement about a statement.
- Notice that the example "a person's eye color" shows up as nominal and ordinal variable. Why?
- In our dataset, `mathscr` (math quiz score, only takes in 11 values from 0 to 1) is an ordinal variable.
  - How to distinguish ordinal variable with continuous variable?
  - A student with math quiz score 8 does not mean his/she is twice as "good" as a student with score 4.
  - But 80 pounds of apples is twice as heavy as 40 pounds of bananas.

Question: Is the variable type fixed?

- We cannot determine the variable type by its name. To accurately categorize a variable, we need to consider how it is recorded.
- A common example is *age.*
  - When considered as continuous/numeric variable: age is recorded an exact value, e.g. 25, 35.5, 80, etc.
  - When considered as categorical variable: age is recorded in categories, e.g. <20, 21-25, $80+$, etc.

4. Data manipulation

- Before analyzing the data, we should spend some time to check that the structure of the data to ensure all the variables are entered properly.
- We should manually tell R to properly restore each variable in the same way as it should be.

This is what the dataset originally looks like:

```r
str(econ)
```

```
## 'data.frame':    814 obs. of  17 variables:
##  $ age     : int  23 23 21 22 22 22 22 22 22 21 ...
##  $ work    : num  15 0 25 30 25 0 20 20 28 22.5 ...
##  $ study   : num  10 22.5 12 40 15 30 25 15 7 25 ...
##  $ econhs  : int  0 1 0 0 1 0 1 0 0 0 ...
##  $ colgpa  : num  3.49 2.1 3.09 2.68 3.75 ...
##  $ hsgpa   : num  3.35 3.22 3.31 3.98 3.89 ...
##  $ acteng  : int  24 23 21 31 28 25 15 28 28 18 ...
##  $ actmth  : int  26 20 24 28 31 30 19 30 28 19 ...
##  $ act     : int  27 24 21 31 32 28 18 32 30 17 ...
##  $ mathscr : int  10 9 8 10 8 10 9 9 6 9 ...
##  $ male    : int  1 1 1 0 1 1 0 1 0 0 ...
##  $ calculus: int  1 0 1 1 1 1 1 1 0 1 ...
##  $ attexc  : int  0 0 1 0 0 1 0 1 1 0 ...
##  $ attgood : int  0 0 0 1 1 0 1 0 0 1 ...
##  $ fathcoll: int  1 0 0 1 0 0 0 1 0 0 ...
##  $ mothcoll: int  1 1 1 1 1 1 0 1 1 0 ...
##  $ score   : num  84.4 57.4 66.4 81.2 95.9 ...
```

4.1 Dealing with Categorical Variables

- So far, all the variables are restores as either `num` or `int`, i.e., R thinks they are all continuous/numeric variables.
- To tell R that some variables are categorical, we use the function `factor()`.
- For binary variables,

```
econ$male <- factor(econ$male)
econ$econhs <- factor(econ$econhs)
econ$calculus <- factor(econ$calculus)
econ$fathcoll <- factor(econ$fathcoll)
econ$mothcoll <- factor(econ$mothcoll)
```

- For ordinal variable,

```
econ$mathscr <- factor(econ$mathscr, ordered = TRUE)
```

This is what the dataset looks like right now:

```
str(econ)
```

```
## 'data.frame':    814 obs. of  17 variables:
##  $ age     : int  23 23 21 22 22 22 22 22 22 21 ...
##  $ work    : num  15 0 25 30 25 0 20 20 28 22.5 ...
##  $ study   : num  10 22.5 12 40 15 30 25 15 7 25 ...
##  $ econhs  : Factor w/ 2 levels "0","1": 1 2 1 1 2 1 2 1 1 1
##  $ colgpa  : num  3.49 2.1 3.09 2.68 3.75 ...
##  $ hsgpa   : num  3.35 3.22 3.31 3.98 3.89 ...
##  $ acteng  : int  24 23 21 31 28 25 15 28 28 18 ...
##  $ actmth  : int  26 20 24 28 31 30 19 30 28 19 ...
##  $ act     : int  27 24 21 31 32 28 18 32 30 17 ...
##  $ mathscr : Ord.factor w/ 10 levels "1"<"2"<"3"<"4"<..: 10 9
##  $ male    : Factor w/ 2 levels "0","1": 2 2 2 1 2 2 1 2 1 1
##  $ calculus: Factor w/ 2 levels "0","1": 2 1 2 2 2 2 2 2 1 2
##  $ attexc  : int  0 0 1 0 0 1 0 1 1 0 ...
##  $ attgood : int  0 0 0 1 1 0 1 0 0 1 ...
##  $ fathcoll: Factor w/ 2 levels "0","1": 2 1 1 2 1 1 1 2 1 1
##  $ mothcoll: Factor w/ 2 levels "0","1": 2 2 2 2 2 2 1 2 2 1
##  $ score   : num  84.4 57.4 66.4 81.2 95.9 ...
```

5. The linear model

- To find out the factors that affects the students' `score`, we consider to fit a linear model:

$$score = \beta_0 + \beta_1 \times X_1 + \cdots + \beta_p \times X_p + \epsilon$$

  - This is a **linear** model, because the left-hand-side has a linear relation with the right-hand-side.
  - The **response/dependent variable** `score`: the variable that we want to predict
  - The **explanatory/independent variables** $X_1, \cdots, X_p$: the variables that we think can explain the variation of the **response/dependent variable**.
    - Each $X_1, \cdots, X_p$ in this model corresponds to either a continuous variable or categorical variable in the dataset.
  - The **error** ϵ: accounts for the variation of `score` that the $X_1, \cdots, X_p$' cannot explain.
  - Because, no model is perfect.

5.1 Assumptions

Linear Regression has the **LINE** assumptions:

- **Linearity (L)**: the response variable and the explanatory variables have a linear relation.
  - Otherwise, there is no point to use Linear Regression!
- **Independence (I)**: the errors ϵ are independently distributed.
  - i.e., incorrectly predicting the score of student A will not affect my prediction for student B.
- **Normality (N)**: the errors ϵ follow Normal distribution, with zero mean and some variance.
- **Equal Variance (E)**: the variance of the errors ϵ is constant.

I-N-E assumptions can be summarized with:

$$\epsilon \overset{iid}{\sim} N(0, \sigma^2)$$

i.e., the errors ϵ are *independently and identically distributed (iid)* and follow Normal distribution.

5.1.1 Fitting a linear model in R

We can fit a linear regression model using the function `lm()`

- Regress `score` against no variable but an intercept:

```
Model_0 = lm(score ~ 1, data = econ) # "1" is intercept
```

- Regress `score` against one variable, `colgpa` (a student's college GPA):

```
Model_1 = lm(score ~ colgpa, data = econ)
```

- Regress `score` against two variables, `colgpa` and `hsgpa` (a student's high school GPA):

```
Model_2 = lm(score ~ colgpa + hsgpa, data = econ)
```

- Regress `score` against all variables in the dataset:

```
Model_full = lm(score ~ ., data = econ)
# "." is the shortcut to include all variables.
```

5.1.2 "Summarize" a fitted linear model using summary()

```r
summary(Model_1)
```

```
##
## Call:
## lm(formula = score ~ colgpa, data = econ)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -41.784  -6.399   0.564   7.553  32.183
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  32.3463     2.0181   16.03   <2e-16 ***
## colgpa       14.3232     0.7051   20.31   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 10.84 on 812 degrees of freedom
## Multiple R-squared:  0.337,  Adjusted R-squared:  0.3361
## F-statistic: 412.6 on 1 and 812 DF,  p-value: < 2.2e-16
```

- Practice: summarize the full model, i.e., `summary(Model_full)`.

5.1.3 Get the Model Coefficients using `coef()`

```r
coef(Model_2)
```

```
## (Intercept)      colgpa       hsgpa
##   19.126435   12.666816    5.343784
```

- Discussion: What can we say about these numbers? Or, do they have a meaning?
- We will talk about interpretation later.

5.2 Interaction terms

- An interaction happens when the effect of an independent variable is affected by the value of another independent variable.
  - E.g., A smart student gets good grades. A hardworking student also gets good grades. A smart and hardworking students gets even better grades. Then, there is *interaction effect* between "being smart" and "being hardworking" on the outcome: grades.
- There are two-factor interactions (2FIs), three-factor interactions (3FIs), etc.
- The higher order interactions are less likely to be significant. They are also harder to interpret.
- We recommend to not go beyond 2FIs unless the literature suggests that certain higher order interaction terms are meaningful.
- Experts' opinion can be helpful to identify meaningful higher-order terms.

5.2.1 Modelling interaction terms using “∗” and “:”

- Consider a model with two independent variables: `colgpa` (continuous, student’s college GPA) and `calculus` (binary, =1 if student took calculus)

```
Model_3 = lm(score ~ colgpa + calculus, data = econ)
```

- To incorporate the interaction term between `colgpa` and `calculus`, we have two ways:

```
Model_3_1 = lm(score ~ colgpa + calculus + colgpa:calculus,
               data = econ) # using ":"
Model_3_2 = lm(score ~ colgpa * calculus,
               data = econ) # using "*"
```

- In R, `A * B` is equivalent to `A + B + A:B`.
- Practice: summarize `Model_3_1` and `Model_3_2` and see if they are the same.

5.3 Model selection

- The data set `econ` has 15 independent variables, hence our linear regression models can contain any combination of these variables and/or their interactions.
- So which model we should choose?
- When fitting linear models, it is important to perform model selection procedures and assess the model fit before interpreting the results.
- In general, a "good" model should:
  - fit the observed data well, i.e., explain the response variable well.
  - not *overfit* the data, i.e., can make good out-of-sample predictions.
- There are 2 major approaches:
  - *Manual selection*: Likelihood Ratio Test (LRT), AIC, BIC, adjusted R^2, etc.
  - *Automatic selection*: forward selection, backward selection, stepwise selection, etc.

5.3.1 Likelihood ratio test

- One of the most common ways to compare models against each other is through the likelihood ratio test (LRT).
- It is used to compare a **full** model vs a **nested** model.
- E.g., a **full** model:

```
Model_full = lm(score ~ ., data = econ)
```

- A **nested** model contains a subset of variables that appear in the full model, e.g.,

```
Model_2 = lm(score ~ colgpa + hsgpa, data = econ)
```

- Use `lrtest()` to perform a LRT between two models:

```
lrtest(Model_2, Model_full)
```

- Discussion: what does the test output say?

5.3.1 Likelihood ratio test (continued)

- The LRT tests if the nested model is **as good as** the full model.
  - Because the full model contains the nested mode, the former should perform no worse than the later.
  - But, is it necessary to make the mode as complicated as the full model?
- From the test results, the p-value < 0.05. So, at 5% significance level, we say that **the nested model is not sufficient to explain the data and the full model is preferred.**
  - Note that we are not saying the full model is the "best", it is only preferred over the nested model.
- If the p-value ≥ 0.05, then we can say that the nested model is **as good as** the full model, and the **simpler is preferred**.

5.3.2 Information criteria

- For nested or non-nested models, we can also perform model selection by
  - Akaike information criterion (AIC)
  - Bayesian information criterion (BIC)
- Models with **smaller** AIC or BIC represents better fit.
- Although both AIC and BIC are similar, research has shown that each are appropriate for different tasks.
- Use `AIC()` and `BIC()`

```
AIC(Model_full)
AIC(Model_2)
BIC(Model_full)
BIC(Model_2)
```

- Discussion: what does the output say?

5.3.3 Stepwise model selection procedure

- When there are many covariates, the built-in stepwise model selection procedure, `step()` may be a better option.
- The `step()` function can evaluate the model on the AIC, or the BIC, where a smaller value represents a better fit.
- We can also specify which direction we want the function to search through: `forward`, `backward` or `both`.
- Exercise: try the following. Which linear model is selected?

```r
## minimal model: intercept only
M0 <- lm(score ~ 1, data = econ)
## maximal model: all main effects and interaction effects
Mfull <- lm(score ~ (.)^2, data = econ)
# stepwise selection
Mstart <- lm(score ~ . - acteng - actmth, data = econ)
Mstep <- step(object = Mstart,
              scope = list(lower = M0, upper = Mfull),
              direction = "both", trace = FALSE)
# summary(Mstep) # model chosen by stepwise selection
```

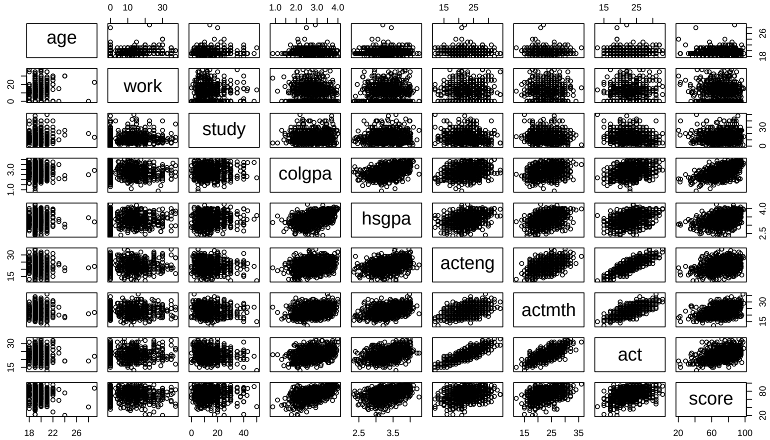
6. Model diagnostics

- After having fitted the model, it is important that we check that the assumptions of our model are satisfied in order to verify that our model is valid.
- Basically, we check the **LINE** assumptions using diagnostic plots.
- Practice: In the following slides, try to regenerate the diagnostic plots. And, what can we say about each plot?

6.1 Check Linearity (L): Scatter plot

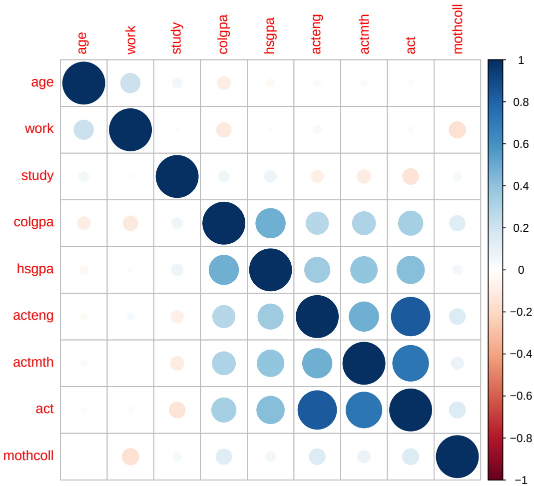
- Scatter plot is always the first step which helps us check the linear relationships among our variables.

```r
pairs(~ age + work + study + colgpa + hsgpa + acteng +
      actmth + act + score, data = econ)
```



6.1 Check Linearity (in terms of Correlation): Heatmap

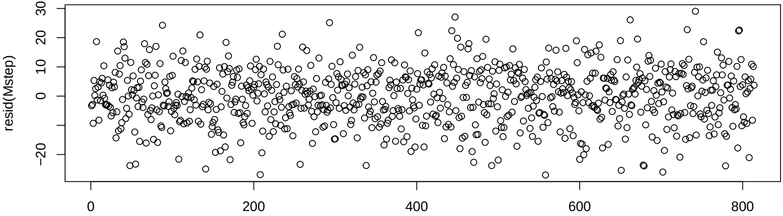
```
tmp <- data.matrix(econ[, c(1:3, 5:9, 16)])
corrplot(cor(tmp), method = "circle")
```



6.2 Check Independence (I): Residual plot

- It is not always possible to assess the independence assumption in practice.
- If data are serially correlated (e.g., time-series, or measurement repeatedly observed from the same object), we may be able to identify any violation of the independence assumption by plotting residuals against their natural ordering.
- If there is no-serial correlation, we should expect the residual plot alike a horizontal band around 0 with no specific pattern.

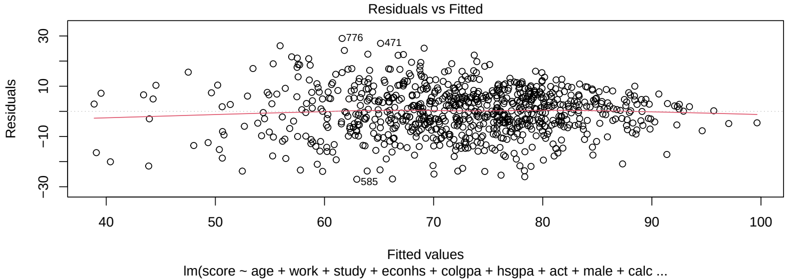
```
plot(resid(Mstep))
```



6.3 Check Equal-Variance (E)

- Plot a scatter plot of residuals and fitted values to check (E).
- If (E) is satisfied, you should see a horizontal band of residuals evenly distributed along with the fitted values.
- In R, function `plot(Model)` can plot all diagnostic plots for a fitted model `Model`. By setting `which=1`, we get the residuals vs fitted values plot.

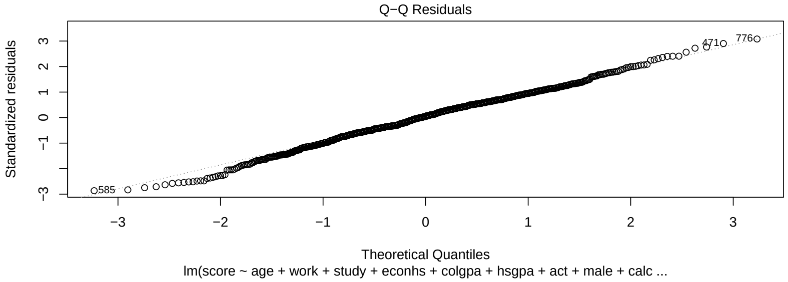
```r
plot(Mstep, which = 1, ask = FALSE)
```



6.4 Check Normality (N)

- Plot a quantile-quantile (QQ) plot to check (N).
- If (N) is satisfied, you should see the dots closed to the straight dashed line.
- In R, set `which = 2` for QQ plot.

```
plot(Mstep, which = 2, ask = FALSE)
```



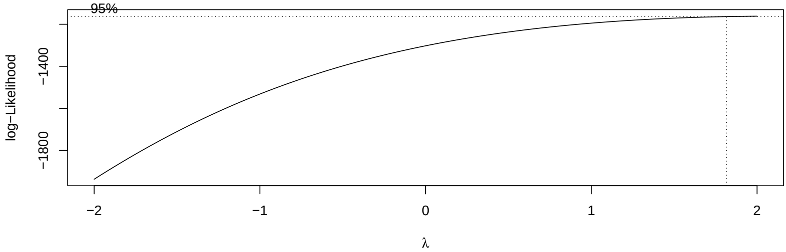
6.4.1 If Normality (N) fails...

- If Normality is violated, we can consider the **power transformation**.
- Power transformation means, when we find the original data Y does not follow Normal distribution, we can raise Y to the power of λ, i.e.,
$$Y \mapsto Y^{\lambda}$$
- If λ is properly chosen, Y^{λ} will follow Normal distribution.

6.4.1 If Normality (N) fails. . . (continued)

- In R, we can use `boxcox()` applied to a fitted model.

```
bc = boxcox(Model_1)
```



- The best power transformation is to take the power of:

```
bc$x[which.max(bc$y)]
```

```
## [1] 2
```

6.5 Check Multicolinearity

- If two explanatory variables are highly correlated, the regression has trouble figuring out whether the change in the response variable is due to one explanatory variable or the other, or both.
- As a result, the estimates for the model coefficients can change a lot from one random sample to another.
- This is known as **variance inflation**.
- We can detect collinearity by checking the **variance inflation factor**.

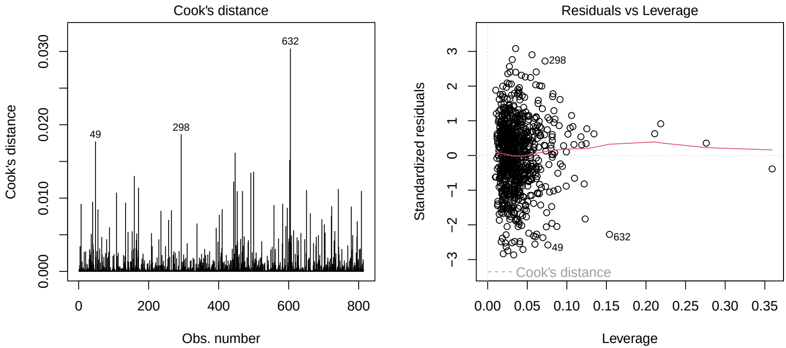
```
X <- model.matrix(Mstep)
VIF <- diag(solve(cor(X[, -1])))
sqrt(VIF)
```

6.6 Checking Outliers and Influential Points

- Outliers are observations which have unusually large residuals compared to the others.
- Influential points are observations which have unusually large **leverage** compared to the others.
- We can use the **Cook's distance** (left) and **Residual vs Leverage** (right) plots to detect outliers and influential plots respectively.
- In R, set `which = 4` in the function `plot()` for **Cook's distance** plot and set `which=5` for **Residual vs Leverage** plot.

6.6 Checking Outliers and Influential Points (Continued)

```
par(mfrow = c(1, 2))
plot(Mstep, which = c(4, 5), ask = FALSE)
```



- Discussion: Outliers and Influential Points are not the same, why?

7. Interpretation of the results

- Multiple regression analysis provides a *ceteris paribus* ("all things being equal") interpretation even though the data have not been collected in a *ceteris paribus* fashion.
- Consider a fitted model below:

$$score = \hat{\beta}_0 + \hat{\beta}_1 \times colgpa + \hat{\beta}_2 \times hsgpa$$

  - $\hat{\beta}$ are the estimates of the model coefficients.
- $\hat{\beta}_1$ quantifies the association of `colgpa` with `score`, holding `hsgpa` fixed.
- Practice: Refit the model and summarize the model output.

7. Interpretation of the results (continued)

- In the model summary, you should find this table:

```
##               Estimate Std. Error   t value     Pr(>|t|)
## (Intercept) 19.126435  3.6904376  5.182701 2.762615e-07
## colgpa      12.666816  0.7987907 15.857490 1.572946e-49
## hsgpa        5.343784  1.2544458  4.259876 2.285231e-05
```

- Practice: What is $\hat{\beta}_1$ and $\hat{\beta}_2$?
- Interpretation:

  *Keeping ‘hsgpa‘ fixed, one unit increase in ‘colgpa‘ is associated with **an average** increase of 12.6668 in ‘score.‘*

- Discussion:
  - How to interpret the association between `hsgpa` and `score`?
  - Why **on average**?
  - Is the association between `colgpa` and `score` “reliable”?
  - What is “reliable”?

7. Interpretation of the results (continued)

- Interpretation of the association between `hsgpa` and `score`:

  *Keeping 'colgpa' fixed, one unit increase in 'hsgpa' is associated with **an average** increase of 5.3438 in 'score.'*

- Linear Regression only tells us an **on average** (or group) effect, the individual differences and variations cannot be explained by Linear Regression Models.
- **Statistical significance**: to tell whether our estimations of the associations between the response and explanatory variables are "reliable".
- We introduce 2 ways to check **statistical significance** of a model estimate: confidence interval (CI) & p-value.

7.1 Confidence Interval (CI) & p-value

- Estimation has errors, which is quantified by the Standard Errors (`Std. Error`) in the model summary.

```
##                Estimate Std. Error   t value     Pr(>|t|)
## (Intercept) 19.126435  3.6904376  5.182701 2.762615e-07
## colgpa      12.666816  0.7987907 15.857490 1.572946e-49
## hsgpa        5.343784  1.2544458  4.259876 2.285231e-05
```

- The $(1-\alpha)\%$ **Confidence Interval (CI)** of $\hat{\beta}$ is the range that the true value of β lies in with $(1-\alpha)\%$ of chance.
  - α is the significance level, usually set as 0.05.
  - Use function `confint` and specify `level = 1-`α to compute the $(1-\alpha)\%$ CI:

```r
confint(Model_2, level=0.95)
```

```
##                 2.5 %    97.5 %
## (Intercept) 11.88250 26.370370
## colgpa      11.09888 14.234757
## hsgpa        2.88144  7.806127
```

7.1 Confidence Interval (CI) & p-value (continued)

- If the estimation of β, i.e., $\hat{\beta}$, lies within the $(1-\alpha)\%$ CI, we say that β **is statistically significant at** $(1-\alpha)\%$ **confidence level**.
- The other way to check statistical significance is the **p-value**, which is given in the last column of

```
##              Estimate Std. Error   t value     Pr(>|t|)
## (Intercept) 19.126435  3.6904376  5.182701 2.762615e-07
## colgpa      12.666816  0.7987907 15.857490 1.572946e-49
## hsgpa        5.343784  1.2544458  4.259876 2.285231e-05
```

- Compare the p-value with the significance level $\alpha = 0.05$.
- If p-value < 0.05, then it is statistically significant.
- Practice: Are $\hat{\beta}_1$ and $\hat{\beta}_2$ statistically significant? Why?

7.2 Prediction Intervals (PI)

- Unlike Confidence Interval, a **prediction interval (PI)** is an estimate of an interval in which a future observation will fall, with a certain probability, given what has already been observed.
  - So, it is a range that we can reasonably expect our model prediction to fall in.
- In R, we can compute the PI using function `predict()` and setting `interval = "prediction"`:

```r
predict(Model, newdata, interval='prediction')
```

8. Statistical vs practical significance

The p-values are commonly used as an indicator of significance/importance. However, we want to remind readers that:

- Statistical inference techniques test for statistical significance.
- Statistical significance means that the effect observed in a sample is very unlikely to occur if the null hypothesis is true.
- Whether this observed effect has practical importance is an entirely different question. The experts in the field of interest determine whether these results have any practical importance.

Some notes about p-values

The ASA's Statement on p-values:

- P-values can indicate how incompatible the data are with a specified statistical model.
- P-values do not measure the probability that the studied hypothesis is true, or the probability that the data were produced by random chance alone.
- Scientific conclusions and business or policy decisions should not be based only on whether a p-value passes a specific threshold.
- Proper inference requires full reporting and transparency
- A p-value, or statistical significance, does not measure the size of an effect or the importance of a result.
- By itself, a p-value does not provide a good measure of evidence regarding a model or hypothesis.

9. Next steps

Linear models are widely used in many literature. However, there are limitations. We encourage you to explore other models such as

- generalized linear models,
- linear mixed effect models, and
- non-parametric statistics model,

to find a good fit for your application.

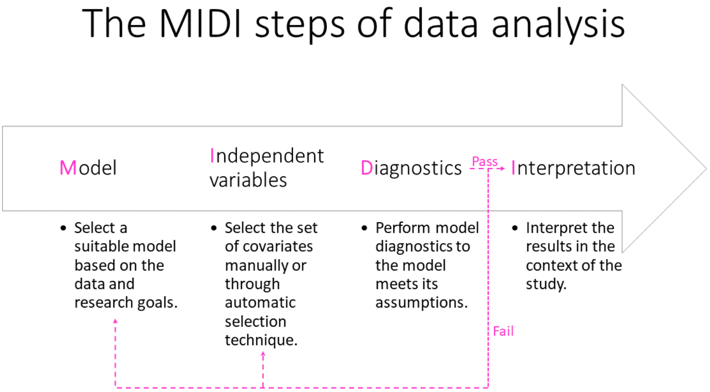


Figure 3: Recommended steps to data analysis

Beyond this workshop

For those who are interested to learn more, the SCSRU hosts statistics seminars and workshops focusing on topics commonly encountered by researchers on campus. Please check our website for future events.

Thank you!

The Statistical Consulting and Survey Research Unit (SCSRU) is the unit through which the Department of Statistics and Actuarial Science provides statistical advice to those working on research problems.