logistic: Name _____ Ar Ah 8/12 → 4/6

1. Formatting:

all margins 2.5cm — informative title

12 pt size — name on all pages

no raw R code or output — all pages numbered

max **10** pages — no blurry plots (**NOT png**)

1/1

2. Introduction/Background:

brief background and statement of scientific question — imprecise

all variables defined

0.75/1

3. EDA:

univariate numerical — bivariate numerical (cor)

univariate graphical — bivariate graphical pairs plots for numerical vars

1.5/2

4. Model fitting:

give mathematical definition of model no ^ in Eq 4

state how model fitted (ie, maximum likelihood) not 'try to'

CLEARLY describe how model selected

define all terms model parameters, all performance measures

1.25/2

5. Model assessment:

CLEARLY state model assumptions: + give PRIMARY references not done?

1. binary outcome 2. independent obs
3. linear relation between logit and linear predictor
4. no multicollinearity 5. no outliers (6. large sample size)

0.5/2

carry out assessment (numerical / graphics):

scatterplots of logit vs. predictors (linearity assumption)

DEFINE -> Cook's distance / standardized residuals (outliers)

vif (to identify multicollinearity)

Define

* Because Cook's D is small on an absolute scale Don't really need to exclude 'outliers': 4/n is a guideline, not a law

5/8

6. Write out final *estimated* model **mathematically**

hat on *response* variable max **2 sig digits** (after decimal) on coefs

7. Plots:

label size (not too small) informative captions

placement explanations

8. Conclusions

1. recap analysis 2. state and interpret main findings

9. Overall presentation (clarity of explanations, appropriate citations / references) :

poor satisfactory good excellent

10. Other comments:

A – no / incomplete / insufficient references

B – cite PRIMARY refs (not course notes, not wikipedia, *etc.*)

C – interpretation (cannot conclude causation, only association)

D – use your OWN WORDS / no apparently unattributed quotations

E – Intro: **1.** Give context; **2.** Clearly state scientific question; **3.** Describe data

F – univariate graphical: histograms not boxplots

G – (mathematical) model misspecified / unclear

H – clearly EXPLAIN / INTERPRET PLOTS (don't just state conclusions)

I – plot size / aspect ratio (make 'pretty')

Other:

logistic: Name AbAk 8/12 → 4/6

1. Formatting:

0.75/1

all margins 2.5cm — informative title

12 pt size — name on all pages

no raw R code or output — all pages numbered

max **10** pages — no blurry plots (**NOT png**)

no cover page / no EPFL logo

2. Introduction/Background:

0.75/1

brief background and statement of scientific question

all variables defined

- too many digits

imprecise / incomplete

1/2

3. EDA: incomplete - ALL VARS

univariate numerical — not done? — bivariate numerical (cor)

univariate graphical — bivariate graphical

4. Model fitting: models not described/explained

1.25/2

give mathematical definition of model

state how model fitted (ie, maximum likelihood)

CLEARLY describe how model selected

define all terms

lasso / AUC / ROC / AIC

5. Model assessment:

CLEARLY state model assumptions: + give PRIMARY references

1.5/2

1. binary outcome 2. independent obs
3. linear relation between logit and linear predictor
4. no multicollinearity 5. no outliers (6. large sample size)

not 'confirm'

carry out assessment (numerical / graphics):

scatterplots of logit vs. predictors (linearity assumption)

DEFINE -> Cook's distance / standardized residuals (outliers)

vif (to identify multicollinearity)

H-L test, performance measures not defined

5.25/8

(ok)

1

6. Write out final *estimated* model **mathematically**

hat on *response* variable max **2 sig digits** (after decimal) on coefs

0.75

7. Plots: +cor mat too small

label size (not too small) informative captions

placement explanations

8. Conclusions

0.5

1. recap analysis 2. state and interpret main findings

be specific

(implicit eda)

spell check

9. Overall presentation (clarity of explanations, appropriate citations / references) :

0.5

poor satisfactory good excellent

10. Other comments:

A – no / incomplete / insufficient references

B – cite PRIMARY refs (not course notes, not wikipedia, *etc.*)

C – interpretation (cannot conclude causation, only association)

D – use your OWN WORDS / no apparently unattributed quotations

E – Intro: **1.** Give context; **2.** Clearly state scientific question; **3.** Describe data

F – univariate graphical: histograms not boxplots

G – (mathematical) model misspecified / unclear

H – clearly EXPLAIN / INTERPRET PLOTS (don't just state conclusions)

I – plot size / aspect ratio (make 'pretty')

Other: - put extra plots in an appendix

you need to include EDA for ALL vars

2.75/4

logistic: Name _____ Ro An 8/12 → 4/6

1. Formatting:

1/1

all margins 2.5cm — informative title

12 pt size — name on all pages

no raw R code or output — all pages numbered

max **10** pages — no blurry plots (**NOT png**)

0.5/1 2. Introduction/Background:

brief background and statement of scientific question — not clear

all variables defined

1.25/2 3. EDA: — write out words

univariate numerical — bivariate numerical (cor)

univariate graphical — bivariate graphical all (numerical) pairs + square

1.5/2 4. Model fitting:

give mathematical definition of model — in terms of logit (before π)

state how model fitted (ie, maximum likelihood)

CLEARLY describe how model selected — what test do you use?

define all terms

5. Model assessment: assess, not 'confirm'

1.25/2 CLEARLY state model assumptions: + give PRIMARY references

1. binary outcome 2. independent obs
3. linear relation between logit and linear predictor
4. no multicollinearity 5. no outliers (6. large sample size)

carry out assessment (numerical / graphics):

scatterplots of logit vs. predictors (linearity assumption)

DEFINE -> Cook's distance / standardized residuals (outliers) — h not defined

vif (to identify multicollinearity)

Define

5.5/8 — Unusual pattern in standardized resids
— is there a spatial relation — Plot with index?

0.75/1 (ok) logit?

6. Write out final *estimated* model **mathematically** — write 0.00 in scientific notation as #.## $\times 10^{-\#}$

hat on *response* variable — max **2 sig digits** (after decimal) on coefs

0.75/1

7. Plots:

label size (not too small) — informative captions

placement — explanations

0.5/1

8. Conclusions (not 'make sure')

1. recap analysis — 2. state and interpret main findings — be specific — not 'meaningful' — 'seems to be'?

0.5/1

9. Overall presentation (clarity of explanations, appropriate citations / references) :

poor — satisfactory — good — excellent

10. Other comments:

A – no / incomplete / insufficient references

B – cite PRIMARY refs (not course notes, not wikipedia, *etc.*)

C – interpretation (cannot conclude causation, only association)

D – use your OWN WORDS / no apparently unattributed quotations

E – Intro: **1.** Give context; **2.** Clearly state scientific question; **3.** Describe data

F – univariate graphical: histograms not boxplots

G – (mathematical) model misspecified / unclear

H – clearly EXPLAIN / INTERPRET PLOTS (don't just state conclusions)

I – plot size / aspect ratio (make 'pretty')

Other: _____

_____

_____

_____

_____

2.5/4

logistic: Name _____ Te Be 7.75/12 → 3.875/6

1. Formatting:

all margins 2.5cm — informative title

12 pt size — name on all pages

no raw R code or output — all pages numbered

max **10** pages — no blurry plots (**NOT png**)

0.75/1

2. Introduction/Background:

brief background and statement of scientific question

all variables defined

0.5/1

3. EDA:

univariate numerical — bivariate numerical (cor)

univariate graphical — bivariate graphical

0.75/2

4. Model fitting:

give mathematical definition of model

state how model fitted (ie, maximum likelihood)

CLEARLY describe how model selected

define all terms

1.25/2

5. Model assessment:

CLEARLY state model assumptions: + give PRIMARY references

1. binary outcome 2. independent obs
3. linear relation between logit and linear predictor
4. no multicollinearity 5. no outliers (6. large sample size)

1/2

carry out assessment (numerical / graphics):

scatterplots of logit vs. predictors (linearity assumption)

DEFINE -> Cook's distance / standardized residuals (outliers)

vif (to identify multicollinearity)

5.25/8

- too many digits
- imprecise
- all vars
- incomplete
- all pairs (numeric)
- incorrect coef interpretation
- stepwise procedure
- AIC: measure of model
- clinically relevant = ?
- what does statistically meaningful mean?
- explain plots, then interpret then conclude
- define

0.75/1

6. Write out final *estimated* model **mathematically**

hat on *response* variable max **2 sig digits** (after decimal) on coefs

7. Plots:

1/1

label size (not too small) informative captions

placement explanations

0.25/1

8. Conclusions

be specific not 'confirmed'

1. recap analysis + EDA 2. state and interpret main findings

9. Overall presentation (clarity of explanations, appropriate citations / references) :

0.5/1

poor satisfactory good excellent

10. Other comments:

A – no / incomplete / insufficient references

B – cite PRIMARY refs (not course notes, not wikipedia, *etc.*)

C – interpretation (cannot conclude causation, only association)

D – use your OWN WORDS / no apparently unattributed quotations

E – Intro: **1.** Give context; **2.** Clearly state scientific question; **3.** Describe data

F – univariate graphical: histograms not boxplots

G – (mathematical) model misspecified / unclear

H – clearly EXPLAIN / INTERPRET PLOTS (don't just state conclusions)

I – plot size / aspect ratio (make 'pretty')

Other: _____

_____

_____

_____

_____

2.5/4

logistic: Name _____ Da Bu 7.25/12 → 3.625/8

1. Formatting:

0.75

all margins 2.5cm — informative title

12 pt size — name on all pages

no raw R code or output — all pages numbered

(Don't need R details)

max 10 pages — no blurry plots (**NOT png**)

0.75

2. Introduction/Background: — too many digits imprecise (predicted LGG (GBM)

brief background and statement of scientific question

all variables defined

1.5

3. EDA: Cramér's V, plx,

univariate numerical — bivariate numerical (cor)

2 univariate graphical — bivariate graphical

4. Model fitting: wald test stat + SE, not just p

1.25

give mathematical definition of model

state how model fitted (ie, maximum likelihood)

2

CLEARLY describe how model selected

define all terms AIC, Wald test

5. Model assessment: not 'ensure validity', not 'verify'

1/2

CLEARLY state model assumptions: + give PRIMARY references

1. binary outcome 2. independent obs
3. linear relation between logit and linear predictor
4. no multicollinearity 5. no outliers (6. large sample size)

carry out assessment (numerical / graphics):

explain plots then interpret then conclude

scatterplots of logit vs. predictors (linearity assumption)

DEFINE -> Cook's distance / standardized residuals (outliers)

vif (to identify multicollinearity)

Define

5.25/8

0.75 Define Y

6. Write out final *estimated* model **mathematically**

hat on *response* variable max **2 sig digits** (after decimal) on coefs

7. Plots:

0.25

label size (not too small) informative captions

placement explanations figure 5 seems out of place; interpret

8. Conclusions very incomplete be specific not correct

0

1. recap analysis 2. state and interpret main findings

9. Overall presentation (clarity of explanations, appropriate citations / references) :

0.5

poor satisfactory good excellent

10. Other comments:

A – no / incomplete / insufficient references

B – cite PRIMARY refs (not course notes, not wikipedia, *etc.*)

C – interpretation (cannot conclude causation, only association)

D – use your OWN WORDS / no apparently unattributed quotations

E – Intro: **1.** Give context; **2.** Clearly state scientific question; **3.** Describe data

F – univariate graphical: histograms not boxplots

G – (mathematical) model misspecified / unclear

H – clearly EXPLAIN / INTERPRET PLOTS (don't just state conclusions)

I – plot size / aspect ratio (make 'pretty')

Other: _____

_____

_____

_____

_____

2/5

logistic: Name _____ Ma Fo 9.5/12 → 4.75/6

1. Formatting:

0.75/1

all margins 2.5cm — informative title

12 pt size — name on all pages

no raw R code or output — all pages numbered

Don't need R fn details

max **10** pages — no blurry plots (**NOT png**)

too many digits

2. Introduction/Background:

0.75/1

imprecise

brief background and statement of scientific question

all variables defined

3. EDA:

2/2

univariate numerical — bivariate numerical (cor)

univariate graphical — bivariate graphical

4. Model fitting: estimated coefs: $\hat{\beta}$'s; incorrect coef interpretation

1.5/2

give mathematical definition of model

state how model fitted (ie, maximum likelihood)

CLEARLY describe how model selected

define all terms

5. Model assessment:

CLEARLY state model assumptions: + give PRIMARY references (all)

1/2

1. binary outcome ✓ 2. independent obs
3. linear relation between logit and linear predictor
4. no multicollinearity ✓ 5. no outliers ✓ (6. large sample size)

carry out assessment (numerical / graphics):

First explain plots then interpret then conclude

scatterplots of logit vs. predictors (linearity assumption)

DEFINE -> Cook's distance / standardized residuals (outliers)

vif (to identify multicollinearity)

Define

6/8

(OK) $\log \frac{p}{1-p}$

1/1

6. Write out final *estimated* model **mathematically**

hat on *response* variable max **2 sig digits** (after decimal) on coefs

7. Plots:

1/1

label size (not too small) informative captions

placement explanations

8. Conclusions not 'confirmed'

0.75

1. recap analysis 2. state and interpret main findings not correct

(implicit EDA)

0.75/1

9. Overall presentation (clarity of explanations, appropriate citations / references) :

poor satisfactory good excellent

10. Other comments:

A – no / incomplete / insufficient references

B – cite PRIMARY refs (not course notes, not wikipedia, *etc.*)

C – interpretation (cannot conclude causation, only association) predicted log-odds

D – use your OWN WORDS / no apparently unattributed quotations

E – Intro: **1.** Give context; **2.** Clearly state scientific question; **3.** Describe data

F – univariate graphical: histograms not boxplots

G – (mathematical) model misspecified / unclear

H – clearly EXPLAIN / INTERPRET PLOTS (don't just state conclusions)

I – plot size / aspect ratio (make 'pretty')

Other: _____

_____

_____

_____

_____

3.5/4

logistic: Name _____ Lu Ga 8.25/12 → 4.125/6

1. Formatting:

- all margins 2.5cm
- 12 pt size
- **no raw R code or output**
- max **10** pages
- informative title
- name on all pages
- all pages numbered
- no blurry plots (**NOT png**)

0.75/1

2. Introduction/Background:

- brief background and statement of scientific question
- all variables defined

0.75/1

3. EDA:

- univariate numerical
- univariate graphical
- bivariate numerical (cor)
- bivariate graphical

1.25/2

4. Model fitting:

- give mathematical definition of model
- state how model fitted (ie, maximum likelihood)
- CLEARLY describe how model selected
- define all terms

1.25/2

5. Model assessment:

- CLEARLY state model assumptions: + give PRIMARY references

1.25/2

1. binary outcome 2. independent obs
3. linear relation between logit and linear predictor
4. no multicollinearity 5. no outliers (6. large sample size)

carry out assessment (numerical / graphics):

- scatterplots of logit vs. predictors (linearity assumption)
- DEFINE -> Cook's distance / standardized residuals (outliers)
- vif (to identify multicollinearity)

5.25/8

← too many digits
imprecise
Don't need stuff above table 2
put in matrix so easier to read
not just heatmap
pairs plots (numerical)
not 'try to model'
no ^ → ^ when estimated values
perfect separation
AIC
explanatory
define

6. Write out final *estimated* model **mathematically**

hat on *response* variable max **2 sig digits** (after decimal) on coefs

7. Plots:

label size (not too small) informative captions

placement explanations

8. Conclusions

1. recap analysis 2. state and interpret main findings

9. Overall presentation (clarity of explanations, appropriate citations / references):

poor satisfactory good excellent

10. Other comments:

A – no / incomplete / insufficient references

B – cite PRIMARY refs (not course notes, not wikipedia, *etc.*)

C – interpretation (cannot conclude causation, only association)

D – use your OWN WORDS / no apparently unattributed quotations

E – Intro: **1.** Give context; **2.** Clearly state scientific question; **3.** Describe data

F – univariate graphical: histograms not boxplots

G – (mathematical) model misspecified / unclear

H – clearly EXPLAIN / INTERPRET PLOTS (don't just state conclusions)

I – plot size / aspect ratio (make 'pretty')

Other: _____

_____

_____

_____

_____

logistic: Name _____ En Go 9.75/12 → 4.875/6

1. Formatting:

all margins 2.5cm — informative title

12 pt size — name on all pages

no raw R code or output — all pages numbered

max **10** pages — no blurry plots (**NOT png**)

2. Introduction/Background:

brief background and statement of scientific question — somewhat imprecise

all variables defined

0.75/1

3. EDA:

univariate numerical — bivariate numerical (cor) (large yr)

univariate graphical — bivariate graphical

2/2

4. Model fitting:

give mathematical definition of model (First in terms of logit then prob)

state how model fitted (ie, maximum likelihood)

CLEARLY describe how model selected use paragraphing

define all terms AIC - max log like model

1.5/2

5. Model assessment:

CLEARLY state model assumptions: + give PRIMARY references

1. binary outcome 2. independent obs
3. linear relation between logit and linear predictor
4. no multicollinearity 5. no outliers (6. large sample size)

+ explain

1.25/2

carry out assessment (numerical / graphics):

scatterplots of logit vs. predictors (linearity assumption)

DEFINE -> Cook's distance / standardized residuals (outliers)

vif (to identify multicollinearity)

Define

6.5/8

6. Write out final *estimated* model **mathematically**

hat on *response* variable max **2 sig digits** (after decimal) on coefs

7. Plots:

label size (not too small) informative captions

placement explanations

8. Conclusions

1. recap analysis 2. state and interpret main findings

9. Overall presentation (clarity of explanations, appropriate citations / references) :

poor satisfactory good excellent

10. Other comments:

A – no / incomplete / insufficient references

B – cite PRIMARY refs (not course notes, not wikipedia, *etc.*)

C – interpretation (cannot conclude causation, only association)

D – use your OWN WORDS / no apparently unattributed quotations

E – Intro: **1.** Give context; **2.** Clearly state scientific question; **3.** Describe data

F – univariate graphical: histograms not boxplots

G – (mathematical) model misspecified / unclear

H – clearly EXPLAIN / INTERPRET PLOTS (don't just state conclusions)

I – plot size / aspect ratio (make 'pretty')

Other: (*) write all all hyp test details (null/alt/ Test Stat / numerical value of tst stat, null dist, conclusion, practical interpretation

GWAS: Name _____ DaGr 8.5/12 → 4.25/6

1. Formatting:

all margins 2.5cm	informative title
12 pt size	name on all pages
no raw R code or output (Don't need R details)	all pages numbered
max **10** pages	no blurry plots (**NOT png**)

1/1

2. Introduction/Background:

brief background and statement of scientific question

all variables defined

1/1

3. EDA / PCA:

relevant histograms / exploratory plots (NO BOXPLOTS) Don't just state conclusion

EXPLAIN RELATION between PCs and population stratification

plot pc2 (y-axis) vs pc1 (x-axis)

1/2

4. Pre-processing / QC steps: CLEARLY EXPLAIN

SNP QC: criteria and reasons

sample QC: criteria and reasons

Hardy-Weinberg equilibrium: DEFINE and say how it relates to quality

measure of LD and how it is used Define What is IBD analysis? kinship coef's Purcell + Reich

1/2

5. Association / post-association analysis: + give PRIMARY references

write out model ***mathematically*** (for a given SNP): CLEARLY define all variables and domains - Model MUST RELATE TO SNP

describe association analysis in words and mathematically what hyp test?

Manhattan plot and explanation not manhattan + explain thresholds

lambda analysis (***including*** mathematical definition of lambda and SQUARE QQ plots)

1.25/2

5.25/8

1/1

6. Identify significant SNPs (include all relevant information)

7. Plots: manhattan plot

0.75/1

label size (not too small) informative captions

placement explanations

8. Conclusions Penn GATH (not penn CATH)

0.75/1

1. recap analysis (incomplete) 2. state and interpret main findings

9. Overall presentation (clarity of explanations, appropriate citations / references) :

0.75/1

poor satisfactory good excellent

10. Other comments:

A – no / incomplete / insufficient references

B – cite PRIMARY refs (not course notes, not wikipedia, *etc.*)

C – interpretation (cannot conclude causation, only association)

D – use your OWN WORDS / no apparently unattributed quotations

E – Intro: **1.** Give context; **2.** Clearly state scientific question; **3.** Describe data

F – univariate graphical: histograms not boxplots

G – (mathematical) model misspecified / unclear

H – clearly EXPLAIN / INTERPRET PLOTS (don't just state conclusions)

I – plot size / aspect ratio (make 'pretty')

Other:

3.25/4

logistic: Name _____ An Is 7.5/12 → 3.75/6

1. Formatting:

0.75

all margins 2.5cm — informative title

12 pt size — name on all pages

no raw R code or output — all pages numbered

max **10** pages — no blurry plots (**NOT png**)

no R formulas

0.5

2. Introduction/Background:

brief background and statement of scientific question — not clear

all variables defined — or explanatory predictor (not 'exploratory') variables

1.5

3. EDA:

2 univariate numerical — bivariate numerical (cor)

univariate graphical — bivariate graphical

1.25 (Don't need figure 1)

4. Model fitting:

2 give mathematical definition of model — BEFORE fitting

state how model fitted (ie, maximum likelihood)

CLEARLY describe how model selected — what are important vars? Don't eliminate vars based on EDA

define all terms AIC

5. Model assessment:

1/2

CLEARLY state model assumptions: + give PRIMARY references

1. binary outcome 2. independent obs
3. linear relation between logit and linear predictor
4. no multicollinearity 5. no outliers (6. large sample size)

carry out assessment (numerical / graphics):

Plots: First explain then interpret then conclude

scatterplots of logit vs. predictors (linearity assumption)

DEFINE -> Cook's distance / standardized residuals (outliers) — Define leverage, Pearson / Deviance resids

vif (to identify multicollinearity)

- Deviance resids QQ; what is the expected distribution?
- H-L test not defined

5/8

0.75/1

6. Write out final *estimated* model **mathematically**

equation

hat on *response* variable max **2 sig digits** (after decimal) on coefs

0.75/1

7. Plots:

label size (not too small) informative captions

placement explanations

0.5/1

8. Conclusions

be specific not 'confirms'

1. recap analysis 2. state and interpret main findings

not 'verifies'

0.5/1

9. Overall presentation (clarity of explanations, appropriate citations / references) :

poor satisfactory good excellent

how do you know that the model is 'important'

10. Other comments:

A – no / incomplete / insufficient references

B – cite PRIMARY refs (not course notes, not wikipedia, *etc.*)

C – interpretation (cannot conclude causation, only association)

D – use your OWN WORDS / no apparently unattributed quotations

E – Intro: **1.** Give context; **2.** Clearly state scientific question; **3.** Describe data

F – univariate graphical: histograms not boxplots

G – (mathematical) model misspecified / unclear

H – clearly EXPLAIN / INTERPRET PLOTS (don't just state conclusions)

I – plot size / aspect ratio (make 'pretty')

Other: _____

_____

_____

_____

_____

2.5/4

(*) Don't need to re-do

discrete: Name Cole 10/12 → ~~5/5~~ 6/6

1. Formatting:

all margins 2.5cm	informative title
12 pt size	name on all pages
no raw R code or output	all pages numbered
max **10** pages	no blurry plots (**NOT png**)

1/1

2. Introduction/Background:

somewhat imprecise

0.75/1 brief statement of scientific question

all variables defined

1.25/2 3. EDA:

cross-tabs mosaic plot 1 plot all 3 vars

4. Testing independence:

2/2 define parameters; give null and alt hyps MATHEMATICALLY

test statistic MATHEMATICALLY and NUMERICALLY

null distribution of test statistic; p-value and conclusion + interpretation

define all terms

5. CMH test: (careful-interpretation)

1.75/2 Explain clearly in words what you are testing

CLEARLY state null and alt hyps mathematically

test statistic MATHEMATICALLY and NUMERICALLY

null distribution of test statistic; p-value and conclusion + interpretation

ASSUMPTION for valid p-value

6.75/8

1/1

6. Woolf test:

null, alt, test stat, null dist of test stat, p-value, conclusion + interpretation

7. Plots:

1/1

label size (not too small) informative captions

placement explanations

0.5/1

8. Conclusions

be specific not entirely correct

1. recap analysis (+ EDA) 2. state and interpret main findings

0.75/1

9. Overall presentation (clarity of explanations, appropriate citations / references) :

poor satisfactory good excellent

10. Other comments:

A – no / incomplete / insufficient references

B – cite PRIMARY refs (not course notes, not wikipedia, *etc.*)

C – interpretation (cannot conclude causation, only association)

D – use your OWN WORDS / no apparently unattributed quotations

E – Intro: **1.** Give context; **2.** Clearly state scientific question; **3.** Describe data

F – univariate graphical: histograms not boxplots

G – (mathematical) model misspecified / unclear

H – clearly EXPLAIN / INTERPRET PLOTS (don't just state conclusions)

I – plot size / aspect ratio (make 'pretty')

Other:

3.25/4

logistic: Name _____ Dilo 9/12 → 4.5/6

1. Formatting:

all margins 2.5cm | informative title

12 pt size | name on all pages

no raw R code or output | all pages numbered

max **10** pages | no blurry plots (**NOT png**)

1/1 (no R formulas, Cook's)

2. Introduction/Background: imprecise 0.75/1

brief background and statement of scientific question

all variables defined

3. EDA: – ALL vars 1.25/2

univariate numerical incomplete | bivariate numerical (cor)

univariate graphical | bivariate graphical pairs plots

4. Model fitting: – interpretation: +/- coef. predicted risk

1.5/2

give mathematical definition of model

state how model fitted (ie, maximum likelihood) most 'relevant' vars?

CLEARLY describe how model selected vague – be specific

define all terms AIC / stepwise

5. Model assessment: assessed, not 'verified'

CLEARLY state model assumptions: + give PRIMARY references

1.5/2 (explanation not entirely correct)

1. binary outcome 2. independent obs
3. linear relation between logit and linear predictor
4. no multicollinearity 5. no outliers (6. large sample size)

carry out assessment (numerical / graphics):

scatterplots of logit vs. predictors (linearity assumption)

DEFINE -> Cook's distance / standardized residuals (outliers)

vif (to identify multicollinearity)

6/8

6. Write out final *estimated* model **mathematically**

hat on *response* variable max **2 sig digits** (after decimal) on coefs

7. Plots:

label size (not too small) informative captions

placement explanations

8. Conclusions

1. recap analysis 2. state and interpret main findings

9. Overall presentation (clarity of explanations, appropriate citations / references) :

poor satisfactory good excellent

10. Other comments:

A – no / incomplete / insufficient references

B – cite PRIMARY refs (not course notes, not wikipedia, *etc.*)

C – interpretation (cannot conclude causation, only association)

D – use your OWN WORDS / no apparently unattributed quotations

E – Intro: **1.** Give context; **2.** Clearly state scientific question; **3.** Describe data

F – univariate graphical: histograms not boxplots

G – (mathematical) model misspecified / unclear

H – clearly EXPLAIN / INTERPRET PLOTS (don't just state conclusions)

I – plot size / aspect ratio (make 'pretty')

Other:

GWAS: Name JuLeF

1. Formatting:

all margins 2.5cm	informative title
12 pt size	name on all pages
no raw R code or output	all pages numbered
max **10** pages	no blurry plots (**NOT png**)

2. Introduction/Background:

brief background and statement of scientific question

all variables defined

3. EDA / PCA:

relevant histograms / exploratory plots (NO BOXPLOTS)

EXPLAIN RELATION between PCs and population stratification

plot pc2 (y-axis) vs pc1 (x-axis)

4. Pre-processing / QC steps: CLEARLY EXPLAIN

SNP QC: criteria and reasons

sample QC: criteria and reasons

Hardy-Weinberg equilibrium: DEFINE and say how it relates to quality

measure of LD and how it is used

5. Association / post-association analysis: + give PRIMARY references

write out model ***mathematically*** (for a given SNP): CLEARLY define all variables and domains - Model MUST RELATE TO SNP

describe association analysis in words and mathematically

Manhattan plot and explanation

lambda analysis (***including*** mathematical definition of lambda and **SQUARE** QQ plots)

0.75 (2) 6. Identify significant SNPs (include all relevant information) 'incomplete' + keep table together + SE or Z

7. Plots:

0.75 label size (not too small) informative captions vertical axis of scree plot incorrect

placement explanations

0.5 8. Conclusions what 'valuable information'? be specific vague

1. recap analysis 2. state and interpret main findings spell check

9. Overall presentation (clarity of explanations, appropriate citations / references) :

0.5 poor satisfactory good excellent

10. Other comments:

A – no / incomplete / insufficient references

B – cite PRIMARY refs (not course notes, not wikipedia, *etc.*)

C – interpretation (cannot conclude causation, only association)

D – use your OWN WORDS / no apparently unattributed quotations

E – Intro: **1.** Give context; **2.** Clearly state scientific question; **3.** Describe data

F – univariate graphical: histograms not boxplots

G – (mathematical) model misspecified / unclear

H – clearly EXPLAIN / INTERPRET PLOTS (don't just state conclusions)

I – plot size / aspect ratio (make 'pretty') PC plots

Other: You need to EXPLAIN REASONS, not just state conclusions – for example – 'SNPs with low MAF will minimally contribute to statistical power, may be prone to errors...' – WHY? You need to EXPLAIN it, not just say it

shape of QQ plot must be SQUARE

(*) Table of Results does not include all top hits

surv: Name _____ Ali Li 7.5/12 → 3.75/6

1. Formatting:

all margins 2.5cm	informative title
12 pt size	name on all pages
no raw R code or output	all pages numbered
max **10** pages	no blurry plots (**NOT png**)

1/1

2. Introduction/Background:

brief background and statement of scientific question — unclear and imprecise

all variables defined — Cox (not cox)

0.75/1

3. EDA: – All vars

univariate numerical — all summary stats in table — bivariate numerical (cor)

univariate graphical — bar charts not hists for categorical vars — bivariate graphical — all pairs

1/2

4. Model fitting:

KM (write out estimator and variance) + log-rank test: state null / alt hyps, value of test stat, give null dist of test stat, p-value, conclusion (reject / do not reject) — mathematically

First write Cox + df model mathematically (no 1) then estimated model

state how model fitted (ie, maximum partial likelihood)

1/2

CLEARLY describe how model selected

define all terms

5. Model assessment: not 'verified'

CLEARLY state Cox PH assumptions: + give PRIMARY references

1.25/2

1. hazards are proportional
2. linear form for covariates
3. no outliers

carry out assessment (graphics / tests) and EXPLAIN:

Schoenfeld residuals (PH assumption) – global test?
Martingale residuals (linear form for continous variables)
Deviance residuals (to identify outliers)

5/8

(ok)

1/1

6. Write out final *estimated* model **mathematically**

hat on *response* variable max **2 sig digits** (after decimal) on coefs

7. Plots:

0.75/1

label size (not too small) informative captions

placement explanations

8. Conclusions

0.25/1

1. recap analysis (implicit EDA) 2. state and interpret main findings

be specific not correct

9. Overall presentation (clarity of explanations, appropriate citations / references):

0.5/1

poor satisfactory good excellent

10. Other comments:

A – no / incomplete / insufficient references

B – cite PRIMARY refs (not course notes, not wikipedia, *etc.*)

C – interpretation (cannot conclude causation, only association)

D – use your OWN WORDS / no apparently unattributed quotations

E – Intro: **1.** Give context; **2.** Clearly state scientific question; **3.** Describe data

F – univariate graphical: histograms not boxplots

G – (mathematical) model misspecified / unclear

H – clearly EXPLAIN / INTERPRET PLOTS (don't just state conclusions)

I – plot size / aspect ratio (make 'pretty')

Other:

2.5/4

logistic: Name _____ Na Mo 8.5/12 → 4.25/6

1. Formatting:

all margins 2.5cm — informative title

12 pt size — name on all pages

no raw R code or output — all pages numbered

max **10** pages — no blurry plots (**NOT png**)

0.75/1 (Don't need R details)

2. Introduction/Background: — Too many digits imprecise

0.75/1

brief background and statement of scientific question

all variables defined

3. EDA:

2/2

univariate numerical — bivariate numerical (cor)

univariate graphical — bivariate graphical

4. Model fitting:

1.5/2

give mathematical definition of model

state how model fitted (ie, maximum likelihood)

CLEARLY describe how model selected — be specific

define all terms AIC

5. Model assessment:

CLEARLY state model assumptions: + give PRIMARY references

1.25/2

1. binary outcome ✓ 2. independent obs
3. linear relation between logit and linear predictor
4. no multicollinearity ✓ 5. no outliers (6. large sample size)

carry out assessment (numerical / graphics):

scatterplots of logit vs. predictors (linearity assumption)

DEFINE -> Cook's distance / standardized residuals (outliers)

vif (to identify multicollinearity)

Define not ensured

First explain plots then interpret & then conclu

6.25/8

(ok) log P/1-P

1/1

6. Write out final *estimated* model **mathematically**

hat on *response* variable max **2 sig digits** (after decimal) on coefs

7. Plots:

1/1

label size (not too small) informative captions

placement explanations

0.5/1

8. Conclusions Use paragraphing & specific

1. recap analysis 2. state and interpret main findings clearly

0.75/1

9. Overall presentation (clarity of explanations, appropriate citations / references) :

poor satisfactory good excellent

10. Other comments:

A – no / incomplete / insufficient references

B – cite PRIMARY refs (not course notes, not wikipedia, *etc.*)

C – interpretation (cannot conclude causation, only association)

D – use your OWN WORDS / no apparently unattributed quotations

E – Intro: **1.** Give context; **2.** Clearly state scientific question; **3.** Describe data

F – univariate graphical: histograms not boxplots

G – (mathematical) model misspecified / unclear

H – clearly EXPLAIN / INTERPRET PLOTS (don't just state conclusions)

I – plot size / aspect ratio (make 'pretty')

Other: _____

_____

_____

_____

_____

2.25/4

(*) Don't need to re-do

discrete: Name _____ Ra Na 10/12 → ~~5/6~~ → 6/6

1. Formatting:

1/1

- no cover page

all margins 2.5cm	informative title
12 pt size	name on all pages
no raw R code or output	all pages numbered
max **10** pages	no blurry plots (**NOT png**)

2. Introduction/Background:

0.75/1

somewhat imprecise

brief statement of scientific question

all variables defined

3. EDA:

1.75/2

cross-tabs

mosaic plot

all 3 vars in 1 plot

4. Testing independence:

1.75/2

define parameters; give null and alt hyps MATHEMATICALLY

test statistic MATHEMATICALLY and NUMERICALLY

null distribution of test statistic; p-value and conclusion + interpretation

(clearly state E;j)

(define all terms)

5. CMH test:

1.5/2

Explain clearly in words what you are testing

CLEARLY state null and alt hyps mathematically

what are E, Var?

test statistic MATHEMATICALLY and NUMERICALLY

null distribution of test statistic; p-value and conclusion + interpretation

ASSUMPTION for valid p-value

6.25/8

6. Woolf test:

null, alt, test stat, null dist of test stat, p-value, conclusion + interpretation

7. Plots:

label size (not too small) informative captions

placement explanations

8. Conclusions

1. recap analysis (implicit EDA) 2. state and interpret main findings (not 'confirmed')

9. Overall presentation (clarity of explanations, appropriate citations / references) :

poor satisfactory good excellent

10. Other comments:

A – no / incomplete / insufficient references

B – cite PRIMARY refs (not course notes, not wikipedia, *etc.*)

C – interpretation (cannot conclude causation, only association)

D – use your OWN WORDS / no apparently unattributed quotations

E – Intro: **1.** Give context; **2.** Clearly state scientific question; **3.** Describe data

F – univariate graphical: histograms not boxplots

G – (mathematical) model misspecified / unclear

H – clearly EXPLAIN / INTERPRET PLOTS (don't just state conclusions)

I – plot size / aspect ratio (make 'pretty')

Other:

3.75/4

logistic: Name _____ Al Ps 8.75/12 → 4.375/6

1. Formatting:

all margins 2.5cm

12 pt size (thought, some text too small)

no raw R code or output

max **10** pages

informative title

name on all pages

all pages numbered

no blurry plots (**NOT png**)

2. Introduction/Background: good (too many digits)

brief background and statement of scientific question

all variables defined

3. EDA: write out words

univariate numerical

univariate graphical

bivariate numerical (cor)

bivariate graphical

4. Model fitting: Do not remove 'influential' points — in magnitude the values are very small

give mathematical definition of model

state how model fitted (ie, maximum likelihood)

CLEARLY describe how model selected

define all terms

5. Model assessment: Don't cyclically remove 'outliers' based on Cook

CLEARLY state model assumptions: + give PRIMARY references

1. binary outcome 2. independent obs
3. linear relation between logit and linear predictor
4. no multicollinearity 5. no outliers (6. large sample size)

carry out assessment (numerical / graphics):

scatterplots of logit vs. predictors (linearity assumption)

DEFINE -> Cook's distance / standardized residuals (outliers)

vif (to identify multicollinearity)

first explain plots, then interpret, then conclude

6.5/8

0.75 / why ≈?

6. Write out final *estimated* model **mathematically**

hat on *response* variable max **2 sig digits** (after decimal) on coefs

0.75 7. Plots:

label size (not too small) informative captions Figure not Fig.

placement explanations

8. Conclusions what does robust mean? How is model useful. ?

0.25 1. recap analysis (+EDA) 2. state and interpret main findings ← be specific

0.5 9. Overall presentation (clarity of explanations, appropriate citations / references) :

poor satisfactory good excellent

10. Other comments:

A – no / incomplete / insufficient references

B – cite PRIMARY refs (not course notes, not wikipedia, *etc.*)

C – interpretation (cannot conclude causation, only association)

D – use your OWN WORDS / no apparently unattributed quotations

E – Intro: **1.** Give context; **2.** Clearly state scientific question; **3.** Describe data

F – univariate graphical: histograms not boxplots

G – (mathematical) model misspecified / unclear

H – clearly EXPLAIN / INTERPRET PLOTS (don't just state conclusions)

I – plot size / aspect ratio (make 'pretty')

Other: _____

_____

_____

_____

_____

2.25/4

logistic: Name _____ MaRa 9.5/12 → 4.75/6

1. Formatting:

all margins 2.5cm — informative title

12 pt size — throughout — name on all pages

no raw R code or output — all pages numbered

max **10** pages — no blurry plots (**NOT png**)

1/1 ↳ you can put results in an appendix

2. Introduction/Background:

brief background and statement of scientific question — imprecise

all variables defined

0.75/1

3. EDA:

univariate numerical — bivariate numerical (cor)

univariate graphical — bivariate graphical — pairs scatterplots (numerical vars)

1.75/2

4. Model fitting:

give mathematical definition of model — Define model first in terms of logit, then prob

state how model fitted (ie, maximum likelihood)

CLEARLY describe how model selected — not 'relies on' + $\ell(\beta)$ looks strange

define all terms

1.75/2

5. Model assessment:

CLEARLY state model assumptions: + give PRIMARY references

1. binary outcome 2. independent obs
3. linear relation between logit and linear predictor
4. no multicollinearity 5. no outliers (6. large sample size)

carry out assessment (numerical / graphics):

scatterplots of logit vs. predictors (linearity assumption)

DEFINE -> Cook's distance / standardized residuals (outliers)

vif (to identify multicollinearity)

Define

7/8

0.75/1 logit p

6. Write out final *estimated* model **mathematically**

hat on *response* variable max **2 sig digits** (after decimal) on coefs

0.75/1 + plot size

7. Plots:

label size (not too small) informative captions

placement explanations

0.5/1 be specific, not entirely correct

8. Conclusions

1. recap analysis 2. state and interpret main findings

0.5/1 spell check

9. Overall presentation (clarity of explanations, appropriate citations / references):

poor satisfactory good excellent

10. Other comments:

A – no / incomplete / insufficient references

B – cite PRIMARY refs (not course notes, not wikipedia, *etc.*)

C – interpretation (cannot conclude causation, only association)

D – use your OWN WORDS / no apparently unattributed quotations

E – Intro: **1.** Give context; **2.** Clearly state scientific question; **3.** Describe data

F – univariate graphical: histograms not boxplots

G – (mathematical) model misspecified / unclear

H – clearly EXPLAIN / INTERPRET PLOTS (don't just state conclusions)

I – plot size / aspect ratio (make 'pretty')

Other: - please display mathematical formulas

- any inline formulas put $\displaystyle ... $

2.5/4

logistic: Name _____ FeRq 8.75/12 → 4.375/6

1. Formatting:

all margins 2.5cm — informative title

12 pt size — name on all pages

no raw R code or output — all pages numbered

max **10** pages — no blurry plots (**NOT png**)

1/1

2. Introduction/Background:

brief background and statement of scientific question — imprecise

all variables defined

0.75/1

3. EDA:

univariate numerical — bivariate numerical (cor)

univariate graphical — bivariate graphical

1.5/2

4. Model fitting: rather incomplete — pairs for numeric

give mathematical definition of model - Intro before equation

state how model fitted (ie, maximum likelihood)

CLEARLY describe how model selected - a little superficial

define all terms AIC

1.25/2

5. Model assessment:

CLEARLY state model assumptions: + give PRIMARY references Cook

1.25/2

1. binary outcome 2. independent obs
3. linear relation between logit and linear predictor
4. no multicollinearity 5. no outliers (6. large sample size)

carry out assessment (numerical / graphics):

scatterplots of logit vs. predictors (linearity assumption)

DEFINE -> Cook's distance / standardized residuals (outliers)

vif (to identify multicollinearity)

Define

First explain plots, then interpret then conclude

5.75

6. Write out final *estimated* model **mathematically**

hat on *response* variable max **2 sig digits** (after decimal) on coefs

7. Plots:

label size (not too small) informative captions

placement explanations

8. Conclusions

1. recap analysis 2. state and interpret main findings

9. Overall presentation (clarity of explanations, appropriate citations / references) :

poor satisfactory good excellent

10. Other comments:

A – no / incomplete / insufficient references

B – cite PRIMARY refs (not course notes, not wikipedia, *etc.*)

C – interpretation (cannot conclude causation, only association)

D – use your OWN WORDS / no apparently unattributed quotations

E – Intro: **1.** Give context; **2.** Clearly state scientific question; **3.** Describe data

F – univariate graphical: histograms not boxplots

G – (mathematical) model misspecified / unclear

H – clearly EXPLAIN / INTERPRET PLOTS (don't just state conclusions)

I – plot size / aspect ratio (make 'pretty')

Other: _____

_____

_____

_____

_____

logistic: Name Ma Si 7.5/12 → 3.75/6

1. Formatting:

all margins 2.5cm — informative title

12 pt size — name on all pages

no raw R code or output — all pages numbered

max **10** pages — no blurry plots (**NOT png**)

2. Introduction/Background:

brief background and statement of scientific question

all variables defined

3. EDA:

univariate numerical — bivariate numerical (cor)

univariate graphical — bivariate graphical

4. Model fitting:

give mathematical definition of model

state how model fitted (ie, maximum likelihood)

CLEARLY describe how model selected

define all terms

5. Model assessment:

CLEARLY state model assumptions: + give PRIMARY references

1. binary outcome 2. independent obs
3. linear relation between logit and linear predictor
4. no multicollinearity 5. no outliers (6. large sample size)

carry out assessment (numerical / graphics):

scatterplots of logit vs. predictors (linearity assumption)

DEFINE -> Cook's distance / standardized residuals (outliers)

vif (to identify multicollinearity)

4.5/8

6. Write out final *estimated* model **mathematically**

hat on *response* variable

max **2 sig digits** (after decimal) on coefs

7. Plots:

label size (not too small)

informative captions

placement

explanations

8. Conclusions

1. recap analysis

2. state and interpret main findings

9. Overall presentation (clarity of explanations, appropriate citations / references) :

poor satisfactory good excellent

10. Other comments:

A – no / incomplete / insufficient references

B – cite PRIMARY refs (not course notes, not wikipedia, *etc.*)

C – interpretation (cannot conclude causation, only association)

D – use your OWN WORDS / no apparently unattributed quotations

E – Intro: **1.** Give context; **2.** Clearly state scientific question; **3.** Describe data

F – univariate graphical: histograms not boxplots

G – (mathematical) model misspecified / unclear

H – clearly EXPLAIN / INTERPRET PLOTS (don't just state conclusions)

I – plot size / aspect ratio (make 'pretty')

Other: _____

_____

_____

_____

_____

GLM: Name _____ NaTa 8.25/12 → 4.125/6

1. Formatting:

all margins 2.5cm — informative title

12 pt size — name on all pages

no raw R code or output — all pages numbered

max **10** pages — no blurry plots (**NOT png**)

0.75/1

2. Introduction/Background:

(too many digits) somewhat imprecise

brief background and statement of scientific question

all variables defined

0.75/1

3. EDA:

univariate numerical — bivariate numerical (cor)

univariate graphical — bivariate graphical

1.5/2

4. Model fitting: somewhat hard to follow + explain all pairs - don't just conclude

give mathematical definition of model Before fitting, + no $\hat{}$ in Eq.!

state how model fitted (ie, maximum likelihood)

CLEARLY describe how model selected Don't select variables from EDA

define all terms & - mathematically, AIC

1/2

5. Model assessment:

CLEARLY state model assumptions: + give PRIMARY references

1.25/2

1. count outcome Poisson 2. independent obs
3. linear relation between log count and linear predictor
4. conditional mean = conditional variance

carry out assessment (numerical / graphics):

relevant scatterplots (linearity assumption)

first explain, then interpret then conclude

5.25/8

(ok) $\log E$

6. Write out final *estimated* model **mathematically**

hat on *response* variable max **2 sig digits** (after decimal) on coefs

7. Plots:

0.75 label size (not too small) informative captions

placement explanations Figure 3

0.5 8. Conclusions be specific

1. recap analysis 2. state and interpret main findings not correct

0.75 (+EDA)

9. Overall presentation (clarity of explanations, appropriate citations / references) :

poor satisfactory good excellent

10. Other comments:

A – no / incomplete / insufficient references

B – cite PRIMARY refs (not course notes, not wikipedia, *etc.*)

C – interpretation (cannot conclude causation, only association)

D – use your OWN WORDS / no apparently unattributed quotations

(2) E – Intro: **1.** Give context; **2.** Clearly state scientific question; **3.** Describe data

F – univariate graphical: histograms not boxplots

G – (mathematical) model misspecified / unclear

H – clearly EXPLAIN / INTERPRET PLOTS (don't just state conclusions)

I – plot size / aspect ratio (make 'pretty')

Other: - Don't need scientific for your results

3/4

surv: Name _____ TyTo 7/12 → 3.5/6

1. Formatting:

0/1

all margins 2.5cm — informative title

12 pt size throughout — name on all pages

no raw R code or output — all pages numbered

max **10** pages — no blurry plots (**NOT png**)

0.75/1

2. Introduction/Background:

brief background and statement of scientific question — unclear and imprecise also incorrect — cannot determine influence

all variables defined

1.5/2

3. EDA:

univariate numerical — bivariate numerical (cor)

univariate graphical + bar charts for categorical vars — bivariate graphical — all (numerical) pairs mathematically

1/2

4. Model fitting:

KM (write out estimator and variance) + log-rank test: state null / alt hyps, value of test stat, give null dist of test stat, p-value, conclusion (reject / do not reject)

first write Cox model mathematically before fitting

state how model fitted (ie, maximum partial likelihood)

CLEARLY describe how model selected

define all terms — Concordance, LRT, Wald, Score + SE + z-stat for coefs in results table

5. Model assessment: — not 'validation'

1.25/2

CLEARLY state Cox PH assumptions: + give PRIMARY references

1. hazards are proportional
2. linear form for covariates
3. no outliers

carry out assessment (graphics / tests) and EXPLAIN: — clearly relate plots to assumptions

Schoenfeld residuals (PH assumption)
Martingale residuals (linear form for continous variables)
Deviance residuals (to identify outliers)

4.5/8

1/1

6. Write out final *estimated* model **mathematically**

hat on *response* variable max **2 sig digits** (after decimal) on coefs

0.75/

7. Plots:

label size (not too small) informative captions

placement explanations

0.25/

8. Conclusions

be specific not correct

1. recap analysis 2. state and interpret main findings

0.5/1

9. Overall presentation (clarity of explanations, appropriate citations / references) : ± EDA somewhat incomplete

poor satisfactory good excellent

10. Other comments:

A – no / incomplete / insufficient references

B – cite PRIMARY refs (not course notes, not wikipedia, *etc.*)

C – interpretation (cannot conclude causation, only association)

D – use your OWN WORDS / no apparently unattributed quotations

E – Intro: **1.** Give context; **2.** Clearly state scientific question; **3.** Describe data

F – univariate graphical: histograms not boxplots

G – (mathematical) model misspecified / unclear

H – clearly EXPLAIN / INTERPRET PLOTS (don't just state conclusions)

I – plot size / aspect ratio (make 'pretty')

Other:

2.5/4

logistic: Name _____ Mc Ul 9/12 → 4.5/6

1. Formatting:

all margins 2.5cm — informative title

12 pt size — name on all pages

no raw R code or output — all pages numbered

max **10** pages — no blurry plots (**NOT png**)

4/

2. Introduction/Background:

0.75/1 brief background and statement of scientific question — imprecise

all variables defined

3. EDA: very incomplete

0.75/1 univariate numerical — summary stats

univariate graphical — all vars

bivariate numerical (cor) — not just heat map

bivariate graphical — all pairs

4. Model fitting:

1.5/2 give mathematical definition of model

state how model fitted (ie, maximum likelihood)

CLEARLY describe how model selected — describe stepwise procedure not R fn

define all terms — AIC: L is max like — LRT - null dist of Test Stat

5. Model assessment:

1.75/2 CLEARLY state model assumptions: + give PRIMARY references

1. binary outcome 2. independent obs
3. linear relation between logit and linear predictor
4. no multicollinearity 5. no outliers (6. large sample size)

carry out assessment (numerical / graphics):

scatterplots of logit vs. predictors (linearity assumption)

DEFINE -> Cook's distance / standardized residuals (outliers)

vif (to identify multicollinearity)

First explain plots then interpret then conclude

- It is always true that a model fits better when you remove outliers — look at magnitude of Cook's (which is small)

5.75/8

1/1

6. Write out final *estimated* model **mathematically**

hat on *response* variable max **2 sig digits** (after decimal) on coefs

1/1

7. Plots:

label size (not too small) informative captions

placement explanations

0.5/1

8. Conclusions

be specific

1. recap analysis (+ EDA) 2. state and interpret main findings

0.75/1

9. Overall presentation (clarity of explanations, appropriate citations / references) :

poor satisfactory good excellent

10. Other comments:

A – no / incomplete / insufficient references

B – cite PRIMARY refs (not course notes, not wikipedia, *etc.*)

C – interpretation (cannot conclude causation, only association)

D – use your OWN WORDS / no apparently unattributed quotations

E – Intro: **1.** Give context; **2.** Clearly state scientific question; **3.** Describe data

F – univariate graphical: histograms not boxplots

G – (mathematical) model misspecified / unclear

H – clearly EXPLAIN / INTERPRET PLOTS (don't just state conclusions)

I – plot size / aspect ratio (make 'pretty')

Other:

3.25/4

surv: Name _____ Thuan Sp 7.75/12 → 3.875/6

1. Formatting:

- too many digits

all margins 2.5cm	informative title
12 pt size	name on all pages
no raw R code or output	all pages numbered
max **10** pages	no blurry plots (**NOT png**)

0.75/1

Don't need R details

2. Introduction/Background: use more paragraphing imprecise

0.75/1

brief background and statement of scientific question

all variables defined

3. EDA:

1.25/

p$<$0.05 not just heatmap

univariate numerical	bivariate numerical (cor)
univariate graphical	bivariate graphical

pairs plots (numeric)

4. Model fitting: mathematically

1.25/2

KM (write out estimator and variance) + log-rank test: state null / alt hyps, value of test stat, give null dist of test stat, p-value, conclusion (reject / do not reject)

Don't need univariate Cox

state how model fitted (ie, maximum partial likelihood)

p$<$ ...

CLEARLY describe how model selected

define all terms

Efron method, concordance, LRT

5. Model assessment: not 'confirm'

1.25/2

CLEARLY state Cox PH assumptions: + give PRIMARY references

1. hazards are proportional
2. linear form for covariates
3. no outliers

carry out assessment (graphics / tests) and EXPLAIN:

First explain plots, then interpret then conclude

Schoenfeld residuals (PH assumption)
Martingale residuals (linear form for continous variables)
Deviance residuals (to identify outliers)

(not 'visual inspection shows...')

5.25/8

1/1

6. Write out final *estimated* model **mathematically**

hat on *response* variable max **2 sig digits** (after decimal) on coefs

0.75/1

7. Plots:

label size (not too small) informative captions

placement explanations

0.25/1

8. Conclusions

be specific not correct (not 'influenced by')

1. recap analysis 2. state and interpret main findings

0.5/1

9. Overall presentation (clarity of explanations, appropriate citations / references) :

poor satisfactory good excellent

10. Other comments:

A – no / incomplete / insufficient references

B – cite PRIMARY refs (not course notes, not wikipedia, *etc.*)

C – interpretation (cannot conclude causation, only association)

D – use your OWN WORDS / no apparently unattributed quotations

E – Intro: **1.** Give context; **2.** Clearly state scientific question; **3.** Describe data

F – univariate graphical: histograms not boxplots

G – (mathematical) model misspecified / unclear

H – clearly EXPLAIN / INTERPRET PLOTS (don't just state conclusions)

I – plot size / aspect ratio (make 'pretty') and legible - you can move them to appendix

Other:

2.5/4

logistic: Name _____ He Za 8.75/12 → 4.375/6

1. Formatting:

all margins 2.5cm

12 pt size

no raw R code or output

max **10** pages

informative title

name on all pages

all pages numbered

no blurry plots (**NOT png**)

- too many digits (p-values)

2. Introduction/Background:

brief background and statement of scientific question

all variables defined

3. EDA:

freq for all values, not just mode

univariate numerical

univariate graphical

bivariate numerical (cor)

bivariate graphical

4. Model fitting:

give mathematical definition of model

state how model fitted (ie, maximum likelihood)

CLEARLY describe how model selected be very specific

define all terms q in AIC

5. Model assessment:

CLEARLY state model assumptions: + give PRIMARY references

1. binary outcome 2. independent obs
3. linear relation between logit and linear predictor
4. no multicollinearity 5. no outliers (6. large sample size)

carry out assessment (numerical / graphics):

scatterplots of logit vs. predictors (linearity assumption)

DEFINE -> Cook's distance / standardized residuals (outliers)

vif (to identify multicollinearity)

Define

1/1

0.5/1

1.25/2

1.75/2

1.75/2

6.25/8

0.75/1

6. Write out final *estimated* model **mathematically**

hat on *response* variable max **2 sig digits** (after decimal) on coefs

7. Plots:

0.75/1

label size (not too small) informative captions

placement explanations

0.25/1

8. Conclusions

be specific

1. recap analysis +EDA 2. state and interpret main findings

0.75/1

9. Overall presentation (clarity of explanations, appropriate citations / references) :

poor satisfactory good excellent

10. Other comments:

A – no / incomplete / insufficient references

B – cite PRIMARY refs (not course notes, not wikipedia, *etc.*)

C – interpretation (cannot conclude causation, only association)

D – use your OWN WORDS / no apparently unattributed quotations

E – Intro: **1.** Give context; **2.** Clearly state scientific question; **3.** Describe data

F – univariate graphical: histograms not boxplots

G – (mathematical) model misspecified / unclear

H – clearly EXPLAIN / INTERPRET PLOTS (don't just state conclusions)

I – plot size / aspect ratio (make 'pretty')

Other: _____

_____

_____

_____

_____

2.5/4