

STAT 112 - Assessment Review

Problem one - Question one on multiple linear regression

Intercept only model with MLR

```
> m1<-lm(academicenvp~1)
> summary(m1)
```

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	60.9947	0.2973	205.2	<2e-16 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 16.24 on 2981 degrees of freedom
(2400 observations deleted due to missingness)

ANOVA

```
> m1=aov(academicenvp~1)
> summary(m1)
```

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Residuals	2981	785785	263.6		

2400 observations deleted due to missingness

MLR model with multiple predictors

```
> m2<-lm(academicenvp~friendlyenvp+FIRSTGEN+leaveUCLA*exclusionaryp)
> summary(m2)
```

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	28.76698	5.64024	5.100	5.51e-07 ***
friendlyenvp	0.48170	0.05790	8.319	1.88e-15 ***
FIRSTGENyes	2.75507	1.59366	1.729	0.0847 .
leaveUCLAYes	-9.41946	4.90059	-1.922	0.0554 .
exclusionaryp	-0.18292	0.08594	-2.128	0.0340 *
leaveUCLAYes:exclusionaryp	0.13495	0.12699	1.063	0.2886

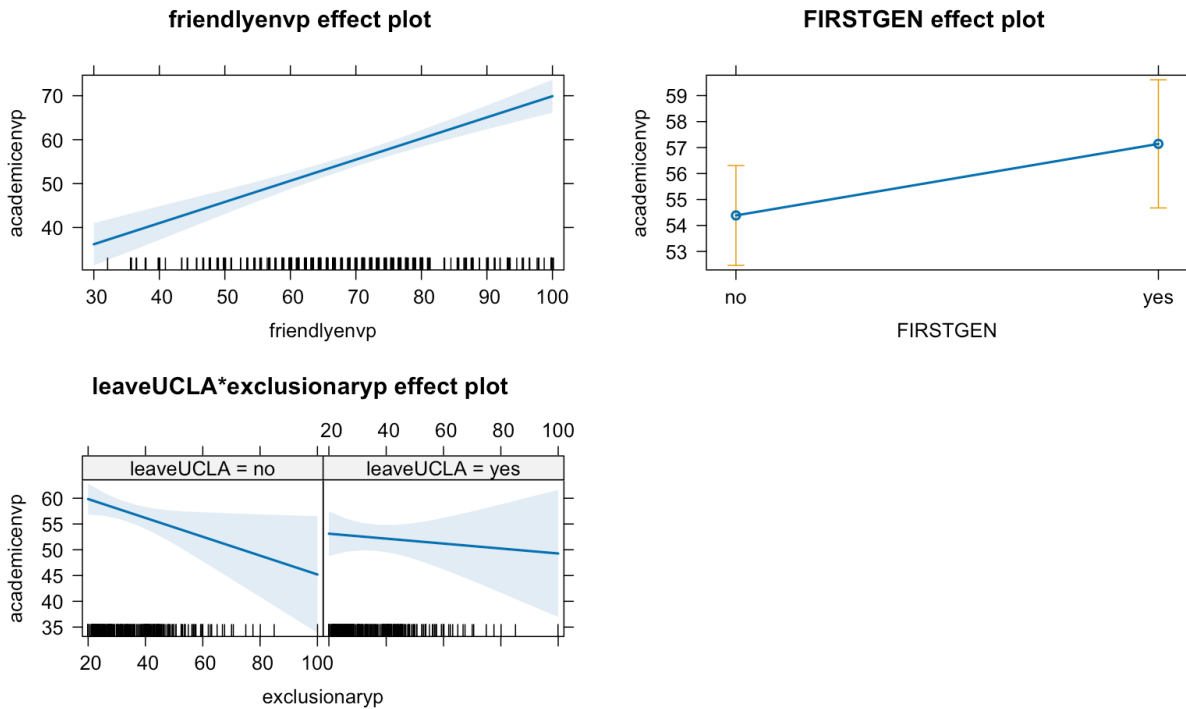
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 14.43 on 358 degrees of freedom
(5018 observations deleted due to missingness)

Multiple R-squared: 0.2416, Adjusted R-squared: 0.231

F-statistic: 22.81 on 5 and 358 DF, p-value: < 2.2e-16

Plot one



Questions to answer about problem one

- What does m1 show you?
- What additional information does the ANOVA model give you?
- Compare the intercept only model with m2, explain how things changed.
- What is residual standard error?
- Compare residual standard error and R^2 from the two models.
- Compare the degrees of freedom from the two models.
- Interpret R^2 in m2.
- Interpret the coefficients of m2 within context
- Interpret the plots given to you under plot one within context.

Problem two - Question two on logistic regression

```
> m3<-glm(leaveUCLA~1,family="binomial")
> summary(m3)
```

Coefficients:

	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	-1.62176	0.03674	-44.14	<2e-16 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 4815.7 on 5376 degrees of freedom
Residual deviance: 4815.7 on 5376 degrees of freedom
(5 observations deleted due to missingness)
AIC: 4817.7

```
> m4<-
glm(leaveUCLA~academicenvp+FIRSTGEN+exclubebehavlocationp*LowFamilyIncomeIndicator
,family="binomial")
> summary(m4)
```

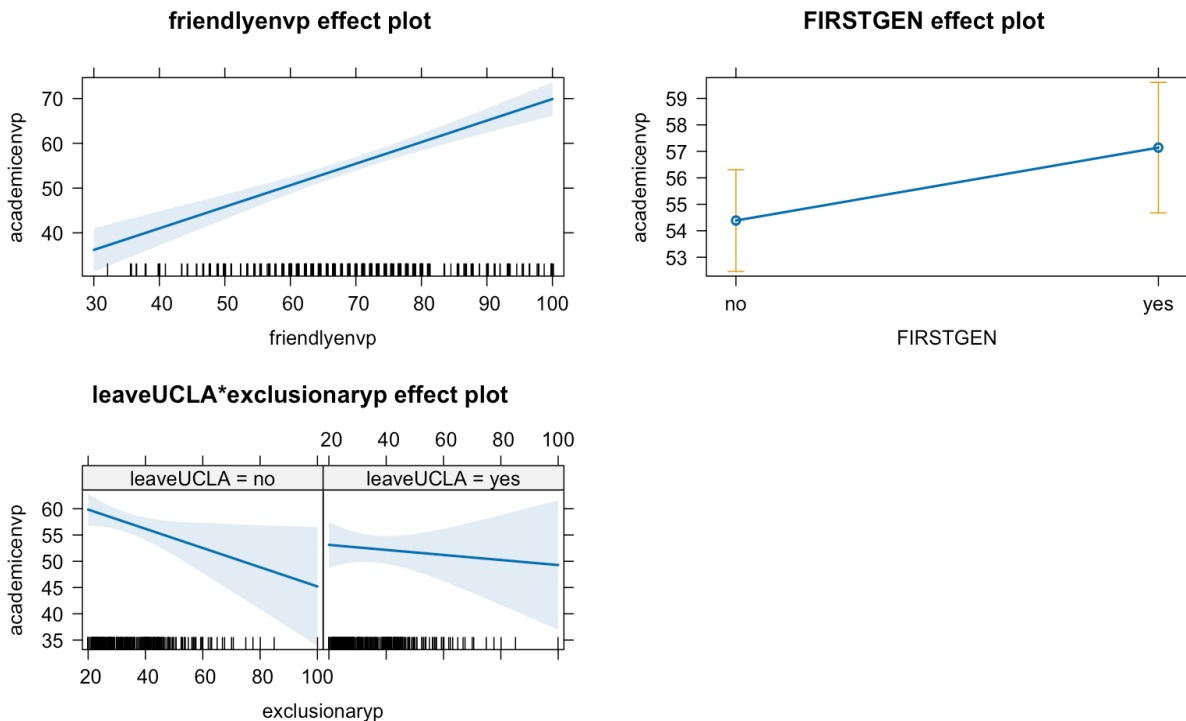
Coefficients:

	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	1.437854	0.401233	3.584	0.000339 ***
academicenvp	-0.045095	0.006084	-7.412	1.24e-13 ***
FIRSTGENyes	0.138669	0.189703	0.731	0.464791
exclubebehavlocationp	0.010421	0.008153	1.278	0.201161
LowFamilyIncomeIndicatorNot Low Income	-0.380383	0.258429	-1.472	0.141046
exclubebehavlocationp:LowFamilyIncomeIndicator Not low-income	0.002398	0.010850	0.221	0.825092

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Null deviance: 991.47 on 857 degrees of freedom
Residual deviance: 907.42 on 852 degrees of freedom
(4524 observations deleted due to missingness)
AIC: 919.42

Plot two



```
> m4<-
multinom(leaveUCLA~academicenvp+FIRSTGEN+exclubehavlocationp*LowFamilyIncomeInd
icator)
> p<-predict(m4,leaveUCLA)
> tab<-table(p,leaveUCLA)
> tab
```

	leaveUCLA	(actual)
predicted	no	yes
no	610	194
yes	21	33

Questions to answer about problem two

- In what ways are m3 and m4 different
- What is the research question we are trying to answer in m4? (Within context)
- Exponentiate and interpret the coefficients for academicenvp and firstgen.
- What was done below? Hint (-0.04509 is log of odds for academicenvp (academic confidence by our student))
 - `exp(-0.04509 * 10)`
 - `[1] 0.637`
- Why did we do it?
- Interpret 0.637 within context.
- Calculate accuracy
- Did the model do better with sensitivity or specificity and why? You want to check the definitions of these terms on google.

Problem three

Review Session: Effectiveness and Interpretation in Predictive Modeling

Scenario

You are consulting on the development of an 18-item subscale designed to measure students' perceptions of effectiveness in predictive modeling and their ability to interpret and communicate findings responsibly. Based on theory, the instrument is expected to have three dimensions (factors):

1. Model Development and Validation – evaluating assumptions, data quality, and generalizability.
2. Interpretation of Findings – understanding uncertainty, context, and theoretical alignment.
3. Communication and Application – translating results accurately for diverse audiences.

Each item is rated on a 5-point Likert scale (1 = Strongly Disagree to 5 = Strongly Agree). The survey was completed by 220 data-science students.

Survey Items

- 1. I evaluate whether model assumptions are satisfied before interpreting results.
- 2. I check prediction accuracy using multiple validation techniques.
- 3. I assess whether my model generalizes to new data.
- 4. I verify data quality before building predictive models.
- 5. I examine model diagnostics to detect potential overfitting.
- 6. I consider whether the predictors have theoretical justification.
- 7. I can clearly explain what a regression coefficient means.
- 8. I interpret statistical significance within the practical context of the study.
- 9. I consider the impact of confounding variables on interpretation.
- 10. I check whether predictor–outcome relationships make substantive sense.
- 11. I verify that results align with theoretical expectations or prior studies.
- 12. I recognize when a model has limited explanatory power.
- 13. I summarize model results in clear language for nontechnical audiences.
- 14. I communicate uncertainty in my findings (e.g., confidence intervals or prediction error).
- 15. I avoid overstating results when model fit is weak.
- 16. I use visuals to clarify predictive findings for decision makers.
- 17. I acknowledge model limitations in reports or presentations.
- 18. I collaborate with others to verify interpretation and presentation of results.

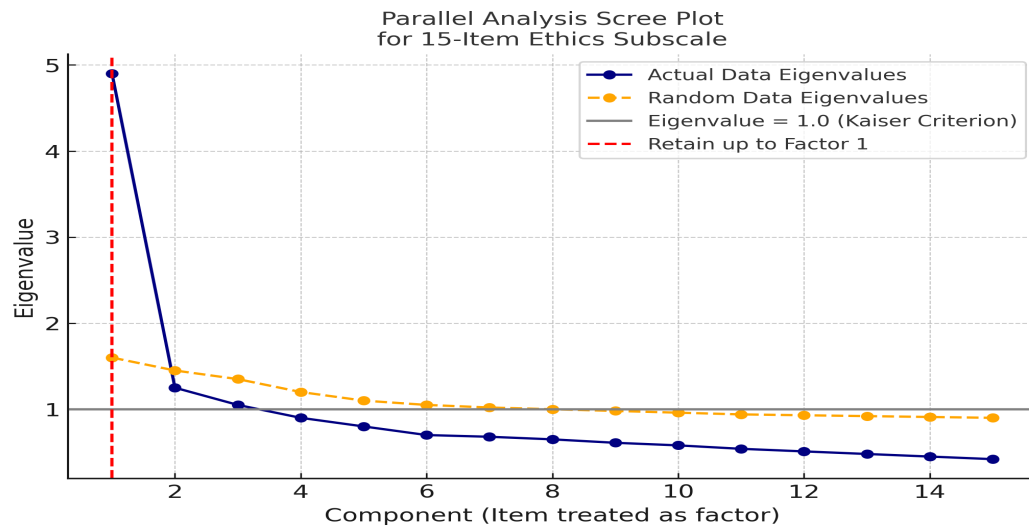
Table 1. Cronbach's Alpha if Item Deleted

Item	Cronbach's α if deleted
1	0.81
2	0.80
3	0.79
4	0.80
5	0.81
6	0.80
7	0.83
8	0.84
9	0.82
10	0.81
11	0.82
12	0.83
13	0.80
14	0.82
15	0.81
16	0.80
17	0.81
18	0.80

Table 2. Item–Total Correlations

Item	Item–Total r
1	0.48
2	0.53
3	0.51
4	0.52
5	0.55
6	0.49
7	0.28
8	0.24
9	0.26
10	0.25
11	0.29
12	0.27
13	0.56
14	0.58
15	0.60
16	0.57
17	0.55
18	0.52

Figure 1. Parallel Analysis Scree Plot



Interpretation: The first three components have eigenvalues above the random data line, suggesting a three-factor structure (Model Development, Interpretation, Communication).

Review Questions

- Q1. What does Table 1 show? Which items will you consider removing because they weaken internal consistency/reliability?
- Q2. What does Table 2 show? Which items will you consider removing and why?
- Q4. Based on your analysis of tables one and two which items will you remove?
- Q5. What does the parallel analysis suggest about the number of underlying factors?
- Q6. Explain what was done in the following plot.

