Name: Key

STAT 530 – Test 1 – Fall 2025

1. For a sample of 54 liver disease patients, several variables: (blood-clotting index, prognostic index, enzyme function, liver function, age, and survival time) were measured.
The following is the **sample covariance matrix** S for this data set.

	blood.clotting	prognostic	enzyme.function	liver.function	age	survival.time
blood.clotting	2.57	2.44	-5.10	0.86	-0.37	220.76
prognostic	2.44	285.70	-8.48	6.68	-8.96	2824.30
enzyme.function	-5.10	-8.48	451.72	9.47	-3.05	4883.67
liver.function	0.86	6.68	9.47	1.15	-2.47	286.77
age	-0.37	-8.96	-3.05	-2.47	123.71	-526.74
survival.time	220.76	2824.30	4883.67	286.77	-526.74	157915.48

(a) Based on this covariance matrix, for a patient with an **especially large enzyme function** value, which of the following is likely to be true?

(A) The patient will likely have a relatively large blood-clotting index and a relatively large liver function.

(B) The patient will likely have a relatively small blood-clotting index and a relatively large liver function.

(C) The patient will likely have a relatively large blood-clotting index and a relatively small liver function.

(D) The patient will likely have a relatively small blood-clotting index and a relatively small liver function.

(b) Based only on this output (without doing extra calculations), can you tell whether the pair of variables "blood-clotting index and survival time" has a stronger linear association than the pair "prognostic index and liver function"? Explain why or why not.

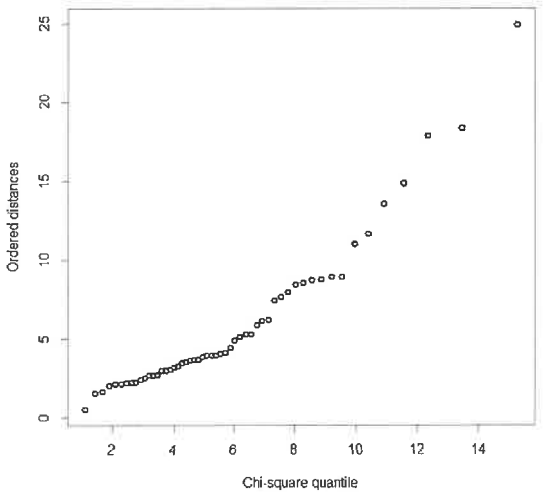
No – with covariance values, we can only tell the direction of the linear association, not the strength.

(c) If we did a principal components analysis on this data set, what is the maximum number of principal components we could obtain? Would we want to retain this maximum number of components in our analysis, and why or why not? The maximum number of PCs here would be $q=6$. But we would not want to retain all 6 PCs because we would not get any dimension reduction.

d) If we wanted to calculate the distances between pairs of observations in this data set, what initial step would it be prudent to do before calculating the distances? Briefly explain your answer.

It would be prudent to scale or standardize the variables first, because that would make all the variables equally important in the calculation of the distances.

e) The chi-square plot for this data set is given below. What is the chi-square plot checking for, and what is your conclusion about the data set? Briefly explain how you arrived at your conclusion.



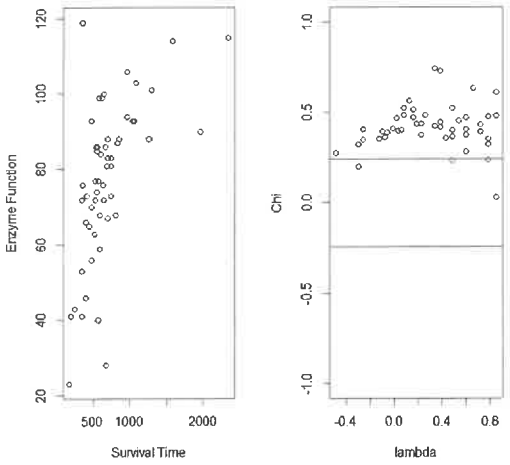
- This is checking for multivariate normality of the data.
- There is a slight bend to the chi-square plot, so perhaps the data are not perfectly multivariate normal (not too far off, though).

f) Calculate the sample correlation coefficient between blood clotting index and liver function, showing your work. Hint: $\text{corr}(X_i, X_j) = \text{cov}(X_i, X_j) / [\text{sd}(X_i)\ \text{sd}(X_j)]$. Briefly interpret what this number tells you.

$$\text{corr} = \frac{0.86}{\sqrt{2.57}\sqrt{1.15}} = 0.50$$

There is a moderately positive linear association between blood clotting index and liver function.

2. The following is the chiplot for the survival time and enzyme function variables in the data set from problem 1. What does the right-hand side of the plot tell you (be as specific as possible)? Briefly explain your answer.



- It indicates a positive association between survival time and enzyme function (not independence) since most of the points are above the central region.

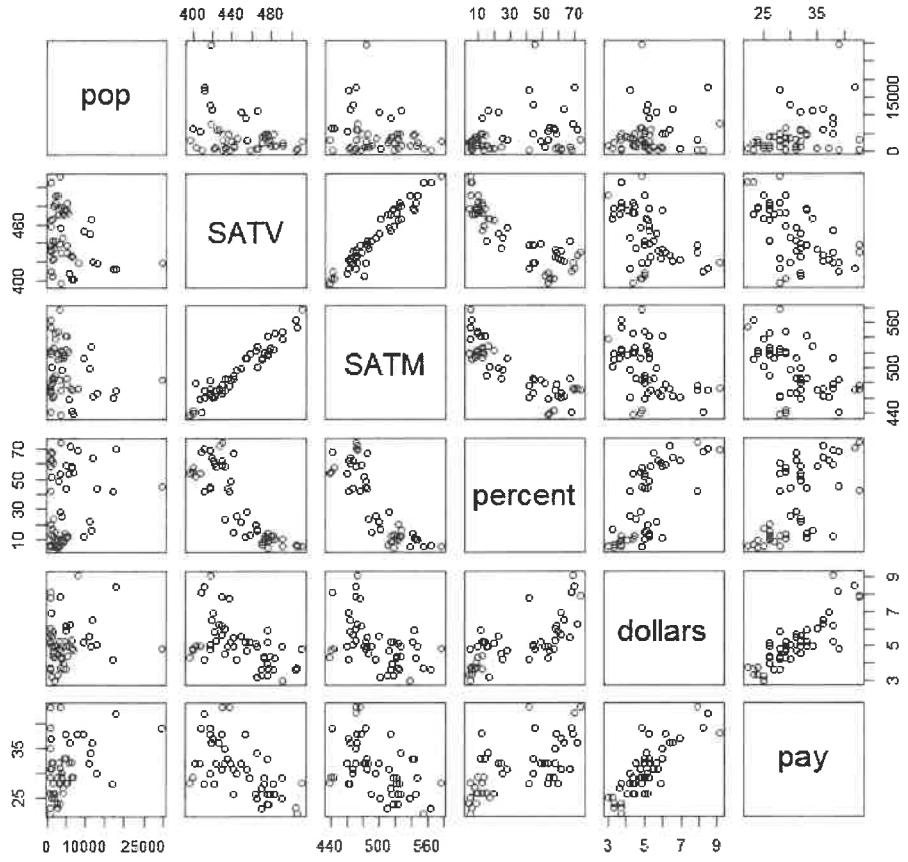
3. Which of the following types of plots or graphs allow us to visualize **three or more** numerical variables at once? Circle all that apply.

Basic bubble plot	Chiplot	Star plot	Bivariate boxplot
Chernoff Faces	Basic scatterplot	Scatterplot matrix	Radar plot

4. Consider the following educational data (originally gathered in the 1990s) on the 50 states plus the District of Columbia. The six variables are:

pop	Population (in 1,000s of people)
SATV	Average score of high-school students in state on SAT *verbal* component
SATM	Average score of high-school students in state on SAT *math* component
percent	Percentage of graduating high-school students in the state who took the SAT exam
dollars	State spending on public education (in $1000s per student)
pay	Average teacher's salary in the state (in $1000s)

State Education Data



(a) A scatterplot matrix for this data set is given on the previous page. Overall, what can be concluded about the pairwise relationships between the variables from the scatterplot matrix? Which two pairs of variables seem to have the strongest association and which two or three pairs seem to have the weakest association? Are any of the prominent associations notably nonlinear; if so, which? Most pairs of variables seem related, except pairs involving population.

- Strong associations: SATV + SATM, dollars + pay, (dollars + percent?)
- Weak associations: pop + percent, pop + dollars
- Nonlinear associations: SATV + percent, SATM + percent

A PCA was done on the data from the 51 states/districts. These are the results:

```
> summary(state.educ.pc,loadings=T)
Importance of components:
                          Comp.1    Comp.2    Comp.3     Comp.4     Comp.5      Comp.6
Standard deviation     1.9773932 0.9677285 0.9439617 0.36953606 0.31007553 0.172192438
Proportion of Variance 0.6516807 0.1560831 0.1485106 0.02275948 0.01602447 0.004941706
Cumulative Proportion  0.6516807 0.8077637 0.9562743 0.97903382 0.99505829 1.000000000

Loadings:
        Comp.1 Comp.2 Comp.3 Comp.4 Comp.5 Comp.6
pop      0.194  0.909  0.282  0.218
SATV    -0.460  0.118 -0.381  0.225 -0.260  0.715
SATM    -0.443  0.243 -0.416  0.205 -0.223 -0.693
percent  0.473 -0.208         0.534 -0.666
dollars  0.402         -0.594  0.406  0.566
pay      0.411  0.236 -0.496 -0.641 -0.342
```

(a) This PCA was done based on the correlation matrix rather than the covariance matrix. If the covariance matrix had been used, would the results be the same? If not, what is a reason for using the correlation matrix in the PCA? The results would not be the same. Using the correlation matrix allows all the variables to contribute equally to the PCA solution, regardless of units of measurement.

(b) In the best way you can, provide simple interpretations of the first three principal components.

PC 1: Measures high-spending, low-performance index (wasted money)

PC 2: Mostly a population index

PC 3: Measures low-spending + low-performance (lack of spending + lack of success)

(c) The state of South Dakota had the following (standardized) data values for the six variables: Pop = -0.769, SATV =1.877, SATM =1.666, percent = -1.194, dollars = -1.050, pay = -1.684. Showing your work, calculate the **score** for South Dakota on the **first** principal component.

$$PC1 = 0.194(-0.769) - 0.46(1.877) - 0.443(1.666) + 0.473(-1.194) + 0.402(-1.05) + 0.411(-1.684) = -3.43$$

(d) What is a reasonable number of principal components to retain in this analysis? There may be different answers that are sensible, but whatever answer you choose, justify it with some numerical reason.

Possible answers:

- 2 PCs since the first 2 PCs explain over 80% of the variance
- 3 PCs since the first 3 eigenvalues (squares of the SDs) will be close to 1.

5. Name one advantage and one limitation of an animation plot.

- Advantage: Can show changing patterns in data over time
- Limitation: Cannot be used in a print publication.

6. An important characteristic of principal components for a data set is that they

(A) are uncorrelated (B) are always positive
(C) follow a chi-square distribution (D) have variance equal to zero

7. What quantities, in a factor analysis, are unobservable characteristics of the individuals?

(A) the communalities (B) the manifest variables
(C) the sample covariance matrix (D) the latent variables

8. As part of a psychology study, volunteers were recruited to assess the subjects of headshot photographs of a sample of individuals. The raters quantified the subjects of the photos based on their perceptions of numerous characteristics (how serious, exciting, calm, independent, sincere, warm, physically attractive, sociable, kind, intelligent, strong, sophisticated, and happy the subject of the photo appeared). These ratings were variables in the data set, and the final variable was the subjects' self-ratings of their own physical attractiveness. In all, there were 114 individuals and 14 variables in the data set.

A factor analysis was conducted on this data set. Partial output from a 2-factor model and full output from a 3-factor model are given on the following page.

Partial 2-factor output:

```
Call:
factanal(x = photos.num, factors = 2, rotation = "varimax")
….
Test of the hypothesis that 2 factors are sufficient.
The chi square statistic is 92.39 on 64 degrees of freedom.
The p-value is 0.0116
```

3-factor output:

```
Call:
factanal(x = photos.num, factors = 3, rotation = "varimax")
Uniquenesses:
     Serious      exciting         calm   independent      sincere         warm      phyattr    sociable         kind
       0.930        0.817        0.899        0.769        0.877        0.514        0.005        0.831        0.005
  intelligent       strong sophisticated        happy        ownPA
       0.646        0.544        0.585        0.469        0.867

Loadings:
              Factor1 Factor2 Factor3
Serious                       0.262
exciting       0.153   0.399
calm           0.312
independent    0.363           0.315
sincere        0.287   0.189
warm           0.649   0.253
phyattr        0.256   0.956   0.126
sociable               0.227   0.329
kind           0.820   0.189  -0.535
intelligent    0.556   0.139   0.160
strong         0.464           0.485
sophisticated  0.595   0.205   0.136
happy          0.666           0.296
ownPA                  0.362

Test of the hypothesis that 3 factors are sufficient.
The chi square statistic is 46.25 on 52 degrees of freedom.
The p-value is 0.698
```

(b) Explain exactly what led the analyst to prefer the 3-factor solution to the 2-factor solution here. Also, attempt to interpret the three factors using as simple language as possible.

The P-value was nonsignificant (0.698) for the 3-factor solution but significant (0.0116) for the 2-factor solution.

- Factor 1: Measures warmth/kindness/happiness (how pleasant subject seems)
- Factor 2: Measures mainly attractiveness (beauty index)
- Factor 3: Index of strong rather than kindness (dominance index?)

(c) Briefly, what was the analyst hoping to accomplish by choosing the "varimax" method when doing this factor analysis?

- Make more interpretable factors by rotating the loadings to be closer to 0 and 1.

(d) Which property is NOT one of the properties of factor loadings that leads to a highly interpretable solution?

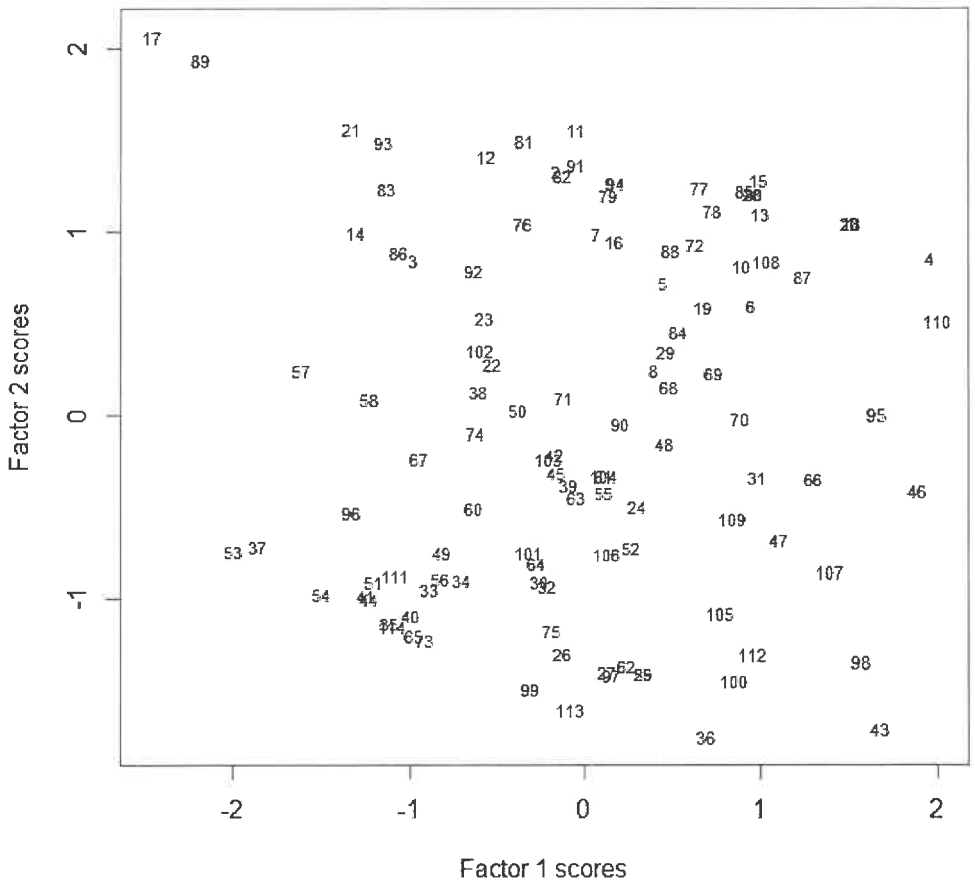
(A) We would like each variable to have a high loading for only one factor.

(B) We would like all of the loadings to have similar values to each other.

(C) We would like each factor to have high loadings on only a few variables.

(D) We would like most of the loadings to be near zero.

(e) The following is a scatterplot of the two factor scores for the subjects in this data set:

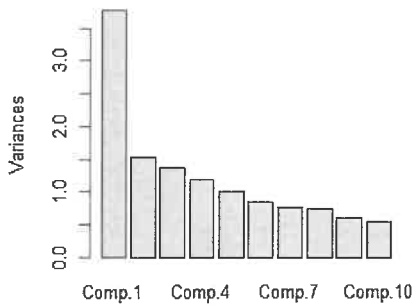


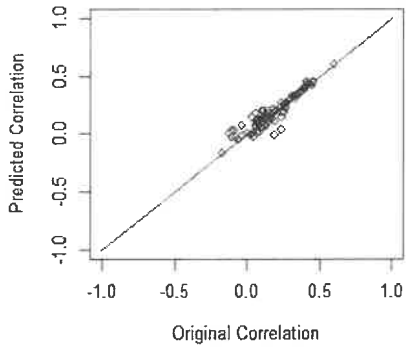
With **reference to your interpretations of the factors from part (b)**, what can you say about Subjects 17 and 89 in this data set based on the plot of factor scores? What can you say about Subject 43?

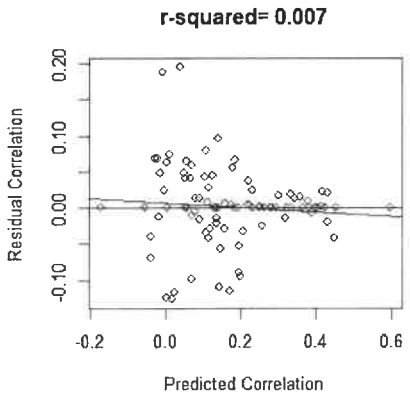
Subjects 17 and 89 may be less warm/kind looking but very physically attractive.

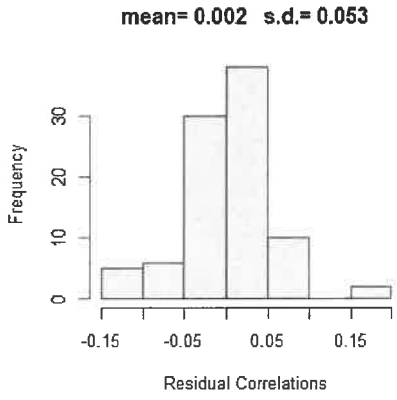
Subject 43 looks warm/kind but not as attractive.

(f) The following gives some diagnostic output for this factor analysis. Discuss the fit of the 3-factor model. Is it near-perfect, or decent but imperfect, or terrible? Justify your answer based on the diagnostic plots.









Seems decent but not perfect. The predicted correlations from the model are sort of close to the original correlations but not perfectly on the 45° line.
- Some residual correlations are a bit far from zero.

Extra credit: Which two statisticians developed a method for letting the data determine the optimal type of transformation of the variable(s) before the analysis?

Box + Cox.