Sociology 505 Winter 2023

```
#Loading required libraries
library(dplyr)
library(ggplot2)
library(modelsummary)
```

```
#importing data into R
Districtdata=read.csv("districtdata2023.csv")
```

Bivariate Models

Part 1

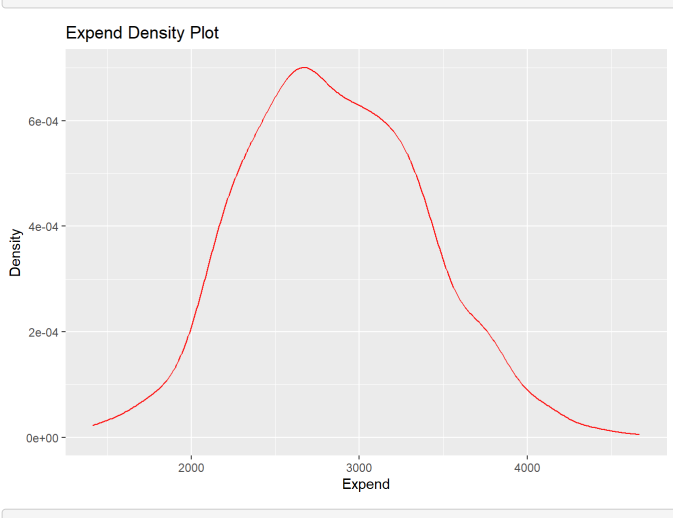
Describe the linear relationship between EXPEND and INCOME (where EXPEND is predicted by INCOME)

- a. Describe/report details about the univariate distributions of the two variables (EXPEND and INCOME); include at least on measure of central tendency and one measure of variability; produce a bivariate plot of association of EXPEND and INCOME.

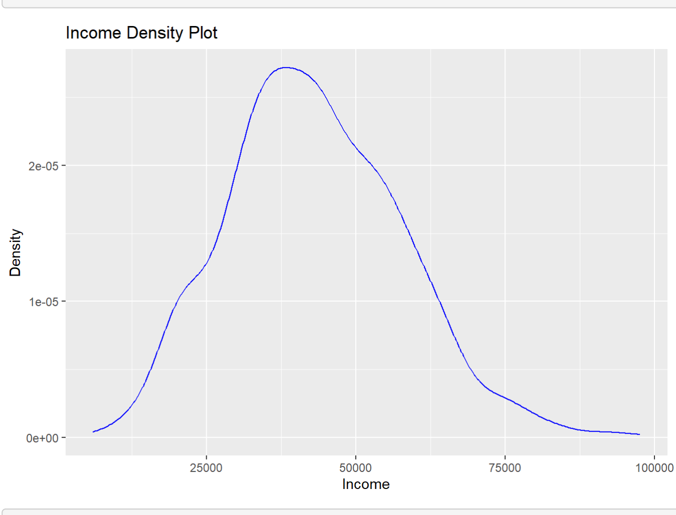
```
#summary of the data (measure of central tendancy:mean and median; measure of variablility standard deviation)
stargazer::stargazer(Districtdata[3:4],type = "text",median = T,min.max = F)
```

```
##
## ============================================
## Statistic  N     Mean     St. Dev.    Median
## --------------------------------------------
## expend    923 2,849.072   537.767   2,815.531
## income    923 43,140.210 14,491.770 42,140.880
## --------------------------------------------
```

```
#plotting distribution of the data
#Density plot for Expend
ggplot(Districtdata)+
  geom_density(aes(expend),col="red")+
  labs(title = "Expend Density Plot",x="Expend",y="Density")
```




```
#Density plot for Income
ggplot(Districtdata)+
  geom_density(aes(income),col="blue")+
  labs(title = "Income Density Plot",x="Income",y="Density")
```




```
#bivariet scatter plot of Income aginist Expend
ggplot(Districtdata,aes(x=income,y=expend))+
  geom_point(color="green")+
  labs(title = "Scatter Plot of Income Aginist Expend",
       x="Income",y="Expend")
```

Both variables are normaly distributed based on the shape of the curve of the density plots. The bivariet scatter plot indicates a positive linear relationship between income and expenditure. This means the more a person earns the more they spend.

- b. Estimate and describe the bivariate linear model relation between the two variables using ML estimation (EXPEND is the outcome variable)

```
#Creating a linear model
Model=glm(expend~income ,data = Districtdata,
          family = gaussian(link = "identity"))
coefficients(Model)
```

```
##  (Intercept)       income
## 1.546360e+03 3.019718e-02
```

$Expend = 1546.36 + 0.0302Income$

Based on the estimated model the when there is no income an individual is expected to spend 1546.36 dollars. when income increases by unit the expenditure increases by 0.0302

- c. Calculate conditional mean of expenditures based on the model if the per capita income is 25000 or if it is 35000

```
paste("2500 Income: Expend",
      round(predict(Model,newdata = data.frame(income=25000)),2))
```

```
## [1] "2500 Income: Expend 2301.29"
```

```
paste("2500 Income: Expend",
      round(predict(Model,newdata = data.frame(income=35000)),2))
```

```
## [1] "2500 Income: Expend 2603.26"
```

Expend 2301.29 Expend 2603.26

- d. look at the data and calculate the raw response residual/deviation for CountyID #73600 and CountyID #41500. What does the deviation tell you about school expenditure in County 73600 and 41500 with respect to the estimated equation?

```
data=data.frame(CountyID=Districtdata$countyid,Residual=residuals(Model))
data[data$CountyID==73600,]
```

	CountyID <int>	Residual <dbl>
736	73600	-315.55

1 row

```
data[data$CountyID==41500,]
```

	CountyID <int>	Residual <dbl>
415	41500	635.8991

1 row

73600ID:315.55 41500ID: 635.9

- e. Calculate the median value of income and run the same model from (b) but select only those cases that are below the median value and again for only those observed cases above the median. What do these two results suggest about the overall equation or model you calculated on the full sample—e.g. how appropriate the systematic part of the model is based on this information? In looking at your plot from part(a) above describe why your results from this exercise are or are not surprising.

```
#values delow the median income
below_income_median=subset(Districtdata,income<median(Districtdata$income))
#values above the median income
above_income_median=subset(Districtdata,income>median(Districtdata$income))

Model1=glm(expend~income,data = below_income_median,
           family = gaussian(link = "identity"))
Model2=glm(expend~income,data = above_income_median,
           family = gaussian(link = "identity"))

modelsummary(list(Model1,Model2))
```

	(1)	(2)
(Intercept)	1551.715	1555.357
	(62.147)	(84.435)
income	0.030	0.030
	(0.002)	(0.002)
Num.Obs.	461	461
R2	0.348	0.460
AIC	6596.3	6624.9
BIC	6608.7	6637.3
Log.Lik.	−3295.164	−3309.448
F	244.993	391.625
RMSE	307.64	317.32

Both the data above and below the median show a linear positive relationship between the variale income and expend

- f. do the same exercise but select if EXPEND is less than EXPEND's median of 2800? Compare differences or similarities in the f's results to (a) and to (e). Explain any differences you might observe.

```
#values delow the median expend
below_expend_median=subset(Districtdata,expend<median(Districtdata$expend))
#values above the median expend
above_expend_median=subset(Districtdata,expend>median(Districtdata$expend))

Model3=glm(expend~income,data = below_expend_median,
           family = gaussian(link = "identity"))
Model4=glm(expend~income,data = above_expend_median,
           family = gaussian(link = "identity"))

modelsummary(list(Model3,Model4))
```

	(1)	(2)
(Intercept)	1858.508	2242.799
	(38.890)	(52.488)
income	0.016	0.020
	(0.001)	(0.001)
Num.Obs.	461	461
R2	0.325	0.474
AIC	6351.6	6395.9
BIC	6364.0	6408.3
Log.Lik.	−3172.786	−3194.970
F	221.490	414.028
RMSE	235.91	247.54

- g. estimate a model in which you predict "expend" simply by a best fitting constant and describe this model; what features of your model and information from " b " above makes you think your model in "b" is better than your model in "g" (or is it???). (Just some intuition as to why you think "a" or "b" is better).

```
Model5=glm(expend~1,data = above_expend_median,
           family = gaussian(link = "identity"))

modelsummary(list(Model,Model5))
```

	(1)	(2)
(Intercept)	1546.360	3284.610
	(32.341)	(15.918)
income	0.030	
	(0.001)	
Num.Obs.	923	461
R2	0.662	0.000
AIC	13229.2	6690.3
BIC	13243.7	6698.6
Log.Lik.	−6611.611	−3343.163
F	1805.442	
RMSE	312.39	341.40

The model in "b" is considered better because it uses multiple predictor variables to predict the response variable and it fits the data better than the model in "g" which is a simple constant model based on the R squared.

Multivariate Models

- a. describe the relations of COLGRAD and INCOME to EXPEND from the results of this new model.

```
#Multivariate model
Model6=glm(expend~income+colgrad,data = Districtdata,
           family = gaussian(link = "identity"))
#model summary
stargazer::stargazer(Model6,type = "text")
```

```
##
## ============================================
##                      Dependent variable:
##                  ---------------------------
##                            expend
## --------------------------------------------
## income                    0.025***
##                           (0.001)
##
## colgrad                   16.693***
##                           (1.633)
##
## Constant                1,482.896***
##                          (31.285)
##
## --------------------------------------------
## Observations                923
## Log Likelihood           -6,562.940
## Akaike Inf. Crit.        13,131.880
## ============================================
## Note:            *p<0.1; **p<0.05; ***p<0.01
```

$Expend = 1482.9 + 0.025Income + 16.693ColGrad$

Keeping percentage of college graduate and income constant the expected expenditure is 1482.9. When income increases by a unit the expected expenditure increases by 0.025 and when percentage of college graduate between age of 25-39 increase by 1% the expeniture increases by 16.7.

- b. explain why there is a difference between the effect of income on expenditure in Part 1 "b" above and what you see now in Problem 1.2.

```
#Comparison of the two models
modelsummary(list(Model,Model6))
```

	(1)	(2)
(Intercept)	1546.360	1482.896
	(32.341)	(31.285)
income	0.030	0.025
	(0.001)	(0.001)
colgrad		16.693
		(1.633)
Num.Obs.	923	923
R2	0.662	0.697
AIC	13229.2	13131.9
BIC	13243.7	13151.2
Log.Lik.	−6611.611	−6561.940
F	1805.442	1056.482
RMSE	312.39	296.02

Based on the models the model includeing college graduate is better based on the R squared. The effect of income on expenditure may be moderated or controlled by the level of college graduates in the population. This means that the relationship between income and expenditure may differ depending on the percentage of college graduates in the population. The effect of income may be stronger or weaker in areas with higher or lower percentages of college graduates, respectively

- c. standardize the two partial derivatives for the two independent variables and interpret their effects

```
#model with sacaled predictors
Model7=glm(expend~scale(income)+scale(colgrad),data = Districtdata,family = gaussian(link = "identity"))
modelsummary(list(Model6,Model7))
```

	(1)	(2)
(Intercept)	1482.896	2849.072
	(31.285)	(9.759)
income	0.025	
	(0.001)	
colgrad	16.693	
	(1.633)	
scale(income)		359.510
		(12.397)
scale(colgrad)		126.760
		(12.397)
Num.Obs.	923	923
R2	0.697	
AIC	13131.9	13131.9
BIC	13151.2	13151.2
Log.Lik.	−6561.940	−6561.940
F	1056.482	1056.482
RMSE	296.02	296.02

The estimated coefficients of the scaled predictors increasesd significantly and income seems to be a better significant contributer to increase in expenditure compared to percentage of college garaduates

- d. THOUGHT QUESTION: If there were basically no change in the magnitude of the coefficient for income what would you know or conclude about the relation between INCOME and COLGRAD?

It would suggest that the relationship between Income and Expend is not affected by the level of Colgrad in the population. It could also suggest that the relationship between Income and Expend is not affected by the level of COLGRAD in the population, or that the relationship between Income and Colgrad is weak or non-existent

- e. THOUGHT or DO QUESTION: If we calculated INCOME as deviation from its mean (i.e. observed income minus mean of income) for each case what would change in your model in "1.2 a" (if anything) and why?

```
#Deviation of income from its mean
Deviation_income_mean=Districtdata$income-mean(Districtdata$income)
#Model using income deviation from mean
Model8=glm(expend~Deviation_income_mean,data = Districtdata,
           family = gaussian(link = "identity"))
#model comparison
modelsummary(list(Model6,Model8))
```

	(1)	(2)
(Intercept)	1482.896	2849.072
	(31.285)	(10.293)
income	0.025	
	(0.001)	
colgrad	16.693	
	(1.633)	
Deviation_income_mean		0.030
		(0.001)
Num.Obs.	923	923
R2	0.697	0.662
AIC	13131.9	13229.2
BIC	13151.2	13243.7
Log.Lik.	−6561.940	−6611.611
F	1056.482	1805.442
RMSE	296.02	312.39

There is a slight difference(0.005) in the coefficent of income and deviation of income. This change might be due to the fact that the coefficient for Income deviation from its mean would represent the change in Expend for a one unit change in Income deviation from the mean and not income in its absolute form.

- f. calculate/describe the variance/covariance matrix for the 3 variables in this problem; also the standardized variance/covariance matrix for the same 3 variables.

```
# variance and covariance of collegrad, expend and income
cov(Districtdata[-2])
```

```
##             colgrad      expend       income
## colgrad    57.66116    2644.553     67801.27
## expend   2644.55253  289193.323   6341750.87
## income  67801.27451 6341750.872 210011359.36
```

```
#variance and covariance of standerdized collegrad, expend and income
cov(scale(Districtdata[-2]))
```

```
##           colgrad    expend    income
## colgrad 1.0000000 0.6476142 0.6161335
## expend  0.6476142 1.0000000 0.8137550
## income  0.6161335 0.8137550 1.0000000
```

Multiple Regression Equation

HOPELESS=f(AGE,CPBONDS,CABONDS,FDISTRES) Where age is in years, cpbonds and cabonds are in counts of people, and family distress is a set of items in a Likert scale from 0 to 6 where 0 is no/low distress and 6 is high distress (note: you can talk about "hopeless units" and "distress units" as continuous scales). Hopeless is a continuous scale ranges from 10 – 30 and is the weighted average of a set of Likert items; higher scores = more hopelessness; one can think of the scale as "hopeless units". Hopelessness scale = linear additive model with independent variables AGE, # of positive supportive peer friends (CPBONDS), # of nonfamily adults(e.g. teachers– CABONDS) who are seen as supportive, and family distress 10pt Likert scale (FDISTRES))

```
AdolMentalHealth2=read.csv("adolmentalhealth2023.csv")
head(AdolMentalHealth2)
```

	risk <chr>	ethnic <chr>	a… <int>	sex <chr>	naturl <chr>	single <chr>	depress <dbl>	anger <dbl>
1	low risk	hispanic, latino	15	0	family of orgin	not single	2.1666667	1.50
2	low risk	asian, pacific islander	17	male	family of orgin	not single	0.6666667	0.50
3	low risk	black, african-american	17	male	family of orgin	not single	1.3333334	0.50
4	low risk	black, african-american	16	male	family of orgin	not single	2.6666667	1.75
5	low risk	caucasian, white, euro-american	16	0	family of orgin	not single	1.0000000	2.25
6	low risk	caucasian, white, euro-american	17	0	family of orgin	not single	1.8333334	3.50

6 rows | 1-9 of 24 columns

```
#Subset of the data
Data=AdolMentalHealth2 %>% select(hopeless,age,cabonds,cpbonds,fdistres)

Model=glm(hopeless~.,data = Data,family = gaussian(link = "identity"))
modelsummary(Model)
```

	(1)
(Intercept)	21.646
	(1.044)
age	0.000
	(0.064)
cabonds	−0.047
	(0.051)
cpbonds	−0.145
	(0.062)
fdistres	0.359
	(0.065)
Num.Obs.	568
R2	0.073
AIC	2147.6
BIC	2173.7
Log.Lik.	−1067.802
F	11.068
RMSE	1.59

a. estimate and interpret the point estimate results from the above multivariate model. Make sure you evaluate the range of magnitude of the marginal effects for each variable and discuss your sense of the likely direction of each effect.

- The estimate for the intercept is 21.65, which represents the expected value of hopeless when all the independent variables are equal to zero. The estimate for age is -0.000107, which means that for every one-year increase in age, the hopelessness score is expected to decrease by 0.000107 units.
- The estimate for cabonds is -0.0467, which means that for every one unit increase in the number of supportive non-family adults, the hopelessness score is expected to decrease by 0.0467 units.
- The estimate for cpbonds is -0.1446, which means that for every one unit increase in the number of positive supportive peer friends, the hopelessness score is expected to decrease by 0.1446 units.
- The estimate for fdistres is 0.3589, which means that for every one unit increase in family distress, the hopelessness score is expected to increase by 0.3589 units.

The range of magnitude of the marginal effects for each variable is relatively small, with the largest effect being seen for fdistres and the smallest for age. However, the direction of each effect is consistent, with a decrease in the number of supportive people and an increase in family distress leading to an increase in hopelessness, and an increase in the number of supportive people and a decrease in family distress leading to a decrease in hopelessness.

b. comment on whether "positive" peers (cpbonds) or number of supportive adults (cabonds) appears more important to an individual's level of hopelessness.

From the model, it appears that the number of positive supportive peer friends (cpbonds) has a stronger effect on an individual's level of hopelessness than the number of supportive non-family adults (cabonds). The estimate for cpbonds is -0.1446, while the estimate for cabonds is -0.0467. This means that for every one unit increase in the number of positive supportive peer friends, the hopelessness score is expected to decrease by 0.1446 units, while for every one unit increase in the number of supportive non-family adults, the hopelessness score is expected to decrease by only 0.0467 units.

c. What is the partial derivative of change in hopelessness with respect to #of positive adult bonds in the 4 variable model.

The partial derivative of change in hopelessness with respect to #of positive adult bonds in the 4 variable model would be -0.1446. The partial derivative is the rate of change of a function with respect to one of its variables while holding the other variables constant. In this case, it represents the change in hopelessness for a unit change in the number of positive adult bonds, with all other variables held constant.

d. compare a model with only age and family distress (call that Model 1) to the model with these two variables plus the two variables counting supportive people (peers and adults) in the model (call that Model 2)…discuss intuitively how you might consider that adding peer information is important to understanding/predicting hopelessness… (i.e. what might be evidence that suggests improvement in the model) (I am looking for both a statistical assessment and also an intuitive feel)

```
Model1=glm(hopeless~age+fdistres,data = Data,family = gaussian(link = "identity"))
modelsummary(list(Model1,Model))
```

	(1)	(2)
(Intercept)	20.969	21.646
	(1.017)	(1.044)
age	−0.003	0.000
	(0.064)	(0.064)
fdistres	0.383	0.359
	(0.063)	(0.065)
cabonds		−0.047
		(0.051)
cpbonds		−0.145
		(0.062)
Num.Obs.	568	568
R2	0.061	0.073
AIC	2150.9	2147.6
BIC	2168.3	2173.7
Log.Lik.	−1071.455	−1067.802
F	18.320	11.068
RMSE	1.60	1.59

When comparing Model 1 (with only age and family distress) to Model 2 (with age, family distress, and the two variables counting supportive people), we can assess the improvement in the model by comparing the residual deviance and AIC values. A lower residual deviance and AIC (2147.6) value in Model 2 than Model 1 (2150.9) would suggest that adding the information about supportive people improves the model's ability to explain the variation in hopelessness. Additionally, we can also look at the coefficients of Model 2 and compare the magnitude of the effect of the new variables (counting supportive people) on hopelessness with the effects of age and family distress. If the new variables have a large and significant effect on hopelessness, it would suggest that adding that information is important in understanding and predicting hopelessness.

An intuitive feel of this is, having a positive peer or supportive adult can have a positive impact on an individual's mental well-being and can help them to cope with the stressors in life, such as family distress. By including these variables in the model, we can better understand the role of social support in predicting hopelessness and can identify individuals who may be at risk due to lack of social support.

e. THOUGHT QUESTION: How might you test a model that says the absolute effect of positive peers and supportive adults have the same effect on hopelessness. Is this a legitimate question? (Hint: use algebra and set their effects to be the same in magnitude). Bonus: if you estimate such a model (kudos if you get it correct and no harm for trying if you get it wrong)

To test a model that says the absolute effect of positive peers and supportive adults have the same effect on hopelessness, one could set their effects to be the same in magnitude in the model. For example, if the current model is: hopeless = b0 + b1age + b2cabonds + b3cpbonds + b4fdistres, then we can test the new model as: hopeless = b0 + b1age + b2cabonds + b3cpbonds + b4fdistres + b5*(cabonds - cpbonds). where b5 is the new coefficient representing the difference between the effects of cabonds and cpbonds. The new model would test the hypothesis that the effects of cabonds and cpbonds are equal in magnitude. It is a legitimate question as it can help to understand the relative strength of the two predictor variables, and compare the effect of cabonds and cpbonds on hopelessness.

f. THOUGHT QUESTION: Would it be reasonable to do the same in comparing the effects of FAMILY DISTRESS and number of POSITIVE PEERS? Yes, No, Maybe… and why.

It would not be reasonable to do the same in comparing the effects of FAMILY DISTRESS and number of POSITIVE PEERS. The family distress and the number of positive peers are two different factors that may affect hopelessness in different ways. Family distress may be a more severe and long-term stressor that can have a more profound impact on hopelessness, while the number of positive peers may be more related to social support and a sense of belonging, which can have a more protective effect on hopelessness. Therefore, comparing the effects of these two variables may not be meaningful and may lead to a misinterpretation of the results.