Basics with SLR Tutorial

For this tutorial, we will work with the dataset `elmhurst` from the `openintro` package in R.

```
library(tidyverse)
library(openintro)
Data<-openintro::elmhurst
```

Type `?openintro::elmhurst` to read the documentation for datasets in R. Always seek to understand the background of your data! The key pieces of information are:

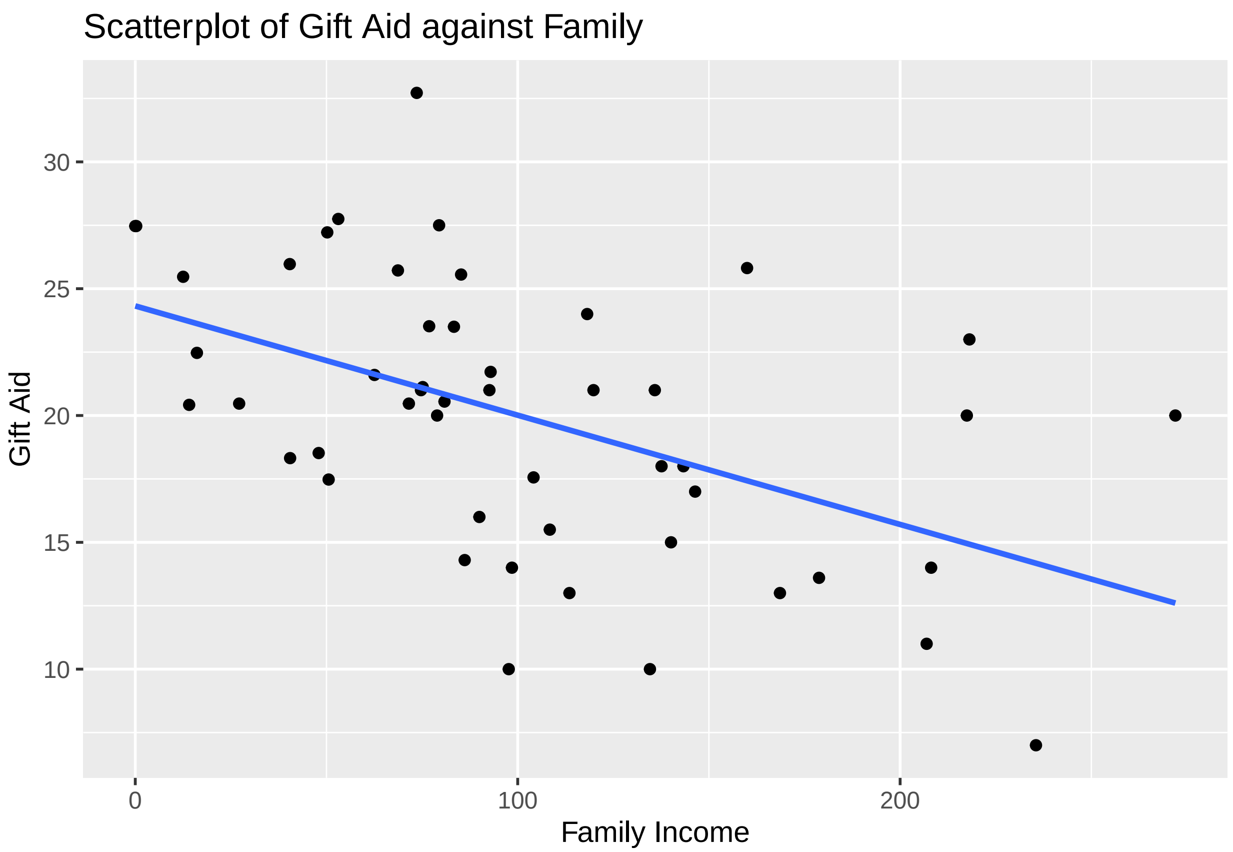
- A random sample of 50 students (all freshman from the 2011 class at Elmhurst College).
- Family income of the student (units are missing).
- Gift aid, in $1000s.

We want to explore how family income may be related to gift aid, in a simple linear regression framework.

Visualization

We should always verify with scatterplot that the relationship is (approximately) linear before proceeding with correlation and simple linear regression!

```
##scatterplot of gift aid against family income
ggplot2::ggplot(Data, aes(x=family_income,y=gift_aid))+
  geom_point()+
  geom_smooth(method = "lm", se=FALSE)+
  labs(x="Family Income", y="Gift Aid", title="Scatterplot of Gift Aid against Family")
```



We note that the observations are fairly evenly scattered on both sides of the regression line, so a linear association exists. We see a negative linear association. As family income increases, the gift aid, on average, decreases.

We also do not see any observation with weird values that may warrant further investigation.

Regression

We use the `lm()` function to fit a regression model:

```
##regress gift aid against family income
result<-lm(gift_aid~family_income, data=Data)
```

Use the `summary()` function to display relevant information from this regression:

```
##look at information regarding regression
summary(result)
```

```
##
## Call:
## lm(formula = gift_aid ~ family_income, data = Data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -10.1128  -3.6234  -0.2161   3.1587  11.5707
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)   24.31933    1.29145  18.831  < 2e-16 ***
## family_income -0.04307    0.01081  -3.985 0.000229 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.783 on 48 degrees of freedom
## Multiple R-squared:  0.2486, Adjusted R-squared:  0.2329
## F-statistic: 15.88 on 1 and 48 DF,  p-value: 0.0002289
```

We see the following values:

- $\hat{\beta}_1$ = -0.0430717. The estimated slope informs us the the predicted gift aid decreases by 0.0430717 thousands of dollars (or \$43.07) per unit increase in family income.
- $\hat{\beta}_0$ = 24.319329. For students with no family income, their predicted gift aid is \$24 319.33. Note: from the scatterplot, we have an observation with 0 family income. We must be careful in not extrapolating when making predictions with our regression. We should only make predictions for family incomes between the minimum and maximum values of family incomes in our data.
- s = 4.7825989, is the estimate of the standard deviation of the error terms. This is reported as residual standard error in R. Squaring this gives the estimated variance.
- F = 15.8772043. This is the value of the ANOVA F statistic. The corresponding p-value is reported. Since this p-value is very small, we reject the null hypothesis. The data support the claim that there is a linear association between gift aid and family income.
- R^2 = 0.2485582. The coefficient of determination informs us that about 24.86% of the variation in gift aid can be explained by family income.

Extract values from R objects

We can actually extract these values that are being reported from `summary(result)`. To see what can be extracted from an R object, use the `names()` function:

```
##see what can be extracted from summary(result)
names(summary(result))
```

```
##  [1] "call"          "terms"         "residuals"     "coefficients"
##  [5] "aliased"       "sigma"         "df"            "r.squared"
##  [9] "adj.r.squared" "fstatistic"    "cov.unscaled"
```

To extract the estimated coefficients:

```
##extract coefficients
summary(result)$coefficients
```

```
##                  Estimate Std. Error   t value     Pr(>|t|)
## (Intercept)   24.31932901 1.29145027 18.831022 8.281020e-24
## family_income -0.04307165 0.01080947 -3.984621 2.288734e-04
```

Notice the information is presented in a table. To extract a specific value, we can specify the row and column indices:

```
##extract slope
summary(result)$coefficients[2,1]
```

```
## [1] -0.04307165
```

```
##extract intercept
summary(result)$coefficients[1,1]
```

```
## [1] 24.31933
```

On your own, extract the values of the residual standard error, the ANOVA F statistic, and R^2.

Prediction

A use of regression models is prediction. Suppose I want to predict the gift aid of a student with family income of 50 thousand dollars (assuming the unit is in thousands of dollars). We use the `predict()` function:

```
##create data point for prediction
newdata<-data.frame(family_income=50)
##predicted gift aid when x=50
predict(result,newdata)
```

```
##        1
## 22.16575
```

This student's predicted gift aid is $22 165.75. Alternatively, you could have calculated this by plugging $x = 50$ into the estimated SLR equation:

```
summary(result)$coefficients[1,1] + summary(result)$coefficients[2,1]*50
```

```
## [1] 22.16575
```

ANOVA table

We use the `anova()` function to display the ANOVA table

```
anova.tab<-anova(result)
anova.tab
```

```
## Analysis of Variance Table
##
## Response: gift_aid
##               Df  Sum Sq Mean Sq F value    Pr(>F)
## family_income  1  363.16  363.16  15.877 0.0002289 ***
## Residuals     48 1097.92   22.87
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The report F statistic is the same as the value reported earlier from `summary(result)`.

The first line of the output gives SS_R, the second line gives SS_{res}. The function doesn't provide SS_T, but we know that $SS_T = SS_R + SS_{res}$.

Again, to see what can be extracted from `anova.tab`:

```
names(anova.tab)
```

```
## [1] "Df"      "Sum Sq"  "Mean Sq" "F value" "Pr(>F)"
```

So SS_T can be easily calculated:

```
SST<-sum(anova.tab$"Sum Sq")
SST
```

```
## [1] 1461.079
```

The R^2 was reported to be 0.2485582. To verify using the ANOVA table:

```
anova.tab$"Sum Sq"[1]/SST
```

```
## [1] 0.2485582
```

Correlation

We use the `cor()` function to find the correlation between two quantitative variables:

```
##correlation
cor(Data$family_income,Data$gift_aid)
```

```
## [1] -0.4985561
```

The correlation is -0.4985561. We have a moderate, negative linear association between family income and gift aid.