Food Expenditures

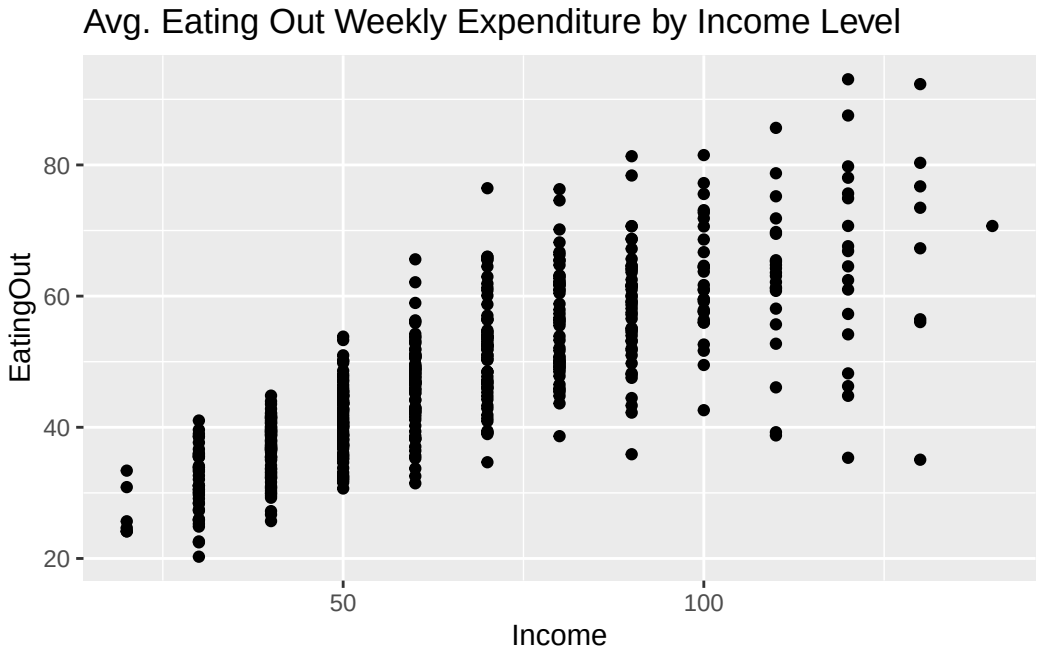
Braden Critchfield and Thomas Olsen

2024-02-16

```
library(tidyverse)
library(lmtest)
library(nlme)
library(multcomp)
library(GGally)
library(MASS)
library(car)
source("predictgls.R")


food <- read_delim("FoodExpenses.txt", delim=" ")
```

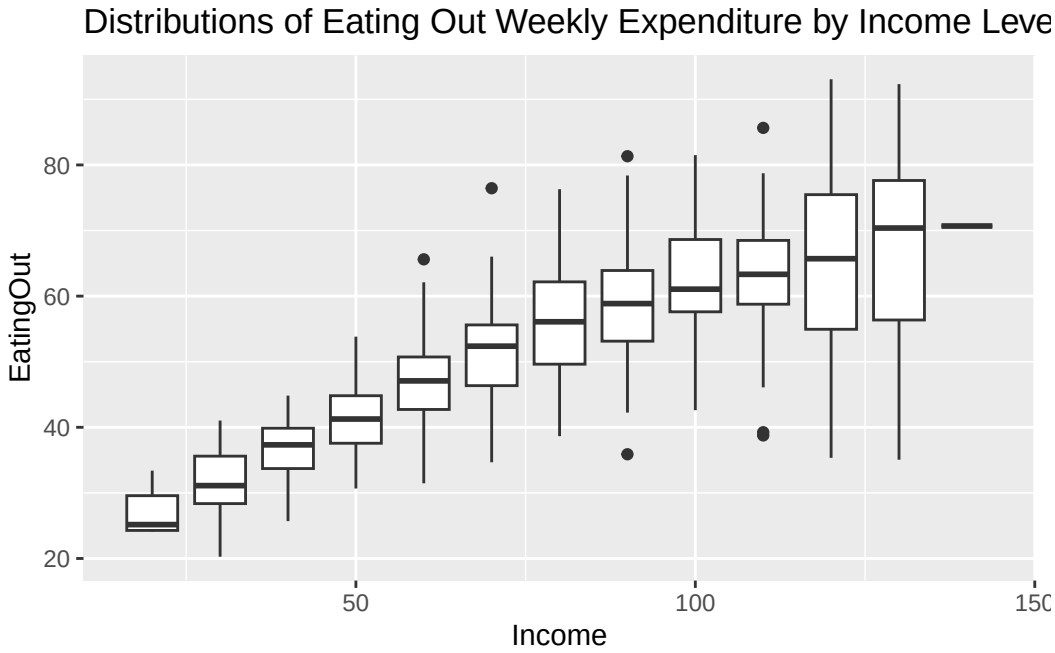
1: Exploratory Plots and Summary Statistics

```
ggplot(data = food, aes(y = EatingOut, x = Income)) +
  geom_point() +
  ggtitle("Avg. Eating Out Weekly Expenditure by Income Level")
```



There seems to be a positive linear relationship between annual household income and average weekly expenditure on food not cooked at home. However, it looks like the variance might grow larger as Income grows larger.

```
ggplot(data = food, aes(y = EatingOut, x = Income, group=Income)) +
  geom_boxplot() +
  ggtitle("Distributions of Eating Out Weekly Expenditure by Income Level")
```



Because Income is a discrete variable, this shows the distributions of Average weekly eating out expenditure per income level. This still shows a positive relationship between the two variables, and again shows how the variance seems to grow as the income increases.

Here are the summary statistics:

```
summary(food)
```

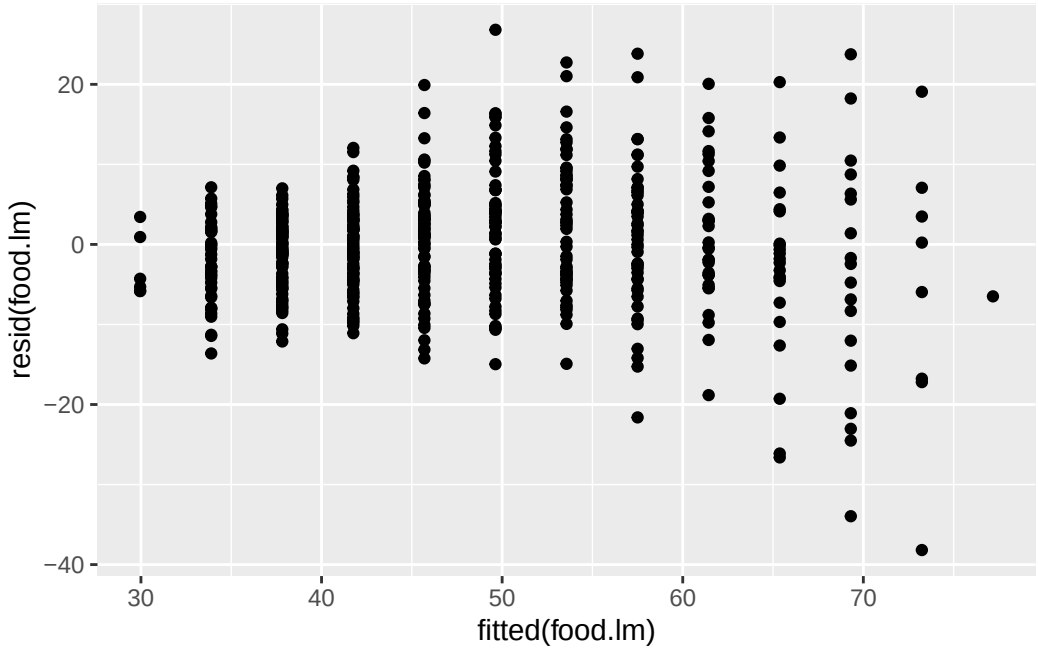
```
     Income         EatingOut
 Min.   : 20.00   Min.   :20.27
 1st Qu.: 45.00   1st Qu.:38.65
 Median : 60.00   Median :46.53
 Mean   : 65.97   Mean   :48.04
 3rd Qu.: 80.00   3rd Qu.:56.40
 Max.   :140.00   Max.   :93.06
```

2: Fit homoskedastic linear model

```
food.lm <- lm(EatingOut ~ ., data = food)
summary(food.lm)
```

```
Call:
lm(formula = EatingOut ~ ., data = food)

Residuals:
    Min      1Q  Median      3Q     Max
-38.182  -4.364   0.154   4.068  26.819

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 22.08512    0.94449   23.38   <2e-16 ***
Income       0.39351    0.01333   29.52   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 7.88 on 521 degrees of freedom
Multiple R-squared:  0.6258,    Adjusted R-squared:  0.6251
F-statistic: 871.3 on 1 and 521 DF,  p-value: < 2.2e-16
```

```
ggplot()+geom_point(mapping=aes(x=fitted(food.lm), y =  resid(food.lm)))
```




```
##bp test
bptest(food.lm)
```

```
    studentized Breusch-Pagan test

data:  food.lm
BP = 66.388, df = 1, p-value = 3.704e-16
```

The equal variance assumption is not met. From the residuals vs. fitted values plot, the spread of the data points grows as income grows. This is confirmed by the Breusch-Pagan test, where we can reject the null hypothesis that there is equal variance and assume the variance is unequal.

Since this assumption is not met, it affects our inference because all of our standard errors will be off, and some of our coefficients could be off as well. This could lead to us concluding variables are significant when they are not, identifying confidence intervals that are wrong, etc.

3: Write down heteroskedastic model

The heteroskedastic linear regression model is written as follows:

$$y \sim N(X\beta, \sigma^2 D(\theta)) \quad d_{ii} = exp(2Income_i\theta)$$

where

$$\mathbf{y} = \begin{bmatrix} EatingOut_1 \\ EatingOut_2 \\ \vdots \\ EatingOut_n \end{bmatrix}$$

$$\mathbf{X} = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix}$$

$$\beta = \begin{bmatrix} \beta_0 \\ \beta_{Income} \end{bmatrix}$$

With this model, we will have an accurate estimation of standard error, so we can test whether income has a significant effect on eating out expenditure, as well as build a confidence interval to see what the estimated effect is.

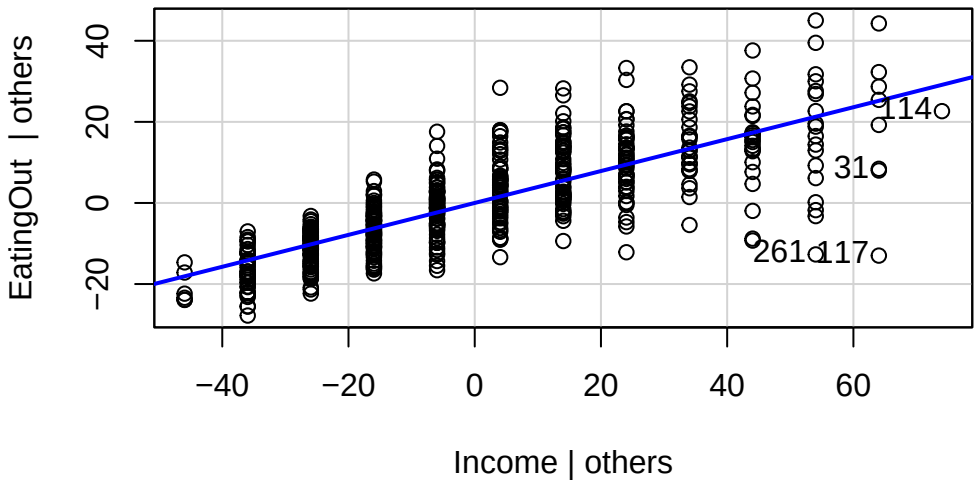
4: Fit Model from Q3 to Eating Out and check assumptions

```
hetero_sked = gls(data = food, EatingOut ~ Income, weights =
                    varExp(form=~Income), method="ML")
coef = hetero_sked$coefficients
```

Linearity

The avplot shows a linear relationship between eating out and income.

```
avPlots(food.lm)
```



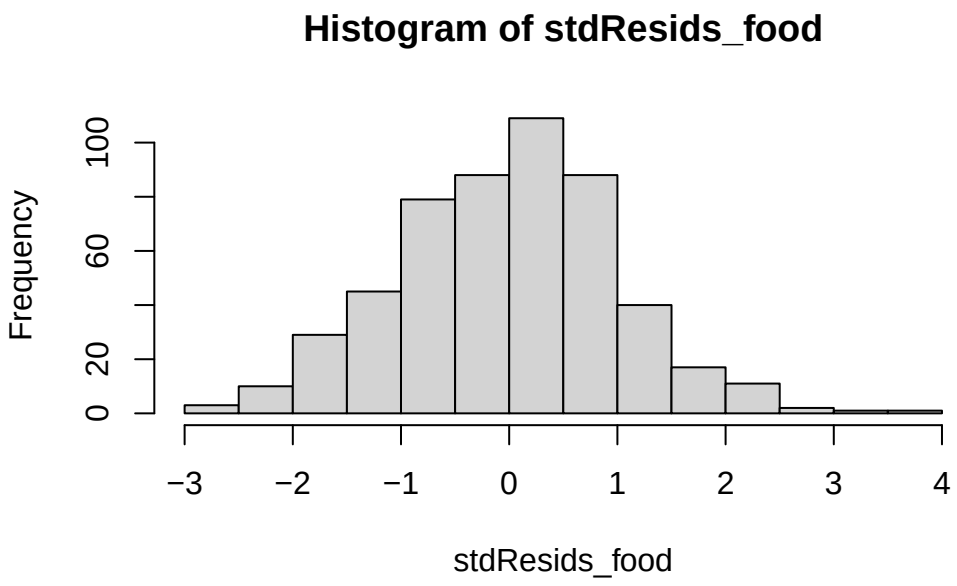
Independence

It is a reasonable assumption that the eating out habits of one person does not affect the eating out habits of another.

Normality

The below histogram show that the residuals are normally distributed.

```
stdResids_food = resid(object=hetero_sked, type="pearson")


hist(stdResids_food)
```



Equal Variance

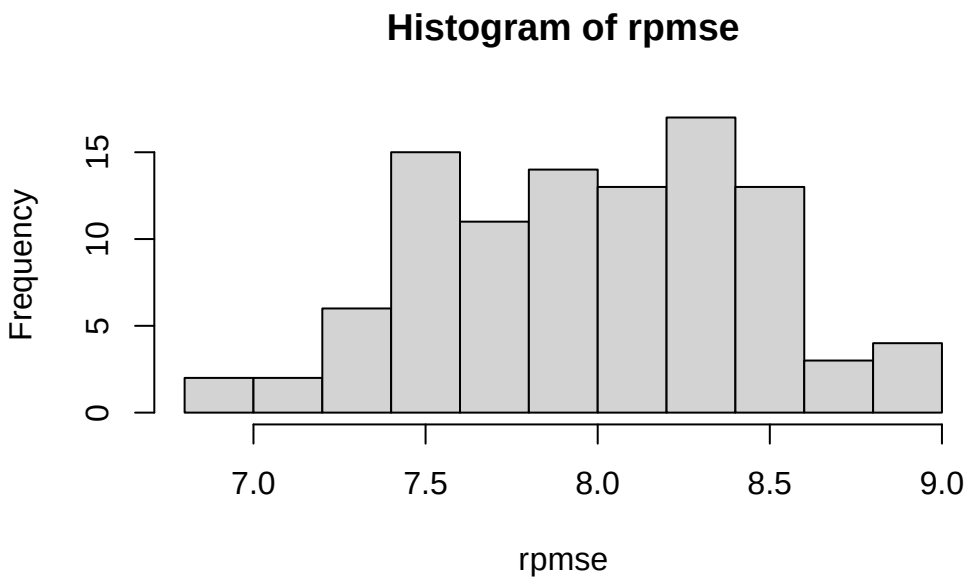
```
bp_test <- bptest(stdResids_food ~ Income, data = food)
```

The above plot shows that the equal variance has been accounted for by the gls model. The Breusch-Pagan test verifies this with p-value of 0.6309398 . Since the p-value is above .05 we fail to reject the null hypothesis that there is heteroskedasticity and conclude that that there is no heteroskedasticity.

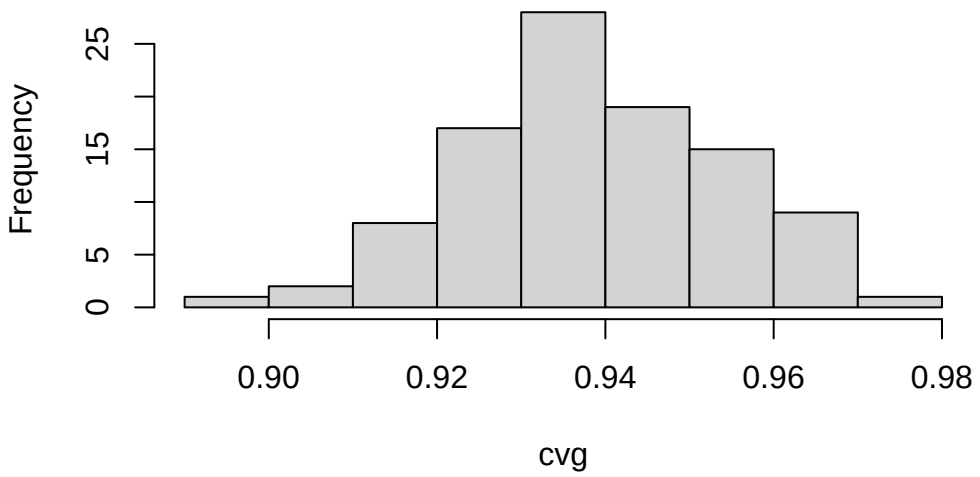
5: Validate predictions via cross-validation

```
n.cv <- 100 #Number of CV studies to run
n.test <- 200 #Number of observations in a test set
rpmse <- rep(x=NA, times=n.cv)
cvg <- rep(x=NA, times=n.cv)
for(cv in 1:n.cv){
  ## Select test observations
  test.obs <- sample(x=1:nrow(food), size=n.test)

  ## Split into test and training sets
  test.set <- food[test.obs,]
  train.set <- food[-test.obs,]
```

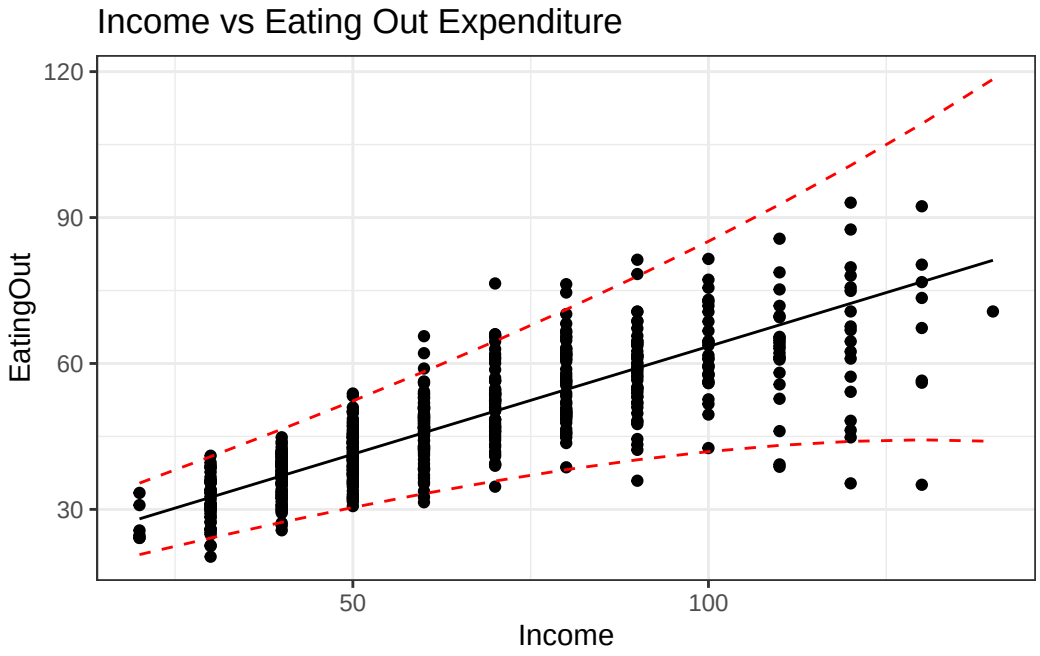
```
  ## Fit a lm() using the training data
  train.gls <- gls(data = train.set, EatingOut ~ Income, weights =
                   varExp(form=~Income), method="ML")

  ## Generate predictions for the test set
  my.preds <- predictgls(train.gls, newdframe=test.set, level=.95)

  ## Calculate RPMSE
  rpmse[cv] <- (test.set[['EatingOut']]-my.preds[,'Prediction'])^2 %>% mean() %>% sqrt()

  ## Calculate Coverage
  cvg[cv] <- ((test.set[['EatingOut']] > my.preds[,'lwr']) & (test.set[['EatingOut']] < my.pr

}
hist(rpmse)
```




```
hist(cvg)
```




```
x = food
x= predictgls(glsobj=hetero_sked, newdframe=x, level=.95)
ggplot() +
geom_point(data=x,
mapping=aes(x=Income, y=EatingOut)) + #Scatterplot
geom_line(data=x,
mapping=aes(x=Income, y=Prediction)) + #Prediction Line
geom_line(data=x,
mapping=aes(x=Income, y=lwr),
color="red", linetype="dashed") + #lwr bound
geom_line(data=x,
mapping=aes(x=Income, y=upr),
color="red", linetype="dashed") + #Upper bound
labs(
title = "Income vs Eating Out Expenditure"
) +
theme_bw()
```



6: Report Beta Hat and Variance Parameters

```
income = hetero_sked$coefficients[2]
conf_income = intervals(hetero_sked, level = .95)
lwr = conf_income$coef[2]
upr = conf_income$coef[6]
```

The coefficient for $\hat{\beta}_{\text{inc}}$ is 0.443. A 95% confidence interval for income is 0.4165066 - 0.4695935. We are 95% confident that for every one increase in Income, Eating Out goes up between 0.4165066 and 0.4695935. Since the interval does not contain 0 we reject the null hypothesis that Income has no affect on average weekly expenditure on food not cooked at home and conclude that there is an affect.

```
intervals(hetero_sked, level = .95)$varStruct[c(1,3)]
```

```
[1] 0.01121274 0.01594923
```

```
theta = conf_income$varStruct[2]
```

The point estimate for theta is 0.013581. The 95% confidence interval for θ in the variance function is (.0112, .0159).

Because the 95% confidence interval for θ contains only positive numbers, as income increases, there is higher variability in the average weekly expenditure on food not cooked at home.

7: Test if the economy is not healthy

```
linCombo = c(0,1) # just a placeholder value to pick off the coeff
# we want

econ = summary(glht(hetero_sked, linfct = t(linCombo), alternative = "less",rhs = .50))

broom::tidy(summary(econ)) %>% kableExtra::kbl(booktabs = T)
```

contrast	null.value	estimate	std.error	statistic	adj.p.value
1	0.5	0.4430501	0.0135114	-4.214959	1.25e-05

H_0 is the economy is healthy $\hat{\beta}_{\text{inc}} = .50$ H_A is the economy is not health $\hat{\beta}_{\text{inc}} < .50$

The p-value is 0.0000125. Since the p-value is

8: Predict your own restaurant spend

```
x <- data.frame(Income = 6.24)
thomas_predict <- predictgls(glsobj=hetero_sked, newdframe=x, level=.95)
c(thomas_predict[["lwr"]], thomas_predict[["upr"]])
```

```
[1] 15.77442 28.13266
```

Thomas and I estimated our first salary to be $62400. At this salary, we are 95% confident we will each spend between $15.77 and $28.13 a week on eating out.