Food Expenditure Data Analysis

Tetsuya Chau

2/6/2022

Problem Background

The question of how much money people spend on “eating out” as their income rises or falls has been a question that has long interested economists. If restaurant food is a “normal” good, one would expect consumption (and spending) to increase as income increases. This relationship has implications for projections of food demand and restaurant ownership as a viable career as incomes grow or as countries develop and become wealthier. However, just because people will spend more at restaurants as their income increases, on average, the relationship is likely more complicated due to high degree of variability for high income individuals. Also of interest is how demand for “quality” and “convenience” changes as income changes, which relates to whether people spend more money on food at home vs. away from home as income increases.

The dataset FoodExpenses.txt contains data from the Food Demand Survey (FooDS) on n=523 households income and associated expenses on “eating out”. Specifically, FoodExpenses.txt contains the following variables:

Variable Name and Description

*Income - Annual household income (in thousands)

*EatingOut - Average weekly expenditure on food not cooked at home.

To analyze this dataset, do the following:

Analysis Questions:

1. Create exploratory plots and calculate summary statistics from the data. Comment on any potential relationships you see between Income and EatingOut.
2. Using a homoskedastic linear model, fit a regression model to EatingOut using Income as the explanatory variable. Determine if the equal variance assumption is met. If it not met, discuss what impact the violation of this assumption could have on an analysis on the relationship between income and food expenditure.
3. Write down a heteroskedastic linear regression model (in matrix and vector form) in terms of population parameters including your specification for the variance function with EatingOut as the response and Income as the explanatory variable. Explain the meaning of any parameters in your model. Explain how statistical inference for your model can be used to answer the effect of income on food expenditure.
4. Fit your model from #3 to EatingOut. Validate the model L-I-N-E assumptions so you will be confident that the statistical inference you perform below will be correct.
5. Validate your predictions based on your model in #3 via cross-validation (any of leave-one-out, Monte Carlo or K-fold). Report your model RPMSE and coverage. Additionally, show your predictions and 95% prediction interval bounds on a scatterplot of income vs. food expenditure.
6. Report $\hat{\beta}_{inc}$ along with a 95% confidence interval for the model in #4. Report any variance parameters (including the variance function parameters) along with appropriate 95% confidence intervals. Correctly interpret all intervals in context.

7. Economists with the National Restaurant Association (which, perhaps unfortunately, shares its acronym with another institution), hypothesize that a “healthy” restaurant economy should see increases of about $0.50 or more per week for each $1000 increase in income. Using your heteroskedastic model, test if the economy is NOT “healthy” for restaurant owners. State your hypotheses, p-value and an appropriate conclusion.
8. Predict how much you will be spending at restaurants for your desired income level upon graduation (meaning at your first job). Report a 95% prediction interval and interpret the interval in context.

```
knitr::opts_chunk$set(warning = FALSE, message = FALSE)
#install.packages("tidyverse")
library(tidyverse)
```

```
## ── Attaching packages ─────────────────────────────────────── tidyverse 1.3.1 ──
```

```
## ✓ ggplot2 3.3.5     ✓ purrr   0.3.4
## ✓ tibble  3.1.6     ✓ dplyr   1.0.7
## ✓ tidyr   1.2.0     ✓ stringr 1.4.0
## ✓ readr   2.1.2     ✓ forcats 0.5.1
```

```
## ── Conflicts ────────────────────────────────────────── tidyverse_conflicts() ──
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
#install.packages("nlme")
library(nlme)
```

```
##
## Attaching package: 'nlme'
```

```
## The following object is masked from 'package:dplyr':
##
##     collapse
```

```
#install.packages("car")
library(car)
```

```
## Loading required package: carData
```

```
##
## Attaching package: 'car'
```

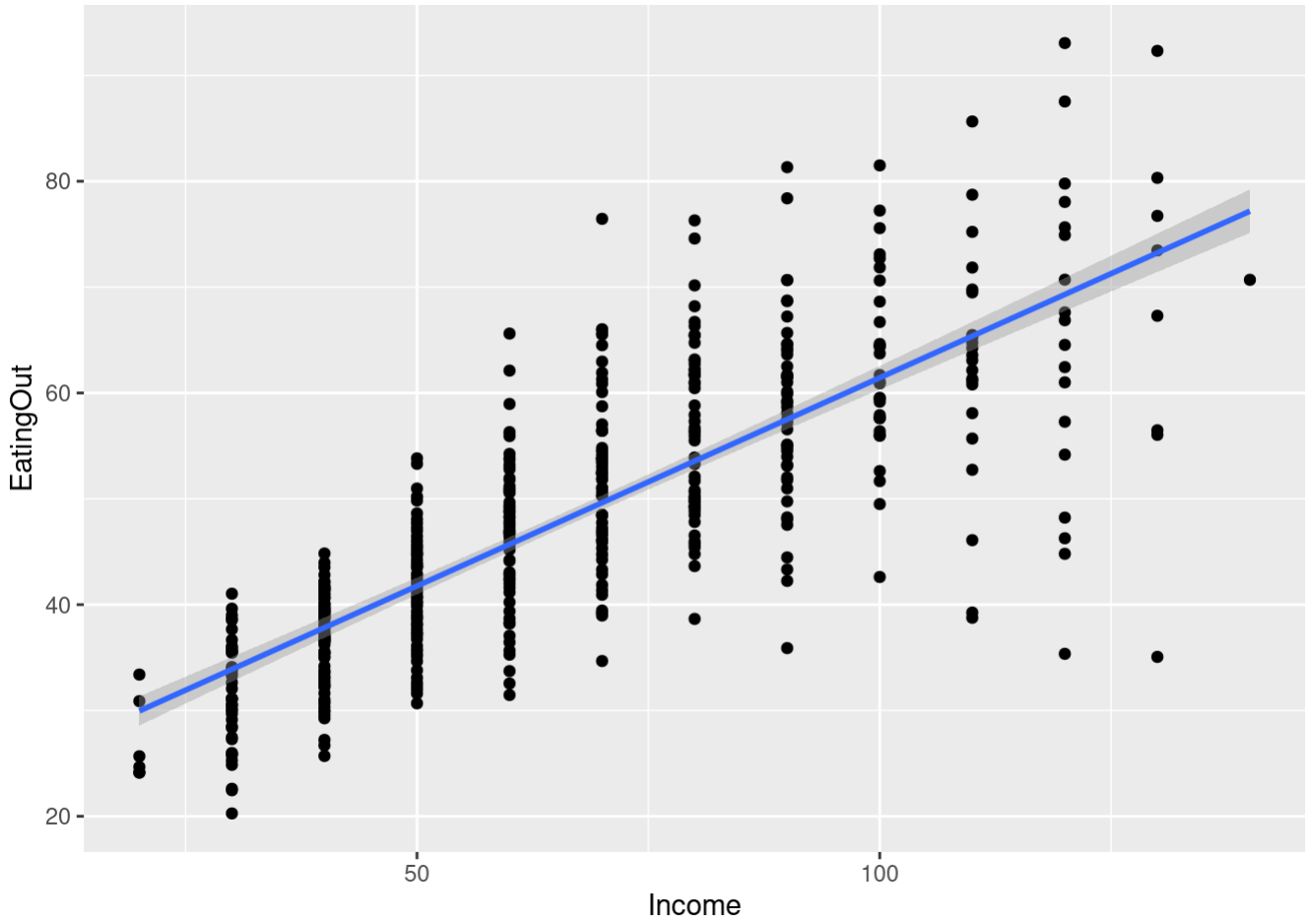
```
## The following object is masked from 'package:dplyr':
##
##     recode
```

```
## The following object is masked from 'package:purrr':
##
##     some
```

```
food <- read.csv("/cloud/project/FoodExpenses.txt", sep="")
```

1. Create exploratory plots and calculate summary statistics from the data. Comment on any potential relationships you see between Income and EatingOut.

```
ggplot(data=food, mapping = aes(x=Income, y=EatingOut))+
 geom_point()+geom_smooth(method="lm")
```



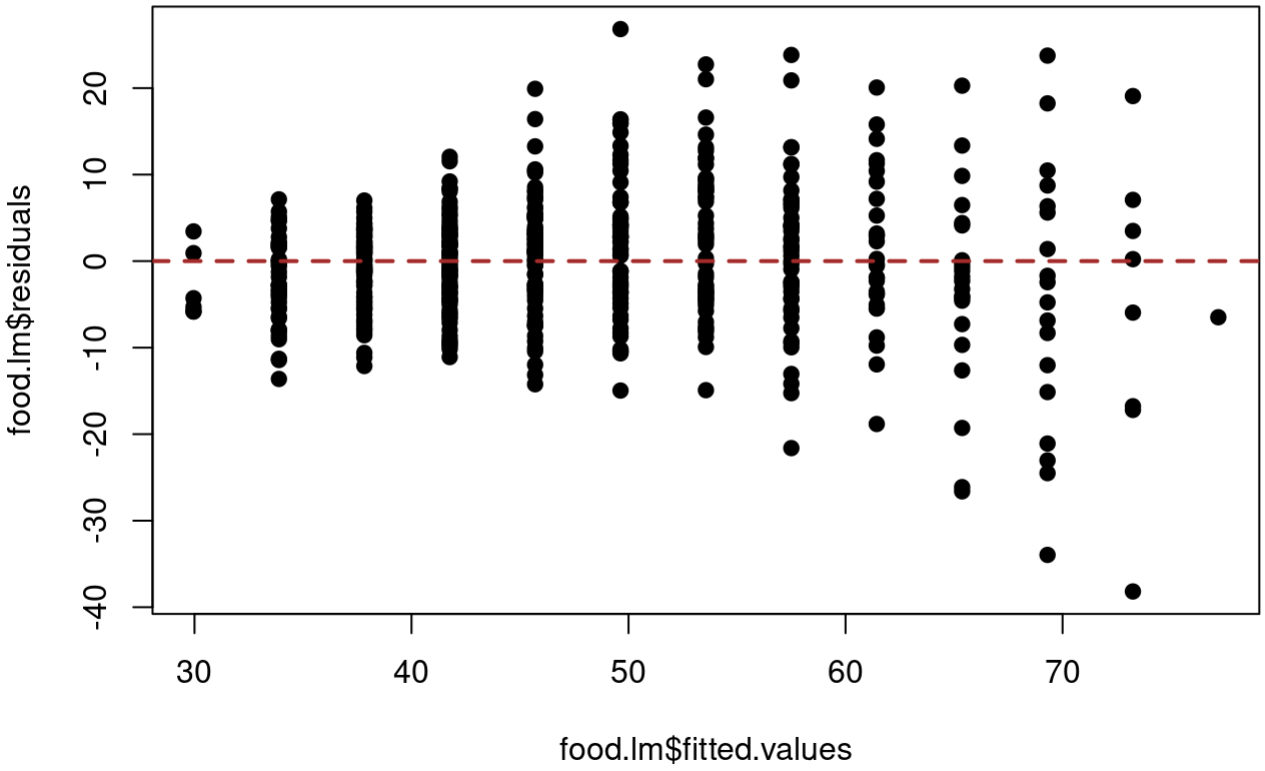

```
food.lm <- lm(EatingOut~Income,data=food)
summary(food.lm)
```

```
##
## Call:
## lm(formula = EatingOut ~ Income, 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
```

Based on the scatter plot of Income vs EatingOut, I was able to observe that there is a positive linear relationship between the two variables. It shows that as a person's income increases, his/her food expenditure increases as well.

2. Using a homoskedastic linear model, fit a regression model to EatingOut using Income as the explanatory variable. Determine if the equal variance assumption is met. If it not met, discuss what impact the violation of this assumption could have on an analysis on the relationship between income and food expenditure.

```
plot(food.lm$fitted.values,food.lm$residuals,pch=19)
abline(0,0,col = "brown",lwd = 2,lty = 2)
```



It looks like the equal variance assumption is not met since the dots in the graph look like they're spread out unevenly. It'll skew the summary outputs of the data such that the values of confidence intervals and hypothesis tests are going to be inaccurate because the variance will be biased.

3. Write down a heteroskedastic linear regression model (in matrix and vector form) in terms of population parameters including your specification for the variance function with EatingOut as the response and Income as the explanatory variable. Explain the meaning of any parameters in your model. Explain how statistical inference for your model can be used to answer the effect of income on food expenditure.

$$y \sim N(X\beta,\ \sigma^2 D(\theta))$$

Where y is the vector of the true values (or observed values). X is the design matrix. Beta is the vector of coefficients. Sigma squared is the variance and D is the diagonal elements,

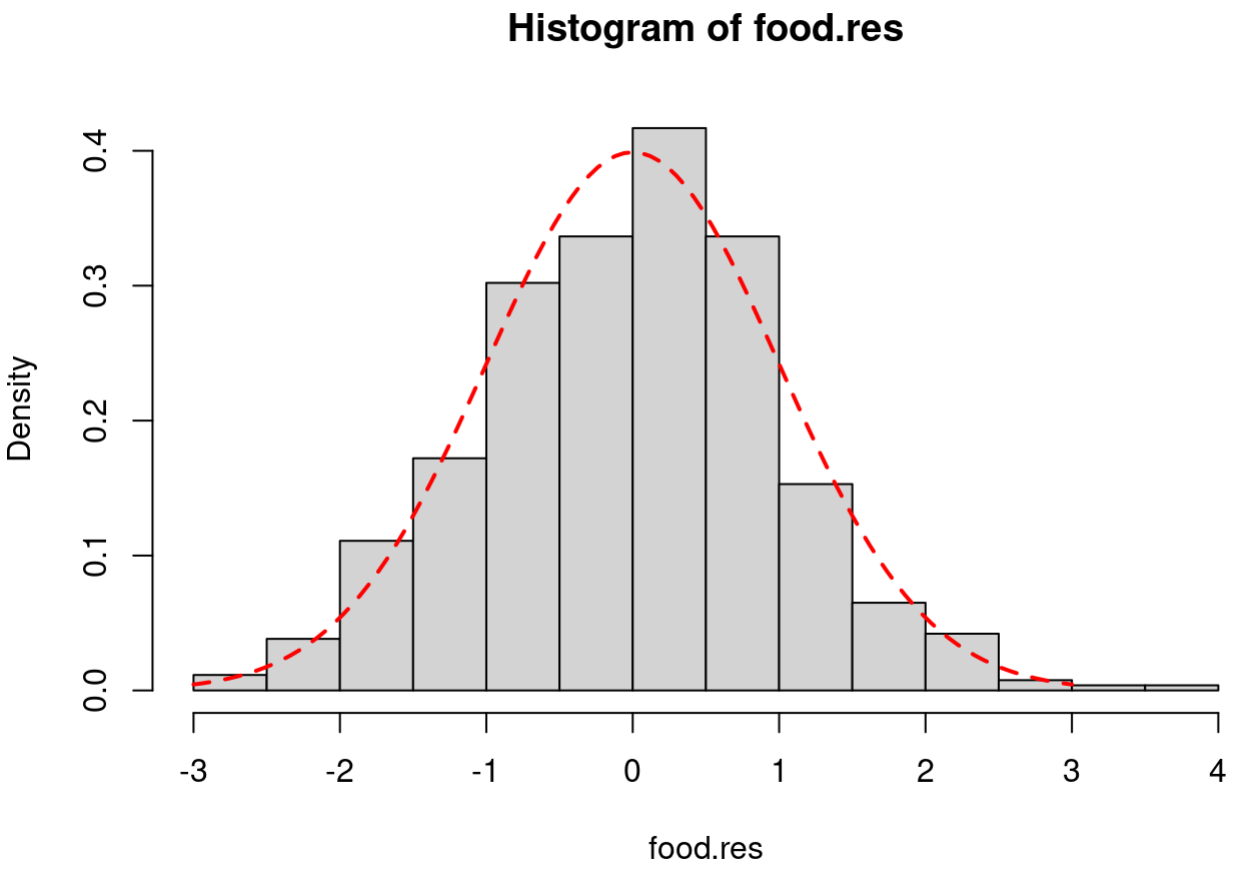
$$d_{ii} = \exp\{2\ income_i\ \theta\}$$

where xi is the income. Theta is the scalar estimated from the data. We can use the model to predict the food expenditure of an individual based on his/her income.

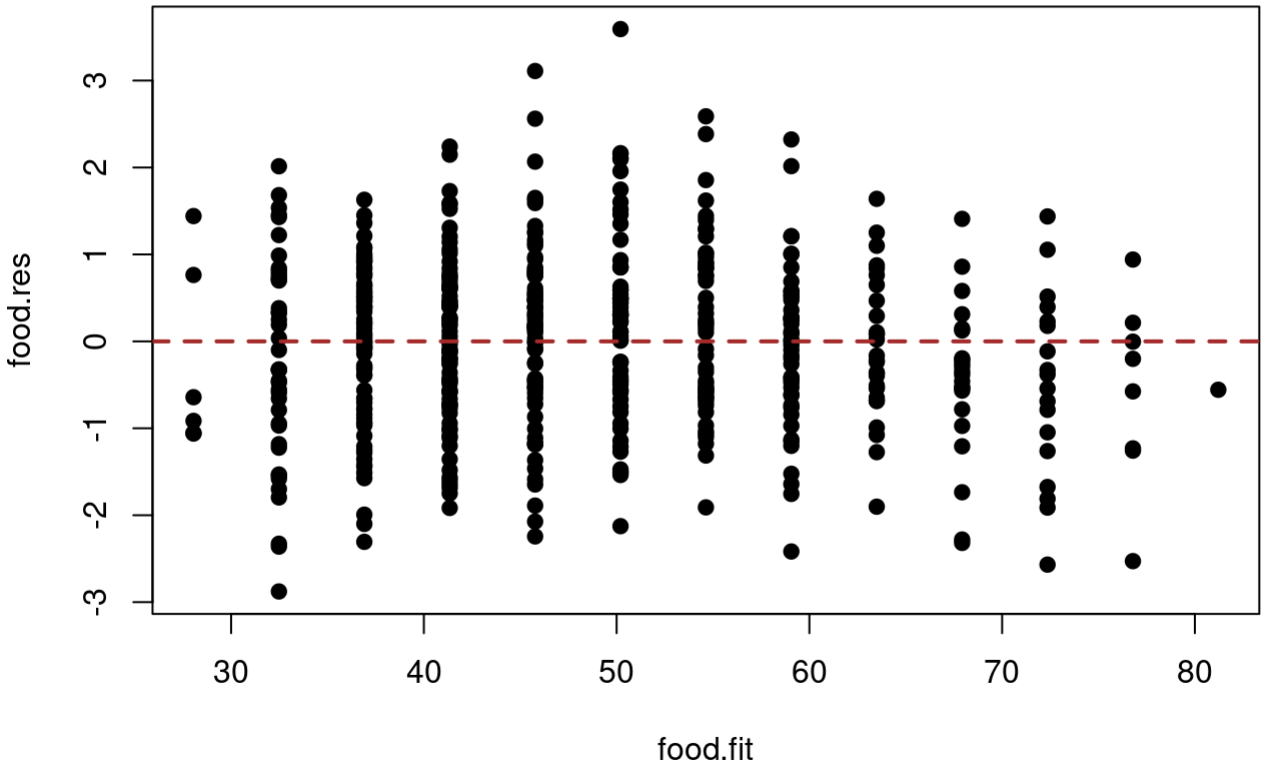
4. Fit your model from #3 to EatingOut. Validate the model L-I-N-E assumptions so you will be confident that the statistical inference you perform below will be correct.

```
food.gls <- gls(EatingOut~Income, data=food, weights=varExp(form=~Income), method="ML")
```

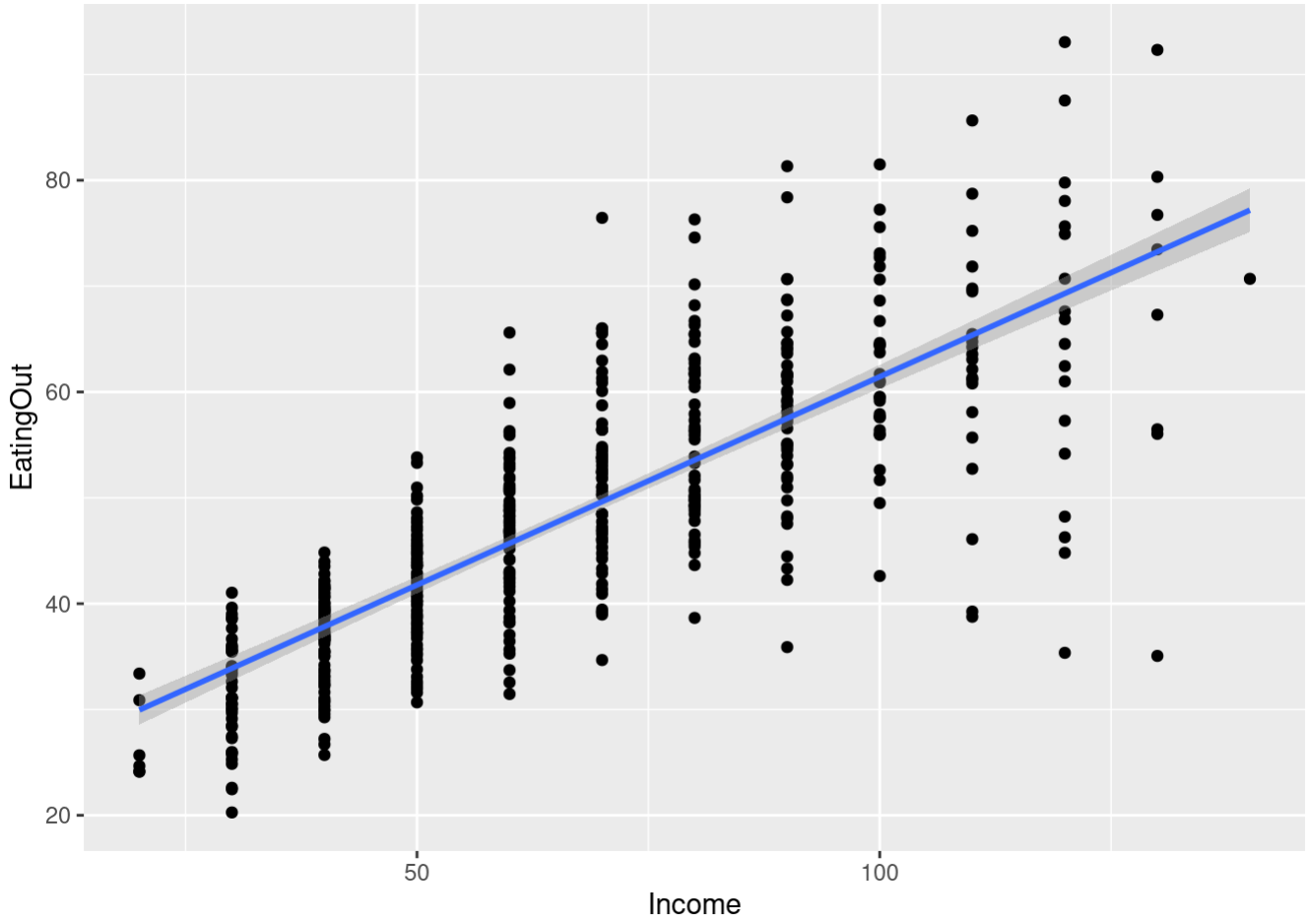
```
food.res <- resid(object=food.gls, type="pearson")
food.fit <- fitted(food.gls)

#Check for normality
hist(food.res,freq = FALSE,breaks = 20)
curve(dnorm,from = -3,to = 3,col = "red",lwd = 2,
lty = 2,add = TRUE)
```




```
#Check for equal variance
plot(food.fit,food.res,pch=19)
abline(0,0,col = "brown",lwd = 2,lty = 2)
```




```
#Check for linearity
ggplot(data=food, mapping = aes(x=Income, y=EatingOut))+
 geom_point()+geom_smooth(method="lm", formula= y~x)
```



List of assumptions and how they are met

Linearity - scatter plot of Income vs EatingOut shows a linear relationship

Independence - we can assume that one person's food expenditure does not affect the other person's food expenditure

Normality - histogram of the residuals shows a bell curve which implies normality

Equal Variance - fitted vs residual plot shows that the data points are scattered evenly (roughly) so we can assume this assumption

5. Validate your predictions based on your model in #3 via cross-validation (any of leave-one-out, Monte Carlo or K-fold). Report your model RPMSE and coverage. Additionally, show your predictions and 95% prediction interval bounds on a scatterplot of income vs. food expenditure.

```
mean_rpmse <- mean(rpmse)
print(mean_rpmse)
```

```
## [1] 7.95971
```

```
mean_cvg <- mean(cvg)
print(mean_cvg)
```

```
## [1] 0.96
```

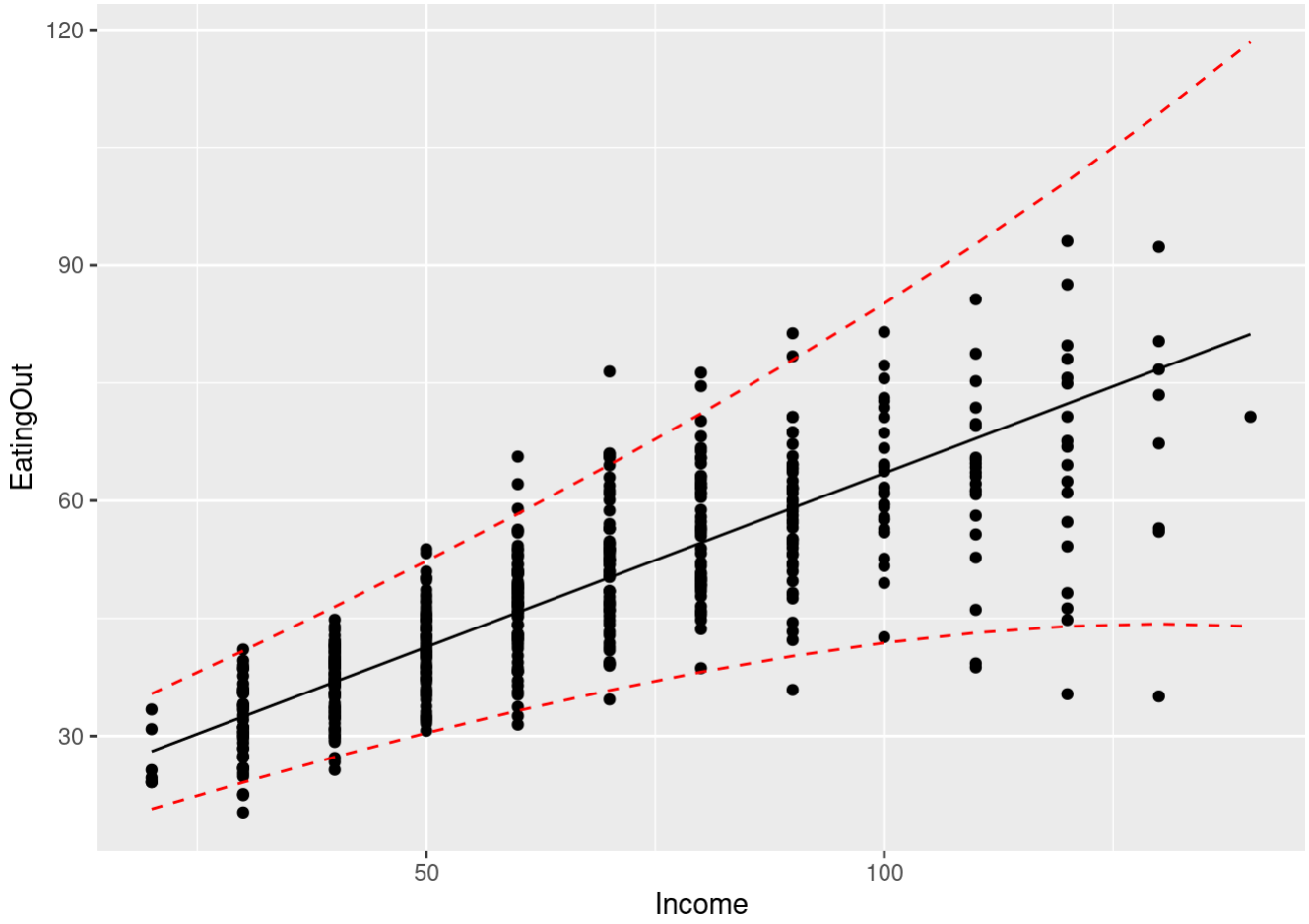
The mean RPMSE is 7.9597102 and the mean coverage is 0.96.

6. Report βˆinc along with a 95% confidence interval for the model in #4. Report any variance parameters (including the variance function parameters) along with appropriate 95% confidence intervals. Correctly interpret all intervals in context.

```
dataPreds <- predictgls(glsobj=food.gls,newdframe=food, level=0.95)
head(dataPreds)
```

```
##     Income EatingOut Prediction  SE.pred      lwr      upr
## 524     30     38.55   32.48041 4.262449 24.10671 40.85411
## 525     80     60.45   54.63291 8.382193 38.16586 71.09996
## 526     30     30.12   32.48041 4.262449 24.10671 40.85411
## 527     50     39.34   41.34141 5.574997 30.38918 52.29365
## 528     60     46.42   45.77191 6.385151 33.22811 58.31572
## 529     30     22.46   32.48041 4.262449 24.10671 40.85411
```

```
ggplot() +
  geom_point(data=dataPreds, mapping=aes(x=Income, y=EatingOut)) + #Scatterplot
  geom_line(data=dataPreds, mapping=aes(x=Income, y=Prediction)) + #Prediction Line
  geom_line(data=dataPreds, mapping=aes(x=Income, y=lwr), color="red", linetype="dashed") + #Lwr
bound
  geom_line(data=dataPreds, mapping=aes(x=Income, y=upr), color="red", linetype="dashed") #Upper
bound
```




```
food.gls
```

```
## Generalized least squares fit by maximum likelihood
##   Model: EatingOut ~ Income
##   Data: food
##   Log-likelihood: -1753.545
##
## Coefficients:
## (Intercept)      Income
##  19.1889086   0.4430501
##
## Variance function:
##  Structure: Exponential of variance covariate
##  Formula: ~Income
##  Parameter estimates:
##      expon
## 0.01358099
## Degrees of freedom: 523 total; 521 residual
## Residual standard error: 2.823682
```

```
(summary(food.gls)$sigma)^2
```

```
## [1] 7.973177
```

```
intervals(food.gls, level = 0.95)
```

```
## Approximate 95% confidence intervals
##
##  Coefficients:
##                  lower       est.      upper
## (Intercept) 17.7220786 19.1889086 20.6557387
## Income       0.4165066  0.4430501  0.4695935
##
##  Variance function:
##            lower       est.     upper
## expon 0.01121277 0.01358099 0.0159492
##
##  Residual standard error:
##    lower     est.    upper
## 2.388064 2.823682 3.338763
```

Beta hat income is 0.4430501

```
confint(food.gls,level=0.95)
```

```
##                  2.5 %     97.5 %
## (Intercept) 17.7254861 20.6523311
## Income       0.4165682  0.4695319
```

We are 95% confident that the true average increase of the food expenditure for every one thousand dollars extra of income is between 0.4165682 and 0.4695319

We are 95% confident that the variance function parameter is between 0.01121277 and 0.0159492. Because the variance parameter is positive, as the income increases, the variance of food expenditure increases on average.

The sigma squared confidence interval is (2.388064, 3.338763).

There is not a way to interpret the sigma squared interval because the variance for every individual observation changes based on the income.

7. Economists with the National Restaurant Association (which, perhaps unfortunately, shares its acronym with another institution), hypothesize that a “healthy” restaurant economy should see increases of about $0.50 or more per week for each $1000 increase in income. Using your heteroskedastic model, test if the economy is NOT “healthy” for restaurant owners. State your hypotheses, p-value and an appropriate conclusion.

$$H_o : \hat{\beta}_{income} = 0.5$$

$$H_a : \hat{\beta}_{income} < 0.5$$

```
#summary(food.gls)$tTable

se <- summary(food.gls)$tTable[2,2]
bta <- coef(food.gls)[2]
test.stat <- (bta-0.5)/se
pvalue <- pt(test.stat, df=n-(1+1), lower.tail=TRUE)
print(pvalue)
```

```
##      Income
## 1.47251e-05
```

Based on the p-value of 1.47251e-05, we can reject the null hypothesis and conclude that the economy is not healthy for restaurant owners.

```
summary(food.gls)$tTable
```

```
##                  Value  Std.Error  t-value       p-value
## (Intercept) 19.1889086 0.74665785 25.69973  1.092569e-94
## Income       0.4430501 0.01351139 32.79087 9.073807e-129
```

8. Predict how much you will be spending at restaurants for your desired income level upon graduation (meaning at your first job). Report a 95% prediction interval and interpret the interval in context.

```
pData = data.frame(Income = c(75))
predictgls(glsobj = food.gls, newdframe = pData, level = 0.95)
```

```
##   Income Prediction  SE.pred     lwr      upr
## 1     75   52.41766 7.830668 37.0341 67.80123
```

We are 95% confident that the amount that I will spend on eating out after graduation per week is between 37.0341 and 67.80123 dollars with an income of 75,000 dollars.