Food Expenditures

Jacob Johnson & Naomi Zubeldia

2023-02-07

1. Exploratory Data Analysis

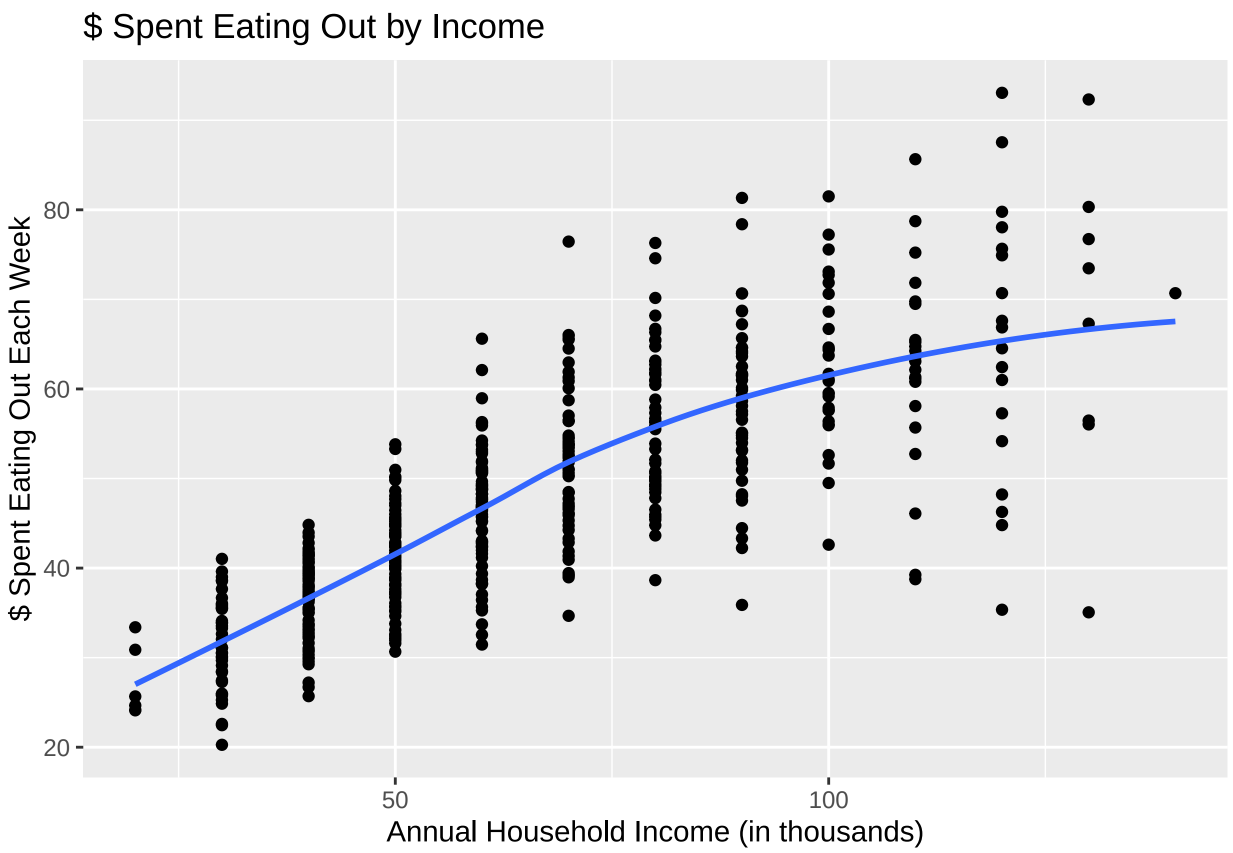
```
## Load needed libraries
library(MASS)
library(car)
library(lmtest)
library(nlme)
library(tidyverse)
library(multcomp)

## Read in the Data
food <- read.table('FoodExpenses.txt', header=TRUE)
```

```
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  
```

To observe the data given and look for a relationship between income and amount of money spent eating out, we created the summary table above. It appears that the income variable ranges from $20k to $140k, which is a pretty large range with the median being closer to the minimum. For the eating out variable, the weekly spending has a closer range than the income but it is still pretty large varying from $20.27 a week to $93.06. It will be interesting to see that relationship between the factors with one having the more drastic spread.

```
#### Exploratory Data Analysis
## Scatterplot of EatingOut by Income
ggplot(food, aes(Income, EatingOut)) +
  geom_point() + geom_smooth(se=FALSE) +
  ggtitle("$ Spent Eating Out by Income") +
  xlab("Annual Household Income (in thousands)") + ylab("$ Spent Eating Out Each Week")
```



From the plot above, we see that annual household income and amount spent eating out each week have a positive linear relationship. As income increases, amount spent eating out also increases roughly linearly. The spread also appears to be rather close at the beginning and starts to spread out in a cone shape, which seems like there may be a potential homoskedastic relationship between the variables.

```r
r <- cor(food[['EatingOut']], food[['Income']])
r
```

```
## [1] 0.7910809
```

The correlation between the two factors is very positive at .7910809. This means that as income increases, amount spent eating out increases with it. However, there may be other influential factors like location of people to restaurants as well as if the people multiple people from the observations came from families that have a preference for eating out.

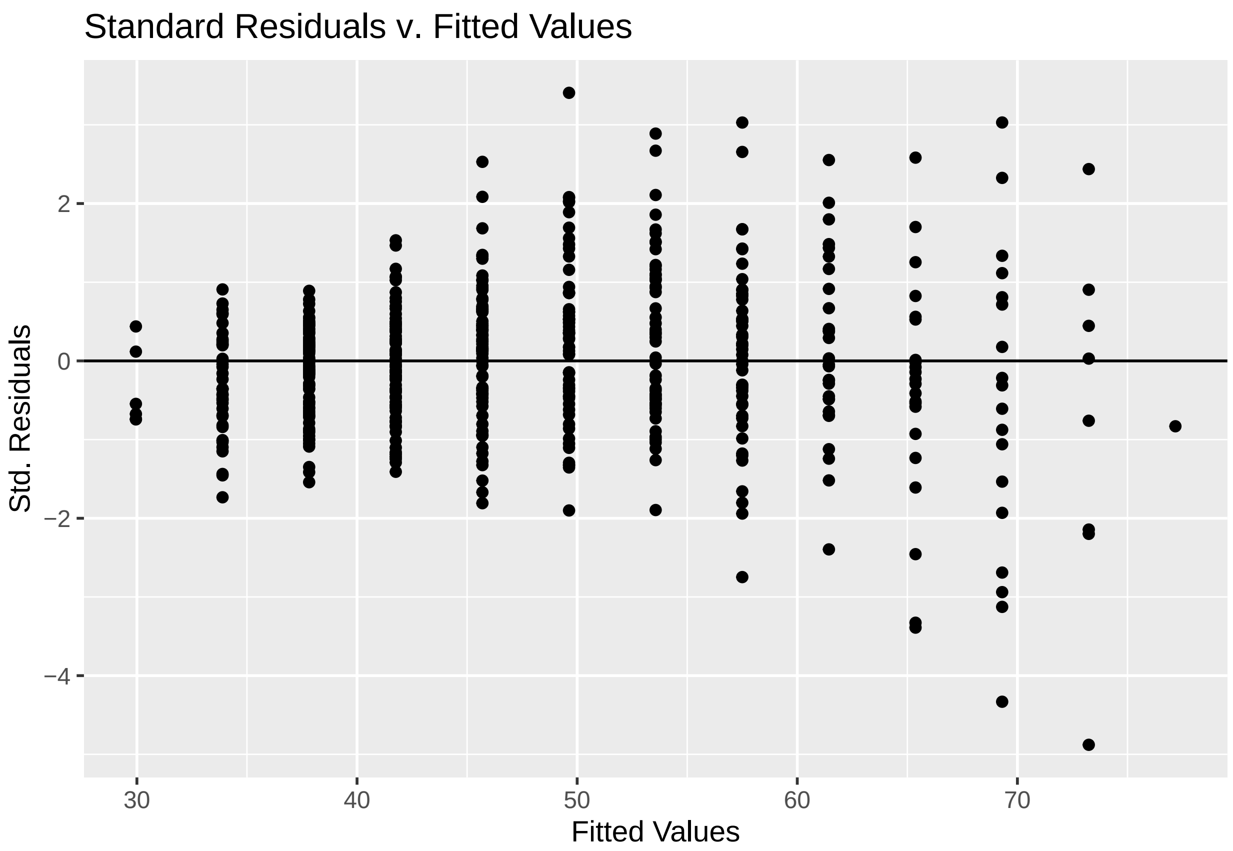
2. Homoskedastic Linear Model

An intuitive first approach is to fit a linear regression model with equal variance and see how well it fits our data.

```r
## Fit a model assuming equal variance
basic.lm <- lm(EatingOut~Income, data=food)
summary(basic.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
```

```
## Check Equal Variance Assumption
#### Fitted Values vs. Residuals Plot
fit.vals <- predict.lm(basic.lm)
residuals <- stdres(basic.lm)
ggplot()+
  geom_point(mapping = aes(x = fit.vals, y = residuals))+
  xlab("Fitted Values") + ylab("Std. Residuals") +
  ggtitle("Standard Residuals v. Fitted Values")+
  geom_hline(mapping = aes(yintercept=0))
```



```
#### BP test for equal variance
bp <- bptest(basic.lm)$p.value
bp
```

```
##           BP
## 3.703729e-16
```

According to the scatter plot above, the spread of the data looks kind of skewed indicating that there may be unequal variance. To officially test this, we did a bp test and found the pvalue to be $3.7037288 \times 10^{-16}$, which is less than the level of significance, which means we reject the null hypothesis that the model has equal variance. We conclude that there is not equal variance about our regression line. The impact of this assumption violation is that all the standard errors calculated for our model are wrong, making our model invalid and we cannot tell how far off our estimates really are.

3. Heteroskedastic Linear Model

Now that we see that a homoskedastic linear model is a bad fit for our data, we will try a heteroskedastic linear model that allows for unequal variance. Specifically, we will account for how the variance of $ spent eating out per week is different for different incomes. The model is detailed below:

$$\boldsymbol{y} \sim MVN(\boldsymbol{X\beta}, \sigma^2\boldsymbol{D}(\theta)) \qquad d_{ii} = \exp\{2\text{Income}_i\theta\}$$

The variables and parameters in our model are explained below.

$\boldsymbol{y}$: A vector of our response variable. In this case, this is the average weekly expenditure on food not cooked at home for the 523 households represented in our data set.

$\boldsymbol{X}$: A model matrix of covariates. In this case, we first have a column of 1's for the intercept (explained below) and a column representing the Annual household income for every household in our data set.

β: A vector of coefficients for each explanatory variable. In our case, these are the coefficients representing the intercept (the expected amount spent in eating out if the income was $0) and the effect of income on $ spent eating out.

σ^2: The variance of the residuals from our model. We can use our model to predict the average amount spent eating out based on a given income, but there is also some variability associated with it. σ^2 is a measure of the variability of our residuals, meaning that σ represents the average error for all of our data points.

$\boldsymbol{D}$: A diagonal matrix of weights that changes the variance around our predicted regression line for a given income. $\mid d_{ii} = \exp\{2\text{Income}_i\theta\}$. The i^{th} diagonal element of the diagonal matrix $\boldsymbol{D}$ is calculated using this formula. This means that the variance of our model (how far the actual values are from our predicted average) will change based on income. Together, $\sigma^2 d_{ii}$ represents the variance around our regression line for a specific income. So, $\sqrt{\sigma^2 d_{ii}}$ is the average distance from the true amount spent eating out to our projected average amount spent for a given income.

θ: A coefficient that represents how the variance of the observations changes based on income. If this value is positive, then the variance increases as income increases, and if it is negative, then the variance decreases as income increases.

4. Model Fit and Assumptions

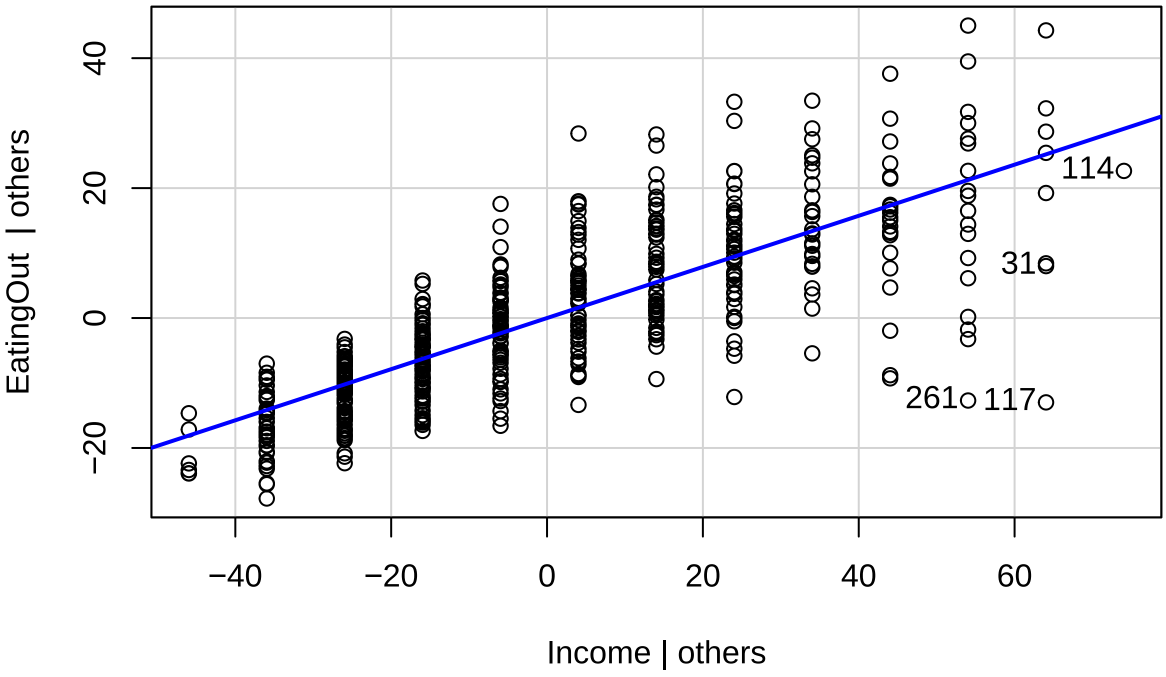
```
## Heteroskedastic Model
food.gls <- gls(model=EatingOut~Income, data=food, weights=varExp(form=~Income), method="ML")

summary(food.gls)
```

```
## Generalized least squares fit by maximum likelihood
##   Model: EatingOut ~ Income
##   Data: food
##        AIC      BIC    logLik
##   3515.089 3532.128 -1753.545
##
## Variance function:
##  Structure: Exponential of variance covariate
##  Formula: ~Income
##  Parameter estimates:
##      expon
## 0.01358099
##
## Coefficients:
##                Value Std.Error  t-value p-value
## (Intercept) 19.18891 0.7466578 25.69973       0
## Income       0.44305 0.0135114 32.79087       0
##
##  Correlation:
##        (Intr)
## Income -0.931
##
```

```
## Standardized residuals:
##         Min          Q1         Med          Q3         Max
## -2.87720034 -0.65501506  0.03262438  0.64060363  3.59255406
##
## Residual standard error: 2.823682
## Degrees of freedom: 523 total; 521 residual
```

After creating a heteroskedastic model, we need to check the LINE assumptions to make sure that our model will be valid

```
## Checking Assumptions
#### Linearity
avPlots(basic.lm)
```



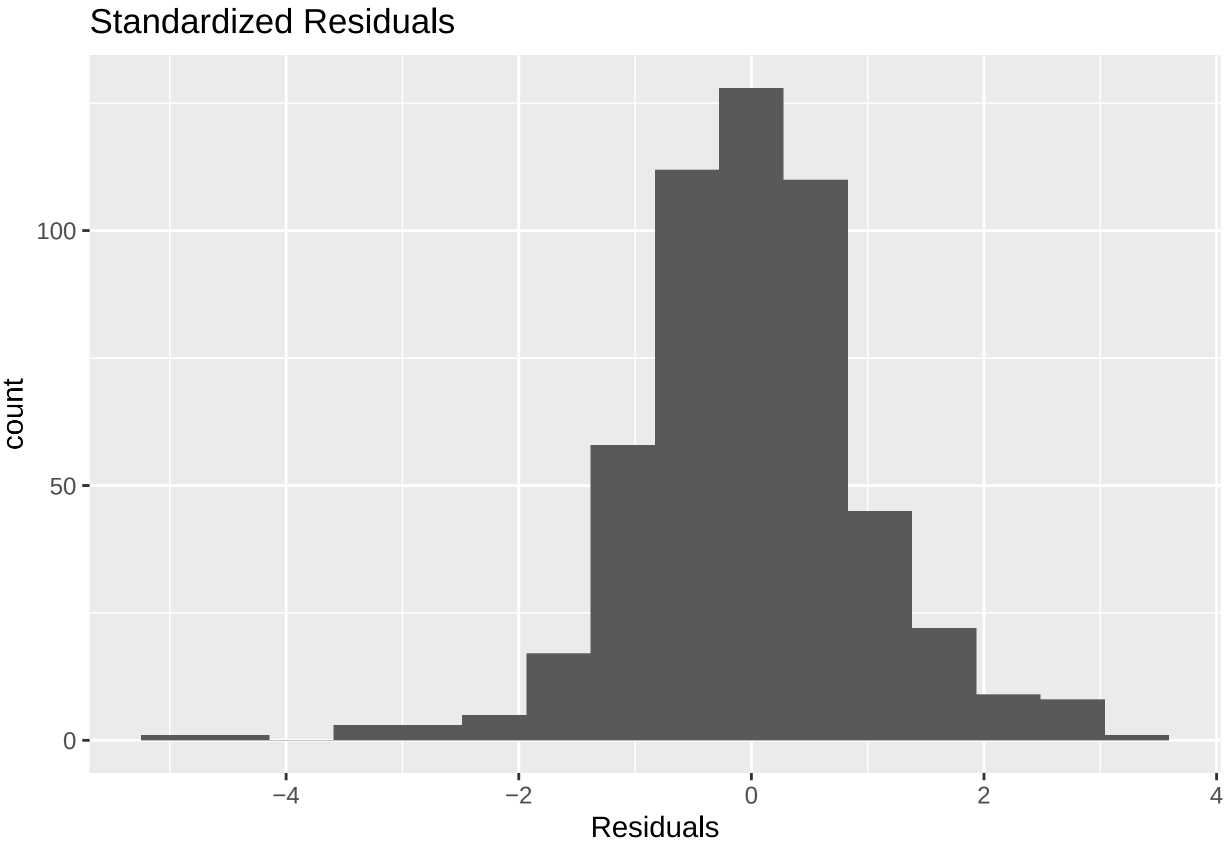
In the added variable plot above, we can see that the to variables have a positive linear relationship, which validates the normality assumption.

To check the independence assumption, we cannot see any reason why the observations are independent. There is nothing in the data to tell us that observations are related, so we assume that this assumption is met.

```
#### Normality
source('predictgls.R')
fit.vals <- predictgls(glsobj=food.gls, level=.95)
```

```
## Warning in predictgls(glsobj = food.gls, level = 0.95): No newdframe provided
## for prediction. Returning fitted values.
```

```
residuals <- resid(object=food.gls, type="pearson")

###### Hist of Standardized Residuals
ggplot()+
  geom_histogram(mapping=aes(x=stdres(basic.lm)),bins=16)+
  xlab("Residuals")+
  ggtitle("Standardized Residuals")
```



```
######
ks <- ks.test(residuals,"pnorm")$p.valu
```

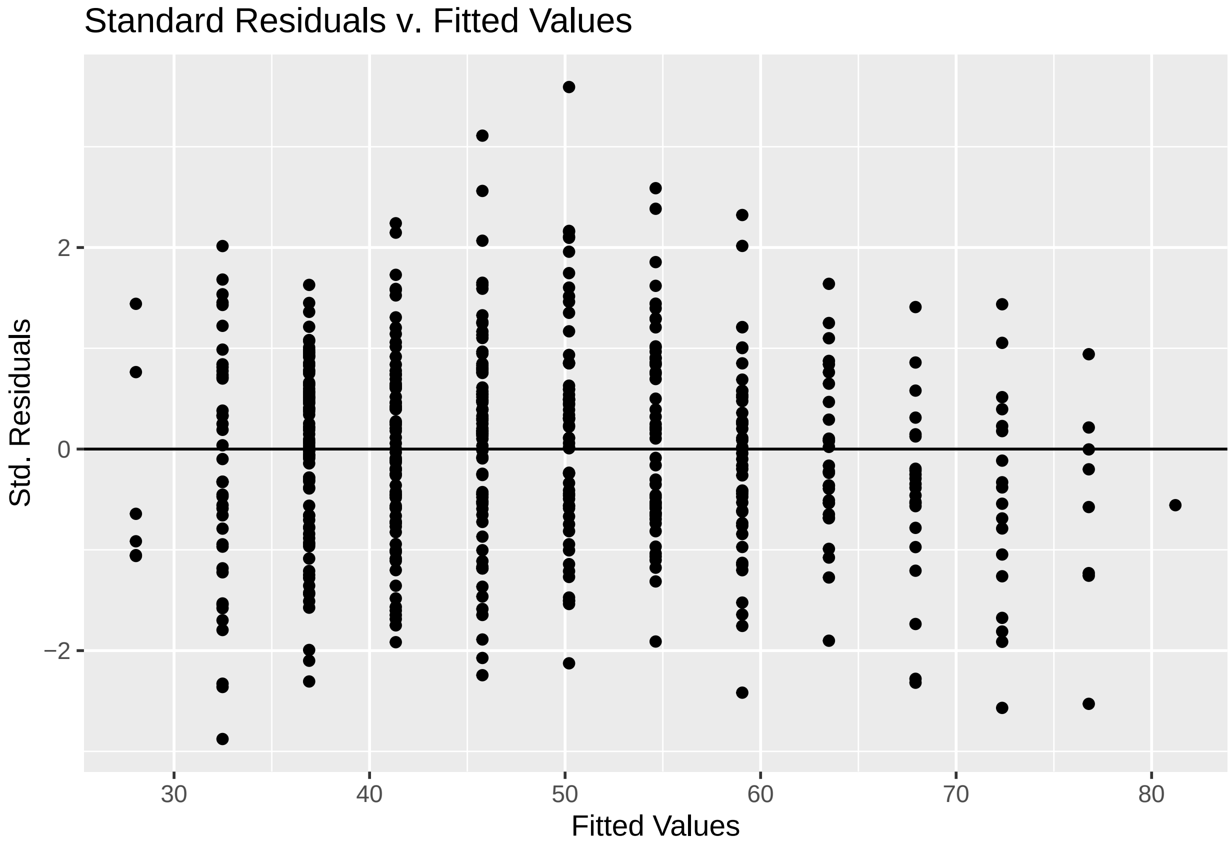
```
## Warning in ks.test.default(residuals, "pnorm"): ties should not be present for
## the Kolmogorov-Smirnov test
```

```
ks
```

```
## [1] 0.8092539
```

To check normality, we plotted the standardized residuals and that plot appears to be normal. To verify normality, we ran a ks test and found the pvalue of 0.8092539 to be larger than the significance level, which means we fail to reject the null hypothesis and assume that the model is normal.

```
#### Equal Variance
###### Fitted Values vs. Residuals Plot
ggplot()+
  geom_point(mapping = aes(x = fit.vals, y = residuals))+
  xlab("Fitted Values") + ylab("Std. Residuals") +
  ggtitle("Standard Residuals v. Fitted Values")+
  geom_hline(mapping = aes(yintercept=0))
```



Lastly, to re-check the equal variance assumption, we plotted the fitted values against the std. residuals of the new model and the spread of the values is much better now. So, the equal variance assumption has now been met.

5. Predictive Diagnostics

Now that we know that our model assumptions are met and that a heteroskedastic regression model can be used, we can test how well our model does at predicting. To do this, we conduct cross validation, where you simulate the prediction process by withholding some of our data as a test set.

```
## Cross Validation
n.cv <- 100 #Number of CV studies to run
n.test <-  round(nrow(food)/10)
rpmse <- rep(x=NA, times=n.cv)
bias <- rep(x=NA, times=n.cv)
wid <- rep(x=NA, times=n.cv)
cvg <- rep(x=NA, times=n.cv)
```

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

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

  ## Calculate bias
  bias[cv] <- mean(my.preds[,'Prediction']-test.set[['EatingOut']])

  ## 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.preds[,'upr'])) %>% mean()

  ## Calculate Width
  wid[cv] <- (my.preds[,'upr'] - my.preds[,'lwr']) %>% mean()

}

## Report the mean of each of our diagnostics
mean(bias)
```

```
## [1] 0.3801853
```

```r
mean(rpmse)
```

```
## [1] 7.804946
```

```r
mean(cvg)
```

```
## [1] 0.9401923
```

```r
mean(wid)
```

```
## [1] 28.94736
```

The first value reported after our cross validation is bias, which is a measure of if we are consistently over or under predicting. We found that our average bias was 0.3801853. This means that on average, our predictions were above the true amount spent eating out by about $0.37. This is incredibly good, meaning that our predictions are unbiased.

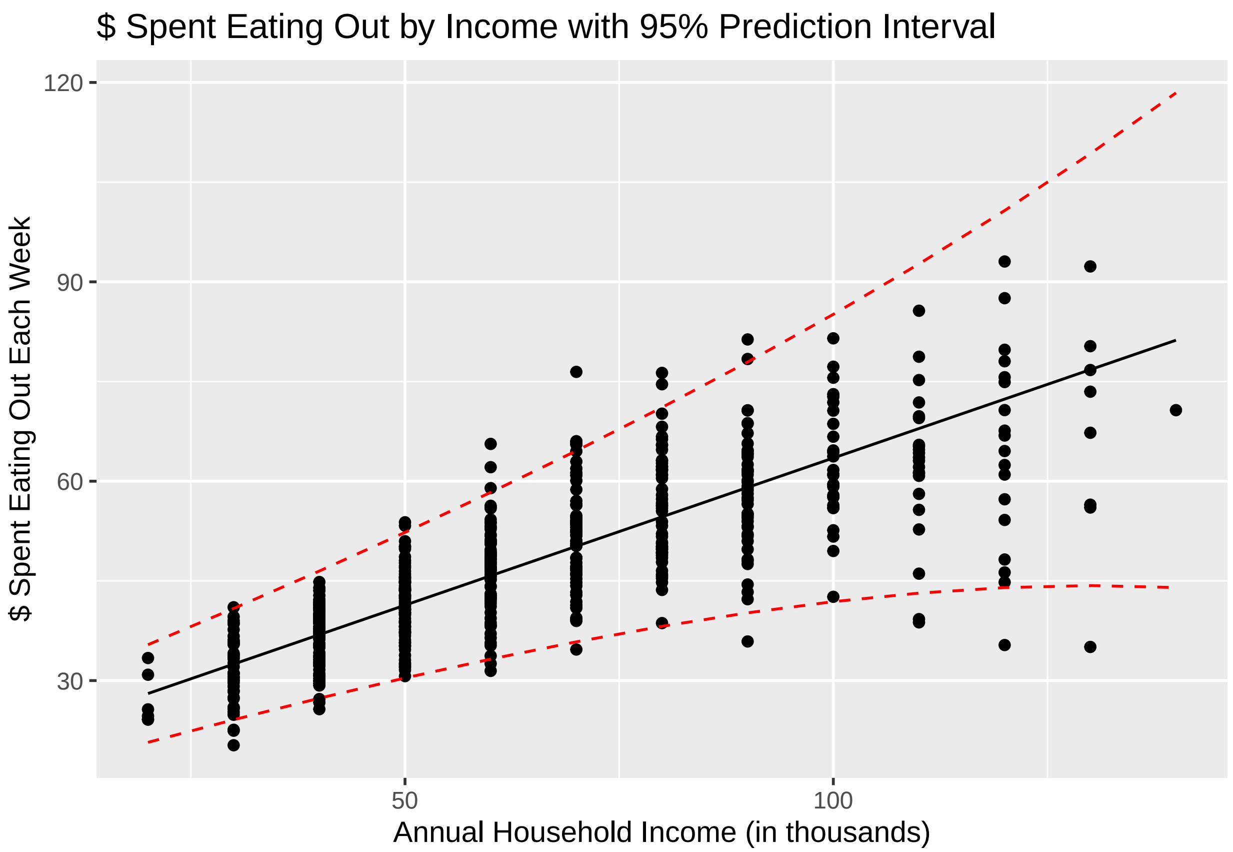
The next value reported from cross validation is the Root Predicted Mean Squared Error (RPMSE). RPMSE is a measure of how far, on average, our predictions are from the true values. This is different than bias in the fact that it treats under and over predicting equally so that we get a measure of how far off we are, but not if our model is biased. The mean RPMSE value from our cross validation was 7.8049464. This means that on average, our predictions were $7.8049464 away from the true amount spent eating out in the data. This shows that our model is helping us predict well because it is about half of the standard deviation of the amount spent eating out on its own (12.86941). If we did not use our model, we would be off by $12.86941 on average, but instead, we are only off by $7.8049464 on average.

The next value reported by our cross validation study is coverage. This is a measure of if our 95% prediction interval contained the true value. After all 523 data points were predicted, 94.0192308% of them were included in the prediction interval. This value is close to 95%, meaning that our prediction intervals are working and doing their job correctly.

And lastly, we reported the width of the prediction intervals mentioned above. On average, these prediction intervals were 28.9473611 wide. We can compare this value to the range of the amount spent eating out reported in the data set. The range of the data is 72.79 and our average width is less than 1/2 of that. This means that our model allows us to predict with much more accuracy than we would have without it.

Now that we know that our model can predict accurately, we can take a look at the predictions and prediction intervals compared to our actual data. Below is a plot with the data, our predictions, and the 95% prediction interval. the 95% prediction interval means that we are 95% confident that the true amount spent Eating Out for a specific household will be within this interval, given a specific annual household income.

```r
## 95% prediction interval
newdat <- predictgls(glsobj=food.gls, newdframe=food, level=.95)

## Plot the prediction intervals
ggplot() +
  geom_point(data=newdat,
             mapping=aes(x=Income, y=EatingOut)) + #Scatterplot
  geom_line(data=newdat,
            mapping=aes(x=Income, y=Prediction)) + #Prediction Line
  geom_line(data=newdat,
            mapping=aes(x=Income, y=lwr),
            color="red", linetype="dashed") + #lwr bound
  geom_line(data=newdat,
            mapping=aes(x=Income, y=upr),
            color="red", linetype="dashed") + #Upper bound
  ggtitle("$ Spent Eating Out by Income with 95% Prediction Interval") +
  xlab("Annual Household Income (in thousands)") + ylab("$ Spent Eating Out Each Week")
```



6. Parameter Inference

Now that we have fitted our model and checked its validity, we can make inference about the estimates from our fitted model. We now have estimates for the coefficients used in the model, and they are explained below.

```
## Beta_hat estimate and confidence interval
#### Get our coefficient estimate
coef(food.gls)
```

```
## (Intercept)      Income
##  19.1889086   0.4430501
```

```
#### Get a 95% confidence interval for beta_hat
confint(food.gls)
```

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

$\hat{\beta}_{inc} = 0.4430501$. We estimate that for every \$1000 increase in annual household income, we expect to see \$0.443 more spent on eating out each week, on average. A 95% confidence interval for this coefficient is (0.4165682 0.4695319). This means that we are 95% confidence that the true average increase in money spent eating out for a \$1000 increase in income is between \$0.416 and \$0.469.

```
## Theta estimate and confidence interval
#### Get our coefficient estimate
coef(food.gls$modelStruct, unconstrained=FALSE)
```

```
## varStruct.expon
##      0.01358099
```

```
#### Get a 95% confidence interval for beta_hat
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 0.01121274 0.01358099 0.01594923
##
##  Residual standard error:
##    lower     est.    upper
## 2.388059 2.823682 3.338769
```

$\hat{\theta} = 0.01358099$. This is our estimate for how the variance changes with income. Because this number is positive, we estimate that the variance increases as income increases.

A 95% confidence interval for this $\hat{theta}$ is (0.01121274, 0.01594923). This means that we are 95% confidence that the true variance parameter θ is between 0.01121274 and 0.01594923. Both of these values are positive, meaning that we are at least 95% confident that the true variance increases on average as income increases.

```
## Sigma estimate and confidence interval
#### Get our coefficient estimate
food.gls$sigma
```

```
## [1] 2.823682
```

```
#### Get a 95% confidence interval for beta_hat
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 0.01121274 0.01358099 0.01594923
```

```
##
##  Residual standard error:
##    lower     est.    upper
## 2.388059 2.823682 3.338769
```

The estimate from our model for σ is 2.8236815. This means that our estimate for the residual variance is 7.9731775. On average, across the whole model, our model predictions are 2.8236815 away from the true values for amount spent eating out. The 95% confidence interval for s (which estimates σ) is (2.388059, 3.338769). This means that we are 95% confident that the overall residual variance in our model (which we can multiply by our function of θ to get the variance given an income level) is between (5.702826, 11.14738) (which is the square of the standard deviation).

7. Hypothesis Testing for the Effect of Income

Now that we have a valid model and understand the effects that are represented in it, we can test to see if the effect for income on eating out is representative of a healthy restaurant economy. The National Restaurant Association estimates a "healthy" restaurant economy should see increases of about $0.50 or more per week for each $1000 increase in income. We can formally test this with the hypotheses below.

$H_0 : \beta_{income} \geq 0.5$: The restaurant economy is healthy and for every $1000 increase in income, the increase in expected amount spent eating out is $0.50 or more.

$H_1 : \beta_{income} < 0.5$: The restaurant economy is not healthy and for every $1000 increase in income, the increase in expected amount spent eating out is less than $0.50.

```
## Hypothesis test for beta_income = 0.5
a <- matrix(0,length(food.gls$coefficients),nrow=1)
a[2] <- 1

## Conduct our hypothesis test
my.test <- glht(food.gls, linfct=a, alternative="less", rhs=0.5)
summary(my.test)
```

```
##
##   Simultaneous Tests for General Linear Hypotheses
##
## Fit: gls(model = EatingOut ~ Income, data = food, weights = varExp(form = ~Income),
##     method = "ML")
##
## Linear Hypotheses:
##          Estimate Std. Error z value   Pr(<z)
## 1 >= 0.5  0.44305    0.01351  -4.215 1.25e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## (Adjusted p values reported -- single-step method)
```

After conducting this test, we get a p-value of 0.0000125. This value is extremely small, meaning that there is a very low chance of seeing results this extreme if the true increase in expected amount spent eating out was at least $0.50. Because this p-value is less that the threshold of 0.05, we reject the null hypothesis and conclude that the restaurant economy is unhealthy because the expected increase in amount spent eating out is less that $0.50 per $1000 increase in household income.

8. Predictions

```
new.x=data.frame(Income=75)
predictgls(glsobj=food.gls, newdframe=new.x, level=.95)
```

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

For a desired income of $75000 our expected weekly spending on eating out is $52.41766. A 95% prediction interval for this value is ($37.0341, $67.80123). This means that we are 95% confident that the true amount that someone with an annual income of $75000 will spend eating out each week is between $37.0341 and $67.80123.