36-402 Homework 9

James “Morgan” Hawkins

Wednesday April 12, 2023

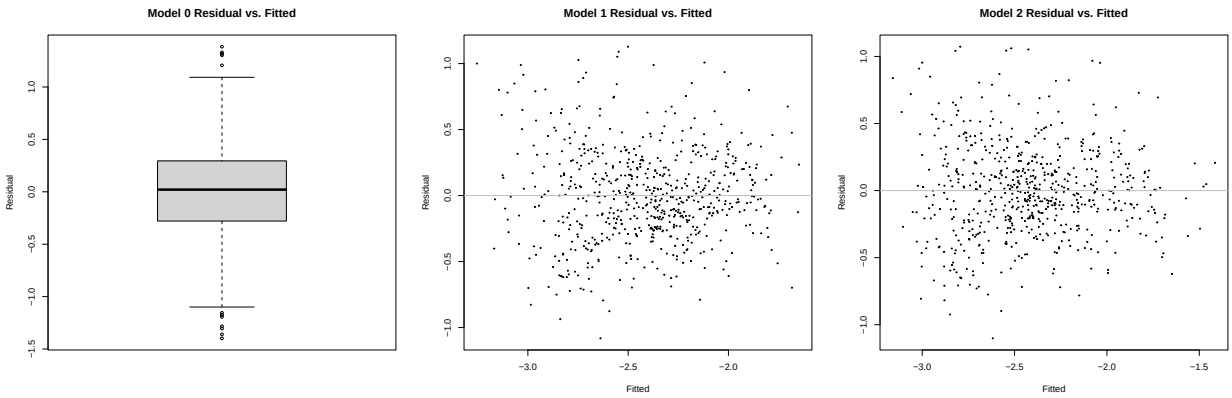
Problem 1

```
mob = read.csv("/Users/morganhawkins/Downloads/mobility.csv") %>% na.omit
```

Problem 1 (a)

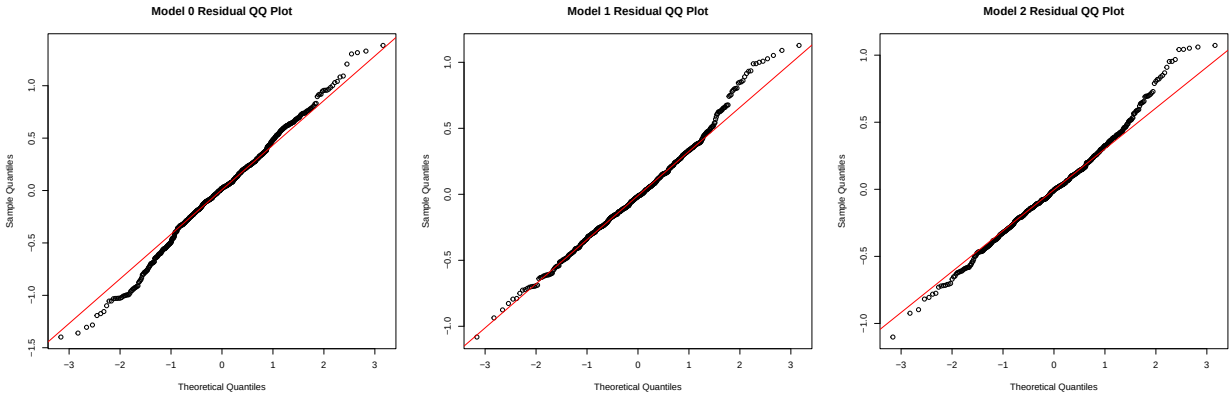
```
#model 0
model.0 = lm(log(Mobility) ~ 1, data = mob)


#model 1
mob$resid.scores = lm(Test_scores ~ Income, data = mob) %>% residuals

model.1 = lm(log(Mobility) ~ Seg_racial + Gini + resid.scores + Social_capital,
             data = mob)

#model 2

model.2 = gam(log(Mobility) ~ s(Seg_racial, k = 6, fx= T) + s(Gini, k = 6, fx= T) +
                s(resid.scores, k = 6, fx= T) + s(Social_capital, k = 6, fx= T),
              data = mob)
```

```
resid.fitted.plot = function(model, main = NA, xlab = "Fitted", ylab = "Residual"){
  fit.values = fitted(model)
  res.values = residuals(model)
  plot(fit.values, res.values, main = main, xlab = xlab, ylab = ylab,
       pch = 19, cex = .25)
  abline(0, 0, col = 'grey')


}

par(mfrow = c(1,3))
boxplot(residuals(model.0), main = "Model 0 Residual vs. Fitted", ylab = "Residual")
resid.fitted.plot(model.1, main = "Model 1 Residual vs. Fitted")
resid.fitted.plot(model.2, main = "Model 2 Residual vs. Fitted")
```



In the plots above we notice that model 0 has residuals with mean 0 as expected. The residuals also appear to be symmetrically distributed with a few outliers on the left and right tail of the distribution. We see that model 1 has residuals that appear to be independent of the fitted values. They appear to have a constant mean of 0 and constant variance. Model 2 has residuals that also appear to have constant mean of 0. However, the residuals from model 2 may have decreasing variance with larger fitted values.

```
par(mfrow = c(1,3))
model.0 %>% residuals %>% qqnorm(main = "Model 0 Residual QQ Plot")
model.0 %>% residuals %>% qqline(col = 'red')

model.1 %>% residuals %>% qqnorm(main = "Model 1 Residual QQ Plot")
model.1 %>% residuals %>% qqline(col = 'red')

model.2 %>% residuals %>% qqnorm(main = "Model 2 Residual QQ Plot")
model.2 %>% residuals %>% qqline(col = 'red')
```



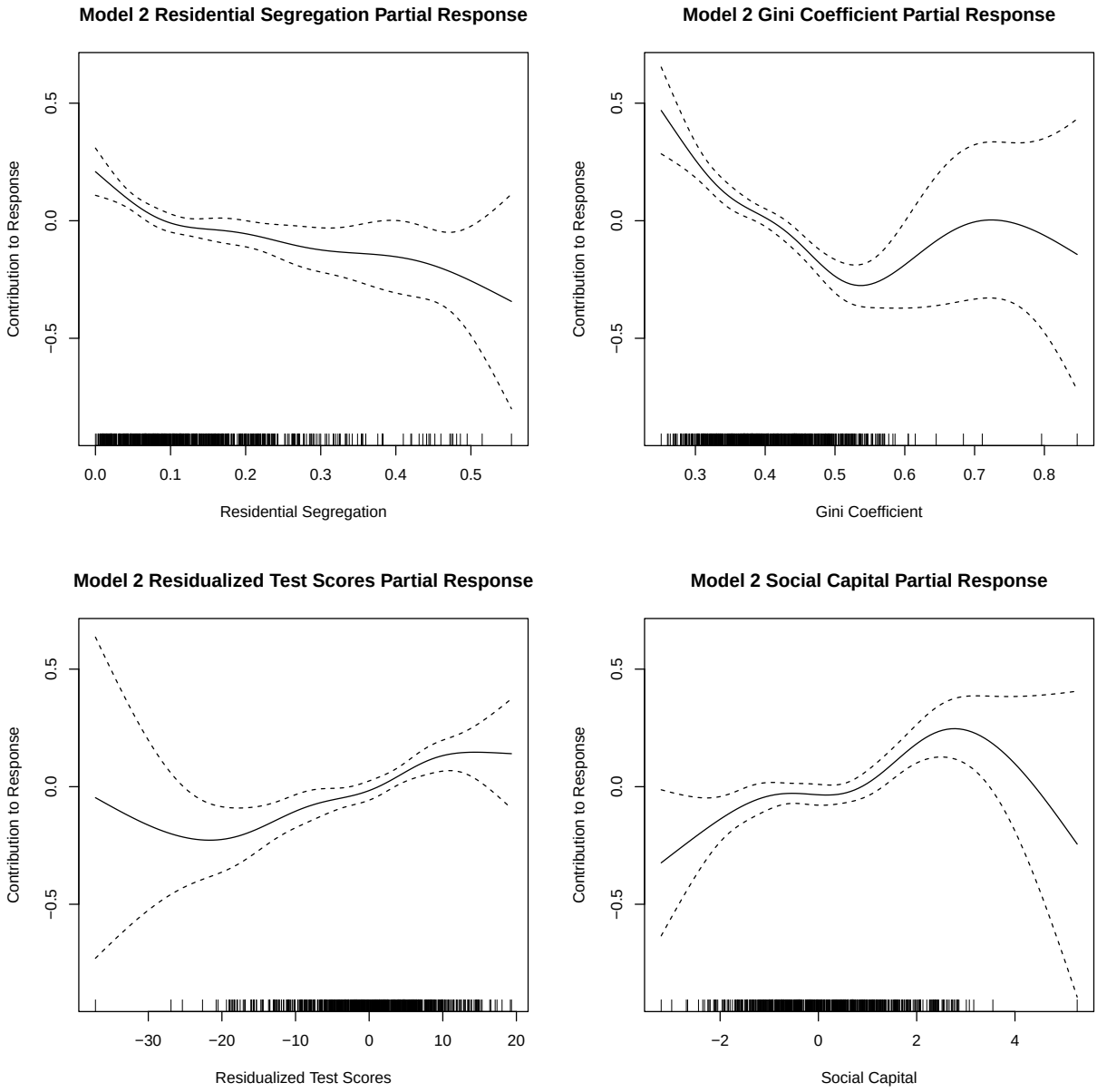
In the QQ plots above we see that all three models have somewhat normally distributed residuals. However, we do see that the residuals do stray from the quantile we would expect at the tails of the distributions. The residual qq plot for model 0 reveals that the distribution of residuals from model 0 likely has a wider left tail than a normal distribution, but a similar right tail to a normal distribution. The opposite is true for models 1 and 2 as we see that the right tail appears to be wider than a normal distribution, but the left tail is similar.

Problem 1 (b)

In order for the F-Test to produce valid results, both variables being measured must be normally distributed so that the sum of squared errors follows a chi-squared distribution. In the QQ plots observed in part A we see that this assumption does not appear to be fully met. However, the true distribution does not appear to stray greatly from the normal distribution.

Problem 1 (c)

```
par(mfrow = c(2,2))
plot.gam(model.2, all.terms = T, select = 1, ylab = "Contribution to Response",
         xlab = "Residential Segregation",
         main = "Model 2 Residential Segregation Partial Response")
plot.gam(model.2, all.terms = T, select = 2, ylab = "Contribution to Response",
         xlab = "Gini Coefficient",
         main = "Model 2 Gini Coefficient Partial Response")
plot.gam(model.2, all.terms = T, select = 3, ylab = "Contribution to Response",
         xlab = "Residualized Test Scores",
         main = "Model 2 Residualized Test Scores Partial Response")
plot.gam(model.2, all.terms = T, select = 4, ylab = "Contribution to Response",
         xlab = "Social Capital",
         main = "Model 2 Social Capital Partial Response")
```



In the partial response functions above we see that residential segregation has a negative relationship with log income mobility. Gini coefficient has an overall negative relationship with log income mobility as well, but it's relationship is less linear. Test scores with household income regressed out has a positive relationship with log income mobility. Social capital appears to have a quadratic shaped relationship with log income mobility. However, the relationship is positive for values of social capital reasonably near the mean

Problem 1 (d)

```
#model 1 vs 0 F test
anova(model.0, model.1, test = "F")
```

```
## Analysis of Variance Table
```

```
##
## Model 1: log(Mobility) ~ 1
## Model 2: log(Mobility) ~ Seg_racial + Gini + resid.scores + Social_capital
##   Res.Df     RSS Df Sum of Sq      F    Pr(>F)
## 1    634 145.389
## 2    630  79.964  4    65.425 128.86 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
#model 2 vs 1 F test
rss.m1 = sum(residuals(model.1)^2)
rss.m2 = sum(residuals(model.2)^2)
df1 = nrow(mob) - 5
df2 = nrow(mob) - 21

f.stat = (rss.m1 - rss.m2)/(df1 - df2)/(rss.m2/df2)
#1-pf(f.stat, df1-df2, df2)

anova(model.1, model.2, test = "F")
```

```
## Analysis of Variance Table
##
## Model 1: log(Mobility) ~ Seg_racial + Gini + resid.scores + Social_capital
## Model 2: log(Mobility) ~ s(Seg_racial, k = 6, fx = T) + s(Gini, k = 6,
##     fx = T) + s(resid.scores, k = 6, fx = T) + s(Social_capital,
##     k = 6, fx = T)
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1    630 79.964
## 2    614 74.007 16    5.9575 3.0892 4.622e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The first test we will run is an f-test between model 0 and model 1. Our null hypthesis is that the residual standard errors is the same between the two models and our alternative hypothesis is that the two models have different residual standard errors. Conducting this test under the assumption that an f test is appropriate with the distribution of our data and our sample size, we reject the null hypothesis and conclude that the predictor variables in model 1 do add a significant amount of predictive power to our model over model 0 ($F_{4,630} = 128.86, p < 2.2e-16$).

The second test we will run is an f test between model 2 and model 1. Our null hypothesis is that model 2 and model 1 have the same error and our alternative hypothesis is that model 2 and model 1 do not have equal residual standard errors. Conducting this test also under the assumption than an f-test is appropriate given the distribution of our errors and our sample size, we reject the null hypothesis and conclude that the linear terms in model 2 do improve our model performance over model 1 ($F_{16,614} = 3.09, p = 4.62e-5$).

Problem 1 (e)

```
pred.obj = predict(model.2, newdata = filter(mob, Name == "Pittsburgh"), se.fit = T)

pred.obj$fit[[1]] + (qnorm(.05)*pred.obj$se.fit)[[1]]
```

```
## [1] -2.739169
```

```
pred.obj$fit[[1]] + (qnorm(.95)*pred.obj$se.fit)[[1]]
```

```
## [1] -2.529207
```

From the code outpur above, we see that the log income mobility in pittsburgh falls in [-2.74, -2.53] with confidence level $\alpha = .1$.