Testing for Linearity Between Major Category & Income

Michael Šòdéké

ABSTRACT

Method

CASE I involved the use of the ANCOVA (one-way) model that helped generate values for the response variable. Parameter estimates, $\hat{\alpha}$ and $\hat{\beta}$, and the degrees of freedom, df, were calculated. The MSE was found after calculating the sum of squares. A hypothesis test was conducted to determine if there truly existed a linear relation between **major category** and **income**. Influence diagnostics were used to identify any data points that effected the outcome of the model. Finally, model validation was explored to determine which parameters were relevant to the model. The model was then adjusted to accommodate for the removal of one of the features.

Conclusion

CASE I revealed a weak linear relation between the regressors and the response variable. However, CASE I resutls only state that no *linear* relation exists, but does not state that no *other* relation exists. More adavnced mathematical modeling techniques are needed to discover what this true relation would be. Adjusted results in CASE II reached a similar conclusion: there is no sign of linearity in the model.

CASE I: PRE-MODEL VALIDATION

INTRODUCTION

CASE I involves the use of the ANCOVA (one-way) model that will help gnerate values for the response variable. From this model, the parameter estimates, $\hat{\alpha}$ and $\hat{\beta}$, and the degrees of freedom, df, can be calculated. Next, the MSE is found after calculating the sum of squares. A hypothesis test is conducted to determine if there truely exists a linear relation between **major category** and **income**. Influence diagnostics will be used to identify any data points that are effecting the outcome of the model. Finally, model validation is explored to determine which parameters are relevant to the model.

ANCOVA ONE-WAY (UNBALANCED) MODEL

Since **major category** is categorical and both **perc college jobs** and **perc non college jobs** are numeric, the best option is to use the ANCOVA (one-way) model with $q = 2$ covariates. The model can be expressed in mathematical notation as

$$y_{ij} = \mu + \alpha_i + \beta x_{ij1} + \beta x_{ij2} + \epsilon_{ij}$$

where:

$$i = 1, 2, ..., k \text{ observations,}$$
$$j = 1, 2, ..., n \text{ features}$$

Here is the model expressed in matrix notation

$$\mathbf{y} = \mathbf{Z}\alpha + \mathbf{X}\beta + \epsilon$$

where:

$$\mathbf{Z} = \begin{bmatrix} 1 & 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & 1 & 0 & \cdots & 0 \\ 1 & 1 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & 0 & 0 & \cdots & 1, \end{bmatrix}, \quad \alpha = \begin{bmatrix} \mu \\ \alpha_1 \end{bmatrix}$$

$$\mathbf{X} = \begin{bmatrix} x_{111} & x_{112} \\ x_{121} & x_{122} \\ \vdots & \vdots \\ x_{kn1} & x_{kn2} \end{bmatrix}, \qquad \beta = \begin{bmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \end{bmatrix}$$

Take note that $\mathbf{Z}$ is assumed to be rank-deficient and $\mathbf{X}$ is full-rank. Therefore $\mathbf{Z}$ will be reparameterized to become full-rank. Below is the ANCOVA model in code.

```
df <- college
df <- college %>% dplyr::select(median, major_category, perc_college_jobs, perc_non_college_jobs)
df <- na.omit(df)
df <- arrange(df, major_category)
df[,2] <- as.factor(df[,2])
lapply(df, head)
df %>% group_by(major_category, .add=TRUE) %>% group_nest()

Z <- as.matrix( ( model.matrix( median ~ major_category, data=df) ) ) # full-rank
X <- as.matrix( cbind(df$perc_college_jobs, df$perc_non_college_jobs) ) # full-rank
colnames(X) <- c("perc_college_jobs", "perc_non_college_jobs")
Y <- df$median
```

ESTIMATING PARAMETERS α & β

In order to generate values for the response variable, $\hat{y}$, α and β need to be estimated. In other words, to find α and β, the estimates $\hat{\alpha}$ and $\hat{\beta}$ must be calculated.

Now that $\mathbf{Z}$ is full-rank, there is no need to use the general inverse of $\mathbf{Z}$, $(\mathbf{Z}'\mathbf{Z})^{-}$. Therefore, $\hat{\alpha}$ can be calculated as follows

$$\hat{\alpha} = (\mathbf{Z}'\mathbf{Z})^{-}(\mathbf{Z}'\mathbf{y}) - (\mathbf{Z}'\mathbf{Z})^{-}(\mathbf{Z}'\mathbf{X}\hat{\beta})$$

The equation for $\hat{\alpha}$ can be used for estimating $\hat{\beta}$

$$\begin{aligned}
\hat{\beta} &= (\mathbf{X}'\mathbf{Z})[(\mathbf{Z}'\mathbf{Z})^{-}(\mathbf{Z}'\mathbf{y}) - (\mathbf{Z}'\mathbf{Z})(\mathbf{Z}'\mathbf{X}\hat{\beta})] + \mathbf{X}'\mathbf{X}\hat{\beta} = \mathbf{X}'\mathbf{y} \\
&= (\mathbf{X}'\mathbf{Z})[(\mathbf{Z}'\mathbf{Z})^{-}(\mathbf{Z}'\mathbf{y})] - (\mathbf{X}'\mathbf{Z})[(\mathbf{Z}'\mathbf{Z})^{-}(\mathbf{Z}'\mathbf{X}\hat{\beta})] + \mathbf{X}'\mathbf{X}\hat{\beta} = \mathbf{X}'\mathbf{y} \\
&= \mathbf{X}'[\mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-}(\mathbf{Z}'\mathbf{y})] + \mathbf{X}'[\mathbf{I} - \mathbf{Z}(\mathbf{Z}'\mathbf{Z})'\mathbf{Z}']\mathbf{X}\hat{\beta} = \mathbf{X}'\mathbf{y} \\
&= \mathbf{X}'(\mathbf{P})\mathbf{y} + \mathbf{X}'[\mathbf{I} - \mathbf{P}]\mathbf{X}\hat{\beta} = \mathbf{X}'\mathbf{y} \\
&= \mathbf{X}'[\mathbf{I} - \mathbf{P}]\mathbf{X}\hat{\beta} = \mathbf{X}'\mathbf{y} - \mathbf{X}'(\mathbf{P})\mathbf{y} \\
&= \mathbf{X}'[\mathbf{I} - \mathbf{P}]\mathbf{X}\hat{\beta} = \mathbf{X}'[\mathbf{I} - \mathbf{P}]\mathbf{y} \\
&= \mathbf{E}_{xx}^{-}\mathbf{e}_{xy}
\end{aligned}$$

For convenience, the ANCOVA model is reconstructed as

$$\mathbf{y} = \mathbf{U}\mathbf{\Gamma} + \epsilon$$

where:

$$\mathbf{U} = \begin{bmatrix} \mathbf{Z}, & \mathbf{X} \end{bmatrix}, \quad \mathbf{\Gamma} = \begin{bmatrix} \hat{\alpha} \\ \hat{\beta} \end{bmatrix}$$

```
P <- Z %*% ginv(t(Z) %*% Z) %*% t(Z)
I <- diag(dim(P)[1])
B_ <- solve(t(X) %*% (I - P) %*% X) %*% (t(X) %*% (I - P) %*% Y); colnames(B_) <- "Estimates"
A_ <- ginv(t(Z) %*% Z) %*% (t(Z) %*% Y) - ginv(t(Z) %*% Z) %*% (t(Z) %*% (X %*% B_))
rownames(A_) <- colnames(Z)
rownames(B_) <- colnames(X)
U <- cbind(Z, X)
GAMMA <- rbind(A_, B_)
Y_ <- U %*% GAMMA
```

DEGREES OF FREEDOM

The degrees of freedom for the error is $df = n - k$. This can be used for generating values for the MSE, σ^2.

```
n <- nrow(U); k <- qr(U)$rank
```

SUM OF SQUARES

Next, the sum of squares are calculated

$$\begin{aligned}
SSE &= SS_{res} = (\mathbf{y} - \hat{\mathbf{y}})'(\mathbf{y} - \hat{\mathbf{y}}) \\
SSR &= SS_{reg} = (\mathbf{y} - \bar{\mathbf{y}})'(\mathbf{y} - \bar{\mathbf{y}}) \\
SST &= SS_{tot} = SS_{res} + SS_{reg} \\
R^2 &= \frac{SS_{res}}{SS_{reg}}
\end{aligned}$$

```
ss.res <- (t(Y) %*% Y) - (t(GAMMA) %*% t(U) %*% Y)
ss.reg <- t(Y_ - mean(Y)) %*% (Y_ - mean(Y))
ss.tot <- ss.res + ss.reg
R.2 <- ss.res/ss.reg
```

MEAN SQUARES

Using the formula for the SS_{res}, calculate the MSE

$$\begin{aligned} MSE &= \sigma^2 \\ &= \frac{SS_{res}}{k\,(n-1)} \end{aligned}$$

Now use the MSE to calculate the standard errors for the estimates in Γ

$$SE_\Gamma = \sqrt{diag(\sigma^2(\mathbf{U}'\mathbf{U})^{-})}$$

```
ms.res <- ss.res/(n-k)
SE.. <- matrix( sqrt( diag( ms.res[1] * solve(t(U) %*% U) ) ) );  colnames(SE..) <- "Std. Error"
```

HYPOTHESIS TEST

To determine if there exists a linear relation between the regressors and the response variable, a hypothesis test is needed, namely a t-test.

The hypothesis test is constructed as follows

$$H_0 : \Gamma = 0 \text{ no linear relation.}$$
$$H_a : \Gamma \neq 0 \text{ linear relation exists.}$$

where:

$$\Gamma = \begin{bmatrix} \alpha \\ \beta_1 \\ \beta_2 \end{bmatrix}$$

Next, construct the t-test

$$t_{test} = \frac{\Gamma - 0}{SE_\Gamma}$$

Reject the null hypothesis, H_0, if $t_{test} \geq t_{1-\alpha,n-k}$.

```
t.B_ <- GAMMA / SE..;  colnames(t.B_) <- "t value"
t.alpha <- qt(1-(0.05/2), df=n-k)
p_val.B_ <- 2*pt(-abs(t.B_), df=n-k);  colnames(p_val.B_) <- "Pr(>|t|)"
p_val.alpha <- 2*pt(-abs(t.alpha), df=n-k)
```

```
cbind(t.B_, p_val.B_)

                                                       t value     Pr(>|t|)
(Intercept)                                         5.49281690 1.600036e-07
major_categoryArts                                 -1.17028901 2.436925e-01
major_categoryBiology & Life Science               -0.05664043 9.549050e-01
major_categoryBusiness                              0.96319248 3.369609e-01
major_categoryCommunications & Journalism          -0.31068316 7.564616e-01
major_categoryComputers & Mathematics              -1.90276400 5.893763e-02
major_categoryEducation                            -1.20920870 2.284358e-01
major_categoryEngineering                          -0.83124453 4.071225e-01
major_categoryHealth                               -0.77897389 4.371904e-01
major_categoryHumanities & Liberal Arts            -2.06184950 4.090272e-02
major_categoryIndustrial Arts & Consumer Services -0.57984119 5.628691e-01
major_categoryInterdisciplinary                    -1.54316186 1.248439e-01
major_categoryLaw & Public Policy                  -1.13011064 2.601867e-01
major_categoryPhysical Sciences                    -0.70983383 4.788806e-01
major_categoryPsychology & Social Work             -1.04710528 2.966917e-01
major_categorySocial Science                       -0.82522474 4.105205e-01
perc_college_jobs                                  -1.12842757 2.608943e-01
perc_non_college_jobs                              -0.22803152 8.199242e-01
```

INFLUENCIAL OBSERVATIONS AND LEVERAGE

To prevent model misspecification, employ influence diagnostic measures to identify outliers effecting the one-way ANCOVA model. These measures include leverages, h_{ii}, residuals, $\hat{\epsilon}_i$, studentized residuals, $\hat{r}_i$, and Cook's Distance, D_i.

These diagnstic tools are calculated below

$$\mathbf{H} = \mathbf{X}(\mathbf{X}'\mathbf{X})\mathbf{X}'$$

$$h_{ii} = diag(\mathbf{H}),\ h_{ii} \geq \frac{2(k+1)}{n}$$

$$\hat{\epsilon}_i = \mathbf{y} - \hat{\mathbf{y}}$$

$$\hat{r}_i = \frac{\hat{\epsilon}_i}{\sqrt{\sigma^2(1-h_{ii})}}$$

$$D_i = \left(\frac{\hat{r}^2}{k+1}\right)\left(\frac{h_{ii}}{1-h_{ii}}\right)$$

```
H <- X %*% solve(t(X) %*% X) %*% t(X)

create.hii <- function(mat=NULL)
{
        vec <- matrix(rep(0, nrow(mat)))
        for (i in 1:nrow(mat))
                for (j in 1:ncol(mat))
                        if ( i == j)
                                vec[i] <- mat[i,j]
```

```
        vec
}

hii <- create.hii(mat=H) # leverage
e_ <- Y - Y_; # residual
r_ <- e_/sqrt(ms.res[1] * (1 - hii)) # studentized residual
D_ <- ((r_^2)/(k+1)) * ((hii)/(1-hii)) # Cook's distance
inf.measures <- cbind(head(Y), head(Y_), head(e_), head(hii), head(r_),  head(D_))
colnames(inf.measures) <- c("yi", "yi_", "ei_", "hii", "ri_", "Di_")
infl <- c( "hii_LEVERAGE"=(2*(k + 1))/n )
```

```
list(infl, inf.measures)

[[1]]
hii_LEVERAGE
   0.2209302

[[2]]
     yi      yi_        ei_         hii         ri_          Di_
1 33000 42216.23  -9216.231 0.012015861 -0.81692462 4.271843e-04
2 58000 45641.31  12358.687 0.008332640  1.09343497 5.287482e-04
3 40000 43851.68  -3851.682 0.006055751 -0.34038706 3.715339e-05
4 65000 45094.40  19905.600 0.009234717  1.76194968 1.522952e-03
5 45000 43943.40   1056.604 0.005921398  0.09336961 2.733135e-06
6 31000 41124.02 -10124.018 0.026722668 -0.90414531 1.181316e-03
```

In FIGURE 1 the *residuals vs. fitted* graph shows a faint pattern that appears to be either *stochastic* or *sinusoidal*. In addition, hardly any outliers exist in the *standardized residuals vs. leverage* graph.

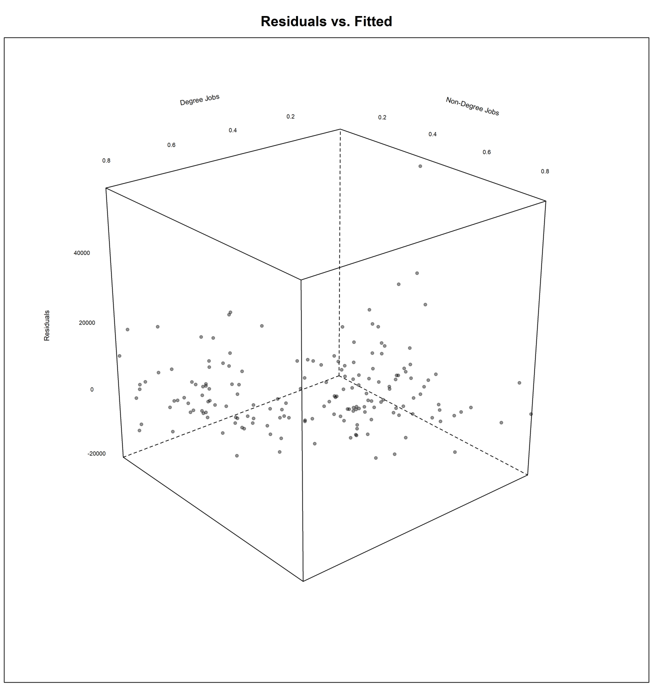

Figure 1:

RESULTS

A comparison of the results from the above procedure, with **r**'s built-in functions, show that they are identical and that the above procedure is correct. Further, these results reveal that a linear relation between the regressors and the response variables is insignificant.

```
summary( lm(median ~ major_category + perc_college_jobs +
            perc_non_college_jobs, data=df) )$coef # test for [linear relation] with t-test
```

```
                                                 Estimate Std. Error
(Intercept)                                    49602.2736   9030.389
major_categoryArts                             -6341.6708   5418.893
major_categoryBiology & Life Science            -267.0807   4715.372
major_categoryBusiness                          4638.8226   4816.091
major_categoryCommunications & Journalism      -2088.8122   6723.287
major_categoryComputers & Mathematics          -9771.8766   5135.622
major_categoryEducation                        -5535.3195   4577.638
major_categoryEngineering                      -3472.6624   4177.666
major_categoryHealth                           -3802.9789   4882.036
major_categoryHumanities & Liberal Arts        -9685.5076   4697.485
major_categoryIndustrial Arts & Consumer Services  -3244.5961   5595.663
major_categoryInterdisciplinary               -18801.7432  12183.909
major_categoryLaw & Public Policy              -7089.2732   6273.079
major_categoryPhysical Sciences                -3608.9103   5084.162
major_categoryPsychology & Social Work         -5585.4128   5334.146
major_categorySocial Science                   -4318.2703   5232.842
perc_college_jobs                              -9849.5558   8728.567
perc_non_college_jobs                          -2473.7562  10848.308
                                                  t value     Pr(>|t|)
(Intercept)                                     5.49281690 1.600036e-07
major_categoryArts                             -1.17028901 2.436925e-01
major_categoryBiology & Life Science           -0.05664043 9.549050e-01
major_categoryBusiness                          0.96319248 3.369609e-01
major_categoryCommunications & Journalism      -0.31068316 7.564616e-01
major_categoryComputers & Mathematics          -1.90276400 5.893763e-02
major_categoryEducation                        -1.20920870 2.284358e-01
major_categoryEngineering                      -0.83124453 4.071225e-01
major_categoryHealth                           -0.77897389 4.371904e-01
major_categoryHumanities & Liberal Arts        -2.06184950 4.090272e-02
major_categoryIndustrial Arts & Consumer Services -0.57984119 5.628691e-01
major_categoryInterdisciplinary                -1.54316186 1.248439e-01
major_categoryLaw & Public Policy              -1.13011064 2.601867e-01
major_categoryPhysical Sciences                -0.70983383 4.788806e-01
major_categoryPsychology & Social Work         -1.04710528 2.966917e-01
major_categorySocial Science                   -0.82522474 4.105205e-01
perc_college_jobs                              -1.12842757 2.608943e-01
perc_non_college_jobs                          -0.22803152 8.199242e-01
```

```
list(Y.values.vs.fitted=head(cbind(Y, Y_)),
     predictors=head(cbind(Z, X)),
     coefficients=cbind(GAMMA, SE.., t.B_, p_val.B_))$coefficients
```

```
                                                Estimates Std. Error
(Intercept)                                    49602.2736   9030.389
major_categoryArts                             -6341.6708   5418.893
major_categoryBiology & Life Science            -267.0807   4715.372
```

```
major_categoryBusiness                                 4638.8226   4816.091
major_categoryCommunications & Journalism             -2088.8122   6723.287
major_categoryComputers & Mathematics                 -9771.8766   5135.622
major_categoryEducation                               -5535.3195   4577.638
major_categoryEngineering                             -3472.6624   4177.666
major_categoryHealth                                  -3802.9789   4882.036
major_categoryHumanities & Liberal Arts               -9685.5076   4697.485
major_categoryIndustrial Arts & Consumer Services     -3244.5961   5595.663
major_categoryInterdisciplinary                      -18801.7432  12183.909
major_categoryLaw & Public Policy                      -7089.2732   6273.079
major_categoryPhysical Sciences                       -3608.9103   5084.162
major_categoryPsychology & Social Work                -5585.4128   5334.146
major_categorySocial Science                          -4318.2703   5232.842
perc_college_jobs                                     -9849.5558   8728.567
perc_non_college_jobs                                 -2473.7562  10848.308
                                                         t value     Pr(>|t|)
(Intercept)                                           5.49281690 1.600036e-07
major_categoryArts                                   -1.17028901 2.436925e-01
major_categoryBiology & Life Science                 -0.05664043 9.549050e-01
major_categoryBusiness                                0.96319248 3.369609e-01
major_categoryCommunications & Journalism            -0.31068316 7.564616e-01
major_categoryComputers & Mathematics                -1.90276400 5.893763e-02
major_categoryEducation                              -1.20920870 2.284358e-01
major_categoryEngineering                            -0.83124453 4.071225e-01
major_categoryHealth                                 -0.77897389 4.371904e-01
major_categoryHumanities & Liberal Arts              -2.06184950 4.090272e-02
major_categoryIndustrial Arts & Consumer Services -0.57984119 5.628691e-01
major_categoryInterdisciplinary                      -1.54316186 1.248439e-01
major_categoryLaw & Public Policy                    -1.13011064 2.601867e-01
major_categoryPhysical Sciences                      -0.70983383 4.788806e-01
major_categoryPsychology & Social Work               -1.04710528 2.966917e-01
major_categorySocial Science                         -0.82522474 4.105205e-01
perc_college_jobs                                    -1.12842757 2.608943e-01
perc_non_college_jobs                                -0.22803152 8.199242e-01
```

```
options(scipen = 999)
list(infl, inf.measures)
options(scipen = 0)
```

```
[[1]]
hii_LEVERAGE
   0.2209302

[[2]]
     yi      yi_        ei_         hii         ri_            Di_
1 33000 42216.23  -9216.231 0.012015861 -0.81692462 0.000427184315
2 58000 45641.31  12358.687 0.008332640  1.09343497 0.000528748242
3 40000 43851.68  -3851.682 0.006055751 -0.34038706 0.000037153395
4 65000 45094.40  19905.600 0.009234717  1.76194968 0.001522951989
5 45000 43943.40   1056.604 0.005921398  0.09336961 0.000002733135
6 31000 41124.02 -10124.018 0.026722668 -0.90414531 0.001181315954
```

Though there exists no linear relation in the model, that does not elimminate other kinds of relations. Note that in FIGURE 2 the data seems to take the form of either a *stochastic* or *sinusoidal* function. A new question then arises, "why do median earnings appear volitle as the percentage of jobs, requiring degrees and those that don't, increases?"

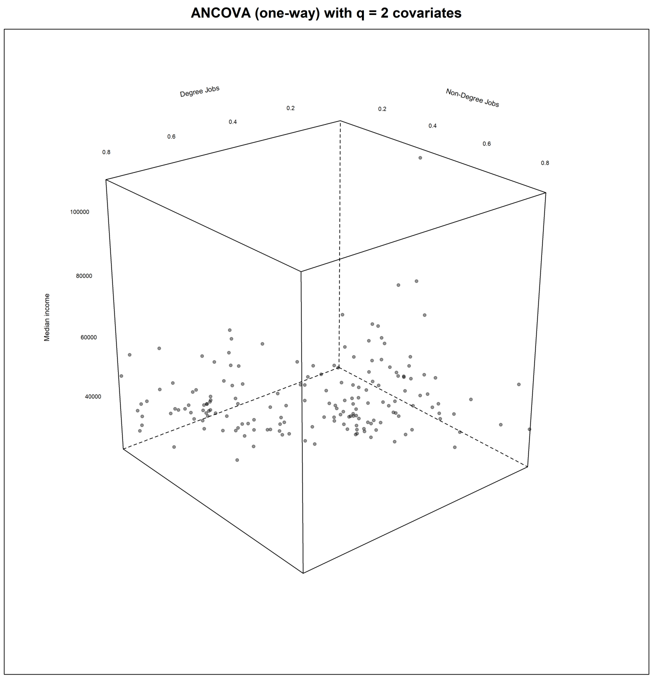

Figure 2:

MODEL VALIDATION

Three new models are compared to determine which features are relevant and which are irrelevant. To compare these models, use the *vif* and *anova* functions in **r**.

```
fit1 <- lm(median ~ major_category, data=df)
fit2 <- lm(median ~ major_category + perc_college_jobs, data=df)
fit3 <- lm(median ~ major_category + perc_college_jobs + perc_non_college_jobs, data=df)
anova(fit1, fit2, fit3)
```

```
Analysis of Variance Table

Model 1: median ~ major_category
Model 2: median ~ major_category + perc_college_jobs
Model 3: median ~ major_category + perc_college_jobs + perc_non_college_jobs
  Res.Df        RSS Df Sum of Sq      F Pr(>F)
1    156 2.0238e+10
2    155 1.9845e+10  1 392727096 3.0486 0.0828 .
3    154 1.9839e+10  1   6698576 0.0520 0.8199
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
fit4 <- aov(median ~ major_category + perc_college_jobs, data = df)
fit5 <- aov(median ~ major_category * perc_college_jobs, data = df)
anova(fit4, fit5)
```

```
Analysis of Variance Table

Model 1: median ~ major_category + perc_college_jobs
Model 2: median ~ major_category * perc_college_jobs
  Res.Df        RSS Df  Sum of Sq      F  Pr(>F)
1    155 1.9845e+10
2    141 1.6418e+10 14 3427295991 2.1024 0.01493 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
vif( lm(median ~ major_category + perc_college_jobs +
                 perc_non_college_jobs, data=df) ) # variance inflation factor
```

```
                          GVIF Df GVIF^(1/(2*Df))
major_category        1.260321 15        1.007742
perc_college_jobs     3.962138  1        1.990512
perc_non_college_jobs 3.876783  1        1.968955
```

The variance inflation factors for **per college jobs** and **perc non college jobs** are moderately high, suggesting a removal of at least one of the two to get a more accurate model. The adjusted model will now be explored in CASE II.

CASE II: POST-MODEL VALIDATION

INTRODUCTION

I began my analysis by first constructing the ANCOVA model with all relevant features that would assist in determining if there is truely an association between college major category and income. However, after performing model validation, I removed **perc non college jobs** from the model to get a more accurate result. I now explore this adjusted model.

ANCOVA ONE-WAY (UNBALANCED) MODEL

The ANCOVA (one-way) model will still bes used, but the covariates are now $q = 1$. In mathematical notation, the model is expressed as

$$y_{ij} = \mu + \alpha_i + \beta x_{ij1} + \epsilon_{ij}$$

where:

$$i = 1, 2, ..., k \text{ observations,}$$
$$j = 1, 2, ..., n \text{ features}$$

Here is the same model in matrix notation

$$\mathbf{y} = \mathbf{Z}\alpha + \mathbf{X}\beta + \epsilon$$

where:

$$\mathbf{Z} = \begin{bmatrix} 1 & 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & 1 & 0 & \cdots & 0 \\ 1 & 1 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & 0 & 0 & \cdots & 1, \end{bmatrix}, \quad \alpha = \begin{bmatrix} \mu \\ \alpha_1 \end{bmatrix}$$

$$\mathbf{X} = \begin{bmatrix} x_{111} \\ x_{121} \\ \vdots \\ x_{kn1} \end{bmatrix}, \qquad \beta = \begin{bmatrix} \hat{\beta}_1 \end{bmatrix}$$

$\mathbf{Z}$ is assumed to be rank-deficient and $\mathbf{X}$ is full-rank. As before, $\mathbf{Z}$ will be reparameterized.

```
df <- college
df <- college %>% dplyr::select(median, major_category, perc_college_jobs)
df <- na.omit(df)
df <- arrange(df, major_category)
df[,2] <- as.factor(df[,2])
lapply(df, head)
df %>% group_by(major_category, .add=TRUE) %>% group_nest()
Z <- as.matrix( ( model.matrix( median ~ major_category, data=df) ) ) # full-rank
X <- as.matrix( cbind(df$perc_college_jobs) ) # full-rank
colnames(X) <- c("perc_college_jobs")
Y <- df$median
```

ESTIMATING PARAMETERS α & β

Calculations for parameters $\hat{\alpha}$ and $\hat{\beta}$ are the same as in CASE I

$$\hat{\alpha} = (\mathbf{Z}^{'}\mathbf{Z})^{-}(\mathbf{Z}^{'}\mathbf{y}) - (\mathbf{Z}^{'}\mathbf{Z})^{-}(\mathbf{Z}^{'}\mathbf{X}\hat{\beta})$$

$$\begin{aligned}
\hat{\beta} &= (\mathbf{X}^{'}\mathbf{Z})[(\mathbf{Z}^{'}\mathbf{Z})^{-}(\mathbf{Z}^{'}\mathbf{y}) - (\mathbf{Z}^{'}\mathbf{Z})(\mathbf{Z}^{'}\mathbf{X}\hat{\beta})] + \mathbf{X}^{'}\mathbf{X}\hat{\beta} = \mathbf{X}^{'}\mathbf{y} \\
&= (\mathbf{X}^{'}\mathbf{Z})[(\mathbf{Z}^{'}\mathbf{Z})^{-}(\mathbf{Z}^{'}\mathbf{y})] - (\mathbf{X}^{'}\mathbf{Z})[(\mathbf{Z}^{'}\mathbf{Z})^{-}(\mathbf{Z}^{'}\mathbf{X}\hat{\beta})] + \mathbf{X}^{'}\mathbf{X}\hat{\beta} = \mathbf{X}^{'}\mathbf{y} \\
&= \mathbf{X}^{'}[\mathbf{Z}(\mathbf{Z}^{'}\mathbf{Z})^{-}(\mathbf{Z}^{'}\mathbf{y})] + \mathbf{X}^{'}[\mathbf{I} - \mathbf{Z}(\mathbf{Z}^{'}\mathbf{Z})^{'}\mathbf{Z}^{'}]\mathbf{X}\hat{\beta} = \mathbf{X}^{'}\mathbf{y} \\
&= \mathbf{X}^{'}(\mathbf{P})\mathbf{y} + \mathbf{X}^{'}[\mathbf{I} - \mathbf{P}]\mathbf{X}\hat{\beta} = \mathbf{X}^{'}\mathbf{y} \\
&= \mathbf{X}^{'}[\mathbf{I} - \mathbf{P}]\mathbf{X}\hat{\beta} = \mathbf{X}^{'}\mathbf{y} - \mathbf{X}^{'}(\mathbf{P})\mathbf{y} \\
&= \mathbf{X}^{'}[\mathbf{I} - \mathbf{P}]\mathbf{X}\hat{\beta} = \mathbf{X}^{'}[\mathbf{I} - \mathbf{P}]\mathbf{y} \\
&= \mathbf{E}_{xx}^{-}\mathbf{e}_{xy}
\end{aligned}$$

A more compact form of the model can be expressed as

$$\mathbf{y} = \mathbf{U}\boldsymbol{\Gamma} + \epsilon$$

where:

$$\mathbf{U} = \left[\begin{array}{cc} \mathbf{Z}, & \mathbf{X} \end{array}\right], \quad \boldsymbol{\Gamma} = \left[\begin{array}{c} \hat{\alpha} \\ \hat{\beta} \end{array}\right]$$

```
P <- Z %*% ginv(t(Z) %*% Z) %*% t(Z)
I <- diag(dim(P)[1])
B_ <- solve(t(X) %*% (I - P) %*% X) %*% (t(X) %*% (I - P) %*% Y); colnames(B_) <- "Estimates"
A_ <- ginv(t(Z) %*% Z) %*% (t(Z) %*% Y) - ginv(t(Z) %*% Z) %*% (t(Z) %*% (X %*% B_))
rownames(A_) <- colnames(Z)
rownames(B_) <- colnames(X)
U <- cbind(Z, X)
GAMMA <- rbind(A_, B_)
Y_ <- U %*% GAMMA
```

DEGREES OF FREEDOM

The degrees of freedom for the error term are the same as in CASE I. Therefore, $df = n - k$

```
n <- nrow(U); k <- qr(U)$rank
```

SUM OF SQUARES

Given the previous calculations, the *sum of squares* can be calculated as such

$$\begin{aligned}
SSE &= SS_{res} = (\mathbf{y} - \hat{\mathbf{y}})^{'} (\mathbf{y} - \hat{\mathbf{y}}) \\
SSR &= SS_{reg} = (\mathbf{y} - \bar{\mathbf{y}})^{'} (\mathbf{y} - \bar{\mathbf{y}}) \\
SST &= SS_{tot} = SS_{res} + SS_{reg} \\
R^2 &= \frac{SS_{res}}{SS_{reg}}
\end{aligned}$$

```
ss.res <- (t(Y) %*% Y) - (t(GAMMA) %*% t(U) %*% Y)
ss.reg <- t(Y_ - mean(Y)) %*% (Y_ - mean(Y))
ss.tot <- ss.res + ss.reg
R.2 <- ss.res/ss.reg
```

MEAN SQUARES

The sum of squared residuals, SSE or SS_{res} can be used to find the MSE

$$\begin{aligned} MSE &= \sigma^2 \\ &= \frac{SS_{res}}{k\,(n-1)} \end{aligned}$$

Now the standard error of Γ, SE_Γ, can be found

$$SE_\Gamma = \sqrt{diag(\sigma^2(\mathbf{U}'\mathbf{U})^{-})}$$

```
ms.res <- ss.res/(n-k)
SE.. <- matrix( sqrt( diag( ms.res[1] * solve(t(U) %*% U) ) ) );  colnames(SE..) <- "Std. Error"
```

HYPOTHESIS TEST

A t-test is used to test for any such linear relation between the regressors and the response variable. The hypothesis test and t-test are constructed below

$$\begin{aligned} H_0 &: \Gamma = 0 \text{ no linear relation.} \\ H_a &: \Gamma \neq 0 \text{ linear relation exists.} \end{aligned}$$

where:

$$\Gamma = \begin{bmatrix} \alpha \\ \beta_1 \end{bmatrix}$$

$$t_{test} = \frac{\Gamma - 0}{SE_\Gamma}$$

The null hypothesis, H_0, is rejected if $t_{test} \geq t_{1-\alpha,n-k}$. In CASE I, results from the t-test revealed no significant *linear* relation between the regressors and the response variable. Therefore, it can be assumed that this current t-test will also reach the same conclusion, but with different results.

```
t.B_ <- GAMMA / SE..;  colnames(t.B_) <- "t value"
t.alpha <- qt(1-(0.05/2), df=n-k)
p_val.B_ <- 2*pt(-abs(t.B_), df=n-k);  colnames(p_val.B_) <- "Pr(>|t|)"
p_val.alpha <- 2*pt(-abs(t.alpha), df=n-k)
```

```
cbind(t.B_, p_val.B_)
                                                    t value     Pr(>|t|)
(Intercept)                                     11.01620338 3.296774e-21
major_categoryArts                              -1.15982207 2.479047e-01
major_categoryBiology & Life Science            -0.06175313 9.508390e-01
major_categoryBusiness                           0.98465353 3.263288e-01
```

```
major_categoryCommunications & Journalism         -0.31194760 7.554996e-01
major_categoryComputers & Mathematics             -1.92375132 5.621843e-02
major_categoryEducation                           -1.22078662 2.240207e-01
major_categoryEngineering                         -0.81882249 4.141448e-01
major_categoryHealth                              -0.79449323 4.281231e-01
major_categoryHumanities & Liberal Arts           -2.06100962 4.097298e-02
major_categoryIndustrial Arts & Consumer Services -0.58638810 5.584677e-01
major_categoryInterdisciplinary                   -1.53196755 1.275689e-01
major_categoryLaw & Public Policy                 -1.12162214 2.637577e-01
major_categoryPhysical Sciences                   -0.70927770 4.792177e-01
major_categoryPsychology & Social Work            -1.04421726 2.980107e-01
major_categorySocial Science                      -0.81217605 4.179362e-01
perc_college_jobs                                 -1.75138343 8.185791e-02
```

INFLUENCIAL OBSERVATIONS AND LEVERAGE

It can be assumed, as in CASE I, that there exits no outliers whose presence would have a huge impact on the model.

```
H <- X %*% solve(t(X) %*% X) %*% t(X)
create.hii <- function(mat=NULL)
{
        vec <- matrix(rep(0, nrow(mat)))
        for (i in 1:nrow(mat))
                for (j in 1:ncol(mat))
                        if ( i == j)
                                vec[i] <- mat[i,j]
        vec
}
hii <- create.hii(mat=H) # leverage
e_ <- Y - Y_; # residual
r_ <- e_/sqrt(ms.res[1] * (1 - hii)) # studentized residual
D_ <- ((r_^2)/(k+1)) * ((hii)/(1-hii)) # Cook's distance
inf.measures <- cbind(head(Y), head(Y_), head(e_), head(hii), head(r_),  head(D_))
colnames(inf.measures) <- c("yi", "yi_", "ei_", "hii", "ri_", "Di_")
infl <- c( "hii_LEVERAGE"=(2*(k + 1))/n )
```

```
list(infl, inf.measures)

[[1]]
hii_LEVERAGE
   0.2209302

[[2]]
     yi      yi_        ei_         hii         ri_          Di_
1 33000 42216.23  -9216.231 0.012015861 -0.81692462 4.271843e-04
2 58000 45641.31  12358.687 0.008332640  1.09343497 5.287482e-04
3 40000 43851.68  -3851.682 0.006055751 -0.34038706 3.715339e-05
4 65000 45094.40  19905.600 0.009234717  1.76194968 1.522952e-03
5 45000 43943.40   1056.604 0.005921398  0.09336961 2.733135e-06
6 31000 41124.02 -10124.018 0.026722668 -0.90414531 1.181316e-03
```

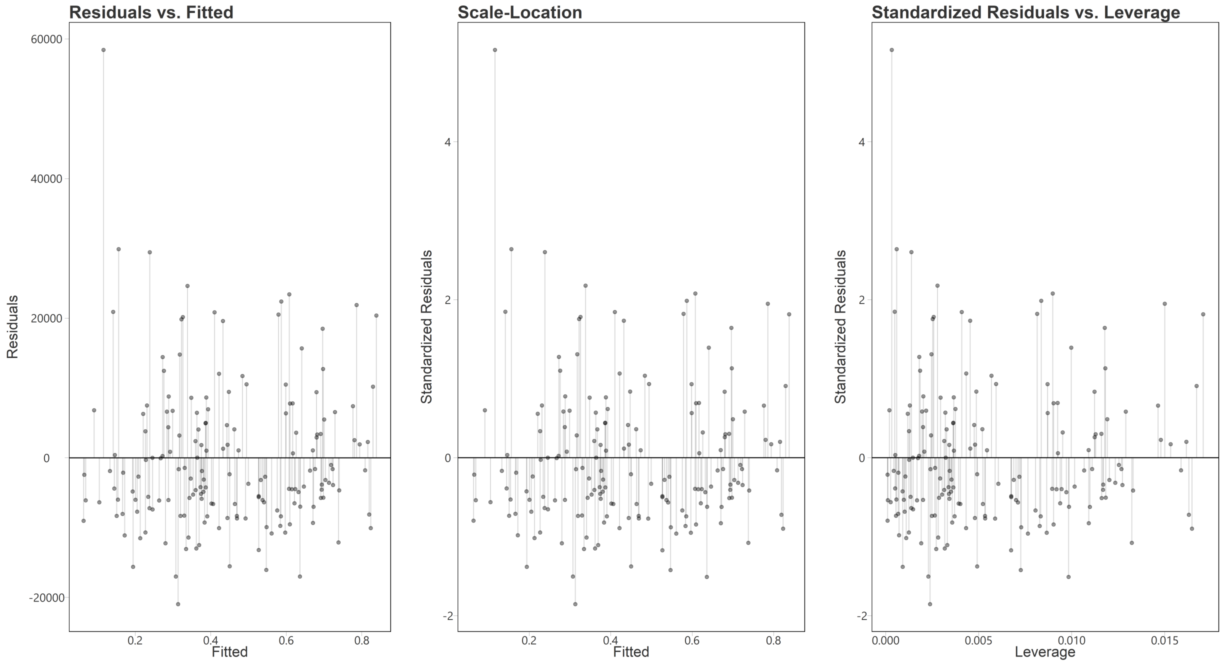

Figure 3:

RESULTS

Again, results show very low signs of linearity between the regressors and the response variable. More sophisticated mathematical modeling techniques will be needed for discovering the **proper** relation between the regressors and response variable.

```
summary( lm(median ~ major_category + perc_college_jobs, data=df) )$coef
```

```
                                              Estimate Std. Error
(Intercept)                                 47797.9954   4338.881
major_categoryArts                          -6247.4680   5386.575
major_categoryBiology & Life Science         -290.2299   4699.841
major_categoryBusiness                       4715.9758   4789.477
major_categoryCommunications & Journalism   -2090.8879   6702.689
major_categoryComputers & Mathematics       -9835.0563   5112.436
major_categoryEducation                     -5568.4043   4561.325
major_categoryEngineering                   -3400.4889   4152.901
major_categoryHealth                        -3861.5159   4860.351
major_categoryHumanities & Liberal Arts     -9645.0093   4679.750
major_categoryIndustrial Arts & Consumer Services  -3270.5064   5577.375
major_categoryInterdisciplinary            -18290.7687  11939.397
major_categoryLaw & Public Policy           -7001.1519   6241.988
major_categoryPhysical Sciences             -3594.7709   5068.214
major_categoryPsychology & Social Work      -5550.6843   5315.641
major_categorySocial Science                -4223.6223   5200.378
perc_college_jobs                           -8169.2759   4664.470
```

```
                                                   t value     Pr(>|t|)
(Intercept)                                    11.01620338 3.296774e-21
major_categoryArts                             -1.15982207 2.479047e-01
major_categoryBiology & Life Science           -0.06175313 9.508390e-01
major_categoryBusiness                          0.98465353 3.263288e-01
major_categoryCommunications & Journalism      -0.31194760 7.554996e-01
major_categoryComputers & Mathematics          -1.92375132 5.621843e-02
major_categoryEducation                        -1.22078662 2.240207e-01
major_categoryEngineering                      -0.81882249 4.141448e-01
major_categoryHealth                           -0.79449323 4.281231e-01
major_categoryHumanities & Liberal Arts        -2.06100962 4.097298e-02
major_categoryIndustrial Arts & Consumer Services -0.58638810 5.584677e-01
major_categoryInterdisciplinary                -1.53196755 1.275689e-01
major_categoryLaw & Public Policy              -1.12162214 2.637577e-01
major_categoryPhysical Sciences                -0.70927770 4.792177e-01
major_categoryPsychology & Social Work         -1.04421726 2.980107e-01
major_categorySocial Science                   -0.81217605 4.179362e-01
perc_college_jobs                              -1.75138343 8.185791e-02
```

```
list(Y.values.vs.fitted=head(cbind(Y, Y_)), predictors=head(cbind(Z, X)),
     coefficients=cbind(GAMMA, SE.., t.B_, p_val.B_))$coefficients
```

```
                                                  Estimates Std. Error
(Intercept)                                      47797.9954   4338.881
major_categoryArts                               -6247.4680   5386.575
major_categoryBiology & Life Science              -290.2299   4699.841
major_categoryBusiness                            4715.9758   4789.477
major_categoryCommunications & Journalism        -2090.8879   6702.689
major_categoryComputers & Mathematics            -9835.0563   5112.436
major_categoryEducation                          -5568.4043   4561.325
major_categoryEngineering                        -3400.4889   4152.901
major_categoryHealth                             -3861.5159   4860.351
major_categoryHumanities & Liberal Arts          -9645.0093   4679.750
major_categoryIndustrial Arts & Consumer Services -3270.5064   5577.375
major_categoryInterdisciplinary                 -18290.7687  11939.397
major_categoryLaw & Public Policy                -7001.1519   6241.988
major_categoryPhysical Sciences                  -3594.7709   5068.214
major_categoryPsychology & Social Work           -5550.6843   5315.641
major_categorySocial Science                     -4223.6223   5200.378
perc_college_jobs                                -8169.2759   4664.470
                                                   t value     Pr(>|t|)
(Intercept)                                    11.01620338 3.296774e-21
major_categoryArts                             -1.15982207 2.479047e-01
major_categoryBiology & Life Science           -0.06175313 9.508390e-01
major_categoryBusiness                          0.98465353 3.263288e-01
major_categoryCommunications & Journalism      -0.31194760 7.554996e-01
major_categoryComputers & Mathematics          -1.92375132 5.621843e-02
major_categoryEducation                        -1.22078662 2.240207e-01
major_categoryEngineering                      -0.81882249 4.141448e-01
major_categoryHealth                           -0.79449323 4.281231e-01
major_categoryHumanities & Liberal Arts        -2.06100962 4.097298e-02
major_categoryIndustrial Arts & Consumer Services -0.58638810 5.584677e-01
major_categoryInterdisciplinary                -1.53196755 1.275689e-01
major_categoryLaw & Public Policy              -1.12162214 2.637577e-01
major_categoryPhysical Sciences                -0.70927770 4.792177e-01
major_categoryPsychology & Social Work         -1.04421726 2.980107e-01
major_categorySocial Science                   -0.81217605 4.179362e-01
perc_college_jobs                              -1.75138343 8.185791e-02
```

```
options(scipen = 999)
list(infl, inf.measures)
options(scipen = 0)
```

```
[[1]]
hii_LEVERAGE
   0.2209302

[[2]]
     yi      yi_        ei_         hii         ri_            Di_
1 33000 42216.23  -9216.231 0.012015861 -0.81692462 0.000427184315
2 58000 45641.31  12358.687 0.008332640  1.09343497 0.000528748242
3 40000 43851.68  -3851.682 0.006055751 -0.34038706 0.000037153395
4 65000 45094.40  19905.600 0.009234717  1.76194968 0.001522951989
5 45000 43943.40   1056.604 0.005921398  0.09336961 0.000002733135
6 31000 41124.02 -10124.018 0.026722668 -0.90414531 0.001181315954
```

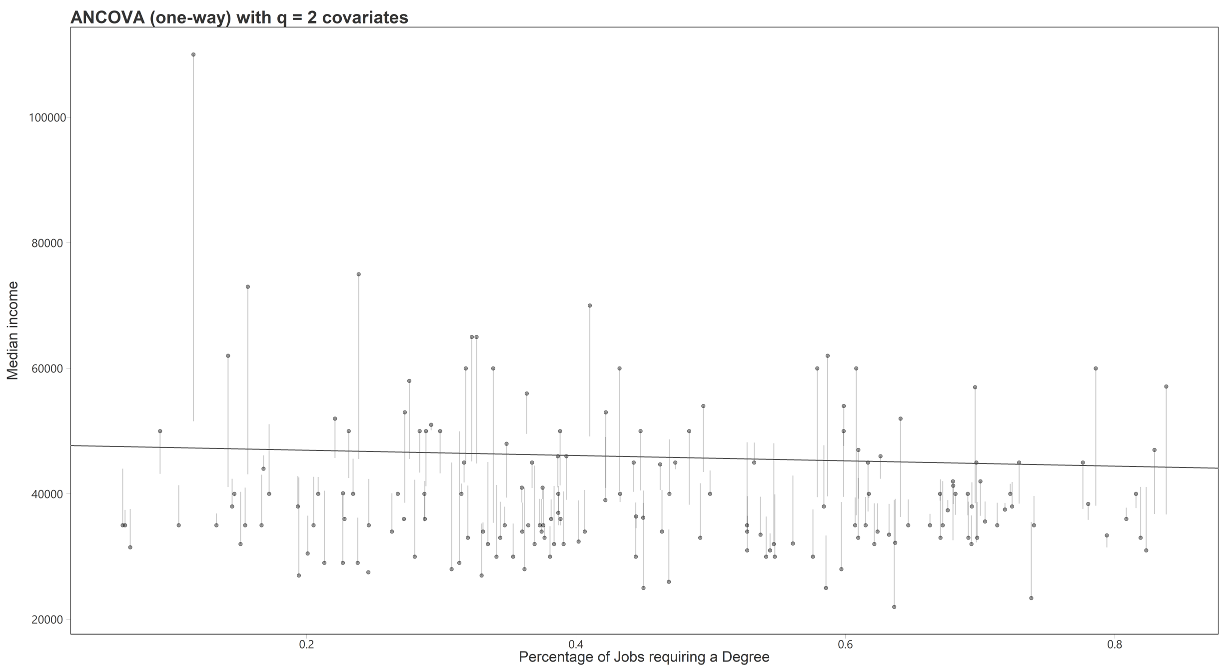

Figure 4:

DISCUSSION

CASE I revealed a weak linear relation between the regressors and the response variable. However, CASE I resutls only state that no *linear* relation exists, but does not state that no *other* relation exists. Such a relation would require more advanced mathematical modeling techiniques to discover what this true relation would be. In addition, the variance inflation factors for **perc college jobs** and **perc non college jobs** were moderatly high. This indicated that at least one of the two features were redundant and should be removed.

After making adjustments to the model, the same procedure in CASE I was conducted a second time in CASE II. However, model validation was skipped, given it served its purpose in CASE I. These results from the hypothesis test reached a similar conclusion: there is no sign of linearity in the model.