STAT 425 Case Study 1

Supanut Wanchai, Kai-Shiang Fan, Jui-Yu Lin

1 Introduction and Data Description

In this case study, we are given a data set to analyze the relationship between the number of practicing physicians by county and the selected county demographic information (CDI) from the years 1990 and 1992. There are 440 observations from the most populous counties in the United States, and each observation has an identification number with a county name and state abbreviation and provides information on 14 variables for a single county. We aim to propose a regression model for predicting the number of the practicing physicians by county from the variables given in the data set. The variables of the data set is shown as Table 1.

2 Data Preprocess

From the data description, we see that the first two columns, ID and $COUNTY$, are unique and serve as the numbers and names assigned to and representing each observation, so we dropped these two columns. The third column, which is $STATE$, is a categorical column, but has a total of 48 unique values. Applying dummy variable to the column would create too much columns and would not be practical, so we decided to drop $STATE$ for our regression model. Besides $STATE$, $REGION$ is also a categorical independent variable, so we did dummy for the column. Since there were 4 unique values , after dropping the original column and one of the dummy columns (here we dropped RE_W), we end up getting 3 new dummy columns of $REGION$, denoted as RE_NE, RE_NC, RE_S.

3 Correlation Check

Our goal for this research is to propose a regression model to predict PHY, and collinearity is an issue we would not want it to happen within the independent variables. In order to avoid the issue, we observed the correlation table (showed in Figure 1) and found high correlation (>0.85) between TP, BED, SC, and $TOTAL\_INC$. Thus, we should only keep one of these variables to avoid collinearity. We decided to keep BED since it has the highest correlation among these four variables to the dependent variable, PHY.

4 Full Model and Model Assumptions Check

4.1 Full model

We run a full model denoted as below:
$PHY = \beta_{0,0} + \beta_{0,1}LA + \beta_{0,2}P18 + \beta_{0,3}P65 + \beta_{0,4}BED + \beta_{0,5}HS + \beta_{0,6}BD + \beta_{0,7}POV + \beta_{0,8}UNEM + \beta_{0,9}CAP\_INC + \beta_{0,10}RE\_NE + \beta_{0,11}RE\_NC + \beta_{0,12}RE\_S$

4.2 Unusual Observations for Full Model

We observed 7 bad high-leverage points out of 35 high-leverage points, 5 outliers, and 1 highly influential data point. The data point 1 is a bad high-leverage point, an outlier, and a highly influential point.

4.3 Model Assumption Checks

4.3.1 Checking Constant Variance Assumption

Breusch-Pagan test (BP test) was constructed and the p-value was less than 2.2×10^{-16} , which is way less than $\alpha = 0.05$. So we rejected the null hypothesis and concluded that constant variance assumption is not satisfied.

4.3.2 Checking Normality Assumption

Kolmogorov-Smirnov test (KS test) was constructed instead of Shapiro-Wilk test (SW test) because the number of observations in this data set is greater than 50. The p-value we get by implying KS test on the full model was 1.033×10^{-7} , which is way less than $\alpha = 0.05$, so we reject the null hypothesis and concluded that the normality assumption is not satisfied.

4.3.3 Conclusion on Model Assumption Checks

Since neither the constant variance assumption and the normality assumption was satisfied, hypothesis testings depending on this assumption such as t-test and F-test could not be interpreted. Before moving on to model selection, we decided to transform the data as follows to meet the model assumptions.

5 Data Transformation, New Full Model and Assumption Check

5.1 Data Transformation

First of all, we looked at the histograms of all the variables and the response (shown in Figure 2). We observed that the distribution of the variables LA and BED, and the response PHY are severely skewed to the right. Therefore, we decided to apply log transformation on them. However, when we tested the model for constant variance and normality assumptions after the transformation, we found out that the model did not satisfy both assumptions. Thus we did another transformation based on those variables we transformed. For the predictor LA and BED, we squared the values which were transformed by log. For the response PHY, we took the reciprocal of the values which were transformed by log. In other words, the transformed columns are denoted as $LAnew = (logLA)^2$, $BEDnew = (logBED)^2$, and $PHYnew = \frac{1}{logPHY}$.

5.2 New Full Model

The new full model is denoted as $PHYnew = \beta_{1,0} + \beta_{1,1}LAnew + \beta_{1,2}P18 + \beta_{1,3}P65 + \beta_{1,4}BEDnew + \beta_{1,5}HS + \beta_{1,6}BD + \beta_{1,7}POV + \beta_{1,8}UNEM + \beta_{1,9}CAP\_INC + \beta_{1,10}RE\_NE + \beta_{1,11}RE\_NC + \beta_{1,12}RE\_S$.

5.3 Constant Variance and Normality Assumption Check for New Full Model

The p-values for BP test and KS test were 0.1853 and 0.1342 respectively, meaning that both the constant variance assumption and the normality assumption are satisfied.

6 Model Selection

6.1 First Reduced Model: Reducing $LAnew$, $P18$, $P65$, POV, $UNEM$

6.1.1 Summary

After transforming the data as mentioned, we looked into the p-values of each of the variables from the summary (showed in Figure 3). $LAnew$, $P18$, $P65$, POV, $UNEM$ are all insignificant variables in terms of the significant level $\alpha = 0.1$, with p-values (rounded to two decimals) of 0.42, 0.96, 0.56, 0.91 and 0.69 respectively, so we decided to drop these 5 variables first. The reduced model is denoted as the following: $PHYnew = \beta_{2,0} + \beta_{2,1}BEDnew + \beta_{2,2}HS + \beta_{2,3}BD + \beta_{2,4}CAP\_INC + \beta_{2,5}RE\_NE + \beta_{2,6}RE\_NC + \beta_{2,7}RE\_S$.

6.1.2 General Linear Test

We adopted General Linear Test to find out whether the first reduced model is better than the full model. By the *anova* table in R (shown in Figure 4), we got a p-value of 0.9413, indicating the reduced model is adequate.

6.1.3 Model Assumption Checks

We then check the model assumptions for the reduced model before doing t-test and F-test. The p-values for BP test and KS test were 0.2305 and 0.09849 respectively, meaning that both the constant variance assumption and the normality assumption are satisfied.

6.2 Second Reduced Model: Reducing HS

6.2.1 Summary

By observing the summary from the aforementioned reduced model (shown in Figure 5), we find that HS has a p-value of 0.05348, whereas all the other variables are significant in terms of significant level $\alpha = 0.05$. We decided to drop this variable and the reduced model is denoted as the following: $PHYnew = \beta_{3,0} + \beta_{3,1}BEDnew + \beta_{3,2}BD + \beta_{3,3}CAP\_INC + \beta_{3,4}RE\_NE + \beta_{3,5}RE\_NC + \beta_{3,6}RE\_S$.

6.2.2 General Linear Test

We adopted General Linear Test to find out whether the second reduced model is better than the first reduced model. By the *anova* table in R (shown in Figure 6), we get a p-value of 0.05348, indicating the second reduced model is adequate.

6.2.3 Assumption Checks

We then check the model assumptions for the second reduced model. The p-values for BP test and KS test were 0.1599 and 0.04783 respectively, meaning that the constant variance assumption is satisfied, but the normality assumption is not satisfied. Thus, we do not accept the second reduced model.

7 Empirical Transformation

7.1 Idea of Empirical Transformation

Since the p-value for the normality assumption check is really close to 0.05, we consider optimizing the transformation we did to PHY. In order to maintain to meet the constant variance assumption, we assume that $Var(Y) \propto [E(Y)]^x$ for $k \in \mathbb{R}, z = E(Y)$.
Then we have $Var(h(Y)) \approx (h'(z))^2 Var(Y)$
$\implies h'(z) = z^{\frac{-x}{2}}$
$\implies h(z) = \int z^{\frac{-x}{2}} dk = \begin{cases} l \times z^{\frac{2-x}{2}}, & \text{for some } l \in \mathbb{R}, x \neq 2 \\ m \times log(z), & x = 2 \end{cases}$
$\implies h(Y) = \begin{cases} Y^{\frac{2-x}{2}}, & x \neq 2 \\ log(Y), & x = 2 \end{cases}$
When we transform Y as above, the constant variance assumption will be satisfied. Therefore, we want to find the parameter of x to satisfy the normality assumption. We decided to test on the power of $PHYnew$ by running a loop of x from -10 to 10 with a 0.1 incremental, where the the newly transformed PHY is denoted as $PHYnew2 = (\frac{1}{logPHY})^{\frac{2-x}{2}}$. The result showed that when x has the value of -0.8, -0.7, -0.6, ..., 1.0, 1.1, 1.2, the model assumptions can be satisfied. We took the median of the values, which was $x = 0.2$, and got 0.9 as the power for our $PHYnew$, denoted as $PHYnew2 = (\frac{1}{logPHY})^{0.9}$.

7.2 Full Model After Empirical Transformation

The full model after empirical transformation is $PHYnew2 = \beta_{4,0} + \beta_{4,1}LAnew + \beta_{4,2}P18 + \beta_{4,3}P65 + \beta_{4,4}BEDnew + \beta_{4,5}HS + \beta_{4,6}BD + \beta_{4,7}POV + \beta_{4,8}UNEM + \beta_{4,9}CAP\_INC + \beta_{4,10}RE\_NE + \beta_{4,11}RE\_NC + \beta_{4,12}RE\_S$.

7.3 Model Assumption Checks

The p-values for BP test and KS test were 0.1669 and 0.1929 respectively, meaning that both the constant variance assumption and the normality assumption are satisfied.

8 Reduced Model after Empirical Transformation: Reducing $LAnew$, $P18$, $P65$, POV, $UNEM$, HS

8.1 Summary

After doing empirical transformation, we looked into the p-values of each of the variables from the summary again (shown in Figure 7). $LAnew$, $P18$, $P65$, POV, $UNEM$ again has high p-values (rounded to two decimals) of 0.42, 0.9, 0.62, 0.94 and 0.69 respectively, so we decided to drop these 5 variables. Different than using 1 as the power of the tranformation to PHY, the variable HS now is insignificant in terms of significant level $\alpha = 0.10$, so we decided to drop it along with the five variables of high p-values. The reduced model is denoted as the following: $PHYnew2 = \beta_{5,0} + \beta_{5,1}BEDnew + \beta_{5,2}BD + \beta_{5,3}CAP\_INC + \beta_{5,4}RE\_NE + \beta_{5,5}RE\_NC + \beta_{5,6}RE\_S$.

8.2 General Linear Test

We adopted General Linear Test to find out whether the empirically transformed reduced model is better than the empirically transformed full model. By the *anova* table in R (shown in Figure 8), we get a p-value of 0.5917, indicating the reduced model is adequate. By adopting t-test, we observed that all the p-values of the predictors in the model now were less than $\alpha = 0.05$. We conclude that no more variables can be dropped at this point since they are all significant.

8.3 Model Assumption Checks

The p-values for BP test and KS test were 0.1359 and 0.0864 respectively, meaning that both the constant variance assumption and the normality assumption can be satisfied, unlike that of before doing the empirical transformation. By satisfying both model assumptions, we can then conduct t-test and F-test on this model.

9 Unusual Observations

We observed 0 bad high-leverage points out of 26 high-leverage points, 2 outliers, and 0 highly influential point.

10 Conclusion

After doing the analysis, we found out that it was sufficient to run the regression model with 6 independent variables which are $BEDnew$, BD, CAP_INC, RE_NE, RE_NC, RE_S against the dependent variable $PHYnew2$, where $BEDnew = (logBED)^2$ and $PHYnew2 = (\frac{1}{logPHY})^{0.9}$ The final model can be denoted as the following: $(\frac{1}{logPHY})^{0.9} = 3.159 \times 10^{-1} - 1.87 \times 10^{-3}BEDnew - 1.104 \times 10^{-3} - 6.774 \times 10^{-7}CAP\_INC + 6.03 \times 10^{-3}RE\_NE + 1.219 \times 10^{-2}RE\_NC + 7.131 \times 10^{-3}RE\_S$ (summary of the final model is shown in Figure 9).

11 Appendix

Variable Number	Variable Name	Variable Abbrev.	Description
1	Identification Number	ID	1-440
2	County	COUNTY	County name
3	State	STATE	Two-letter state abbreviation
4	Land Area	LA	Land area (square miles)
5	Total Population	TP	Estimated 1990 population
6	Percent of population aged 18–24	P18	Percent of 1990 CDI population aged 18-24
7	Percent of population 65 or older	P65	Estimated 1990 population aged or older
8	Number of active physicians	PHY	Number of professionally active nonfederal physicians during 1990
9	Number of hospital beds	BED	Total number of beds, cribs and bassinets during 1990
10	Total serious crimes	SC	Total number of serious crimes in 1990 as reported by law enforcement agencies
11	Percent high school graduates	HS	Percent of adult population who completed 12 or more years of school
12	Percent bachelor's degrees	BD	Percent of adult population who with bachelor's degrees
13	Percent below poverty level	POV	Percent of 1990 CDI population with income below poverty level
14	Percent unemployment	UNEM	Percent of 1990 CDI labor force that is unemployed
15	Per capita income	CAP_ INC	Per capita income of 1990 CDI population (dollars)
16	Total personal income	TOTAL_ INC	Total personal income of 1990 CDI population (in millions of dollars)
17	Geographic region	REGION	Geographic region classification that is used in the US Bureau of the Census: 1=NE, 2=NC, 3=S, 4=W

Table 1: CDI Data Set Variables

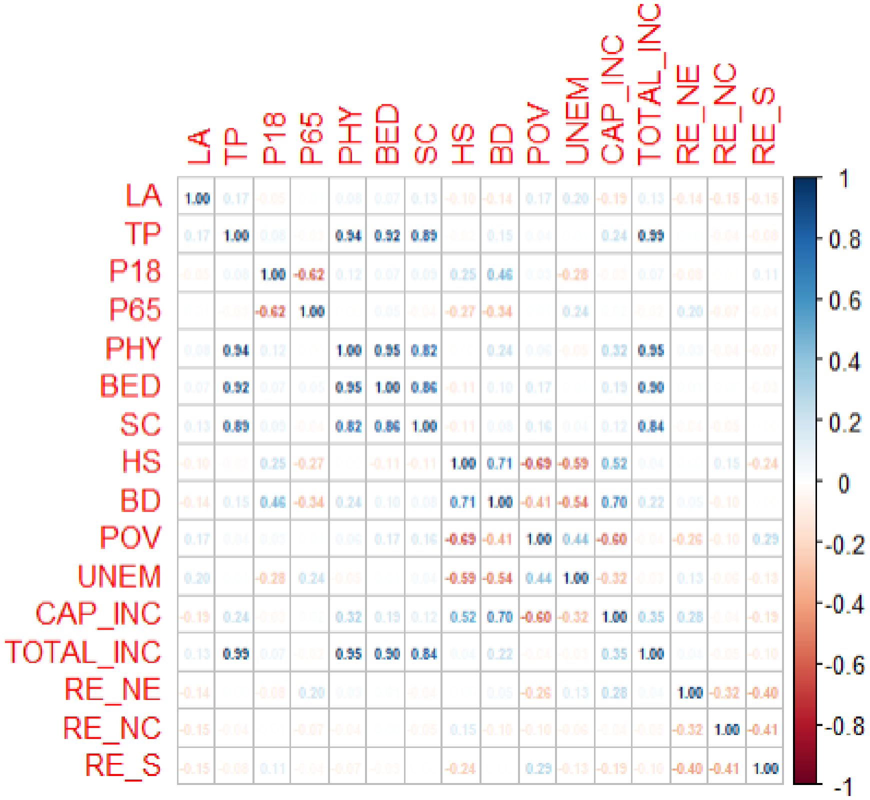


Figure 1: Correlation Table

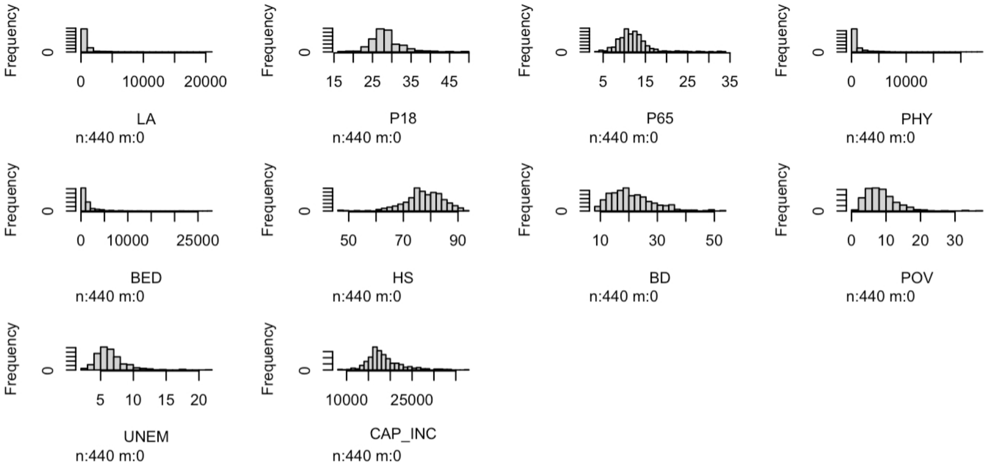
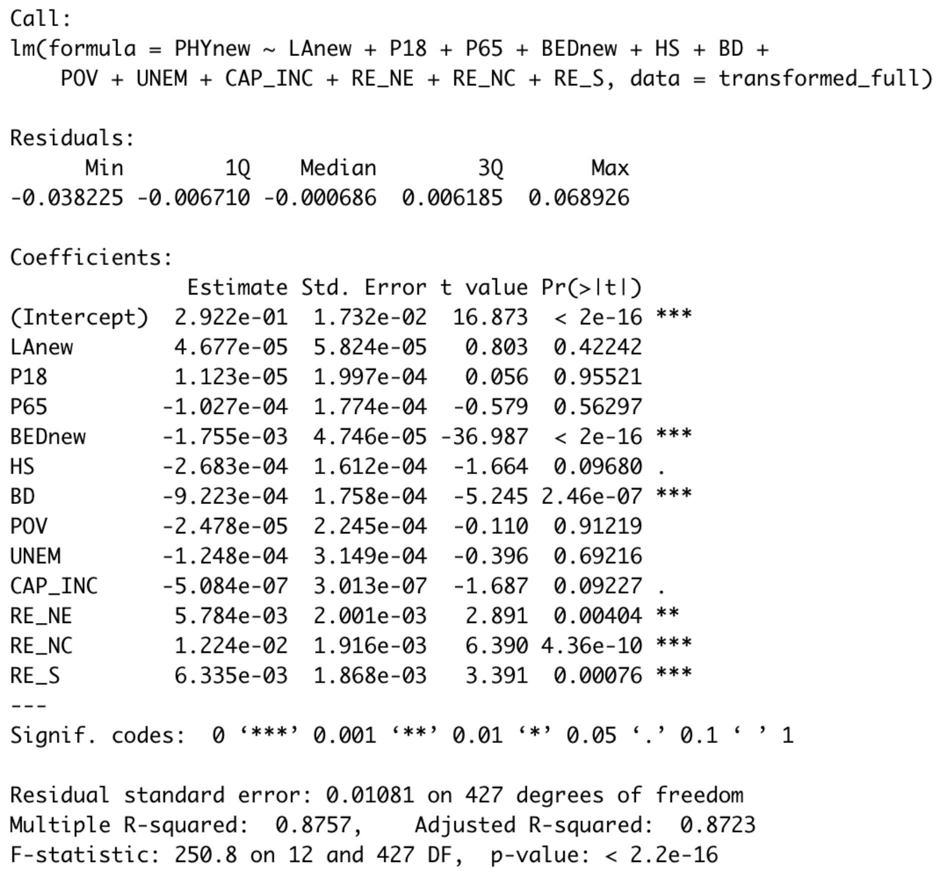


Figure 2: Histograms of all Variables



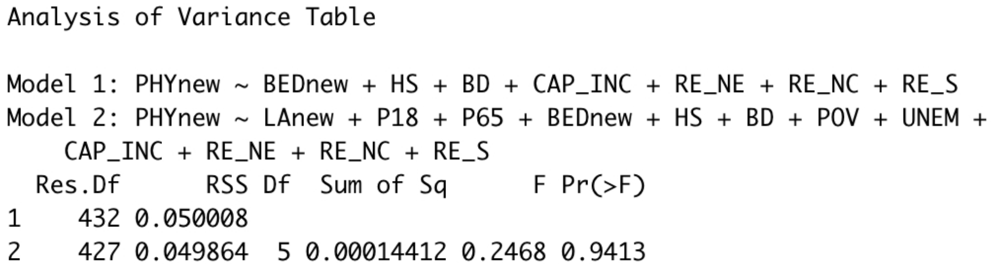
```
Call:
lm(formula = PHYnew ~ LAnew + P18 + P65 + BEDnew + HS + BD +
    POV + UNEM + CAP_INC + RE_NE + RE_NC + RE_S, data = transformed_full)

Residuals:
      Min        1Q    Median        3Q       Max
-0.038225 -0.006710 -0.000686  0.006185  0.068926

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.922e-01  1.732e-02  16.873  < 2e-16 ***
LAnew        4.677e-05  5.824e-05   0.803  0.42242
P18          1.123e-05  1.997e-04   0.056  0.95521
P65         -1.027e-04  1.774e-04  -0.579  0.56297
BEDnew      -1.755e-03  4.746e-05 -36.987  < 2e-16 ***
HS          -2.683e-04  1.612e-04  -1.664  0.09680 .
BD          -9.223e-04  1.758e-04  -5.245 2.46e-07 ***
POV         -2.478e-05  2.245e-04  -0.110  0.91219
UNEM        -1.248e-04  3.149e-04  -0.396  0.69216
CAP_INC     -5.084e-07  3.013e-07  -1.687  0.09227 .
RE_NE        5.784e-03  2.001e-03   2.891  0.00404 **
RE_NC        1.224e-02  1.916e-03   6.390 4.36e-10 ***
RE_S         6.335e-03  1.868e-03   3.391  0.00076 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.01081 on 427 degrees of freedom
Multiple R-squared:  0.8757,    Adjusted R-squared:  0.8723
F-statistic: 250.8 on 12 and 427 DF,  p-value: < 2.2e-16
```

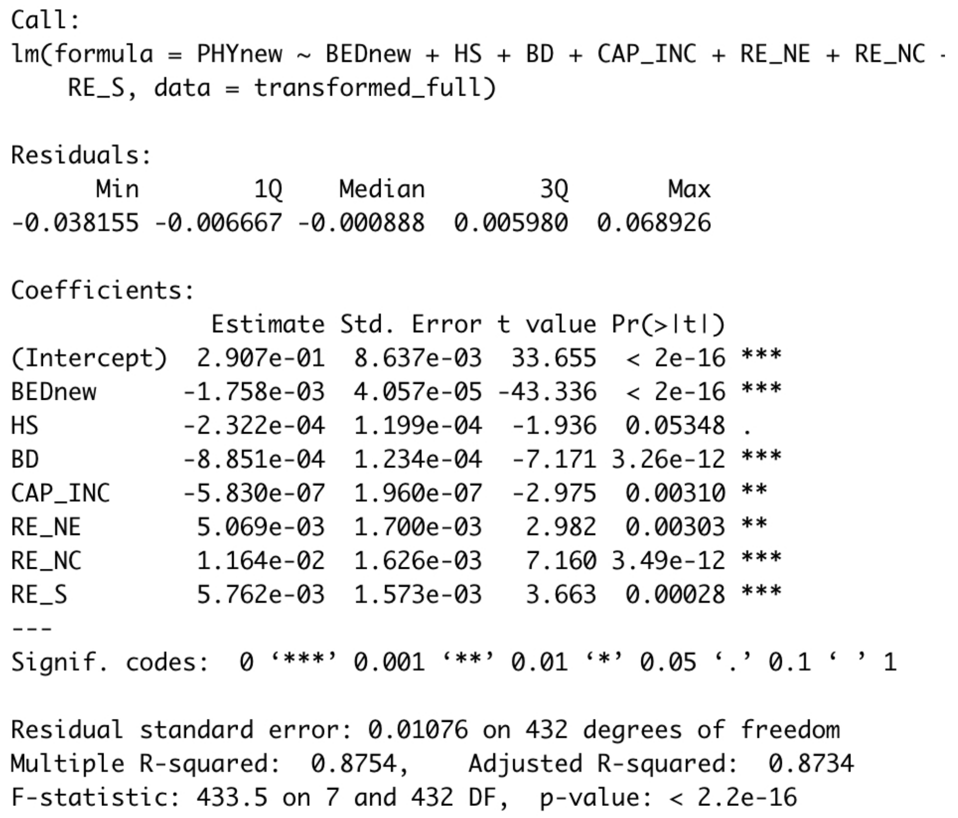
Figure 3: Summary of Transformed Full Model



```
Analysis of Variance Table

Model 1: PHYnew ~ BEDnew + HS + BD + CAP_INC + RE_NE + RE_NC + RE_S
Model 2: PHYnew ~ LAnew + P18 + P65 + BEDnew + HS + BD + POV + UNEM +
    CAP_INC + RE_NE + RE_NC + RE_S
  Res.Df      RSS Df  Sum of Sq      F Pr(>F)
1    432 0.050008
2    427 0.049864  5 0.00014412 0.2468 0.9413
```

Figure 4: Anova Table of First Reduced Model



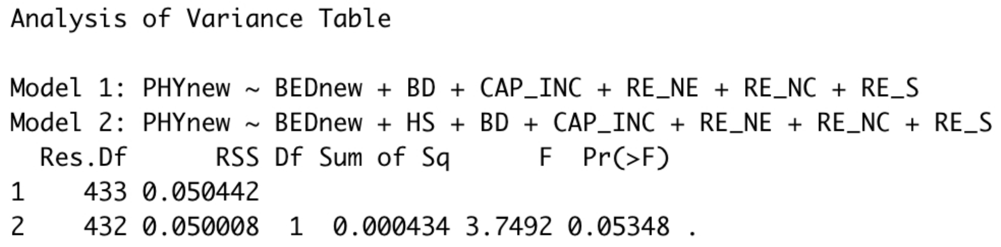
```
Call:
lm(formula = PHYnew ~ BEDnew + HS + BD + CAP_INC + RE_NE + RE_NC -
    RE_S, data = transformed_full)

Residuals:
      Min        1Q    Median        3Q       Max
-0.038155 -0.006667 -0.000888  0.005980  0.068926

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.907e-01  8.637e-03  33.655  < 2e-16 ***
BEDnew      -1.758e-03  4.057e-05 -43.336  < 2e-16 ***
HS          -2.322e-04  1.199e-04  -1.936  0.05348 .
BD          -8.851e-04  1.234e-04  -7.171 3.26e-12 ***
CAP_INC     -5.830e-07  1.960e-07  -2.975  0.00310 **
RE_NE        5.069e-03  1.700e-03   2.982  0.00303 **
RE_NC        1.164e-02  1.626e-03   7.160 3.49e-12 ***
RE_S         5.762e-03  1.573e-03   3.663  0.00028 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.01076 on 432 degrees of freedom
Multiple R-squared:  0.8754,    Adjusted R-squared:  0.8734
F-statistic: 433.5 on 7 and 432 DF,  p-value: < 2.2e-16
```

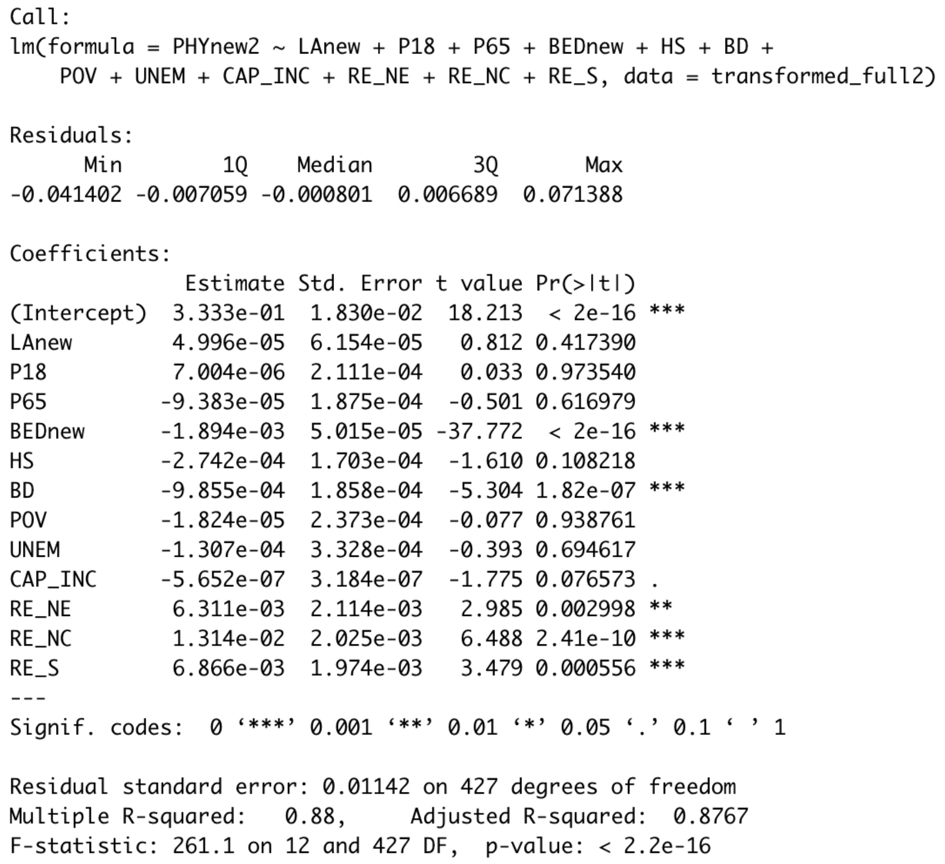
Figure 5: Summary of First Reduced Model



```
Analysis of Variance Table

Model 1: PHYnew ~ BEDnew + BD + CAP_INC + RE_NE + RE_NC + RE_S
Model 2: PHYnew ~ BEDnew + HS + BD + CAP_INC + RE_NE + RE_NC + RE_S
  Res.Df      RSS Df Sum of Sq      F  Pr(>F)
1    433 0.050442
2    432 0.050008  1  0.000434 3.7492 0.05348 .
```

Figure 6: Anova Table of Second Reduced Model



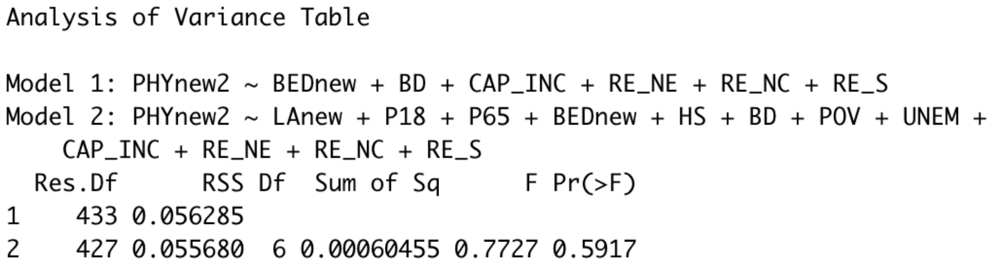
```
Call:
lm(formula = PHYnew2 ~ LAnew + P18 + P65 + BEDnew + HS + BD +
    POV + UNEM + CAP_INC + RE_NE + RE_NC + RE_S, data = transformed_full2)

Residuals:
      Min        1Q    Median        3Q       Max
-0.041402 -0.007059 -0.000801  0.006689  0.071388

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  3.333e-01  1.830e-02  18.213  < 2e-16 ***
LAnew        4.996e-05  6.154e-05   0.812 0.417390
P18          7.004e-06  2.111e-04   0.033 0.973540
P65         -9.383e-05  1.875e-04  -0.501 0.616979
BEDnew      -1.894e-03  5.015e-05 -37.772  < 2e-16 ***
HS          -2.742e-04  1.703e-04  -1.610 0.108218
BD          -9.855e-04  1.858e-04  -5.304 1.82e-07 ***
POV         -1.824e-05  2.373e-04  -0.077 0.938761
UNEM        -1.307e-04  3.328e-04  -0.393 0.694617
CAP_INC     -5.652e-07  3.184e-07  -1.775 0.076573 .
RE_NE        6.311e-03  2.114e-03   2.985 0.002998 **
RE_NC        1.314e-02  2.025e-03   6.488 2.41e-10 ***
RE_S         6.866e-03  1.974e-03   3.479 0.000556 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.01142 on 427 degrees of freedom
Multiple R-squared:   0.88,     Adjusted R-squared:  0.8767
F-statistic: 261.1 on 12 and 427 DF,  p-value: < 2.2e-16
```

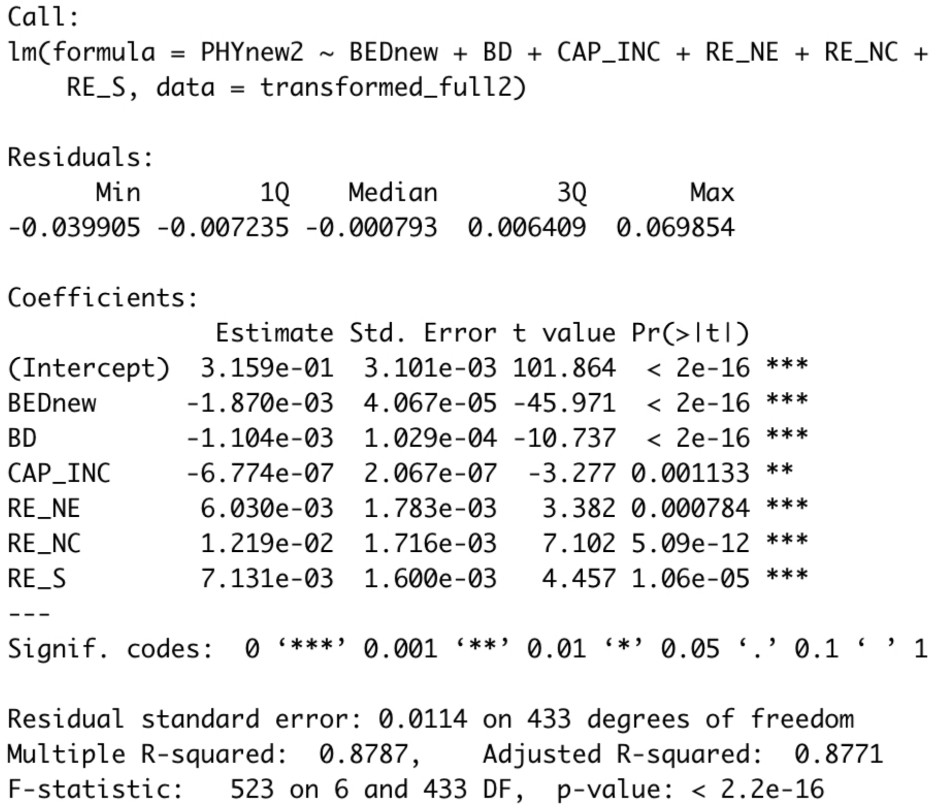
Figure 7: Summary of the Empirically Transformed Full Model



```
Analysis of Variance Table

Model 1: PHYnew2 ~ BEDnew + BD + CAP_INC + RE_NE + RE_NC + RE_S
Model 2: PHYnew2 ~ LAnew + P18 + P65 + BEDnew + HS + BD + POV + UNEM +
    CAP_INC + RE_NE + RE_NC + RE_S
  Res.Df      RSS Df  Sum of Sq      F Pr(>F)
1    433 0.056285
2    427 0.055680  6 0.00060455 0.7727 0.5917
```

Figure 8: Anova Table of the Reduced Model after Empirical Transformation



```
Call:
lm(formula = PHYnew2 ~ BEDnew + BD + CAP_INC + RE_NE + RE_NC +
    RE_S, data = transformed_full2)

Residuals:
      Min        1Q    Median        3Q       Max 
-0.039905 -0.007235 -0.000793  0.006409  0.069854 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept)  3.159e-01  3.101e-03 101.864  < 2e-16 ***
BEDnew      -1.870e-03  4.067e-05 -45.971  < 2e-16 ***
BD          -1.104e-03  1.029e-04 -10.737  < 2e-16 ***
CAP_INC     -6.774e-07  2.067e-07  -3.277 0.001133 ** 
RE_NE        6.030e-03  1.783e-03   3.382 0.000784 ***
RE_NC        1.219e-02  1.716e-03   7.102 5.09e-12 ***
RE_S         7.131e-03  1.600e-03   4.457 1.06e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.0114 on 433 degrees of freedom
Multiple R-squared:  0.8787,	Adjusted R-squared:  0.8771 
F-statistic:   523 on 6 and 433 DF,  p-value: < 2.2e-16
```

Figure 9: Summary of the Final Model