5-14-2019

Theory of Linear Models for Estimating Regression Parameters with Applications to Two-Factor Studies with Unequal Sample Sizes

Zenan Sun
Bridgewater State University

Theory of Linear Models for Estimating Regression Parameters with Applications to Two-Factor Studies with Unequal Sample Sizes

Zenan Sun

Submitted in Partial Completion of the
Requirements for Departmental Honors in Mathematics

Bridgewater State University

May 14, 2019

Dr. Kevin Rion, Thesis Advisor
Dr. Irina Seceleanu, Committee Member
Dr. Wanchunzi Yu, Committee Member

Abstract:

In this thesis we explored some topics in regression analysis. In particular, we studied what linear regression is from a matrix theory perspective, and applied analysis of variance in a setting with two factors and unbalanced sample sizes. In addition, we applied Box-Cox variable transformation as a solution when the regression model violated the normality and equal variance (also called homoscedasticity) assumption. Our main goal is to use these theories to construct models and investigate questions related to lifetime earnings of people living in America by using real data. In doing so, we used the statistical software R to perform calculation involved in variable selection models, to identify and quantify relationships between variables as well as to test hypotheses.

Introduction to Linear Models from Matrix Theory Perspective

We will restrict attention to Simple Linear Regression to illustrate the main ideas. Simple linear regression models arise as an attempt to represent the relationship between two real valued variables x and y in the form

$$y = \beta_0 + \beta_1 x.$$

After observing n datapoints $(\mathrm{x_I}, \mathrm{y_I})$, the goal is to find the β_0, and β_1 for which $y_i = \beta_0 + \beta_1 x_i$ holds for all i. As a matrix equation this is

$$\mathbf{Y}\text{=}\mathbf{X}\boldsymbol{\beta}$$

where

$$\mathbf{Y} = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix}, \mathbf{X} = \begin{bmatrix} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix} \text{, and } \boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix}.$$

We call $\mathbf{X}$ the design matrix, and its column $(x_1, \dots, x_n)^{\mathrm{T}}$ is chosen in the design of a study. But $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta}$ usually has no solution, so we use the vector $\widehat{\mathbf{Y}}$ in the range of $\mathbf{X}$ that is closest to $\mathbf{Y}$ as an approximate solution. As we will illustrate below, this choice of $\widehat{\mathbf{Y}}$ as an approximate solution has an illuminating geometric interpretation for which the standard regression identity: $\mathrm{SSTotal} = \mathrm{SSRegression} + \mathrm{SSResiduals}$, is seen as a Pythagorean Theorem in n spaces.

We will use ||·|| to denote the norm $x \mapsto \sqrt{\sum x_i^2}$, and denote the span of a set of vectors $\mathbf{B}$ as $\langle \mathbf{B} \rangle$, and use ⦹ to denote a direct sum of orthogonal spaces of vectors.

Procedure for acquiring $\widehat{\mathbf{Y}}$ as an approximate solution to $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta}$.

Step 1: Let $\mathbb{1}$ be the vector $(1,\dots,1)^{\mathrm{T}} \in \mathbb{R}^n$, and $\mathbf{X} = (x_1,\dots,x_n)^{\mathrm{T}}$. Orthogonalize $\{\mathbb{1},\mathbf{x}\}$ to obtain a basis $\{\mathbb{1},\mathbf{x}-\bar{\mathbf{x}}\cdot\mathbb{1}\ \}$ of orthogonal vectors spanning the same 2-dimensional subset of $\mathbb{R}^{\mathrm{n}}$as $\{\mathbb{1},\mathbf{x}\}$ does. Since the range of $\mathbf{X}$ is the span of its column, this gives us

$$Range(\mathbf{X}) = \langle\{\mathbb{1},\mathbf{x}-\bar{\mathbf{x}}\cdot\mathbb{1}\}\rangle = \langle\mathbb{1}\rangle \circledperp \langle\mathbf{x}-\bar{\mathrm{x}}\cdot\mathbb{1}\rangle.$$

Step 2: Project $\mathbf{y}$ orthogonally onto $\langle\mathbb{1}\rangle \circledperp \langle\mathbf{x}-\bar{\mathbf{x}}\cdot\mathbb{1}\rangle$ to obtain $\widehat{\mathbf{Y}}$. It is achieved by the Hat Matrix $\mathbf{H} = \mathbf{X}(\mathbf{X}'\cdot\mathbf{X})^{-1}\mathbf{X}'$, and it is named because it puts the hat on $\mathbf{Y}$ (see page 29, 70 of [3]):

$$\widehat{\mathbf{Y}} = \mathbf{HY}$$

Note that as long s the matrix $\mathbf{X}$ has rank 2, the matrix $\mathbf{X}'\mathbf{X}$ is invertible. We will always satisfy this requirement since we choose the x vector and it would only lie in the span of $\mathbb{1}$ all our observations are made of the same level of x. In this case, that would be a vertical line.

In figure 1 that follows, $\mathbf{Y}$ is vectors $\overrightarrow{\mathbf{ED}}$ and $\widehat{\mathbf{Y}}$ is vector $\overrightarrow{\mathbf{EH}}$. This orthogonal projection makes $\widehat{\mathbf{Y}}$ be the closest vector in the span of $\mathbb{1}$ and $\mathbf{x}-\bar{\mathbf{x}}\cdot\mathbb{1}$ to $\mathbf{Y}$. Equivalently $\widehat{\mathbf{Y}}$ minimizes the length of the residual vector $\mathbf{Y}-\widehat{\mathbf{Y}}$ (vector $\overrightarrow{\mathbf{HD}}$ in figure 1).

Step 3: We obtain a representation of $\widehat{\mathbf{Y}}$ as a linear combination of $\mathbb{1}$ and $\mathbf{x}-\bar{\mathrm{x}}\cdot\mathbb{1}$ as follows (Theorem 1.6.7, Page 34, [3])

$$\begin{aligned}\widehat{\mathbf{Y}} &= \mathrm{Proj}^{\perp}(\mathbf{Y}|\langle\mathbb{1}\rangle \circledperp \langle\mathbf{x}-\bar{\mathrm{x}}\cdot\mathbb{1}\rangle) \\ &= \mathrm{Proj}^{\perp}(\mathbf{Y}|\langle\mathbb{1}\rangle) \circledperp \mathrm{Proj}^{\perp}(\mathbf{Y}|\langle\mathbf{x}-\bar{\mathrm{x}}\cdot\mathbb{1}\rangle) \\ &= \bar{y}\cdot\mathbb{1} \circledperp \widehat{\beta_1}\cdot(\mathbf{x}-\bar{\mathrm{x}}\cdot\mathbb{1})\end{aligned}$$

And we can conclude that $\widehat{\mathbf{Y}} = \bar{y}\cdot\mathbb{1} + \widehat{\beta_1}\cdot(\mathbf{x}-\bar{\mathrm{x}}\cdot\mathbb{1})$ is the closest element of $\langle\{\mathbb{1},\mathbf{x}\}\rangle$ to $\mathbf{Y}.$ In figure 1, this is the claim $\overrightarrow{\mathbf{EH}} = \overrightarrow{\mathbf{EF}} + \overrightarrow{\mathbf{EG}}$

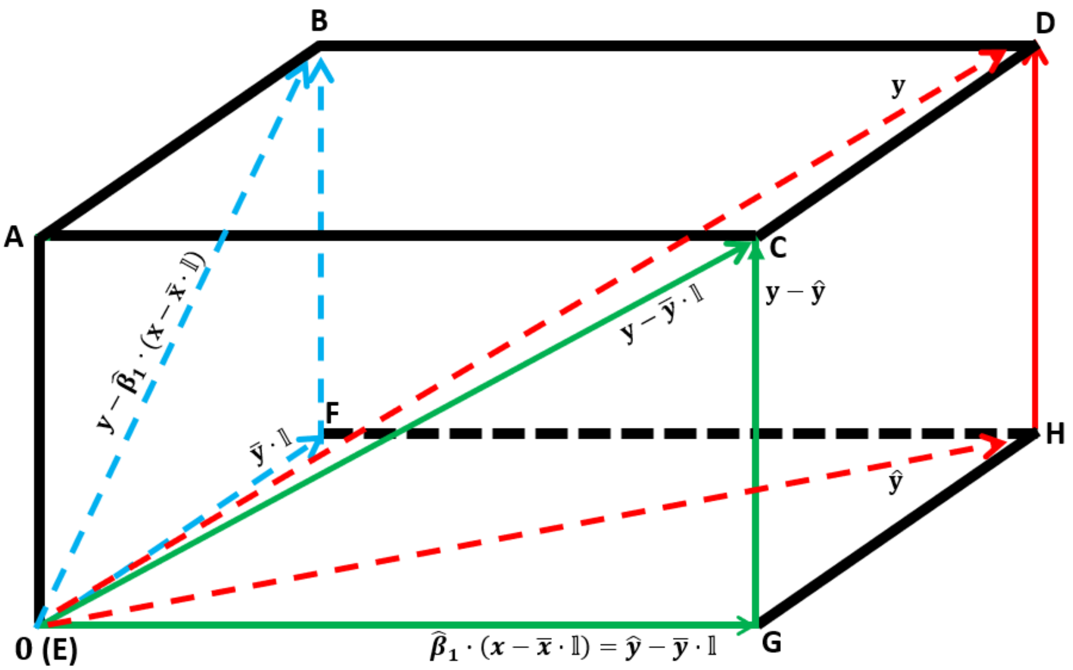

B
D
$\mathbf{y}$
A
C
$\mathbf{y} - \hat{\mathbf{y}}$
$\mathbf{y} - \bar{\mathbf{y}} \cdot \mathbb{1}$
$\mathbf{y} - \hat{\boldsymbol{\beta}}_1 \cdot (\mathbf{x} - \bar{\mathbf{x}} \cdot \mathbb{1})$
F
$\bar{\mathbf{y}} \cdot \mathbb{1}$
H
$\hat{\mathbf{y}}$
0 (E)
$\hat{\boldsymbol{\beta}}_1 \cdot (\boldsymbol{x} - \bar{\boldsymbol{x}} \cdot \mathbb{1}) = \hat{\boldsymbol{y}} - \bar{\boldsymbol{y}} \cdot \mathbb{1}$
G

Figure 1: Linear regression model as solution to projection problem

Note that:

- The regression identity $SSTotal = SSRegression + SSResiduals$ can be interpreted as a Pythagorean theorem in n-space for the triangle with vertices $\mathrm{E}, \mathrm{G}, \mathrm{C}$:

$$\left\|\overrightarrow{\mathbf{EC}}\right\|^2 = \left\|\overrightarrow{\mathbf{EG}}\right\|^2 + \left\|\overrightarrow{\mathbf{GC}}\right\|^2 \text{equivalently,}$$

$$\|\mathbf{Y} - \bar{\mathbf{y}} \cdot \mathbb{1}\|^2 = \left\|\hat{\mathbf{Y}} - \bar{\mathbf{y}} \cdot \mathbb{1}\right\|^2 + \left\|\mathbf{Y} - \hat{\mathbf{Y}}\right\|^2$$

Methodology

1. Two-Factor Studies with Unequal Sample Sizes

In the case of two-factor with equal sample sizes, a Pythagorean style theorem still holds but on more orthogonal terms. Then we can still identify statistics for using in estimation and testing assertions about the mean response to each combination of factor levels. However, an unbalanced sample sizes setting makes the study more complicated. In particular orthogonality is lost, and so we fail to use the conceptually clear and clean approach to find estimators and their distributions.

To get around this, we can use a generalized linear approach: Two-Factor Analysis with Unequal Sample Sizes (Chapter 23, [2]). This ANOVA model still follows the rules that the observations are normally distributed, and the variance of each group is the same.

Suppose we have two factors A and B with mean effects α and β. The factor-fixed effects model for two-factor ANOVA with interaction is:

$$Y_{ijk} = \mu_{..} + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \varepsilon_{ijk}$$

for $a, b \in \mathbb{N}$, $i = 1, \dots, a$ and $j = 1, \dots, b$; $n_{i,j} \in \mathbb{N}$, $k = 1, \dots, n_{i,j}$. Here:

- $\mu_{..}$ is a mean of the whole sample (the grand mean)
- α_i and β_j are constants subject to the restriction $\sum \alpha_i = 0$ and $\sum \beta_j = 0$
- $(\alpha\beta)_{ij}$ are interaction constants subject to the restrictions:

$$\textstyle\sum_i (\alpha\beta)_{ij} = 0 \qquad \text{for each } j = 1, \dots, b$$

$$\textstyle\sum_j (\alpha\beta)_{ij} = 0 \qquad \text{for each } i = 1, \dots, a$$

- And ε_{ijk} are independent $N(0, \sigma^2)$

Therefore, for each (i, j, k), the expected responses to the k^{th} observation of the treatment for which A is set to level i and B is set to level j is:

$$\mathbb{E}[Y_{ijk}] = \mu_{..} + \alpha_i + \beta_j + (\alpha\beta)_{ij}$$

By the constraints listed above, the expected overall response when A is set to level i is α_i, and the expected overall response to setting B with level j is β_j, and the $(\alpha\beta)_{ij}$ term is the mean interaction influence that the combination of setting A to level i and setting B to level j has on the response Y.

Also, since $\sum \alpha_i = 0$ and $\sum \beta_i = 0$, the first sum has $(a - 1)$ and the second sum has $(b - 1)$ degrees of freedom. Consequently, the last term α_a and β_b can be written as:

$$\alpha_a = -\alpha_1 - \alpha_2 - \cdots - \alpha_{a-1}$$

$$\beta_b = -\beta_1 - \beta_2 - \cdots - \beta_{b-1}$$

Similarly, we can write these equations for the interaction parameters:

$$(\alpha\beta)_{ib} = -(\alpha\beta)_{i1} - (\alpha\beta)_{i2} - \cdots - (\alpha\beta)_{i,b-1}$$

$$(\alpha\beta)_{aj} = -(\alpha\beta)_{1j} - (\alpha\beta)_{2j} - \cdots - (\alpha\beta)_{a-1,j}$$

Therefore, there are only $(a - 1)(b - 1)$ free interaction effects, and $(a - 1) + (b - 1)$ factor effects need to be estimated.

Example: Suppose we collect data on the responses to two treatment factors A and B, where A has 2 levels, and B has 3. So we need to estimate one of α_1 and α_2, two of β_1, β_2, β_3, and two of

$(\alpha\beta)_{11}, (\alpha\beta)_{12}, (\alpha\beta)_{13}, (\alpha\beta)_{21}, (\alpha\beta)_{22}$, and $(\alpha\beta)_{23}$. We will show how to find estimators of $\alpha_1, \beta_1, \beta_2, (\alpha\beta)_{11}$, and $(\alpha\beta)_{12}$.

Once we do this, we can obtain the others as follow:

$$\alpha_2 = -\alpha_1$$

$$\beta_3 = -\beta_1 - \beta_2$$

$$(\alpha\beta)_{13} = -(\alpha\beta)_{11} - (\alpha\beta)_{12}$$

$$(\alpha\beta)_{21} = -(\alpha\beta)_{11}$$

$$(\alpha\beta)_{22} = -(\alpha\beta)_{12}$$

$$(\alpha\beta)_{23} = -(\alpha\beta)_{13} = (\alpha\beta)_{11} + (\alpha\beta)_{12}$$

We begin by expressing the dependence of the k^{th} response Y has to the treatment for which A is set to level i and B is set to level j in terms of indicator functions X_1, X_2, and X_3 that are defined after the full regression model:

$$Y_{ijk} = \mu_{\cdot\cdot} + \underbrace{\alpha_1 X_{ijk1}}_{\text{Factor } A \text{ main effect}} + \underbrace{\beta_1 X_{ijk2} + \beta_2 X_{ijk3}}_{\text{Factor } B \text{ main effect}} + \underbrace{(\alpha\beta)_{11} X_{ijk1} X_{ijk2} + (\alpha\beta)_{12} X_{ijk1} X_{ijk3}}_{\text{Fractor } AB \text{ interation effect}} + \varepsilon_{ijk}$$

Here $X_{ijk1}, X_{ijk2}, X_{ijk3}$ are indicator functions that depend on the treatment.

Specifically,

$$X_{ijk1} = \begin{cases} 1, & \text{if A is set to level 1 (i=1)} \\ -1, & \text{if A is set to level 2 (i=2)} \end{cases}$$

$$X_{ijk2} = \begin{cases} 1, & \text{if } B \text{ is set to level 1 (j=1)} \\ -1, & \text{if } B \text{ is set to level 3 (j=3)} \\ & 0, \text{otherwise} \end{cases}$$

$$X_{ijk3} = \begin{cases} 1, & \text{if } B \text{ is set to level 2 (j=2)} \\ -1, & \text{if } B \text{ is set to level 3 (j=3)} \\ & 0, \text{otherwise} \end{cases}$$

Next, by applying the expectation operator to this equation we obtain the following equations for the expected values $\mu_{ij} = \mathbb{E}[Y|A = i, B = j]$.

$$\mu_{11} = \mu_{..} + \alpha_1 \cdot 1 + \beta_1 \cdot 1 + \beta_2 \cdot 0 + (\alpha\beta)_{11} \cdot 1 \cdot 1 + (\alpha\beta)_{12} \cdot 1 \cdot 0 = \mu_{..} + \alpha_1 + \beta_1 + (\alpha\beta)_{11}$$

$$\mu_{12} = \mu_{..} + \alpha_1 \cdot 1 + \beta_1 \cdot 0 + \beta_2 \cdot 1 + (\alpha\beta)_{11} \cdot 1 \cdot 0 + (\alpha\beta)_{12} \cdot 1 \cdot 1 = \mu_{..} + \alpha_1 + \beta_2 + (\alpha\beta)_{12}$$

$$\mu_{13} = \mu_{..} + \alpha_1 \cdot 1 + \beta_1 \cdot (-1) + \beta_2 \cdot (-1) + (\alpha\beta)_{11} \cdot 1 \cdot (-1) + (\alpha\beta)_{12} \cdot 1 \cdot (-1)$$

$$= \mu_{..} + \alpha_1 - \beta_1 - \beta_2 - (\alpha\beta)_{11} - (\alpha\beta)_{12}$$

$$\mu_{21} = \mu_{..} + \alpha_1 \cdot (-1) + \beta_1 \cdot 1 + \beta_2 \cdot 0 + (\alpha\beta)_{11} \cdot (-1) \cdot 1 + (\alpha\beta)_{12} \cdot (-1) \cdot 0$$

$$= \mu_{..} - \alpha_1 + \beta_1 - (\alpha\beta)_{11}$$

$$\mu_{22} = \mu_{..} + \alpha_1 \cdot (-1) + \beta_1 \cdot 0 + \beta_2 \cdot 1 + (\alpha\beta)_{11} \cdot (-1) \cdot 0 + (\alpha\beta)_{12} \cdot (-1) \cdot 1$$

$$= \mu_{..} - \alpha_1 + \beta_2 - (\alpha\beta)_{12}$$

$$\mu_{23} = \mu_{..} + \alpha_1 \cdot (-1) + \beta_1 \cdot (-1) + \beta_2 \cdot (-1) + (\alpha\beta)_{11} \cdot (-1) \cdot (-1) + (\alpha\beta)_{12} \cdot (-1) \cdot (-1)$$

$$= \mu_{..} - \alpha_1 - \beta_1 - \beta_2 + (\alpha\beta)_{11} + (\alpha\beta)_{12}$$

Assembling these information lead to the augmented matrix below, which can be row reduced to yield estimators of $\alpha_1, \beta_1, \beta_2, (\alpha\beta)_{11}$, and $(\alpha\beta)_{12}$

$$\left(\begin{array}{cccccc|c}
\mu_{..} & \alpha_1 & \beta_1 & 0 & (\alpha\beta)_{11} & 0 & \mu_{11} \\
\mu_{..} & \alpha_1 & 0 & \beta_2 & 0 & (\alpha\beta)_{12} & \mu_{12} \\
\mu_{..} & \alpha_1 & -\beta_1 & -\beta_2 & -(\alpha\beta)_{11} & -(\alpha\beta)_{12} & \mu_{13} \\
\mu_{..} & -\alpha_1 & \beta_1 & 0 & -(\alpha\beta)_{11} & 0 & \mu_{21} \\
\mu_{..} & -\alpha_1 & 0 & \beta_2 & 0 & -(\alpha\beta)_{12} & \mu_{22} \\
\mu_{..} & -\alpha_1 & -\beta_1 & -\beta_2 & (\alpha\beta)_{11} & (\alpha\beta)_{12} & \mu_{23}
\end{array}\right)$$

After solving this matrix, we get these parameters:

$$\begin{pmatrix} \alpha_1 \\ \beta_1 \\ \beta_2 \\ (\alpha\beta)_{11} \\ (\alpha\beta)_{12} \end{pmatrix} = \begin{pmatrix} \mu_{1.} - \mu_{..} \\ \mu_{.1} - \mu_{..} \\ \mu_{.2} - \mu_{..} \\ \mu_{11} - \mu_{1.} - \mu_{.1} + \mu_{..} \\ \mu_{12} - \mu_{1.} - \mu_{.1} + \mu_{..} \end{pmatrix}$$

Each mean occurring in the vector on the right-hand side of the previous equation is then estimated with the corresponding sample means.

Specifically,

- $\hat{\mu}_{ij} = \bar{y}_{ij.} := \frac{1}{n_{ij}} \sum_{k=1}^{n_{ij}} y_{ijk}$
- $\hat{\mu}_{i.} = \frac{1}{b} \sum_{j=1}^{b} \bar{y}_{ij.}$
- $\hat{\mu}_{.j} = \frac{1}{a} \sum_{i=1}^{a} \bar{y}_{ij.}$, and
- $\hat{\mu}_{..} = \frac{1}{a} \sum_{i=1}^{a} \hat{\mu}_{i.} = \frac{1}{ab} \sum_{i=1}^{a} \sum_{j=1}^{b} \bar{y}_{ij.}$

From this, we obtain

$$\begin{pmatrix} \hat{\alpha}_1 \\ \hat{\beta}_1 \\ \hat{\beta}_2 \\ \widehat{(\alpha\beta)}_{11} \\ \widehat{(\alpha\beta)}_{12} \end{pmatrix} = \begin{pmatrix} \hat{\mu}_{1.} - \hat{\mu}_{..} \\ \hat{\mu}_{.1} - \hat{\mu}_{..} \\ \hat{\mu}_{.2} - \hat{\mu}_{..} \\ \hat{\mu}_{11} - \hat{\mu}_{1.} - \hat{\mu}_{.1} + \hat{\mu}_{..} \\ \hat{\mu}_{12} - \hat{\mu}_{1.} - \hat{\mu}_{.1} + \hat{\mu}_{..} \end{pmatrix}$$

Note that, in the case of a balanced dataset, theses formulas reduce to

$$\begin{pmatrix} \hat{\alpha}_1 \\ \hat{\beta}_1 \\ \hat{\beta}_2 \\ \widehat{(\alpha\beta)}_{11} \\ \widehat{(\alpha\beta)}_{12} \end{pmatrix} = \begin{pmatrix} \bar{y}_{1..} - \bar{y}_{...} \\ \bar{y}_{.1.} - \bar{y}_{...} \\ \bar{y}_{.2.} - \bar{y}_{...} \\ \bar{y}_{11.} - \bar{y}_{1..} - \bar{y}_{.1.} + \bar{y}_{...} \\ \bar{y}_{12.} - \bar{y}_{1..} - \bar{y}_{.1.} + \bar{y}_{...} \end{pmatrix}$$

2. Box-Cox Transformation

When the explanatory variables are quantitative and we wish to regress the response Y on the explanatory variables X_1 and X_2 it is fairly common that the regression assumptions are not satisfied in that either the response variable Y is not normally distributed or it is not a linear function of the levels of X_1 and X_2. In such settings, a transformation may help. The Box-Cox method (Chapter 3.9, [2]) of choosing a transformation is well-studied and we introduce it here. Such transformations can be used to reduce skewness in the distributions of the errors ε_{ij}, stabilize the unequal error variance[1], and reduce the nonlinearity of the association between Y and X_1 and X_2.

By this method we will obtain a power-transformed variable $Y' = Y^{\lambda}$ where the power λ is chosen with the intention that there be parameters β_0, and β_1 so that for each factor X_i holds.

$$Y_i' = (Y_i)^{\lambda} = \beta_0 + \beta_1 X_i + \varepsilon_i$$

For example:

$$\lambda = -2 \implies Y' = \frac{1}{Y^2}$$

$$\lambda = 0 \quad \implies Y' = \ln Y \qquad \text{(by definition)}$$

$$\lambda = 0.5 \implies Y' = \sqrt{Y}.$$

[1] The function $(x_1, x_2) \longmapsto Var(Y|X_1 = x_1, X_2 = x_2)$ is called the scedastic function. So the assumption that all the error variables have the same variance is commonly called an assumption of **homoscedasticity**, and when that assumption fails the family of error variables is said to be **heteroscedastic**.

Since we also want this transformation to correct for the unequal variances, we further suppose the error variances are equal, say to σ^2 and include an estimation of that variance as part of the goal of determining the parameters. The method of Box-Cox uses maximum likelihood estimators for each of the parameters λ, β_0, β_1, and σ^2; however, an exact solution for $\hat{\lambda}$ is not typically desired since for example, a value of $\lambda = 0.4372$ will not typically perform much differently than a nicer value like $\lambda = 0.5$.

In practice there is a simple procedure to find an estimate $\hat{\lambda}$ of λ prior to attempting to find the others. First, standardized values W_i of the Y_i^λ variables are introduced for which the magnitude of the sum of squared errors (SSE) of these standardized variables is minimized at λ. Then a sequence of values of λ is selected, the SSE of these standardized variables is calculated for each of those values, and graph is produced displaying the SSE as a function of the putative value of. Lastly, we can just look at the graph and suggest a “nice” value for $\hat{\lambda}$ that is close to the number that would produce a global minimum.

For example if we think λ is between -2 and 3 we could use $\lambda \in \{-2.00, -1.75, -1.50, \dots, 2.75, 3.00\}$. If the relationship between the SSE and λ is as in the figure 2, then we could choose $\hat{\lambda}$ to be 1.5.

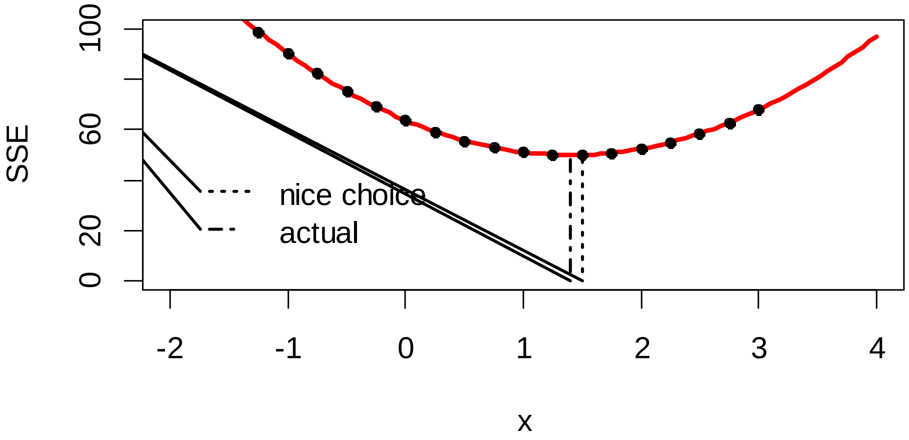


Figure 2: Maximum likelihood of λ

Specifically, the standardized values are defined as so. For each i, the standardized value W_i is calculated from a proposed value of λ, the response Y_i to factor i, and the geometric mean of all the response variables:

$$W_i = \begin{cases} K_1\left(Y_i^{\lambda} - 1\right) & \lambda \neq 0 \\ K_2(\ln Y_i) & \lambda = 0 \end{cases}$$

where

$$K_2 = \left(\prod_{i=1}^{n} Y_i\right)^{\frac{1}{n}}$$

$$K_1 = \frac{1}{\lambda K_2^{\lambda - 1}}$$

The best estimation $\hat{\lambda}$ will be the number that minimizes SSE of these standardized values, but as stated above, we usually select a “nice” $\hat{\lambda}$ that is approximately the SSE minimizing number.

3. Tukey Multiple Comparison Procedure

In a multifactor setting, the ANOVA test only tells if *some* treatment effects are significantly different. It does not tell us which treatments are different. To find which means are different, a post-hoc analysis is performed. One method for determining which effects significantly different is called the Tukey Test (Tukey's Honest Significant Difference Test). It is based on the studentized range distribution and allows us to determine which group means were responsible for the rejection of the ANOVA hypothesis test by conducting a family of hypothesis tests on all pairwise differences in means:

$$\mathrm{H}_0: \mu_i = \mu_j$$
$$\mathrm{H}_1: \mu_i \neq \mu_j$$

Here μ_i and μ_j are the mean responses to treatments groups $i \neq j = 1,2,3, \dots, r$ when there are r treatment groups.

If the total sample size is N, the test is based on the studentized range statistic:

$$\mathrm{q} = \frac{\bar{Y}_{largest} - \bar{Y}_{smallest}}{\sqrt{\mathrm{MSE/N}}}$$

At significance level α, the critical number for this studentized range statistic is denoted by $q(1-\alpha; r; N-r)$. For each pairwise comparison we would like to make, we can calculate the test statistic

$$q^* = \frac{\sqrt{2}|\mu_i - \mu_j|}{\sqrt{\mathrm{MSE}\left(\frac{1}{n_i} + \frac{1}{n_j}\right)}}$$

and compare its value with that of $q(1-\alpha; r; N-r)$.

Specifically, if the observed value of $q^* > q(1-\alpha; r; N-r)$ then we reject the null hypothesis and state that the difference between the two groups' mean is statistically significant at level α.

Data Background

Our data comes from Stanford's DeepSolar Project. They constructed this database by gathering information from American Community Survey (ACS). It recorded various community data in 2015. This database provides valuable resources for socioeconomic analysis, as well as insight for education prospect.

Research Question

Based on a real market data set, we come up with a question: "Are city rates of bachelor level or higher education associated with per capita income of that city? Does it also depend on the employment rate of the city?"

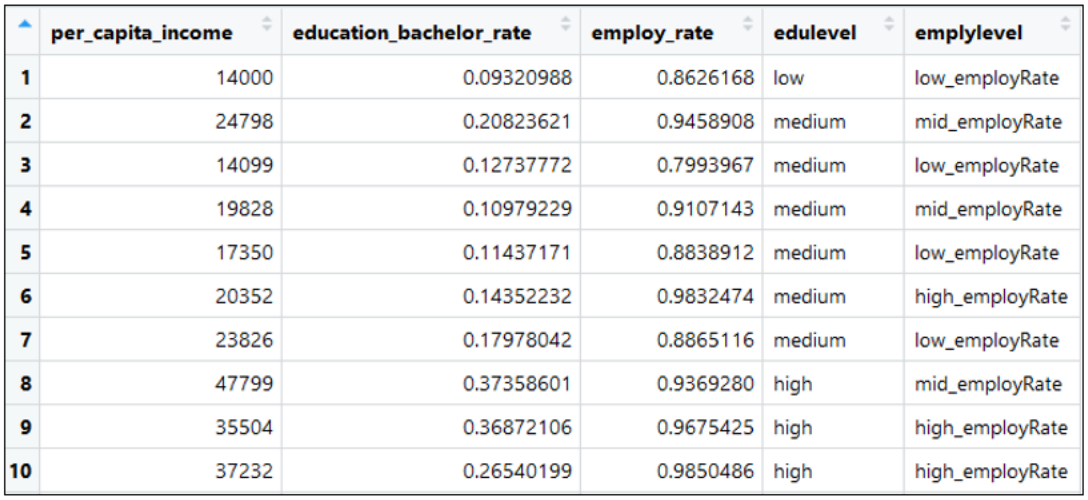
Analysis

In our dataset, we have quantitative variables "bachelor education rate" and "employment rate". We first set these two variables to categorical type and both with three levels: low, medium and high. We use the first and third quantile as the bound of these levels:

	Bachelor Education Rate (R1)	Employment Rate (R2)
Low	$\mathrm{R1} < 0.096$	$\mathrm{R2} < 0.888$
Medium	$0.096 \le \mathrm{R1} < 0.245$	$0.888 \le \mathrm{R2} < 0.949$
High	$0.245 \le \mathrm{R1}$	$0.949 \le \mathrm{R2}$

Figure 3: Set two categorical variables "Bachelor Education Rate" and "Employment Rate"

We use two-factor ANOVA interaction unbalanced model to compare the between-group means. Figure 4 shows the first ten rows of our modified data for this model. Each row represents observations on a unique city, totally 71555 observations.



	per_capita_income	education_bachelor_rate	employ_rate	edulevel	emplylevel
1	14000	0.09320988	0.8626168	low	low_employRate
2	24798	0.20823621	0.9458908	medium	mid_employRate
3	14099	0.12737772	0.7993967	medium	low_employRate
4	19828	0.10979229	0.9107143	medium	mid_employRate
5	17350	0.11437171	0.8838912	medium	low_employRate
6	20352	0.14352232	0.9832474	medium	high_employRate
7	23826	0.17978042	0.8865116	medium	low_employRate
8	47799	0.37358601	0.9369280	high	mid_employRate
9	35504	0.36872106	0.9675425	high	high_employRate
10	37232	0.26540199	0.9850486	high	high_employRate

Figure 4: Modified data for two-factor ANOVA model

Let Y be the per capita income, α_i be the effect of education rate, and β_j be the effect of employment rate. We have a full regression model:

$$Y_{ijk} = \mu_{..} + \alpha_i + \beta_j + \varepsilon_{ijk}$$

where $i = 1,2,3$; $j = 1,2,3$; and $k = 1,2,3, \ldots ,71555$.

We conduct three pairs of hypotheses for our ANOVA model:

- H_0: the means of all education level groups are equal
- H_1: the means of at least one education level groups are different

- H_0: the means of all employment level groups are equal
- H_1: the means of at least one employment level groups are different

- H_0: there is no interaction effect between education level and employment level
- H_1: there is interaction effect between education level and employment level

Before we run the ANOVA test in R, we first need to examine the normality and homoscedasticity assumptions. From the following residuals plot and histogram, we notice that the variances in each group are not equal and the response variable Y has a right skewed distribution.

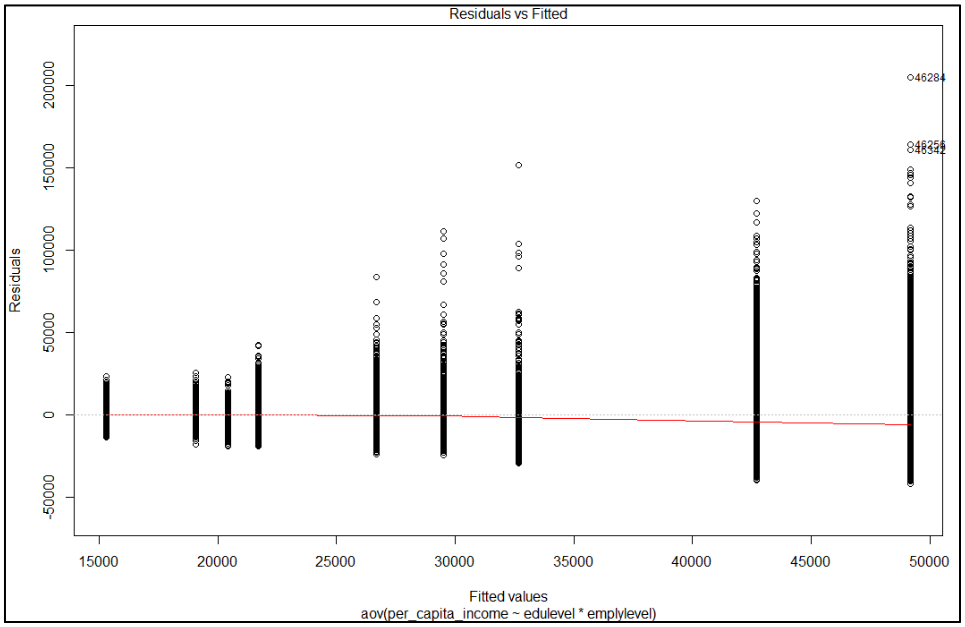


Figure 5: Residuals versus fitted plot

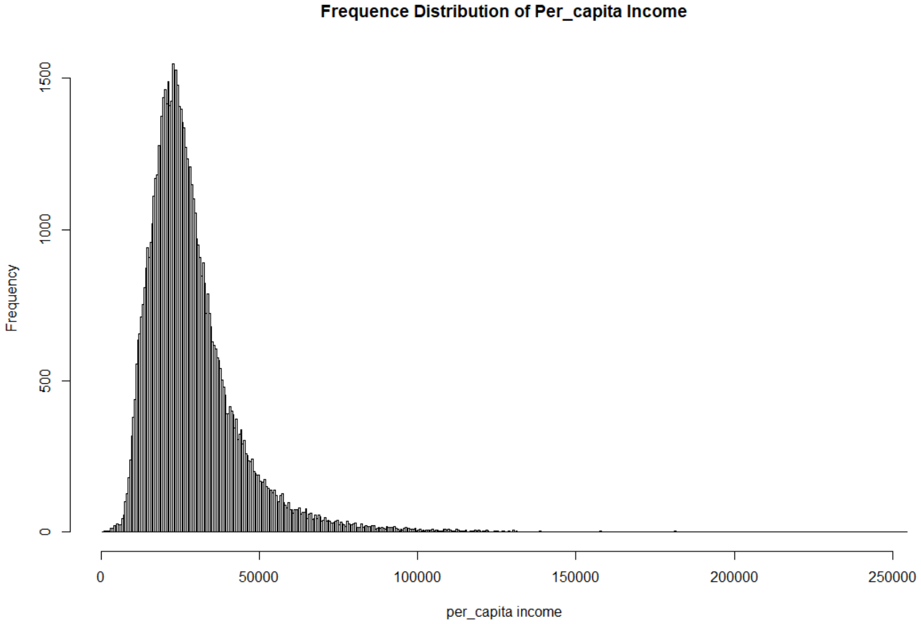


Figure 6: Frequency distribution of per capita income

To improve our model, we use Box-Cox transformation to make a better fit. Figure7 is the graph showing the estimation of λ by the maximum likelihood method. λ has the best estimation at -0.02, and so we select $\lambda = 0$ be our choice. That is, we use log transformation for the response variable.

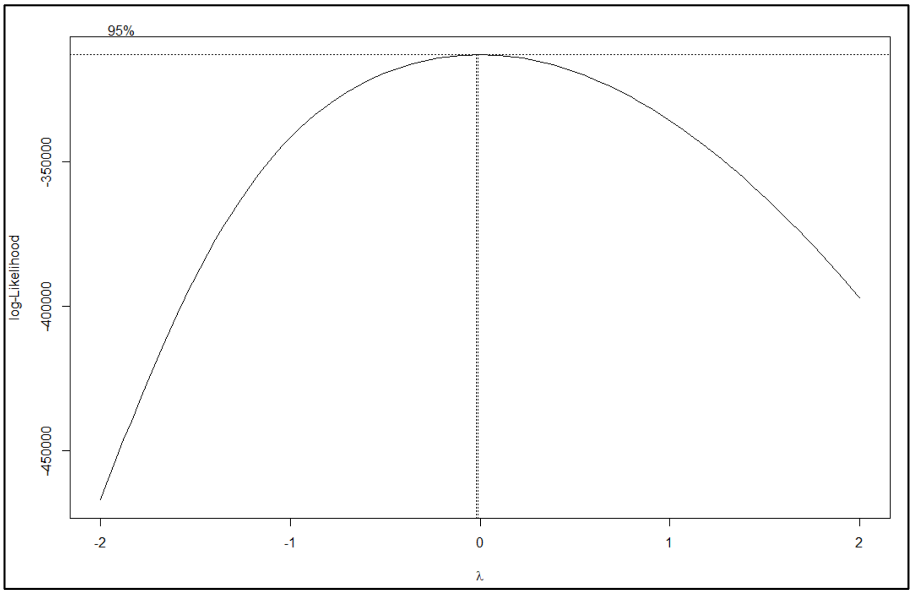


Figure 7: Box Cox transformation for two-factor ANOVA

Therefore, we update our full regression model as follow:

$$\ln(Y_{ijk}) = \mu_{..} + \alpha_i + \beta_j + \varepsilon_{ijk}$$

From figure 8 and 9, we can see that all most every group has similar variance if we ignore the outliers, and it is more likely to be a normal distribution.

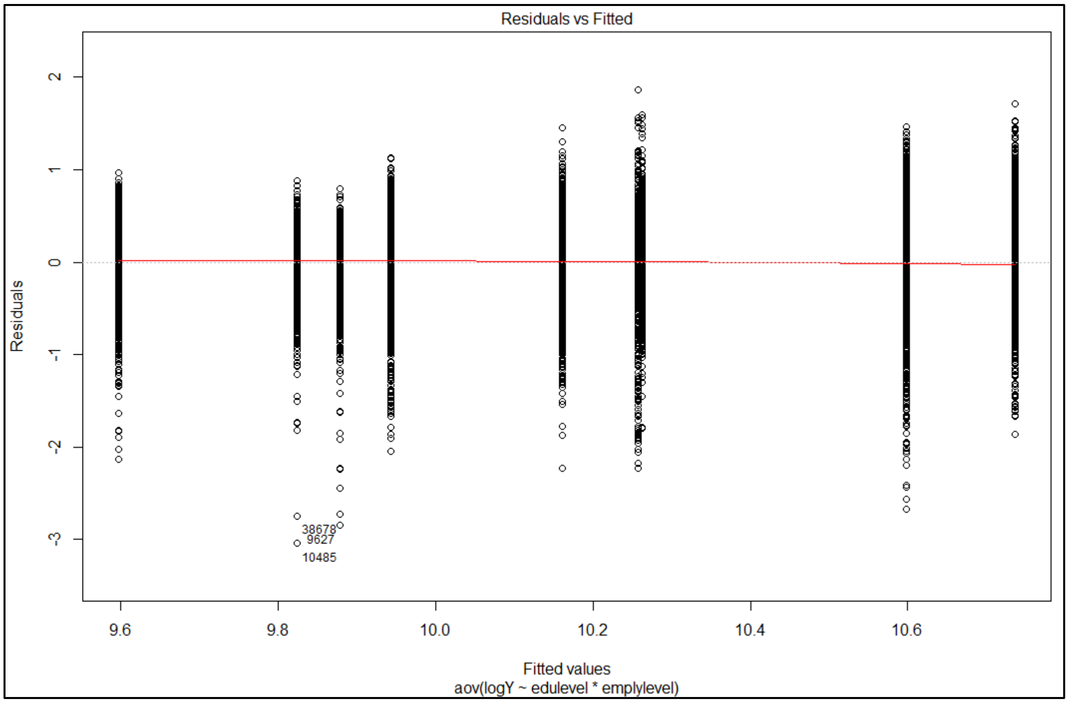


Figure 8: Updated Residuals versus fitted plot

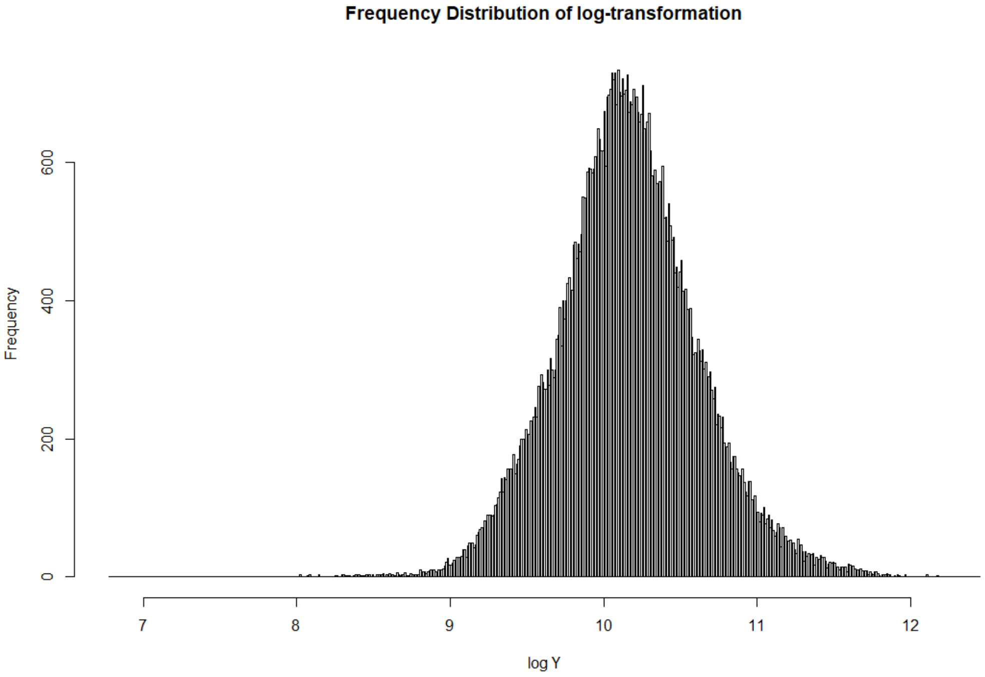


Figure 9: Histogram of updated model

Results

Figure 10 is the box plot of our new transformed model. From this graph we found that, as the level of education level increases, the per capita income will also increase; and for each education level, the per capita income raising along with the changing of employment level.

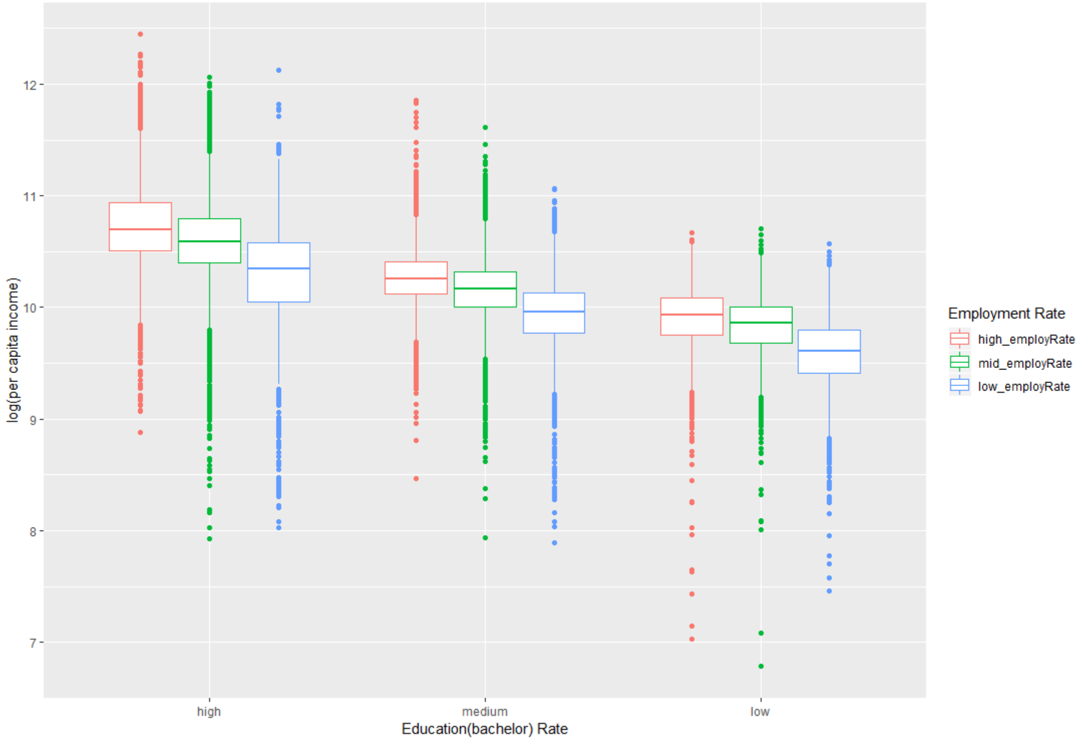
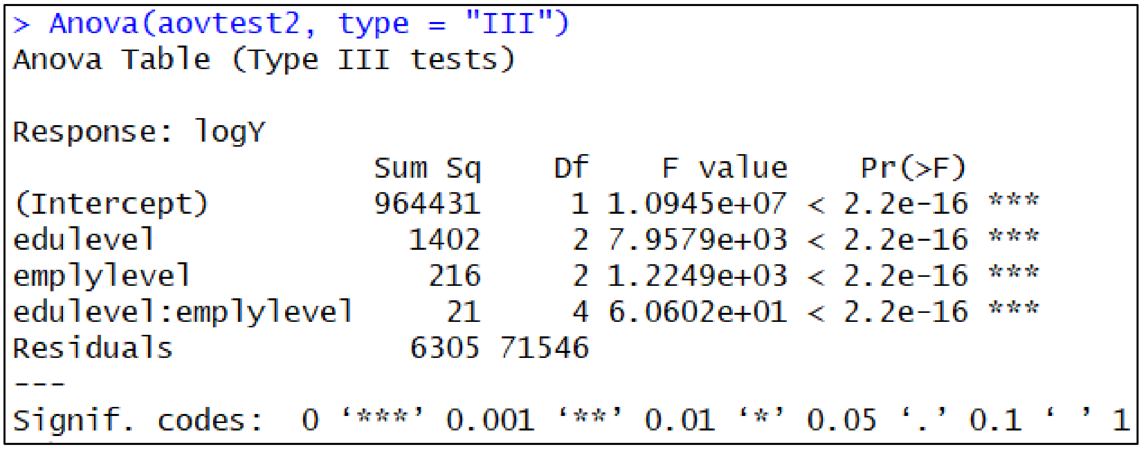


Figure 10: Box plot for updated ANOVA model

Next, we perform ANOVA test to confirm the result we had from box plot. From the ANOVA result table (figure 11):

- the p-value of education level is less than 0.05, which indicates that the levels of education rate are significantly associated with the changing of per capita income.
- the p-value of employment level is less than 0.05, which also indicates that the levels of employment rate are significantly associated with the changing of per capita income.
- the p-value for the interaction between education level and employment level is less 0.05, and so the per capita income depends on both variables.



```
> Anova(aovtest2, type = "III")
Anova Table (Type III tests)

Response: logY
                     Sum Sq    Df    F value    Pr(>F)    
(Intercept)          964431     1 1.0945e+07 < 2.2e-16 ***
edulevel               1402     2 7.9579e+03 < 2.2e-16 ***
emplylevel              216     2 1.2249e+03 < 2.2e-16 ***
edulevel:emplylevel      21     4 6.0602e+01 < 2.2e-16 ***
Residuals              6305 71546                         
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Figure 11: ANOVA result table

At last we show the interaction plot between two factors and the Tukey test. Since the p-value are all extremely small and close to 0, it shows that all levels’ means are significant different to each other.

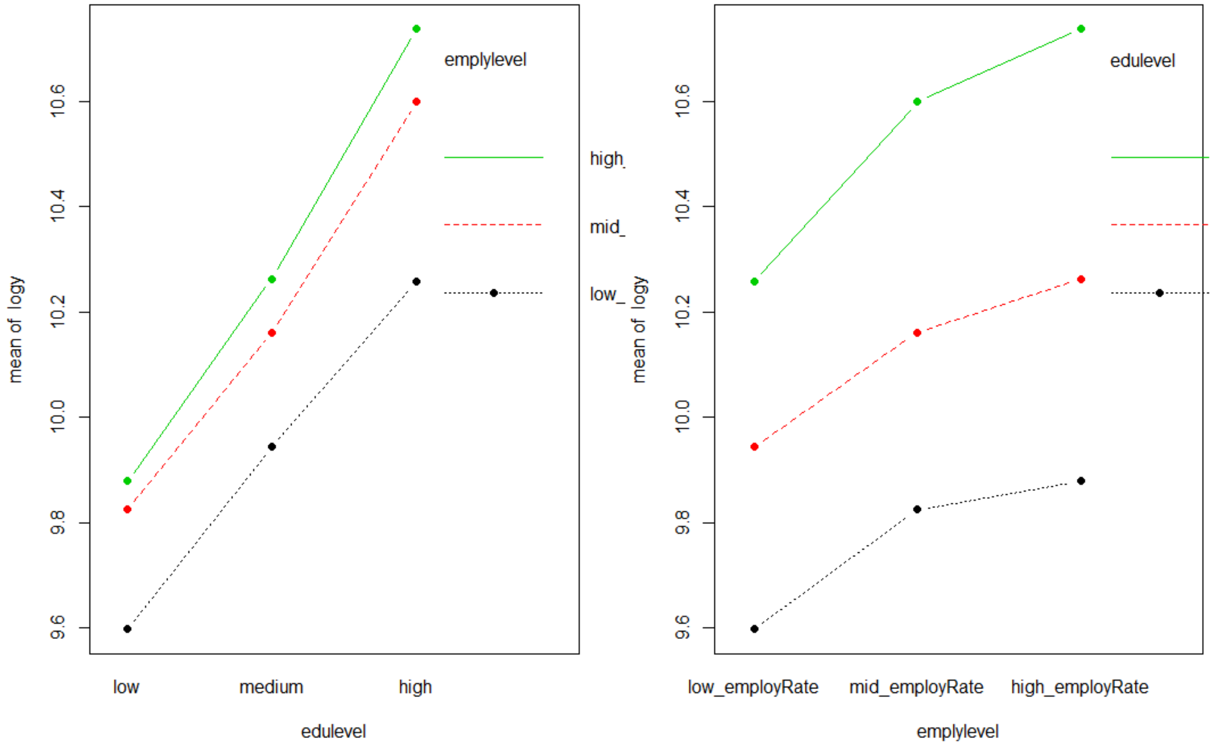
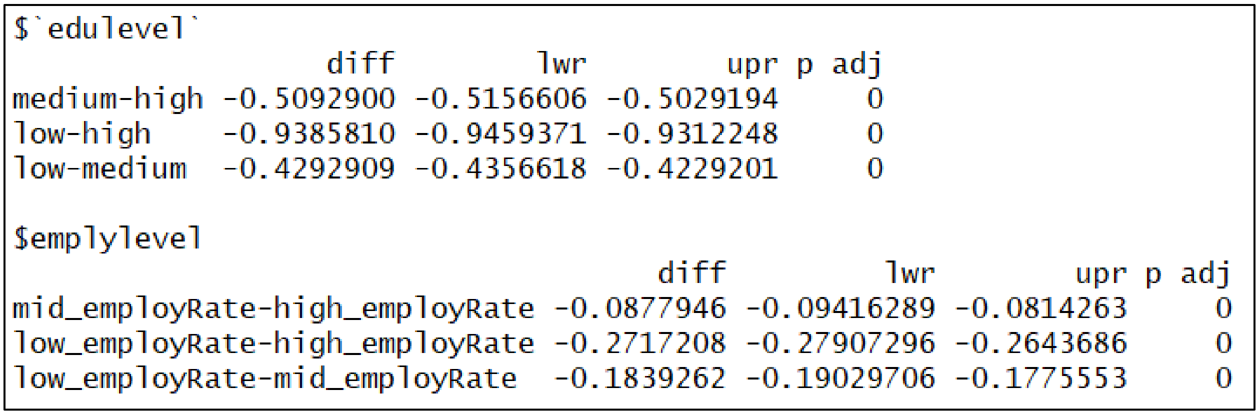


Figure 12: Interaction plot



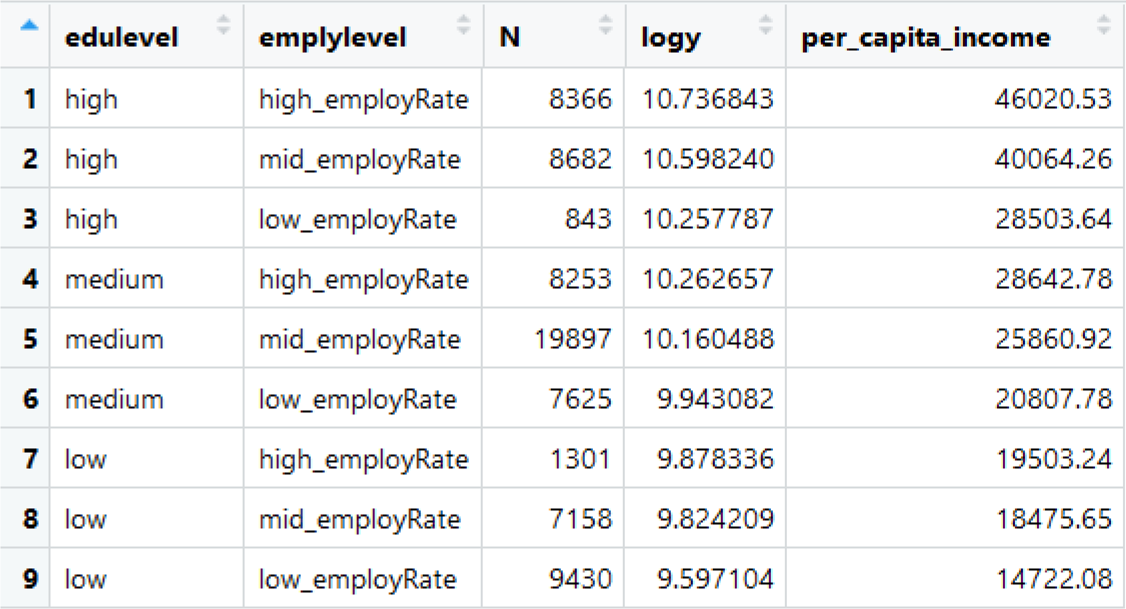
```
$`edulevel`
                  diff        lwr        upr p adj
medium-high -0.5092900 -0.5156606 -0.5029194     0
low-high    -0.9385810 -0.9459371 -0.9312248     0
low-medium  -0.4292909 -0.4356618 -0.4229201     0

$emplylevel
                                   diff         lwr        upr p adj
mid_employRate-high_employRate -0.0877946 -0.09416289 -0.0814263     0
low_employRate-high_employRate -0.2717208 -0.27907296 -0.2643686     0
low_employRate-mid_employRate  -0.1839262 -0.19029706 -0.1775553     0
```

Figure 13: Tukey HSD result table

Conclusion

From the ANOVA test, interaction plot and Tukey test showing above, we conclude that that all levels of education rate are significantly associated with the changing of per capita income and it also depends on the employment rate. Education rate level and employment rate level both have positive relation to the per capita income.

Figure 14 is the summary statistics table. The first two columns represent different levels of two factors. The third column "N" counts the size of each group. "logy" is the mean of transformed response variable, and the last column is the fitted value of per capita income. On average, increasing education rate by 1 level will cause the per capita income changes $10314.5; increasing employment rate by 1 level will cause the per capita income changes $5022.2.



	edulevel	emplylevel	N	logy	per_capita_income
1	high	high_employRate	8366	10.736843	46020.53
2	high	mid_employRate	8682	10.598240	40064.26
3	high	low_employRate	843	10.257787	28503.64
4	medium	high_employRate	8253	10.262657	28642.78
5	medium	mid_employRate	19897	10.160488	25860.92
6	medium	low_employRate	7625	9.943082	20807.78
7	low	high_employRate	1301	9.878336	19503.24
8	low	mid_employRate	7158	9.824209	18475.65
9	low	low_employRate	9430	9.597104	14722.08

Figure 14: Summary statistics table

Bibliography

1. DOBSON, A. J. (2018). *Introduction to generalized linear models*. Place of publication not identified: CHAPMAN & HALL CRC.
2. Kutner, M. H., Nachtsheim, C. J., Neter, J., & Li, W. (2005). *Applied linear statistical models*. New York: McGrawHill Education.
3. Stapleton, J. H. (2009). *Linear statistical models*. Hoboken, NJ: Wiley.