Linear regression - HW1

David Redek, Filip Kulla

October 2020

Introduction

We are given the official census data from US by county. The complete dataset contains information on 3148 counties in 51 American states. For a given county, the dataset contains information about percentage of population under 21-years-old enrolled to educational institutions, median earnings of an adult, racial composition, median age of the population, percentage of children raised in a single parent household and percentage of obese adults. We will analyse this dataset in an effort to answer questions about a) the median income in Texas, b) obesity rates in Texas, Oklahoma, Arizona, and New Mexico, c) dependence of percentage of children enrolled to educational institutes and the percentage of children raised in single parent household in Texas and d) wealth of Georgia, Alabama, Mississippi, and Louisiana.

Income in Texas

In this section, we are interested in the median income in the counties of Texas, and its relation to the racial composition of the county. The question that we want to answer is whether median income in the counties of Texas where non–Hispanic whites are a majority (that is, more than 50% of the total population) is significantly higher than in other Texas counties.

We will think of median incomes in the counties of Texas where non–Hispanic whites are a majority and median incomes in the counties of Texas where non–Hispanic whites are not a majority as of two independent random samples $\boldsymbol{X}$ and $\boldsymbol{Y}$. In this section $\boldsymbol{X} = (X_1, \ldots, X_n)$ will denote median incomes in countries of Texas where non–Hispanic whites are a majority while $\boldsymbol{Y} = (Y_1, \ldots, Y_m)$ will denote median incomes in countries of Texas where non–Hispanic whites are not a majority. To form some preliminary ideas about our data let us look at basic descriptive statistics in Table 1. There are 182 counties, where non–Hispanic whites are majority and 72 where they are not.

	Min.	**1st Qu.**	**Median**	**Mean**	**3rd Qu.**	**Max.**	**Sd.**
$\boldsymbol{X}$	15501	22360	24647	25246	27137	44497	4596
$\boldsymbol{Y}$	14206	19047	22114	22205	24921	38465	4341

Table 1: Descriptive statistics for the median income in the counties of Texas where non–Hispanic whites are a majority ($\boldsymbol{X}$) and in the counties of Texas where non–Hispanic whites are not a majority ($\boldsymbol{Y}$).

To visualise our data we will use a boxplot in Figure 1.

Descriptive statistics as well as the boxplot seem to suggest that the median income in the counties of Texas where non–Hispanic whites are a majority might be higher than the median income in in the counties of Texas where non–Hispanic whites are not a majority.

From normal Q-Q plots (Figure 2), especially the one for the counties where non–Hispanic whites are a majority, it seems that we cannot believe in normality of our samples. Therefore, instead of trusting in homoscedasticity and using Student's t-test, which would be also a valid option because the assumption of

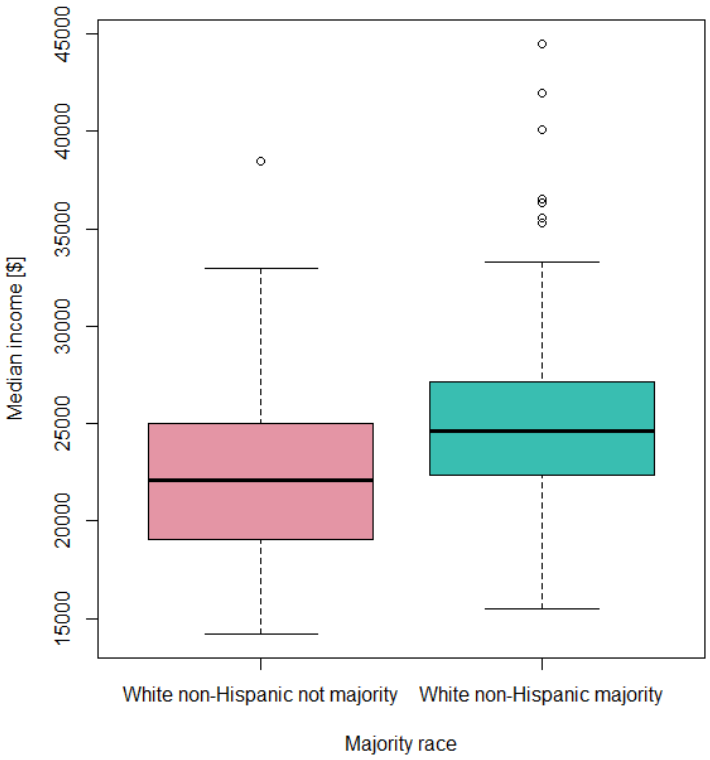

Figure 1: Boxplots of income dependent on racial composition.

equal variances does not seem to be violated too badly and given equal variances Student's t-test is asymptotically correct even without normality, we will use **Welch´s t-test**.

Model: $\mathcal{F} = \{F_X \in \mathcal{L}^2_+, F_Y \in \mathcal{L}^2_+\}$.

Tested parameters: expected values $\mu_X = \mathsf{E}\, X_i$ and $\mu_Y = \mathsf{E}\, Y_i$.

Null hypothesis: $H_0 : \mu_X - \mu_Y \leq 0$.

Alternative hypothesis: $H_1 : \mu_X - \mu_Y > 0$.

Test statistic:

$$Z_{n,m} = \frac{\overline{X_n} - \overline{Y_m}}{\sqrt{\frac{S_X^2}{n} + \frac{S_Y^2}{m}}},$$

where $\overline{X_n}$, $\overline{Y_m}$ are sample means and S_X^2, S_Y^2 are sample variances of $\boldsymbol{X}$ and $\boldsymbol{Y}$.

Critical region:

$$H_0 \text{ is rejected} \Leftrightarrow Z_{n,m} \geq t_f(1-\alpha),$$

where $t_f(1-\alpha)$ is $(1-\alpha)$-quantile of t-distribution with f degrees of freedom, where

$$f = \frac{\left(\frac{S_X^2}{n} + \frac{S_Y^2}{m}\right)^2}{\frac{(S_X^2)^2}{n^2(n-1)} + \frac{(S_Y^2)^2}{m^2(m-1)}}.$$

P-value (asymptotic): $1 - T(z)$, where z is observed value of $Z_{n,m}$ and T is cumulative distribution function of t-distribution with f degrees of freedom.

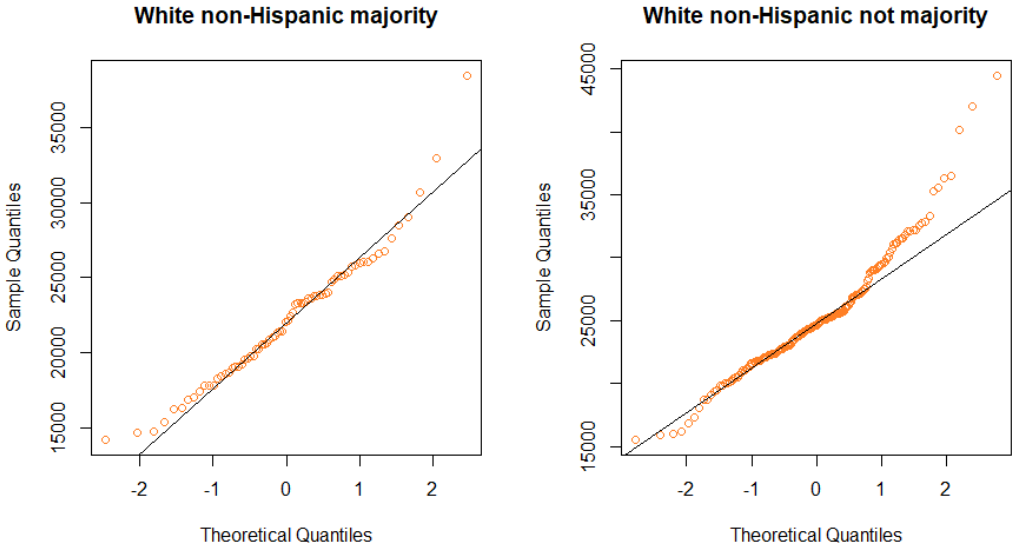

Figure 2: Q-Q plots of income dependent on racial composition.

Confidence interval (asymptotic): $\left(\overline{X_n} - \overline{Y_m} - t_f(1-\alpha)\sqrt{\frac{S_X^2}{n} + \frac{S_Y^2}{m}}, \infty\right)$.

For our data, we have $n = 182$ and $m = 72$. Observed value of $Z_{n,m}$ is 4.95, f is 137.34, p-value is 1.079×10^{-6} and confidence interval for the quantity $\mu_X - \mu_Y$ is $(2023.25, \infty)$.

Since p-value is lower than specified $\alpha = 0.05$ we reject the null hypothesis that $\mu_X - \mu_Y \leq 0$ in favour of the alternative hypothesis that $\mu_X - \mu_Y > 0$. In other words, we have proved that median income in the counties of Texas where non–Hispanic whites are a majority is significantly higher than in other Texas counties. Furthermore, thanks to the confidence interval, we can infer that the expected median income in the counties of Texas where non–Hispanic whites are a majority is higher than in other Texas counties by at least 2023$.

Further we would like to investigate relation between racial composition of county and its median income, again in the state of Texas. To visualize our data on racial composition and median income, we can plot scatter plots for percentage of certain race and median income in dollars (Figure 3).

Now for the sake of simplicity we will choose percentage of just one race and try to quantify its relation to median income in the state of Texas. For this analysis, we will choose the percentage of white not Latino people in each county. By looking at the scatter plot for this case, we might suggest, that it holds true that the higher the percentage of white not Latino population, the higher the median income. To justify this hypothesis we will try to fit a linear model.

Let´s think of our data as of random sample $(Y_i, X_i)^T, i = 1, \ldots, n$, where Y represents median income and X percentage of white not Latino population in a given county. In our case $n = 254$, which is the number of counties in Texas.

From the scatterplot it seems that it might be possible to model $\mathsf{E}[Y|X = x]$ as a line $\beta_0 + \beta_1 x$. From the diagnostics plots (Figure 4 and Figure 5) it seems that the assumptions of a linear model are more or less satisfied, however, it seems that the most important one might be slightly compromised because for low fitted

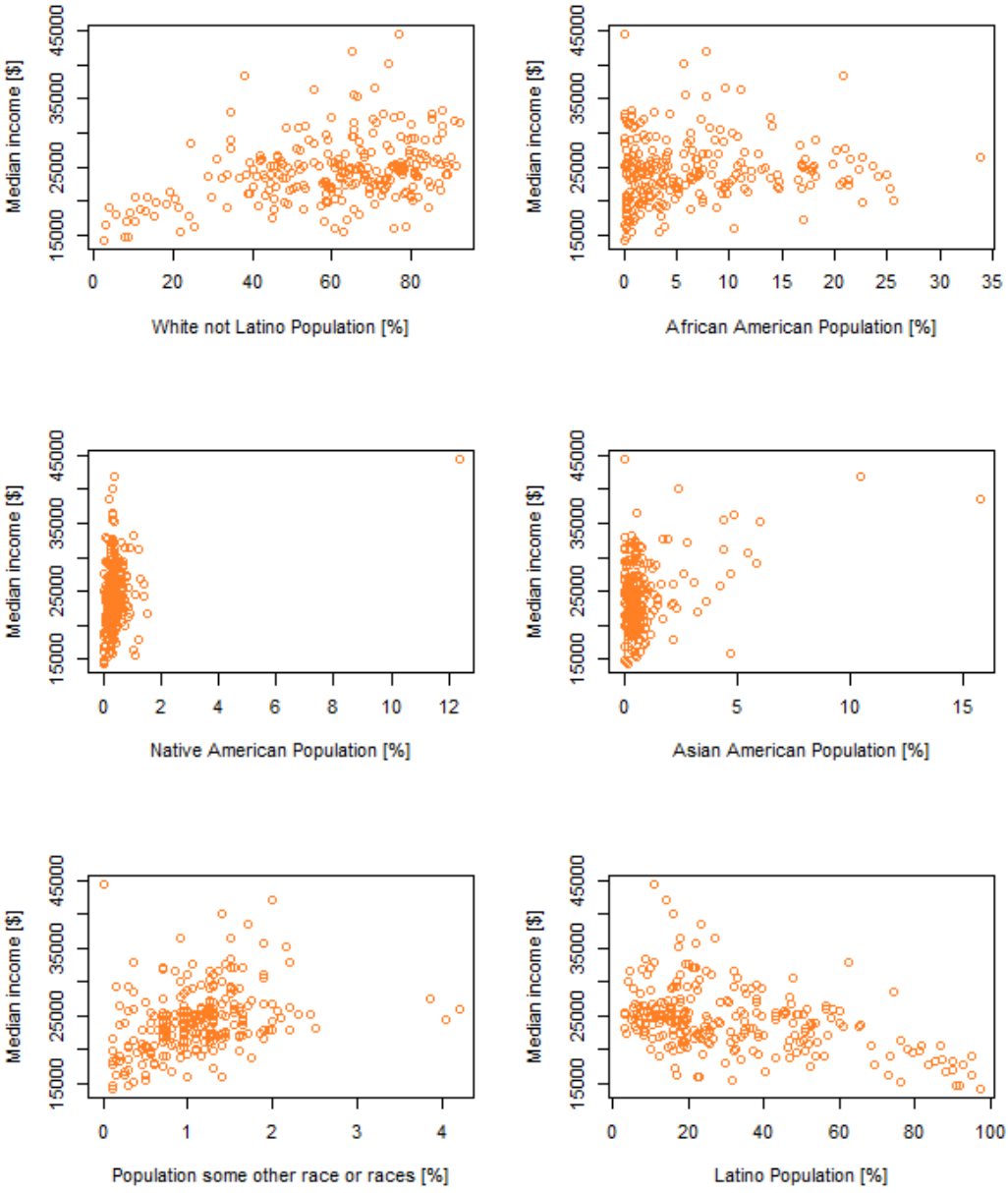

Figure 3: Scatter plots for percentage of certain race and median income in dollars.

values, the residuals are not around 0 but below 0. This problem is evident from Figure 6. It is obvious that for counties with small percentage of white not Latino population our regression line overestimates the median income. Maybe it would be possible to solve this issue by modeling $\mathsf{E}[Y|X = x]$ as $\beta_0 + \beta_1 log(x)$. From diagnostics plots for this model (Figure 7 and Figure 8) we see that we have probably chosen correct regression function since in Residuals vs. Fitted values/log(White not Latino Population) graphs, residuals are scattered around 0. The assumption of homoscedasticity might be a compromised for approximately 15 datapoints with the lowest fitted values. The assumption of normality seems to be a little compromised for both models to the same degree. From Figure 9 we see that both regression lines are very similar once the percentage of white not Latino population is sufficiently high. If the percentage is small, the second regression line models the data seemingly better. We will choose the second model and will try to make inference about β_1.

Tested parameter: β_1.

Null hypothesis: $\beta_1 = 0$.

Alternative hypothesis: $\beta_1 \neq 0$

Test statistic:

$$T_1 = \frac{\widehat{\beta_1}}{\sqrt{MS_e \; v_{11}}},$$

where v_{11} is element of matrix $V = (X^\mathsf{T} X)^{-1}$, where X is the model matrix, MS_e denotes residual mean square and $\widehat{\beta}$ is LSE estimator of β.

Critical region:

$$H_0 \text{ is rejected} \Leftrightarrow |T_1| \geq t_{n-2}(1 - \alpha/2),$$

where $t_{n-2}(1 - \alpha/2)$ is $(1 - \alpha/2)$-quantile of t_{n-2}-distribution.

P-value (asymptotic): $2(1 - F_n(|t|))$, where t is observed value of T_1 and F_n is cumulative distribution function of t_{n-2}-distribution.

For our data $\widehat{\beta} = (10750.7, 3423.4)$, the observed value of the test statistic T_1 is 7.34 and calculated p-value is 2.95×10^{-12}. Thus we can reject the null hypothesis with sufficient statistical significance. Hence, we have found that it is not enough to model $\mathsf{E}[Y|X = x]$ as a constant, but need the term $log(x)$ and thus have found a trend in the data. The interpretation of β in our model is that β_0 is conditional expectation of median income if 1 percent of population is white not Latino and β_1 is is increase in conditional expectation of median income if logarithm of percentage of white not Latino population increases by 1. In this case the interpretation is not so nice as in the first model where β_0 would represent expected median income in a population in which are no white not Latino people. Furthermore, we could also deny that $\beta_1 \leq 0$ with p-value of 1.48×10^{-12}. Our data are in accordance with assertion that with increasing percentage of white not Latino population the median income increases.

Obesity in Texas, Arizona, Oklahoma and New Mexico

In this section we will investigate the obesity rates in four south-western states - Texas, Oklahoma, Arizona, and New Mexico. We are interested in

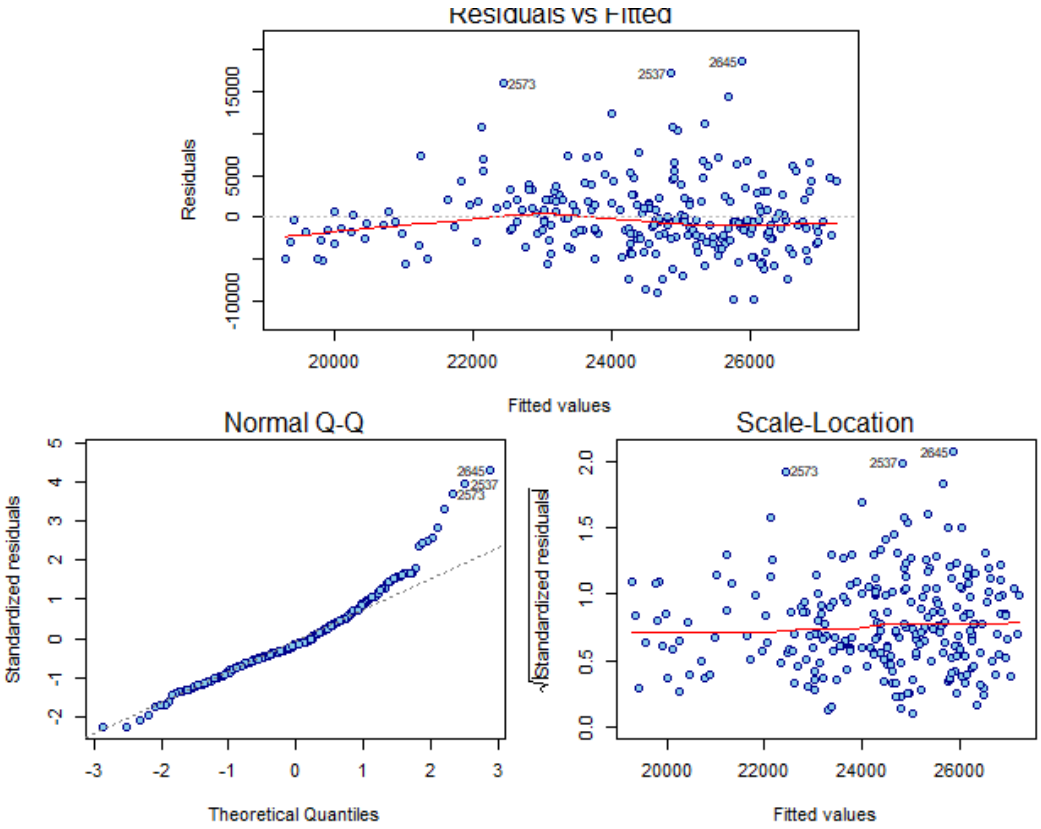

Figure 4: Diagnostics plot for linear model assuming a line.

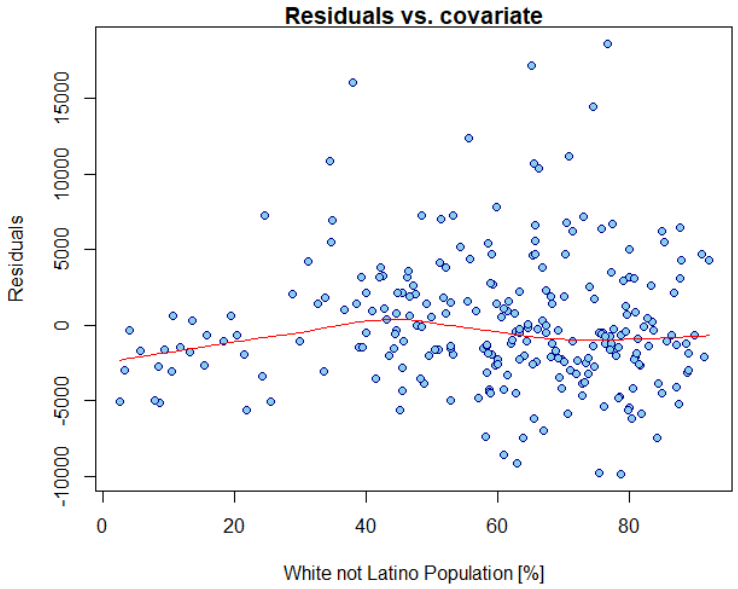

Figure 5: Fitted vs covariate for linear model assuming a line.

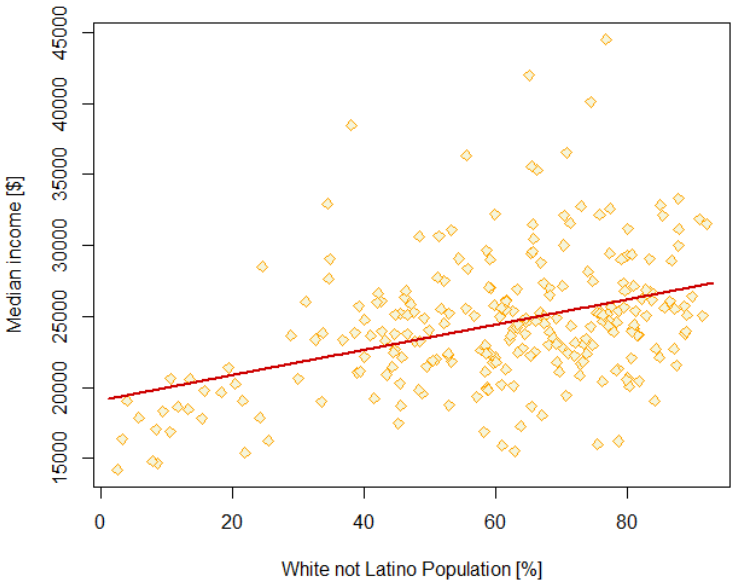

Figure 6: Regression line for linear model assuming a line.

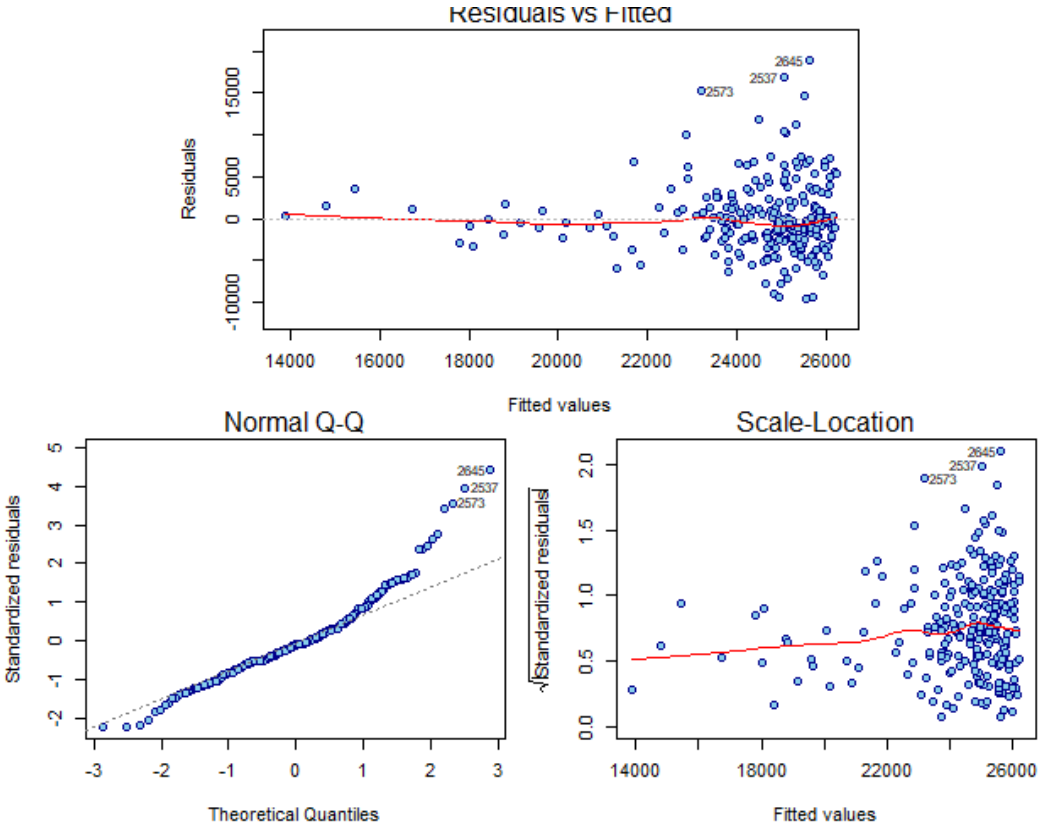

Figure 7: Diagnostics plot for linear model assuming log.

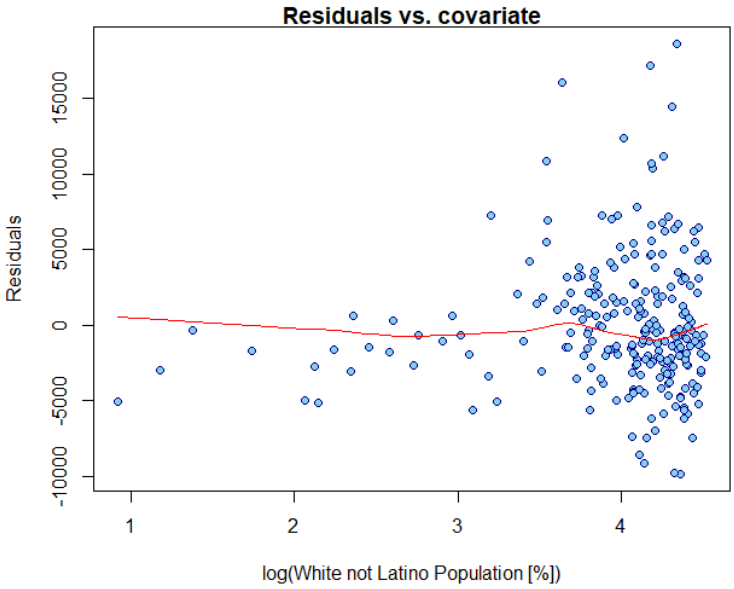

Figure 8: Fitted vs covariate for linear model assuming log.

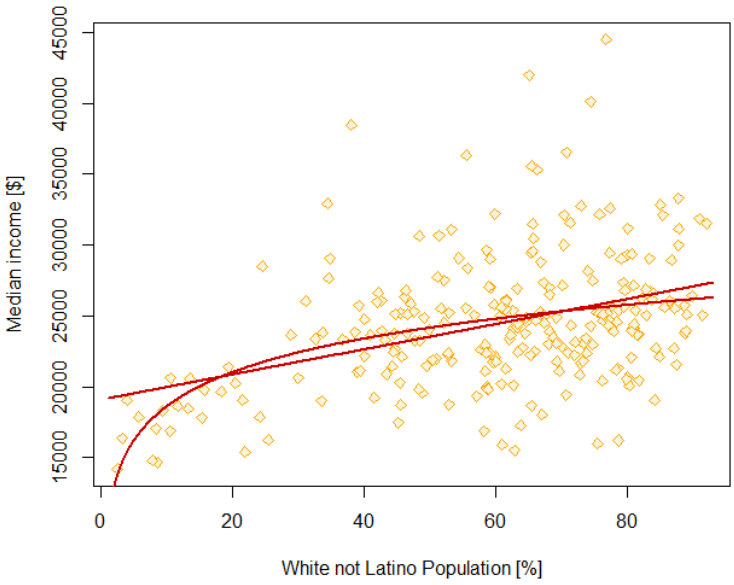

Figure 9: Comparison of 2 regression lines.

whether the obesity rates depend on a given state or whether it can be said that the obesity rates in these four states are the same.

To get some idea about our data we will look at basic descriptive statistics.

	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	Sd.
Texas	0.2340	0.2870	0.2955	0.2981	0.3108	0.3720	0.0197
Arizona	0.2010	0.2375	0.2710	0.2781	0.3155	0.3420	0.0459
Oklahoma	0.2890	0.3160	0.3300	0.3339	0.3510	0.3930	0.0232
New Mexico	0.1380	0.2340	0.2530	0.2540	0.2670	0.3580	0.0432

Table 2: Descriptive statistics for the obesity rates in the counties of Texas, Arizona, Oklahoma and New Mexico.

Furthermore, we will visualize our data using a boxplot (Figure 10).

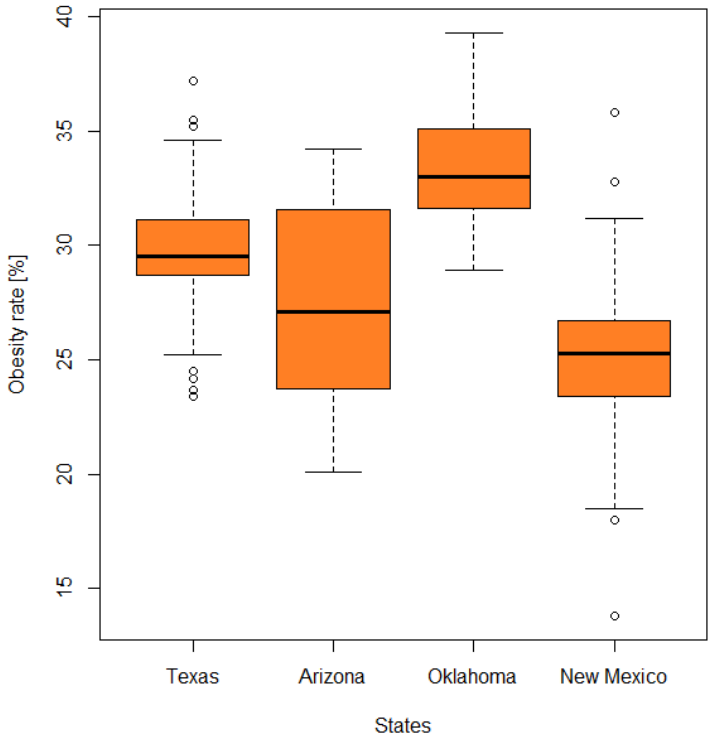

Figure 10: Boxplots visualising obesity rates in four American states.

From the descriptive statistics and the boxplots it seems that the expected obesity rates in Texas, Arizona, Oklahoma and New Mexico are not the same. This is especially evident when we look at the boxes corresponding to Texas, Oklahoma and New Mexico and notice that they do not even intersect (3rd Qu. for New Mexico $<$ 1st Qu. for Texas $<$ 3rd Qu. for Texas $<$ 1st Qu. for Oklahoma). It seems that it could be the case that the expected obesity rate in New Mexico is less than the expected obesity rate in Texas and that this expected rate is less than the expected obesity rate in Oklahoma.

Furthermore, from normal Q-Q plots (Figure 11) it seems that we cannot completely believe in normality of our samples. What is worse, however, is that from the table and the boxplot it seems that the variances in individual states are not the same. If the variances were the same, to test whether expected obesity rates in given states are equal, we could use standard ANOVA, because ANOVA is asymptotically correct even without normality. In our situation, however, we will use **generalised ANOVA**.

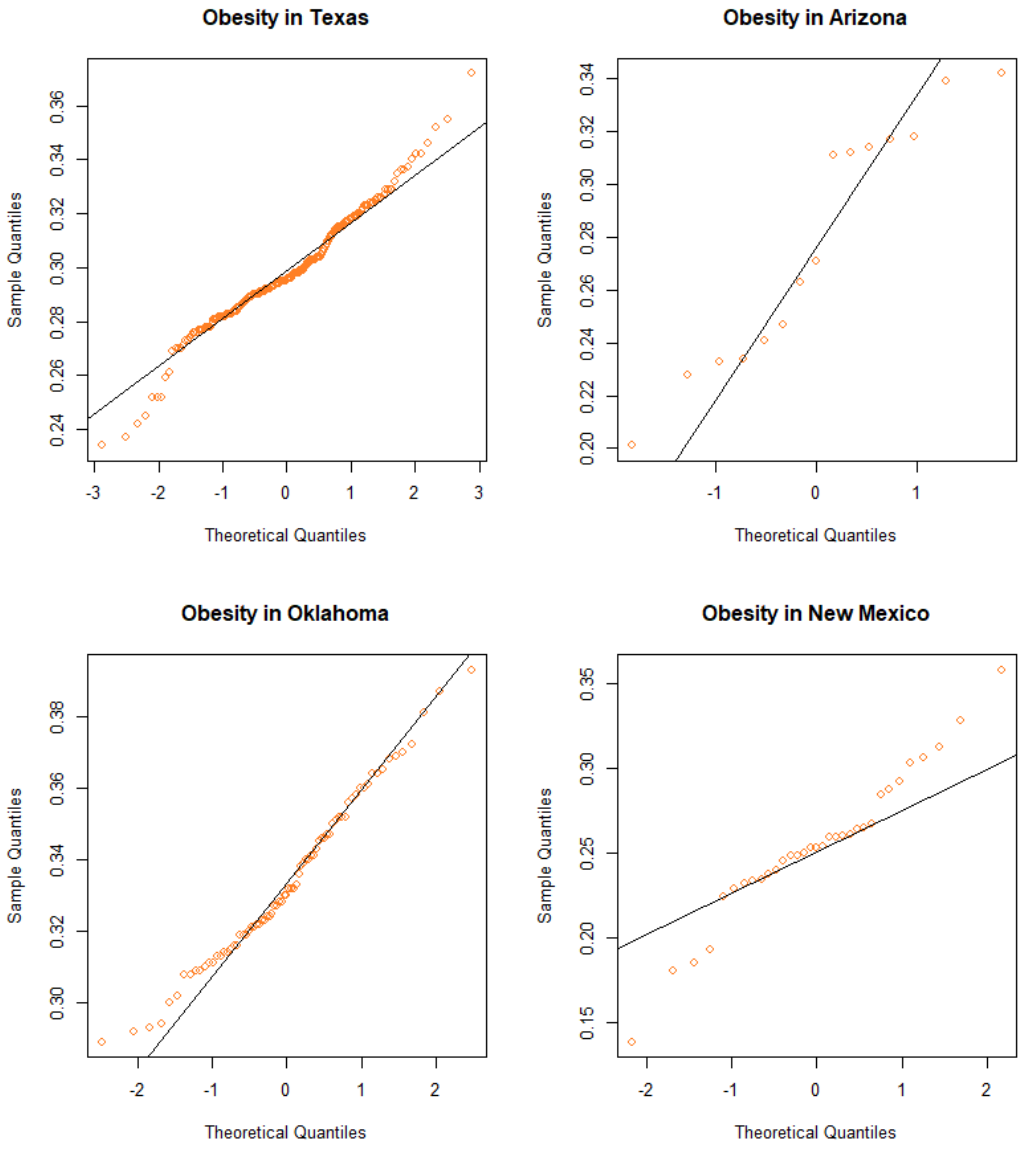

Figure 11: $Q-Q$ plots of obesity rates in four American states.

For every $i \in \{1,2,3,4\}$ we have a random sample $\mathbf{Y}_i = (Y_{i1}, \ldots, Y_{in_i})$ with expected value $\mathsf{E}\, Y_{i1} = \mu_i$ and variance $\mathrm{var} Y_{i1} = \sigma_i^2$. Let F_i denote cumulative distribution function of Y_{i1}. In our situation every $\mathbf{Y}_i$ represents one state.

Model: $\mathcal{F} = \{F_i \sim (\mu_i, \sigma_i^2), \mu_i \in \mathbb{R}, 0 < \sigma_i^2 < \infty, i \in \{1,2,3,4\}\}$.

Tested parameters: expected values $\mu_1, \mu_2, \mu_3, \mu_4$.

Null hypothesis: $H_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4$.

Alternative hypothesis: $H_1 : \exists\, i \neq j \in \{1,2,3,4\} : \mu_i \neq \mu_j$.

Test statistic:

$$F_w = \frac{\sum_{i=1}^{K} w_i(\overline{Y_{i+}} - \overline{Y_w})^2}{K-1} \frac{1}{1+2\Lambda(K-2)},$$

where $K = 4$ is number of independent random samples $\boldsymbol{Y}_i$, $w_i = \frac{n_i}{S_i^2}$ is weight assigned to the i-th group, S_i^2 is sample variance of $\boldsymbol{Y}_i$, $\overline{Y_{i+}}$ is sample mean of $\boldsymbol{Y}_i$, $\overline{Y_w} = \frac{\sum_{i=1}^{K} w_i \overline{Y_{i+}}}{\sum_{i=1}^{K} w_i}$ is an estimation of common expectation under H_0 and

$$\Lambda = \frac{\sum_{i=1}^{K} \frac{1}{n_i - 1} \left(1 - \frac{w_i}{\sum_{j=1}^{K} w_j}\right)^2}{K^2 - 1}$$

is some correction. Under H_0 in model $\mathcal{F}$ (that means without homoscedasticity and normality) it holds that $(K-1)F_w \xrightarrow{D} \chi^2_{K-1}$, when $min(n_1, \ldots, n_K) \to \infty$ and at the same time $\frac{n_i}{N} \to \lambda_i > 0$, $i \in \{1, \ldots, K\}$ ($N = \sum_{i=1}^{K} n_i$).

Critical region:

$$H_0 \text{ is rejected} \Leftrightarrow (K-1)F_w \geq \chi^2_{K-1}(1-\alpha),$$

where $\chi^2_{K-1}(1-\alpha)$ is $(1-\alpha)$-quantile of χ^2_{K-1}.

P-value (asymptotic): $1 - F_{\chi^2_{K-1}}((K-1)f_w)$, where f_w is observed value of F_w and $F_{\chi^2_{K-1}}$ is cumulative distribution function of χ^2_{K-1}.

For our data, we have $n_1 = 254, n_2 = 15, n_3 = 77, n_4 = 33$, which is number of counties in Texas, Arizona, Oklahoma and New Mexico. Observed value of F_w is 64.29 and p-value is numerically zero. Hence, we reject the null hypothesis that the expected obesity rates in Texas, Arizona, Oklahoma and New Mexico are the same in favour of the alternative hypothesis that they are not equal. Our data are in accordance with the assertion that the obesity rates in our states depend on a particular state.

Now we might be interested to find out which states differ significantly. The problem with (generalised) ANOVA is that although it says that expected values are not the same, it does not indicate which couple is different. In comparison, Bonferroni method, although generally weak, does indicate problematic couples. Nevertheless, it often happens that although ANOVA rejects equality of the expected values, Bonferroni method cannot find a couple which differs at the adjusted significance level. In our situation, however, as was discussed above, Texas, Oklahoma and New Mexico seem to be very different. We will test equality of expected obesity rates for every couple of states at the adjusted significance level $\alpha_0 = \alpha/6 = 0.05/6$. If we reject this equality for a particular couple, we will declare expected obesity rates in these 2 countries to be different at the overall significance level of $\alpha = 0.05$. As in the first section, we will use Welch´s t-test. Since we test equality of expected obesity rates, p-value will be calculated as $2\,(1 - T(|z|))$, where z is observed value of $Z_{n,m}$ and T is cumulative distribution function of t-distribution with f degrees of freedom.

Table 3 displays individual tests. Thus, we declare expected obesity rates in Texas - Oklahoma, Texas - New Mexico, Arizona - Oklahoma and Oklahoma - New Mexico to be different at the overall significance level of $\alpha = 0.05$.

	$Z_{n,m}$	**P-value (asymptotic)**
Texas - Arizona	1.6854	0.1136
Texas - Oklahoma	-12.234	$< 2.2\times10^{-16}$
Texas - New Mexico	5.8	1.607×10^{-6}
Arizona - Oklahoma	-4.5981	0.0003249
Arizona - New Mexico	1.7176	0.09789
Oklahoma - New Mexico	10.03	1.689×10^{-12}

Table 3: Results of Welch´s t-test for every couple of states.

Finally, we would like to order our 4 states from "the least obese" one to "the most obese" one. What we can do, is that we can construct a confidence interval for μ_i, for every $i \in \{1, 2, 3, 4\}$. From these one-dimensional confidence intervals we will construct 4-dimensional confidence set for $(\mu_1, \mu_2, \mu_3, \mu_4)^\mathsf{T}$, as the Cartesian product of our intervals, so that this set covers $(\mu_1, \mu_2, \mu_3, \mu_4)^\mathsf{T}$ with the specified probability of $(1 - \alpha)$. Let $I(Y_i)$ be a $(1 - \alpha)^{\frac{1}{4}} \times 100\%$ confidence interval for μ_i. Let $M = I(Y_1) \times I(Y_2) \times I(Y_3) \times I(Y_4)$. Then

$$\mathsf{P}\left((\mu_1, \mu_2, \mu_3, \mu_4)^\mathsf{T} \in M\right) = \mathsf{P}\left(\bigcap_{i=1}^{4} \left[\mu_i \in I(Y_i)\right]\right) = \prod_{i=1}^{4} \mathsf{P}\left(\mu_i \in I(Y_i)\right) = (1 - \alpha).$$

Now, if every point $(x_1, x_2, x_3, x_4)^\mathsf{T} \in M$ satisfies that $x_1 < x_2 < x_3 < x_4$, then it is reasonable to infer that also $(\mu_1, \mu_2, \mu_3, \mu_4)^\mathsf{T}$ satisfies that $\mu_1 < \mu_2 < \mu_3 < \mu_4$.

Let us calculate $(1-\alpha)^{\frac{1}{4}} \times 100\%$ asymptotic confidence intervals based on the central limit theorem. We know that for a random sample $\boldsymbol{X} = (X_1, \ldots, X_n)$, $(1 - \beta) \times 100\%$ asymptotic confidence interval for $\mathsf{E}\, X_1$ is of the form

$$\left(\overline{X_n} - u_{1-\frac{\beta}{2}} \frac{S_n}{\sqrt{n}}, \overline{X_n} + u_{1-\frac{\beta}{2}} \frac{S_n}{\sqrt{n}}\right),$$

where $\overline{X_n}$ is sample mean of $\boldsymbol{X}$ and S_n^2 is sample variance of $\boldsymbol{X}$. Therefore, $(1 - \alpha)^{\frac{1}{4}} \times 100\%$ asymptotic confidence interval for the expected obesity rate in New Mexico is $(0.2353, 0.2727)$, for the expected obesity rate in Arizona it is $(0.2486, 0.3076)$, for the expected obesity rate in Texas it is $(0.2951, 0.3012)$ and for the expected obesity rate in Oklahoma it is $(0.3273, 0.3405)$.

Since 3 of these one-dimensional intervals are disjoint we can infer that the expected obesity rate in New Mexico $<$ the expected obesity rate in Texas $<$ the expected obesity rate in Oklahoma. Because confidence intervals for New Mexico and Arizona intersect we cannot claim that New Mexico is "less obese" than Arizona or vice versa. Similarly for Arizona and Texas. To conclude, we inferred that the expected obesity rate in New Mexico $<$ the expected obesity rate in Texas $<$ the expected obesity rate in Oklahoma, and that the expected obesity rate in Arizona $<$ the expected obesity rate in Oklahoma.

Dependence of children school enrollment and single parent households

In this section we would like to investigate whether the percentage of children enrolled to educational institutes in the counties of Texas depends on the percentage of children raised in a single parent household.

We will think of percentage of school enrollment and percentage of single parent households as of a random sample $(X_1,Y_1)^T,\ldots,(X_n,Y_n)^T$. In this section X_i will denote percentage of school enrollment while Y_i will denote percentage of single parent households in a certain Texas county. Texas has 254 counties, but we have one missing value for percentage of single parent household (in county of Loving), so we will omit this observation and only use the remaining 253 counties. So in the context of this section let $n = 253$ To form some preliminary ideas about our data let us look at basic descriptive statistics in Table 4.

	Min.	**1st Qu.**	**Median**	**Mean**	**3rd Qu.**	**Max.**	**Sd.**
$\boldsymbol{X}$	38.95	71.30	74.30	73.91	76.80	91.60	5.65
$\boldsymbol{Y}$	0.00	26.00	32.30	31.54	37.80	56.50	8.71

Table 4: Descriptive statistics for $\boldsymbol{X}$ and $\boldsymbol{Y}$, where $\boldsymbol{X} = (X_1,\ldots,X_{253})$ denotes percentage of school enrollment and $\boldsymbol{Y} = (Y_1,\ldots,Y_{253})$ denotes percentage of single parent households.

By just looking at a scatter plot (Figure 12), we can't really see and suggest dependency of observed values, therefore we would like to test independence of X_i and Y_i. For this we will use test based on **Pearson correlation coefficient**.

Furthermore, by looking at Q–Q plots (Figure 13), we cannot really assume normality of our sample. Therefore we cannot be sure of (X_i,Y_i) being from bivariate normal distribution, required by the model on which Pearson correlation coefficient test is based. Despite that, test based on Pearson correlation coefficient works just fine as an asymptotic test, which we will discuss later.

Model: $\mathcal{F} = \{F_X \in \mathcal{L}^2_+, F_Y \in \mathcal{L}^2_+\}$.
Null hypothesis: $H_0 : X_i$ and Y_i are independent.
Alternative hypothesis: $H_1 : X_i$ and Y_i are not independent.
Test statistic:

$$T_n = \sqrt{n-2}\frac{\widehat{\rho_n}}{\sqrt{1-\widehat{\rho_n^2}}}$$

where $\widehat{\rho_n}$ is sample correlation coefficient

$$\widehat{\rho_n} = \frac{S_{XY}}{S_X S_Y} = \frac{\sum_{i=1}^n (X_i - \overline{X_n})(Y_i - \overline{Y_n})}{\sqrt{\sum_{i=1}^n (X_i - \overline{X_n})^2 \sum_{i=1}^n (Y_i - \overline{Y_n})^2}}$$

Critical region:

$$H_0 \text{ is rejected} \Leftrightarrow |T_n| \geq t_{n-2}(1-\alpha/2),$$

where $t_{n-2}(1-\alpha/2)$ is $(1-\alpha/2)$-quantile of t_{n-2}–distribution.

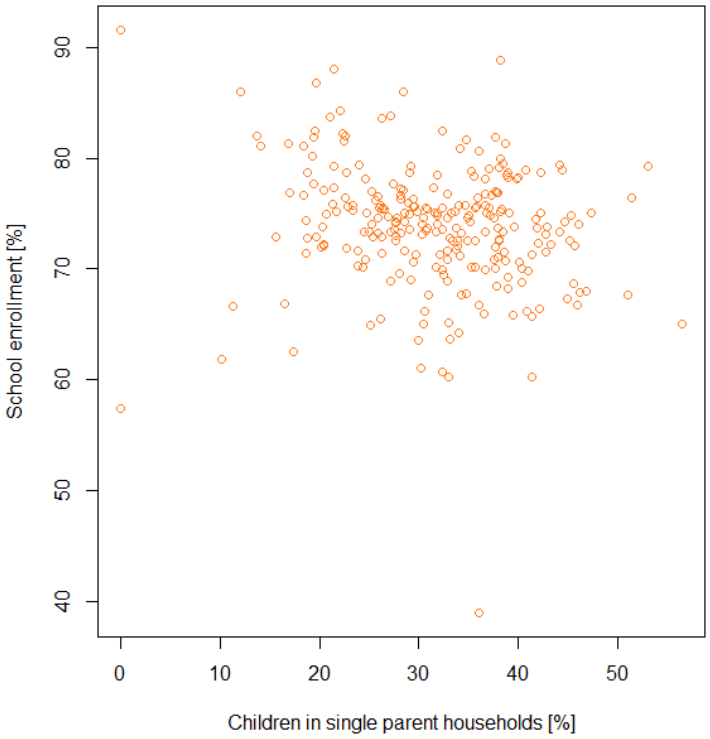

Figure 12: Scatter plot of percentage of children in single parent household and percentage of children school enrollment.

P-value (asymptotic): $2(1 - F_n(|t|))$, where t is observed value of T_n and F_n is cumulative distribution function of t_{n-2}–distribution.

We know that under the assumptions of $(X_i,Y_i)^T, i = 1, \ldots, n$ being a random sample from bivariate distribution with finite, non–singular variance matrix and X_i, Y_i being **independent** (under H_0), test statistic T_n has asymptotically standard normal distribution. In other words

$$T_n = \sqrt{n-2}\frac{\widehat{\rho_n}}{\sqrt{1-\widehat{\rho_n^2}}} \xrightarrow{n\to\infty} N(0,1)$$

Thus we will not assume normality of data and we will consider this test as an asymptotic test. Out of caution, we will also use, more conservatively, quantiles of t_{n-2} distribution, as in standard exact test.

For our data, observed value of T_n is -3.05 and corresponding p–value 0.0026. That means, that we reject the null hypothesis that the percentage of children enrolled to educational institutes is independent with percentage of children raised in single parent household, both in state of Texas, in favour of the alternative hypothesis that they are not independent.

We will quantify the relation between school enrollment and single parent

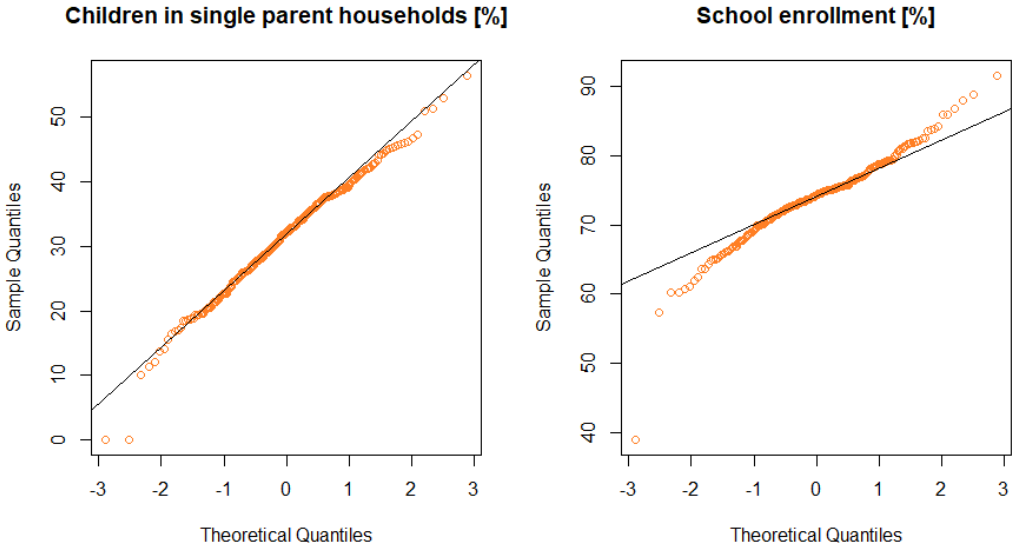

Figure 13: Q-Q plots of percentage of children in single parent household and percentage of children school enrollment.

households by correlation coefficient

$$\rho_{X_i,Y_i} = \frac{\mathsf{cov}(X_i,Y_i)}{\sqrt{\mathsf{var}X_i \mathsf{var}Y_i}}$$

Consistent estimate of this value is sample correlation coefficient $\widehat{\rho_n}$, defined earlier. Observed value of sample correlation coefficient is $\widehat{\rho_n} = -0.189$.

Next we would like to construct a confidence interval for this numerical characteristic. Accurate distribution of $\widehat{\rho_n}$ is too complicated, however using Δ–method we can derive asymptotic distribution. Since correlation is shift invariant, for simplicity we can assume that the random variables are centered and we will center our dataset for this purpose as well.

Let's denote

$$T_n = \frac{1}{n}\begin{pmatrix} \sum_{i=1}^n X_i \\ \sum_{i=1}^n Y_i \\ \sum_{i=1}^n X_i^2 \\ \sum_{i=1}^n Y_i^2 \\ \sum_{i=1}^n X_i Y_i \end{pmatrix}, \mu = \begin{pmatrix} 0 \\ 0 \\ \mathrm{cov}(X_i,X_i) \\ \mathrm{cov}(Y_i,Y_i) \\ \mathrm{cov}(X_i,Y_i) \end{pmatrix}, \Sigma = \begin{pmatrix} \mathrm{cov}(X_1,X_1) & \dots & \mathrm{cov}(X_1,X_1Y_i) \\ \mathrm{cov}(Y_1,X_1) & \dots & \mathrm{cov}(Y_1,X_1Y_i) \\ \vdots & \ddots & \vdots \\ \mathrm{cov}(X_1Y_1,X_1) & \dots & \mathrm{cov}(X_1Y_!,X_1Y_1) \end{pmatrix}.$$

From the central limit theorem we know, that

$$\sqrt{n}\,(T_n - \mu) \xrightarrow[n\to\infty]{d} \mathsf{N}_5(\vec{0},\Sigma),$$

Now we will use Δ–method twice.

Let's define function $g : \mathbb{R}^5 \to \mathbb{R}^3$ as $g(a,b,c,d,e) = (c - a^2, d - b^2, e - ab)$. Clearly, g is differentiable. Now, from Δ–method we know that

$$\sqrt{n}(g(T_n) - g(\mu)) \xrightarrow[n\to\infty]{d} \mathsf{N}_3(\vec{0}, Dg(\mu)\Sigma Dg(\mu)^T).$$

Now let's define function $h : \mathbb{R}^3 \to \mathbb{R}$ as $h(a,b,c) = \frac{c}{\sqrt{ab}}$. Clearly, g is differentiable. Now, again from Δ–method we know that

$$\sqrt{n}(h(g(T_n)) - h(g(\mu))) \xrightarrow[n\to\infty]{d} \mathsf{N}(0, Dh(g(\mu))Dg(\mu)\Sigma Dg(\mu)^T Dh(g(\mu))^T).$$

We chose functions g,h in a way, that

$$h\left(g\left(\frac{1}{n}\begin{pmatrix}\sum_{i=1}^n X_i\\ \sum_{i=1}^n Y_i\\ \sum_{i=1}^n X_i^2\\ \sum_{i=1}^n Y_i^2\\ \sum_{i=1}^n X_iY_i\end{pmatrix}\right)\right) = h\begin{pmatrix}S_X^2\\ S_Y^2\\ S_{XY}\end{pmatrix} = \widehat{\rho_n},$$

and

$$h(g(\mu)) = \rho_{XY}.$$

Let's denote

$$\begin{aligned}
A &= Dh(g(\mu))Dg(\mu)\Sigma Dg(\mu)^T Dh(g(\mu))^T\\
&= \frac{1}{4}\mathrm{cov}(X^2,X^2)\mathrm{cor}(X,Y)^2\frac{1}{\mathrm{var}(X)^2} + \frac{1}{2}\mathrm{cov}(X^2,Y^2)\mathrm{cor}(X,Y)^2\frac{1}{\mathrm{var}(X)\mathrm{var}(Y)}\\
&+ \frac{1}{4}\mathrm{cov}(Y^2,Y^2)\mathrm{cor}(X,Y)^2\frac{1}{\mathrm{var}(Y)^2} - \mathrm{cov}(X^2,XY)\mathrm{cor}(X,Y)\frac{1}{\sqrt{\mathrm{var}(X)^3}\mathrm{cor}(X,Y)rt\mathrm{var}(Y)}\\
&- \mathrm{cov}(Y^2.XY)\mathrm{cor}(X,Y)\frac{1}{\sqrt{\mathrm{var}(X)}\sqrt{\mathrm{var}(Y)^3}} + \mathrm{cov}(XY,XY)\frac{1}{\mathrm{var}(X)\mathrm{var}(Y)}\mathrm{cor}(X,Y).
\end{aligned}$$

Now we can write

$$\sqrt{n}\frac{(\widehat{\rho_n} - \rho)}{\sqrt{A}} \xrightarrow[n\to\infty]{d} \mathsf{N}(0,1).$$

Now using Cramer-Slutsky theorem, where $\widehat{A}$ is a consistent estimate of A, we can write

$$\sqrt{n}\frac{(\widehat{\rho_n} - \rho)}{\sqrt{\widehat{A}}} \xrightarrow[n\to\infty]{d} \mathsf{N}(0,1).$$

Estimate $\widehat{A}$ can be obtained easily by using sample covariance and sample variance, but we will not state it separately in order not to prolong this section.

From here we can construct an asymptotic confidence interval for parameter ρ_{XY} as

$$\left(\widehat{\rho_n} - \sqrt{\widehat{A}}\frac{u_{1-\frac{\alpha}{2}}(1-\widehat{\rho_n}^2)}{\sqrt{n}}, \widehat{\rho_n} + \sqrt{\widehat{A}}\frac{u_{1-\frac{\alpha}{2}}(1-\widehat{\rho_n}^2)}{\sqrt{n}}\right).$$

For our data we have observed confidence interval $(-0.194, -0.183)$. That means, that asymptotically with probability of 95%, this interval covers the real value of correlation coefficient of school enrollment and single parent household percentages. That would mean, that with rising percentage of children raised

in single parent household we would probably observe decline in percentage of children enrolled to educational institutes in the counties of Texas.

Alternatively we could assume normality of our random sample and our confidence interval would simplify to just

$$\left(\widehat{\rho_n} - \frac{u_{1-\frac{\alpha}{2}}(1-\widehat{\rho_n}^2)}{\sqrt{n}}, \widehat{\rho_n} + \frac{u_{1-\frac{\alpha}{2}}(1-\widehat{\rho_n}^2)}{\sqrt{n}}\right),$$

which for our data would be $(-0.193, -0.185)$.

Furthermore we could use the knowledge of this confidence interval for hypothesis testing. Let $(\eta_L(X), \eta_U(X))$ be a confidence interval for parameter θ_X (of some random sample X) with confidence level of $1-\alpha$. Let's assume a test of hypothesis $H_0 : \theta_X = \theta_0$ against $H_1 : \theta_x \neq \theta_0$. From duality of hypothesis testing and confidence intervals we know, that test based on decision rule

$$\begin{aligned} \text{we refuse } H_0, &\text{ if } \theta_0 \notin (\eta_L(X), \eta_U(X)), \\ \text{we do not refuse } H_0, &\text{ if } \theta_0 \in (\eta_L(X), \eta_U(X)), \end{aligned}$$

has significance level α.

We can also use this knowledge in context of our problem. Our hypothesis is H_0 : X_i and Y_i are independent. We also know that independence of two random variables implies their zero correlation. Therefore if our null hypothesis was $\widetilde{H_0} : \rho_{XY} = 0$ and we could reject this hypothesis in favor of $\widetilde{H_1} : \rho_{XY} \neq 0$ by a decision rule based on confidence interval. This approach might also imply the dependency of the two random variables.

Wealth in southern states

In the last section we investigate income in southern states of the USA - Georgia, Alabama, Mississippi and Louisiana. We divide counties in each state into three groups based on the level of their median income by the following rule:

- rich if median income of the county is at least \$30 000,
- medium if median income is at least \$20 000, but less than \$30 000,
- poor if median income is less than \$20 000.

We are interested in whether the proportion of counties classified as poor, medium or rich is the same in all four states. We have two categorical random variables – X and Y. Random variable X has four categories, that give us the location of the county – Georgia, Alabama, Mississippi or Louisiana, the Y has three categories that give classification based on median income – rich, medium or poor. We will use the $\boldsymbol{\chi}^2$ **Test of Independence** to assess whether the income classification of county is independent with the location of that particular county. If we reject the hypothesis of independence, then the proportion of counties classified by median income is not the same in the four states.

More generally, let's assume a random sample $\boldsymbol{X} = (X_1,Y_1)^T, \ldots, (X_N,Y_N)^T$ of size N, where X, Y are both categorical random variables, $X \in \{1, \ldots, J\}$, $Y \in \{1, \ldots, K\}$. Let's denote number of instances classified into the $j-th$ category of X and $k-th$ category of Y as

$$n_{jk} = \sum_{i=1}^{N} \mathbb{1}\{X_i = j, Y_i = k\}, \qquad j = 1, \ldots, J, \quad k = 1, \ldots, K$$

Using this notation we will put together a contingency table (Table 5) filled with values n_{jk}:

	Georgia	**Louisiana**	**Alabama**	**Mississippi**	Σ
Rich	3	22	5	3	33
Medium	58	126	54	61	299
Poor	6	11	5	18	40
$\sum$	67	159	64	82	372

Table 5: Contingency table of rich/medium/poor counties in a given state.

We will also visualize this table in Figure 14.

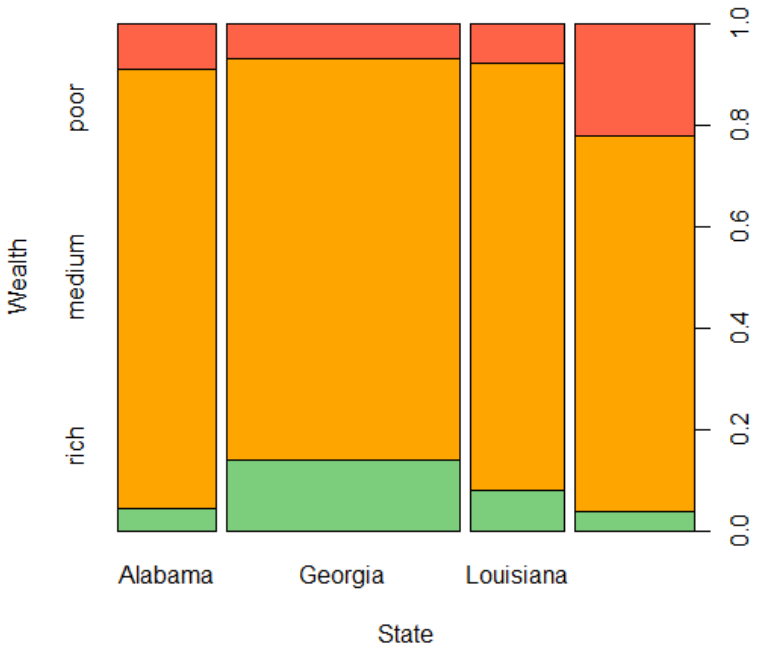

Figure 14: Plot showing wealth classification by state.

Null hypothesis: X and Y are independent.
Alternative hypothesis: X and Y are not independent.

Test statistic:

$$\chi^2 = \sum_{j=1}^{J}\sum_{k=1}^{K} \frac{\left(n_{jk} - \frac{n_{j+}n_{+k}}{N}\right)^2}{\frac{n_{j+}n_{+k}}{N}}.$$

Critical region:

$$H_0 \text{ is rejected} \Leftrightarrow \chi^2 \geq \chi^2_{(J-1)(K-1)}(1-\alpha),$$

where $\chi^2_{(J-1)(K-1)}(1-\alpha)$ is $(1-\alpha)$-quantile of $\chi^2_{(J-1)(K-1)}$.

P-value (asymptotic): $1 - F_{\chi^2_{(J-1)(K-1)}}(c)$, where c is observed value of χ^2 and $F_{\chi^2_{(J-1)(K-1)}}$ is cumulative distribution function of $\chi^2_{(J-1)(K-1)}$.

For our data, we have $N = 372$, since there are totally 372 counties in our four observed states. Observed value of χ^2 is 21.76 and p-value is 0.0013.

Since p-value is lower than the specified $\alpha = 0.05$, we reject the null hypothesis that X and Y are independent in favour of the alternative hypothesis that they are not. We have proved that the proportion of counties classified as poor/medium/rich in the four southern states is not the same.

Since we rejected the null hypothesis that the proportion of counties classified as poor/medium/rich in the four southern states is the same, we would like to know how it is different and to somehow quantify the differences in wealth between our states. For each state we will calculate how many percent of its counties are rich/medium/poor (Table 6). From the table we can see that the proportion of rich counties in Georgia is higher than in any other state, while the proportion of poor counties is, simultaneously, lower than in any other state. Similarly, in Mississippi almost one quarter of counties are poor, which is the highest proportion from all the states. At the same time, proportion of rich states in Mississippi is the lowest from all the states. Thus, it is reasonable to infer that Georgia could be the most wealthy state and Mississippi the least wealthy one.

	Georgia	**Louisiana**	**Alabama**	**Mississippi**
Rich[%]	13.8	7.8	4.5	3.7
Medium[%]	79.2	84.4	86.6	74.4
Poor[%]	6.9	7.8	9	22

Table 6: Proportion of rich/medium/poor counties in a given state.

Finally, we will investigate whether it is true that the proportion of rich counties in Georgia is higher than the proportion of rich counties in Alabama by at least 5 percentage points? In this situation it is natural to think of counties in Georgia and counties in Alabama as of two independent random samples $\boldsymbol{X} = (X_1, \ldots, X_n)$ and $\boldsymbol{Y} = (Y_1, \ldots, Y_m)$ from alternative distributions with parameters p_1 and p_2, where p_1 represents the proportion of rich counties in Georgia and p_2 represents the proportion of rich counties in Alabama. We want to test whether $p_1 \geq p_2 + 0.05$.

Model: $\mathcal{F} = \{F_X \sim Alt(p_1), F_Y \sim Alt(p_2)\}$.

Tested parameters: probabilities p_1 and p_2.

Null hypothesis: $H_0 : p_1 - p_2 \leq 0.05$.

Alternative hypothesis: $H_1 : p_1 - p_2 > 0.05$.

Test statistic:

$$T_d = \frac{(\widehat{p_1} - \widehat{p_2}) - 0.05}{\sqrt{\frac{\widehat{p_1}(1-\widehat{p_1})}{n} + \frac{\widehat{p_2}(1-\widehat{p_2})}{m}}},$$

where $\widehat{p_1}$ is sample mean of $\boldsymbol{X}$ and $\widehat{p_2}$ is sample mean of $\boldsymbol{Y}$.

Critical region:

$$H_0 \text{ is rejected} \Leftrightarrow T_d \geq u_{1-\alpha},$$

where $u_{1-\alpha}$ is $(1-\alpha)$-quantile of $N(0,1)$.

P-value (asymptotic): $1 - \phi(t_d)$, where t_d is observed value of T_d and ϕ is cumulative distribution function of $N(0,1)$.

For our data, we have $n = 159$ and $m = 67$, corresponding with number of counties in Georgia and Alabama. Observed value of T_d is 1.17, p-value is 0.12.

Since p-value is too high, we cannot reject H_0. That means that our data are not sufficiently different from the distribution they would have under the null hypothesis and we cannot prove that that the proportion of rich counties in Georgia is higher than the proportion of rich counties in Alabama by more than 5 percentage points.

Conclusion

To conclude, we have found that median income in the counties of Texas where non–Hispanic whites are a majority is significantly higher than in other Texas counties and that the data are in accordance with assertion that with increasing percentage of white not Latino population the median income increases. Furthermore, we found out that obesity rates in Texas, Arizona, Oklahoma and New Mexico are not the same and that New Mexico is "less obese" than Texas, which is "less obese" than Oklahoma. Moreover, we found that the percentage of children enrolled to educational institutes in the counties of Texas does depend on the percentage of children raised in a single parent household. Finally, we proved that the proportion of counties classified as poor/medium/rich in Alabama, Georgia, Louisiana and Mississippi is not the same.