Assignment 4
MingMing Z — Student No. 21058539

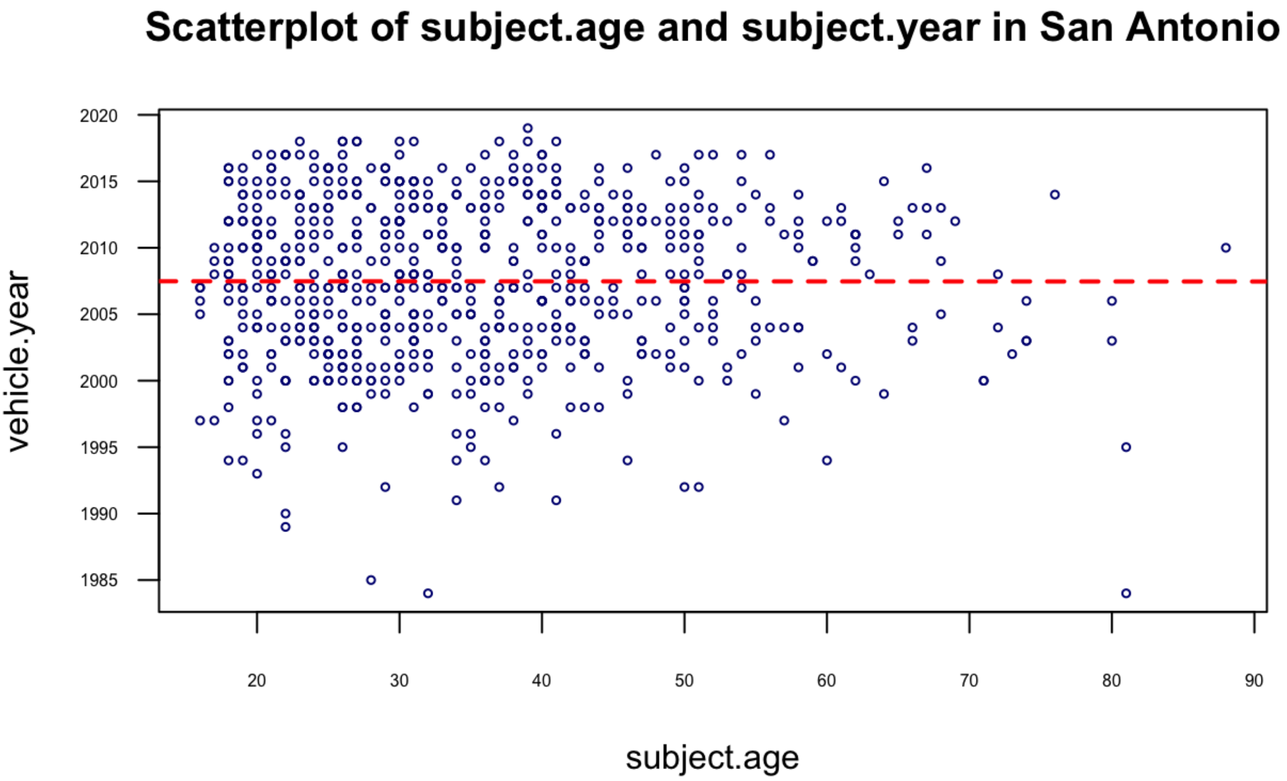
Analysis 1

1a: My ID number is 21058539. I will analyze the data for San Antonio.

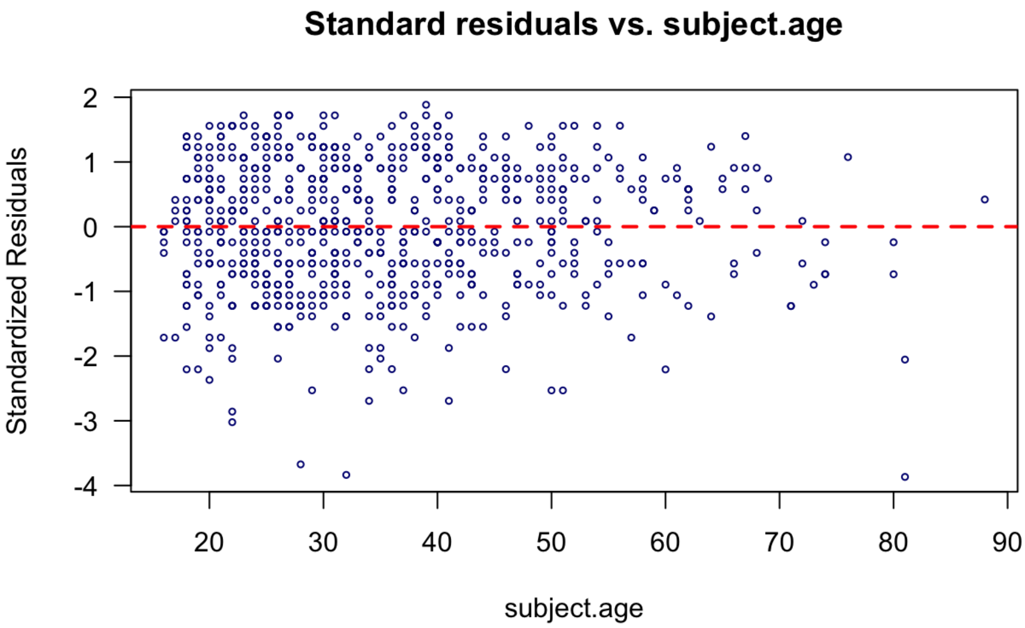
1b: The least squares estimate of α is 2007, with 95% confidence interval [2006.173, 2008.796]. The least squares estimate of β is -0.000259, with 95% confidence interval [-0.0349, 0.0344].

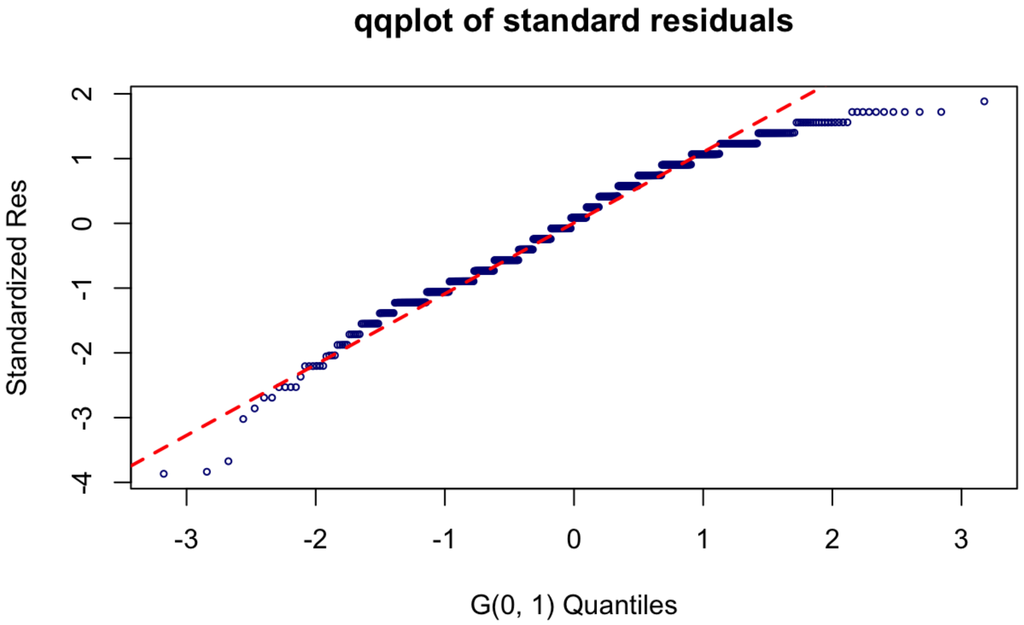
1c: In the context of this study, α represents the mean value of the `vehicle.year` in the study population of individuals for whom the `subject.age` is zero.

1d: Scatterplot:



1e: Residual and Q-Q Plots:
(See Next Page)





1f: The linear model assumes that the Gaussian model is a good fit for the response variates with standard deviation σ which does not depend on the covariates, and that the expected value of the response variate `vehicle.year` is a linear function of the explanatory variate `subject.age`. If these hold, we would expect to see the residuals to follow a Gaussian distribution, with the residual plot to show a random scatter of points and the qqplot of standard residuals to resemble a straight line. For my sample, we observe that the residual plot is very similar to the original scatterplot, implying nonlinearity since residuals reflect the deviations from the linear model. Additionally, the qqplot is slightly concave down and implying asymmetry and negative skew, not Gaussian. Overall, the linear model does not seem suitable for my sample.

1g: An estimate of the mean value of `vehicle.year` for all stops in the study population of people aged 30 years is 2007.477, with 95% confidence interval [2006.976, 2007.977].

1h: The p-value of a test of $H_0 : \beta = 0$ is 0.988. This was calculated using t-distribution.

1i: Based on the results of Analysis 1h, I conclude that the changes in the `subject.age` doesn't have much of an impact on the `vehicle.age` of the traffic stops in San Antonio (no evidence of a linear relationship), agreeing with the proposed null hypothesis. The p-value is quite large, so it implies that there's little to no evidence that the null hypothesis ($\beta = 0$) is false.

Analysis 2

2a: My ID number is 21058539. I will analyze the data for San Antonio and for Analyses 2d-2h I will analyze `subject.race`.

2b: The observed value of the test statistic is $\lambda(\theta_0) = -2\log\left(\frac{L(\theta_0)}{L(\hat{\theta})}\right) = 0.968$, and the resulting p-value is $P(W \geq \lambda(\theta_0)), \quad W \sim \chi^2_1, = 0.325$. The p-value was calculated using the chi-squared distribution.

2c: Based on the results from Analysis 2b, I conclude that we have no evidence against the null hypothesis $\theta_0 = 34.1$ based on the observed data ($P > 0.1$). A limitation of my analysis is that we assumed an exponential model, if the chosen model does not reasonably fit the data then the test can be inaccurate.

2d: The sample mode for `subject.race` is Hispanic.

2e:

Sample Statistic	Hispanic	Non-Hispanic
Size	335	336
Mean	3.51	3.48
Median	3.50	3.43
Standard deviation	0.358	0.366

2f: To test $H_0 : \mu_0 = \mu_1$ we use an unpaired test since the data points of being Hispanic and Non-Hispanic are independent (not measured from the same subjects, so the sample size for Hispanic and non-Hispanic is not the same). The observed value of the test statistic is calculated by letting log transformed ages of Hispanics in San Antonio follow $Y_{1i} \sim G(\mu_1, \sigma_1)$, and non-Hispanics follow $Y_{2i} \sim G(\mu_2, \sigma_2)$, then $d = \frac{|y_1 - y_2 - 0|}{s_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}$, where y_1 represents the sample mean of log-transformed age for subjects that are Hispanic in San Antonio, y_2 represents the sample mean of log-transformed age for subjects that are not Hispanic in San Antonio, n_1 is the sample size of subjects that are Hispanic in San Antonio, and n_2 is the sample size of subjects that are not Hispanic in San Antonio.
In particular $s_p = \sqrt{\frac{(n_1-1)s_1^2+(n_2-1)s_2^2}{n_1+n_2-2}}$, which is the pooled estimator for standard deviation, with s_1 being the sample standard deviation of log-transformed age for subjects that are Hispanic in San Antonio, s_2 represents the sample standard deviation of log-transformed age for subjects that are not Hispanic in San Antonio. The value of the test statistic for my sample is 1.24.
To calculate the p-value we use the t-distribution, where $p = 2[1 - P(T \leq d)], \quad T \sim t_{n_1+n_2-2}$, and the resulting p-value is 0.215.

2g: One assumption upon which the results in Analysis 2f rely is that both the data for Hispanic and Non-Hispanic log transformed ages followed a Gaussian model with independence: $Y_{1i} \sim G(\mu_1, \sigma_1)$ for i = 1,...,335 independently, $Y_{2i} \sim G(\mu_2, \sigma_2)$ for i = 1,...,336 independently. I do think this assumption is reasonable for my sample, because the age's of subjects should not be affected by another in either cases, so they are independent. Additionally we can see the mean and median for both cases are similar, following what is expected of a Gaussian distribution.

2h: Based on the results of my analysis, I conclude there is little to no evidence against the null hypothesis ($p > 0.1$) so there's no significant difference between the mean log-transformed ages of Hispanics and non-Hispanics in San Antonio.

Analysis 3

a. My ID is 21058539, I chose San Antonio for Analysis 1 and Analysis 2. I would really like to score 85+ on this assignment!

b. I guess we can technically make a prediction, but it's not a realistic age, so it's not in our dataset, meaning we would need extrapolate. This leads to problems of: model assumptions may no longer hold, and we have no way to check them, and our predictions may not make sense. Nonetheless, I used `predict()` in R to obtain these values: predicted value is 2007.381, 95% prediction intervals is [1989.94 2024.822], but I would advise against treating it as a proper prediction.

c. Well, `subject.age.log` is just the log transformed age of subjects, so naturally there would be a relationship since they come from the same variate. When you perform a linear regression with `subject.age` as the explanatory variate for `subject.age.log`, you are modelling how age on a logarithmic scale changes as age itself changes. This analysis could be useful if we want to see a general relationship of log transformed data, but otherwise for prediction purposes, it's not really useful.

d. What you're suggesting is something called **p-hacking**, manipulating data (omitting stops and changing values) to obtain a small p-value (usually to be below a certain threshold). This is unethical and unhelpful to our analysis, so I would advise against it, but if we're feeling devious, omitting stops that agree with the null hypothesis $H_0 : \mu_0 = \mu_1$ would likely decrease the p-value.

Appendix: R Code

Include your R code here.

```
# setting up
mydata <- read.csv("~/Desktop/School/Stat 231/stat231f24dataset21058539.csv")
dim(mydata)
colnames(mydata)
sum(is.na(mydata))

# beginning of analysis
SanAntonio <- subset(mydata, city=="sa")
NewOrleans <- subset(mydata, city=="no")

# 1b.
mod <- lm(vehicle.year ~ subject.age, SanAntonio)
mod
summary(mod)
confint(mod, level = 0.95)

# 1d.
plot(SanAntonio$subject.age, SanAntonio$vehicle.year, xlab =
      "subject.age", ylab = "vehicle.year",
      main = "Scatterplot of subject.age and subject.year in San Antonio",
      pch = 1, cex = 0.5, col = "navy",las = 1, cex.axis = 0.5)
abline(coef(mod), lwd = 2, lty = 2, col = "red")

# 1e.
stdres <- rstandard(mod)
mean(stdres)
sd(stdres)
# scatterplot
plot(SanAntonio$subject.age, stdres, main = "Standard residuals vs. subject.age",
     xlab = "subject.age", ylab = "Standardized Residuals",
     pch = 1, col = "navy", cex = 0.5, las = 1)
abline(h = 0, lty = 2, col = "red", lwd = 2)
# qq plot
qqnorm(stdres, main = "qqplot of standard residuals", xlab = "G(0, 1) Quantiles",
  ylab = "Standardized Res", pch = 1, col = "navy", cex = 0.5)
qqline(stdres, lty = 2, col = "red", lwd = 2)

# 1g.
predict(mod, newdata = data.frame(subject.age = 30),
    interval = "confidence", level = 0.95)

# 1h.
coef(summary(mod))
```

```
# 2b.
ExpRLF <- function(theta, n, thetahat) {
  exp(n * log(thetahat/theta) + n * (1 - thetahat/theta))
}
n <- length(log(SanAntonio$subject.age))
thetahat <- mean(SanAntonio$subject.age)
n
thetahat
teststat <- -2 * log(ExpRLF(34.1, n, thetahat))
teststat
p <- 1 - pchisq(teststat, df = 1)
p


# 2d.
mydata$subject.age.log <- log(mydata$subject.age)
table(SanAntonio$vehicle.colour)
table(SanAntonio$vehicle.make)
table(SanAntonio$subject.race)

SanAntonio$race.binary <- 1 - as.numeric(SanAntonio$subject.race == "hispanic")
table(SanAntonio$race.binary)
his <- subset(mydata, SanAntonio$race.binary == 0)
nhis <- subset(mydata, SanAntonio$race.binary == 1)
summary(his$subject.age.log)
sd(his$subject.age.log)
summary(nhis$subject.age.log)
sd(nhis$subject.age.log)

# 2e.
test.results <- t.test(his$subject.age.log, nhis$subject.age.log, var.equal = TRUE)
test.results
test.results$statistic
test.results$p.value

# 3b.
predict(mod, newdata = data.frame(subject.age = 400),
    interval = "prediction", level = 0.95)

# 3c.
summary(lm(subject.age.log ~ subject.age, data = mydata))
```