```
library(MASS)
library(car)
library(rgl)
library(glmnet) # to use LASSO
library(nlme)

data <- read.table("asthma.txt")
# if our urban been numerical variabel, 0/1, then R will use it as numeric, not factor
# but we expected to use factor
lm <- lm(asthma ~ urban + age + pollution + sunny + tobacco + income + education, data
= data)
summary(lm)

par(mfrow=c(2,2)); plot(lm)

data_new  <- data[-92, ] # as example how to remove outlier (if it will be outlier)

# ASSUMPTIONS -----------------------------------------------------------------
# Assumptions required for estimation:
# - Residuals have zero mean and are homoscedastic:
# The residuals vs fitted diagnostic plot shows that the residuals are evenly
distributed on both sides of the zero line and show no specific
# mean or variance pattern against the fitted values
# Assumptions required for inference:
# - Residuals are normally distributed: The residuals Q-Q plot shows that the empirical
quantiles
# are close to normal quantiles, except for the left tail which seems a bit heavy

# to check this assumtpion we can run shapiro test (to check normality),
# and look on graphics of model
shapiro.test(lm$residuals)

# Can we affirm at 90% confidence level that the age has a positive effect on asthma
prevalence?
# ANSWER:
# The estimate of the coefficient associated to age is positive and the p-value of the
significance test is below 0.1 so we
# can affirm that the age has a positive effect at the 90%
# confidence level

# Additionally,
# provide an 95% confidence interval for the mean difference between the asthma
prevalence in an urban province
# and in a non-urban one

summary(lm)
confint(lm, parm="urbanYes", level = 0.95)

# After having reduced the model M0, if appropriate
# Reduce model step by step, removing by one unsignificant feature (with highest p-
value)

lm2 <- lm(asthma ~ urban + age + pollution + sunny + tobacco + income, data = data)
summary(lm2) # removed education

lm3 <- lm(asthma ~ urban + pollution + sunny + tobacco + income, data = data)
summary(lm3) # removed age
# now everything significant


# Update it by introducing a compound-Symmetry Correlation Structure using
# the region as a grouping factor (model M1).
fitS <- gls(asthma ~ urban + pollution + sunny + income + tobacco,
            correlation = corCompSymm(form = ~1| region_id), data = data)

summary(fitS)

plot(fitS)
```

```
# Provide a 99% confidence interval for the
# parameters ρ and σ of the compound symmetry.

intervals(fitS, which = "var-cov", level = 0.99)

# From the possibly reduced version of the model M0, update it now by introducing a
random intercept related
# to the regional grouping factor (model M2). What do you observe? Provide the estimate
of the standard
# deviation of the random intercept along with the one of the error term.
library(lme4)

# Fit the mixed effects model
fitM2 <- lme(asthma ~ urban + pollution + sunny + income + tobacco,
            random = ~1|region_id,
            data = data)

summary(fitM2) # we don't see any signifficant difference between lsat two models

# Random effects:
#  Formula: ~1 | region_id
#           (Intercept) Residual
# StdDev:    17.48762 3.426692
# 17.48762 - estimate of the standard deviation of the random intercept
# 3.426692 - standart deviation one of the error term
```