Analysis of Ecosystems homework week 6

Multiple regression and ANOVA

The solution

03 March 2019

```
pacman::p_load(s20x, pander, plyr, tidyverse, gridExtra, multcomp)
```

Data preparation

Identification

```
mpg <- mpg %>%
  mutate(trans = ifelse((substring(trans,1,4)=='auto'),'auto', 'manual'),
         drv = case_when(
          drv == "f" ~ "front-wheel",
          drv == "4" ~ "four-wheel",
          TRUE ~ "rear-wheel"    )) %>%
  mutate(drv = as.factor(drv)  )
str(mpg)
```

```
## Classes 'tbl_df', 'tbl' and 'data.frame':    234 obs. of  11 variables:
##  $ manufacturer: chr  "audi" "audi" "audi" "audi" ...
##  $ model       : chr  "a4" "a4" "a4" "a4" ...
##  $ displ       : num  1.8 1.8 2 2 2.8 2.8 3.1 1.8 1.8 2 ...
##  $ year        : int  1999 1999 2008 2008 1999 1999 2008 1999 1999 2008 ...
##  $ cyl         : int  4 4 4 4 6 6 6 4 4 4 ...
##  $ trans       : chr  "auto" "manual" "manual" "auto" ...
##  $ drv         : Factor w/ 3 levels "four-wheel","front-wheel",..: 2 2 2 2 2 2 2 1 1 1 ...
##  $ cty         : int  18 21 20 21 16 18 18 18 16 20 ...
##  $ hwy         : int  29 29 31 30 26 26 27 26 25 28 ...
##  $ fl          : chr  "p" "p" "p" "p" ...
##  $ class       : chr  "compact" "compact" "compact" "compact" ...
```

`cty` will be the response variable in all tests. Multiple regression will test `cty` against `displ` as a continuous variable and `trans` as a categorical variable. ANOVA will test `cty` against `drv`.

Assumptions

Before fitting a model, one needs to ensure that the distribution of the response variable fits a normal distribution:

```
# Distribution plots
  raw.d.gg <-
    ggplot(mpg, aes(x=cty)) + theme_bw(14) +
      geom_histogram(aes(y=..density..),
                     binwidth=1,
                     colour="black", fill="lightgreen") +
      geom_density(alpha=0.2, fill="lightgreen") +
      stat_function(data=mpg,
```

```
                          fun = dnorm,
                          args=list(mean=mean(mpg$cty),
                                    sd=sd(mpg$cty)),
                          colour="blue", size=1.1)
  log.d.gg <-
    ggplot(mpg, aes(x=log(cty))) + theme_bw(16) +
      geom_histogram(aes(y=..density..),
                     binwidth=0.1,
                     colour="black", fill="lightgreen") +
      geom_density(alpha=0.2, fill="lightgreen") +
      stat_function(data=mpg,
                          fun = dnorm,
                          args=list(mean=mean(log(mpg$cty)),
                                    sd=sd(log(mpg$cty))),
                          colour="blue", size=1.1)
# Q-Q plots
  raw.QQ.gg <-
         ggplot(mpg, aes(sample=cty)) + theme_bw(14) +
            stat_qq(size=4,  bg="#43a2ca", col="black", pch=21) +
            stat_qq_line(size=1.5, color="blue")
  log.QQ.gg <-
         ggplot(mpg, aes(sample=log(cty))) + theme_bw(14) +
            stat_qq(size=4, bg="#43a2ca", col="black", pch=21) +
            stat_qq_line(size=1.5, color="blue")

grid.arrange(raw.d.gg,  log.d.gg,
             raw.QQ.gg, log.QQ.gg, nrow = 2)
```

The untransformed data are skewed right (Fig. 1), but log transformation improves the fit between the distribution of the cty variable and the theoretical normal distribution. QQ plot of the log transformed data confirm the better fit.

Multiple linear regression

Graphing

```
prov.col2 <- viridis::magma(n=2, begin=0.25, end=0.75, direction = 1)
ggplot(mpg, aes(x=displ, y=log(cty))) + theme_bw(20) +
  geom_smooth(aes(color=trans),
              se=FALSE, method="lm",
              size=1.5) +
  geom_point(aes(fill=trans, shape=trans),
             size=4,  col="black") +
  scale_shape_manual(name="Transmission\ntype",
                     values = c(21, 24)) +
  scale_fill_manual(name="Transmission\ntype",
                     values = prov.col2) +
  scale_color_manual(name="Transmission\ntype",
                     values = prov.col2) +
  labs(x="Engine size (L)",
       y="City fuel economy\n(log mpg)")
```

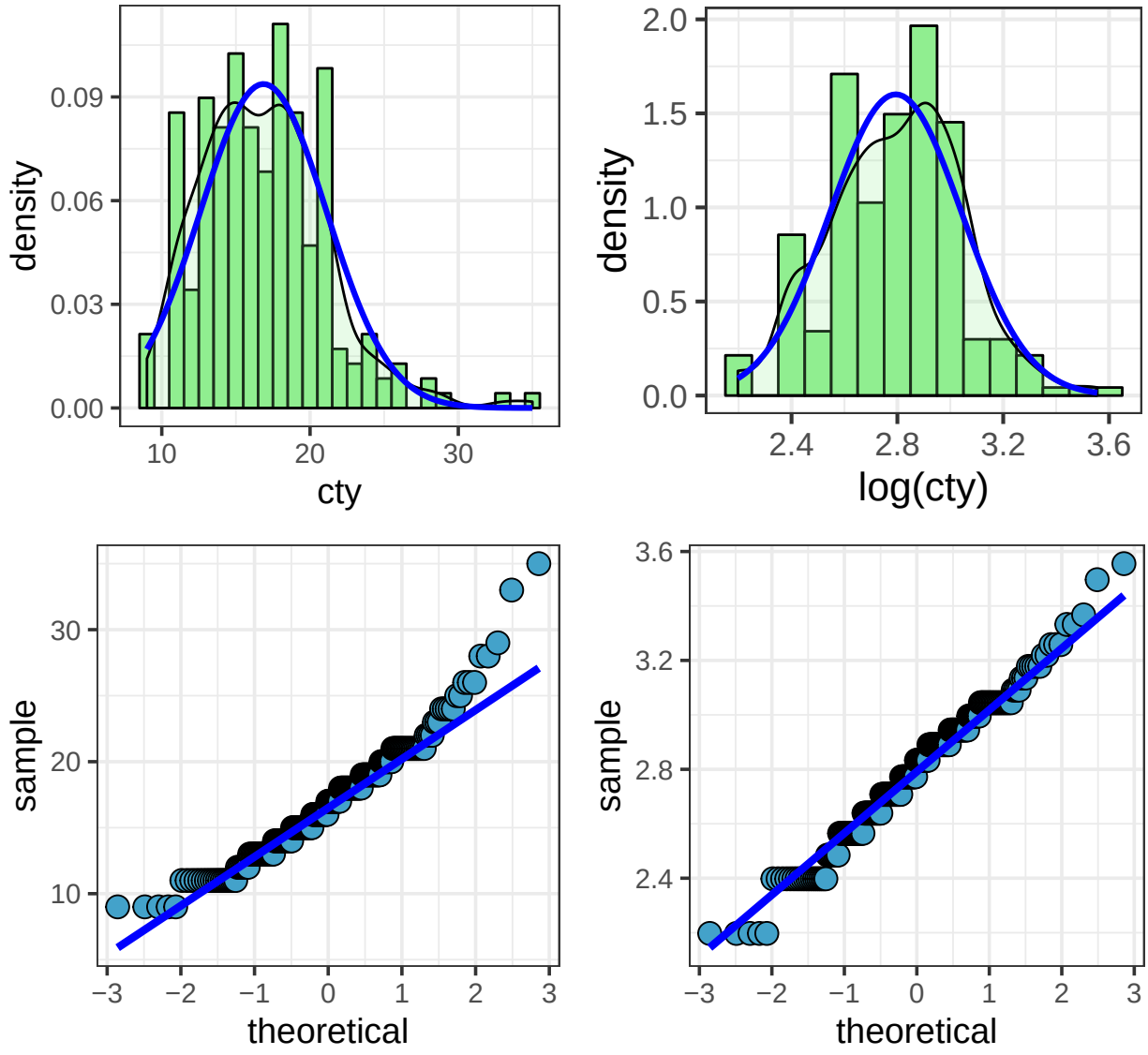

Figure 1: TOP: Distribution of the `cty` variable, before log transformation (L) and after log transformation (R). BOTTOM: Q-Q plots for the above.

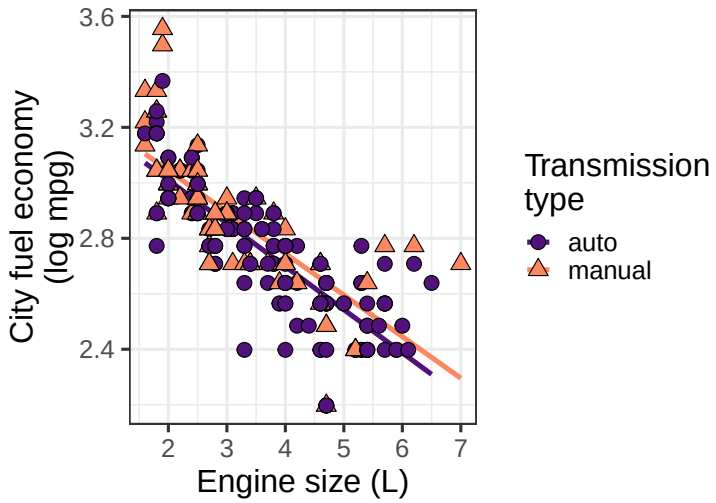

Figure 2: The linear relationship between city fuel economy (`cty`) against engine size (`displ`) and `trans`mission type. `cty` log transformed.

Hypothesis statement

H_0: No relationship between city fuel economy and transmission type or engine size.

H_1: City fuel economy declines with engine size; manual transmissions generally get better fuel economy at a given engine size.

Fit a linear model

```
m1 <- lm(log(cty) ~ displ + trans, mpg)
car::Anova(m1, type="2")
```

```
## Anova Table (Type II tests)
##
## Response: log(cty)
##           Sum Sq  Df  F value  Pr(>F)
## displ     8.5101   1 421.0206 < 2e-16 ***
## trans     0.0837   1   4.1399 0.04303 *
## Residuals 4.6692 231
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
summary(m1)
```

```
##
## Call:
## lm(formula = log(cty) ~ displ + trans, data = mpg)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.43595 -0.08326 -0.00414  0.07358  0.49142
##
## Coefficients:
```

```
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  3.314368   0.030109 110.080   <2e-16 ***
## displ       -0.153840   0.007498 -20.519   <2e-16 ***
## transmanual  0.041856   0.020571   2.035    0.043 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1422 on 231 degrees of freedom
## Multiple R-squared:  0.6772, Adjusted R-squared:  0.6744
## F-statistic: 242.3 on 2 and 231 DF,  p-value: < 2.2e-16
```

```
gvlma::gvlma(m1)
```

```
##
## Call:
## lm(formula = log(cty) ~ displ + trans, data = mpg)
##
## Coefficients:
## (Intercept)        displ  transmanual
##     3.31437     -0.15384      0.04186
##
##
## ASSESSMENT OF THE LINEAR MODEL ASSUMPTIONS
## USING THE GLOBAL TEST ON 4 DEGREES-OF-FREEDOM:
## Level of Significance =  0.05
##
## Call:
##  gvlma::gvlma(x = m1)
##
##                      Value   p-value                   Decision
## Global Stat        64.6941 2.984e-13 Assumptions NOT satisfied!
## Skewness            0.3202 5.715e-01    Assumptions acceptable.
## Kurtosis           25.9326 3.535e-07 Assumptions NOT satisfied!
## Link Function      35.3841 2.707e-09 Assumptions NOT satisfied!
## Heteroscedasticity  3.0572 8.038e-02    Assumptions acceptable.
```

Interpretation

The model fit these data well, with approximately 68% of variance explained. Results of `gvlma` analysis of the multiple regression model show that assumptions for both skewness and heteroscedasticity were met, indicating that the model was robust. Thus, we can reject that null hypothesis, and have statistical evidence for both parts of the alternative hypothesis: both terms were significant, and the *t* statistics for `displ` and manual `transmissions` were -20 and 2, respectively.

ANOVA on categorical predictor variables

Graphing

```
prov.col3 <- viridis::magma(n=3, begin=0.25, end=0.75, direction = 1)

ggplot(mpg, aes(x=drv, y=cty)) + theme_bw(20) +
```

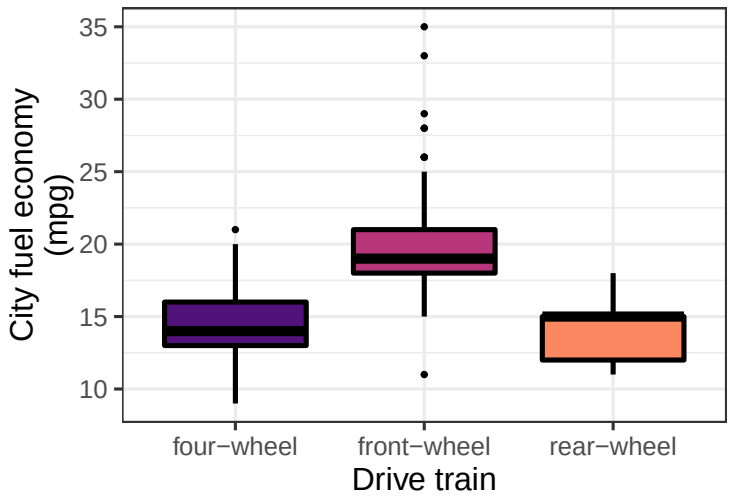

Figure 3: The relationship between city fuel economy (`cty`) and drive train type. `cty` log transformed in analysis, but not in the graph.

```
  geom_boxplot(aes(fill=drv),
               size=1.5,  col="black") +
  scale_fill_manual(values = prov.col3, guide=FALSE) +
  labs(x="Drive train",
       y="City fuel economy\n(mpg)")
```

Hypothesis statement

H_0: No relationship between city fuel economy and drive train.

H_1: Relative to front wheel drive vehicles, city fuel economy will be lower for four wheel drive vehicles and lowest for rear wheel drive vehicles.

Fit an ANOVA model

```
m2 <- lm(log(cty) ~ drv, mpg)
anova(m2)
```

```
## Analysis of Variance Table
##
## Response: log(cty)
##            Df Sum Sq Mean Sq F value    Pr(>F)
## drv         2 6.6599  3.3300  98.549 < 2.2e-16 ***
## Residuals 231 7.8055  0.0338
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
m2.tuk <- glht(m2, linfct = mcp(drv="Tukey"))
summary(m2.tuk)
```

```
##
##   Simultaneous Tests for General Linear Hypotheses
```

```
##
## Multiple Comparisons of Means: Tukey Contrasts
##
##
## Fit: lm(formula = log(cty) ~ drv, data = mpg)
##
## Linear Hypotheses:
##                                Estimate Std. Error t value Pr(>|t|)
## front-wheel - four-wheel == 0  0.336928   0.025433  13.248   <1e-05 ***
## rear-wheel - four-wheel == 0  -0.009895   0.040984  -0.241    0.968
## rear-wheel - front-wheel == 0 -0.346823   0.040870  -8.486   <1e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## (Adjusted p values reported -- single-step method)
```

Interpretation

As expected, there was significant variability in city fuel economy among the drive train types (Fig. 3): front wheel drive vehicles were consistently greater than both rear wheel drive ($t = 13.2$, $P < 0.001$) and four wheel drive vehicles ($t = 8.5$, $P < 0.001$). However, there was no statistically-significant difference between rear wheel drive and four wheel drive vehicles ($t =$ -0.24, $P = 0.97$).