Analysis of Ecosystems homework week 5

Intro to statistical inference

The solution

02 March 2019

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

Data preparation

```
## Classes 'tbl_df', 'tbl' and 'data.frame':    234 obs. of  12 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         : chr  "f" "f" "f" "f" ...
##  $ 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" ...
##  $ origin      : Factor w/ 2 levels "Foreign","USA": 1 1 1 1 1 1 1 1 1 1 ...
```

The t test

Assumptions

- Independent samples
- Normal distribution
- Equal variance within groups

Distribution

Graphing

```
# 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)
```

Interpretation

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.

Variance

Test

One option is the base `var.test`:

```
vt <- var.test(log(cty) ~ origin, mpg, ratio=1)
pander(vt)
```

Table 1: F test to compare two variances: `log(cty)` by `origin` (continued below)

Test statistic	num df	denom df	P value	Alternative hypothesis
1.201	132	100	0.336	two.sided

ratio of variances
1.201

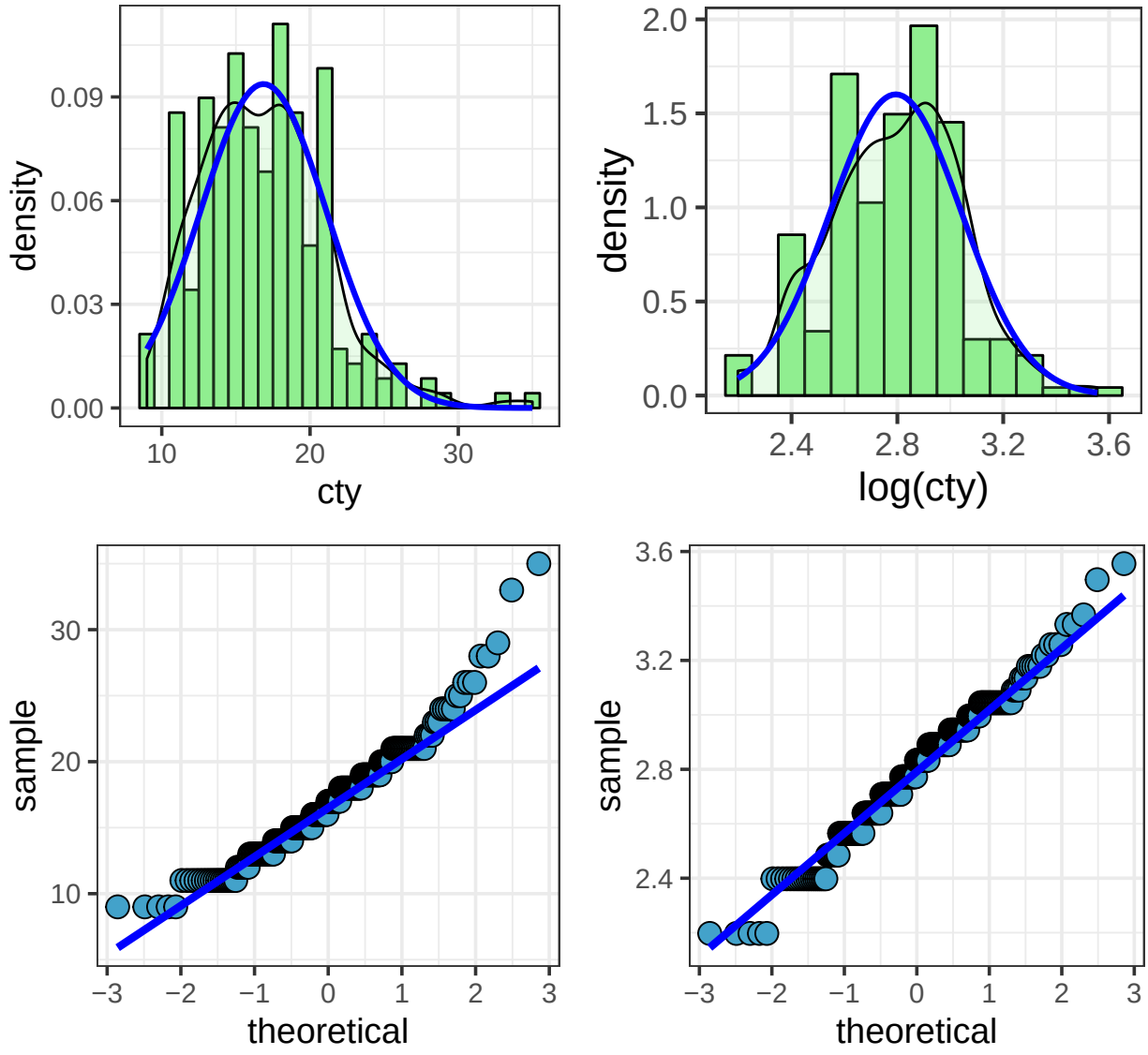

Figure 1: TOP: Distribution of the `cty` variable, before log transformation (L) and after log transformation (R). BOTTOM: Q-Q plots for the above.

Another option is a Bartlett's test using the \texttt{ols_test_bartlett} function in the `olsrr` package:

```
mpg$Lcty <- log(mpg$cty)
olsrr::ols_test_bartlett(mpg, Lcty, group_var = origin)
```

```
##
##     Bartlett's Test of Homogenity of Variances
## ------------------------------------------------
## Ho: Variances are equal across groups
## Ha: Variances are unequal for atleast two groups
##
##         Test Summary
##  ----------------------------
##  DF            =    1
##  Chi2          =    0.9424139
##  Prob > Chi2   =    0.3316578
```

Interpretation

The variance ratio of the log transformed `cty` variable among the two `origin` groups is 1.2, and not statistically different than 1. Thus these data meet the assumption of equal variance.

The linear model

Assumptions

- Linear relationship between variables
- Independent samples
- Normal distribution
- *Homoscedasticity*–Homogeneous variance along regression gradient

Distribution

Graphing

Figure 1 already shows us that the `cty` variable meets the assumptions of normality. To test the unique assumption of the linear model—equal variance along the regression gradient—we can look at the residuals of the regression model. This requires us to fit the model to assess the assumption. In a sense, then, homoscedasticity is more of an assumption of the *model results* than of the data themselves, although obviously the model results are unique to the data.

```
m1 <- lm(log(cty) ~ displ, mpg)
ggplot(m1) + theme_bw(24) +
  geom_hline(yintercept = 0, size=1.5) +
  geom_smooth(aes(x=.fitted, y=.resid),
              color="red", se=FALSE) +
  geom_smooth(aes(x=.fitted, y=.resid), se=FALSE,
              method="lm", color="lightblue", lty=2 ) +
  geom_point(aes(x=.fitted, y=.resid), size=4,
             bg="#43a2ca", col="black", pch=21)
```

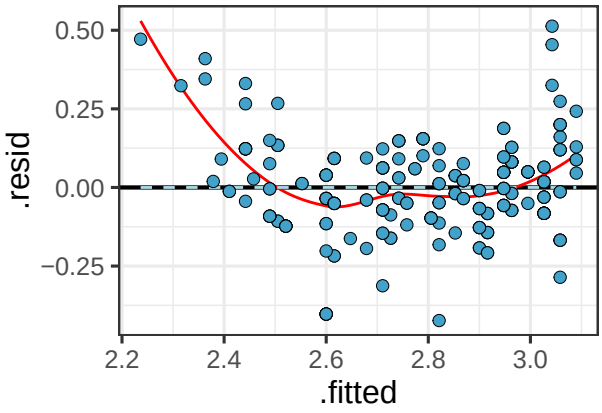

Figure 2: Residual plot for the linear model fitting `cty` against `displ`.

Statistical tests

Several statistical procedures test the alternative hypothesis that variance is not constant. The `car` package gives us a test for non-constant variance, typically known also as a *Breusch-Pagan* test, although the `car` formulation returns a χ^2 test statistic:

```
car::ncvTest(m1)
```

```
## Non-constant Variance Score Test
## Variance formula: ~ fitted.values
## Chisquare = 13.437, Df = 1, p = 0.00024671
```

The `lmtest` package also provides a function to perform a Breusch-Pagan test, `bptest` :

```
lmtest::bptest(m1)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  m1
## BP = 6.7485, df = 1, p-value = 0.009383
```

Finally, package `olsrr` includes several tests for heteroscedasticity; here we run an F test:

```
olsrr::ols_test_f(m1)
```

```
##
##  F Test for Heteroskedasticity
##  -----------------------------
##  Ho: Variance is homogenous
##  Ha: Variance is not homogenous
##
##  Variables: fitted values of log(cty)
##
##        Test Summary
##  --------------------------
##  Num DF     =    1
##  Den DF     =    232
##  F          =    6.889512
##  Prob > F   =    0.009246334
```

These data fail each test. What can be done about non-constant variance, or heteroscedasticity? The conventional approach is a *Box-Cox transformation*:

```
caret::BoxCoxTrans(mpg$Lcty)
```

```
## Box-Cox Transformation
##
## 234 data points used to estimate Lambda
##
## Input data summary:
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   2.197   2.639   2.833   2.794   2.944   3.555
##
## Largest/Smallest: 1.62
## Sample Skewness: -0.0245
##
## Estimated Lambda: 1.1
## With fudge factor, no transformation is applied
```

The Box-Cox transformation actually suggested we are within the "fudge factor" and should not perform the transformation. So we won't. (In my view this is the best approach. I like a procedure that doesn't feel like it has to do something, and clearly lets us know when we're within the margins of acceptability and don't need to change anything. I don't like to depend on a P value.)

Interpretation

The previous log transformation took care of our normality assumption. The residual plot (Fig. 2) is ok but not great—a linear regression fit to the residuals has $\beta_0 = 0$, which is good (blue broken line in Fig. 2), but there is some curvilinearity at the lowest end of the plot (red line in Fig. 2).

Fit a linear model

Test

```
anova(m1) %>%
  pander(.)
```

Table 3: Analysis of Variance Table

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
displ	1	9.712	9.712	474.1	5.411e-58
Residuals	232	4.753	0.02049	NA	NA

Plot

```
ggplot(mpg, aes(x=displ, y=log(cty))) + theme_bw(20) +
  geom_smooth(se=FALSE, method="lm",
              size=1.5, color="blue") +
  geom_point(size=4,  pch=21,
             bg="#43a2ca", col="black") +
  labs(x="Engine size (L)",
```

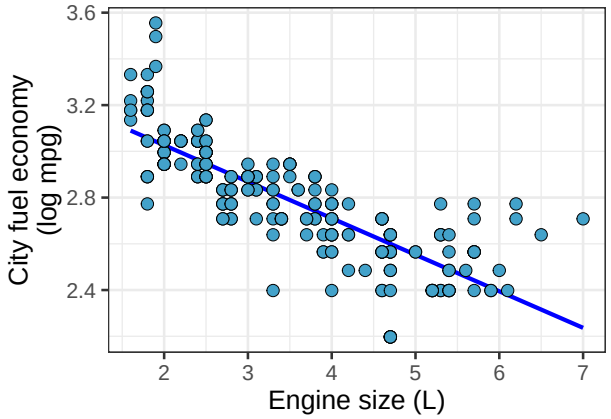

Figure 3: The linear relationship between `cty` and `displ`, the former log transformed.

```
       y="City fuel economy\n(log mpg)")
```

Evaluate the linear model

Model fit

```
summary(m1)
```

```
##
## Call:
## lm(formula = log(cty) ~ displ, data = mpg)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.42329 -0.08219 -0.00317  0.08111  0.51292
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  3.342686   0.026879  124.36   <2e-16 ***
## displ       -0.158030   0.007258  -21.77   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1431 on 232 degrees of freedom
## Multiple R-squared:  0.6714, Adjusted R-squared:   0.67
## F-statistic: 474.1 on 1 and 232 DF,  p-value: < 2.2e-16
```

The model fits the data well, explaining about 71% of variation between the variables as determined by R^2=0.71.

Model results

- Degrees of freedom: 232
- Total Sum of Squares: 14.5
- Test statistic for the overall model: $F_{1/232}$=573.8
- Test statistic for the dependent variable: t = -21.8

- Results of significance test: $P < 0.001$

Interpretation

There is a strong negative linear relationship between `cty` and `displ`, indicating that in general, as engine size increases, fuel economy declines.