OLS in R Tutorial

Robert Schnitman

1. Purpose

This is a tutorial on how to estimate and interpret OLS regressions in R.

2. Set Up

First, let's load the *mtcars* dataset.

```
data(mtcars) # Load data.
```

3. Model Results

Next, we'll set up a model with *lm()* and estimate it with *summary()*.

In *lm()*, we need two basic inputs: the formula and data. The formula is based on the format of `y ~ x1 + x2 + ...`, where *y* is your dependent variable and the *x* terms are the covariates. The function *summary()* is general-purpose, meaning we can apply it to any object. The output of *summary()* differs depending on the input. For example, if we executed *summary(mtcars)*, we will obtain summary statistics for each of the variables. In contrast, applying the function on a model will produce an ANOVA table, coefficient table, and model fit statistics based on what we specified in *lm()*.

For this model, we want to analyze how an automobile's weight (*wt*) and horsepower (*hp*) influence its miles per gallon (*mpg*).

```
mymodel <- lm(formula = mpg ~ wt + hp, data = mtcars) # Save model.
summary(mymodel)                                      # Print model results!
```

```
##
## Call:
## lm(formula = mpg ~ wt + hp, data = mtcars)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -3.941 -1.600 -0.182  1.050  5.854
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 37.22727    1.59879  23.285  < 2e-16 ***
## wt          -3.87783    0.63273  -6.129 1.12e-06 ***
## hp          -0.03177    0.00903  -3.519  0.00145 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.593 on 29 degrees of freedom
## Multiple R-squared:  0.8268, Adjusted R-squared:  0.8148
## F-statistic: 69.21 on 2 and 29 DF,  p-value: 9.109e-12
```

```r
### Notes ###
# Alternatively, we could have typed "summary(lm(mpg ~ wt + hp, mtcars))".
# This format is not preferable when chaining 3 or more functions with multiple inputs.
# Readability is important! You never know when you need to come back to a file!
##################
```

So, a 1-ton increase in a car's weight *decreases mpg* by 3.88. Horsepower also seems to decrease *mpg*: the coefficient is -0.03. If the covariates equal 0, the expected *mpg* is 37.23.

These terms are statistically significant at the 5% level (determined by the p-value column, *Pr(>|t|)*). According to the R-squared of 82.68%, our model as a whole strongly explains changes in *mpg*.

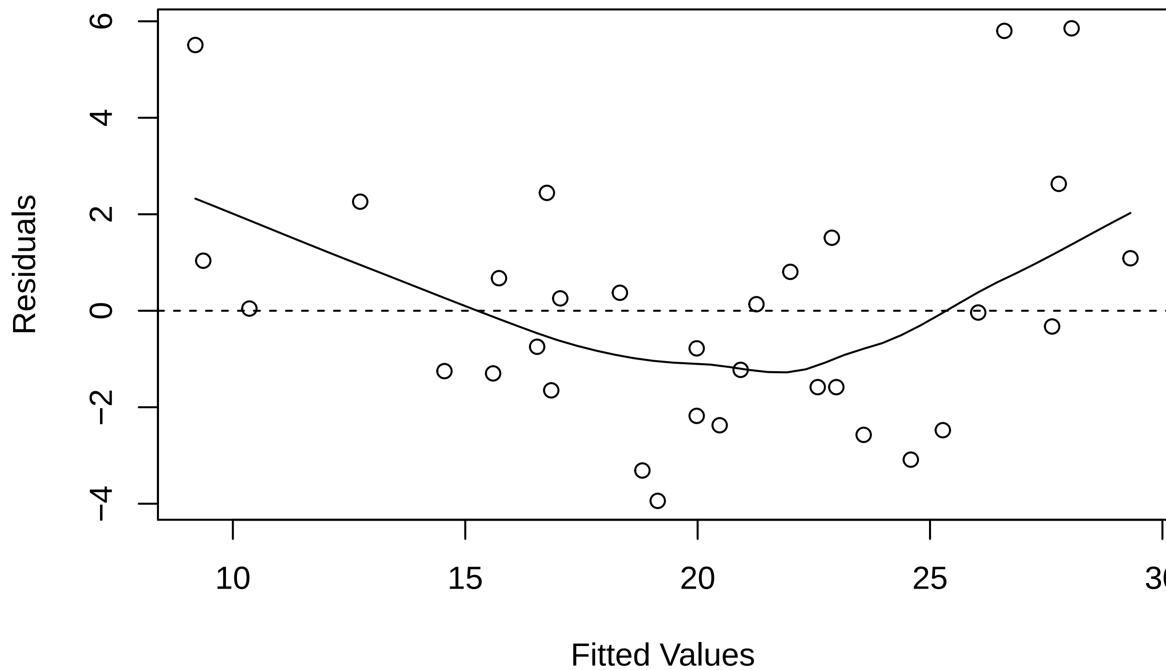
4. Diagnostics

How else can we diagnose our the results of our model? How do we know its biased? We will use *scatter.smooth()* to examine whether the residuals are 0 on average.

```r
fit <- predict(mymodel)                # predict() calculates our fitted values, yhat.
res <- resid(mymodel)                  # resid() computes our residuals, y - yhat.

scatter.smooth(y = res,                # Are the residuals 0 on average?
               x = fit,
               xlab = 'Fitted Values', # X-axis label.
               ylab = 'Residuals')     # Y-axis label.

abline(a = 0, b = 0, lty = 2)          # Set dash line at 0 for comparison.
```

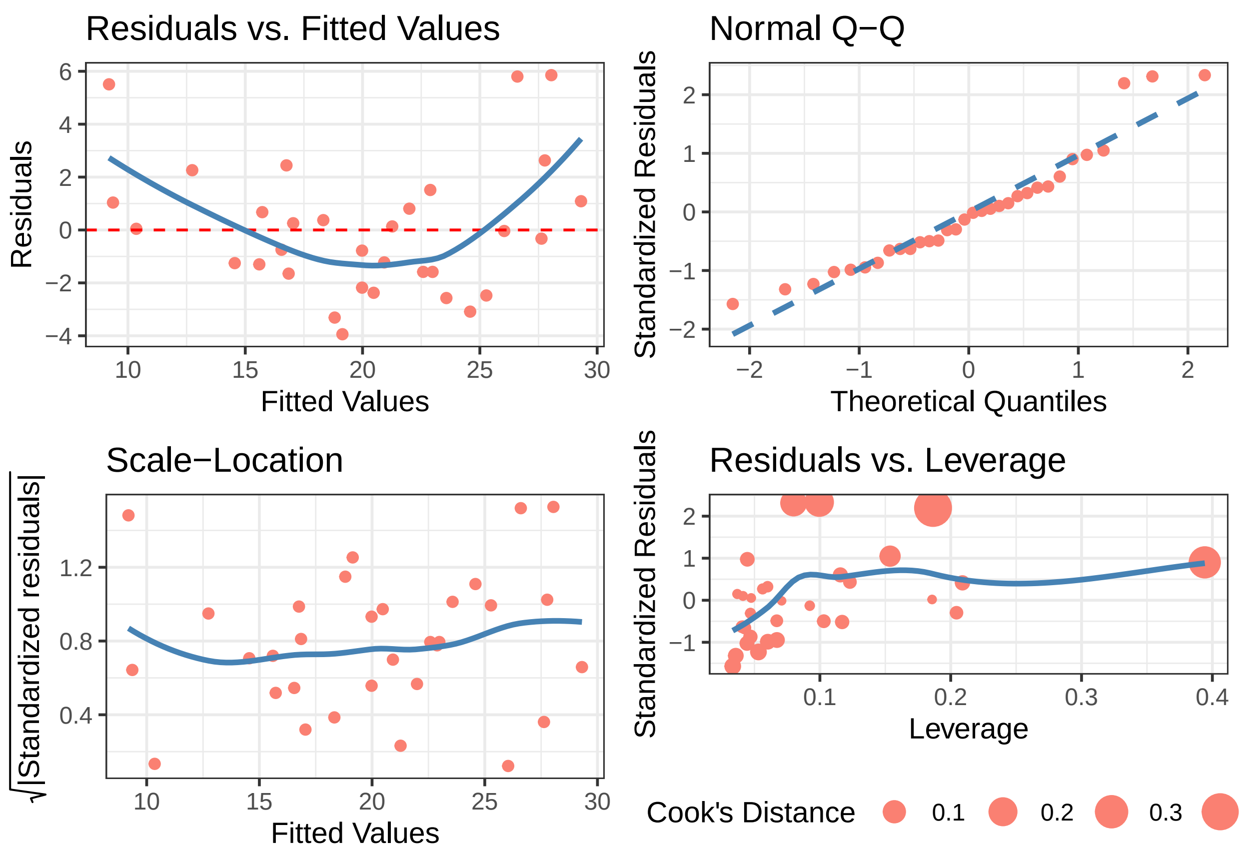


In an ideal situation, the smoothed curve would be flat at the y-intercept of 0. However, our model residuals indicate that we are experiencing some heteroskedasticity–our variance is not constant. As such, while our model is able to explain variations in the dependent variable well overall, we also tend to overestimate *mpg* at certain levels (rememeber that residuals = actual - prediction, so paired values below the 0 line in the above plot indicates overestimation).

What else could we do to achieve an unbiased while maintaining statistical significance obtained in the previous model results?

5. Next Steps

Further lessons will discuss how to improve model performance, such as including more variables and applying logarithmic functions, as well as plotting multiple graphs on a grid.



End of Document