Diagnostics and Transformations

Simple linear regression – Stat 230

In this class, you will work with your group to explore the adequacy of simple linear regression models and how transformations can help remedy some violations of those assumptions. While you work through this activity, make sure that all group members are engaged and contribute ideas, and also follow the code. While the R Manual has some useful R code for today's activities (especially for review), new code is given as needed.

Assessing SLR model assumptions

Task 1. To begin, review the assumptions/conditions necessary for valid inference from simple linear regression models. List the necessary assumptions/conditions and give at least one way to check each.

Task 2. Extended Activity 32 in chapter 2 of your textbook has you examine the fit of a simple linear regression model relating the brain weights (in grams) and body weights (in kilograms) of 30 species of mammal. You can load the data using the code shown below. For the questions, see the original problem in your textbook.

```
weight <- read.csv("https://aloy.rbind.io/kuiper_data/Weights.csv")
```

Residual plots in R

To construct quick residual plots, I recommend using functions from the {car} package, so you will need to add `library(car)` to your setup code chunk. For the sample code below, `fm` refers to a fitted regression model object.

Standardized residuals vs. fitted values

```
residualPlot(fm, quadratic = FALSE, type = "rstandard")
```

Normal Q-Q plot of standardized residuals

```
qqPlot(fm, type = "rstandard", distribution = "norm")
```

By default, the Q-Q plot has an "envelope" added to the plot in an effort to help you assess normality. If you find this distracting, then add the argument `envelope = FALSE` to the `qqPlot()` call. In addition, if you prefer filled points to hollow points, add the argument `pch = 16` to your plotting commands from the {car} package.

An alternative approach to residual plots in R

If you prefer to stay in the {ggformula} plotting universe, then I recommend *augmenting* your data set to include key diagnostic information. To do this, you can use the `augment()` function in the {broom} package. Again, letting `fm` denote our fitted regression object, we can create a new data frame

```r
aug_data <- augment(fm)
```

This new data set will have the columns used to fit your model as well as `.resid` (residuals) and `.std.resid` (standardized residuals), as well as a few more.

Task 3. In Task 2 you selected transformation(s) to help "linearize" the relationship between brain weights and body weights. While this is helpful from a statistical perspective, it's often preferred to plot the fitted model on the **original scale** of the data. To do this, we need to backtransform the model.

- Start by creating a scatterplot on the original scale.
- Then, add a `gf_lm` layer, specifying in the transformations in the `formula` and *if y is transformed* how to backtransform *y* via the `backtrans` argument.

For example, if we log-transformed both `x` and `y`, then we pass `log(y) ~ log(x)` in as the formula to `gf_lm()` and `backtrans = exp` to backtransform `y`. Below is the full call where `df` is the data set with columns `x` and `y`.

```r
gf_point(y ~ x, data = df) |>
  gf_lm(formula = log(y) ~ log(x), backtrans = exp, interval = "confidence")
```

Use this idea to plot the fitted model relating brain weights and body weights on the original scale.

Task 4. Extended Activity 33 in chapter 2 of your textbook provides the `RegrTrans` data set which contains four pairs of variables (`X1`, `Y1`; `X2`, `Y2`; `X3`, `Y3`; `X4`, `Y4`) to help you practice exploring and applying transformations. For each pair of variables, complete parts (a)-(c) given in your textbook.

```r
regr_trans <- read.csv("https://aloy.rbind.io/kuiper_data/RegrTrans.csv")
```

Exploring outliers and influential points

Now, let's consider how outliers and influential points can impact the fitted regression line. To do this, we'll explore a data set containing the bid price and coupon rate of U.S. Treasury bonds.

Here's a little background on Treasury bonds:

> US Treasury bonds are among the least risky investments, in terms of the likelihood of your receiving the promised payments. In addition to the primary market auctions by the Treasury, there is an active secondary market in which all outstanding issues can be traded. You would expect to see an increasing relationship between the coupon of the bond, which indicates the size of its periodic payment (twice a year), and the current selling price. The … data set of coupons and bid prices [are] for US Treasury bonds maturing between 1994 and 1998… The bid prices are listed per 'face value' of $100 to be paid at maturity. Half of the coupon rate is paid every six months. For example, the first one listed pays $3.50 (half of the 7% coupon rate) every six months until maturity, at which time it pays an additional $100. (Siegel, 1997 pp. 384–385)

To begin, load the data.

```
bonds <- read.csv("https://aloy.rbind.io/data/bonds.csv")
```

Task 5. Fit a simple linear regression model to predict the bid price given the coupon rate of U.S. Treasury bonds. Write down the fitted model equation.

Task 6. Create a scatterplot of bid price against the coupon rate and superimpose the fitted regression line. Do you see any potential outliers? If so, identify what rows of the data set correspond to these outliers. (Hint: Click on the name of the data set in the "Environment" tab to open a spreadsheet of the data.)

Task 7. Create a plot of the standardized residuals vs. the fitted values. Do you see any potential outliers? If so, identify what rows of the data set correspond to these outliers. Are they the same rows? (Hint: use the `augment()` function from {broom} here and look at the `.std.resid` column to find potential outliers as flagged by the standardized residuals.)

Task 8. Create an index plot of the leverage values for each observation. You can do this from the augmented data frame, but a quicker way is to use the `infIndexPlot()` function in the {car} package. `infIndexPlot()` requires that you pass in the fitted regression model as well as the influence diagnostic you want to plot. Notice that it calls the leverage `"hat"`. Identify the high leverage cases (the row numbers are printed for "auto detected" high leverage points).

```
infIndexPlot(fm, vars = "hat")
```

Task 9. Compare the cases (row numbers) you identified in Task 6 and Task 7. Which, if any, of these points to you expect to have high values of Cook's distance? Why?

Task 10. Create an index plot of the Cook's distance values for each observation. You can do this from the augmented data frame, but a quicker way is to use the `infIndexPlot()` function in the {car} package by setting `vars = "Cook"`. Are there any influential points identified by Cook's distance? If so, identify these points by their row number.

```
infIndexPlot(fm, vars = "Cook")
```

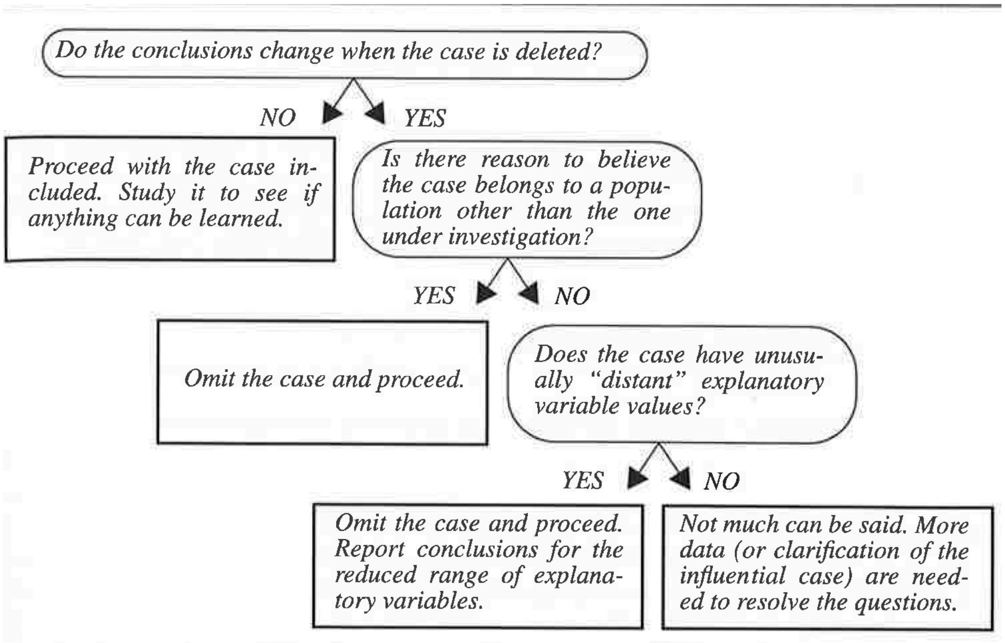
Task 11. Refit the model without the influential points you identified in Task 9. To do this, add a `subset` argument to `lm()` as seen below. How much did the regression coefficients change?

```
# Fill in the blank with row numbers to delete seperated by commas
no_influential <- lm(BidPrice ~ CouponRate, data = bonds, subset = -c(___))
```

Task 12. Are we done? Check the residual plots and recreate the influence index plots you considered in Tasks 7 and 9. Do any additional points cause concern after your have already deleted a few points?

Should we delete the outliers/influential points?

As I discussed in the daily prep videos, it is not always wise to delete outliers and influential points. While this can improve the fit of your model, you might be deleting important information about your population. Of course, if a data entry error was made, then it should be fixed and the analysis should be rerun. In other situations, I recommend following the advice from the flow chart that Ramsey and Schafer include in *The Statistical Sleuth*.



Do the conclusions change when the case is deleted?

NO

Proceed with the case included. Study it to see if anything can be learned.

YES

Is there reason to believe the case belongs to a population other than the one under investigation?

YES

Omit the case and proceed.

NO

Does the case have unusually "distant" explanatory variable values?

YES

Omit the case and proceed. Report conclusions for the reduced range of explanatory variables.

NO

Not much can be said. More data (or clarification of the influential case) are needed to resolve the questions.