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

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 and predictors

```
residualPlots(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

```
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 1. 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.

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

(a) Create a scatterplot of y versus x with a regression line, a plot of the residuals vs. the explanatory variable, a plot of the residuals vs. predicted (or "fitted") values ($\hat{y}$), and either a normal Q-Q plot or a histogram of the residuals.

(b) Try various transformations of the explanatory and response variables to create a better linear regression model. Hint: Notice that both the x and y variables are right skewed and have outliers, both may need a transformation.

Task 2. In Task 1 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 `xvar` and `yvar`.

```
gf_point(yvar ~ xvar, data = df) |> # change variable names and data set name here
  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 3. 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).

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

(a) Create a scatterplot of y versus x with a regression line, a plot of the residuals vs. the explanatory variable, a plot of the residuals vs. predicted (or "fitted") values ($\hat{y}$), and either a normal Q-Q plot or a histogram of the residuals.

(b) Discuss a curve that would better fit the data than the regression line. Notice that the plot of the residuals vs. the explanatory variable emphasizes the patterns in the residuals much better than does the scatterplot of y vs. x.

(c) Try various transformations of the explanatory and response variables to create a better linear regression model. Hint: Notice that both the x and y variables are right skewed and have outliers, both may need a transformation.