Know your data and how to handle it correctly: statistical assumptions

Daiva & Luke

2015-01-09

Welcome to our Statistical Assumptions workshop

Purpose:

To teach the statistical assumptions of linear regression and show how you test data to see if they satisfy the assumptions. Knowing how to check these assumptions is part of “best practices” in data analysis.

Significance:

It is very important to check that your data satisfies linear regression assumptions. If your data does not meet these criteria, the use of linear regression is inappropriate. Other methods can be used, but…

Caveat (again): We aren't here to teach statistics

Need help with stats? Use these resources!

- U of T Statistical Consulting Services (click here)
- `http://www.stackoverflow.com`
- `http://stats.stackexchange.com`
- Helpful statistical tests flowchart (PDF on GitHub)
- Very helpful webpage on regression diagnostics: `http://www.ats.ucla.edu/stat/sas/webbooks/reg/chapter2/sasreg2.htm` (Note: Goes into much more detail than what is covered in this workshop)

Notes and help during this workshop

Go to this website:

`https://etherpad.mozilla.org/dnsWorkshops`

Linear Regression

- Used to test associations between independent and dependent variables
- Based on a linear relationship: $y = mx + b$

How about: $y = X\beta + \varepsilon$?

- y = dependent variable(s), m = slope, x = independent variable, b = error terms (covariates)

Some Linear Regression Assumptions

- Model is good (i.e. linear relationship between predictors and outcome variable)
- Residuals[1] have a normal distribution
- Residuals are homoscadastic (have equal/constant variance)

[1] Residual (aka the error term) = Observed - expected

Brief aside: assumptions/diagnostics we are not covering in this workshop

- Independence (residuals of one observation are not associated with residuals of another)
- Errors in variables (predictor variables are measured without error)
- Influence (i.e. outliers)
- Collinearity (predictors that are linearly related – affects estimate of regression coefficients)
- Very helpful webpage on regression diagnostics that covers these: `http://www.ats.ucla.edu/stat/sas/webbooks/reg/chapter2/sasreg2.htm`

How to check assumptions

- Model fit: Plot residuals vs. predicted fit (check pattern)
- Distribution of residuals: Normal probability plot
- Variance of residuals: Plot residuals vs. predicted fit (check spread of points)

Model fit

```
proc sgplot data=playing;
    scatter x=weight y=length1;
run;
```

Model fit

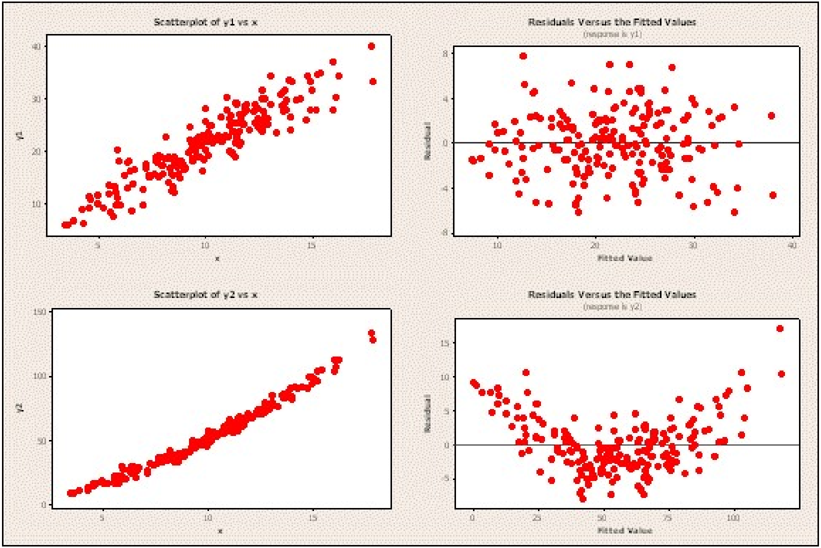

Figure 1

Residual distribution

```
proc reg data=playing;
    model height=weight;
    output out=resid residual=r predicted=fit;
run;
quit;

goptions reset=all;
proc univariate data=resid normal;
    var r;
    qqplot r / normal(mu=est sigma=est);
run;
```

Residual distribution

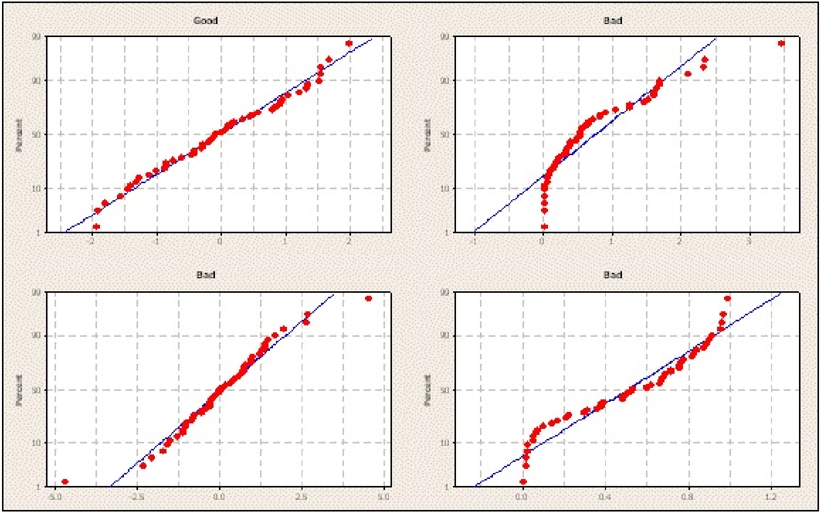

Figure 2:

Residual variance

```
proc reg data=playing;
    model height=weight;
    plot r.*p.;
run;
quit;
```

Residual variance

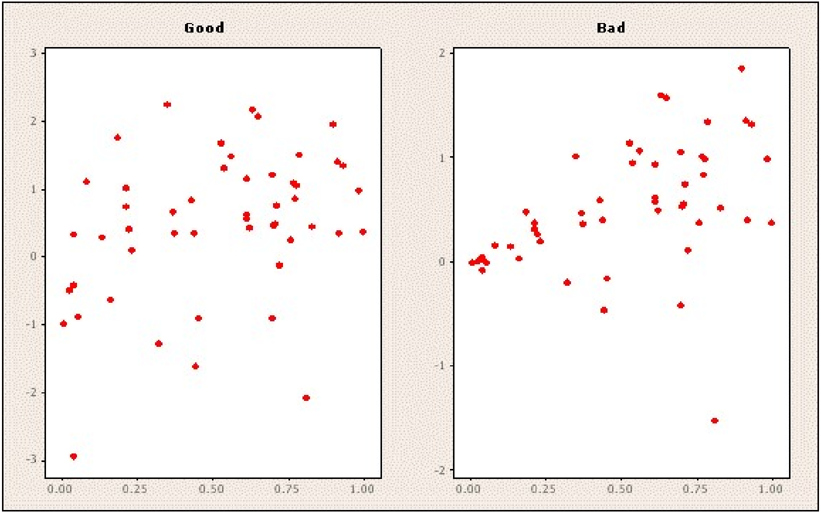

Figure 3:

What do you do if your data does not meet this assumptions?

- Try transforming the data (log, square root)
- Use a non-parametric statistical test if can not obtain normal distribution of residuals after attempting a transformation

Main Exercise

1. Download the Statistical Tests Flowchart from GitHub (.pdf)
2. Download datafile1 (.csv) from GitHub
3. Perform assumptions check using your statistical analysis software
4. Write a report summary of results (text file) for this datafile and conclude whether or not linear regression is appropriate for this data.
5. Push your report summary to the GitHub
6. Download datafile2 (.csv) from GitHub
7. Perform assumptions check using your statistical analysis software
8. Write a report summary of results (text file) for this datafile and conclude whether or not linear regression is appropriate for this data.
9. Push your report summary to the GitHub