Statistical Methods for Insurance: Linear Models

Di Cook & Souhaib Ben Taieb, Econometrics and Business Statistics, Monash University

W4.C2

Overview of this class

- Fitting a linear model to olympic medal tally
- Review of linear regression
- READING: Ch 5, Diez, Barr, Cetinkaya-Rundel

Modeling Olympic medal counts

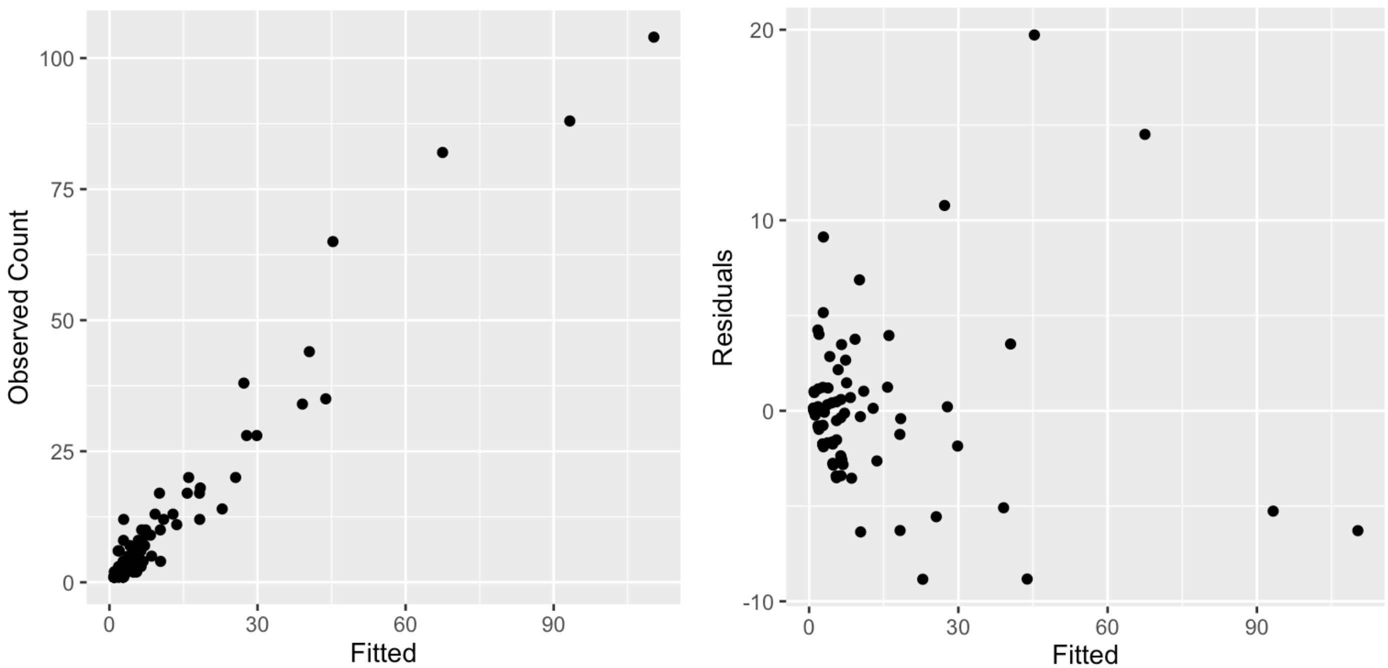
term	estimate	std.error	statistic	p.value
(Intercept)	1.28	0.67	1.91	0.06
Total.2008	0.91	0.04	20.80	0.00
Pop_std	-0.51	0.54	-0.94	0.35
GDP_std	1.27	0.85	1.50	0.14

$$M_{2012} = \beta_0 + \beta_1 M_{2008} + \beta_2 Population + \beta_3 GDP + \varepsilon$$

Model summary

```
#>   null.deviance df.null logLik AIC BIC deviance df.residual
#> 1         30252      84   -242 494 507     1486          81
```

Fit and residuals



5/22

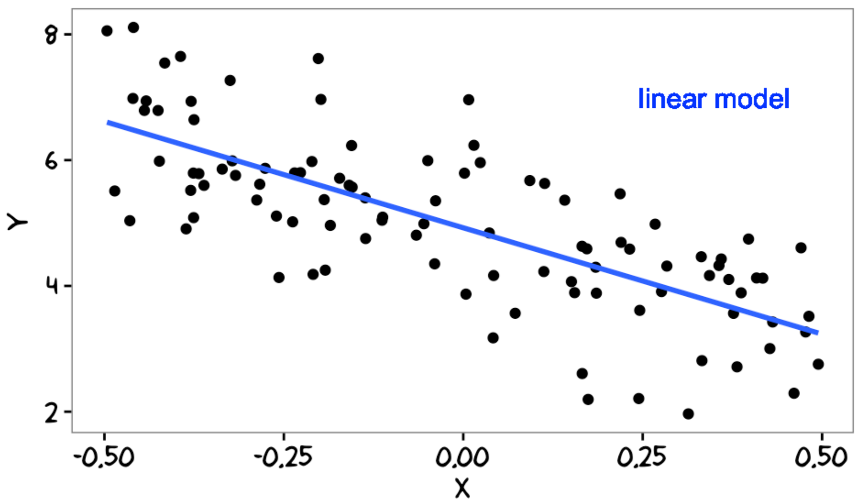
Make plots interactive

Make plots interactive

Simple linear model

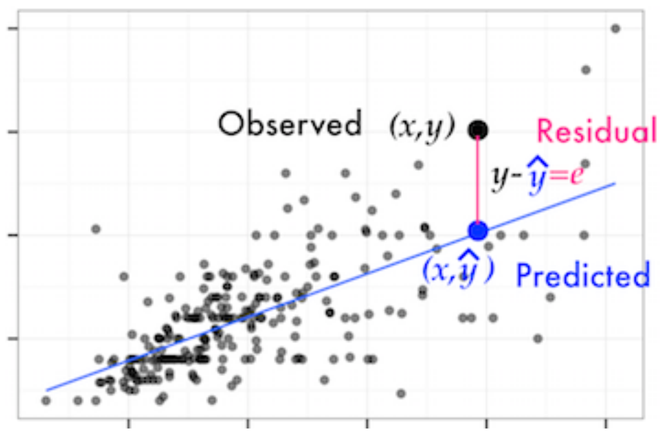
$$Y = \beta_0 + \beta_1 X + \varepsilon$$

- Explains how response variable (Y) changes in relation to explanatory variable (X), on average.
- Use line to predict value of Y for a given value of X



Observed, fitted, residuals

- Observed value is Y (a point on plot)
- Fitted value is $\hat{Y}$, a value that lies on the line
- Residual is the difference between observed and fitted, $e = Y - \hat{Y}$



9/22

Fitting process

- Minimizing the sum of squared residuals produces the best fitting line.
- Minimizes $\sum e^2$
- Line that is closest to the points, as a whole.

Parameter interpretation

- Line of best fit: $\hat{Y} = b_0 + b_1 X$
- b_0 is the intercept of the line with y-axis
- b_1 is the slope of the line

Calculating manually

Given standard deviation of X, s_x, standard deviation of Y, s_y, and the correlation, r, between the two, the slope is computed by

$$b_1 = r\frac{s_y}{s_x}$$

and given the sample means $\bar{X}, \bar{Y}$

$$b_0 = \bar{Y} - b_1\bar{X}$$

YOUR TURN

Is the point $\bar{X}, \bar{Y}$ on the regression line?

Prediction

For given X values, plug these into the model equation to predict Y,

$$\hat{Y} = b_0 + b_1 X$$

Goodness of fit

- R^2 is the proportion of variation in Y that is explained by X. Computed by

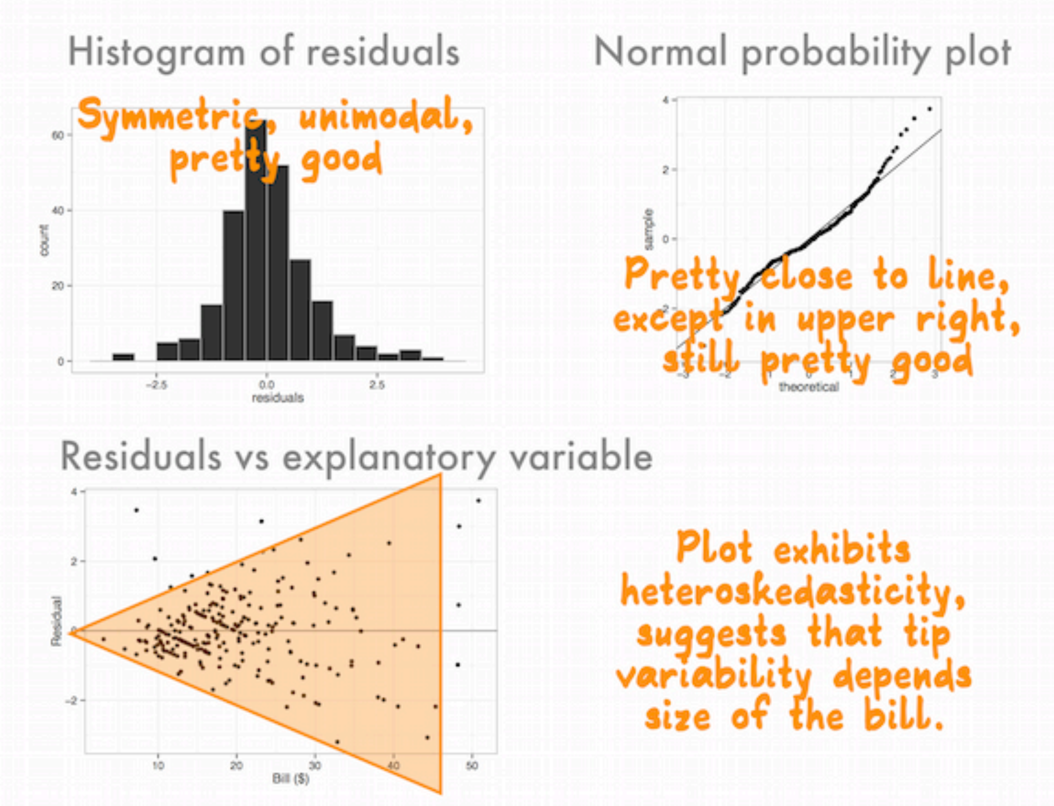
$$R^2 = 1 - \frac{\sum e^2}{\sum Y^2}$$

- Deviance: up to a constant, minus twice the maximized log-likelihood

Reading residual plots

- Make a histogram and normal probability plot of the residuals - for a good fit the shape should be pretty symmetric and bell-shaped
- Plot the residuals against the fitted values - for a good fit should be just a random splatter, no patterns

Residual plots



17/22

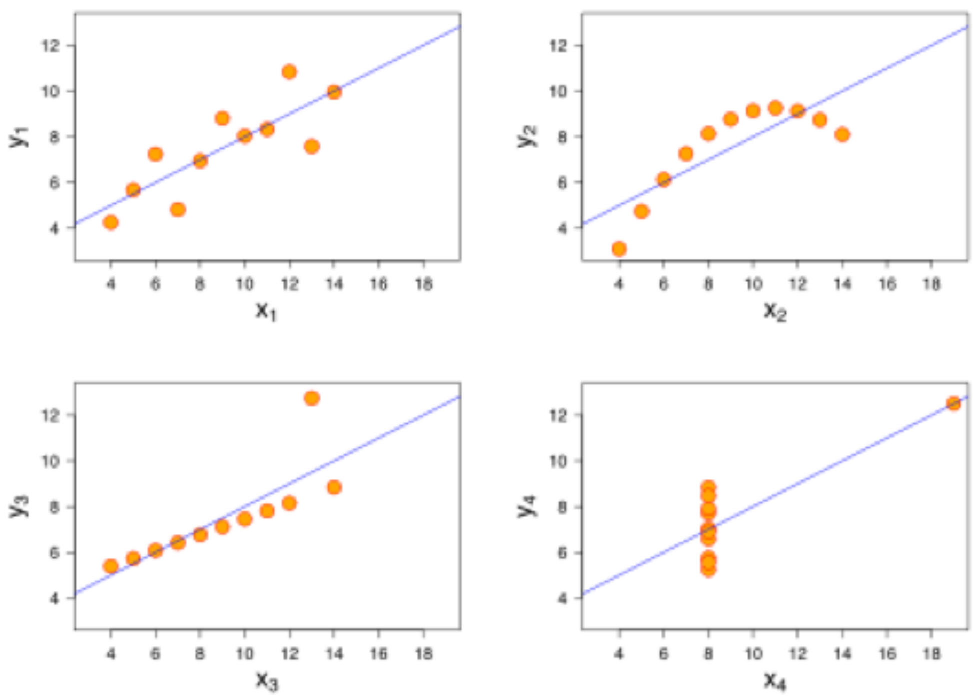
Diagnostics

- Influential points: leverage (diagonal elements of hat matrix, values $> 2p/n$ would indicate cases with high influence), cooksd (Cooks distance, measures the change in the residual when the case is removed)
- Collinearity between explanatory variables (multiple regression): variance inflation factor

Cautions

- Association is not causation
- Linear association only
- Extrapolation outside the range of the data is not recommended

Anscombe's quartet



Always plot the data, because very different patterns can lead to the same correlation.

Resources

- Statistics online textbook, Diez, Barr, Cetinkaya-Rundel
- Ancombe's quartet

Share and share alike