Unit 6: Introduction to linear regression

2. Outliers and inference for regression

Sta 101 - Spring 2015

Duke University, Department of Statistical Science

April 1, 2015

Dr. Çetinkaya-Rundel

Slides posted at http://bitly.com/sta101sp15

Announcements

- Project 2: *https://stat.duke.edu/courses/Spring15/sta101.001/projects/project2.html*

(1) R^2 assesses model fit -- higher the better

- R^2: percentage of variability in y explained by the model.
- For single predictor regression: R^2 is the square of the correlation coefficient, R.

```
cor(murder$annual_murders_per_mil, murder$perc_pov)^2
```

```
[1] 0.7052275
```

- For all regression: $R^2 = \frac{SS_{reg}}{SS_{tot}}$

```
m1 = lm(annual_murders_per_mil ~ perc_pov, data = murder)
```

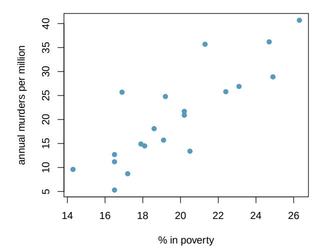
```
Analysis of Variance Table

Response: annual_murders_per_mil
          Df  Sum Sq Mean Sq F value    Pr(>F)
perc_pov   1 1308.34 1308.34  43.064 3.638e-06 ***
Residuals 18  546.86   30.38
```

$$R^2 = \frac{\textit{explained variabilty}}{\textit{total variability}} = \frac{SS_{reg}}{SS_{tot}} = \frac{1308.34}{1308.34 + 546.86} = \frac{1308.34}{1855.2} \approx 0.71$$

Clicker question

R^2 for the regression model for predicting annual murders per million based on percentage living in poverty is roughly 71%. Which of the following is the correct interpretation of this value?



(a) 71% of the variability in percentage living in poverty is explained by the model.

(b) 84% of the variability in the murder rates is explained by the model, i.e. percentage living in poverty.

(c) 71% of the variability in the murder rates is explained by the model, i.e. percentage living in poverty.

(d) 71% of the time percentage living in poverty predicts murder rates accurately.

Inference for regression uses the T distribution

- Use a T distribution for inference on the slope, with degrees of freedom $n - 2$
  - Degrees of freedom for the slope(s) in regression is $df = n - p - 1$ where p is the number of predictors (explanatory variables) in the model.
- Hypothesis testing for a slope: $H_0 : \beta_1 = 0$; $H_A : \beta_1 \neq 0$
  - $T_{n-2} = \frac{b_1 - 0}{SE_{b_1}}$
  - p-value = P(observing a slope at least as different from 0 as the one observed if in fact there is no relationship between x and y
- Confidence intervals for a slope:
  - $b_1 \pm T^\star_{n-2} SE_{b_1}$

Conditions for regression

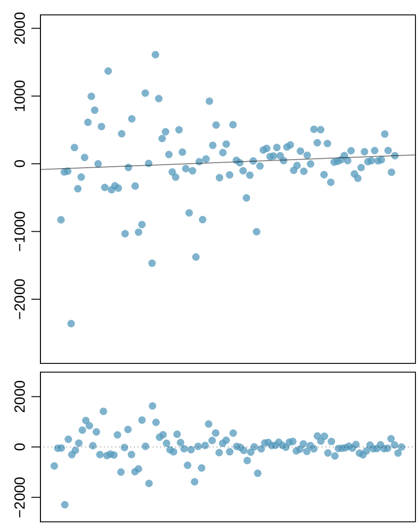
- Linearity → randomly scattered residuals around 0 in the residuals plot -- important regardless of doing inference
- Nearly normally distributed residuals → histogram or normal probability plot of residuals -- important for inference
- Constant variability of residuals (*homoscedasticity*) → no fan shape in the residuals plot -- important for inference
- Independence of residuals (and hence observations) → depends on data collection method, often violated for time-series data -- important for inference

Checking conditions

Clicker question

What condition is this linear model obviously and definitely violating?

(a) Linear relationship
(b) Non-normal residuals
(c) Constant variability
(d) Independence of observations

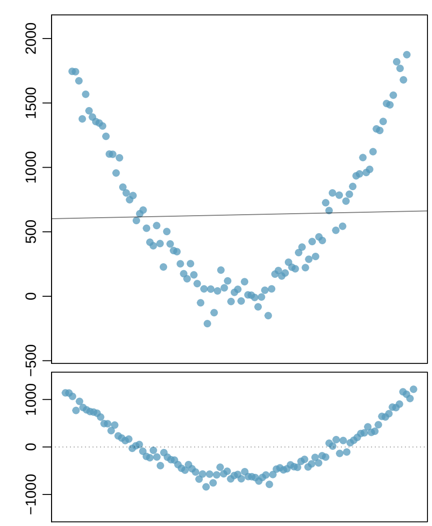


Checking conditions

Clicker question

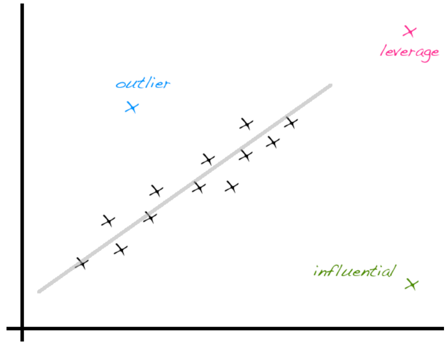
What condition is this linear model obviously and definitely violating?

(a) Linear relationship
(b) Non-normal residuals
(c) Constant variability
(d) Independence of observations



Type of outlier determines how it should be handled

- *Leverage* point is away from the cloud of points horizontally, does not necessarily change the slope
- *Influential* point changes the slope (most likely also has high leverage) -- run the regression with and without that point to determine



- *Outlier* is an unusual point without these special characteristics (this one likely affects the intercept only)
- If clusters (groups of points) are apparent in the data, it might be worthwhile to model the groups separately.

Application exercise: 6.2 Linear regression

See course website for details

Summary of main ideas

1. R^2 assesses model fit -- higher the better
2. Inference for regression uses the T distribution
3. Conditions for regression
4. Type of outlier determines how it should be handled