

Unit 6: Introduction to linear regression

2. Outliers and inference for regression

Sta 101 - Spring 2016

Duke University, Department of Statistical Science

Dr. Çetinkaya-Rundel

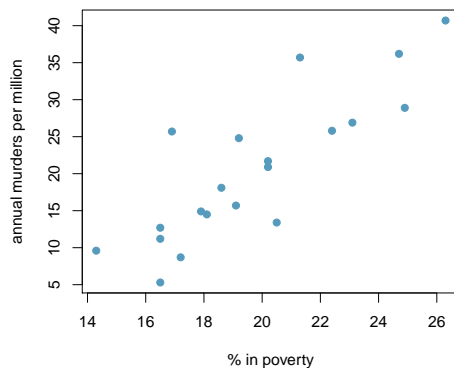
Slides posted at http://bit.ly/sta101_s16

Clicker question

Suppose you want to predict annual murder count (per million) for a series of districts that were not included in the dataset. For which of the following districts would you be most comfortable with your prediction?

A district where % in poverty =

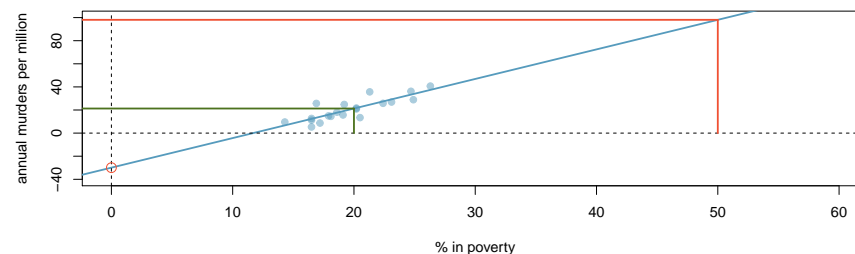
- (a) 5%
- (b) 15%
- (c) 20%
- (d) 26%
- (e) 40%



- ▶ PA 6 opens today, due Apr 10, Sun
- ▶ PS 6 due tonight
- ▶ RA 7 (last RA!) on Monday
- ▶ Project questions?
 - If you want to see sample posters from previous years, stop by office hours
 - Most important advice: Sketch out a meeting / working plan with your team **TODAY**

A note about the intercept

Sometimes the intercept might be an extrapolation: useful for adjusting the height of the line, but meaningless in the context of the data.



By hand: $\widehat{murder} = -29.91 + 2.56 \text{ poverty}$

The predicted number of murders per million per year for a county with 20% poverty rate is:

$$\widehat{murder} = -29.91 + 2.56 \times 20 = 21.29$$

In R:

```
# load data
murder <- read.csv("https://stat.duke.edu/~mc301/data/murder.csv")
# fit model
m_mur_pov <- lm(annual_murders_per_mil ~ perc_pov, data = murder)
# create new data
newdata <- data.frame(perc_pov = 20)
# predict
predict(m_mur_pov, newdata)
```

```
1
21.28663
```

4

- ▶ Regression models are useful for making predictions for new observations not include in the original dataset.
- ▶ If the model is good, the predictions should be close to the true value of the response variable for this observation, however it may not be exact, i.e. \hat{y} might be different than y .
- ▶ With any prediction we can (and should) also report a measure of uncertainty of the prediction.

5

A *prediction interval* for y for a given x^* is

$$\hat{y} \pm t_{n-2}^* s \sqrt{1 + \frac{1}{n} + \frac{(x^* - \bar{x})^2}{(n-1)s_x^2}}$$

where s is the standard deviation of the residuals, and x^* is a new observation.

- ▶ Interpretation: We are XX% confident that \hat{y} for given x^* is within this interval.
- ▶ The width of the prediction interval for \hat{y} increases as
 - x^* moves away from the center
 - s (the variability of residuals), i.e. the scatter, increases
- ▶ Prediction level: If we repeat the study of obtaining a regression data set many times, each time forming a XX% prediction interval at x^* , and wait to see what the future value of y is at x^* , then roughly XX% of the prediction intervals will contain the corresponding actual value of y .

6

By hand:

Don't worry about it...

In R:

```
# predict
predict(m_mur_pov, newdata, interval = "prediction", level = 0.95)
```

```
fit      lwr      upr
1 21.28663 9.418327 33.15493
```

We are 95% confident that the annual murders per million for a county with 20% poverty rate is between 9.52 and 33.15.

7

(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 .

```
murder %>%
  summarise(r_sq = cor(annual_murders_per_mil, perc_pov)^2)
```

```
  r_sq
1 0.7052275
```

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

```
anova(m_mur_pov)
```

Analysis of Variance Table

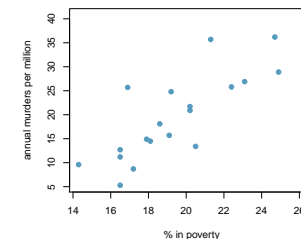
```
Response: annual_murders_per_mil
      Df Sum Sq Mean Sq F value    Pr(>F)
perc_pov  1 1308.34 1308.34  43.064 3.638e-06 ***
Residuals 18  546.86   30.38
```

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

8

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.

9

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 $df = n - k - 1$ where k is the number of slopes being estimated 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:
 - $b_1 \pm T_{n-2}^* SE_{b_1}$
 - In R:

```
confint(m_mur_pov, level = 0.95)
```

```
      2.5 %      97.5 %
(Intercept) -46.265631 -13.536694
perc_pov      1.740003   3.378776
```

10

Conditions for regression

Important regardless of doing inference

- ▶ Linearity → randomly scattered residuals around 0 in the residuals plot – important regardless of doing inference

Important for inference

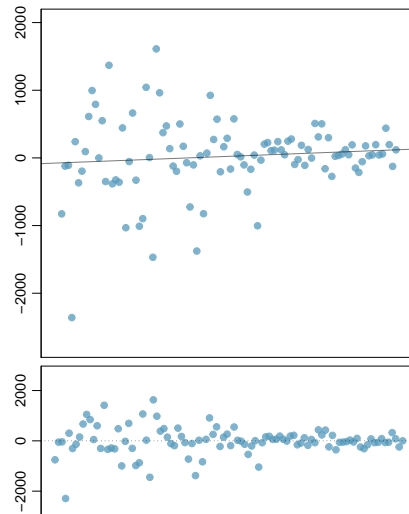
- ▶ Nearly normally distributed residuals → histogram or normal probability plot of residuals
- ▶ Constant variability of residuals (*homoscedasticity*) → no fan shape in the residuals plot
- ▶ Independence of residuals (and hence observations) → depends on data collection method, often violated for time-series data

11

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

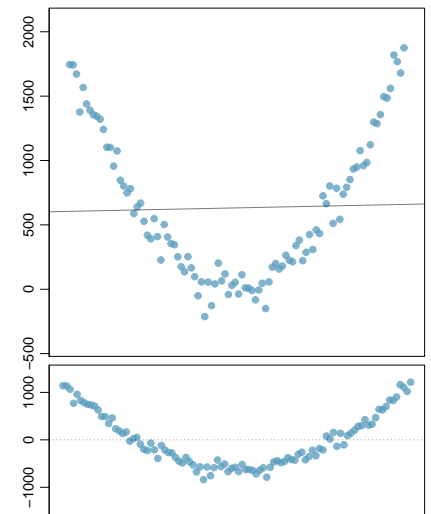


12

Clicker question

What condition is this linear model obviously and definitely violating?

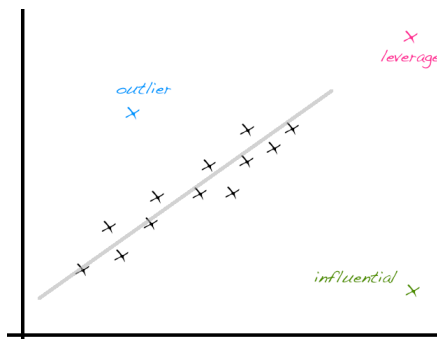
- (a) Linear relationship
- (b) Non-normal residuals
- (c) Constant variability
- (d) Independence of observations



13

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

14

15

1. Predict, but don't extrapolate
2. Predicted values also have uncertainty around them
3. R^2 assesses model fit – higher the better
4. Inference for regression uses the t -distribution
5. Conditions for regression
6. Type of outlier determines how it should be handled