Lecture 7: Linear Regression

Heidi Perry, PhD

Hack University

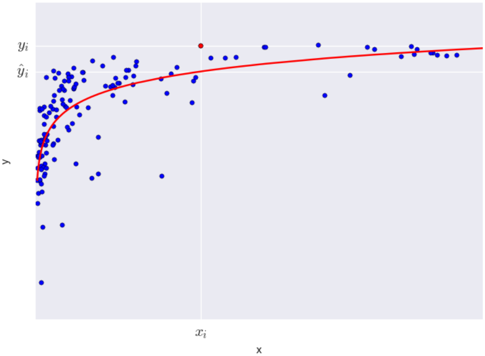
heidiperryphd@gmail.com

11/1/2016

Overview

Regression Model

"All models are wrong, but some are useful." - George Box



model: $f(x)$

$$y_i \approx f(x_i)$$

$$f(x_i) = \beta_0 + \beta_1 \times log(x_i)$$

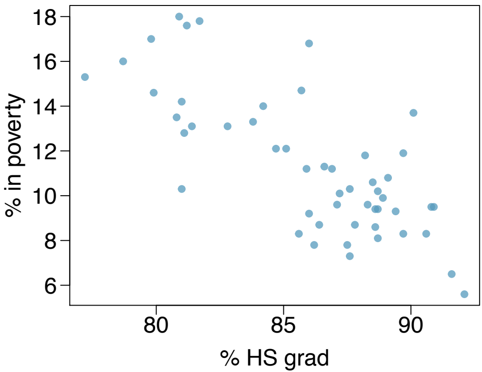
$$y_i = f(x_i) + \epsilon_i$$

$$E[\epsilon_i] = 0$$

$$\hat{y}_i = f(x_i)$$

Poverty vs. HS graduate rate

The *scatterplot* below shows the relationship between HS graduate rate in all 50 US states and DC and the % of residents who live below the poverty line (income below $23,050 for a family of 4 in 2012).



Response variable?

% in poverty

Explanatory variable?

% HS grad

Relationship?

linear, negative, moderately strong

Estimate the correlation

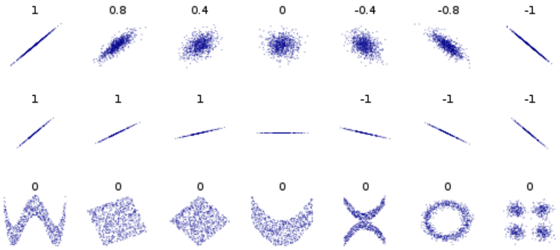
Which is closest? 0.6; -0.75; -0.1; 0.02; -1.5

Correlation

Correlation Coefficient

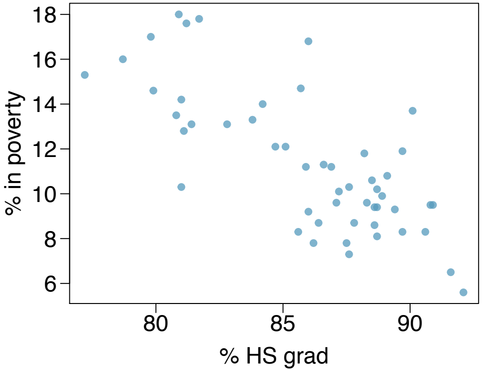
Also known as Pearson's [product-moment] coefficient measures the linear correlation between two [numerical] random variables X and Y.

$$\rho_{X,Y} = corr(X, Y) = \frac{cov(X, Y)}{\sigma_X \sigma_Y} = \frac{E[(X - \mu_x)(Y - \mu_y)]}{\sigma_x \sigma_y}$$



Poverty vs. HS graduate rate

The *scatterplot* below shows the relationship between HS graduate rate in all 50 US states and DC and the % of residents who live below the poverty line (income below $23,050 for a family of 4 in 2012).



Response variable?

% in poverty

Explanatory variable?

% HS grad

Relationship?

linear, negative, moderately strong

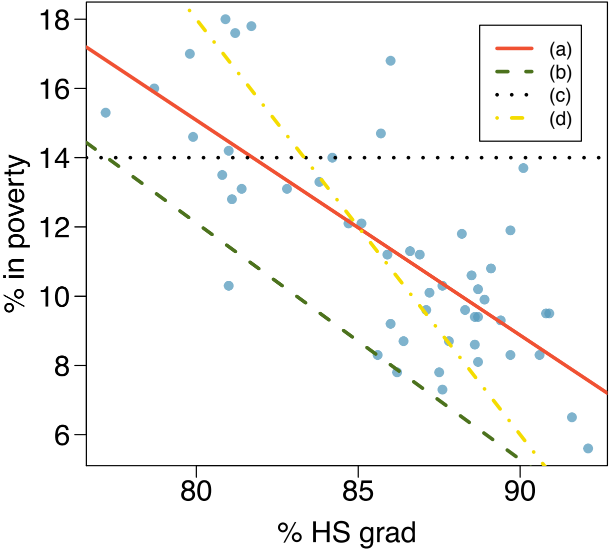
Estimate the correlation

-0.75

Eyeballing the line

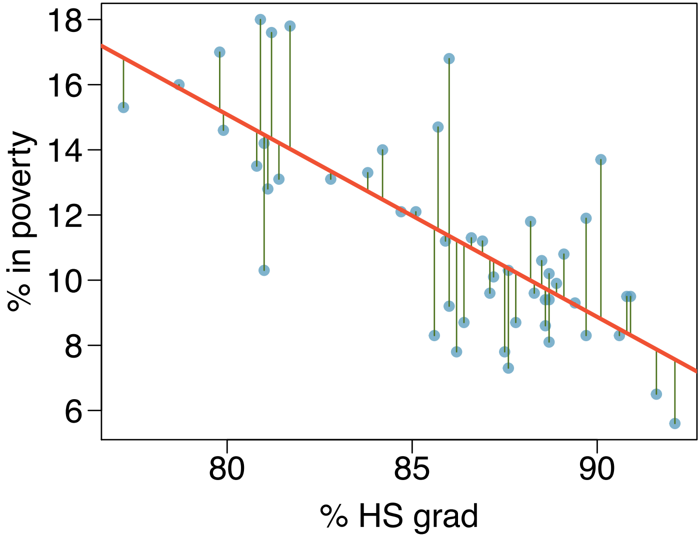
Which of the following appears to be the line that best fits the linear relationship between % in poverty and % HS grad? Choose one.

(a)



Residuals

Residuals are the leftovers from the model fit: Data = Fit + Residual

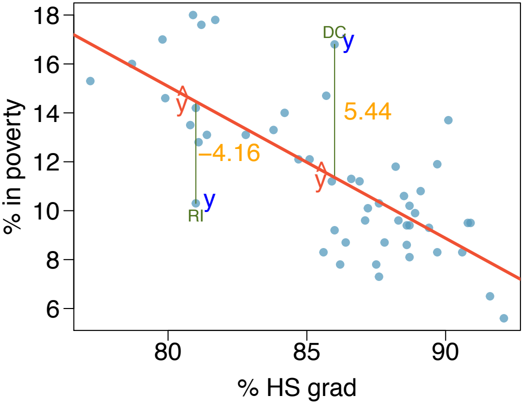


Residuals (cont.)

Residual

Residual is the difference between the observed (y_i) and predicted $\hat{y}_i$.

$$e_i = y_i - \hat{y}_i$$



- % living in poverty in DC is 5.44% more than predicted.
- % living in poverty in RI is 4.16% less than predicted.

A measure for the best line

- We want a line that has small residuals:
  1. Option 1: Minimize the sum of magnitudes (absolute values) of residuals

$$|e_1| + |e_2| + \cdots + |e_n|$$

  2. Option 2: Minimize the sum of squared residuals – *least squares*

$$e_1^2 + e_2^2 + \cdots + e_n^2$$

- Why least squares?
  1. Most commonly used
  2. Easier to compute (continuous function!) by hand and using software
  3. In many applications, a residual twice as large as another is usually more than twice as bad

The least squares line

$$\hat{y} = \beta_0 + \beta_1 x$$

predicted y

intercept

slope

explanatory variable

Notation:

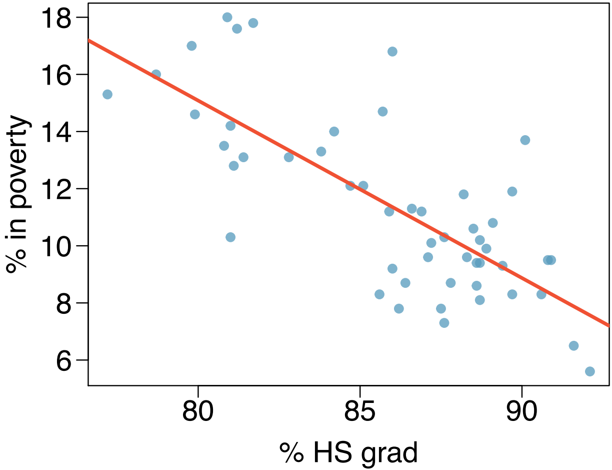
- Intercept:
  - Parameter: β_0
  - Point estimate: b_0
- Slope:
  - Parameter: β_1
  - Point estimate: b_1

Output from statsmodels OLS regression model

```
Out[11]:
<class 'statsmodels.iolib.summary.Summary'>
"""
                            OLS Regression Results
==============================================================================
Dep. Variable:                Poverty   R-squared:                       0.558
Model:                            OLS   Adj. R-squared:                  0.549
Method:                 Least Squares   F-statistic:                     61.81
Date:                Sat, 29 Oct 2016   Prob (F-statistic):           3.11e-10
Time:                        17:02:01   Log-Likelihood:                -108.74
No. Observations:                  51   AIC:                             221.5
Df Residuals:                      49   BIC:                             225.3
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [95.0% Conf. Int.]
------------------------------------------------------------------------------
Intercept     64.7810      6.803      9.523      0.000        51.111    78.451
Graduates     -0.6212      0.079     -7.862      0.000        -0.780    -0.462
==============================================================================
Omnibus:                        3.534   Durbin-Watson:                   1.977
Prob(Omnibus):                  0.171   Jarque-Bera (JB):                2.653
Skew:                           0.540   Prob(JB):                        0.265
Kurtosis:                       3.289   Cond. No.                     2.01e+03
==============================================================================
```

Regression line

$$\widehat{\% \text{ in poverty}} = 64.68 - 0.62 \ \% \ HS \ grad$$

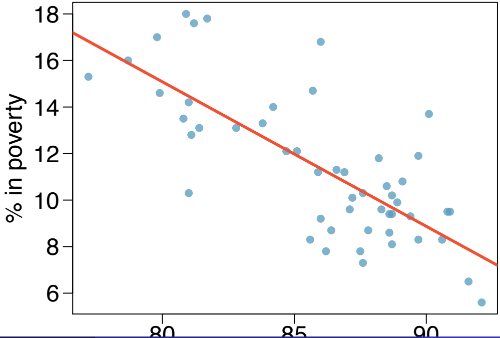


Slope

	coef	std err	t	P>|t|	[95.0% Conf. Int.]	
Intercept	64.7810	6.803	9.523	0.000	51.111	78.451
Graduates	-0.6212	0.079	-7.862	0.000	-0.780	-0.462

Slope

For each additional % point in HS graduate rate, we would expect the % living in poverty to be lower on average by 0.62% points.

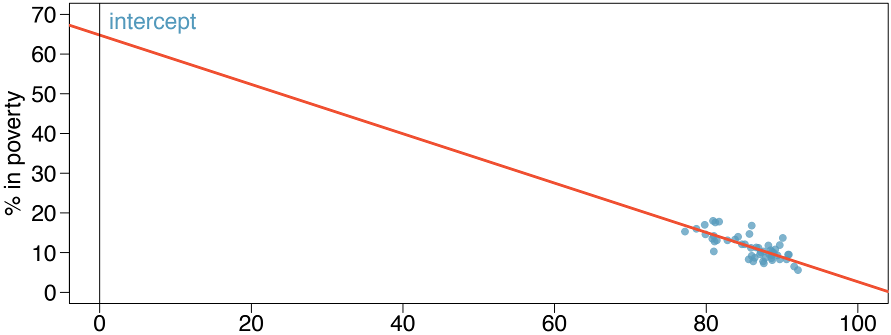


Intercept

	coef	std err	t	P>|t|	[95.0% Conf. Int.]	
Intercept	64.7810	6.803	9.523	0.000	51.111	78.451
Graduates	-0.6212	0.079	-7.862	0.000	-0.780	-0.462

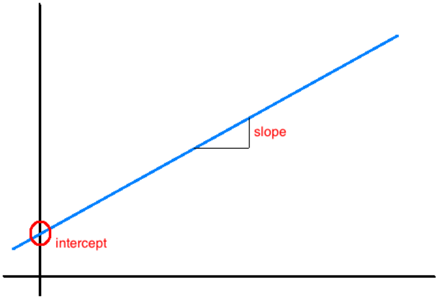
Intercept

The intercept is where the regression line intersects the *y*-axis; the value of the response parameter if the explanatory parameter is zero.



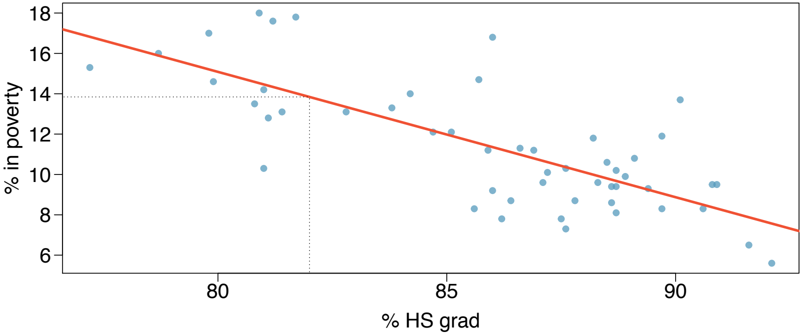
Interpretation of slope and intercept

- *Intercept:* When $x = 0$, y is expected to equal the intercept.
- *Slope:* For each unit in x, y is expected to increase / decrease on average by the slope.



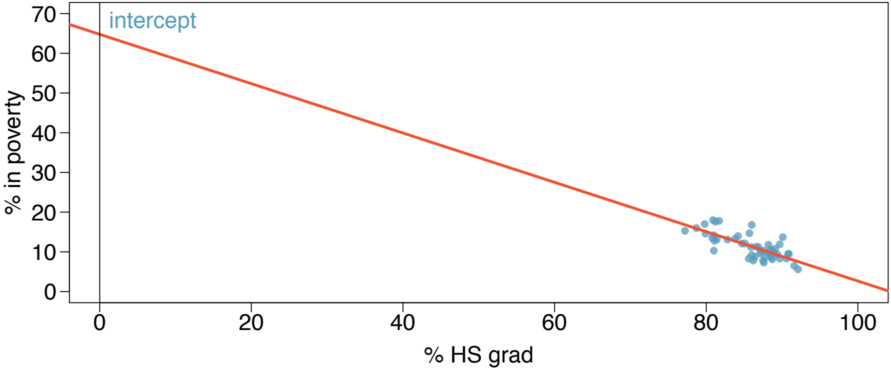
Prediction

- Using the linear model to predict the value of the response variable for a given value of the explanatory variable is called *prediction*, simply by plugging in the value of x in the linear model equation.
- There will be some uncertainty associated with the predicted value.

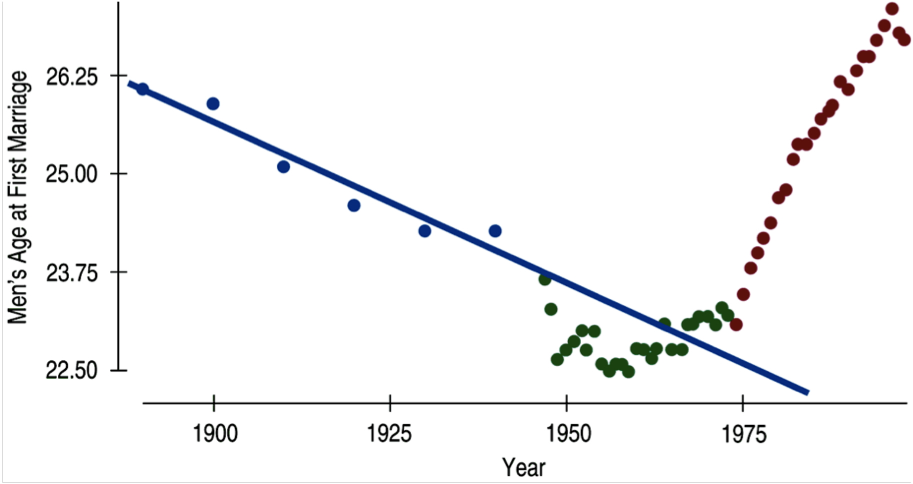


Extrapolation

- Applying a model estimate to values outside of the realm of the original data is called *extrapolation*.
- Sometimes the intercept might be an extrapolation.



Examples of extrapolation

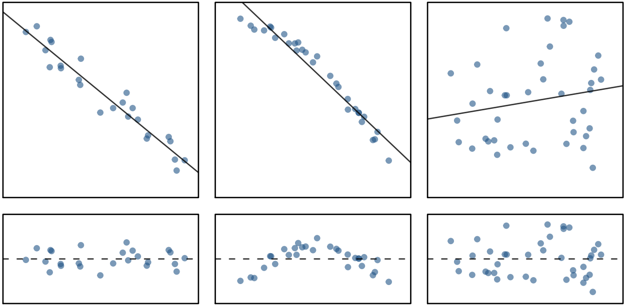


Conditions for the least squares line

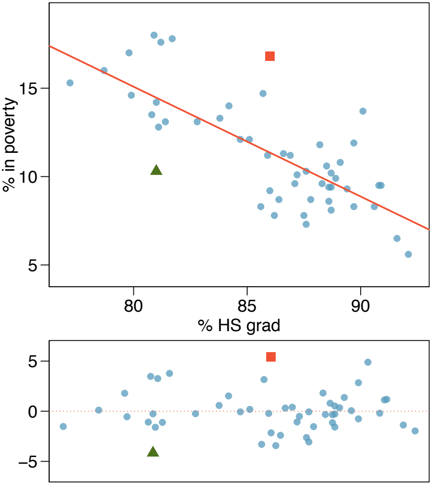
1. Linearity
2. Nearly normal residuals
3. Constant variability

Conditions: (1) Linearity

- The relationship between the explanatory and the response variable should be linear.
- See OpenIntro Statstistics Supplement for a quick introduction to fitting non-linear models.
- Check using a scatterplot of the data, or a *residuals plot*.



Anatomy of a residuals plot



RI:

$$\% \ HS \ grad = 81 \qquad \% \ in \ poverty = 10.3$$

$$\% \ \widehat{in \ poverty} = 64.68 - 0.62 * 81 = 14.46$$

$$e = \% \ in \ poverty - \% \ \widehat{in \ poverty} = 10.3 - 14.46 = -4.16$$

DC:

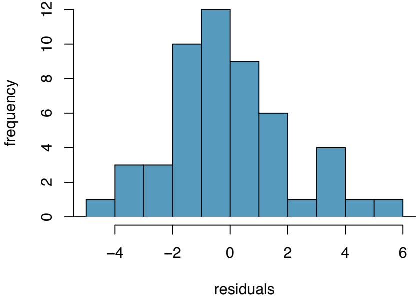
$$\% \ HS \ grad = 86 \qquad \% \ in \ poverty = 16.8$$

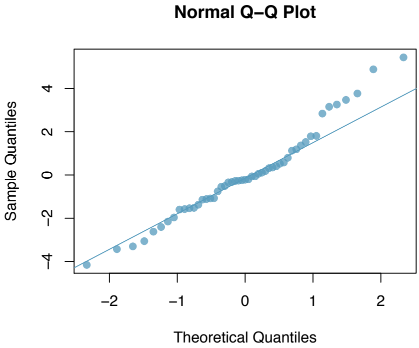
$$\% \ \widehat{in \ poverty} = 64.68 - 0.62 * 86 = 11.36$$

$$e = \% \ in \ poverty - \% \ \widehat{in \ poverty} = 16.8 - 11.36 = 5.44$$

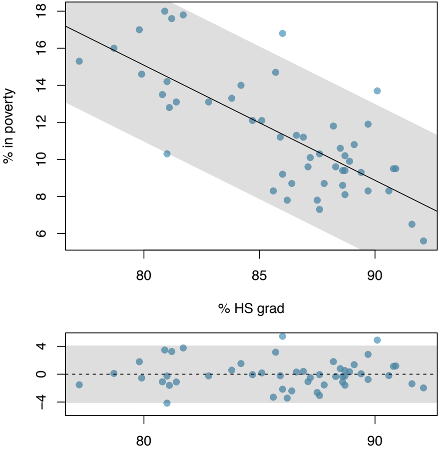
Conditions: (2) Nearly normal residuals

- The residuals should be nearly normal.
- This condition may not be satisfied when there are unusual observations that don't follow the trend of the rest of the data.
- Check using a histogram or normal probability plot of residuals.





Conditions: (3) Constant variability

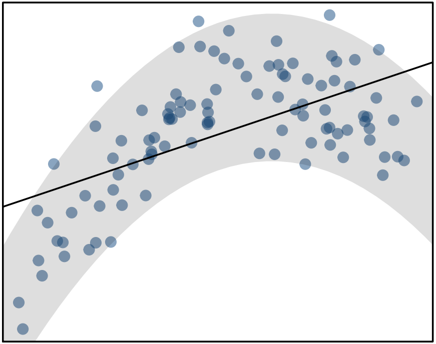


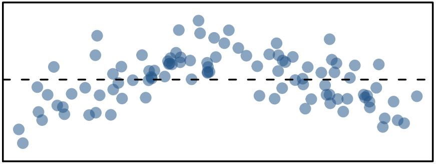
- The variability of points around the least squares line should be roughly constant.
- This implies that the variability of residuals around the 0 line should be roughly constant as well.
- Also called *homoscedasticity*.
- Check using a histogram or normal probability plot of residuals.

Checking conditions

What condition is this linear model obviously violating?

(a) Constant variability
(b) Linear relationship
(c) *Linear relationship*
(d) Normal residuals
(e) No extreme outliers

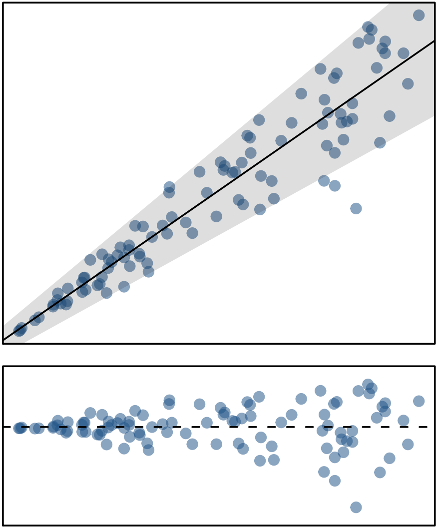




Checking conditions

What condition is this linear model obviously violating?

(a) Constant variability
(b) *Constant variability*
(c) Linear relationship
(d) Normal residuals
(e) No extreme outliers

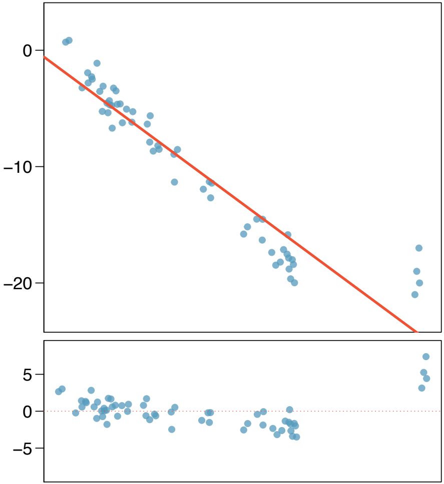


R^2

- The strength of the fit of a linear model is most commonly evaluated using R^2.
- R^2 is calculated as the square of the correlation coefficient.
- It tells us what percent of variability in the response variable is explained by the model.
- The remainder of the variability is explained by variables not included in the model or by inherent randomness in the data.
- For the model we've been working with, $R^2 = -0.62^2 = 0.38$.

Types of outliers

How do outliers influence the least squares line in this plot?

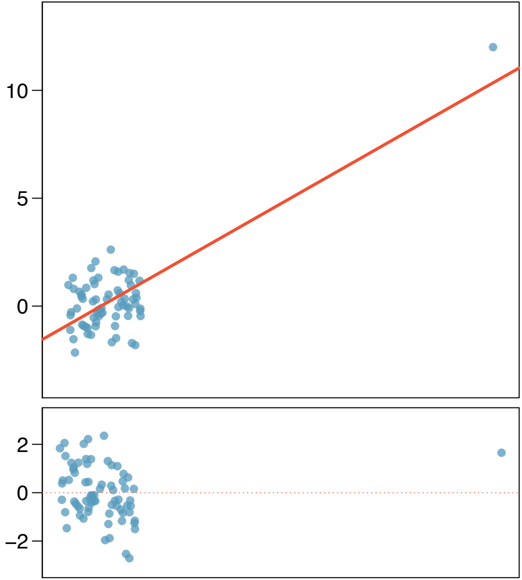


To answer this question think of where the regression line would be with and without the outlier(s). Without the outliers the regression line would be steeper, and lie closer to the larger group of observations. With the outliers the line is pulled up and away from some of the observations in the larger group.

Types of outliers

How do outliers influence the least squares line in this plot?

Without the outlier there is no evident relationship between x and y.

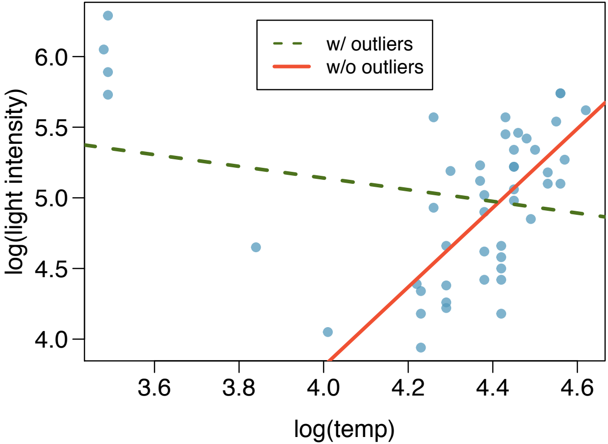


Some terminology

- *Outliers* are points that lie away from the cloud of points.
- Outliers that lie horizontally away from the center of the cloud are called *high leverage* points.
- High leverage points that actually influence the slope of the regression line are called *influential* points.
- In order to determine if a point is influential, visualize the regression line with and without the point. Does the slope of the line change considerably? If so, then the point is influential. If not, then its not an influential point.

Influential points

Data are available on the log of the surface temperature and the log of the light intensity of 47 stars in the star cluster CYG OB1.



References

David Diez, Christopher Barr, & Mine Çetinkaya-Rundel (2015)
OpenIntro Statistics, OpenIntro

Recommended Reading

OpenIntro Statistics, Chapters 7-8
Data Science from Scratch, Chapters 14-16
Art of Data Science, Chapter 7

Articles for discussion:
WHY YOU SHOULD STOP WORRYING ABOUT DEEP LEARNING AND DEEPEN YOUR UNDERSTANDING OF CAUSALITY INSTEAD
Spurious Correlations

Exercise

Lesson6_LinearRegression.ipynb