Linear Regression

September 30, 2013

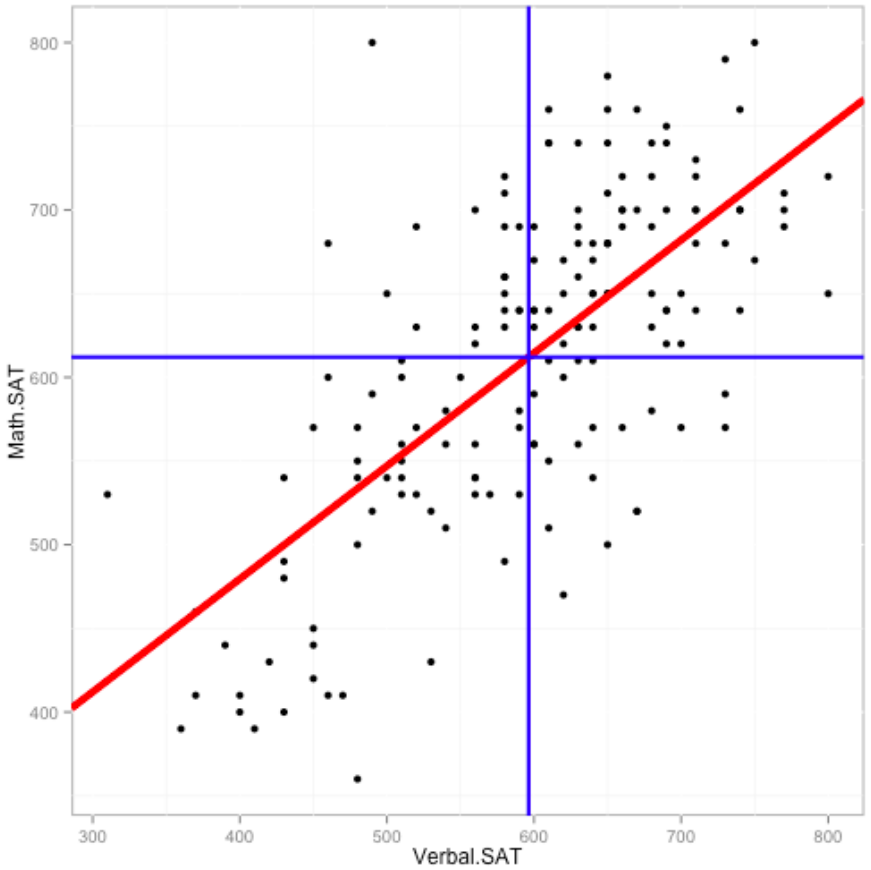
Jason Bryer
epsy530.bryer.org

The Linear Model

Math and verbal SAT Scores.

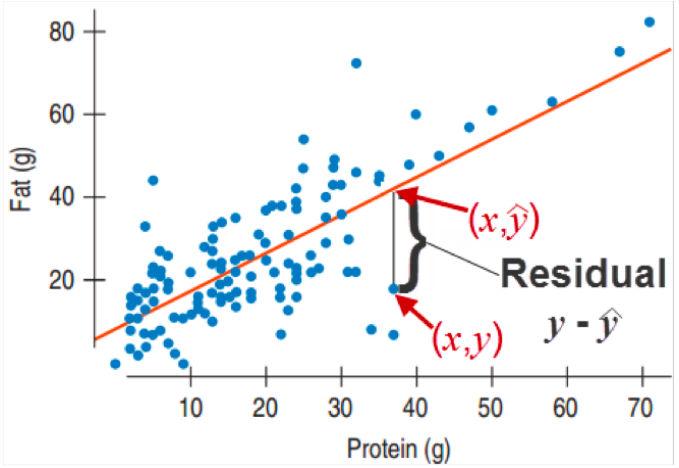
```
cor(sat$Math.SAT, sat$Verbal.SAT)
```

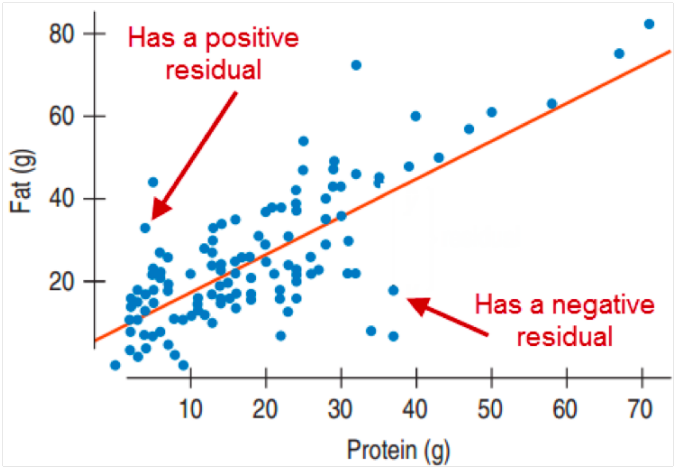
```
[1] 0.6846
```



Residuals

- $\hat{y}$ is the predicted value.
- The residual is defined by $y - \hat{y}$
- That is, the residual is the difference between the observed value and the predicted value.
- Squaring all residuals makes them possitive.
- The line of best fit then, or least squares line, minimizes the squares of the residuals.





The Line of Best Fit

Line from algebra:

- $y = mx + b$

Line of best fit:

- $\hat{y} = b_0 + b_1 x$
- b_1 is the slope.
- b_0 is the y-intercept. That is, the value of $\hat{y}$ when $x = 0$.

The Line of Best Fit: SAT Scores

```
(sat.lm = lm(Math.SAT ~ Verbal.SAT, data = sat))
```

```
Call:
lm(formula = Math.SAT ~ Verbal.SAT, data = sat)

Coefficients:
(Intercept)   Verbal.SAT
    209.554        0.675
```

- $b_1 = 0.675$ (slope) - for each 1-point increase in Verbal SAT, we expect the Math SAT score to increase by 0.675.
- $b_0 = 210$ (y-intercept)

Slope and Correlation

$$b_1 = r \frac{S_y}{S_x}$$

- Since the standard deviations are always positive, the slope and the correlation always have the same sign.
- The correlation has no units, but the slope has units of y over units of x.

The y-intercept

The y-intercept and the slope are related by:

$$\overline{y} = b_0 + b_1\overline{x}$$

- The point corresponding to the mean of x and y will always fall on the line of best fit.
- Given the mean of x, the mean of y, and the slope, we can find the y-intercept.

$$b_0 = \overline{y} - b_1\overline{x}$$

Assumptions for Using Regression

The line of best fit is also called the least squares line or the regression line. Only use the regression line to make predictions if:

- The variable must be Quantitative.
- The relationship is Straight Enough.
- There should be no Outliers.

Correlation and Prediction

- A new male student joins the class. How tall is he in inches? Best guess would be the mean ($\hat{z}_{in} = 0$).
- What if you also know his GPA was 3.94 ($z_{GPA} = 2$)?
  Best guess for height would not change.
- What would your guess for height be if you knew the student's shoe size had z = 2?
  - The correlation is positive($0 < z_{in} < 2$).
  - Since $b_0 = b_1\overline{x} - \overline{y}$ and the means for z-scores are both 0, this gives $b_0 = 0$.
  - Since the standard deviations are both 1, $b_1 = r\frac{S_y}{S_x}$ gives $b_1 = r$. Substituting into $\hat{y} = b_0 + b_1 x$ gives:

$$\hat{z}_y = rz_x$$

Regression to the mean

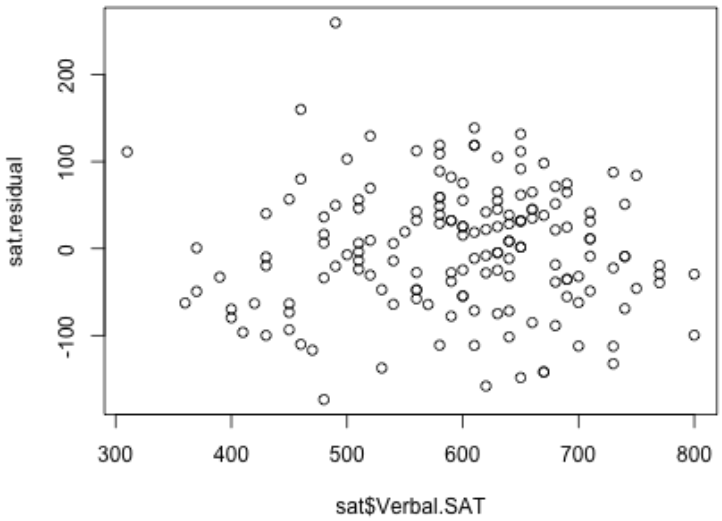
- Galton's Discovery: Tall parents have tall children, but the children's heights are likely to be closer to the mean than the parent's heights.
- Since $-1 \leq r \leq 1$, rz_x is smaller in absolute value than the z_x. This is called regression to the mean.
- The greater the deviation of a random variable from its mean, the greater the probability that the next measured variate will deviate less fars.
- In other words, an extreme event is likely to be followed by a less extreme event.

Residuals Revisited

The regression model is a good model if the residual scatterplot has no interesting features.

- No direction
- No shape
- No bends
- No outliers
- No identifiable pattern

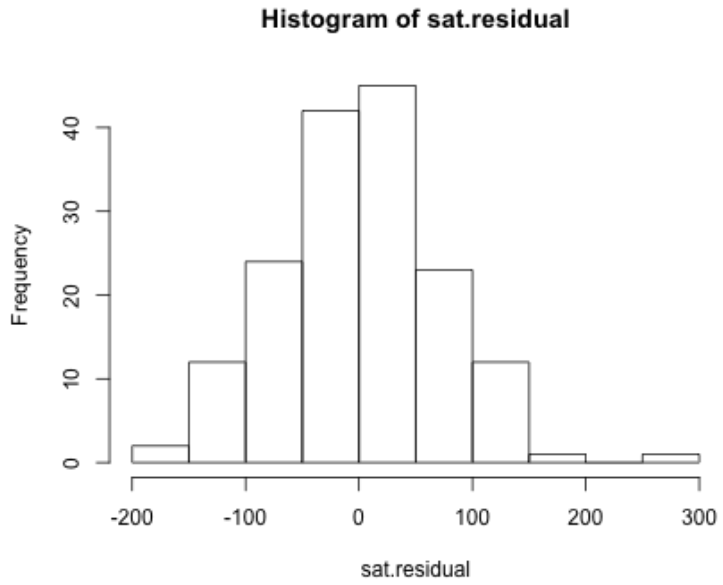
```
sat.residual = resid(sat.lm)
plot(sat$Verbal.SAT, sat.residual)
```



The Residual Standard Deviation

- Since the mean of the residuals is 0, the standard deviation of the residuals is a measure of how small the residuals are.
- Equal Variance Assumption: A good model will have the spread of the residuals consistent and small

```
hist(sat.residual)
```



Comparing the Variation of y with the Variation of the Residuals

$r = -1$ or $r = 1$

- The residuals are all 0. There is no variation of the residuals.

$r = 0$

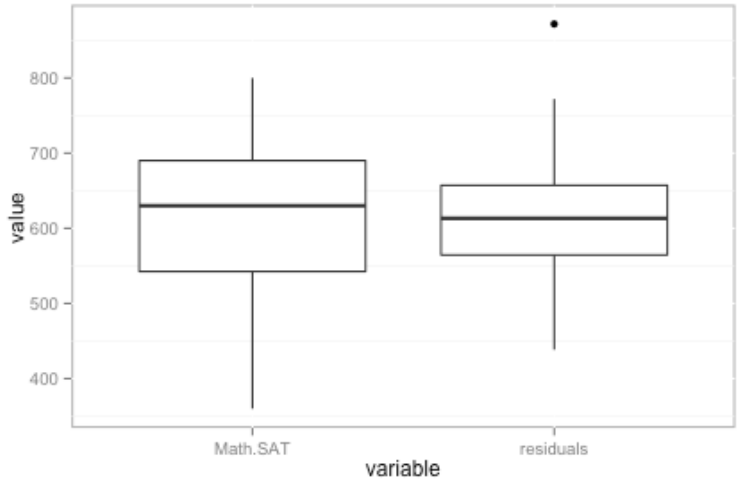
- The regression line is horizontal through the mean.
- The residuals are the y values minus the mean.
- The variation of the residuals would be the same as the variation of the original y values.

Comparing the Variation of y with the Variation of the Residuals: General r

- r^2 (written R^2) gives the fraction ofthe data's variation accounted for by the model
- 47% of the variability in the Math score is accounted for by the variation in Verbal score.
- 53% of the variability in the Math score is left in the residuals (i.e unaccounted for by other factors).

```
cor(sat$Verbal.SAT, sat$Math.SAT)^2
```

```
[1] 0.4687
```



When is R-Squared Big Enough

R^2 provides us with a measure of how useful the regression line is as a prediction tool.

- If R^2 is close to 1, then the regression line is useful.
- If R^2 is close to 0, then the regression line is not useful.
- What "close to" means depends on who is using it.
- Good Practice: Always report R^2 and let the researcher decide.

Beware of Just Switching x and y

- Switching x and y in the regression equation and solving for x does not give the equation of the regression line in reverse.
- Instead, you must start over with all the computations.
- This is no big deal if you use a computer or calculator, since the data is already entered.

Conditions to Check For

- Quantitative Variable Condition: Regression analysis cannot be used for qualitative variables.
- Straight Enough Condition: The scatterplot should indicate a relatively straight pattern.
- Outlier Condition: Outliers dramatically influence the fit of the least squares line.
- Does the Plot Thicken? Condition: The data should not become more spread out as the values of x increase. The spread should be relatively consistent for all x.

Conditions on the Scatterplot of the Residuals

- There should be no bends.
- There should be no outliers.
- There should be no changes in the spread from one part of the plot to another.

Causation and Regression

Never report out a cause and effect relationship based solely on regression analysis.

- Even when correlation is high and the model was reasonably linear, we would need a scientific explanation to conclude cause and effect. Regression analysis alone can never prove cause and effect.

What Can Go Wrong?

- Don't fit a straight line to a nonlinear relationship.
  If there are curves and bends in the scatterplot, don't use regression analysis.
- Don't ignore outliers.
  Instead report them out and think twice before using regression analysis.
- Don't invert the regression.
  Switching x and y does not mean just solving for x in the least squares line. You must start over.