Review of Introduction to Probability and Statistics

Chris Mack, http://www.lithoguru.com/scientist/statistics/review.html

Homework #4 Solutions

1. Below is a table of data showing the evaporation coefficient of burning fuel droplets in an engine as a function of surrounding air velocity.

x: Air Velocity (cm/s)	y: Evaporation Coefficient (mm^2/s)
20	0.18
60	0.37
100	0.35
140	0.78
180	0.56
220	0.75
260	1.18
300	1.36
340	1.17
380	1.65

   a. Calculate the mean and variance for x and y
   b. Calculate the covariance of x and y
   c. From the results of (a) and (b), calculate the linear regression coefficient, r
   d. From the above results, calculate the least-squares estimates for the slope and intercept of a straight-line fit of the data
   e. Plot the data in Excel, and use the linear trendline function to display the best fit line and equation. How does the Excel best fit line compare to your answer in part (d)

Solution:

(a) From the accompanying Excel spreadsheet,

$$\bar{x} = 200\frac{cm}{s}, s_x^2 = 14667\left(\frac{cm}{s}\right)^2, \bar{y} = 0.835\frac{mm^2}{s}, s_y^2 = 0.2375\left(\frac{mm^2}{s}\right)^2$$

(b) From the accompanying Excel spreadsheet, *cov*(x,y) = 56.16 (cm/s)(mm^2/s)

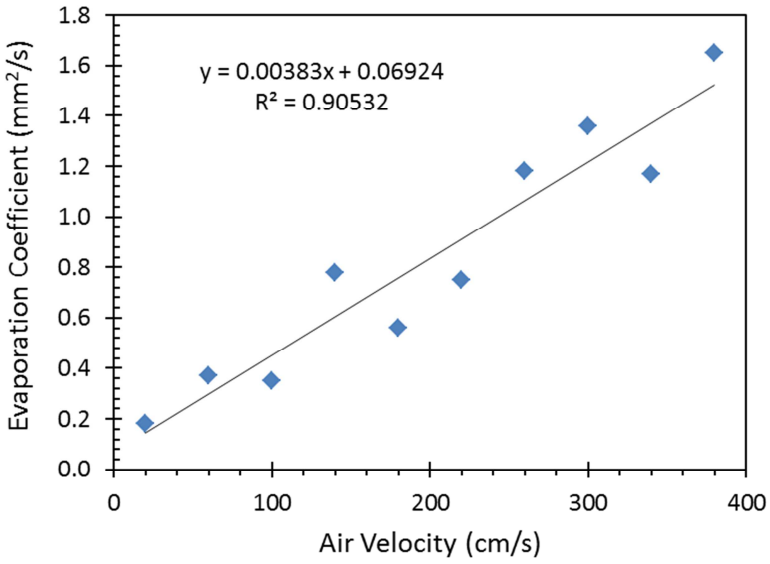
(c)

$$r = \frac{cov(x,y)}{s_x s_y} = \frac{56.1555}{\sqrt{(14667)(0.2375)}} = 0.9515$$

(d) $slope = \frac{cov(x,y)}{s_x^2} = \frac{56.1555}{14666.7} = 0.003829\left(\frac{mm^2}{s}\right)/\left(\frac{cm}{s}\right)$ (strange units!)

(e) $intercept = \bar{y} - slope \cdot \bar{x} = 0.835 - 0.003829\,(200) = 0.0692$ (mm^2/s)

(f)



2. A random sample of 120 students from an incoming freshman college class were selected for a study to determine whether the students GPA at the end of the freshman year can be predicted from their ACT test score. Assuming a first-order (straight line) regression model is appropriate,

    a. Obtain the least-squares estimate of the slope and intercept and state the regression function.
    b. Plot the regression function with the data.
    c. For an ACT score of 30, what is the expected mean freshman year GPA?
    d. Check and discuss all assumptions that went into your least-squares regression.

Note: every statistic reported should always include a confidence interval. For this problem, use a 95% confidence interval.

Solution:

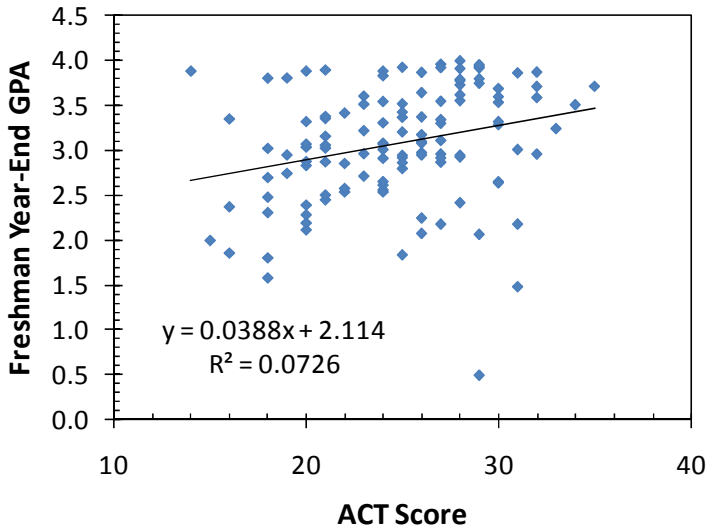
a) Using the Excel LINEST function (see accompanying spreadsheet),

$\hat{y} = b + ax = 2.11 + 0.0388x$
b = 2.11, with standard error = 0.321
a = 0.0388, with standard error = 0.0128 (GPA/ACT score)

With 118 degrees of freedom, the 95% confidence interval critical t- value = 1.9803 (almost identical with the z-critical value of 1.96, since the sample size is so large). Thus, the 95% confidence intervals for the regression coefficients are:

b: (1.48, 2.75)
a: (0.0135, 0.0641), GPA/ACT score

b) The regression scatterplot (from Excel):



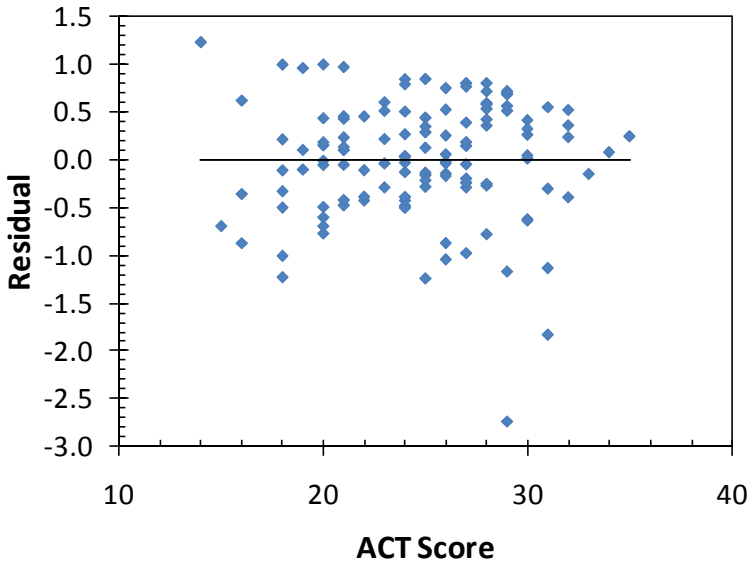
c) The expected (mean) y-value for x = 30:

$\hat{y} = 2.11 + 0.0388(30) = 3.28$

The standard error of this estimate is 0.088, so that the 95% CI for this predicted GPA is (3.10, 3.45).

d) Checking the assumptions:

1. The model is perfect (the real behavior is linear)
   a. Plot e_i vs. $\hat{y}$ or vs. each predictor variable.



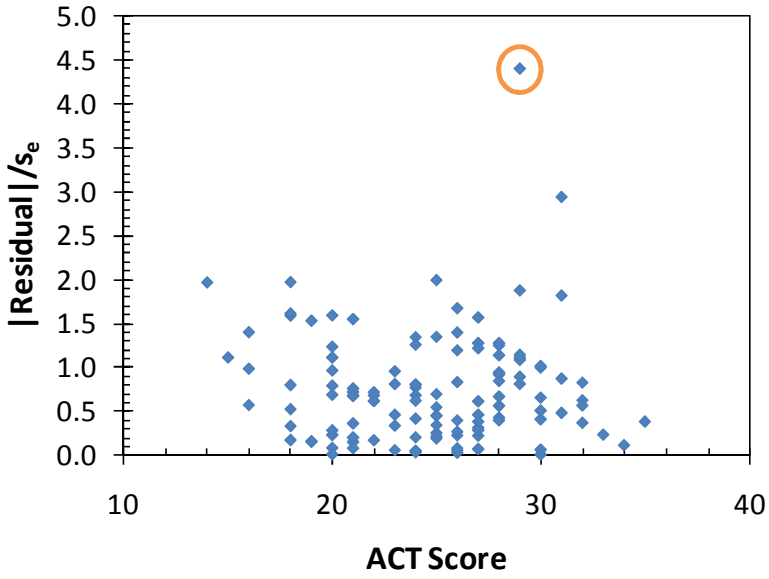
No discernable trend in the residuals can be seen. Without further information, it is not possible to look for other predictor variables that might account of some of the unmodeled variation.

Note, however, that GPA has a maximum value of 4.0. It is possible that the linear trend is saturating as the GPA approaches 4.0, so that linear behavior is only expected for the lower ACT scores.

2. All the residuals ε_i are independent
    a. Plot e_i vs. time/sequence, look for trend
    b. Think about your experimental design – any place for data non-independence to creep in?

Without further information, it is hard to check for independence. However, it seems likely that most of the student did not know each other when taking the ACT tests (which are generally administered at the same time), and thus these scores are likely to be independent.

3. All residuals ε_i have the same variance (they are iid)
    a. Plot e_i vs. $\hat{y}$ or vs. x, look for change in spread
    b. Plot $|e_i|$ or e_i^2 vs. $\hat{y}$ or vs. x, look for change in spread
    c. Check for outliers



The spread in residuals seems fairly uniform as a function of the predictor variable x. There is one potential outlier (circled in orange) where the residual is 4.5 standard deviations in magnitude. This point was included in all statistical analyses.

4. $\varepsilon \sim N(0,\sigma_\varepsilon)$
    a. Generate a normal probability plot of residuals – is it a straight line?

We did not discuss this in our review class, but we will during the full-semester course!

5. The values of each x_i are known exactly (no error in x values)
    a. Think about your experiment, do the x_i have uncertainty?

It is assumed that the ACT scores were accurately recorded and transmitted. Note, however, that if a student were to retake the ACT multiple times under identical conditions, chances are there would be some variability in the resulting score.