More than 21 statements about statistics

September 2018

(Updated: November 05 2018 13:00)

Here are some statements or questions about statistics, mainly about regression. Most of them seem to have an obvious answer.

But is the 'correct' answer the obvious one?

Ponder whether each statement is true, false or does its truth depend on unstated conditions?

1. Health and Weight

Suppose you are studying how some measure of health is related to weight. You are looking at a multiple regression of health on height and weight but you observe that what you are really interested in is the relationship between health and excess weight relative to height. What you should do is to compute the residuals of weight on height and replace weight in the model with this new variable. The resulting coefficient of 'excess weight' will give a better estimate of the effect of excess weight.

2. Measurement error: confounding factor or causal variable

Suppose you are planning to obtain observational data on the relationship between Health and Coffee (measured in grams of caffeine consumed per day). Suppose you want to control for a possible confounding factor ‘Stress’. In this kind of study it is more important to make sure that you measure coffee consumption accurately or would it be more important to try to measure ‘stress’ accurately. Either way, explain why.

3. Biases in class size surveys?

A survey of students at York reveals that the average class size of the classes they attend is 130. A survey of faculty shows an average class size of 30. The students must be exaggerating their class sizes or the faculty under-reporting.

4. Biases in wealth surveys?

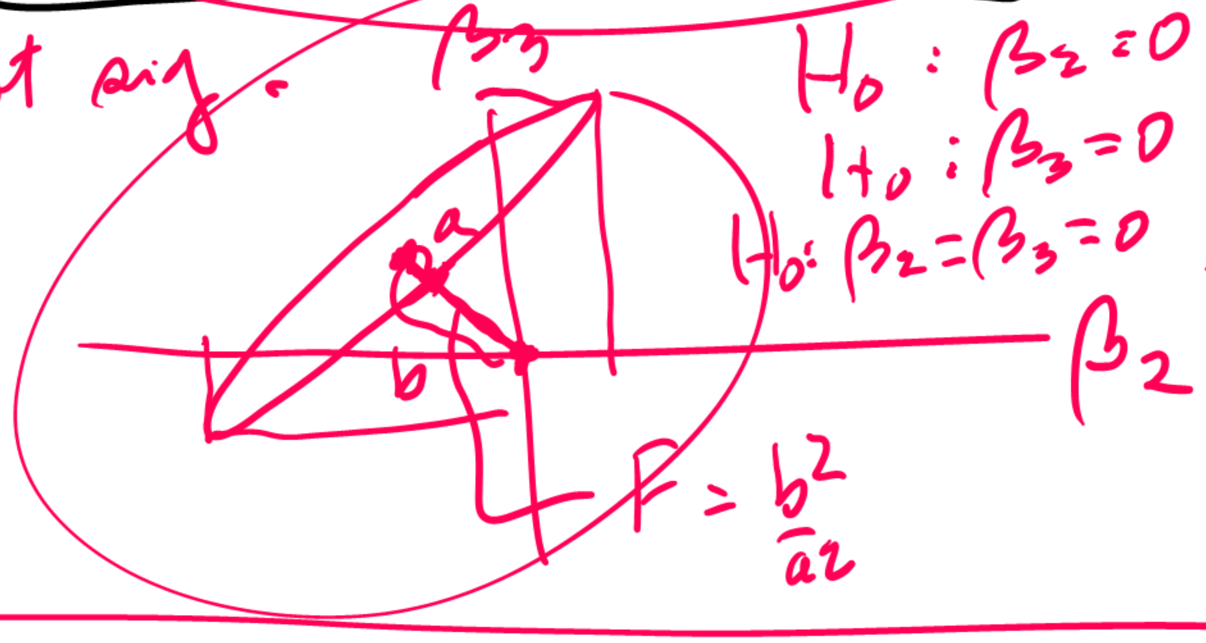
A survey of Canadian families yielded average ‘equity’ (i.e. total owned in real estate, bonds, stocks, etc. minus total owed) of $48,000. Aggregate government data of the total equity in the Canadian population shows that this figure must be much larger, in fact more than twice as large. This shows that respondents much tend to dramatically underreport their equity.

5. Dropping variables to simplify a model

In a multiple regression of Y on three predictors, X1, X2 and X3, suppose the coefficients of each of X2 and X3, are not significant. It is safe to drop these two variables and perform a regression on X1 alone. Dropping the variables with non-significant coefficients should result in a model that fits almost as well as the original model.

	Coef	. . . .	p-value
Int	1.3		0.012
X_1	5.6		0.005
X_2	0.2		0.84
X_3	−9.1		0.72

not sig.



$H_0 : \beta_2 = 0$ ns

$H_0 : \beta_3 = 0$ ns

$H_0 : \beta_2 = \beta_3 = 0$ hs

$F = \frac{b^2}{a^2}$

$$Y = \beta_0 + \beta_1 X + \beta_2 Z + \varepsilon$$

confounding

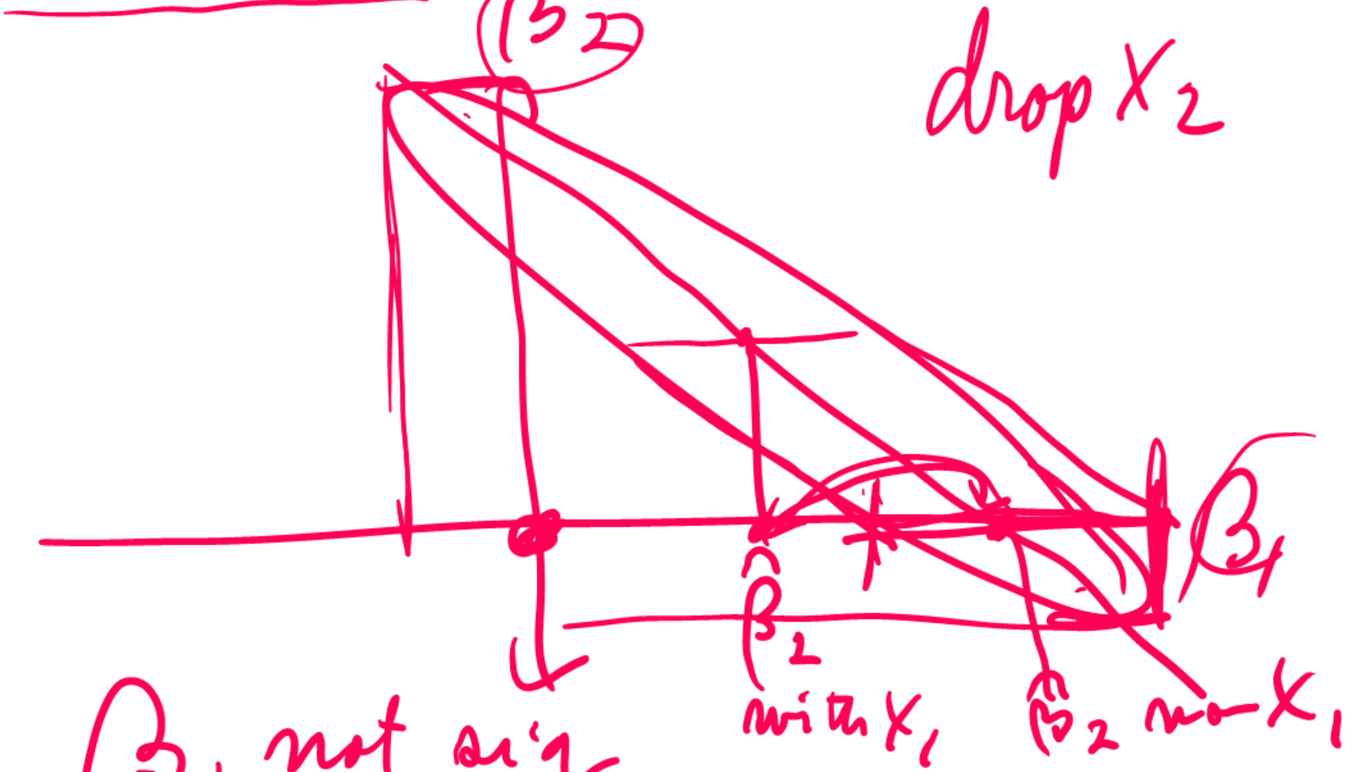
Must include Z even if not significant

6. Comparing two groups

If smoking really is bad for your health, you expect that a comparison of a group of people who have quit smoking with a group that has continued to smoke will reveal that the group quitting is, on average, healthier than the group that continued.

7. Dropping a non-significant predictor

In a multiple regression, if you drop a predictor whose effect is not significant, the coefficients of the other predictors should not change very much, nor should the p-values associated with them.



drop X_2

β_1 not sig

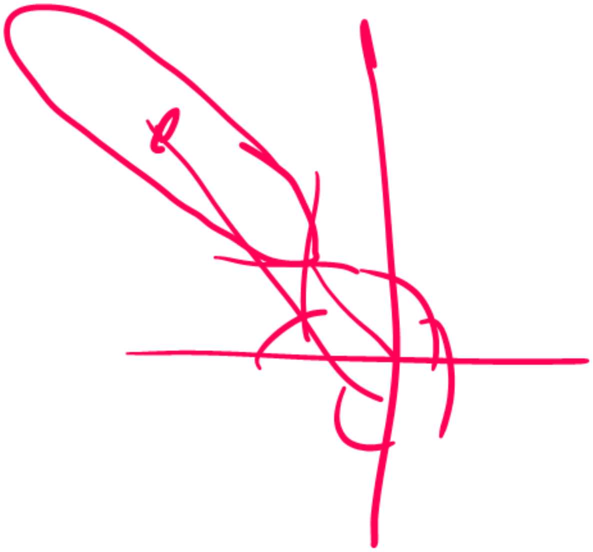
Drop X_2 β_1 becomes highly sig.

8. Interpretation of MLE

We use maximum likelihood to estimate parameters because the parameter value with the highest likelihood is the value that has the highest probability of being correct. ‘Likelihood’ is just a different word for ‘probability’.

9. Forward stepwise or backward stepwise regression?

If you want to reduce the number of predictor variables in a model, forward stepwise regression will do a good job of identifying which variables you should keep. What about backward stepwise regression?



10. Non-significant interaction

In a regression model with two predictors X1 and X2, and an interaction term between the two predictors, it is dangerous to interpret the ‘main’ effects of X1 and X2 without further qualification. However, it is okay to do so if the interaction term is not significant.

11. Comparing best and worst outcomes

In a model to assess the effect of a number of treatments on some outcome, we can estimate the difference between the best treatment and the worse treatment by using the difference in the mean outcomes.

12. Interaction and collinearity

In general we don't need to worry about interactions between variables unless there is a correlation between them.

13. Confounding factor and association with predictor

In general, a variable, Z, cannot be a *confounding factor* for the effect of another variable X_1 on Y unless Z is associated with X_1.

14. Confounding factor and association with response

In general, a variable, Z, cannot be a *confounding factor* for the effect of another variable X_1 on Y unless Z is associated with Y.

15. Importance of predictors

In a multiple regression, the predictor that is most important is the one with the smallest p-value.

16. Imputing a missing grade

You need to impute a mid-term grade for a student who missed the mid-term with a valid excuse. You plan to somehow use the grade on the final exam to impute a mid-term grade. Discuss the relative consequences of using

a. the predicted mid-term grade based on the regression equation of the mid-term grades on the final grades,
b. the student's raw grade on the final,
c. using the student's z-score on the final to impute the score on the mid-term with the same z-score, and
d. use the regression equation of the final on the mid-term to calculate the mid-term grade that would have predicted the student's actual final grade.

If you had to choose one of these four, which would you choose and why?

If you have a better solution for the previous problem, what is it and why?

17. p-values and error rate in publications

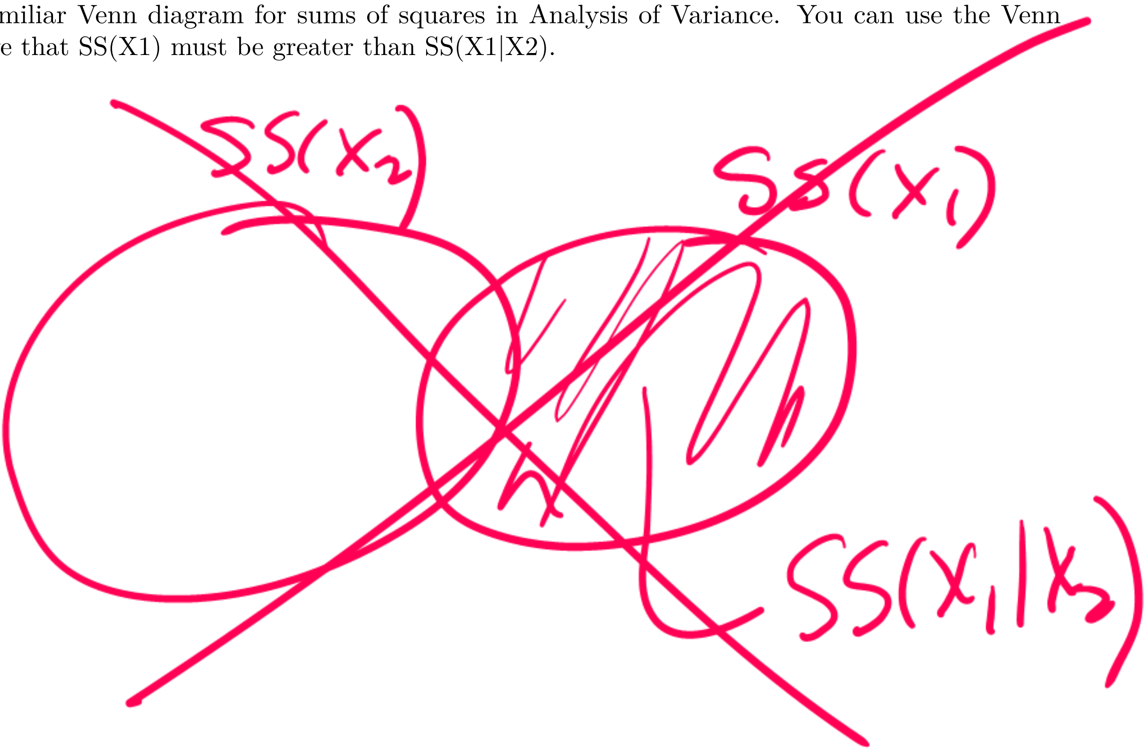
If all scientists used a p-value of 0.05 to decide which results to publish, that would ensure that at most 5% of published results would be incorrect.

18. Significance with added variables

If a variable X1 is not significant in a regression of Y on X1 then it will be even less significant in a regression of Y on both X1 and X2 where X2 is another variable. This follows since there is less variability left to explain in a model that already includes X2 than in a model that does not.

19. Extra sums of squares

Consider the familiar Venn diagram for sums of squares in Analysis of Variance. You can use the Venn diagram to prove that SS(X1) must be greater than SS(X1|X2).



20. Dropping redundant variables

The best way to deal with high collinearity between predictors is to drop predictors that are not significant.

21. AIC

AIC is useful to identify the best model among a set of models that you have selected after exploring your data if the models are not nested within each other.

22. Comparing groups

A recent study showed that people who sleep more than 9 hours per night on average have a higher chance of premature death than those who sleep fewer than 9 hours. This does not necessarily mean that sleeping more than 9 hours on average is bad for your health because the sample might not have been representative.

23. Error rate and posterior probability

Suppose a screening test for steroid drug use has a specificity of 95% and a sensitivity of 95%. This means that the test is incorrect 5% of the time. Therefore, if John takes the test and the result is ‘positive’ (i.e. the test indicates that John takes steroid drugs) the probability that he does not take steroid drugs is only 5%.