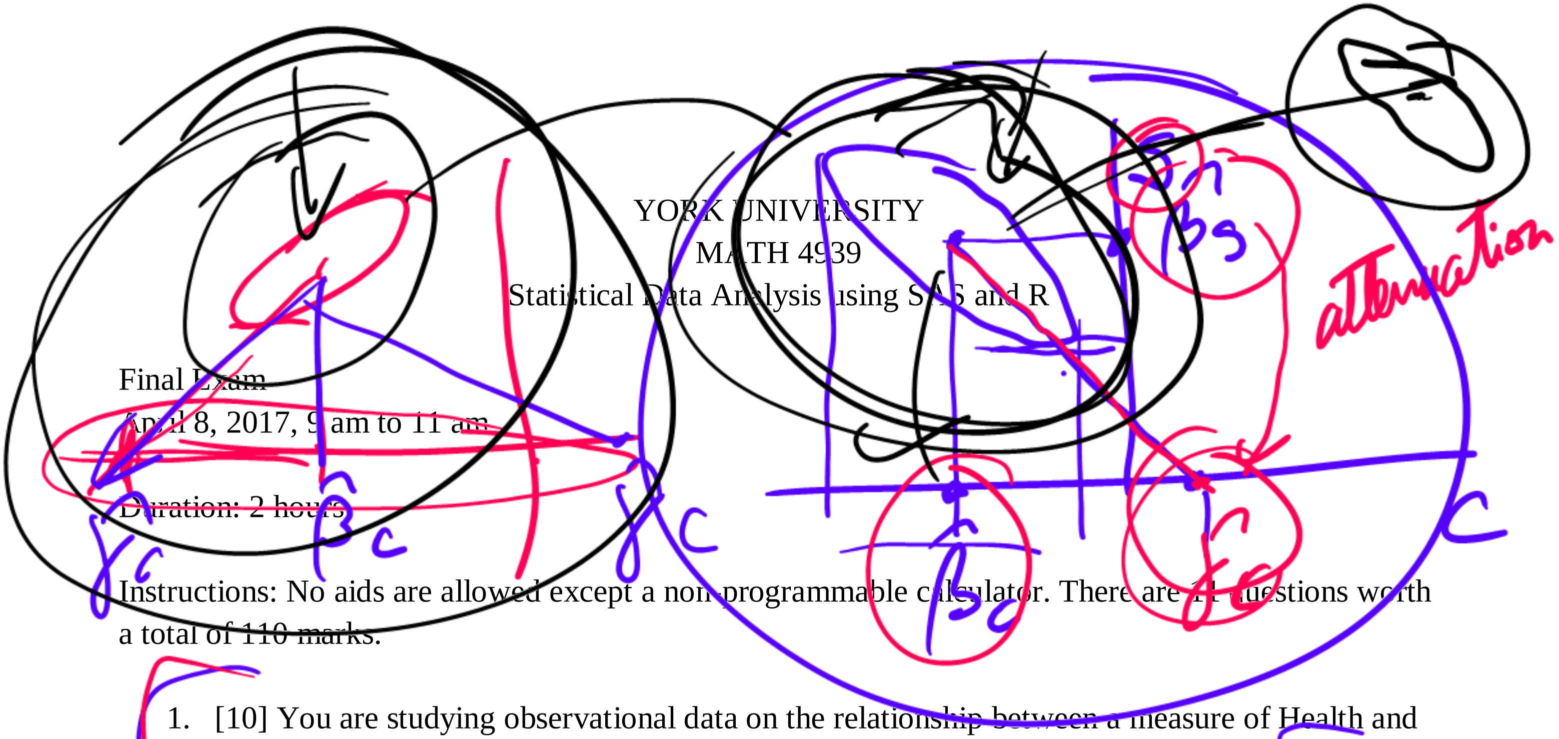
YORK UNIVERSITY
MATH 4939
Statistical Data Analysis using SAS and R

Final Exam
April 8, 2017, 9 am to 11 am

Duration: 2 hours

Instructions: No aids are allowed except a non-programmable calculator. There are 11 questions worth a total of 110 marks.



1. [10] You are studying observational data on the relationship between a measure of Health and coffee consumption (measured in grams of caffeine consumed per day). Suppose you want to control for a possible confounding factor 'Stress'. Describe the consequences of measurement error in coffee consumption? Describe the consequences of measurement error in Stress? Compare the relative impact of each source of measurement error if a) your goal is a predictive inference and b) if your goal is causal inference on the health effects of coffee consumption. What are the consequences for the probability of Type I error and of Type II error.

   S

   C

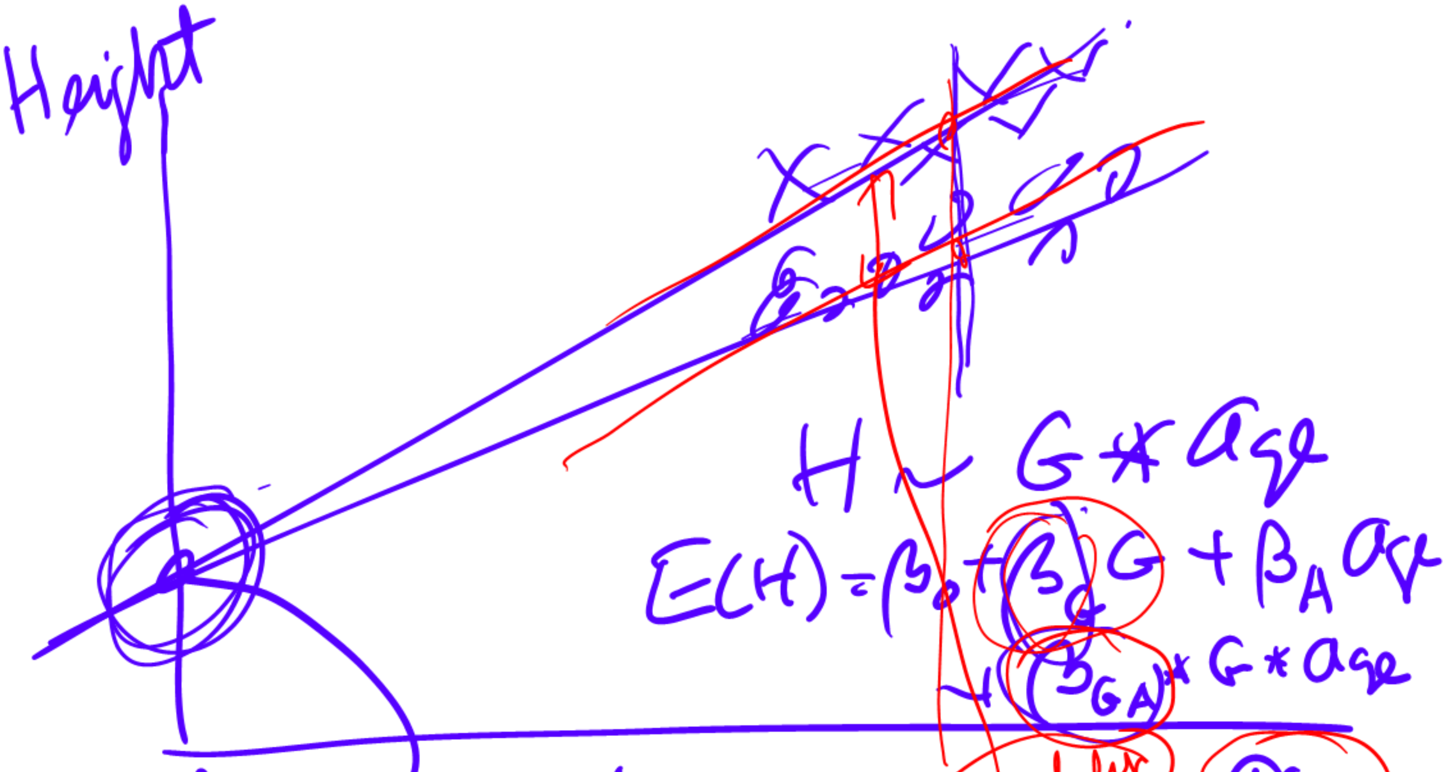
2. [10] A survey of Canadian families yielded average 'equity' (i.e. total owned in real estate, bonds, stocks, etc. minus total owed) of $48,000. Aggregate government data of the total equity in the Canadian population shows that this figure must be much larger, in fact more than three times as large. Does this show that respondents tend to dramatically underreport their equity or are there other plausible explanations consistent with honest responses by respondents? Explain briefly.

3. [10] Discuss situations when a) it would be important to include a variable that is not significant and b) it would be important to exclude a variable that is highly significant?

4. [10] In a regression model with two predictors X1 and X2, and an interaction term between the two predictors, we know that it is dangerous to interpret the `main' effects of X1 and X2 when the model includes an interaction term but is it safe to do so provided the interaction term is not significant. Discuss in a way a client would understand. Consider using an example and/or a sketch.

5. [10] Discuss the relevance of Simpson's Paradox for causal inference.

6. Write a brief essay with illustrations explaining the following issues concerning the interpretation of p-values in the output produced by the 'summary' method applied to a regression model.



$H \sim G * age$

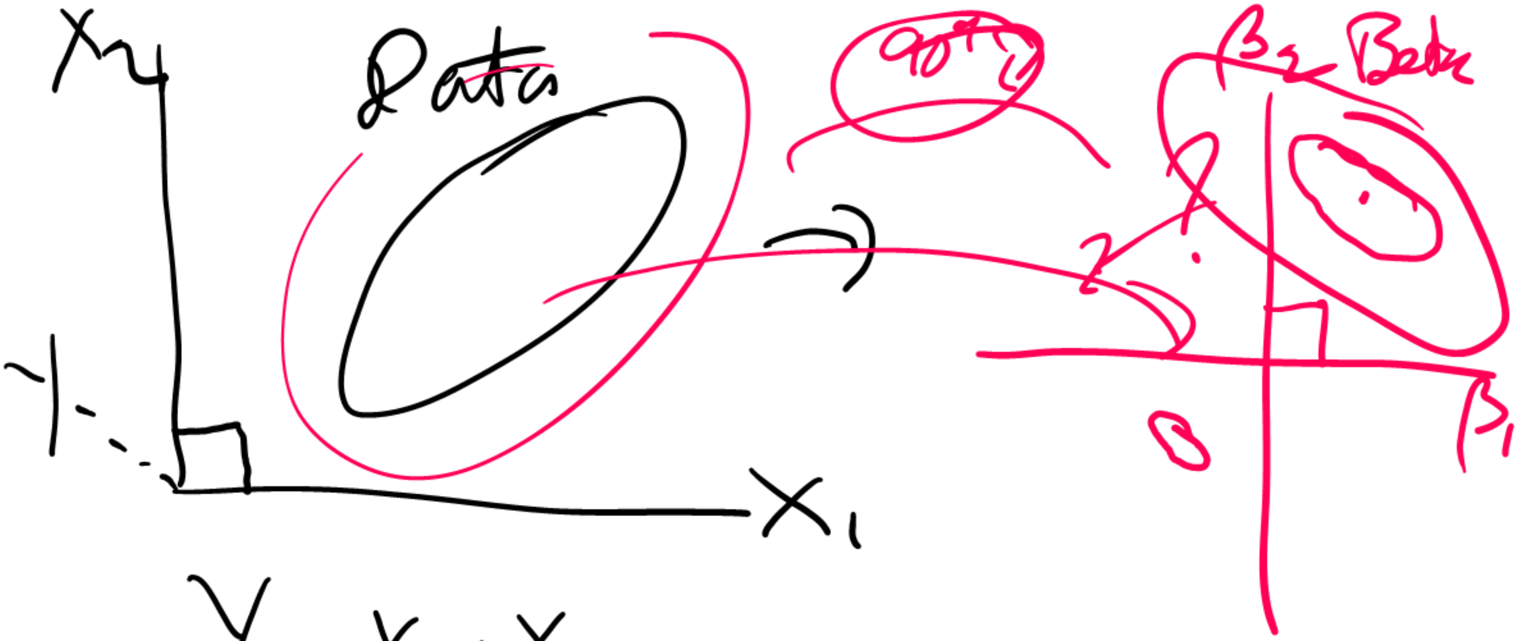
$E(H) = \beta_0 + \beta_G G + \beta_A age + \beta_{GA} * G * age$

highly sig

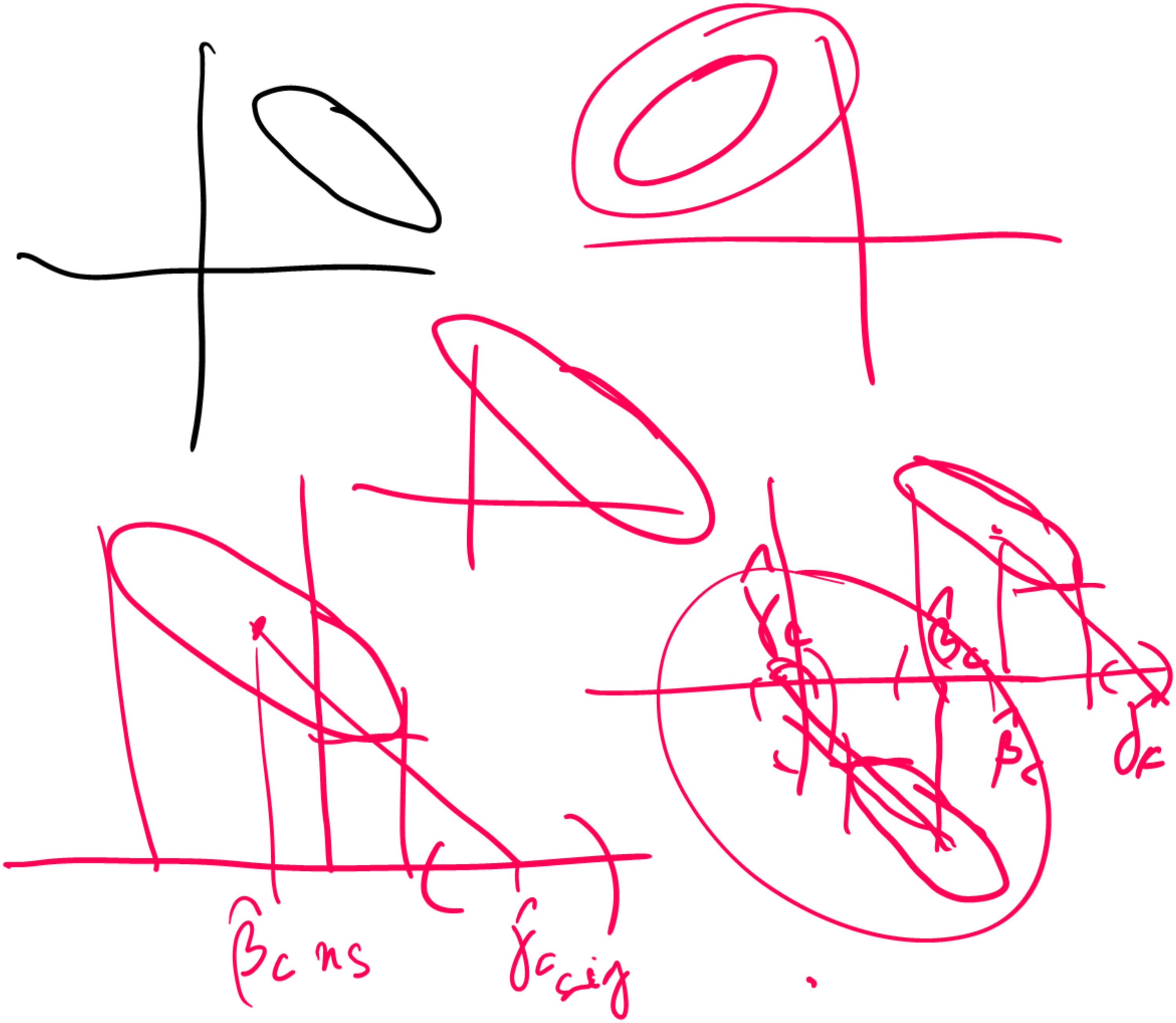
$\beta_G = \text{gap} \mid age = 0$

Must refit

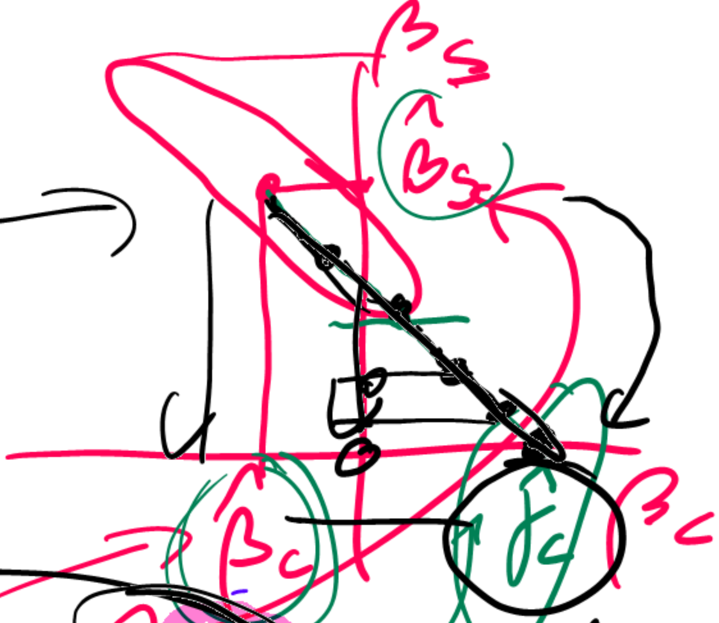
$E(H) = \beta_0 + \beta_G G + \beta_{age} age$



$Y \sim X_1 + X_2$



S



BTUYOC

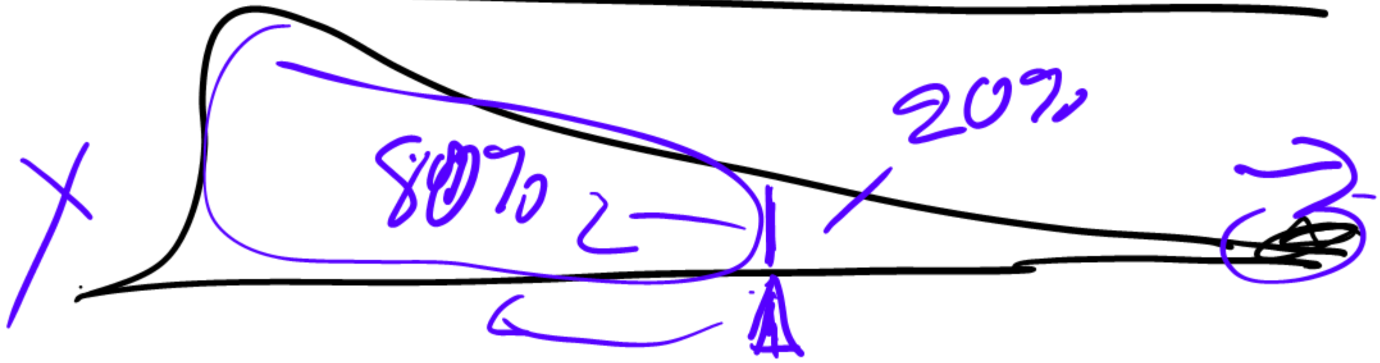
$$\hat{Y} = \hat\beta_0 + \hat\beta_c C + \hat\beta_s S$$

$$\hat{Y} = \hat\gamma_0 + \hat\gamma_c C$$

bias toward uncontrolled value of coef.

attenuation

2



$$P(\bar{X} < \mu) = ?$$

$$E(\bar{X}) = \mu$$

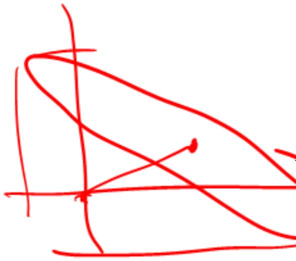
Paul Krugman

Pop-weighted density

density

\$10 X win $E(X) = 5.$

a. [5] Why should p-values for main effects that are marginal to an interaction be interpreted with caution. Are there any situation in which it is legitimate to use them? Explain briefly.



b. [5] If two main effects that are not marginal to any interaction have non-significant p-values, is it appropriate to drop both terms? Explain briefly.

7. [10] Suppose a test for mononucleosis (a disease) has a specificity (probability of a negative test result if someone does not have the disease) and a sensitivity (probability of positive test result is someone does have the disease) of 95%.

   a. Does this mean that the test will be wrong 5% of the time? Prove or disprove.

   b. If you take the test and the result is positive, does this mean that the probability that you have mononucleosis is 95%. Prove or disprove.

8. Consider the following output

```
> library(car)
> library(spida2)
> head(Prestige)
                      education income women prestige type
gov.administrators        13.11  12351 11.16     68.8 prof
general.managers          12.26  25879  4.02     69.1 prof
accountants               12.77   9271 15.70     63.4 prof
purchasing.officers       11.42   8865  9.11     56.8 prof
chemists                  14.62   8403 11.68     73.5 prof
physicists                15.64  11030  5.13     77.6 prof
> tab(Prestige, ~ type)
type
   bc  prof    wc Total
   44    31    23    98
> # women is the percentage of women in an occupation
> # type has three levels: prof, wc and bc for
> #       professional, white collar and blue collar respectively
> # education is the mean years of education for an occupation
> # income is the mean income for an coccupation
> fit <- lm(income ~ (women + education + type)^2, Prestige)
> summary(fit)

Call:
lm(formula = income ~ (women + education + type)^2, data = Prestige)
```

$Y \sim (A + B + C)^2$

$A * A : A$

income ~ (women + I(women ^ 2)) * type

arith.

```
Residuals:
    Min      1Q  Median      3Q     Max 
-8370.8  -967.7    38.1   641.3 15133.4 

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)  
(Intercept)         301.455   3607.274   0.084   0.9336  
women                23.746     83.827   0.283   0.7776  
education           700.898    415.143   1.688   0.0949 .
typeprof          -2347.177   6296.157  -0.373   0.7102  
typewc            -4494.487   8394.279  -0.535   0.5937  
women:education      -8.276     10.302  -0.803   0.4240  
women:typeprof       -5.102     63.458  -0.080   0.9361  
women:typewc         12.656     40.847   0.310   0.7572  
education:typeprof  369.593    532.130   0.695   0.4892  
education:typewc    406.283    800.590   0.507   0.6131  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2797 on 88 degrees of freedom
Multiple R-squared:  0.603,	Adjusted R-squared:  0.5623 
F-statistic: 14.85 on 9 and 88 DF,  p-value: 2.314e-14
```

0% 100% (a)

(a) - (b)

$\sum \bigcirc \times 100 = \boxed{\text{answer}}$

$\frac{\partial E(Y)}{\partial \text{women}}$

a. [5] Estimate the gender gap (the estimated difference between a job that is 0% female and a job that is 100% female), for a white collar job with 9 years of education.

b. [5] Estimate the gap between professional and white collar jobs among jobs that are 50% female with 12 years of education.

c. [5] Sketch the estimated response function as a function of education for professional occupations that have 10% women and for occupations that have 50% women. Use a scale for education going from 10 to 20. Indicate the height of each response curve at 10 years and at 20 years of education.

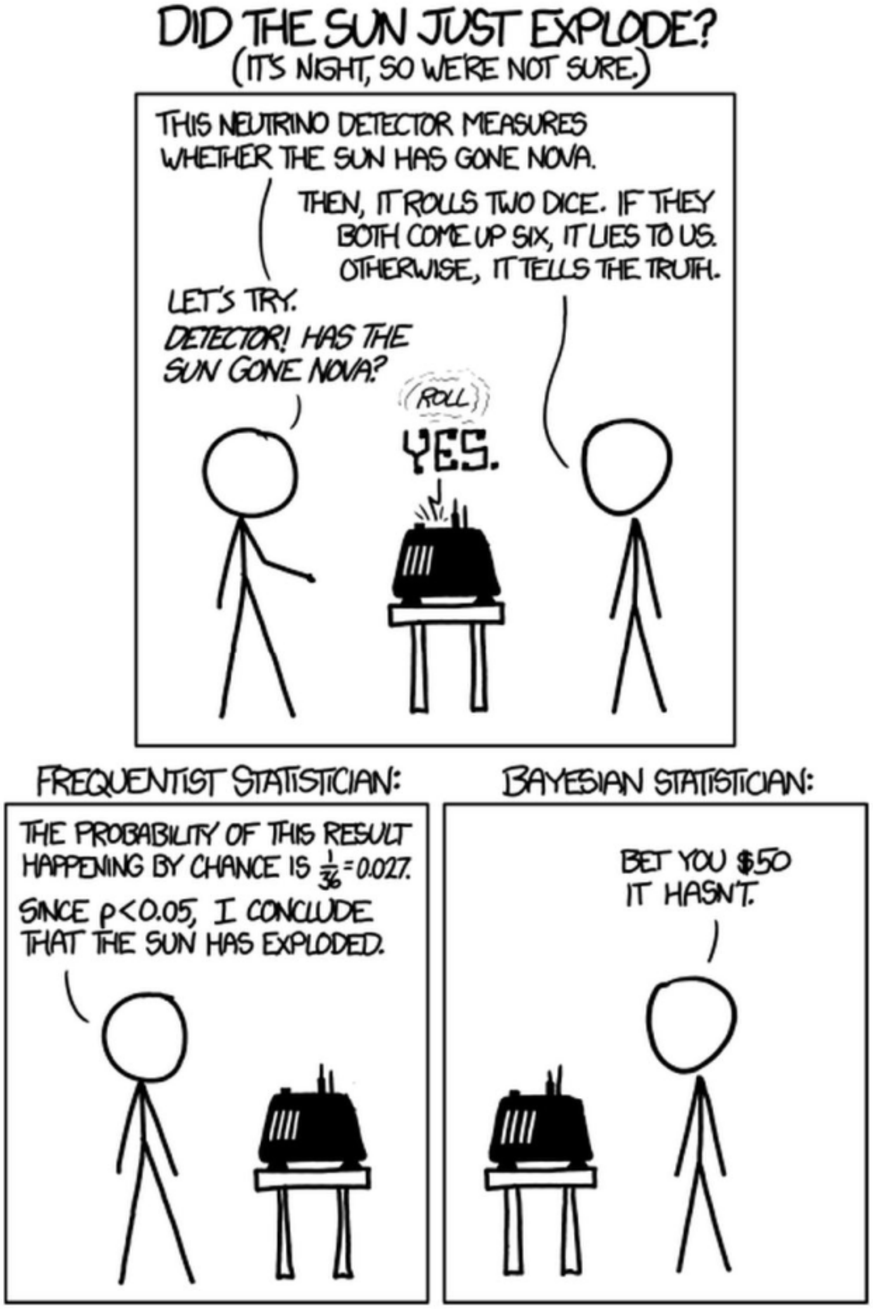
9. [10] If all scientists used 0.05 as a level of ‘alpha’ to decide which results are significant and worthy of publication, that would ensure that only approximately 5% of published results would be wrong and we wouldn’t face the current ‘crisis of reproducibility’ revealing that, in some fields, a very large proportion of published results cannot be reproduced. Is this statement correct? Discuss.

10. [5] If `x` is a list in R discuss the difference between `x[1]` and `x[[1]]`.

11. Consider the following XKCD cartoon:

a) [5] How did the frequentist statistician compute a p-value and what is its interpretation?

b) [5] What relevant probability would the Bayesian statistician use and how did he or she arrive at it?


DID THE SUN JUST EXPLODE?
(IT'S NIGHT, SO WE'RE NOT SURE.)
THIS NEUTRINO DETECTOR MEASURES WHETHER THE SUN HAS GONE NOVA.
THEN, IT ROLLS TWO DICE. IF THEY BOTH COME UP SIX, IT LIES TO US. OTHERWISE, IT TELLS THE TRUTH.
LET'S TRY.
DETECTOR! HAS THE SUN GONE NOVA?
(ROLL)
YES.
FREQUENTIST STATISTICIAN:
THE PROBABILITY OF THIS RESULT HAPPENING BY CHANCE IS 1/36 = 0.027. SINCE p<0.05, I CONCLUDE THAT THE SUN HAS EXPLODED.
BAYESIAN STATISTICIAN:
BET YOU $50 IT HASN'T.

HAVE A GOOD SUMMER