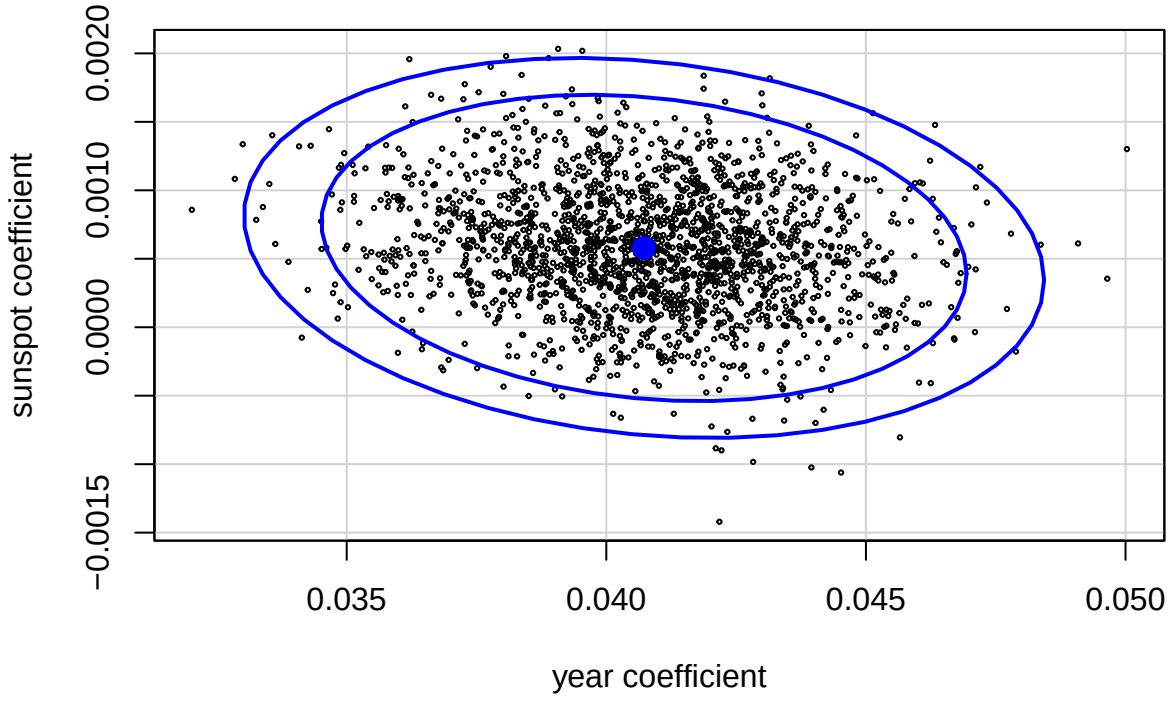
Homework 6

Gianni Spiga

2023-02-24

Problem Set 4

Question 13



ci.B1	0.0357146	0.0457214
ci.B2	-0.0002937	0.0014721

When testing the null hypothesis $H_0 : \beta_i = 0,\ i = 1, 2$ we find that the coefficient for year is significant at $\alpha = 0.05$ and the coefficient for sunspot is not. Our confidence interval for sunspots (β_2) contains zero, which we can see from both the interval and the ellipse. These results line up with our previous ones in Homework 2 when we performed formal tests. In testing for the overall regression effect with bootstrap, we get a p-value of 0.0004, matching with the similar conclusion of problem 7 in problem set 1 as well.

Problem 14

(Intercept)	-9.8093027	-7.0000900
NPreg	0.0603103	0.1860543
PGC	0.0278946	0.0424328
DBP	-0.0235534	-0.0030377
Tricep	-0.0129038	0.0141417
Serum.Insulin	-0.0029581	0.0005747
BMI	0.0601292	0.1192727
Pedigree	0.3588507	1.5315088
Age	-0.0034272	0.0331652

LB	UB
-9.809303	0.0603103
-9.809303	0.0603103
-9.809303	0.0603103
-9.809303	0.0603103
-9.809303	0.0603103
-9.809303	0.0603103
-9.809303	0.0603103
-9.809303	0.0603103
-9.809303	0.0603103

Though some of the confidence intervals for the slopes differ above, we can see that they are very similar to each other. The CI's for the standard MLE approximation is on the left, and the bootstrap intervals are on the right. Regardless of the width of the interval, all of them lead to the same conclusion for rejecting or failing to reject their respective null hypothesis.

Problem Set 5

Question 1

One of the advantages of the log link is that it stabilizes the variance of data with a constant coefficient of variation. By doing so, one could run ordinary least squares on the log-transformed data. However, the intercepts would be biased by the offset -0.5*v where v is the coefficient of variation. The canonical inverse link is just not as practical.

Question 2

a.)

$$E(Y) = E(\mu(\epsilon + 1)) = E(\mu * 1) = \mu Var(Y) = Var(\mu(\epsilon + 1)) = \mu^2 Var(\epsilon) = \mu^2\sigma^2 v = \frac{var(y)}{E(y)^2} = \frac{\mu^2\sigma^2}{\mu^2} = \sigma^2$$

Thus v is constant.

b.)

We could either fit the Gamma regression model to the data, assuming we woul use the log link. However, we could also used the log link and perform an OLS for the data. However, performing this ordinary least suares would be biased by the offset mentioned in question 1.

Question 8

a.)

The GLM which would be suitable is the logistic regression model predicting the binary response Y, using the logit link. We could combine samples from the US and Japan and create a column that identifies whether or not a observation is from the United States or Japan.

b.)

Intuitively, the logit link is the best link of choice since it is the canonical link for logistic regression. If there was a different link choice hypothesized, a goodness-of-link test would be appropriate to help identify the best link.

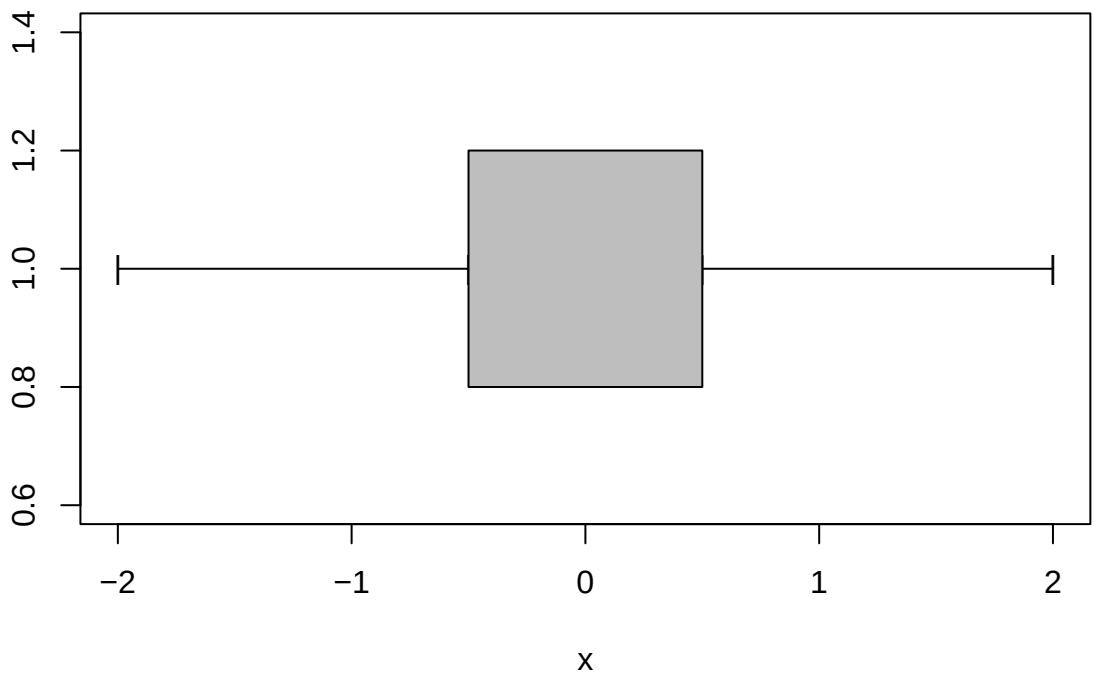
c.)

We would be curious to test that the interaction between whether the woman is from Japan or the US and the variable fat. Since the investigator is curious in testing that the US and Japan are the same, we would test this in our alternative hypothesis. Call the coefficient for the mentioned interaction β_4, we test:

$$H_0 : \beta_4 \neq 0 H_a : \beta_4 = 0$$

However, testing if a value is strictly equal to zero is very challenging to perform. Instead we would need to discuss with the investigator if we could expand the hypothesis into a a small interval centered around zero. This way, we would have a better ability to approximate with the hypothesis.

d.)



Let the grey region be our alternative hypothesis, and all points outside of this region be our null. Our p-value would be the minimum α level needed to include our point in this interval. In this example, the region is from -0.5 to 0.5, but this arbitrary and can be picked by the investigator at any level.

e.)

We could embed a new model with a quadratic term or a another non-linear term. Doing this, we could compare the models via a likelihood ratio test and return the results to the investigator.

f.)

We can imagine the model as the following:

$$E(Y|X) = f(x)$$

We could use a GAM with a non-linear function on $f(x)$ to measure the complex relationship of the probability to the expectation of Y. However, this would make interpretation complex of the probabilities. Another option would be picking a smoothing kernel estimation to understand the relationship between the predictors and the response, in this case fat and age on whether or not one is diagnosed with breast cancer.