Bios 601 H1 V2

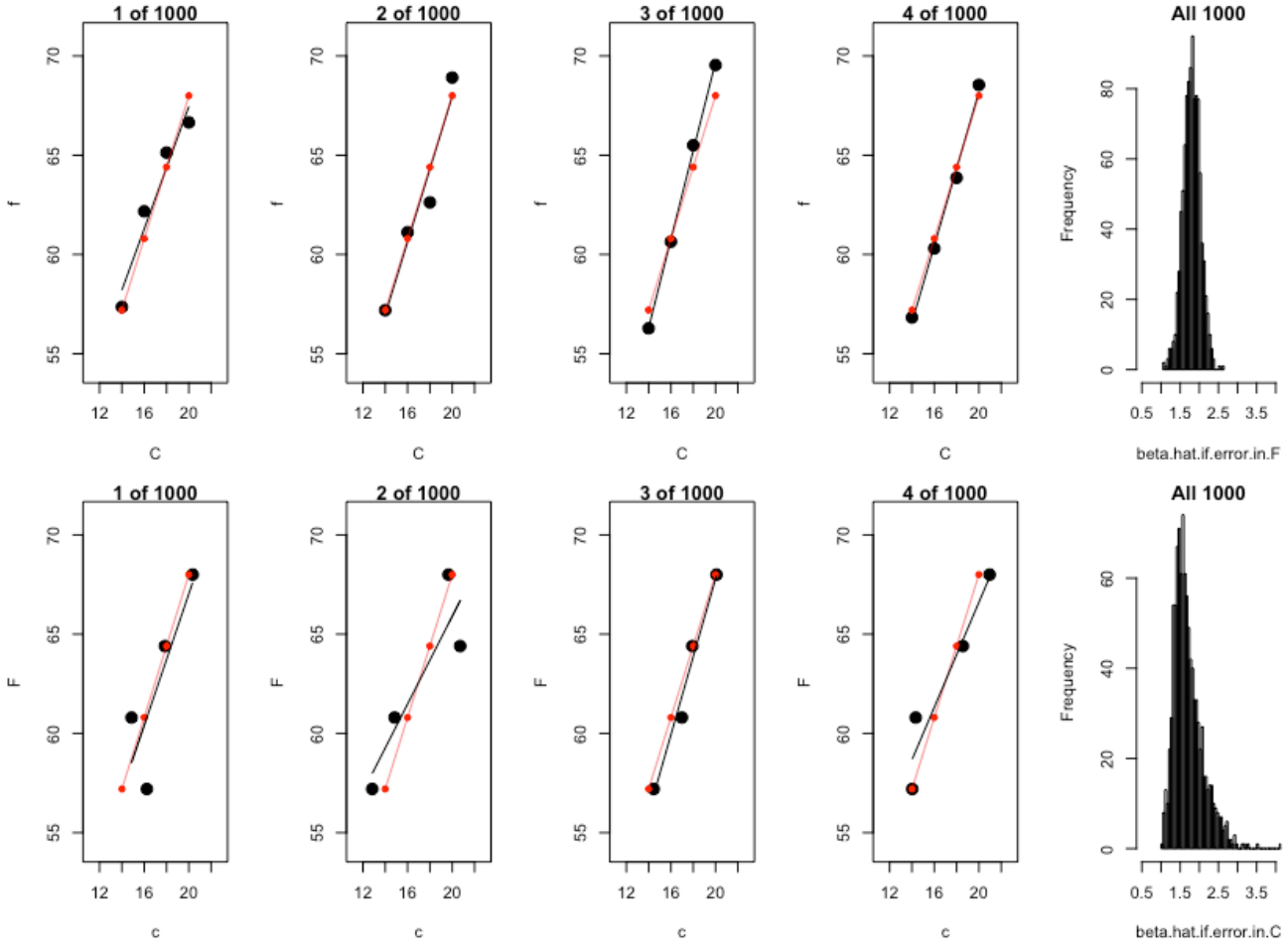
Jiewen Liu 260825295

September 11, 2021

Q8

8.a & 8.b

I follow the codes provided, and the figure produced is given below. The red dots represent the measurements without any errors. In the first row, the there are random errors $\sigma_F \sim N(0,1)$ in data. In the first row, the there are random errors $\sigma_C \sim N(0,1)$ in data.



For errors in measurements of F, we see from the last graph in the first row that the empirical variance and the theoretical one are quite close. We can't reject the null hypothesis. There is no evidence to say our b_1 is biased.

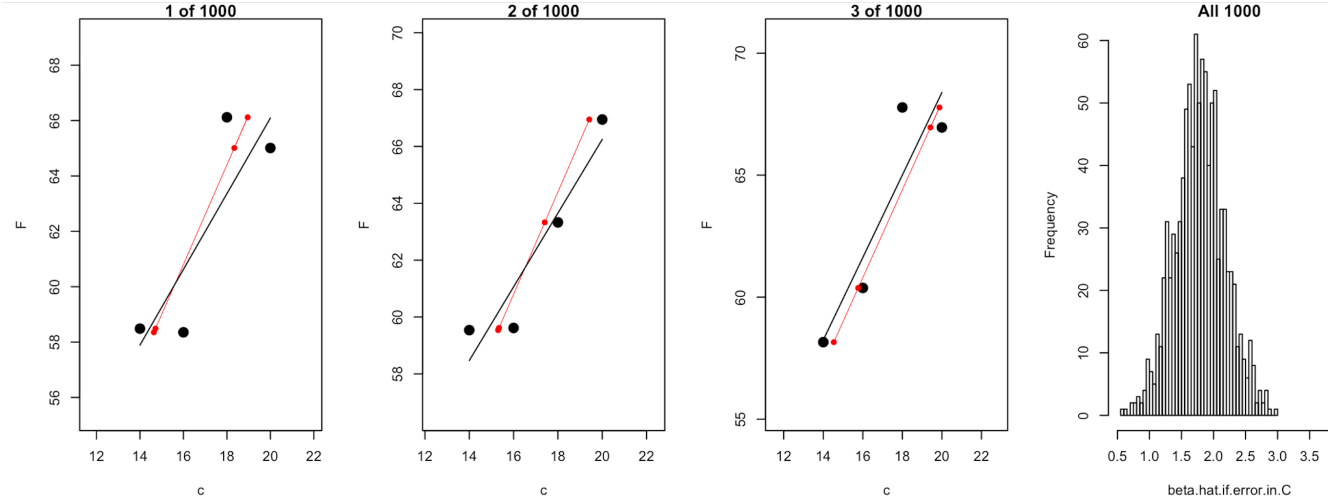
Empirical Variance	0.05
Theoretical Variance	0.05
95% C.I.	$(1.79, 1.81)$

For errors in measurements of C, we see from the last graph in the second row that the empirical variance and the theoretical one are quite different.

Empirical Variance	0.16
Theoretical Variance	0.05
95% C.I.	$(1.69, 1.74)$

8.c

The new figure is generated and shown below.



Empirical Variance	0.15
Theoretical Variance	0.05
95% C.I.	$(1.78, 1.83)$

My findings line up with the predictions in the Notes. The Berkon error doesn't necessarily flatten the curve on average.

Q9

The R plots with the variance of the random error in the measurement ranging from 0 22 is given in the Figure 1.

We can see the results in Figure 2. With error's variance increasing from 0 to 22 (0.1 gap), the red dots are the true theoretical ICCs, blue dots are fitted slopes (we see is decays as true ICC increases), purple dots are predicted attenuation effects / ICCs ($=\frac{\beta}{\hat{\beta}_{Attenuated}} = \frac{1.8}{Fitted\_Slope}$), the green dots are estimatedd de-attenuated ones.

If we recover the $\hat{\beta}$ using the true theoretical ICC from $\hat{\beta}_{Deattenuated} = \frac{\hat{\beta}_{Attenutaed}}{ICC}$, we still see the variance is increased but the expectation seems to be preserved. Quite interesting. It could be possible to get the confidence interval by simulation. A random part in simulation makes things a bit messy.

The flattening of the curve is not warranted, especially in the discreet cases with not many observations and not much variance of measurement error.

Q17

17.a

The pedometer and accelerometers just slightly over-count the number of steps in the our experiment composed of 500 & 1500 steps trials. For smartphone applications, some of them over-count and some of them under-count. There is also distinguishable difference across different applications. However, for wearable devices, they seem to all under-count the number of steps. Some of them even deviate from the true number quite much.

17.b

Wearable devices. The evidence is quite strong. With many ones having large SD, a test will quickly reject the null $H_0 : \mu = \mu_{\text{WearableDevice}}$.

Q20

```
AGE.1999 = AGE.1986+13 -> 35 35 35 35 36 38 39 39 39 40 40 40 40 41 41 41 42 42 43 43 43
43 43 44 44 44 44 45 45 46 46 47 47 48 49 50 51 51 52 55

FIT = lm(AGE.1999~AGE.1986)
FIT$coefficient  -> (Intercept): 13,   AGE.1986:1

age.1986 = AGE.1986 + 5*sample(c(-1,1), 40, replace=TRUE)
fit = lm(AGE.1999~age.1986)
fit$coefficient -> (Intercept): 28.2578332,  age.1986: 0.4877677

fit = lm((AGE.1999-age.1986)~age.1986)
fit$coefficient -> (Intercept): 28.2578332,   age.1986: -0.5122323
```

20.b

Every one ages by 13.

20.c

Those who are originally younger age most, and those who are originally elder age least.

Explanation: For the case of classical error, our target is to find a line that minimizes the sum of squared all points' distances to the it. We see the best line has the slope which bumps up the observed X''s $Y[X']$ when error ϵ is negative. Overestimate it and thus pull it toward the actual X where $X' = X + \epsilon$. The reverse applies when the error term is positive. The best slope suppress it down toward the actual X in order to counter the effect of positive measurement error. Assume the slope is positive if no error. The measured X' falls around true true X. A smaller slope value, i.e. a more flattened curve, is preferred in order to bump up X' if it falls on the left side and suppress it if it falls on the right size. The curve is flattened over and over again as more error introduced.

For algebraic things, we could look back our deduction and simulation results from Q9.

20.d

$AGE.1999 - age.1986$ is how much a person ages in the intervening of 13 years. I regard it as the other way to view who ages more or less. The negative slope is quite intuitive. When classical measurement errors are introduced, those who report low ages are more likely from the younger group, vice versa. The optimal prediction based on MSE needs to give the prediction toward the true Ys generated by their true Xs. Therefore, as the observed age X' increases, the less aging in 13 years intervention is anticipated.

Q24

Just my bold guesses. Not well furnished.

I think we may create pseudo-population and use bootstrapping. For example, assume the name Campbell has 25% probability of being the name of boy and 80% prob of being the name of girl. We could replicate the data value by 100 times. Label 25 of them as male and the remaining 75 as female. By doing so, The 1000 samples are expanded into two groups (roughly 50,000 man and 50,000 women). Do bootstrapping. We could get estimated mean, the variance and thus interval estimator achieved. (Remember to divide the bootstrap variance by extra 2 or ? in our case cuz bootstrap doesn't give you estimated sample variance directly).

Another computational heavy way I could imagine is to iterate all possible "male and female combination" like $\hat{\mu}_{Male} = \sum \frac{\#\text{ of dead males}}{\#ofmales} P(P_1 = "M", P2 =' M' \ldots)$?

Q26

For the case of Berkson error, it is given the error is independent of the observed X' instead of true X. It is equal likely that the observed X' is the overestimation or underestimation of the true X. Consequently, the best slope is the same compared to the case of no error. The true Xs around X' counteracts and thus no extra effect on the expected β.

$$\beta_{1\_\text{BerksonError}} = \frac{Cov(X', Y)}{Var(X')} \tag{1}$$

$$= \frac{Cov(X + \epsilon, Y)}{Var(X')} \tag{2}$$

$$= \frac{Cov(X, Y)}{Var(X')} + \frac{Cov(\epsilon, Y)}{Var(X')} \tag{3}$$

$$= \underbrace{\frac{Cov(X, Y)}{Var(X)}}_{\beta_1} * \frac{Var(X)}{Var(X')} + \frac{Cov(\epsilon, Y)}{Var(X')} \tag{4}$$

$$= \beta_1 * \frac{Var(X)}{Var(X')} + \frac{Cov(\epsilon, \beta_0 + \beta_1 X)}{Var(X')} \tag{5}$$

$$(X = X' - \epsilon) \implies = \beta_1 * \frac{Var(X' - \epsilon)}{Var(X')} + \frac{Cov(\epsilon, \beta_1(X' - \epsilon))}{Var(X')} \tag{6}$$

$$= \beta_1 * \frac{Var(X') + Var(\epsilon) - 2Cov(X', \epsilon)}{Var(X')} + \beta_1 * \frac{Cov(\epsilon, X') - Cov(\epsilon, \epsilon)}{Var(X')} \tag{7}$$

$$(\epsilon \perp X') \implies = \beta_1 * \frac{Var(X') + Var(\epsilon)}{Var(X')} + \beta_1 * \frac{-Var(\epsilon)}{Var(X')} \tag{8}$$

$$= \beta_1 * (\frac{Var(X') + Var(\epsilon) - Var(\epsilon)}{Var(X')}) \tag{9}$$

$$= \beta_1 \tag{10}$$

It's unbiased! Verification through R simulation are given in Figure.3.

```
esti_beta = rep(NA,1000000)
for(i in 1:1000000){
  err = rnorm(4,0,10)
  X = x - err
  Y <- 32 + 1.8*X
  esti_beta[i] = cov(x,Y)/var(x)
}
summary(esti_beta)
    Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
```

```
-16.6255  -0.9109   1.8001   1.8002   4.5139  22.4679
hist(esti_beta,breaks=200)
```

Q27

```
age = NULL
AGE = NULL
half.decades.min = 30
half.decades.max = 70
DA = 5
for(half.decade in seq(half.decades.min,half.decades.max,DA)){
  AGE = c(AGE,rep(half.decade+0:(DA-1),each=DA))
  BP = 100 + 1*AGE
  age = c(age,rep(half.decade+0:(DA-1),DA))
}
plot(age,BP)
points(AGE,BP,col='blue',pch=19)
FIT =lm(BP~AGE)
fit =lm(BP~age)
FIT$coefficients -> 1
fit$coefficients -> 0.9881423
cov(age-AGE,age)
err = AGE-age
print(cov(err,AGE)/var(AGE)) -> 0.01185771
print(1*(1-cov(err,AGE)/var(AGE))) ->  0.9881423
```

27.a

Hi professor James! To save your time of reading it, I'll follow exactly what is shown in class on 5 or 10 age bands. But the same things applies just with some substitution of numbers. Let's first see the algebraic stuffs to understand the expected $\hat{\beta}$ goes down. Recall from what we've done in Q26 and write:

$$\beta_{1\_\text{ShuffledAge}} = \frac{Cov(X', Y)}{Var(X')} \tag{11}$$

$$= \beta_1 * \frac{Var(X') + Var(\epsilon) - 2Cov(X', \epsilon)}{Var(X')} + \beta_1 * \frac{Cov(\epsilon, X') - Cov(\epsilon, \epsilon)}{Var(X')} \tag{12}$$

$$(\epsilon \not\perp \mathbf{X}')!! \implies = \beta_1 * (1 - \frac{Cov(X', \epsilon)}{Var(X')}) \tag{13}$$

$$= \beta_1 * (1 - \frac{Cov(X, \epsilon) + Var(\epsilon)}{Var(X')}) \tag{14}$$

good

In class, if the coarsening is set to 5, we see the slope goes from 1 to 0.9881423. Assume and use discrete uniform random error here. Notice our simulation gives the expected result since we create 5 objects of each age and add 0, 1, 2, 3 and 4 respectively.

The independence between the observed X' and the error ϵ is lost. Every true X is rounded down to its 5 multiples and a uniform random error from $\{0, 1, 2, 3, 4\}$ is added to the $X_{Rounded}$. Total error is

$$\epsilon = \epsilon_{total} = \epsilon_{original} + (X - X_{Rounded}.)$$

In each 5-interval, for example, the coarsened 100 is more likely to be bumped up and 104 is more likely to be lowered down. There exists a covariance of the observed X' sand total errors. So in British's Covid cases, if we know the way of shuffling and the total number of people of each age respectively, we could calculated the expected "flattening" effect by such kind of error introduced. **The covariance which causes the bias in verified above**

using R, we see the "cov(err,age)" pulls things down.

Going back to the blood pressure case with 0.2 slope, definitely we could get expected flattening effect by simulation or directly calculating "cov(err,age)".

27.b

Change the DA = 10. We see the covariance increases and the slope is further flattened. It is not hard to imagine why. The likelihood for true 100 to be bumped up goes up from (5-1)/5 to (10-1)/10, which makes the co-fluctuation more likely.

I think it behaves more like Berkson error. The error measurement is not independent of Y, i.e larger the error indicates higher probability for smaller true X. The level of coarsening is subject to our control.

Q28

28.a

CFR: for VOC+ is 4,000/500,000 = 0.8%, for VOC- is 2,000/400,000 = 0.5%.

28.b

The adjusted CFR: for VOC+ is (4,000-500,000*0.1*0.005)/(500,000*0.9) = 0.83%, for VOC- is still 0.5%. Thus, cthe corrected case fatality ratio is $\frac{0.0083}{0.005} = 1.6$

28.c

Yes. It is close.

28.d

It is closer to the traditional CFR. JH's use of CRF is still based on VOC+ (tested through SGTF). Also authors of the paper adopt the number of positive test results as the denominator. The asymptomatic and undiagnosed subjects are not included, which makes IFR lower, which is also mentioned in the paper.

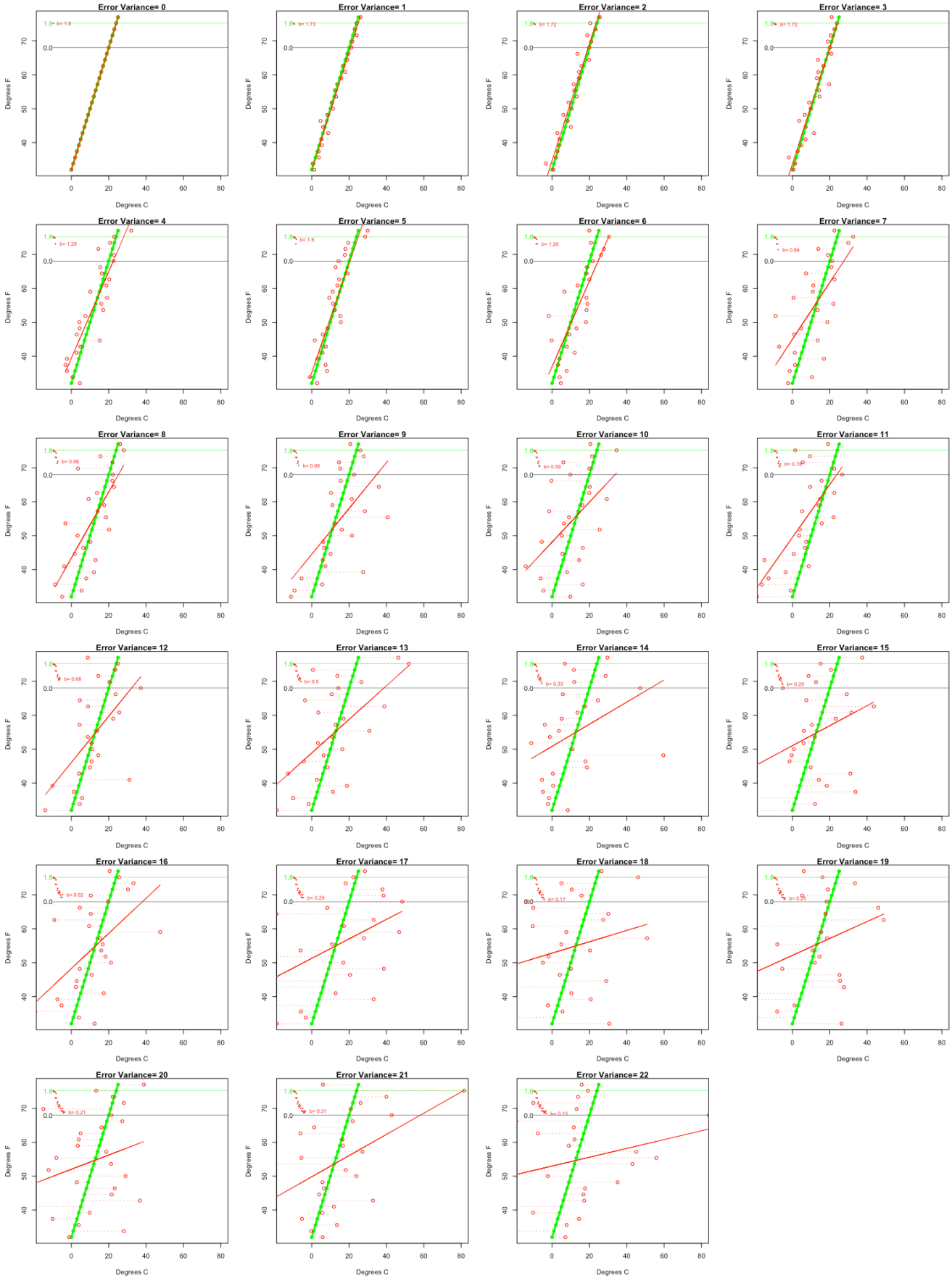


Figure 1: Effect of ICC with Increasing Error Variance

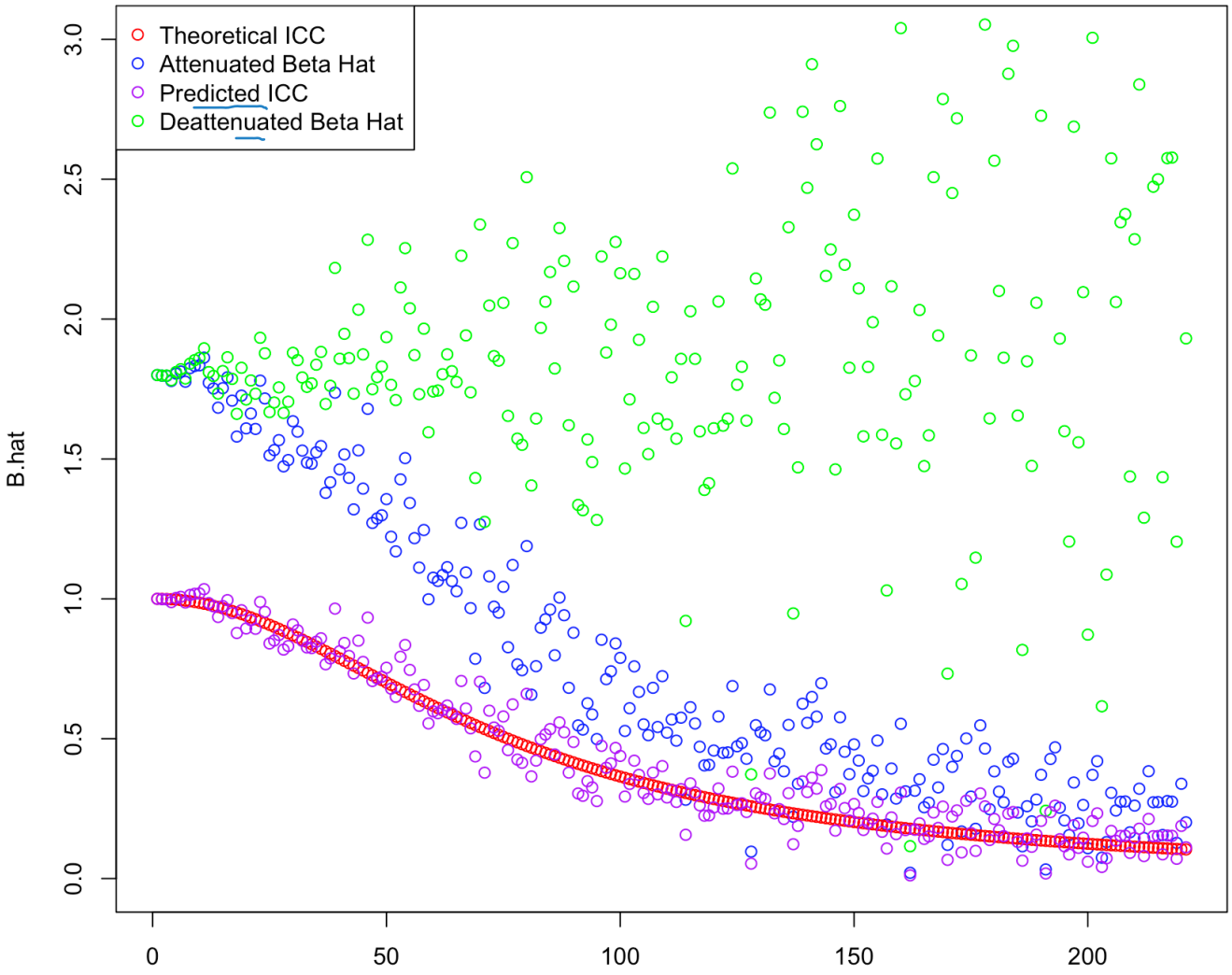


Figure 2: Plotting of Simulation Results

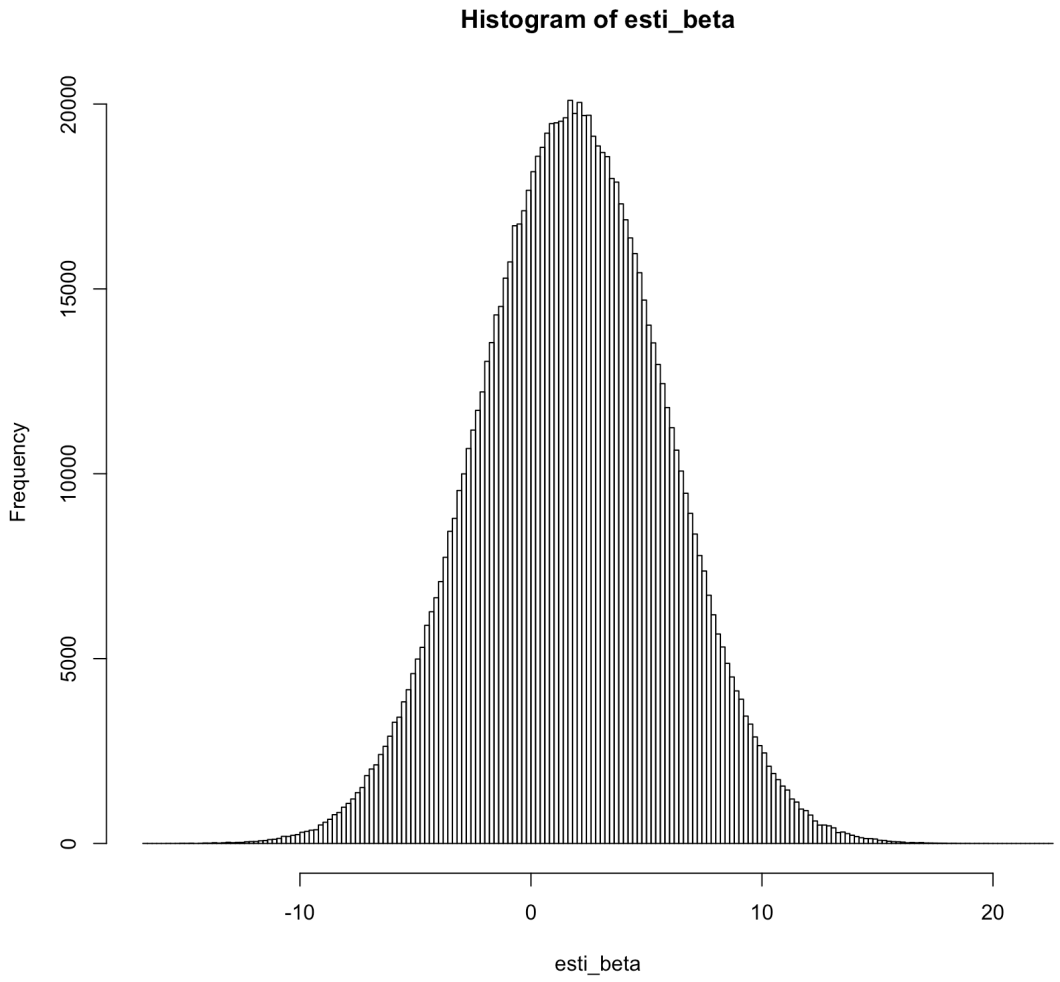


Figure 3: Simulations Results of Fitted Slopes from 100,000 runs

Which type of error ?