ECON312 Problem Set 1B: question 5

Futing Chen, Hongfan Chen, Will Parker

05/14/2020

Contents


```
library(tidyverse)
library(knitr)

library(readxl)
```

Load in data

```
sheets <- excel_sheets("PS1_Q5_Data.xlsx")


dataset_list <- list()

for (s in seq(1:length(sheets))) {
  dataset_list[[s]] <- readxl::read_excel("PS1_Q5_Data.xlsx", sheet = s) %>%
    mutate(dataset_num = s) %>%
    select(Y, X1, X2, dataset_num)
}
```

A: Pre-test estimator

```
sample_params <- function(df){


  dataset_num <- filter(df, row_number() ==1)$dataset_num

  n <- df %>% nrow()

  mu_1 <- mean(df$X1)
```

```
mu_2 <- mean(df$X2)

sigma2_1 <- var(df$X1)
sigma2_2 <- var(df$X2)
rho <- cov(df$X1, df$X2)/sqrt(sigma2_1*sigma2_2)

m_1 <- lm(data = df, formula = formula(Y ~ X1 + X2))

sigma2_epsilon <- mean(m_1$residuals^2)

beta_1_hat <- m_1$coefficients[["X1"]]

beta_1 <- 1
beta_2 <- 1

if (is.na(m_1$coefficients[["X2"]]) == FALSE){
  t_beta_2 <- summary(m_1)$coefficients[["X2", "t value"]]
} else { t_beta_2 <- 0}


m_2 <- lm(data = df, formula = formula(Y ~ X1))

beta_1_tilda <- m_2$coefficients[["X1"]]


if (abs(t_beta_2) > 1.964) {
  beta_1_star <- beta_1_hat

  Q_xx <- matrix(nrow =2, c(sigma2_1, sqrt(sigma2_1*sigma2_2)*rho, sqrt(sigma2_1*sigma2_2)*rho, sigma2

  std_err_beta_1_star <- sqrt(sigma2_epsilon*solve(Q_xx)[[1,1]]/n)

  analytic_bias <- 0
}  else {
  beta_1_star <- beta_1_tilda

  analytic_bias <- (1)*(rho*sqrt(sigma2_2/sigma2_1))
  std_err_beta_1_star <- sqrt((1/n)*((beta_2^2*(1-rho^2)*sigma2_2)/sigma2_1 + sigma2_epsilon/sigma2_1)
}



output <- tibble(dataset_num,
                 mu_1,
                 mu_2,
                 sigma2_1,
                 sigma2_2,
                 rho,
                 sigma2_epsilon,
                 t_beta_2,
                 beta_1_hat,
                 beta_1_tilda,
                 beta_1_star,
```

```
                   std_err_beta_1_star,
                   analytic_bias) %>%
    mutate(empiric_bias = beta_1_star -beta_1)

  return(output)
}
```

Test that function is working

```
summary(lm("Y~ X1 + X2", dataset_list[[2]]))
```

```
##
## Call:
## lm(formula = "Y~ X1 + X2", data = dataset_list[[2]])
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.2904 -0.6041 -0.0148  0.5814  2.1644
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.97843    0.88558   1.105   0.2720
## X1           0.95717    0.08796  10.882   <2e-16 ***
## X2           0.17665    0.08065   2.190   0.0309 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.8824 on 97 degrees of freedom
## Multiple R-squared:  0.564,  Adjusted R-squared:  0.555
## F-statistic: 62.75 on 2 and 97 DF,  p-value: < 2.2e-16
```

```
summary(lm("Y~ X1", dataset_list[[2]]))
```

```
##
## Call:
## lm(formula = "Y~ X1", data = dataset_list[[2]])
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.28018 -0.66271 -0.08795  0.53837  2.14983
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.90289    0.90189   1.001    0.319
## X1           0.96544    0.08956  10.779   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.8993 on 98 degrees of freedom
## Multiple R-squared:  0.5425, Adjusted R-squared:  0.5378
## F-statistic: 116.2 on 1 and 98 DF,  p-value: < 2.2e-16
```

```r
sample_params(dataset_list[[2]])
```

```
## # A tibble: 1 x 14
##   dataset_num  mu_1   mu_2 sigma2_1 sigma2_2    rho sigma2_epsilon t_beta_2
##         <int> <dbl>  <dbl>    <dbl>    <dbl>  <dbl>          <dbl>    <dbl>
## 1           2  10.0 0.0412     1.02     1.21 0.0429          0.755     2.19
## # ... with 6 more variables: beta_1_hat <dbl>, beta_1_tilda <dbl>,
## #   beta_1_star <dbl>, std_err_beta_1_star <dbl>, analytic_bias <dbl>,
## #   empiric_bias <dbl>
```

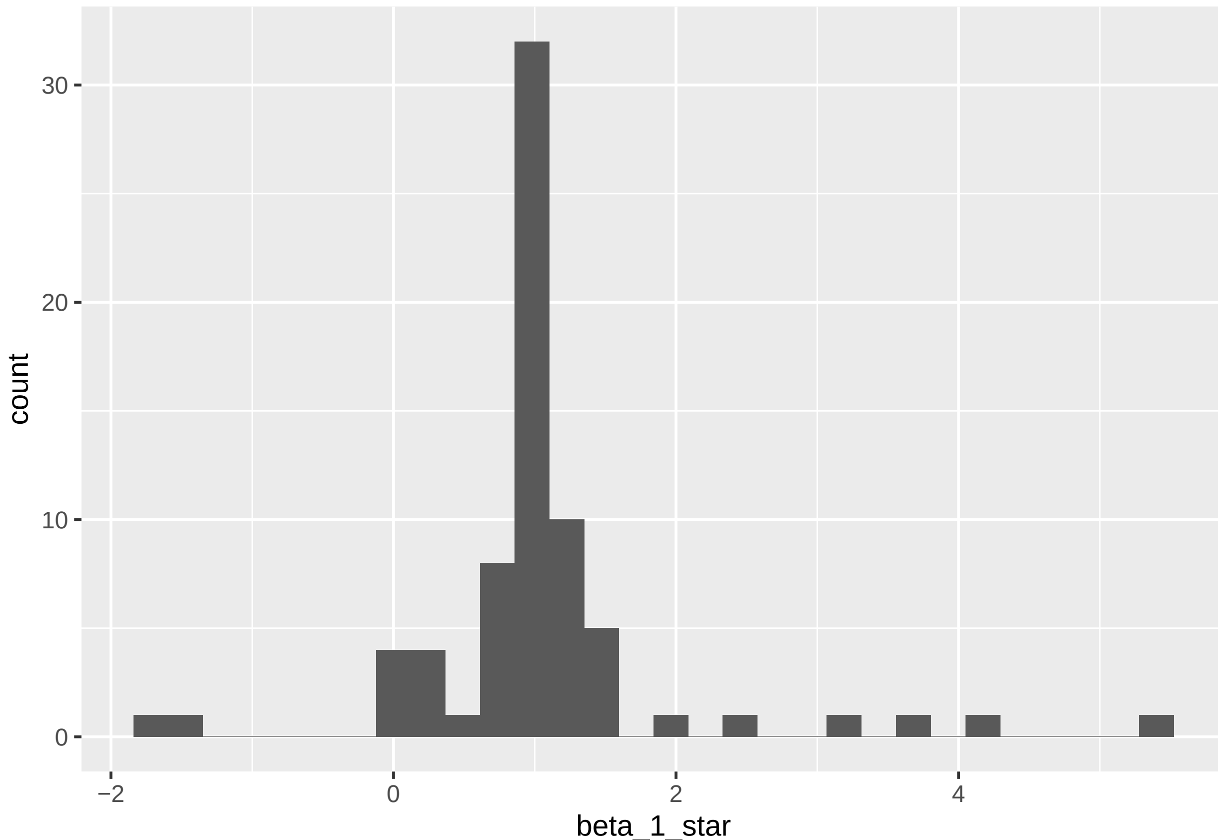
```r
results <- map_dfr(dataset_list, sample_params)

results %>%
  kable(col.names = c("Dataset", "$\\mu_1$", "$\\mu_2$", "$\\sigma^2_1$", "$\\sigma_2^2$", "$\\rho$", "$
```

Dataset	μ_1	μ_2	σ_1^2	σ_2^2	ρ	σ_e^2	t_{β_2}	$\hat{\beta}_1$	$\tilde{\beta}_1$	β_1^*	$se(\beta_1^*)$	analytic	empiric
1	0.10	-0.16	1.18	1.00	-0.14	0.75	0.38	0.92	0.91	0.91	0.12	-0.13	-0.09
2	10.02	0.04	1.02	1.21	0.04	0.76	2.19	0.96	0.97	0.96	0.09	0.00	-0.04
3	0.00	-0.10	0.10	0.93	-0.05	0.84	-0.44	1.58	1.59	1.59	0.41	-0.14	0.59
4	10.02	-0.04	0.09	1.01	-0.04	0.80	1.09	0.98	0.97	0.97	0.45	-0.12	-0.03
5	-0.57	-0.07	9.47	1.09	0.12	0.83	1.00	0.94	0.94	0.94	0.04	0.04	-0.06
6	9.27	-0.02	8.14	1.60	-0.02	1.01	0.32	1.02	1.02	1.02	0.06	-0.01	0.02
7	0.03	-0.19	0.92	0.88	0.01	9.51	-0.63	1.18	1.17	1.17	0.34	0.01	0.17
8	9.93	-0.01	0.95	0.98	0.01	9.90	-2.00	1.24	1.24	1.24	0.32	0.00	0.24
9	-0.02	0.11	0.10	1.08	-0.04	10.86	-1.37	0.72	0.78	0.78	1.09	-0.13	-0.22
10	10.01	-0.09	0.12	0.86	0.09	9.52	0.59	0.13	0.18	0.18	0.91	0.24	-0.82
11	0.08	0.09	9.61	1.05	-0.01	9.38	-0.03	1.00	1.00	1.00	0.10	0.00	0.00
12	9.89	0.05	11.27	0.82	-0.02	9.89	1.47	1.06	1.06	1.06	0.10	-0.01	0.06
13	0.08	0.02	1.11	1.10	0.02	102.48	-0.78	1.00	0.99	0.99	0.96	0.02	-0.01
14	10.15	0.13	1.10	1.25	0.04	134.33	-0.26	2.44	2.43	2.43	1.11	0.05	1.43
15	0.04	0.03	0.10	1.07	-0.03	92.24	0.58	5.45	5.40	5.40	3.06	-0.10	4.40
16	9.99	-0.12	0.09	1.38	-0.05	99.28	-0.67	0.25	0.36	0.36	3.42	-0.19	-0.64
17	0.63	-0.08	10.44	0.89	0.00	110.50	-0.88	1.08	1.08	1.08	0.33	0.00	0.08
18	10.69	0.02	9.17	1.35	-0.12	105.48	0.11	0.52	0.51	0.51	0.34	-0.05	-0.49
19	-0.06	-0.06	1.18	1.18	1.00	1.29	0.00	1.08	1.08	1.08	0.10	1.00	0.08
20	9.92	-0.08	1.14	1.14	1.00	1.48	0.00	0.88	0.88	0.88	0.11	1.00	-0.12
21	0.04	0.14	0.09	0.95	1.00	0.98	0.00	0.73	0.73	0.73	0.32	3.16	-0.27
22	10.03	0.09	0.08	0.85	1.00	0.98	0.00	1.09	1.09	1.09	0.34	3.16	0.09
23	0.11	0.04	11.04	1.10	1.00	1.09	0.00	0.97	0.97	0.97	0.03	0.32	-0.03
24	9.60	-0.13	11.22	1.12	1.00	0.86	0.00	0.98	0.98	0.98	0.03	0.32	-0.02
25	0.01	0.01	1.03	1.03	1.00	9.34	0.00	1.15	1.15	1.15	0.30	1.00	0.15
26	9.97	-0.03	1.14	1.14	1.00	9.13	0.00	1.03	1.03	1.03	0.28	1.00	0.03
27	-0.01	-0.04	0.12	1.19	1.00	11.85	0.00	-0.02	-0.02	-0.02	1.00	3.16	-1.02
28	9.94	-0.19	0.13	1.32	1.00	10.15	0.00	0.87	0.87	0.87	0.88	3.16	-0.13
29	0.08	0.02	9.13	0.91	1.00	8.70	0.00	0.94	0.94	0.94	0.10	0.32	-0.06
30	9.75	-0.08	6.65	0.66	1.00	7.76	0.00	1.06	1.06	1.06	0.11	0.32	0.06
31	0.08	0.08	0.92	0.92	1.00	107.84	0.00	1.01	1.01	1.01	1.08	1.00	0.01
32	10.10	0.10	0.75	0.75	1.00	96.66	0.00	-1.39	-1.39	-1.39	1.13	1.00	-2.39
33	-0.05	-0.17	0.09	0.86	1.00	104.87	0.00	1.12	1.12	1.12	3.48	3.16	0.12
34	9.96	-0.12	0.10	1.00	1.00	102.50	0.00	0.06	0.06	0.06	3.21	3.16	-0.94
35	-0.80	-0.25	9.99	1.00	1.00	102.70	0.00	1.03	1.03	1.03	0.32	0.32	0.03
36	10.06	0.02	9.56	0.96	1.00	82.33	0.00	0.82	0.82	0.82	0.29	0.32	-0.18
37	0.08	0.13	1.21	1.05	0.54	1.09	0.71	0.97	1.02	1.02	0.12	0.50	0.02
38	10.10	0.11	1.17	1.15	0.65	0.92	-1.41	1.12	1.01	1.01	0.12	0.64	0.01

Dataset	μ_1	μ_2	σ_1^2	σ_2^2	ρ	σ_e^2	t_{β_2}	$\hat{\beta}_1$	$\tilde{\beta}_1$	β_1^*	$se(\beta_1^*)$	analytic	empiric
39	0.02	0.08	0.13	1.23	0.64	1.01	0.81	0.52	0.71	0.71	0.37	2.01	-0.29
40	10.01	0.03	0.10	1.02	0.59	0.82	-0.27	0.90	0.84	0.84	0.38	1.83	-0.16
41	0.32	-0.18	9.49	0.82	0.47	1.06	0.44	0.95	0.96	0.96	0.04	0.14	-0.04
42	10.24	-0.02	9.36	0.79	0.39	1.14	-1.04	1.01	1.00	1.00	0.04	0.11	0.00
43	0.14	0.03	0.95	0.97	0.47	9.38	-1.03	1.30	1.12	1.12	0.33	0.48	0.12
44	9.91	0.06	1.10	0.96	0.55	6.45	-1.47	1.53	1.28	1.28	0.25	0.52	0.28
45	0.03	0.15	0.10	1.02	0.47	9.51	-0.29	1.43	1.27	1.27	1.01	1.50	0.27
46	10.01	0.03	0.10	0.99	0.62	9.48	0.01	0.20	0.21	0.21	1.01	1.94	-0.79
47	0.06	-0.04	9.65	0.84	0.49	11.00	1.17	0.98	1.05	1.05	0.11	0.14	0.05
48	9.55	-0.10	11.89	1.32	0.64	10.18	0.39	1.05	1.08	1.08	0.10	0.21	0.08
49	0.00	0.16	0.89	0.96	0.51	112.69	-0.15	0.13	0.03	0.03	1.13	0.53	-0.97
50	10.13	0.02	1.04	0.69	0.45	79.51	0.71	-0.29	0.02	0.02	0.88	0.36	-0.98
51	0.04	0.01	0.09	1.16	0.37	80.31	-0.40	1.92	1.43	1.43	2.99	1.33	0.43
52	10.00	-0.16	0.09	1.00	0.33	134.49	-0.50	4.77	4.07	4.07	3.97	1.11	3.07
53	-0.31	-0.12	7.32	0.91	0.39	109.05	-1.17	0.88	0.69	0.69	0.39	0.14	-0.31
54	10.38	0.09	11.73	1.06	0.47	104.18	0.07	1.23	1.24	1.24	0.30	0.14	0.24
55	0.02	-0.04	0.76	0.86	-0.52	0.90	1.92	1.18	1.05	1.05	0.14	-0.55	0.05
56	10.01	-0.06	1.00	0.82	-0.48	1.04	-0.67	1.11	1.15	1.15	0.13	-0.43	0.15
57	0.04	-0.08	0.11	1.12	-0.53	0.92	2.05	1.59	1.20	1.59	0.35	0.00	0.59
58	10.02	-0.13	0.11	1.00	-0.52	0.76	0.22	1.58	1.55	1.55	0.37	-1.57	0.55
59	0.05	-0.08	11.28	1.16	-0.44	0.91	-0.51	0.98	0.99	0.99	0.04	-0.14	-0.01
60	10.13	0.00	10.28	0.88	-0.45	1.07	-0.30	1.00	1.00	1.00	0.04	-0.13	0.00
61	-0.13	0.10	0.86	0.74	-0.37	10.73	-0.18	0.21	0.24	0.24	0.36	-0.34	-0.76
62	10.05	-0.11	0.87	0.82	-0.48	10.22	-0.09	0.75	0.77	0.77	0.35	-0.47	-0.23
63	0.08	-0.11	0.11	0.80	-0.45	9.32	-1.68	2.94	3.72	3.72	0.95	-1.20	2.72
64	9.99	-0.02	0.10	1.03	-0.39	13.47	-0.44	1.74	1.97	1.97	1.23	-1.30	0.97
65	0.43	-0.05	8.14	0.93	-0.47	8.02	-0.32	0.91	0.93	0.93	0.10	-0.16	-0.07
66	10.34	0.06	9.43	0.98	-0.52	9.48	-0.99	1.01	1.07	1.07	0.10	-0.17	0.07
67	-0.12	0.10	1.30	1.08	-0.66	99.34	-0.25	1.16	1.36	1.36	0.88	-0.60	0.36
68	10.12	0.00	1.12	0.87	-0.48	86.29	-1.89	-0.24	0.68	0.68	0.88	-0.42	-0.32
69	-0.06	0.15	0.10	0.87	-0.54	96.46	0.10	-1.51	-1.72	-1.72	3.17	-1.61	-2.72
70	10.01	-0.14	0.11	1.02	-0.47	89.17	1.33	5.16	3.08	3.08	2.92	-1.45	2.08
71	-0.31	0.21	8.23	0.99	-0.55	101.82	0.60	1.28	1.14	1.14	0.35	-0.19	0.14
72	10.10	-0.03	10.83	0.98	-0.59	89.17	0.48	1.02	0.92	0.92	0.29	-0.18	-0.08

Distribution of β_1^* across the 72 samples

```
results %>%
  ggplot(aes(x  = beta_1_star)) +
  geom_histogram()
```



Sampling distribution for the pre-test estimator

If $|t|_{\hat{\beta}_2} > 1.96$, then β_1^* has the typical OLS asymptomic variance, i.e. for $\hat{\beta} = (\hat{\beta}_1, \hat{\beta}_2)$

$$\sqrt{n}(\hat{\beta}_n - \beta) \xrightarrow{d} N(0, \sigma_\epsilon^2 * E[X'X]^{-1})$$

In terms of the model parameters, we can write

$$E[X'X] = (\begin{matrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{matrix})$$

$$\sqrt{N}(\hat{\beta} - \beta) \ \xrightarrow{d} \ \mathcal{N}\left(0, \frac{1}{(1-\rho^2)\sigma_1^2\sigma_2^2}\begin{pmatrix} \sigma_2^2 & -\rho\sigma_1\sigma_2 \\ -\rho\sigma_1\sigma_2 & \sigma_1^2 \end{pmatrix}\sigma_\varepsilon^2\right). \tag{1}$$

Thus, we obtain that

$$\hat{\beta}_1 \sim \mathcal{N}\left(\beta_1, \frac{\sigma_\varepsilon^2}{N(1-\rho^2)\sigma_1^2}\right) \tag{2}$$

If $|t|_{\hat{\beta}_2} \leq 1.96$, then asymptomic variance of $\tilde{\beta}_1$ is more complex

Note that since $Var(X_1X_2) = (1-\rho^2)\sigma_1^2\sigma_2^2$. We obtain that

$$
\begin{aligned}
Var(\tilde{\beta}_1) &= \frac{1}{N} Var(X_1)^{-1} Var(X_1(X_2\beta_2 + \varepsilon)) Var(X_1)^{-1} \\
&= \frac{1}{N} Var(X_1)^{-1} \Big[Var(X_1 X_2 \beta_2) + Var(X_1 \varepsilon) \Big] Var(X_1)^{-1} \\
&= \frac{1}{N} \left[\frac{\beta_2^2 (1-\rho^2)\sigma_2^2}{\sigma_1^2} + \frac{\sigma_\varepsilon^2}{\sigma_1^2} \right]. \qquad (3)
\end{aligned}
$$

Thus, we obtain that

$$
\tilde{\beta}_1 \sim \mathcal{N}\left(\beta_1 + \frac{\rho\sigma_2}{\sigma_1}\beta_2, \frac{1}{N}\left[\frac{\beta_2^2(1-\rho^2)\sigma_2^2}{\sigma_1^2} + \frac{\sigma_\varepsilon^2}{\sigma_1^2} \right] \right) \qquad (4)
$$

B

Analytic Bias of β_1^*

If $|t|_{\hat{\beta}_2} > 1.96$, then $\beta_1^* = \hat{\beta}_1$ which is unbiased, i.e.

$$
E[\hat{\beta}_1] = \beta_1
$$

If $|t|_{\hat{\beta}_2} \leq 1.96$, then $\beta_1^* = \tilde{\beta}_1$ which has the standard missing variable bias

$$
E[\tilde{\beta}_1] = \beta_1 + \beta_2 \frac{Cov(X_1, X_2)}{Var(X_1)}
$$

based on the data geneterating process we know

$$
\begin{aligned}
Cov(X_1, X_2) &= \rho\sigma_1\sigma_2 \\
Var(X_1) &= \sigma_1^2
\end{aligned}
$$

So by solving we have the bias

$$
E[\tilde{\beta}_1] = \beta_1 + \beta_2 \frac{\rho\sigma_2}{\sigma_1}
$$

So then the $E[\beta_1^*]$

$$
E[\beta_1^*] = P(|t|_{\hat{\beta}_2} > 1.96) * \beta_1 + P(|t|_{\hat{\beta}_2} \leq 1.96) * (\beta_1 + \beta_2 \frac{\rho\sigma_2}{\sigma_1})
$$

The expected bias is then

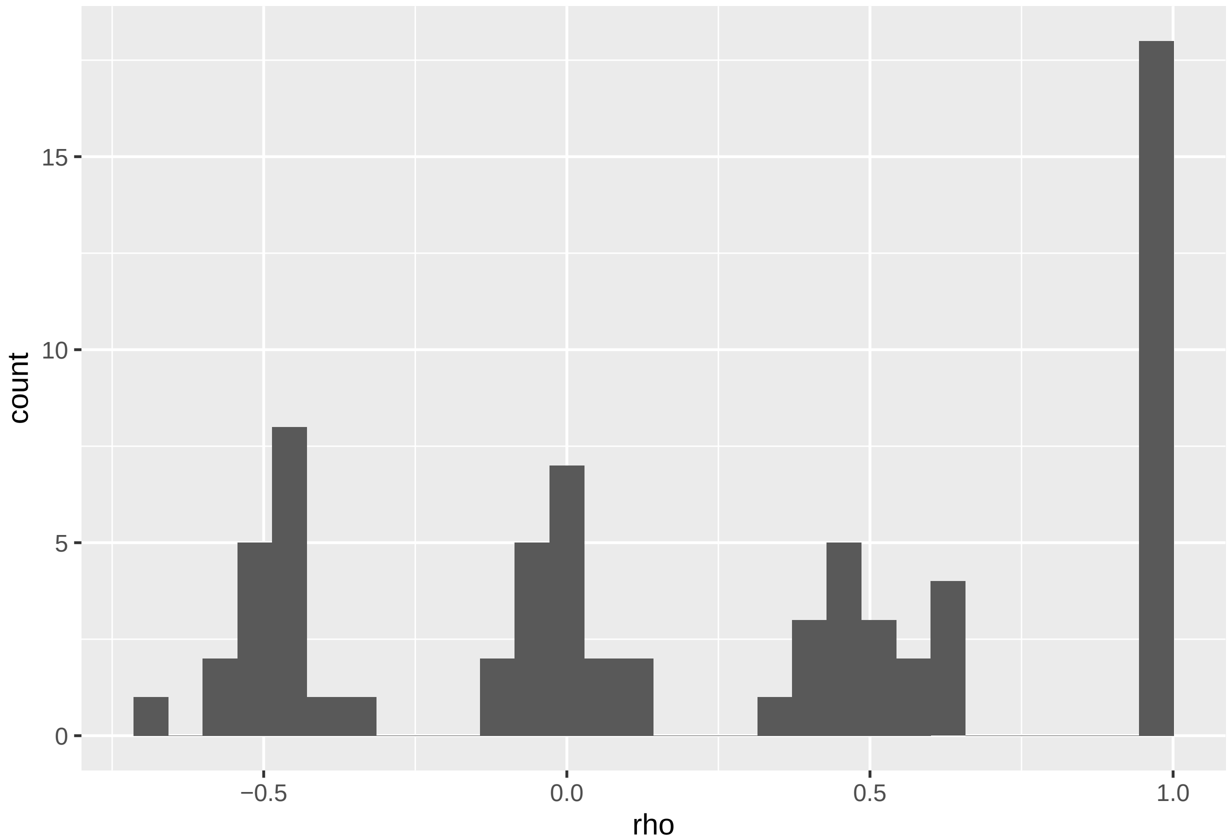
$$
E[\beta_1^* - \beta_1] = P(|t|_{\hat{\beta}_2} \leq 1.96) * (\beta_2 \frac{\rho\sigma_2}{\sigma_1})
$$

Relationship of parameters to observed bias

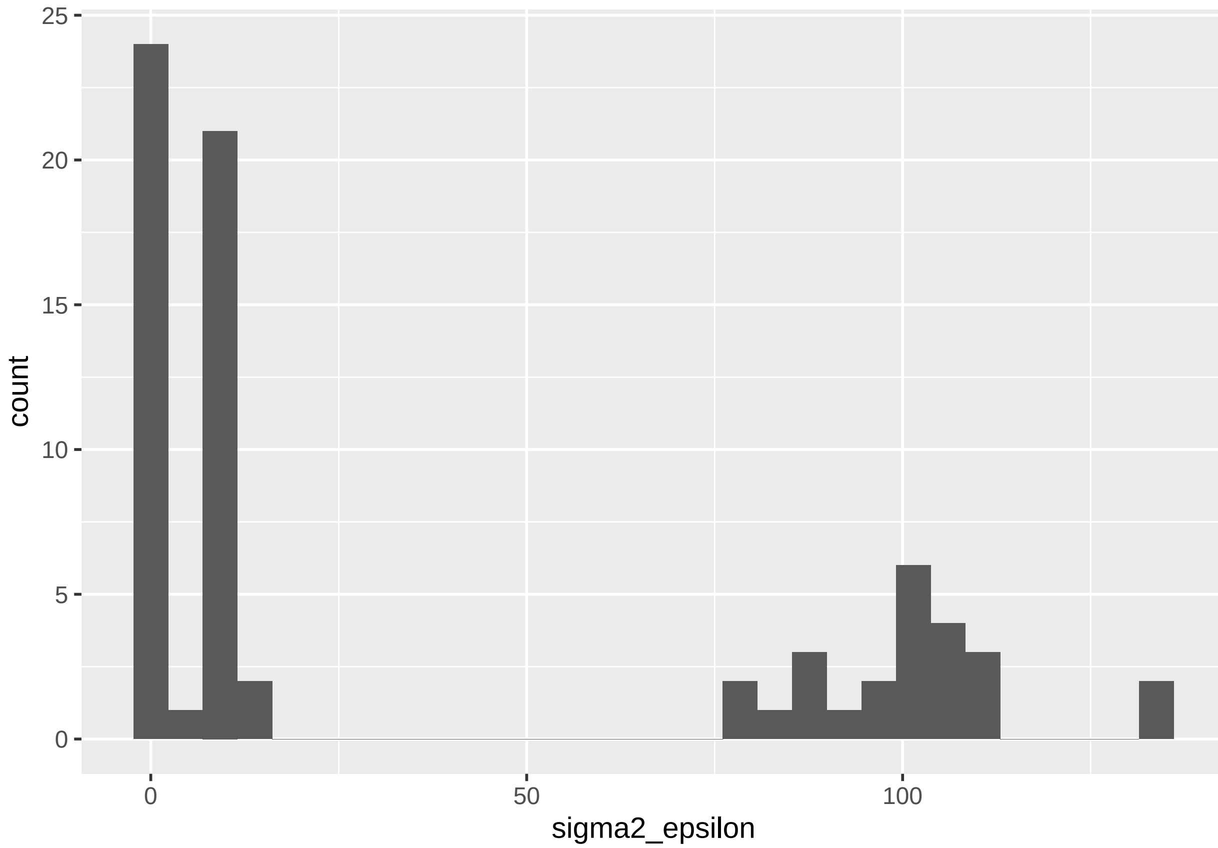
Based on our dervied expressions, higher values of σ_ϵ^2, σ_2^2, and ρ should be correlated with higher bias of the pre-test estimator. Lower values of σ_1^2
are correlated with lower bias of the pre-test estimator. We made several plots to illustrate this.

Parameter distribution in the datasets

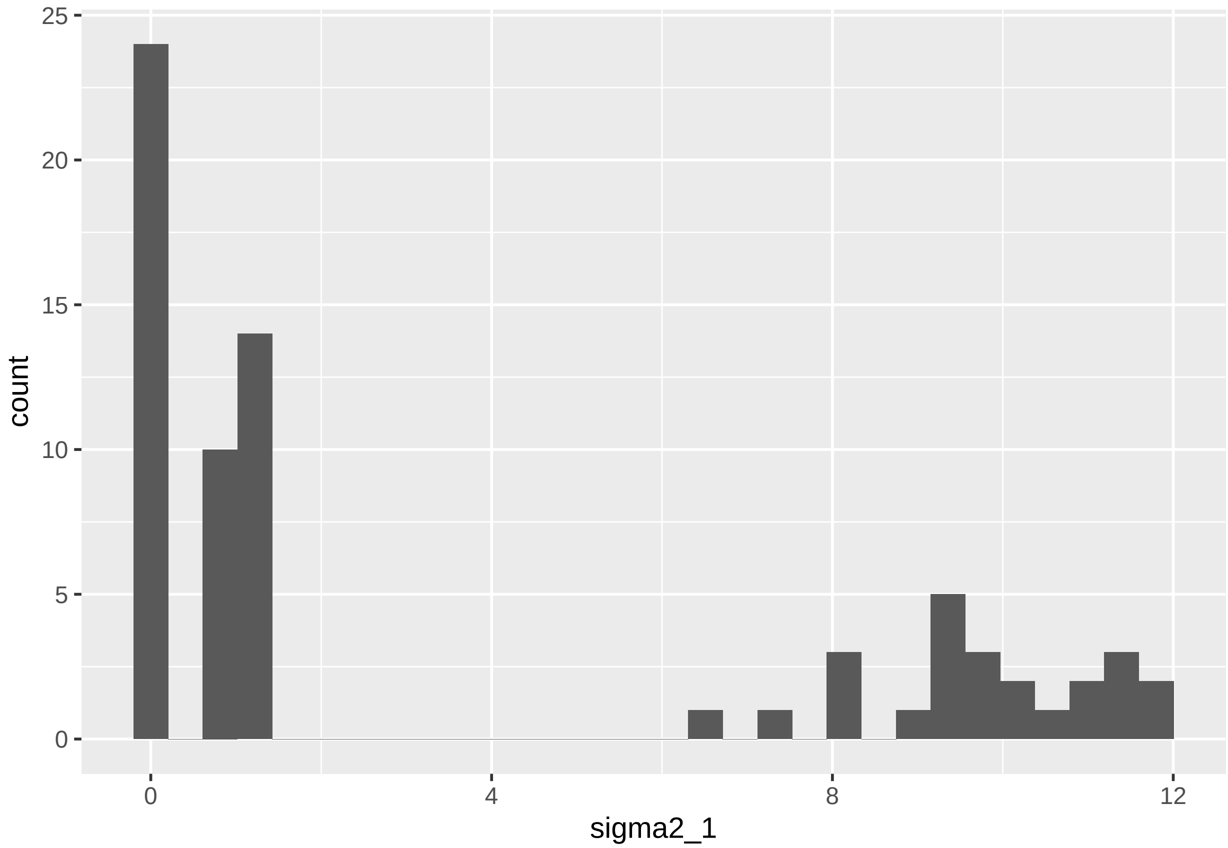
```
results %>%
  ggplot(aes(x = rho)) +
  geom_histogram()
```



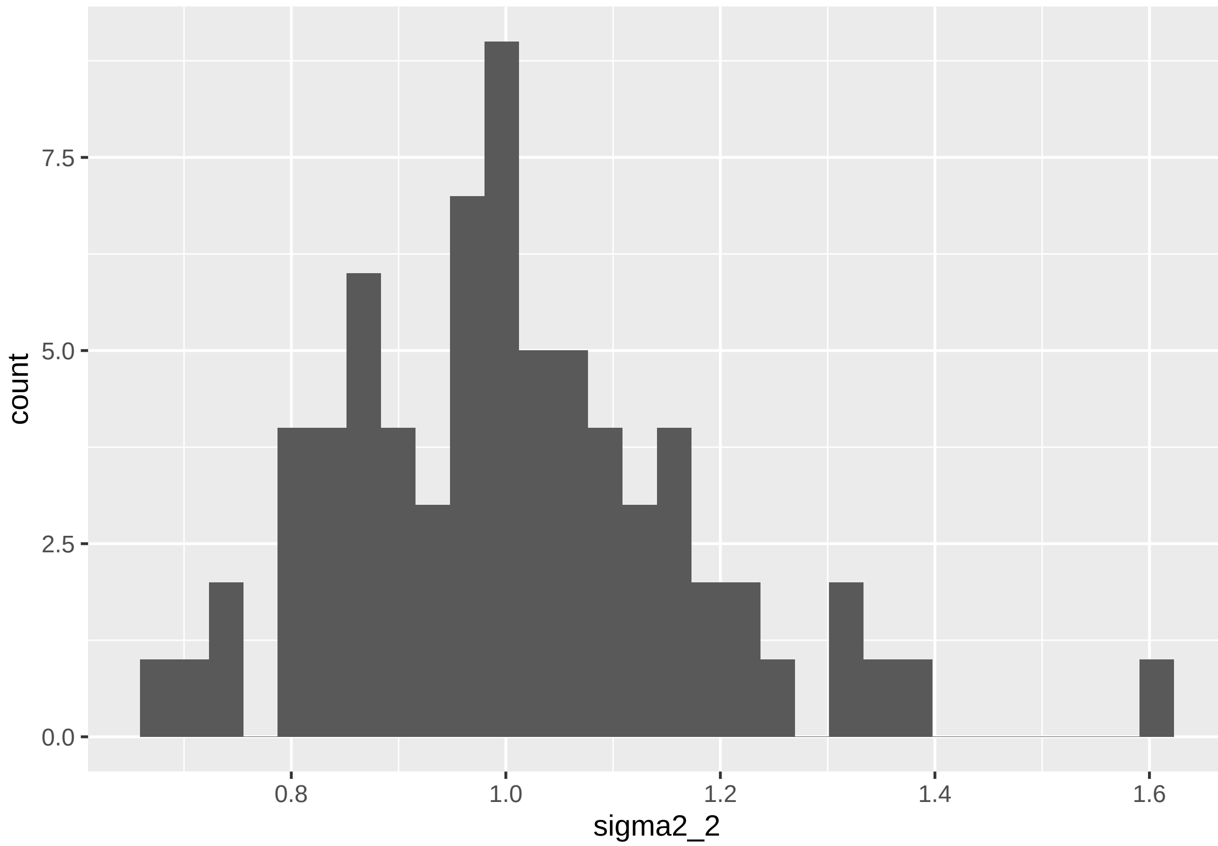

```
results %>%
  ggplot(aes(x = sigma2_epsilon)) +
  geom_histogram()
```



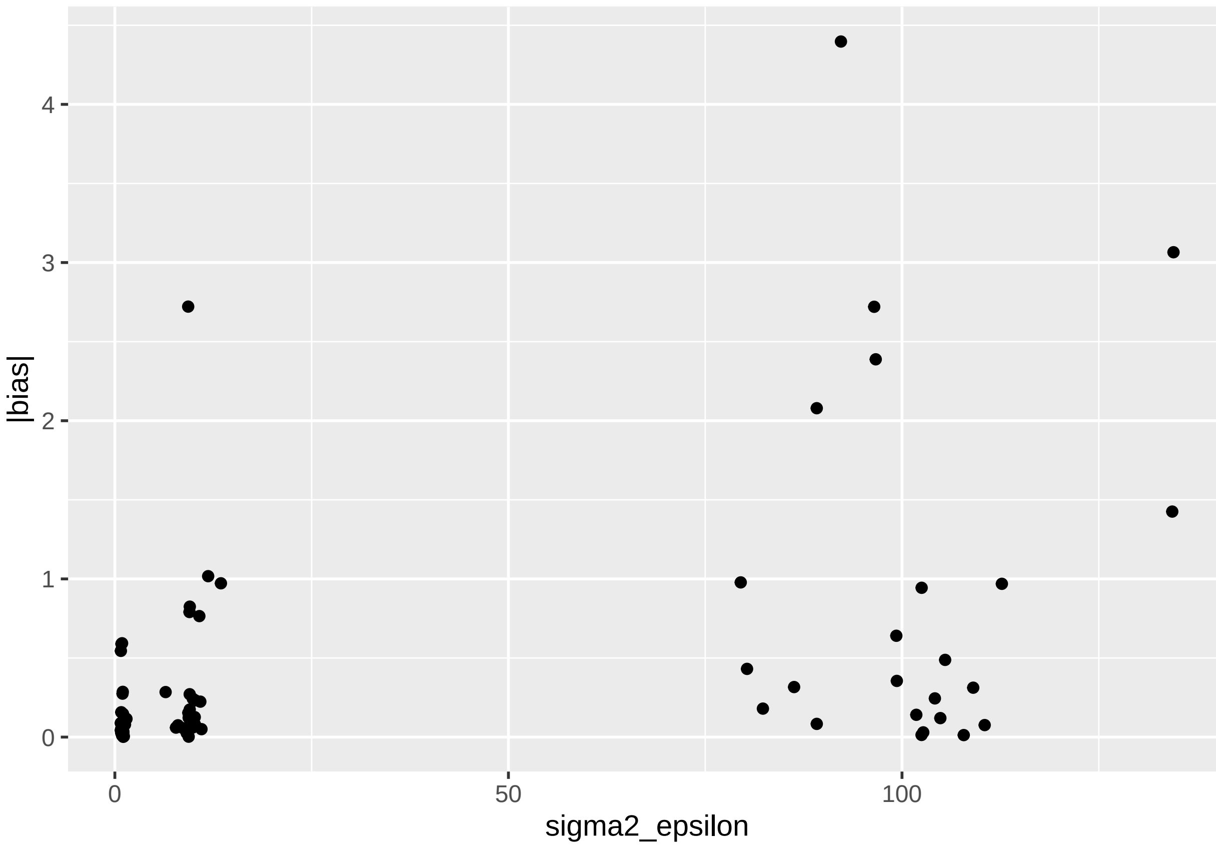

```r
results %>%
  ggplot(aes(x = sigma2_1)) +
  geom_histogram()
```




```
results %>%
  ggplot(aes(x = sigma2_2)) +
  geom_histogram()
```

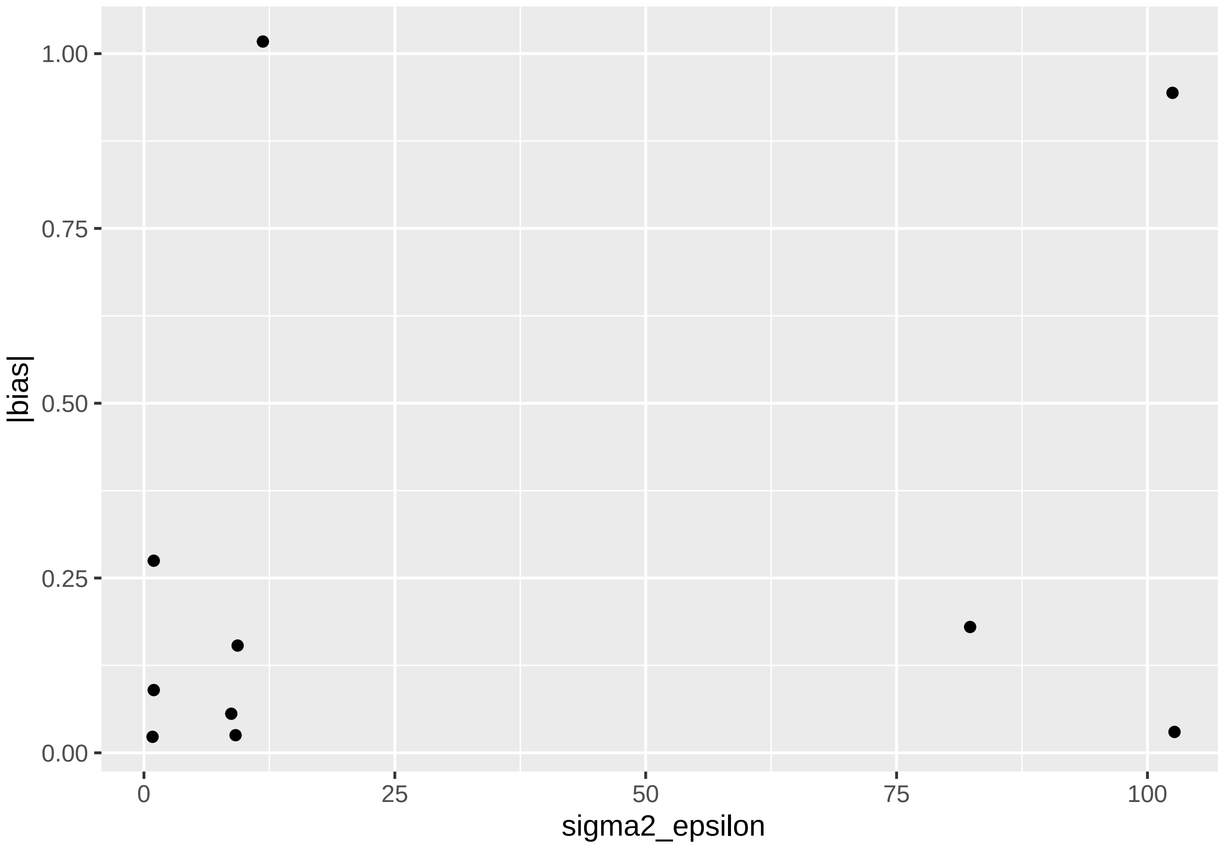



```r
results %>%
  ggplot(aes(x = sigma2_epsilon, y = abs(empiric_bias))) +
  geom_point() + labs(x = "sigma2_epsilon", y = "|bias|")
```



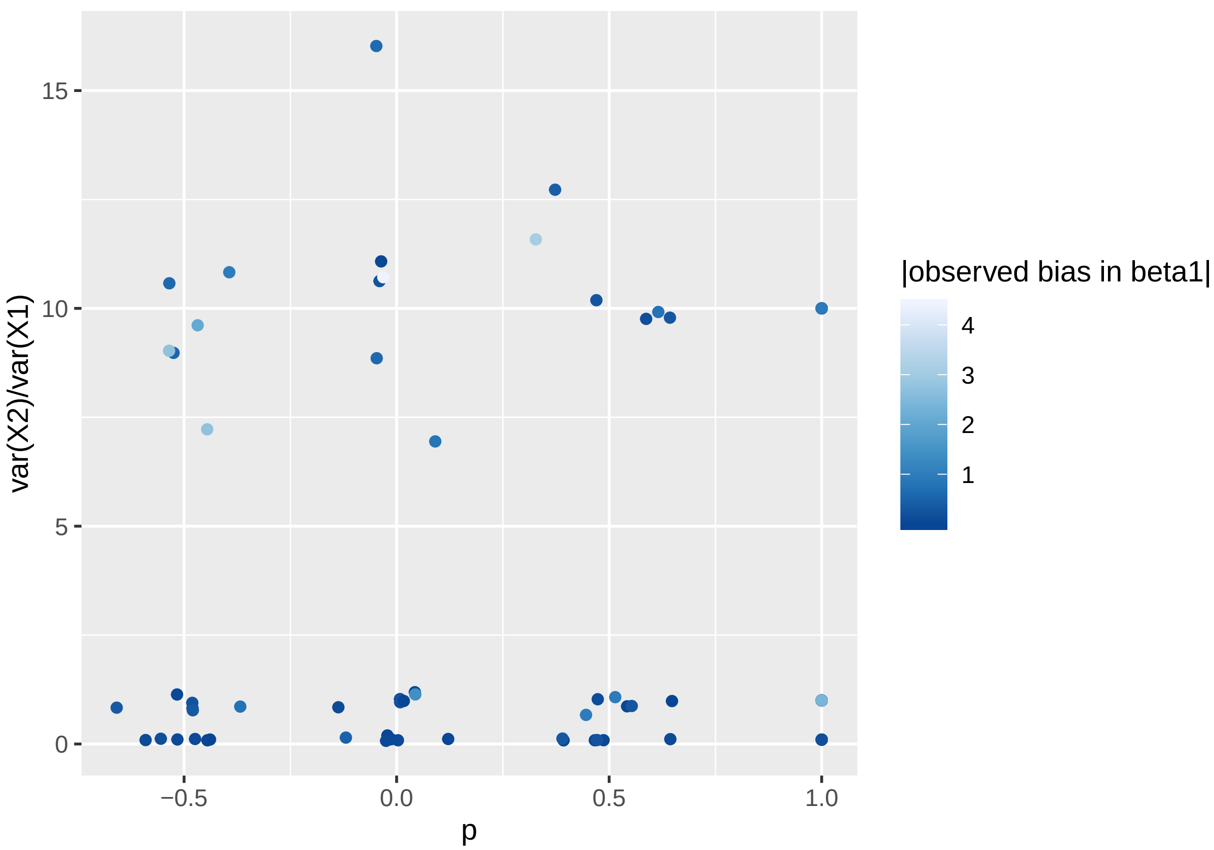
fixing $\rho = 1$ and estimating bias as a function of σ^2_ϵ

```
results %>%
  filter(rho ==1) %>%
  ggplot(aes(x = sigma2_epsilon, y = abs(empiric_bias))) +
  geom_point() + labs(x = "sigma2_epsilon", y = "|bias|")
```



fixing $\rho = 1$ and estimating bias as a function of σ^2_ϵ

```
results %>%
  ggplot(aes(x = rho, y = sigma2_2/sigma2_1,  color = abs(empiric_bias))) +
  geom_point() + labs(x = "p", y = "var(X2)/var(X1)", color = "|observed bias in beta1|") +
  scale_color_distiller()
```



C: Bayesian approach

A bayesian would assume a prior distribution for $\theta = (\beta_0, \beta_1, \beta_2)$, e.g.

$$P(\theta) = N(0, \Sigma)$$

Then compute the posterior distribution of $P(\theta|(\mathbf{X}, \mathbf{Y}))$ via bayes formula

$$P(\theta|(\mathbf{X}, \mathbf{Y})) = \frac{1}{Z} f(\theta|(\mathbf{X}, \mathbf{Y})) * P(\theta)$$

Where

$$Z = \int_X \int_Y f(\theta|(\mathbf{X}, \mathbf{Y})) dY dX$$

Then the Bayesian could do testing of any specific hypothesis on β_1 or β_2 with corresponding posterior marginal probability distribution, e.g. for the hypothesis that β_1 is greater than 1

$$P(\beta_1 > 1) = \int_1^\infty P(\beta_1|(\mathbf{X}, \mathbf{Y})) dX_1$$