ST 503 Homework 3

Eric Warren

2024-02-02

Contents

1 Problem 1

Using the `teengamb` data, fit a model with gamble as the response and the other variables as predictors.

```
# install.packages(c("faraway", "tidyverse")) # Uncomment if you do not have packages
library(faraway)
library(tidyverse)
(
  gambling <- as_tibble(teengamb) %>%
    mutate(sex = ifelse(sex == 0, "Male",
                        ifelse(sex == 1, "Female", "Error")))
)
```

```
## # A tibble: 47 x 5
##    sex    status income verbal gamble
##    <chr>   <int>  <dbl>  <int>  <dbl>
##  1 Female     51   2          8   0
##  2 Female     28   2.5        8   0
##  3 Female     37   2          6   0
##  4 Female     28   7          4   7.3
```

```
##  5 Female     65   2         8  19.6
##  6 Female     61   3.47      6   0.1
##  7 Female     28   5.5       7   1.45
##  8 Female     27   6.42      5   6.6
##  9 Female     43   2         6   1.7
## 10 Female     18   6         7   0.1
## # i 37 more rows
```

```
# Check to make sure the number of "Error" values is 0; want it to be TRUE
nrow(gambling %>% filter(sex == "Error")) == 0
```

```
## [1] TRUE
```

```
# Fit the model
gamblingFit <- lm(gamble ~ ., gambling)
```

1.1 Part A

Which variables are statistically significant?

We can tell that a predictor is statistically significant by looking at its corresponding p-value for its partial slope. If the p-value is less than equal some α value then we say it is statistically significant. In this case we will say $\alpha = 0.05$ so if the p-value for a predictor's partial slope is less than or equal to 0.05 then we can say it is statistically significant.

```
summary(gamblingFit)
```

```
##
## Call:
## lm(formula = gamble ~ ., data = gambling)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -51.082 -11.320  -1.451   9.452  94.252
##
## Coefficients:
##             Estimate Std. Error t value  Pr(>|t|)
## (Intercept)  0.43732   14.75429   0.030    0.9765
## sexMale     22.11833    8.21111   2.694    0.0101 *
## status       0.05223    0.28111   0.186    0.8535
## income       4.96198    1.02539   4.839 0.0000179 ***
## verbal      -2.95949    2.17215  -1.362    0.1803
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 22.69 on 42 degrees of freedom
## Multiple R-squared:  0.5267, Adjusted R-squared:  0.4816
## F-statistic: 11.69 on 4 and 42 DF,  p-value: 0.000001815
```

As we can see, the two predictors that show this quality are `sex` and `income`.

1.2 Part B

What interpretation should be given to the coefficient for `sex`?

A nice thing in `R` when you are using a factor like `sex` is that it can give you an easy interpretation of the predictor. By looking at our summary in **Part A**, we can see that it says "sexMale" with an estimate of 22.1183301. A reasonable interpretation that we can make is that if the sex of a teen gambler in Britain is a male then we expect (or predict) that they will gamble an additional 22.1183301 pounds in a year (as opposed to their female counterparts).

1.3 Part C

Fit a model with just income as a predictor and use an F-test to compare it to the full model.

Here we are going to see if we fit a model with just `income` if it predicts statistically as well as the full model we made. If the resulting p-value is less than equal some α value then we say that it is statistically significant to conclude that the "full model" is a better model to use than the reduced (as in at least one predictor in the full model not in the reduced model is significant – partial slope is not zero). In this case we will say $\alpha = 0.05$ so if the p-value for a predictor's partial slope is less than or equal to 0.05 then we can say it is statistically significant.

```
# Make reduced model
incomeGamblingFit <- lm(gamble ~ income, gambling)

# Do ANOVA test to see if we can reduce model
anova(incomeGamblingFit, gamblingFit)
```

```
## Analysis of Variance Table
##
## Model 1: gamble ~ income
## Model 2: gamble ~ sex + status + income + verbal
##   Res.Df   RSS Df Sum of Sq      F  Pr(>F)
## 1     45 28009
## 2     42 21624  3    6384.8 4.1338 0.01177 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

We can see from our F-test that the resulting p-value is 0.0117721 which means that we have statistically significant evidence to conclude that at least one predictor in the full model not in the reduced model is significant and that the "full model" is a better model to use than the reduced.

2 Problem 2

Here we are going to use the `sat` data.

```
(sat <- tibble(sat))
```

```
## # A tibble: 50 x 7
##    expend ratio salary takers verbal  math total
##     <dbl> <dbl>  <dbl>  <int>  <int> <int> <int>
##  1   4.40  17.2   31.1      8    491   538  1029
```

```
##  2   8.96  17.6   48.0     47    445   489   934
##  3   4.78  19.3   32.2     27    448   496   944
##  4   4.46  17.1   28.9      6    482   523  1005
##  5   4.99  24     41.1     45    417   485   902
##  6   5.44  18.4   34.6     29    462   518   980
##  7   8.82  14.4   50.0     81    431   477   908
##  8   7.03  16.6   39.1     68    429   468   897
##  9   5.72  19.1   32.6     48    420   469   889
## 10   5.19  16.3   32.3     65    406   448   854
## # i 40 more rows
```

2.1 Part A

Fit a model with total sat score as the response and expend, ratio and salary as predictors. Test the hypothesis that $\beta_{Salary} = 0$. Test the hypothesis that $\beta_{salary} = \beta_{ratio} = \beta_{expend} = 0$. Do any of these predictors have an effect on the response?

First let us take a look at the hypothesis that $\beta_{Salary} = 0$. We are going to fit our model and see if our resulting p-value is less than equal some α value then we say that it is statistically significant to conclude that salary is a significant predictor (beta is not equal to 0). In this case we will say $\alpha = 0.05$ so if the p-value for a predictor's partial slope is less than or equal to 0.05 then we can say it is statistically significant.

```
# Fit the model
satFit <- lm(total ~ salary + ratio + expend, sat)

# Check the predictors
summary(satFit)
```

```
##
## Call:
## lm(formula = total ~ salary + ratio + expend, data = sat)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -140.911  -46.740   -7.535   47.966  123.329
##
## Coefficients:
##             Estimate Std. Error t value         Pr(>|t|)
## (Intercept) 1069.234    110.925   9.639 0.00000000000129 ***
## salary        -8.823      4.697  -1.878           0.0667 .
## ratio          6.330      6.542   0.968           0.3383
## expend        16.469     22.050   0.747           0.4589
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 68.65 on 46 degrees of freedom
## Multiple R-squared:  0.2096, Adjusted R-squared:  0.1581
## F-statistic: 4.066 on 3 and 46 DF,  p-value: 0.01209
```

Looking at our summary , we can see that the p-value for `salary` is about 0.0667 which is greater than 0.05. So we do not have significant evidence that $\beta_{Salary} \neq 0$ and instead have statistical evidence to think that $\beta_{Salary} = 0$ is a possibility since we failed to reject our null hypothesis.

Now we are going to look at our hypothesis that $\beta_{salary} = \beta_{ratio} = \beta_{expend} = 0$. To do this, we are going to look at our full model of all the variables and our reduced model of the intercept and see what p-value we get from the corresponding F-test. If the p-value we get is less than or equal to 0.05 then we can say it is statistically significant.

```
# Make intercept only model
interceptOnlySAT <- lm(total ~ 1, sat)

# Do F-test
anova(interceptOnlySAT, satFit)
```

```
## Analysis of Variance Table
##
## Model 1: total ~ 1
## Model 2: total ~ salary + ratio + expend
##   Res.Df    RSS Df Sum of Sq      F  Pr(>F)
## 1     49 274308
## 2     46 216812  3     57496 4.0662 0.01209 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

We can see from our F-test that the resulting p-value is 0.0120861 (less than 0.05) which means that we have statistically significant evidence to conclude that at least one predictor in the full model not in the reduced (intercept only) model is significant and that the "full model" is a better model to use than the reduced. Therefore, our hypothesis that $\beta_{salary} = \beta_{ratio} = \beta_{expend} = 0$ is proven to be statistically significantly false (rejected).

Lastly, we will answer: Do any of these predictors have an effect on the response? In this case, we are going to see if our resulting p-value is less than equal some α value then we say that predictor is statistically significant (beta is not equal to 0). In this case we will say $\alpha = 0.05$ so if the p-value for a predictor's partial slope is less than or equal to 0.05 then we can say it is statistically significant.

```
summary(satFit)
```

```
##
## Call:
## lm(formula = total ~ salary + ratio + expend, data = sat)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -140.911  -46.740   -7.535   47.966  123.329
##
## Coefficients:
##             Estimate Std. Error t value         Pr(>|t|)
## (Intercept) 1069.234    110.925   9.639 0.00000000000129 ***
## salary        -8.823      4.697  -1.878           0.0667 .
## ratio          6.330      6.542   0.968           0.3383
## expend        16.469     22.050   0.747           0.4589
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 68.65 on 46 degrees of freedom
## Multiple R-squared:  0.2096, Adjusted R-squared:  0.1581
## F-statistic: 4.066 on 3 and 46 DF,  p-value: 0.01209
```

However, when we look at each predictor individually, we get conflicting results with our F-test. That is, with our F-test we showed that at least one predictor was statistically significant (its beta value is not 0); however, when looking at it individually we can see that none of them are statistically significant (and all have a possibility to have its partial slope be zero). What I would do is remove one variable at a time starting with the highest p-value and making a new model and check partial slopes until either all the remaining predictor(s) have a p-value less than or equal to 0.05 or we have an intercept-only model (no predictors) left. My guess is with the `salary` variable so close to 0.05, it could eventually become significant. Hint: This happens – see below:

```
summary(lm(total ~ salary, sat))
```

```
##
## Call:
## lm(formula = total ~ salary, data = sat)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -147.125  -45.354    4.073   42.193  125.279
##
## Coefficients:
##             Estimate Std. Error t value             Pr(>|t|)
## (Intercept) 1158.859     57.659  20.098 < 0.0000000000000002 ***
## salary        -5.540      1.632  -3.394              0.00139 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 67.89 on 48 degrees of freedom
## Multiple R-squared:  0.1935, Adjusted R-squared:  0.1767
## F-statistic: 11.52 on 1 and 48 DF,  p-value: 0.001391
```

2.2 Part B

Now add takers to the model. Test the hypothesis that $\beta_{salary} = 0$. Compare this model to the previous one using an F-test. Are the F-test and t-test here equivalent?

First we will make the new model. Let us take a look at the hypothesis that $\beta_{Salary} = 0$. We are going to fit our model and see if our resulting p-value is less than equal some α value then we say that it is statistically significant to conclude that salary is a significant predictor (beta is not equal to 0). In this case we will say $\alpha = 0.05$ so if the p-value for a predictor's partial slope is less than or equal to 0.05 then we can say it is statistically significant.

```
# Fit the model
satFitUpdated <- lm(total ~ salary + ratio + expend + takers, sat)

# Check the predictors
summary(satFitUpdated)
```

```
##
## Call:
## lm(formula = total ~ salary + ratio + expend + takers, data = sat)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
```

```
## -90.531 -20.855  -1.746  15.979  66.571
##
## Coefficients:
##              Estimate Std. Error t value             Pr(>|t|)
## (Intercept) 1045.9715    52.8698  19.784 < 0.0000000000000002 ***
## salary         1.6379     2.3872   0.686                0.496
## ratio         -3.6242     3.2154  -1.127                0.266
## expend         4.4626    10.5465   0.423                0.674
## takers        -2.9045     0.2313 -12.559 0.000000000000000261 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 32.7 on 45 degrees of freedom
## Multiple R-squared:  0.8246, Adjusted R-squared:  0.809
## F-statistic: 52.88 on 4 and 45 DF,  p-value: < 0.00000000000000022
```

Looking at our summary , we can see that the p-value for `salary` is about 0.496 which is greater than 0.05. So we do not have significant evidence that $\beta_{Salary} \neq 0$ and instead have statistical evidence to think that $\beta_{Salary} = 0$ is a possibility since we failed to reject our null hypothesis.

Now we are going to compare it to our previous model without taker. Here we are going to do a F-test to see if $\beta_{Takers} = 0$.

```
anova(satFit, satFitUpdated)
```

```
## Analysis of Variance Table
##
## Model 1: total ~ salary + ratio + expend
## Model 2: total ~ salary + ratio + expend + takers
##   Res.Df    RSS Df Sum of Sq      F               Pr(>F)
## 1     46 216812
## 2     45  48124  1    168688 157.74 0.0000000000000002607 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Here with a really small p-value we can see that we have statistically significant evidence that $\beta_{Takers} \neq 0$. Also note that the t-test of individually looking at `takers` in the full model and doing the F-test provide the same result in terms of p-values. This makes sense since a t-statistic squared (following a t-distribution with x degrees of freedom) is the same a F-statistic (following a F-distribution with 1, x degrees of freedom). In this case we are looking at a t-statistic following a t-distribution with 45 degrees of freedom and a F-statistic following a F-distribution with 1, 45 degrees of freedom, which matches the above claim.

3 Problem 3

The dataset `prostate` comes from a study on 97 men with prostate cancer who were due to receive a radical prostatectomy. Fit a model with lpsa as the response and lcavol as the predictor. Record the residual standard error and the R^2. We are going to fit the model and record these important quantities below. Please note for this problem when we say R^2 we are really looking at the R^2.

```
#Make the model
prostateModel1 <- lm(lpsa ~ lcavol, prostate)
```

```
# Standard error
(fit1SE <- summary(prostateModel1)$sigma)
```

```
## [1] 0.7874994
```

```
# R-squared
(fit1R2 <- summary(prostateModel1)$r.squared)
```

```
## [1] 0.5394319
```

Now add lweight, svi, lbph, age, lcp, pgg45 and gleason to the model one at a time. For each model record the residual standard error and the R^2.

First we will add lweight.

```
#Make the model
prostateModel2 <- lm(lpsa ~ lcavol + lweight, prostate)

# Standard error
(fit2SE <- summary(prostateModel2)$sigma)
```

```
## [1] 0.7506469
```

```
# R-squared
(fit2R2 <- summary(prostateModel2)$r.squared)
```

```
## [1] 0.5859345
```

Now add svi.

```
#Make the model
prostateModel3 <- lm(lpsa ~ lcavol + lweight + svi, prostate)

# Standard error
(fit3SE <- summary(prostateModel3)$sigma)
```

```
## [1] 0.7168094
```

```
# R-squared
(fit3R2 <- summary(prostateModel3)$r.squared)
```

```
## [1] 0.6264403
```

Now add lbph.

```
#Make the model
prostateModel4 <- lm(lpsa ~ lcavol + lweight + svi + lbph, prostate)

# Standard error
(fit4SE <- summary(prostateModel4)$sigma)
```

```
## [1] 0.7108232
```

```r
# R-squared
(fit4R2 <- summary(prostateModel4)$r.squared)
```

```
## [1] 0.6366035
```

Now add age.

```r
#Make the model
prostateModel5 <- lm(lpsa ~ lcavol + lweight + svi + lbph + age, prostate)

# Standard error
(fit5SE <- summary(prostateModel5)$sigma)
```

```
## [1] 0.7073054
```

```r
# R-squared
(fit5R2 <- summary(prostateModel5)$r.squared)
```

```
## [1] 0.6441024
```

Now add lcp.

```r
#Make the model
prostateModel6 <- lm(lpsa ~ lcavol + lweight + svi + lbph + age + lcp, prostate)

# Standard error
(fit6SE <- summary(prostateModel6)$sigma)
```

```
## [1] 0.7102135
```

```r
# R-squared
(fit6R2 <- summary(prostateModel6)$r.squared)
```

```
## [1] 0.645113
```

Now add pgg45.

```r
#Make the model
prostateModel7 <- lm(lpsa ~ lcavol + lweight + svi + lbph + age + lcp + pgg45, prostate)

# Standard error
(fit7SE <- summary(prostateModel7)$sigma)
```

```
## [1] 0.7047533
```

```r
# R-squared
(fit7R2 <- summary(prostateModel7)$r.squared)
```

```
## [1] 0.6544317
```

Lastly add gleason.

```
#Make the model
prostateModel8 <- lm(lpsa ~ lcavol + lweight + svi + lbph + age + lcp + pgg45 + gleason, prostate)

# Standard error
(fit8SE <- summary(prostateModel8)$sigma)
```

```
## [1] 0.7084155
```

```
# R-squared
(fit8R2 <- summary(prostateModel8)$r.squared)
```

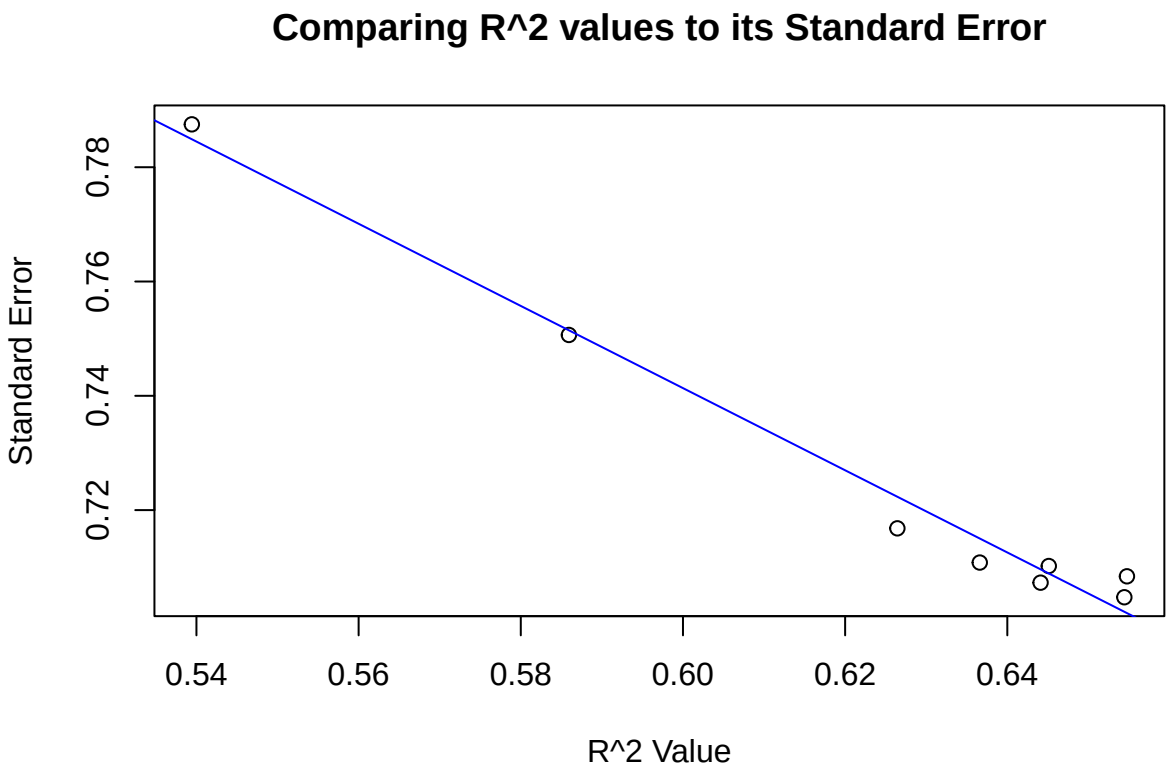
```
## [1] 0.6547541
```

Now we will plot the trends in these two statistics and comment on any patterns you see.

```
# Make vector of r^2 values
r2 <- c(fit1R2, fit2R2, fit3R2, fit4R2, fit5R2, fit6R2, fit7R2, fit8R2)

# Make vector of SE values
se <- c(fit1SE, fit2SE, fit3SE, fit4SE, fit5SE, fit6SE, fit7SE, fit8SE)

# Plot the trend
plot(r2,
     se,
     main = "Comparing R^2 values to its Standard Error",
     xlab = "R^2 Value",
     ylab = "Standard Error"
     )
abline(lm(se ~ r2), col = "blue")
```



It seems here that as the R^2 gets higher the standard error gets lower. It seems to have this negative linear property with each other. This makes sense given the R^2 formula will include the sum of squares residual (and the lower residual amount makes a higher R^2). It is also cool to note that this pattern seems to be close to a negative linear relationship, at least it seems that way in the graph.

4 Problem 4

Consider the analysis of variance (ANOVA) model: $y_i = \mu + \alpha_i + e_{ij}$ where i = 1, 2; j = 1, ... , 5; and we assume e_{ij} are IID N(0, σ^2). Find the probability distribution of the BLUE of $\mu + \alpha_1$.

First, we know that this type of model follows a Gauss-Markov Model (implying the errors follow a normal distribution). Due to this, we can assume that our BLUE of $\mu + \alpha_1$ will also follow a normal distribution as well. Thus if we call $\hat{\theta}$ our estimator, we know that our BLUE will follow a $N(\mathbf{E}(\hat{\theta}), \mathbf{Var}(\hat{\theta}))$

Next we can see if this estimable. We can see that $c^T = \begin{pmatrix} 1 & 1 & 0 \end{pmatrix}$ and $\hat{\beta} = \begin{pmatrix} \mu \\ \alpha_1 \\ \alpha_2 \end{pmatrix}$ which we can easily prove is in the column space of X^T – see previous lecture materials where we say that

$$X^T = \begin{pmatrix} 1 & ... & 1 & 1 & ... & 1 \\ 1 & ... & 1 & 0 & ... & 0 \\ 0 & ... & 0 & 1 & ... & 1 \end{pmatrix}$$

So we know that $c^T\hat{\beta}$ is estimable.

Next, we can find the expected value of the $\mu + \alpha_1$. Note from lecture we can say the best estimator is $\frac{y_{1.}}{5}$ which are all the values averaged out that contain μ and α_1. We can find the expected value of that to be $\mathbf{E}(\frac{y_{1.}}{5}) = \mathbf{E}(\frac{y_{11}}{5})+\mathbf{E}(\frac{y_{12}}{5})+\mathbf{E}(\frac{y_{13}}{5})+\mathbf{E}(\frac{y_{14}}{5})+\mathbf{E}(\frac{y_{15}}{5}) = \frac{\mu+\alpha_1}{5}+\frac{\mu+\alpha_1}{5}+\frac{\mu+\alpha_1}{5}+\frac{\mu+\alpha_1}{5}+\frac{\mu+\alpha_1}{5} = 5*\frac{\mu+\alpha_1}{5} = \mu+\alpha_1$. So our expected value is $\mu + \alpha_1$.

Now we can find our variance. Using our estimator $\frac{y_{1.}}{5}$ we will find its variance. Note all covariance is equal to 0 because of the iid condition so please note for variance formula simplification, I skip showing all the covariances since they are equal to 0. We can find the variance to be $\mathbf{Var}(\frac{y_{1.}}{5}) = \frac{1}{25}\mathbf{Var}(y_{1.}) = \frac{1}{25}[\mathbf{Var}(y_{11}) + \mathbf{Var}(y_{12}) + \mathbf{Var}(y_{13}) + \mathbf{Var}(y_{14}) + \mathbf{Var}(y_{15})] = \frac{1}{25}[\sigma^2 + \sigma^2 + \sigma^2 + \sigma^2 + \sigma^2] = \frac{1}{25}[5\sigma^2] = \frac{\sigma^2}{5}$. So we found our variance to be $\frac{\sigma^2}{5}$.

Therefore, by knowing if we call $\hat{\theta}$ our estimator that our BLUE will follow a $N(\mathbf{E}(\hat{\theta}), \mathbf{Var}(\hat{\theta}))$, we can plug our above results in to see that our BLUE will follow a $N(\mu + \alpha_1, \frac{\sigma^2}{5})$ probability distribution.