STATGR5205 Midterm - Fall 2017 - October 18

Name:_____

UNI:_____

The GU5205 midterm is closed notes and closed book. Calculators are allowed. Tablets, phones, computers and other equivalent forms of technology are strictly prohibited. Students are not allowed to communicate with anyone with the exception of the TA and the professor. If students violate these guidelines, they will receive a zero on this exam and potentially face more severe consequences. Students must include all relevant work in the handwritten problems to receive full credit.

Problem 1 [65 pts]

Consider the simple linear regression model

$$Y_i = \beta_0 + \beta_1 x_i + \epsilon_i, \quad \imath = 1, \ldots, n, \quad \epsilon_i \overset{iid}{\sim} N(0, \sigma^2), \tag{1}$$

and least squares estimators

$$\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}} \qquad \text{and} \qquad \hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{x}.$$

For this problem, you can use the following results:

$$E[\hat{\beta}_0] = \beta_0, \qquad E[\hat{\beta}_1] = \beta_1, \qquad Var[\hat{\beta}_0] = \sigma^2 \left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}} \right), \qquad Var[\hat{\beta}_1] = \frac{\sigma^2}{S_{xx}}. \tag{2}$$

For this exercise, use the scalar form of the simple linear regression model, i.e., don't use matrices.

Part A (5 pts)

Under model (1), prove that $\hat{\beta}_0 - \hat{\beta}_1$ is an unbiased estimator of $\beta_0 - \beta_1$. Note that you can directly use the relations from (2).

Part B (20 pts)

Under model (1), derive an expression for $Cov(\hat{\beta}_0, \hat{\beta}_1)$, where $\hat{\beta}_0$ and $\hat{\beta}_1$ are the least squares estimators. Simplify the result as much as possible. Note that you can directly use the relations from (2).

Part C (10 pts)

Under model (1), derive an expression for $Var[\hat{\beta}_0 - \hat{\beta}_1]$, where $\hat{\beta}_0$ and $\hat{\beta}_1$ are the least squares estimators. Simplify the result as much as possible. Note that you can directly use the relations from (2).

Note: if you cannot complete Part B, then express the solution to Part C in terms of $Cov(\hat{\beta}_0, \hat{\beta}_1)$.

Part D (15 pts)

Although not a very useful or common approach, we now consider a testing procedure to see if the intercept statistically differs from the slope, i.e., consider testing the null/alternative pair

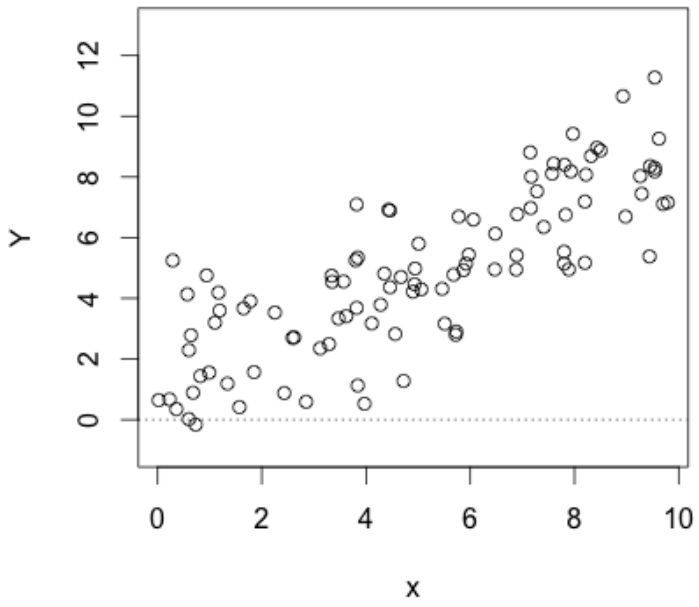
$$H_0 : \beta_0 = \beta_1 \qquad \text{versus} \qquad H_A : \beta_0 \neq \beta_1.$$

Write down **both** the T-statistic and F-statistic for testing the above null/alternative pair. When constructing the F-statistic, also identify the full and reduced models. Write your solution on pages 4 and 5.

Note: when specifying the full and reduced models, you do not have to derive the maximum likelihood estimators but make sure to identify them.

Part E (15 pts)

Consider the following toy dataset displayed in the scatter plot below. Let the predictor variable be assigned as `x`, the response as `Y` and assign `n` as the sample size. Note that there are $n = 100$ cases in this dataset.



Using the `R` code and output displayed on pages 6, 7 and 8, test the if the intercept statistically differs from the slope, i.e., test the null/alternative pair:

$$H_0 : \beta_0 = \beta_1 \qquad \text{versus} \qquad H_A : \beta_0 \neq \beta_1.$$

To receive full credit, compute **both** the T-statistic and F-statistic for testing the above null/alternative pair. Also compute the correct p-value and state the statistical conclusion. Write the solution on the top of page 7.

Note:

```
1-pt(t.calc,98)=0.1185218       1-pf(f.calc,1,98)=0.2370437
1-pt(t.calc,99)=0.1185074       1-pf(f.calc,1,99)=0.2370147
```

R code and Output:

```
> # Means and S.xx
> mean(x)
[1] 5.0423

> mean(Y)
[1] 4.9012

> sum((x-mean(x))^2)
[1] 834.4460
```

```
> # Model 1 with Summary and ANOVA #-----------------------------------------------

> summary(lm(Y~x))

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   1.1708     0.3144   3.725 0.000327 ***
x             0.7398     0.0541  13.676  < 2e-16 ***

Residual standard error: 1.563 on 98 degrees of freedom
Multiple R-squared:  0.6562,Adjusted R-squared:  0.6527
F-statistic:   187 on 1 and 98 DF,  p-value: < 2.2e-16

> anova(lm(Y~x))

          Df Sum Sq Mean Sq F value    Pr(>F)
x          1 456.71  456.71  187.03 < 2.2e-16 ***
Residuals 98 239.31    2.44

> # Model 2 with ANOVA #----------------------------------------------

> anova(lm(Y~x-1))

          Df  Sum Sq Mean Sq F value    Pr(>F)
x          1 2825.00 2825.00  1023.8 < 2.2e-16 ***
Residuals 99  273.18    2.76

> # Model 3 with ANOVA #----------------------------------------------

> x.new <-  x+1
> anova(lm(Y~x.new))

          Df Sum Sq Mean Sq F value    Pr(>F)
x.new      1 456.71  456.71  187.03 < 2.2e-16 ***
Residuals 98 239.31    2.44

> # Model 4 with ANOVA #----------------------------------------------

> x.new <-  x+1
> anova(lm(Y~x.new-1))

          Df  Sum Sq Mean Sq F value    Pr(>F)
x.new      1 2855.42 2855.42  1164.5 < 2.2e-16 ***
Residuals 99  242.76    2.45
```

Problem 2 [35 pts]

Consider the following study examining the effects of different amounts of THC, the major ingredient in marijuana, injected directly in the brain. The response variable (Y) is locomotor activity. In this approach, the researchers run an ANCOVA model (or multiple linear regression model) on the pos-injection scores, partialling out pre-injection differences. Such a procedure would adjust for the fact that much of the variability in post-injection activity could be accounted for by the variability in pre-injection activity. Note that variables D_1 through D_4 are indicator variables representing the different dosage levels and x_i is the continuous variable pre-injection activity.

control		.1 micro g (D_1)		.5 micro g (D_2)		1 micro g (D_3)		2 micro g (D_4)	
X	Y	X	Y	X	Y	X	Y	X	Y
Pre	Post	Pre	Post	Pre	Pos	Pre	Pos	Pre	Pos
4.34	1.30	1.55	0.93	7.18	5.10	6.94	2.29	4.00	1.44
3.50	0.94	10.56	4.44	8.33	4.16	6.10	4.75	4.10	1.11
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
⋮	⋮	⋮	⋮	⋮	⋮	7.35	2.35	⋮	⋮
⋮	⋮	⋮	⋮	6.30	4.84			⋮	⋮
1.90	0.93	9.58	4.22					5.54	2.93
$n_1 = 10$		$n_2 = 10$		$n_3 = 9$		$n_4 = 8$		$n_5 = 10$	

The statistical model used in our setting is:

$$Y_i = \beta_0 + \beta_1 D_{i1} + \beta_2 D_{i2} + \beta_3 D_{i3} + \beta_4 D_{i4} + \beta_5 x_i + \epsilon_i$$

$$i = 1, \ldots, 47, \qquad \epsilon_i \overset{iid}{\sim} N(0, \sigma^2),$$

where

$$D_{i1} = \begin{cases} 1 & \text{if .1 micro grams} \\ 0 & \text{if otherwise} \end{cases}, \quad D_{i2} = \begin{cases} 1 & \text{if .5 micro grams} \\ 0 & \text{if otherwise} \end{cases}$$

$$D_{i3} = \begin{cases} 1 & \text{if 1 micro grams} \\ 0 & \text{if otherwise} \end{cases}, \quad D_{i4} = \begin{cases} 1 & \text{if 2 micro grams} \\ 0 & \text{if otherwise} \end{cases}$$

and x_i is the respondent's pre-injection locomotor activity.

Part A (25 pts)

Run a single hypothesis testing procedure to see if THC dosage levels statistically influence post-locomotor activity, after controlling for pre-locomotor activity. To receive full credit, state the correct null/alternative pair, compute the test statistic and identify the correct P-value. To complete this exercise, use the `R` code & output displayed on Page 11. Assume $\alpha = 0.05$ significance.

R code and Output:

```
# Full model SSE and Summary #-----------------------------------------------

> full.model <- lm(Y~D1+D2+D3+D4+X,data=THC.Data)
> sum(residuals(full.model)^2)

[1] 20.12544

> summary(full.model)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.37384    0.27689  -1.350   0.1844
D1           0.61834    0.32855   1.882   0.0669 .
D2           1.45653    0.35656   4.085   0.0002 ***
D3           0.66599    0.34231   1.946   0.0586 .
D4           0.21998    0.31456   0.699   0.4883
X            0.43466    0.04918   8.838 4.84e-11 ***
---
Residual standard error: 0.7006 on 41 degrees of freedom
Multiple R-squared:  0.8042,Adjusted R-squared:  0.7803
F-statistic: 33.67 on 5 and 41 DF,  p-value: 1.704e-13

> # Reduced models SSE values #-----------------------------------------------

> reduced.1 <- lm(Y~X,data=THC.Data)
> sum(residuals(reduced.1)^2)

[1] 29.34931

> reduced.2 <- lm(Y~D1+D2+D3+D4,data=THC.Data)
> sum(residuals(reduced.2)^2)

[1] 58.4661

> # P-values #----------------------------------------------

> 1-pf(f.calc,1,41)
[1] 0.03606229

> 1-pf(f.calc,3,41)
[1] 0.006554796

> 1-pf(f.calc,4,41)
[1] 0.003256544
```

Part B (10 pts)

Run a Bonferroni procedure to test which THC dosage levels statistically influence post-locomotor activity, after controlling for pre-locomotor activity, i.e., simultaneously test the null hypotheses $H_0 : \beta_1 = 0, \quad H_0 : \beta_2 = 0, \quad H_0 : \beta_3 = 0, \quad H_0 : \beta_4 = 0$. To receive full credit, **circle** the correct R output and briefly identify which THC dosage levels statistically influence post-locomotor activity. Assume 95% family-wise error rate.

R code and Output:

```
> confint(model.1,level=1-.05/2)
                1.25 %   98.75 %
(Intercept) -1.0180992 0.2704152
D1          -0.1461084 1.3827809
D2           0.6269190 2.2861442
D3          -0.1304642 1.4624432
D4          -0.5119099 0.9518764
X            0.3202275 0.5490896

> confint(model.1,level=1-.05/4)
               0.625 %  99.375 %
(Intercept) -1.0972852 0.3496011
D1          -0.2400666 1.4767392
D2           0.5249510 2.3881122
D3          -0.2283567 1.5603357
D4          -0.6018672 1.0418337
X            0.3061627 0.5631544

> confint(model.1,level=1-.05/8)
               0.312 %  99.688 %
(Intercept) -1.1720667 0.4243826
D1          -0.3287988 1.5654713
D2           0.4286546 2.4844086
D3          -0.3208042 1.6527832
D4          -0.6868210 1.1267874
X            0.2928803 0.5764369

> confint(model.1,level=1-.05/10)
                0.25 %   99.75 %
(Intercept) -1.1953778 0.4476938
D1          -0.3564587 1.5931312
D2           0.3986367 2.5144265
D3          -0.3496223 1.6816013
D4          -0.7133031 1.1532696
X            0.2887398 0.5805773
```