STAT GR5205 Final - Fall 2018 - December 17

Name:_____ **UNI:**_____

The GR5205 final exam 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. Students must include all relevant work in the handwritten problems to receive full credit. The exam time is 180 minutes and once the exam time has expired, please do not discuss the midterm until after you leave the room (or log off of Zoom). For convenience, tear off the equation sheet for reference. Also note that the last two pages are scratch paper. The online students **do not have to scan the equation sheet** when submitting the final exam.

By signing below, you are acknowledging that if you do not follow the guidelines, then you will receive a score of zero on the midterm and potentially face more severe consequences. Signing the line below is also worth 5 points.

Signature: _____ **Date:** _____

Problem	**Points**	**Student's Score**
Signature	5	
1.i	10	
1.ii	5	
1.iii	5	
2.i	5	
2.ii	5	
2.iii	5	
2.iv	5	
2.v	5	
2.vi	5	
3.i	5	
3.ii	5	
3.iii	5	
4.i	10	
5.i	10	
5.ii	5	
5.ii	5	
Total	100	

Problem 1 [10 pts]

The article "Truth and DARE: Tracking Drug Education to Graduation" (*Social Problems* [1994]:448-456) compared the drug use of 288 randomly selected high school seniors exposed to drug education programs (DARE) and 335 randomly selected high school seniors who were not exposed to such a program. Data for marijuana use are given in the following table:

	Number who use marijuana	Number who do not use marijuana	Sample size
Exposed to DARE	141	147	288
Not exposed to DARE	181	154	335

In this setting, we use **simple logistic regression** to study the relationship between marijuana usage and whether or not the respondents were exposed to DARE. Both the independent and dependent variables are represented as dichotomous, i.e.,

$$y = \begin{cases} 1 & \text{use marijuana} \\ 0 & \text{don't use marijuana} \end{cases} \qquad x = \begin{cases} 1 & \text{not exposed to DARE} \\ 0 & \text{exposed to DARE} \end{cases}$$

The maximum likelihood estimate of the vector $\boldsymbol{\beta} = \begin{pmatrix} \beta_0 & \beta_1 \end{pmatrix}^T$ is:

$$\widehat{\boldsymbol{\beta}} = \begin{pmatrix} -0.0417 & 0.2032 \end{pmatrix}^T$$

The Hessian matrix of the log-likelihood function evaluated at $\widehat{\boldsymbol{\beta}}$ is:

$$\mathbf{G} = \mathbf{G}(\widehat{\boldsymbol{\beta}}) = \begin{pmatrix} -155.1747 & -83.2060 \\ -83.2060 & -83.2060 \end{pmatrix}$$

Perform the following tasks:

1.i Fill out the missing entries of the logistic regression summary table displayed on the next page:

	Estimate	Std. Error	z value	Pr($>|z|$)
(Intercept)				0.724
x				0.207

1.ii Interpret the estimated slope $\hat{\beta}_1$ in terms of the application. Also construct a 95% confidence interval for β_1. Does the program DARE have a significant impact on whether or not a respondent uses marijuana? For full credit, write down the correct null/alternative pair, confidence interval and statistical conclusion.
Note: $z_{.05/2} = 1.96$

1.iii Construct a 95% confidence interval for the true proportion of people who have used marijuana given they were not exposed to DARE.

Problem 2 [10 pts]

It is at least part of the folklore that repeated experience with any standardized test leads to better scores, even without any intervening study. Suppose that we obtain eight subjects and give them a standardized admissions exam every Saturday morning for 2 weeks. The data follow:

Subject	1	2	3	4	5	6	7	8
Exam 1	550	440	610	650	400	700	490	580
Exam 2	580	470	610	670	450	710	510	590

Our research question is:

Does the data suggest that the exam scores differ after the repeated trial?

The motivation behind this problem is the **repeated measures design**, where respondents are measured multiple times in a statistical experiment. The above example is a very basic application of repeated measures, which is also equivalent with the **paired (not pooled) two-sample t-test**. In our setting, we will use the linear regression model to study this data set. To do so, consider **stacking** exam scores ($\mathbf{Y}$) and using an indicator variable for exam ($\mathbf{x}_1$), i.e.,

$$\mathbf{Y} = \begin{pmatrix} 550 & 440 & \cdots & 580 & 580 & 470 & \cdots & 590 \end{pmatrix}^T$$

$$\mathbf{x}_1 = \begin{pmatrix} 0 & 0 & \cdots & 0 & 1 & 1 & \cdots & 1 \end{pmatrix}^T$$

To analyze this dataset correctly, the experimenter should also incorporate subject variability into the model to account for how respondents perform differently on exams.

Perform the following tasks:

2.i Suppose that we use the simple linear regression model and **do not** control for subject variability. In this case, the model is $Y_i = \beta_0 + \beta_1 x_{i1} + \epsilon_i$. Using the `R output` displayed on pages 8-9, test if the exam scores statistically changed after the repeated trial. For full credit, write down the correct null/alternative pair, test-statistic, p-value and statistical conclusion. Use $\alpha = .05$

2.ii Now suppose that we **do account for subject variability**. Write down the multiple linear regression model and the correct design matrix for this scenario.

2.iii Using the `R output` displayed on pages 8-9, test if students performance statistically differs, after taking into account exam variability. **Note: this is not our research question.** For full credit,

a. Write down the correct null/alternative pair.

b. Compute the f-statistic by hand using the **general linear f-stat formula**. Show all steps in this calculation.

c. Identify the p-value and write down the statistical conclusion.

Note: use $\alpha = .05$. **Hint:** This is a balanced design, i.e., Type I SS = Type III SS.

2.iv Using the `R output` displayed on pages 8-9, test if the exam scores statistically differ after the repeated trial, taking into account subject variability. **Note: this is our research question.** For full credit, write down the correct null/alternative pair, test-statistic, p-value and statistical conclusion. Use $\alpha = .05$

R code, summary output and p-value:

```
# Store data vectors #---------------------------------------------

Scores <- c(550,440,610,650,400,700,490,580,580,470,610,670,450,710,510,590)
Student <- rep(1:8,2)
Exam <- c(rep("Exam1",8),rep("Exam2",8))

# Model_1 Summary and ANOVA #-------------------------------------------

> model_1 <- lm(Scores~Exam)
> summary(model_1)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   552.50      34.68  15.929 2.29e-10 ***
ExamExam2      21.25      49.05   0.433    0.671
---
Residual standard error: 98.1 on 14 degrees of freedom
Multiple R-squared:  0.01323,Adjusted R-squared:  -0.05726
F-statistic: 0.1877 on 1 and 14 DF,  p-value: 0.6715

> anova(model_1)
Analysis of Variance Table

Response: Scores
          Df Sum Sq Mean Sq F value Pr(>F)
Exam       1   1806  1806.2  0.1877 0.6715
Residuals 14 134738  9624.1

# Model_2 Summary and ANOVA #-----------------------------------------------
> model_2 <- lm(Scores~Student_fac+Exam)
> summary(model_2)

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   554.375      8.234  67.326 4.19e-11 ***
Student_fac2 -110.000     10.979 -10.019 2.11e-05 ***
Student_fac3   45.000     10.979   4.099 0.004580 **
Student_fac4   95.000     10.979   8.653 5.50e-05 ***
Student_fac5 -140.000     10.979 -12.752 4.22e-06 ***
Student_fac6  140.000     10.979  12.752 4.22e-06 ***
Student_fac7  -65.000     10.979  -5.920 0.000587 ***
Student_fac8   20.000     10.979   1.822 0.111294
ExamExam2      21.250      5.489   3.871 0.006123 **
---
```

```
Residual standard error: 10.98 on 7 degrees of freedom
Multiple R-squared:  0.9938,Adjusted R-squared:  0.9868
F-statistic: 140.7 on 8 and 7 DF,  p-value: 4.902e-07

> anova(model_2)
Analysis of Variance Table

Response: Scores
            Df Sum Sq Mean Sq F value    Pr(>F)
Student_fac  7 133894 19127.7 158.689 3.576e-07 ***
Exam         1   1806  1806.3  14.985  0.006123 **
Residuals    7    844   120.5


# Model_3 Summary and ANOVA #-----------------------------------------------
> model_3 <- lm(Scores~Student+Exam)
> summary(model_3)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   530.27      61.13   8.675 9.11e-07 ***
Student         4.94      11.02   0.448    0.661
ExamExam2      21.25      50.51   0.421    0.681
---
Residual standard error: 101 on 13 degrees of freedom
Multiple R-squared:  0.02824,Adjusted R-squared:  -0.1213
F-statistic: 0.1889 on 2 and 13 DF,  p-value: 0.8301

> anova(model_3)
Analysis of Variance Table

Response: Scores
          Df Sum Sq Mean Sq F value Pr(>F)
Student    1   2050  2050.3  0.2009 0.6614
Exam       1   1806  1806.2  0.1770 0.6809
Residuals 13 132687 10206.7
```

2.v In a few sentences, describe how the results from **Part 2.i** to **Part 2.iv** change, i.e., describe the difference in significance in relation to R^2, SSE and overall model specification.

2.vi In this problem part, we compare Type I versus Type III sums of squares. Suppose we add some noise into our model. This might represent recording the respondents' height in the study, which clearly doesn't contribute to their test scores. We will compare the four ANOVA tables shown below:

ANOVA.1: `anova(lm(Scores∼Exam+Student_fac))`

ANOVA.2: `anova(lm(Scores∼Student_fac+Exam))`

ANOVA.3: `anova(lm(Scores∼noise+Exam+Student_fac))`

ANOVA.4: `anova(lm(Scores∼noise+Student_fac+Exam))`

Notice from the `R output` on page 11, the Type I SS are the same for `ANOVA.1` and `ANOVA.2`, regardless of the order that the variables are entered into the model. However, this relationship does not hold for `ANOVA.3` and `ANOVA.4`.

In a few sentences, describe why the order of the variables changes the Type I SS in the presence of the continuous noise variable. Please use complete sentences and any relevant computations to support your argument.

```
#  ANOVA 1 #-----------------------------------------------
> anova(lm(Scores~Exam+Student_fac))
Analysis of Variance Table

Response: Scores
            Df Sum Sq Mean Sq F value    Pr(>F)
Exam         1   1806  1806.2  14.985  0.006123 **
Student_fac  7 133894 19127.7 158.689 3.576e-07 ***
Residuals    7    844   120.5

#  ANOVA 2 #-----------------------------------------------
> anova(lm(Scores~Student_fac+Exam))
Analysis of Variance Table

Response: Scores
            Df Sum Sq Mean Sq F value    Pr(>F)
Student_fac  7 133894 19127.7 158.689 3.576e-07 ***
Exam         1   1806  1806.3  14.985  0.006123 **
Residuals    7    844   120.5

#  Add noise (weight) #-------------------------------------------------

> noise <- rnorm(16,mean=70,sd=3)

#  ANOVA 3 #-----------------------------------------------
> anova(lm(Scores~noise+Exam+Student_fac))
Analysis of Variance Table

Response: Scores
            Df Sum Sq Mean Sq F value    Pr(>F)
noise        1  23053 23052.8 179.222 1.075e-05 ***
Exam         1   1470  1470.5  11.432   0.01484 *
Student_fac  7 111249 15892.7 123.556 4.660e-06 ***
Residuals    6    772   128.6

#  ANOVA 4 #-----------------------------------------------
> anova(lm(Scores~noise+Student_fac+Exam))
Analysis of Variance Table

Response: Scores
            Df Sum Sq Mean Sq F value    Pr(>F)
noise        1  23053 23052.8 179.222 1.075e-05 ***
Student_fac  7 110888 15841.1 123.156 4.706e-06 ***
Exam         1   1831  1831.2  14.237  0.009254 **
Residuals    6    772   128.6
```

Problem 3 [?? pts]

In a small-scale experimental study of the relation between degree of brand liking (Y) and moisture content (X_1) and sweetness (X_2) of the product, the following results were obtained from the experiment based on a completely randomized design.

Case	Y	X_1	X_2
1	64.00	4.00	2.00
2	73.00	4.00	4.00
3	61.00	4.00	2.00
4	76.00	4.00	4.00
5	72.00	6.00	2.00
6	80.00	6.00	4.00
7	71.00	6.00	2.00
8	83.00	6.00	4.00
9	83.00	8.00	2.00
10	89.00	8.00	4.00
11	86.00	8.00	2.00
12	93.00	8.00	4.00
13	88.00	10.00	2.00
14	95.00	10.00	4.00
15	94.00	10.00	2.00
16	100.00	10.00	4.00

Also consider the following quantities:

$$(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y} = \begin{pmatrix} 37.650 \\ 4.425 \\ 4.375 \end{pmatrix}$$

$$\mathbf{Y}^T\big(\mathbf{I} - \frac{1}{n}\mathbf{J}\big)\mathbf{Y} = 1967$$

$$\mathbf{Y}^T\big(\mathbf{I} - \mathbf{H}\big)\mathbf{Y} = 94.3$$

$$(\mathbf{X}^T\mathbf{X})^{-1} = \begin{pmatrix} 1.2375 & -0.0875 & -0.1875 \\ -0.0875 & 0.0125 & 0.0000 \\ -0.1875 & 0.0000 & 0.0625 \end{pmatrix}$$

3.i Fill out the missing entries of the multiple linear regression summary table and standard ANOVA table, both displayed below:

Summary Table

	Estimate	Std. Error	t value	Pr($>\|z\|$)
(Intercept)				1.20×10^{-8}
$X1$				1.78×10^{-9}
$X2$				2.01×10^{-5}

ANOVA Table

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Regression					2.658×10^{-9}
Residuals			BLANK	BLANK	BLANK
Total			BLANK	BLANK	BLANK

3.ii Run the relevant test to see if sweetness (X_2) is significantly related to the degree of brand liking after moisture is held constant. To receive full credit, show all relevant steps of the testing procedure.

3.iii Consider estimating parameters β_0, β_1 and β_2 simultaneously using a Bonferroni procedure with familywise confidence level 95%. Construct the Bonferoni interval for β_2. **Note:** In practice I want to construct three intervals but to save time on the exam, I am having students construct only one of the three intervals. Also note that one of the critical values below is correct.

$$t_{0.025,13} = 2.1604, \quad t_{0.0125,13} = 2.5326, \quad t_{0.0083,13} = 2.7459$$
$$t_{0.025,14} = 2.1447, \quad t_{0.0125,14} = 2.5096, \quad t_{0.0083,14} = 2.7178$$

Problem 4 [?? pts]

Consider the least squares estimated **multiple** linear regression model

$$\hat{\mathbf{Y}} = \mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{H}\mathbf{Y}, \tag{1}$$

where $\mathbf{H}$ is the hat-matrix and $\mathbf{X}$ is the design matrix of dimensions $(n \times p)$. Using properties of the hat-matrix, prove that any vector in the column space of $\mathbf{X}$ is orthogonal to the residual vector $\mathbf{e}$.

Problem 5 [?? pts]

Suppose that respondents are randomly allocated into two distinct groups of size n_1 and n_2. Also suppose that some continuous measurement (Y) is recorded from each case. In our setting, we consider the linear model:

$$Y_i = \beta_1 x_{i1} + \beta_2 x_{i2} + \epsilon_i, \quad i = 1, \ldots, n, \quad \epsilon_i \overset{iid}{\sim} N(0, \sigma^2), \tag{2}$$

where

$$x_{i1} = \begin{cases} 1 & \text{Group 1} \\ 0 & \text{Group 2} \end{cases} \qquad x_{i2} = \begin{cases} 1 & \text{Group 2} \\ 0 & \text{Group 1} \end{cases}$$

Note that $\mathbf{x}_1^T \mathbf{x}_2 = \sum x_{i1} x_{i2} = 0$.

5.i Using the least squares equation, compute and simplify an expression for $\widehat{\boldsymbol{\beta}}$.

5.ii Using the least squares equation, compute and simplify an expression for var$[\widehat{\boldsymbol{\beta}}]$.

5.ii In this setting, is the following identity true?

$$\sum_{i=1}^{n} e_i = \sum_{i=1}^{n} (y_i - \hat{y}_i) = 0$$

Justify your answer in a few sentences.

Formula Page

Simple linear regression model

$$Y_i = \beta_0 + \beta_1 x_i + \epsilon_i \qquad i = 1, .., n \qquad \epsilon_i \sim N(0, \sigma^2)$$

Regression through the origin model

$$Y_i = \beta x_i + \epsilon_i \qquad i = 1, .., n \qquad \epsilon_i \sim N(0, \sigma^2)$$

Multiple linear regression model

$$Y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_{p-1} x_{i,p-1} + \epsilon_i \qquad i = 1, .., n \qquad \epsilon_i \sim N(0, \sigma^2)$$

Or

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon} \qquad \boldsymbol{\epsilon} \sim MN(\mathbf{0}, \sigma^2\mathbf{I})$$

Sums of squares

- Simple linear regression

$$S_{xy} = \sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) = \sum_{i=1}^{n} x_i y_i - \frac{1}{n}\Big(\sum_{i=1}^{n} x_i\Big)\Big(\sum_{i=1}^{n} y_i\Big)$$

$$S_{xx} = \sum_{i=1}^{n}(x_i - \bar{x})^2 = \sum_{i=1}^{n} x_i^2 - \frac{1}{n}\Big(\sum_{i=1}^{n} x_i\Big)^2$$

$$SST = S_{yy} = \sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{n} y_i^2 - \frac{1}{n}\Big(\sum_{i=1}^{n} y_i\Big)^2$$

$$SSR = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 = \hat{\beta}_1^2 S_{xx}$$

$$SSE = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

- Multiple linear regression

$$SST = S_{yy} = \sum_{i=1}^{n}(y_i - \bar{y})^2 = \mathbf{Y}^T\big(\mathbf{I} - \frac{1}{n}\mathbf{J}\big)\mathbf{Y}$$

$$SSR = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 = \mathbf{Y}^T\big(\mathbf{H} - \frac{1}{n}\mathbf{J}\big)\mathbf{Y}$$

$$SSE = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = \mathbf{Y}^T\big(\mathbf{I} - \mathbf{H}\big)\mathbf{Y}$$

Additive identity: $SST = SSR + SSE$

Mean squares: $MSR = SSR/(p-1)$ $\hat{\sigma}^2 = MSE = SSE/(n-p)$

Estimation

Model	Coefficients	Variance
$Y_i = \beta_0 + \beta_1 x_i + \epsilon_i$	$\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}}$ $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$	$\hat{\sigma}_{\hat{\beta}_1} = \frac{MSE}{S_{xx}}$
$Y_i = \beta x_i + \epsilon_i$	$\hat{\beta} = \frac{\sum y_i x_i}{\sum x_i^2}$	$\hat{\sigma}_{\hat{\beta}} = \frac{MSE}{\sum x_i^2}$
$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$	$\hat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}$	$\hat{\boldsymbol{\Sigma}}_{\hat{\boldsymbol{\beta}}} = MSE(\mathbf{X}^T\mathbf{X})^{-1}$

Hat-values:

Model	Hat-values/matrix	Properties
$Y_i = \beta_0 + \beta_1 x_i + \epsilon_i$	$h_{ij} = \frac{1}{n} + \frac{1}{S_{xx}}(x_i - \bar{x})(x_j - \bar{x})$ $i, j = 1, 2, \ldots, n$	$h_{ij} = h_{ji}$ $\sum_{j=1}^n h_{ij} = 1$ $\sum_{j=1}^n h_{ij} x_j = x_i$ $\sum_{j=1}^n {h_{ij}}^2 = h_{ii}$
$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$	$\mathbf{H} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T$	$\mathbf{H}^T = \mathbf{H}$ $\mathbf{HX} = \mathbf{X}$ $\mathbf{H}^2 = \mathbf{H}$

Expectation and covariance properties for random vectors

- Let $E[\mathbf{Y}] = \mu$, $Var[\mathbf{Y}] = \mathbf{\Sigma}$ and $\mathbf{A}$ be a matrix of scalars. Then

$$E[\mathbf{W}] = E[\mathbf{AY}] = \mathbf{A}E[\mathbf{Y}] = \mathbf{A}\mu,$$

$$Var[\mathbf{W}] = Var[\mathbf{AY}] = \mathbf{A}Var[\mathbf{Y}]\mathbf{A}^T = \mathbf{A\Sigma A}^T.$$

- Let A be symmetric, then the *quadratic form* $\mathbf{Y}^T\mathbf{AY}$ has expectation

$$E[\mathbf{Y}^T\mathbf{AY}] = tr(\mathbf{A\Sigma}) + \mu^T\mathbf{A}\mu,$$

Inferential procedures for slope (simple linear regression)

Test statistic	Confidence interval
$t_{calc} = \frac{\hat{\beta}_1 - \beta_{10}}{\sqrt{\frac{MSE}{S_{xx}}}}$	$\hat{\beta}_1 \pm t_{\alpha/2,n-2}\sqrt{\frac{MSE}{S_{xx}}}$

ANOVA Table (simple and multiple linear regression)

Source	Degrees of freedom	Sums of squares	Mean square	F-ratio
Regression	$p-1$	SSR	MSR	$f_{calc} = \frac{MSR}{MSE}$
Error	$n-p$	SSE	MSE	
Total	$n-1$	SST		

General linear f-statistic (simple and multiple linear regression)

$$f_{calc} = \left(\frac{SSE_R - SSE_F}{df_R - df_F}\right) \div \left(\frac{SSE_F}{df_F}\right)$$

Inferential procedure for linear combination of $\boldsymbol{\beta}$

- For known vector $\mathbf{c} \in \mathbb{R}^p$, define the linear parameter ψ and its estimator $\hat{\psi}$ by

$$\psi = \mathbf{c}^T\boldsymbol{\beta} \quad \text{and} \quad \hat{\psi} = \mathbf{c}^T\hat{\boldsymbol{\beta}}$$

- T-statistic for testing $H_0 : \psi = \psi_0$

$$t_{calc} = \frac{\hat{\psi} - \psi_0}{\hat{\sigma}_{\hat{\psi}}} = \frac{\mathbf{c}^T\hat{\boldsymbol{\beta}} - \psi_0}{\sqrt{\mathbf{c}^T(\widehat{\text{var}}[\hat{\boldsymbol{\beta}}])\mathbf{c}}} \qquad T_{calc} \sim t(df = n - p)$$

Maximum likelihood and likelihood ratio test

- Likelihood function:

$$\mathcal{L}(\theta; y_1, y_2, \ldots, y_n) = f(y_1, y_2, \ldots, y_n|\theta) = f(y_1|\theta) \times f(y_2|\theta) \times \cdots \times f(y_n|\theta)$$

- Maximum likelihood estimator:

$$\hat{\theta} = \underset{\theta}{\text{argmax}}\mathcal{L}(\theta; y_1, y_2, \ldots, y_n) = \underset{\theta}{\text{argmax}}\mathcal{L}(\theta)$$

- Consider the null alternative pair $H_0 : \theta \in \Theta_0$ versus $H_A : \theta \in \Theta_0^C$. The likelihood ratio test statistic is

$$\lambda(y_1, y_2, \ldots, y_n) = \frac{\underset{\theta\in\Theta_0}{\max}\mathcal{L}(\theta; y_1, y_2, \ldots, y_n)}{\underset{\theta\in\Theta}{\max}\mathcal{L}(\theta; y_1, y_2, \ldots, y_n)}$$

  Reject when $\lambda(y_1, y_2, \ldots, y_n) \leq c$, for some $0 \leq c \leq 1$.

Prediction

$$\hat{y}_h = \hat{\beta}_0 + \hat{\beta}_1 x_h \qquad \hat{y}_h = \mathbf{X}_h^T\widehat{\boldsymbol{\beta}}$$

	Confidence interval for EY_h	Prediction interval for future value $Y_{h(new)}$
Simple	$\hat{y}_h \pm t_{\alpha/2,n-2}\sqrt{MSE\left(\frac{1}{n} + \frac{(x_h-\bar{x})^2}{S_{xx}}\right)}$	$\hat{y}_h \pm t_{\alpha/2,n-2}\sqrt{MSE\left(1 + \frac{1}{n} + \frac{(x_h-\bar{x})^2}{S_{xx}}\right)}$
Multiple	$\hat{y}_h \pm t_{\alpha/2,n-p}\sqrt{MSE(\mathbf{X}_h^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}_h)}$	$\hat{y}_h \pm t_{\alpha/2,n-p}\sqrt{MSE(1 + \mathbf{X}_h^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}_h)}$

Correlation

Sample correlation coefficient	Coefficient of determination
$r = \frac{S_{xy}}{\sqrt{(S_{xx})(S_{yy})}}$	$r^2 = 1 - \frac{SSE}{SST}$ (R^2 is the same for multiple regression)

Inferential procedure for linear correlation

$$t_{calc} = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$$

Diagnostics

$$e_i = Y_i - \hat{Y}_i \qquad \mathbf{e} = \mathbf{Y} - \hat{\mathbf{Y}}$$

$$d_i = Y_i - \hat{Y}_{i(i)} = \frac{e_i}{1 - h_{ii}}$$

$$(DFFITS)_i = \frac{\hat{Y}_i - \hat{Y}_{i(i)}}{\sqrt{MSE_{(i)}h_{ii}}}, \qquad i = 1, \ldots, n$$

$$(DFBETAS)_{k(i)} = \frac{\hat{\beta}_k - \hat{\beta}_{k(i)}}{\sqrt{MSE_{(i)}c_{kk}}}, \qquad k = 0, 1, \ldots, p-1.$$

Extra sums of squares

$$SSR(x_1|x_2) = SSE(x_2) - SSE(x_1, x_2)$$

$$SSR(x_1|x_2) = SSR(x_1, x_2) - SSR(x_2)$$

Type I sums of squares decomposition of SSR

$$SSR(x_1, x_2, \ldots, x_{p-1}) = SSR(x_1) + SSR(x_2|x_1) + SSR(x_3|x_1, x_2) + \cdots + SSR(x_{p-1}|x_1, \ldots, x_{p-2})$$

Qualitative predictors

A regression model with $p-1$ predictor variables contains **additive effects** if the response function can be written in the form:

$$E[Y] = f_1(x_1) + f_2(x_2) + \cdots + f_{p-1}(x_{p-1}),$$

where $f_1, f_2, \ldots, f_{p-1}$ can be any functions.

Model selection

$$C_p = \frac{SSE_p}{MSE(x_1, x_2, \ldots, x_{P-1})} - (n - 2p)$$

$$AIC_p = n \log(SSE_p) - (n \log(n) - 2p)$$

Logistic regression model

$$\begin{aligned} E[Y_i] = p_i &= F_L(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_{p-1} x_{i,p-1}) \\ &= \frac{\exp(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_{p-1} x_{i,p-1})}{1 + \exp(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_{p-1} x_{i,p-1})} \end{aligned}$$

Logit transformation

$$F_L^{-1}(EY_i) = \log\left(\frac{EY_i}{1 - EY_i}\right) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_{p-1} x_{i,p-1}$$

Asymptotic approximate distribution of MLEs

$$\hat{\theta}_{MLE} \xrightarrow{d} N(\theta, \mathcal{I}^{-1}(\theta))$$

- $\mathcal{I}(\theta)$ is the *Fisher information*

$$\mathcal{I}(\theta) = -E\left[\frac{\partial^2}{\partial \theta^2} \log f(y; \theta)\right].$$

- Let $\mathbf{G}$ (Hessian) denote the matrix of second-order partial derivatives of the log-likelihood function, the derivatives being taken with respect to the parameters $\boldsymbol{\beta}$.

$$\mathbf{G}_{p \times p} = [g_{ij}] \quad i, j = 0, 1, \ldots, p-1.$$

- The estimated approximate variance-covariance matrix of the estimated regression coefficients ($\hat{\boldsymbol{\beta}}$) is the negative inverse of the Hessian matrix evaluated at the maximum likelihood estimates:

(3) $$\hat{\boldsymbol{\Sigma}}_{\hat{\boldsymbol{\beta}}} = \big[-\mathbf{G}(\hat{\boldsymbol{\beta}}) \big]^{-1}$$

$(1-\alpha)100\%$ confidence interval for $E[Y_h] = p_h$

$$CI = (L^*, U^*) \qquad L^* = \frac{\exp(L)}{1+\exp(L)}, \qquad U^* = \frac{\exp(U)}{1+\exp(U)},$$

where

$$L = \mathbf{X}_h^T \hat{\boldsymbol{\beta}} - z_{\alpha/2} s_h, \qquad U = \mathbf{X}_h^T \hat{\boldsymbol{\beta}} + z_{\alpha/2} s_h,$$

$$s_h^2 = \mathbf{X}_h^T \hat{\boldsymbol{\Sigma}}_{\hat{\boldsymbol{\beta}}} \mathbf{X}_h,$$

Other metrics

- Mean square error (inference definition)

$$MSE(\hat{\theta}) = E[(\hat{\theta} - \theta)^2] = Var(\hat{\theta}) + (E[\hat{\theta}] - \theta)^2$$

- The Kullback-Leibler (KL) divergence

$$KL(f,g) = \int \log\left\{ \frac{f(\mathbf{y})}{g(\mathbf{y})} \right\} f(\mathbf{y}) d\mathbf{y} = \int \log\{f(\mathbf{y})\} f(\mathbf{y}) d\mathbf{y} - \int \log\{g(\mathbf{y})\} f(\mathbf{y}) d\mathbf{y}.$$

Normal density

$$\frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}(y-\mu)^2 \right), \quad -\infty < y < \infty$$

Inverse of 2×2 matrix

$$\mathbf{A} = \begin{pmatrix} a & b \\ c & d \end{pmatrix} \qquad \mathbf{A}^{-1} = \frac{1}{ad-bc} \begin{pmatrix} d & -b \\ -c & a \end{pmatrix}$$