학번: 이름:

회귀분석 - 중간고사

1. 다음의 각 문제에 대한 올바른 답에 O표시 하시오.

   (a) 상수를 포함하는 단순선형회귀모형의 잔차들의 합은 0이다. (TRUE / FALSE)

   (b) 잔차들은 서로 독립이다. (TRUE / FALSE)

   (c) 잔차들의 분산은 동일하다. (TRUE / FALSE)

   (d) 원점을 지나는 단순선형회귀모형의 분석에서 잔차와 설명변수의 곱의 합, i.e., $\sum_{i=1}^{n} e_i X_i$, 은 0이다. (TRUE / FALSE)

   (e) 상수를 포함하는 단순선형회귀모형의 분석에서 적합된 모형은 항상 $(\overline{X}, \overline{Y})$를 지난다. (TRUE / FALSE)

   (f) $Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \cdots + \beta_d X^d + \epsilon$의 모형식은 모수에 대하여 선형이다. (TRUE / FALSE)

   (g) 상수를 포함하는 단순선형회귀모형, i.e., $Y_i = \beta_0 + \beta_1 X_i + \epsilon_i, \quad i = 1, 2, \cdots, n$, 에서 반응변수와 설명변수를 각각 $Y_i^* = 1 - Y_i$, $X_i^* = 20 - X_i$로 변환하여 분석하였을 때, 기울기의 추정값은 변하지 않고 동일하다. (TRUE / FALSE)

   (h) 두 단순선형회귀모형들

   $$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i, \quad i = 1, 2, \cdots, n \quad \text{(모형1)}$$
   $$Y_i = \beta_0^* + \beta_1^* (X_i - \overline{X}) + \epsilon_i, \quad i = 1, 2, \cdots, n \quad \text{(모형2)}$$

   을 자료에 적합시켰을 때, β_1^*의 추정값은 β_1의 추정값과 동일하다. (TRUE/FALSE)

   (i) 상수항이 0이 아닌 단순선형회귀분석에서 잔차와 예측값의 표본공분산은 상황에 따라 양수 또는 음수가 된다. (TRUE/FALSE)

2. 'hweight' 자료에 단순선형회귀모형을 적합하였다.

```
> str(dat)
'data.frame': 696 obs. of  3 variables:
 $ weight: num  60.8 69.3 74.3 46.3 68.1 ...
 $ height: num  169 171 175 172 176 ...
 $ gender: Factor w/ 2 levels "M","F": 1 1 1 1 1 1 1 1 1 1

> fit <- lm( weight ~ gender, data = dat)

> summary(fit)
...
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  69.5001     0.4712  147.49   <2e-16 ***
genderF     -15.9998     0.6626  -24.15   <2e-16 ***
...
```

- 각각 남자와 여자의 평균 몸무게를 적으시오.

3. 상수항을 포함하는 단순선형회귀모형의 분석에서 회귀제곱합 SSR의 정의는 다음과 같다.

$$SSR = \sum_{i=1}^{n} \left(\hat{Y}_i - \overline{Y}\right)^2$$

   (a) 아래 등식이 성립함을 증명하시오.

$$SSR = \hat{\beta}_1^2 \sum_{i=1}^{n} \left(X_i - \overline{X}\right)^2$$

   (b) 위 등식을 사용하여 $E(MSR)$을 계산하시오.

4. 상수항을 포함하는 단순선형회귀모형에서 반응변수 Y의 표본평균 $\overline{Y}$와 기울기의 최소제곱 추정량 $\hat{\beta}_1$의 공분산 $Cov(\overline{Y}, \hat{\beta}_1)$을 계산하시오.

5. 상수항을 포함하는 단순선형회귀분석에서 결정계수(R^2)는 반응변수와 설명변수 간의 표본상관계수의 제곱(r^2)과 같음을 증명하시오.

6. 상수항을 포함하는 단선선형회귀분석에서 가설 $H_0 : \beta_1 = 0$에 대한 t 검정통계량을 제곱하면 분산분석표의 F 검정통계량과 동일함을 보여라.

7. 중선형회귀모형

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

   에서

   (a) 오차제곱합을 벡터와 행렬을 이용하여 표현하시오.

   (b) 최소제곱추정량 $\hat{\boldsymbol{\beta}}$을 계산하시오 (간략하게 도출 과정을 포함할 것. 만약, 결과만 적으면 부분 점수만 부여함)

   (c) 최소제곱추정량 $\hat{\boldsymbol{\beta}}$의 평균과 분산을 계산하시오. (간략하게 도출 과정을 포함할 것. 만약, 결과만 적으면 부분 점수만 부여함)

   (d) 행렬 $\boldsymbol{I} - \frac{1}{n}\boldsymbol{J}$이 멱등행렬임을 보이시오.

8. 'bookprice' 자료에 다음의 단순선형회귀모형을 적합하였다.

```
> dat <- bookprice
> str(dat)
'data.frame': 20 obs. of  2 variables:
 $ price: num  27 15 14 15 9.5 20 22 20 16 24 ...
 $ pages: int  637 336 336 430 164 533 529 509 419 596 ...
> fit <- lm(formula = price ~ pages, data = dat)
> summary(fit)

Call:
lm(formula = price ~ pages, data = dat)

Residuals:
    Min      1Q  Median      3Q     Max
-2.9228 -0.7875 -0.1059  0.9603  2.4975

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 2.191079   0.859491   2.549   0.0201 *
pages       0.035026   0.001922  18.223 4.77e-13 ***
---
Residual standard error: 1.433 on 18 degrees of freedom
Multiple R-squared:  0.9486, Adjusted R-squared:  0.9457
F-statistic: 332.1 on 1 and 18 DF,  p-value: 4.769e-13

> round( qt( 0.05, 15:20 ), 3 )
[1] -1.753 -1.746 -1.740 -1.734 -1.729 -1.725
> round( qt( 0.025, 15:20 ), 3 )
[1] -2.131 -2.120 -2.110 -2.101 -2.093 -2.086
> sapply( dat, mean )
  price   pages
 16.725 414.950
> sapply( dat, var )
      price       pages
   37.82829 29249.52368
```

(a) 적합된 회귀식을 적으시오.

(b) β_1이 모형에 포함되는 것이 적절한지 t 검정을 다음 순서에 따라 실시하시오. (유의수준 $\alpha = 0.05$을 사용할 것)

i. 귀무가설과 대립가설

ii. 검정통계량의 분포

iii. 검정통계량의 관측값

iv. p 값

v. 결론

(c) β_0에 대한 95% 신뢰구간을 구하시오.

(d) $pages = 400$에서 반응변수 $price$의 평균에 대한 95% 신뢰구간을 구하시오.

(e) $pages = 400$에서 관측될 수 있는 새로운 반응변수 $price$에 대한 95% 예측구간을 구하시오.

9. 'adsale' 자료에 다음의 중선형회귀모형을 적합하였다.

```
> dat <- adsale
> str(dat)
'data.frame': 10 obs. of  3 variables:
 $ sale : int  39 42 45 47 50 50 52 55 57 60
 $ ad   : int  4 6 6 8 8 9 9 10 12 12
 $ media: Factor w/ 2 levels "TV","Newspaper": 1 2 1 2 1 2 1 1 2 1
> fit <- lm(formula = sale ~ ., data = dat)
> anova(fit)
Analysis of Variance Table

Response: sale
          Df Sum Sq Mean Sq F value    Pr(>F)
ad         1 378.50  378.50    3269 1.313e-10 ***
media      1  16.79   16.79     145 6.212e-06 ***
Residuals  7   0.81    0.12
> qf( c(0.95, 0.975), 2, 7 )
[1] 4.737414 6.541520
> qf( c(0.95, 0.975), 2, 8 )
[1] 4.458970 6.059467
> qf( c(0.95, 0.975), 1, 7 )
[1] 5.591448 8.072669
> qf( c(0.95, 0.975), 1, 8 )
[1] 5.317655 7.570882
```

(a) 설명변수 ad가 포함된 모형에 설명변수 $media$가 추가될 때, 모형에 의해 추가적으로 설명되는 변동의 증가분을 적으시오.

(b) 다음의 분산분석표를 작성하시오.

요인	제곱합	자유도	평균제곱	F비
회귀				
오차				
전체				

(c) 잔차표준오차 s값을 구하시오.

(d) 결정계수 R^2 값을 구하고 그 의미를 설명하시오.

(e) (b)에서 계산한 F 통계치를 사용하여 검정을 실시하시오. (유의수준 $\alpha = 0.05$을 사용할 것)

i. 귀무가설과 대립가설

ii. 검정통계량의 분포

iii. 기각역

iv. 검정통계량의 관측값

v. 결론