UNIVERSITY OF TORONTO
Faculty of Arts and Science

December 2016 EXAMINATIONS

STA302/1001H1F

Duration - 3 hours

Examination Aids: Scientific Calculator

STA 302/1001	Last Name (Print): ______________________
Fall 2016	
Final Exam	First Name (Print): ______________________
12/12/2016	
Time Limit: 3 hours	Student Number: ______________________

Check TWO: STA302 □ STA1001 □ L0101 □L0201 □L0501 □

This exam contains 16 pages (including this cover page) and 6 problems. Check to see if any pages are missing. Enter all requested information on the top of this page.

- This is a closed-book exam. You are only allowed to use a scientific calculator and the formulae from the last page of the exam.
- SLR stands for 'Simple Linear Regression';
  MLR stands for 'Multiple Linear Regression';
  MLE stands for 'Maximum Likelihood Estimator';
  LSE stands for 'Least Squares Estimator'.
- You are required to show your work on each problem on this exam. Please carry all possible precision through a numerical question, and give your final answer to four (4) decimals, unless they are trailing zeroes or otherwise indicated.
- You may use a benchmark of $\alpha = 5\%$ for all inference, unless otherwise indicated.
- Do not write in the table to the right.

Problem	Points	Score
1	10	
2	15	
3	10	
4	15	
5	30	
6	20	
Total:	100	

1. (10 points) **Multiple Choice:** Answer the following questions by circling all *correct* answers.

   I. Circling all correct statement(s) about the probability distributions of b_1 and β_1 in a SLR model:

      A. Both are Normally distributed.

      B. $b_1 \sim N(0, \sigma^2)$, no distribution for β_1 since it is non-random.

      C. $b_1 \sim N(\beta_1, \sigma^2 / \sum_i (X_i - \bar{X})^2)$, $\beta_1 \sim N(0, \sigma^2 / \sum_i (X_i - \bar{X})^2)$

      D. $b_1 \sim N(\beta_1, \sigma^2 / \sum_i (X_i - \bar{X})^2)$ and no distribution for β_1 since it is non-random.

   II. Which statistic in the following is used to identify problems of multicollinearity.?

      A. Cook's Distance

      B. DFBETA

      C. Adjusted R-squared

      D. Variance Inflation Factor

   III. For a SLR model, what are the least assumptions we need to show that the ordinary least squares estimator (OLS) is a BLUE (best linear unbiased estimator):

      A. The linear form, $Y_i = \beta_0 + \beta_1 X_i + \epsilon_i, \forall i$.

      B. $E(\epsilon_i) = 0, \forall i$

      C. $Cov(\boldsymbol{\epsilon}) = \sigma^2 \mathbf{I}$.

      D. $\epsilon_i \sim N(0, \sigma^2)$ independently. note error doesnt have to be normal under gauss

   IV. Which of the following is necessarily true of high leverage points?

      A. They have a large effect on the slope.

      B. They are far from the sample mean of X.

      C. They are outliers or influential points.

      D. They make R^2 higher.

   V. Which of the following statistics are not influence metrics?

      A. Residuals

      B. DFFITS

      C. DFBETAS

      D. Cook's distance

   VI. Circling all correct statements in the following:

      A. The LSE of slope and intercept in a SLR model are uncorrelated.

      B. In linear regression, the MLE and the LSE are the same for regression coefficients estimation.

      C. LSE are BLUE, there are no estimators with lower variance than the LSE.

      D. LSE are considered as linear estimators.

   VII. A transformation on Y does NOT help in which of the following cases?

      A. Non-constant variance.

      B. Non-Normal residuals. Text

      C. Correlation between residuals.

      D. A non-linear relationship between X and Y.

2. (15 points) Short answer questions.

(2.a) (2 pts) To obtain the least squares estimators of $\boldsymbol{\beta}$ in a MLR model, the error terms $\epsilon_i, i = 1, \ldots, n$, must be I.I.D. $N(0, \sigma^2)$ distributed. Is this statement true or false, give a brief and clear justification of your answer.

False.

To obtain the least squares estimators of $\boldsymbol{\beta}$ in a MLR model, the only assumption we need is the model form of $\mathbf{Y} = \boldsymbol{X\beta + \epsilon}$.

(2.b) (3 pts) In SLR setting, what is the probability distribution of $b_0 + b_1\bar{X}$? (b_0, b_1 are the least squares estimators of β_0, β_1 respectively).

$b_0 + b_1\bar{X} \sim N(\beta_0 + \beta_1\bar{X}, \sigma^2/n)$

- Both b_0 and b_1 are linear combination of $\mathbf{Y}$ and normally distributed, so is $b_0 + b_1\bar{X}$.
- $E(b_0 + b_1\bar{X}) = E(b_0) + E(b_1)\bar{X} = \beta_0 + \beta_1\bar{X}$.
- $Var(b_0 + b_1\bar{X}) = Var(\bar{Y}) = \sigma^2/n$

(2.c) (2 pts) Residuals $e_i, i = 1, \ldots, n$, are independent. True or false, justify your answer.

False.

$$Cov(\mathbf{e}) = \sigma^2\textbf{(I-H)}$$

This implies that for $i \neq j$

$$Cov(e_i, e_j) = -\sigma^2 h_{ij} = -\sigma^2\Big(\frac{1}{n} + \frac{(X_i - \bar{X})(X_j - \bar{X})}{\sum_{k=1}^{n}(X_k - \bar{X})^2}\Big) \neq 0$$

The covariance is close to 0 when n is very large.

(2.d) (4 pts) For MLR model in matrix form, is it true that $\boldsymbol{e^T\hat{Y}}$=0 where **e** is the column vector of residuals and $\boldsymbol{\hat{Y}}$ is the column vector of fitted values? (Give a clear justification of your answer.)

True.

$$\begin{aligned}
\mathbf{\hat{Y}} &= \mathbf{Xb} = \mathbf{HY} \\
\boldsymbol{e^T}\mathbf{\hat{Y}} &= [\textbf{(I-H)Y}]^T HY \\
&= \boldsymbol{Y^T(I-H)^T} HY \\
&= \boldsymbol{Y^T(I-H)HY} \\
&= \boldsymbol{Y^T(H-HH)Y} \\
&= \boldsymbol{Y^T(H-H)Y} \\
&= 0
\end{aligned}$$

(2.e) (2 pts) Is R^2 always greater than adjusted R^2? Explain.

True.

$$\begin{aligned}
R^2 &= 1 - \frac{SSE}{SSTO} = 1 - \frac{n-1}{n-1}\frac{SSE}{SSTO} \\
R^2_{adj} &= 1 - \frac{n-1}{n-p}\frac{SSE}{SSTO}
\end{aligned}$$

Since $\frac{n-1}{n-p} \geq 1$ so $R^2 \geq R^2_{adj}$

(2.f) (2 pts) In a SLR model, if we increase the standard deviation of X's, we would get a more accurate estimate of the slope. Is this statement true? Justify your answer.

True.

Increasing the standard deviation of X's is the same as increasing $S_{xx} = \sum_i^n (X_i - \bar{X})^2$. And since $Var(b_1) = \sigma^2/S_{xx}$, so the larger S_{xx}, the smaller $Var(b_1)$. i.e, we have more accurate estimate of the slope.

3. (10 points) Answer the following questions for a simple linear regression model (SLR).

(3.a) (5 pts) In a SLR model, SSR=$\sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2$, and $MSR = SSR/1$. Show that

$$E(MSR) = \sigma^2 + \beta_1^2 \sum_{i=1}^{n}(X_i - \bar{X})^2$$

and explain how this is related to the construction of the analysis of variance F-test.

$\hat{Y}_i = b_0 + b_1 X_i = (\bar{Y} - b_1\bar{X}) + b_1 X_i = \bar{Y} + b_1(X_i - \bar{X})$

$$SSR = \sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2 = \sum_{i=1}^{n}(\bar{Y} + b_1(X_i - \bar{X}) - \bar{Y})^2 = b_1^2 \sum_{i=1}^{n}(X_i - \bar{X})^2$$

$$\begin{aligned} E(MSR) = E(SSR/1) &= \sum_{i=1}^{n}(X_i - \bar{X})^2 E(b_1^2) = \sum_{i=1}^{n}(X_i - \bar{X})^2 (Var(b_1) + E(b_1)^2) \\ &= \sum_{i=1}^{n}(X_i - \bar{X})^2 (\sigma^2 / \sum_{i=1}^{n}(X_i - \bar{X})^2 + \beta_1^2) \\ &= \sigma^2 + \beta_1^2 \sum_{i=1}^{n}(X_i - \bar{X})^2 \end{aligned}$$

We know that MSE is an unbiased estimator of σ^2. If $\beta_1 = 0$ then MSR is also an unbiased estimator of σ^2. This implies that the ratio of MSR and MSE close to 1 suggests β_1 is 0 and far greater than 1 in favour of $\beta_1 \neq 0$. That is, using the F test statistic, $F^* = MSR/MSE$, we can test

$$H_0 : \beta_1 = 0 \qquad vs \qquad H_a : \beta_1 \neq 0$$

We reject H_0 when F^* is large.

(3.b) (5 pts) State the SLR model in matrix form, defining all matrices and vectors. Include the standard Normal error assumption.

$$\mathbf{Y}_{n\times 1} = \mathbf{X}_{n\times 2}\boldsymbol{\beta}_{2\times 1} + \boldsymbol{\epsilon}_{n\times 1}$$

That is

$$\begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix} = \begin{bmatrix} 1 & X_1 \\ 1 & X_2 \\ \vdots & \\ 1 & X_n \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} + \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix}$$

The standard Normal error assumption is $\boldsymbol{\epsilon}_{n\times 1} \sim N(\mathbf{0}_{n\times 1}, \sigma^2 \mathbf{I}_{n\times n})$ where $\mathbf{I}_{n\times n}$ is the n by n identity matrix.

4. (15 points) Answer the following questions for a multiple linear regression model (MLR).

(4.a) For the multiple linear regression model $\mathbf{Y}$=$\mathbf{X}\boldsymbol{\beta}+\boldsymbol{\epsilon}$, the least squares estimators are $\mathbf{b} = \boldsymbol{(X'X)^{-1}X'Y}$ and the residuals are $\mathbf{e} = \mathbf{Y}$-$\mathbf{Xb}$. We further assume that $\boldsymbol{\epsilon} \sim N(\mathbf{0}, \sigma^2\mathbf{I})$.

- (2 pts) Show variance-covariance of $\mathbf{b}$: Var($\mathbf{b}$)=$\sigma^2\boldsymbol{(X'X)^{-1}}$

$$\begin{aligned} Var(\mathbf{b}) &= Var(\boldsymbol{(X'X)^{-1}X'Y}) \\ &= \boldsymbol{(X'X)^{-1}X'}Var(Y)[\boldsymbol{(X'X)^{-1}X'}]' \\ &= \sigma^2\boldsymbol{(X'X)^{-1}X'IX(X'X)^{-1}} \\ &= \sigma^2\boldsymbol{(X'X)^{-1}} \end{aligned}$$

- (2 pts) Show SSR=$\sum_{i=1}^n(\hat{Y}_i - \bar{Y})^2 = \sum_{i=1}^n \hat{Y}_i^2 - n\bar{Y}^2 = \boldsymbol{Y'(H - \frac{1}{n}J)Y}$ where $\mathbf{J} = \mathbf{11'}$ is a matrix of 1 everywhere.

$$SSR = \sum_{i=1}^n(\hat{Y}_i - \bar{Y})^2 = \sum_{i=1}^n(\hat{Y}_i^2 - 2\bar{Y}\hat{Y}_i + \bar{Y}^2)$$

$$\begin{aligned} SSR = \sum_{i=1}^n(\hat{Y}_i - \bar{Y})^2 &= \sum_{i=1}^n(\hat{Y}_i^2 - 2\bar{Y}\hat{Y}_i + \bar{Y}^2) \\ &= \sum_{i=1}^n\{\hat{Y}_i^2 - 2\bar{Y}[\bar{Y} + b_1(X_i - \bar{X})] + \bar{Y}^2\} \\ &= \sum_{i=1}^n \hat{Y}_i^2 - n\bar{Y}^2 + b_1\sum_{i=1}^n(X_i - \bar{X}) \\ &= \sum_{i=1}^n \hat{Y}_i^2 - (\sum_i Y_i)^2/n \\ &= \boldsymbol{(HY)'(HY)} - \frac{1}{n}\boldsymbol{(1'Y)'1'Y} \\ &= \boldsymbol{Y'HY} - \boldsymbol{Y'\frac{1}{n}JY} = \boldsymbol{Y'(H - \frac{1}{n}J)Y} \end{aligned}$$

- (4 pts) Show $\mathbf{e}$=$\mathbf{(I\text{-}H)}\boldsymbol{\epsilon}$ and Var($\mathbf{e}$)=$\mathbf{(I\text{-}H)}\boldsymbol{\sigma^2}$ where $\mathbf{H} = \boldsymbol{X(X'X)^{-1}X'}$.

$$\mathbf{e} = \mathbf{Y} - \boldsymbol{\hat{Y}} = \boldsymbol{(X\beta + \epsilon) - HY} = \boldsymbol{(X\beta + \epsilon) - H(X\beta + \epsilon) = \epsilon - H\epsilon} = \mathbf{(I\text{-}H)}\boldsymbol{\epsilon}$$

$$\begin{aligned} Var(\mathbf{e}) &= Var(\mathbf{(I\text{-}H)}\boldsymbol{\epsilon}) = \mathbf{(I\text{-}H)}Var(\boldsymbol{\epsilon})\mathbf{(I\text{-}H)}' \\ &= \sigma^2(I - H)(I - H) = \sigma^2(I - H)(I - H) \\ &= \sigma^2(I - H - H + HH) = \sigma^2(I - 2H + H) \\ &= \sigma^2(I - H) \end{aligned}$$

(4.b) We will now generalize the MLR model to include the case where the variance-covariance matrix of $\boldsymbol{\epsilon}$ is the $n \times n$ matrix Σ (no restriction on Σ and it is a valid variance-covariance matrix). We will assume that $E(\boldsymbol{\epsilon}) = 0$. To obtain the generalized least square estimator, the quantity, $Q = \boldsymbol{(Y - X\beta)'\Sigma^{-1}(Y - X\beta)}$ is minimized with respect to $\boldsymbol{\beta}$.

- (5 pts) Show $\mathbf{b} = \boldsymbol{(X'\Sigma^{-1}X)^{-1}X'\Sigma^{-1}Y}$

$$
\begin{aligned}
Q &= \boldsymbol{(Y - X\beta)'\Sigma^{-1}(Y - X\beta)} \\
&= \boldsymbol{(Y' - \beta'X')\Sigma^{-1}(Y - X\beta)} \\
&= \boldsymbol{Y'\Sigma^{-1}Y - Y'\Sigma^{-1}X\beta + \beta'X'\Sigma^{-1} + \beta'X'\Sigma^{-1}X\beta}
\end{aligned}
$$

$$
\begin{aligned}
\Rightarrow \partial Q/\partial\boldsymbol{\beta} &= \boldsymbol{-Y'\Sigma^{-1}X - (X'\Sigma^{-1}Y)' + 2\beta'X'\Sigma^{-1}X} \\
&= \boldsymbol{-Y'\Sigma^{-1}X + 2\beta'X'\Sigma^{-1}X}
\end{aligned}
$$

Note that Σ is symmetric matrix, so is Σ^{-1}.

$$
\frac{\partial Q}{\partial\boldsymbol{\beta}} = 0 \Leftrightarrow \boldsymbol{Y'\Sigma^{-1}X = b'X'\Sigma^{-1}X} \Leftrightarrow \boldsymbol{X'\Sigma^{-1}Y} = \boldsymbol{X'\Sigma^{-1}Xb}
$$

From last equality, we have $\mathbf{b} = \boldsymbol{(X'\Sigma^{-1}X)^{-1}X'\Sigma^{-1}Y}$.

- (2 pts) In this case, $\mathbf{H} = \boldsymbol{X(X'\Sigma^{-1}X)^{-1}X'\Sigma^{-1}}$, show $\mathbf{H}$ is idempotent.

$$
\begin{aligned}
\mathbf{HH} &= \boldsymbol{X(X'\Sigma^{-1}X)^{-1}X'\Sigma^{-1}\ X(X'\Sigma^{-1}X)^{-1}X'\Sigma^{-1}} \\
&= \boldsymbol{X(X'\Sigma^{-1}X)^{-1}X'\Sigma^{-1}} \\
&= H
\end{aligned}
$$

So we conclude that in this case $\mathbf{H}$ is idempotent.

5. (30 points) Analysis of a data set which consists of 654 observations on children with age from 3 to 19. Forced Expiratory Volume (FEV), which is a measure of lung capacity, is the variable in interest. Age and height are two continuous predictors.

```
> summary(mod)

Call:
lm(formula = log(fev) ~ log(age) + height, data = a2)

Residuals:
     Min       1Q   Median       3Q      Max
-0.62937 -0.08648  0.01346  0.09536  0.44077

                                       B = -2.17 / A
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -2.170548    (A)        (B)  < 2e-16 ***
log(age)     0.194570    (C)        (D) 3.65e-09 ***
height       0.043314    (E)        (F)  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

                                G = sqrt(MSE) = sqrt(0.022)
Residual standard error: (G) on 651 degrees of freedom
Multiple R-squared:  0.8063,Adjusted R-squared:  (H)     H = 1 - mse/mst = 1 - 0.022 /
F-statistic:  (I) on (J) and 651 DF,  p-value: < 2.2e-16   ((45.753+12.721+14.052) / 653) = 0.80
                               J = df_ssreg = 2
> anova(mod)  I = msreg / mse = ((45.753 + 12.721) / 2) / 0.022 =  1328
Analysis of Variance Table

Response: log(fev)
           Df Sum Sq Mean Sq F value    Pr(>F)
log(age)    1 45.753  45.753 2119.72 < 2.2e-16 ***
height      1 12.721  12.721  589.37 < 2.2e-16 ***
Residuals 651 14.052   0.022
---        p = 2, df_rss = n -p-1 = 651 n = 654
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

> sqrt(diag(vcov(mod)))  # find square root of diagonal elements on  MSE*(X'X)^(-1)
(Intercept)    log(age)      height
 0.06415289  0.03252900    0.00178416
```

(5.a) (10 pts) Some values have been replaced with letters (A through J) in above **R** output, fill in those values.

$A = 0.06415289$ $B = -2.170548/0.06415289 = -33.83399$

$C = 0.03252900$ $D = 0.194570/0.03252900 = 5.981432$

$E = 0.00178416$ $F = 0.043314/0.00178416 = 24.27697$

$G = \sqrt{0.022} = 0.148324$ $H = 1 - \frac{MSE * 653}{SSTO} = 0.8019193$

$I = MSR/MSE = 1328.955$ $J = 2$

(5.b) (2 pts) Write down the estimated regression model.

$$\widehat{\log(fev)} = -2.170548 + 0.194570\ \log(age) + 0.043314\ height$$

(5.c) (2 pts) Interpret the meaning of the slope of height in terms of the original variables.

It is estimated that, on average, FEV increases by 4.4266% ($e^{0.043314} - 1$) with each one unit increases in height.

Or, it is estimated that, on average, FEV will increases by a factor of 1.044266 ($e^{0.043314}$) with each one unit increases in height.

(5.d) (4 pts) Find the simultaneous confidence intervals for the 3 regression coefficients with family confidence coefficients at $1 - 5\%$. Use the Bonferroni method. Choose the correct critical values in the following

$$t_{1-0.05/6,651} = 2.400; \quad t_{1-0.05/3,651} = 2.132; \quad t_{1-0.05/2,651} = 1.963$$

General formula:

$$b_k \pm t_{1-0.05/6,651} s(b_k) = b_k \pm 2.4 s(b_k)$$

$$\beta_0 : -2.170548 \pm 2.4 * 0.064153 = (-2.32452631,\ -2.01656893)$$

$$\beta_1 : 0.194570 \pm 2.4 * 0.032529 = (0.11649481,\ 0.27264593)$$

$$\beta_2 : 0.043314 \pm 2.4 * 0.001784 = (0.03903146,\ 0.04759608)$$

(5.e) (2 pts) What does it mean for the intervals in (5.d) to be "simultaneous"?

Simultaneously, all the confidence limits capture the their respective true values of β_0, β_1 and β_2 at least 95% of the time in repeated samples of size 654.

(5.f) (2 pts) The Bonferroni method is "conservative". Explain what this means in relation to your answer to (5.d).

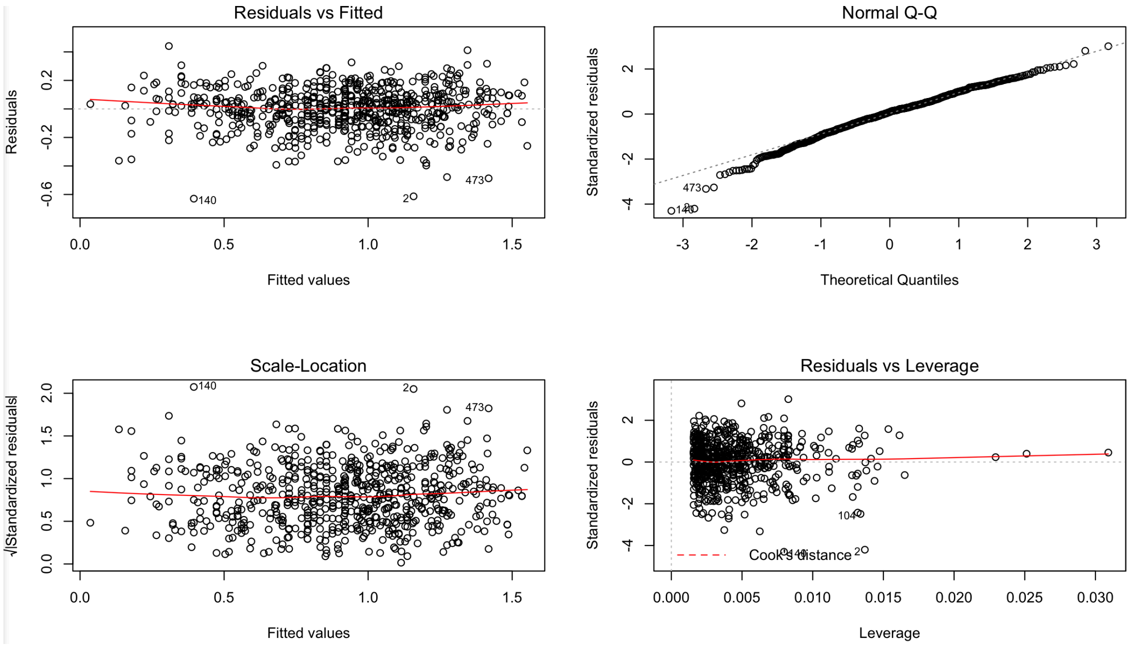
The probability that all CIs constructed in the manner in (5.d) capture the true of β_0, β_1 and β_2 is **at least** of 95% (**probably more**).

(5.g) (4 pts) For p value with 3.65e-09 in the summary output, what are the null and alternative hypotheses? What is the test statistic and what do you conclude?

$$\log(fev) = \beta_0 + \beta_1 \ \log(age) + \beta_2 \ height + \epsilon$$

The null and alternative hypotheses $H_0 : \beta_1 = 0 \ vs \ H_a : \beta_1 \neq 0$, and the test statistics: $t^* = b_1/s(b_1)$ where $s^2(b_1)$ is the second element on the diagonal of $MSE(X^TX)^{-1}$. The p-value for this test is less than 0.001, so we have strong evidence that the slope of $\log(age)$ is nonzero.

(5.h) (2 pts) The diagnostics for the fitted model is given in the following plot. Does the linearity, constant-variance and normality of error terms look fine? Does there exist any influential points?



From the residual and scale-location plots, both suggest that the linearity and constant-variance assumptions are satisfied. From the normal QQ-plot, we observed heavy left tail with several observations that have large residuals, but overall the normality assumption looks fine.

From the residual versus leverage plot, we don't see any observations lie out of Cook distance of 1, we conclude that we do not have influential points in the data.

(5.i) (2 pts) Compare the residual plot and the scale-location plot,what's the difference between residual and standardized residual?

Residual is denoted by e_i which is defined as the vertical distance between observation and fitted value. i.e. $e_i = Y_i - \hat{Y}_i$. While the standardized residual is defined as

$$e_i^* = \frac{e_i}{\sqrt{MSE}}$$

6. (20 points) Duncan's Occupational Prestige Data: A data includes the prestige and other characteristics of 45 U. S. occupations in 1950. Variables in the data:

- type: types of occupation: bc= blue-collar; prof=professional; wc=white-colloar.
- income: percent of males in occupation earning 3500 or more in 1950.
- education: percent of males in occupation in 1950 who were high-school graduates.
- prestige: percent of raters in NORC study rating occupation as excellent or good in prestige.

A: type
B: type + income
C: type + income + education

Three models are fitted to the data:

$$model\ A : prestige = \beta_{0A} + \beta_{1A}\ I_{Prof} + \beta_{2A}\ I_{wc} + \epsilon$$

$$model\ B : prestige = \beta_{0B} + \beta_{1B}\ I_{Prof} + \beta_{2B}\ I_{wc} + \beta_{3B}\ income + +\epsilon$$

$$model\ C : prestige = \beta_{0C} + \beta_{1C}\ education + \beta_{2C}\ I_{Prof} + \beta_{3C}\ I_{wc} + \beta_{4C}\ income + \epsilon$$

The estimated models for above from R are in the following.

```
> with(Duncan,tapply(prestige,type,mean))
      bc     prof       wc
22.76190 80.44444 36.66667

## =========== Model A ===========##
> summary(modelA)
Call:
lm(formula = prestige ~ type, data = Duncan)

Coefficients:   A = 22.761
              Estimate Std. Error t value Pr(>|t|)
(Intercept)    (A)      3.466    (B)     6.08e-08 ***
typeprof       57.683    5.102   11.305 2.54e-14 ***
typewc         (C)      7.353    (D)     0.0655 .
---          C=  36.66 - 22.76 = 13.9
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 15.88 on 42 degrees of freedom
Multiple R-squared:  0.7574,Adjusted R-squared:  0.7459
F-statistic: 65.57 on 2 and 42 DF,  p-value: 1.207e-13

> anova(modelA)
Analysis of Variance Table

Response: prestige
          Df  Sum Sq Mean Sq F value    Pr(>F)
type       2  33090 16545.0  65.571 1.207e-13 ***
Residuals 42  10598   252.3
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
## =========== Model B ===========##
> summary(modelB)
Call:
lm(formula = prestige ~ type + income, data = Duncan)

Coefficients:
            Estimate Std. Error t value      Pr(>|t|)
(Intercept)  6.70386    3.22408   2.079        0.0439 *
typeprof    33.15567    4.83190   6.862 0.00000002583 ***
typewc      -4.27720    5.54974  -0.771        0.4453
income       0.67579    0.09377   7.207 0.00000000843 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 10.68 on 41 degrees of freedom
Multiple R-squared:  0.893,Adjusted R-squared:  0.8852
F-statistic:   114 on 3 and 41 DF,  p-value: < 2.2e-16

> anova(modelB)
Analysis of Variance Table

Response: prestige
          Df Sum Sq Mean Sq F value         Pr(>F)
type       2  33090 16545.0 145.095      < 2.2e-16 ***
income     1   5922  5922.4  51.938 0.000000008428 ***
Residuals 41   4675   114.0
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

## =========== Model C ===========##
> summary(modelC)
Call:
lm(formula = prestige ~ income + type + education, data = Duncan)

Coefficients:
             Estimate Std. Error t value     Pr(>|t|)
(Intercept)  -0.18503    3.71377  -0.050      0.96051
income        0.59755    0.08936   6.687 0.0000000512 ***
typeprof     16.65751    6.99301   2.382      0.02206 *
typewc      -14.66113    6.10877  -2.400      0.02114 *
education     0.34532    0.11361   3.040      0.00416 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 9.744 on 40 degrees of freedom
Multiple R-squared:  0.9131,Adjusted R-squared:  0.9044
F-statistic:   105 on 4 and 40 DF,  p-value: < 2.2e-16
```

```
> anova(modelC)
Analysis of Variance Table

Response: prestige
          Df  Sum Sq Mean Sq  F value    Pr(>F)
income     1 30664.8 30664.8 322.9617 < 2.2e-16 ***
type       2  8347.6  4173.8  43.9585 7.991e-11 ***
education  1   877.2   877.2   9.2388  0.004164 **
Residuals 40  3798.0    94.9
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

(6.a) (1 pts) Type of occupations is a factor variable with 3 levels, how many dummy variables do we need to distinguish them?

$$answer = 2$$

(6.b) (4 pts) Fill in those 4 missing values in the summary output of model A (A through D).

$A = \underline{22.76190}$ $B = \underline{6.567196}$

$C = \underline{13.90477}$ $D = \underline{1.891034}$

(6.c) (2 pts) Let μ_{bc} be the mean of prestige when occupation type is blue-collar and μ_{prof} be the mean of prestige when occupation type is professional. Want to test

$$H_0 : \mu_{prof} - \mu_{bc} = 0 \quad vs \quad H_a : \mu_{prof} - \mu_{bc} \neq 0$$

What is the equivalent test in model A output? What can you conclude?

It is equivalent to the test of

$$H_0 : \beta_{1A} = 0 \quad vs \quad H_a : \beta_{1A} \neq 0$$

The p-value for the t-test is less than 0.001, we conclude that we have strong evidence that the mean difference between the professional and blue-collar groups is non-zero.

(6.d) (3 pts) For the F statistic with observed value 114 on 3 and 41 DF in the summary output of model B, what are the null and alternative hypotheses? What is the test statistic and what do you conclude?

It is equivalent to the test of

$$H_0 : \beta_{1B} = \beta_{2B} = \beta_{3B} = 0 \quad vs \quad H_a : not\ all\ \beta_k\ in\ H_0\ are\ zero.$$

$$\text{Test statistic:} \quad F^* = \frac{MSR}{MSE}$$

The p-value for the F-test is less than 0.001, we conclude that we have strong evidence that at least one of β_k as specified in the null hypothesis is nonzero.

(6.e) (2 pts) Find the extra sum of squares, $SSR(income|type)$ and the associated degree of freedom of it?

$$SSR(income|type) = \underline{5922} \qquad d.f. = \underline{1}$$

(6.f) (4 pts) Perform a partial F-test to test the hypothesis that education and income are useful predictors given the type of occupation is already in the model. If you cannot perform this test, state what you are missing in order to do it. If you can, give a test statistic (with df), using the following given information and make a conclusion in words.

$$F_{0.95,1,40} = 4.084746;\ F_{0.95,2,40} = 3.231727;\ F_{0.95,3,40} = 2.838745;\ F_{0.95,4,40} = 2.605975$$

$$F_{0.95,1,40} = 4.084746;\ F_{0.95,2,40} = 3.231727;\ F_{0.95,3,40} = 2.838745;\ F_{0.95,4,40} = 2.605975$$

$$F_{0.975,1,40} = 5.423937;\ F_{0.975,2,40} = 4.050992;\ F_{0.975,3,40} = 3.463260;\ F_{0.975,4,40} = 3.126114$$

$$F_{0.975,1,42} = 5.403859;\ F_{0.975,2,42} = 4.032710;\ F_{0.975,3,42} = 3.445689;\ F_{0.975,4,42} = 3.108870$$

Solution:

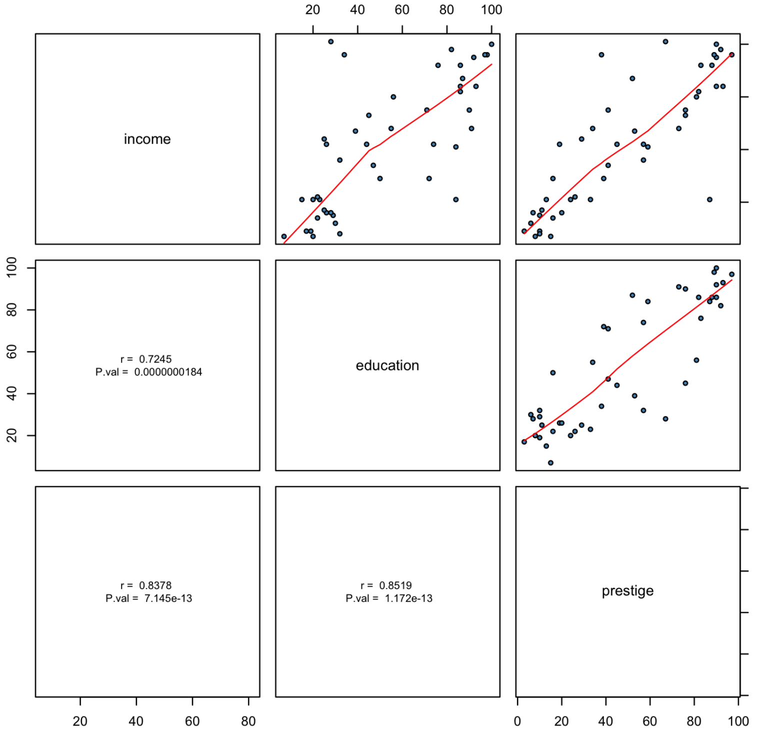
$$\begin{aligned} F^* &= \frac{SSR(education, income|type)/2}{SSE/(n-5)} = \frac{(SSR_C - SSR_A)/2}{SSE_C/40} \\ &= \frac{[(30664.8 + 8347.6 + 877.2) - 33090]/2}{94.9} \\ &= 35.825079 \end{aligned}$$

Since $35.825079 > 3.231727 = F_{0.95,2,40}$, so we reject the null hypothesis and conclude that education and income are useful predictors given the type of occupation is already in the model.

(6.g) (2 pts) Is it possible to perform a partial F-test to test the hypothesis that income and type interact together to predict prestige? That is, testing whether the coefficient of income*type is zero or not. If you can, give a test statistic (with df), using the above given information and make a conclusion in words. If you cannot perform this test, state what you are missing in order to do it.

Solution: No.
In order to perform the partial F-test to test the hypothesis that income and type interact together to predict prestige. We need to run a model with **type**, **income** and **type*income** as predictor variables in the MLR model to predict prestige.

(6.h) (2 pts) A pairwise scatter plot and correlation test are performed among all the non-quantitative variables in the data. Given the extra information, do you have any concerns about model C?



Answer:
Multicollinearity exists in model C since both income and education variables are correlated not just to prestige (the dependent variable), but also to strongly correlated $(r = 0.7245, p.val < 0.001)$ to each other.

If severe multicollinearity exists, the coefficient estimates are unstable and it is difficult to interpret them. It can also increase the variance of the coefficient estimates and make the estimates very sensitive to minor changes in the model and might cause the coefficients to switch signs, and makes it more difficult to specify the correct model.

Further investigation such as VIF calculation is needed.

Some formulae (SLR and MLR):

$$b_1 = \frac{\Sigma(X_i - \bar{X})(Y_i - \bar{Y})}{\Sigma(X_i - \bar{X})^2} = \frac{\Sigma X_i Y_i - n\bar{X}\bar{Y}}{\Sigma X_i^2 - n\bar{X}^2} \qquad b_0 = \bar{Y} - b_1\bar{X}$$

$$Var(b_1) = \frac{\sigma^2}{\Sigma(X_i - \bar{X})^2} \qquad Var(b_0) = \sigma^2\left(\frac{1}{n} + \frac{\bar{X}^2}{\Sigma(X_i - \bar{X})^2}\right)$$

$$Var(\hat{Y}_h) = \sigma^2\left(\frac{1}{n} + \frac{(X_h - \bar{X})^2}{\Sigma(X_i - \bar{X})^2}\right) \qquad \sigma^2\{pred\} = Var(Y_h - \hat{Y}_h) = \sigma^2\left(1 + \frac{1}{n} + \frac{(X_h - \bar{X})^2}{\Sigma(X_i - \bar{X})^2}\right)$$

$$SSTO = \Sigma(Y_i - \bar{Y})^2 \qquad SSE = \Sigma(Y_i - \hat{Y}_i)^2 \qquad SSR = \Sigma(\hat{Y}_i - \bar{Y})^2 = b_1^2\Sigma(X_i - \bar{X})^2$$

$$r = \frac{\Sigma(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\Sigma(X_i - \bar{X})^2\Sigma(Y_i - \bar{Y})^2}} \qquad Cov(b_0, b_1) = -\frac{\sigma^2\bar{X}}{\Sigma(X_i - \bar{X})^2}$$

Working-Hotelling coefficient: $W = \sqrt{pF(1-\alpha; p, n-p)}$

$$R^2 = \frac{SSR}{SSTO} = 1 - \frac{SSE}{SSTO} \qquad R^2_{adj} = 1 - \frac{(n-1)MSE}{SSTO}$$

$$\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y} \qquad Cov(\mathbf{b}) = \sigma^2(\mathbf{X}'\mathbf{X})^{-1}$$

$$\hat{\mathbf{Y}} = \mathbf{Xb} = \mathbf{HY} \qquad \mathbf{e} = \mathbf{Y} - \hat{\mathbf{Y}} = (\mathbf{I} - \mathbf{H})\mathbf{Y}$$

$$\mathbf{H} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-\mathbf{1}}\mathbf{X}' \qquad SSE = \mathbf{Y}'(\mathbf{I} - \mathbf{H})\mathbf{Y}$$

$$SSR = \mathbf{Y}'\left(\mathbf{H} - \frac{1}{n}\mathbf{J}\right)\mathbf{Y} \qquad SSTO = \mathbf{Y}'\left(\mathbf{I} - \frac{1}{n}\mathbf{J}\right)\mathbf{Y}, J = \mathbf{11'}$$

$$\sigma^2\{\hat{Y}_h\} = \sigma^2\mathbf{X_h'}(\mathbf{X}'\mathbf{X})^{-\mathbf{1}}\mathbf{X_h} \qquad \sigma^2\{pred\} = \sigma^2\left(1 + \mathbf{X_h'}(\mathbf{X}'\mathbf{X})^{-\mathbf{1}}\mathbf{X_h}\right)$$