UNIVERSITY OF TORONTO
Faculty of Arts and Science

December 2016 EXAMINATIONS

STA302/1001H1F

Duration - 3 hours

Examination Aids: Scientific Calculator

STA 302/1001
Fall 2016
Final Exam
12/12/2016
Time Limit: 3 hours

Last Name (Print): ____________________

First Name (Print): ____________________

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.

   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.

      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
only assumption is in the form of Y = Xb + e

(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).

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

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

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

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

yes

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.

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

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 $\mathbf{e}=\mathbf{Y}\text{-}\mathbf{Xb}$. We further assume that $\boldsymbol{\epsilon}\sim N(\mathbf{0},\sigma^2\mathbf{I})$.

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

- (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.

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

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

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

Coefficients:              A = 0.064
             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

Residual standard error: (G) on 651 degrees of freedom
Multiple R-squared:  0.8063,Adjusted R-squared:  (H)
F-statistic:  (I) on (J) and 651 DF,  p-value: < 2.2e-16

> anova(mod)
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
---
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                 variance covariance matrix of betas
 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.

$A =$ ____________________ $B =$ ____________________

$C =$ ____________________ $D =$ ____________________

$E =$ ____________________ $F =$ ____________________

$G =$ ____________________ $H =$ ____________________

$I =$ ____________________ $J =$ ____________________

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

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

with age constant, each k-fold increase in height, results in e^0.04 factor change in y

(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 value in the following

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

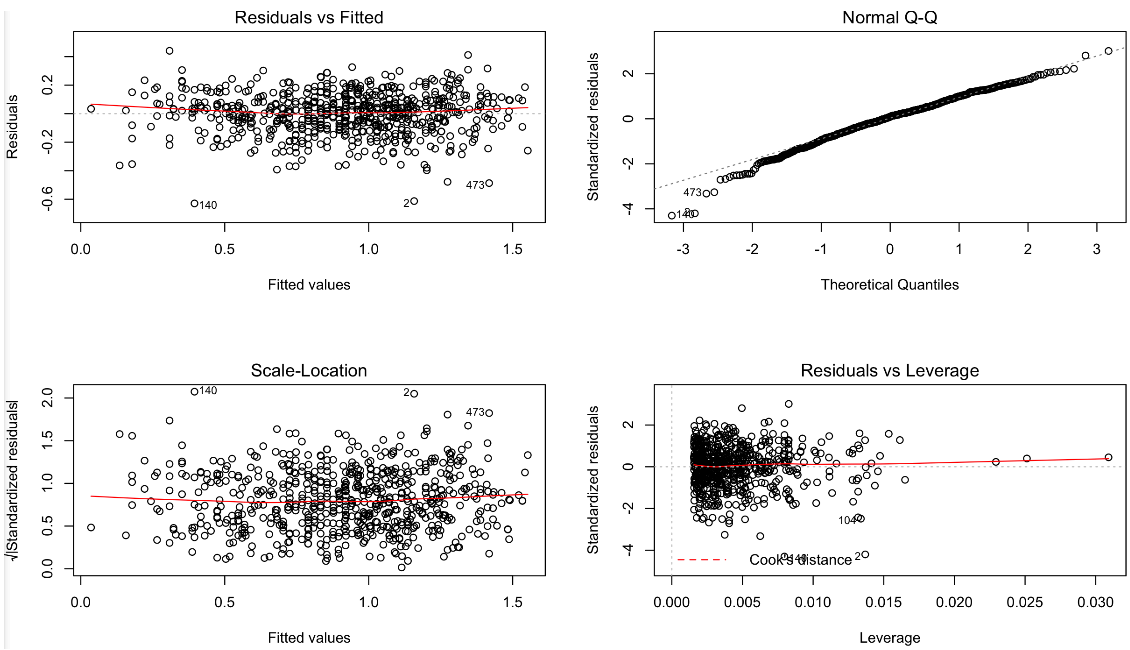
(5.e) (2 pts) What does it mean for the intervals in (5.d) to be "simultaneous"?

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

(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?

H_0: beta_1 = 0
test statistic = D
t = betahat / se(betahat) , se(betahat) is second elemnt on diagonal of mse(X'X)^-1
conclude reject null, slope for 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?



linearity: residual plot scatter around the mean
constant variance: yes. residual plot shows random scatter
normality: yes qqplot, there is some unusual point i guess, heavy left tail, but overall normality assumption looks fine
influential: all within cook's distance

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

residual plot higher variance around mean of fitted value
because var(residual) = mse * (1 - h_ii), higher leverage, away from the mean, lower variance
the scale-location is spread out more evenly

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.

   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:
            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 .
---
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?

2

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

$A =$ __________ $B =$ __________

$C =$ __________ $D =$ __________

(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?

(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?

H_0: beta_k = 0 H_alpha: any of beta_k != 0
F^* = msreg/mse

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

RSS_{type} - RSS_{type + income} = 10598 - 4675 = 5923

1

$SSR(income|type) =$ ____________________ $d.f. =$ ____________________

(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.

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

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

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

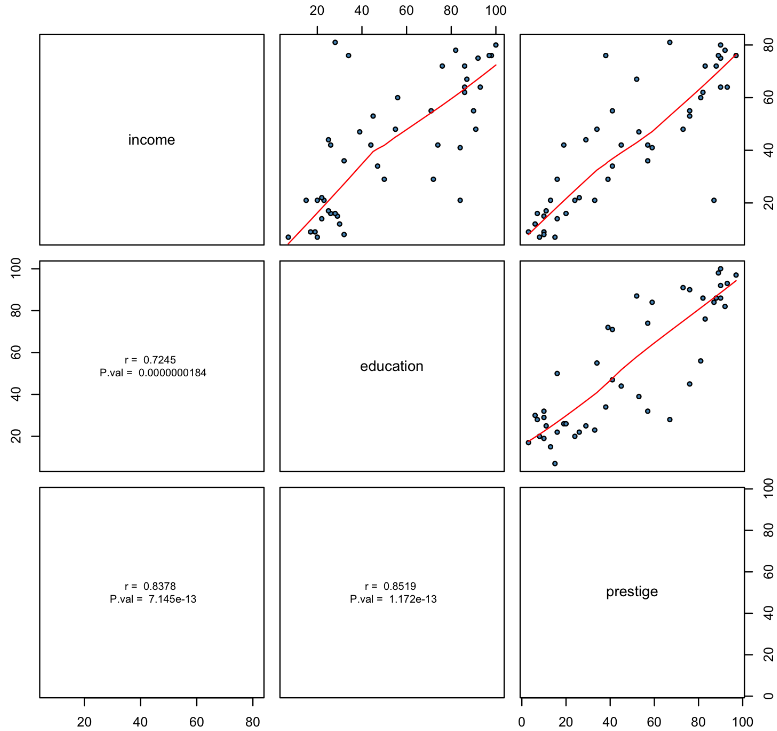
$$F_{0.975,1,42} = 5.403859;\ F_{0.975,2,42} = 4.032710;\ F_{0.975,3,42} = 3.445689;\ F_{0.975,4,42} = 3.108870$$

RSS_{type} = 10598 RSS_{type+edu+inc} = 3798
F_partial = ((10598 - 3798) / 2) / (3798 / 40) = 35.8 (2, 40)
reject

(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.

no
add model with prestige ~ income*type

(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?



income and education is correlated to each other, r = 0.72
if severe, estimates unstable and difficult to interpret them
can increase variance of coefficients estimates sensitive to change in model

need VIF to quantify it

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

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

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

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

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

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

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