Biostatistics Methods I - EXAM 2

TANVIR KHAN, UNI: tk2886

PROBLEM 1A

P-1 } 5-1 = 4 Dummy Variables need to create to analyze the effect of variable region.

P represents the number of predictors

PROBLEM 1b

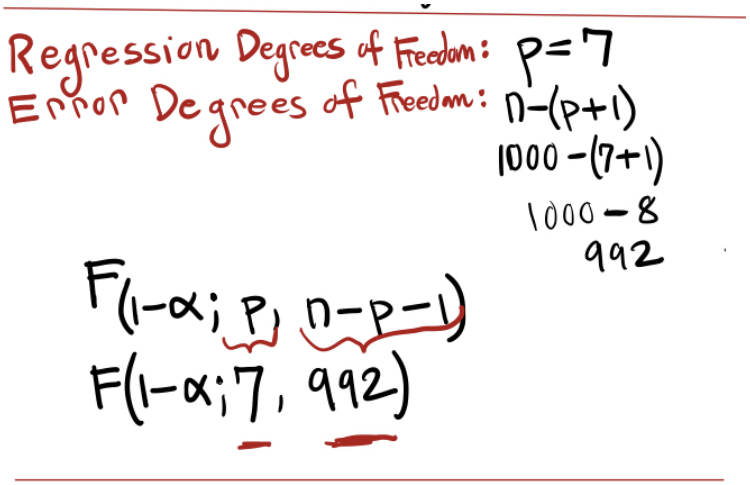
0 Dummy Variables needed to create to analyze the effect of variable Age.

PROBLEM 1c

$$E(Y) = \beta_0 + \beta_1(\text{Age}) + \beta_2 I(\text{Region} = \text{East Europe}) + \beta_3 I(\text{Region} = \text{Asia}) + \beta_4 I(\text{Region} = \text{Sub Saharan Africa}) + \beta_5 I(\text{Region} = \text{South America}) + \beta_6 I(\text{Sex} = \text{Female}) + \beta_7 I(\text{Sex} = \text{Declined to Answer}) + \varepsilon_i$$

I Removed Region = North America and Sex = Male, one less for each categorical variable

PROBLEM 1d


Regression Degrees of Freedom: p=7
Error Degrees of Freedom: n-(p+1)
1000-(7+1)
1000-8
992
F(1-α; p, n-p-1)
F(1-α; 7, 992)

PROBLEM 1e

Model 1:

$$Y_i = \beta_0 + \beta_1(\text{Age}) + \beta_2 I(\text{Region} = \text{East Europe}) + \beta_3 I(\text{Region} = \text{Asia}) + \beta_4 I(\text{Region} = \text{Sub Saharan Africa}) + \beta_5 I(\text{Region} = \text{South America}) + \beta_6 I(\text{Sex} = \text{Female}) + \beta_7 I(\text{Sex} = \text{Declined to Answer}) + \varepsilon_i$$

I Removed Region = North America and Sex = Male, one less for each categorical variable

Model 2:

$$Y_i = \beta_0 + \beta_1(\text{Age}) + \beta_2 I(\text{Sex} = \text{Female}) + \beta_3 I(\text{Sex} = \text{Declined to Answer})$$

$df_S = n - p_S - 1 = 1000 - 3 - 1 = 996$

$df_L = n - p_L - 1 = 1000 - 7 - 1 = 992$

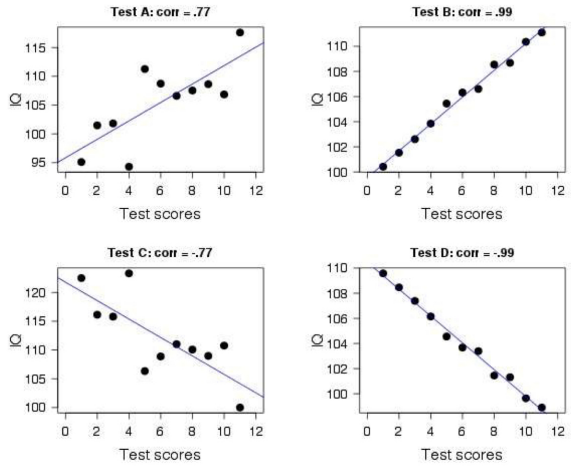
$F_{df_L - df_S,\, df_L}$

$F_{4,\,992}$

PROBLEM 2a

Problem 2

Researchers investigated the relationships between the results of four different visual tests (Test A, B, C, and D) and IQ. They randomly selected 44 subjects and split them into 4 groups of equal size. Each subject had his/her IQ evaluated and was then given one of the four tests. For each test, the researchers made a scatter plot of the subjects' test scores and their IQs and computed the corresponding sample correlation.



a) [Select one correct answer.] Which of the following statements correctly describes the above figure? (2 points)

I. Higher scores on Tests C and D correspond to higher IQ.
II. Higher scores on Tests A and C correspond to higher IQ.
III. The relationship between scores on Test A and IQ is stronger than the relationship between scores on Test D and IQ.
IV. Subjects with similar scores on Test A have a larger spread of IQs than subjects with similar scores on Test B.

Problem 2b

Researchers fit a simple linear regression relating IQ to test score using data from one of the four groups of subjects. Here is part of the regression output:

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	121.7847	2.9582	41.169	$1.47e^{-11}$
Test score	-1.6031	0.4362	_____	_____

Model Information:

Residual standard error: 4.574 with 9 degrees of freedom
R-squared: 0.5929
Adjusted R-squared: 0.5538
F-statistic: 13.51 with 1 and 9 DF and a p-value of 0.00511

b) [Select one correct answer.] Which test's scores did researchers use as the predictor in the regression? (4 points)

I. Test A
II. Test B
III. Test C
IV. Test D

Problem 2c

2c:

We know researchers fit a simple linear regression.

Compute & Interpret 95% CI

For the slope from the regression output above:

- A $(1-\alpha)100\%$ Confidence Interval for the true slope is given by:

$$\hat{\beta}_1 \pm t_{n-2,1-\alpha/2} \cdot se(\hat{\beta}_1),$$

where $se(\hat{\beta}_1) = \sqrt{MSE / \sum_{i=1}^{n}(x_i - \bar{x})^2}$

$$\hat{\beta}_1 \pm t_{n-2,1-\alpha/2} \cdot se(\hat{\beta}_1)$$

$$-1.6031 \pm t_{11-2,1-0.05/2} \cdot 0.4362$$

$$t_{9,0.975}$$

$$qt(.975, 9)$$

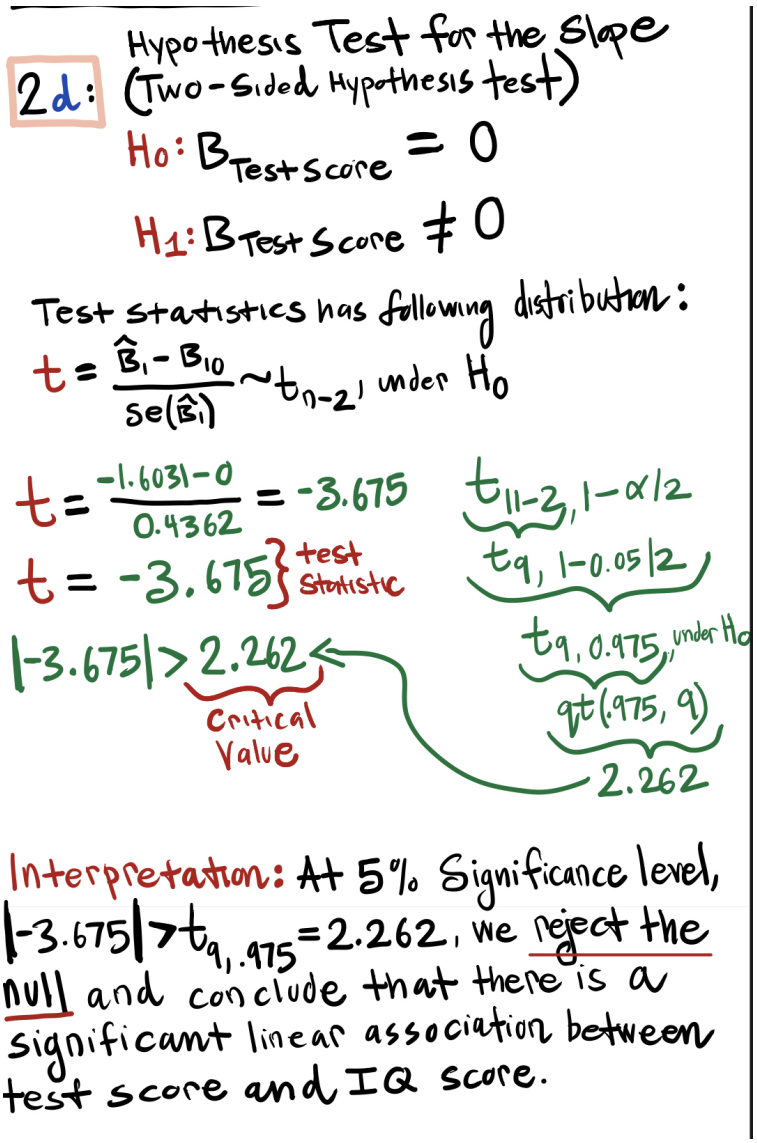
$$\rightarrow 2.26$$

$$-1.6031 - 2.26 \cdot 0.4362 \} -2.5889$$

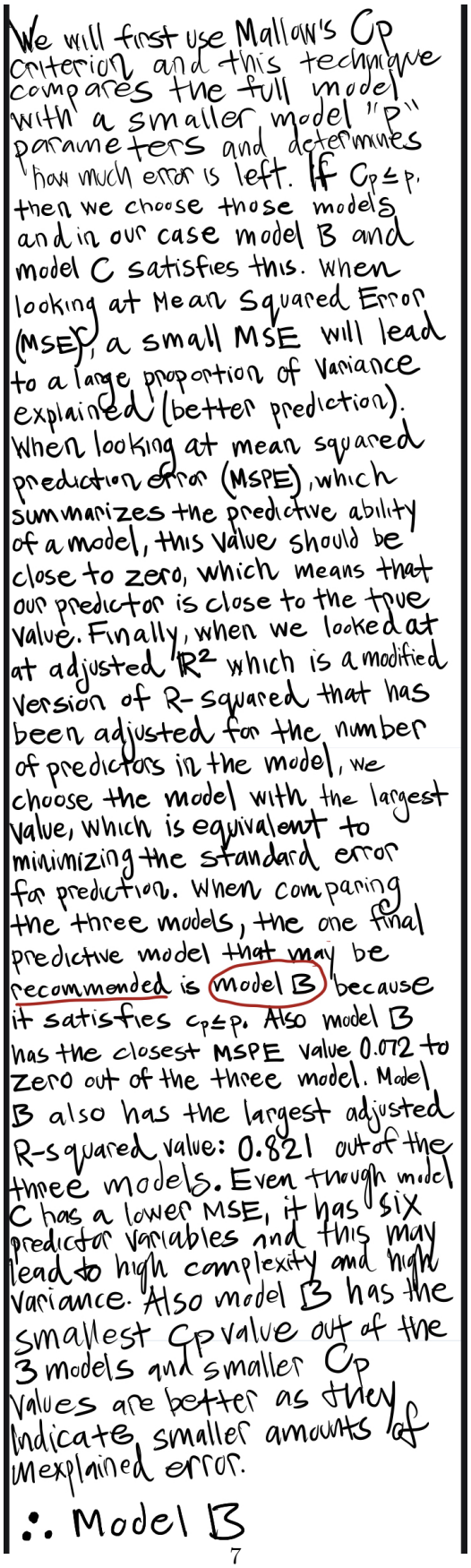
$$-1.6031 + 2.26 \cdot 0.4362 \} -0.62$$

Interpretation: With 95% Confidence Interval, we estimate that the IQ score decreases somewhere between 0.62 and 2.5889 for each additional 1 point increase in test score.

Problem 2d


2d: Hypothesis Test for the Slope (Two-Sided Hypothesis test)
$H_0: \beta_{Test\,score} = 0$
$H_1: \beta_{Test\,score} \neq 0$
Test statistics has following distribution:
$t = \frac{\hat{\beta}_1 - \beta_{10}}{se(\hat{\beta}_1)} \sim t_{n-2}$, under H_0
$t = \frac{-1.6031 - 0}{0.4362} = -3.675$
$t = -3.675$ } test statistic
$|-3.675| > 2.262$
Critical Value
$t_{11-2, 1-\alpha/2}$
$t_{9, 1-0.05/2}$
$t_{9, 0.975}$, under H_0
qt(.975, 9)
2.262
Interpretation: At 5% Significance level, $|-3.675| > t_{9,.975} = 2.262$, we reject the null and conclude that there is a significant linear association between test score and IQ score.

Problem 3



We will first use Mallow's Cp criterion and this technique compares the full model with a smaller model "p" parameters and determines how much error is left. If $C_p \leq p$, then we choose those models and in our case model B and model C satisfies this. When looking at Mean Squared Error (MSE), a small MSE will lead to a large proportion of variance explained (better prediction). When looking at mean squared prediction error (MSPE), which summarizes the predictive ability of a model, this value should be close to zero, which means that our predictor is close to the true value. Finally, when we looked at at adjusted R^2 which is a modified version of R-squared that has been adjusted for the number of predictors in the model, we choose the model with the largest value, which is equivalent to minimizing the standard error for prediction. When comparing the three models, the one final predictive model that may be recommended is model B because it satisfies $C_p \leq p$. Also model B has the closest MSPE value 0.072 to zero out of the three model. Model B also has the largest adjusted R-squared value: 0.821 out of the three models. Even though model C has a lower MSE, it has six predictor variables and this may lead to high complexity and high variance. Also model B has the smallest Cp value out of the 3 models and smaller Cp values are better as they indicate smaller amounts of unexplained error.

∴ Model B