Test 2

STAT 292 Applied Statistics 2A

19/06/2020

Section A

1. (d)

2. (c)

3. (d)

4. (c)

5. (b)

6. (b)

7. (a)

8. (d)

Section B

9.(a)

Model equation:

$Y_{ij} = \mu_i + E_{ij}$, for observation j under treatment i,

where Y_{ij} is the score out of 100 each 10-year old scored in one of the tests,

μ_i is the mean of scores for each of the tests,

and E_{ij} is the error term of the scores of each individual under the different tests.

Hypothesis:

$\mathcal{H}_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4$, all population means are equal

$\mathcal{H}_\mathcal{A} : At\ least\ one\ population\ mean\ is\ different$

9.(b)

The assumptions for a one-way ANOVA is that the data come independently from a normal distribution with constant variance. This is usually represented by the errors (E_{ij}) because the data being independent, normally distributed and constantly variable $if and only if$ the errors are also independent, normally distributed and constantly variable. Mathematically this is written as: $E_{ij} \sim iid\ N(0, \sigma^2)$.

Independence is the assumption that ANOVA is most sensitive to so it's paramount to ensure that our data is independent. To ensure independence, we have to carefully design the experiment to make sure we're not introducing extraneous variability and bias. In this experiment the kids are randomly selected and even though there are four different tests, the tests in the same group are identical. This helps to support our assumption of independence, when carrying out an ANOVA test we have to assume independence or else the validity of the results of our test becomes questionable.

Constant variance is the assumption that each group has the same variance as one another. A way to check for a violation of this assumption is to look at a scatter plot of residuals vs predicted values. If the plot

shows a fanning or funneling pattern then our assumption is violated, a fanning/funneling pattern shows that as the predicted values increase the variance also increases or decreases so the data cannot have equal variance between groups. In this experiment there seems to be a fairly level band, no huge violations of constant variance so our assumption holds. The Levene's test for homogeneity of variance has a high p-value so we don't reject the null null hypothesis which also supports our assumption of constant variance.

Normally distributed is the assumption that even though the different groups may have came from different distributions, all those distributions must be normally distributed. Assuming that the errors come from a normal distributed is the same as assuming the data comes from different normal distributions. To check for normality we look to the QQ plot which is a quantile plot of the residuals. If the data doesn't deviate too much from the straight diagonal line then our assumption holds. In this QQ plot we can see that the points are all closely conforming to the line with very little deviations so our assumption still holds.

9.(c)

Test statistic: F-statistic = 2.01

Degrees of freedom:

Treatment df $= p - 1 = 4 - 1 = 3$

Error df $= n - p = 40 - 4 = 36$

Total df $= n - 1 = 40 - 1 = 39$

p-value: 0.1293

Statistical conclusion:

With a p-value of 0.1293, we fail to reject $\mathcal{H}_0$ at the 5% significance level.

Interpretation:

We don't reject $\mathcal{H}_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4$ so there isn't sufficient evidence to say that there is a difference between the tests.

9.(d)

Failure to reject the $\mathcal{H}_0$ when at least one of the actual means are different is an example of a type II error, essentially a false negative. This can often be attributed by the test having low statistical power$(1 - \beta)$. Increasing our significance level can increase our power making it easier to reject $\mathcal{H}_0$ thus reducing our chance of a type II error. Other ways of increasing the power is by increasing the sample size or by having a balanced experiement. Increasing our power by increasing our α however has the adverse effect of increasing susceptibility to type I errors (rejecting $\mathcal{H}_0$ when $\mathcal{H}_0$ is true).

10.(a)

Theoretical model:

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_p x_p + E,$$

where Y is the response variable,

β_0 is the intercept,

βx is the effect (slope) of the explanatory variable x from $\beta_1 x_1$ up to $\beta_p x_p$,

and E is the error term.

Fitted model:

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + ... + \hat{\beta}_p x_p,$$

where $\hat{Y}$, $\hat{\beta}_0$ and $\hat{\beta}x$ are all estimations of the true positions

10.(b)

This is a multiple linear regression model as we have more than one predictor for Y. The assumptions for multiple linear regression is the same for simple linear regression. ~~In that~~ $E \sim iid\ N(0, \sigma^2)$, in other words the errors must be independent, normally distributed and equally variable. The errors must also need to be independent of x.

5/5

each of them!

10.(c)

The reason why logs are generally used in linear regression models is that it can sometimes eleviate the effects of assumption violations. In this experiment we can see from the SAS output that there is an extreme case of funneling in our studentised residuals plot. This is quite a clear case of unconstant variance, taking the logs instead is more appropriate. The logs studentised residuals plot also has slight funneling so our assumption still does not quite hold but the variations in the varying variances is a lot less than if we didn't take the logs.

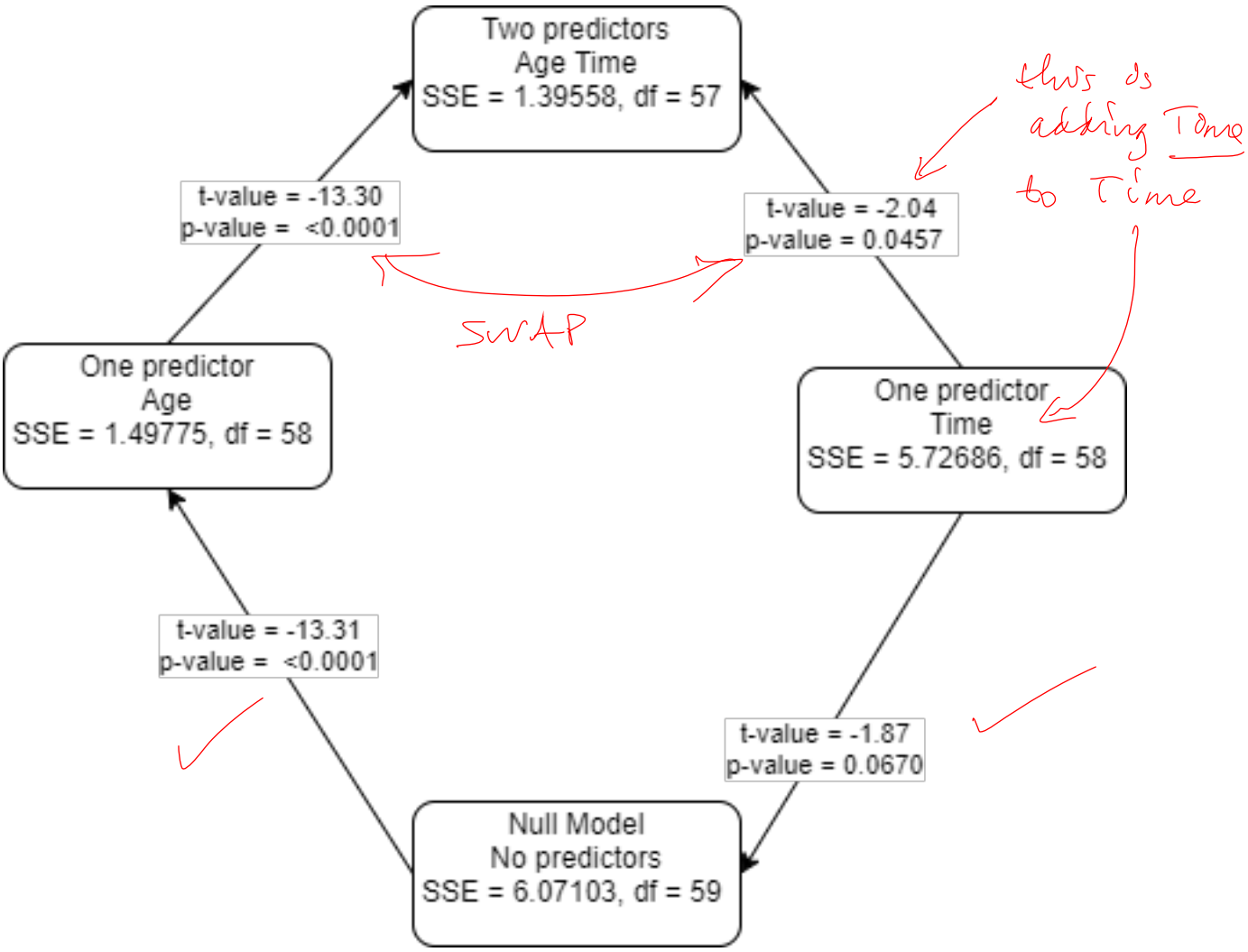
4/8

10.(d)

give other checks too ---

Complete Model	Predictors	p-value for Age	p-value for Time
1	Age+Time	<0.0001	0.0457
2	Age	<0.0001	-
3	Time	-	0.0670

8/8 → for later pages



Regression of logTES on the predictor Age:

Hypothesis:

$\mathcal{H}_0$: $\beta_1 = 0$, no linear relationship

$\mathcal{H}_\mathcal{A}$: $\beta_1 \neq 0$, there is a linear relationship

t-statistic $= -13.31$ with df(58)

p-value $= < 0.0001$

Statistical conclusion:

At the 5% significance level we reject the null hypothesis in favour of the alternate hypothesis. There is a significant linear relation between Age and logTES under the 5% significance level.

Regression of logTES on the predictor Time:

Hypothesis:

$\mathcal{H}_0$: $\beta_1 = 0$, no linear relationship

$\mathcal{H}_\mathcal{A}$: $\beta_1 \neq 0$, there is a linear relationship

t-statistic $= -1.87$ with df(58)

p-value = 0.0670

Statistical conclusion:

At the 5% significance level we fail to reject the null hypothesis. There is not enough evidence to suggest that there is a significant linear relation between Time and logTES under the 5% significance level.

Regression of logTES on the two predictors Age and Time:

Test of Age, given that the term $\beta_2 x_2$ for Time is included in the model:

Hypothesis:

$\mathcal{H}_0$: $\beta_1 = 0$, no linear relationship

$\mathcal{H}_\mathcal{A}$: $\beta_1 \neq 0$, there is a linear relationship

t-statistic = -13.30 with df(57)

p-value = < 0.0001

Statistical conclusion:

At the 5% significance level we reject the null hypothesis in favour of the alternate hypothesis. There is a significant linear relation between Age, *given that the predictor Time is already included in the model*, and logTES under the 5% significance level.

Test of Time, given that the term $\beta_1 x_1$ for Age is included in the model:

Hypothesis:

$\mathcal{H}_0$: $\beta_2 = 0$, no linear relationship

$\mathcal{H}_\mathcal{A}$: $\beta_2 \neq 0$, there is a linear relationship

t-statistic = -2.04 with df(57)

p-value = 0.0457

Statistical conclusion:

At the 5% significance level we reject the null hypothesis in favour of the alternate hypothesis. There is a significant linear relation between Time, *given that the predictor Age is already included in the model*, and logTES under the 5% significance level.

Test of Age and Time simultaneously:

Hypothesis:

$\mathcal{H}_0$: $\beta_1 = \beta_2 = 0$, no linear relationship

$\mathcal{H}_\mathcal{A}$: *At least one of* β_1 *and* β_2 *is non* $-$ *zero*, there is a linear relationship

t-statistic = 9.77 (squareroot of our F-value, $\sqrt{95.48}$)

p-value = < 0.0001

Statistical conclusion:

At the 5% significance level we reject the null hypothesis in favour of the alternate hypothesis. There is a significant linear relation between both predictors and logTES under the 5% significance level.

10.(e)

Under the 5% significance level our tests for each model gives us evidence that simple linear regression of Age, linear regression of Age given Time, linear of Time given Age and the multiple linear regression have significant linear relationships between the explanatory variables with the response variable. However there

isn't sufficient evidence to determine a significant linear relationship between Y and x in the simple linear regression for Time.

In other words Age with Time has a significant linear relation with log(Total Error Score), Age given Time has a significant linear relationship with log(Total Error Score), Time given Age has a significant linear relationship with log(Total Error Score), Age by itself has a significant linear relationship with log(Total Error Score) but Time by itself does not have a significant linear relation with log(Total Error Score).

4/6

Since a perfect Total Error Score is 0 (lower scores > higher scores) it makes sense that the slopes for all our models are negative. This means that as Age and Time increase the TES decreases. → also need conditional interpretation

Choosing a model with the best predictors means comparing effects of one predictor versus another, then all the predictors simultaneously then a predictor given that another predictor is already in the model. From our analysis it appears that Age is much more influential on TES than Time as all the models including Age has significant linear relationships, but Time by itself is not significant. This dissonance between the simple linear regression and the multiple linear regression in regards to Time can be explained by the fact Age is involved in all the multiple linear regression analysis. Adding Time when Age is already included doesn't have as big of an effect as adding Age when Time is included, if our test has more precision (such as a 1% significance level) then we wouldn't have rejected the null hypothesis.

have you done this conclusively?

2/2 ← given deaths, for arrows pointing up to "Age + Time"

2/4