Assignment 4

STAT 292 Applied Statistics 2A

8/06/2020

1. Comprehension Test

(a)

Model equation

$Y_{ijk} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + E_{ijk}$,

where μ is the overall mean averaged over α and β,

α_i is the ethnic groups at level i,

β_j is the sex at level j,

$(\alpha\beta)_{ij}$ is the interaction parameter between ethnic groups and sex of the individual,

and E_{ijk} is the error term.

Assumptions

The sum-to-zero constraints:

$\sum_{i=1}^{a} \alpha_i = 0$, $\sum_{j=1}^{b} \beta_j = 0$, for each: $j = 1, 2, ..., b$, $\sum_{i=1}^{a} (\alpha\beta)_{ij} = 0$, and for each: $i = 1, 2, ..., a$, $\sum_{j=1}^{b} (\alpha\beta)_{ij} = 0$

Which essentially means that all the parameters must add to zero. It's a constraint to mitigate the problem of having more parameters than predictions, allowing us to arrive at unique estimates of the parameters.

The ANOVA assumptions:

We also have to staisfy the ANOVA assumption that the values of Y are independent, normally distributed and equally variable within groups. The assumptions are usually stated in terms of the errors because the errors and the Y scores are identically formulated due to the Y values being independent, normally distributed and equally variable within groups *if and only if* the error components in the model are themselves independent, normally distributed and equally variable of each other. Mathematically we write this as: $E_{ijk} \sim iid\ N(0, \sigma^2)$.

The assumption of independence:

Independence is checked during the experiment design phase. But we're assuming independence else the validity of our results and the validity of our data becomes questionable.

The assumption of normally distributed errors:

Normal distribution of errors can be checked with our QQ plot, the QQ plot shows the residuals (errors) are fitting the straight line so we can assume our errors come from a normal distribution.

The assumption of equally variable errors:

To determine whether our assumption of equal variances hold, we look to our residuals vs predicted values plot. We can see from our plot that there isn't any funneling so there's no clear signs of varying variation so our assumption holds.

Null and alternative hypotheses

$\mathcal{H}_0$: There is no interaction (all $(\alpha\beta)_{ij} = 0$)

$\mathcal{H}_\mathcal{A}$: There is some interaction (at least one $(\alpha\beta)_{ij} \neq 0$)

ANOVA Table (if relevant), p-values

Source	DF	Type I SS	Mean Square	F Value	Pr > F
Ethnicity	2	3060.640086	1530.320043	48.79	<.0001
Sex	1	275.113176	275.113176	8.77	0.0070
Ethnicity*Sex	2	29.685435	14.842718	0.47	0.6289

Statistical conclusions

Our Ethnicity×Sex p-value is 0.63 so we fail to reject $\mathcal{H}_0$ at the 5% and 1% significance levels.

Interpretation

There isn't sufficient evidence to say that there is interaction between ethnicity and sex on their effect on comprehension. Since we haven't found a significant interaction, we then perform a main effects test.

Main effects null and alternative hypotheses

Main effects hypothesis for factor A(Ethnicity):

$\mathcal{H}_0$: There is no significant main effect of factor A (all $\alpha_i = 0$)

$\mathcal{H}_\mathcal{A}$: There is a significant main effect of factor A (at least one $\alpha_i \neq 0$)

Main effects hypothesis for factor B(Sex):

$\mathcal{H}_0$: There is no significant main effect of factor B, after allowing for factor A (all $\beta_j = 0$)

$\mathcal{H}_\mathcal{A}$: There is a significant main effect of factor B, after allowing for factor A (at least one $\beta_j \neq 0$)

Main effects statistical conclusions

The main effects test for factor A(ethnicity) results in a very small p-value of < 0.0001, so we reject the null hypothesis (that there is no significant main effect of factor A), in favour of the alternate hypothesis (that there is a significant main effect of Factor A) at both the 5% and 1% significance levels.

After allowing for ethnicity, the main effects test for factor B(sex) shows that there's also a significant main effect of sex ($p = 0.0070$) at the 5% and 1% levels of confidence.

Main effects interpretation

Despite our test telling us that there's no interaction between ethnicity and sex on an individuals comprehension, our main effects test tells us that both ethnicity and sex(after accounting for ethnicity) are significant factors.

Interaction plots interpretation

The near parallel lines in our interaction plot supports our p-value, that ethnicity and sex do not interact in their effect on comprehension. From our interaction plots we can also see that in all ethnicities, females have a higher average comprehension than males and that E1 has a higher average comprehension than E2 who has a higher comprehension than E3. We can also note that there's a greater difference between ethnicities than there are between sex, so the differing ethnic groups have a greater effect on comprehension than the differing sexes.

(b)

The one-way ANOVA with sex being the sole factor yields a p-value of 0.3292, so we reject the notion that both sexes have equal population means at the 5% significance level but not at the 1% significance level. In other words average comprehension differs between sexes with 95% confidence but not with 99% confidence.

Doing the main effects test for sex also gives us the conclusion that sex has an effect on comprehension. We got a p-value of 0.0070 an incredibly small value so unlike the one-way ANOVA we have a much stronger case even though the conclusion is roughly the same at the 5% level, the main effects test gives us more concrete evidence and we can also reject $\mathcal{H}_0$ at the 1% level. However this could largely be due to already taking ethnicity into account whereas the one-way ANOVA omits this factor entirely.

The mean of the squared errors in the two-way ANOVA are much lower than those of the one-way ANOVA, this is because the one-way ANOVA omits a useful factor (ethnicity). Only taking sex into consideration means it's the only factor we're observing for explanation of variance, but we see from the main effects test of the ethnicity factor that ethnicity even without sex has a much greater effect on comprehension.

2. Invertebrates in Mussel Clumps

(a)

Linearity

Linearity is shown on the scatter plot as the points form a positive linear pattern (Y increases as X increases and absence of curvature) with no outliers.

Constant variance

Constant variance is shown as the points display no signs of funneling, there's no widening of points as our x values grow larger.

(b)

Model equation

Theoretical model:

$Y = \beta_0 + \beta_1 x + E,$

where Y is the number of different species,

x is the log_{10}(Area),

β_0 is the intercept,

β_1 is the slope(effect) of x,

and E is the error term.

Fitted model:

$\hat{Y} = \hat{\beta_0} + \hat{\beta_1 x_1},$

or $\hat{Y} = -25.64136 + 11.20214x$

Assumptions

The assumptions for the simple linear regression model is similiar to the ANOVA assumptions in that $E \sim iid\ N(0, \sigma^2)$, the errors must come independentally from a normal distribution with constant variance.

On top of this we also have assumptions that the errors are independent of x and that there is a linear relationship between x and Y.

Independence, normal distribution, constant variance of errors:

The RStudent(studentized residuals vs predicted value) plot shows that there is funneling, there's a widening on the right indicating that as our predicted values get larger so does our variance thus our assumption of constant variance is violated. A log transformation of the Y scores may be appropriate.

The QQ plot shows that the errors mostly fit the line with one major deviation at the left tail. Our assumption of normality holds.

A Cook's Distance plot is a way of detecting outliers and values that have high leverage, in other words values that have a lot of influence on the regression line. In this plot we can see that there's one point that has greater influence than the others but not to a significant degree. Our choice of thresholds for Cook's distance varies, 1 is usually a good number when we have a large n (with large n, 1 is usually used as a threshold because the median is close to 1), others say $\frac{4}{n}$ or even $\frac{4}{(n-k-1)}$ (where k is the number of explanatory variables) are also good thresholds. With a threshold of 1, all the cook's distances are well below so there's no need to regard any point as influential or concerning. However in the case of $\frac{4}{n} = \frac{4}{25} = 0.16$ as marked by the horizontal line shows us that observation 21 is a concerning point and may significantly influence our regression. All other observations are below the line (0.16) but you also might want to keep an eye on observation 2, though not over our threshold it is significantly above the other points.

Independence cannot be checked but is always assumed. The results from tests done on a model that doesn't have independence are unreliable.

Linear relationship between x and Y:

This was already explained in **2.(a)**

Null and alternative hypotheses

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

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

ANOVA Table (if relevant), p-values

F-test statistic:

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	868.50179	868.50179	117.85	<.0001
Error	23	169.49821	7.36949		
Corrected Total	24	1038.00000			

t-test statistic:

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	-25.64136	3.78287	-6.78	<.0001
logArea	1	11.20214	1.03189	10.86	<.0001

Statistical conclusions

We can use either the t-test statistic or the F-test statistic to get our p-value. The F-test statistic gives us a value of 117.85 with df(1, 23) and the t-test gives us 10.86 with df(23). Both tests give us a p-value of < 0.0001 so at both the 1% and 5% significance levels we reject $\mathcal{H}_0$: $\beta_1 = 0$ in favour of $\mathcal{H}_\mathcal{A} : \beta_1 \neq 0$.

Interpretation

The model is valid if there's a linear relationship between x and Y, in other words if you reject $\mathcal{H}_0$ then it's valid because the test gives evidence that there is a linear relationship. Our p-value is < 0.0001, an incredibly small value so we reject $\mathcal{H}_0$ which supports the validity of our model. Concluding that the explanatory variable of $log_{10}(Area)$ and the response variable of number of different species fits a linear regression.

3. Coarse Woody Debris in Lakes

(a)

CWD.BASA vs. RIP.DENS:

Weak positive linear correlation. CWD.BASA increases as RIP.DENS increases, no curvature but nowhere near a perfect straight line.

CWD.BASA vs. L10CABIN:

Weak negative linear correlation. CWD.BASA decreases as L10CABIN increases, no curvature but nowhere near a perfect straight line.

RIP.DENS vs. L10CABIN:

Also possible weak negative linear correlation as we can see that RIP.DENS is decreasing as L10CABIN is increasing but there's also possible outliers.

(b)

i. Regression of CWD.BASA on the predictor RIP.DENS

Model equation

Theoretical model:

$Y = \beta_0 + \beta_1 x + E$,

where Y is the coarse woody debris,

x is the riparian tree density,

β_0 is the intercept,

β_1 is the slope(effect) of x,

and E is the error term.

Fitted model:

$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 x_1$,

or $\hat{Y} = -77.09908 + 0.11552x$

Hypothesis

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

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

Conclusions

The t-test gives us a t statistic of 4.93, p-value of 0.0002 with df(14) and the F-test gives us an F statistic of 24.30, p-value of 0.0002 with df(1, 14). With an incredibly small p-value of 0.0002 we reject the null hypothesis in favour of the alternative hypothesis. There is a linear relationship between riparian tree density and coarse woody debris.

ii. Regression of CWD.BASA on the predictor L10CABIN

Model equation

Theoretical model:

$Y = \beta_0 + \beta_1 x + E$,

where Y is the coarse woody debris,

x is the cabin density,

β_0 is the intercept,

β_1 is the slope(effect) of x,

and E is the error term.

Fitted model:

$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 x_1$,

or $\hat{Y} = 121.96875 - 93.30142x$

Hypothesis

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

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

Conclusions

The t-test gives us a t statistic of -5.10, p-value of 0.0002 with df(14) and the F-test gives us an F statistic of 26.00, p-value of 0.0002 with df(1, 14). With an incredibly small p-value of 0.0002 we reject the null hypothesis in favour of the alternative hypothesis. There is a linear relationship between cabin density and coarse woody debris.

iii. Regression of CWD.BASA on the two predictors RIP.DENS and L10CABIN

Model equation

Theoretical model:

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

where Y is the coarse woody debris,

x_1 is the riparian tree density,

x_2 is the cabin density,

β_0 is the intercept,

β_1 is the slope(effect) of x_1,

β_2 is the slope(effect) of x_2,

and E is the error term.

Fitted model:

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

or $\hat{Y} = 18.16485 + 0.06572 x_1 - 56.26481 x_2$

Assumptions

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 also need to be independent of x.

Studentized residuals vs. Predicted value:

This plot is used to find signs of non-constant variance by identifying inconsistent spread of data. From this plot, we see that the residuals are fairly level without signs of funnelling. There is one potential outlier that may influence our results.

QQ plot of residuals:

The QQ plot is used to identify deviations from normality. If the points on the plot conform to the straight diagonal line then we conclude that our errors come from a normal distribution. We can see that the plots do form a straight diagonal line, with slight deviations but nothing to be concerned about.

Cook's distance for Y:

Cook's distance is used to find points of interest, points that are outliers or have high leverage. These points have a potential to influence the regression significantly more than the other points. A threshold number of 1 is commonly used when n is sufficiently large but I prefer to use $\frac{4}{n}$ as the threshold. So for $\frac{4}{16} = 0.25$ we can see a particular point of interest at observation 2 highlighting a possibility that this point can heavily influence our regression.

Errors independent of each other and of x:

We always assume independence. Always.

Hypothesis

Test of riparian tree density, given that the term $\beta_2 x_2$ for cabin density is included in the model:

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

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

Test of cabin density, given that the term $\beta_1 x_1$ for riparian tree density is included in the model:

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

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

Test of cabin density and riparian tree density simultaneously:

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

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

Conclusions

RIP.DENS, given that L10CABIN is included, has a t-value of 2.33 with df(13) and a p-value of 0.0367. We reject the null hypothesis at the 5% significance level but not at the 1% significance level. Rejecting the null hypothesis means that even after L10CABIN is included, adding RIP.DENS will improve the model's fit.

If RIP.DENS is already included in the model, the test for L10CABIN gives us a t-statistic of -2.50 with df(13) and a p-value of 0.0267. So at the 5% significance level we reject the null hypothesis, but not at the 1% level. Under the 5% significance level, L10CABIN is a useful predictor even after adding RIP.DENS.

Both those tests don't reveal any meaningful insight into our predictors, since with our p-values being similar with the same conclusions. It essentially says that both predictors are useful after adding one another so we don't have any evidence to which one would make a better predictor, possibly it's both. To test this we do a test of the two predictors simultaneously.

In the model where RIP.DENS and L10CABIN are tested simultaneously, we can see that our p-value is super small (< 0.0001). We reject the null hypothesis at both the 5% and the 1% levels unlike the previous two tests which we only reject at the 5% level. The null hypothesis for this test states that there is no relationship between Y and the x predictors, in other words the values of Y predicted are no closer than the Y values you would expect by chance. Rejecting the null hypothesis instead says that the Y values are correlated with the x predictors that were tested. When building a multiple linear regression model doing a test such as this allows us to extract the important features and omit the unneccesary ones. With our conclusion, we choose to keep both RIP.DENS and L10CABIN as useful predictors.

(c)

The regression of CWD.BASA on the two predictors RIP.DENS and L10CABIN. The question asks whether human habitation, after allowing for riparian density, has any effect on coarse woody debris. The other two tests only test if one of the x predictors is related to Y, what we want is to observe the effect of one x on Y after adding the other x. The test in particular is the test of cabin density, given that the term $\beta_1 x_1$ for riparian tree density is included in the model. From the results of our test we're 95% confident that human habitants still have an effect on coarse woody debris after allowing for riparian density.

4.Age of Teeth

(a)

$Y = \alpha_i + \beta_i x + E,$

where Y is the estimated age,

α_i is the effect of Method at level i,

β_i is the effect of the covariate true age when Method is at level i,

and E is the error term

(b)

Assumptions

The ANCOVA model has the same assumptions as an ANOVA model. $E \sim iid\ N(0, \sigma^2)$, all the errors have to come independently from a normal distribution with constant variance and also be independent of x.

Studentized residuals vs. Predicted values:

There's a slight curved pattern but not too explicit. Most of the points do seem to be fairly level with a possible outlier but nothing that would violate our assumption.

QQ plot of residuals:

The points do seem to conform to a straight line but there's a bit of a staircase pattern but there aren't any huge deviations from normality.

Histogram of residuals:

We don't have enough data to make any reasonable judgements based on the histogram but from what we do have there seems to be some right-skewness. Doesn't support our assumption of normality.

Cook's distance:

With a threshold of $\frac{4}{20} = 0.2$ we identify two possible influential points, observation 8 and observation 17. Observation 8 is a point to keep an eye on but it's near enough to the threshold that it might be ok to keep. Observation 17 on the other hand is a significant distance above the other points and might be worthwhile to delete.

Independence:

The teeth are randomly allocated which is a factor in maintaining independence, we assume everything else is independent too.

Validity:

The diagnostic graphs shows many points of concern but not overwhelmingly, I believe ANCOVA would be appropriate.

Hypothesis

$\mathcal{H}_0 : \beta_1 = \beta_2$, no interaction between covariate and factor(s)

$\mathcal{H}_\mathcal{A} : \beta_1 \neq \beta_2$ (at least one β_i is different), there is an interaction between covariate and factor(s)

Statistical conclusions

Our interaction test gives us an F-value of 1.18 and a p-value of 0.2938. At the 5% significance level we fail to reject the null hypothesis establishing that there is no interaction between the covariate and the factor. We then move onto the main effects test.

Main effects hypothesis

Test of main effects of covariate:

$\mathcal{H}_0$: $\beta = 0$

$\mathcal{H}_\mathcal{A}$: $\beta \neq 0$

Test of main effects of factor A:

$\mathcal{H}_0$: *all* $\alpha_i = 0$, no main effect of factor A after allowing for covariate

$\mathcal{H}_0$: *at least one* $\alpha_i \neq 0$, there is a main effect of factor A after allowing for covariate

Main effects statistical conclusions

For the covariate we reject the null hypothesis at the 5% level with a p-value of < 0.0001 and conclude that there is a main effect of the covariate on Y. For factor A, after allowing for the covariate, we also reject the null hypothesis with a p-value of 0.0490 but only marginally.

Interpretation

Even though there's no interaction between the true age and the method on the estimated age, we find that the true age is a huge factor on the estimated age and that the Method is also a useful predictor after accounting for the true age. From our test we conclude to keep both the covariate and the A factor.