第 3 章例题写法

Method and Assumption Checks

The histogram of the data looked reasonably normal and so a null model was fitted.

All model assumptions appear to be satisfied - could be verified with the histogram of the data. [1]

Our final model is

$$Exam_i = \beta_0 + \epsilon_i \text{ (or } Exam_i = \mu + \epsilon_i)$$

, where $\epsilon_i \sim iid\ N(0, \sigma^2)$

Executive Summary

We were interested in building a model to describe exam marks.

We estimate the expected exam mark to be between 49.8 and 55.9 (out of 100).

We have no reason to believe that the expected exam mark differs from the historical average value of 55 (out of 100) ($P\text{-}value = 0.17$).

第 4 章例题写法

Method and Assumption Checks

The scatter plot of exam mark vs assignment mark suggested curvature in the relationship.

We began with a linear model to describe exam marks with assignment marks. The residual plot from the fit of a simple linear model showed fairly constant scatter but had strong curvature. So, a quadratic term was added to the linear model.

All model assumptions look satisfied once we added the quadratic term to the linear model.

Our final model is

$$Exam_i = \beta_0 + \beta_1 \times Assign_i + \beta_2 \times Assign_i^2 + \epsilon_i,$$

where $\epsilon_i \sim iid\ N(0, \sigma^2)$.

Our model explained 55% of the variability in the students' final exam marks.

Executive Summary

We were interested in building a model to estimate exam marks with assignment marks.

The relationship between expected exam mark and assignment score modelled was quadratic.

Here, for a one mark increase in assignment score, the increase in expected exam score was greater as assignment score increased. For example, there was little difference in expected exam score for those getting 7 or 8 in assignments, but a much bigger difference for those getting 17 or 18.[1]

For assignment marks of 0, 10 and 20, the estimate expected exam marks were between 16.6 to 40.3, 34.3 to 41.1, and 73.6 to 84.8, respectively.

[1] What could be causing this? Cheating on assignments? Over-zealous lab demonstrators giving too many hints?

第 5 章例题写法

Method and Assumption Checks

We wish to explain exam marks with attendance, a two-level factor. So, we have fitted a linear model with a single explanatory dummy variable. (Note, this is equivalent to conducting a two-sample t-test).

Four non-attending students did unusually well (i.e., large positive residuals), but since the sample size was large, this will be of little consequence. Hence, all model assumptions were satisfied.

Our final model is

$$Exam_i = \beta_0 + \beta_1 \times Attend.Yes_i + \epsilon_i,$$

where $\epsilon_i \sim iid\ N(0, \sigma^2)$. Here $Attend.Yes_i = 1$ if the student regularly attended (i.e., answered "Yes"), otherwise it is zero (i.e. "No").

Our model explained a small 15% of the variability in the students' final exam marks.

Executive Summary

We wanted to quantify the relationship between exam marks and attendance.

There was strong evidence that exam marks were higher for students who attend class versus students who didn't attend class ($P\text{-}value \approx 10^{-6}$). We estimate that regular attendance could increase their expected exam mark between 9.5 to 21.6 exam marks (out of 100).

The expected exam marks of non-attendees and attendees are between 37.2 to 47.3, and 54.4 to 61.2, respectively.

The predicted exam marks for individual non-attendees and attendees are between 7.7 to 76.7, and 23.5 to 92.1, respectively.

Our model only explains 15% of the variability in the students' final exam marks, this would not be very good for prediction. We can see this in how wide our prediction intervals are.

第 6 章例题写法

Method and Assumption Checks

The scatter plot of age vs price showed clear nonlinearity and an increase in variability with price.

Residuals from a simple linear model showed failed the equality of variance and no-trend assumptions, and so the prices were log transformed. A simple linear model fitted to logged price satisfied all assumptions.

Our final model is

$$log(Price_i) = \beta_0 + \beta_1 \times Age_i + \epsilon_i,$$

where $\epsilon_i \sim iid\ N(0, \sigma^2)$.

Our model explained 82% of the variability in the logged Mazda prices.

Executive Summary

We wanted to see how Mazda car prices decrease with age.

There was clear evidence the price of the cars was exponentially decreasing as the cars got older ($P\text{-}value \approx 0$).

We estimate that the median price for new Mazda cars (in 1991) was between A$23,600 to A$30,400 (to the nearest A$100).

We estimate that each additional year in age results in depreciation of between 13.9% to 16.3%.

第 8 章例题写法

Method and Assumption Checks

As we have two explanatory variables, one numeric and one factor, we have fitted a linear model that used different intercept and slopes for each attendance group (i.e., interaction model). We could not drop the interaction term ($P\text{-}value = 0.043$).

All model assumptions were satisfied.

Our final model is

$$Exam_i = \beta_0 + \beta_1 \times Test_i + \beta_2 \times Attend_i + \beta_3 \times Attend_i \times Test_i + \epsilon_i,$$

where $Attend_i = 1$ if student i is a regular attender, otherwise 0, and $\epsilon_i \sim iid\ N(0, \sigma^2)$.

Our model explained a modest 63% of the variability in students' exam marks.

Executive Summary

We wanted to quantify students' exam marks relationship with attendance and test marks.[1]

[1]Since there are different slopes in the two groups, we need to discuss each slope individually.

There was a clear linear relationship between test and exam scores, but this relationship differed between students who attended and who did not attend lectures.

We estimate that each additional test mark (out of 20) obtained by a non-attending student would increase their expected exam mark by between 1.8 to 3.7.

For regular attenders, the increase is an additional 0.04 to 2.2 expected exam marks per test mark.

第 9 章例题写法

Method and Assumption Checks

To explain language score, we first fitted the model with explanatory variables teaching method, IQ, and their interaction. But, the interaction term was not significant (*P-value* = 0.37). The model was refitted with the interaction term removed.

All model assumptions were satified. [*Optional:* The students should be acting independent of each other as they were randomly allocated to the method taught and they students are measured under test conditions.]

Our final model is

$$lang_i = \beta_0 + \beta_1 \times IQ_i + \beta_2 \times method.method2_i + \beta_3 \times method.method3_i + \epsilon_i,$$

where:

- $method.method2_i$ is set to one if student i received method 2, otherwise it is zero,
- $method.method3_i$ is set to one if student i received method 3, otherwise it is zero,
- and $\epsilon_i \sim iid\ N(0, \sigma^2)$.

Here method 1 is our baseline.

The final model was also refitted with method 2 as the baseline. **Note:** When we change the baseline (to level 2), the values of the dummy variables switch, so that $method.method2_i$ becomes $method.method1_i$. Hence, $method.method1_i$ is set to one if student i received method 1, otherwise it is zero.

Our model explains almost 80% of the variation in language score.

Executive Summary

We were interested in comparing the effectiveness of three teaching methods on language scores acheived by students. We also wanted to see how this was effected by students IQ's.

We found that the effects of the teaching methods are the same regardless of IQ and the effect of IQ is the same regardless of teaching method.

In particular teaching method 2 is significantly better than the other two methods. Also, both methods 1 and 2 are significantly better than method 3.

Not surprisingly, students with higher IQ tended to score higher.

With 95% confidence:

- For students experiencing the same teaching method, we estimate that the expected language test score increases by between 1.6 and 4.7 marks for each additional 10 IQ points,
- For students with the same IQ, we estimate that the expected language test score for students taught using method 2 is between 4.1 and 15.7 marks higher than those taught using method 1,
- For students with the same IQ, we estimate that the expected language test score for students taught using method 1 is between 8.3 and 20.0 marks higher than those taught using method 3,
- For students with the same IQ, we estimate that the expected language test score for students taught using method 3 is between 18.3 and 29.7 marks lower than those taught using method 2.

第 10 章例题写法

Method and Assumption Checks

Looking at the pairs plot, we saw that `bwt` was related to a number of our explanatory variables. So, we want to construct a multiple linear regression model with `bwt` as the response variable.

We had issues with Cooks Distance for observations 239 and 820. We dropped these observations from the final model. Then, all model assumptions were satisfied.

Our final model is

$$\begin{aligned} bwt_i = \beta_0 + \beta_1 \times gestation_i + \beta_2 \times OD_i + \beta_3 \times height_i \ + \beta_4 \times bmi_i + \beta_5 \times not.first.born_i + \\ \beta_6 \times smokes_i + \beta_7 \times gestation_i \times OD_i + \epsilon_i, \end{aligned}$$

where $\epsilon_i \ iid \sim N(0, \sigma)$. Here our dummy variables correspond to whether the baby was overdue, not the first born, and whether the mother smokes respectively.

Our model only explains about 31% of the variability in a baby's birthweight.

Executive Summary.

We wanted to build a model to explain the birth weight of babies.

Keeping all other varaiables constant:

- A child has a higher expected birth-weight the longer its gestation time — up to a 42 weeks — then it starts decreasing in size the longer it stays unborn. We estimated an expected increase of 0.57 to 0.71 ounces per gestation day. After 42 weeks this will decrease by about -0.61 to -1.13 ounces per gestational day [NOTE: it might have been better had we changed the OD baseline].
- We estimated that for each additional inch of mother's height the baby's birthweight increases by 0.94 to 1.64 ounces, on average.
- We estimated that for each unit change in a mothers BMI the baby'sb irthweight increases by 0.08 to 0.63 ounces, on average.
- If the mother smokes this reduces the baby's birthweight by 6.17 to 9.79 ounces, on average.
- Not being first born seems to reduce the baby's birthweight by 1.48 to 5.52 ounces, on average.

第 11 章例题写法

Methods and Assumption Checks

The boxplot of `days` by `group` indicated that males living with 8 uninterested females have shorter lives compared to their counterparts in other groups. So, we fitted a One-way ANOVA model to these data.

The model assumptions seem satisfied.

Our final model is

$$\text{days}_i = \beta_0 + \beta_1 \times Group2_i + \beta_2 \times Group3_i + \beta_3 \times Group4_i + \beta_4 \times Group5_i + \epsilon_i,$$

where $GroupX_i$ is 1 if the ith male fruitfly is in group X and 0 otherwise, and $\epsilon_i \sim iidN(0, \sigma^2)$.

Alternatively, our final model could be written as

$$\text{days}_{ij} = \mu + \alpha_i + \epsilon_{ij},$$

where μ is the overall mean survival time and α_i is the effect of being in the ith group and $\epsilon_{ij} \sim iid\ N(0, \sigma^2)$.

Our model explained 31% of variability in male fruitfly longevity.

Executive Summary

Researchers were interested in how sexual activity affects male fruitfly longevity.

We see that the effect of Group 5, males with 8 uninterested females, seems markedly different from all the others.

In particular group 5 males, on average, lived fewer days than:

- Group 1 males (living alone) by between 13 to 36 fewer days.
- Group 2 males (living with one interested female) by between 14 to 38 fewer days.
- Group 3 males (living with eight interested females) by between 13 to 36 fewer days.
- Group 4 males (living with one uninterested female) by between 6 to 30 fewer days.

On a lighter note these male fruit flies are fine if no females are about or if they are there they need to be 'interested' in them — otherwise they die earlier (they 'drop like flies'). It's tempting to make similar inference about the human species but that may be going too far!

第 12 章例题写法

Methods and Assumption Checks

We have two explanatory factors, `Pass.test` and `Attend`, and one numeric response `Exam`. The interaction plots indicated different slopes between the levels. So, we fitted a two-way ANOVA model with interaction between `Pass.test` and `Attend`. The interaction term was significant ($P\text{-}value = 0.04297$).

The model assumptions seem satisfied.

Our final model is

$$\text{Exam}_i = \beta_0 + \beta_1 \times Pass.Test.pass_i + \beta_2 \times Attend.Yes_i + \beta_3 \times Pass.test.pass_i \times Attend.Yes_i + \epsilon_i,$$

where $Pass.Test.pass_i$ and $Attend.Yes_i$ are dummy variables that takes the value 1 if the student passed the test and if the student regularly attended lectures respectively, otherwise they are 0, and $\epsilon_i \sim iid\ N(0, \sigma^2)$.

Alternatively, our final model could be written as

$$\text{Exam}_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + \epsilon_{ijk},$$

where μ is the overall mean exam mark, α_i is the effect of whether the student passed the test, β_j is the effect of whether the student regularly attended lectures, γ_{ij} is the interaction effect for the combination of a student passing the test and attendance, and $\epsilon_{ijk} \sim iid\ N(0, \sigma^2)$.

Our model explained 39% of variability in students' exam marks.

Executive Summary

Is regular attendance in class and passing the test is associated with exam mark?

We have evidence that the effect that passing the test has on exam marks depends on whether a student attends regularly or not.

We estimate for those who attended regularly, those who passed the test got more than 15 to 34 more marks in the exam than those who did not pass the test.

We estimate for those who did not attend regularly, those who passed the test got more than 2 to 24 more marks in the exam than those who did not pass the test.

We estimate that for those who passed the test, those who regularly attended got between 6 and 24 more marks in the exam than those who didn't attend regularly.

It suffices to say that those who did not pass the test and do not attend regularly are very unlikely to pass the course.

(Note — this agrees with what we discovered in Case Study 5.1.)