Leland Randles DATA606

Homework, Chapter 8

8.2 Baby weights, Part II. Exercise 8.1 introduces a data set on birth weight of babies. Another variable we consider is `parity`, which is 0 if the child is the first born, and 1 otherwise. The summary table below shows the results of a linear regression model for predicting the average birth weight of babies, measured in ounces, from `parity`.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	120.07	0.60	199.94	0.0000
parity	-1.93	1.19	-1.62	0.1052

(a) Write the equation of the regression line.

(b) Interpret the slope in this context, and calculate the predicted birth weight of first borns and others.

(c) Is there a statistically significant relationship between the average birth weight and parity?

(a) average birth weight = 120.07 – 1.93 * parity

(b) If a child is not the first born, their birth weight will be 1.93 ounces less than the first born child. The predicted weight for a first-born child is 120.07 ounces and the predicted weight for a child which is not first-born is 118.14 ounces.

(c) No. The p-value is greater than a 0.05 or 0.01 threshold.

8.4 Absenteeism. Researchers interested in the relationship between absenteeism from school and certain demographic characteristics of children collected data from 146 randomly sampled students in rural New South Wales, Australia, in a particular school year. Below are three observations from this data set.

	eth	sex	lrn	days
1	0	1	1	2
2	0	1	1	11
⋮	⋮	⋮	⋮	⋮
146	1	0	0	37

The summary table below shows the results of a linear regression model for predicting the average number of days absent based on ethnic background (`eth`: 0 - aboriginal, 1 - not aboriginal), sex (`sex`: 0 - female, 1 - male), and learner status (`lrn`: 0 - average learner, 1 - slow learner).[18]

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	18.93	2.57	7.37	0.0000
eth	-9.11	2.60	-3.51	0.0000
sex	3.10	2.64	1.18	0.2411
lrn	2.15	2.65	0.81	0.4177

(a) Write the equation of the regression line.

(b) Interpret each one of the slopes in this context.

(c) Calculate the residual for the first observation in the data set: a student who is aboriginal, male, a slow learner, and missed 2 days of school.

(d) The variance of the residuals is 240.57, and the variance of the number of absent days for all students in the data set is 264.17. Calculate the R^2 and the adjusted R^2. Note that there are 146 observations in the data set.

(a) average # days absent = 18.93 – 9.11 * eth + 3.1 * sex + 2.15 * lrn

(b) The ethnic slope indicates that a non-aboriginal student will have 9.11 less average days absent, a male will have 3.1 more average days absent than a female, and a slow learner will have 2.15 more average days absent than an average worker.

(c) 18.93 – 9.11 * 0 + 3.1 * 1 + 2.15 * 1 = 18.93 + 3.1 + 2.15 = 24.18; 2 minus 24.18 = -22.18

(d) $R^2 = 1 – (240.57 / 264.17) = 0.0893$
Adjusted $R^2 = 1 – ((240.57 / 264.17) * ((146 – 1) / (146 – 3 – 1))) = 0.0701$

8.8 Absenteeism, Part II. Exercise 8.4 considers a model that predicts the number of days absent using three predictors: ethnic background (`eth`), gender (`sex`), and learner status (`lrn`). The table below shows the adjusted R-squared for the model as well as adjusted R-squared values for all models we evaluate in the first step of the backwards elimination process.

	Model	Adjusted R^2
1	Full model	0.0701
2	No ethnicity	-0.0033
3	No sex	0.0676
4	No learner status	0.0723

Which, if any, variable should be removed from the model first?

Learner status should be removed first because the model without it has an R^2 value which is higher than the R^2 value for the full model.

8.16 Challenger disaster, Part I. On January 28, 1986, a routine launch was anticipated for the Challenger space shuttle. Seventy-three seconds into the flight, disaster happened: the shuttle broke apart, killing all seven crew members on board. An investigation into the cause of the disaster focused on a critical seal called an O-ring, and it is believed that damage to these O-rings during a shuttle launch may be related to the ambient temperature during the launch. The table below summarizes observational data on O-rings for 23 shuttle missions, where the mission order is based on the temperature at the time of the launch. *Temp* gives the temperature in Fahrenheit, *Damaged* represents the number of damaged O-rings, and *Undamaged* represents the number of O-rings that were not damaged.

Shuttle Mission	1	2	3	4	5	6	7	8	9	10	11	12
Temperature	53	57	58	63	66	67	67	67	68	69	70	70
Damaged	5	1	1	1	0	0	0	0	0	0	1	0
Undamaged	1	5	5	5	6	6	6	6	6	6	5	6

Shuttle Mission	13	14	15	16	17	18	19	20	21	22	23
Temperature	70	70	72	73	75	75	76	76	78	79	81
Damaged	1	0	0	0	0	1	0	0	0	0	0
Undamaged	5	6	6	6	6	5	6	6	6	6	6

(a) Each column of the table above represents a different shuttle mission. Examine these data and describe what you observe with respect to the relationship between temperatures and damaged O-rings.

(b) Failures have been coded as 1 for a damaged O-ring and 0 for an undamaged O-ring, and a logistic regression model was fit to these data. A summary of this model is given below. Describe the key components of this summary table in words.

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	11.6630	3.2963	3.54	0.0004
Temperature	-0.2162	0.0532	-4.07	0.0000

(c) Write out the logistic model using the point estimates of the model parameters.

(d) Based on the model, do you think concerns regarding O-rings are justified? Explain.

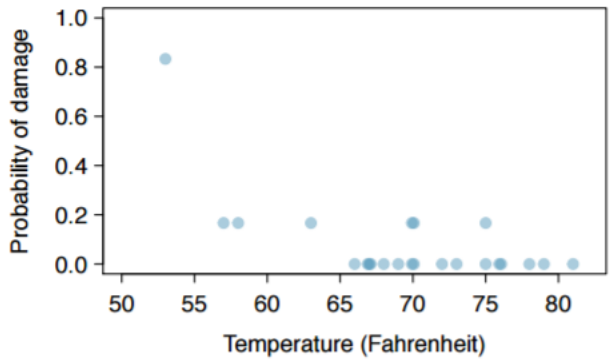
(a) The largest number (by a large margin) of damaged O-rings were observed for the launch at 53 degrees. There was also one damaged O-ring for 57, 58 and 63 degrees, respectively. For all remaining temperatures from 66 to 81 (29 observations total), there were 3 damaged O-rings.

(b) The intercept tells us how many damaged O-rings we would get at a temperature of zero, though we would never use the model for zero because it is way outside of our range of valules. The temperature slope tells us that for each 1-degree rise in temperature, the number of damaged O-rings would drop by -0.2162. The standard error gives us a measure of the variability for the intercept and slope. The z-scores indicates how many standard deviations each element is from the mean. Lastly, the right-most column tells us the p-value for the statistic.

(c) # of damaged O-rings = 11.663 – 0.2162 * temperature

(d) Yes, because the p-values are very small

8.18 Challenger disaster, Part II. Exercise 8.16 introduced us to O-rings that were identified as a plausible explanation for the breakup of the Challenger space shuttle 73 seconds into takeoff in 1986. The investigation found that the ambient temperature at the time of the shuttle launch was closely related to the damage of O-rings, which are a critical component of the shuttle. See this earlier exercise if you would like to browse the original data.



(a) The data provided in the previous exercise are shown in the plot. The logistic model fit to these data may be written as

$$\log\left(\frac{\hat{p}}{1-\hat{p}}\right) = 11.6630 - 0.2162 \times Temperature$$

where $\hat{p}$ is the model-estimated probability that an O-ring will become damaged. Use the model to calculate the probability that an O-ring will become damaged at each of the following ambient temperatures: 51, 53, and 55 degrees Fahrenheit. The model-estimated probabilities for several additional ambient temperatures are provided below, where subscripts indicate the temperature:

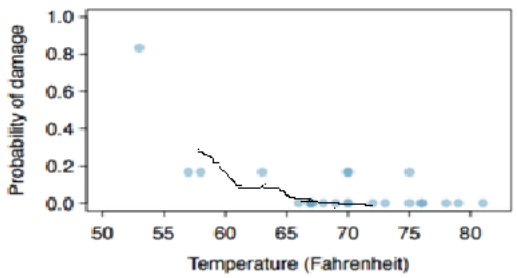
$$\hat{p}_{57} = 0.341 \qquad \hat{p}_{59} = 0.251 \qquad \hat{p}_{61} = 0.179 \qquad \hat{p}_{63} = 0.124$$
$$\hat{p}_{65} = 0.084 \qquad \hat{p}_{67} = 0.056 \qquad \hat{p}_{69} = 0.037 \qquad \hat{p}_{71} = 0.024$$

(b) Add the model-estimated probabilities from part (a) on the plot, then connect these dots using a smooth curve to represent the model-estimated probabilities.

(c) Describe any concerns you may have regarding applying logistic regression in this application, and note any assumptions that are required to accept the model's validity.

(a) 51 degrees = $e^{0.6368} / (1 + e^{0.6368}) = 0.654$
53 degrees = $e^{0.2044} / (1 + e^{0.2044}) = 0.551$
55 degrees = $e^{-0.228} / (1 + e^{-0.228}) = 0.443$

(b)



(c) The assumption for logistic regression are 1) each predictor x_i is linearly related to logit(p_i) if all other predictors are held constant, and 2) each outcome Y_i is independent of the other

outcomes. Regarding the first assumption, there are no other predictors, which could mean we are overlooking other predictors which are collinear or confounding. Furthermore, assumption one is in question because the relationship does not appear to be linear. Finally, it is not clear to me if the same shuttle is being used repeatedly or if there is a newly constructed shuttle each launch. If the same shuttle is used multiple times and I am not convinced assumption 2 holds.