Practice Exercises: Lesson 6.3

STAT 1201
Introduction to Probability and Statistics

Exercises

9.1 Baby weights, Part I. The Child Health and Development Studies investigate a range of topics. One study considered all pregnancies between 1960 and 1967 among women in the Kaiser Foundation Health Plan in the San Francisco East Bay area. Here, we study the relationship between smoking and weight of the baby. The variable `smoke` is coded 1 if the mother is a smoker, and 0 if not. The summary table below shows the results of a linear regression model for predicting the average birth weight of babies, measured in ounces, based on the smoking status of the mother.[9]

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	123.05	0.65	189.60	0.0000
smoke	-8.94	1.03	-8.65	0.0000

The variability within the smokers and non-smokers are about equal and the distributions are symmetric. With these conditions satisfied, it is reasonable to apply the model. (Note that we don't need to check linearity since the predictor has only two levels.)

(a) Write the equation of the regression model.

(b) Interpret the slope in this context, and calculate the predicted birth weight of babies born to smoker and non-smoker mothers.

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

9.2 Baby weights, Part II. Exercise 9.1 introduces a data set on birth weight of babies. Another variable we consider is `parity`, which is 1 if the child is the first born, and 0 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 model.

(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?

[9] Child Health and Development Studies, Baby weights data set.

9.3 Baby weights, Part III. We considered the variables `smoke` and `parity`, one at a time, in modeling birth weights of babies in Exercises 9.1 and 9.2. A more realistic approach to modeling infant weights is to consider all possibly related variables at once. Other variables of interest include length of pregnancy in days (`gestation`), mother's age in years (`age`), mother's height in inches (`height`), and mother's pregnancy weight in pounds (`weight`). Below are three observations from this data set.

	bwt	gestation	parity	age	height	weight	smoke
1	120	284	0	27	62	100	0
2	113	282	0	33	64	135	0
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
1236	117	297	0	38	65	129	0

The summary table below shows the results of a regression model for predicting the average birth weight of babies based on all of the variables included in the data set.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-80.41	14.35	-5.60	0.0000
gestation	0.44	0.03	15.26	0.0000
parity	-3.33	1.13	-2.95	0.0033
age	-0.01	0.09	-0.10	0.9170
height	1.15	0.21	5.63	0.0000
weight	0.05	0.03	1.99	0.0471
smoke	-8.40	0.95	-8.81	0.0000

(a) Write the equation of the regression model that includes all of the variables.

(b) Interpret the slopes of `gestation` and `age` in this context.

(c) The coefficient for `parity` is different than in the linear model shown in Exercise 9.2. Why might there be a difference?

(d) Calculate the residual for the first observation in the data set.

(e) The variance of the residuals is 249.28, and the variance of the birth weights of all babies in the data set is 332.57. Calculate the R^2 and the adjusted R^2. Note that there are 1,236 observations in the data set.

Chapter exercises

9.19 Multiple regression fact checking. Determine which of the following statements are true and false. For each statement that is false, explain why it is false.

(a) If predictors are collinear, then removing one variable will have no influence on the point estimate of another variable's coefficient.

(b) Suppose a numerical variable x has a coefficient of $b_1 = 2.5$ in the multiple regression model. Suppose also that the first observation has $x_1 = 7.2$, the second observation has a value of $x_1 = 8.2$, and these two observations have the same values for all other predictors. Then the predicted value of the second observation will be 2.5 higher than the prediction of the first observation based on the multiple regression model.

(c) If a regression model's first variable has a coefficient of $b_1 = 5.7$, then if we are able to influence the data so that an observation will have its x_1 be 1 larger than it would otherwise, the value y_1 for this observation would increase by 5.7.

(d) Suppose we fit a multiple regression model based on a data set of 472 observations. We also notice that the distribution of the residuals includes some skew but does not include any particularly extreme outliers. Because the residuals are not nearly normal, we should not use this model and require more advanced methods to model these data.

9.20 Logistic regression fact checking. Determine which of the following statements are true and false. For each statement that is false, explain why it is false.

(a) Suppose we consider the first two observations based on a logistic regression model, where the first variable in observation 1 takes a value of $x_1 = 6$ and observation 2 has $x_1 = 4$. Suppose we realized we made an error for these two observations, and the first observation was actually $x_1 = 7$ (instead of 6) and the second observation actually had $x_1 = 5$ (instead of 4). Then the predicted probability from the logistic regression model would increase the same amount for each observation after we correct these variables.

(b) When using a logistic regression model, it is impossible for the model to predict a probability that is negative or a probability that is greater than 1.

(c) Because logistic regression predicts probabilities of outcomes, observations used to build a logistic regression model need not be independent.

(d) When fitting logistic regression, we typically complete model selection using adjusted R^2.

9.21 Spam filtering, Part I. Spam filters are built on principles similar to those used in logistic regression. We fit a probability that each message is spam or not spam. We have several email variables for this problem: `to_multiple`, `cc`, `attach`, `dollar`, `winner`, `inherit`, `password`, `format`, `re_subj`, `exclaim_subj`, and `sent_email`. We won't describe what each variable means here for the sake of brevity, but each is either a numerical or indicator variable.

(a) For variable selection, we fit the full model, which includes all variables, and then we also fit each model where we've dropped exactly one of the variables. In each of these reduced models, the AIC value for the model is reported below. Based on these results, which variable, if any, should we drop as part of model selection? Explain.

Variable Dropped	AIC
None Dropped	1863.50
`to_multiple`	2023.50
`cc`	1863.18
`attach`	1871.89
`dollar`	1879.70
`winner`	1885.03
`inherit`	1865.55
`password`	1879.31
`format`	2008.85
`re_subj`	1904.60
`exclaim_subj`	1862.76
`sent_email`	1958.18

See the next page for part (b).