1. 40 pts. The data for this problem come from accident investigations involving Ford Explorer vehicles with a specific type of tire. These tires were suspected of failing catastrophically, causing fatal accidents. The data are from a national data base of fatal accidents. Each accident is categorized as being caused by tire failure (tire = 1) or note (tire = 0), whether the vehicle was a Ford Explorer (ford = 1) or note (ford = 0), and the age of tires (age, values from 0 to 5 in years). In all questions in the problem, the response variable is tire (1 = caused by tire failure or 0 = caused by something else).

   (a) Here are the counts of accidents for each combination of cause of failure and type of vehicle. I have also calculated the probability that the accident was caused by tires for each type of vehicle.

Vehicle	Cause of Failure		Probability
Type	Other	Tire	
Ford	500	22	0.0421
Other	1794	5	0.00278

   Calculate the odds ratio that fills in the blank in this sentence:
   The odds that an accident is caused by tire failure for Ford Explorers is _____ times that for other vehicles
   Show your calculations:

It is known that older tires are more likely to fail, so a more careful evaluation of Ford Explorer tires will use the model

$$\log\left(\frac{\pi_i}{1-\pi_i}\right) = \beta_0 + \beta_1 \text{ ford}_i + \beta_2 \text{ age}_i, \tag{1}$$

where π_i is the probability of tire failure for tire i, and ford_i and age_i are the values of the ford variable (values of 0 or 1) and tire age variable for tire i.

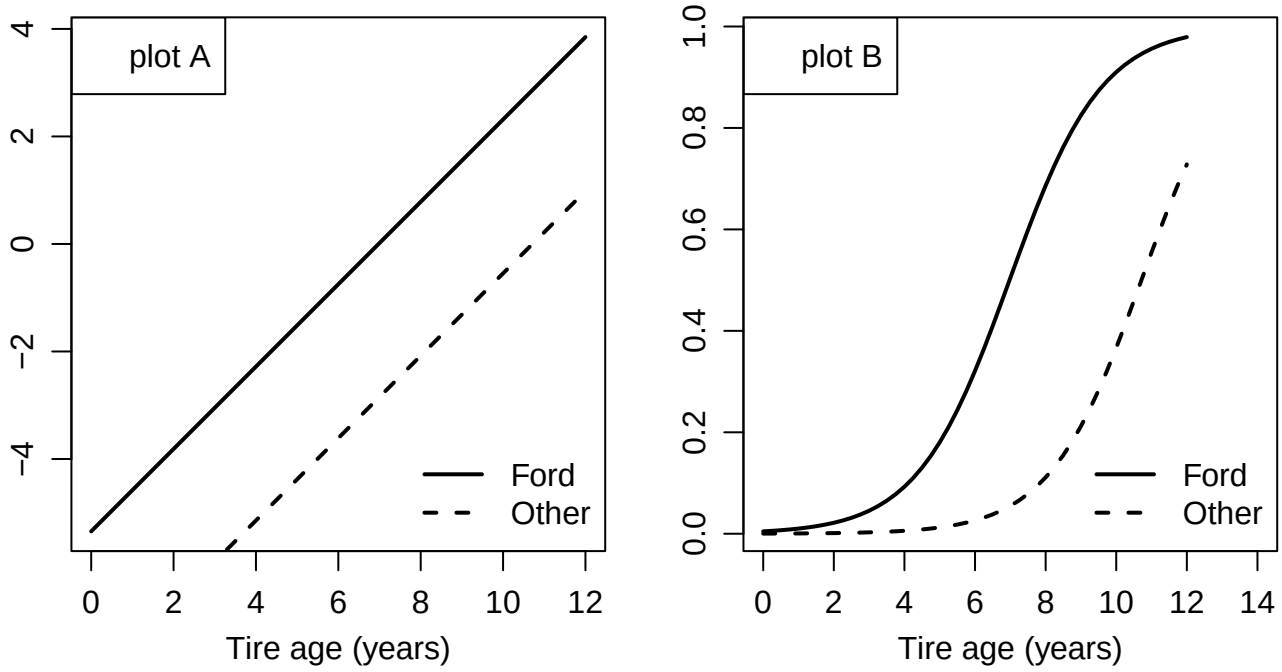
   (b) 3 pts. Circle the more appropriate name for equation (1) and briefly explain your answer.

linear regression, **logprobability regression**, or **logistic regression**.

   (c) Give two reasons why it is wrong to fit the usual multiple linear regression model to these data.

(d) Here are two plots of the fitted model: A on the left and B on the right. In both, the X axis is tire age. The Y axis has not been labeled. It could be odds, log odds, or probability. Circle the correct choice for each plot:

Plot	The Y axis label is:		
A	odds	log odds	probability
B	odds	log odds	probability



Hint for all remaining questions in problem 1: Combine multiple regression concepts and the interpretation that is appropriate for the response being modeled.

Equation (1) was fit to the data. Here are the estimated coefficients and their standard errors.

Coefficient	Estimate	s.e.
β_0	-8.208	0.765
β_1	2.863	0.503
β_2	0.766	0.169

The investigators also fit the model

$$\left(\frac{\pi_i}{1-\pi_i}\right) = \beta_0 + \beta_1 \text{ ford}_i. \tag{2}$$

The estimated coefficients and their standard errors for equation (2) are:

Coefficient	Estimate	s.e.
β_0	-5.883	0.448
β_1	2.759	0.498

(d) Calculate the odds ratio that fills in the blank in this sentence:
The odds that an accident is caused by tire failure for Ford Explorers is _____ times that for other vehicles, when you compare the same age tires.
Show your calculations:

(e) Calculate the t (or z) statistic that tests the null hypothesis that the odds ratio in question 1d = 1. Show your work

The investigators fit two additional models to evaluate assumptions made by model (1):

$$\log\left(\frac{\pi_i}{1-\pi_i}\right) = \beta_0 + \beta_1 \text{ ford}_i + \beta_2 \text{ age}_i + \beta_3(\text{ age}_i)^2, \quad (3)$$

and

$$\log\left(\frac{\pi_i}{1-\pi_i}\right) = \beta_0 + \beta_1 \text{ ford}_i + \beta_2 \text{ age}_i + \beta_3 \text{ age}_i \times \text{ ford}_i, \quad (4)$$

. Log-likelihood statistics for these two models and model (1) are:

Model	Log likelihood
(1)	-113.88
(3), i.e., (1) plus $(\text{age}_i)^2$	-112.70
(4), i.e., (1) plus $\text{age}_i \times \text{ford}_i$	-113.72

(f) Calculate the test statistic (i.e., the drop in deviance) for comparing model (3) to model (1). Show your work

(g) The p-value for the test in question 1f is 0.12. Write an appropriate conclusion for this test.

2. 50 pts. Cost of road construction, part 1

In the 1980's, the State of Florida investigated bid rigging in road construction. They suspected that road construction companies were agreeing to submit artificially high bids to increase their profits on state funded road construction jobs. This practice is called bid rigging. The state compiled information on 279 road construction projects. The state believed that the bidding was fair on 194 of these projects and was rigged on 85 of these projects. The state recorded:

logbid: log transformed contract price, in \$, of the lowest bid
loglength: log transformed length of the project, in miles
logdays: log transformed estimated number of days to complete the project
status: an indicator variable: 0 if the bidding was fair and 1 if it was suspected to be rigged.

There are 279 projects with complete information on these variables.

You do not need to back transform answers for any parts of this problem. All answers may be in terms of log-scale quantities.

Most results from fitting model (5) are in the computer packet:

$$\text{logbid}_i = \beta_0 + \beta_1 \text{status}_i + \beta_2 \text{loglength}_i + \beta_3 \text{logdays}_i + \varepsilon_i \tag{5}$$

The sequential (type I) and partial (type III) sums-of-squares (SS) and associated p-values for model (5) are:

	sequential		partial	
Term	SS	p-value	SS	p-value
status	2.60	0.0074	4.03	0.0009
loglength	2.24	0.013	9.25	< 0.0001
logdays	407.02	< 0.0001	407.02	< 0.0001

(a) Calculate the error degrees of freedom when fitting model (5). Show your work.

(b) Predict the value of logbid for a road project with loglength = 2, logdays = 5, and status = 0. Show your work.

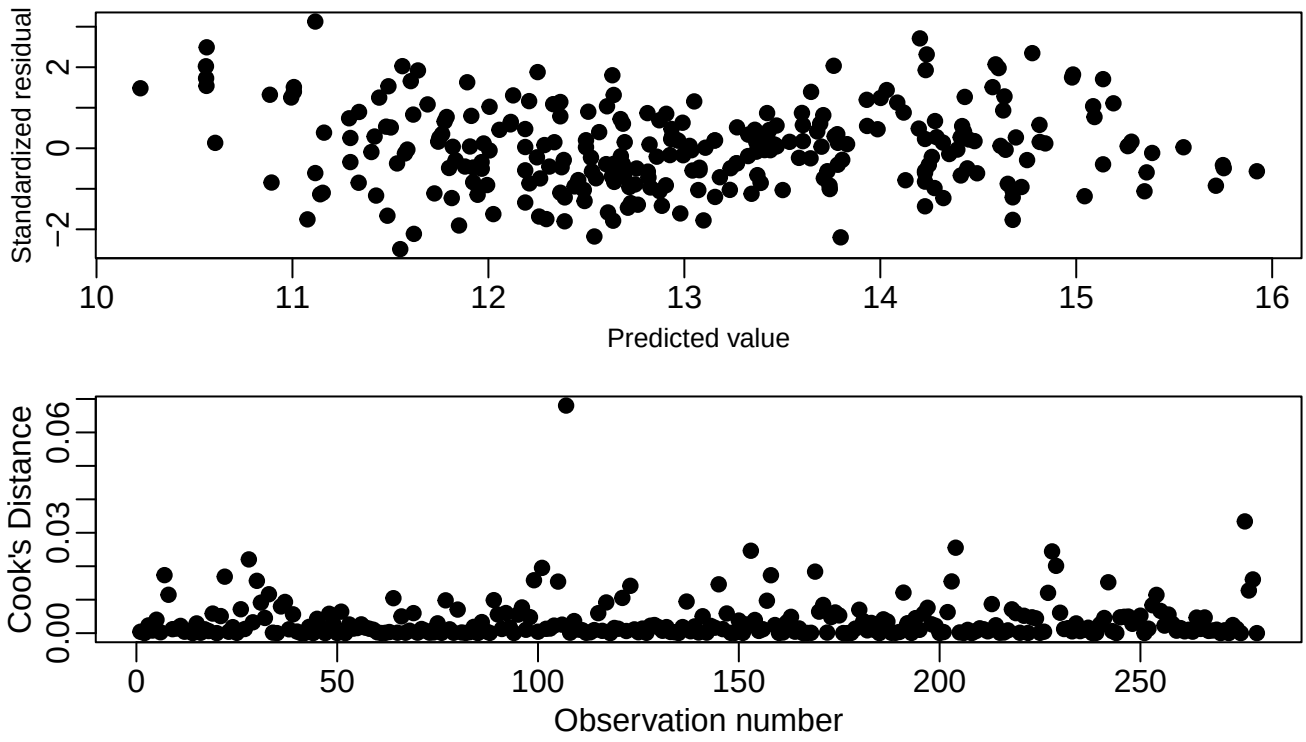
(c) Estimate the difference in logbid price between a project with loglength = 2 and logdays = 5 that was suspected to be rigged and a project with loglength = 2 and logdays = 5 that was believed to be fair.

(d) Is the estimate in question 2c significantly different from 0? I.e., is the p-value for that test < 0.05? Briefly explain your answer. If you need additional information, briefly describe what you need.

(e) You want to test the null hypothesis that $\beta_1 = 0$, i.e., there is no difference between the two levels of status, when compared at the same values of loglength and logdays. What are the appropriate sums-of-squares and p-value for this test?

(f) Briefly explain, as if to your major professor, why the sequential and partial sums-of-squares for logdays are identical.

Here are diagnostic plots of standardized residuals and Cook's distance.



(a) Are there any concerns about unusually large or unusually small residuals? Answer yes or no. If yes, circle those observations on the relevant plot.

(b) Are there any concerns about influential observations? Answer yes or no. If yes, circle those observations on the relevant plot.

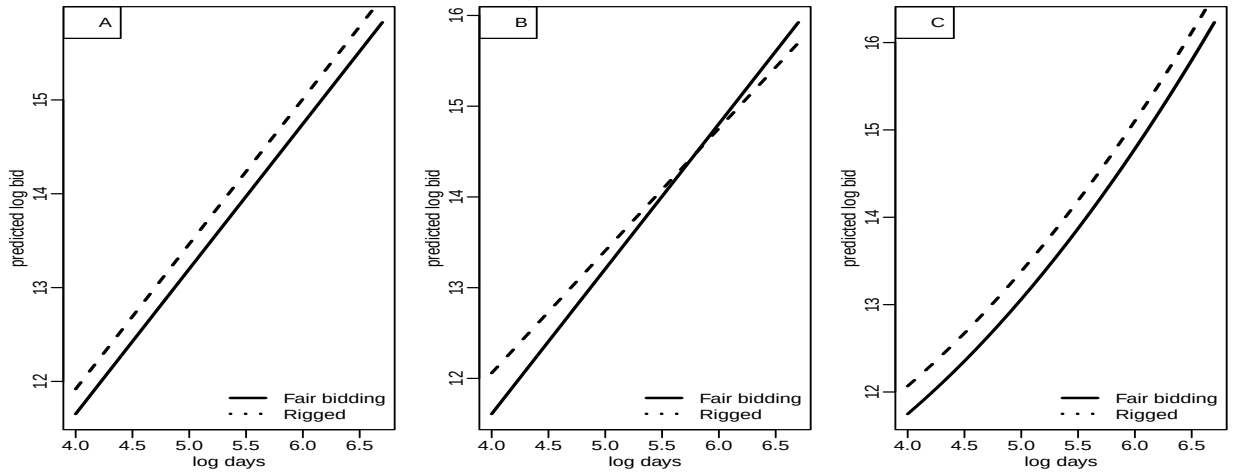
Consider the model :

$$\text{logbid}_i = \beta_0 + \beta_1 \text{loglength}_i + \beta_2 \text{logdays}_i + \beta_3 \text{status}_i + \beta_4 \text{loglength}_i \times \text{status}_i + \beta_5 \text{logdays}_i \times \text{status}_i + \varepsilon_i \quad (6)$$

Estimated coefficients and their standard errors are:

	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\beta}_4$	$\hat{\beta}_5$
Estimate	5.04	0.074	1.59	1.47	−0.01	−0.25
Std. Err.	0.0261	0.0017	0.051	0.543	0.029	0.112

(i) Imagine a collection of projects, all for 10 miles of road, i.e. loglength=log(10) but different estimated numbers of days (from 30 to 900). Then consider what the plot of Y = predicted logbid against X = log days would look like. Here are three possible plots for model (6).



Which plot best shows the relationship between logdays and predicted logbid for model (6)? Your answer is the letter (A, B, or C) and a brief explanation for your choice.

(j) The estimated status effect (difference between status = 1 and status = 0) in model (6) is 1.47 This seems really large when the observed average log bids are 13.09 for status = 1 and 12.88 for status = 0. That difference is 0.21. Briefly explain, as if to your major professor, the interpretation of the status coefficient in model (6). Your explanation should make it clear why the estimated status effect (1.47) is so different from the observed difference (0.21).

3. 25 pts. Cost of road construction, part II

The state is very interested in the difference in the logbid amount between projects with suspected rigging (status = 1) and those believed fair (status = 0). The state wants to estimate this difference after adjusting for relevant features of a project that are associated with the bid amount. We will define the **the RigEffect** as this difference, holding constant other relevant variables. The 10 potential relevant features are:

ldays: log of the projected number of days
llen: log of the roadway length
nbids: number of bids on the project.
pasph: percent of costs attributable to asphalt
pbase: percent of costs attributable to base material
pexc: percent of costs attributable to excavation
pmob: percent of costs attributable to mobilization
pstru: percent of costs attributable to structures
ptraf: percent of costs attributable to traffic control
sub: indicator for subcontractor. 0 = no subcontractor, 1 = project included a subcontract

The complete list of variables has **logbid** (the response variable) **status** (whether or not bid rigging was suspected), and 10 other features of the project. There are 217 projects with complete information for all these variables.

The investigators use model selection to choose an appropriate set of the 10 "other features" (i.e., not including status) to predict logbid for new observations. Here is information about a few of these models. An * indicates that variable (column name) was included in that model. Model numbers are in the first column. AICc is the corrected Akaike Information Criterion, BIC is the Bayesian Information Criterion, and PRESS is the Predicted Residual Error Sum-of-Squares.

#	ldays	llen	nbids	pasph	pbase	pexc	pmob	pstru	ptraf	sub	AICc	BIC	PRESS
1	*	*		*		*					373.7	394.0	67.28
2	*	*		*		*	*			*	375.2	398.8	67.92
3	*	*		*	*	*	*			*	375.9	403.0	68.08
4	*	*									392.0	405.9	75.79
5	*	*	*	*	*	*	*	*	*	*	382.7	423.6	70.25

(a) If you use AICc as the criterion, which is the most appropriate model? Report the number of that model and briefly explain your choice.

(b) Is model 2 a reasonable alternative to the model chosen in question 3a? Briefly explain your answer.

(c) Explain, as if to your major professor, why model 4 has a large PRESS statistic, compared to the other models.

(d) When residuals are calculated using model 5 fit to all observations, the error sums-of-squares of the residuals is 66.36. Explain, as if to your major professor, why the PRESS is larger than the error sums-of-squares.

4. 5 pts. Imagine a conversation with a student just starting this course. They want to know how to be most successful in the course. What are the two most useful pieces of advice you can give them?

That's all - good luck on the rest of your exams and have a good break.