

The first five questions are worth five points each. In each case, circle **True** or **False**. If you answer **False**, make a small change to the statement to make it true or briefly (in one or two sentences) explain why it is false.

**Problem 1 (5 points)**

**True** or **False** —In multiple regression, the fit of seven points  $(X_i, Y_i)$  with a degree six polynomial

$$Y = \hat{\beta}_0 + \hat{\beta}_1 X + \hat{\beta}_2 X^2 + \cdots + \hat{\beta}_6 X^6$$

will have an  $R^2$  of 1.

Solution: **True**. A degree 6 polynomial has 7 free parameters. With only 7 data points, all residuals will be 0 and the fit will go through each data point.

**Problem 2 (5 points)**

**True** or **False** —In a multiple regression setting, if a point has a large leverage value, dropping the point from the analysis must result in a large change to the estimated regression coefficients.

Solution: **False**. Leverage is not influence. A point with large leverage has values of explanatory variables far from other points. It has the *potential* to be influential, but it may not if the predicted value from the other points alone is close to the actual value.

**Problem 3 (5 points)**

**True** or **False** —Using AIC as a criterion, if forward selection and backward elimination each end with the same model, then this must be the model with the lowest AIC over all possible models that use some of the full set of variables.

Solution: **False**. Forward selection and backward elimination are heuristic search methods and may not find the actual global optimal model.

**Problem 4 (5 points)**

**True** or **False** —Sixteen dairy cattle have their average milk yield measured during each of four different diets, a total of 64 means overall. In a two-way ANOVA with cow, diet, and a cow/diet interaction as the explanatory variables and yield as the response variable, a plot of residuals versus fitted values would be useful in seeing if a transformation of the data would be helpful.

Solution: **False**. This is a two-way ANOVA where there is no replication. Each cow provides one measurement per diet. The saturated model with both main effects (cow and diet) plus the cow diet interaction term is equivalent to a model that allows for a different mean for each cow and diet, and thus would fit the data exactly. All residuals would be 0, so the residual plot would be pretty meaningless.

**Problem 5 (5 points)**

**True** or **False** —A garage collects data on cars that come in for repair. One variable is the number of tires that need to be replaced. Other variables include the mileage, the car manufacturer (make), the type of car, the year it was built, and the color. If we wished to model the number of tires that need to be replaced as a function of the other variables, logistic regression would be a reasonable model choice.

Solution: **False**. In logistic regression, response variables are 0 or 1. In this problem, the response is a count from 0 to 4.

The following five problems are worth a total of 45 points and are based on the poison data set.

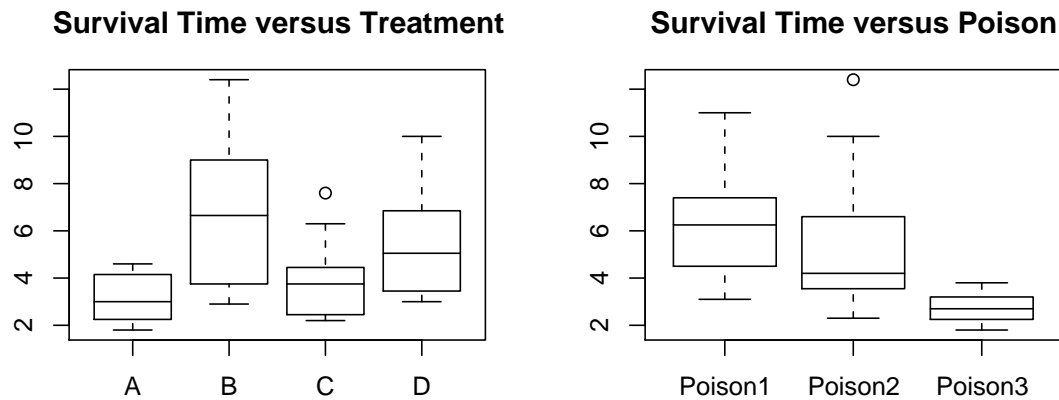
In an animal experiment, animals are randomly assigned to four treatment groups. After receiving the treatment, each animal was given one of three poisons and the survival time in hours was recorded.

Here is a summary of the data.

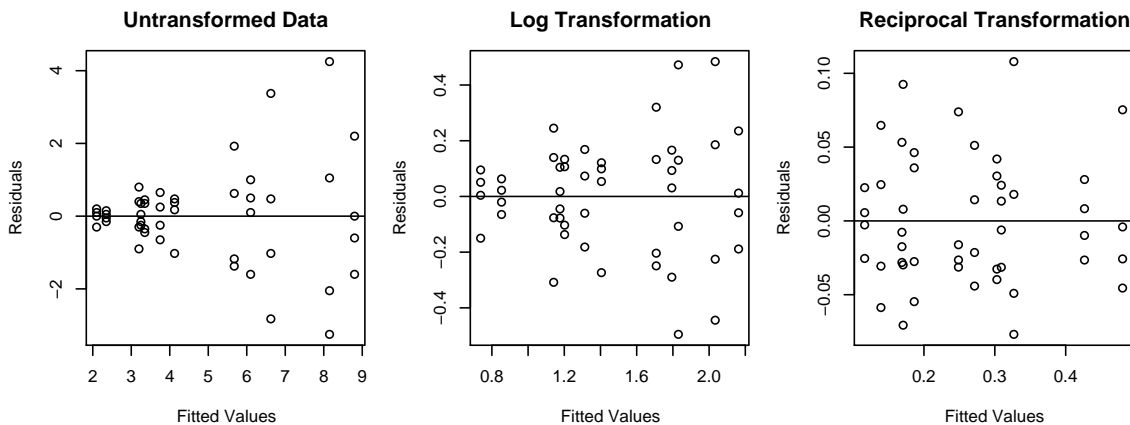
```

'data.frame':      48 obs. of  3 variables:
 $ poison      : Factor w/ 3 levels "Poison1","Poison2",...: 1 1 2 2 3 3 1 1 2 2 ...
 $ treatment   : Factor w/ 4 levels "A","B","C","D": 1 1 1 1 1 1 1 1 1 1 ...
 $ survivalTime: num  3.1 4.6 3.6 4 2.2 1.8 4.5 4.3 2.9 2.3 ...
    
```

One graphical summary of the data is side-by-side boxplots, one for treatment, and one for poison.



We can fit a saturated model and examine the residual plots to check if the fit is adequate. Alternative models look at the log transformation or the reciprocal transformation.



**Problem 6 (5 points)**

Explain why a statistician may prefer the reciprocal transformation for this data, other than the fact that the reciprocal of a survival time has a simple interpretation, the *death rate*.

Solution: This transformation results in data for which the constant variance assumption is most reasonable.

Consider for the next several problems the reciprocal transformed data. Here is an ANOVA table for the saturated model and an ANOVA table for an additive model.

**Model with an interaction**

## Analysis of Variance Table

Response: 1/survivalTime

	Df	Sum Sq	Mean Sq	F value	Pr(>F)	
treatment	3	0.20396	0.06799	28.4100	1.336e-09	***
poison	2	0.34863	0.17432	72.8419	2.217e-13	***
treatment:poison	6	0.01567	0.00261	1.0911	0.3864	
Residuals	36	0.08615	0.00239			

---

Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

**Model without an interaction**

## Analysis of Variance Table

Response: 1/survivalTime

	Df	Sum Sq	Mean Sq	F value	Pr(>F)	
treatment	3	0.20396	0.06799	28.045	4.063e-10	***
poison	2	0.34863	0.17432	71.906	2.740e-14	***
Residuals	42	0.10182	0.00242			

---

Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

**Problem 7 (15 points)**

Carry out an extra-sum-of-squares  $F$ -test to see if the saturated model fits significantly better than the additive model. In this analysis, (a) state null and alternative hypotheses in words; (b) report a test statistic; (c) report a  $p$ -value and the reference distribution from which this  $p$ -value is computed; (d) summarize the results of this test in the context of the problem, without statistical jargon.

Solution:

- Null hypothesis — there is no interaction between poison and treatment. Alternative hypothesis — there is an interaction between poison and treatment.
- $F = 1.0911$ .
- $p = 0.3864$  from an  $F$  distribution with 6 and 36 degrees of freedom.
- The data is consistent with a no interaction between poison and treatment. In other words, there is no reason to think that the effects of the different poisons on survival time are different for different treatments.

**Problem 8 (5 points)**

Suppose that we had an additional variable  $ID$  that was a unique identifier for each of the 48 animals. For each possible pair of the explanatory factors  $ID$ ,  $treatment$ ,  $poison$ , (there are three pairs), say if these factors are crossed or nested.

Solution:  $ID$  is nested within  $treatment$ , because each animal has only one treatment, not all treatments. $ID$  is nested within  $poison$ , because each animal is given only one poison, not all poisons.

The variables *poison* and *treatment* are crossed because there are measurements taken for each of the twelve combinations.

### Problem 9 (5 points)

Explain why the investigators could not use a repeated measures design for this experiment and measure each animal with each poison.

Solution: Once an animal has died from the effects of one poison, it cannot be killed again.

Here is a summary of the regression coefficients from an additive model. You may also wish to consult the ANOVA for the additive model on the previous page.

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	0.26972481	0.01740770	15.494572	5.641416e-19
treatmentB	-0.16574024	0.02010068	-8.245505	2.548206e-10
treatmentC	-0.05721354	0.02010068	-2.846349	6.812240e-03
treatmentD	-0.13567043	0.02010068	-6.749545	3.319768e-08
poisonPoison2	0.04698667	0.01740770	2.699190	9.971370e-03
poisonPoison3	0.19964249	0.01740770	11.468633	1.607680e-14

### Problem 10 (15 points)

Please answer these questions.

- What is the estimated death rate of an animal using Treatment A and Poison 3?
- How strong is the evidence that the treatments have different effects on death rate?
- Find a 95% confidence interval for the difference in *death rates* for Poison 1 and Poison 2 (averaged across all four treatments).

Solution:

- The intercept is the fitted death rate for Treatment A and Poison 1. The other coefficients are for changes from this reference combination. We simply need to add the intercept to the coefficient for the difference between Poison 3 and Poison 1, or  $0.2697 + 0.1996 = 0.4694$  deaths per hour.
- From the ANOVA table from the additive model on the previous page, the  $p$ -value for the treatment is very small (about  $4 \times 10^{-10}$ ), so there is strong evidence that treatment effects are different. In addition, the box plots on the previous page show large differences among treatments.

We really have to be a bit careful in using the ANOVA table. The sums of squares are *sequential*, so the  $F$  value is from a comparison of a model with an intercept only with a model that contains a treatment factor. However, because the design is balanced, the order of *treatment* and *poison* does not matter and we would get the same  $F$  statistic and  $p$ -value if the order were switched.

- This confidence interval is centered at the coefficient for Poison 2. The  $t$  multiplier comes from the 0.975 quantile of a  $t$  distribution with 42 degrees of freedom, which is a little more than 2.

$$0.047 \pm 2.02 \times 0.0174 = 0.047 \pm 0.0351$$

deaths per hour.

**There are three remaining problems worth 30 points that are from the AFLP marker data set.**

An AFLP marker is a kind of genetic marker that is present or absent in each individual. This data set contains two AFLP markers (from a much larger experiment) for each of 111 varieties of alfalfa. Other variables include yield, a quantitative variable, and population, a categorical variable. If the yield is a good predictor of a marker, it is possible that the marker is located close to a gene that affects yield.

What follows are the summaries of two separate logistic regressions, one for the first marker and one for the second marker. In each case, the population is a blocking variable that should be controlled for, but is not of direct interest. The interaction between yield and population is far from statistical significance (analysis not shown). Individuals from the same population are more closely related, genetically, and so we might expect their AFLP markers to be more similar to one another.

### Logistic regression fit of marker 1

```
'data.frame':      111 obs. of  4 variables:
 $ marker1   : int  0 0 0 0 0 0 0 0 0 1 ...
 $ marker2   : int  0 0 0 0 0 1 1 1 1 0 ...
 $ yield     : num  313 331 334 348 364 ...
 $ population: Factor w/ 6 levels "pop1","pop2",...: 1 1 1 1 1 1 1 1 1 1 ...
```

	Estimate	Std. Error	z value	Pr(> z )
(Intercept)	5.55114202	2.023750605	2.74299709	0.006088122
yield	-0.01745276	0.006038893	-2.89005922	0.003851693
populationpop2	-0.40289272	0.741366888	-0.54344580	0.586822900
populationpop3	0.44471353	0.927831915	0.47930398	0.631722391
populationpop4	0.05291225	0.631051626	0.08384774	0.933177490
populationpop5	1.21390786	0.779191417	1.55790713	0.119255266
populationpop6	-0.18142107	0.661007811	-0.27446132	0.783730127

### Logistic regression fit of marker 2

	Estimate	Std. Error	z value	Pr(> z )
(Intercept)	1.527117013	2.034132111	0.75074623	0.4528054
yield	-0.003595686	0.006056134	-0.59372627	0.5526952
populationpop2	0.029595591	0.723331950	0.04091564	0.9673631
populationpop3	7.214729200	10.079482562	0.71578369	0.4741249
populationpop4	0.334526783	0.623625423	0.53642262	0.5916665
populationpop5	2.358102635	1.155718544	2.04037795	0.0413127
populationpop6	0.076253341	0.642793717	0.11862801	0.9055701

#### Problem 11 (10 points)

For marker 1, what change in the odds ratio of the presence of a marker do you expect to see if the yield were to increase by one unit? (*Hint: take the exponential of the change in the log odds ratio. See page 586.*)

Solution:  $e^{-0.0175} = 0.983$ .

#### Problem 12 (10 points)

For the logistic regression fit of the first marker versus explanatory variables, find a 95% confidence interval for the change in the odds ratio that the marker will be present after an increase in yield of 1 unit.

Solution: On the log scale, the 95% confidence interval is  $-0.0036 \pm 1.96 \times 0.0061$ . Taking exponentials, we get the interval from 0.971 to 0.994. We are 95% confident that the multiplicative effect of yield on the odds ratio is in this interval.

**Problem 13 (10 points)**

A 95% confidence interval for the change in odds ratio for marker 2 after an increase in yield of 1 unit is (0.98, 1.01). For which marker is there greater evidence that the marker is located close to a gene that has a quantitative effect on yield? If this marker is present, would you expect a larger or smaller yield? Explain.

Solution: An odds ratio of one corresponds to no change in the odds. The 95% confidence interval for marker 1 is strictly below one, so there is evidence that an increase in yield would make it less likely that marker 1 is present. Inverting this, presence of marker 1 might predict a lower yield, so marker 1 may be close to a gene with a negative effect on yield. (The sign of the coefficient on the log-scale is negative.) In contrast, marker 2 has a confidence interval that includes one, so the observed data is consistent with yield having no predictive power about the presence of marker 2 (and inversely, marker 2 not being close to a gene that affects yield).