Questions (1) through (7) apply to the fish PCB data below.

Fish polychlorinated-biphenyl (PCB) measurements are usually taken on fish filets or whole fish. Amrhein et al. (1999, *Environ. Toxic. and Chem.*) collected data to determine the best predictors of whole-fish PCB concentrations in two Lake Michigan fishes.

For 213 fish, the variables measured were:

- **wfpcb**: percent PCB in the whole fish
- **length**: fish length
- **ffat**: percent lipids in the fish filet
- **fpcb**: percent PCB in the fish filet
- **type**: fish type = 1 for rainbow trout; fish type = 0 for coho salmon

The following model is fit to the data:

log(wfpcb) ~ length + log(ffat) + log(fpcb) + type

```
Call: lm(formula = log(wfpcb) ~ length + log(ffat) + log(fpcb) + type)

Coefficients:
               Value Std. Error  t value Pr(>|t|)
(Intercept)   0.3161   0.1703     1.8569   0.0647
     length   0.0066   0.0024     2.7665   0.0062
  log(ffat)  -0.3296   0.0550     ------   ------
  log(fpcb)   0.7790   0.0420    18.5654   ------
       type  -0.0904   0.0662    -1.3653   ------

Residual standard error: 0.3089 on --- degrees of freedom
Multiple R-Squared: 0.8584
F-statistic: 315.3 on - and --- degrees of freedom, the p-value is 0

           Df Sum of Sq  Mean Sq  F Value     Pr(F)
length      -  53.77050    ----- 563.4016     ----
log(ffat)   -   2.98354  2.98354  31.2612     ----
log(fpcb)   -  63.42949 63.42949 664.6075     ----
type        -   0.17790  0.17790   1.8640     ----
Residuals ---  19.85132  0.09544
```

1. Write a 1-sentence interpretation of the coefficient of log of percent lipids in the fish filet (*log(ffat)*) for rainbow trout. [5 points]

2. Perform a hypothesis test to determine if the log of percent PCB in whole fish (*log(wfpcb)*) is negatively associated with the log of percent lipids in filet (*log(ffat)*) for rainbow trout. Give hypotheses, rejection region and conclusion of your test. Use α=0.05. [10 points]

3. Give a confidence interval for the coefficient of the log of percent PCB in filet (*log(fpcb)*). [5 points]

4. Another researcher uses the model output given and a Bonferonni multiple comparison procedure to create confidence intervals for all 5 parameters in the model. What can you say about the interval that he calculates for the log of percent PCB in filet (*log(fpcb)*)? [5 points]

   (a) It will be narrower than the interval in problem (3) above.

   (b) It will be wider than the interval in problem (3) above.

   (c) It will be the same as the interval in problem (3) above.

   (d) There is not enough information to answer this question.

5. Perform a test to determine the joint significance of fish type (*type*) and log of percent PCB in filet (*log(fpcb)*) in the model. Give the hypotheses, rejection region and conclusion for your test. [10 points]

6. **True or False**. *Circle one.* The ANOVA table provided in the Splus output above allows us to compare the larger model, *log(wfpcb) ~ length + log(ffat) + log(fpcb) + type*, to the reduced model, *log(wfpcb)~ length + type.* [5 points]

7. The study authors recenter the length, percent PCB in filet (*fpcb*), and percent lipids in filet (*ffat*) variables at their respective means, take log transforms as before, and re-run the model. What does the intercept mean in this case? [5 points]

Questions 8 through 11 refer to the indoor air pollution data discussed below.

In indoor air quality studies, a crucial input to exposure models is an estimate of the ventilation rate in a room, referred to as the *air exchange rate.* This is the rate at which the air in the room is replaced with fresh air, and is usually reported in units of "air changes per hour."

The air exchange rate can be estimated based on experimental data as follows. A tracer gas is released in a room for a short period of time at a fixed concentration, and then the subsequent decay in tracer gas concentrations over time is recorded, as seen in Figure 1.

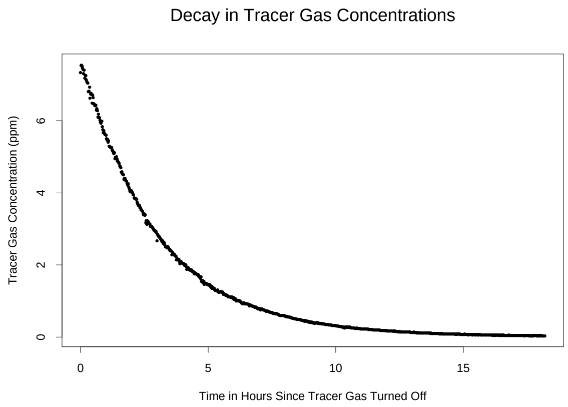

Figure 1: Plot of measured tracer gas concentrations versus time in hours.

Civil engineers use the following equation to describe the concentrations during the decay period:

$$Y = ae^{bX} \tag{1}$$

where Y = concentration of tracer gas (in ppm), X = time in hours (where $X = 0$ hours corresponds to time when the tracer gas is shut off and the decay period begins), and b is the estimated air exchange rate.

8. How could you reformulate the relationship in (1) so that you could estimate b using simple linear regression? Write out the linear regression equation. [10 points]

9. Interpret the meaning of the parameter a in the model. [5 points]

10. For the indoor air pollution data, use the model you wrote in problem (8) to interpret the effect on the concentration in ppm of a one-hour change in time. Give your answer in a sentence in terms of model parameter(s). [5 points]

11. For the indoor air pollution data, a naive researcher performs a regression of Y in ppm on X in hours. What can be said about the model she fits? *Circle all that apply.* [5 points]

    (a) Fitted values of Y would overshoot the true mean of Y at some values of X.

    (b) Bias would be seen in the difference between the fitted values and the mean response.

    (c) The residuals of the regression would have a horn shape; that is, the residuals would have steadily increasing variance as X increases.

    (d) The model would overestimate the tracer gas concentration for low values of X and for high values of X.

12. A researcher wants to know the optimum level of fertilization to maximize vertical growth (in cm) of a plant. 100 seedlings were divided into 4 groups, 25 plants each. Each was planted in a similar pot, given a prespecified level of fertilizer (100 mg, 500 mg, 1000 mg, 2000 mg) and grown under identical conditions for several weeks.

    The researcher hires 2 consultants, telling each one to fit a model that describes plant growth as a function of fertilizer level. Consultant 1 fits a one-way ANOVA model to the data. Consultant 2 fits a simple linear regression model to the data.

    What is similar about the two analyses? *Circle all that apply.* [5 points]

    (a) Both approaches would use the same null hypothesis to investigate the relationship between growth and fertilizer level.

    (b) Both approaches would use the same alternative hypothesis to investigate the relationship between growth and fertilizer level.

    (c) Both models have equal ability to detect linear differences in means of height at different fertilizer levels, if they exist.

    (d) Both models have the same assumptions about the random error term.

    (e) Both models would yield the same predictions of plant height at a fertilizer level of 1000 mg.

13. R^2 *Circle all that apply.* [5 points]

    (a) provides reliable guidance in selecting between numerous multivariate models.

    (b) gives an accurate measure of the magnitude of the slope of the regression line.

    (c) is a function of the residual sum of squares for the regression.

    (d) can be large even when the simple linear regression model is inadequate.

14. Suppose you want to check the predictive power of a model. List two ways to evaluate the accuracy of the model for prediction. [5 points]

15. A Bayesian statistician from Duke and a frequentist statistician from Stanford are each presented with a challenge to model Y as a linear function of 10 X variables. Each works separately to choose the best multivariate linear regression model using model selection among all subsets as described in Chapter 12 of *Statistical Sleuth.* What is similar about their approaches and results? [5 points]

    (a) At the end of the analysis, both will assess the average estimate of each of the X variables using a weighted average that represents the strength of belief in each of the models considered.

    (b) Both will use a model-fitting criterion that incorporates a penalty for the number of parameters.

    (c) Both will arrive at the same best model.

    (d) At the end of the analysis, both will assess the probabilities of different models in the model space.

    (e) Both will use the method of least squares to fit the different subset models.

16. Consider any linear regression which does not explain 100% of the variance in the dependent variable. For a particular X value, the prediction interval for an individual Y value will always be wider than the confidence interval for the mean value of Y. **Is this statement TRUE or FALSE? Explain why.** [5 points]

Questions 17 through 23 refer to the birthweight dataset described below.

In the following study, we are interested in finding out about factors related to low birth weights for babies. Researchers used records on 189 births at a US hospital, and looked into medical histories to determine age, weight, smoking status during pregnancy, and their history of hypertension.

The following variables are measured:

- **low**: low=1 if birth weight less than 2.5 kg; low=0 if birth weight greater than or equal to 2.5 kg.
- **age**: age of the mother in years
- **lwt**: weight of mother (in pounds) at last menstrual period
- **smoke**: smoke=1 if smoker during pregnancy; smoke=0 if non-smoker during pregnancy
- **ht**: history of hypertension (1 = Yes, 0 = No)

Questions 17 through 18 refer to the 2 by 2 table below.

For the birthweight data, the researcher begins a simple analysis on the relationship between smoking and low birthweight by creating a 2 by 2 table:

Table 1: Table of Birthweight and Smoker Counts

	Birthweight < 2.5 kg	Birthweight ≥ 2.5 kg
Smoker	30	44
Non-smoker	29	86

17. Using Table 1, give an estimate of the odds of a low birthweight baby given that the mother is a smoker. [5 points]

18. Using Table 1, which of the following is true? *Circle all that apply.* [5 points]

    (a) The probability of a low birthweight baby for smokers is 2.0 times that of non-smokers.

    (b) The odds of a low birthweight baby for smokers is 2.0 times that of non-smokers.

    (c) The odds ratio of a low birthweight baby for smokers is 2.0 times that of non-smokers.

    (d) The odds ratio of a low birthweight baby for non-smokers is 2.0 times that of non-smokers.

    (e) None of the above.

The following model is fit to the birthweight data ($n = 189$): **low** $\sim$ **age** $+$ **lwt** $+$ **smoke** $+$ **ht**. A portion of Splus output is given below.

```
Call: glm(formula = low ~ age + lwt + smoke + ht, family = "binomial",
data = birthwt,link = "logit")

Coefficients:
                  Value  Std. Error   t value
(Intercept)  1.76579622 1.044622092  1.690368
        age -0.03567799 ----------- -1.800202
        lwt -0.01694696 0.006552933 -2.586165
      smoke  0.67893947 0.330933555  2.051588
         ht  1.78777193 0.682605189  2.619042

    Null Deviance: 234.672 on 188 degrees of freedom

Residual Deviance: 215.6843 on 184 degrees of freedom
```

19. Using Wald's test, test the significance of the age effect after accounting for mother's weight (*lwt*), smoking status (*smoke*) and history of hypertension (*ht*). Write down hypotheses and give the p-value for your test. Compare your p-value to $\alpha = 0.05$ to make your conclusion. [5 points]

20. Use the model output above to determine how much the odds of having a low birthweight baby change for a smoker relative to a non-smoker. Give your result in a sentence. [5 points]

21. Based on the model above, what is the estimated odds of a low birthweight baby for a mother of age 34 and weight 128 pounds and who is a non-smoker and does not have a history of hypertension? [5 points]

22. Compare the odds of a low birthweight baby for these two women: the one from the previous problem (21) and an otherwise identical woman who does smoke and does have a history of hypertension. Write your result in a sentence. [5 points]