UNIVERSITY OF MASSACHUSETTS
Department of Mathematics and Statistics
Applied Statistics
August 2017

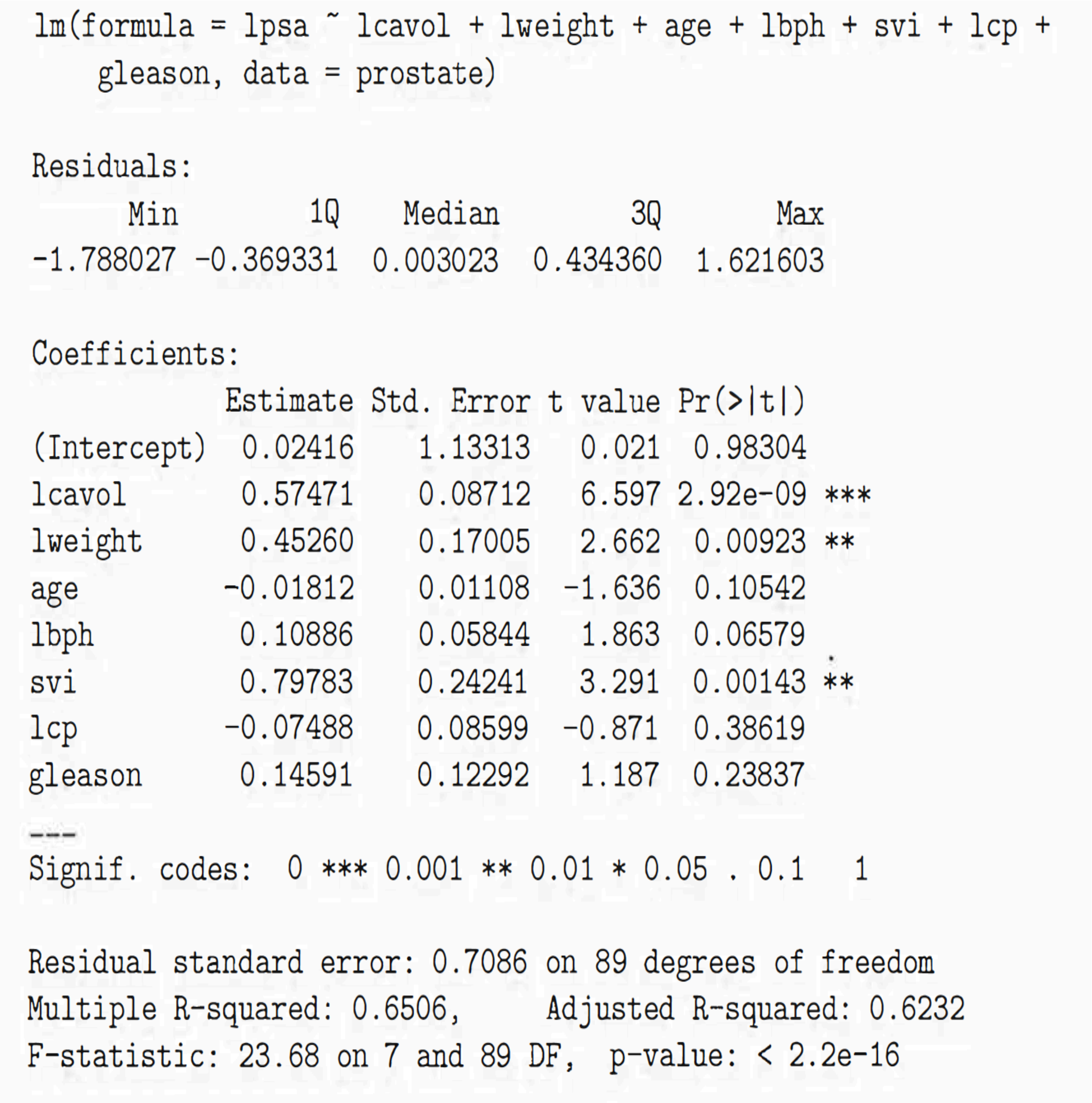
Work all problems. 60 points are needed to pass at the Masters Level and 75 to pass at the Ph.D. level.

1. Data were collected on a sample of 97 men who were due to receive a radical prostatectomy. The measured variables included *lcavol*=log(cancer volume), *lweight*=log(prostate weight), *age*, *lbph*=log(benign prostatic hyperplasia amount), *svi*=indicator of seminal vesicle invasion, *lcp*=log(capsular penetration), *gleason*=Gleason score and *lpsa*=log(prostate specific antigen). The goal of the study was to relate *lpsa* to the other variables, which represent indications of the severity of the patients prostate cancer. The output of a multiple regression analysis appears in the next page.

   (a) (6pts) Is there a significant linear relationship between *lpsa* and the predictors? Explain your answer.

   (b) (6pts) Are there predictors that do not seem to contribute to prediction of *lpsa*? Which predictors would you consider removing from the model? Explain.

   (c) (6pts) The model states that the presence of seminal vesicle invasion (*svi*=l) has a constant effect on the level of *lpsa* regardless of the level of any of the other predictors. How would you test this hypothesis against the alternative that the effect of *svi* on *lpsa* varies with *age*? State models for the null and alternative hypotheses, describe how to compute a test statistic and state the distribution of your test statistic under the null hypothesis.



```
lm(formula = lpsa ~ lcavol + lweight + age + lbph + svi + lcp +
    gleason, data = prostate)

Residuals:
      Min        1Q    Median        3Q       Max
-1.788027 -0.369331  0.003023  0.434360  1.621603

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.02416    1.13313   0.021  0.98304
lcavol       0.57471    0.08712   6.597 2.92e-09 ***
lweight      0.45260    0.17005   2.662  0.00923 **
age         -0.01812    0.01108  -1.636  0.10542
lbph         0.10886    0.05844   1.863  0.06579 .
svi          0.79783    0.24241   3.291  0.00143 **
lcp         -0.07488    0.08599  -0.871  0.38619
gleason      0.14591    0.12292   1.187  0.23837
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 0.7086 on 89 degrees of freedom
Multiple R-squared: 0.6506,     Adjusted R-squared: 0.6232
F-statistic: 23.68 on 7 and 89 DF,  p-value: < 2.2e-16
```

2. Dose response data were collected where the response of interest is the number of bacterial colonies on a plate and the doses are 0,10,33,100,333, and 1000. The data were in the form $\left(x_i, Y_{ij}\right)$, *i*=1, …, 6, *j*=1,…,3, where *Y* was the number of colonies on a plate and $x = \log\,(1 + dose)$. The goal was to determine the relationship of the transformed dose x on the response Y. One scientist argued that the data should be reduced to $(x_i, \bar{Y}_{i\cdot})$ where $\bar{Y}_{i\cdot} = \frac{1}{3}\Sigma_{j=1}^{3} Y_{ij}$, and a least squares line should be estimated. Another argued that the data should be treated as a one-way ANOVA using the six levels of x as treatments. Their results are shown in the next page.

   a) (7pts) Assume throughout that $Var\left(Y_{ij}\right) = \sigma^2$ for all *i,j*. If the model in the first analysis is true, does this analysis provide an unbiased estimate of σ^2?

   b) (6pts) Suppose that the linear regression model $Y_{ij} = \beta_0 + \beta_1 x_i + \epsilon_{ij}$ is correct and that the error terms are i.i.d. with zero means and constant variance. Does the regression analysis of $\bar{Y}_{i\cdot}$ vs x_i provide unbiased estimates of the regression coefficients?

   c) (6pts) Is there enough information in the printed output to test whether the linear model in (b) is inadequate to describe the data? If so, give a numerical expression for the test statistic and its distribution if the linear regression is correct. If not, identify the information needed.

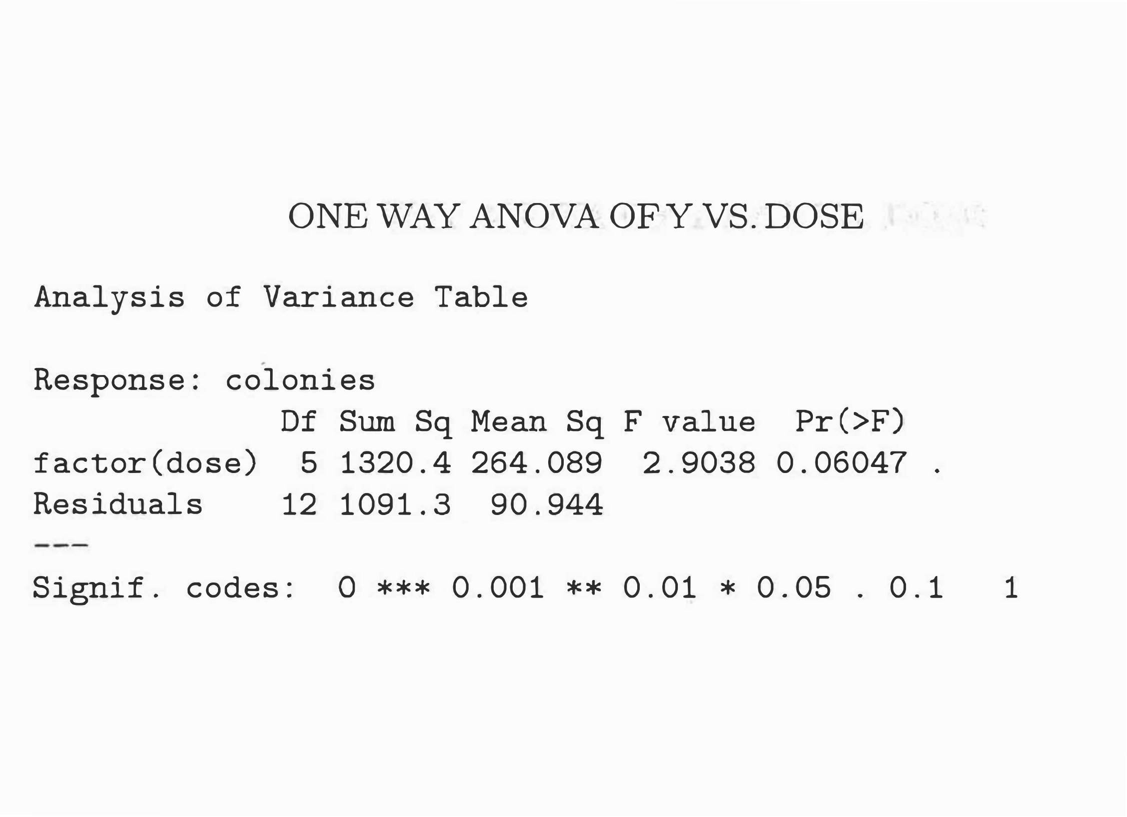
LINEAR REGRESSION OF Y-BAR VS. LOG(1 + DOSE)

```
Call:
lm(formula = colonies.avg ~ x)

Residuals:
     1      2      3      4      5      6 
 1.843 -7.235 -3.272 11.786  3.587 -6.709 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)  
(Intercept)   19.823      6.562   3.021   0.0391 *
x              2.396      1.461   1.640   0.1764  
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1 

Residual standard error: 8.112 on 4 degrees of freedom
Multiple R-squared:  0.4019,    Adjusted R-squared:  0.2524 
F-statistic: 2.688 on 1 and 4 DF,  p-value: 0.1764
```



ONE WAY ANOVA OF Y VS. DOSE

```
Analysis of Variance Table

Response: colonies
             Df Sum Sq Mean Sq F value  Pr(>F)  
factor(dose)  5 1320.4 264.089  2.9038 0.06047 .
Residuals    12 1091.3  90.944                  
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1 
```

3. A sample of 12 observations of $(x_1, x_{2,}Y)$ is displayed in the figure, with points (x_2, Y) labeled by their $x_1$values. The least squares regression line $y = \hat{\beta}_0 + \hat{\beta}_2 x_2$ is also displayed in the figure.

   (a) (7pts) Suppose that x_1 is a quantitative variable. How would you decide if the simple linear regression model
   $$Y_i = \beta_0 + \beta_2 x_{2i} + \epsilon_i, i = 1, \ldots ,12$$
   describes the data adequately? Assume that the error terms are i.i.d. $N(0, \sigma^2)$. Explain how to compute any test statistics and give their distributions.

   (b) (7pts) Suppose instead that the variable x_1 is a qualitative variable labeling groups, which were chosen at random from some population. Propose a random effects model, which is a generalization of the simple linear regression model, and describe how to test whether the random effects model fits the data better than the simple linear regression model of (a).

   (c) (7pts) Based on your examination of the graph, under the assumptions of either (a) or (b), would the regression coefficients of x_2 be positive if $x_1$were included in the model? Why?

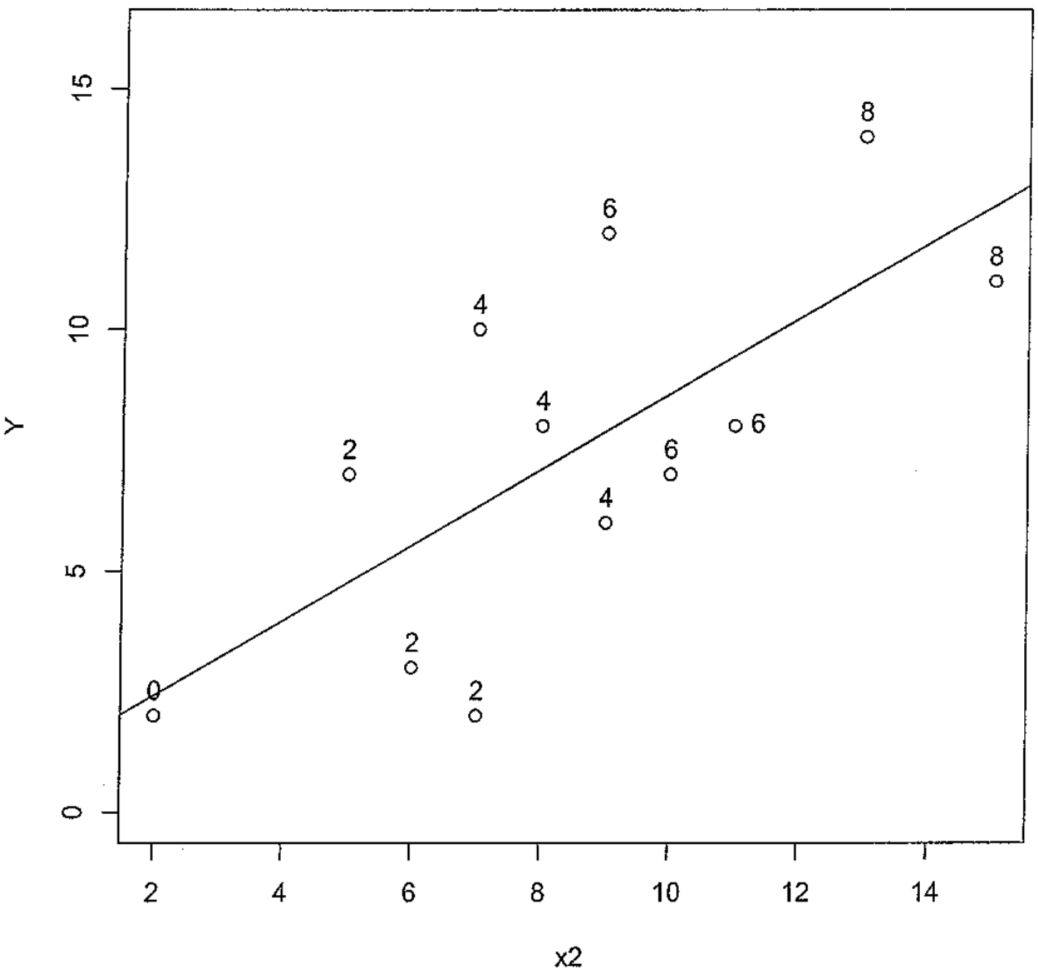


Figure 1: Plot of Y vs. x_2 with points labeled by x_1 values.

4. A recent article, *Predictors of Student Productivity in Biomedical Graduate School Applications*, reported on whether a graduate student's GRE score was predictive of the number of first-author articles that student would publish. The subjects were the three cohorts of students who entered the Biological and Biomedical Sciences Program at the University of North Carolina from 2008 to 2010. The predictor variable was the student's GRE score. The response was the number of first-authored publications by the student through July 12, 2016. The article found that GRE score was not strongly predictive of publication number.

   The article did not account for matriculation year. But since some of the students matriculated only in 2010, they may not have had time for many publications to appear by July 12, 2016. Those who matriculated in 2010 might be expected to have fewer publications than those who matriculated earlier. Suppose that's true, and also that the expected number of publications is a linear function of GRE score and matriculation year and that for any combination of GRE score and matriculation year, the publication numbers are Normally distributed. We want to estimate the predictive value of GRE after accounting for matriculation year.

   (a) (6pts) Explain why modeling publication number as Normal is questionable. Give a more plausible model. Despite its questionability, we assume Normality for the rest of this problem.

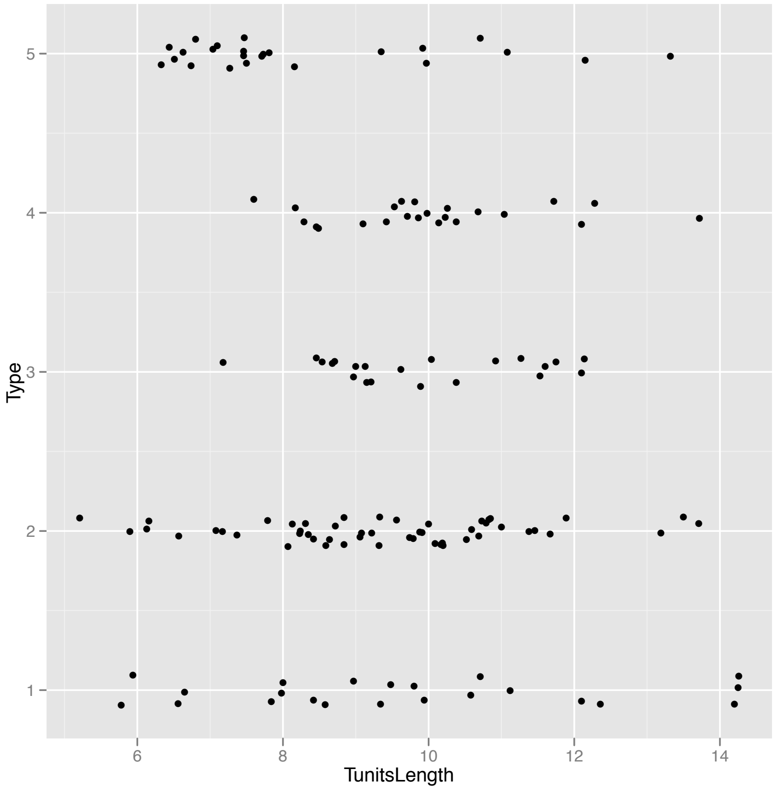
   (b) (6pts) Write a linear model that describes the expected publication number as a linear function of GRE and matriculation year. What parameter in your model are the investigators most interested in?

   (c) (6pts) Added variable plots, also known as partial regression plots, are one way of displaying the relationship between a response and a predictor after accounting for another predictor. Describe an added variable plot for the GRE/publication problem. Say what would go on each axis of the plot. What, in the added variable plot, would correspond to the parameter of interest in your linear model?

   (d) (6pts) Statisticians sometimes calculate a t or F statistic as part of the inference for the parameter of interest. Choose either t or F and say where the estimated standard deviation $\hat{\sigma}$ of the residuals comes into its calculation.

   (e) (6pts) How would $\hat{\sigma}$ likely change if year is omitted from the model? How would that likely change the t or F statistic?

5. A linguist taught introductory Russian to a large class of students at UMass. Each student wrote an essay and the linguist measured something called *T-units* on each essay. The students had various previous exposure to Russian—some had no exposure; others had parents who spoke Russian; still others had other amounts of exposure. There were five exposure groups all together and the linguist wanted to know whether previous exposure was related to T-units. The figure shows the data. T-units is on the x-axis; group is on the y-axis. The y-value has been jittered slightly for clarity.



To compare the groups, the linguist conducted an ANOVA. The ANOVA found strong differences only between groups 5 and 3 and groups 5 and 4.

(a) (6pts) What aspect of the T-units distribution does the ANOVA compare from group to group?

(b) (6pts) Judging from the graph, what differences between groups in terms of the T-units distributions did the ANOVA miss?