Stat 306 Midterm exam March 1, 2016

Instructions: Please write your name and student number on the front page of the exam booklet. A 2-sided sheet of notes and a non-programmable calculator may be used. The 4 problems may be done in any order. A partial t table is given below.

ν	5	10	20	30	50	100	195	200	1000
$t_{\nu,0.975}$	2.571	2.228	2.086	2.042	2.009	1.984	1.972	1.972	1.962

(14) 1. SAT is a scholastic aptitude test in the US for high school students applying to universities; there are Math and Verbal components. let y=Math SAT, x=Verbal SAT. For a sample of size $n = 1000$ in the year 2005, summary statistics are $\overline{x} = 597.2$, $s_x = 98.6$, $\overline{y} = 614.3$, $s_y = 96.8$, $r_{xy} = 0.685$. Suppose the scatterplot is oval-shaped suggesting homoscedaticity.

(a) What are the intercept and slope of the least squares line?

(b) What is the residual standard deviation $\hat{\sigma}$?

(c) Consider the subpopulation in 2005 with a verbal SAT score of (around) 500. What is a point estimate of the mean math SAT score for this subpopulation and what is an approximate 95% confidence interval? (If you did not solve (b), assume that $\hat{\sigma} = 75$ for computing the confidence interval.)

(8) 2. Intepretations.

(a) In question 1, would the equation be valid to make predictions in the year 2015. Explain your reasoning briefly in one sentence.

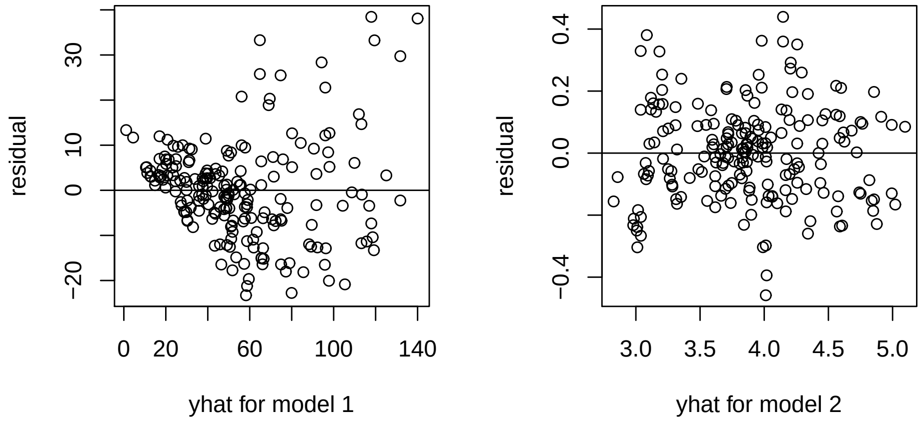
(b) Consider the least squares equation in question 1. Does this equation imply that if one spends extra effort in studying for the verbal SAT, then one should improve his or her score for the math SAT? Explain your reasoning briefly in one sentence.

(10) 3. Consider data (x_i, y_i), $i = 1, \ldots, n$, with $n > 2$. A least squares line is fit with intercept $\hat{\beta}_0$ and slope $\hat{\beta}_1$. Let $e_i = y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i)$ be the ith residual.

1. If $e_1 > 0$, is (x_1, y_1) above or below the least squares line?
2. Is e_1 a linear or non-linear function of $y_1, \ldots, y_n$. Explain your reasoning in one sentence.
3. Let $S(\beta_0, \beta_1) = \sum_{i=1}^n (y_i - \beta_0 - \beta_1 x_i)^2$. Suppose $(\hat{\beta}_0, \hat{\beta}_1)$ is the solution to the system of two equations: $\frac{\partial S}{\partial \beta_0} = 0$ and $\frac{\partial S}{\partial \beta_1} = 0$. Obtain the partial derivatives, and based on the definition of $(\hat{\beta}_0, \hat{\beta}_1)$ in the preceding sentence, then show that $\bar{e} = n^{-1} \sum_{i=1}^n e_i = 0$.

(18) 4. Consider a data set of prices of $n = 202$ new cars in 2015, where several manufacturers were selected and the car type is compact, coupe, sedan, or wagon. The price is MSRP (manufacturer suggested retail price). The aim is to find a prediction equation of MSRP (in thousands of dollars) based on explanatory variables: Displacement (Disp), Horsepower (Hp), Tanksize (Tank), and Brand Category (4 categories labelled as G1=economical, G2=upscale, G3=luxury, Porsche). Two models, respectively with MSRP and ln(MSRP) as response variables, are fitted. Some outputs and residual plots from R are shown below.

```
model1=lm(MSRP~Brand+Disp+Hp+Tank, data=cardat)
model2=lm(lnMSRP~Brand+Disp+Hp+Tank, data=cardat)
Variable    |  Model1 edited output               |   Model2 edited output
            |  Estimate  SE     t value Pr(>|t|)  |   Estimate  SE   t value Pr(>|t|)
(Intercept)| -16.858    6.762  -2.493  0.0135    |   2.6301  0.0943  27.890  < 2e-16
BrandG1     | -24.870    4.484  -5.546 9.4e-08    |  -0.5404  0.0625  -8.642 1.99e-15
BrandG2     | -26.344    4.138  -6.366 1.4-09     |  -0.2700  0.0577  -4.679 5.38e-06
BrandG3     | -13.133    3.945  -3.329  0.0010    |  -0.0892  0.0550  -1.621    0.107
Disp        |   1.006    1.108   0.908  0.3648    |   0.0362  0.0154   2.346    0.020
Hp          |   0.165    0.013  12.619  < 2e-16   |   0.0021  0.0002  11.390  < 2e-16
Tank        |   0.580    0.077   7.553 1.61e-12   |   0.0111  0.0011  10.360  < 2e-16
            |  Residual SD: 11.08 on 195 dof      |  Residual SD: 0.1545 on 195 dof
            |  Multiple R^2:  0.8810              |  Multiple R^2:  0.9222
            |  Adjusted R^2:  0.8773              |  Adjusted R^2:  0.9198
```



(a) What is a noticeable pattern in one of the residual plots.

(b) Based on the above, which of the two models is better? Give at least 2 reasons for your choice.

(c) What is an approximate 95% confidence interval for the coefficient of BrandG1 in model 2?

(d) Interpret $\hat{\beta}_{\mathrm{BrandG1}}$ and the interval in (c).

(e) Two lines of the data files are:

```
MSRP    Brand   Disp  Hp Tank
54.590  G3       3.5 300 65.0
37.900  G2       2.5 204 66.0
```

What are in their corresponding rows of the **X** matrix (which has 7 columns), for model y~Brand+Disp+Hp+Tank.