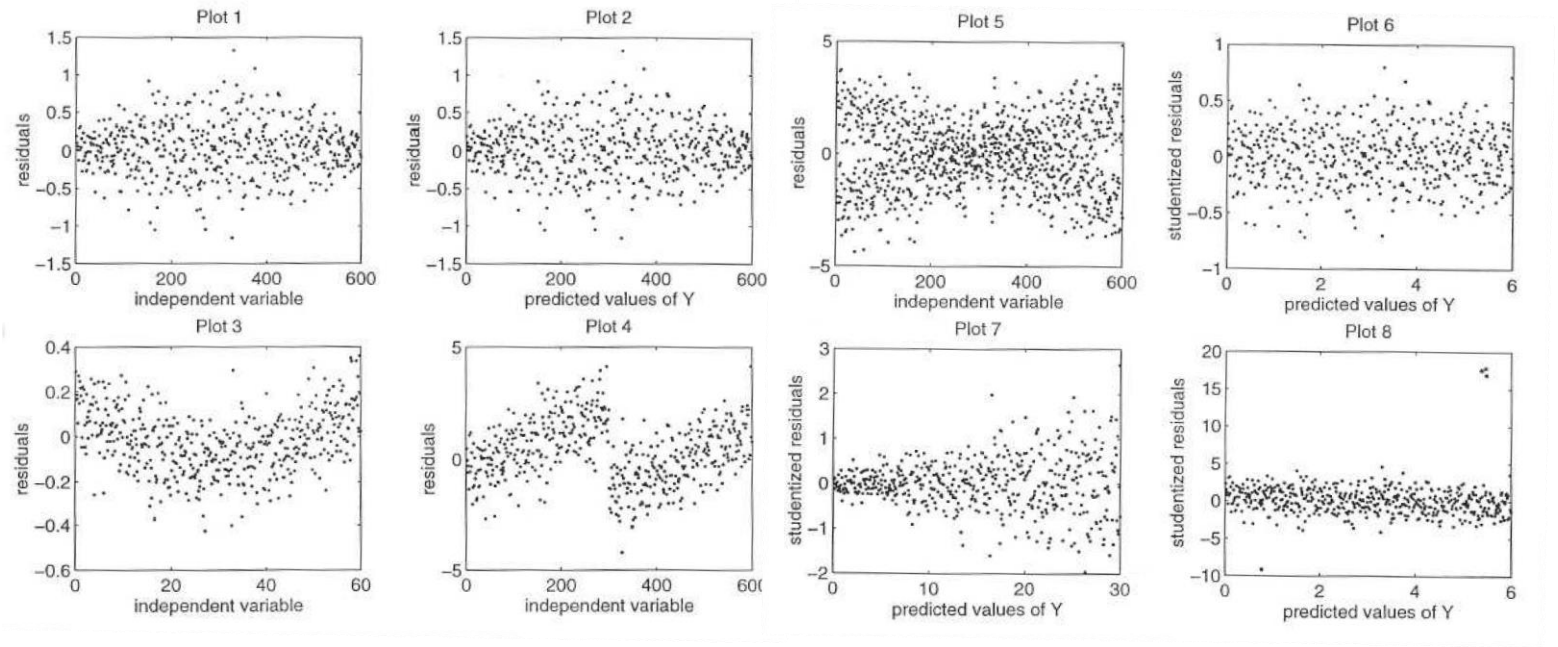
STAT 615/415

Regression **FINAL EXAM** **Name:**

1. (22 points) (By hand)
   The following R code was written to study relation between variables y, x_1, and x_2.
   Unfortunately, coffee was spilled on the output, and some parts of it became unreadable. Restore the missing parts in the 11 empty boxes.

```
> attach(DATASET)
> summary(DATASET)
     x1                x2                    y
 No    :10   Min.      :0.04564    Min.      :0.9589
 Yes   :10   1st Qu.   :0.45671    1st Qu.   :1.5509
             Median    :0.64683    Median    :1.8745
             Mean      :0.65205    Mean      :1.9655
             3rd Qu.   :0.82315    3rd Qu.   :2.3329
             Max.      :1.12577    Max.      :3.3721

> reg = lm( y ~ x1 + x2, data=DATASET )
> anova(reg)
Analysis of Variance Table

Response:     y
             Df  Sum Sq      Mean Sq      F value      Pr(>F)
 x1           1  2.0937      .......      .......      .......     ***
 x2          1     4.9187    [        ]   [        ]   [        ]  ***
 Residuals   17    0.0939    [        ]

> summary(reg)

Coefficients:
              Estimate    Std.   Error      t value      Pr(>|t|)
 (Intercept)    3.40071        0.04405     .......      .......      ***
 x1Yes         -0.32623        0.03494     [        ]   [        ]   ***
 x2            -1.95094        0.06538     .......      .......      ***

Residual standard error:  [        ]  on 17 degrees of freedom
F-statistic:  [        ]  on  [        ]  and  [        ]  DF, p-value:      1.11e-16

> predict(reg, data.frame(x1="No", x2=5))
[        ]
```

2. (16 points) Multiple linear regression models are studied for eight different sets of data. Below are various kinds of residual plots. For each case, give a brief diagnostics of the regression model and suggest a way to improve the model, if necessary.



3. (32 points) (Use R for data analysis)

Based on our known *"mtcars"* dataset, we'll now build a model for predicting variable *qsec*, the time in seconds that it takes a car to drive 1/4 mile from a full stop.

(a) Select the **optimal** subset of independent variables for the prediction of the *qsec* using your favorite variable selection method. Show all methods used and justify your conclusion.

(b) Is there any indication of nonlinear relations? Explain how you reach this conclusion.

(c) Test significance of interaction between the weight (*wt*) and transmission (*am*, 0 = automatic, 1 = manual). (Hint: Use the "best" model according to part (a)).

(d) **(only for 615 students)** Construct a 90% prediction interval for *qsec* of a car that has 6 cylinders (*cyl*) and 150 horsepower (*hp*).

4. (10 points) (By hand)

A student fitted a linear regression function for a class assignment. The student plotted the residuals e_i against responses Y_i and found positive relation. When the residuals were plotted against the fitted values $\hat{Y}_i$, the student found no relation.

(a) How could the differences arise? Which is the more meaningful plot?

(b) Support your answer by deriving the sample covariance Cov(e_i, Y_i) and Cov(e_i , $\hat{Y}_i$). Feel free to use any Formulae and any results that we derived in class.

5. (25 points) (Use R for data analysis)

An experiment was conducted to evaluate the effect of vitamin C on tooth growth. Sixty guinea pigs received various doses of vitamin C by one of two delivery methods, orange juice or ascorbic acid. Results of this experiment are in dataset ‘‘ToothGrowth’’ which is already loaded in R. You can look at it with commands ***attach(ToothGrowth), names(ToothGrowth), summary(ToothGrowth), ToothGrowth***.
The data set contains the following variables:

len: Tooth length
supp: supplement or delivery method (OJ = orange juice, VC = ascorbic acid)
dose: Dose in milligrams/day

(a) Fit a linear regression model that can be used to predict the tooth length based on the dose and the delivery method of vitamin C.
(b) Is delivery method significant?
(c) Is there a significant interaction between the dose and the delivery method, at α = 5% level?
(d) Write two regression equations explicitly, one equation for each delivery method.
(e) What percent of the total variation of the tooth length is explained by this regression?