STAT 408 Final

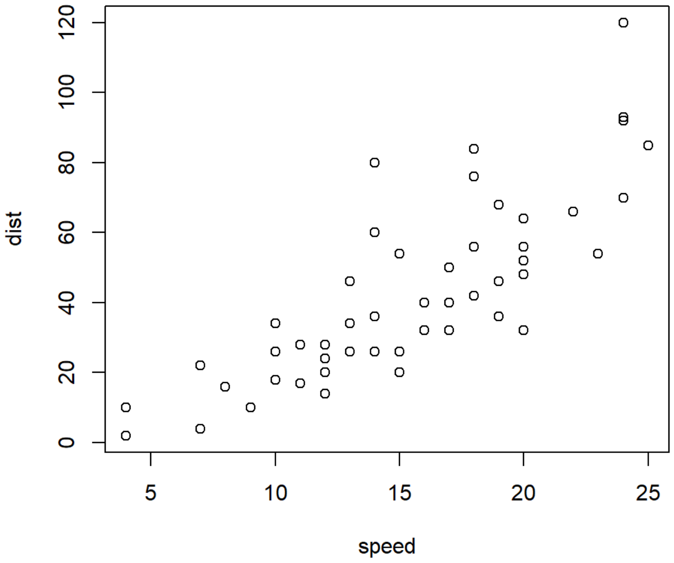
55 points

1. (5 points) What is collinearity in the linear model? What are the potential impacts of collinearity on the estimation and inference of linear models?

2. The “cars” dataset records the speed and stopping distances for 50 cars.

speed: speed (mph)
dist: stopping distance (ft)

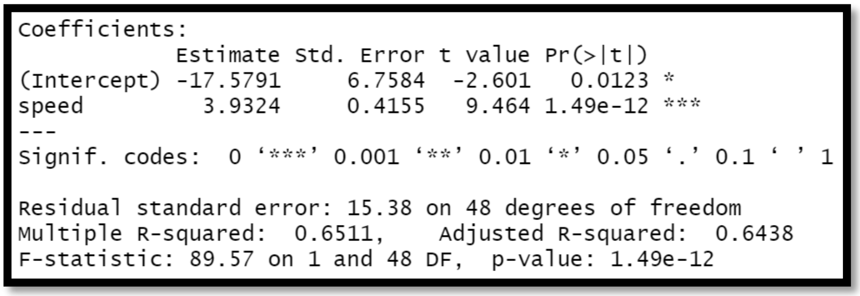
a. (5 points) Below is the plot of distance vs. speed. Do you think the relationship between stopping distance and speed is linear? What type of model do you want to build?



Suppose we fit four different models to model distance vs. speed:

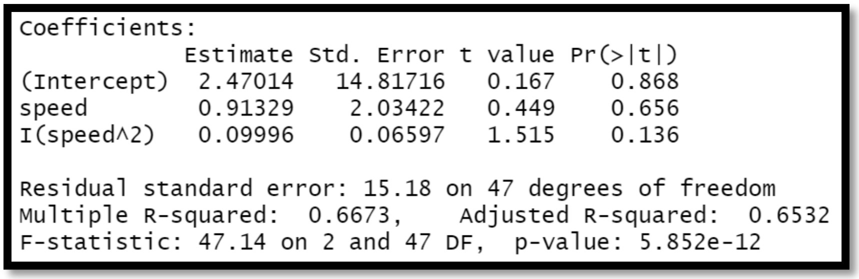
i) summary(lm(dist~speed, data = cars))
ii) summary(lm(dist~speed+I(speed^2), data = cars))
iii) summary(lm(dist~I(speed^2), data = cars))
iv) summary(lm(sqrt(dist)~speed, data = cars))

The model fitting results are shown by following the same order:



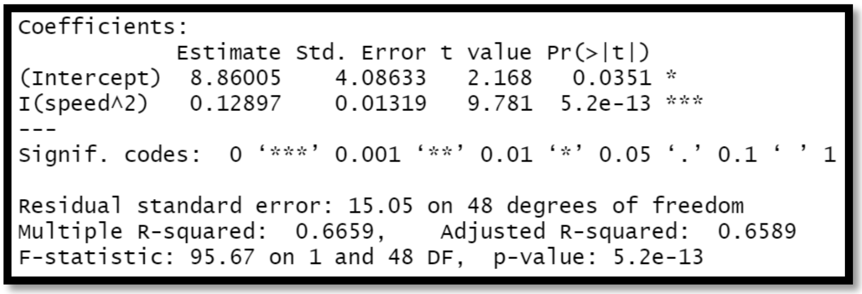
```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -17.5791     6.7584  -2.601   0.0123 *
speed         3.9324     0.4155   9.464 1.49e-12 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 15.38 on 48 degrees of freedom
Multiple R-squared:  0.6511,    Adjusted R-squared:  0.6438
F-statistic: 89.57 on 1 and 48 DF,  p-value: 1.49e-12
```



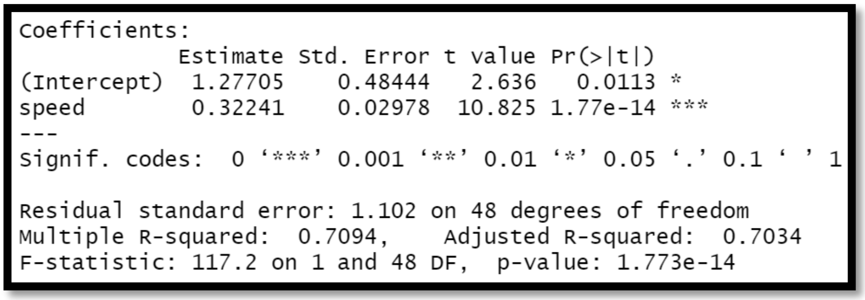
```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.47014   14.81716   0.167    0.868
speed        0.91329    2.03422   0.449    0.656
I(speed^2)   0.09996    0.06597   1.515    0.136

Residual standard error: 15.18 on 47 degrees of freedom
Multiple R-squared:  0.6673,    Adjusted R-squared:  0.6532
F-statistic: 47.14 on 2 and 47 DF,  p-value: 5.852e-12
```



```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  8.86005    4.08633   2.168   0.0351 *
I(speed^2)   0.12897    0.01319   9.781  5.2e-13 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 15.05 on 48 degrees of freedom
Multiple R-squared:  0.6659,    Adjusted R-squared:  0.6589
F-statistic: 95.67 on 1 and 48 DF,  p-value: 5.2e-13
```



```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.27705    0.48444   2.636   0.0113 *
speed        0.32241    0.02978  10.825 1.77e-14 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.102 on 48 degrees of freedom
Multiple R-squared:  0.7094,    Adjusted R-squared:  0.7034
F-statistic: 117.2 on 1 and 48 DF,  p-value: 1.773e-14
```

b. (5 points) Use mathematical notations to write down the four models. What are the underlining assumptions in terms of the specific form of each model? (Hint: if they are linear, quadratic, or others; the constrain on response and predictors…)

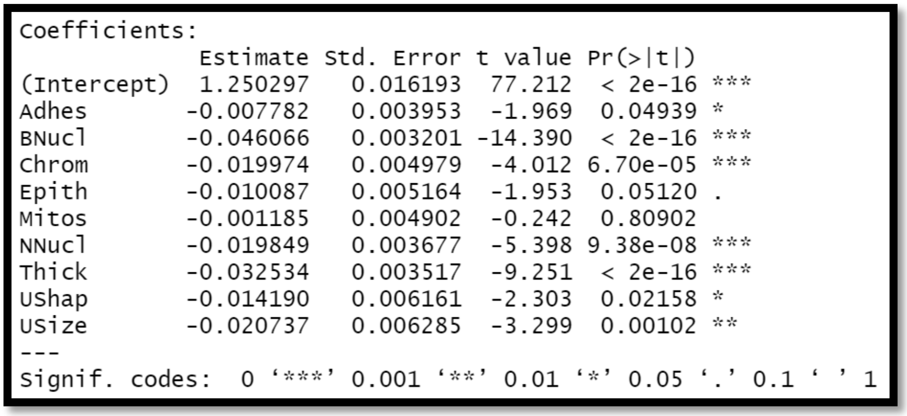
c. (5 points) Based on the model fitting results, which model you prefer? Give your reason. Under your best model, how can we interpret the model parameters and the relationship between response and predictors?

3. Dataset wbca comes from a study of breast cancer in Wisconsin. There are 681 cases of potentially cancerous tumors, of which 238 are malignant. Determining whether a tumor is malignant is traditionally determined by an invasive surgical procedure. The purpose of this study was to determine whether a new procedure called fine needle aspiration which draws only a small sample of tissue could be effective in determining tumor status.

Class: 0 if malignant, 1 if benign
Adhes: marginal adhesion
BNucl: bare nuclei
Chrom: bland chromatin
Epith: epithelial cell size
Mitos: mitoses
NNucl: normal nucleoli
Thick: clump thickness
UShap: cell shape uniformity
USize: cell size uniformity

The predictor values are determined by a doctor observing the cells and rating them on a scale from 1 (normal) to 10 (most abnormal) with respect to the particular characteristic. We fit a logistic regression to predict the malignant tumor:

```
summary(lmod <- glm(Class~., data = wbca))
```

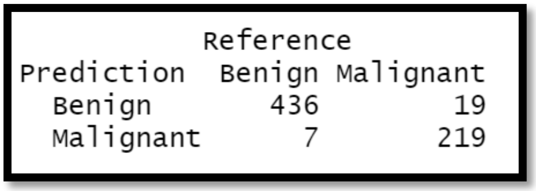


```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept)  1.250297   0.016193  77.212  < 2e-16 ***
Adhes       -0.007782   0.003953  -1.969  0.04939 *  
BNucl       -0.046066   0.003201 -14.390  < 2e-16 ***
Chrom       -0.019974   0.004979  -4.012 6.70e-05 ***
Epith       -0.010087   0.005164  -1.953  0.05120 .  
Mitos       -0.001185   0.004902  -0.242  0.80902    
NNucl       -0.019849   0.003677  -5.398 9.38e-08 ***
Thick       -0.032534   0.003517  -9.251  < 2e-16 ***
UShap       -0.014190   0.006161  -2.303  0.02158 *  
USize       -0.020737   0.006285  -3.299  0.00102 ** 
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

a. (5 points) If we want to perform model selection using backward selection, what operations do you need to do in the first step?

b. (5 points) Write down the logistic model fitted by the code, but only include the significant predictors (at 5% level). Interpret the meaning of BNucl's parameter in this model.

c. (5 points) Suppose that a cancer is classified as benign if p>0.5 and malignant otherwise. Based on this prediction rule and our model output, we have the following confusion matrix:



```
           Reference
Prediction  Benign Malignant
  Benign       436        19
  Malignant      7       219
```

Calculate the overall accuracy, true positive rate, true negative rate, and precision. Treat benign as positive and malignant as negative.

d. (5 points) Do you think the four measurements calculated in (c) is unbiased for estimating the model prediction performance? Give your reason. Show other potentially unbiased solutions for evaluating model prediction performance.

4. Suppose that we have a dataset with a categorical response variable having 4 classes. The 4 classes are labeled as Y=0, Y=1, Y=2, and Y=3. We use a multinomial logistic regression to fit the data.

a. (5 points) In the model, what random variable do we use to model the 4-class response? Write down the probability distribution of this random variable.

b. (5 points) Suppose that there are p predictors $X_1, X_2, \ldots, X_p$ in the data. How do we model the probabilities of 4 classes in terms of all predictors? Write down the mathematical form.

c. (5 points) With the probabilities of 4 classes, how do we obtain the final class label for each observation? Write down the mathematical form.