PSTAT126 Final Examination

Instructions

It is an open book exam. You can use any reference materials including notes and books. All work must be your own. You should not show or discuss your exam with anyone before submission. Evidence of collaboration with other students will be severely penalized. Anyone caught cheating will receive an automatic F on the exam. In addition the incident will be reported, and dealt with according to University's Academic Dishonesty regulations. If you have any questions, please visit the zoom link for office hours or send me an email.

1. [10] Give brief discussion of the multiple linear regression model. Write down the definition of this model with all assumptions, illustrate possible applications in practice, specify a R function for fitting this model.

2. [10] Write a brief explanation of the Box-Cox transformation including the purpose of the transformation, the formula, and how to find this transformation in R.

3. [10] A student stated: "Adding independent variables to a regression model can never reduce R^2, so we should include all available independent variables in the model." Do you agree and why? Can you suggest one objective way to select variables to be include in a model.

4. [25 total, 5 points each] The model $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \epsilon$ is fitted to a data set with 20 observations.

   (a) Supply values for the missing entries in the following ANOVA table.

Table 1: ANOVA Table

Source	SS	d.f.	MS	F
Regression				10
Error			5	
Total				

   (b) Is the overall regression significant?

   (c) What proportion of the Y-variability is explained by the regression model?

   (d) Another model $Y = \beta_0 + \beta_1 X_1 + \epsilon$ is fitted to the same data set. Supply values for the missing entries in the following ANOVA table.

Table 2: ANOVA Table

Source	SS	d.f.	MS	F
Regression				
Error	120			
Total				

   (e) Use extra sum of squares principle to test if variables X_2 and X_3 are important. State your hypothesis.

5. [45 total, (c) and (e) have 10 points and others have 5 points] To investigate factors that may affect gasoline consumption, 20 cars were randomly selected. The following variables were recorded: `VOL`: Cubic feet of cab space, `HP`: Engine horsepower, `MPG`: miles per gallon, `SP`: Top speed (mph), `WT`: Vehicle weight (100 lb). Data and a simple analysis using R is shown below.

```
> mpg <- read.table("mpg.dat", header=T)
> mpg
    VOL HP MPG  SP  WT
1   92  55 56.0  97 20.0
2   92  70 49.0 105 20.0
3   92  53 46.5  96 20.0
4   50  62 59.2  98 22.5
5   89  73 41.1 103 22.5
6   99  92 40.9 113 22.5
7   89  73 40.4 103 22.5
8   91  78 38.9 106 22.5
9  103  90 42.2 109 25.0
10 106  95 32.2 106 30.0
11  92 102 32.2 109 30.0
12 102  93 31.5 105 30.0
13  99 100 31.5 108 30.0
14 111 100 31.4 108 30.0
15 103  98 31.4 107 30.0
16  86 130 31.2 120 30.0
17 101 115 33.7 109 35.0
18 101 115 32.6 109 35.0
19 101 115 31.3 109 35.0
20 124 115 31.3 109 35.0

> pairs(a)
> fit <- lm(MPG~VOL+HP+SP+WT, data=mpg)
> summary(fit)
Call:
lm(formula = MPG ~ VOL + HP + SP + WT)

Residuals:
   Min     1Q Median     3Q    Max
-4.781 -3.347 -1.405  3.632  6.205

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 213.86212  168.48781   1.269   0.2237
VOL          -0.18126    0.08842  -2.050   0.0583 .
HP            0.42248    0.83426   0.506   0.6199
SP           -1.39853    1.81344  -0.771   0.4526
WT           -1.74851    2.00709  -0.871   0.3974
```

```
---
Signif. codes:  0 .***. 0.001 .**. 0.01 .*. 0.05 ... 0.1 . . 1

Residual standard error: 4.29 on 15 degrees of freedom
Multiple R-squared: 0.807,      Adjusted R-squared: 0.7555
F-statistic: 15.68 on 4 and 15 DF,  p-value: 3.09e-05

> (fit1 <- step(fit, direction=''backward''))
Start:  AIC=62.5
MPG ~ VOL + HP + SP + WT

       Df Sum of Sq    RSS    AIC
- HP    1      4.72 280.80  60.84
- SP    1     10.95 287.02  61.28
- WT    1     13.97 290.05  61.49
<none>              276.08  62.50
- VOL   1     77.34 353.42  65.44

Step:  AIC=60.84
MPG ~ VOL + SP + WT

       Df Sum of Sq    RSS    AIC
<none>              280.80  60.84
- SP    1     90.85 371.65  64.44
- VOL   1    102.51 383.30  65.06
- WT    1    174.85 455.65  68.52

Call:
lm(formula = MPG ~ VOL + SP + WT, data = mpg)

Coefficients:
(Intercept)          VOL           SP           WT
   129.1801      -0.1964      -0.4869      -0.7394

> summary(fit1)

Call:
lm(formula = MPG ~ VOL + SP + WT)

Residuals:
    Min      1Q  Median      3Q     Max
-4.5135 -3.5167 -0.8465  3.4304  6.9058

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 129.18006   20.15791   6.408 8.66e-06 ***
```

```
VOL          -0.19640    0.08126  -2.417  0.02797 *
SP           -0.48691    0.21401  -2.275  0.03700 *
WT           -0.73938    0.23424  -3.156  0.00611 **
---
Signif. codes:  0 .***. 0.001 .**. 0.01 .*. 0.05 ... 0.1 . . 1

Residual standard error: 4.189 on 16 degrees of freedom
Multiple R-squared: 0.8037,     Adjusted R-squared: 0.7669
F-statistic: 21.84 on 3 and 16 DF,  p-value: 6.686e-06

> plot(fitted(fit1), residuals(fit1), xlab=''Fitted Response'',
     ylab=''Residuals'')
> abline(h=0)
> title(``Residuals vs fitted'')

> qqnorm(residuals(fit1), ylab=''Residuals'', main=")
> qqline(residuals(fit1))
> title(``QQ-plot of residuals'')
```

(a) Comment on the preliminary plots.

(b) For the saved R object `fit`, write down the model that is being fitted including assumptions.

(c) Explain the variable selection procedure performed by the `step` function including the direction, the criterion, the outcome of each step and the final model.

(d) From the output of `summary(fit)`, none of the variables is significant at 5% level. However, from the output of `summary(fit1)`, all remaining variables are significant at 5% level. Is this possible? Explain why or why not.

(e) What are your conclusions based on the final model `fit1`?

(f) Comment on the residual plots.

(g) Are you satisfied with the model `fit1`? If yes, explain why. If not, explain why and provide some suggestions that may improve the fit.

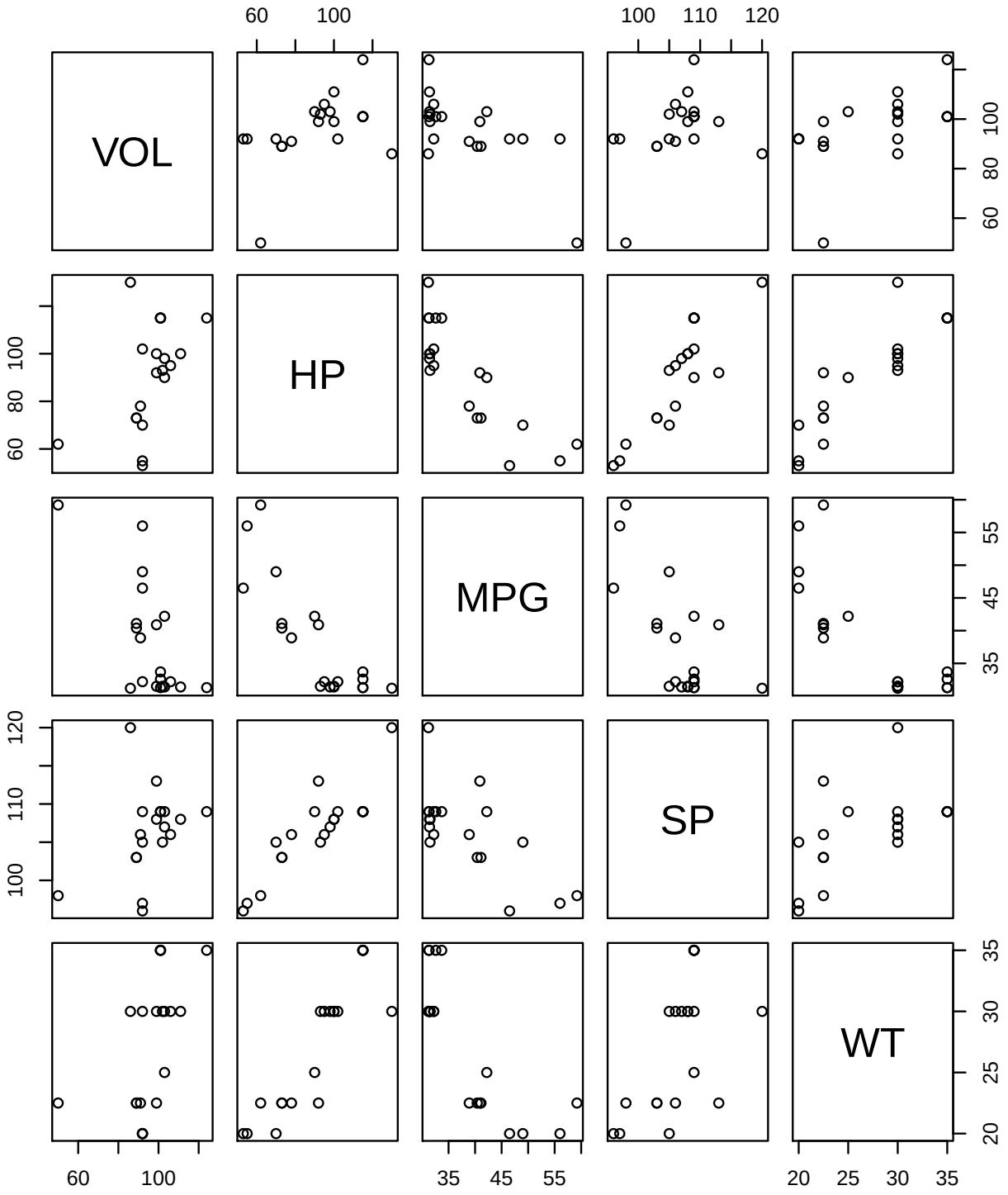

Figure 1: Scatter plots for MPG data.

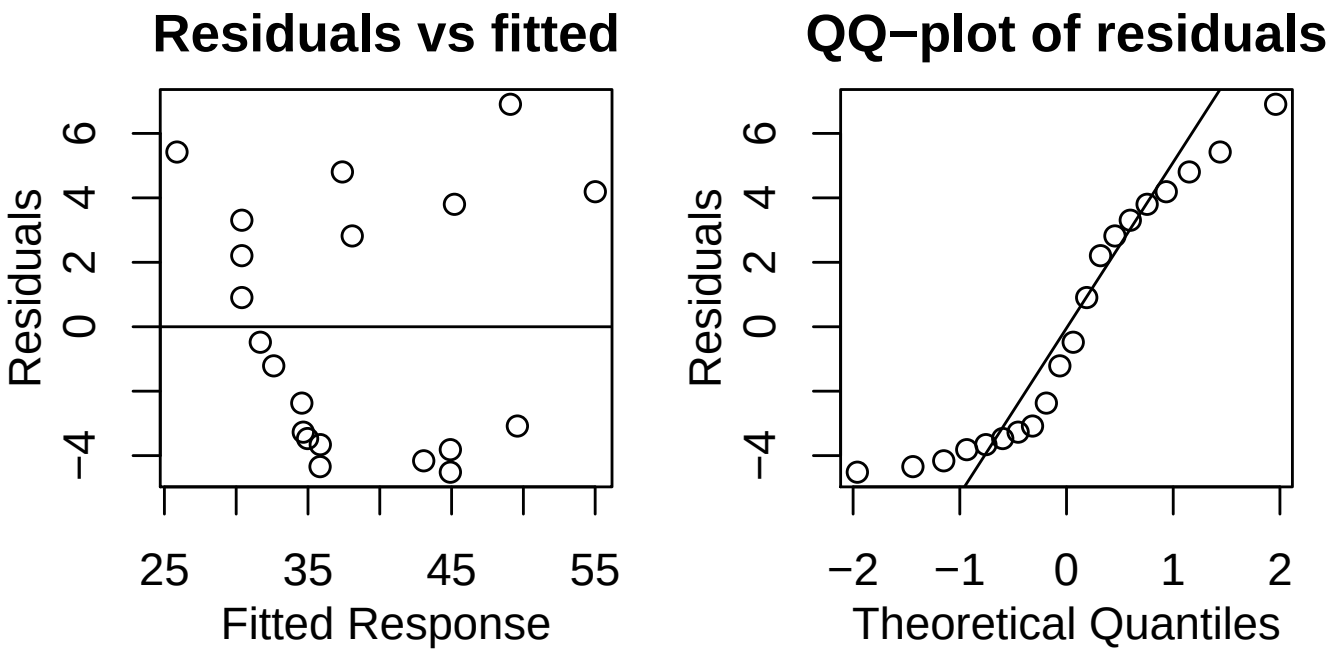

Figure 2: Diagnostic plots for MPG data.