

HW1

Tianqi Wu

2/1/2020

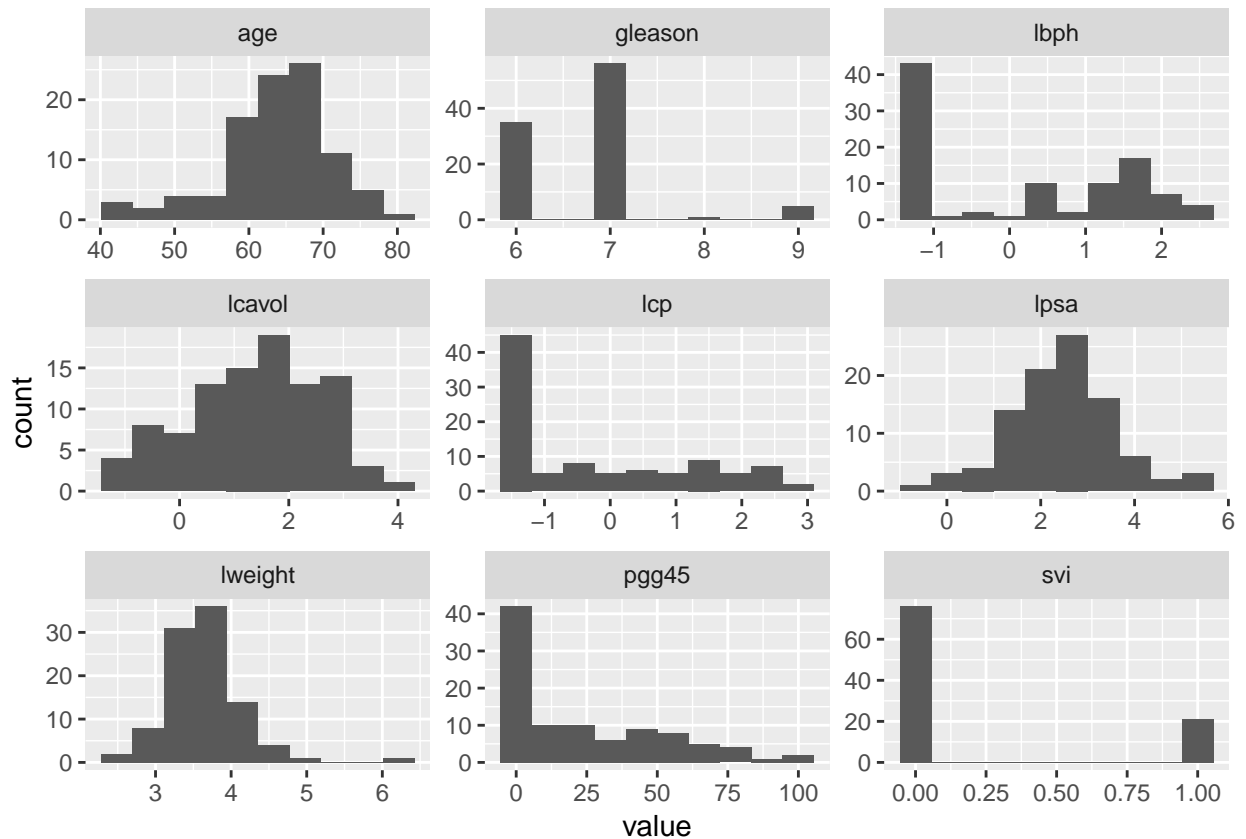
Problem 1

From the numerical and graphical summary, we can see that *lbph* (log(benign prostatic hyperplasia amount)), *lcp* (log(capsular penetration)) and *pgg45* (percentage Gleason scores 4 or 5) are quite skewed to the right. Most of men have small value for those three variables. *svi* is binary which takes value between 0 and 1. *gleason* is also categorical variable which takes value among 6,7,8 and 9.

```
summary(prostate,digits=3)
```

```
##      lcavol      lweight      age      lbph
##  Min.   :-1.347   Min.    :2.37   Min.    :41.0   Min.    :-1.39
##  1st Qu.: 0.513   1st Qu.:3.38   1st Qu.:60.0   1st Qu.: -1.39
##  Median : 1.447   Median :3.62   Median :65.0   Median : 0.30
##  Mean   : 1.350   Mean   :3.65   Mean   :63.9   Mean   : 0.10
##  3rd Qu.: 2.127   3rd Qu.:3.88   3rd Qu.:68.0   3rd Qu.: 1.56
##  Max.   : 3.821   Max.   :6.11   Max.   :79.0   Max.   : 2.33
##      svi      lcp      gleason      pgg45
##  Min.   :0.000   Min.   :-1.386   Min.    :6.00   Min.    : 0.0
##  1st Qu.:0.000   1st Qu.: -1.386   1st Qu.:6.00   1st Qu.: 0.0
##  Median :0.000   Median : -0.799   Median :7.00   Median :15.0
##  Mean   :0.216   Mean   : -0.179   Mean   :6.75   Mean   :24.4
##  3rd Qu.:0.000   3rd Qu.: 1.179   3rd Qu.:7.00   3rd Qu.:40.0
##  Max.   :1.000   Max.    :2.904   Max.    :9.00   Max.   :100.0
##      lpsa
##  Min.   :-0.431
##  1st Qu.: 1.732
##  Median : 2.592
##  Mean   : 2.478
##  3rd Qu.: 3.056
##  Max.   : 5.583
```

```
ggplot(gather(prostate), aes(value)) +
  geom_histogram(bins=10) +
  facet_wrap(~key, scales = 'free')
```



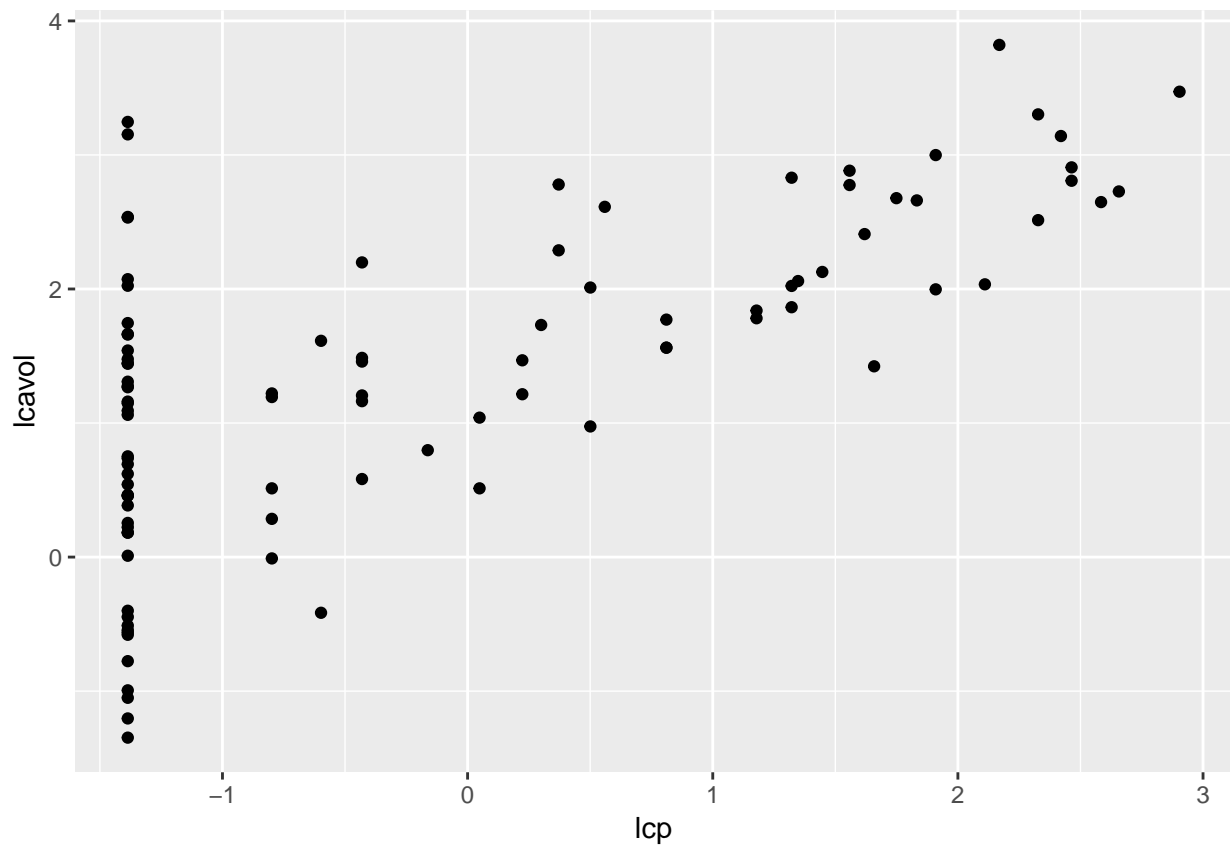
If we run a correlation analysis on the data. We could find that *gleason* (Gleason score) and *pgg45* (percentage Gleason scores 4 or 5) are strongly correlated with value of 0.752. It is not surprising since both of them are related to Gleason score.

```
round(cor(prostate),3)
```

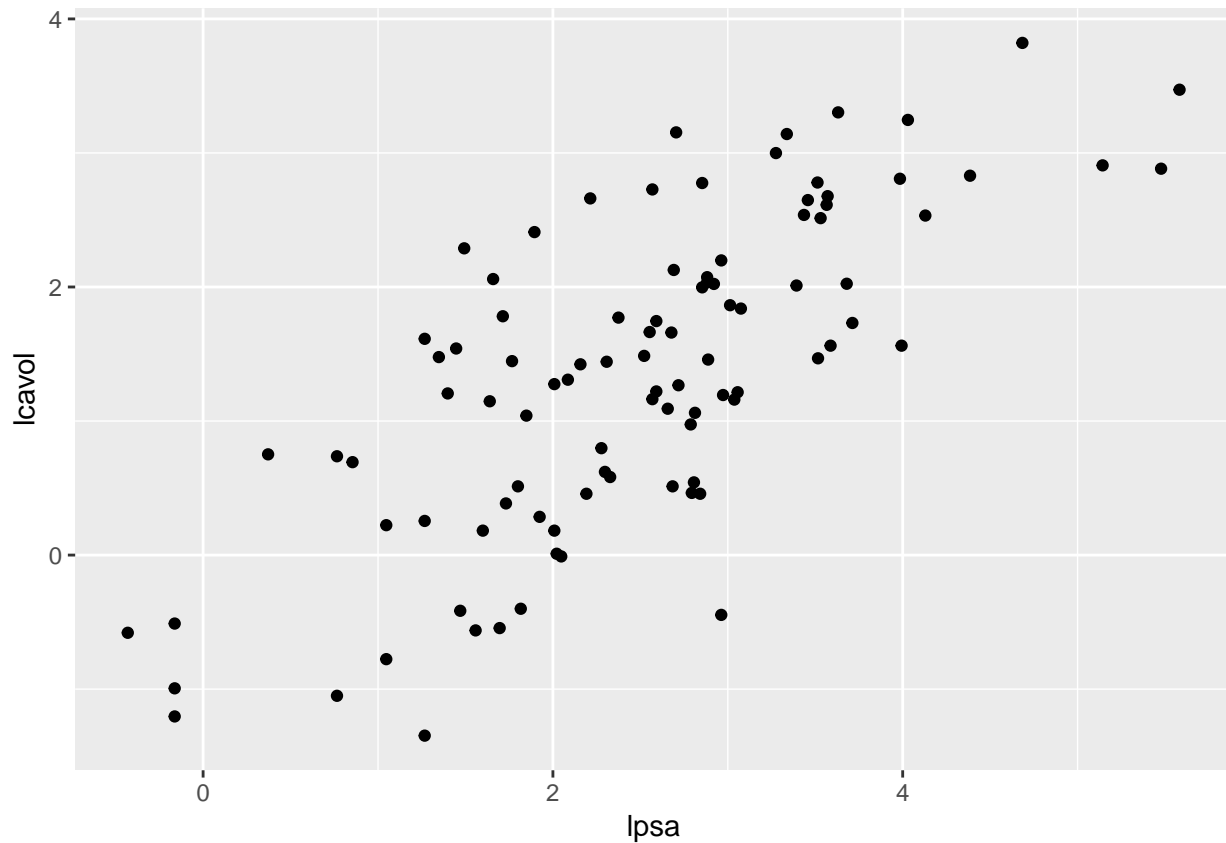
```
##      lcavol lweight  age  lbph   svi   lcp gleason pgg45  lpsa
## lcavol  1.000  0.194 0.225  0.027  0.539  0.675   0.432  0.434  0.734
## lweight 0.194  1.000 0.308  0.435  0.109  0.100  -0.001  0.051  0.354
## age     0.225  0.308 1.000  0.350  0.118  0.128   0.269  0.276  0.170
## lbph    0.027  0.435 0.350  1.000 -0.086 -0.007   0.078  0.078  0.180
## svi     0.539  0.109 0.118 -0.086  1.000  0.673   0.320  0.458  0.566
## lcp     0.675  0.100 0.128 -0.007  0.673  1.000   0.515  0.632  0.549
## gleason 0.432 -0.001 0.269  0.078  0.320  0.515   1.000  0.752  0.369
## pgg45   0.434  0.051 0.276  0.078  0.458  0.632   0.752  1.000  0.422
## lpsa    0.734  0.354 0.170  0.180  0.566  0.549   0.369  0.422  1.000
```

If we assume that *lcavol* (log(cancer volume)) is the dependent variable. We could find that *lcp* (log(capsular penetration)) and *lpsa* (log(prostate specific antigen)) are highly correlated to *lcavol*. Scatter plots of those two are shown below and it is clear that they both are positively correlated to *lcavol*.

```
ggplot(prostate,aes(x=lcp,y=lcavol))+geom_point()
```



```
ggplot(prostate,aes(x=lpsa,y=lcavol))+geom_point()
```

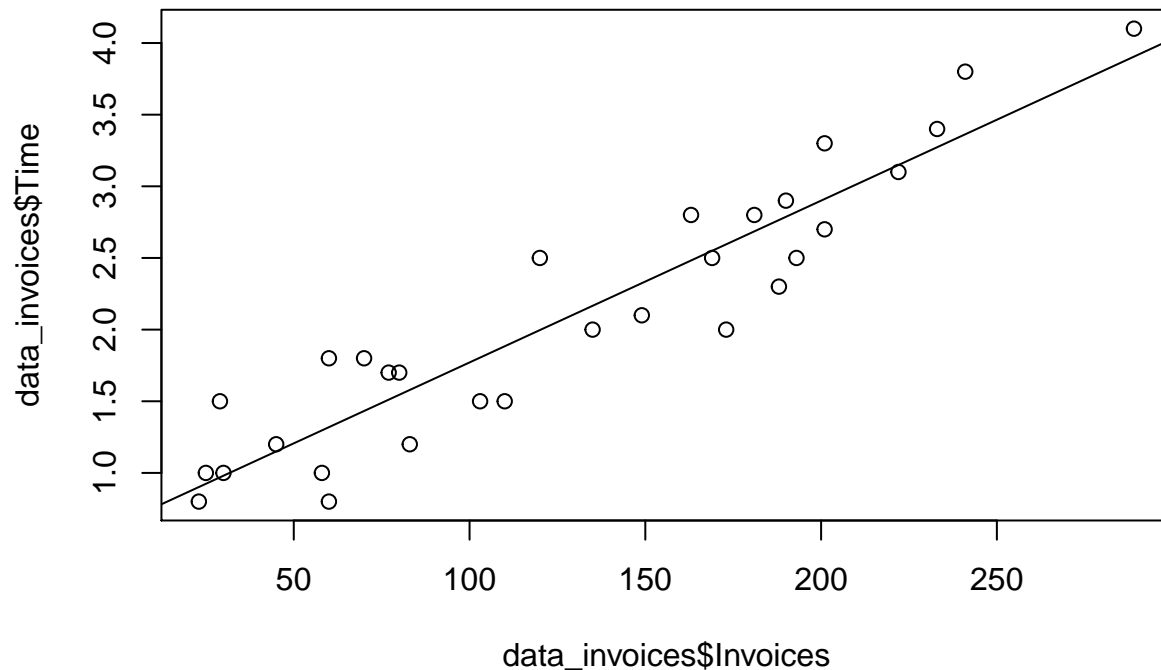


Problem 3

3(a)

Linear regression model fits the data well since most of the points are near the regression line. The processing time increases as number of invoices increases. There is no obvious outlier.

```
data_invoices = read.csv('invoices.txt',sep='\t')
plot(data_invoices$Invoices,data_invoices$Time)
lm_invoices = lm(Time~Invoices,data_invoices)
abline(lm_invoices)
```



3(b)

The 95% confidence interval for the start-up time is (0.391,0.892)

```
predict(lm_invoices,newdata=data.frame(Invoices=0),interval='confidence')
```

```
##          fit          lwr          upr
## 1 0.6417099 0.3912496 0.8921701
```

3(c)

Since $p\text{-value}=0.1257 > 0.05$, we do not have enough evidence to reject the null hypothesis that the average processing time for an additional invoice is 0.01 hours.

```
mycoef = summary(lm_invoices)$coefficients
2*pt((0.01-mycoef[2,1])/mycoef[2,2], 28)
```

```
## [1] 0.1257402
```

3(d)

The point estimate is 2.109 and 95% prediction interval is (1.986,2.232) for the time taken to process 130 invoices.

```
predict(lm_invoices,newdata=data.frame(Invoices=130),interval='confidence')
```

```
##          fit          lwr          upr
## 1 2.109624 1.986293 2.232954
```

Problem 5

5(a)

R-squared is 0.2792 which means that 27.92% of the variance in the data can be explained by the linear regression model. Adjusted R-squared is 0.2341 which means that 23.41% of the variance in the data can be explained by the linear regression model. R-squared always increases as number of predictors increases, adjusted R-squared penalizes additional predictors that are not helpful to explain the variance in data. Relatively low adjusted R-squared means that the linear regression model does not fit data well.

```
data_indicators = read.csv('indicators.txt', sep='\t')
lm_indicators = lm(PriceChange~LoanPaymentsOverdue, data=data_indicators)
summary(lm_indicators)

##
## Call:
## lm(formula = PriceChange ~ LoanPaymentsOverdue, data = data_indicators)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -4.6541 -3.3419 -0.6944  2.5288  6.9163
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)      4.5145     3.3240   1.358   0.1933
## LoanPaymentsOverdue -2.2485     0.9033  -2.489   0.0242 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.954 on 16 degrees of freedom
## Multiple R-squared:  0.2792, Adjusted R-squared:  0.2341
## F-statistic: 6.196 on 1 and 16 DF,  p-value: 0.02419
```

5(b)

The 95% confidence interval for the slope is (-4.163,-0.333). Since the confidence interval is negative, there is evidence of a significant negative linear association.

```
confint(lm_indicators)

##              2.5 %      97.5 %
## (Intercept) -2.532112 11.5611000
## LoanPaymentsOverdue -4.163454 -0.3335853
```

5(c)

$E(Y|X = 4) = -4.479$ with 95% confidence interval (-6.648, -2.310). Since the confidence interval does not include 0%, it is not a feasible value.

```
predict(lm_indicators, newdata=data.frame(LoanPaymentsOverdue=4), interval='confidence')

##      fit      lwr      upr
## 1 -4.479585 -6.648849 -2.310322
```