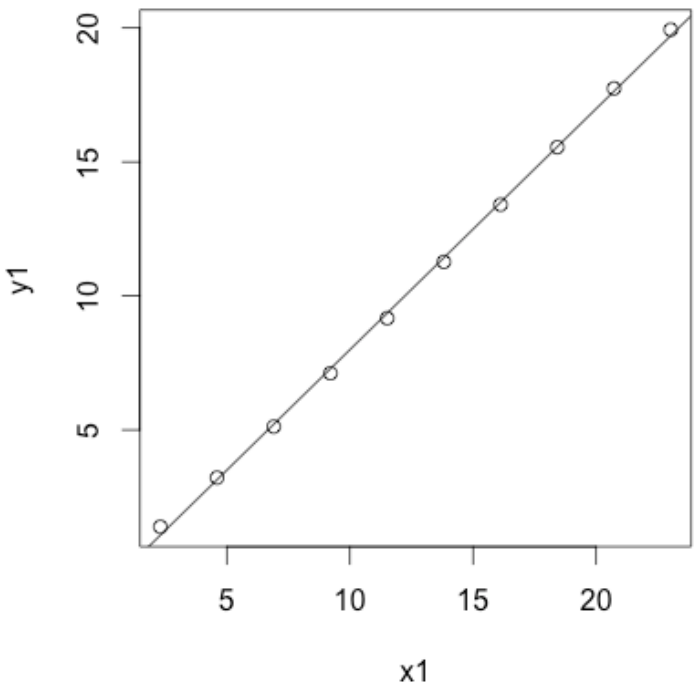
MSIA401 Stat
HW1
Shawn Xiang Li
Prof. Tamhane's book:

2.1 (a) since y= P(x) = 1/ln(x) is not a linear model. So we need to transform it by log both y and x.
(b) Plot Y vs. X, get the following graph:

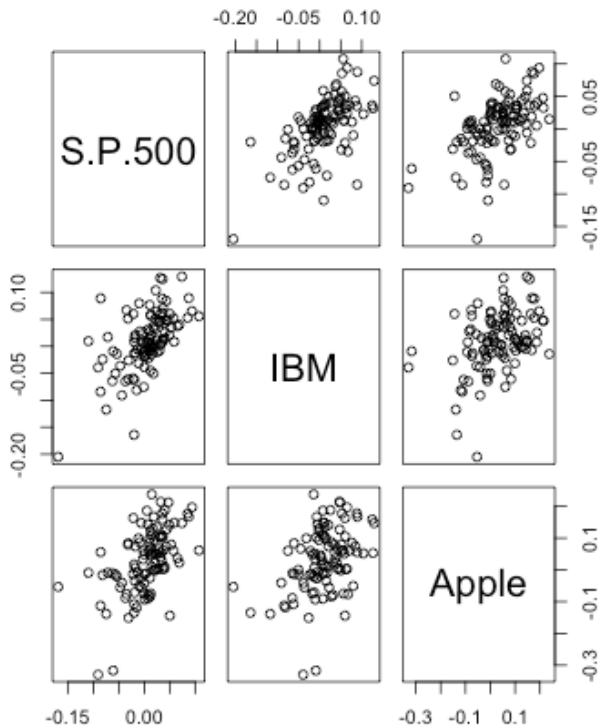


	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	-0.999290	0.125841	-7.941	4.61e-05	***
x1	0.899492	0.008808	102.123	9.44e-14	

95%CI of slope is [0.8791808, 0.9198032], the slope coefficient is within 95%CI of slope. So, the slope coefficient is close to what is predicated by the prime number theorem.

MSIA401 Stat
HW1
Shawn Xiang Li

2.2 (a)



Based on the scatter plots, SP500, IBM and Apple don't have a linear relationship with each other.

(b)

```
SP500 VS Apple
Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.002587   0.003862   -0.67    0.504
apple        0.232712   0.036081    6.45  3.8e-09 ***
```

```
SP500 VS IBM
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.0001799  0.0035770  -0.050     0.96
IBM          0.4792907  0.0636926   7.525 2.15e-11
```

After ran regression, we can get beta(Apple) = 0.232712 and beta(SP500) = 0.4792907. It tells that IBM had a higher expected return relative to SP500.

MSIA401 Stat
HW1
Shawn Xiang Li

(c) SD(apple) = 0.103104; SD(IBM) = 0.05557105; SD(SP500) = 0.04457853
correlation matrix is below:

	S.P.500	IBM	Apple
S.P.500	1.0000000	0.5974779	0.5382317
IBM	0.5974779	1.0000000	0.4147253
Apple	0.5382317	0.4147253	1.0000000

Based on above results, Beta ^{hat} = $r*S_y/S_x$..
Beta hat (Apple) = 0.5382317* SD(Apple)/SD(SP500) = 1.24485579
Beta hat (IBM) = 0.5974779* SD(IBM)/SD(SP500) = 0.7448086
These two results is pretty close to the two in (b).

(d)
SD(IBM) is much higher than SD(Apple). It means the volatility of IBM was much higher than Apple's. And we found out the IBM stock had a higher expected return relative to SP500 and also brought more volatilities to SP500. It shows a higher expected return is accompanied by higher volatility of the stock relative to S&P500.

2.3 Price elastics equals dy/y/dx/y, which is coefficient of x. So price elastics = beta 1 in linear model. After run regression of each type of meat, we get:

Demand of Chuck VS. Price of Chuck

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	5.8899	0.2871	20.513	< 2e-16
price_chuck	-1.3687	0.3199	-4.278	9.44e-05

Demand of Porter VS. Price of Porter

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	9.1123	0.5136	17.742	< 2e-16	***
price_porter	-2.6565	0.2752	-9.654	1.23e-12	

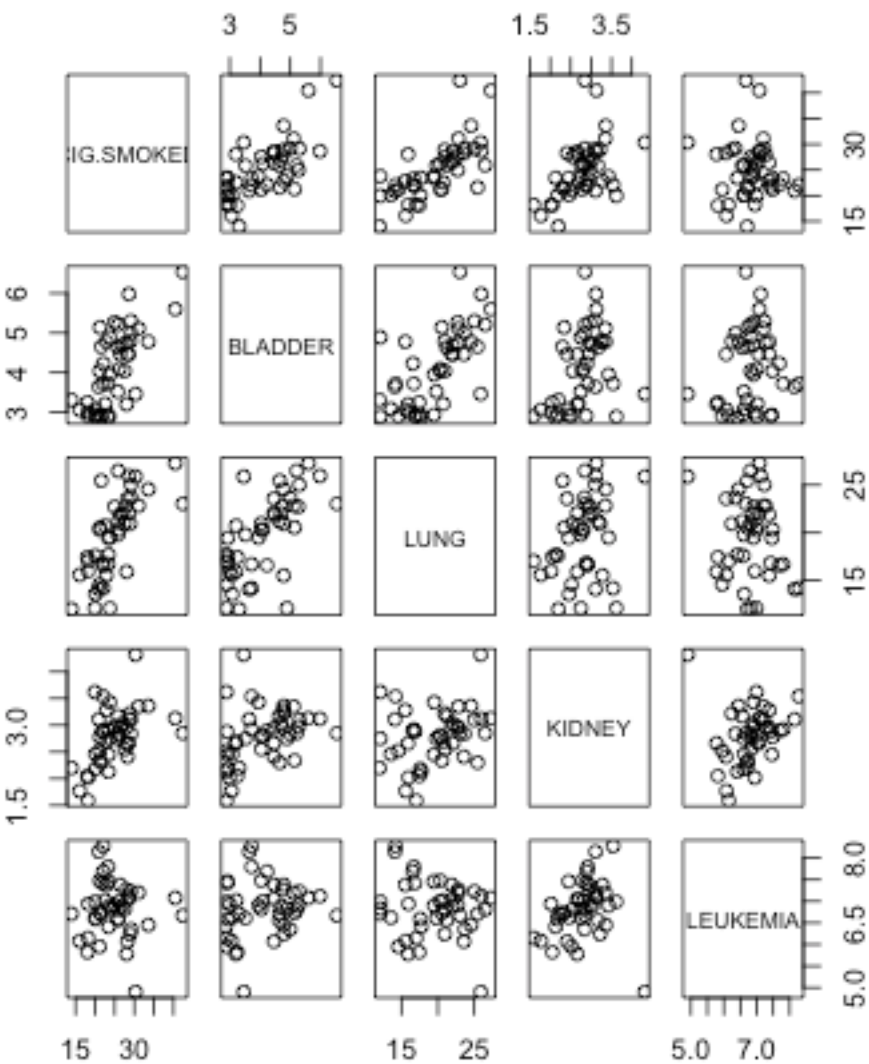
Demand of Ribeye VS. Price of Ribeye

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	7.6627	0.7537	10.167	2.39e-13	***
price_ribeye	-1.4460	0.3731	-3.876	0.000335	

Price elasticity of Chuck = -1.3687; Price elasticity of Porter = -2.6565
Price elasticity of Ribeye = -1.4460
For the data result, price elasticities for all three are not in right order. Since price elasticity of ribeye is between price elasticity of Chuck and Porter.

After price is increased by 10% for each cut:
Demand of chuck approximately decrease 13.687%
Demand of porter approximately decrease 26.565%
Demand of porter approximately decrease 14.46%

MSIA401 Stat
HW1
Shawn Xiang Li
2.4



From the scatter plot, the first row shows the plots of smoking VS each type of cancer. We can tell Bladder, Lung and Kidney cancer has kind of linear relationship with smoking. Leukemia cancer has nonlinear relationship with smoking. There exits outliners within each plot.

	CIG.SMOKED	BLADDER	LUNG	KIDNEY	LEUKEMIA
CIG.SMOKED	1.00000000	0.7036219	0.6974025	0.4873896	-0.06848123
BLADDER	0.70362186	1.0000000	0.6585011	0.3588140	0.16215663
LUNG	0.69740250	0.6585011	1.0000000	0.2827431	-0.15158448
KIDNEY	0.48738962	0.3588140	0.2827431	1.0000000	0.18871294
LEUKEMIA	-0.06848123	0.1621566	-0.1515845	0.1887129	1.00000000

MSIA401 Stat
HW1
Shawn Xiang Li

From above correlation matrix, we can see the bladder cancer death is most highly correlated to cigarette smoking.

Textbook:

2.10 Y: husband's height; X: wife's height

(a) Cov(Y, X) = Cor(Y,X)*sqrt[var(Y)*var(X)] = 69.41294

```
> corXY<-cor(height[,1],height[,2]);
> corXY;
[1] 0.7633864
> varY<-var(height[,1]);
> varY;
[1] 99.21042
> varX<-var(height[,2]);
> varX;
[1] 83.3364
> covXY = corXY * sqrt(varX*varY);
> covXY;
[1] 69.41294
```

(b) Cov(Y, X) = 10.75904

```
>newY <- height[,1]*0.393701;
>newX<-height[,2]*0.393701;
> corXY1<-cor(newY,newX);
> corXY1;
[1] 0.7633864
> varY1<-var(newY);
> varY1;
[1] 15.37766
> varX1<-var(newX);
> varX1;
[1] 12.91718
> covXY1 = corXY1 * sqrt(varX1*varY1);
> covXY1;
[1] 10.75904
```

(c) correlation coefficient = cor(Y,X) = 0.7633864 from (a) R result.
(d) cor(Y,X) = 0.7633864 from (b), which is same to measure in CM.
(e) the correlation between husband heights and wife height is 1. They are perfectly c orrelated.
(f) Response variable Y is husbands' height; explained variable X is wife's height.

MSIA401 Stat
HW1
Shawn Xiang Li

So the Linear model would be $Y = \beta_0 + \beta_1 X$ + error. Since the model is linear, it's possible to convert the model to $X = \beta'_0 + \beta'_1 Y$ + error. It's not different either choosing X or Y to be the response variable.

(g) H0: $\beta 1 = 0$; Ha : $\beta 1 \neq 0$

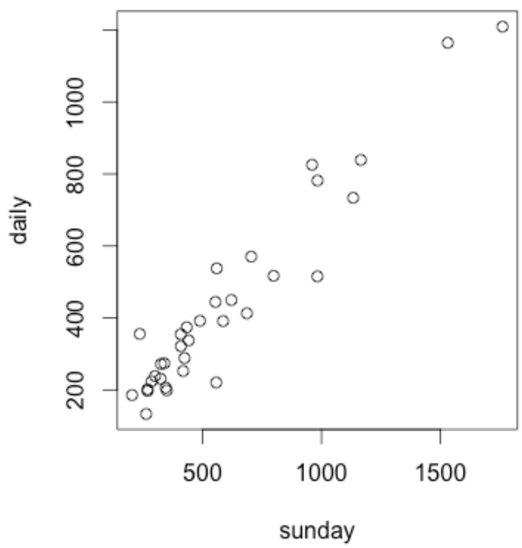
```
Call:
lm(formula = height[, 1] ~ height[, 2])

Residuals:
    Min      1Q  Median      3Q     Max 
-16.7438  -4.2838  -0.1615   4.2562  17.7500 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 37.81005   11.93231   3.169  0.00207 ** 
height[, 2]  0.83292    0.07269  11.458  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 6.468 on 94 degrees of freedom
Multiple R-squared:  0.5828,	Adjusted R-squared:  0.5783 
F-statistic: 131.3 on 1 and 94 DF,  p-value: < 2.2e-16
```

After run the linear regression, we can get the P-value <0.05 (set $\alpha = 0.05$). So, we reject H0: $\beta 1 = 0$ and we conclude that there is a association between Y and X.

(h) H0: $\beta 0 = 0$; Ha : $\beta 0 \neq 0$

We can get the P-value <0.05 (set $\alpha = 0.05$). So, we reject H0: $\beta 0 = 0$ and we conclude that the intercept is not zero.

MSIA401 Stat
HW1
Shawn Xiang Li
2.12 (a)



The scatter plot shows there is a linear trend between Daily and Sunday circulation.

(b) fit data into a linear model: Sunday = $\beta 0+ \beta 1$*Daily+error

```
Call:
lm(formula = sunday ~ daily)

Residuals:
    Min      1Q  Median      3Q     Max
-255.19  -55.57  -20.89   62.73  278.17

Coefficients:
            Estimate   Std. Error t value Pr(>|t|)
(Intercept) 13.83563   35.80401   0.386    0.702
daily        1.33971    0.07075  18.935   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 109.4 on 32 degrees of freedom
Multiple R-squared:  0.9181,	Adjusted R-squared:  0.9155
F-statistic: 358.5 on 1 and 32 DF,  p-value: < 2.2e-16
```

(c) 95%CI for $\beta 0$: [-59.094743, 86.766003]; 95%CI for $\beta 1$: [1.195594 1.483836]

(d) set H0: $\beta 1=0$; Ha : $\beta 1 \neq 0$
After run regression, we get:
P-value is less than 0.05, so we reject H0. We conclude that there is a significant relationship between Sunday and Daily circulation.
(e) R^2 = 91.81%

MSIA401 Stat
HW1
Shawn Xiang Li

(f) Given daily = 500000, 95% CI for Sunday circulation is [644195.1, 723191]
(g) Given daily = 500000, 95% Predictive Interval for Sunday circulation is [457336.7, 910049.3]. Although PI is very like CI in (f), by thereon, PI is always longer than CI. The theorem is proved by the results here.

(h) Given daily circulation = 2000000, 95 CI for Sunday circulation is [2463926, 2922604]. 95% PI is [2373463, 3013068]. Intervals here are more shorter than the on in (g). It's more accurate.