

Homework 1

Yi Yang (yy3089)

October 9, 2021

1 P1

1.1 a

$$\begin{aligned}
 (n-1)S^2 + n\bar{X}^2 &= \sum_{i=1}^n (X_i - \bar{X})^2 + n\bar{X}^2 \\
 &= \sum_{i=1}^n X_i^2 - 2 \sum_{i=1}^n (X_i \bar{X}) + \sum_{i=1}^n \bar{X}^2 + n\bar{X}^2 \\
 &= \sum_{i=1}^n X_i^2 - 2\bar{X} \sum_{i=1}^n (X_i) + \sum_{i=1}^n \bar{X}^2 + n\bar{X}^2 \\
 &= \sum_{i=1}^n X_i^2 - 2\bar{X} * n\bar{X} + n\bar{X}^2 + n\bar{X}^2 \\
 &= \sum_{i=1}^n X_i^2
 \end{aligned} \tag{1.1}$$

1.2 b

$$\begin{aligned}
 \mathbb{E}S^2 &= \frac{1}{n-1} \left(\sum_{i=1}^n \mathbb{E}X_i^2 - n\mathbb{E}\bar{X}^2 \right) \\
 \mathbb{E}X_i^2 &= D(X_i) + (\mathbb{E}X_i)^2 = \sigma^2 + \mu^2, \quad i = 1, \dots, n \\
 \mathbb{E}\bar{X}^2 &= D(\bar{X}) + (\mathbb{E}\bar{X})^2 = \sigma^2/n + \mu^2 \\
 \mathbb{E}S^2 &= \frac{1}{n-1} [n(\sigma^2 + \mu^2) - n(\sigma^2/n + \mu^2)] = \sigma^2
 \end{aligned} \tag{1.2}$$

So S^2 is an unbiased estimator of σ^2 .

1.3 c

$$\begin{aligned}
 D(\bar{X}) &= D\left(\frac{1}{n} \sum_{i=1}^n X_i\right) = \frac{1}{n^2} \sum_{i=1}^n D(X_i) = \frac{1}{n} \sigma^2 \\
 D(X_i - \bar{X}) &= D(X_i - \frac{1}{n} \sum_{i=1}^n X_i) = \left(\frac{n-1}{n}\right)^2 \sigma^2 + (n-1)\left(\frac{1}{n}\right)^2 \sigma^2 = \frac{n-1}{n} \sigma^2
 \end{aligned} \tag{1.3}$$

We need to prove that $Cov(\bar{X}, X_i - \bar{X}) = 0$:

$$\begin{aligned}
D(X_i - \bar{X} + \bar{X}) &= D(X_i) = \sigma^2 \\
D(X_i - \bar{X} + \bar{X}) &= D(X_i - \bar{X}) + D(\bar{X}) + 2Cov(\bar{X}, X_i - \bar{X}) \\
&= \frac{n-1}{n}\sigma^2 + \frac{1}{n}\sigma^2 + 2Cov(\bar{X}, X_i - \bar{X}) = \sigma^2
\end{aligned} \tag{1.4}$$

Thus, $Cov(\bar{X}, X_i - \bar{X}) = 0$.

1.4 d

In section c we have proved that \bar{X} is independent of $X_i - \bar{X}$ for any i , so \bar{X} is independent of $f(X_1 - \bar{X}, X_2 - \bar{X}, \dots, X_n - \bar{X})$.

$$S^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2 = g(X_1 - \bar{X}, X_2 - \bar{X}, \dots, X_n - \bar{X}) \tag{1.5}$$

Thus, \bar{X} is independent of S^2 .

2 P2

Let sample covariance $Q(X, Y) = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})$:

$$\begin{aligned}\hat{\beta}_1 &= \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2} \\ \hat{\beta}_0 &= \bar{y} - \hat{\beta}_1 \bar{x}\end{aligned}\tag{2.1}$$

For $R^2 = \frac{\sum_{i=1}^n (\hat{y}_i - \bar{y})^2}{\sum_{i=1}^n (y_i - \bar{y})^2}$:

$$\begin{aligned}\sum_{i=1}^n (\hat{y}_i - \bar{y})^2 &= \sum_{i=1}^n (\bar{y} - \hat{\beta}_1 \bar{x} + \hat{\beta}_1 x_i - \bar{y})^2 \\ &= \hat{\beta}_1^2 \sum_{i=1}^n (x_i - \bar{x})^2\end{aligned}\tag{2.2}$$

$$\begin{aligned}R^2 &= R^2 = \frac{\sum_{i=1}^n (\hat{y}_i - \bar{y})^2}{\sum_{i=1}^n (y_i - \bar{y})^2} \\ &= \hat{\beta}_1^2 \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{\sum_{i=1}^n (y_i - \bar{y})^2}\end{aligned}\tag{2.3}$$

$$\begin{aligned}r^2 &= \frac{(\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}))^2}{\sum_{i=1}^n (x_i - \bar{x})^2 \sum_{i=1}^n (y_i - \bar{y})^2} \\ &= \hat{\beta}_1 \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (y_i - \bar{y})^2} \\ &= \hat{\beta}_1 \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}) \sum_{i=1}^n (x_i - \bar{x})^2}{\sum_{i=1}^n (x_i - \bar{x})^2 \sum_{i=1}^n (y_i - \bar{y})^2} \\ &= \hat{\beta}_1^2 \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{\sum_{i=1}^n (y_i - \bar{y})^2}\end{aligned}\tag{2.4}$$

Thus, we can conclude that $R^2 = r^2$.

3 P3

3.1 a

```
1 set.seed(1)
2 x = rnorm(100, mean=0, sd=1)
```

3.2 b

```
1 eps = rnorm(100, mean = 0, sd = 0.25)
```

3.3 c

```
1 y = -1 + 0.5*x + eps
2 length(y)
```

In this question, $length(y) = 100, \beta_0 = -1, \beta_1 = 0.5$

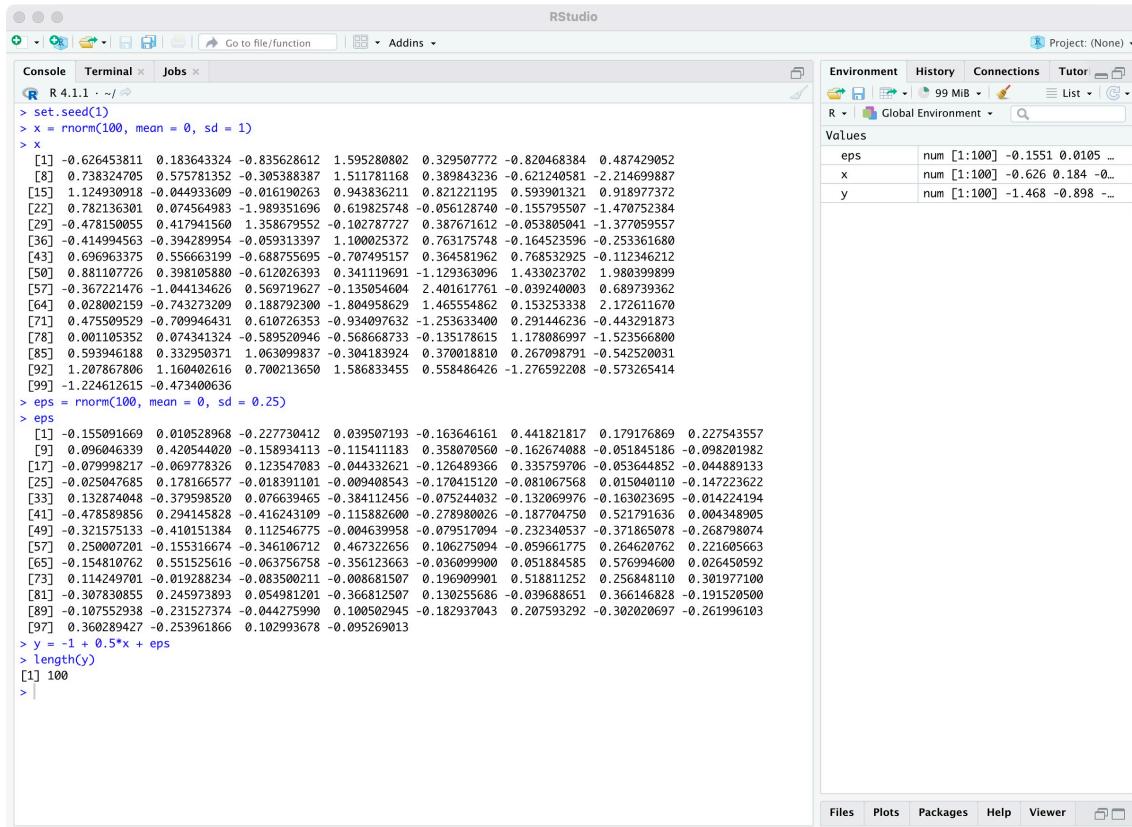


Figure 3.1: Screenshot for '3-c'

3.4 d

```
1 plot(x, y, xlab = "x", ylab = "y", main = "Y-X Relationship")
```

There is some linear relationship between Y and X.

3.5 e

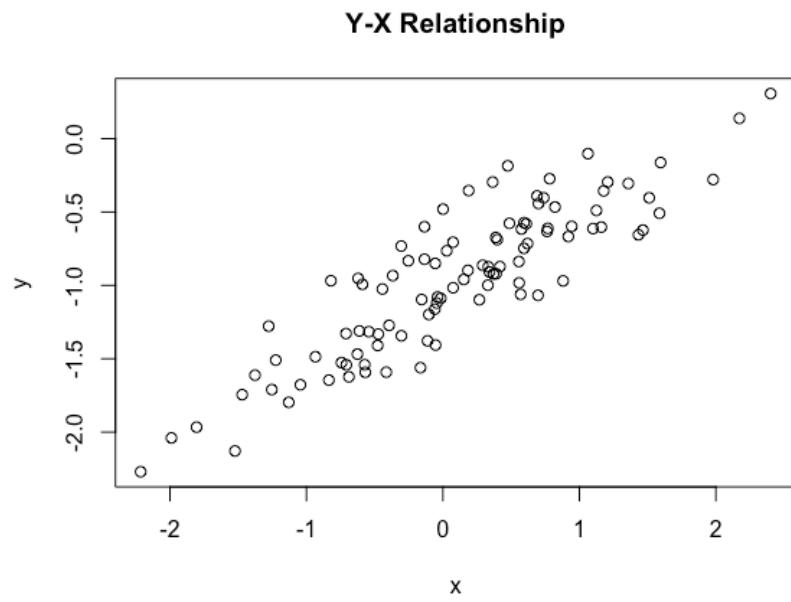


Figure 3.2: Screenshot for '3-d'

```

1 lm.fit = lm(y ~ x)
2 summary(lm.fit)
3 plot(x, y, xlab = "x", ylab = "y", main = "Y-X Relationship")
4 abline(lm.fit)

```

$\hat{\beta}_0 = -1.00942$, $\hat{\beta}_1 = 0.49973$, for both $\hat{\beta}_1, \hat{\beta}_0, Pr(\geq |t|) < 2e-16$, which means that we can reject the hypothesis that $\hat{\beta}_1 = 0, \hat{\beta}_0 = 0$. The p -value is less than $2e-16$ and R^2 -statistic is 0.7784, which means we can reject the hypothesis that there is no linear relationship between Y and X.

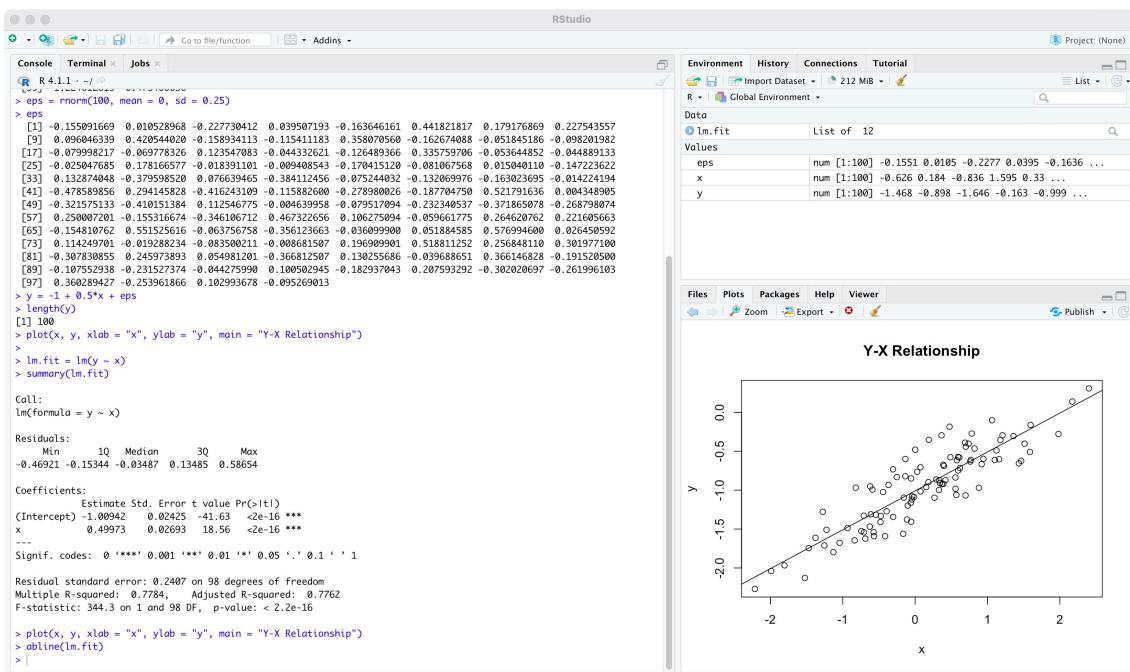


Figure 3.3: Screenshot for '3-e'

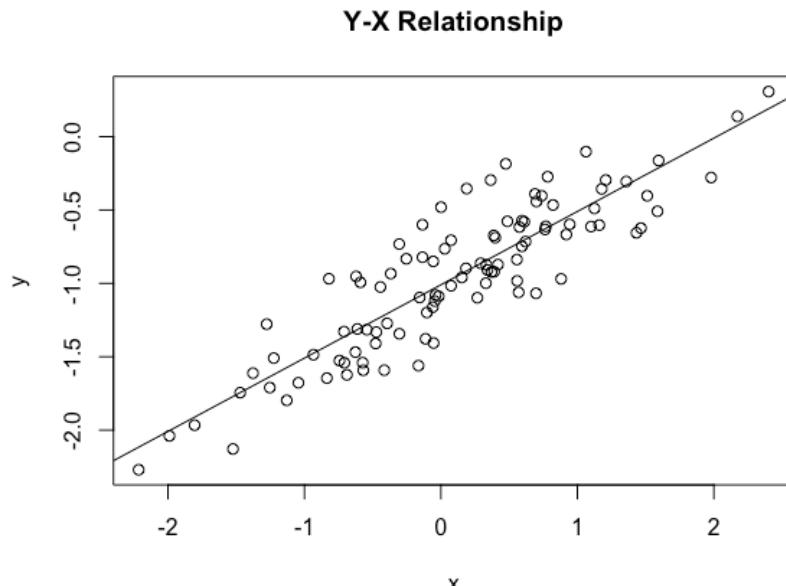


Figure 3.4: Screenshot for 'Least Squares Linear Model'

3.6 f

```
1 par(col='black')
2 plot(x, y, xlab = "x", ylab = "y", main = "Y-X Relationship")
3 abline(lm.fit, col='red', lty=3)
4 abline(a=-1, b=0.5, col='green', lty=1)
5 legend('topleft', inset=0.05, c('least-squares', 'population-regression')\
6 , lty = c(3, 1), col = c('red', 'green'), bty="o")
```

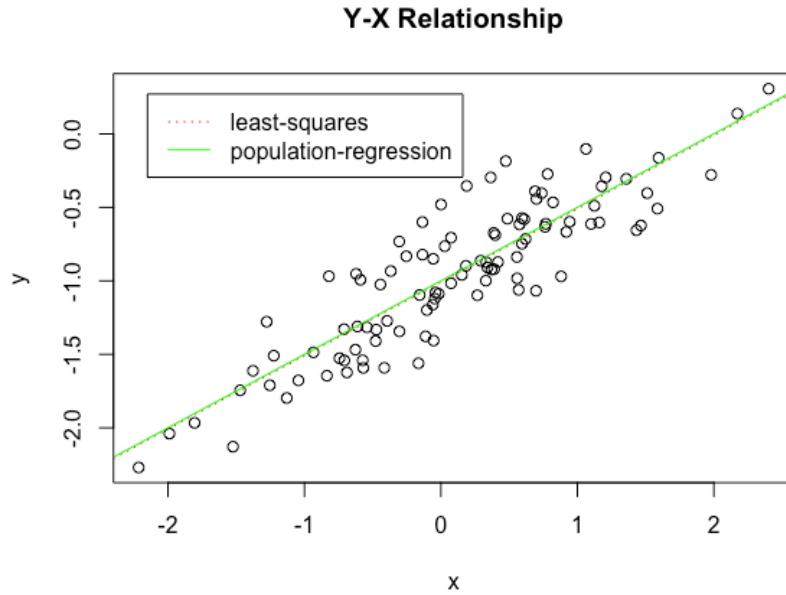


Figure 3.5: Screenshot for '*Least-Squares and Population-Regression model*'

3.7 g

```
1 poly_fit = lm(y ~ poly(x, 2))
2 summary(poly_fit)
3 xlims = range(x)
4 x.grid = seq(from=xlims[1], to=xlims[2])
5 predictions = predict(poly_fit, newdata = list(x = x.grid), se=TRUE)
6 se.bands = cbind(predictions$fit+2*predictions$se.fit \
7 , predictions$fit-2*predictions$se.fit)
8 plot(x, y, xlab = "X", ylab = "Y", main = "Polynomial Relationship" \
9 , col = "black")
10 lines(x.grid, predictions$fit, lwd = 2, col = "green", lty = 1)
11 matlines(x.grid, se.bands, lwd = 1, col="red", lty = 3)
12 legend('topleft', inset = 0.05, c("polynomial-regression", "SE-wrapper"), \
13 lwd = c(2, 1), lty = c(1, 3), col = c("green", "red"), bty = "o")
```

From the t -test result, we can find that the $Pr(>|t|)$ for the X^2 term is 0.164, which means there is a possibility of 0.164 that we accept the hypothesis that there is no relationship between Y and X^2 . And the polynomial regression result is very close to the linear model's least squares result.

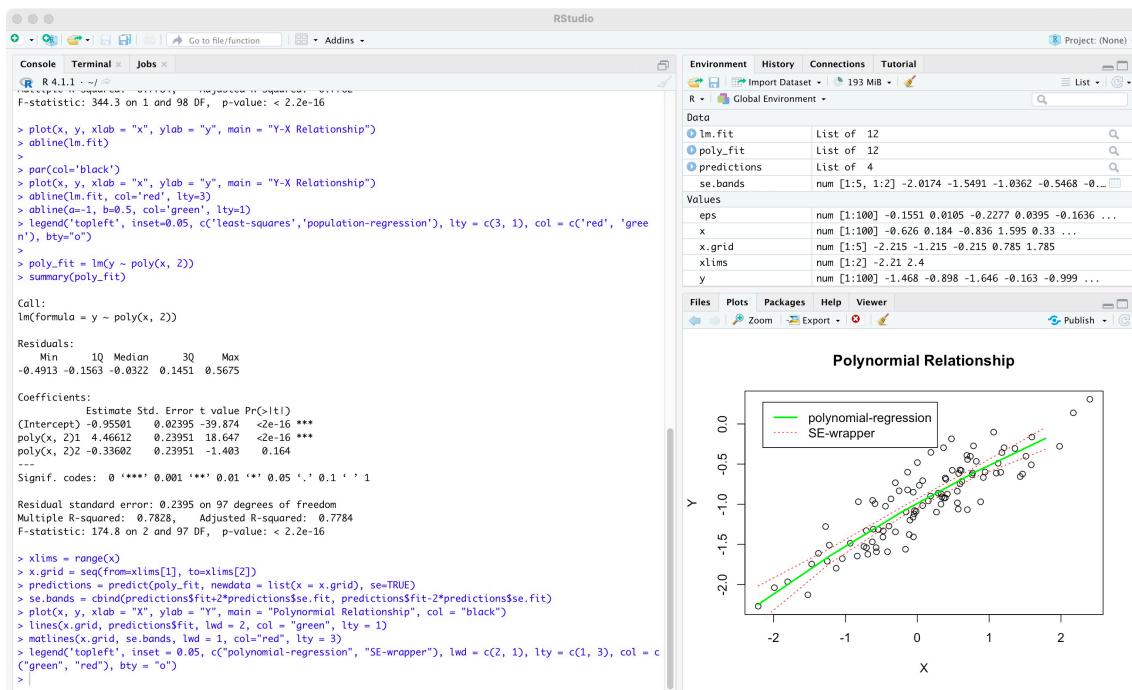


Figure 3.6: Screenshot for '3-g'

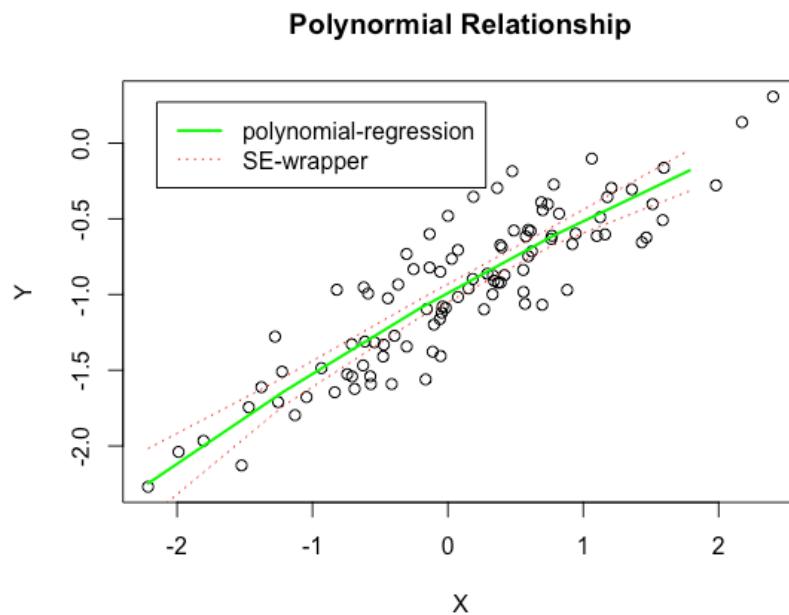


Figure 3.7: Screenshot for 'Polynomial Regression Model'

3.8 h

```

1 set.seed(1)
2 x = rnorm(100, mean = 0, sd = 1)
3 eps = rnorm(100, mean = 0, sd = 0.01)
4 y = -1 + 0.5*x + eps
5 length(y)
6
7 plot(x, y, xlab = "x", ylab = "y", main = "Y-X Relationship")
8 lm.fit = lm(y ~ x)
9 summary(lm.fit)
10
11 plot(x, y, xlab = "x", ylab = "y", main = "Y-X Relationship")
12 abline(lm.fit)
13 par(col='black')
14 plot(x, y, xlab = "x", ylab = "y", main = "Y-X Relationship")
15 abline(lm.fit, col='red', lty=3)
16 abline(a=-1, b=0.5, col='green', lty=1)
17 legend('topleft', inset=0.05, c('least-squares','population-regression')\
18 , lty = c(3, 1), col = c('red', 'green'), bty="o")

```

$\hat{\beta}_0 = -0.10003769, \hat{\beta}_1 = 0.4999894$, R-statistic result = 0.9995. This result is much better than that in section e.

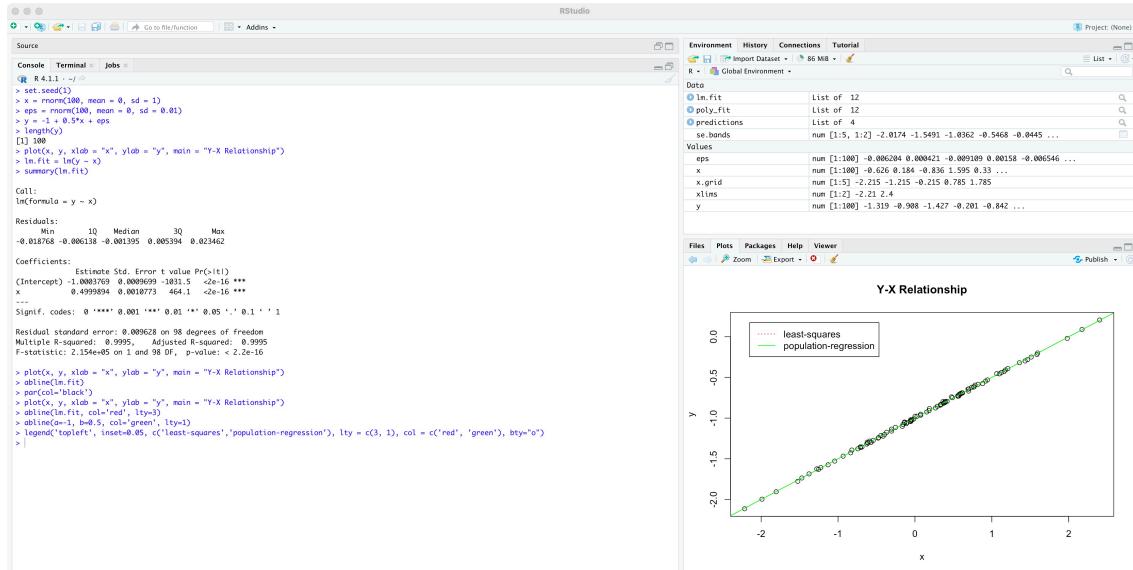


Figure 3.8: Screenshot for '3-h'

3.9 i

```

1 set.seed(1)
2 x = rnorm(100, mean = 0, sd = 1)
3 eps = rnorm(100, mean = 0, sd = 0.81)
4 y = -1 + 0.5*x + eps
5 length(y)

```

```

6
7 plot(x, y, xlab = "x", ylab = "y", main = "Y-X Relationship")
8 lm.fit = lm(y ~ x)
9 summary(lm.fit)
10
11 plot(x, y, xlab = "x", ylab = "y", main = "Y-X Relationship")
12 abline(lm.fit)
13 par(col='black')
14 plot(x, y, xlab = "x", ylab = "y", main = "Y-X Relationship")
15 abline(lm.fit, col='red', lty=3)
16 abline(a=-1, b=0.5, col='green', lty=1)
17 legend('topleft', inset=0.05, c('least-squares', 'population-regression')\
18 , lty = c(3, 1), col = c('red', 'green'), bty="o")

```

$\hat{\beta}_0 = -0.103053, \hat{\beta}_1 = 0.49914$, R-statistic result = 0.2503. This result is much worse than that in section e.

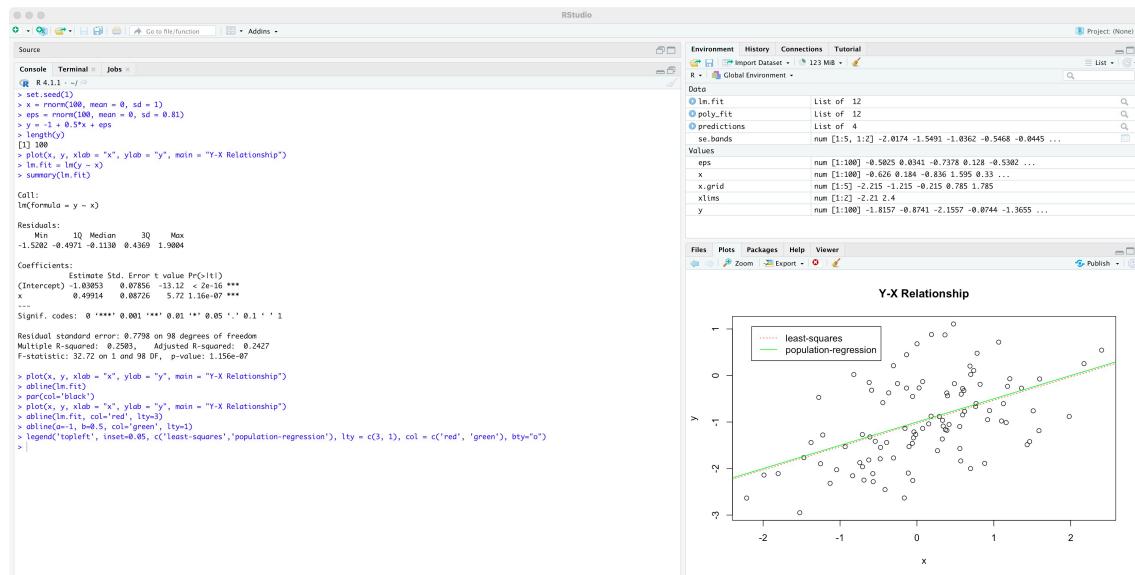


Figure 3.9: Screenshot for '3-i'

3.10 j

for sd = 0.25:

```

1 confint(lm.fit)
              2.5 %      97.5 %
2 (Intercept) -1.0575402 -0.9613061
3 x            0.4462897  0.5531801

```

for sd = 0.01:

```

1 confint(lm.fit)
              2.5 %      97.5 %
2 (Intercept) -1.0023016 -0.9984522
3 x            0.4978516  0.5021272

```

for sd = 1:

```
1 confint(lm.fit)
2                   2.5 %      97.5 %
3 (Intercept) -1.2301607 -0.8452245
4 x            0.2851588  0.7127204
```

When there is less noise, the confidence intervals for both β_0 and β_1 are smaller which means the estimations are more close to true value; When there is more noise, the confidence intervals for both β_0 and β_1 are larger which means the estimations are less close to true value;

4 P4

For TV:

```
1 adv = read.csv("advertising.csv", header = T, na.strings = '?')
2 adv = na.omit(adv)
3 TV_fit = lm(adv$Sales ~ adv$TV)
4 TV_confint = confint(TV_fit, level = 0.92)
5 plot(adv$TV, adv$Sales, xlab = "TV\u222aAdvertising", ylab = "Sales",
6       main="Sales-TV\u222aAdvertising")
7 abline(TV_fit, col = 'green', lty = 1)
8 abline(a = TV_confint[1, 1], TV_confint[2, 1], col = 'red', lty = 3)
9 abline(a = TV_confint[1, 2], TV_confint[2, 2], col = 'red', lty = 3)
10 legend('topleft', inset = 0.05, c('least-squares',
11      '0.92\u2225condifence\u2225interval'), lty = c(1, 3), col = c('green', 'red'))
12 bty = 'o')
13 TV_confint
14                 4 %         96 %
15 (Intercept) 6.40721844 7.54242454
16 adv$TV        0.05212914 0.05880041
```

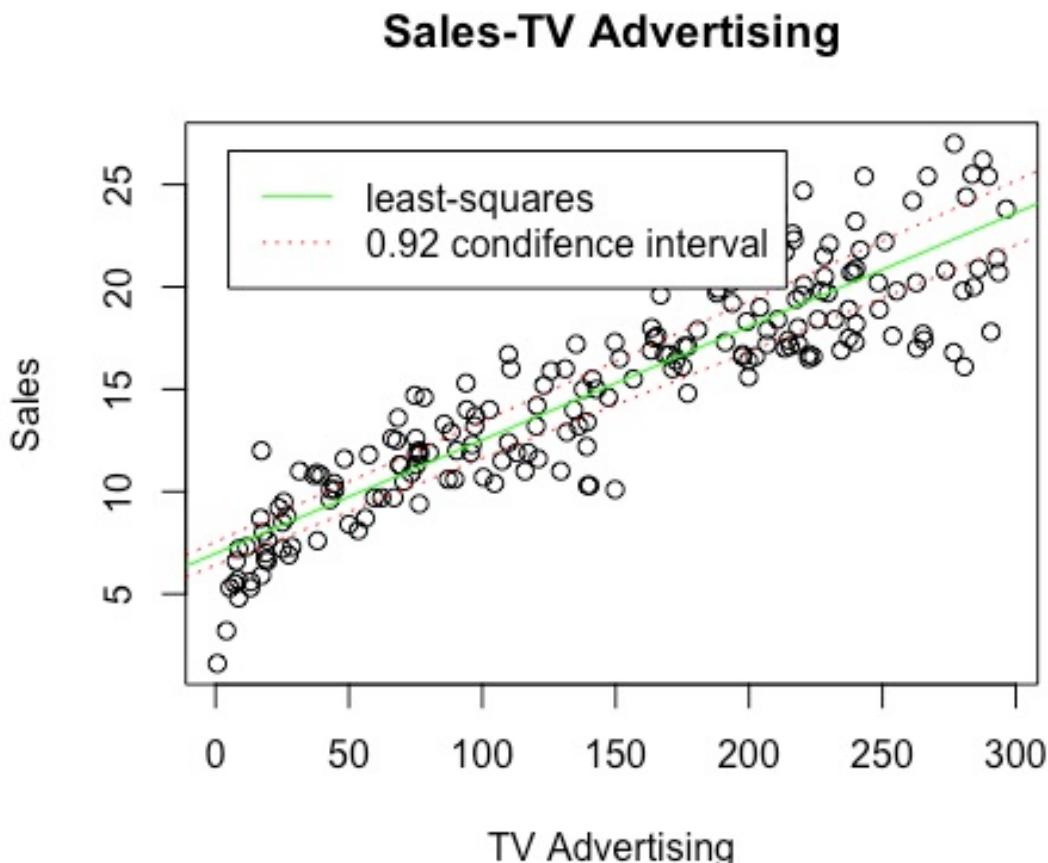


Figure 4.1: Screenshot for 'Sales-TV'

For Radio:

```
1 adv = read.csv("advertising.csv", header = T, na.strings = '?')
2 adv = na.omit(adv)
3 Radio_fit = lm(adv$Sales ~ adv$Radio)
4 Radio_confint = confint(Radio_fit, level = 0.92)
5 plot(adv$Radio, adv$Sales, xlab = "Radio_Advertising", ylab = "Sales",
6       main="Sales-Radio_Advertising")
7 abline(Radio_fit, col = 'green', lty = 1)
8 abline(a = Radio_confint[1, 1], Radio_confint[2, 1], col = 'red', lty = 3)
9 abline(a = Radio_confint[1, 2], Radio_confint[2, 2], col = 'red', lty = 3)
10 legend('topleft', inset = 0.05, c('least-squares',
11      '0.92 confidence interval'), lty = c(1, 3), col = c('green', 'red'))
12 bty = 'o')
13 Radio_confint
14
15        4 %      96 %
15 (Intercept) 11.08577062 13.38567
16 adv$Radio     0.08273333  0.16613
```

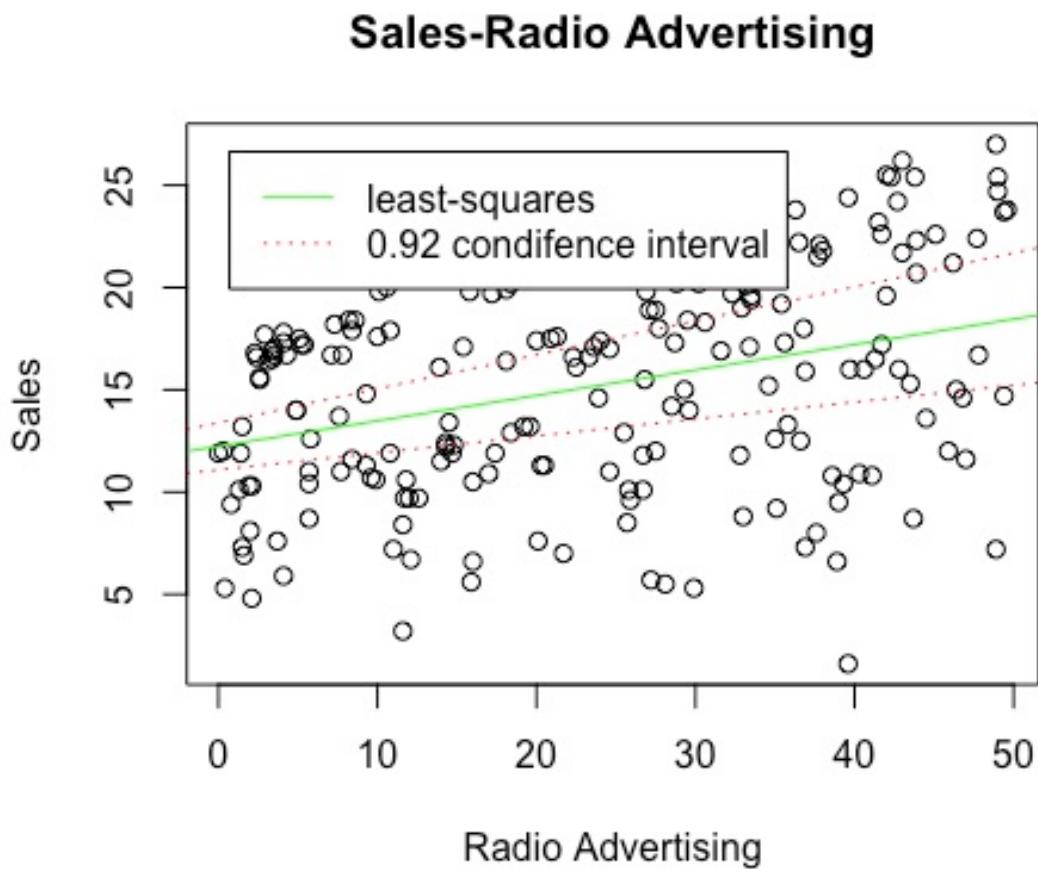


Figure 4.2: Screenshot for 'Sales-Radio'

For Newspaper:

```
1 adv = read.csv("advertising.csv", header = T, na.strings = '?')
```

```

2 adv = na.omit(adv)
3 Newspaper_fit= lm(adv$Sales ~ adv$Newspaper)
4 Newspaper_confint = confint(Newspaper_fit, level = 0.92)
5 plot(adv$Newspaper, adv$Sales, xlab = "Newspaper_Advertising",
6      ylab = "Sales", main="Sales-Newspaper_Advertising")
7 abline(Newspaper_fit, col = 'green', lty = 1)
8 abline(a = Newspaper_confint[1, 1], Newspaper_confint[2, 1], col = 'red',
9 abline(a = Newspaper_confint[1, 2], Newspaper_confint[2, 2], col = 'red',
10 legend('topleft', inset = 0.05, c('least-squares',
11       '0.92 confidence interval'), lty = c(1, 3), col = c('green', 'red'))
12       bty = 'o')
13 Newspaper_confint
14
15 (Intercept) 12.83634111 15.08275619
16 adv$Newspaper 0.00836363 0.06828436

```

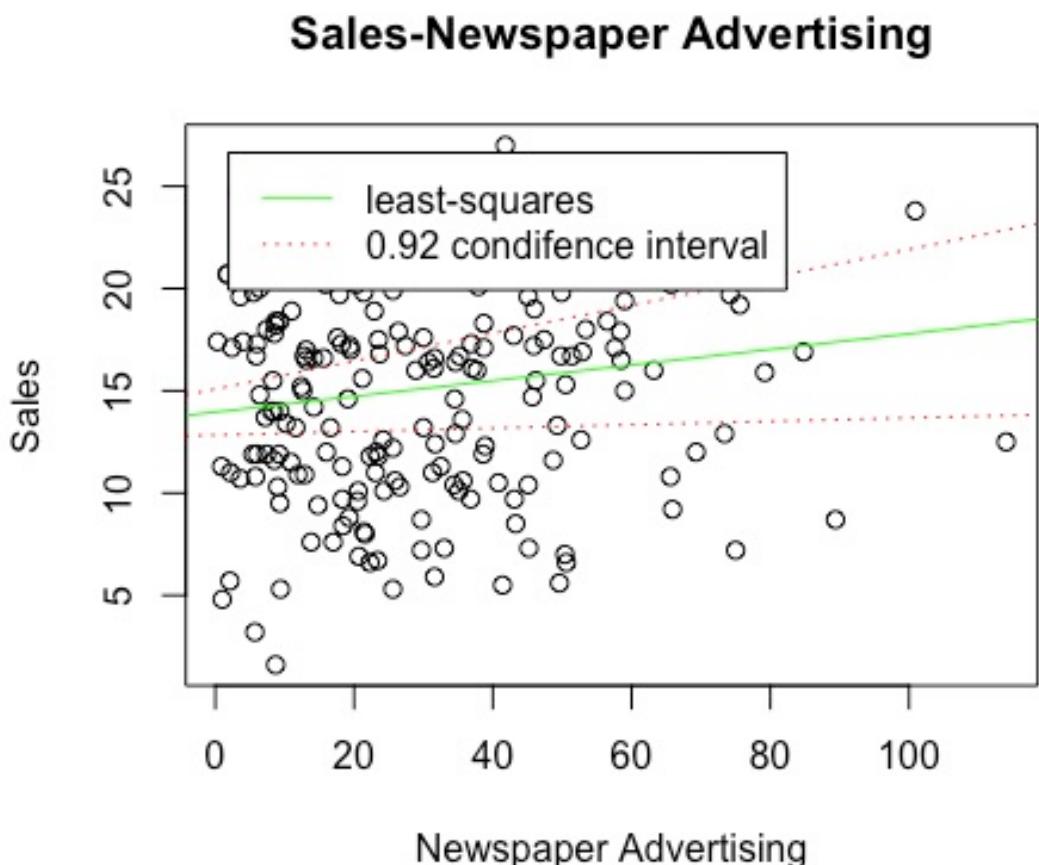


Figure 4.3: Screenshot for '*Sales-Newspaper*'

5 P5

5.1 a

```
1 auto = read.csv("auto-mpg.csv", header = T, na.strings = "?",
2                   stringsAsFactors = T)
3 auto = na.omit(auto)
4 names(auto)
5 pairs(auto[1:9])
```

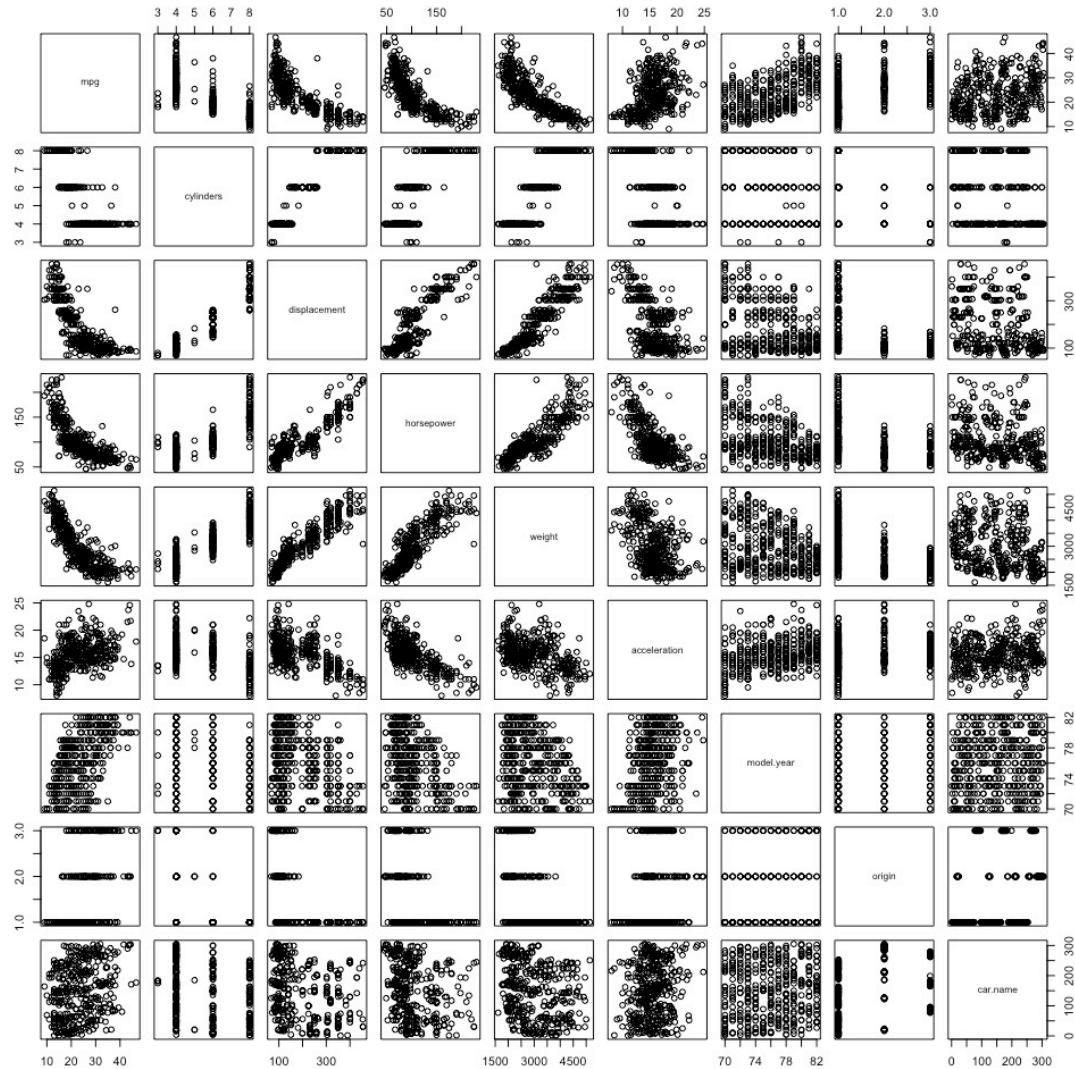


Figure 5.1: Screenshot for 'Pairs'

5.2 b

```
1 Cor = cor(auto[1:8])
```

5.3 c

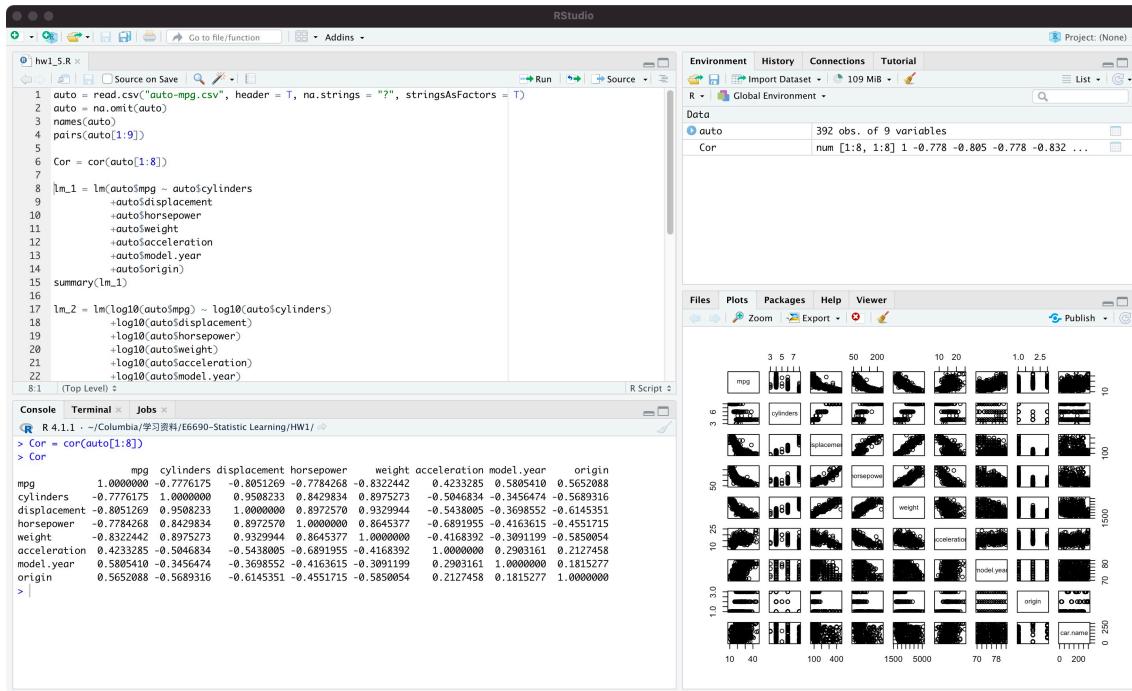


Figure 5.2: Screenshot for 'Cor'

```

1 lm_1 = lm(auto$mpg ~ auto$cylinders
2           + auto$displacement
3           + auto$horsepower
4           + auto$weight
5           + auto$acceleration
6           + auto$model.year
7           + auto$origin)
8 summary(lm_1)

```

- As the p -value for this linear-regression is less than $2e - 16$, we should reject the null hypothesis that there is no linear relationship between Mpg and all variables. The $R - statistic = 0.8215$ shows that the linear regression fits very well.
- The $Pr(< |t|)$ for displacement, weight, model year and origin are very small, which means that they have a statistically significant relationship to the response.
- The coefficient of year suggest that there is a positive relationship between year and mpg. 1 increase in year will lead to 0.75 increase in mpg.

5.4 d

- Different transformation on variables will lead to different statistical significance for each variable.
- Different transformation on variables will lead to different coefficient for each variable, which means they will have different relationship with the response.

RStudio interface showing the results of the lm_1 model. The console output shows the model formula, coefficients, residuals, and summary statistics. The environment pane shows the global environment with objects like auto, lm_1, lm_2, lm_3, and lm_4. The files pane shows the project structure and files.

```

> lm_1 = lm(auto$mpg ~ auto$cylinders + auto$displacement + auto$horsepower + auto$weight + auto$acceleration + auto$model.year + auto$origin)
> summary(lm_1)

Call:
lm(formula = auto$mpg ~ auto$cylinders + auto$displacement +
    auto$horsepower + auto$weight + auto$acceleration + auto$model.year +
    auto$origin)

Residuals:
    Min      1Q   Median      3Q     Max 
-9.5903 -2.1565  0.1169  1.8690 13.0604 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  17.218435  4.644294 -3.707 0.00024 ***
auto$cylinders -0.493376  0.323282 -1.526 0.12780    
auto$displacement 0.019896  0.007515  2.647 0.00844 **  
auto$horsepower -0.016951  0.013787 -1.236 0.21963    
auto$weight     -0.006474  0.000652 -9.929 < 2e-16 ***
auto$acceleration 0.0809576 0.098845  0.815 0.41548    
auto$model.year  0.570773  0.050973 14.729 < 2e-16 ***
auto$origin      1.426141  0.278136  5.127 4.67e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.328 on 384 degrees of freedom
Multiple R-squared:  0.8215, Adjusted R-squared:  0.8182 
F-statistic: 252.4 on 7 and 384 DF, p-value: < 2.2e-16

```

Figure 5.3: Screenshot for 'X'

RStudio interface showing the results of the lm_2 model. The console output shows the model formula, coefficients, residuals, and summary statistics. The environment pane shows the global environment with objects like auto, lm_1, lm_2, lm_3, and lm_4. The files pane shows the project structure and files.

```

> lm_2 = lm(log10(auto$mpg) ~ log10(auto$cylinders) + log10(auto$displacement) + log10(auto$horsepower) + log10(auto$weight) + log10(auto$acceleration) + log10(auto$model.year) + log10(auto$origin))
> summary(lm_2)

Call:
lm(formula = log10(auto$mpg) ~ log10(auto$cylinders) + log10(auto$displacement) +
    log10(auto$horsepower) + log10(auto$weight) + log10(auto$acceleration) +
    log10(auto$model.year) + log10(auto$origin))

Residuals:
    Min      1Q   Median      3Q     Max 
-0.179356 -0.030825  0.000238  0.026708 0.171685 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  0.067485  0.281523 -0.240  0.81068    
log10(auto$cylinders) -0.082815  0.061429 -1.348  0.17841    
log10(auto$displacement) 0.006625  0.056974  0.116  0.98748    
log10(auto$horsepower) -0.294389  0.057652 -5.106 5.18e-07 ***
log10(auto$weight)     -0.569666  0.082397 -6.914 1.98e-11 ***  
log10(auto$acceleration) -0.179239  0.059538 -3.011 0.00278 **  
log10(auto$model.year)  2.243989  0.131661 17.044 < 2e-16 ***
log10(auto$origin)     0.044848  0.018821  2.383 0.01767 *  
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.04935 on 384 degrees of freedom
Multiple R-squared:  0.8903, Adjusted R-squared:  0.8883 
F-statistic: 445.3 on 7 and 384 DF, p-value: < 2.2e-16

```

Figure 5.4: Screenshot for 'logX'

```

R 4.1.1 - ~/Columbia/学习资料/E6690-Statistic Learning/HW1/
> lm_3 = lm(sqrt(auto$mpg) ~ sqrt(auto$cylinders) + sqrt(auto$displacement) +
+   +sqrt(auto$horsepower) + sqrt(auto$weight) + sqrt(auto$acceleration) +
+   +sqrt(auto$model.year) + sqrt(auto$origin))
> summary(lm_3)

Call:
lm(formula = sqrt(auto$mpg) ~ sqrt(auto$cylinders) + sqrt(auto$displacement) +
  sqrt(auto$horsepower) + sqrt(auto$weight) + sqrt(auto$acceleration) +
  sqrt(auto$model.year) + sqrt(auto$origin))

Residuals:
    Min      1Q Median      3Q     Max 
-0.98667 -0.17288 -0.00315  0.16145  1.02245 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  0.949286  0.847481  -2.300 0.021979 *  
sqrt(auto$cylinders) -0.108552  0.141968  -0.765 0.444964  
sqrt(auto$displacement) 0.019707  0.021182  0.930 0.352752  
sqrt(auto$horsepower) -0.090896  0.028428  -3.197 0.001582 *** 
sqrt(auto$weight)       -0.061414  0.007292  -8.422 7.48e-16 *** 
sqrt(auto$acceleration) -0.107258  0.077948  -1.392 0.164699  
sqrt(auto$model.year)   1.266015  0.079308  15.963 < 2e-16 *** 
sqrt(auto$origin)       0.272324  0.070883  3.842 0.000143 *** 
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.2964 on 384 degrees of freedom
Multiple R-squared:  0.8662, Adjusted R-squared:  0.8638 
F-statistic: 355.1 on 7 and 384 DF, p-value: < 2.2e-16

> 

```

Figure 5.5: Screenshot for ' \sqrt{X} '

```

R 4.1.1 - ~/Columbia/学习资料/E6690-Statistic Learning/HW1/
> lm_4 = lm(auto$mpg^2 ~ auto$cylinders^2 +
+   +auto$displacement^2 +
+   +auto$horsepower^2 +
+   +auto$weight^2 +
+   +auto$acceleration^2 +
+   +auto$model.year^2 +
+   +auto$origin^2)
> summary(lm_4)

Call:
lm(formula = auto$mpg^2 ~ auto$cylinders^2 + auto$displacement^2 +
  auto$horsepower^2 + auto$weight^2 + auto$acceleration^2 +
  auto$model.year^2 + auto$origin^2)

Residuals:
    Min      1Q Median      3Q     Max 
-483.45 -141.87 -19.62 103.58 1042.84 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  0.878e+03  2.928e+02 -6.412 4.22e-10 *** 
auto$cylinders -1.436e+01  2.038e+01 -0.704 0.48157    
auto$displacement 1.328e+00  4.738e-01  2.802 0.00534 **  
auto$horsepower -3.587e-01  8.693e-01 -0.413 0.68009    
auto$weight     -3.522e-01  4.111e-02 -8.567 2.62e-16 *** 
auto$acceleration 9.278e+00  6.232e+00  1.489 0.13740    
auto$model.year  4.081e+00  3.214e+00 12.698 < 2e-16 *** 
auto$origin      9.509e+01  1.754e+01  5.422 1.04e-07 *** 
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 209.8 on 384 degrees of freedom
Multiple R-squared:  0.7292, Adjusted R-squared:  0.7243 
F-statistic: 147.8 on 7 and 384 DF, p-value: < 2.2e-16

> 

```

Figure 5.6: Screenshot for ' X^2 '

6 P6

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2} \quad (6.1)$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

To calculate $\hat{\beta}_0$ and $\hat{\beta}_1$:

$$\bar{x} = \frac{1}{20} \sum_{i=1}^{20} x_i = \frac{1}{20} * 8.552 = 0.4276 \quad (6.2)$$

$$\bar{y} = \frac{1}{20} \sum_{i=1}^{20} y_i = \frac{1}{20} * 398.2 = 19.91$$

$$\begin{aligned} \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}) &= \sum_{i=1}^{20} (x_i y_i - \bar{x} y_i - x_i \bar{y} + \bar{x} \bar{y}) \\ &= \sum_{i=1}^{20} (x_i y_i) - \bar{x} \sum_{i=1}^{20} y_i - \bar{y} \sum_{i=1}^{20} x_i + 20 \bar{x} \bar{y} \\ &= 216.6 - 0.4276 * 398.2 - 19.91 * 8.552 + 20 * 0.4276 * 19.91 \\ &= 46.33 \end{aligned} \quad (6.3)$$

$$\begin{aligned} \sum_{i=1}^n (x_i - \bar{x})^2 &= \sum_{i=1}^{20} (x_i^2 - 2x_i \bar{x} + \bar{x}^2) \\ &= \sum_{i=1}^{20} x_i^2 - 2\bar{x} \sum_{i=1}^{20} x_i + 20 \bar{x}^2 \\ &= 5.196 - 2 * 0.4276 * 8.552 + 20 * 0.4276^2 \\ &= 1.54 \end{aligned}$$

Thus:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2} = \frac{46.33}{1.54} = 30.10 \quad (6.4)$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} = 19.91 - 30.10 * 0.4276 = 7.04$$

$$\begin{aligned} RSS &= \sum_{i=1}^n (y_i - \hat{y}_i)^2 = \sum_{i=1}^n (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2 = \sum_{i=1}^n (y_i - \bar{y} + \hat{\beta}_1 \bar{x} - \hat{\beta}_1 x_i)^2 \\ &= \sum_{i=1}^n (y_i - \bar{y})^2 - 2\hat{\beta}_1 \sum_{i=1}^n (y_i - \bar{y})(x_i - \bar{x}) + \hat{\beta}_1^2 \sum_{i=1}^n (x_i - \bar{x})^2 \\ &= \sum_{i=1}^{20} y_i^2 - 2\bar{y} \sum_{i=1}^{20} y_i + 20 \bar{y}^2 - 2\hat{\beta}_1 \sum_{i=1}^{20} (y_i - \bar{y})(x_i - \bar{x}) + \hat{\beta}_1^2 \sum_{i=1}^{20} (x_i - \bar{x})^2 \\ &= \sum_{i=1}^{20} y_i^2 - 2\bar{y} \sum_{i=1}^{20} y_i + 20 \bar{y}^2 - 2\hat{\beta}_1 \sum_{i=1}^{20} (y_i - \bar{y})(x_i - \bar{x}) + \hat{\beta}_1 \sum_{i=1}^{20} (y_i - \bar{y})(x_i - \bar{x}) \\ &= \sum_{i=1}^{20} y_i^2 - 2\bar{y} \sum_{i=1}^{20} y_i + 20 \bar{y}^2 - \hat{\beta}_1 \sum_{i=1}^{20} (y_i - \bar{y})(x_i - \bar{x}) \\ &= 9356 - 2 * 19.91 * 398.2 + 20 * 19.91^2 - 30.10 * 46.33 = 33.31 \end{aligned} \quad (6.5)$$

$$\hat{\sigma}^2 = \frac{RSS}{n-2} = 33.31/18 = 1.85 \quad (6.6)$$

$$R^2 = 1 - \frac{RSS}{\sum_{i=1}^n (y_i - \bar{y})^2} = 1 - \frac{33.31}{9356 - 2 * 19.91 * 398.2 + 20 * 19.91^2} = 0.9767$$

For $x = 0.5$:

$$\hat{y} = 7.04 + 30.10 * 0.5 = 22.09 \quad (6.7)$$

7 P7

Assume that $p = 6$, $n = 45$.

$$F = \frac{(TSS - RSS)/p}{RSS/(n-p-1)} = \frac{(11.62 - 8.95)/6}{8.95/(45-6-1)} = 1.89 \quad (7.1)$$

```
1 pf(1.89, 6, 38, lower.tail=F)
2 [1] 0.1078249
```

The p -value for null hypothesis is 0.1078249