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

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

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

1.3 c

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

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 $f(X_1 - \bar{X}, X_2 - \bar{X}, ..., X_n - \bar{X})$.

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

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

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

3.2 b

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

3.3 c

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

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

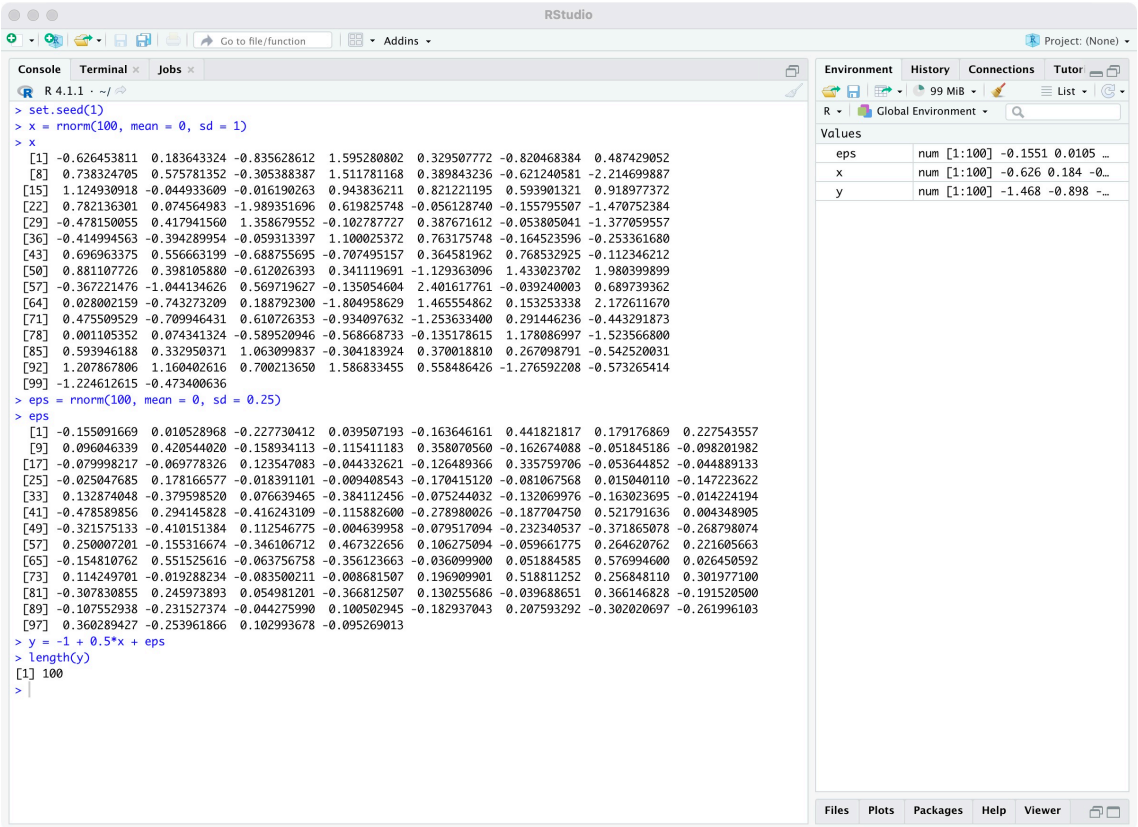

Figure 3.1: Screenshot for *'3-c'*

3.4 d

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

There is some linear relationship between Y and X.

3.5 e

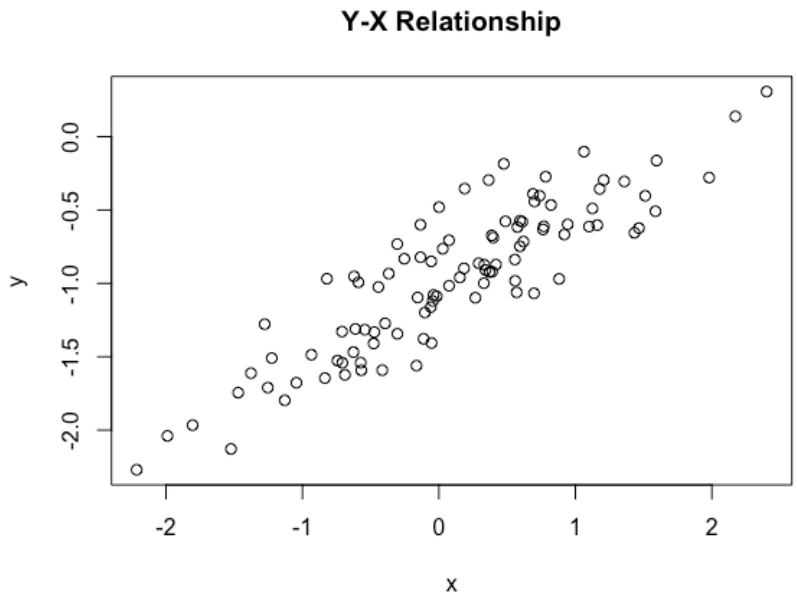

Figure 3.2: Screenshot for *'3-d'*

```
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 $\hat{\beta}_1, \hat{\beta}_0$, $Pr(\geqslant |t|) < 2e-16$, 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 their is no linear relationship between Y and X.

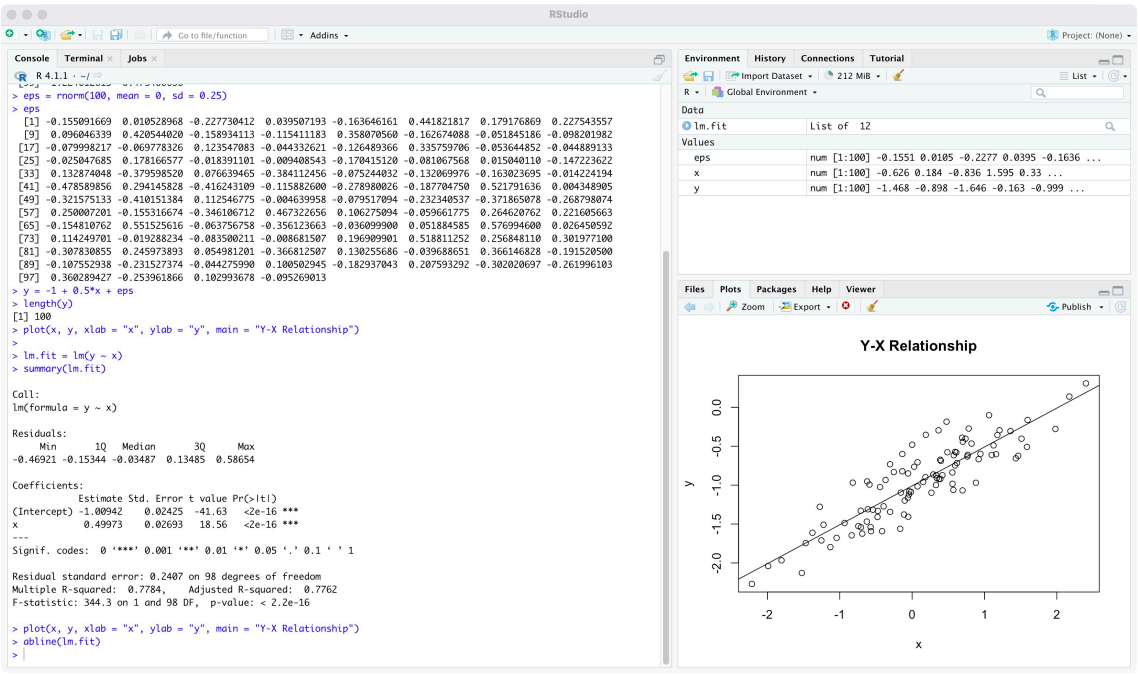

Figure 3.3: Screenshot for *'3-e'*

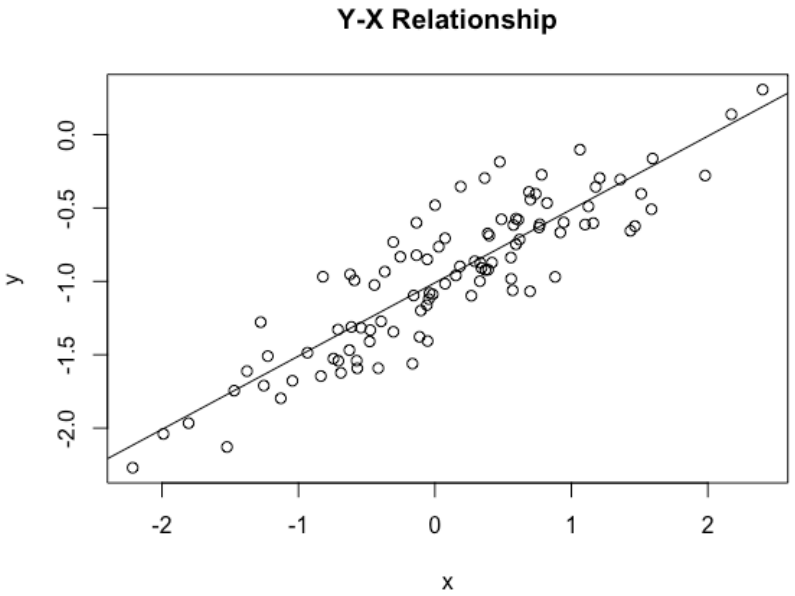

Figure 3.4: Screenshot for *'Least Squares Linear Model'*

3.6 f

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

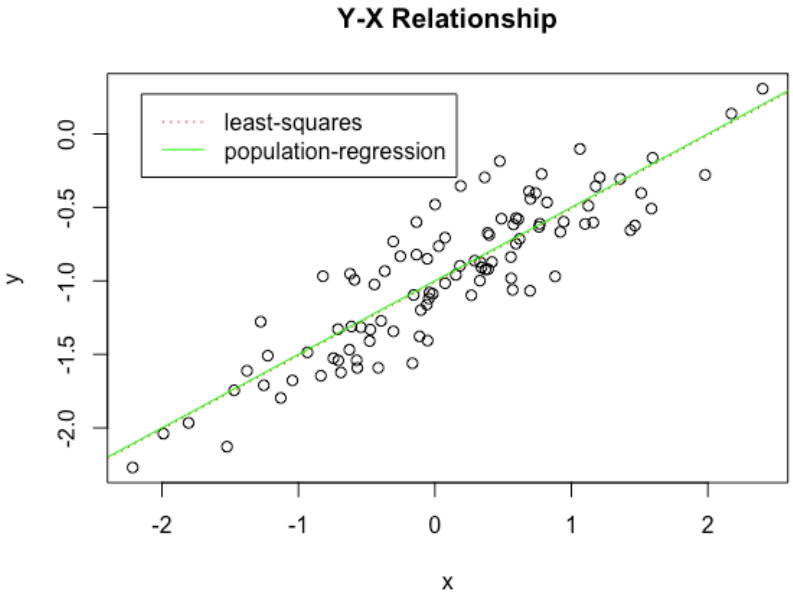

Figure 3.5: Screenshot for *'Least-Squares and Population-Regression model'*

3.7 g

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

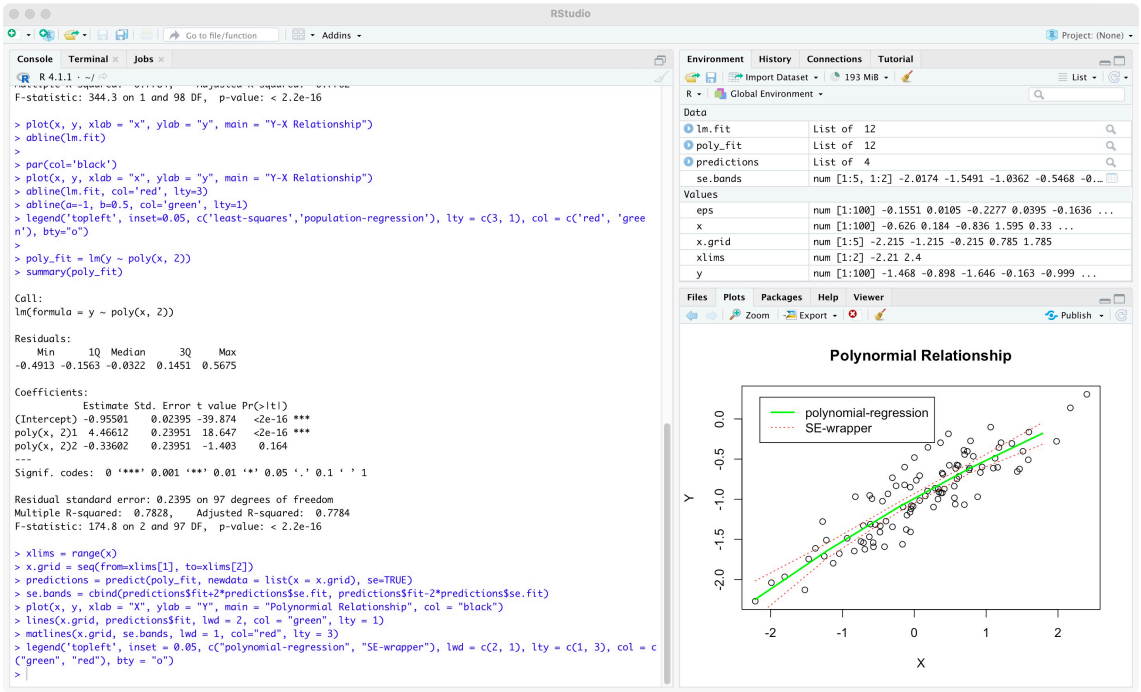

Figure 3.6: Screenshot for *'3-g'*

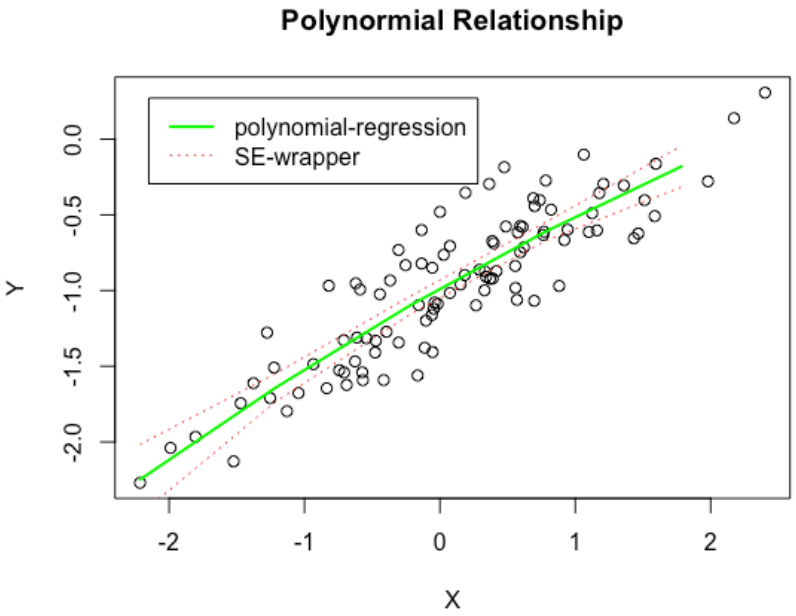

Figure 3.7: Screenshot for *'Polynomial Regression Model'*

3.8 h

```
set.seed(1)
x = rnorm(100, mean = 0, sd = 1)
eps = rnorm(100, mean = 0, sd = 0.01)
y = -1 + 0.5*x + eps
length(y)

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

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

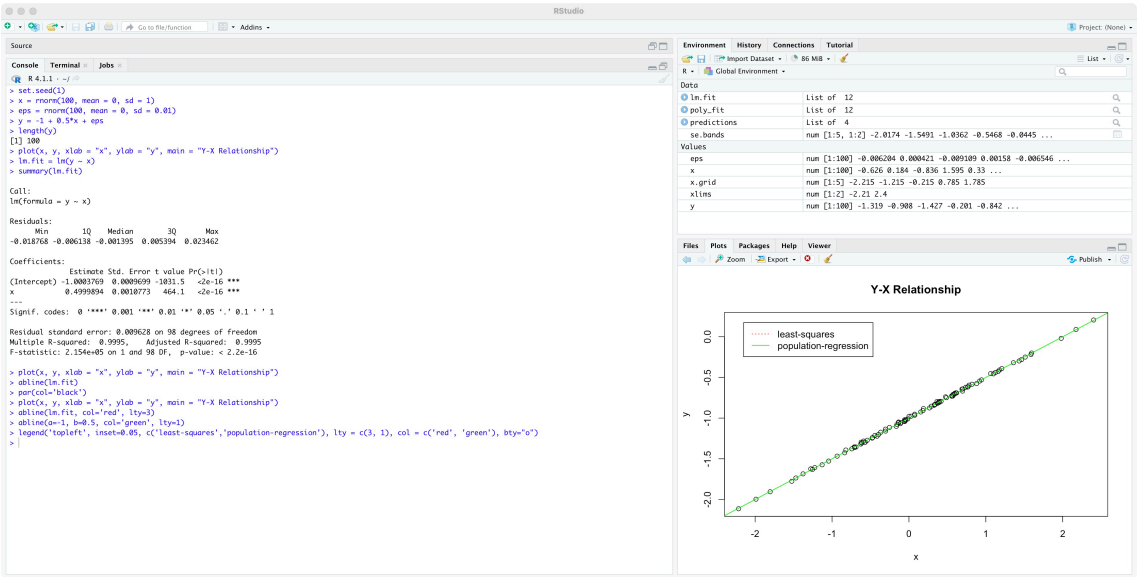

Figure 3.8: Screenshot for *'3-h'*

3.9 i

```
set.seed(1)
x = rnorm(100, mean = 0, sd = 1)
eps = rnorm(100, mean = 0, sd = 0.81)
y = -1 + 0.5*x + eps
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.

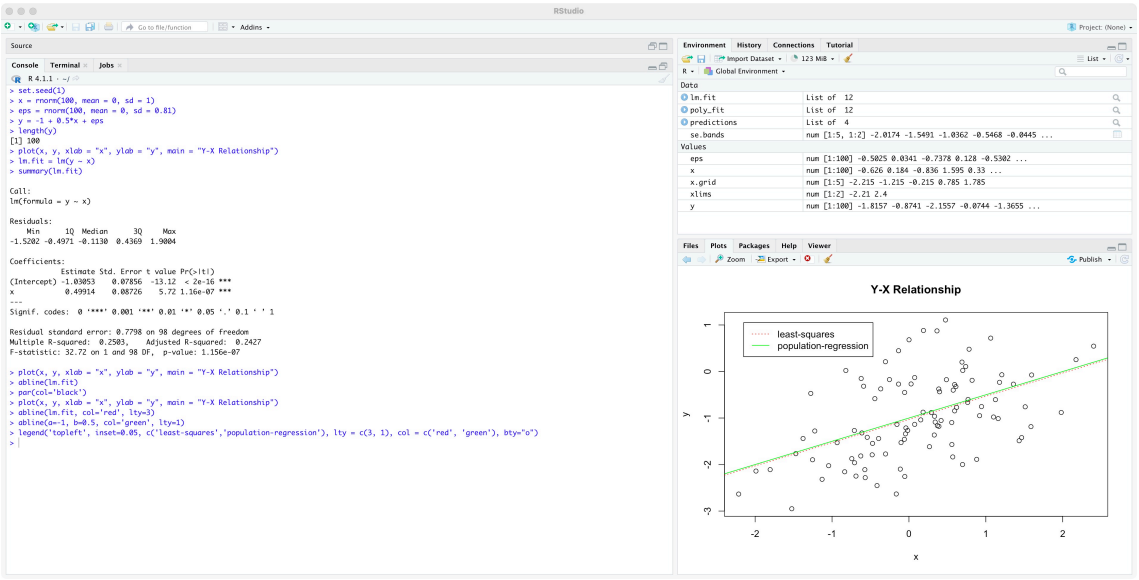

Figure 3.9: Screenshot for *'3-i'*

3.10 j

for sd = 0.25:

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

for sd = 0.01:

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

for sd = 1:

```
confint(lm.fit)
                    2.5 %     97.5 %
(Intercept) -1.2301607 -0.8452245
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:

```
adv = read.csv("advertising.csv", header = T, na.strings = '?')
adv = na.omit(adv)
TV_fit= lm(adv$Sales ~ adv$TV)
TV_confint = confint(TV_fit, level = 0.92)
plot(adv$TV, adv$Sales, xlab = "TV Advertising", ylab = "Sales",
     main="Sales-TV Advertising")
abline(TV_fit, col = 'green', lty = 1)
abline(a = TV_confint[1, 1], TV_confint[2, 1], col = 'red', lty = 3)
abline(a = TV_confint[1, 2], TV_confint[2, 2], col = 'red', lty = 3)
legend('topleft', inset = 0.05, c('least-squares',
       '0.92 condifence interval'), lty = c(1, 3), col = c('green', 'red')
       bty = 'o')
TV_confint
                    4 %       96 %
(Intercept) 6.40721844 7.54242454
adv$TV      0.05212914 0.05880041
```

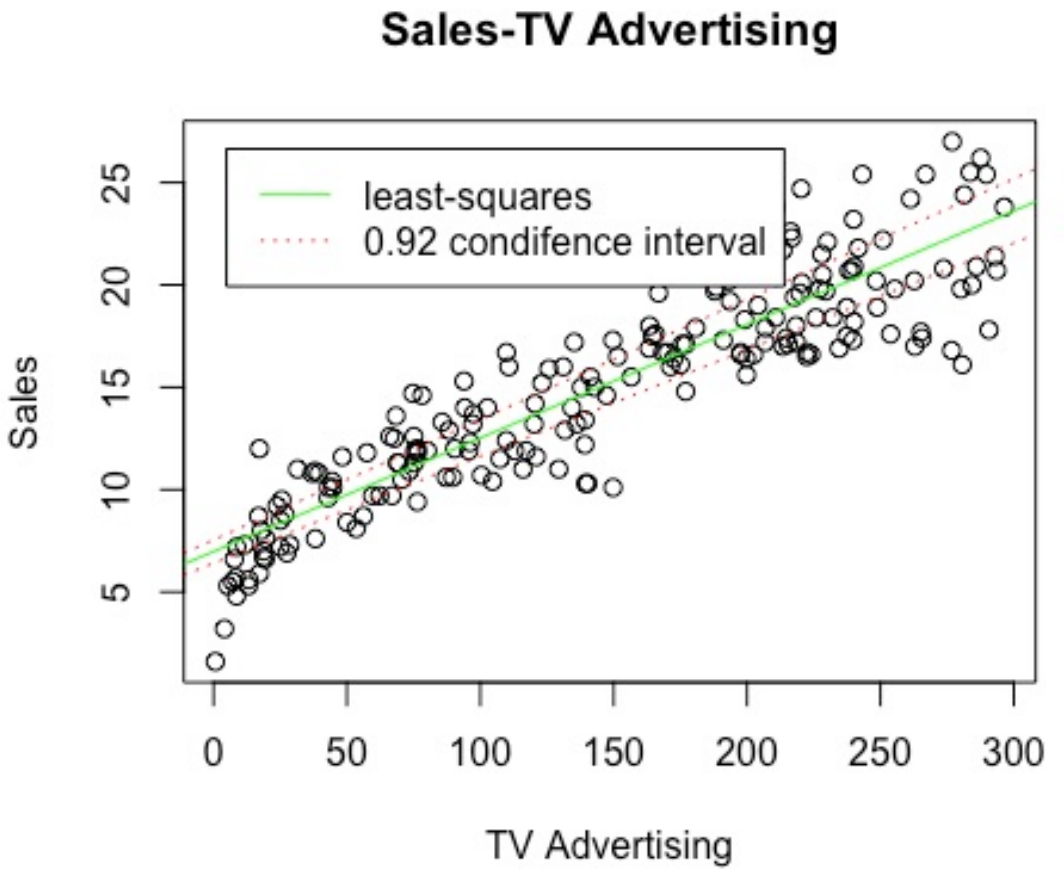

Figure 4.1: Screenshot for *'Sales-TV'*

For Radio:

```
adv = read.csv("advertising.csv", header = T, na.strings = '?')
adv = na.omit(adv)
Radio_fit= lm(adv$Sales ~ adv$Radio)
Radio_confint = confint(Radio_fit, level = 0.92)
plot(adv$Radio, adv$Sales, xlab = "Radio Advertising", ylab = "Sales",
     main="Sales-Radio Advertising")
abline(Radio_fit, col = 'green', lty = 1)
abline(a = Radio_confint[1, 1], Radio_confint[2, 1], col = 'red', lty = 3)
abline(a = Radio_confint[1, 2], Radio_confint[2, 2], col = 'red', lty = 3)
legend('topleft', inset = 0.05, c('least-squares',
       '0.92 condifence interval'), lty = c(1, 3), col = c('green', 'red')
       bty = 'o')
Radio_confint
                      4 %      96 %
(Intercept) 11.08577062 13.38567
adv$Radio    0.08273333  0.16613
```

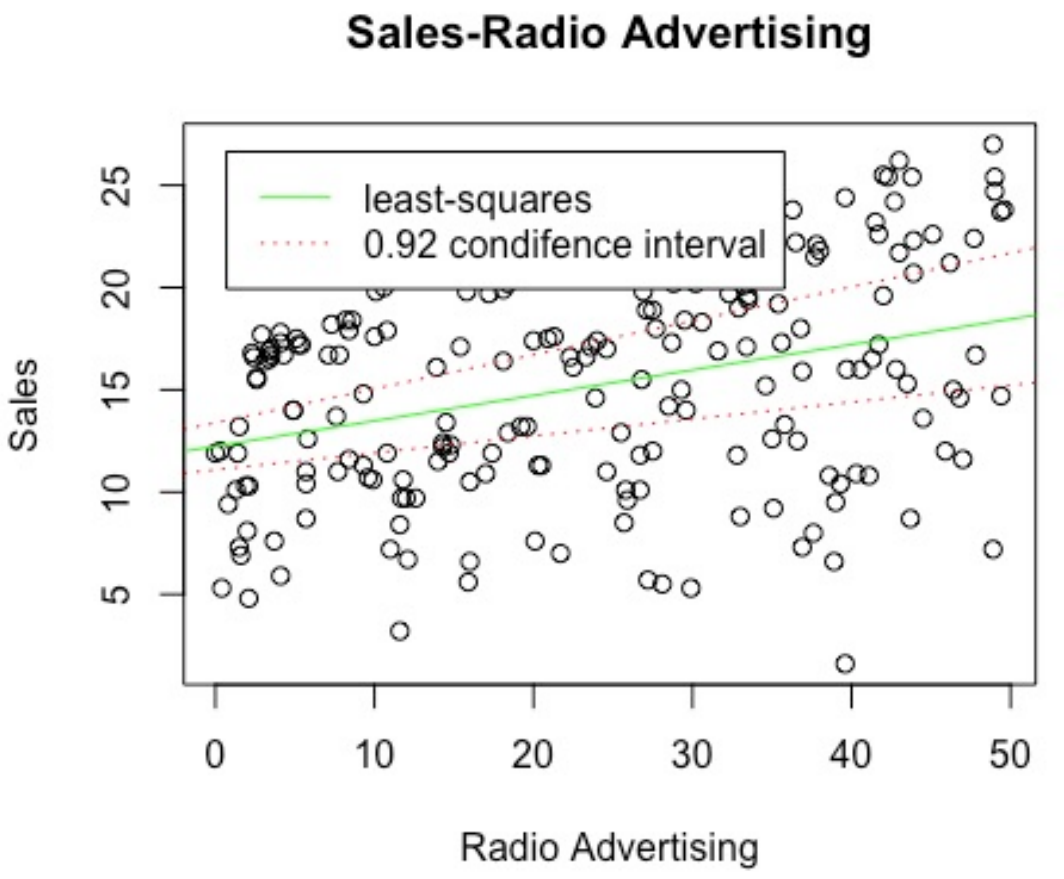

Figure 4.2: Screenshot for *'Sales-Radio'*

For Newspaper:

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

```
adv = na.omit(adv)
Newspaper_fit= lm(adv$Sales ~ adv$Newspaper)
Newspaper_confint = confint(Newspaper_fit, level = 0.92)
plot(adv$Newspaper, adv$Sales, xlab = "Newspaper␣Advertising",
     ylab = "Sales", main="Sales-Newspaper␣Advertising")
abline(Newspaper_fit, col = 'green', lty = 1)
abline(a = Newspaper_confint[1, 1], Newspaper_confint[2, 1], col = 'red',
abline(a = Newspaper_confint[1, 2], Newspaper_confint[2, 2], col = 'red',
legend('topleft', inset = 0.05, c('least-squares',
       '0.92␣condifence␣interval'), lty = c(1, 3), col = c('green', 'red')
       bty = 'o')
Newspaper_confint
                        4 %         96 %
(Intercept)    12.83634111  15.08275619
adv$Newspaper   0.00836363   0.06828436
```

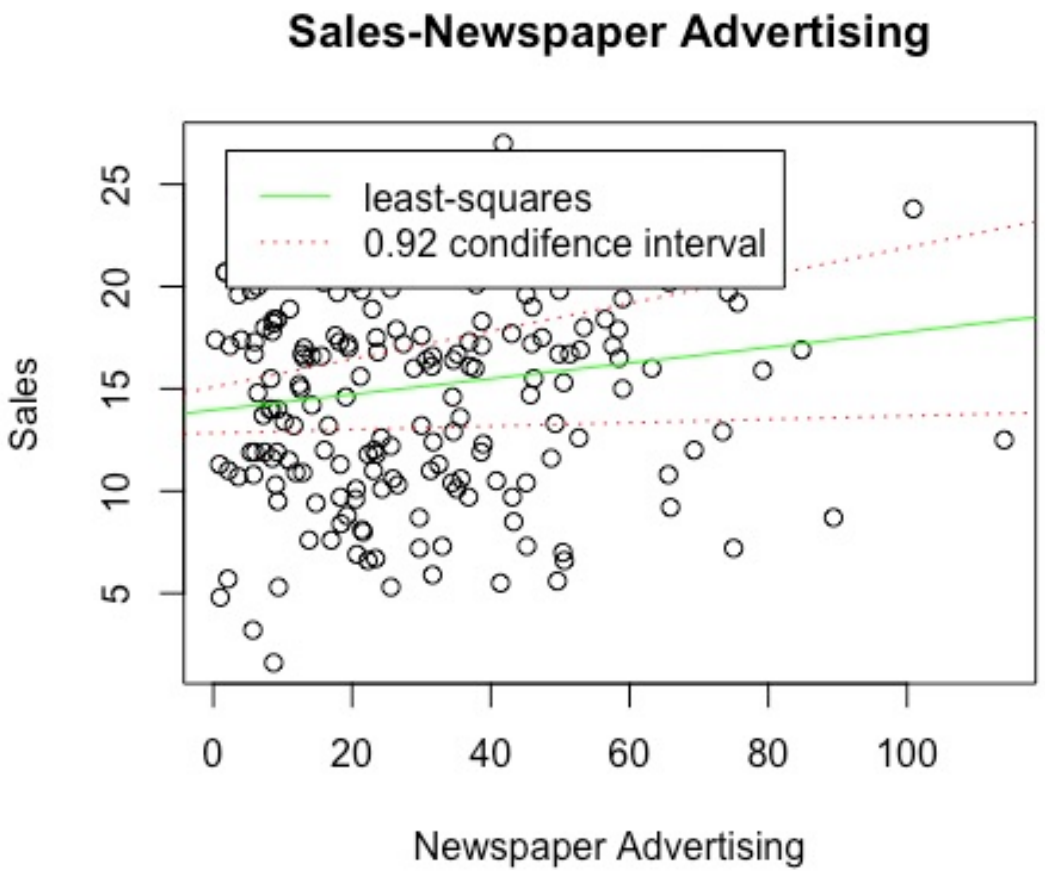

Figure 4.3: Screenshot for *'Sales-Newspaper'*

5 P5

5.1 a

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

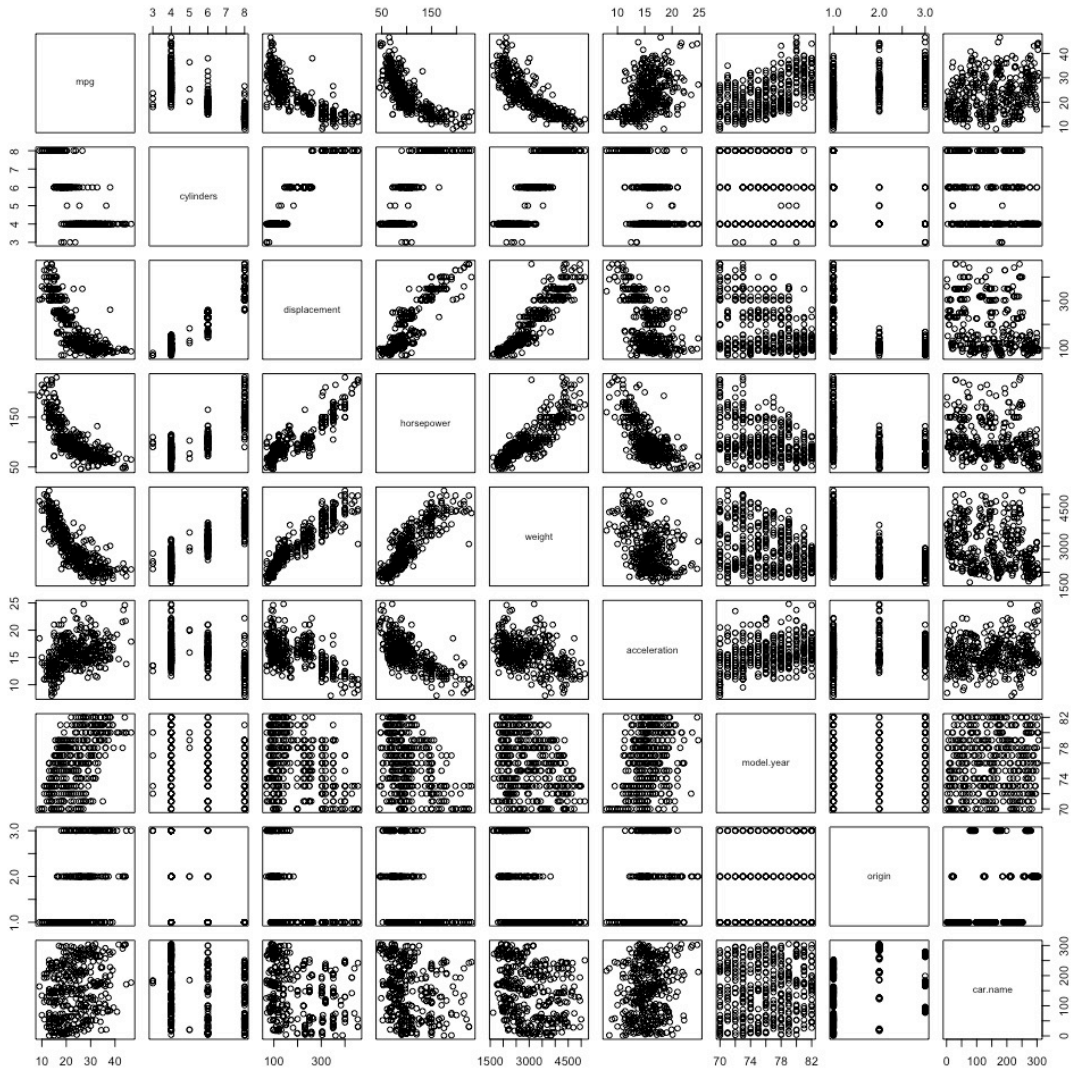

Figure 5.1: Screenshot for *'Pairs'*

5.2 b

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

5.3 c

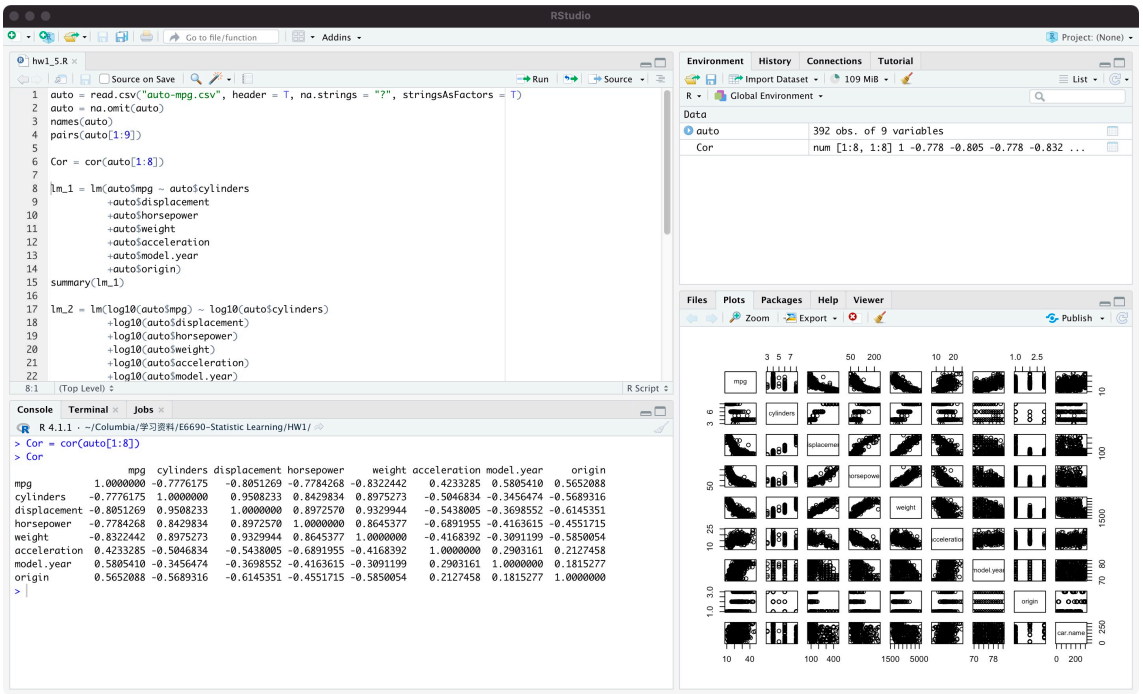

Figure 5.2: Screenshot for *'Cor'*

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

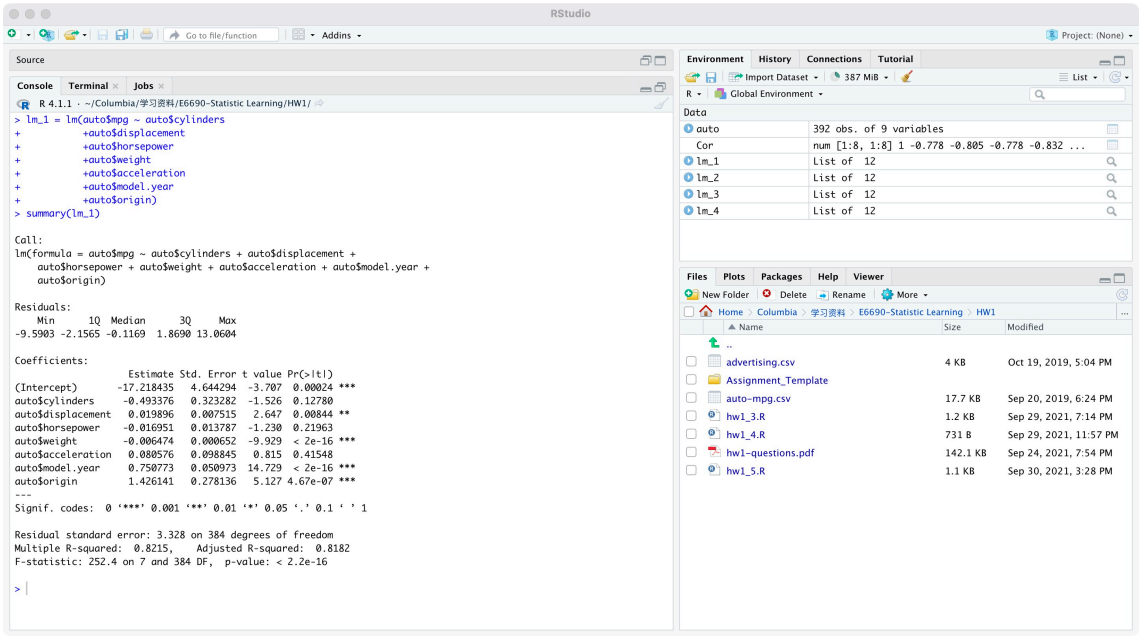

Figure 5.3: Screenshot for *'X'*

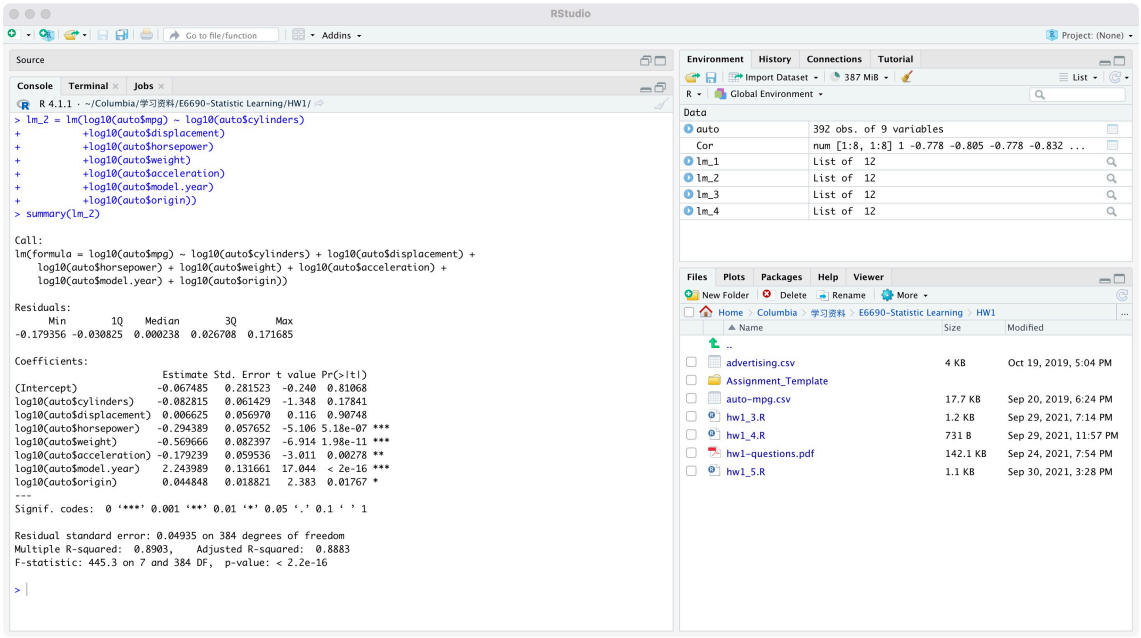

Figure 5.4: Screenshot for *'$logX$'*

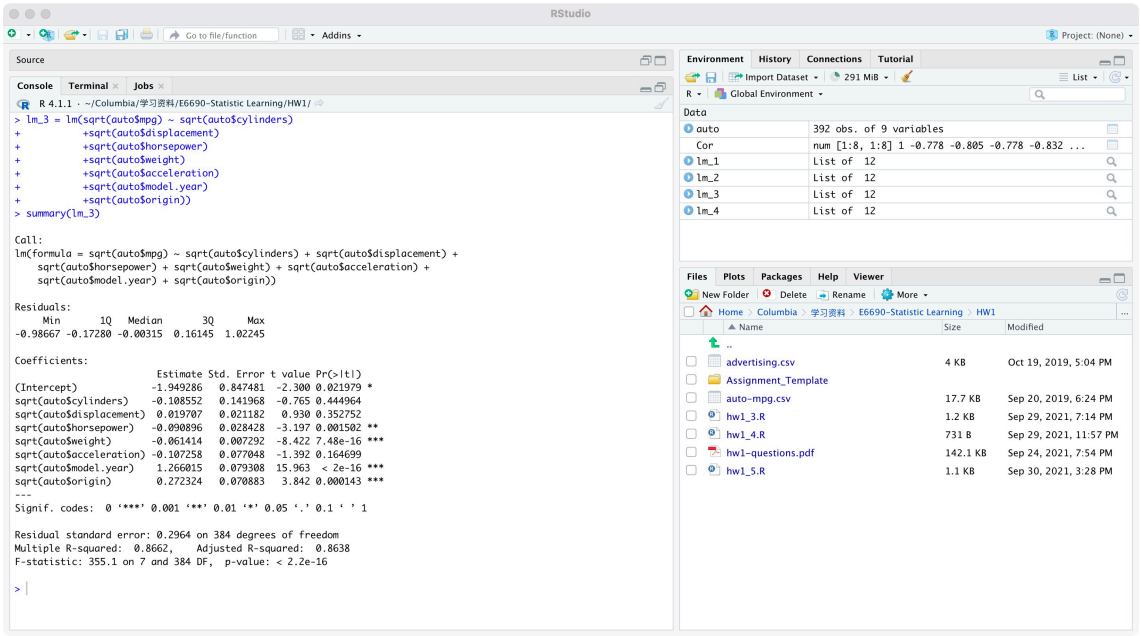

Figure 5.5: Screenshot for '$\sqrt{X}$'

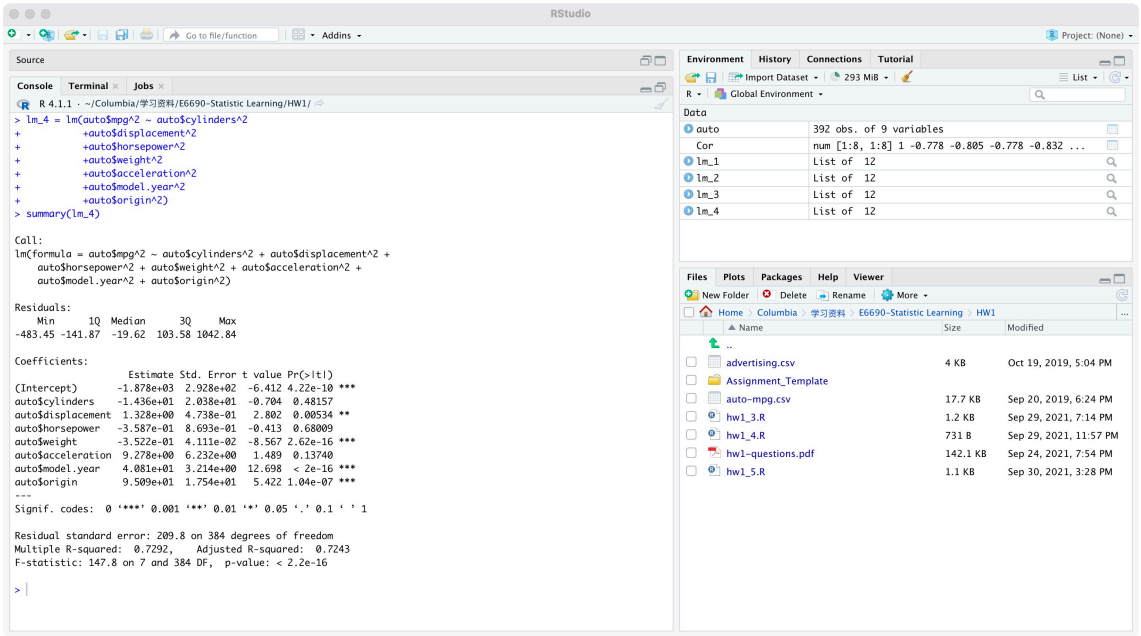

Figure 5.6: Screenshot for 'X^2'

6 P6

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sum_{i=1}^{n}(x_i-\bar{x})^2} \tag{6.1}$$
$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1\bar{x}$$

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

$$\begin{aligned}\bar{x} &= \frac{1}{20}\sum_{i=1}^{20} x_i = \frac{1}{20} * 8.552 = 0.4276\\ \bar{y} &= \frac{1}{20}\sum_{i=1}^{20} y_i = \frac{1}{20} * 398.2 = 19.91\end{aligned} \tag{6.2}$$

$$\begin{aligned}\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y}) &= \sum_{i=1}^{20}(x_i y_i - \bar{x}y_i - x_i\bar{y} + \bar{x}\bar{y})\\ &= \sum_{i=1}^{20}(x_i y_i) - \bar{x}\sum_{i=1}^{20} y_i - \bar{y}\sum_{i=1}^{20} x_i + 20\bar{x}\bar{y}\\ &= 216.6 - 0.4276 * 398.2 - 19.91 * 8.552 + 20 * 0.4276 * 19.91\\ &= 46.33\\ \sum_{i=1}^{n}(x_i-\bar{x})^2 &= \sum_{i=1}^{20}(x_i^2 - 2x_i\bar{x} + \bar{x}^2)\\ &= \sum_{i=1}^{20} x_i^2 - 2\bar{x}\sum_{i=1}^{20} x_i + 20\bar{x}^2\\ &= 5.196 - 2 * 0.4276 * 8.552 + 20 * 0.4276^2\\ &= 1.54\end{aligned} \tag{6.3}$$

Thus:

$$\begin{aligned}\hat{\beta}_1 &= \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sum_{i=1}^{n}(x_i-\bar{x})^2} = \frac{46.33}{1.54} = 30.10\\ \hat{\beta}_0 &= \bar{y} - \hat{\beta}_1\bar{x} = 19.91 - 30.10 * 0.4276 = 7.04\end{aligned} \tag{6.4}$$

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

$$\begin{aligned}\hat{\sigma}^2 &= \frac{RSS}{n-2} = 33.31/18 = 1.85\\ R^2 &= 1 - \frac{RSS}{\sum_{i=1}^{n}(y_i-\bar{y})^2} = 1 - \frac{33.31}{9356 - 2 * 19.91 * 398.2 + 20 * 19.91^2} = 0.9767\end{aligned} \tag{6.6}$$

For $x = 0.5$:

$$\hat{y} = 7.04 + 30.10 * 0.5 = 22.09 \tag{6.7}$$

7 P7

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

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

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

The p-value for null hypothesis is 0.1078249