Homework 1

E6690: Statistical Learning for Bio & Info Systems

P1. Let

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \quad \text{and} \quad S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2.$$

Show that:

(a) (2pt) $\sum_{i=1}^{n} X_i^2 = (n-1)S^2 + n\bar{X}^2$

Answer: By definition of S^2, we have

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

The desired result follows.

(b) (2pt) If $X_1, X_2, ..., X_n$ are independent and identically distributed (i.i.d.), the S^2 is an unbiased estimator of σ^2, i.e., $\mathbb{E}S^2 = \sigma^2$

Answer: Follows easily from (a):

$$\begin{aligned}(n-1)\mathbb{E}S^2 &= n\mathbb{E}X_1^2 - n\mathbb{E}\bar{X}^2 \\ &= n\mathbb{E}X_1^2 - n\frac{\mathbb{E}(\sum X_i)^2}{n^2} \\ &= n\mathbb{E}X_1^2 - \frac{1}{n}(n\mathbb{E}X_1^2 - n(n-1)\mathbb{E}X_1) \\ &= (n-1)\sigma^2.\end{aligned}$$

In the following, in addition to the above, assume that X_i-s have normal/Gaussian distribution $\mathcal{N}(\mu, \sigma^2)$.

(c) (3pt) Prove that $\bar{X}$ is independent of $X_i - \bar{X}$, $i = 1, 2, \ldots, n$.
(Hint: Both $\bar{X}$ and $X_i - \bar{X}$ are normal.)

Answer: Recall that $X_1, X_2, ..., X_n \sim \mathcal{N}(\mu, \sigma^2)$ and independent. Since $\bar{X}$ and $X_i - \bar{X}$ are linear combinations of X_i, they are jointly normal, and therefore, for their independence it is enough to show that $\bar{X}$ and $X_i - \bar{X}$ are uncorrelated:

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \sim \mathcal{N}(\mu, \sigma^2/n)$$

and

$$X_i - \bar{X} \sim \mathcal{N}(0, \left(\frac{n-1}{n}\right)\sigma^2).$$

Then,

$$\begin{aligned}\mathsf{Cov}(\bar{X}, X_i - \bar{X}) &= \mathsf{Cov}(\bar{X}, X_i) - \mathsf{Var}(\bar{X}) \\ &= \frac{1}{n}\mathsf{Cov}\left(X_i + \sum_{i \neq j} X_j, X_i\right) - \frac{\sigma^2}{n} \\ &= \frac{1}{n}\mathsf{Var}(X_i) - \frac{\sigma^2}{n} = 0.\end{aligned}$$

Implying that $\bar{X}$ is independent of $X_i - \bar{X}$, $i = 1, 2, \ldots, n$.

(d) (3pt) Show (prove) that the sample mean, $\bar{X}$, is independent of the sample variance, S^2.

Answer: According to (a), S^2 is a function of $X_i - \bar{X}$, $i = 1, 2, ..., n$. We proved in (b) that $\bar{X}$ is independent of $X_i - \bar{X}$. Therefore it can be concluded that S^2 is independent of $\bar{X}$.

P2. (10pt) Show (prove) that in the case of simple linear regression between Y and X, the R^2 statistic is equal to the square of the correlation coefficient between X and Y (r^2). For simplicity, you may assume that $\bar{y} = \bar{x} = 0$. Recall that

$$R^2 = \frac{\sum_{i=1}^n (\hat{y}_i - \bar{y})^2}{\sum_{i=1}^n (y_i - \bar{y})^2} \quad \text{and} \quad r = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2 \sum_{i=1}^n (y_i - \bar{y})^2}}.$$

Answer: For a linear regression, $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$, where

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2} = \frac{\sum_{i=1}^n x_i y_i}{\sum_{i=1}^n x_i^2} \quad \text{and} \quad \hat{\beta}_0 = \frac{\bar{y}\sum_{i=1}^n x_i^2 - \bar{x}\sum_{i=1}^n x_i y_i}{\sum_{i=1}^n x_i^2 - n\bar{x}^2} = 0.$$

Now,

$$\begin{aligned}\sum_{i=1}^n (\hat{y}_i - \bar{y})^2 &= \sum_{i=1}^n (\hat{\beta}_1 x_i)^2 \\ &= \left(\frac{\sum_{i=1}^n x_i y_i}{\sum_{i=1}^n x_i^2}\right)^2 \cdot \sum_{i=1}^n x_i^2 \\ &= \frac{\left(\sum_{i=1}^n x_i y_i\right)^2}{\sum_{i=1}^n x_i^2}.\end{aligned}$$

Finally,

$$R^2 = \frac{\sum_{i=1}^n (\hat{y}_i - \bar{y}_i)^2}{\sum_{i=1}^n (y_i - \bar{y})^2} = \frac{\frac{\left(\sum_{i=1}^n x_i y_i\right)^2}{\sum_{i=1}^n x_i^2}}{\sum_{i=1}^n y_i^2} = \left(\frac{\sum_{i=1}^n x_i y_i}{\sqrt{\sum_{i=1}^n x_i^2 \sum_{i=1}^n y_i^2}}\right)^2 = r^2.$$

P3. (20pt; each bullet 2pt) Create some simulated data and fit simple linear regression models to it. Make sure to use `set.seed(1)` prior to starting part (a) to ensure consistent results.

(a) Using the `rnorm()` function, create a vector, `x`, containing 100 observations drawn from a $\mathcal{N}(0,1)$ distribution. This represents a feature, X.

Answer:

```
> set.seed(1)
> x <- rnorm(100)
```

(b) Using the `rnorm()` function, create a vector, `eps`, containing 100 observations drawn from a $\mathcal{N}(0,0.25)$ distribution.

Answer:

```
> eps <- rnorm(100, 0, sqrt(0.25))
```

(c) Using `x` and `eps`, generate a vector `y` according to the model

$$Y = -1 + 0.5X + \epsilon.$$

What is the length of the vector `y`? What are the values of β_0 and β_1 in this linear model?

Answer:

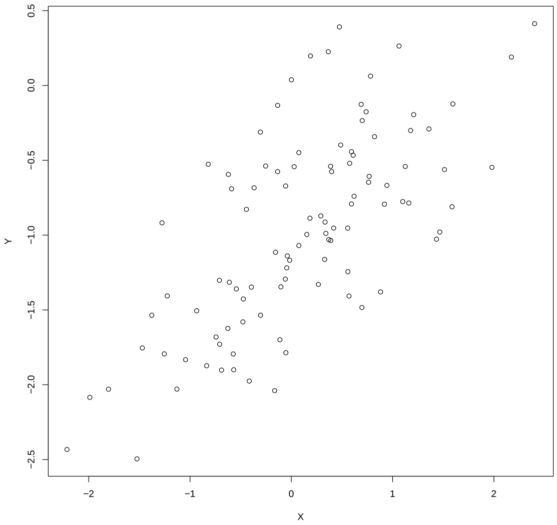
```
> y <- -1 + 0.5*x + eps
> length(y)

## 100
```

The values of β_0 and β_1 are -1 and 0.5, respectively.

(d) Create a scatterplot displaying the relationship between `x` and `y`. Comment on what you observe.

Answer:

```
> pdf(file="4-d.pdf",width = 10,height = 10)
> plot(x, y,xlab = "X",ylab = "Y")
> dev.off()
```



It appears that there is a linear relationship between X and Y.

(e) Fit a least squares linear model to predict `y` using `x`. Comment on the model obtained. How do $\hat{\beta}_0$ and $\hat{\beta}_1$ compare to β_0 and β_1?

Answer:

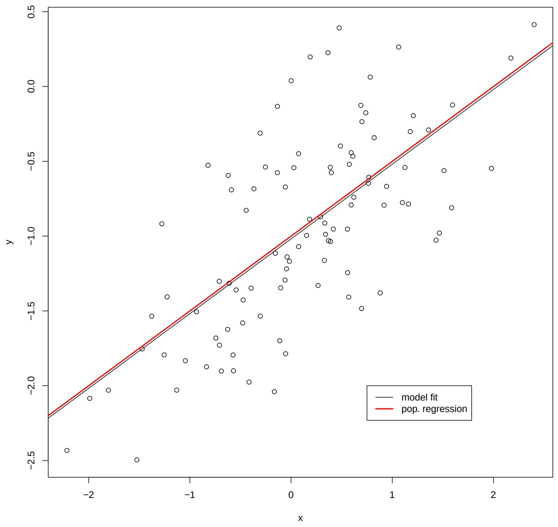
```
> lm.fit = lm(y~x)
> summary(lm.fit)
##Call:
##lm(formula = y ~ x)
##
##Residuals:
##     Min       1Q   Median       3Q      Max
##-0.93842 -0.30688 -0.06975  0.26970  1.17309
##
##Coefficients:
##            Estimate Std. Error t value Pr(>|t|)
##(Intercept) -1.01885    0.04849 -21.010  < 2e-16 ***
##x            0.49947    0.05386   9.273 4.58e-15 ***
##---
##Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
##
##Residual standard error: 0.4814 on 98 degrees of freedom
##Multiple R-squared:  0.4674,Adjusted R-squared:  0.4619
##F-statistic: 85.99 on 1 and 98 DF,  p-value: 4.583e-15
```

We can reject null hypothesis since the p-value is close to zero and the F-statistics is large. In this case, coefficients are very close to what we had at the beginning as β_0 and β_1. Now we have $\hat{\beta}_0 = -1.01885$ and $\hat{\beta}_1 = 0.49947$. Hence, our estimates are very close to the true values of $\beta_0 = -1$ and $\beta_1 = 0.5$.

(f) Display the least squares line on the scatterplot obtained in (d). Draw the population regression line on the plot, in a different color. Use the `legend()` command to create an appropriate legend.

Answer:

```
> pdf(file="4-f.pdf",width = 10,height = 10)
> plot(x, y)
> abline(lm.fit, lwd=1, col=1)
> abline(-1, 0.5, lwd=2, col=2)
> legend(0.75,-2, legend = c("model fit", "pop. regression"),
> col=1:2,lwd=1:2, cex=1)
> dev.off()
```



(g) Now fit a polynomial regression model that predicts y using x and x^2. Is there evidence that the quadratic term improves the model fit? Explain your answer.

Answer:

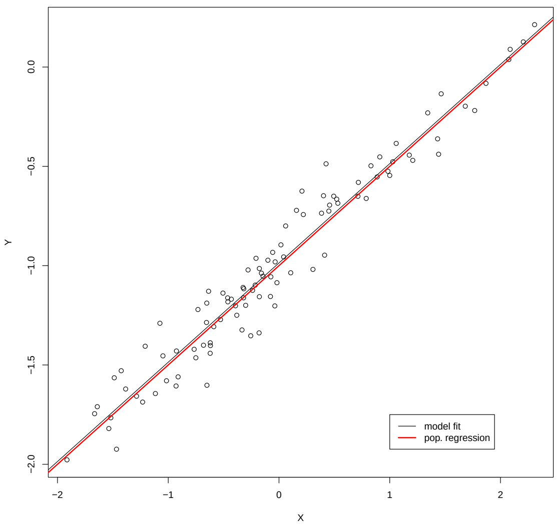
```
> lm.fit_sq = lm(y~x+I(x^2))
> summary(lm.fit_sq)

##Call:
##lm(formula = y ~ x + I(x^2))
##
##Residuals:
##     Min       1Q   Median       3Q      Max
##-0.98252 -0.31270 -0.06441  0.29014  1.13500
##
##Coefficients:
##            Estimate Std. Error t value Pr(>|t|)
##(Intercept) -0.97164    0.05883 -16.517  < 2e-16 ***
##x            0.50858    0.05399   9.420  2.4e-15 ***
##I(x^2)      -0.05946    0.04238  -1.403    0.164
##---
##Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
##
##Residual standard error: 0.479 on 97 degrees of freedom
##Multiple R-squared:  0.4779,Adjusted R-squared:  0.4672
##F-statistic:  44.4 on 2 and 97 DF,  p-value: 2.038e-14
```

According to p-value of t-statistics, there is not a relationship between y and x^2. We also have a slight increase in R^2 and decrease in RSE.

(h) Repeat (a)-(f) after modifying the data generation process in such a way that there is less noise in the data. The model in (c) should remain the same. You can do this by decreasing the variance of the normal distribution used to generate the error term ϵ in (b). Describe your results.

Answer:

```
> set.seed(1)
> x1 = rnorm(100)
> eps1 = rnorm(100, 0, 0.125)
> y1 = -1 + 0.5*x1 + eps1
> lm.fit1 = lm(y1~x1)
> summary(lm.fit1)

##Call:
##lm(formula = y1 ~ x1)
##
##Residuals:
##     Min       1Q   Median       3Q      Max
##-0.23461 -0.07672 -0.01744  0.06742  0.29327
##
##Coefficients:
##            Estimate Std. Error t value Pr(>|t|)
##(Intercept) -1.00471    0.01212  -82.87   <2e-16 ***
##x1           0.49987    0.01347   37.12   <2e-16 ***
##---
##Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
##
##Residual standard error: 0.1203 on 98 degrees of freedom
##Multiple R-squared:  0.9336,Adjusted R-squared:  0.9329
##F-statistic:  1378 on 1 and 98 DF,  p-value: < 2.2e-16

> pdf(file="4-h.pdf",width = 10,height = 10)
> plot(x1, y1,xlab="X",ylab = "Y")
> abline(lm.fit1, lwd=1, col=1)
> abline(-1, 0.5, lwd=2, col=2)
> legend(1,-1.75, legend = c("model fit", "pop. regression"), col=1:2,lwd=1:2, cex=1)
> dev.off()
```



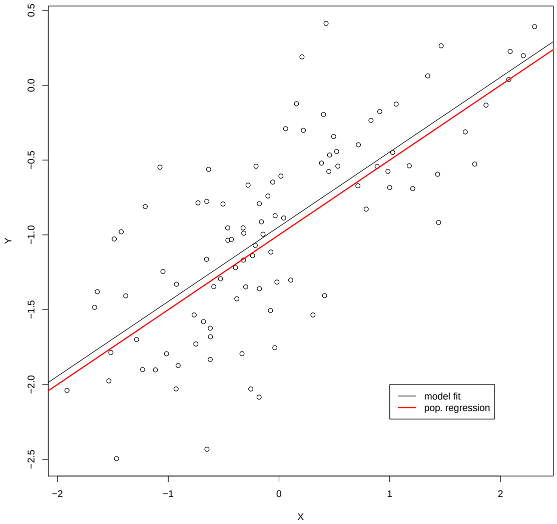
The error in R^2 and RSE decreased.

(i) Repeat $(a) - (f)$ after modifying the data generation process in such a way that there is more noise in the data. The model in (c) should remain the same. You can do this by increasing the variance of the normal distribution used to generate the error term ϵ in (b). Describe your results.

Answer:

```
> set.seed(1)
> x2 = rnorm(100)
> eps2 = rnorm(100, 0, 1.)
> y2 = -1 + 0.5*x2 + eps2
> lm.fit2 = lm(y2~x2)
> summary(lm.fit2)

##Call:
##lm(formula = y2 ~ x2)
##
##Residuals:
##    Min      1Q  Median      3Q     Max
##-1.8768 -0.6138 -0.1395  0.5394  2.3462
##
##Coefficients:
##            Estimate Std. Error t value Pr(>|t|)
##(Intercept) -1.03769    0.09699 -10.699  < 2e-16 ***
##x2           0.49894    0.10773   4.632 1.12e-05 ***
##---
##Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
##
##Residual standard error: 0.9628 on 98 degrees of freedom
##Multiple R-squared:  0.1796,Adjusted R-squared:  0.1712
##F-statistic: 21.45 on 1 and 98 DF,  p-value: 1.117e-05
```

```
> pdf(file="4-i.pdf",width = 10,height = 10)
> plot(x2, y2,xlab="X",ylab = "Y")
> abline(lm.fit2, lwd=1, col=1)
> abline(-1, 0.5, lwd=2, col=2)
> legend(1,-2.25, legend = c("model fit", "pop. regression"), col=1:2,lty=1:2, cex=0.7)
> dev.off()
```



The error in R^2 and RSE increased.

(j) What are the confidence intervals for β_0 and β_1 based on the original data set, the noisier data set, and the less noisy data set? Comment on your results.

Answer:

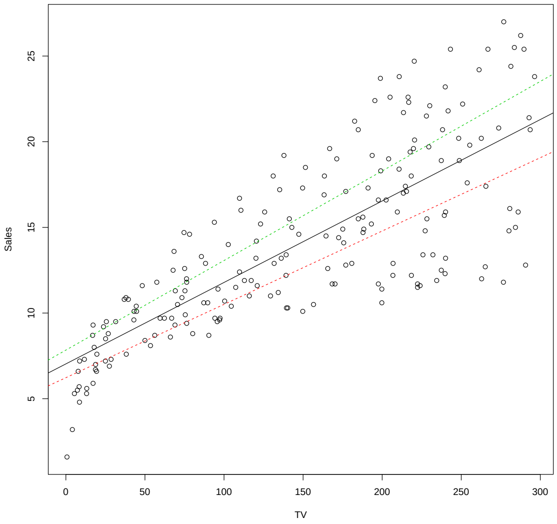
```
> confint(lm.fit)
##                  2.5 %     97.5 %
##(Intercept) -1.1150804 -0.9226122
##x            0.3925794  0.6063602

> confint(lm.fit1)
##                  2.5 %     97.5 %
##(Intercept) -1.0287701 -0.9806531
##x1           0.4731449  0.5265901

> confint(lm.fit2)
##                  2.5 %     97.5 %
##(Intercept) -1.2301607 -0.8452245
##x2           0.2851588  0.7127204
```

All the intervals are close to β_0 and β_1. But the second one is the narrowest and the third one is the widest one.

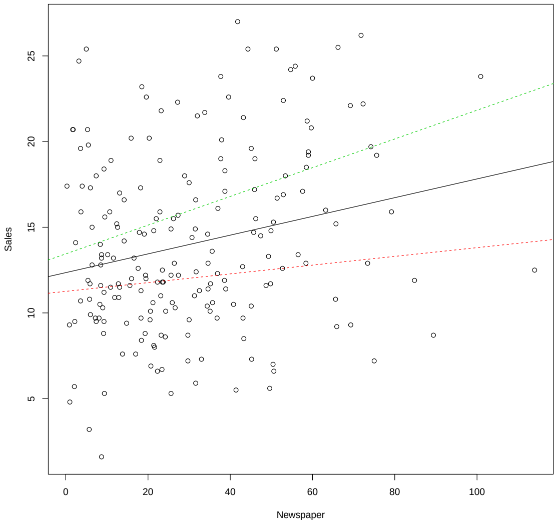
P4. (10pt) Using R and Advertising data set, find 92% confidence intervals for β_0 and β_1 for three single-feature linear regressions of Sales versus Newspaper, TV and Radio, respectively. Then, create a scatterplot for each of them with the 92% confidence interval lines, i.e., draw the lines that correspond to the ends of confidence intervals for (β_0, β_1). The answer should include the R code and graphs.

Answer:

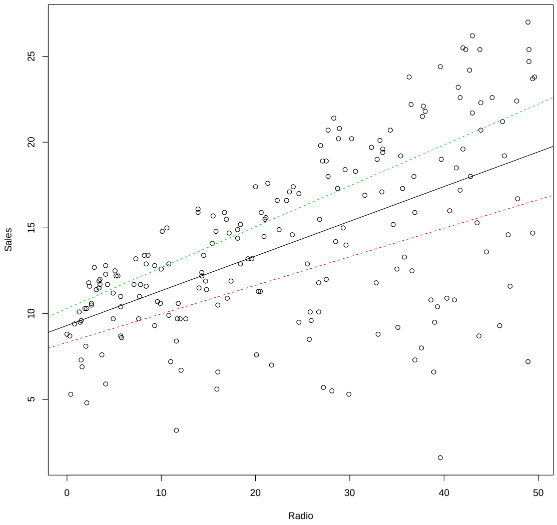
```
> #For TV
> lm.fitTV<-lm(advertising$Sales~advertising$TV)
> summary(lm.fitTV)
> pdf(file="tv.pdf",width = 10,height = 10)
> plot(advertising$TV,advertising$Sale,xlab = "TV",ylab = "Sales")
> abline(lm.fitTV)
> conf_TV<-confint(lm.fitTV,level = 0.92)
> abline(coef=conf_TV[,1],lty=2,col=2)
> abline(coef=conf_TV[,2],lty=2,col=3)
> dev.off()
```



```
> #For Newspaper
> lm.fitNewspaper<-lm(advertising$Sales~advertising$Newspaper)
> summary(lm.fitNewspaper)
> pdf(file="Newspaper.pdf",width = 10,height = 10)
> plot(advertising$Newspaper,advertising$Sale,xlab = "Newspaper",ylab = "Sales")
> abline(lm.fitNewspaper)
> conf_TV<-confint(lm.fitNewspaper,level = 0.92)
> abline(coef=conf_TV[,1],lty=2,col=2)
> abline(coef=conf_TV[,2],lty=2,col=3)
> dev.off()
```



```
> #For Radio
> lm.fitRadio<-lm(advertising$Sales~advertising$Radio)
> summary(lm.fitRadio)
> pdf(file="Radio.pdf",width = 10,height = 10)
> plot(advertising$Radio,advertising$Sale,xlab = "Radio",ylab = "Sales")
> abline(lm.fitRadio)
> conf_TV<-confint(lm.fitRadio,level = 0.92)
> abline(coef=conf_TV[,1],lty=2,col=2)
> abline(coef=conf_TV[,2],lty=2,col=3)
> dev.off()
```

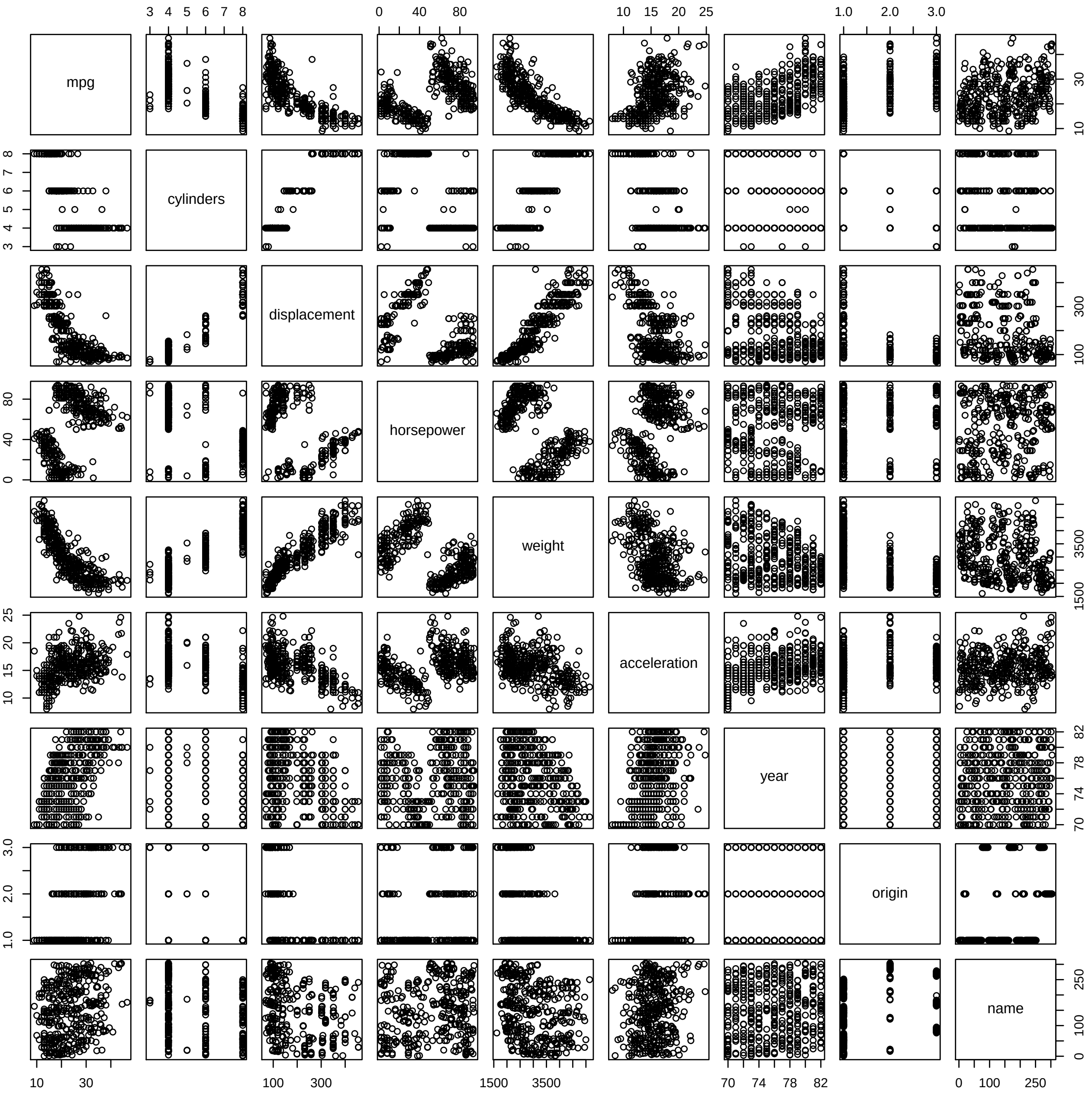


P5. Consider the `Auto` data set:

(a) (5pt) Produce a scatterplot matrix which includes all of the variables in the data set.

Answer:

```
> Auto<-read.csv(file.choose(),header=T)
> sapply(Auto,class)
> Auto<-Auto[!Auto$horsepower=="?",]
> Auto$horsepower<-as.numeric(Auto$horsepower)
> pdf(file="5-a.pdf",width = 10,height = 10)
> pairs(Auto)
> dev.off()
```



(b) (5pt) Compute the matrix of correlations between the variables using the function `cor()`. You will need to exclude the `name` variable, which is qualitative.

Answer:

```
> correlation_b<-cor(Auto[sapply(Auto, is.numeric)])

##                    mpg  cylinders displacement horsepower     weight acceleration       year     origin
##mpg           1.0000000 -0.7776175   -0.8051269 -0.7784268 -0.8322442    0.4233285  0.5805410  0.5652088
##cylinders    -0.7776175  1.0000000    0.9508233  0.8429834  0.8975273   -0.5046834 -0.3456474 -0.5689316
##displacement -0.8051269  0.9508233    1.0000000  0.8972570  0.9329944   -0.5438005 -0.3698552 -0.6145351
```

```
##horsepower   -0.7784268  0.8429834    0.8972570  1.0000000  0.8645377   -0.6891955 -0.4163615 -0.4551715
##weight       -0.8322442  0.8975273    0.9329944  0.8645377  1.0000000   -0.4168392 -0.3091199 -0.5850054
##acceleration  0.4233285 -0.5046834   -0.5438005 -0.6891955 -0.4168392    1.0000000  0.2903161  0.2127458
##year          0.5805410 -0.3456474   -0.3698552 -0.4163615 -0.3091199    0.2903161  1.0000000  0.1815277
##origin        0.5652088 -0.5689316   -0.6145351 -0.4551715 -0.5850054    0.2127458  0.1815277  1.0000000
```

(c) (5pt) Use the `lm()` function to perform a multiple linear regression with `mpg` as the response and all other variables except `name` as the predictors. Use the `summary()` function to print the results. Comment on the output. For instance:

i. Is there a relationship between the predictors and the response?

ii. Which predictors appear to have a statistically significant relationship to the response?

iii. What does the coefficient for the `year` variable suggest?

Answer:

```
> lm.fit1 = lm(mpg~.-name, data=Auto)
> summary(lm.fit1)

##Call:
##lm(formula = mpg ~ . - name, data = Auto)
##
##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 ***
##cylinders     -0.493376   0.323282  -1.526  0.12780
##displacement   0.019896   0.007515   2.647  0.00844 **
##horsepower    -0.016951   0.013787  -1.230  0.21963
##weight        -0.006474   0.000652  -9.929  < 2e-16 ***
##acceleration   0.080576   0.098845   0.815  0.41548
##year           0.750773   0.050973  14.729  < 2e-16 ***
##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
```

i. Yes, according to the small p-value and the F-statistics greater than 1, we reject the null hypothesis. So there is a relationship between the predictors and the rsponse.

ii. By comparing p-values in each predictors t-statistics, we can see that displacement, weight, year, and origin have a statistically significant relationship and cylinders, horsepower, and acceleration do not.

iii. The regression coefficient for year is 0.7508, which suggests that each year, mpg increases by the coefficient. It means that cars are using less fuel each year.

(d) (5pt) Try a few different transformations of the variables, such as $\log(X)$, $\sqrt{X}$, X^2. Comment on your findings.

Answer:

```
> lm.fit2 = lm(log(Auto$mpg)~log(Auto$displacement)+sqrt(Auto$year)+
     I(Auto$weight^2))
> summary(lm.fit2)

##Call:
##lm(formula = log(Auto$mpg) ~ log(Auto$displacement) + sqrt(Auto$year) +
##    I(Auto$weight^2))
##
##Residuals:
##     Min       1Q   Median       3Q      Max
##-0.46341 -0.06791  0.00109  0.06958  0.45817
##
##Coefficients:
##                        Estimate Std. Error t value Pr(>|t|)
##(Intercept)             1.334e-01  3.184e-01   0.419    0.676
##log(Auto$displacement) -2.401e-01  2.930e-02  -8.193 3.74e-15 ***
##sqrt(Auto$year)         5.099e-01  3.190e-02  15.984  < 2e-16 ***
##I(Auto$weight^2)       -2.583e-08  2.846e-09  -9.078  < 2e-16 ***
##---
##Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
##
##Residual standard error: 0.1257 on 388 degrees of freedom
##Multiple R-squared:  0.8645,Adjusted R-squared:  0.8634
##F-statistic: 824.9 on 3 and 388 DF,  p-value: < 2.2e-16
```

Based on F-statistics and p-value, H_0 is rejected. It appears that with this transformation of variables, we have less RSE.

P6. (10pt) A data set has $n = 20$,

$$\sum_{i=1}^{20} x_i = 8.552, \quad \sum_{i=1}^{20} y_i = 398.2, \quad \sum_{i=1}^{20} x_i^2 = 5.196, \quad \sum_{i=1}^{20} y_i^2 = 9356, \quad \text{and} \quad \sum_{i=1}^{20} x_i y_i = 216.6.$$

Calculate $\hat{\beta}_0$, $\hat{\beta}_1$ and $\hat{\sigma}^2$. What is the fitted value when $x = 0.5$? Compute R^2.

Answer: We have

$$\bar{x} = \frac{1}{20}\sum_{i=1}^{20} x_i = \frac{8.552}{20} = 0.4276$$

and

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

Therefore,

$$\hat{\beta_1} = \frac{\sum_{i=1}^{20}(y_i - \bar{y})(x_i - \bar{x})}{\sum_{i=1}^{20}(x_i - \bar{x})^2} = \frac{\sum_{i=1}^{20}(y_i x_i - y_i\bar{x} - \bar{y}x_i + \bar{y}\bar{x})}{\sum_{i=1}^{20}(x_i^2 + \bar{x}^2 - 2x_i\bar{x})} = 30.1005$$

and

$$\hat{\beta_0} = \frac{\bar{y}\sum_{i=1}^{n} x_i^2 - \bar{x}\sum_{i=1}^{n} x_i y_i}{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2} = 7.039.$$

The linear regression is $\hat{y} = 7.039 + 30.1005 \cdot x$. In particular for $x = 0.5$, we have $\hat{y} = 22.089$.

Observe that

$$R^2 = \frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} = \frac{\sum_{i=1}^{20}(7.037 + 30.103x_i - \bar{y})^2}{\sum_{i=1}^{20}(y_i^2 + \bar{y}^2 - 2y_i\bar{y})} = 0.97668.$$

Since $R^2 = 1 - \frac{\text{RSS}}{\text{TSS}}$, we have RSS$= 33.29$.

Finally,

$$\hat{\sigma}^2 = \frac{\text{RSS}}{n-p-1} = \frac{33.29}{20-1-1} = 1.849.$$

P7. (10pt) The multiple linear regression model

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5 + \beta_6 x_6$$

is fitted to a data set of $n = 45$ observations. The total sum of squares is TSS $= 11.62$, and the residual sum of squares is RSS $= 8.95$. What is the p-value for the null hypothesis

$$\mathcal{H}_0 : \quad \beta_1 = \beta_2 = \beta_3 = \beta_4 = \beta_5 = \beta_6 = 0 \quad ?$$

Answer: Note that under H_0, we have

$$\frac{\frac{\text{TSS}-\text{RSS}}{p}}{\frac{\text{RSS}}{n-p-1}} \sim F_{p,n-p-1}.$$

In our case $p = 6$, $n - p - 1 = 38$ and

$$\frac{\frac{\text{TSS-RSS}}{p}}{\frac{\text{RSS}}{n-p-1}} = 1.889.$$

To find $\mathbb{P}[F \geq 1.889]$, we use R:

```
> pf(1.889,6,38,lower.tail = F)
```

```
## 0.1080044
```

Therefore, $p-$value$= 0.1080044$.

Extra Credit

Under normal assumptions we can compute the distributions of a lot of quantities explicitly.

E1. (5pt) *Chi-squared distribution.* Let $X_1, X_2, \ldots, X_n$ be independent standard normal random variables and recall that Chi-squared random variable with n degrees of freedom is defined as $\chi_n^2 = X_1^2 + X_2^2 + \cdots + X_n^2$. Prove that the density of χ_n^2 is given by

$$g_n(x) = \frac{1}{\Gamma(n/2)2^{n/2}} x^{n/2-1} e^{-x/2},$$

where $\Gamma(x)$ is the gamma function. (Hint: Prove first for $n = 1, 2$, and then use the mathematical induction.)

E2. (5pt) Let $X_1, X_2, \ldots, X_n$ be independent normal random variables $\mathcal{N}(\mu, \sigma^2)$. Prove that

$$\frac{(n-1)S^2}{\sigma^2} \overset{d}{=} \chi_{n-1}^2,$$

where $\overset{d}{=}$ stands for equality in distribution.
(Hint: Derive the moment generating function of χ_n^2 and use problem **P1.**(a) and (d).)

E3. (5pt) *Student's t distribution.* Let t_n be student's t variable, defined as

$$t_n = \frac{Z}{\sqrt{\chi_n^2/n}},$$

where $Z \sim \mathcal{N}(0, 1)$. Prove that t_n has the density

$$f_n(t) = \frac{\Gamma((n+1)/2)}{\sqrt{\pi n}\Gamma(n/2)} \cdot \frac{1}{(1 + t^2/n)^{(n+1)/2}},$$

where $\Gamma(x)$ is the gamma function. Show that for large values of n, $f_n(t)$ is approximately normal, $f_n(t) \approx e^{-t^2}/\sqrt{2\pi}$. (Hint: First show that the conditional density (distribution) of t_n given $\chi_n^2 = x$ is normal with mean 0 and variance $\sqrt{n/x}$. Then, use problem **E1.** to integrate this conditional density.)

E4. (5pt) *F (Fisher) distribution.* Let U and V be two independent Chi-squared random variables with degrees of freedom n_1 and n_2, and define the random variable, $F \equiv F(n_1, n_2)$, as

$$F = \frac{U/n_1}{V/n_2}.$$

Show that the density of F is given by

$$f_{n_1,n_2}(w) = \frac{(n_1/n_2)^{n_1/2}\Gamma[(n_1+n_2)/2]w^{(n_1/2)-1}}{\Gamma[n_1/2]\Gamma[n_2/2][1 + (n_1 w/n_2)]^{(n_1+n_2)/2}}.$$

(Hint: Compute first the distribution of F given V.)