

Stats 203 - Problem Set 1

Courtesy to Lee Shoa Long Clarke

January 31, 2010

1a. By definition $\sum_{i=1}^n e_i = \sum_{i=1}^n y_i - \hat{y}_i$. Using the definition of \bar{y} we can rewrite this:

$$\begin{aligned}\sum_{i=1}^n y_i - \hat{y}_i &= \sum_{i=1}^n y_i - \sum_{i=1}^n \hat{y}_i \\ &= n\bar{y} - \sum_{i=1}^n \hat{y}_i\end{aligned}$$

Since $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$, we can substitute to get:

$$\begin{aligned}\sum_{i=1}^n y_i - \hat{y}_i &= n\bar{y} - \sum_{i=1}^n \hat{\beta}_0 + \hat{\beta}_1 x_i \\ &= n\bar{y} - n\bar{y} + n\hat{\beta}_1 \bar{x} - \sum_{i=1}^n \hat{\beta}_1 x_i \\ &= n\hat{\beta}_1 \bar{x} - \sum_{i=1}^n \hat{\beta}_1 x_i \\ &= n\hat{\beta}_1 \bar{x} - n\hat{\beta}_1 \bar{x} \\ &= 0\end{aligned}$$

1b. No. The fact that $\sum_{i=1}^n e_i = 0$ is a consequence of how we estimate \hat{y}_i (least squares estimation). Whereas the assumption that ϵ_i are i.i.d. normal with mean 0 is based on our belief that there is not an inherent bias in our measurement of Y . Even if the error is not iid $N(0, \sigma^2)$, the least square estimation will still give us $\sum_{i=1}^n e_i = 0$.

2a. Let $X = \begin{pmatrix} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix}$, $Y = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}$, $\beta = \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix}$.

$$\hat{\beta} = \arg \min_{\beta} L(\beta) = \arg \min_{\beta} (Y - X\beta)^T (Y - X\beta)$$

Hence $\hat{\beta}$ satisfies

$$\begin{aligned}\frac{\partial L}{\partial \beta} &= 2X^T(Y - X\beta) = 0 \\ \hat{\beta} &= (X^T X)^{-1} X^T Y \\ \hat{Y} &= X\hat{\beta} = X(X^T X)^{-1} X^T Y\end{aligned}$$

By substitution we can show that the residual vector \mathbf{r} can be expressed as a linear transformation of Y .

$$\begin{aligned}\mathbf{r} &= Y - \hat{Y} \\ &= Y - (X^T X)^{-1} X^T Y \\ &= (I_n - (X^T X)^{-1} X^T) Y\end{aligned}$$

- 2b.** Define matrix $P = (X^T X)^{-1} X^T$, P is an projection matrix which project data to the space spanned by columns of X . Given $\hat{\beta} = PY$ and $r = (I_n - P)Y$,

$$\begin{aligned}\text{Cov}(\hat{\beta}, r) &= \text{Cov}(PY, (I_n - P)Y) \\ &= PCov(Y, Y)(I_n - P)^T \\ &= P\sigma^2 I_n (I_n - P) \\ &= \sigma^2 P(I_n - P^T) \\ &= \sigma^2 P(I_n - P) \quad \text{since P is symmetric} \\ &= \sigma^2 (P - PP) \\ &= \sigma^2 (P - P) \quad \text{since P is idempotent} \\ &= 0\end{aligned}$$

In addition, Y is normally distributed and linear combination of normal r.v. is still normally distributed, hence $\hat{\beta}$ and r are normal random variables. Since uncorrelated normal random variables are independent, $\hat{\beta}$ and r are independent.

- 2c.** For simplicity, let's rewrite the model using matrix notation, $Y = X\beta + \epsilon$. Since each ϵ_i is distributed $N(0, \sigma^2)$, we see that

$$\sum_{i=1}^n \epsilon_i^2 = \epsilon^T \epsilon \sim \sigma \chi_n^2$$

We can rewrite this,

$$\begin{aligned}\epsilon^T \epsilon &= (Y - X\beta)^T (Y - X\beta) \\ &= (Y - X\beta + X\hat{\beta} - X\hat{\beta})^T (Y - X\beta + X\hat{\beta} - X\hat{\beta}) \\ &= (Y - X\hat{\beta} + X(\hat{\beta} - \beta))^T (Y - X\hat{\beta} + X(\hat{\beta} - \beta)) \\ &= (Y - X\hat{\beta})^T (Y - X\hat{\beta}) + (\hat{\beta} - \beta)^T X^T X (\hat{\beta} - \beta) + 2(\hat{\beta} - \beta)^T X^T (Y - X\hat{\beta}) \\ &= (Y - X\hat{\beta})^T (Y - X\hat{\beta}) + (\hat{\beta} - \beta)^T X^T X (\hat{\beta} - \beta)\end{aligned}$$

The last equality follows from the fact that $(\hat{\beta} - \beta)^T X^T (Y - X\hat{\beta}) = (\hat{\beta} - \beta)^T (X^T Y - X^T X \hat{\beta}) = 0$ by using the definition of $\hat{\beta}$ as a linear transformation of Y as done above. Now, using $\text{Var}(\hat{\beta}) = \sigma^2 (X^T X)^{-1}$ and $E[\hat{\beta}] = \beta$, we see

$$(\hat{\beta} - \beta)^T X^T X (\hat{\beta} - \beta) = \sigma^2 (\hat{\beta} - \beta)^T \text{Var}^{-1}(\hat{\beta}) (\hat{\beta} - \beta) \sim N(0, \sigma^2 I_2).$$

and since $\hat{\beta}$ is a linear transformation of Y and $Y \sim N(\beta_0 + X\beta_1, \sigma^2)$

$$(\hat{\beta} - \beta)^T X^T X (\hat{\beta} - \beta) \sim \sigma^2 \chi_2^2$$

As in 2(b), $\hat{\beta}$ and r are independent, we know that $(\hat{\beta} - \beta)^T X^T X (\hat{\beta} - \beta)$ and $(Y - X\hat{\beta})^T (Y - X\hat{\beta})$ are independent. In addition, we have proven that

$$(\hat{\beta} - \beta)^T X^T X (\hat{\beta} - \beta) \sim \sigma^2 \chi_2^2$$

$$(Y - X\hat{\beta})^T (Y - X\hat{\beta}) \sim \chi_n^2$$

Therefore $(Y - X\hat{\beta})^T (Y - X\hat{\beta})$ must be distributed as $\sigma^2 \chi_{n-2}^2$. Rigorous proof could be found by using characteristic function.

RABE 3.4 First, create a linear model for each equation:

```
# Model equation 3.52
> lm352 <- lm(F~P1+P2)

# Model equation 3.53
> lm353 <- lm(F~P1)

# Model equation 3.54
> lm354 <- lm(F~P2)

# Model equation 3.55
> lm355 <- lm(P1~P2)

# Model equation 3.56
> lm356 <- lm(P2~P1)

# Print the coefficients
> lm352$coefficients
(Intercept)      P1      P2
-14.5005376  0.4883376  0.6720356
> lm353$coefficients
(Intercept)      P1
-22.342436  1.260516
> lm354$coefficients
(Intercept)      P2
-1.853547  1.004267
> lm355$coefficients
(Intercept)      P2
25.8980504  0.6803307
> lm356$coefficients
(Intercept)      P1
-11.668873  1.149014
```

- (a) According to the above models, $\hat{\beta}'_1 = 1.260516$, $\hat{\beta}_1 = 0.4883376$, $\hat{\beta}_2 = 0.6720356$, and $\hat{\alpha}_1 = 1.149014$. We can see that $\hat{\beta}'_1 = 1.260516 \approx 0.4883376 + 0.6720356 \times 1.149014 = 1.260515223$.
- (b) According to the above models, $\hat{\beta}'_2 = 1.004267$, $\hat{\beta}_2 = 0.6720356$, $\hat{\beta}_1 = 0.4883376$, and $\hat{\alpha}_2 = 0.6803307$. We can see that $\hat{\beta}'_2 = 1.004267 \approx 0.6720356 + 0.4883376 \times 0.6803307 = 1.004266661$.

RABE 3.14 (a) For this problem, our general model includes all parameters. $H_0: \beta_{Female} = 0$, given the the other parameters. $H_a: \beta_{Female} \neq 0$, given the other parameters.

```
# create general linear model using all parameters
> cig.full <- lm(Sales~Age+HS+Income+Black+Female+Price)
```

```
# create nested model where the Female coefficient is 0
> cig.minusFemale <- lm(Sales~Age+HS+Income+Black+Price)
```

We test if the general model fits the data better than the nested model using an F-statistic. Where under the null model, $F \sim F_{1,44}$.

```
# calculate the F-statistic of the nested model versus the full model
> anova(cig.minusFemale,cig.full)
Analysis of Variance Table

Model 1: Sales ~ Age + HS + Income + Black + Price
Model 2: Sales ~ Age + HS + Income + Black + Female + Price
  Res.Df  RSS Df Sum of Sq    F Pr(>F)
1      45 34954
2      44 34926  1    28.453 0.0358 0.8507
```

$F = 0.0358$, with a p-value of 0.8507. Thus, we cannot reject the null model that $\beta_{Female} = 0$.

- (b) $H_0: \beta_{HS} = 0, \beta_{Female}$, given the other parameters. $H_a: \beta_{HS} \neq 0$ or $\beta_{Female} \neq 0$, given the other parameters.

```
# create general linear model using all parameters
> cig.minusFemale <- lm(Sales~Age+HS+Income+Black+Price+Female)
# create nested model using all parameters except Female and HS
> cig.minusFemaleMinusHS <- lm(Sales~Age+Income+Black+Price)
```

We test if the general model fits the data better than the nested model using an F-statistic. Where under the null model, $F \sim F_{2,44}$. $F = 0.021$, with a p-value of 0.979. Thus, we cannot reject the null hypothesis that $\beta_{HS} = 0$ at the 5% level.

- (c) For this problem, I assume we are using the model where $\beta_{Female} = 0$ and $\beta_{HS} = 0$.

```
> summary(cig.minusFemaleMinusHS)

Call:
lm(formula = Sales ~ Age + Income + Black + Price)

Residuals:
    Min       1Q   Median       3Q      Max
-46.784 -11.810  -5.380   5.758 132.789

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 55.329580  62.395293   0.887   0.3798
Age          4.191538   2.195535   1.909   0.0625 .
Income       0.018892   0.006882   2.745   0.0086 **
Black        0.334162   0.312098   1.071   0.2899
Price       -3.239941   0.998778  -3.244   0.0022 **
```

Under this model, $\hat{\beta}_{Income} = 0.018892$ and $s.e.(\hat{\beta}_{Income}) = 0.006882$. We can calculate the 95% confidence interval using a t-distribution with $51 - 2 = 49$ degrees of freedom.

```
# get critical value
> qt(0.975, 49)
```

```

[1] 2.009575
# calculate confidence interval
> c(0.018892 - 0.006882*2.009575, 0.018892 + 0.006882*2.009575)
[1] 0.005062105 0.032721895

```

Thus, the 95% confidence interval of β_{Income} is 0.005 to 0.033.

- (d) I assume “above regression equation” refers to the model where $\beta_{Female} = 0$ and $\beta_{HS} = 0$. If we remove Income, we are left modeling Sales with the variables Age, Black, and Price.

```

# Create linear model and calculate R-squared
> cig.AgeBlackPrice <- lm(Sales~Age+Black+Price)
> summary(cig.AgeBlackPrice)

```

```

Call:
lm(formula = Sales ~ Age + Black + Price)

```

```

Residuals:
    Min       1Q   Median       3Q      Max
-32.607 -15.896  -5.519   5.159 133.851

```

```

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  72.8741    66.2420   1.100  0.2769
Age           5.4900     2.2882   2.399  0.0204 *
Black         0.3794     0.3326   1.141  0.2598
Price        -2.7818     1.0510  -2.647  0.0110 *
---

```

```

Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1 1

```

```

Residual standard error: 29.42 on 47 degrees of freedom
Multiple R-squared:  0.2088,    Adjusted R-squared:  0.1583
F-statistic: 4.135 on 3 and 47 DF,  p-value: 0.01108

```

$R^2 = 0.2088$, and so 20.88% of the variation in Sales can be accounted for by this model.

- (e) # Model Sales using Price, Age, and Income
- ```

> cig.priceAgeIncome <- lm(Sales~Price+Age+Income)
> summary(cig.priceAgeIncome)

```

```

Call:
lm(formula = Sales ~ Price + Age + Income)

```

```

Residuals:
 Min 1Q Median 3Q Max
-50.430 -13.853 -4.962 6.691 128.947

```

```

Coefficients:
 Estimate Std. Error t value Pr(>|t|)
(Intercept) 64.248227 61.933008 1.037 0.30487
Price -3.399234 0.989172 -3.436 0.00124 **
Age 4.155909 2.198699 1.890 0.06491 .
Income 0.019281 0.006883 2.801 0.00737 **

```

```

Signif. codes: 0 *** 0.001 ** 0.01 * 0.05 . 0.1 1

```

```

Residual standard error: 27.61 on 47 degrees of freedom
Multiple R-squared: 0.3032, Adjusted R-squared: 0.2588
F-statistic: 6.818 on 3 and 47 DF, p-value: 0.0006565

```

Modeling Sales with Price, Age, and Income, we get  $R^2 = 0.3032$ , and so 30.32% of the variation in Sales can be accounted for by these variables.

```

(f) # Model Sales by Income
 > cig.income <- lm(Sales~Income)
 > summary(cig.income)

Call:
lm(formula = Sales ~ Income)

Residuals:
 Min 1Q Median 3Q Max
-54.550 -15.772 -6.517 4.491 144.628

Coefficients:
 Estimate Std. Error t value Pr(>|t|)
(Intercept) 55.362454 27.743082 1.996 0.0516 .
Income 0.017583 0.007283 2.414 0.0195 *

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 30.63 on 49 degrees of freedom
Multiple R-squared: 0.1063, Adjusted R-squared: 0.08808
F-statistic: 5.829 on 1 and 49 DF, p-value: 0.01954

```

Modeling Sales with Income we get  $R^2 = 0.1063$ , and so 10.63% of the variation in Sales can be accounted for by Income.

**RABE 4.7** The data from Cigarette.txt is held in the object cig.

- (a) **Age** - Positive. It may be likely that older people tend to smoke more.
- HS** - Negative. A less educated population may smoke more.
- Income** - Negative. Poorer populations tend to smoke more.
- Black** - Positive. Perhaps smoking is more prevalent in black populations.
- Female** - Negative. Men smoke more than women.
- Price** - Negative. As price increases, demand decreases.

```

(b) # get the pairwise correlation coefficients matrix
 > cor(cig)

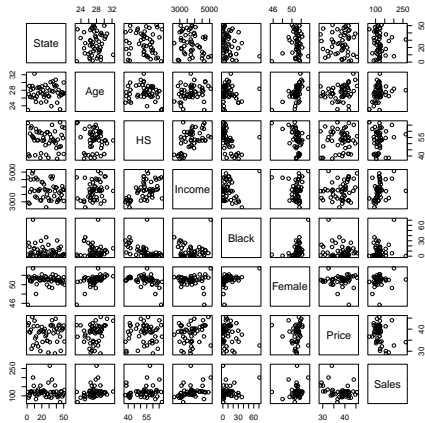
```

| State  | State | Age         | HS          | Income      | Black       | Female      | Price       | Sales       |
|--------|-------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|
|        | 1     | NA          | NA          | NA          | NA          | NA          | NA          | NA          |
| Age    | NA    | 1.00000000  | -0.09891626 | 0.25658098  | -0.04033021 | 0.55303189  | 0.24775673  | 0.22655492  |
| HS     | NA    | -0.09891626 | 1.00000000  | 0.53400534  | -0.50171191 | -0.41737794 | 0.05697473  | 0.06669476  |
| Income | NA    | 0.25658098  | 0.53400534  | 1.00000000  | 0.01728756  | -0.06882666 | 0.21455717  | 0.32606789  |
| Black  | NA    | -0.04033021 | -0.50171191 | 0.01728756  | 1.00000000  | 0.45089974  | -0.14777619 | 0.18959037  |
| Female | NA    | 0.55303189  | -0.41737794 | -0.06882666 | 0.45089974  | 1.00000000  | 0.02247351  | 0.14622124  |
| Price  | NA    | 0.24775673  | 0.05697473  | 0.21455717  | -0.14777619 | 0.02247351  | 1.00000000  | -0.30062263 |
| Sales  | NA    | 0.22655492  | 0.06669476  | 0.32606789  | 0.18959037  | 0.14622124  | -0.30062263 | 1.00000000  |

```

 # construct the scatter plots
 > pairs(cig)

```



- (c) The correlation coefficients agree well with the scatter plots. However, there are a few outliers distort the correlation to be a bit larger than expected.
- (d) The biggest difference between my expectation and the results is that I hypothesized a negative relationship between Income and Sales. However the two variables appear to be positively correlated (although very weakly), with a correlation coefficient of 0.32606789.

```
(e) > cig.full <- lm(Sales~Age+HS+Income+Black+Female+Price)
> summary(cig.full)

Call:
lm(formula = Sales ~ Age + HS + Income + Black + Female + Price)

Residuals:
 Min 1Q Median 3Q Max
-48.398 -12.388 -5.367 6.270 133.213

Coefficients:
 Estimate Std. Error t value Pr(>|t|)
(Intercept) 103.34485 245.60719 0.421 0.67597
Age 4.52045 3.21977 1.404 0.16735
HS -0.06159 0.81468 -0.076 0.94008
Income 0.01895 0.01022 1.855 0.07036
Black 0.35754 0.48722 0.734 0.46695
Female -1.05286 5.56101 -0.189 0.85071
Price -3.25492 1.03141 -3.156 0.00289 **
```

The one discrepancy between my expectations and the estimated coefficients is that Income coefficient is positive, where I speculated a negative relationship between Income and Sales.

- (f) First, correlation coefficients are measures of covariance that have been standardized by the sample standard deviations of both variables, so that the resulting values always range between

-1 and 1. Regression coefficients, however, can potentially have any real value. Second, the correlation coefficients represent a comparison of a single predictor variable with the response. The regression coefficients, however, are calculated given all the parameters in the model, and thus will change depending on which parameters are included. Hence it is the partial correlation between predictor and response adjusted by other predictors.

- (g) In question 3.14, the model selection process could easily drop both Female and HS from the full regression. In the correlation matrix, we find that some predictors are correlated such as Female and HS which may lead to collinearity in the full regression. The assumption of our F test may be violated so it would be not wise to simply drop Female HS in this case. More analysis is needed to reach a conclusion.