Homework 2 Key (100 pts + extra credit 10 pts)

4/13/2018

Problem 1 (8 + 6 = 14 pts)

a. Let us write the coefficients fitted by OLS in the first model as $\hat{\beta}_0$, $\hat{\beta}_1$. Then $RSS_1 = \sum(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{1i})^2$. The OLS estimated residual sum of squares from the second model is

$$RSS_{12} = \min_{\tilde{\beta}_0, \tilde{\beta}_1, \tilde{\beta}_2} \sum (y_i - \tilde{\beta}_0 - \tilde{\beta}_1 x_{1i} - \tilde{\beta}_2 x_{i2})^2$$

Therefore RSS_{12} is no larger than any residual sum of squares from the second model with a particular choice of $\tilde{\beta}_0$, $\tilde{\beta}_1$, and $\tilde{\beta}_2$. By letting $\tilde{\beta}_0 = \hat{\beta}_0$, $\tilde{\beta}_1 = \hat{\beta}_1$ and $\tilde{\beta}_2 = 0$, we can show that

$$RSS_{12} = \min_{\tilde{\beta}_0, \tilde{\beta}_1, \tilde{\beta}_2} \sum (y_i - \tilde{\beta}_0 - \tilde{\beta}_1 x_{1i} - \tilde{\beta}_3 \times x_{i2})^2 \leq \sum (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{1i} - 0 \times x_{i2})^2 = RSS_1$$

b. Since $R^2 = 1 - RSS/TSS$, and $TSS = \sum(y_i - \bar{y})^2$ is the same for both models, using the result from part (a) we know that R^2 from the second model is no less than the R^2 from the first model.

Problem 2 (14 pts)

Table 3.4 of the text displays a summary from the multiple linear regression of number of units sold on radio, TV, and newspaper advertising budgets. The null hypothesis of each p-value is that the coefficient for each of `TV`, `Radio`, `newspaper` is zero. By definition, the p-value is the probability of obtaining a result equal or more extreme than the ones obtained under the null hypothesis for the model:

$$\texttt{sales} = \beta_0 + \beta_1 \texttt{TV} + \beta_2 \texttt{Radio} + \beta_3 \texttt{newspaper}.$$

In this particular example, the p-values for `intercept`, `TV`, `Radio` are all less than 0.0001, which indicates that we can reject the null hypothesis (at any level $\alpha > 0.001$) that each of these predictors do not to have a (linear) association with sales under this model.

More specifically, we can conclude from the p-values that assuming there truly is no association between TV/radio and sales, the probability of seeing such a strong association between TV/radio and sales is less than 0.0001.

The large p-value of 0.8599 for `newspaper` indicates that we do not have enough evidence to reject the null hypothesis (at any common level α, for example $\alpha = 0.05$) that changes in `newspaper` does not tend to have any association with `sales` (i.e. we can conclude we do not have enough evidence to find a linear relationship between `sales` and `newspaper` under the model considered above).

Problem 3 (20 pts)

From definition, $\tilde{X} = \begin{bmatrix} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix}$, so $\tilde{X}^T\tilde{X} = \begin{bmatrix} 1 & \cdots & 1 \\ x_1 & \cdots & x_n \end{bmatrix} \begin{bmatrix} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix} = \begin{bmatrix} n & \Sigma x_i \\ \Sigma x_i & \Sigma x_i^2 \end{bmatrix}$

Then $(\tilde{X}^T X)^{-1} = \dfrac{1}{n \Sigma x_i^2 - (\Sigma x_i)^2} \begin{bmatrix} \Sigma x_i^2 & -\Sigma x_i \\ -\Sigma x_i & n \end{bmatrix}$

$$\tilde{X}^T Y = \begin{bmatrix} 1 & \cdots & 1 \\ x_1 & \cdots & x_n \end{bmatrix} \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} \Sigma y_i \\ \Sigma x_i y_i \end{bmatrix}$$

So $\begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \end{bmatrix} = \dfrac{1}{n \Sigma x_i^2 - (\Sigma x_i)^2} \begin{bmatrix} \Sigma x_i^2 & -\Sigma x_i \\ -\Sigma x_i & n \end{bmatrix} \begin{bmatrix} \Sigma y_i \\ \Sigma x_i y_i \end{bmatrix}$

since $n \Sigma x_i^2 - (\Sigma x_i)^2 = n(\Sigma x_i^2 - n\bar{x}^2) = n \Sigma (x_i - \bar{x})^2$

$$\begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \end{bmatrix} = \frac{1}{n \Sigma (x_i - \bar{x})^2} \begin{bmatrix} \Sigma x_i^2 \Sigma y_i - \Sigma x_i \Sigma x_i y_i \\ -\Sigma x_i \Sigma y_i + n \Sigma x_i y_i \end{bmatrix} = \frac{1}{n \Sigma (x_i - \bar{x})^2} \begin{bmatrix} n\bar{y} \Sigma x_i^2 - n\bar{x} \Sigma x_i y_i \\ -n^2 \bar{x}\bar{y} + n \Sigma x_i y_i \end{bmatrix}$$

Therefore

$$\hat{\beta}_0 = \frac{1}{\cancel{n} \Sigma (x_i - \bar{x})^2} \cdot \left(\cancel{n}\bar{y} \Sigma x_i^2 - \cancel{n}\bar{x} \Sigma x_i y_i \right)$$

$$= \frac{1}{\Sigma (x_i - \bar{x})^2} \left(\bar{y} \Sigma x_i^2 - \bar{y}(n\bar{x}^2) + \bar{y}(n\bar{x}^2) - \bar{x} \Sigma x_i y_i \right)$$

$$= \frac{1}{\Sigma (x_i - \bar{x})^2} \left(\bar{y} (\Sigma x_i^2 - n\bar{x}^2) - \bar{x} (\Sigma x_i y_i - n\bar{x}\bar{y}) \right)$$

Again since $\Sigma (x_i - \bar{x})^2 = \Sigma x_i^2 - n\bar{x}^2$, $\Sigma (x_i - \bar{x})(y_i - \bar{y}) = \Sigma x_i y_i - n\bar{x}\bar{y}$

$$\hat{\beta}_0 = \bar{y} - \frac{\Sigma (x_i - \bar{x})(y_i - \bar{y})}{\Sigma (x_i - \bar{x})^2} \bar{x} = \bar{y} - \hat{\beta}_1 \bar{x} \quad \text{since}$$

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

Problem 4 (9 + 9 + 9 = 27 pts)

a. We fit the multiple linear model with all covariates (except name) as predictors for `mpg`. For the variable `origin`, we use *American* (`origin` = 1) as the baseline and produced two dummy variables: `originEuropean` (1 if `origin` = 2, 0 otherwise) and `originJapanese` (1 if `origin` = 3, 0 otherwise).

```
library(ISLR)
data(Auto)
Auto$origin <-  c("American", "European", "Japanese")[Auto$origin]
fit.ml <- lm(mpg ~ . - name, data = Auto)
summary(fit.ml)
```

```
##
## Call:
## lm(formula = mpg ~ . - name, data = Auto)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9.0095 -2.0785 -0.0982  1.9856 13.3608
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)    -1.795e+01  4.677e+00  -3.839 0.000145 ***
## cylinders      -4.897e-01  3.212e-01  -1.524 0.128215
## displacement    2.398e-02  7.653e-03   3.133 0.001863 **
## horsepower     -1.818e-02  1.371e-02  -1.326 0.185488
## weight         -6.710e-03  6.551e-04 -10.243  < 2e-16 ***
## acceleration    7.910e-02  9.822e-02   0.805 0.421101
## year            7.770e-01  5.178e-02  15.005  < 2e-16 ***
## originEuropean  2.630e+00  5.664e-01   4.643 4.72e-06 ***
## originJapanese  2.853e+00  5.527e-01   5.162 3.93e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.307 on 383 degrees of freedom
## Multiple R-squared:  0.8242, Adjusted R-squared:  0.8205
## F-statistic: 224.5 on 8 and 383 DF,  p-value: < 2.2e-16
```

The multiple linear regression model indicates that there is a negative association between `mpg` and `cylinders, horsepower, weight`, whereas the relationship is positive between `mpg` and `displacement, acceleration, year, originEuropean and origin Japanese`.

The following predictors appear to have a statistically significant relationship to the response: `displacement, weight, year, originEuropean and origin Japanese` for any commonly used level α.

The coefficient for the `year` variable suggests that if `year` is increased by one unit (with all other predictors fixed) then `mpg` is expected to increases by 0.7770 `mpg` on average.

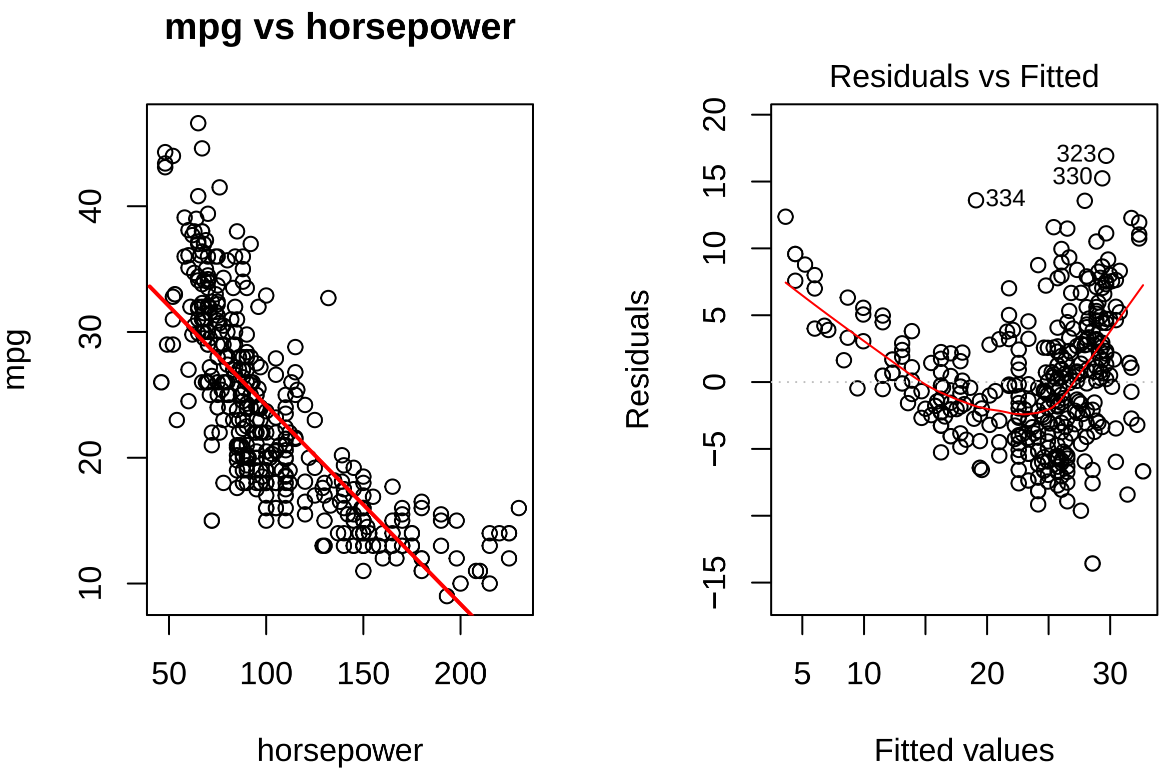
b. We start with the most simple linear model below. It can be seen that the residual plot shows a strong patten: we tend to over-estimate observations with large mpg and under-estimate observations with small mpg.

```
par(mfrow = c(1, 2))
fit0 <- lm(mpg ~ horsepower, data = Auto)
summary(fit0)
```

```
##
```

```
## Call:
## lm(formula = mpg ~ horsepower, data = Auto)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -13.5710  -3.2592  -0.3435   2.7630  16.9240
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 39.935861   0.717499   55.66   <2e-16 ***
## horsepower  -0.157845   0.006446  -24.49   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.906 on 390 degrees of freedom
## Multiple R-squared:  0.6059, Adjusted R-squared:  0.6049
## F-statistic: 599.7 on 1 and 390 DF,  p-value: < 2.2e-16
```

```r
new <- data.frame(horsepower=40:240)
plot(Auto$horsepower, Auto$mpg, xlab = "horsepower", ylab = "mpg",
     main = "mpg vs horsepower")
lines(new$horsepower, predict(fit0, new), col = "red", lwd = 2)
plot(fit0, which = 1)
```

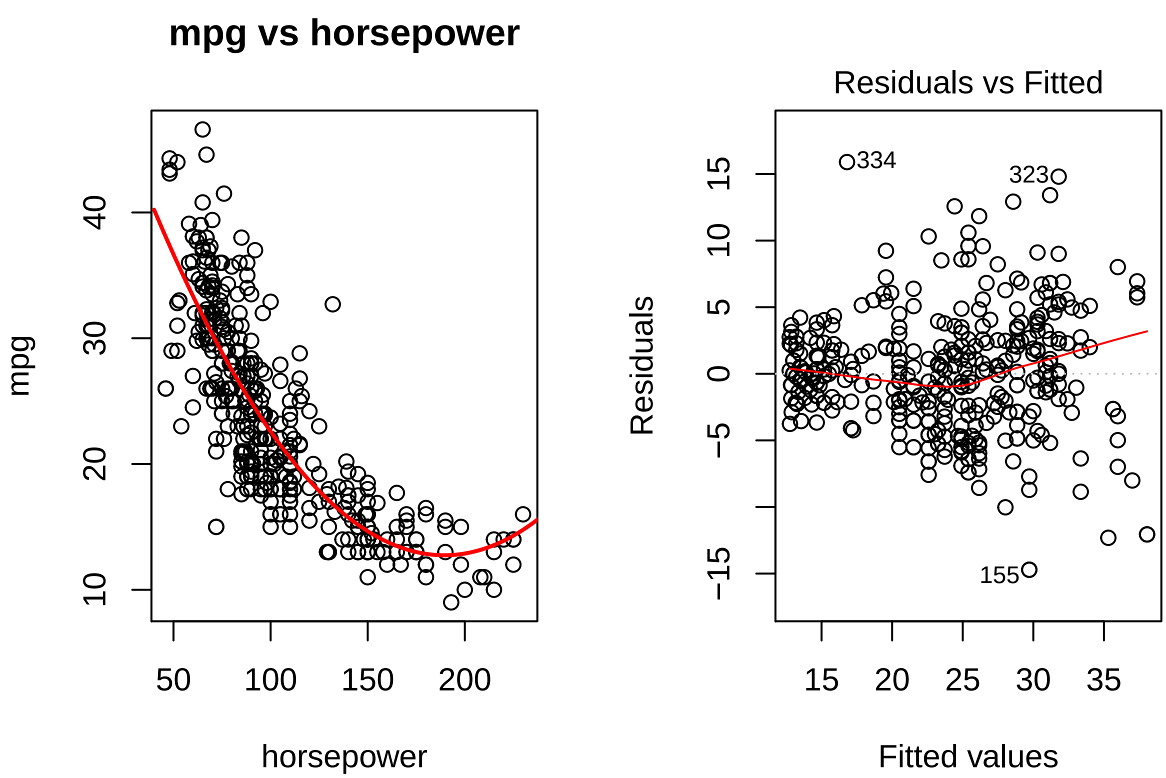


Based on the scatter plot, it seems reasonable to estimate `mpg` as a quadratic function of `horsepower`. It can be seen from the residual plot that this looks much more like a 'null plot'.

```r
par(mfrow = c(1, 2))
fit1 <- lm(mpg ~ horsepower + I(horsepower^2), data = Auto)
summary(fit1)
```

```
##
## Call:
## lm(formula = mpg ~ horsepower + I(horsepower^2), data = Auto)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -14.7135  -2.5943  -0.0859   2.2868  15.8961
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)     56.9000997  1.8004268   31.60   <2e-16 ***
## horsepower      -0.4661896  0.0311246  -14.98   <2e-16 ***
## I(horsepower^2)  0.0012305  0.0001221   10.08   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.374 on 389 degrees of freedom
## Multiple R-squared:  0.6876, Adjusted R-squared:  0.686
## F-statistic:   428 on 2 and 389 DF,  p-value: < 2.2e-16
```

```r
new <- data.frame(horsepower = 40:240)
plot(Auto$horsepower, Auto$mpg, xlab = "horsepower", ylab = "mpg",
     main = "mpg vs horsepower")
lines(new$horsepower, predict(fit1, new), col = "red", lwd = 2)
plot(fit1, which = 1)
```



At this point, it is fine to then experiment with models containing higher-order polynomial terms of `horsepower` or try adding other common transformations of `horsepower` such as log and square root transformation. You should be able to find additional predictors beyond the quadratic term to be not significant and does not change the fitted curve much. So this is a fine stopping point and you can draw conclusions based on the quadratic model.

c. The fitted model is as follows. As can be seen from the table, all the regression coefficients appear to have a statistical significant relationship to `mpg`.

```
fit.c <- lm(mpg ~ horsepower * origin, data = Auto)
summary(fit.c)
```

```
##
## Call:
## lm(formula = mpg ~ horsepower * origin, data = Auto)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -10.7415  -2.9547  -0.6389   2.3978  14.2495
##
## Coefficients:
##                             Estimate Std. Error t value Pr(>|t|)
## (Intercept)                34.476496   0.890665  38.709  < 2e-16 ***
## horsepower                 -0.121320   0.007095 -17.099  < 2e-16 ***
## originEuropean             10.997230   2.396209   4.589 6.02e-06 ***
## originJapanese             14.339718   2.464293   5.819 1.24e-08 ***
## horsepower:originEuropean -0.100515   0.027723  -3.626 0.000327 ***
## horsepower:originJapanese -0.108723   0.028980  -3.752 0.000203 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.422 on 386 degrees of freedom
## Multiple R-squared:  0.6831, Adjusted R-squared:  0.679
## F-statistic: 166.4 on 5 and 386 DF,  p-value: < 2.2e-16
```

To interpret each of the coefficients,

- On average, an American vehicle with 0 horsepower is expected to have `mpg` to be 34.4765 (0 horsepower is unrealistic, so we could also say an American vehicle is expected to have `mpg` to be $34.4765 - 0.1213 \times$ `horsepower`.)
- For an American vehicle, miles per gallon is expected to decrease by 0.1213 mpg per unit increase in engine horsepower.
- For vehicles with the same horsepower, European vehicles are expected to have 10.9972 higher mpg than American vehicles, and Japanese vehicles are expected to have 14.3397 higher mpg than American vehicles.
- For a European vehicle, miles per gallon is expected to decrease by $0.1213 + 0.1005 = 0.2218$ mpg per unit increase in engine horsepower.
- For a Japanese vehicle, miles per gallon is expected to decrease by $0.1213 + 0.1087 = 0.2300$ mpg per unit increase in engine horsepower.

Problem 5 (4 + 4 + 6 + 6 + 5 = 25 pts)

a. This model would be:

$\texttt{balance} = \beta_0 + \beta_1\texttt{income} + \beta_2\texttt{student=graduate} + \beta_3\texttt{student=undergraduate}$

To interpret each of the coefficients,

- β_0 represents the average credit card balance for non-students with 0 income.

- β_1 represents the difference in average credit card balance associated with a one-unit increase in income, comparing subjects with the same student status.
- β_2 represents the difference in average credit card balance comparing graduate students with non-students who have equal incomes.
- β_3 represents the difference in average credit card balance comparing undergraduate students with non-students who have equal incomes.

b. This model would be:

$\texttt{balance} = \beta_0 + \beta_1\texttt{income} + \beta_2\texttt{student=graduate} + \beta_3\texttt{student=not student}$

We interpret the coefficients here very similarly. The only difference is in our baseline (or reference) group:

- β_0 represents the average credit card balance for undergraduates with 0 income.
- β_1 represents the difference in average credit card balance associated with a one-unit increase in income, comparing subjects with the same student status.
- β_2 represents the difference in average credit card balance comparing graduate students with undergraduates who have equal incomes.
- β_3 represents the difference in average credit card balance comparing undergraduate students with undergraduates who have equal incomes.

c. This model would be:

$\texttt{balance} = \beta_0 + \beta_1\texttt{income} + \beta_2\texttt{student=graduate} + \beta_3\texttt{student=undergraduate} + \beta_4\texttt{income} \times \texttt{student=graduate} + \beta_5\texttt{income} \times \texttt{student=undergraduate}$

To interpret each of the coefficients,

- β_0 represents the average credit card balance for non-students with 0 income.
- β_1 represents the difference in average credit card balance associated with a one-unit increase in income, comparing non-students only.
- Now, we can write the difference in average credit card balance comparing comparing graduate students with non-students who have equal incomes as $\beta_2 + \beta_4\texttt{income}$. Therefore, β_2 and β_4 represent the intercept and slope of the line which defines this difference.
- Similarly, we can write the difference in average credit card balance comparing comparing undergraduate students with non-students who have equal incomes as $\beta_3 + \beta_5\texttt{income}$. Therefore, β_3 and β_5 represent the intercept and slope of the line which defines this difference.
- We can also write the difference in average credit card balance associated with a one-unit increase in income, comparing graduate students only, as $\beta_1 + \beta_4$. Comparing this with the interpretation of β_1 above, we can consider β_4 as the additional association between credit card balance and income induced by being a graduate student vs. a non-student.
- Similarly, we can write the difference in average credit card balance associated with a one-unit increase in income, comparing undergraduate students only, as $\beta_1 + \beta_5$. Comparing this with the interpretation of β_1 above, we can consider β_5 as the additional association between credit card balance and income induced by being an undergraduate student vs. a non-student.

d. This model would be:

$\texttt{balance} = \beta_0 + \beta_1\texttt{income} + \beta_2\texttt{student=graduate} + \beta_3\texttt{student=non student} + \beta_4\texttt{income} \times \texttt{student=graduate} + \beta_5\texttt{income} \times \texttt{student=non student}$

We interpret the coefficients here very similarly. The only difference is in our baseline (or reference) group:

- β_0 represents the average credit card balance for undergraduates with 0 income.

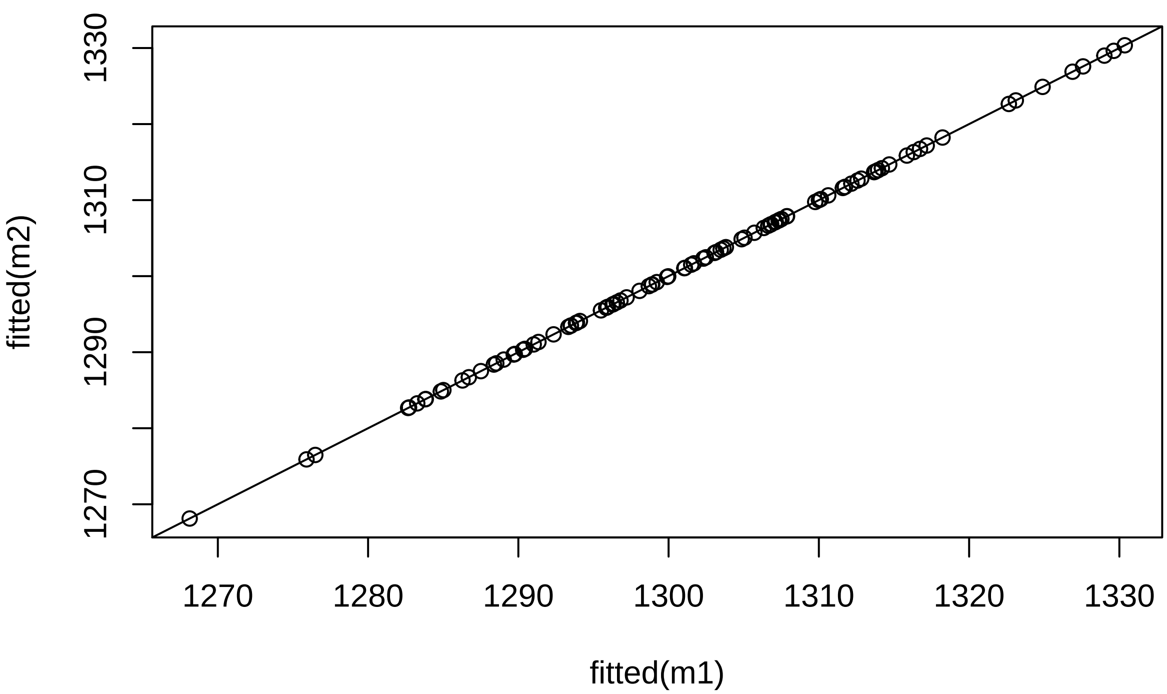
- β_1 represents the difference in average credit card balance associated with a one-unit increase in income, comparing undergraduates only.
- Now, we can write the difference in average credit card balance comparing comparing graduate students with undergraduates who have equal incomes as $\beta_2 + \beta_4$`income`. Therefore, β_2 and β_4 represent the intercept and slope of the line which defines this difference.
- Similarly, we can write the difference in average credit card balance comparing comparing non-students students with undergraduates who have equal incomes as $\beta_3 + \beta_5$`income`. Therefore, β_3 and β_5 represent the intercept and slope of the line which defines this difference.
- We can also write the difference in average credit card balance associated with a one-unit increase in income, comparing graduate students only, as $\beta_1 + \beta_4$. Comparing this with the interpretation of β_1 above, we can consider β_4 as the additional association between credit card balance and income induced by being a graduate student vs. an undergraduate.
- Similarly, we can write the difference in average credit card balance associated with a one-unit increase in income, comparing non-students only, as $\beta_1 + \beta_5$. Comparing this with the interpretation of β_1 above, we can consider β_5 as the additional association between credit card balance and income induced by being a non-student vs. an undergraduate.

e. The R code below simulates the linear models in parts a and b, and compares their predictions.

```r
set.seed(124)
income = rchisq(100, df = 10000)
student = sample(c("graduate", "undergraduate", "not student"), size = 100, replace = TRUE)
balance = 300 + 0.1*income + 3*(student=="graduate") + 4*(student=="undergraduate") + rnorm(100)

m1 = lm(balance ~ income + (student=="graduate") + (student=="undergraduate"))
m2 = lm(balance ~ income + (student=="graduate") + (student=="not student"))

plot(fitted(m1), fitted(m2))
abline(0,1)
```



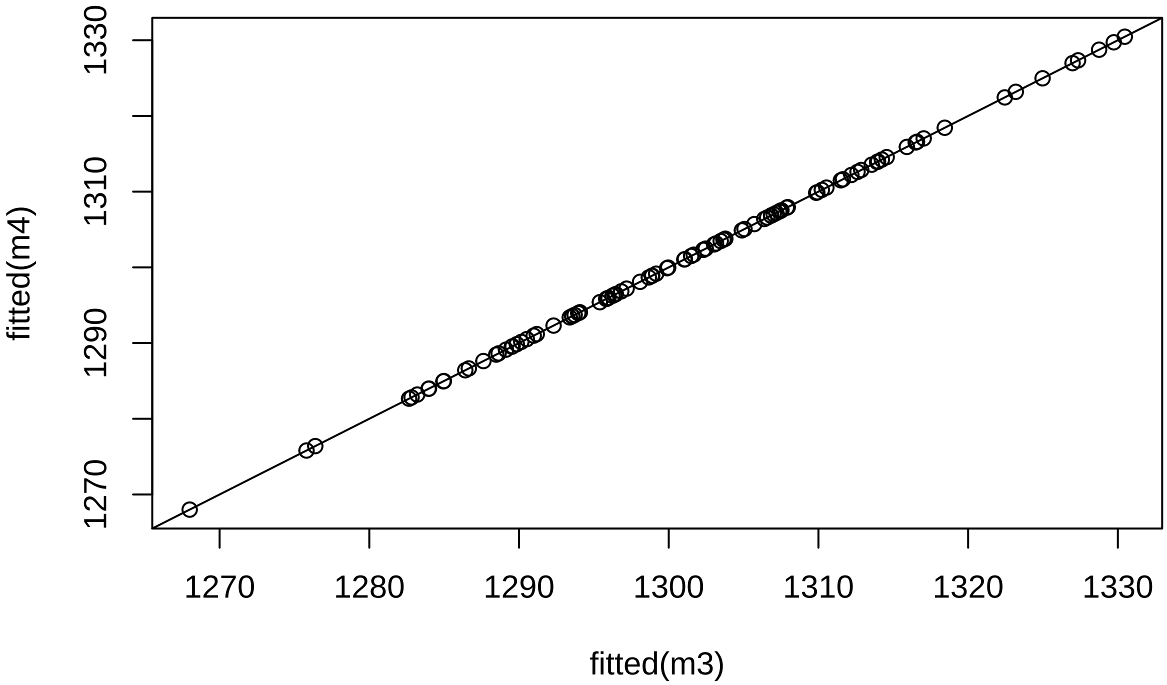
The fitted values lie exactly on the diagonal line, and therefore are equal.

The R code below simulates the linear models in parts c and d, and compares their predictions.

```r
set.seed(124)
income = rchisq(100, df = 10000)
student = sample(c("graduate", "undergraduate", "not student"), size = 100, replace = TRUE)
balance = 300 + 0.1*income + 3*(student=="graduate") + 4*(student=="undergraduate") + rnorm(100)

m3 = lm(balance ~ income + (student=="graduate") + (student=="undergraduate") +
          income*(student=="graduate") + income*(student=="undergraduate"))
m4 = lm(balance ~ income + (student=="graduate") + (student=="undergraduate") +
          income*(student=="graduate") + income*(student=="not student"))

plot(fitted(m3), fitted(m4))
abline(0,1)
```



Again, the fitted values lie exactly on the diagonal line, and therefore are equal.

Problem 6 (extra credit 10 pts)

$$Var(\hat{\beta}_1) = Var\left(\frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2}\right)$$

$$= \left(\frac{1}{\sum (x_i - \bar{x})^2}\right)^2 Var\left(\sum (x_i - \bar{x})(y_i - \bar{y})\right)$$

$$= \left(\frac{1}{\sum (x_i - \bar{x})^2}\right)^2 Var\left(\sum (x_i - \bar{x}) y_i - \bar{y} \sum (x_i - \bar{x})\right)$$

(since $\sum (x_i - \bar{x}) = 0$)

$$= \left(\frac{1}{\sum (x_i - \bar{x})^2}\right)^2 \cdot Var\left(\sum (x_i - \bar{x}) y_i\right)$$

$$= \left(\frac{1}{\sum (x_i - \bar{x})^2}\right)^2 \cdot \sum_{i=1}^{n} \left[Var\left((x_i - \bar{x}) y_i\right)\right]$$

[since $Var(y_i) = Var(\beta_0 + \beta_1 x + \varepsilon_i)$
$= Var(\varepsilon_i)$]

$$= \left(\frac{1}{\sum (x_i - \bar{x})^2}\right)^2 \cdot \sum_{i=1}^{n} \left((x_i - \bar{x})^2 \, Var(y_i)\right)$$

$$= \left(\frac{1}{\sum (x_i - \bar{x})^2}\right)^2 \cdot \sum \left((x_i - \bar{x})^2 \sigma^2\right) = \frac{\sum (x_i - \bar{x})^2}{\left(\sum (x_i - \bar{x})^2\right)^2} \sigma^2 = \frac{\sigma^2}{\sum (x_i - \bar{x})^2}$$

$$Var(\hat{\beta}_0) = Var(\bar{y} - \hat{\beta}_1 \bar{x}) = Var(\bar{y}) + Var(\hat{\beta}_1 \bar{x}) - 2cov(\bar{y}, \hat{\beta}_1 \bar{x})$$

$$= Var(\tfrac{1}{n} \Sigma y_i) + \bar{x}^2 Var(\hat{\beta}_1) - 2cov(\bar{y}, \hat{\beta}_1 \bar{x})$$

$$= \frac{1}{n^2} \cdot \sigma^2 + \bar{x}^2 \cdot \frac{\sigma^2}{\Sigma(x_i - \bar{x})^2} - 2cov(\bar{y}, \hat{\beta}_1 \bar{x})$$

$$= \sigma^2 \left[\frac{1}{n^2} + \frac{\bar{x}^2}{\Sigma(x_i - \bar{x})^2} \right] - 2cov(\bar{y}, \hat{\beta}_1 \bar{x})$$

Now, we show the covariance term is 0

$$Cov(\bar{y}, \hat{\beta}_1 \bar{x}) = \bar{x}\, cov(\bar{y}, \hat{\beta}_1) = \bar{x}\, cov\left(\frac{1}{n} \Sigma y_i \,,\, \frac{\Sigma(x_i - \bar{x})(y_i - \bar{y})}{\Sigma(x_i - \bar{x})^2} \right)$$

$$= \bar{x} \cdot \frac{1}{n} \cdot \frac{1}{\Sigma(x_i - \bar{x})^2} Cov\left(\Sigma y_i \,,\, \Sigma(x_i - \bar{x}) y_i + \underbrace{\Sigma(x_i - \bar{x}) \bar{y}}_{=0} \right)$$

$$= \bar{x} \cdot \frac{1}{n} \cdot \frac{1}{\Sigma(x_i - \bar{x})^2} Cov\left(\Sigma y_i \,,\, \Sigma(x_i - \bar{x}) y_i \right)$$

since $cov(y_i, y_j) = 0$ if $i \neq j$ $\Rightarrow$

$$= \bar{x} \cdot \frac{1}{n} \cdot \frac{1}{\Sigma(x_i - \bar{x})^2} \Sigma \left[cov(y_i \,,\, (x_i - \bar{x}) y_i) \right]$$

$$= \bar{x} \cdot \frac{1}{n} \cdot \frac{1}{\Sigma(x_i - \bar{x})^2} \underbrace{\Sigma \left[(x_i - \bar{x}) \sigma^2 \right]}_{=0}$$

$$= 0$$

Therefore, $Var(\hat{\beta}_0) = \sigma^2 \left[\frac{1}{n^2} + \frac{\bar{x}^2}{\Sigma(x_i - \bar{x})^2} \right]$