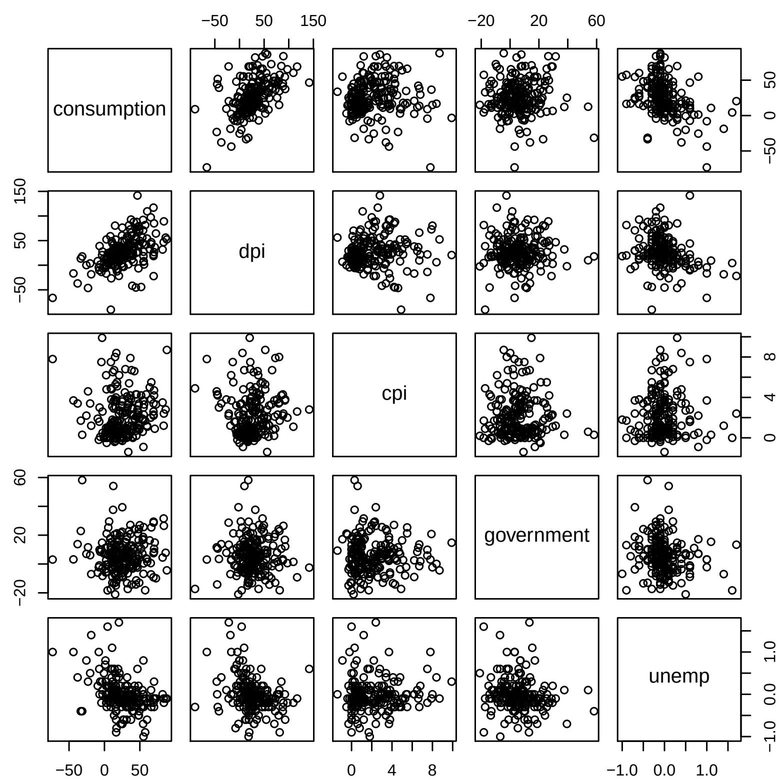
Solutions to Selected Computer Lab Problems and Exercises in Chapter 9 of *Statistics and Data Analysis for Financial Engineering, 2nd ed.* by David Ruppert and David S. Matteson

Problem 1. No outliers are seen in the scatterplots.



Changes in `consumption` show a positive relationship with changes in `dpi` and a negative relationship with changes in `unemp`, so these two variables should be most useful for predicting changes in `consumption`. The correlations between the predictors (changes in the variables other than `consumption`) are weak and collinearity will not be a serious problem.

Problem 2. Changes in `dpi` and `unemp` are highly significant and so are useful for prediction. Changes in `cpi` and `government` have large p-values and do not seem useful.

```
Call:
lm(formula = consumption ~ dpi + cpi + government + unemp)

Residuals:
```

```
    Min      1Q  Median      3Q     Max
-60.626 -12.203  -2.678   9.862  59.283

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  14.752317   2.520168   5.854 1.97e-08 ***
dpi           0.353044   0.047982   7.358 4.87e-12 ***
cpi           0.726576   0.678754   1.070    0.286
government   -0.002158   0.118142  -0.018    0.985
unemp       -16.304368   3.855214  -4.229 3.58e-05 ***
---

Residual standard error: 20.39 on 198 degrees of freedom
Multiple R-squared: 0.3385,     Adjusted R-squared: 0.3252
F-statistic: 25.33 on 4 and 198 DF,  p-value: < 2.2e-16
```

Problem 3. No, the AOV table contains sums of squares and mean squares that are not in the summary, but these are not needed for variable selection.

```
> anova(fitLm1)
Analysis of Variance Table

Response: consumption
            Df Sum Sq Mean Sq F value    Pr(>F)
dpi          1  34258   34258 82.4294 < 2.2e-16 ***
cpi          1    253     253  0.6089    0.4361
government   1    171     171  0.4110    0.5222
unemp        1   7434    7434 17.8859 3.582e-05 ***
Residuals  198  82290     416
```

Problem 4. First changes in `government` is removed and then changes in `cpi`.

```
> fitLm2 = stepAIC(fitLm1)
Start:  AIC=1228.98
consumption ~ dpi + cpi + government + unemp

             Df Sum of Sq    RSS    AIC
- government  1    0.1387  82291   1227
```

```
- cpi         1       476  82767   1228
<none>                     82290   1229
- unemp       1      7434  89724   1245
- dpi         1     22500 104790   1276

Step:  AIC=1226.98
consumption ~ dpi + cpi + unemp

        Df Sum of Sq    RSS    AIC
- cpi    1       476  82767   1226
<none>                82291   1227
- unemp  1      7604  89895   1243
- dpi    1     22542 104833   1274

Step:  AIC=1226.15
consumption ~ dpi + unemp

        Df Sum of Sq    RSS    AIC
<none>                82767   1226
- unemp  1      7381  90148   1241
- dpi    1     22932 105699   1274
> summary(fitLm2)

Call:
lm(formula = consumption ~ dpi + unemp)

Residuals:
    Min      1Q  Median      3Q     Max
-60.892 -12.660  -3.065   9.737  59.374

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  16.28476    1.91084   8.522 3.79e-15 ***
dpi           0.35567    0.04778   7.444 2.84e-12 ***
unemp       -16.01489    3.79216  -4.223 3.66e-05 ***
---

Residual standard error: 20.34 on 200 degrees of freedom
```

```
Multiple R-squared: 0.3347,     Adjusted R-squared: 0.3281
F-statistic: 50.31 on 2 and 200 DF,  p-value: < 2.2e-16
```

Problem 5. AIC decreased by 2.83 which is not a huge improvement. Dropping variables increases the log-likelihood (which increases AIC) and decreases the number of variables (which decreases AIC). The decrease due to dropping variables is limited; it is twice the number of deleted variables. In this case, the maximum possible decrease in AIC from dropping variables is 4 and is achieved only if dropping the variables does not change the log-likelihood, so we should not have expected a huge decrease. Of course, when there are many variables then a huge decrease in AIC is possible if a very large number of variables can be dropped.

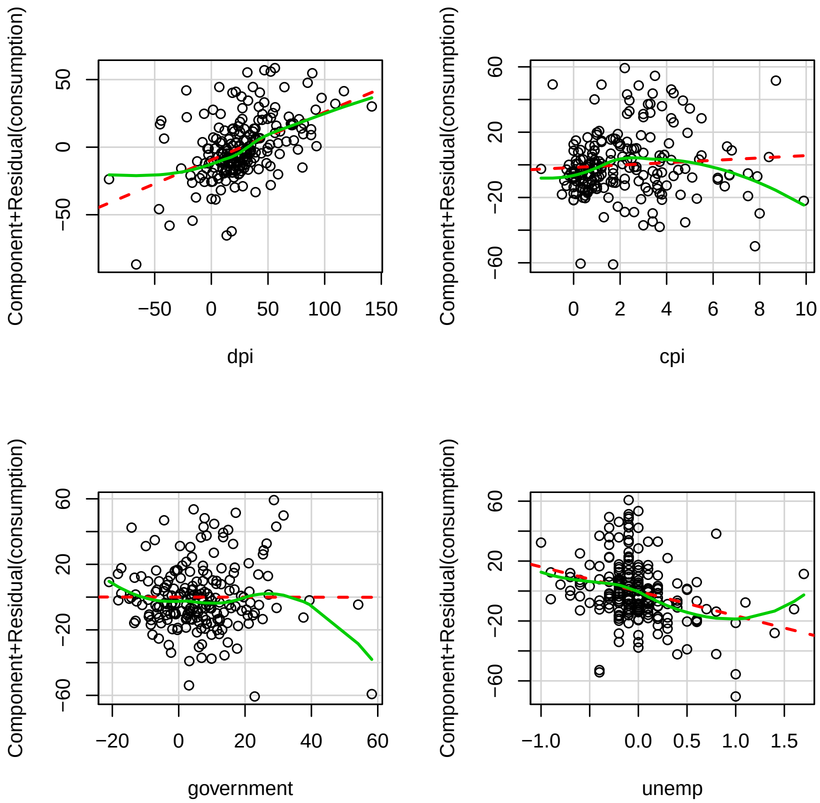
```
> AIC(fitLm1)
[1] 1807.064
> AIC(fitLm2)
[1] 1804.237
> AIC(fitLm1)-AIC(fitLm2)
[1] 2.827648
```

Problem 6.

```
> vif(fitLm1)
       dpi        cpi government      unemp
  1.100321   1.005814   1.024822   1.127610
> vif(fitLm2)
     dpi    unemp
1.095699 1.095699
```

There was little collinearity in the original model, since all four VIFs are near their lower bound of 1. Since there was little collinearity to begin with, it could not be much reduced.

Problem 7. The least-squares lines for `government` and `cpi` are nearly horizontal, which agrees with the earlier result that these variables can be dropped. The lowess curves are close to the least-squares lines, at least relative to the random variation in the partial residuals, and this indicates that the effects of `dpi` and `unemp` on `consumption` are linear.



Exercise 1a. $E(Y_i|X_i = 1) = \beta_0 + \beta_1 = 1.4 + 1.7 = 3.1$

$\text{SD}(Y_i|X_i = 1) = \text{SD}(\epsilon_i) = \sqrt{0.3} = 0.5477.$

$P(Y_i \le 3|X_i = 1) =$ `pnorm(3,mean=3.1,sd=sqrt(.3))` $= 0.4276$

Exercise 1b. $E(Y_i) = 3.1$

$\text{Var}(Y_i) = (1.7)^2(.7) + .3 = 2.323$

$P(Y_i \le 3) =$ `pnorm(.3,mean=3.1,sd=2.323)` $= 0.1140$

Exercise 2. The likelihood is

$$\prod_{i=1}^{n} \left[\frac{1}{\sqrt{2\pi}\sigma} \exp\left\{ -\frac{1}{2\sigma^2}(Y_i - \beta_0 - \beta_1 X_i)^2 \right\} \right]$$

$$= \left(\frac{1}{\sqrt{2\pi}\sigma} \right)^n \exp\left\{ -\frac{1}{2\sigma^2} \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 X_i)^2 \right\}$$

Inspection of the argument of the exponential function shows that, regardless of the value of σ, the likelihood is maximized by minimizing $\sum_{i=1}^{n}(Y_i - \beta_0 - \beta_1 X_i)^2$.

Exercise 3. This is a particularly simple calculation.

$$\text{var}(\widehat{\beta}|X_1, \ldots, X_n) = \sum_{i=1}^{n} w_i^2 \text{var}(Y_i|X_1, \ldots, X_n) = \sigma^2 \sum_{i=1}^{n} w_i^2.$$

Exercise 4a. The correlation is 0.974 and the VIFs are both 19.4. Below are two methods for calculating these results.

```
> # first method
> options(digits = 3)
> x=seq(1,15,length=30)
> corr = cor(x,x^2)
> vif = 1 /(1-corr^2)
> corr
[1] 0.974
> vif
[1] 19.4

> # second method
> xsq = x^2
> fit = lm(x~xsq)
> summary(fit)

Call:
lm(formula = x ~ xsq)

Residuals:
   Min     1Q Median     3Q    Max
-2.231 -0.689  0.269  0.845  1.044

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  3.17135    0.27829    11.4    5e-12 ***
xsq          0.05928    0.00261    22.7   <2e-16 ***
---
Residual standard error: 0.982 on 28 degrees of freedom
Multiple R-squared: 0.948,      Adjusted R-squared: 0.947
F-statistic:  515 on 1 and 28 DF,  p-value: <2e-16

> fit2 = lm(rnorm(30)~x+xsq)
> library(faraway)
> vif(fit2)
   x  xsq
19.4 19.4
```

Exercise 4b. The correlation is 0 and the VIFs are both 1. Below are two methods for calculating these results.

```
> options(digits = 3)
> # first method
> x=seq(1,15,length=30)
> x = x - mean(x)
```

```
> corr = cor(x,x^2)
> vif = 1 /(1-corr^2)
> corr
[1] 3.49e-17
> vif
[1] 1
>
> # second method
> xsq = x^2
> fit = lm(x~xsq)
> summary(fit)

Call:
lm(formula = x ~ xsq)

Residuals:
   Min     1Q Median     3Q    Max
  -7.0   -3.5    0.0    3.5    7.0

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.83e-16   1.19e+00       0        1
xsq          2.74e-17   5.07e-02       0        1

Residual standard error: 4.33 on 28 degrees of freedom
Multiple R-squared: 9.23e-33,   Adjusted R-squared: -0.0357
F-statistic: 2.58e-31 on 1 and 28 DF,  p-value: 1

> fit2 = lm(rnorm(30)~x+xsq)
> library(faraway)
> vif(fit2)
  x xsq
  1   1
>
```

Exercise 5a. $R^2 = 0.65^2 = 0.423$

Exercise 5b. 57.7

Exercise 5c. 42.3

Exercise 5d. 1.60

Exercise 7. No, one can accept that β_1 and β_2 are both 0. It is quite possible that X and X^2 are nearly collinear with high VIFs. In that case, β_1 and β_2 could both be non-zero and yet have large p-values. One might delete β_2 which has the larger p-value and then recompute the p-value for β_1.

Exercise 10a. 1.042

Exercise 10b. 0.935

Exercise 10c. $0.152 \pm (1.96)(0.012)$

Exercise 10d. Yes, `optim` converged since `mle$convergence` equals 0