# Regression Analysis, Fall 2001: Homework # 3 Solutions

October 1, 2001

---

1. Problem 3

$$\underline{Y} = X\underline{\beta} + \underline{\epsilon},\ \underline{Y} = (Y_1, Y_2, \ldots, Y_n)',\ X = \begin{bmatrix} 1 & X_1 \\ 1 & X_2 \\ \vdots & \vdots \\ 1 & X_n \end{bmatrix}$$

$\underline{\beta} = (\beta_0, \beta_1), \quad \underline{\epsilon} = (\epsilon_1, \epsilon_2, \ldots, \epsilon_n)$

In matrix notation, $\underline{\hat{\beta}} = (X'X)^{-1}X'Y$. We have:

$$X'X = \begin{bmatrix} 1 & 1 & \ldots & 1 \\ X_1 & X_2 & \ldots & X_n \end{bmatrix} \begin{bmatrix} 1 & X_1 \\ 1 & X_2 \\ \vdots & \vdots \\ 1 & X_n \end{bmatrix} = \begin{bmatrix} n & \sum X_i \\ \sum X_i & \sum X_i^2 \end{bmatrix}$$

$$\begin{aligned} (X'X)^{-1}X' &= \frac{1}{n\sum X_i^2 - (\sum X_i)^2} \begin{bmatrix} \sum X_i^2 & -\sum X_i \\ -\sum X_i & n \end{bmatrix} \begin{bmatrix} 1 & 1 & \ldots & 1 \\ X_1 & X_2 & \ldots & X_n \end{bmatrix} \\ &= \frac{1}{n\sum (X_i - \bar{X})^2} \begin{bmatrix} \sum X_i^2 - X_1 \sum X_i & \ldots & \sum X_i^2 - X_n \sum X_i \\ -\sum X_i + nX_1 & \ldots & -\sum X_i + nX_n \end{bmatrix} \end{aligned}$$

$\Rightarrow$

$$(X'X)^{-1}X'Y = \frac{1}{n\sum (X_i - \bar{X})^2} \begin{bmatrix} Y_1 \sum X_i^2 - X_1 Y_1 \sum X_i + \ldots + Y_n \sum X_i^2 - X_n Y_n \sum X_i \\ -Y_1 \sum X_i + nX_1 Y_1 - \ldots - Y_n \sum X_i + nX_n Y_n \end{bmatrix}$$

$\Rightarrow$

$$\begin{aligned} \hat{\beta}_1 &= \frac{1}{n\sum (X_i - \bar{X})^2} \left(-\sum X_i \sum Y_i + n \sum X_i Y_i\right) \\ &= \frac{n(\sum X_i Y_i - n\bar{X}\bar{Y})}{n\sum (X_i - \bar{X})^2} \\ &= \frac{\sum (X_i - \bar{X})(Y_i - \bar{Y})}{\sum (X_i - \bar{X})^2} \\ \hat{\beta}_0 &= \frac{1}{n\sum (X_i - \bar{X})^2} \left(\sum X_i^2 \sum Y_i - \sum X_i \sum X_i Y_i\right) \\ &= \frac{\sum X_i^2 \sum Y_i - \sum Y_i (\sum X_i)^2 + \sum Y_i (\sum X_i)^2 - \sum X_i \sum X_i Y_i}{n\sum (X_i - \bar{X})^2} \\ &= \frac{n\bar{Y}\sum (X_i - \bar{X})^2}{n\sum (X_i - \bar{X})^2} - \frac{n\bar{X}[\sum X_i Y_i - n\bar{X}\bar{Y}]}{n\sum (X_i - \bar{X})^2} \\ &= \bar{Y} - \hat{\beta}_1 \bar{X} \end{aligned}$$

---

2. Problem 4: RPD 3.12

   (a) Use Y $= \log \frac{Q_o}{Q_p}$
   Here is the raw result of $Y \sim X_1 + X_2 + X_3 + X_4$:

```
> summary(lm(Y~X1+X2+X3+X4))

Call: lm(formula = Y ~ X1 + X2 + X3 + X4)
Residuals:
      1       2       3       4       5       6        7       8        9
 -0.028 0.03143 0.01892 0.01694 -0.0153 -0.0286 -0.02467 0.02501 -0.04129
      10
 0.04556

Coefficients:
              Value Std. Error t value Pr(>|t|)
(Intercept)  0.0873  0.3119     0.2799  0.7907
         X1 -0.0354  0.0093    -3.7899  0.0128
         X2  0.0047  0.0048     0.9919  0.3668
         X3 -0.0019  0.0042    -0.4567  0.6670
         X4 -0.1201  0.0238    -5.0433  0.0040

Residual standard error: 0.04119 on 5 degrees of freedom
Multiple R-Squared: 0.9287
F-statistic: 16.29 on 4 and 5 degrees of freedom, the p-value is 0.004502
```

X1 (area of watershed) and X4 (peak intensity of rainfall) seem to be the most important predictors of Y (peak water flow); theirs are the only two regression coefficients which seem to be significantly different from zero according to the $t$-tests SPLUS provides. In assessing the p-value, we should take into consideration a reduction of significant level when more variables are at present. If the test level is at $\alpha = 0.05$, Bonferroni's correction (divide $\alpha$ by the number of tests, 4) results in 0.0125 cut-off for each of the four variables. Although the p-value for X1 is 0.0128, which is slightly higher than what we need, its significant level far exceeds X2 and X3. I still conclude that X1 should be taken into consideration as a serious predictor.

Also, X4 turns out to be very significant in this full model, even though in the simple regression with X4 as the only predictor, the p-value and the multiple R-Squared did not seem significant.

Together, X1 and X4 in the full model seem to be significant.

(b) Model without intercept: All the matrix notation remains unchanged except that the vector **X** does not contain a column of 1's any more since we assume $\beta_0$ is not present. It is straight forward to check that the dimensions of the matrices in the normal equations do not violate the matrix multiplication requirements (inner dimensions are equal).

$$\hat{Y} = X\hat{\beta} = X(X'X)^{-1}X'Y$$

where:

- $X = [X_1\, X_2\, X_3\, X_4]$
- $X' = [X_1\, X_2\, X_3\, X_4]'$ Note that the prime indicates transpose.
- $Y = ln(Q_o/Q_p)$
- $\hat{\beta} = (X'X)^{-1}X'Y$
- $\hat{Y} = X\hat{\beta}$

  Here are the results of the regression without an intercept:

```
Call: lm(formula = Y ~ X1 + X2 + X3 + X4 - 1)
Residuals:
      Min       1Q     Median      3Q     Max
 -0.04001 -0.02546 -0.0001699 0.02573 0.04582
```

```
Coefficients:
     Value Std. Error t value Pr(>|t|)
X1 -0.0330  0.0038    -8.6977  0.0001
X2  0.0058  0.0028     2.0524  0.0860
X3 -0.0008  0.0005    -1.3721  0.2191
X4 -0.1215  0.0214    -5.6841  0.0013

Residual standard error: 0.0379 on 6 degrees of freedom
Multiple R-Squared: 0.9889
F-statistic: 133.3 on 4 and 6 degrees of freedom, the p-value is 5.468e-06
```

The significant levels of both X1 and X4 increase (decrease in p-values of the hypotheses that their slopes equal 0). Without the intercept, p-values of X2 and X3 also decrease, but not to standard levels of significance.

(c) Only using X1 and X4, the regression without the intercept is:

```
Call: lm(formula = Y ~ X1 + X4 - 1)
Residuals:
      Min       1Q   Median      3Q     Max
 -0.05797 -0.03781 0.004037 0.02578 0.05984

Coefficients:
      Value Std. Error  t value Pr(>|t|)
X1  -0.0332   0.0039    -8.5248   0.0000
X4  -0.1243   0.0122   -10.1563   0.0000

Residual standard error: 0.04342 on 8 degrees of freedom
Multiple R-Squared: 0.9805
F-statistic: 201.3 on 2 and 8 degrees of freedom, the p-value is 1.441e-07
```

The changes in the coefficients of X1 and X4 are negligible. We have seen before that X1 and X4 are the two significant predictors for our model (with or without intercept). Since the regression coefficients of X2 and X3 are not significantly different from zero, it is plausible that dropping them does not produce much change in the remaining regression coefficients. *See also problem 3.13(b) below: it may also be true that the columns of $Z = [X2, X3]$ are nearly orthogonal to the columns of $X = [1, X1, X4]$.*

(d) All the standard errors are reduced when the intercept is removed from the model. Furthermore, when X2 and X3 are removed from the model, the standard errors of X1 and X4 are also reduced. As a result, we can eliminate predictors which do not seem to be significant for a linear model to improve our estimate of the coefficients of the ones which are significant. *However one must be careful of the* bias-variance tradeoff *we are making with this sort of reduction. We may be decreasing the standard errors of the coefficients of the remaining variables when we drop some variables, but this may come at the expense of increasing the bias in the estimates of those remaining coefficients; see problem 3.13(a) below.*

---

3. Problem 5: RPD 3.13

(a) Find E($\hat{\beta}$): Using the LS estimates $\hat{\beta} = (X'X)^{-1}X'Y$ from the model Y = X$\beta$ + $\epsilon$, $\hat{\beta} = (X'X)^{-1}X'Y$, but knowing that the true expected value of $Y$ is $E[Y] = X\beta + Z\gamma$ from the model

Y = X$\beta$ + Z$\gamma$ + $\epsilon$, we get

$$\begin{aligned} E(\hat{\beta}) &= (X'X)^{-1}X'E(Y) \\ &= (X'X)^{-1}X'[X\beta + Z\gamma] \\ &= \beta + (X'X)^{-1}X'Z\gamma \end{aligned}$$

Since E($\hat{\beta}$) is not $\beta$, $\hat{\beta}$ is a biased estimate of $\beta$.

(b) Condition under which $\hat{\beta}$ is unbiased: From part (a), E($\hat{\beta}$) = $\beta$ if and only if $(X'X)^{-1}X'Z\gamma = 0$. This can happen in one of three ways:

- The true value of $\gamma$ is a vector of all zero's (note: this violates the assumptions in the problem statement, i.e. the $Z\gamma$ part wasn't really in the model after all);
- The true values of the $\gamma$'s are set up "*just right*" so that the vector $\gamma$ is orthogonal to all the columns of $X'Z$ (again this basically violates the assumptions in the problem statement, and is very unlikely in any case);
- The columns of $X$ are orthogonal to the columns of $Z$, so that $X'Z\gamma = 0$ for all vectors $\gamma$.

The third is condition most often cited underwhich $E[\hat{\beta}]$ would be unbiased even though some other regressors were left out of the regression.

---

4. Problem 6: Antique Clocks

(a) Test the hypothesis that the auction price of a clock increases as the number of bidders increases, when age is held constant.
The original model is:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \epsilon.$$

Hypothesis:

$$H_0 : \beta_2 = 0$$

$$H_1 : \beta_2 > 0$$

This is a one-tailed test. The summary of the fitted model is:

```
Call: lm(formula = Price ~ Age + Bidders)
Residuals:
    Min     1Q Median    3Q   Max
 -206.5 -117.3  16.66 102.6 213.5

Coefficients:
                 Value Std. Error    t value   Pr(>|t|)
(Intercept) -1338.9513   173.8095    -7.7036     0.0000
        Age    12.7406     0.9047    14.0820     0.0000
    Bidders    85.9530     8.7285     9.8474     0.0000

Residual standard error: 133.5 on 29 degrees of freedom
Multiple R-Squared: 0.8923
F-statistic: 120.2 on 2 and 29 degrees of freedom, the p-value is 9.215e-15

Correlation of Coefficients:
        (Intercept)     Age
    Age -0.8759
Bidders -0.6701      0.2537
```

95% CI of $\hat{\beta}_2$ is (85.9530 $\pm$ 1.69*8.7285) = (71.2018,100.7042). Thus, we reject the hypothesis that $\beta_2 = 0$ since this interval clearly does not contain the point 0.

(b) Interpret the estimates of the $\beta$ coefficient.

- $\hat{\beta}_0$ = -1338.95, which means that if both Age and Bidders equal 0 then price is -1338.95, a value which does not make much sense in the real world.
- $\hat{\beta}_1$ = 12.74, which means that if Bidder is set to a constant then one unit increase in Age $X_1$ results in 12.74 unit increase in Price Y.
- $\hat{\beta}_2$ = 85.95, which means that if Age is set to a constant then one unit increase in Bidders $X_2$ results in 85.95 unit increase in Price Y.

(c) Compare the values of $R^2$ and adjusted $R^2$. In the full model, the multiple R-Squared is 0.8923. The calculated adjusted $R^2$ is $1 - \frac{(1-R^2)(n-1)}{(n-p-1)} = 1 - \frac{(1-0.8923)(32-1)}{(32-2-1)} = 0.8848724$; This value is a little smaller (but not greatly smaller) than the multiple R-Squared. The adjustment does not change my interpretation of the usefulness of the model. It is true that $R^2_{adjusted}$ gives a quantity which is more comparable than $R^2$ over models involving different number of parameters but the change in $R^2_{adjusted}$ is not greatly noticeable and as a result, I do not think that my inference on the model changes accordingly.

(d) Construct an overall F test for the usefulness of the model. As reported, the F-statistic is 120.2 on 2 and 29 degrees of freedom. The p-value is neglegible. This is an alternative test for the two-tailed Student's t. The small p-value indicates that there is no evidence to conclude that the $\beta_i's$ are zeros.