Regression and ANOVA

Problem Set 4

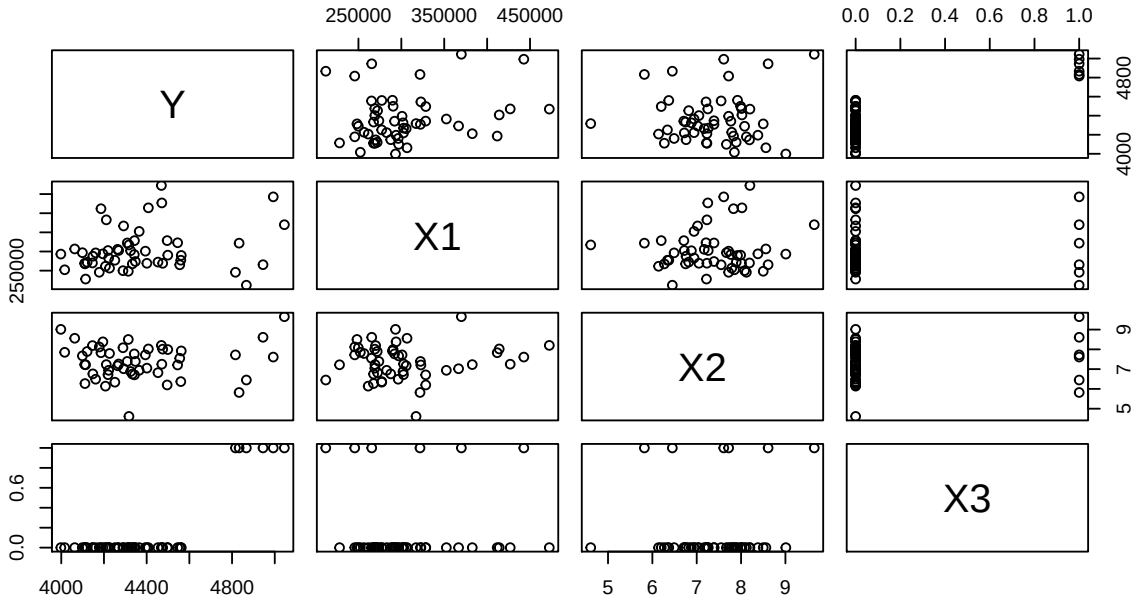
Mohamed Salem

October 29, 2019

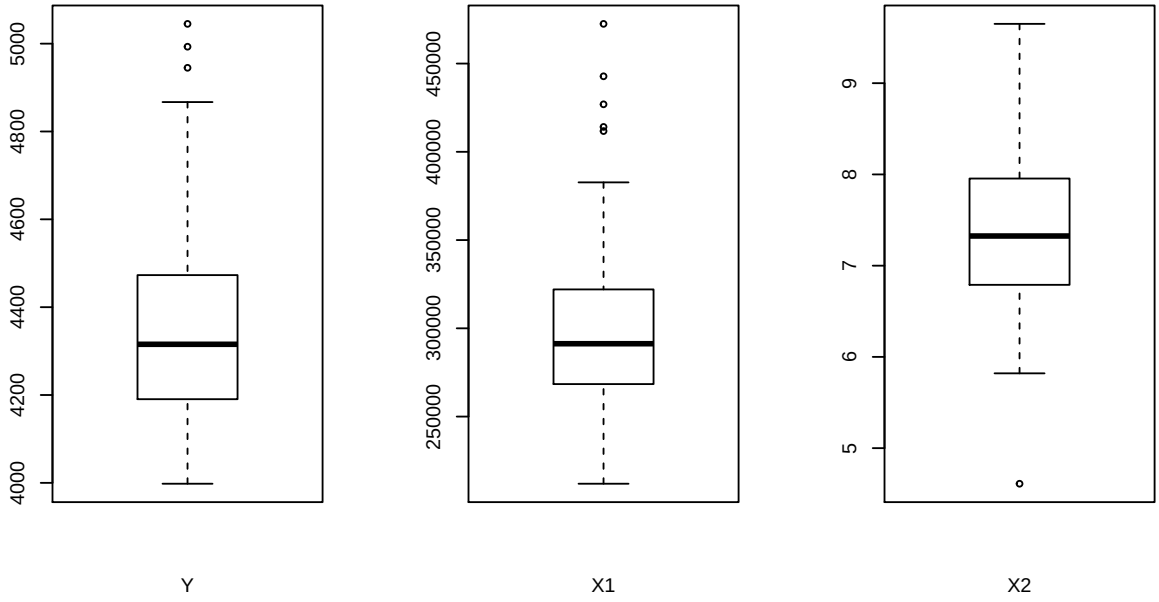
Problem 1:

(a) We begin by importing the data and carrying out some descriptive analyses.

```
# Importing the Data
datahw4 <- read.csv("D:/Vtech/Regression and ANOVA/PS4/datahw4.txt",
    sep = "")
pairs(datahw4)
```




```
par(mfrow = c(1, 3))
boxplot(datahw4$Y, xlab = "Y")
boxplot(datahw4$X1, xlab = "X1")
boxplot(datahw4$X2, xlab = "X2")
```



From our charts, we expect the variable X3 to have strong predictive power as levels of Y exceeding a certain value all have X3=1. There does not seem to be a clearly defined linear relationship between Y and either of X1 and X2. We will now fit our model and observe the results.

	Df	Sum Sq	Mean Sq	F-Value	Pr(>F)
x	3	2176606.2	725535.39	35.337	0
Residuals	48	985529.7	20531.87		

Estimated Regression Function:

$$Y_i = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2 + \hat{\beta}_3 X_3$$

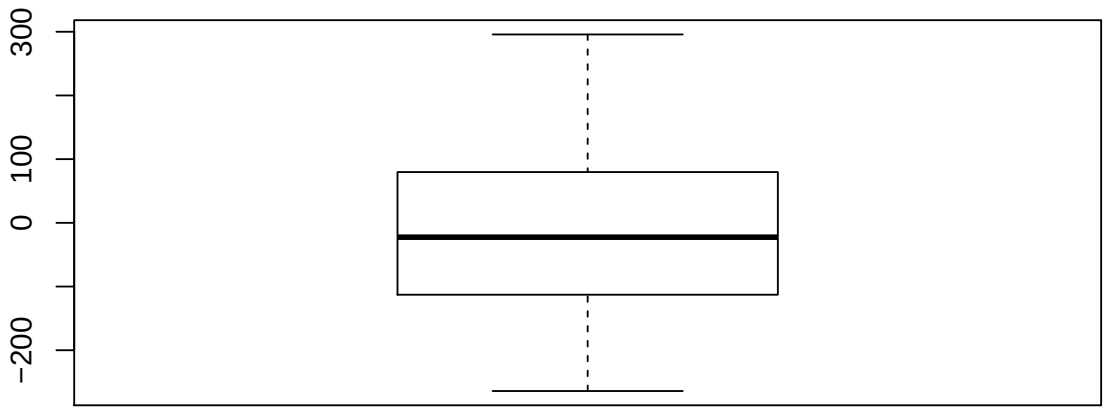
$$Y_i = 4149.887 + 0.001 X_1 + \text{-}13.166 X_2 + 623.554 X_3$$

Here, we interpret our Beta's as follows:

$\hat{\beta}_1$: The number of additional labor hours required/added when we have a unit increase in the number of cases shipped (X_1), while holding the indirect costs of the total labor hours as a percentage (X_2), and whether or not the week has a holiday (X3), constant.

$\hat{\beta}_2$: The number of additional labor hours required/added when we have a unit increase in the indirect costs of the total labor hours as a percentage (X_2), while holding the number of cases shipped (X_1) and whether or not the week has a holiday (X3), constant.

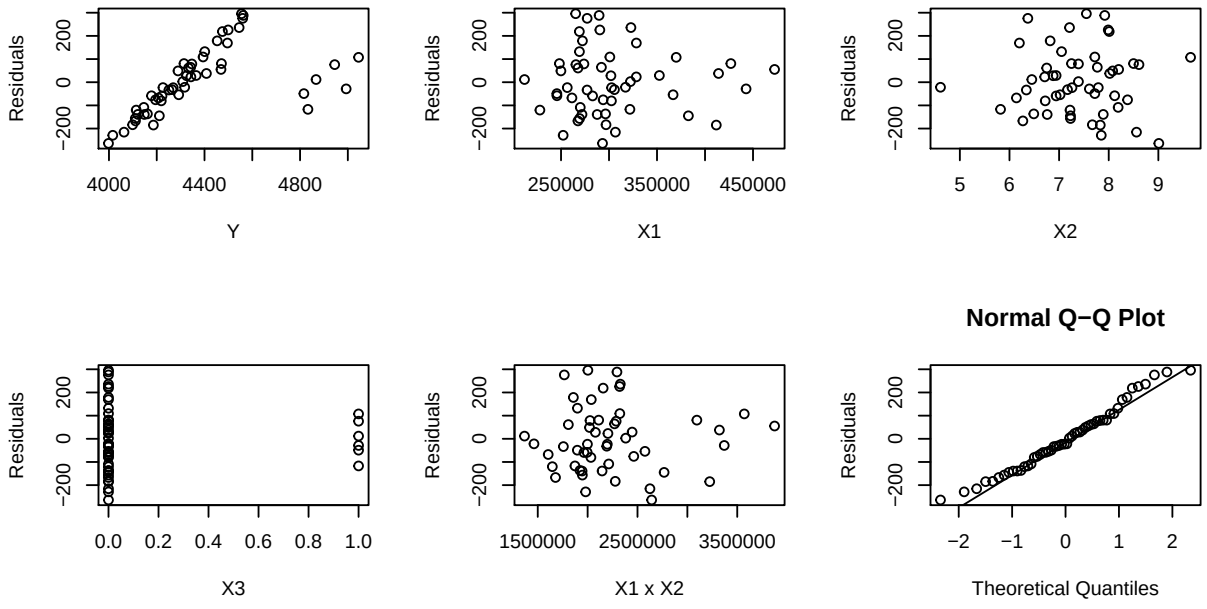
$\hat{\beta}_3$: The number of additional labor hours required/added when we have a holiday in the week (X_3), while holding the number of cases shipped (X_1) and the indirect costs of the total labor hours as a percentage (X_2), constant.



Residuals

The above residual box plot helps us understand how our residuals are distributed. We see that the distrubution of the residuals seems to show no skewness. we observe no outliers and the medin seems to be close to the mean which is always zero.

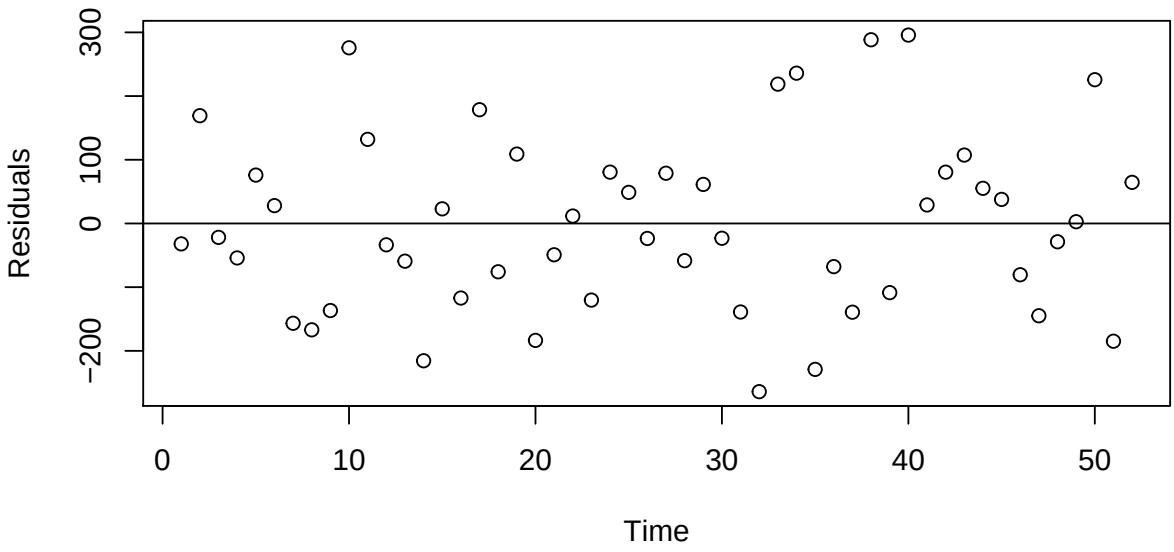
(c)



Our Normal QQ plot shows that we may have a violation of the assumption of normality of the residuals , as the distribution of the residuals seems to have thick tails. There does not seem to be a linear or polynomial

relationship between the residuals and any of our covariates, but we may have a problem of non-constant variance, given the residual plot with X3. We notice that some residuals seem to be far off from the linear relationship with the observed response variable Y, we will probably want to investigate these points further.

(d) Now we prepare a timeplot of the residuals



From the time-plot, the residuals do not seem to show any time dependence.

(e) Now we conduct a Brown-Forsythe test for constant variance Where the hypotheses are:

$$H_0 : constant\ variance$$

$$H_a : non - constant\ variance$$

And our BF test statistic is:

$$t_{BF} = \frac{\bar{d}_1 - \bar{d}_2}{s.\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

$$s^2 = \frac{\sum_i (d_{i1} - \bar{d}_1)^2 + \sum_i (d_{i2} - \bar{d}_2)^2}{n - p - 1}$$

The Brown-Forsythe test provides evidence against non-constant variance, where we observe a p-value of 0.0052 which supports the null hypothesis of constant variance if we set our type I error to a value of $\alpha = 0.01$; therefore we will assume that the constant variance assumption holds.

To test whether there is a regression relation, I will apply five tests: 3 individual t-tests, testing that each of the slope coefficients is not equal to zero; 1 F-test to test that not all coefficients are equal to zero; 1 ANoVA linear lack of fit test to test whether a linear structure is appropriate.

We begin with the t-tests:

$$H_0 : \hat{\beta}_1 = 0$$
$$H_a : \hat{\beta}_1 \neq 0$$
$$reject\ H_0\ if \quad |t^*| > t_{n-2,\,\alpha/2}$$

Our t-test produces a t-statistic equal to 2.1590228, which is greater than the critical t-value at $\alpha = 0.05$ which is equal to 2.0106348, therefore we reject our null hypothesis of no existence of a linear relationship between number of cases shipped and labor hours.

Our t-test produces a t-statistic equal to -0.5701616, which is less than the critical t-value at $\alpha = 0.05$ which is equal to 2.0106348, therefore we fail to reject our null hypothesis of no existence of a linear relationship between the indirect costs of the total labor hours as a percentage and labor hours.

Our t-test produces a t-statistic equal to 9.954423, which is greater than the critical t-value at $\alpha = 0.05$ which is equal to 2.0106348, therefore we reject our null hypothesis of no existence of a linear relationship between whether there is a holiday within the week and labor hours.

Next, we construct and check the ANOVA table:

	Df	Sum Sq	Mean Sq	F-Value	Pr(>F)
x	3	2176606.2	725535.39	35.337	0
Residuals	48	985529.7	20531.87		

We observe that the F-statistic is associated with a p-value very close to zero, which supports the existence of a linear relationship between our covariates and our response variable based on our hypotheses below:

$$H_0 : c\hat{\beta} = 0$$
$$H_a : c\hat{\beta} \neq 0$$
$$reject\ H_0\ if \quad F^* > F_{1,\,n-p-1}$$

(g) Now we apply the Bonferroni adjustment to the confidennce interval to obtain the family-wise confidencce intervals:

The table below displays the Bonferroni adjusted confidence intervals:

b0	b1	b2	b3
3664.732	-0.0001173	-70.45161	468.1559
4635.043	0.0016915	44.11957	778.9531

We observe that zero is in the confidence interval for $\hat{\beta}_2$ which would imply that the coefficient for X_2 could be zero.

(h) Next, we will obtain the analysis of variance table that decomposes the regression sum of squares into extra sums of squares associated with X1; with X3, given X1; and with X2, given X1 and X3:

	df.R - df.F	Extra Sum Sq	F-Value	Pr(>F)
Total Model	3	2176606	35.34	0
X1	1	136366	2.25	0.1399
X3|X1	1	2033565	100.43	0
X2|X1,X3	1	6675	0.33	0.5683
Total	51	3162136		

(i) From our previous construction of the extra sum of squares ANOVA we observe that:

$$F^* = \frac{SSRes(X_1, X_2) - SSRes(X_1, X_2, X_3)/((n-2)-(n-4))}{SSRes(X_1, X_2, X_3)/(n-4)} = 0.33 \quad \sim F_{2,\ 48}$$

And our hypotheses are:

$$H_0 : \beta_2 = 0$$
$$H_a : \beta_2 \neq 0$$

The associated p-value with an F-statistic of 0.33 is 0.57, therefore we fail to reject the null hypothesis at $\alpha = 0.05$.

(j)

	R.sq
X1	0.9568753
X2	0.9963964
X1,X2	0.9550645
X1|X2	0.0414814
X2|X1	0.0018924
X1,X2,X3	0.6883342

(k) Fitting the standardized regression model:

	Df	Sum Sq	Mean Sq	F-Value	Pr(>F)
x	2	0.6883342	0.3441671	54.11	0
Residuals	49	0.3116658	0.0063605		

(l) Coefficients of partial determintion between all pairs of predictor variables:

	R.sq
X1|X2,X3	0.0885161
X2|X1,X3	0.0067270
X3|X1,X2	0.6736704

For the resulting standardized coefficients, we know the follwoing identity holds:

$$\beta_k = (\frac{S_y}{S_k})\beta_{K\ std}$$

Therefore the standardized coefficients are simply scaled versions of the original coefficients; and thus we can consider the standardized regression coeffecients to reflect the effect of one predictor variable when the others are held constant.

(m) Using the above identity to transform our coefficients, we obtain the following result:

Beta Transformed	Beta Original
0.0007871	0.0007871
-13.1660192	-13.1660192
623.5544807	623.5544807

(n)

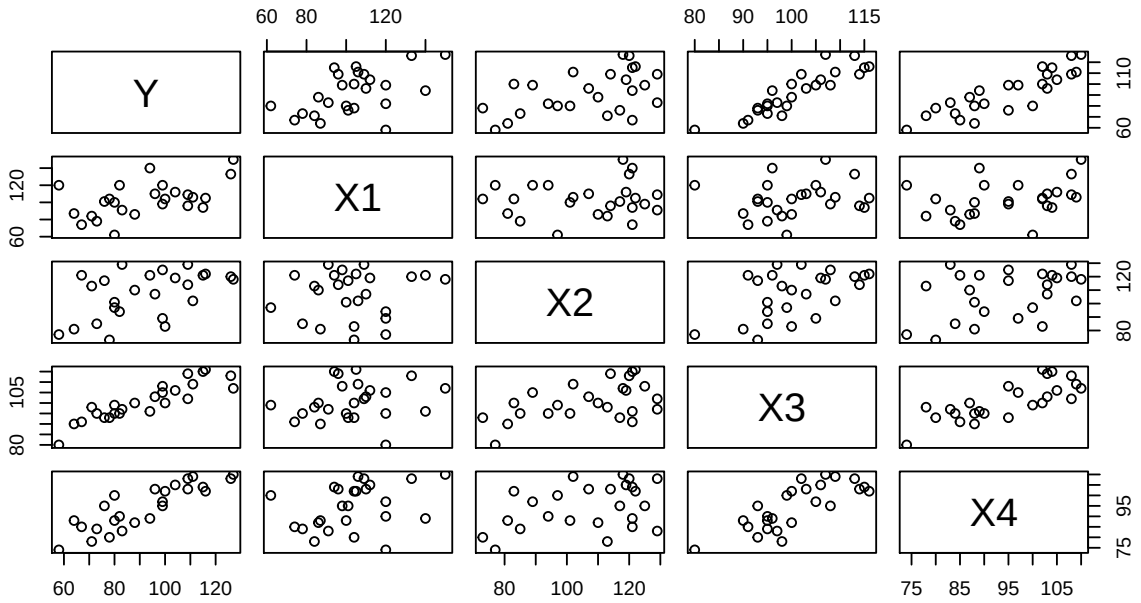
SSR X1	136366.2
SSR X1|X2	130697.2

We observe that the two figures are not equal here, however the difference is not substantial.

Problem 2:

(a)

```
# Importing the Data
data3hw4 <- read.csv("D:/Vtech/Regression and ANOVA/PS4/data3hw4.txt",
    sep = "")
pairs(data3hw4)
```




```
corr_mat <- cor(data3hw4[, 2:5])
tbl_corr <- (kable(corr_mat))
tbl_corr <- kable_styling(tbl_corr, position = "center")
tbl_corr <- column_spec(tbl_corr, 1, border_left = T)
tbl_corr <- column_spec(tbl_corr, 5, border_right = T)
```

	X1	X2	X3	X4
X1	1.0000000	0.1022689	0.1807692	0.3266632
X2	0.1022689	1.0000000	0.5190448	0.3967101
X3	0.1807692	0.5190448	1.0000000	0.7820385
X4	0.3266632	0.3967101	0.7820385	1.0000000

The scatterplots suggest that there may be a linear relationship between our response variable Y, and all

our predictor variables. However, we notice from the correlation matrix that we have a strong correlation between X3 and X4, as well as a medium strength correlation between X2 and X3, both of which may lead to problems of multicollinearity. To check whether we have serious multicollinearity we compute the variance inflation factor as follows:

```
lmfitx1 <- lm(X1 ~ X4 + X2 + X3, data = data3hw4)
lmfitx2 <- lm(X2 ~ X1 + X4 + X3, data = data3hw4)
lmfitx3 <- lm(X3 ~ X1 + X2 + X4, data = data3hw4)
lmfitx4 <- lm(X4 ~ X1 + X2 + X3, data = data3hw4)

VIF_X1 <- 1/(1 - summary(lmfitx1)$r.squared)
VIF_X2 <- 1/(1 - summary(lmfitx2)$r.squared)
VIF_X3 <- 1/(1 - summary(lmfitx3)$r.squared)
VIF_X4 <- 1/(1 - summary(lmfitx4)$r.squared)

vif_mat <- data.frame()
vif_mat[1, 1] <- VIF_X1
vif_mat[2, 1] <- VIF_X2
vif_mat[3, 1] <- VIF_X3
vif_mat[4, 1] <- VIF_X4
rownames(vif_mat) <- c("VIF X1", "VIF X2", "VIF X3", "VIF X4")
colnames(vif_mat) <- c("")
tbl_vif <- (kable(vif_mat))
tbl_vif <- kable_styling(tbl_vif, position = "center")
tbl_vif <- column_spec(tbl_vif, 1, border_left = T)
tbl_vif <- column_spec(tbl_vif, 2, border_right = T)
```

VIF X1	1.138043
VIF X2	1.369512
VIF X3	3.016549
VIF X4	2.834776

Since all our variance inflation factors are less than 10, we conclude that we do not have a severe multicollinearity problem.

(b) After fitting the multiple linear model, at first glance, it seems that the covariate X2 should not be icluded in the model.

```
lmfit2 <- lm(Y ~ ., data = data3hw4)
summary(lmfit2)
```

```
##
## Call:
## lm(formula = Y ~ ., data = data3hw4)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.9779 -3.4506  0.0941  2.4749  5.9959
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -124.38182    9.94106 -12.512 6.48e-11 ***
## X1             0.29573    0.04397   6.725 1.52e-06 ***
## X2             0.04829    0.05662   0.853  0.40383
```

```
## X3             1.30601    0.16409   7.959 1.26e-07 ***
## X4             0.51982    0.13194   3.940  0.00081 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.099 on 20 degrees of freedom
## Multiple R-squared:  0.9629, Adjusted R-squared:  0.9555
## F-statistic: 129.7 on 4 and 20 DF,  p-value: 5.262e-14
```

(c) Now we will use adjusted R^2 to find the four best subset regression models:

```
library(leaps)
b <- regsubsets(Y ~ ., data = data3hw4, nbest = 16)
rs <- summary((b))
rsort <- arrange(data.frame(cbind(rs$which, rs$adjr)), desc(rs$adjr2))
names(rsort) <- c("Intercept", "X1", "X2", "X3", "X4", "R.sq.adj")
subs_mat <- head(rsort, 4)
tbl_subs <- (kable(subs_mat))
tbl_subs <- kable_styling(tbl_subs, position = "center")
tbl_subs <- column_spec(tbl_subs, 1, border_left = T)
tbl_subs <- column_spec(tbl_subs, 6, border_right = T)
```

Intercept	X1	X2	X3	X4	R.sq.adj
1	1	0	1	1	0.9560482
1	1	1	1	1	0.9554702
1	1	0	1	0	0.9269043
1	1	1	1	0	0.9246779

Based on R^2-adjusted of all possible models, we conclude that the models with the best fit are the ones displayed above.

(d) We could also check the following criteria: Mallow's C_{p}, Akaike's Information Criterion (AIC), Schwarz Bayesian Information Criterion (BIC), or the Prediction Sum of Squares Criterion (PRESS). The first three of these criteria offer ways of rewarding a model for producing smaller residuals (i.e: accounting for a greater portion of the variation in the model), and penalize the model for incorporating many parameters. The fial method, on the other hand (PRESS), is based on the model's predictive power for one observation that is left out, with the process repeated for some (usually all) individual observations.

(e) Using backward elimination to find the best subset of predictor variables to predict job proficiency, we find:

```
b <- regsubsets(Y ~ ., data = data3hw4, nbest = 16, method = "backward")
rs <- summary((b))
rsort <- arrange(data.frame(cbind(rs$which, rs$adjr)), desc(rs$adjr2))
names(rsort) <- c("Intercept", "X1", "X2", "X3", "X4", "R.sq.adj")
subs_mat <- head(rsort, 1)
tbl_subs <- (kable(subs_mat))
tbl_subs <- kable_styling(tbl_subs, position = "center")
tbl_subs <- column_spec(tbl_subs, 1, border_left = T)
tbl_subs <- column_spec(tbl_subs, 6, border_right = T)
```

Intercept	X1	X2	X3	X4	R.sq.adj
1	1	0	1	1	0.9560482

(f) Using forward selection to find the best subset of predictor variables to predict job proficiency, we find:

```
b <- regsubsets(Y ~ ., data = data3hw4, nbest = 16, method = "forward")
rs <- summary((b))
rsort <- arrange(data.frame(cbind(rs$which, rs$adjr)), desc(rs$adjr2))
names(rsort) <- c("Intercept", "X1", "X2", "X3", "X4", "R.sq.adj")
subs_mat <- head(rsort, 1)
tbl_subs <- (kable(subs_mat))
tbl_subs <- kable_styling(tbl_subs, position = "center")
tbl_subs <- column_spec(tbl_subs, 1, border_left = T)
tbl_subs <- column_spec(tbl_subs, 6, border_right = T)
```

Intercept	X1	X2	X3	X4	R.sq.adj
1	1	0	1	1	0.9560482

(g) Using forward selection to find the best subset of predictor variables to predict job proficiency, we find:

```
b <- regsubsets(Y ~ ., data = data3hw4, nbest = 16, method = "seqrep")
rs <- summary((b))
rsort <- arrange(data.frame(cbind(rs$which, rs$adjr)), desc(rs$adjr2))
names(rsort) <- c("Intercept", "X1", "X2", "X3", "X4", "R.sq.adj")
subs_mat <- head(rsort, 1)
tbl_subs <- (kable(subs_mat))
tbl_subs <- kable_styling(tbl_subs, position = "center")
tbl_subs <- column_spec(tbl_subs, 1, border_left = T)
tbl_subs <- column_spec(tbl_subs, 6, border_right = T)
```

Intercept	X1	X2	X3	X4	R.sq.adj
1	1	0	1	1	0.9560482

(h) we observe that in this specific case, all three methods, backward elimination, forward selection, and stepwise regression, have all produced the same result in terms of which model is the best in terms of adjusted R^{2}.