# Assignment #2

Matt Massey

## Question 1.

```
mfgcost = read.csv("hw2_mfgcost.csv", header = T)

# MLR model
mfg.lm = lm(COST ~ PAPER + MACHINE + OVERHEAD + LABOR, data = mfgcost)
result.mfg = summary(mfg.lm)
result.mfg
##
## Call:
## lm(formula = COST ~ PAPER + MACHINE + OVERHEAD + LABOR, data = mfgcost)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -18.691  -7.407  -1.978   6.675  22.516
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 51.72314   21.70397   2.383   0.0262 *
## PAPER        0.94794    0.12002   7.898 7.30e-08 ***
## MACHINE      2.47104    0.46556   5.308 2.51e-05 ***
## OVERHEAD     0.04834    0.52501   0.092   0.9275
## LABOR       -0.05058    0.04030  -1.255   0.2226
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 11.08 on 22 degrees of freedom
## Multiple R-squared:  0.9988, Adjusted R-squared:  0.9986
## F-statistic:  4629 on 4 and 22 DF,  p-value: < 2.2e-16

# 95% confidence interval of true manufacturing cost as a product of machine hours
ci = 2.47104 + c(-1, 1) * qt(p = 0.025, df = 22, lower.tail = F) * 0.46556

cat("True marginal cost associated with total machine hours per month: ", 2.47104, "\n")
## True marginal cost associated with total machine hours per month:  2.47104

cat("95% confidence interval: ", ci)
## 95% confidence interval:  1.505528 3.436552
```

**1A.** COST = 51.72314 + 0.9794(PAPER) + 2.47104(MACHINE) + 0.04834(OVERHEAD) - 0.05058(LABOR)

COST = total manufacturing cost per month in thousands of dollars
PAPER = total production of paper per month in tons
MACHINE = total machine hours used per month
OVERHEAD = total variable overhead costs per month in thousands of dollars
LABOR = total direct labor hours used each month

**1B.** The true marginal cost associated with total machine hours per month is given by the regression coefficient for the MACHINE predictor variable, which is 2.47104. The 95% confidence interval for this estimate ranges from 1.505528 to 3.436552. This translates to a true marginal cost of $2,471.04 for every machine hour used per month, with a 95% confidence interval of $1,505.53 to $3,436.55 for every machine hour used per month.

**1C.** The $R^2$ value indicates that 99.88% of the variability of manufacturing cost per month (in $1000) is explained by this regression model.

## Question 2.

```
# read data
wheat = read.csv("hw2_wheat.csv", header = T)

# given in hw2, but not used?
n = dim(wheat)[1]

# multi linear model using two predictor variables - EXCHRATE and PRICE
wheat.lm = lm(SHIPMENT ~ EXCHRATE + PRICE, data = wheat)
result.wheat = summary(wheat.lm)
result.wheat
##
## Call:
## lm(formula = SHIPMENT ~ EXCHRATE + PRICE, data = wheat)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1608.0  -537.3    24.6   513.1  3491.1
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  3361.932    633.194   5.309 4.53e-07 ***
## EXCHRATE        1.869      4.223   0.443  0.65877
## PRICE       -2413.837    846.480  -2.852  0.00505 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 798.3 on 132 degrees of freedom
## Multiple R-squared:  0.08801,    Adjusted R-squared:  0.07419
## F-statistic: 6.369 on 2 and 132 DF,  p-value: 0.002287

#calculate F critical value
fcv = qf(p=0.05, df1=2, df2=132, lower.tail=F)
cat("F0.05,2,132 critical value for MLR: ", fcv, "\n\n")
## F0.05,2,132 critical value for MLR:  3.064761

#calculate t critical value
tcv = qt(p=0.025, df=132, lower.tail=F)
cat("t0.05,132 critical values for EXCHRATE: ", -1*tcv, " & ", tcv)
## t0.05,132 critical values for EXCHRATE:  -1.978099  &  1.978099
```

### 2A.

SHIPMENT = 3361.932 + 1.869(EXCHRATE) - 2413.837(PRICE)

SHIPMENT = U.S. wheat export shipments
EXCHRATE = $B_1$ = the real index of weighted-average exchange rates of the U.S. dollar
PRICE = $B_2$ = the per-bushel real price of no. 1 red winter wheat

### 2B.

To test the significance of the overall fit of the regression model, we will use a right-tailed test with a null hypothesis ($H_0$) of $B_1=B_2=0$, an alternative hypothesis ($H_A$) of $B_k\neq0$ for some k (at least one predictor not equal to 0), and a significance level of $\alpha=0.05$. If the P value is at or below the significance level, or the F value is beyond the F critical value, we will reject the null hypothesis. Otherwise we will fail to reject the null hypothesis.

Here we see that P=0.002287, which is well below the significance level of 0.05, so we can reject the null hypothesis. The F value of the model (6.369) and the F critical value (3.064761) yields the same rejection of $H_0$. We can conclude that at least one predictor variable is not equal to 0 and that the overall fit of the MLR

model relating U.S. wheat export shipments to exchange rate and price per bushel of wheat (predictors) is significant.

## 2C.

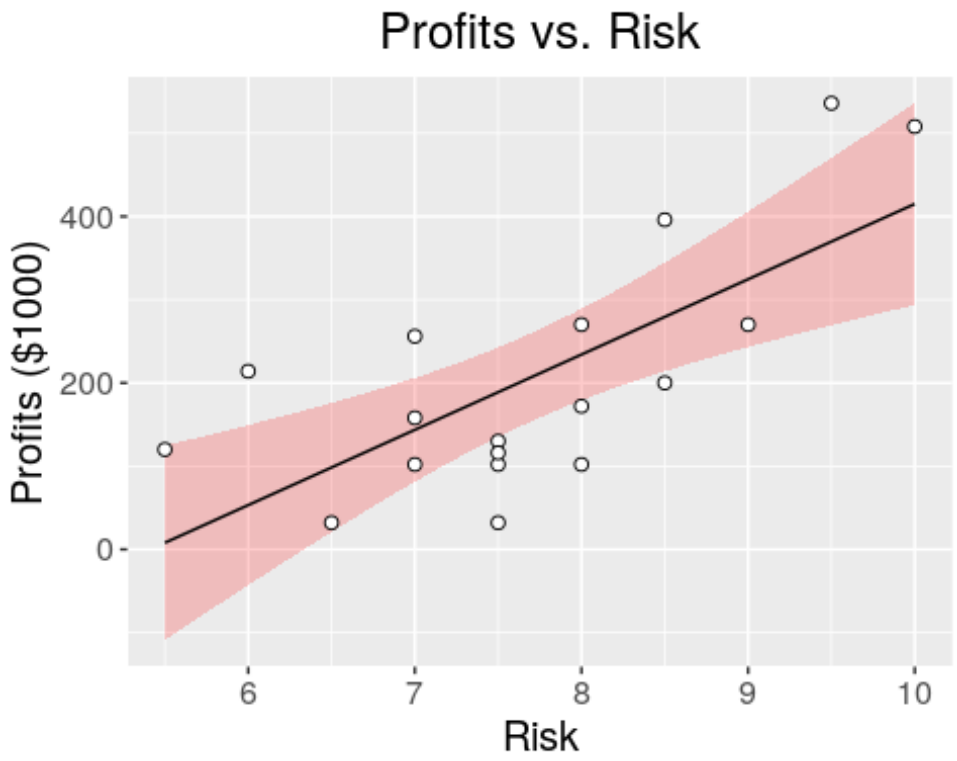
Given that the predictor PRICE is accounted for and constant, we can test the significance of a relationship between SHIPMENT and EXCHRATE using a two-tailed hypothesis test with a null hypothesis ($H_0$) of $B_1=0$, an alternative hypothesis ($H_A$) of $B_1\neq0$, and a significance level of $\alpha=0.05$. If the P value is at or below the significance level, or the t value is beyond the t critical values, we will reject the null hypothesis. Otherwise we will fail to reject the null hypothesis.

Here we see that P=0.65877 and is much greater than the significance level of 0.05. We also see that the t value of EXCHRATE is 0.443, which is well within the range of calculated critical values of ±1.978099. We fail to reject the null hypothesis. This implies that changing the exchange rate, while keeping price constant, does not significantly change the mean response of shipments, however, this does not mean that exchange rate has no effect.

**Question 3.**

**3A.**

```
# import ggplot2
library(ggplot2)

# read file and create dataframe
RDS = read.csv("hw2_RDS.csv",header = T)

# add new dataframe column of RD^2
RDS_1 = data.frame(RDS,RDS$RD^2)
colnames(RDS_1)[4]<-"RD_SQ"

# extract individual columns from RDS_1 dataframe
Risk = RDS_1$RISK
Profit = RDS_1$PROFIT
Rd = RDS_1$RD

# create plot from RDS_1 dataframe with datapoints, MLR
ggplot(RDS_1, aes(RISK, PROFIT)) + geom_smooth(method="lm", col=1, color='red', size=0.5,
fill='red', alpha=0.2) + geom_point(aes(RISK, PROFIT), shape=21, size=2, fill='white') +
labs(x = "Risk", y = "Profits ($1000)") + theme(text = element_text(size = 15),
plot.title=element_text(hjust=0.5)) + ggtitle("Profits vs. Risk")
## Warning: Duplicated aesthetics after name standardisation: colour
## `geom_smooth()` using formula 'y ~ x'
```



```
ggplot(RDS_1, aes(RD, PROFIT)) + geom_smooth(method="lm",col=1, color='red', size=0.5,
fill='red', alpha=0.2) + geom_point(aes(RD, PROFIT), shape=21, size=2, fill='white') + labs(x
= "Research & Development ($100)", y = "Profits ($1000)") + theme(text = element_text(size =
15), plot.title=element_text(hjust=0.5)) + ggtitle("Profits vs. Research & Development")
## Warning: Duplicated aesthetics after name standardisation: colour
## `geom_smooth()` using formula 'y ~ x'
```


```
RDS_1.lm = lm(PROFIT~ RISK + RD, data = RDS_1)
RDS_1.result = summary(RDS_1.lm)
RDS_1.result
##
## Call:
## lm(formula = PROFIT ~ RISK + RD, data = RDS_1)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -16.555 -11.496  -2.318   6.133  27.060
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -453.1763    23.5061 -19.279 5.37e-12 ***
## RISK          29.3090     3.6686   7.989 8.76e-07 ***
## RD             4.5100     0.1538  29.333 1.16e-14 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 14.34 on 15 degrees of freedom
## Multiple R-squared:  0.9915, Adjusted R-squared:  0.9904
## F-statistic: 879.1 on 2 and 15 DF,  p-value: 2.852e-16
```

The two plots both show a positive association of company profits with both predictor variables, risk and research and development. We also see very low P values and high t values of risk and research and development, indicating a significant effect of both predictors on profits, when accounting for the effects of the other variables. The $R^2$ value also shows that the model accounts for 99.15% of the variability in the data.

### 3B.

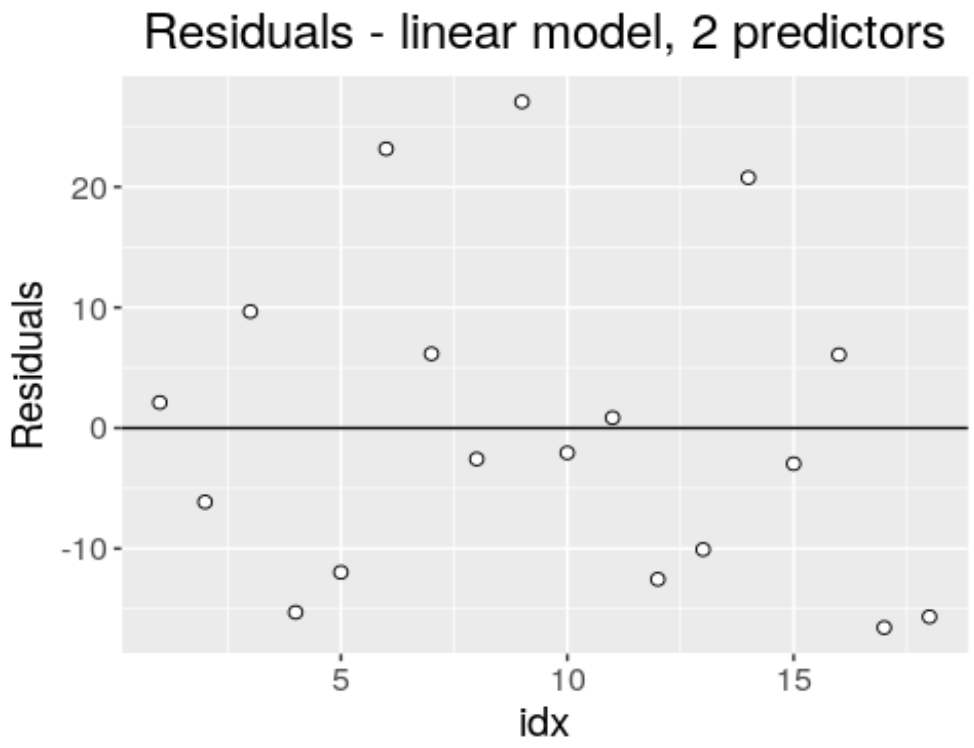
```
RDS_2.lm = lm(PROFIT~ RISK + RD + RD_SQ, data = RDS_1)
RDS_2.result = summary(RDS_2.lm)
RDS_2.result
##
## Call:
## lm(formula = PROFIT ~ RISK + RD + RD_SQ, data = RDS_1)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -7.3566 -2.0076 -0.0788  2.7391  5.1709
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -2.454e+02  1.481e+01 -16.567 1.36e-10 ***
## RISK         2.325e+01  9.884e-01  23.522 1.18e-12 ***
## RD           1.014e+00  2.324e-01   4.365 0.000647 ***
## RD_SQ        1.757e-02  1.152e-03  15.248 4.10e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.538 on 14 degrees of freedom
## Multiple R-squared:  0.9995, Adjusted R-squared:  0.9994
## F-statistic:  9708 on 3 and 14 DF,  p-value: < 2.2e-16
```

The scatter plot of Profit vs RD in 3A shows a very good fit with the MLR model. However, there is a VERY slight curvature to the data points and it is worth at least comparing the MLR with a polynomial regression model.

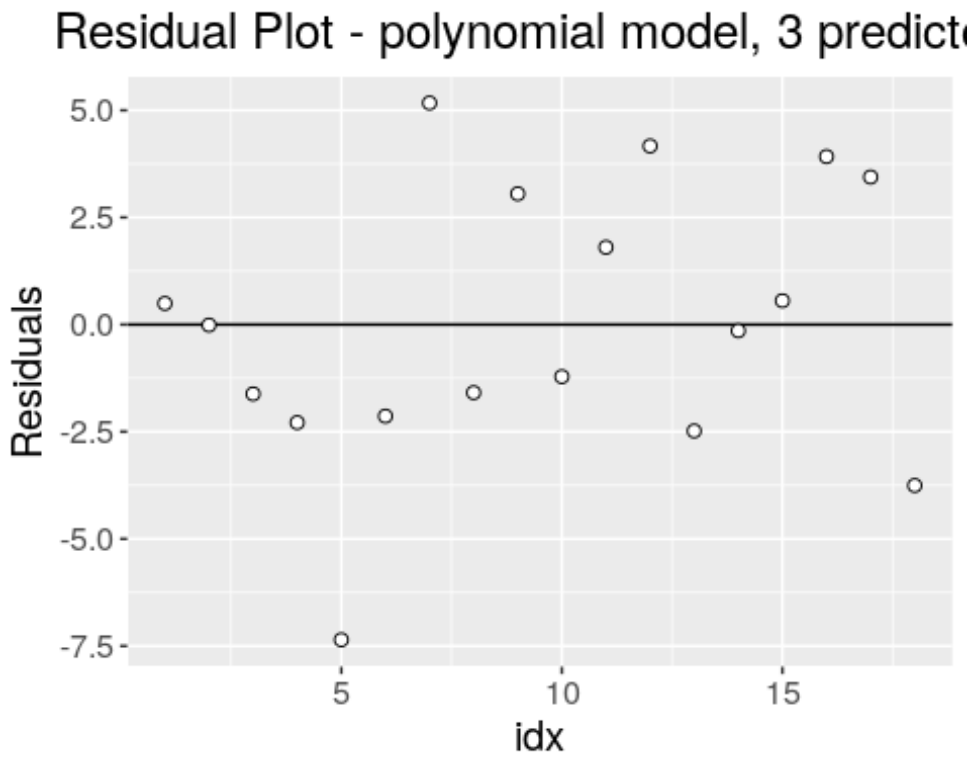
### 3C.

```
# original regression model with two predictors (Question 3A)
RDS_1.result
##
## Call:
## lm(formula = PROFIT ~ RISK + RD, data = RDS_1)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -16.555 -11.496  -2.318   6.133  27.060
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -453.1763    23.5061 -19.279 5.37e-12 ***
## RISK          29.3090     3.6686   7.989 8.76e-07 ***
## RD             4.5100     0.1538  29.333 1.16e-14 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 14.34 on 15 degrees of freedom
## Multiple R-squared:  0.9915, Adjusted R-squared:  0.9904
## F-statistic: 879.1 on 2 and 15 DF,  p-value: 2.852e-16


# modified regression model with three predictors (Question 3B)
RDS_2.result
##
## Call:
## lm(formula = PROFIT ~ RISK + RD + RD_SQ, data = RDS_1)
##
## Residuals:
```

```
##     Min      1Q  Median      3Q     Max
## -7.3566 -2.0076 -0.0788  2.7391  5.1709
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -2.454e+02  1.481e+01 -16.567 1.36e-10 ***
## RISK         2.325e+01  9.884e-01  23.522 1.18e-12 ***
## RD           1.014e+00  2.324e-01   4.365 0.000647 ***
## RD_SQ        1.757e-02  1.152e-03  15.248 4.10e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.538 on 14 degrees of freedom
## Multiple R-squared:  0.9995, Adjusted R-squared:  0.9994
## F-statistic:  9708 on 3 and 14 DF,  p-value: < 2.2e-16

# F critical values
fcv_1 = qf(p=0.05, df1=2, df2=15, lower.tail=F)
cat("F0.05,2,15 critical value for original MLR: ", fcv_1, "\n")
## F0.05,2,15 critical value for original MLR:  3.68232

fcv_2 = qf(p=0.05, df1=3, df2=14, lower.tail=F)
cat("F0.05,3,14 critical value for polynomial model: ", fcv_2)
## F0.05,3,14 critical value for polynomial model:  3.343889

# residuals and plots
res.RDS_1 = RDS_1.result$residuals
res.RDS_2 = RDS_2.result$residuals
res = data.frame(idx= c(1:length(res.RDS_1)), res.RDS_1, res.RDS_2)

ggplot(res, aes(idx, res.RDS_1)) + geom_hline(yintercept=0) + geom_point(aes(idx, res.RDS_1),
shape=21, size=2, fill='white') + labs(x = "idx", y = "Residuals") + theme(text =
element_text(size = 15), plot.title=element_text(hjust=0.5)) + ggtitle("Residuals - linear
model, 2 predictors")
```



```
ggplot(res, aes(idx, res.RDS_2)) + geom_hline(yintercept=0) + geom_point(aes(idx, res.RDS_2),
shape=21, size=2, fill='white') + labs(x = "idx", y = "Residuals") + theme(text =
element_text(size = 15), plot.title=element_text(hjust=0.5)) + ggtitle("Residual Plot -
polynomial model, 3 predictors")
```




Comparison of the overall significance for each MLR model both show very high F values of 879.1 (original model with RISK and RD) versus 9708 (polynomial model with RISK, RD, $RD^2$). The polynomial model has a higher F value, but both models are very high and much higher than their F critical values of 3.68232 (MLR) and 3.343889 (polynomial model).

Comparison of the $R^2$ values for both models also show extremely high values of 0.9915 for the original MLR and 0.9995 for the polynomial model, indicating that both account for over 99% of the variability of company profit. The polynomial model has a slightly higher $R^2$ value, but this is expected because of the additional parameter. The adjusted $R^2$ values are 0.9904 (MLR) versus 0.9994 (polynomial) provide a better comparison since these normalize the number of parameters, and the polynomial model is still slightly higher.

Comparison of the residual plots for both models shows some difference. Both residual plots show a slight positive skew to the data. The original MLR shows a bigger range of residuals from -16.555 to 27.060 and a median of -2.318, whereas the polynomial model shows a smaller range from -7.3566 to 5.1709 and a median closer to zero at -0.0788.

## Question 4.

### 4A.

```
bank = read.csv("hw2_bank.csv", header = T)

bank_1 = lm(SALARY ~ EDUCAT + EXPER + MONTHS + MALES, data = bank)
result_bank_1 = summary(bank_1)
result_bank_1
##
## Call:
## lm(formula = SALARY ~ EDUCAT + EXPER + MONTHS + MALES, data = bank)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1238.66  -352.62   -24.76   280.08  1569.23
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 3526.4221   327.7254  10.760  < 2e-16 ***
## EDUCAT        90.0203    24.6936   3.645 0.000451 ***
## EXPER          1.2690     0.5877   2.159 0.033562 *
## MONTHS        23.4062     5.2009   4.500 2.07e-05 ***
## MALES        722.4607   117.8216   6.132 2.41e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 507.4 on 88 degrees of freedom
## Multiple R-squared:  0.5109, Adjusted R-squared:  0.4886
## F-statistic: 22.98 on 4 and 88 DF,  p-value: 5.072e-13

# F critical value
fcv_bank1 = qf(p=0.05, df1=4, df2=88, lower.tail=F)
cat("F0.05,4,88 critical value for MLR: ", fcv_bank1, "\n")
## F0.05,4,88 critical value for MLR:  2.475277
```

To test the significance of the overall fit of the regression model, we will use a right-tailed test with a null hypothesis ($H_0$) of $B_1=B_2=B_3=B_4=0$, an alternative hypothesis ($H_A$) of $B_k\neq0$ for some k (at least one predictor not equal to 0), and a significance level of $\alpha=0.05$. If the P value is at or below the significance level, or the F value is beyond the F critical value, we will reject the null hypothesis. Otherwise we will fail to reject the null hypothesis.

Here we see that the F value of the model is 22.98 and is well beyond the F critical value of 2.475277, and the P value of 5.072e-13 is well below the significance level of 0.05. We reject the null hypothesis and can conclude that at least one predictor variable is not equal to 0 and that the overall fit of the MLR model relating starting salaries to years of schooling, work experience, number of months hired after 1969, and gender is significant.

### 4B.

```
tcv_bank1 = c(-1,1) * qt(p=0.025, df=88, lower.tail=F)
cat("t0.05,88 critical values for EXCHRATE: ", tcv_bank1)
## t0.05,88 critical values for EXCHRATE:  -1.98729 1.98729
```

We want to test whether salaries are different for male and female workers on average, given that the other predictors are accounted for and constant. For this significance test we will use a two-tailed test with a null hypothesis ($H_0$) of $B_4=0$, an alternative hypothesis ($H_A$) of $B_4\neq0$, and a significance level of $\alpha=0.05$. If the P value is at or below the significance level, or the t value is beyond the t critical values, we will reject the null hypothesis. Otherwise we will fail to reject the null hypothesis.

Here we see that P=2.41e-08 and is much lower than the significance level of 0.05. We also see that the t value is 6.132, which is well outside the range of calculated t critical values of ±1.98729. We reject the null hypothesis. This implies that gender (male or female), while keeping other predictors constant, significantly changes the mean response of beginning salary and supports the accusation that Harris Bank has discriminated against female employees.

## 4C.

```
# calculate point estimates for male and female
Education = 12
Experience = 10
Hired = 15
Male = 1
Female = 0

# male
salary_m = 3526.4221 + (90.0203*Education) + (1.2690*Experience) + (23.4062*Hired) +
(722.4607*Male)
cat("Mean salary of male: ", salary_m, "\n")
## Mean salary of male:  5692.909

# female
salary_f = 3526.4221 + (90.0203*Education) + (1.2690*Experience) + (23.4062*Hired) +
(722.4607*Female)
cat("Mean salary of female: ", salary_f)
## Mean salary of female:  4970.449
```

Using the MLR model above, we can predict the mean salaries for male and female employees by using the following regression equation and parameters:

SALARY = 3526.4221 + (90.0203**EDUCAT) + (1.2690**EXPER) + (23.4062**MONTHS) + (722.4607**MALES)

EDUCAT = 12 (years of education)
EXPER = 10 (years of experience)
MONTHS = 15 (time hired)

We find that the mean salary predicted by the MLR model is $5692.91 for males and $4970.45 for females.

## Question 5.

```
bank_2 = lm(SALARY ~ EDUCAT + EXPER + MONTHS + MALES + EDUCAT*EXPER, data = bank)

result_bank_2 = summary(bank_2)
result_bank_2
##
## Call:
## lm(formula = SALARY ~ EDUCAT + EXPER + MONTHS + MALES + EDUCAT *
##     EXPER, data = bank)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1195.73  -380.22    -7.68   273.68  1522.97
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  3006.1675   490.6598   6.127 2.54e-08 ***
## EDUCAT        134.4704    39.8138   3.377   0.0011 **
## EXPER           5.6792     3.1641   1.795   0.0761 .
## MONTHS         22.4205     5.2177   4.297 4.50e-05 ***
## MALES         687.6297   119.6969   5.745 1.33e-07 ***
## EDUCAT:EXPER   -0.3643     0.2569  -1.418   0.1597
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 504.5 on 87 degrees of freedom
## Multiple R-squared:  0.5219, Adjusted R-squared:  0.4945
## F-statistic:    19 on 5 and 87 DF,  p-value: 1.001e-12
```

### 5A.

The adjusted $R^2$ value for the new MLR model (using new predictor EDUCAT*EXPER) is 0.4945, which is slightly higher than the original model with an adjusted $R^2$ value of 0.4886. The new model with the new indicator does appear to be a slightly better fit.

### 5B.

```
# T test of new variable, df=93-5-1
tcv_bank2 = c(-1,1) * qt(p=0.025, df=87, lower.tail=F)
cat("t0.05,87 critical values for EDUCAT:EXPER: ", tcv_bank2)
## t0.05,87 critical values for EDUCAT:EXPER:  -1.987608 1.987608
```

We want to test whether the new predictor, EDUCAT*EXPER, significantly affects salaries, given that the other predictors are accounted for and constant. For this significance test we will use a two-tailed test with a null hypothesis ($H_0$) of $B_5=0$, an alternative hypothesis ($H_A$) of $B_5\neq0$, and a significance level of $\alpha=0.05$. If the P value is at or below the significance level, or the t statistic is beyond the t critical values, we will reject the null hypothesis. Otherwise we will fail to reject the null hypothesis.

Here we see that P value for the new predictor is 0.1597 and is higher than the significance level of 0.05. We also see that the t value of -1.418 falls within the critical values of ±1.987608. We fail to reject the null hypothesis and conclude that changing the new predictor EDUCAT*EXPER does not significantly change the mean salary while holding the other predictors constant. However, this does not mean that there is no effect of the new predictor on the response.

### 5C.

The new variable, EDUCAT*EXPER, does not significantly affect the regression model. Both models show a strong association of salary with education, experience, hiring date, and gender. Adding the new variable explains slightly more variation of the data according to their $R^2$ and adjusted $R^2$ values, however, that could be due to the addition of the new parameter into the equation (thus raising the $R^2$ a priori). In addition, the new variable does not significantly change the salary when accounting for the other predictors.