MATH 204 HW4 **260762008 Tianze Lin**

12.20

a. $y=\beta_0+\beta_1*x_1+\beta_2*x_2+\varepsilon$

b. β_1 represents the change in revenue (y) for every 1-tweet increase in the tweet rate (x1), holding PN-ratio (x2) constant.

c. β_2 represents the change in revenue (y) for every 1-tweet increase in the PN-ratio (x2), holding tweet rate (x1) constant.

d. The R^2 is 0.945. This implies 94.5% of variance in y (revenue) is explained by independent variables of our regression model. The R_a^2 value is .940. This implies that the least squares model has explained about 94.0% of the total sample variation in y values (revenue), after adjusting for sample size and number of independent variables in the model.

e. H_0: $\beta_1=\beta_2=0$

H_a: At least one of the two model coefficients is nonzero

$F_C=\frac{R^2/k}{(1-R^2)/[n-(k+1)]}=\frac{0.945/2}{(1-0.945)/[24-(2+1)]}=180.41$

$F_\alpha=92.82$

```
> qt(1-0.05,2,24-3)
[1] 92.81763
```

Our test statistic is larger than F_α, so we reject the null hypothesis. There is no evidence that both model coefficients are zero when $\alpha = 0.05$.

f. Since $\alpha = 0.01$ exceeds the p-value (0.0001), the data provide strong evidence that both model coefficients are nonzero. Therefore, the overall model appears to be statistically useful in predicting revenue.

12.24

a. H_0: $\beta_1=\beta_2=\beta_3=\beta_4=0$

H_a: At least one model coefficient is nonzero

$\alpha = 0.10$ exceeds the p-value (0.091), so we reject the null hypothesis. Therefore, the data provide evidence that at least one model coefficient is nonzero.

b. The R_a^2 value is .629. This implies that the least squares model has explained about 62.9% of the total sample variation in y values (grafting efficiency), after adjusting for sample size and number of independent variables in the model.

c. s is the mean squared error. It means the average of the squares of the differences between the estimator and what is estimated is 11.2206.

d. $\hat{\beta}_3 + t_{\alpha/2} \cdot S_{\hat{\beta}_3} = 0.4330 + t_{0.05}^{9-(4+1)} \cdot 0.3054 = 1.08$

$\hat{\beta}_3 - t_{\alpha/2} \cdot S_{\hat{\beta}_3} = 0.4330 - t_{0.05}^{9-(4+1)} \cdot 0.3054 = -0.22$

The 90% confidence interval for β_3 is $(-0.22, 1.08)$

e. H_0: $\beta_4=0$

H_a: $\beta_4\neq0$

Since $\alpha = 0.10$ is smaller than the p-value (0.503), we cannot reject the null hypothesis. Therefore, the data does not provide evidence that β_4 is nonzero. The reaction temperature doesn't appear to be statistically useful in predicting grafting efficiency.

12.28

a.

```
> bubble <- read_csv("D:/学习/McGill/U0/Winter/Math204/Assignment/A4/BUBBLE2.csv")
Parsed with column specification:
cols(
  Label = col_character(),
  MassFlux = col_integer(),
  HeatFlux = col_double(),
  Diameter = col_double(),
  Density = col_double()
)
> head(bubble)
# A tibble: 6 x 5
  Label  MassFlux HeatFlux Diameter Density
  <chr>     <int>    <dbl>    <dbl>   <dbl>
1 P4-145      406    0.150    0.640   13103
2 P4-148      406    0.290    1.02    29117
3 P4-149      406    0.370    1.15   123021
4 P4-150      406    0.620    1.26   165969
5 P4-151      406    0.860    0.910  254777
6 P4-152      406    1.00     0.680  347953
> bubble_a <- lm(Diameter~MassFlux+HeatFlux, data=bubble)
> summary(bubble_a)

Call:
lm(formula = Diameter ~ MassFlux + HeatFlux, data = bubble)

Residuals:
     Min       1Q   Median       3Q      Max 
-0.34129 -0.23205  0.04017  0.15505  0.32121 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept)  1.0884141  0.1837825   5.922  2.8e-05 ***
MassFlux    -0.0002343  0.0001737  -1.348    0.198    
HeatFlux    -0.0800181  0.1877160  -0.426    0.676    
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.2441 on 15 degrees of freedom
Multiple R-squared:  0.1177,	Adjusted R-squared:  7.021e-06 
F-statistic:     1 on 2 and 15 DF,  p-value: 0.3911
```

H_0: $\beta_1=\beta_2=0$

H_a: At least one model coefficient is nonzero

Choose $\alpha = 0.10$. Since $\alpha = 0.10$ is smaller than the p-value (0.3911), we cannot reject the null hypothesis. Therefore, the data does not provide evidence that β_1 or β_2 is nonzero. The mass flux and heat flux don't appear to be statistically useful in predicting diameter.

b.

```
> bubble_b <- lm(Density~MassFlux+HeatFlux, data=bubble)
> summary(bubble_b)

Call:
lm(formula = Density ~ MassFlux + HeatFlux, data = bubble)

Residuals:
   Min     1Q Median     3Q    Max 
-42636 -19706  -9202  26264  40453 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept)  -1030.03   21237.16  -0.049   0.9620    
MassFlux       -57.90      20.08  -2.884   0.0114 *  
HeatFlux    332037.09   21691.71  15.307 1.46e-10 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 28200 on 15 degrees of freedom
Multiple R-squared:  0.9418,	Adjusted R-squared:  0.934 
F-statistic: 121.3 on 2 and 15 DF,  p-value: 5.474e-10
```

H_0: $\beta_1=\beta_2=0$

H_a: At least one model coefficient is nonzero

Choose $\alpha = 0.10$. Since $\alpha = 0.10$ is greater than the p-values (5.474e-10), we reject the null hypothesis. Therefore, the data provides strong evidence that β_1 and β_2 are nonzero. The mass flux and heat flux appear to be statistically useful in predicting density.

c.

Density is better predicted by mass flux and heat flux.

12.40

a.

```
> boiler <- read_csv("D:/学习/McGill/U0/Winter/Math204/Assignment/A4/BOILERS.csv")
Parsed with column specification:
cols(
  ManHours = col_integer(),
  Capacity = col_integer(),
  Pressure = col_integer(),
  Boiler = col_integer(),
  Drum = col_integer()
)
```

```
> boiler_model <- lm(ManHours~Capacity+Pressure+Boiler+Drum,data=boiler)
> summary(boiler_model)

Call:
lm(formula = ManHours ~ Capacity + Pressure + Boiler + Drum,
    data = boiler)

Residuals:
     Min       1Q   Median       3Q      Max 
-1612.66  -549.18   -12.38   406.97  2768.66 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept) -3.783e+03  1.205e+03  -3.139 0.003711 ** 
Capacity     8.749e-03  9.035e-04   9.684 6.86e-11 ***
Pressure     1.926e+00  6.489e-01   2.969 0.005723 ** 
Boiler       3.444e+03  9.117e+02   3.778 0.000675 ***
Drum         2.093e+03  3.056e+02   6.849 1.12e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 894.6 on 31 degrees of freedom
Multiple R-squared:  0.903,	Adjusted R-squared:  0.8904 
F-statistic: 72.11 on 4 and 31 DF,  p-value: 2.977e-15
```

The prediction equation is $E(y)=-3783+0.008749\beta_1+1.926\beta_2+3444\beta_3+2093\beta_4$.

b. H_0: $\beta_1=\beta_2=\beta_3=\beta_4=0$

H_a: At least one model coefficient is nonzero

Since $\alpha = 0.01$ is greater than the p-value (2.977e-15), we reject the null hypothesis. Therefore, the data provides strong evidence that at least one model coefficient is nonzero. The model appears to be statistically useful in predicting hours.

c.

```
> predict(boiler_model,newdata=data_frame(Capacity=150000,Pressure=500,Boiler=1,Drum=0),interval="confidence",se.fit=T,level=0.95)
$fit
       fit     lwr      upr
1 1936.412 1448.65 2424.174

$se.fit
[1] 239.1562

$df
[1] 31

$residual.scale
[1] 894.6032
```

The 95% confidence interval for E(y) is (1449,2424). It means if we were to repeat our experiment multiple times, i.e. collect the data repeatedly in the same way, and we were to compute a confidence interval using the same recipe for each data set, then approximately 95% of our calculated intervals (1449,2424) would contain the true E(y).

d. We should use prediction interval.

```
> predict(boiler_model,newdata=data_frame(Capacity=150000,Pressure=500,Boiler=1,Drum=0),interval="prediction",se.fit=T,level=0.95)
$fit
       fit      lwr      upr
1 1936.412 47.78441 3825.039

$se.fit
[1] 239.1562

$df
[1] 31

$residual.scale
[1] 894.6032
```

The prediction interval is (47.79,3825).

12.58

a. $E(y)=\beta_0+\beta_1*x_1+\beta_2*x_2+\beta_3*x_3+\beta_4*x_1*x_3+\beta_5*x_2*x_3$

b.

```
> aswells <- read_csv("D:/学习/McGill/U0/Winter/Math204/Assignment/A4/ASWELLS.csv")
Parsed with column specification:
cols(
  WELLID = col_integer(),
  UNION = col_character(),
  VILLAGE = col_character(),
  LATITUDE = col_double(),
  LONGITUDE = col_double(),
  `DEPTH-FT` = col_integer(),
  YEAR = col_integer(),
  `KIT-COLOR` = col_character(),
  ARSENIC = col_integer()
)

> aswells_a <- lm(ARSENIC~LATITUDE+LONGITUDE+`DEPTH-FT`+LATITUDE*`DEPTH-FT`+LONGITUDE*`DEPTH-FT`,data=aswells)

> summary(aswells_a)

Call:
lm(formula = ARSENIC ~ LATITUDE + LONGITUDE + `DEPTH-FT` + LATITUDE * 
    `DEPTH-FT` + LONGITUDE * `DEPTH-FT`, data = aswells)

Residuals:
    Min      1Q  Median      3Q     Max 
-175.75  -65.04  -23.02   29.82  480.01 

Coefficients:
                      Estimate Std. Error t value Pr(>|t|)  
(Intercept)           10845.07   67720.06   0.160   0.8729  
LATITUDE              -1279.76    1053.11  -1.215   0.2252  
LONGITUDE               217.40     814.50   0.267   0.7897  
`DEPTH-FT`            -1549.22     985.58  -1.572   0.1170  
LATITUDE:`DEPTH-FT`     -11.00      11.86  -0.927   0.3547  
LONGITUDE:`DEPTH-FT`     19.98      11.20   1.783   0.0755 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 103.1 on 321 degrees of freedom
  (1 observation deleted due to missingness)
Multiple R-squared:  0.1372,	Adjusted R-squared:  0.1238 
F-statistic: 10.21 on 5 and 321 DF,  p-value: 4.306e-09
```

The least squares prediction equation is

$E(y)=10845.07-1279.76*x_1+217.40*x_2-1549.22*x_3-11*x_1*x_3+19.98*x_2*x_3$

c. H_0: $\beta_4=0$

H_a: $\beta_4\neq0$

Since $\alpha = 0.05$ is smaller than the p-value (0.3507), the data doesn't provide strong evidence that β_4 is nonzero. Therefore, we fail to reject the null hypothesis when $\alpha = 0.05$. The interaction between latitude and depth will not affect the arsenic level.

d. H_0: $\beta_5=0$

H_a: $\beta_5 \neq 0$

Since $\alpha = 0.05$ is smaller than the p-value (0.0755), the data doesn't provide strong evidence that β_5 is nonzero. Therefore, we fail to reject the null hypothesis when $\alpha = 0.05$. The interaction between longitude and depth will not affect the arsenic level.

e. We fail to reject the null hypotheses in c and d. Therefore, the arsenic level is not affected by the interaction between latitude and depth and the interaction between longitude and depth.

12.98

a. Set x_1 is 1 if the slice is in Group B and 0 if the slice is not in Group B.

Set x_2 is 1 if the slice is in Group C and 0 if the slice is not in Group C.

$E(y)=\beta_0+\beta_1*x_1+\beta_2*x_2$

b.

```
> library(tidyverse)
> sand <- read_csv("D:/学习/McGill/U0/Winter/Math204/Assignment/A4/SAND.csv")
Parsed with column specification:
cols(
  PermA = col_double(),
  PermB = col_double(),
  PermC = col_double()
)
```

```
> summary(sand)
     PermA             PermB            PermC       
 Min.   : 55.20   Min.   : 50.4   Min.   : 52.20  
 1st Qu.: 62.20   1st Qu.:109.0   1st Qu.: 67.97  
 Median : 70.45   Median :139.3   Median : 78.65  
 Mean   : 73.62   Mean   :128.5   Mean   : 83.07  
 3rd Qu.: 81.28   3rd Qu.:146.9   3rd Qu.: 95.25  
 Max.   :122.40   Max.   :150.0   Max.   :129.00  
```

β_0=mean(PermA)=73.62

β_1= mean(PermB)- mean(PermA)=128.5-73.62=54.88

β_2= mean(PermC)- mean(PermA)=83.07-73.62=9.45

c.

Change the file into the following format.

	A	B	C
1	Group	Permeability	
2	PermA	55.4	
3	PermA	57.2	
4	PermA	59.7	
5	PermA	57.9	
6	PermA	59.9	
7	PermA	59.3	
8	PermA	59.9	
9	PermA	58.3	
10	PermA	56.2	
11	PermA	57.4	
12	PermA	58.4	
13	PermA	[illegible]	

```
> sand <- read_csv("D:/学习/McGill/U0/Winter/Math204/Assignment/A4/SAND2.csv")
Parsed with column specification:
cols(
  Group = col_character(),
  Permeability = col_double()
)

> sand_model <- lm(Permeability~Group,data=sand)
> summary(sand_model)

Call:
lm(formula = Permeability ~ Group, data = sand)

Residuals:
    Min      1Q  Median      3Q     Max
-78.137 -13.723  -1.797  17.163  48.777

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   73.623      1.910  38.550  < 2e-16 ***
GroupPermB    54.914      2.701  20.332  < 2e-16 ***
GroupPermC     9.447      2.701   3.498 0.000541 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 19.1 on 297 degrees of freedom
Multiple R-squared:  0.6141,    Adjusted R-squared:  0.6115
F-statistic: 236.3 on 2 and 297 DF,  p-value: < 2.2e-16
```

The output coefficients of this model are approximately equal to the estimated values of β_0, β_1, and β_2. This shows our β estimate in part b are correct.

12.132

a.

If the researchers' belief is true, then the expected sign of β_2 in the model is negative.

b.

SUMMARY OUTPUT

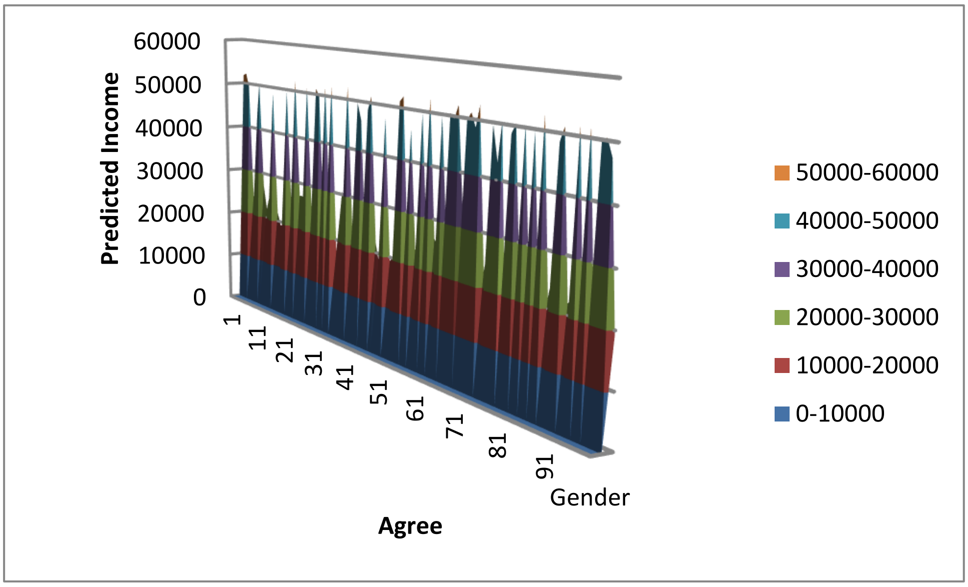
Regression Statistics	
Multiple R	0.874641
R Square	0.764997
Adjusted R	0.757653
Standard E	7737.365
Observatic	100

ANOVA

	df	SS	MS	F	ignificance F
Regressioı	3	1.87E+10	6.24E+09	104.1683	4.48E-30
Residual	96	5.75E+09	59866814		
Total	99	2.45E+10			

	Coefficients	tandard Err	t Stat	P-value	Lower 95%	Upper 95%	ower 95.0%	Jpper 95.0%
Intercept	-21657.1	31779.75	-0.68147	0.497212	-84739.4	41425.22	-84739.4	41425.22
Gender	25482.28	1551.535	16.42391	1.23E-29	22402.5	28562.05	22402.5	28562.05
Agree	37154.58	19257.06	1.9294	0.056634	-1070.39	75379.55	-1070.39	75379.55
Agree^2	-7056.46	2903.202	-2.43058	0.016932	-12819.3	-1293.65	-12819.3	-1293.65

Based on the E(y), we can get this plot.



c.

$E(y)=\beta_0+\beta_1*x_1+\beta_2*x_1^2+\beta_3*x_2+\beta_4*x_2^2+\beta_5*x_1*x_2$

d.

SUMMARY OUTPUT

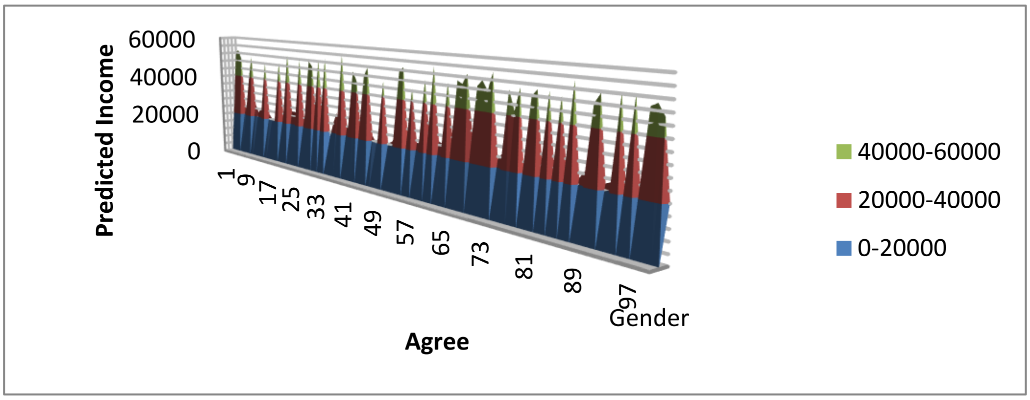
Regression Statistics	
Multiple R	0.876963
R Square	0.769064
Adjusted R	0.748814
Standard E	7710.378
Observatio	100

ANOVA

	df	SS	MS	F	ignificance F
Regressior	5	1.88E+10	3.76E+09	79.09233	3.86E-32
Residual	95	5.65E+09	59449934		
Total	100	2.45E+10			

	Coefficients	tandard Err	t Stat	P-value	Lower 95%	Upper 95%	ower 95.0%	Jpper 95.0%
Intercept	-10460.66	32830.52	-0.318626	0.75071	-75637.49	54716.17	-75637.49	54716.17
Gender	0	0	65535	#NUM!	0	0	0	0
Gender^2	43937.36	14350.92	3.06164	#NUM!	15447.18	72427.54	15447.18	72427.54
Agree	27643.45	20550.37	1.345156	0.181777	-13154.19	68441.1	-13154.19	68441.1
Agree^2	-5230.977	3218.935	-1.625064	0.107462	-11621.37	1159.417	-11621.37	1159.417
G_A	-5615.256	4341.083	-1.293515	0.198969	-14233.4	3002.885	-14233.4	3002.885

Based on the E(y), we can get this plot.



e.

H_0: $\beta_4=\beta_5=0$

f.

H_0: $\beta_4=\beta_5=0$

H_a: at least one is nonzero

```
> wagap <- read_csv("D:/学习/McGill/U0/Winter/Math204/Assignment/A4/WAGAP.csv")
Parsed with column specification:
cols(
  Gender = col_integer(),
  Agree = col_double(),
  G_A = col_double(),
  Income = col_integer()
)
> wagap_a <- lm(Income~Agree+I(Agree^2)+Gender,data=wagap)
> wagap_c <- lm(Income~Agree+Gender+Agree*Gender+I(Agree^2)+I(Gender^2),data=wagap)
```

```
> anova(wagap_a,wagap_c)
Analysis of Variance Table

Model 1: Income ~ Agree + I(Agree^2) + Gender
Model 2: Income ~ Agree + Gender + Agree * Gender + I(Agree^2) + I(Gender^2)
  Res.Df        RSS Df Sum of Sq      F Pr(>F)
1     96 5747214158
2     95 5647743687  1  99470471 1.6732  0.199
```

Since α = 0.10 is smaller than the p-value (0.199), we fail to reject the null hypothesis when α = 0.10. Therefore, the second-order term (x_2^2) of first model didn't appear to be statistically useful in predicting income.

12.140

a. In step one of the stepwise regression, 8 different one-variable models are fitted to the data.

b. The "best" one-variable model must have the smallest p-value. In this case, comparing to other models, the model contains x1 must have smallest p-value.

c. In step two of the stepwise regression, 7 different two-variable models are fitted to the data.

d. The beta coefficient is the degree of change in the outcome variable for every 1-unit of change in the predictor variable. In this case, β_1(-0.28) means for every 1-unit increase in the x1(company role of estimator), if we hold x8(previous accuracy) constant, the y(effort) will decrease by 0.28. β_2(0.27) means for every 1-unit increase in the x8(previous accuracy), if we hold x1(company role of estimator) constant, the y(effort) will increase by 0.27.

e. There are two reasons. First, the result of stepwise procedure is a model containing only those terms with t-values that are significant at the specified a level. Thus, only several of the large number of independent variables remain. However, this doesn't mean that all the independent variables that are important in predicting y have been identified or that the unimportant independent variables have been eliminated. Second, when we choose the variables to be included in the stepwise regression, we often omit high-order terms. Consequently, we may have initially omitted several important terms from the model. Source: P127 128 of textbook