MATH 204 HW3 Tianze Lin 260762008

11.36

```
> tweet <- read_csv("D:/学习/McGill/U0/Winter/Math204/Assignment/A3/TWEETS.csv")
Parsed with column specification:
cols(
  TweetRate = col_double(),
  `Revenue (millions)` = col_double()
)
> head(tweet)
# A tibble: 6 x 2
  TweetRate `Revenue (millions)`
      <dbl>                <dbl>
1      1366                142
2      1213                 77.0
3       582                 61.0
4       310                 32.0
5       455                 31.0
6       290                 30.0
> t.model <- lm(`Revenue (millions)`~TweetRate,data=tweet)

> summary(t.model)

Call:
lm(formula = `Revenue (millions)` ~ TweetRate, data = tweet)

Residuals:
    Min      1Q  Median      3Q     Max
-36.751  -2.302   2.468   5.083  33.270

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 1.150154   3.676108   0.313    0.757
TweetRate   0.078767   0.007938   9.923 2.22e-09 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 13.32 on 21 degrees of freedom
Multiple R-squared:  0.8242,	Adjusted R-squared:  0.8159
F-statistic: 98.47 on 1 and 21 DF,  p-value: 2.217e-09
```

Revenue=0.078767*TweetRate+1.150154=0.078767*100+1.150154=9.026854
The estimation of a movie's opening weekend revenue change as the tweet rate for the movie increases by an average of 100 tweets per hour is 9.026854 million.

11.70

Ley y be the blood lactate level and x be the perceived recovery.

$\hat{y}=B_1x+B_0$

H0: $B_1=0$

Ha: $B_1\neq 0$

```
> box <- read_csv("D:/学习/McGill/U0/Winter/Math204/Assignment/A3/BOXING2.csv")
Parsed with column specification:
cols(
  LACTATE = col_double(),
  RECOVERY = col_integer()
)
```

SUMMARY OUTPUT									
Regression Statistics									
Multiple R	0.570183								
R Square	0.325108								
Adjusted R	0.276902								
Standard E	4.279658								
Observatic	16								
ANOVA									
	df	SS	MS	F	Significance F				
Regressioi	1	123.5208	123.5208	6.744069	0.021102849				
Residual	14	256.4167	18.31548						
Total	15	379.9375							
	Coefficients	andard Err	t Stat	P-value	Lower 95%	Upper 95%	ower 90.0%	Jpper 90.0%	
Intercept	2.796667	4.983797	0.561152	0.583565	-7.89251377	13.48585	-5.98134	11.57468	
X Variable	2.566667	0.988345	2.596935	0.021103	0.446877947	4.686455	0.825885	4.307448	

The p-value is 0.0211, which is smaller than our α(0.10). Therefore, we reject the null hypothesis, which means there is evidence to prove that the blood lactate level is linearly related to perceived recovery.

11.76

Let x be the time and y be the mass of the spill.

$\hat{y}=B_1x+B_0$

H0: $B_1=0$

Ha: $B_1\neq0$

SUMMARY OUTPUT								
Regression Statistics								
Multiple R	0.92376344							
R Square	0.853338893							
Adjusted R Square	0.846355031							
Standard Error	0.857257302							
Observations	23							
ANOVA								
	df	SS	MS	F	Significance F			
Regression	1	89.79419524	89.7942	122.1872461	3.25999E-10			
Residual	21	15.43269171	0.73489					
Total	22	105.226887						
	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%	Lower 95.0%	Upper 95.0%
Intercept	5.220695364	0.295977687	17.6388	4.55174E-14	4.605176069	5.83621466	4.605176069	5.83621466
X Variable 1	-0.114022801	0.010315226	-11.0538	3.25999E-10	-0.135474489	-0.092571113	-0.135474489	-0.092571113

The p-value is 3.26E-10, which is smaller than 0.05. Therefore, we reject the null hypothesis. When α=0.05, there is sufficient evidence to indicate that the mass of the spill tends to diminish linearly as elapsed time increases.

From the table above, the 95% confidence interval for B1 is (-0.135,-0.092).

11.92

a. The correlation between height and average earning for people in sales occupations is 0.41. This number is positive, which means the average earning will increase as the height increases.

b. r^2=0.41*0.41=0.1681

r^2 is 0.1681, which means the percentage of variance in y (averages earnings) explained by the regression model is 16.81%.

c.

H_0: $\rho=0$

Ha: $\rho>0$

d.

Test statistic t_c=r*sqrt((n-2)/(1-r^2)=0.41*sqrt(115/(1-0.41^2))=4.82

e.

Critical value:

```
> qt(1-0.01,117-2)
[1] 2.359212
```

$t_c > t_\alpha$

At the significance level α=0.01, we reject the null hypothesis. Therefore, there is evidence to indicate that average earnings and height are positively related.

f. I select managers.

The correlation between height and average earning for people in sales occupations is 0.35. This number is positive, which means the average earning will increase as the height increases.

r^2=0.35*0.35=0.1225

r^2 is 0.1225, which means the percentage of variance in y (averages earnings) explained by the regression model is 12.25%.

H_0: $\rho=0$

Ha: $\rho>0$

Test statistic t_c=r*sqrt((n-2)/(1-r^2)=0.35*sqrt(453/(1-0.35^2))=7.95

```
> qt(1-0.01,455-2)
[1] 2.334608
```

$t_c > t_\alpha$

At the significance level α=0.01, we reject the null hypothesis. Therefore, there is evidence to indicate that average earnings and height are positively related.

11.120

```
> data <- read_csv("D:/学习/McGill/U0/Winter/Math204/Assignment/A3/TWEETS.csv")
Parsed with column specification:
cols(
  TweetRate = col_double(),
  `Revenue (millions)` = col_double()
)

> model <- lm(`Revenue (millions)`~TweetRate,data=data)
```

```
> predict(model,newdata=data_frame(TweetRate=c(150)),se.fit=T,interval=c("prediction"),level=0.90)
$fit
       fit       lwr      upr
1 12.96524 -10.53561 36.46608

$se.fit
[1] 3.032267

$df
[1] 21

$residual.scale
[1] 13.31651
```

The 90% prediction interval for the revenue of a movie with a tweet rate of 150 tweets per hour is (-10.53,36.46). It means we are 90% confident that with a tweet rate of 150 tweets per hour, the revenue of the movie will fall between -10.53 and 36.46

11.124

$\hat{y}=B_1x+B_0$

H0: $B_1=0$

Ha: $B_1\neq0$

```
> wb <- read_csv("D:/学习/McGill/U0/Winter/Math204/Assignment/A3/GMAC.csv")
Parsed with column specification:
cols(
  `WLB-SCORE` = col_double(),
  HOURS = col_integer()
)
> head(wb)
# A tibble: 6 x 2
  `WLB-SCORE` HOURS
        <dbl> <int>
1        75.2    50
2        65.0    45
3        49.6    50
4        44.5    55
5        70.1    50
6        54.7    60

> model <- lm(`WLB-SCORE`~HOURS,data=wb)
> summary(model)

Call:
lm(formula = `WLB-SCORE` ~ HOURS, data = wb)

Residuals:
    Min      1Q  Median      3Q     Max 
-35.477  -8.412  -0.652   8.121  33.525 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 62.49851    1.41351   44.22   <2e-16 ***
HOURS       -0.34673    0.02761  -12.56   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 12.28 on 2085 degrees of freedom
Multiple R-squared:  0.07033,	Adjusted R-squared:  0.06988 
F-statistic: 157.7 on 1 and 2085 DF,  p-value: < 2.2e-16
```

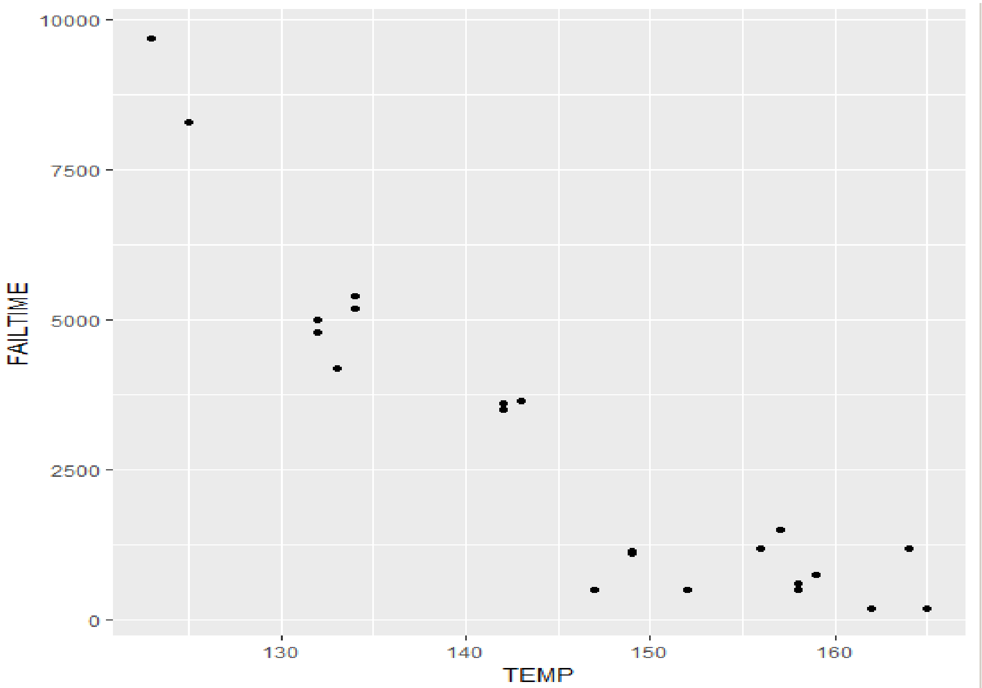
ŷ= -0.34673x+62.49851

Choose α=0.05. The p-value is smaller than α, so we reject the null hypothesis, which means there is evidence to prove that the work-life balance scale score is linearly related to average number of hours worked per week. r^2 is 0.07033, which means the percentage of variance in y (WLB Score) explained by the regression model is 7.033%.

12.78

a.

```
> wafer <- read_csv("D:/学习/McGill/U0/Winter/Math204/Assignment/A3/WAFER.csv")
Parsed with column specification:
cols(
  TEMP = col_integer(),
  FAILTIME = col_integer()
)
> head(wafer)
# A tibble: 6 x 2
   TEMP FAILTIME
  <int>    <int>
1   165      200
2   162      200
3   164     1200
4   158      500
5   158      600
6   159      750
> ggplot(wafer)+geom_point(aes(TEMP,FAILTIME))
```



Curvilinear relationship appears to exist between failure time and solder temperature.

b.

```
> wafer_quad <- lm(FAILTIME~TEMP+I(TEMP^2),data=wafer)
> summary(wafer_quad)

Call:
lm(formula = FAILTIME ~ TEMP + I(TEMP^2), data = wafer)

Residuals:
     Min       1Q   Median       3Q      Max 
-1260.49  -475.70   -15.57   528.45  1131.69 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept) 154242.914  21868.474   7.053 1.03e-06 ***
TEMP         -1908.850    303.664  -6.286 4.92e-06 ***
I(TEMP^2)        5.929      1.048   5.659 1.86e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 688.1 on 19 degrees of freedom
Multiple R-squared:  0.9415,	Adjusted R-squared:  0.9354 
F-statistic: 152.9 on 2 and 19 DF,  p-value: 1.937e-12
```

$\hat{y}$=154242.914-1908.850x+5.929x^2

c.

H0: B_2=0

Ha: B_2>0

The p-value is 1.86e-05, which is smaller than α(0.05), so we reject the null hypothesis. There is evidence to prove the upward curvature in the relationship between failure time and solder temperature.

12.162

a.

```
> wafer <- read_csv("D:/学习/McGill/U0/Winter/Math204/Assignment/A3/WAFER.csv")
Parsed with column specification:
cols(
  TEMP = col_integer(),
  FAILTIME = col_integer()
)

> wafer_line <- lm(FAILTIME~TEMP,data=wafer)
> summary(wafer_line)

Call:
lm(formula = FAILTIME ~ TEMP, data = wafer)

Residuals:
    Min      1Q  Median      3Q     Max 
-2195.5  -719.0    12.9   371.8  2406.8 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 30855.91    2713.28   11.37 3.49e-10 ***
TEMP         -191.57      18.49  -10.36 1.74e-09 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1099 on 20 degrees of freedom
Multiple R-squared:  0.8429,	Adjusted R-squared:  0.8351 
F-statistic: 107.3 on 1 and 20 DF,  p-value: 1.741e-09
```

Fitting the straight-line model to the data, we get E(y)=30855.91-191.57x

b.

ŷ=30855.91-191.57*152=1737.27

When x= 152, y = 500.

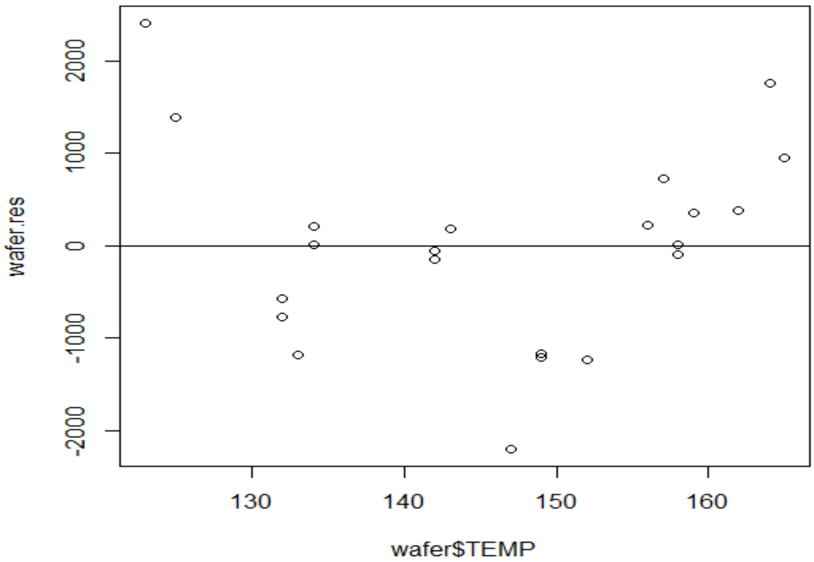
Residual=y-ŷ=500-1737.27=-1237.27

The residual for a microchip manufactured at a temperature of 152 degree is -1237.27.

c.

```
> wafer.res <- resid(wafer_line)

> plot(wafer$TEMP,wafer.res)
> abline(0,0)
```



It has a U-shape and looks like a quadratic function.

d.

Yes. This plot has a clear shape, which means our straight-line model has room for improvement. The U-shape also indicates that failure time and solder temperature are curvilinearly related.