High-Dimensional Data Analysis

Homework#1

Student name: Lin Ting-Yu

1.Full in all values

Simple regression									
	$\widehat{\beta_0}$	$\widehat{\beta_1}$	$se(\widehat{\beta_0})$	$se(\widehat{\beta_1})$	t_0	t_1	Rss	P_0	P_1
TV(x_1)	7.033	0.048	0.458	0.003	15.360	17.668	2102.531	≈ 0	≈ 0
radio(x_2)	9.312	0.202	0.563	0.020	16.542	9.921	3618.48	≈ 0	≈ 0
newspaper(x_3)	12.351	0.055	0.621	0.017	19.876	3.3	5134.805	≈ 0	0.001148196
CC	2.529	0.860	0.345	0.025	7.335	34.321	779.557	≈ 0	≈ 0
H_0	$H_0: \beta_{TV} = \beta_{radio} = \beta_{newspaper} = 0$								

Multiple regression				
	$\widehat{\beta_i}$	$se(\widehat{\beta_i})$	t_0	p
Intercept	2.938889369	0.311908	9.422	≈ 0
TV(x_{i1})	0.045764645	0.001395	32.809	≈ 0
Radio(x_{i2})	0.188530017	0.008611	21.893	≈ 0
Newspaper(x_{i3})	-0.001037493	0.005871	-0.177	0.86
Rss	556.8253			
F	F=570.2707			
H_0	$H_0: \beta_{TV} = \beta_{radio} = \beta_{newspaper} = 0$			

2.Use mathematical formulas to explain how to calculate all the values.

For simple linear regression

TV:

$$\overline{x_1} = \frac{\sum_{i=1}^{200} x_i}{200} = 147.0425\,, \overline{y_{sales}} = \frac{\sum_{i=1}^{200} y_i}{200} = 14.0225$$

$$S_{x_1x_1} = \sum_{i=1}^{200}(x_i - \overline{x_1})^2 = 1466819\,,\ S_{x_1y} = \sum_{i=1}^{200}(x_i - \overline{x_1})(y_i - \overline{y_{sales}}) = 69727.65$$

$$\Rightarrow \widehat{\beta_1} = \frac{S_{x_1y}}{S_{x_1x_1}} = \frac{69727.65}{1466819} = 0.04753664\ ,\qquad \widehat{\beta_0} = \overline{y_{sales}} - \widehat{\beta_1} \times \overline{x_1} = 7.032594$$

$$RSS = \sum_{i=1}^{200}(y_i - \hat{y}_i)^2 = \Sigma\left(y_i - \widehat{\beta_0} - \widehat{\beta_1} \times x_i\right)^2 = 2102.531$$

$$\widehat{\sigma^2} = \frac{1}{198}\sum_{i=1}^{200}\left(y_i - \widehat{\beta_0} - \widehat{\beta_1} \times x_i\right)^2 = 10.61884 \quad \Rightarrow\ \hat{\sigma} = 3.258656$$

$$\Rightarrow se\left(\widehat{\beta_0}\right) = \hat{\sigma}\sqrt{\frac{1}{n} + \frac{\overline{x_1}^2}{S_{x_1x_1}}} = 0.4578429\,, se\left(\widehat{\beta_1}\right) = \sqrt{\frac{\hat{\sigma}^2}{S_{x_1x_1}}} = 0.002690607$$

$$\Rightarrow t - statistic\ t_0 = \frac{\widehat{\beta_0}}{se\left(\widehat{\beta_0}\right)} = \frac{7.032594}{0.002690607} = 15.36028\,, p - value = Pr\left(\left|t_{df=198}\right| > |t|\right) \approx 0$$

$$t - statistic\ t_1 = \frac{\widehat{\beta_1}}{se\left(\widehat{\beta_1}\right)} = \frac{0.04753664}{0.002690607} = 17.66763, p - value = Pr\left(\left|t_{df=198}\right| > |t|\right) \approx 0$$

radio:

$$\overline{x_2} = \frac{\sum_{i=1}^{200} x_i}{200} = 23.264\,, \overline{y_{sales}} = \frac{\sum_{i=1}^{200} y_i}{200} = 14.0225$$

$$S_{x_2x_2} = \sum_{i=1}^{200}(x_i - \overline{x_2})^2 = 43.86512\,,\ S_{x_2y} = \sum_{i=1}^{200}(x_i - \overline{x_2})(y_i - \overline{y_{sales}}) = 8882.502$$

$$\Rightarrow \widehat{\beta_1} = \frac{S_{x_2y}}{S_{x_2x_2}} = \frac{8882.502}{43.86512} = 0.2024958\ ,\qquad \widehat{\beta_0} = \overline{y_{sales}} - \widehat{\beta_1} \times \overline{x_2} = 9.311638$$

$$RSS = \sum_{i=1}^{200}(y_i - \hat{y}_i)^2 = \Sigma\left(y_i - \widehat{\beta_0} - \widehat{\beta_1} \times x_i\right)^2 = 3618.48$$

$$\widehat{\sigma^2} = \frac{1}{198}\sum_{i=1}^{200}\left(y_i - \widehat{\beta_0} - \widehat{\beta_1} \times x_i\right)^2 = 18.27515 \quad \Rightarrow\ \hat{\sigma} = 4.274944$$

$$\Rightarrow se\left(\widehat{\beta_0}\right) = \hat{\sigma}\sqrt{\frac{1}{200} + \frac{\overline{x_2}^2}{S_{x_2x_2}}} = 0.5629005\,, se\left(\widehat{\beta_1}\right) = \sqrt{\frac{\hat{\sigma}^2}{S_{x_2x_2}}} = 0.02041131$$

$$\Rightarrow t - statistic\ t_0 = \frac{\widehat{\beta_0}}{se\left(\widehat{\beta_0}\right)} = \frac{9.311638}{0.5629005} = 16.54225, p - value = Pr\left(\left|t_{df=198}\right| > |t|\right) \approx 0$$

$$t - statistic\ t_1 = \frac{\widehat{\beta_1}}{se\left(\widehat{\beta_1}\right)} = \frac{0.2024958}{0.02041131} = 9.920765, p - value = Pr\left(\left|t_{df=198}\right| > |t|\right) \approx 0$$

newspaper:

$$\overline{x_3} = \frac{\sum_{i=1}^{200} x_i}{200} = 30.554\,, \overline{y_{sales}} = \frac{\sum_{i=1}^{200} y_i}{200} = 14.0225$$

$$S_{x_3x_3} = \sum_{i=1}^{200}(x_i - \overline{x_3})^2 = 94387.36\,,\ S_{x_3y} = \sum_{i=1}^{200}(x_i - \overline{x_3})(y_i - \overline{y_{sales}}) = 5162.337$$

$$\Rightarrow \widehat{\beta_1} = \frac{S_{x_3y}}{S_{x_3x_3}} = \frac{5162.337}{94387.36} = 0.0546931\ ,\qquad \widehat{\beta_0} = \overline{y_{sales}} - \widehat{\beta_1} \times \overline{x_3} = 12.35141$$

$$RSS = \Sigma(y_i - \hat{y_i})^2 = \Sigma\left(y_i - \widehat{\beta_0} - \widehat{\beta_1} \times x_i\right)^2 = 5134.805$$

$$\widehat{\sigma^2} = \frac{1}{198}\Sigma\left(y_i - \widehat{\beta_0} - \widehat{\beta_1} \times x_i\right)^2 = 25.93336 \quad \Rightarrow\ \hat{\sigma} = 5.09248$$

$$\Rightarrow se\left(\widehat{\beta_0}\right) = \hat{\sigma}\sqrt{\frac{1}{200} + \frac{\overline{x_3}^2}{S_{x_3x_3}}} = 0.6214202\,, se\left(\widehat{\beta_1}\right) = \sqrt{\frac{\hat{\sigma}^2}{S_{x_3x_3}}} = 0.01657572$$

$$\Rightarrow t - statistic\ t_0 = \frac{\widehat{\beta_0}}{se\left(\widehat{\beta_0}\right)} = \frac{12.35141}{0.6214202} = 19.8761\,, p - value = Pr\left(\left|t_{df=198}\right| > |t|\right) \approx 0$$

$$t - statistic\ t_1 = \frac{\widehat{\beta_1}}{se\left(\widehat{\beta_1}\right)} = \frac{0.0546931}{0.01657572} = 3.299591\,, p - value = Pr\left(\left|t_{df=198}\right| > |t|\right)$$

$$= 0.001148196$$

CC:

$$\overline{CC} = \frac{\sum_{i=1}^{200} CC_i}{200} = 13.37186\,, \overline{y_{sales}} = \frac{\sum_{i=1}^{200} y_i}{200} = 14.0225$$

$$S_{CC} = \sum_{i=1}^{200}(CC_i - \overline{CC})^2 = 6277.578\,,\ S_{CCy} = \sum_{i=1}^{200}(CC_i - \overline{CC})(y_i - \overline{y_{sales}}) = 5395.632$$

$$\Rightarrow \widehat{\beta_1} = \frac{S_{xCCy}}{S_{CC}} = \frac{5395.632}{6277.578} = 0.8595085\ ,\qquad \widehat{\beta_0} = \overline{y_{sales}} - \widehat{\beta_1} \times \overline{CC} = 2.529271$$

$$RSS = \sum_{i=1}^{200}(y_i - \hat{y_i})^2 = \sum_{i=1}^{200}\left(y_i - \widehat{\beta_0} - \widehat{\beta_1} \times CC_i\right)^2 = 779.5574$$

$$\widehat{\sigma^2} = \frac{1}{198}\sum_{i=1}^{200}\left(y_i - \widehat{\beta_0} - \widehat{\beta_1} \times CC_i\right)^2 = 3.9837158 \quad \Rightarrow\ \hat{\sigma} = 1.984227$$

$$\Rightarrow se\left(\widehat{\beta_0}\right) = \hat{\sigma}\sqrt{\frac{1}{200} + \frac{\overline{x_3}^2}{S_{CC}}} = 0.3447996\,, se\left(\widehat{\beta_1}\right) = \sqrt{\frac{\hat{\sigma}^2}{S_{CC}}} = 0.02504352$$

$$\Rightarrow t - statistic\ t_0 = \frac{\widehat{\beta_0}}{se\left(\widehat{\beta_0}\right)} = \frac{2.529271}{0.3447996} = 7.335481\,, p - value = Pr\left(\left|t_{df=198}\right| > |t|\right)$$

$$\approx\ 5.549561e - 12$$

$$t - statistic\ t_1 = \frac{\widehat{\beta_1}}{se\left(\widehat{\beta_1}\right)} = \frac{0.8595085}{0.02504352} = 34.32059\,, p - value = Pr\left(\left|t_{df=198}\right| > |t|\right) \approx 0$$

For multiple linear regression

$$\widehat{\boldsymbol{\beta}} = (\boldsymbol{X^T X})^{-1}\boldsymbol{X^T Y}\ , where\ \widehat{\boldsymbol{\beta}} = \begin{bmatrix} \widehat{\beta_0} \\ \widehat{\beta_1} \\ \widehat{\beta_2} \\ \widehat{\beta_3} \end{bmatrix}_{4\times 1}, \boldsymbol{X} = \begin{bmatrix} 1 & x_{1,1} & x_{1,2} & x_{1,3} \\ 1 & x_{2,1} & x_{2,2} & x_{2,3} \\ \vdots & & \vdots\ \vdots\ \vdots & \\ 1 & x_{200,1} & x_{200,2} & x_{200,3} \end{bmatrix}_{200\times 4}, \boldsymbol{Y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_{200} \end{bmatrix}_{200\times 1}$$

$$\widehat{\beta_0} = 2.938889369\ , \widehat{\beta_1} = 0.045764645\ , \widehat{\beta_2} = 0.188530017\ , \widehat{\beta_3} = -0.001037493$$

$$Let\{(x_i, y_i)|i = 1,2,\dots,200\}, where\ x_i = (x_{i1}, x_{i2}, x_{i3}), y_i$$

$$= \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \varepsilon_i, \varepsilon_i \sim N(0, \sigma^2)$$

$$\widehat{\boldsymbol{Y}} = \boldsymbol{X}(\boldsymbol{X^T X})^{-1}\boldsymbol{X^T Y} = \boldsymbol{X}\widehat{\boldsymbol{\beta}} \quad \Longrightarrow \mathrm{SSR_{eg}} = \left(\boldsymbol{y} - \boldsymbol{X}\,\widehat{\boldsymbol{\beta}}\right)^{\mathrm{T}}\left(\boldsymbol{y} - \boldsymbol{X}\,\widehat{\boldsymbol{\beta}}\right) = 556.8253$$

$$\mathrm{SST} = \sum_{\mathrm{i}=1}^{200}(y_i - \bar{y}) = 5417.149\ ,$$

To test $\mathrm{H_0}: \beta_1 = \beta_2 = \beta_3 = 0$ with $F = \frac{(SST - SSR_{eg})/3}{SSR_{eg}/196} = \frac{(5417.149 - 556.8253)/3}{556.8253/196} = 570.2707$ with

$$p - value = Pr\left(F_{(3,196)} > F\right)$$

$$se\left(\widehat{\beta_J}\right) = \sqrt{\frac{SSR_{eg}}{196}\boldsymbol{C_{jj}}}\ , \text{where } \boldsymbol{C} = (\boldsymbol{X^T X})^{-1}$$

$$se\left(\widehat{\beta_0}\right) = \sqrt{\frac{SSR_{eg}}{196}\boldsymbol{C_{11}}} = \sqrt{2.840945 \times 0.0342445} = 0.3119082$$

$$se\left(\widehat{\beta_1}\right) = \sqrt{\frac{SSR_{eg}}{196}\boldsymbol{C_{22}}} = \sqrt{2.840945 \times 6.848908 \times 10^{-7}} = 0.001394897$$

$$se\left(\widehat{\beta_2}\right) = \sqrt{\frac{SSR_{eg}}{196}\boldsymbol{C_{33}}} = \sqrt{2.840945 \times 2.610165 \times 10^{-5}} = 0.008611234$$

$$se\left(\widehat{\beta_3}\right) = \sqrt{\frac{SSR_{eg}}{196}\boldsymbol{C_{44}}} = \sqrt{2.840945 \times 1.213285 \times 10^{-5}} = 0.00587101$$

To test $\mathrm{H_0}: \beta_0 = \beta_1 = \beta_2 = \beta_3 = 0$ with $t_i = \frac{\widehat{\beta_i}}{\sqrt{\frac{SSR_{eg}}{196}\boldsymbol{C_{jj}}}}$

$$\mathrm{t_0} = \frac{2.938889369}{0.3119082} = 9.422288\ , p - value = Pr\left(\left|t_{df=196}\right| > |t|\right) \approx 0$$

$$t_1 = \frac{0.045764645}{0.001394897} = 32.80862\ , p - value = Pr\left(\left|t_{df=196}\right| > |t|\right) \approx 0$$

$$t_2 = \frac{0.188530017}{0.008611234} = 21.8935\ , p - value = Pr\left(\left|t_{df=196}\right| > |t|\right) \approx 0$$

$$t_3 = \frac{-0.001037493}{0.00587101} = -0.1767146\ , p - value = Pr\left(\left|t_{df=196}\right| > |t|\right) = 0.8599132$$

3.Explain how to test H_0:There is no relationship between sales and 3 budgets by compound covariate.State your conclusions.

Suppose that H_0:There is no relationship between sales and 3 budgets by CC(compound covariate) then it is equivalent to test $H_0: \gamma_1 = 0$.

$$\text{sales} = \gamma_0 + \gamma_1 CC + \varepsilon\ , \varepsilon \sim N(0, \sigma^2)$$

$$\widehat{\gamma_0} = 2.529271\ , \widehat{\gamma_1} = 0.8595085$$

$$t - statistic = \frac{\widehat{\gamma_1}}{se(\widehat{\gamma_1})} = \frac{0.8595085}{0.02504352} = 34.32059, p - value\ \ Pr\left(\left|t_{df=198}\right| > |34.32059|\right) \approx 0 < 0.05$$

$\Rightarrow$ reject H_0:There is no relationship between sales and 3 budgets ,there exists relationship bwtween sales and three budgets.

Appendix

R-code&their outputs

```
> data=read.table("C:\\Users\\teresa\\Desktop\\Advertising.csv",sep=",",header=T)
> t=data[1:200,2]
> r=data[1:200,3]
> n=data[1:200,4]
> s=data[1:200,5]
> mean(s)
[1] 14.0225
> ####simple#regression#TV
> mean(TV)
[1] 147.0425
> st=sum((t-mean(TV))^2)
```

```
> st
[1] 1466819
> sts=sum((t-mean(TV))*(s-mean(s)))
> sts
[1] 69727.65
> bt1tv=sum((t-mean(TV))*(s-mean(s)))/st
> bt1tv
[1] 0.04753664
> bt0tv=mean(s)-bt1tv*mean(TV)
> bt0tv
[1] 7.032594
> sigma2_hat_tv=(sum((s-bt0tv-bt1tv*t)^2))/198
> sigma2_hat_tv
[1] 10.61884
> sigma2_hat_tv^0.5
[1] 3.258656
> se_bt0tv=(sigma2_hat_tv^0.5)*((1/200+mean(TV)^2/st)^0.5)
> se_bt0tv
[1] 0.4578429
> se_bt1tv=(sigma2_hat_tv/st)^0.5
> se_bt1tv
[1] 0.002690607
> t0tv=bt0tv/se_bt0tv
> t0tv
[1] 15.36028
> t1tv=bt1tv/se_bt1tv
> t1tv
[1] 17.66763
```

```
> ptv1=(1-pt(t1tv,198))*2
> ptv1
[1] 0
> ptv0=(1-pt(t0tv,198))*2
> ptv0
[1] 0
> (1-pt(t1tv,198))
[1] 0
> ####simple#regression#radio
> mean(radio)
[1] 23.264
> sr=sum((r-mean(radio))^2)
> sr
[1] 43865.12
> srs=sum((r-mean(radio))*(s-mean(s)))
> srs
[1] 8882.502
> bt1ra=sum((r-mean(radio))*(s-mean(s)))/sr
> bt1ra
[1] 0.2024958
> bt0ra=mean(sales)-bt1ra*mean(radio)
> bt0ra
[1] 9.311638
> sigma2_hat_ra=(1/198)*(sum((s-bt0ra-bt1ra*r)^2))
> sigma2_hat_ra
[1] 18.27515
> sigma2_hat_ra^0.5
[1] 4.274944
```

```
> se_bt0ra=(sigma2_hat_ra^0.5)*((1/200+mean(radio)^2/sr)^0.5)
> se_bt0ra
[1] 0.5629005
> se_bt1ra=(sigma2_hat_ra/sr)^0.5
> se_bt1ra
[1] 0.02041131
> t0ra=bt0ra/se_bt0ra
> t0ra
[1] 16.54225
> t1ra=bt1ra/se_bt1ra
> t1ra
[1] 9.920765
> pra1=(1-pt(t1ra,198))*2
> pra1
[1] 0
> pra0=(1-pt(t0ra,198))*2
> pra0
[1] 0
> ####simple#regression#newspaper
> mean(newspaper)
[1] 30.554
> sn=sum((n-mean(newspaper))^2)
> sn
[1] 94387.36
> sns=sum((n-mean(newspaper))*(s-mean(s)))
> sns
[1] 5162.337
> bt1ne=sum((n-mean(newspaper))*(s-mean(s)))/sn
```

```
> bt1ne
[1] 0.0546931
> bt0ne=mean(sales)-bt1ne*mean(newspaper)
> bt0ne
[1] 12.35141
> sigma2_hat_ne=(1/198)*(sum((s-bt0ne-bt1ne*n)^2))
> sigma2_hat_ne
[1] 25.93336
> sigma2_hat_ne^0.5
[1] 5.09248
> se_bt0ne=(sigma2_hat_ne^0.5)*((1/200+mean(newspaper)^2/sn)^0.5)
> se_bt0ne
[1] 0.6214202
> se_bt1ne=(sigma2_hat_ne/sn)^0.5
> se_bt1ne
[1] 0.01657572
> t0ne=bt0ne/se_bt0ne
> t0ne
[1] 19.8761
> t1ne=bt1ne/se_bt1ne
> t1ne
[1] 3.299591
> pnews1=(1-pt(t1ne,198))*2
> pnews1
[1] 0.001148196
> pnews0=(1-pt(t0ne,198))*2
> pnews0
[1] 0
```

```
> ####simple#regression#cc
> cc=bt1tv*t+bt1ra*r+bt1ne*n
> mean(cc)
[1] 13.37186
> sccs=sum((cc-mean(cc))*(s-mean(s)))
> sccs
[1] 5395.632
> scc=sum((cc-mean(cc))^2)
> scc
[1] 6277.578
> bt1cc=sccs/scc
> bt1cc
[1] 0.8595085
> bt0cc=mean(s)-bt1cc*mean(cc)
> bt0cc
[1] 2.529271
> sigma2_hat_cc=sum((s-bt0cc-bt1cc*cc)^2)/198
> sigma2_hat_cc
[1] 3.937158
> sigma2_hat_cc^0.5
[1] 1.984227
> se_bt0cc=(sigma2_hat_cc^0.5)*(1/200+mean(cc)/scc^0.5)
> se_bt0cc
[1] 0.3447996
> se_bt1cc=(sigma2_hat_cc/scc)^0.5
> se_bt1cc
[1] 0.02504352
> t0cc=bt0cc/se_bt0cc
```

```
> t0cc
[1] 7.335481
> t1cc=bt1cc/se_bt1cc
> t1cc
[1] 34.32059
> pcc1=(1-pt(t1cc,198))*2
> pcc1
[1] 0
> pcc0=(1-pt(t0cc,198))*2
> pcc0
[1] 5.549561e-12
> x=as.matrix(cbind(intercept=rep(1,200),data[,2:4]))
> y=as.matrix(data[,5])
> solve(t(x)%*%x)%*%t(x)%*%y
                              [,1]
intercept    2.938889369
TV                0.045764645
radio           0.188530017
newspaper -0.001037493
> sst=sum((s-mean(s))^2)
> ssres=sum((s-2.938889369-0.045764645*t-0.188530017*r+0.001037493*n)^2)
> ssres
[1] 556.8253
> msr=(sst-ssres)/3
> mse=ssres/196
> F=msr/mse
> F
[1] 570.2707
```

```
> AA=solve(t(x)%*%x)
> se0=(mse*AA[1,1])^0.5
> se0
[1] 0.3119082
> se1=(mse*AA[2,2])^0.5
> se1
[1] 0.001394897
> se2=(mse*AA[3,3])^0.5
> se2
[1] 0.008611234
> se3=(mse*AA[4,4])^0.5
> se3
[1] 0.00587101
> t0=2.938889369/se0
> t0
[1] 9.422288
> t1=0.045764645/se1
> t1
[1] 32.80862
> t2=0.188530017/se2
> t2
[1] 21.8935
> t3=-0.001037493/se3
> t3
[1] -0.1767146
> pp0=(1-pt(t0,198))*2
> pp0
[1] 0
```

```
> pp1=(1-pt(t1,198))*2
> pp1
[1] 0
> pp2=(1-pt(t2,198))*2
> pp2
[1] 0
> pp3=(1-pt(t3,198))*2
> pp3
[1] 0.8599132
```