HW5

Friday, June 13, 2025 8:51 AM

1. An analyst wanted to fit the regression model

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \varepsilon_i, \quad i = 1, \ldots, n$$

by the method of least squares when it is known that $\beta_3 = -2$. How can the analyst obtain the desired fit by using a multiple regression computer program?

The best way would be to move the $-2x_{i3}$ to the left hand side. Then we would create $\tilde{y}_i = y_i + 2x_{i3}$ and then regress $\tilde{y}_i$ against the remaining covariates.

2. A hospital administrator wished to study the relation between patient satisfaction (Y) and patient's age (x_1), severity of illness (x_2), and anxiety (x_3). The data (46 patients) are available in `PatientSatisfaction.txt`.

   (a) Obtain a scatterplot matrix and correlation matrix. What information do these provide?

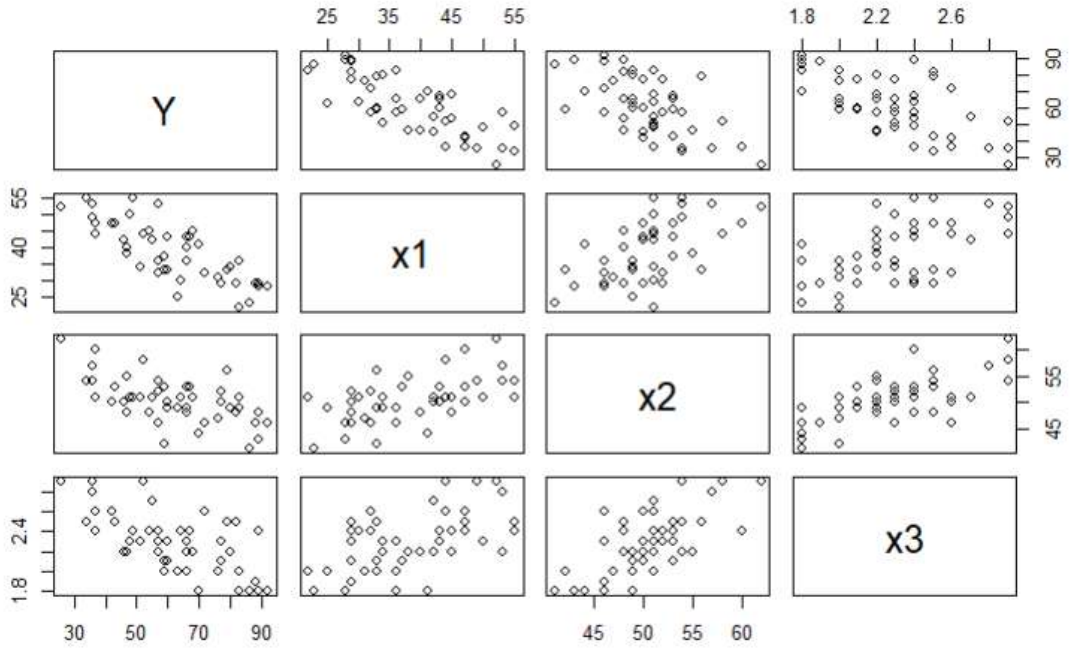
   (b) Fit a first-order regression model for three predictor variables to the data and state the estimated regression function. How is $\hat{\beta}_2$ interpreted here?

   (c) Test whether there is a regression relation, using $\alpha = 0.10$. State the hypotheses, decision rule, p-value, and conclusion. What does your test imply about β_1, β_2, and β_3?

   (d) Find the coefficient of multiple determination R^2 and interpret it here.

   (e) Obtain a 90% confidence interval for the mean satisfaction for a 35-year-old patient with severity index 45 and anxiety index 2.2.

```
> ps<-read.table("C:\\Users\\jacob\\OneDrive\\Documents\\Stats\\8050\\HW5\\PatientSatisfaction.txt",header=
T)
> y<-ps[,1]
> x1<-ps[,2]
> x2<-ps[,3]
> x3<-ps[,4]
> n<-length(y)
> p<-4
> pairs(ps, labels=c("Y","x1","x2","x3"))
```



It looks like there is some linearity involved in these plots

```
> model<-lm(y~x1+x2+x3)
> modcof<-model$coefficients
> cat("E[Y_i]=",modcof[1],"+",modcof[2],"x_i1+",modcof[3],"x_i2+",modcof[4],"x_i3")
E[Y_i]= 158.4913 + -1.141612 x_i1+ -0.4420043 x_i2+ -13.47016 x_i3
```

b2 can be interpreted as if holding x1 and x3, what is the slope corresponding to x2

```
> rse<-summary(model)$sigma
> sumsq<-anova(model)$"Sum Sq"
> f<-sum(sumsq)/(p-1)/rse^2
> 1-pf(f,p-1,n-p)
[1] 4.9305e-13
```

H0: \beta1=\beta2=\beta3=0

Ha: Not all \beta_i=0

Since 4.9*10^-13<0.1, we reject H0

```
> R2a<-1-rse^2/(rse^2*(n-p)+sum(sumsq))/(n-1)
> R2a
[1] 0.9998724
```

This is pretty large, so there is a good chance that there is some correlation between these variables

```
> v<-vcov(model)
> ciR<-drop(qt(0.95,n-p)*sqrt(t(c(1,35,45,2.2)) %*% v %*% c(1,35,45,2.2)))
> cimean<-predict(model,newdata=data.frame(x1=35,x2=45,x3=2.2))#We could get CI from this function by itse
lf
> cat("CI=(",cimean-ciR,",",cimean+ciR,")")
CI=( 64.52854 , 73.49204 )
```

3. An endocrinologist explored the relationship between steroid level (Y) and age (x) in 27 healthy females (age 8–25). Data: `SteroidLevels.txt`.

   (a) Fit a model: $y = \beta_0 + \beta_1 x_i' + \beta_{11}(x_i')^2 + \varepsilon_i$, where $x_i' = x_i - \bar{x}$. Plot the data and fitted model. Comment on fit. Compute R^2.

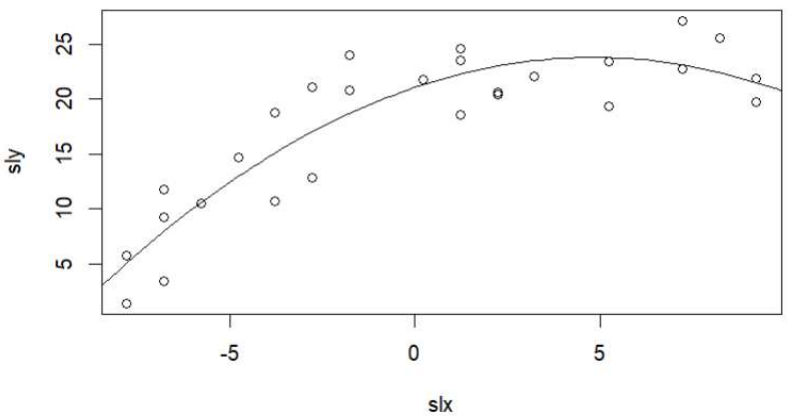
   (b) Test whether or not there is a regression relation ($\alpha = 0.01$). State hypotheses, decision rule, p-value, and conclusion.

   (c) Predict the steroid level of a female at age 15 with 99% prediction interval. Interpret your interval.

   (d) Test if the quadratic term can be dropped from the model ($\alpha = 0.01$). State hypotheses, decision rule, and conclusion.

   (e) Express the fitted function obtained in (a) in terms of the original variable x.

```
> sl<-read.table("C:\\Users\\jacob\\OneDrive\\Documents\\Stats\\8050\\HW5\\SteroidLevels.txt",header=T)
> sly<-sl[,1]
> slx<-sl[,2]-mean(sl[,2])
> slx2<-slx^2
> slmodel<-lm(sly~slx+slx2)
> slcof<-slmodel$coefficients
> plot(slx,sly)
> curve(slcof[1]+slcof[2]*x+slcof[3]*x^2,-10,10,add=T)
> summary(slmodel)$r.squared
[1] 0.8143372
```




```
> slrse<-summary(slmodel)$sigma
> slsumsq<-anova(slmodel)$"Sum Sq"
> slf<-sum(slsumsq)/(2)/slrse^2
> 1-pf(slf,2,length(sly)-3)
[1] 2.173525e-10
```

H0: \beta1=\beta2=\beta3=0

Ha: Not all \beta_i=0

Since 2.2*10^-10<0.1, we reject H0

```
> predict(slmodel,newdata=data.frame(slx=15,slx2=225),interval="prediction",level=0.99)
       fit       lwr      upr
1 11.51424 -4.007162 27.03564
```

H0: \beta1=\beta2=\beta3=0

Ha: Not all \beta_i=0

Since 2.2*10^-10<0.01, we reject H0

```
> sltval<-summary(slmodel)$coefficients[,"t value"]
> quadt<-sltval[3]
> 2*min(1-pt(quadt,length(sly)-3),pt(quadt,length(sly)-3))
[1] 3.70783e-05
```

H0: \beta2=0

Ha: \beta_2\neq0

Since 3.7*10^-5<0.01, we reject H0

expand(21.09416+1.13736·(x+7.896606·10^−16)−0.11840·(x+7.896606·10^−16)^2);

$$-\left(0.1184\,x^2\right)+1.1373599999999997\,x+21.09416$$

4. Refer to the patient satisfaction data in 2.

   (a) Obtain the ANOVA table that decomposes regression sum of squares into extra sums of squares for: x_2; x_1 given x_2; x_3 given x_2 and x_1.

   (b) Test whether x_3 can be dropped given that x_1 and x_2 are retained (use F^*, $\alpha = 0.025$).

   (c) Test whether x_2 and x_3 can be dropped given that x_1 is retained ($\alpha = 0.025$). Give p-value.

   (d) Test whether $\beta_1 = -1.0$ and $\beta_2 = 0$ ($\alpha = 0.025$). State hypotheses, full and reduced models, decision rule, and conclusion.

```
> fit1<-lm(y~x2)
> fit2<-lm(y~x1+x2)
> fit3<-lm(y~x3+x2)
> fit4<-lm(y~x1)
> anova(fit1,fit2,fit3,fit4)
Analysis of Variance Table

Model 1: y ~ x2
Model 2: y ~ x1 + x2
Model 3: y ~ x3 + x2
Model 4: y ~ x1
  Res.Df    RSS Df Sum of Sq      F    Pr(>F)
1     44 8509.0
2     43 4613.0  1    3896.0 36.317 3.348e-07 ***
3     43 7106.4  0   -2493.4
4     44 5093.9 -1    2012.5
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

> SSE.R1 <- anova(lm(y~x1+x2))[3,2]
> SSE.F <- anova(lm(y~x1+x2+x3))[4,2]
> F.star1<- ((SSE.R1-SSE.F)/1)/(SSE.F/(n-p))
> print(paste("Test statistic=",F.star1,"p-value=", 1-pf(F.star1,1,n-p)))
[1] "Test statistic= 3.59973485144707 p-value= 0.0646781268944826"
```

H0: x3=0
Ha: x3\neq0
0.065>0.025 so accept H0

```
> SSE.R2 <- anova(lm(y~x1))[2,2]
> SSE.F <- anova(lm(y~x1+x2+x3))[4,2]
> F.star2<- ((SSE.R2-SSE.F)/2)/(SSE.F/(n-p))
> print(paste("Test statistic=",F.star2,"p-value=", 1-pf(F.star2,2,n-p)))
[1] "Test statistic= 4.17680314099523 p-value= 0.0221611821078052"
```

H0: x3=0=x2
Ha: x2 and x3 are not 0
0.022<0.025 so reject H0

```
> SSE.R3<-anova(lm(y+x1~x3))[2,2]
> F.star3<- ((SSE.R3-SSE.F)/1)/(SSE.F/(n-p))
> print(paste("Test statistic=",F.star1,"p-value=", 1-pf(F.star1,1,n-p)))
[1] "Test statistic= 3.59973485144707 p-value= 0.0646781268944826"
```

H0: \beta2\neq 0 and \beta1\neq -1
Ha: \beta1=-1 and \beta2=0
0.065>0.025 so fail to reject H0

5. Suppose a dataset with sample size $n = 25$ and 3 potential regressors (predictors). Use the adjusted R^2 to perform variable selection.

   (a) Find adjusted R^2 for the null and full models. Given: Residual standard error = 40, R^2 of full model = 0.6.

   (b) Use forward selection to choose the final model.

   (c) Did this procedure fail to find the true best model? If yes, explain.

   (d) What is the AIC of the chosen model from question 2?

Adjusted R^2 for Candidate Models

Table 1: *

Model	Adjusted R^2
$E(Y) = \beta_0$	—
$E(Y\|X_1) = \beta_0 + \beta_1 X_1$	0.488
$E(Y\|X_2) = \beta_0 + \beta_2 X_2$	0.331
$E(Y\|X_3) = \beta_0 + \beta_3 X_3$	0.491
$E(Y\|X_1, X_2) = \beta_0 + \beta_1 X_1 + \beta_2 X_2$	0.653
$E(Y\|X_1, X_3) = \beta_0 + \beta_1 X_1 + \beta_3 X_3$	0.452
$E(Y\|X_2, X_3) = \beta_0 + \beta_2 X_2 + \beta_3 X_3$	0.427
$E(Y\|X_1, X_2, X_3)$	—

a) $RSE = \sqrt{\frac{SSE}{n-p}} \Rightarrow 40^2 \cdot 21 = 33600$

$0.6 = 1 - \frac{33600}{SSTO} \Rightarrow SSTO = 84000$

$R_0^2 = 1 - \frac{SSTO}{SSTO} = 0$

$R_F^2 = 1 - \frac{MSE}{SSTO/n-1} = 1 - \frac{33600/21}{84000/24} = 0.543$

b) R_3^2 is largest so add to model

$R_{1|3}^2$ is larger so add to model

R_F^2 is larger so add to model

c) No because R_{12}^2 is the largest and we chose R_F^2

d) $SSE_F = 33600$

w $SSE_F = 33600$

$AIC = 25\log(33600) - 25\log(25) + 2 \cdot 4$

$= 188.085$