ISyE 6414, "Regression Analysis", Summer 17'

Homework 2 Solutions

June 01, 2017

Problem 1: Paddy Soil Adhesion

Pan and Lu (1998)[1] provide measurements of adhesion on 43 pairs of samples of paddy soil to steel and rubber. From 1974 to 1983, during the rice-growing season, the adhesion of soils to steel and to rubber were measured in situ simultaneously in paddy fields in South China. As steel and rubber have long been the most important materials used for wetland running gears such as wheel and track, it is expected that the adhesion to them would be roughly the same. The adhesion was measured with an adhesometer.

Data set `paddy.csv|dat|mat|xlsx` has two columns: (1) adhesion to steel, and (2) adhesion to rubber. Both measurements are given in kPa.

(a) If adhesion to steel is considered as independent variable x, fit the linear regression where the response variable y is adhesion to rubber. Report the coefficients and R^2.

Solution

Here we fit a linear model of the form $\mathbf{y} = \mathbf{X}\cdot \boldsymbol{\beta} + \epsilon$ where $\mathbf{y} \in \mathbb{R}^n$is the response vector of observations, $\mathbf{X} \in \mathbb{R}^{nxp}$ is the data matrix (including the intercept column) and $\boldsymbol{\beta} \in \mathbb{R}^p$is the vector of coefficients to be estimated. Finally $\epsilon \sim MVN\left(\mathbf{0}, \sigma^{\mathbf{2}}\mathbf{I}_{\mathrm{n}}\right)$ is the vector of random errors. According to the data set, $n = 43$, $p = 2$.

Using Least Squares, we get that $\hat{\boldsymbol{\beta}} = \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{y}$ and the corresponding R^2 is given by $R^2 = 1 - \frac{SS_E}{SS_T}$ where $SS_E = \left(\mathbf{y} - \mathbf{X}\cdot\widehat{\boldsymbol{\beta}}\right)^T\left(\mathbf{y} - \mathbf{X}\cdot\widehat{\boldsymbol{\beta}}\right)$ and $SS_T = \left(\mathbf{y} - \bar{y}\mathbf{1}\right)^T\left(\mathbf{y} - \bar{y}\mathbf{1}\right)$ and $\mathbf{1}$ is a vector of 1 of dimension nx1.

Using a matlab code (see attached code at the end of this document) we obtained the following results:

$$\hat{y} = 0.3974 + 0.7693\cdot x \tag{1}$$

$$R^2 = 1 - \frac{\left(\mathbf{y} - \mathbf{X}\cdot\widehat{\boldsymbol{\beta}}\right)^T\left(\mathbf{y} - \mathbf{X}\cdot\widehat{\boldsymbol{\beta}}\right)}{\left(\mathbf{y} - \bar{y}\mathbf{1}\right)^T\left(\mathbf{y} - \bar{y}\mathbf{1}\right)} = 1 - \frac{6.7232}{21.7430} = 0.6908 \tag{2}$$

[1] Pan, J. Z. and Lu, Z. X. (1998). Relationship between paddy soil adhesion to steel and to rubber. *Journal of Terramechanics*, **35**, 155–158.

(b) If the adhesions to steel and rubber are comparable, the regression should have $\beta_0 = 0$ and $\beta_1 = 1$ as population parameters. Test $H_0 : \beta_0 = 0$ versus $H_1 : \beta_0 > 0$, and $H_0 : \beta_1 = 1$ versus $H_1 : \beta_1 < 1$, both at the level $\alpha = 0.05$..

Solution

Since $\hat{\boldsymbol{\beta}} = \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{y}$ and $\mathbf{y} = \mathbf{X}\cdot\boldsymbol{\beta} + \epsilon$ we have that

$$\mathbf{y} \sim MVN\left(\mathbf{X}\cdot\boldsymbol{\beta}, \sigma^2\mathbf{I}_n\right) \tag{3}$$

Therefore, $\hat{\boldsymbol{\beta}} = \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\left(\mathbf{X}\cdot\boldsymbol{\beta} + \boldsymbol{\epsilon}\right) = \boldsymbol{\beta} + \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\boldsymbol{\epsilon}$ which implies that:

$$\hat{\boldsymbol{\beta}} \sim MVN_p\left(\boldsymbol{\beta}, \sigma^2\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\right) \tag{4}$$

From (4) we can see that for $j = 1, ..., p$ $Var\left(\beta_j\right) = \sigma^2\left(\mathbf{X}^T\mathbf{X}\right)^{-1}_{jj}$ where $\left(\mathbf{X}^T\mathbf{X}\right)^{-1}_{jj}$ if the $j-$th diagonal element of $\left(\mathbf{X}^T\mathbf{X}\right)^{-1}$.

Using (4) and the last result, we can see that $\mathbf{a}^T\hat{\boldsymbol{\beta}} \sim MVN_p\left(\mathbf{a}^T\boldsymbol{\beta}, \sigma^2\mathbf{a}^T\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{a}\right)$ for any $a \in \mathbb{R}^p$. Thus for $j = 1, ..., p$:

$$\frac{\hat{\beta}_j - \beta_j}{\sigma\sqrt{\left(\mathbf{X}^T\mathbf{X}\right)^{-1}_{jj}}} \sim N\left(0, 1\right) \tag{5}$$

Also, we have that:

$$SS_E = \left(\mathbf{X}\cdot\boldsymbol{\beta} + \boldsymbol{\epsilon}\right)^T\left(\mathbf{I}_n - \mathbf{X}\left(\mathbf{X^TX}\right)^{\mathbf{-1}}\mathbf{X}^T\right)\left(\mathbf{X}\cdot\boldsymbol{\beta} + \boldsymbol{\epsilon}\right) = \boldsymbol{\epsilon}^T\left(\mathbf{I}_n - \mathbf{X}\left(\mathbf{X^TX}\right)^{\mathbf{-1}}\mathbf{X}^T\right)\boldsymbol{\epsilon} \tag{6}$$

Where $\mathbf{I}_n - \mathbf{X}\left(\mathbf{X^TX}\right)^{\mathbf{-1}}\mathbf{X}^T$ is an idempotent symmetric Matrix of rank $n - p$. Therefore:

$$\frac{SS_E}{\sigma^2} \sim \chi^2_{n-p} \tag{7}$$

It can also be shown that $\frac{SS_E}{\sigma^2}$ and $\frac{\hat{\beta}_j - \beta_j}{\sigma\sqrt{\left(\mathbf{X}^T\mathbf{X}\right)^{-1}_{jj}}}$ are independent Random Variables (by the properties of the Normal Distribution). Combining (5) and (7) with the last statement, we have that:

$$\frac{\hat{\beta}_j - \beta_j}{\hat{\sigma}\sqrt{\left(\mathbf{X}^T\mathbf{X}\right)^{-1}_{jj}}} \sim t_{n-p} \tag{8}$$

where $\sqrt{\hat{\sigma} = \frac{SS_E}{n-p}}$.

From equation (8) we obtain our test statistic for this question. In particular, we test:

$$\begin{aligned} H_0 :& \quad \beta_0 = 0 \\ H_1 :& \quad \beta_0 > 0 \end{aligned} \tag{9}$$

$$\begin{aligned} H_0 :& \quad \beta_1 = 0 \\ H_1 :& \quad \beta_1 < 1 \end{aligned} \tag{10}$$

So the test statistics are respectively:

$$T_0 = \frac{\hat{\beta}_0}{\hat{\sigma}\sqrt{\left(\mathbf{X}^T\mathbf{X}\right)^{-1}_{11}}} \sim t_{n-p} \tag{11}$$

$$T_1 = \frac{\hat{\beta}_1 - 1}{\hat{\sigma}\sqrt{\left(\mathbf{X}^T\mathbf{X}\right)^{-1}_{22}}} \sim t_{n-p} \tag{12}$$

Since we have one sided tests, the corresponding p_{values} are given respectively by the following expressions:

$$Pr\left\{t_{n-p} > T_0\right\} \tag{13}$$

$$Pr\left\{t_{n-p} < T_1\right\} \tag{14}$$

Using the implemented code in matlab we obtain $Pr\left\{t_{n-p} > T_0\right\} = 0.0015$ and $Pr\left\{t_{n-p} < T_1\right\} = 0.0032$, consequently, at the $\alpha = 0.05$ level, we reject both Null hypothesis.

(c) What adhesion with rubber do you predict in paddy soil for which adhesion to steel was 2. Find 95% prediction interval for a single response.

Solution

In this question we use the model obtained in (1) together with the new data $\mathbf{x}_{new} = [1 \quad 2]^T$ and we obtain:

$$\hat{y}_{new} = 0.3974 + 0.7693 \cdot x = 1.9631 \tag{15}$$

The corresponding 95% confidence interval for prediction is given by:

$$\left[\hat{y}_{new} - \sqrt{\frac{SS_E}{n-p}} \cdot t_{1-\frac{\alpha}{2},n-p}\sqrt{1 + \mathbf{x}^T_{new}\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{x}_{new}}\ ,\ \hat{y}_{new} + \sqrt{\frac{SS_E}{n-p}} \cdot t_{1-\frac{\alpha}{2},n-p}\sqrt{1 + \mathbf{x}^T_{new}\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{x}_{new}}\right] \tag{16}$$

which lead us to obtain $[1.1025, 2.7697]$ as the desired interval.

Matlab Code for Problem 1

```
%% ISyE 6414 Summer 2017 HW#2 Solutions
%%%% Problem 1 Paddy %%%%
load paddy.dat
x = paddy(:,1);
y = paddy(:,2);
n = length(x);
vecones=ones(n,1);
X=[vecones x];
p = size(X,2); %p=2 number of parameters (beta0, beta1)
%% Part (a)
% estimators of coefficients beta1 and beta0
betas = inv(X'*X)*X'*y;
b0 =betas(1);   %    0.3974
b1=betas(2); %       0.7693
% predictive equation (regression equation)
yhat = b0 + b1 * x;
%residuals
res = y - yhat;
% ANOVA Identity
SST = sum(  (y - mean(y)).^2 ) %the same as SST=21.7430
SSE = sum(  (y - yhat).^2 ) %which is sum(res.^2)= 6.7232
Rsq=1-SSE/SST; % R^2=0.6908

%% Part (b)
%  Standard error of coefficient estimators
sigma_hat=SSE/(n-p);
Var_mat=inv(X'*X);
Var_b0= Var_mat(1,1); % 0.0970
Var_b1=Var_mat(2,2);  % 0.0394
se_b0=sqrt(Var_b0); % 0.3115
se_b1=sqrt(Var_b1); % 0.1985
df=n-p; % degrees of freedom fot t distribution = 41
beta0_hyp=0; % value of Null hypothesis for beta0
beta1_hyp=1; % value of Null hypothesis for beta1

% intercept  H0: beta0 = 0  vs H1: beta0 > 0
t_0=(b0-beta0_hyp)/(sqrt(sigma_hat)*se_b0); % 3.1509
pt_0 = 1- tcdf(t_0, df)  %  0.0015
% we reject H0:beta0=0 at the 0.05 level.

% slope H0: beta1=1  vs H1: beta1 < 1
t_1=(b1-beta1_hyp)/(sqrt(sigma_hat)*se_b1); % -2.8693
pt_1 =  tcdf(t_1, df)  %    0.0032
% we reject H0:beta1=1 at the 0.05 level.

%% Part (c)
newx = [1 2]; %   New steel measurement = 2
ypred = betas' * newx'    %     1.9361
syp = sqrt(sigma_hat) * sqrt(1+newx*inv(X'*X)*newx') %s for prediction yhat = 0.4128
%intervals CI and PI
alpha = 0.05;

% prediction interval
lbyp = ypred - tinv(1-alpha/2, n-p) * syp;
rbyp = ypred + tinv(1-alpha/2, n-p) * syp;
%print the intervals
[lbyp rbyp]       %       1.1025    2.7697
```

Problem 2: Prostate Cancer Data.

Data set `prost.csv|dat|mat|xlsx` comes from the study by Stamey et al. (1989)[2] that examined the relationship between the level of serum prostate specific antigen (Yang polyclonal radioimmunoassay) and a number of histological and morphometric measures in 97 patients who were about to receive a radical prostatectomy. The with first 8 columns (`lcavol - pgg45`) are predictors, and the 9th column (`lpsa`) is the response.

x_1	`lcavol`	logarithm of cancer volume
x_2	`lweight`	logarithm of prostate weight
x_3	`age`	patient's age
x_4	`lbph`	logarithm of benign prostatic hyperplasia amount
x_5	`svi`	seminal vesicle invasion, 0 – no, 1 – yes.
x_6	`lcp`	logarithm of capsular penetration
x_7	`gleason`	Gleason score
x_8	`pgg45`	percentage Gleason scores 4 or 5
y	`lpsa`	logarithm of prostate specific antigen

Table 1: Columns in file `prost.csv|dat|mat|xlsx`. First 8 fields are predictors, and the last is the response to be modeled.

(a) Using forward stepwise variable selection propose a parsimonious linear model. If you are using MATLAB, use `stepwise`; for R, use `step`.

Solution

Using MATLAB (see the details of the code in the next section), using `stepwise`, we obtain the following model:

$$\hat{y} = -0.7772 + 0.526 \cdot x_1 + 0.662 \cdot x_2 + 0.6655 \cdot x_5 \tag{17}$$

The details of the `stepwise` procedure in MATLAB are provided in Fig. 1.

(b) Find the solution to the following problem:

Mr. Smith (a new patient) has response $y = 2.3$ and covariates:

$x_1 = 1.4$, $x_2 = 3.7$, $x_3 = 65$, $x_4 = 0.1$, $x_5 = 0$, $x_6 = -0.16$, $x_7 = 7$, and $x_8 = 30$.

How close to the measured response $y = 2.3$ does the regression proposed in from (a) predict y for Mr. Smith? Denote this prediction by $\hat{y}_p$. Calculate the residual $r = \hat{y}_p - y$.

Hint: In calculating $\hat{y}_p$ you should use only covariates x_i suggested by stepwise selection procedure in (a).

Solution

Here we used the model resulting from (a) detailed in equation (17). We have that:

[2]Stamey, T. A., Kabalin, J. N., McNeal, J. E., Johnstone, I. M., Freiha, F., Redwine, E. A. and Yang, N. (1989). Prostate specific antigen in the diagnosis and treatment of adenocarcinoma of the prostate: II. Radical prostatectomy treated patients. *Journal of Urology*, **141**, 5, 1076–1083.

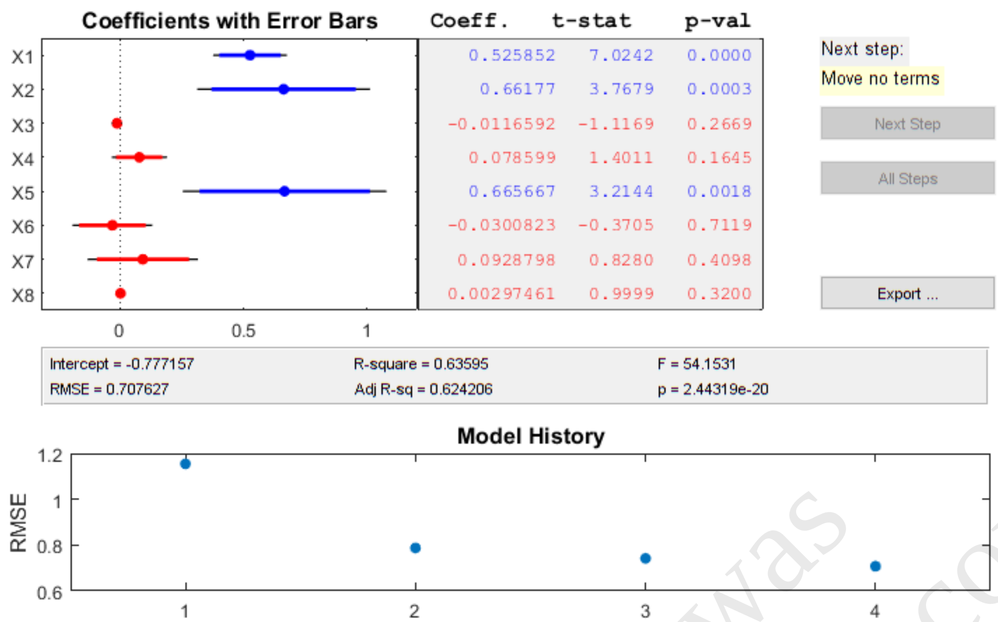

Figure 1: Stepwise summary for Problem 2 from MATLAB.

$$\hat{y}_p = -0.7772 + 0.526 \cdot 1.4 + 0.662 \cdot 3.7 + 0.6655 \cdot 0 = 2.4076 \tag{18}$$

Computing $r = \hat{y}_p - y$ where $y = 2.3$, we get that $r = 0.1076$.

(c) The best, in max R^2 sense, single predictor for y is x_1 – the logarithm of cancer volume. Fit the regression using x_1 as the predictor. What is $\hat{y}_p$ for Mr. Smith based on this regression? Find a 95% prediction interval for y_p. Is $y = 2.3$ in the interval?

Solution

Fitting the model as indicated, we get:

$$\hat{y} = 1.5073 + 0.7193 \cdot x_1 \tag{19}$$

With a corresponding R^2 is 0.5394.

Now, using $x_1 = 1.4$, we get $\hat{y}_p$ is given by:

$$\hat{y} = 1.5073 + 0.7193 \cdot 1.4 = 2.5123 \tag{20}$$

Using equations (6) and (16), for $\alpha = 0.05$ we have that $n-p = 95$, $SS_E = 58.9148$, $\sqrt{1 + \mathbf{x}_{new}^T \left(\mathbf{X}^T\mathbf{X}\right)^{-1} \mathbf{x}_{new}} = 1.0052$, so we get that the 95% prediction interval for $\hat{y}$ is $[0.9429, 4.0858]$ which clearly contains 2.3.

Matlab Code for Problem 2

```
%% ISyE 6414 Summer 2017 HW#2 Solutions
%%%% Problem 2 Prostate Cancer data %%%%
clear all

load prost.dat
x = prost(:,2:9);
y = prost(:,10);
n = size(x,1);
vecones=ones(n,1);
X=[vecones x];
p = size(X,2); %p=10 number of parameters

%% (a) Stepwise procedure

model = stepwise(x,y);
% as a result, only x1,x2 and x5 are included in the model.

%% (b)

X_step=[X(:,1) x(:,1) x(:,2) x(:,5)]; % matrix of selected variables.
betas=inv(X_step'*X_step)*X_step'*y; % coefficients for reduced model.
xnew=[1 1.4 3.7 0]; % using intercept and x1, x2, x5
y_pred=xnew*betas; % 2.4076
y_obs=2.3; % the observed response

res=(y_obs-y_pred); % -0.1076

%% (c)

X_best=[X(:,1) x(:,1)];
betas_best=inv(X_best'*X_best)*X_best'*y; %b0=1.5073 b1=0.7193
yhat = X_best*betas_best;
%residuals
res = y - yhat;
% Computation of R^2
SST = sum(  (y - mean(y)).^2 ) %the same as SST=127.9177
SSE = sum(  (y - yhat).^2 ) %which is sum(res.^2)= 58.9148
Rsq=1-SSE/SST; % R^2=0.5394

% Computation of y_hat
xnew=[1 1.4];
y_hat=xnew*betas_best; % 2.5123

% Computation of the prediction interval
[n p]=size(X_best);
df=n-p; % 95
sigma_hat=sqrt(SSE/df); % 0.7875
syp = sigma_hat * sqrt(1+xnew*inv(X_best'*X_best)*xnew') %s for prediction yhat = 0.7916
%intervals CI and PI
alpha = 0.05;

% prediction interval
lbyp = y_hat - tinv(1-alpha/2, df) * syp;
rbyp = y_hat + tinv(1-alpha/2, df) * syp;
%print the intervals

[lbyp rbyp]       %       0.9429    4.0858
% We can see that 2.3 is contained in the predicted interval for x1=1.4
```