

Linear Regression Models - Final Exam  
Bootstrapping

Matthew Chesnes  
Linear Regression Models  
Kenyon College

May 8, 2001

## Section 2.1

Problem 2.1 considers a regression function where the error terms do not have equal variance. Instead, the variance of each observation is a function of the level of the  $X$  variable, such that,  $\sigma^2\{\varepsilon_i\} = 0.8 * X_i$ . The method of ordinary least squares will therefore lead to biased estimators due to problems of heteroscedasticity. Thus, an alternative method called bootstrapping will be attempted.

To generate random  $Y$  values for each level of  $X$ , I used a combination of SAS and excel. The following is a copy of my SAS code:

```
FILENAME indata 'h:\senior\regression\thedata.txt';
DATA one;
  INFILE indata;
  INPUT X;
  Exp_Y = 20+10*X;
  Var_Err = 0.8*X;
  Std_Err = sqrt(Var_Err);
  weight = 1/Var_Err;
DATA TWO;
  SET one;
  DO j=1 TO 200;
    e=Std_Err*RANNOR(0);
    Y = 20+10*X+e;
    OUTPUT;
  END;
PROC SORT;
  BY j;
PROC PRINT;
RUN;
```

I simply generated random  $Y$  values for each level of  $X$  and used normally distributed error terms with variance defined above. This yields the following output.

OBS	X	EXP_Y	VAR_ERR	STD_ERR	WEIGHT	J	Gen_Error	Y_Pred
1	10	120	8	2.82843	0.12500	1	1.5161	121.516
2	20	220	16	4.00000	0.06250	1	2.0235	222.024
3	30	320	24	4.89898	0.04167	1	4.8917	324.892
4	40	420	32	5.65685	0.03125	1	-1.6436	418.356
5	50	520	40	6.32456	0.02500	1	-0.8759	519.124
6	10	120	8	2.82843	0.12500	2	0.2959	120.296
7	20	220	16	4.00000	0.06250	2	-0.3799	219.620
8	30	320	24	4.89898	0.04167	2	-4.7587	315.241
9	40	420	32	5.65685	0.03125	2	10.7476	430.748
10	50	520	40	6.32456	0.02500	2	-1.4343	518.566
...								
991	10	120	8	2.82843	0.12500	199	-3.1024	116.898
992	20	220	16	4.00000	0.06250	199	-3.3619	216.638
993	30	320	24	4.89898	0.04167	199	4.7224	324.722
994	40	420	32	5.65685	0.03125	199	-1.0149	418.985
995	50	520	40	6.32456	0.02500	199	-4.5398	515.460
996	10	120	8	2.82843	0.12500	200	-3.8132	116.187
997	20	220	16	4.00000	0.06250	200	2.2146	222.215
998	30	320	24	4.89898	0.04167	200	1.0022	321.002
999	40	420	32	5.65685	0.03125	200	-4.3385	415.661
1000	50	520	40	6.32456	0.02500	200	9.5228	529.523

So, for the first repetition, (observations 1 through 5), I have outputted the X levels, the expected value of Y based on the true regression relation, and the variance and standard error of each observation which is given. Then I generated weights which are based on the variance of each observation, such that,

$$w_i = \frac{1}{\hat{s}_i^2}. \quad (1)$$

J is an indexing variable which I later sort by to allow for computing the coefficients on each repetition. Gen\_Error is the error term that is generated from a normal distribution with mean 0 and variance of  $0.8 * X_i$ .

To compute the ordinary and weighted least squares coefficients, I copied the above output into excel and computed my coefficients such that<sup>1</sup>,

$$b_1 = \frac{\sum(X_i - \bar{X})(Y_i - \bar{Y})}{\sum(X_i - \bar{X})^2}. \quad (2)$$

$$b_{1w} = \frac{\sum w_i X_i Y_i - \frac{\sum w_i X_i \sum w_i Y_i}{\sum w_i}}{\sum w_i X_i^2 - \frac{(\sum w_i X_i)^2}{\sum w_i}}. \quad (3)$$

Using these two equations and the data generated from SAS, I determined that the estimated coefficients for  $\beta_1$  and  $\beta_{1w}$  for the first repetition were,

$$b_1 = 12.5008 \quad (4)$$

$$b_{1w} = 9.9357 \quad (5)$$

I then did this same process a total of 200 times, each time generating new error terms and therefore new estimates of  $\beta_1$  and  $\beta_{1w}$ . I found the following means and variances of each estimated coefficient.

$$\bar{b}_1 = 12.5047 \quad \sigma_{b_1}^2 = 0.0382 \quad (6)$$

$$\bar{b}_{1w} = 10.0032 \quad \sigma_{b_{1w}}^2 = 0.0179 \quad (7)$$

The unweighted coefficient is biased because we know that the true value of  $\beta_1$  is 10. This is due to the hetercedasticity that has been added into the model. The weighed estimate however appears to be unbiased and has lower variability then the unweighted estimate. This make intuitive sense since the weighted method reduces the influence of those observations with the highest variance which should yield unbiased estimators that have lower variability.

## Section 2.2 : Milage Study

For part 2 of this problem, I developed the regression model 7.66a and determined  $\hat{X}_{max}$  using the following SAS code.

```
FILENAME indata 'p:\data\math\regression\data\ch07pr31.dat';
DATA data;
  INFILE indata;
  INPUT MPG MPH;
  MPHP = MPH - 47.5;
  MPHP2 = MPHP**2;
PROC REG;
  MODEL MPG = MPHP MPHP2;
RUN;
```

<sup>1</sup>Obtained from Neter: equations 2.2 and 10.26a