

# 36-707: Regression Analysis Homework 2

## Solutions

Fall 2007

**Problem 1**  $Y_i = \beta X_i + \epsilon_i$

a) Find the LS estimator of  $\beta$ .

$$RSS = \sum_{i=1}^n (Y_i - \beta X_i)^2$$

$$\frac{\partial RSS}{\partial \beta} = \sum_{i=1}^n (-2X_i)(Y_i - \beta X_i)$$

$$0 = \sum_{i=1}^n X_i Y_i - \beta \sum_{i=1}^n X_i^2$$

$$\hat{\beta} = \frac{\sum_{i=1}^n X_i Y_i}{\sum_{i=1}^n X_i^2}$$

b) Find an expression for the hat matrix.

In the case where there is no intercept,  $X$  is an  $n \times 1$  matrix.  $X^T = (X_1, X_2, \dots, X_n)$ .  
Then

$$\begin{aligned} H &= X(X^T X)^{-1} X^T \\ &= X(\sum_{i=1}^n X_i^2)^{-1} X^T \\ &= \frac{X X^T}{\sum_{i=1}^n X_i^2} \end{aligned}$$

c) Describe the column space.

The column space of  $X_{n \times 1}$  is all linear combinations of  $X$ .

d) Find the mean and variance of  $\hat{\beta}$  assuming the model is correct.

$$\begin{aligned} E[\hat{\beta}] &= E\left[\frac{\sum_{i=1}^n X_i Y_i}{\sum_{i=1}^n X_i^2}\right] \\ &= \frac{\sum_{i=1}^n X_i EY_i}{\sum_{i=1}^n X_i^2} \\ &= \frac{\sum_{i=1}^n X_i \beta X_i}{\sum_{i=1}^n X_i^2} \\ &= \beta \end{aligned}$$

$$\begin{aligned} V[\hat{\beta}] &= V\left[\frac{\sum_{i=1}^n X_i Y_i}{\sum_{i=1}^n X_i^2}\right] \\ &= \frac{\sum_{i=1}^n X_i^2 VY_i}{(\sum_{i=1}^n X_i^2)^2} \\ &= \frac{\sum_{i=1}^n X_i^2 \sigma^2}{(\sum_{i=1}^n X_i^2)^2} \\ &= \frac{\sigma^2}{\sum_{i=1}^n X_i^2} \end{aligned}$$

e) Find the mean and variance of  $\hat{\beta}$  assuming the model is not correct.

$$\begin{aligned} E[\hat{\beta}] &= E\left[\frac{\sum_{i=1}^n X_i Y_i}{\sum_{i=1}^n X_i^2}\right] \\ &= \frac{\sum_{i=1}^n X_i EY_i}{\sum_{i=1}^n X_i^2} \\ &= \frac{\sum_{i=1}^n X_i r(X_i)}{\sum_{i=1}^n X_i^2} \end{aligned}$$

The variance does not change.

**Problem 2** If  $\hat{\theta}$  is an estimate of a parameter  $\theta$ , show that  $E(\hat{\theta} - \theta)^2 = bias^2 + variance$ .

$$\begin{aligned} E(\hat{\theta} - \theta)^2 &= E(\hat{\theta} - E\hat{\theta} + E\hat{\theta} - \theta)^2 \\ &= E[(\hat{\theta} - E\hat{\theta})^2] + E[(E\hat{\theta} - \theta)^2] + 2E[(\hat{\theta} - E\hat{\theta})(E\hat{\theta} - \theta)] \\ &= variance + E[(E\hat{\theta} - \theta)^2] + 2E[(\hat{\theta} - E\hat{\theta})(E\hat{\theta} - \theta)] \\ &= variance + (E\hat{\theta} - \theta)^2 + 2(E\hat{\theta} - \theta)E[\hat{\theta} - E\hat{\theta}] \end{aligned}$$

Where we use that  $E\hat{\theta}$  and  $\theta$  are constants.

$$\begin{aligned} &= variance + bias^2 + 2(E\hat{\theta} - \theta)(E\hat{\theta} - E\hat{\theta}) \\ &= variance + bias^2 \end{aligned}$$

**Problem 3** Let  $Y \sim N(\theta, 1)$ . Let  $\hat{\theta} = aY$  where  $a$  is a constant.

a) Find the bias, variance and mean-squared error of  $\hat{\theta}$ .

$$\begin{aligned} \text{bias}(\hat{\theta}) &= E\hat{\theta} - \theta \\ &= E[aY] - \theta \\ &= aEY - \theta \\ &= a\theta - \theta \\ &= (a - 1)\theta \end{aligned}$$

$$\begin{aligned} \text{Var}(\hat{\theta}) &= \text{Var}(aY) \\ &= a^2 \text{Var}(Y) \\ &= a^2 \cdot 1 \\ &= a^2 \end{aligned}$$

$$\begin{aligned} \text{MSE}(\hat{\theta}) &= \text{variance} + \text{bias}^2 \\ &= a^2 + (a - 1)^2\theta^2 \end{aligned}$$

b) What value of  $a$  minimizes the mean squared error.

To minimize  $a^2 + (a - 1)^2\theta^2$  with respect to  $a$ , we set  $\frac{\partial}{\partial a}[a^2 + (a - 1)^2\theta^2] = 0$ .

$$\begin{aligned} \frac{\partial}{\partial a}[a^2 + (a - 1)^2\theta^2] &= 2a + 2(a - 1)\theta^2 \\ &= 2a(1 + \theta^2) - 2\theta^2 \end{aligned}$$

$$\begin{aligned} 0 &= a(1 + \theta^2) - \theta^2 \\ a(1 + \theta^2) &= \theta^2 \\ a &= \frac{\theta^2}{1 + \theta^2} \end{aligned}$$

Then MSE is minimized when  $a = \frac{\theta^2}{1 + \theta^2}$ .

(c) Suppose we predict a new observation  $Y^* \sim N(\theta, 1)$  and we use  $\hat{Y} = \hat{\theta}$  as the prediction. Find the prediction error.

Below I present the bias-variance decomposition of the prediction error. Note that this decomposition is true for any regression model.

$$\begin{aligned}
 E[(Y^* - \hat{\theta})^2] &= E[(Y^* - \theta + \theta - \hat{\theta})^2] \\
 &= E[(Y^* - \theta)^2] + E[(\theta - \hat{\theta})^2] + 2E[(Y^* - \theta)(\theta - \hat{\theta})] \\
 &= \text{Var}[Y^*] + \text{MSE}(\hat{\theta}) + 2E[(Y^* - \theta)(\theta - \hat{\theta})] \\
 &= \text{Var}[Y^*] + \text{MSE}(\hat{\theta}) + 2E[Y^*\theta - \theta^2 - Y^*\hat{\theta} + \theta\hat{\theta}] \\
 &= \text{Var}[Y^*] + \text{MSE}(\hat{\theta}) + 2E[Y^*]\theta - 2\theta^2 - 2E[Y^*\hat{\theta}] + 2\theta E\hat{\theta} \\
 &= \text{Var}[Y^*] + \text{MSE}(\hat{\theta}) + 2\theta^2 - 2\theta^2 - 2EE[Y^*\hat{\theta}] + 2\theta E\hat{\theta} \\
 &= \text{Var}[Y^*] + \text{MSE}(\hat{\theta}) - 2E[\hat{\theta}E[Y^*]] + 2\theta E\hat{\theta} \\
 &= \text{Var}[Y^*] + \text{MSE}(\hat{\theta}) - 2\theta E[\hat{\theta}] + 2\theta E\hat{\theta} \\
 &= \text{Var}[Y^*] + \text{MSE}(\hat{\theta})
 \end{aligned}$$

So in this case, we have that the prediction error is just  $1 + a^2 + (a - 1)^2\theta^2$ .

**Problem 4** Problem 3.1 parts 1 and 3.

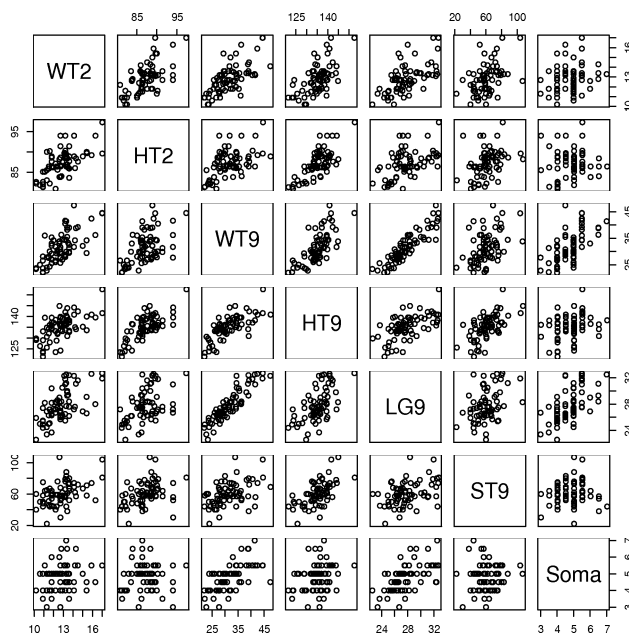


Figure 1: 3.1.1 Scatterplot matrix of the variables.

3.1.1. There appears to be positive correlation between most of the variables. Only the variables compared with Soma lack the positive relationship; most of these appear to be just random scatter with the exception of perhaps WT9 vs Soma and LG9 vs Soma which both appear to have a positive relationship.

Below is the correlation matrix for all three height variables (2, 9 and 18):

	[, 1]	[, 2]	[, 3]
[1, ]	1.0000000	0.7383562	0.6633351
[2, ]	0.7383562	1.0000000	0.8078083
[3, ]	0.6633351	0.8078083	1.0000000

3.1.3. Using R, we get the following output:

Coefficients:

	Estimate	Std. Error	t value	Pr(> t )	
(Intercept)	8.8590417	2.3764431	3.728	0.000411	***
HT2	-0.0792535	0.0354034	-2.239	0.028668	*
WT2	-0.0409358	0.0754343	-0.543	0.589244	
HT9	-0.0009613	0.0260735	-0.037	0.970704	
WT9	0.1280506	0.0203544	6.291	3.2e-08	***
ST9	-0.0092629	0.0060130	-1.540	0.128373	

Residual standard error: 0.5791 on 64 degrees of freedom  
Multiple R-Squared: 0.5211, Adjusted R-squared: 0.4837  
F-statistic: 13.93 on 5 and 64 DF, p-value: 3.309e-09

From this, we deduce:

- $\hat{\sigma} = 0.579$
- $R^2 = 0.52$
- Overall Anova Table:

Analysis of Variance Table

Response: Soma

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Regression	5	23.35552	4.6711	13.93	3.309e-09
Residuals	64	21.4623	0.3353		

- Overall F Test:  $F = 13.93$  with 5 and 64 degrees of freedom resulting in a p-value  $< 0.001$ . From this we know that at least one coefficient is statistically different from 0.
- T-statistic to test each  $\beta_j, j = 0, 1..5$ .  
The hypotheses are as follows:  
 $H_0 : \beta_j = 0, j = 0, 1..5$   
 $H_1 : \beta_j \neq 0, j = 0, 1..5$

Then from the above output, we find that the coefficients for WT9, HT2 and the intercept have large t-values and small p-values. Therefore, we reject the null hypotheses that these coefficients equal 0 and conclude that the intercept, HT2, and WT9 are significant variables in this model. However, the other coefficients have

small t-values and large p-values indicating that we don't have evidence to reject the null hypothesis and cannot conclude that these coefficients are indeed different from zero. At this point, WT2, HT9, and ST9 don't seem to be adding any new information to the model.

**Problem 5** Problem 3.5 parts 1 and 2.

3.5.1.

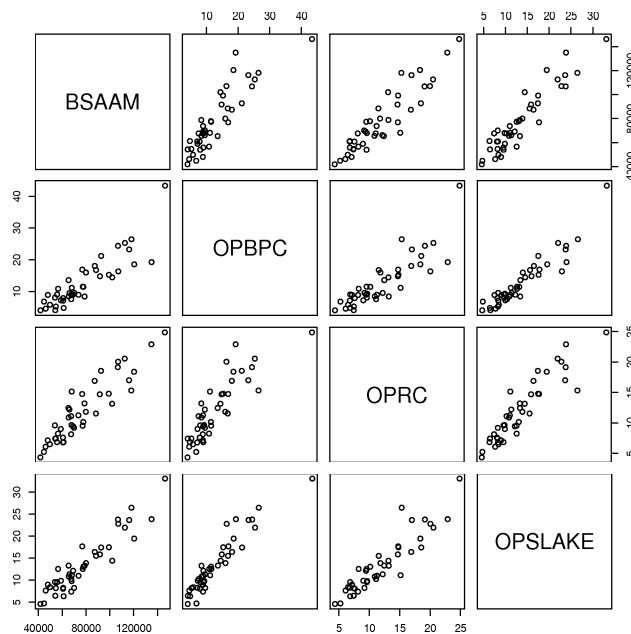


Figure 2: 3.5.1 Scatterplot matrix of OPBPC + OPRC + OPSLAKE.

Since all plots in the scatterplot matrix appear to have a strong positive relationship, we would expect all numbers in the correlation matrix to be close to 1 (large and positive).

Below is the correlation matrix for BSAAM, OPBPC,OPRC,and OPSLAKE respectively. As expected, the numbers are all large and positive.

```

      [,1]      [,2]      [,3]      [,4]
[1,] 1.0000000 0.8857478 0.9196270 0.9384360
[2,] 0.8857478 1.0000000 0.8647073 0.9433474
[3,] 0.9196270 0.8647073 1.0000000 0.9191447
[4,] 0.9384360 0.9433474 0.9191447 1.0000000

```

The regression summary is as follows:

Coefficients:

	Estimate	Std. Error	t value	Pr(> t )	
(Intercept)	22991.85	3545.32	6.485	1.1e-07	***
OPBPC	40.61	502.40	0.081	0.93599	
OPRC	1867.46	647.04	2.886	0.00633	**
OPSLAKE	2353.96	771.71	3.050	0.00410	**

---

Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 8304 on 39 degrees of freedom  
 Multiple R-Squared: 0.9017, Adjusted R-squared: 0.8941  
 F-statistic: 119.2 on 3 and 39 DF, p-value: < 2.2e-16

The t-values column tests that the corresponding slope parameter = 0 vs  $\neq$  0. Therefore, a high t-value will result in a low p-value which means that we can reject that the coefficient for that parameter = 0. In the context of this problem, we see that the intercept, OPRC, and OPSLAKE have coefficients that  $\neq$  0 however, it's possible that the coefficient for OPBPC = 0.

### 3.5.2.

$H_0$ : BSAAM is independent of OPBPC,OPRC, and OPSLAKE

$H_1$ : BSAAM is dependent on at least one of OPBPC,OPRC, or OPSLAKE

As we see above, the F-statistics = 119.2 with 3 and 39 DF for a p-value <  $2.2e - 16$ . This means that at least one of the coefficients is significant in predicting BSAAM. Therefore, we can reject independence and say that BSAAM is dependent on at least one of the variables: OPBPC,OPRC, or OPSLAKE.

**Problem 6** Problem 4.3.

4.3.1. Some of the variables are labeled as “aliased” or missing because since A and D are completely defined by T1 and T2 we are using the same information twice. No new information is introduced by adding A and D to a model that already contains T1 and T2.

4.3.2. M1: Time  $\sim$  T1 + T2

Coefficients:

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	144.36944	170.54410	0.847	0.398
T1	5.46206	0.43327	12.607	<2e-16 ***
T2	2.03455	0.09434	21.567	<2e-16 ***

Residual standard error: 1143 on 258 degrees of freedom  
Multiple R-Squared: 0.9091, Adjusted R-squared: 0.9083  
F-statistic: 1289 on 2 and 258 DF, p-value: < 2.2e-16

M2: Time  $\sim$  A + D

Coefficients:

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	144.3694	170.5441	0.847	0.398
A	7.4966	0.3654	20.514	< 2e-16 ***
D	1.7138	0.2548	6.726	1.12e-10 ***

Residual standard error: 1143 on 258 degrees of freedom  
Multiple R-Squared: 0.9091, Adjusted R-squared: 0.9083  
F-statistic: 1289 on 2 and 258 DF, p-value: < 2.2e-16

M3: Time  $\sim$  T2 + D

Coefficients:

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	144.3694	170.5441	0.847	0.398
T2	7.4966	0.3654	20.514	<2e-16 ***
D	5.4621	0.4333	12.607	<2e-16 ***

Residual standard error: 1143 on 258 degrees of freedom  
 Multiple R-Squared: 0.9091, Adjusted R-squared: 0.9083  
 F-statistic: 1289 on 2 and 258 DF, p-value: < 2.2e-16

**M4: Time ~ T1 + T2 + A + D**

Coefficients: (2 not defined because of singularities)

	Estimate	Std. Error	t value	Pr(> t )	
(Intercept)	144.36944	170.54410	0.847	0.398	
T1	5.46206	0.43327	12.607	<2e-16	***
T2	2.03455	0.09434	21.567	<2e-16	***
A	NA	NA	NA	NA	
D	NA	NA	NA	NA	

Residual standard error: 1143 on 258 degrees of freedom  
 Multiple R-Squared: 0.9091, Adjusted R-squared: 0.9083  
 F-statistic: 1289 on 2 and 258 DF, p-value: < 2.2e-16

4.3.2. Some aspects of the fitted regressions are the same: the intercept, Residual standard error, R-squared, and F-statistic are the same for all models. However, the estimates for the coefficients of T1, T2, A and D differ from model to model.

4.3.3.  $M3 = 7.5T2 + 5.46D = 7.5T2 + 5.46(T1 - T2) = 5.46T1 + (7.5 - 5.46)T2 = 5.46T1 + 2.04T2 = M1$

**Problem 7** Problem 4.5. Use the bootstrap to estimate confidence intervals of the coefficients in the fuel data.

Below are my confidence intervals. Yours should be similar for large B.

Parameter	Lower Bound	Upper Bound
$\beta_0$	-2014609	1192938
Income	-24.64855	34.16920
Miles	-0.8926299	10.7678069
Drivers	-0.1157588	1.4521802
Tax	-40675.892	-4427.707
Pop	-0.6912027	0.6736347
MPC	-17.75594	132.32868

OR

Parameter	Lower Bound	Upper Bound
$\beta_0$	-146.16	772.48
Income	-9.58	-2.66
log(Miles)	-5.86	32.58
Dlic	0.12	0.78
Tax	-10.37	0.01

**Problem 8** Problem 4.8.

4.8.1 This regression formula tells us that we would expect a male to have a salary of \$24,697 and a female to have a salary of \$24,697 - \$3,340 = \$21,357.

4.8.2 The coefficient of the estimated variable sex could have changed signs if there is some kind correlation between sex and the number of years employed by the college.

**Problem 9** Problem 5.1.

5.1.1.

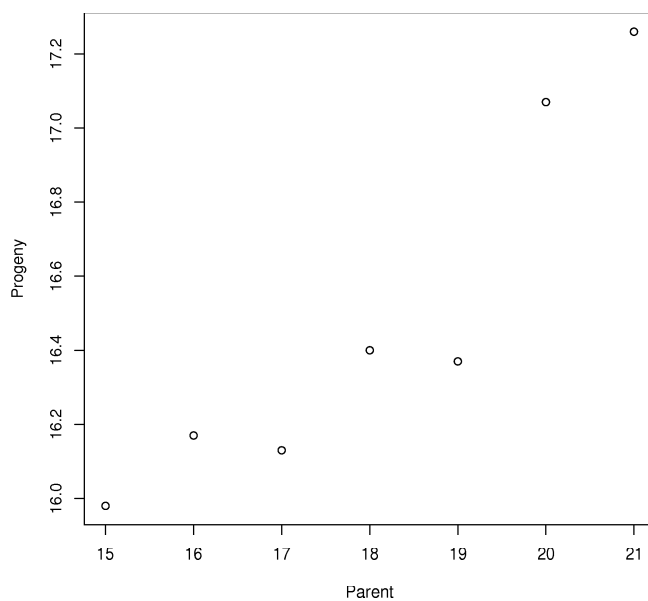


Figure 3: 5.1.1 Scatterplot of Progeny vs. Parent.

5.1.2.

If we assume the standard deviations given are population values, we can compute the weighted least squares regression:

Let  $\sigma^2 = 1$ . Then  $SD = \frac{1}{\sqrt{w_i}} \Rightarrow w_i = \frac{1}{SD^2}$ .

Coefficients:

	Estimate	Std. Error	t value	Pr(> t )	
(Intercept)	12.79642	0.68112	18.787	7.87e-06	***
Parent	0.20480	0.03815	5.368	0.00302	**

Residual standard error: 0.11 on 5 degrees of freedom

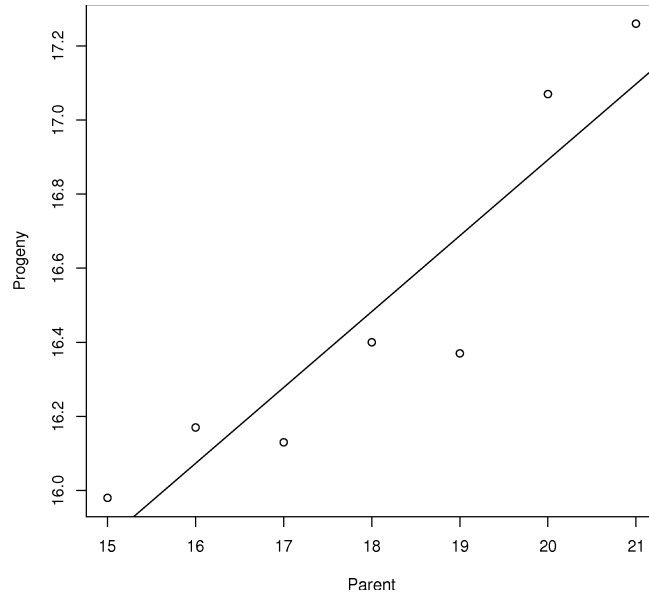


Figure 4: 5.1.1 Regression line added to scatterplot of Progeny vs. Parent.

Multiple R-Squared: 0.8521,      Adjusted R-squared: 0.8225  
 F-statistic: 28.81 on 1 and 5 DF,   p-value: 0.003021

Analysis of Variance Table

Response: Progeny

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Parent	1	0.34872	0.34872	28.811	0.003021 **
Residuals	5	0.06052	0.01210		

5.1.3.

$$H_0 : \beta_1 = 1$$

$$H_1 : \beta_1 < 1$$

$$T = (\hat{\beta}_1 - 1) / se(\hat{\beta}_1) = (0.20480 - 1) / 0.03815 = -20.84$$

Since T is so large, the p-value is essentially 0. Therefore, we can reject the null hypothesis that  $\beta_1 = 1$  and conclude that  $\beta_1 < 1$ . We can reject the idea of perfect

inheritance and conclude that they are reverting toward the average ancestral type.

#### 5.1.4

a) Estimation of the intercept and slope:

Taking exceptionally large seeds from a small plant and exceptionally small seeds from a large plant is going to make the distribution of seeds inaccurate. It will decrease the range so if we think about this in terms of linear regression, the line will be tilted downward. Points near the bottom of true regression plot will be pushed upwards since we are looking at exceptionally large seeds. Points near the top of the true regression plot will be pushed downward since we are looking at exceptionally small seeds. Therefore, we would expect a decreased slope and an increased intercept.

b) Estimates of the error:

Since the estimated regression line is built using data that isn't entirely accurate (looking at exceptionally large and small seeds for small and large plants respectively does not accurately portray the true distribution of seeds), we would expect the errors to be artificially lower than the truth.

#### **Problem 10** Problem 5.3.

5.3.1. Here is my code:

```
q6.fun<-function(x,y){  
  
  lin = lm(y ~ x)  
  rss = resid(lin)  
  sigma = summary(lin)$sigma  
  nonpar = loess(y ~ x)  
  G = sum((fitted(lin) - fitted(nonpar))^2)/sigma  
  
  B = 999  
  l = 0  
  n = length(rss)  
  
  rss_star = rep(0,n)
```

```

for (i in 1:B){
  for (j in 1:n){
    rand = round(runif(1,1,n))
    rss_star[j] = rss[rand]
  }

  y.lin.star <- lin$fitted+rss_star
  lin_star = lm(y.lin.star ~ x)
  sigma_star = summary(lin_star)$sigma
  nonpar_star = loess(y.lin.star ~ x)
  G_star<-sum((fitted(lin_star)-fitted(nonpar_star))^2)/sigma_star

  if (G_star > G){l = l+1}
}

sig<-1/B
return(sig = sig)
}

```

### 5.3.2.

$H_0$  : Parametric model is appropriate

$H_1$  : Parametric model is not appropriate

Significance Level = 0.001. Since this is a simulation, everyone's answer will be slightly different but should be close to zero.

Since the significance level is small, we reject the null hypothesis. We conclude that simple linear regression is not appropriate for the regression of Height on Dbh.