

BS1a: Applied Statistics, 1st Assessed Practical, Michaelmas Term 09

1. INTRODUCTION TO THE DATA

(Nemli *et al.* 2005) <http://dx.doi.org/10.1016/j.buildenv.2004.12.008> consider the effect of construction-process on the surface roughness of particle-board. The file `board.txt` contains the data from Tables 1 and 2 of that paper.

```
> bd<-read.table('board.txt',header=T)
> bd
  DensityT Pressure Shelling DensityA Thickness Roughness
A      0.6       30       32    0.572     18.30     71.19
B      0.7       30       32    0.685     18.33     57.16
C      0.6       35       32    0.594     18.05     67.30
D      0.7       35       32    0.709     18.07     50.25
E      0.6       30       45    0.602     18.06     61.72
F      0.7       30       45    0.713     18.08     48.32
G      0.6       35       45    0.615     17.93     58.41
H      0.7       35       45    0.725     17.95     39.65
I      0.7       30       32    0.716     18.03     42.96
J      0.6       35       32    0.630     17.93     58.11
```

In these data each row gives data for a particular piece of board. The first 3 columns of data, `DensityT`, `Pressure` and `Shelling` give the settings used in manufacturing the boards. The last three columns `DensityA`, `Thickness` and `Roughness` are properties of the finished board.

`DensityT`: Target density (g/cm^3)
`Pressure`: Pressure (kg/cm^2)
`Shelling`: Shelling ratio % (see below)
`DensityA`: Actual achieved density (g/cm^3)
`Thickness`: Thickness (mm)
`Roughness`: surface roughness (μm)

The first column A-J of row names simply names the particular board. Wood particles used for boards I and J were approximately 70% pine, 20% beech and 10% poplar. For boards A-H, particles consist of approximately 70% beech, 20% pine and 10% poplar. The boards were all made by taking soft wood-particle mats ($280 \times 210 \times 1.8 \text{ cm}^3$ thick) and pressing them using pressures of 30 and 35 kg/cm^2 at a high temperature. The particle-board is constructed to have a hard dense surface, or shell, and a pulpy interior. The shelling ratio is the fraction of the board which is shell, expressed as a percentage.

2. QUESTIONS

- (1) Read the description of the data on the previous page carefully. How do the variables predict `Roughness`? In order to answer this question, you should
 - (a) construct a normal linear model for the response `Roughness`, and carry out diagnostic checks, explaining what action you took in response to the results obtained at this stage,
 - (b) carry out model selection, commenting on the effects of correlation on parameter significance,

- (c) briefly interpret your final model.
- (2) Consider now **Thickness** as a response with **DensityT**, **Pressure**, **Shelling** and **DensityA** potentially explanatory. Carry out variable selection. Interpret your final model for the response **Thickness**.
- (3) Finally, consider **DensityA** as a response, with **Pressure**, **Shelling** and **DensityT** potentially explanatory. The authors' assert that "Particle boards produced with 45% shelling ratio and pressed under 35 kg/cm^2 had higher density values than those of the panels pressed under 30 kg/cm^2 and produced with 32% shelling ratio, statistically ($p \leq 0.01$)". Verify this statement.

3. SAMPLE ANSWERS

There are a number of ways to deal acceptably with the questions above. Here are some sample answers which are somewhat more detailed than we might expect from a student (as they attempt to cover all the bases). The small sample size poses some problems.

3.1. Modeling roughness. We note the variables come in two groups, the settings for board-construction (**DensityT**, **Pressure** and **Shelling**) which are independent variables, and the properties of the boards after construction (**DensityA**, **Thickness** and **Roughness**) which are all dependent. We would expect the former to predict the latter. However, the two dependent variables **DensityA** and **Thickness** may respond to unrecorded features of the board construction process, not informed by the independent variables, and hence help predict **Roughness**. We will begin, as the question suggests, using all the variables to predict **Roughness**.

```
> #read the data and make a pairs-plot
> bd<-read.table('board.txt',header=T)
> pairs(Roughness~DensityA+Thickness,data=bd,
+       col=1+(bd$DensityT==0.7),
+       pch=15*(bd$Pressure==35),
+       cex=1+(bd$Shelling==45))
> round(cor(bd[,-6]),2)
```

	DensityT	Pressure	Shelling	DensityA	Thickness
DensityT	1.00	-0.20	0.00	0.95	0.14
Pressure	-0.20	1.00	0.00	-0.03	-0.66
Shelling	0.00	0.00	1.00	0.11	-0.42
DensityA	0.95	-0.03	0.11	1.00	-0.14
Thickness	0.14	-0.66	-0.42	-0.14	1.00

In the pairs plot Figure 1 the binary variables are represented using color, symbol and symbol size. **DensityA** seems to predict **Roughness** fairly well (pairs plot). **Thickness** goes linearly with **DensityA** with intercept controlled by **DensityT**.

DensityT and **DensityA** are correlated (`cor()`). Dependent variable **Thickness** is correlated with **Pressure** and **Shelling**, but **Pressure** and **Shelling** are uncorrelated (`cor()` matrix). Note that **Shelling** is actually orthogonal to **Pressure** and **DensityT**, so the **Shelling** parameter estimate will be independent of the **Pressure** and **DensityT** parameter estimates.

```
> #Fit NUMBER 1: linear model with given formula (includes intercept)
> bd.lm1<-lm(Roughness~DensityT+Pressure+Shelling+DensityA+Thickness,
```

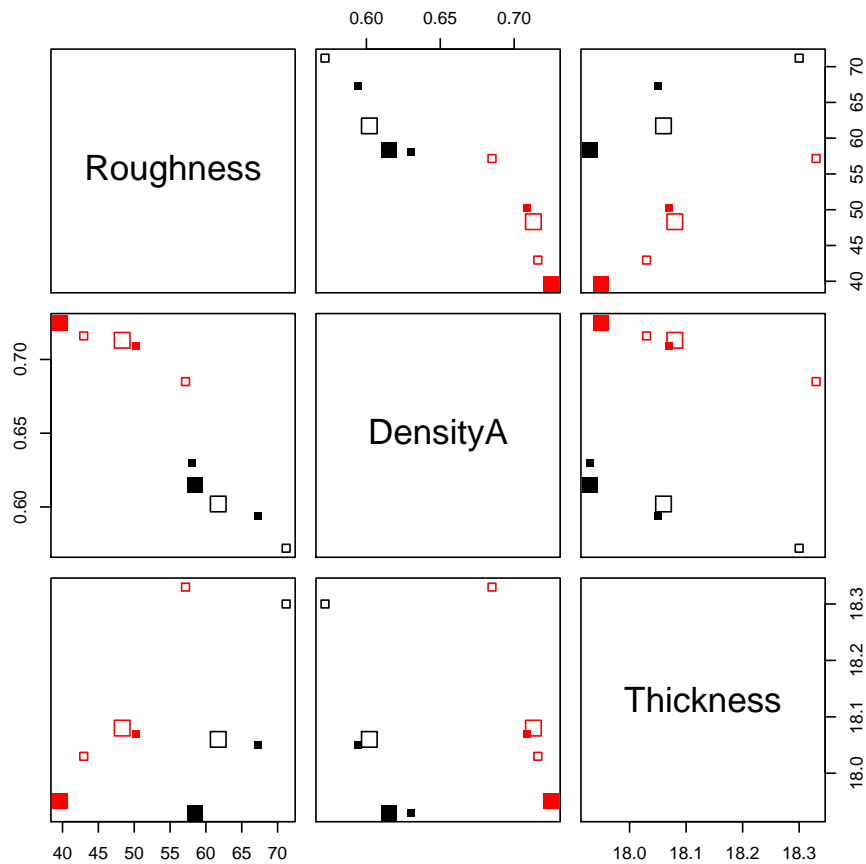


FIGURE 1. Pairs plot for all the data. Low/high values (Black/Red) DensityT, (Empty/Full) Pressure, (small/big) Shelling.

```

+           data=bd)
> #use diagnostics to look at quality of fit
> par(mfrow=c(2,2))
> p<-6; n<-10
> qqnorm(rstudent(bd.lm1),pch=row.names(bd),main='');
> qqline(rstudent(bd.lm1))
> plot(bd.lm1$fitted.values,rstudent(bd.lm1),xlab='fitted values',
+      ylab='studentised residuals',pch=row.names(bd))
> plot(hatvalues(bd.lm1),ylab='leverage',xlab='data row index',
+      pch=row.names(bd))
> plot(cooks.distance(bd.lm1),ylab='Cooks distance',
+      xlab='data row index',pch=row.names(bd))

```

Referring to Figure 2, it is clear that boards I and J are unusual. There is no strong misfit, but they have relatively high leverage and Cook's distance. Returning to the data we note that these boards are special (composed of different wood).

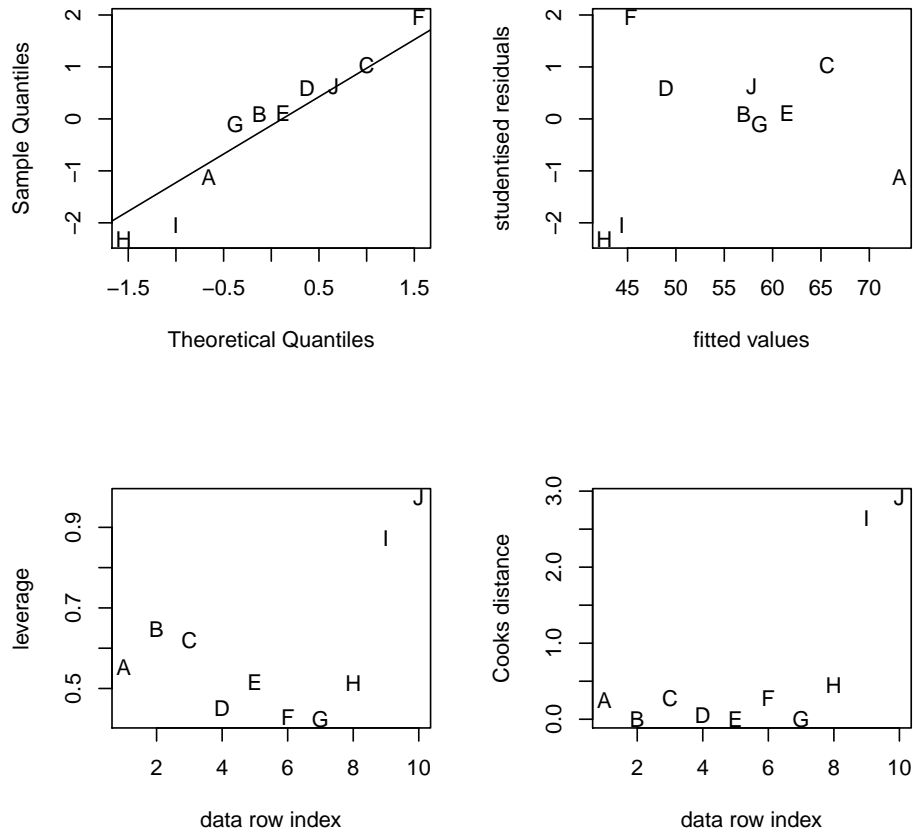


FIGURE 2. Diagnostics plots for FIT #1.

Rather than dropping these boards from the analysis we add an indicator variable marking them as distinct.

```
> bdw<-bd
> bdw$Wood<-c(0,0,0,0,0,0,0,0,1,1)
> bdw
```

	DensityT	Pressure	Shelling	DensityA	Thickness	Roughness	Wood
A	0.6	30	32	0.572	18.30	71.19	0
B	0.7	30	32	0.685	18.33	57.16	0
C	0.6	35	32	0.594	18.05	67.30	0
D	0.7	35	32	0.709	18.07	50.25	0
E	0.6	30	45	0.602	18.06	61.72	0
F	0.7	30	45	0.713	18.08	48.32	0
G	0.6	35	45	0.615	17.93	58.41	0
H	0.7	35	45	0.725	17.95	39.65	0
I	0.7	30	32	0.716	18.03	42.96	1
J	0.6	35	32	0.630	17.93	58.11	1

At the cost of an extra parameter (for the new variable Wood) we can offset the response for these two boards. Alternatively we could drop these boards from the analysis. This would be a less attractive option. We repeated the diagnostics and found I and J were no longer exceptional. Fitting all variables and testing for no linear relation (parameters for all variables bar intercept equal zero) we find evidence for a linear relation ($p = 0.004456$).

```
> summary(bdw.lm1<-lm(Roughness~DensityT+Pressure+Shelling+
+                      DensityA+Thickness+Wood,data=bdw))
```

...

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	-117.2613	321.7404	-0.364	0.740
DensityT	-487.9249	223.8428	-2.180	0.117
Pressure	-1.3464	0.8406	-1.602	0.208
Shelling	-0.9583	0.3820	-2.509	0.087 .
DensityA	289.3781	201.5355	1.436	0.247
Thickness	21.1745	16.9376	1.250	0.300
Wood	-16.0958	6.6433	-2.423	0.094 .

...

Residual standard error: 1.779 on 3 degrees of freedom

F-statistic: 48.49 on 6 and 3 DF, p-value: 0.004456

None of the variables appear to be explanatory (p -values for t -tests for parameters equal zero are all each > 0.05). However, this is due to masking of significance by correlation as we now show.

We now carry out model selection. We find we can explain the response Roughness using either the dependent or the independent variables. When we use one, the others are not needed.

Consider the null model dropping DensityA and Thickness:

```
> summary(bdw.lm0a<-lm(Roughness~DensityT+Pressure+Shelling+Wood,data=bdw))
```

...

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	224.5190	13.0581	17.194	1.22e-05 ***
DensityT	-166.4917	12.3264	-13.507	3.98e-05 ***
Pressure	-0.9712	0.2465	-3.939	0.010967 *
Shelling	-0.7269	0.1039	-6.999	0.000918 ***
Wood	-10.9400	1.6538	-6.615	0.001188 **

...

Residual standard error: 1.91 on 5 degrees of freedom

F-statistic: 62.49 on 4 and 5 DF, p-value: 0.0001859

```
> anova(bdw.lm0a,bdw.lm1)
```

Analysis of Variance Table

Model 1: Roughness ~ DensityT + Pressure + Shelling + Wood

Model 2: Roughness ~ DensityT + Pressure + Shelling + DensityA + Thickness + Wood

	Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
1	5	18.2328				

```
2      3  9.4894  2      8.7433 1.3821 0.3755
```

There is no evidence to reject the model `Roughness ~ DensityT + Pressure + Shelling + Wood` ($p = 0.3755$). We see from the `summary()` output that no individual variable can now be removed ($p > 0.05$ for all individual t -tests).

There is an alternative null model based on precisely the variables we have just dropped.

```
> summary(bdw.lm0b<-lm(Roughness~DensityA+Thickness,data=bdw))
```

```
...
```

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	-217.888	108.010	-2.017	0.08346 .
DensityA	-153.843	13.927	-11.046	1.11e-05 ***
Thickness	20.712	5.882	3.521	0.00971 **

```
...
```

Residual standard error: 2.443 on 7 degrees of freedom

F-statistic: 74.36 on 2 and 7 DF, p-value: 1.926e-05

```
> anova(bdw.lm0b,bdw.lm1)
```

Analysis of Variance Table

Model 1: `Roughness ~ DensityA + Thickness`

Model 2: `Roughness ~ DensityT + Pressure + Shelling + DensityA + Thickness + Wood`

	Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
1	7	41.792				
2	3	9.489	4	32.303	2.553	0.2336

Again, there is no evidence against `Roughness ~ DensityA + Thickness` ($p = 0.2336$) and this time we can even omit `Wood`, presumably because the dependent explanatory variables summarise all important aspects of the board-setup that actually impact the finished board, and therefore roughness. The two final (non-nested) models are

$$\text{Roughness} \sim \text{DensityT} + \text{Pressure} + \text{Shelling} + \text{Wood}$$

in which `Roughness` decreases with all variables (and the I,J board types have reduced roughness, and

$$\text{Roughness} \sim \text{DensityA} + \text{Thickness}$$

in which `Roughness` decreases with density, and increases with thickness. We might compare the two models on the basis of their AIC. However, n is not large compared to p here, so the approximation we used to justify its use is unreliable. Using `step(bdw.lm1)` on the original full model finds no variable to drop which improves the AIC, and leaves the full model.

3.2. A Model for Board Thickness. An exploratory approach starting with all the variables led to the model `Thickness ~ DensityT+DensityA+Pressure`. The first step showed the same issue as before:

```
> summary(bdw.lm1<-lm(Thickness~DensityT+DensityA+Pressure+Shelling+Wood,data=bdw))
```

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
--	----------	------------	---------	----------

```
(Intercept) 18.747728 1.529279 12.259 0.000254 ***
DensityT    4.214918 6.262769 0.673 0.537822
DensityA   -3.772859 5.642339 -0.669 0.540317
Pressure   -0.020204 0.022666 -0.891 0.423101
Shelling   -0.007146 0.010696 -0.668 0.540672
Wood       -0.082996 0.191668 -0.433 0.687329
```

correlation is masking the significance of variables. I use the AIC to make a quick automatic search:

```
> step(bdw.lm1)
Start: AIC=-56.1
Thickness ~ DensityT + DensityA + Pressure + Shelling + Wood
.
.
.
Step: AIC=-57.8
Thickness ~ DensityT + DensityA + Pressure
```

	Df	Sum of Sq	RSS	AIC
<none>			0.014	-57.796
- Pressure	1	0.003	0.017	-57.652
- DensityT	1	0.082	0.096	-40.470
- DensityA	1	0.087	0.100	-40.002

Now removing the variable `Pressure` only just increases the AIC (from -57.796 to -57.652, so the important variables here are `DensityA` and `DensityT`. On reflection there is a natural physical model: the difference between the target and achieved density might be explained by a corresponding linear variation in the thickness (since mass per unit area of board is fixed). This motivates the model `Thickness ~ I(DensityT - DensityA)` where `I(DensityT - DensityA)` is a new variable equal to the difference in densities.

```
> summary(bdw.lm0<-lm(Thickness~I(DensityT-DensityA),data=bdw))
...
```

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	18.11508	0.02271	797.499	< 2e-16 ***
I(DensityT - DensityA)	6.89863	1.24947	5.521	0.00056 ***

```
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1 1
```

```
Residual standard error: 0.06767 on 8 degrees of freedom
Multiple R-squared: 0.7921, Adjusted R-squared: 0.7661
F-statistic: 30.48 on 1 and 8 DF, p-value: 0.0005595
```

```
> anova(bdw.lm0,bdw.lm1)
Analysis of Variance Table
```

```
Model 1: Thickness ~ I(DensityT - DensityA)
Model 2: Thickness ~ DensityT + DensityA + Pressure + Shelling + Wood
Res.Df      RSS Df Sum of Sq      F Pr(>F)
```

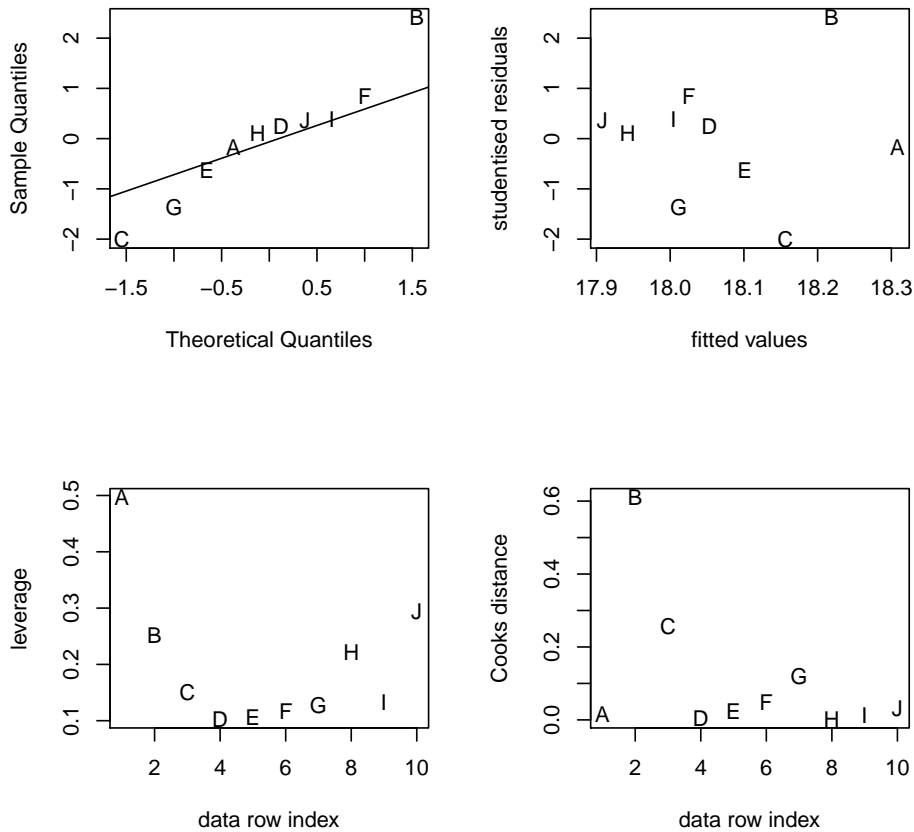


FIGURE 3. Diagnostic plots for the model $\text{Thickness} \sim \text{I}(\text{DensityT} - \text{DensityA})$.

```

1      8 0.036630
2      4 0.011026  4  0.025604 2.3222 0.2173
>

```

The p-value for the test to drop all variables except the intercept and the density gap favors the null. The model $\text{Thickness} \sim \text{I}(\text{DensityT} - \text{DensityA})$ was in effect proposed *ab initio* on physical grounds and we may ignore the tests and AIC stepping which led to it.

Outlier checks are interesting here. The relevant figures are shown in Figure 3. The qq-plot (top left) looks poor, and *B* has a large studentised residual (2.4), but recall these residuals are $t(n - p - 1) = t(7)$ under the NLM, so we are not at “large n ”, and might expect extra weight in the tails (the std.dev of $t(\nu)$ is $\sqrt{\nu/(\nu - 2)} \simeq 1.18$ here so *B* is just on two deviations). There may be increasing variance with increasing fitted values (top right). The leverages (lower left) have changed since we added *Wood*, and *A* is now important. Our index for the CD (lower right) is $8/(10 - 2) = 1.33$, based on a misfit threshold of 2, suggests no problem. There is at least no major problem with the fit.

3.3. A Model for Density. We now fit with response `DensityA`.

Let μ_2 be the mean achieved density for boards pressed at the high pressure and shelling levels, and let μ_1 be the mean at the low levels. Under the null these two means are equal. We cannot (at least directly) compare the achieved densities (using a t -test for the difference in means of the two groups of `DensityA`-values) because `DensityT` and `Wood` varies within each group. However, fitting $\text{DensityA} \sim \text{DensityT} + \text{Pressure} + \text{Shelling} + \text{Wood}$ we get

```
> bdw.lm1<-lm(DensityA~DensityT+Pressure+Shelling+Wood,data=bdw)
> (bdw.sum<-summary(bdw.lm1))
```

...

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)	
(Intercept)	-0.2634615	0.0284558	-9.259	0.000247	***
DensityT	1.1083333	0.0268613	41.261	1.58e-07	***
Pressure	0.0038333	0.0005372	7.135	0.000839	***
Shelling	0.0018269	0.0002263	8.071	0.000473	***
Wood	0.0330000	0.0036038	9.157	0.000260	***

...

with all variables apparently explanatory. If we take two pressings with pressures p_1 and p_2 , shelling ratios s_1 and s_2 and other variables equal, then the difference in the means is

$$\mu_2 - \mu_1 = \beta_P(p_2 - p_1) + \beta_S(s_2 - s_1).$$

To cut to the answer, since the differences $\delta_P = p_2 - p_1$ and $\delta_S = s_2 - s_1$ are both positive, and the confidence intervals for β_P, β_S are positive and do not overlap zero at 1%, it follows that the 1% CI for $\mu_2 - \mu_1$ is positive and doesn't overlap zero, and the claim is supported. In more detail, under the null, $\hat{\mu}_2 - \hat{\mu}_1 \sim N(0, V)$ with

$$V = \delta_P^2 \text{var}(\hat{\beta}_P) + \delta_S^2 \text{var}(\hat{\beta}_S) - 2\delta_P \delta_S \text{cov}(\hat{\beta}_P, \hat{\beta}_S).$$

Since $\text{cov}(\hat{\beta}_P, \hat{\beta}_S) = 0$ we have

$$\frac{\hat{\mu}_2 - \hat{\mu}_1}{s \sqrt{\delta_P^2 (X^T X)_{33}^{-1} + \delta_S^2 (X^T X)_{44}^{-1}}} \sim t(n - p),$$

with $n = 10$ and $p = 5$. Computing a p-value for the one sided test for $\mu_2 - \mu_1 < 0$,

```
> #reading off summary() output using cov=0
> den<-sqrt(25*0.0005372^2+169*0.0002263^2)
> num<-5*0.0038333+13*0.0018269
> 1-pt(num/den,n-p) #one-sided test
[1] 5.974306e-05
```

The p-value for negative $\mu_2 - \mu_1$ under the null is very small, so the claim “Particle boards produced with 45% shelling ratio and pressed under 35 kg/cm^2 had higher density values than those of the panels pressed under 30 kg/cm^2 and produced with 32% shelling ratio, statistically ($p \leq 0.01$)” is not contradicted.

REFERENCES

Nemli, Gkay, Ozturk, Ibrahim, and Aydin, Ismail. 2005. Some of the parameters influencing surface roughness of particleboard. *Building and environment*, 40(10), 1337 – 1340.