

1. A study was made of the effectiveness of Cognitive Behavioral Therapy on treatment of anorexia nervosa. Measurements were made of the weights of the patients before treatment and after treatment. See the following display of the data and *R*-input and output to answer the questions.

Before	80.50	84.90	81.50	82.60	79.90	88.70	94.90	76.30	...
After	82.20	85.60	81.40	81.90	76.40	103.6	98.40	93.40	...

```

< anorexia <- read.table("anorexia.data",header=TRUE)
< names(anorexia)
[1] "Before" "After" < attach(anorexia)
< t.test(After,Before,alternative='greater',paired=TRUE)

```

```

data: After and Before
t = 2.2156, df = 28, p-value = 0.01751
alternative hypothesis: true difference in means is greater than 0
95 percent confidence interval:
0.6981979      Inf
sample estimates:
mean of the differences      3.006897

```

- (a) What are the null and alternative hypotheses being tested here?

$$H_0: \mu_{\text{after}} \leq \mu_{\text{before}}$$

$$H_a: \mu_{\text{after}} > \mu_{\text{before}}$$

- (b) What conclusions regarding the effectiveness of Cognitive Behavioral Therapy can you draw from the *R* output above. Your conclusions should include, but not be limited to, a conclusion for the hypothesis test.

With a *p*-value of .01751 we reject H_0 and conclude that the mean weight after Cognitive Behavioral Therapy is greater than the mean weight before the therapy. The 95% confidence interval suggest that the average increase was at least .698 pounds.

- (c) What *p*-value would be associated with the following *R* input?

```
> t.test(After,Before,paired=TRUE)
```

This is a two sided alternative, so the *p*-value would double to .03502.

- (d) What *p*-value would be associated with the following *R* input?

```
> t.test(After-Before,alternative='greater')
```

This test is equivalent to the test we performed, thus the *p*-value would be identical: .01751.

2. Consider the following data descriptions concerning the energy content of municipal waste and the regression run on it, with summary output following.

Energy Energy content (kcal/kg)
Plastic Percent plastic composition by weight
Paper Percent paper composition by weight
Garbage Percent garbage composition by weight
Water Percent water composition by weight

```
> waste <- read.table("municipal-waste.data",header=TRUE)
> attach(waste)
> mod <- lm(energy ~ plastic + paper + garbage + water)
> summary(mod)
```

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	2245.09	177.89	12.62	2.4e-12
Plastics	22.92	2.82	10.24	2.0e-10
Paper	7.64	2.31	3.30	0.0029
Garbage	4.30	1.92	2.24	0.0340
Water	-37.36	1.83	-20.37	.0005

Residual standard error: 31.5 on 25 degrees of freedom

Multiple R-Squared: 0.964, Adjusted R-Squared: 0.958

F-statistic: 168 on 4 and 25 DF, p-value <2e-16

- (a) What percentage of the variability in *Energy* is explained by the model?

96.4% of the variability.

- (b) What are the null and alternative hypotheses being tested by the p -value 0.0029?

$$H_0 : \beta_{paper} = 0$$

$$H_a : \beta_{paper} \neq 0$$

- (c) What are the models implicitly being tested by this hypothesis test?

Null model:

$$\text{energy}_i = \beta_0 + \beta_{pl} \text{plastic}_i + \beta_g \text{garbage}_i + \beta_w \text{water}_i + \epsilon_i$$

Alternative model:

$$\text{energy}_i = \beta_0 + \beta_{pl} \text{plastic}_i + \beta_{pa} \text{paper}_i + \beta_g \text{garbage}_i + \beta_w \text{water}_i + \epsilon_i$$

- (d) What are the null and alternative hypotheses being tested by the p -value <2e-16?

$$H_0 : \beta_{plastic} = \beta_{paper} = \beta_{garbage} = \beta_{water} = 0$$

H_a : at least one slope is non-zero

- (e) What are the models implicitly being tested by this hypothesis test?

Null model:

$$\text{energy}_i = \beta_0 + \epsilon_i$$

Alternative model:

$$\text{energy}_i = \beta_0 + \beta_{pl} \text{plastic}_i + \beta_{pa} \text{paper}_i + \beta_g \text{garbage}_i + \beta_w \text{water}_i + \epsilon_i$$

- (f) What would be p -value associated with the anova test in the following R commands? `> new.mod <- lm(energy ~ plastic + garbage + water)`
`> anova(mod,new.mod)`

This F -test is equivalent to the t -test in the *paper* row of the summary output above, hence the p -value would be 0.0029.

- (g) If sample A had 10% more (in absolute terms) plastic than sample B , what is the predicted energy content of sample A compared to sample B ?

Sample A would have $10 \cdot 22.92 = 229.2$ *kcal/kg* more energy than sample B .

- (h) What energy content would this model predict for a sample that was purely water? Comment on the reliability of this prediction.

Such a sample is 100% water, 0% everything else, so:

$$\text{energy} = 2245.09 - 37.36(100) = -1490.91$$

Clearly this is absurd. It is unlikely that the data used to construct this model contains a trash sample that was 100% water, or even anything near to this, hence this is a wild extrapolation from the data, and we are not surprised that the prediction is absurd.

- (i) Estimate the values of the 95% confidence interval for the *Garbage* parameter. Explain how you arrived at your answer.

Roughly 2 times the standard error above and below the estimate, hence:

$$(4.30 - 2 \cdot 1.92, 4.30 + 2 \cdot 1.92) = (.46, 8.14)$$

- (j) Would any of the confidence intervals produced by the following R command contain 0? How do you know?

`< confint(mod)`

No, since a 95% CI containing zero is equivalent to not rejecting $H_0 : \beta = 0$ at $\alpha = .05$, and all of the p -values are less than .05.

- (k) Suppose that an extra variable, *Refined petroleums* giving the percent gasoline, oil, etc. composition of the trash was added to the model. What would happen to the value of Multiple R-squared.

Multiple R-squared **always** goes up when extra terms are added to the model.

3. Complete the table below, filling in *RSS* for Residual Sum of Squares, *TSS* for Total Sum of Squares, and *RegSS* for Regression or Model Sum of Squares. Then below the table write an equation relating these quantities (a correct equation of course!). Then write an equation relating these quantities (or at least some of them) and r^2 (again a correct equation is preferred).

$\sum_{i=1}^n (\hat{y}_i - \bar{y})^2$	RegSS
$\sum_{i=1}^n (y_i - \bar{y})^2$	TSS
$\sum_{i=1}^n (y_i - \hat{y})^2$	RSS

4. Consider a study of Diabetes using the following variables:

glyhb Glycosated hemoglobin, the response variable
chol Total cholesterol
ratio Cholesterol/HDL ratio
age Age in years
gender Male = 1, Female = 0
weight Weight in pounds

```
< diabetes <- read.table("diabetes.data",header=TRUE)
< attach(diabetes)
< mod <- lm(glyhb ~ chol + ratio + age + gender + weight)
< summary(mod)
```

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	0.453964	0.670124	0.677	0.4985
chol	0.004281	0.002624	1.632	0.1036
ratio	0.280721	0.068754	4.083	5.42e-05
age	0.038744	0.006346	6.106	2.52e-09
gendermale	-0.066613	0.206245	-0.323	0.7469
weight	0.006618	0.002614	2.532	0.0117

Residual standard error: 1.969 on 382 degrees of freedom
Multiple R-Squared: 0.2166, Adjusted R-squared: 0.2063
F-statistic: 21.12 on 5 and 382 DF, p-value: < 2.2e-16

```
< mod.1 | lm(glyhb ~ ratio + age + weight)
< summary(mod.1)
```

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	1.062952	0.555563	1.913	0.0565
ratio	0.329918	0.061099	5.400	1.17e-07
age	0.040601	0.006206	6.542	1.95e-10
weight	0.006293	0.002605	2.415	0.0162

Residual standard error: 1.971 on 384 degrees of freedom
Multiple R-Squared: 0.2107, Adjusted R-squared: 0.2045
F-statistic: 34.16 on 3 and 384 DF, p-value: < 2.2e-16

- (a) RSS for the model *mod.1* was 1493.135. How could you compute this from the summary output?

Residual standard error is $\sqrt{\frac{RSS}{df}}$, so

$$(1.971)^2 = \frac{RSS}{384}$$

which is easily solved for *RSS*.

- (b) RSS for the model *mod* was 1480.920. How could you compute the *F*-statistic from the anova test command below from these and the summaries above?

```
< anova(mod,mod.1)
```

The anova test uses

$$F = \frac{(RSS_{smallmodel} - RSS_{bigmodel}) / (\text{difference in number of terms in models})}{RSS_{bigmodel} / \text{df big model}}$$
$$= \frac{(1493.135 - 1480.920) / 2}{1480.920 / 382}$$

- (c) On the back of the test is a plot produced by the command

```
< plot(mod.1$fitted.values,abs(mod.1$residuals))
```