
A researcher has two primary research questions: (1) Is there a difference in weight loss, on the average, between the low-fat and low-carbohydrate diets? (2) Is there a difference in mean weight loss for the low-fat and low-carbohydrate diets compared to the mean of just the Mediterranean diet group?

(a) (1 pt) It is appropriate to test a a few primary research questions using an individual CI or individual test for each. This way, the full power of the study is applied to these questions.

(b) (3 pts) An ANOVA was fit to these data (R-code is in the Appendix). The ANOVA table is:

```
Analysis of Variance Table

Df  Sum Sq Mean Sq  F value Pr(>F)
Group   2  216.9 108.430  3.2358  0.04086 *
Residuals 269  9013.9  33.509
```

The assumptions were checked using normal probability and residual vs. fits plots.

- Based on the normal probability plot, the residuals look normal.
- Based on the residual vs. fits plot, the residuals appear to have constant variance. Furthermore, using our rule of thumb, the largest SD (7.0 for Low-Carb) is not twice more than the smallest SD (4.11 for Low-Fat), which also substantiates our constant variance assumption.
- The humans who participated in this study were all employees from a single workplace. Furthermore, the participating employees from this workplace were recruited by the researchers, so there is potential for voluntary response and non-response bias. Therefore, these data are not a random sample from the population of all humans, or of Americans, or of all employees.
- Because the employees were randomly assigned to different groups, the groups are independent.

(c) (3 pts) Testing the SECOND research question:

i. Using the notation $\mu_1$ for the mean weight-loss of the low-carbohydrate group, $\mu_2$ for the mean weight loss of the low-fat group and $\mu_3$ being the mean weight loss of the Mediterranean group, the parameter $\gamma$ being tested is the linear combination

$$\gamma = \frac{1}{2}\mu_1 + \frac{1}{2}\mu_2 - \mu_3.$$ 

ii. The mean weight loss is 2.54 points (this is an estimate for $\gamma$). A 95% CI is [0.73, 4.35] (see R-code in the Appendix). weight

iii. With 95% confidence, the Low Carb and Low Fat diets give 0.73 to 4.35 kg more weight loss, on the average, compared to the Mediterranean diet.

(d) (2 pts) It is appropriate to test for all pairwise differences among the group means in addition to the primary research questions because the ANOVA suggests that there is some difference in the mean weight loss among the diets ($p=0.041, F = 3.2$). Testing all pairwise comparisons will be performed using a multiple comparison procedure to maintain a family-wise significance level of 95%. Although the average weight loss of the Low-Carb diet was more than either of the other two groups in this study, the evidence only suggests that the Low Carb diet causes a higher weight loss than the Low-Fat diet in a larger population ($p=0.033$, Tukeys). The evidence fails to suggest
a difference in the mean weight loss between the Mediterranean diet and either the Low Carb
\((p = 0.566)\) or Low Fat \((p=0.277)\) diets.

(e) \((2 \text{ pts})\) \textit{Scope of Inference} Because this was a randomized experiment, then the evidence does suggest that the Low-Carb diet \textit{caused} more weight loss compared to the Low Fat diet. Because the humans in this study were not a random sample of all humans, and were not even a random sample of the employees from the single workplace, it is unclear what population these results can be extended to beyond the participants themselves.

2. Regarding exercise 27 on page 234 re: lifespans of mammals.

(a) \((1 \text{ pt})\) Lifespans and log(Lifespan) are plotted as functions of metabolic rate and log(metabolic rate) using scatterplots in the Appendix. These plots suggest that we log-transform both lifespan and metabolic rate.

(b) \((1 \text{ pt})\) An SLR was fit to the Lifespans as a function of Metab (without any transforms), giving the regression equation

\[
\hat{\mu}\{\text{lifespan} | \text{metabolic rate}\} = 10.3 + 0.000387(\text{metabolic rate}).
\]

This line was added to the appropriate scatterplot in the Appendix. The graph shows that the regression line does not fit the untransformed data well.

(c) \((1 \text{ pts})\) As suggested by the scatterplots, residual plots (in the Appendix) confirm that regression fit in \#2b is poor. The assumptions of normality, linearity and constant variance all appear to be suspect. Independence of the data is also suspect, but we would need more information to assess this assumption.

(d) \((2 \text{ pts})\) SLRs were fit to each of the 3 candidate models (i.e., \(y \) vs \(\log_{10}(x)\): \(\log_{10}(y) \) vs \(x\) and \(\log_{10}(y) \) vs \(\log_{10}(x)\)), and residual and normal probability plots were generated for each model. These plots show that the SLR that models log(Lifespan) as a function of log(metabolic rate) best satisfies the regression assumptions.

(e) \((1 \text{ pts})\) Of the 4 models considered, the best SLR model is

\[
\mu\{\log_{10}(\text{Life}|\text{Metab})\} = \beta_0 + \beta_1 \log_{10}(\text{Metab}).
\]

We have to exponentiate both sides to get this written as \(\mu\{\text{Life}|\text{Metab}\}\). First, under the assumption of normality, then

\[
\mu\{\log_{10}(\text{Life}|\text{Metab})\} = \text{Median}\{\log_{10}(\text{Life}|\text{Metab})\} = \log_{10}(\text{Median}\{\text{Life}|\text{Metab}\}).
\]

So

\[
\log_{10}(\text{Median}\{\text{Life}|\text{Metab}\}) = \beta_0 + \beta_1 \log_{10}(\text{Metab})
\]

\[
\text{Median}\{\text{Life}|\text{Metab}\} = 10^{\beta_0 + \beta_1 \log_{10}(\text{Metab})} = 10^{\beta_0} \cdot \text{Metab}^{\beta_1}
\]

The fit SLR equation is

\[
\hat{\mu}\{\log_{10}(\text{Life})|\text{Metab}\} = -0.092 + 0.382 \log_{10}(\text{Metab}).
\]

Plugging these estimate for \(\beta_0\) and \(\beta_1\) into the last equation we get that:

\[
\text{Median}\{\text{Life}|\text{Metab}\} = 0.809 \times \text{Metab}^{0.382}
\]

(f) \((2 \text{ pts})\) To test whether there is an association between metabolic rate and lifespan at a significance level of \(\alpha = 1\%\), we perform the following 6 steps:

- \(H_0 : \beta_1 = 0\) vs. \(H_a : \beta_1 \neq 0\).
We checked the SLR assumptions above and found no evidence to suggest that either linearity, normality or constant variance are violated.

This is a $t$-test, with test statistic (from the Appendix) $t = 18.11$

The two-sided $p$-value is tiny, $2 \times P(t > 18.11) < 2 \times 10^{-16}$

Because the $p$-value is tiny, we reject $H_0$.

There is strong evidence to suggest that there is an association between life span and metabolism. The estimated equation that describes this association is $\hat{\text{Median}}\{\text{Life}|\text{Metab}\} = 0.809 \times \text{Metab}^{0.382}$, with a 95% CI $[0.34, 0.42]$ for the exponent of $\text{Metab}$.

(g) (2 pts) Scope of Inference: This was an observational study, so we may only state that increased metabolism is associated with longer life spans. It is not clear how one would conduct an experiment to assess whether faster metabolic rates cause longer life spans. Because it is not clear how mammals were chosen to be included in the data set, and it is not clear whether randomization played a role, then it is not clear whether the association that we found can be generalized to the larger population of all mammals.

3. CIs and PIs for lifespans of mammals:

(a) (2 pts) An individual 95% CI for the median lifespan of a mammal with a metabolic rate of 100 (i.e., $\log_{10} x = 2$) was found by fitting the model (i.e., the “shift trick”)

$$\log_{10}(\text{Median}\{\text{Life}|\text{Metab}\}) = \beta_0 + \beta_1 (\log_{10}(\text{Metab}) - 2)$$

$$\text{Median}\{\text{Life}|\text{Metab}\} = 10^{\beta_0 + \beta_1 \log_{10}(\text{Metab})} = 10^{\beta_0} + (\text{Metab}/100)^{\beta_1}$$

(see Apendix). A 95% CI for $\beta_0$ is $[0.623, 0.721]$. A 95% CI for $10^{\beta_0}$ is $[10^{0.623}, 10^{0.721}] = [4.19, 5.27]$. Hence we are 95% confident that the true median lifespan for mammals with a metabolic rate of 100 is between 4.2 and 5.3 years.

(b) (1 pt) A band of Workman-Hotelling CIs that maintain a family-wise confidence level of 95% has been added to the scatterplot.

(c) (1 pt) A band of individual 95% PIs for lifespan has been added to the scatterplot.

(d) (1 pt) The Workman-Hotelling 95% CI for the mean lifespan of mammals with a metabolic rate of 100 ($\log_{10} x = 2$) is LARGER than the individual 95% CI we found in #3a for the median lifespan of a mammal with a metabolic rate of 100 because the Workman-Hotelling CIs maintains a 95% confidence level over all metabolic rates.

(e) (2 pts) A 95% PI for lifespan for a mammal with a metabolic rate of 100 is $[1.74, 12.68]$ years. Hence we are 95% confident that an individual mammal with a metabolism of 100 has a lifespan between 1.7 and 12.7 years.

Appendix

Housekeeping

```r
source("http://www.math.montana.edu/parker/courses/STAT411/diagANOVA.r")
library(Sleuth3)
library(gmodels)
```
Problem 1

Get, summarize and plot the data:

```r
summary(ex0623)
```

```r
tapply(ex0623$WtLoss24,ex0623$Group,mean)
tapply(ex0623$WtLoss24,ex0623$Group,sd)
```

Descriptive stats, fit an ANOVA, perform Tukeys
m1 = lm(WtLoss24 ~ Group, data = ex0623)

anova(m1)

## Analysis of Variance Table
##
## Response: WtLoss24
## Df Sum Sq Mean Sq F value Pr(>F)
## Group 2 216.9 108.430 3.2358 0.04086 *
## Residuals 269 9013.9 33.509
## ---
## Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

TukeyHSD(aov(WtLoss24 ~ Group, data = ex0623))

## Tukey multiple comparisons of means
## 95% family-wise confidence level
##
## Fit: aov(formula = WtLoss24 ~ Group, data = ex0623)
##
## $Group
## diff  lwr  upr  p adj
## Low-Fat-Low-Carbohydrate -2.1828035 -4.224771 -0.1408361 0.0329364
## Mediterranean-Low-Carbohydrate -0.8849083 -2.932082 1.1622656 0.5656813
## Mediterranean-Low-Fat 1.2978952 -0.697418 3.2932084 0.2771180

Assess model fit - the residual plots are gorgeous, indicating that the ANOVA assumptions of normality and constant variance are met. This is because the book says that they generated these data.
Estimate the linear combination associated with the primary research question.

```r
estimable(lm(WtLoss24 ~ Group, data=ex0623), c(1/2, 1/2, -1), conf.int=.95)
```

```r
## Estimate Std. Error t value DF Pr(>|t|) Lower.CI
## (0.5 0.5 -1) 2.537036 0.9185056 2.762135 269 0.006138259 0.728662
## Upper.CI
## (0.5 0.5 -1) 4.34541
```

Problem #2

Inspect data:
### CommonName Species Mass

<table>
<thead>
<tr>
<th>CommonName</th>
<th>Species</th>
<th>Mass</th>
</tr>
</thead>
<tbody>
<tr>
<td>Length:95</td>
<td>Length:95</td>
<td>Min. : 0.0036</td>
</tr>
<tr>
<td>Class:character</td>
<td>Class:character</td>
<td>1st Qu.: 0.0990</td>
</tr>
<tr>
<td>Mode:character</td>
<td>Mode:character</td>
<td>Median : 1.9800</td>
</tr>
<tr>
<td></td>
<td>Mean : 69.6215</td>
<td></td>
</tr>
<tr>
<td></td>
<td>3rd Qu.: 11.9000</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Max. :3000.0000</td>
<td></td>
</tr>
</tbody>
</table>

### Metab Life

<table>
<thead>
<tr>
<th>Metab</th>
<th>Life</th>
</tr>
</thead>
<tbody>
<tr>
<td>Min. : 5.17</td>
<td>Min. : 0.8</td>
</tr>
<tr>
<td>1st Qu.: 39.05</td>
<td>1st Qu.: 3.5</td>
</tr>
<tr>
<td>Median : 320.00</td>
<td>Median : 8.0</td>
</tr>
<tr>
<td>Mean : 5173.49</td>
<td>Mean : 12.3</td>
</tr>
<tr>
<td>3rd Qu.: 1695.00</td>
<td>3rd Qu.: 15.5</td>
</tr>
<tr>
<td>Max. :165000.00</td>
<td>Max. : 75.0</td>
</tr>
</tbody>
</table>

### dim(ex0826)

```
# [1] 95 5
```

Plot the data 4 different ways:

```
par(mfrow=c(2,2))
plot(ex0826$Metab,ex0826$Life)
abline(10.3,3.87e-4,col="red")  # adding the regression line
plot(log10(ex0826$Metab),ex0826$Life)
plot(log10(ex0826$Metab),log10(ex0826$Life))
plot(ex0826$Metab,log10(ex0826$Life))
```
The SLR model we investigate first is

\[ \mu\{\text{Life}|\text{Metab}\} = \beta_0 + \beta_1 \text{Metab} \]

\[
m2 = \text{lm}(\text{Life} - \text{Metab}, \text{data}=\text{ex0826}) \]

\[
\text{summary}(m2) \]

---

\[
\text{Call:} \quad \text{lm(formula = Life ~ Metab, data = ex0826)} \]

\[
\text{Residuals:} \quad \begin{array}{cccc}
\text{Min} & 1Q & \text{Median} & 3Q & \text{Max} \\
-20.205 & -6.885 & -2.341 & 3.598 & 61.775 \\
\end{array}
\]

\[
\text{Coefficients:} \quad \begin{array}{cccc}
\text{Estimate} & \text{Std. Error} & t value & \text{Pr(>|t|)} \\
(Intercept) & 1.030e+01 & 1.124e+00 & 9.161 & 1.22e-14 *** \\
\text{Metab} & 3.873e-04 & 5.740e-05 & 6.748 & 1.26e-09 *** \\
\end{array}
\]

---

\[
\text{Signif. codes:} \quad 0 \ '***' \ 0.001 \ '***' \ 0.01 \ '**' \ 0.05 \ '*' \ 0.1 \ '.' \ 1 \]

\[
\text{Residual standard error:} \quad 10.57 \text{ on} \ 93 \text{ degrees of freedom} \]

\[
\text{Multiple R-squared:} \quad 0.3287, \text{ Adjusted R-squared:} \quad 0.3215 \]

\[
\text{F-statistic:} \quad 45.53 \text{ on} \ 1 \text{ and} \ 93 \text{ DF,} \quad p\text{-value:} \quad 1.262e-09
\]

It is clear just from these scatterplots that we should log-transform both lifespan and metabolic rate.
m3=lm(log10(Life) ~ log10(Metab),data=ex0826)
summary(m3)

## Call:
## lm(formula = log10(Life) ~ log10(Metab), data = ex0826)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
##-0.48440 -0.13773 -0.00900  0.09498  0.65186
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -0.09221   0.05799  -1.59   0.115
## log10(Metab)  0.38215   0.02111  18.11  <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2156 on 93 degrees of freedom
## Multiple R-squared: 0.779, Adjusted R-squared: 0.7766
## F-statistic: 327.8 on 1 and 93 DF, p-value: < 2.2e-16

confint(m3)

##                2.5 %   97.5 %
## (Intercept)  -0.20736 0.0294751
## log10(Metab)  0.34023 0.42406207

I checked all 4 diagnostic plots (commented out in the R-code below) to confirm what via residual plots what the scatterplots suggest. Here I only show the diagnostic plots for the model:

\[ \mu \{ \log_{10}(Life | Metab) \} = \beta_0 + \beta_1 \log_{10}(Metab) \]

: diagANOVA(m2)
In this sample of size \( n = 95 \), correlation of the residuals in the qq-plot is \( r = 0.873901 \)

In the following table, if \( r < \text{critical}\_r \), then the qq-plot suggests the residuals are not normal:

<table>
<thead>
<tr>
<th>( n )</th>
<th>( \text{critical}_r )</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.832</td>
</tr>
<tr>
<td>2</td>
<td>0.880</td>
</tr>
<tr>
<td>3</td>
<td>0.911</td>
</tr>
<tr>
<td>4</td>
<td>0.929</td>
</tr>
<tr>
<td>5</td>
<td>0.941</td>
</tr>
<tr>
<td>6</td>
<td>0.949</td>
</tr>
<tr>
<td>7</td>
<td>0.960</td>
</tr>
<tr>
<td>8</td>
<td>0.966</td>
</tr>
<tr>
<td>9</td>
<td>0.971</td>
</tr>
<tr>
<td>10</td>
<td>0.976</td>
</tr>
</tbody>
</table>

```r
#diagANOVA(lm(Life ~ log10(Metab), data=ex0826))
#diagANOVA(lm(log10(Life) ~ Metab, data=ex0826))
diagANOVA(m3)
```
In this sample of size n=95, correlation of the residuals in the qq-plot is r=0.983399

In the following table, if r < critical.r, then the qq-plot suggests the residuals are not normal:

<table>
<thead>
<tr>
<th>n</th>
<th>critical.r</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.832</td>
</tr>
<tr>
<td>2</td>
<td>0.880</td>
</tr>
<tr>
<td>3</td>
<td>0.911</td>
</tr>
<tr>
<td>4</td>
<td>0.929</td>
</tr>
<tr>
<td>5</td>
<td>0.941</td>
</tr>
<tr>
<td>6</td>
<td>0.949</td>
</tr>
<tr>
<td>7</td>
<td>0.960</td>
</tr>
<tr>
<td>8</td>
<td>0.966</td>
</tr>
<tr>
<td>9</td>
<td>0.971</td>
</tr>
<tr>
<td>10</td>
<td>0.976</td>
</tr>
</tbody>
</table>

Problem #3

Implementing the “shift-trick” to assess the median lifespan of mammals with a metabolic rate of 100.

```r
m4<-lm(log10(Life) ~ I(log10(Metab)-2), data=ex0826)
summary(m4)
```

```r
## Call:
## lm(formula = log10(Life) ~ I(log10(Metab) - 2), data = ex0826)
##
## Residuals:
##   Min     1Q Median     3Q    Max
##  -0.419  -0.22   0.000  -0.22   0.42
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  3.270069   0.033018  98.804  < 2e-16 ***
## I(log10(Metab) - 2) 0.202110   0.033018   6.150 6.36e-09 ***
## ---
## Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
##
## Residual standard error: 0.2758 on 93 degrees of freedom
## Multiple R-squared:  0.9745, Adjusted R-squared:  0.9738
## F-statistic: 3222 on 1 and 93 DF,  p-value: < 2.2e-16
```
## Min 1Q Median 3Q Max
## -0.48440 -0.13773 -0.00900 0.09498 0.65186

## Coefficients:
## Estimate Std. Error t value Pr(>|t|)
## (Intercept) 0.67209 0.02488 27.01 <2e-16 ***
## I(log10(Metab) - 2) 0.38215 0.02111 18.11 <2e-16 ***
## ---
## Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

## Residual standard error: 0.2156 on 93 degrees of freedom
## Multiple R-squared: 0.779, Adjusted R-squared: 0.7766
## F-statistic: 327.8 on 1 and 93 DF, p-value: < 2.2e-16

confint(m4)

## 2.5 % 97.5 %
## (Intercept) 0.6226789 0.7214959
## I(log10(Metab) - 2) 0.3402304 0.4240621

10^c(0.6226789, 0.7214959)

## [1] 4.194487 5.266182

The Workman-Hotelling and PI bands:

new = data.frame(Metab = seq(5,165000,length=100))
est.mean.ses = predict(m3,newdata=new,se.fit=T,interval="confidence")
head(data.frame(new,est.mean.ses$fit))

## Metab fit lwr upr
## 1 5.000 0.1749037 0.0861267 0.2636806
## 2 1671.616 1.1395044 1.0870534 1.1919555
## 3 3338.232 1.2542935 1.1940336 1.3145534
## 4 5004.848 1.3215033 1.2559765 1.3870300
## 5 6671.465 1.3692067 1.2997091 1.4387043
## 6 8338.081 1.4062156 1.3335272 1.4789040

cnf.BAND.WH.low <- est.mean.ses$fit[,1] - sqrt(2*qf(.95,2,93))*est.mean.ses$se.fit
cnf.BAND.WH.hi <- est.mean.ses$fit[,1] + sqrt(2*qf(.95,2,93))*est.mean.ses$se.fit

plot(Life ~ Metab, data = ex0826)
lines(new$Metab, 10^est.mean.ses$fit[,1], lty=1, lwd=2)  # add the fitted line
lines(new$Metab, 10^conf.BAND.WH.low, lty=4, lwd=2, col=4)  # lower WH CL
lines(new$Metab, 10^conf.BAND.WH.hi, lty=4, lwd=2, col=4)  # upper WH CL

# Now add the PIs:
pred.ses = predict(m3,newdata=new,se.fit=T,interval="prediction")
head(data.frame(new,est.mean.ses$fit))

## Metab fit lwr upr
## 1 5.000 0.1749037 0.0861267 0.2636806
## 2 1671.616 1.1395044 1.0870534 1.1919555
## 3 3338.232 1.2542935 1.1940336 1.3145534
## 4 5004.848 1.3215033 1.2559765 1.3870300
## 5 6671.465 1.3692067 1.2997091 1.4387043
## 6 8338.081 1.4062156 1.3335272 1.4789040
lines(new$Metab, 10^pred.ses$fit[,2], lty=2, lwd=2, col=3) # lower ind. PL
lines(new$Metab, 10^pred.ses$fit[,3], lty=2, lwd=2, col=3) # upper ind. PL
legend("bottomright",legend=c("Predicted lifespan","Family of 95% CIs for median lifespan of all mammals","Individual 95% PIs for one mammal"),lty=c(1,4,2),col=c(1,4,3),lwd=2)

# Get PI for mammals with Metab = 100
predict(m3,newdata=data.frame(Metab=100),se.fit=T,interval="prediction")

## $fit
## fit    lwr     upr
## 1 0.6720874 0.2410635 1.103111

## $se.fit
## [1] 0.02488086

## $df
## [1] 93

## $residual.scale
## [1] 0.215622

10^c(0.2410635,1.103111)

## [1] 1.742062e+00 1.267976e+01