STAT 516 sp 2025 exam 02

75 minutes, no calculators or notes allowed

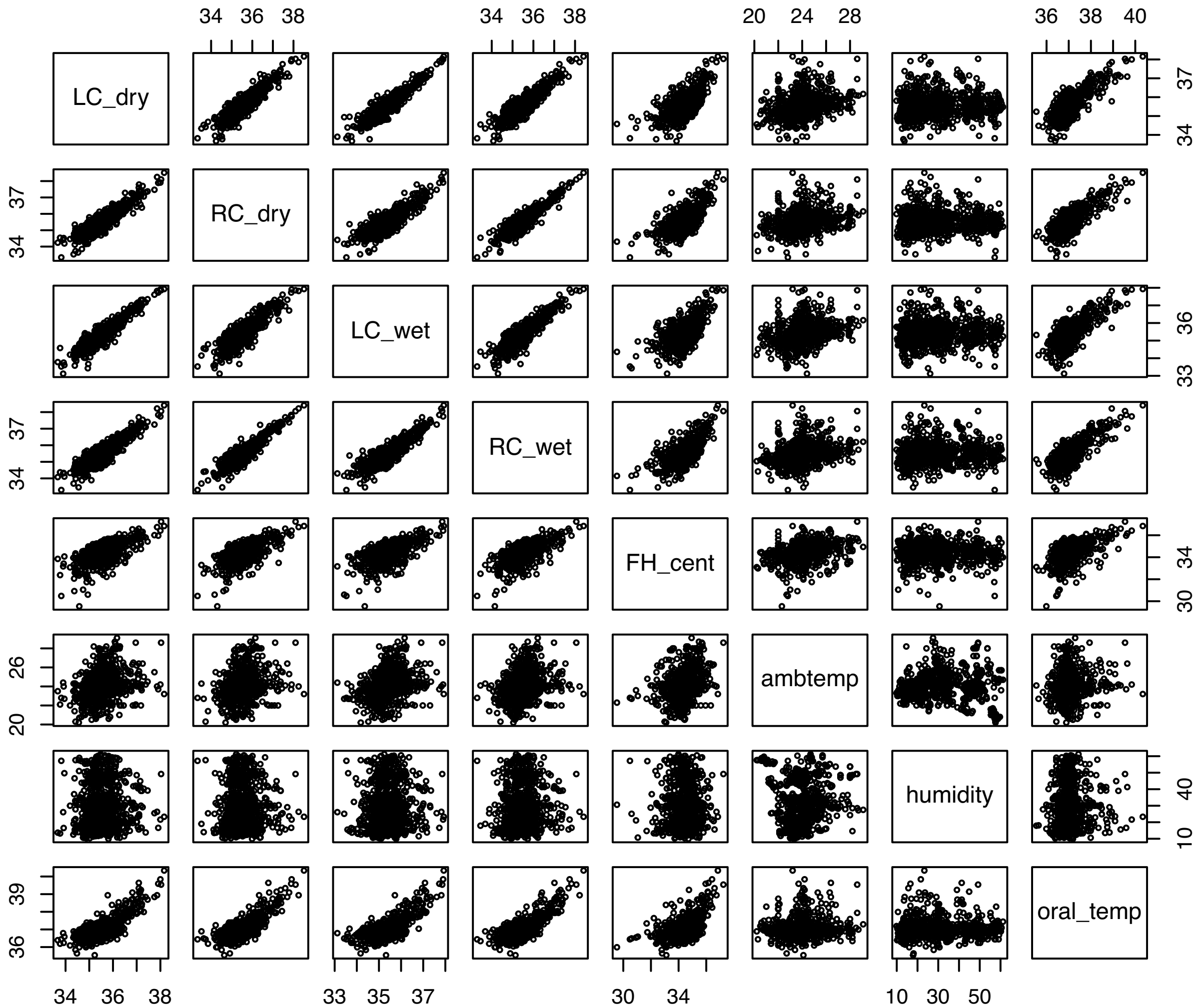
1. Multiple linear regression

In a study of the accuracy of infrared thermography (IRT) to determine humans' body temperatures from a thermal image of the face, the oral temperatures (regarded as the correct temperatures) of 933 subjects were recorded as well as the temperature readings from IRT at various regions of the subjects faces. Also recorded were the humidity level and the ambient temperature of the environment in which the IRT measurements were taken as well as the distance of the subject from the infrared camera. The table below describes the variables in the data set:

Variable	Description
`LC_Dry`	IRT temperature at dry area of left canthus
`LC_Wet`	IRT temperature at wet area of left canthus
`RC_Dry`	IRT temperature at dry area of right canthus
`RC_Wet`	IRT temperature at wet area of right canthus
`FH_cent`	IRT temperature at center of forehead
`ambtemp`	The ambient temperature
`humidity`	The humidity level
`distance`	Distance of the subject to the thermal camera
`oral_temp`	The subject's temperature as measured with an oral thermometer (the response)

Study carefully the R code and its output below:

```
plot(data, cex=.5)
```




```r
lm1 <- lm(oral_temp ~ LC_wet + FH_cent + ambtemp + humidity, data = data)
summary(lm1)
```

```

Call:
lm(formula = oral_temp ~ LC_wet + FH_cent + ambtemp + humidity,
    data = data)

Residuals:
    Min      1Q  Median      3Q     Max
-1.3257 -0.2249 -0.0386  0.1951  1.6777

Coefficients:
```

```
              Estimate Std. Error t value Pr(>|t|)
(Intercept) 16.1919941  0.6874924  23.552  < 2e-16 ***
LC_wet       0.5471070  0.0242316  22.578  < 2e-16 ***
FH_cent      0.0838784  0.0190259   4.409 1.16e-05 ***
ambtemp     -0.0597091  0.0093124  -6.412 2.29e-10 ***
humidity     0.0006504  0.0009033   0.720    0.472
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.3593 on 928 degrees of freedom
Multiple R-squared:  0.5074,	Adjusted R-squared:  0.5053
F-statistic:   239 on 4 and 928 DF,  p-value: < 2.2e-16
```

```r
lm2 <- lm(oral_temp ~ LC_wet + LC_dry + RC_wet + RC_dry
          + FH_cent + ambtemp + humidity, data = data)
summary(lm2)
```

```
Call:
lm(formula = oral_temp ~ LC_wet + LC_dry + RC_wet + RC_dry +
    FH_cent + ambtemp + humidity, data = data)

Residuals:
     Min       1Q   Median       3Q      Max
-1.13412 -0.20883 -0.02978  0.19202  1.68678

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) 13.1700628  0.6634732  19.850  < 2e-16 ***
LC_wet       0.0199131  0.0479429   0.415    0.678
LC_dry       0.2459987  0.0577504   4.260 2.26e-05 ***
RC_wet       0.2326471  0.0505648   4.601 4.79e-06 ***
RC_dry       0.2035901  0.0494070   4.121 4.12e-05 ***
FH_cent      0.0056129  0.0181624   0.309    0.757
ambtemp     -0.0537898  0.0084738  -6.348 3.42e-10 ***
humidity     0.0009555  0.0008222   1.162    0.245
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.3261 on 925 degrees of freedom
Multiple R-squared:  0.5956,	Adjusted R-squared:  0.5926
F-statistic: 194.6 on 7 and 925 DF,  p-value: < 2.2e-16
```

```r
library(car)
```

```
Warning: package 'car' was built under R version 4.4.1
```

```
Loading required package: carData
```

```r
vif(lm1)
```

```
  LC_wet  FH_cent  ambtemp humidity
1.626828 1.619943 1.154916 1.017395
```

```r
vif(lm2)
```

```
  LC_wet   LC_dry   RC_wet   RC_dry  FH_cent  ambtemp humidity
7.733028 9.891932 8.289578 7.914788 1.792566 1.161182 1.023502
```

Note that two models were fit: In the first model, only one of the four variables `LC_Dry`, `LC_Wet`, `RC_Dry`, and `RC_Wet` were included, whereas in the second model, all four of these variables were included as predictors.

(a) Report the value of R^2 for both models, and explain why it is higher for one model than for the other.

The first model had $R^2 = 0.5074$ and the second had $R^2 = 0.5956$.

The value of R^2 always increases when more predictors are added to the model.

(b) Report the p-value for testing the significance of `LC_Wet` in both models. Does one come to the same conclusion regarding the importance of this variable for predicting a subject's oral temperature?

In the first model it is very small, but in the second model it is 0.678.

From the first model we would conclude that it is important, but from the second model we would conclude that it is unimportant.

(c) Study carefully the figure displaying scatterplots for every pair of variables in the data set. How can this scatterplot help you understand your observation from part (b)? Give a detailed answer.

The four variables LC_Wet, LC_Dry, RC_Wet, and RC_Dry are highly correlated which each other. In consequence it is hard to distinguish the effect that each one has on the response. Including all these variables decreases the accuracy with which we estimate the effect of each one.

(d) Name two strategies we talked about in class for selecting a set of variables to keep in the model.

Forward stepwise, backward stepwise, best subsets, Lasso....

(e) Give one reason why one might not want to include all available variables in one's model.

For the reason that including more variables increases the variances of one's regression coefficient estimators.

(f) Explain the output of `vif(lm1)` and `vif(lm2)`. What is a "VIF" and why did the VIF change for the variable `LC_wet` from the first to the second model?

A "VIF" is a variance inflation factor. It gives the factor by which the variance with which we estimate the effect of a covariate is multiplied owing to its correlations with the other covariates. LC_wet has a much higher VIF in the second model due to the inclusion of additional covariates highly correlated with it.

2. One-way ANOVA

An experiment studied the effect of temperature on the failure time of a kind of sheathed tubular heater. At each of four temperatures, 1520°, 1620°, 1660°, and 1708°, the number of hours until failure was recorded for six heaters. The data are tabulated here:

Temperature	Failure time (hrs)
1520°	1953,2135,2471,4727,6134,6314
1620°	1190,1286,1550,2125,2557,2845
1660°	651,837,848,1038,1361,1543
1708°	511,651,651,652,688,729

Consider fitting the one-way ANOVA model to these data. Let

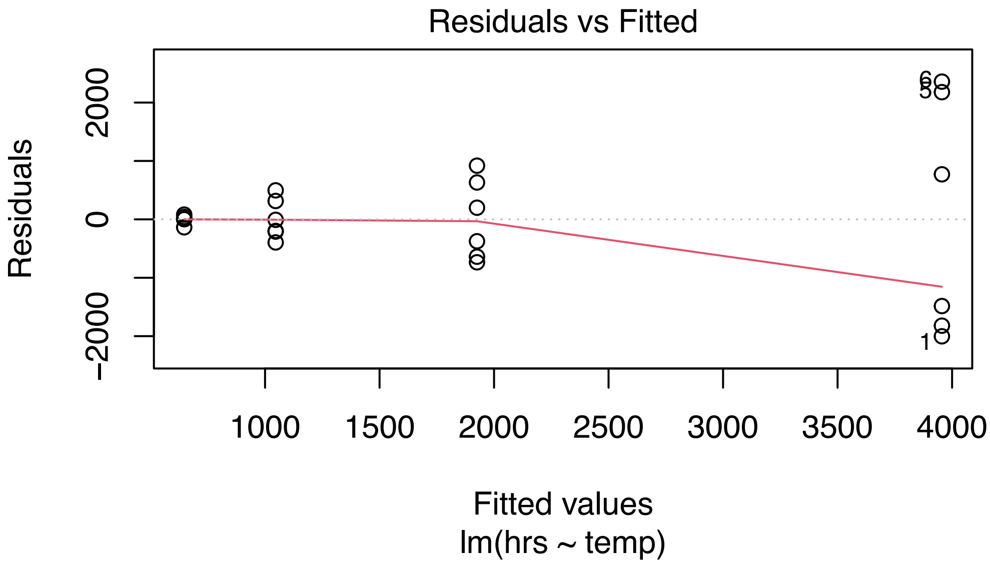
$$Y_{ij} = \mu + \tau_i + \varepsilon_{ij}$$

for $i = 1, \dots, a$, $j = 1, \dots, n_i$, where the ε_{ij} are independent Normal$(0, \sigma^2)$ random variables.

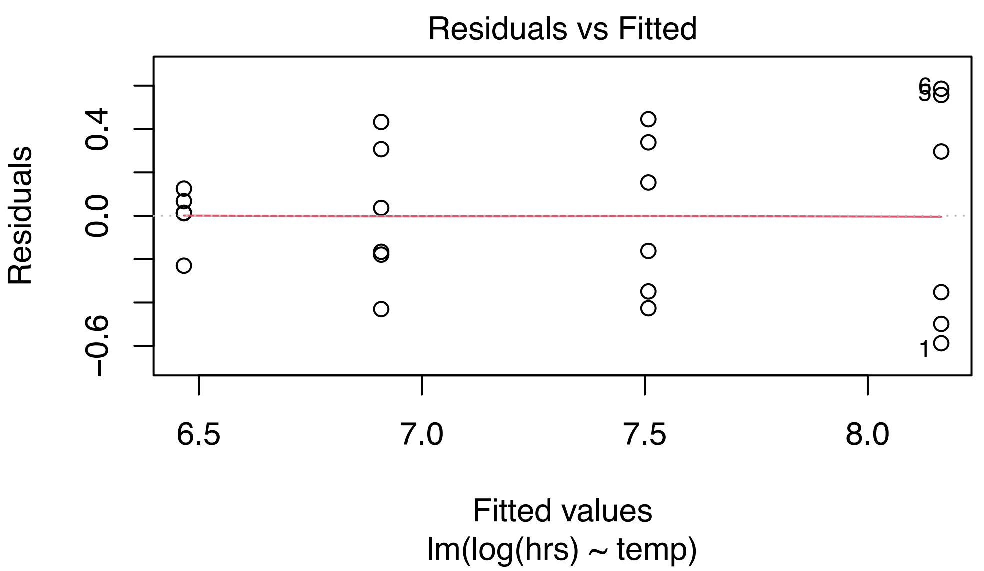
The R code below reads in the data and fits two one-way ANOVA models: One using the original response values and one using the natural log of the response values. Residuals versus fitted values plots for the two models are shown.

```
hrs <- c(1953,2135,2471,4727,6134,6314,
         1190,1286,1550,2125,2557,2845,
         651,837,848,1038,1361,1543,
         511,651,651,652,688,729)
temp <- as.factor(c(rep(1520,6),rep(1620,6),rep(1660,6),rep(1708,6)))

lm_hrs <- lm(hrs~temp)
plot(lm_hrs,which = 1)
```




```r
lm_loghrs <- lm(log(hrs)~temp)
plot(lm_loghrs,which = 1)
```




```r
summary(lm_loghrs)
```

```
Call:
lm(formula = log(hrs) ~ temp)
```

```
Residuals:
     Min       1Q   Median       3Q      Max 
-0.58769 -0.25978  0.01279  0.29893  0.58571 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)   8.1648     0.1507  54.169  < 2e-16 ***
temp1620     -0.6567     0.2132  -3.081  0.00589 ** 
temp1660     -1.2559     0.2132  -5.892 9.20e-06 ***
temp1708     -1.6983     0.2132  -7.967 1.24e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.3692 on 20 degrees of freedom
Multiple R-squared:  0.7823,	Adjusted R-squared:  0.7497 
F-statistic: 23.96 on 3 and 20 DF,  p-value: 7.912e-07
```

```r
TukeyHSD(aov(log(hrs) ~ temp),conf.level = .99)
```

```
  Tukey multiple comparisons of means
    99% family-wise confidence level

Fit: aov(formula = log(hrs) ~ temp)

$temp
                diff       lwr         upr     p adj
1620-1520 -0.6567415 -1.413104  0.09962083 0.0277405
1660-1520 -1.2558618 -2.012224 -0.49949948 0.0000508
1708-1520 -1.6983332 -2.454696 -0.94197085 0.0000007
1660-1620 -0.5991203 -1.355483  0.15724202 0.0488199
1708-1620 -1.0415917 -1.797954 -0.28522935 0.0004796
1708-1660 -0.4424714 -1.198834  0.31389095 0.1950191
```

contains zero

do not contain zero.

(a) Explain carefully why the model which uses the natural log of the responses will probably yield more reliable inferences.

The residuals from the model using log(hrs) as the response have a more equal spread across the treatment groups.

(b) Use the R output to compute the mean of the natural log of the observed failure times in the 1620° temperature group.

$\bar{Y}_{1620.} = 8.1648 + (-0.6567) = \begin{array}{r} 8.1648 \\ -0.6567 \\ \hline 7.5081 \end{array} = 7.5081$

(c)

Source	Df	SS	MS	F	p-value
Treatment	viii	xiv	vi	xi	xii
Error	i	x	iii		
Total	ix	xiii			

$F_{stat} = \frac{MS_{TRT}}{MS_{Error}}$

$MS_{TRT} = MS_{Error} * F_{stat}$

Fill the blank ANOVA table with numerals from among (i)–(xx) to indicate which of the below values belong where (more values are listed than are needed):

(i) 20	(ii) 0.3692	(iii) $(0.3692)^2$	(iv) 1.24×10^{-7}
(v) 0.7497	(vi) $(23.96)(0.3692)^2$	(vii) 8.1648	(viii) 3
(ix) 23	(x) $20(0.3692)^2$	(xi) 23.96	(xii) 7.912×10^{-7}
(xiii) $20(0.3692)^2 + 3(23.96)(0.3692)^2$	(xiv) $3(23.96)(0.3692)^2$	(xv) 0.7823	(xvi) $(23.96)^2$
(xvii) $20(23.96)(0.3692)^2$	(xviii) $20(0.3692)$	(xix) 24	(xx) 17

(d) Based on the model with the natural log of the responses, is there evidence to conclude that the temperature is related to the failure time? Explain your answer carefully.

Yes, since the p-val in the ANOVA table is very small, we reject

$H_0: \mu_1 = \mu_2 = \mu_3 = \mu_4$,

where $\mu_1, \mu_2, \mu_3, \mu_4$ represent the mean failure times at the four temperatures.

~~(e) What do we learn from the Tukey multiple comparisons of means part of the R output? Give a detailed answer.~~

(e) ~~(f)~~ In the experiment, under which temperature did the sheathed tubular heaters last the longest, on average, before failing? Based on the R output, can we conclude that under this temperature, the mean failure time was statistically significantly greater than the other means? Explain your answer.

The greatest observed average failure time was at 1520°, the lowest temp.

From the Tukey's pairwise comparisons at 99% confidence, we cannot say that this mean is greater than that at 1620, but it is greater than the mean at the other two temperatures.

(f) (5) If one wished only to compare the mean failure times at the temperatures 1520° and 1620°, one would construct the confidence interval $\bar{Y}_{1.} - \bar{Y}_{2.} \pm 0.4446392$, where the margin of error involves a quantile from a t-distribution. With Tukey's method, however, the confidence interval for comparing these means is constructed as $\bar{Y}_{1.} - \bar{Y}_{2.} \pm 0.5966148$. Explain the difference between the two intervals and explain the reason for the difference.

The Tukey interval is wider because the Tukey intervals are calibrated to capture the true differences between all pairs of means simultaneously, whereas the t-interval is calibrated to capture only one single difference in means. So the former is wider than the latter.

(g) (4) What additional plot should one generate in order to ensure that the data from this experiment satisfies the assumptions of the one-way ANOVA model?

A normal QQ plot of the residuals.

3. Two-way factorial design

In order to understand how the temperature and salinity of water effect the growth of shrimp raised in aquariums, three aquiriums were set to each combination of temperatures (25° and 35° Celcius) and salinity levels (10%, 25%, and 40%) and the weight gain of the shrimp over a period of four weeks recorded for each aquarium. The experiment resulted in the data tabulated below:

Temperature	Salinity	Weight gain	$\bar{Y}_{ij.}$
25°	10%	86,52,73	70.33
	25%	544,371,482	465.67
	40%	390,290,397	359.00
35°	10%	439,436,349	408.00
	25%	249,245,330	274.67
	40%	247,277,205	243.00

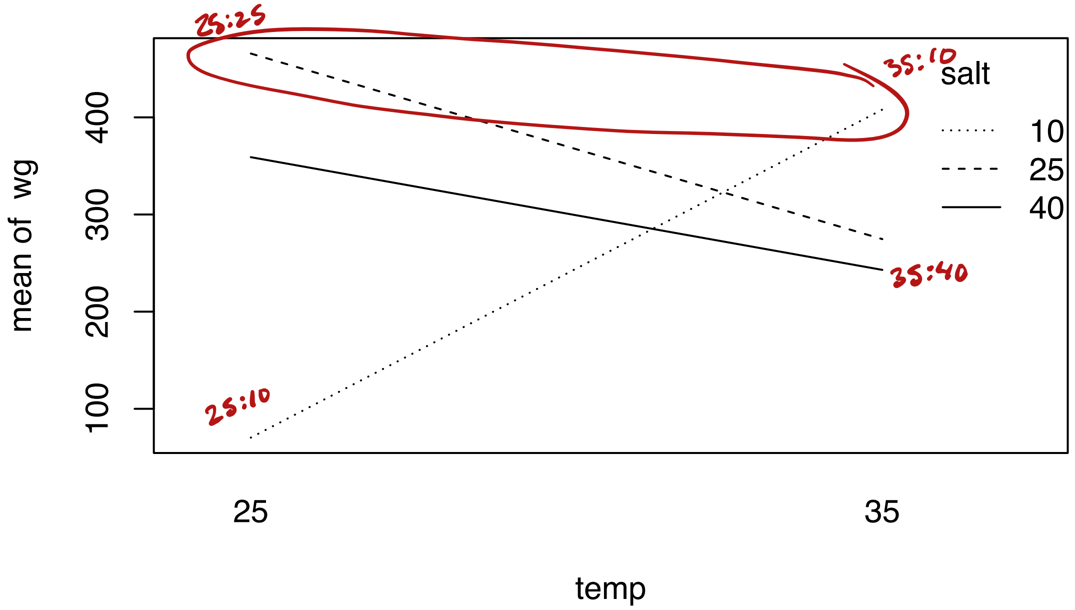
Consider the following model, assuming that the assumptions are satisfied: Let

$$Y_{ijk} = \mu + \tau_i + \gamma_j + (\tau\gamma)_{ij} + \varepsilon_{ijk},$$

$i = 1, \ldots, a$, $j = 1, \ldots, b$, and $k = 1, \ldots, n_{ij}$, where the ε_{ijk} are independent Normal$(0, \sigma^2)$ random variables. Let i index the temperature and j index the salinity level. Consider the R code below and its output:

```r
wg <- c(86,52,73,544,371,482,390,290,397,
        439,436,349,249,245,330,247,277,205)
temp <- as.factor(c(rep(25,9),rep(35,9)))
salt <- as.factor(c(1,1) %x% c(rep(10,3),rep(25,3),rep(40,3)))

interaction.plot(temp,salt,wg)
```




```r
lm_shrimp <- lm(wg ~ temp + salt + temp:salt)

TukeyHSD_out <- TukeyHSD(aov(lm_shrimp))
TukeyHSD_out$`temp:salt`
```

```
                  diff        lwr       upr        p adj
35:10-25:10  337.66667  188.36777 486.96557 7.262611e-05
25:25-25:10  395.33333  246.03443 544.63223 1.455431e-05
35:25-25:10  204.33333   55.03443 353.63223 6.247420e-03
25:40-25:10  288.66667  139.36777 437.96557 3.297825e-04
35:40-25:10  172.66667   23.36777 321.96557 2.060247e-02
25:25-35:10   57.66667  -91.63223 206.96557 7.812446e-01
35:25-35:10 -133.33333 -282.63223  15.96557 9.059335e-02
25:40-35:10  -49.00000 -198.29890 100.29890 8.713239e-01
35:40-35:10 -165.00000 -314.29890 -15.70110 2.757719e-02
35:25-25:25 -191.00000 -340.29890 -41.70110 1.028917e-02
25:40-25:25 -106.66667 -255.96557  42.63223 2.300708e-01
35:40-25:25 -222.66667 -371.96557 -73.36777 3.185672e-03
25:40-35:25   84.33333  -64.96557 233.63223 4.476993e-01
35:40-35:25  -31.66667 -180.96557 117.63223 9.766950e-01
35:40-25:40 -116.00000 -265.29890  33.29890 1.680994e-01
```

yes diff

no diff

```r
anova(lm_shrimp)
```

```
Analysis of Variance Table

Response: wg
          Df Sum Sq Mean Sq F value    Pr(>F)
temp       1    470     470  0.1587  0.697379
salt       2  51537   25768  8.6953  0.004633 **
temp:salt  2 245463  122732 41.4144 4.106e-06 ***
Residuals 12  35562    2964
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

$a = 2$ $\quad a-1 = 1$

$b = 3$ $\quad b-1 = 2$

$(a-1)(b-1) = 2$

$n = 3$

$ab(n-1) = 2 \cdot 3 \cdot 2 = 12$

(a) Fill in the missing values in the above ANOVA table (five values have been removed).

(b) Give the value of $\hat{\sigma}^2$.

$\hat{\sigma}^2 = 2964.$

(c) Give the value from the ANOVA table which reflects the ratio of the variation in the responses owing to the effect of the different temperatures over the variation owing to random differences from aquarium to aquarium.

This is $F_{temp} = \dfrac{MS_{temp}}{MS_{Error}} = 0.1587.$

(d) Give the value of $2\sum_{j=1}^{3} 3(\bar{Y}_{.j.} - \bar{Y}_{...})^2$, which appears in the ANOVA table.

This is $SS_{salt} = 51537.$

(e) Can one say that one temperature is better than the other? Explain your answer. What would you say if a shrimp supplier asked, "At which temperature should I keep my aquariums?"

It's not that simple: At low salinity (10%), it seems that a warmer temperature is better, but at higher salinities (25%, 40%) a cooler temperature is better. So one must first ask what salinity the supplier uses.

(f) If someone said that the temperature is irrelevant to the growth rate of shrimp because of the p-value 0.697379 appearing in the table, what would you say in response?

Temperature is important because of its interaction with salinity, which is statistically significant. That means temperature, even though it appears not to have a "main effect", plays nonetheless an important role.

(g) Give an interpretation to the value 4.106×10^{-6} appearing in the ANOVA table.

This is the p-value which tells us that there is a significant interaction between temperature and salinity.

(h) Based on the R output, can you recommend a single best combination of temperature and salinity for fostering the growth of shrimp? If so, what is it; if not, why not?

No. It seems there may be no difference between 25° × 25% salinity and 35° × 10% salinity, which are both contending for having the greatest mean. This is according to the Tukey output.

(i) Based on the R output, can you identify a single worst combination of temperature and salinity for fostering the growth of shrimp? If so, what is it; if not, why not?

Yes. 25° × 10% salinity seems to be the worst scenario, from the Tukey output, this mean is lower than all the others.

(j) Suppose one of the aquariums had started leaking during the experiment so that the weight gain of the shrimp in this aquarium had to be excluded from the analysis, resulting in only two values for one of the temperature and salinity combinations. What do we call the situation in which the number of replicates is not the same for all combinations of factor levels? How does this complicate the analysis?

The experiment would in this case be <u>unbalanced</u>.

Some formulas get more complicated, and one must be careful to use Type III instead of Type I (the default in R) sums of squares.