Key for HW #11

1.
Check these assumptions.
   i) The straight line relationship is correct.
   ii) Errors $e_i$ are independent.
   iii) Errors $e_i$ have homogeneous variance.
   iv) Errors $e_i$ have normal distribution

(a) The variance in $Y$ is not equal for all $X$, but increases with increasing $X$.
(b) The relationship between $X$ and $Y$ is not linear.
(c) The residuals are not normally distributed.
(d) The very large residual (middle top) is an outlier, and it is of concern. It suggests either a non-normal distribution of residuals or that the variance of residuals is not homogeneous.

2.
   a.
   The number of added nutrients should be the explanatory variable ($X$), as this was controlled by the experimenters. The response variable of interest ($Y$) is the number of plant species supported.

R code:

```r
dat=read.csv("17q14GrasslandNutrientsPlantSpecies.csv")
plot(dat[,1],dat[,2],xlab="Number of nutrients added",ylab="Number of plant species")
```
b. Regression Result:

|                | Estimate | Std. Error | t value | Pr(>|t|) |
|----------------|----------|------------|---------|----------|
| (Intercept)    | 34.110   | 2.599      | 13.12   | 1.08e-06 *** |
| nutrients      | -3.339   | 1.098      | -3.04   | 0.0161 *  |

---

Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 5.336 on 8 degrees of freedom
Multiple R-squared: 0.536,   Adjusted R-squared: 0.478
F-statistic: 9.241 on 1 and 8 DF,  p-value: 0.01607

R code:
```r
fit=lm(species~nutrients,data=dat)
summary(fit)
```

b = -3.339,  SE = 1.10.
According to this result, we can say that 3.34 species are lost on average for each nutrient added.

c. 

![Graph showing the relationship between number of nutrients added and number of plant species.](image)

R code:
```r
plot(dat[,1],dat[,2],xlab="Number of nutrients added",ylab="Number of plant species")
fit=lm(dat[,2]~dat[,1])
abline(fit)
```
We can use a coefficient of determination (also called 'multiple R-squared' in the R output), which we can calculate as $\frac{SS(\text{nutrients})}{SS(\text{total})}$, $r^2$.

\[ r^2 = 0.536 \]

d.

H0: There is no treatment effect ($\beta_1 = 0$)

HA: There is a treatment effect ($\beta_1 \neq 0$)

We'll use F-test.

**ANOVA result:**

<table>
<thead>
<tr>
<th></th>
<th>Df</th>
<th>Sum Sq</th>
<th>Mean Sq</th>
<th>F value</th>
<th>Pr(&gt;F)</th>
</tr>
</thead>
<tbody>
<tr>
<td>nutrients</td>
<td>1</td>
<td>263.11</td>
<td>263.112</td>
<td>9.2406</td>
<td>0.01607 *</td>
</tr>
<tr>
<td>Residuals</td>
<td>8</td>
<td>227.79</td>
<td>28.474</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

---

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0

R code:

```r
fit = lm(species ~ nutrients, data = dat)
anova(fit)
```

Since the p-value is smaller than 0.05, we can reject H0. It means that adding more nutrients reduces the number of plant species.
3. (a)

R code:

```r
dat=read.csv("17q15EarthwormsAndNitrogen.csv")
plot(dat[,1],dat[,2],xlab="number of earthworm species", ylab="Total nitrogen content(%)")
```

b.

\[
 b_1 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2} = \frac{2.453}{86.359} = 0.028 \text{ nitrogen per species of earthworm}
\]

\[
 b_0 = \bar{y} - b_1 \bar{x} = 0.215 - 0.028 \cdot 2.205 = 0.152 \text{ nitrogen}
\]
R code:
plot(dat[,1],dat[,2],xlab="number of earthworm species", ylab="Total nitrogen content(%)")
fit=lm(nitrogen~worm.species,data=dat)
abline(fit)

c1.
The unit of estimate of slope, b1, is %Nitrogen per earthworm species.

c2.
H0: There is no effect of the number of earthworm species on nitrogen content (β1 = 0)
Ha: There is an effect (β1 ≠ 0)

ANOVA table:

<table>
<thead>
<tr>
<th></th>
<th>d.f.</th>
<th>SS</th>
<th>MS</th>
<th>F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Reg</td>
<td>1</td>
<td>0.0697</td>
<td>0.0697</td>
<td>8.69</td>
</tr>
<tr>
<td>Error</td>
<td>37</td>
<td>0.2967</td>
<td>0.008</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>38</td>
<td>0.366</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

SS Total = ∑(yi - ȳ)^2 = 0.366
SS Reg = β1 ∑(xi - x̄)(yi - ȳ) = 0.028 · 2.453 = 0.069681
SS Error = SS Total - SS Reg
            = 0.366 - 0.069681
            = 0.296688
MS Error = SS Error / 37 = 0.296688 / 37 = 0.008019
F = MS Reg / MS Error = 0.069681 / 0.008019 = 8.69
F ~ F distribution with df₁ = 1 and df₂ = 37
p - value = P(F > 8.69)
            = 1 - pf(8.69,137), using R
            = 0.005513917

Since the p-value is very small (p~0.005), there is strong evidence to reject H0. It means that there is strong evidence that the number of earthworm species increases the nitrogen content.

s_e^2 = MS Error = 0.008019
s_e = √0.008019 = 0.09
d1. 
\[ \hat{y}_{\text{pred}} = 0.152 + 0.028X = 0.152 + 0.028 \cdot 5 = 0.294 \]

d2. 
\[ SE_{\hat{y}_{\text{pred}}} = s_e \sqrt{\frac{1}{n} + \frac{(x^* - \bar{x})^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2} + 1} \]
\[ = \sqrt{0.008019} \sqrt{\frac{1}{39} + \frac{(5 - 2.205)^2}{86.359}} + 1 \]
\[ = 0.095 \]

95% CI for the new value
\[ 0.294 \pm t_{0.025,37} \cdot 0.095 \]
\[ \Rightarrow 0.294 \pm 2.026 \cdot 0.095 \]
\[ \therefore [0.102, 0.485] \]

e. 
\[ SE_{b_1} = \frac{s_e}{\sqrt{\sum_{i=1}^{n} (x_i - \bar{x})^2}} = \frac{\sqrt{0.008019}}{\sqrt{86.359}} = 0.0096 \]

f. 
95% CI for \( \beta_1 \)
\[ b_1 \pm t_{0.025,37} \cdot SE_{b_1} \]
\[ 0.028 \pm 2.03 \cdot 0.0096 \]
\[ \therefore [0.009, 0.047] \]

g. 
Check these assumptions.
   i) Independence of observations between and within samples.
   ii) In each treatment, observations come from a normal distribution or the sample size is large.
   iii) Equal variances.
This experimental design is randomized design.

R code:
fit=lm(nitrogen~worm.species,data=dat)
par(mfrow=c(2,2))
plot(fit)

When we look at these plots, the assumptions seem to be met. The normal quantile plot of the residuals (top right) looks straight enough and suggests that the normal assumption (ii) is met. The plot of the residual values versus predicted values (top left) shows a somewhat homogeneous variance of residuals around the horizontal line. The variance might be a bit higher on the right-hand side, but nothing to be worried about.
According to this scatter plot, the relation between body weight and nitrogen balance is not linear. So, it’s not appropriate to apply a regression analysis. Furthermore, the researcher measured nitrogen balance three times for each four male swine. So, this sample is not independent. Thus, the independence assumption is violated.