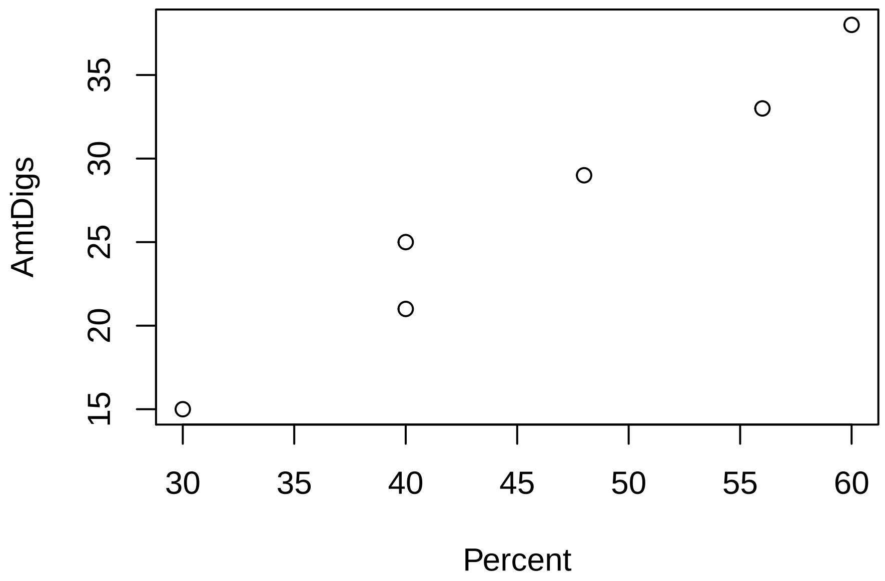
BTRY 6020 Homework I Solution

Question 1

a) See R codes below.

```r
# Input the data
x <- c(30, 40, 40, 48, 56, 60)
y <- c(15, 25, 21, 29, 33, 38)
# Plot scatterplot
plot(x, y, xlab="Percent", ylab="AmtDigs")
```



Since the data appear approximately linear in the graph above, linear regression over the given range of x-values in the data appears to be appropriate.

For the rest of the problems, we will use results from the table Figure 1 followed by verification using R.

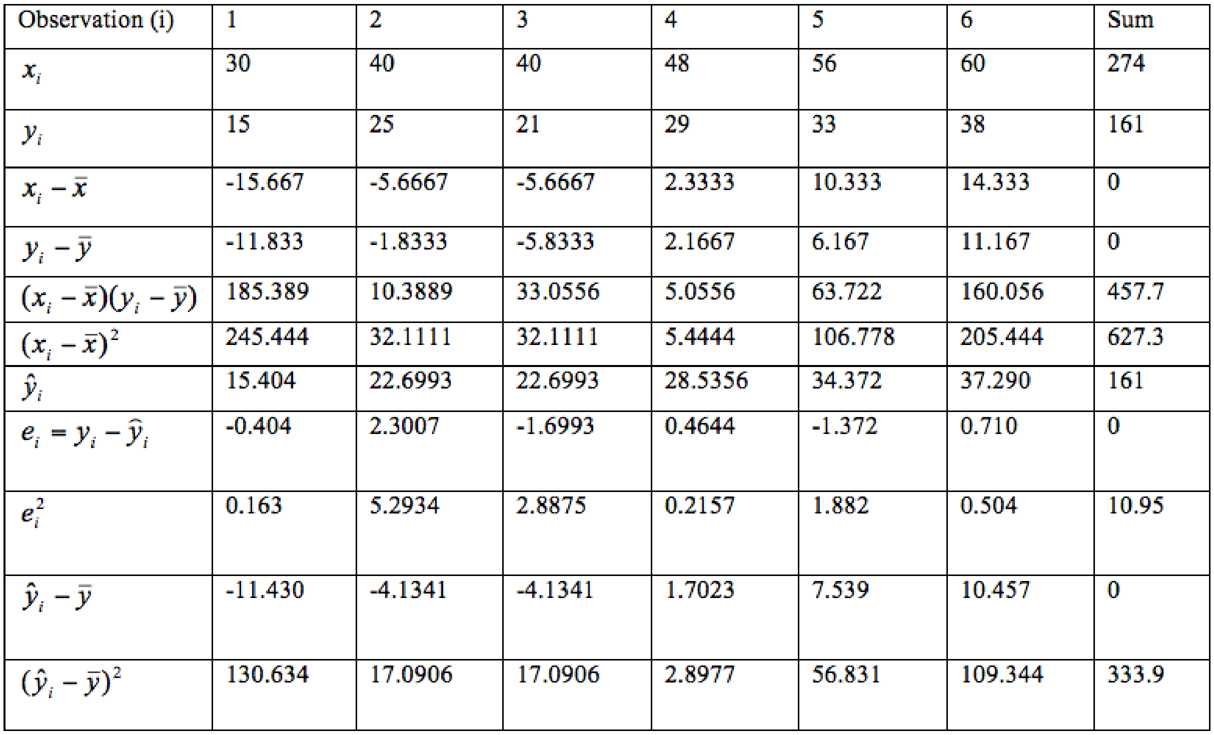
b) $\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}} = \frac{457.7}{627.3} = 0.730$

We estimate that for every percent increase in the amount of detergent-solubles in a 300 gram amount of feed, a white-tail deer will digest an additional .730 grams of these detergent-solubles.

c) $\hat{\beta}_o = \bar{y} - b_1\bar{x} = 26.83 - 0.73 \times 45.67 = -6.51$

The interpretation here is that when the percent detergent-solubles in the 300 grams of feed is 0, then -6.51 grams of them should be digested. Since negative grams is not possible, the estimated intercept is not realistic.

```r
# Check b) and c) in R
m <- lm(y~x)
coef(m)
```



Observation (i)	1	2	3	4	5	6	Sum
x_i	30	40	40	48	56	60	274
y_i	15	25	21	29	33	38	161
$x_i - \bar{x}$	-15.667	-5.6667	-5.6667	2.3333	10.333	14.333	0
$y_i - \bar{y}$	-11.833	-1.8333	-5.8333	2.1667	6.167	11.167	0
$(x_i - \bar{x})(y_i - \bar{y})$	185.389	10.3889	33.0556	5.0556	63.722	160.056	457.7
$(x_i - \bar{x})^2$	245.444	32.1111	32.1111	5.4444	106.778	205.444	627.3
$\hat{y}_i$	15.404	22.6993	22.6993	28.5356	34.372	37.290	161
$e_i = y_i - \hat{y}_i$	-0.404	2.3007	-1.6993	0.4644	-1.372	0.710	0
e_i^2	0.163	5.2934	2.8875	0.2157	1.882	0.504	10.95
$\hat{y}_i - \bar{y}$	-11.430	-4.1341	-4.1341	1.7023	7.539	10.457	0
$(\hat{y}_i - \bar{y})^2$	130.634	17.0906	17.0906	2.8977	56.831	109.344	333.9

Figure 1: Table of Different Quantities

```
## (Intercept)           x
##   -6.482465    0.729543
```

d) The predicted value $\hat{y}_i$ for each x_i can be found in the table. From the values on the table, $SSR = 333.9$.

```
# Compute fitted values
fits <- fitted(m)
fits
```

```
##        1        2        3        4        5        6
## 15.40383 22.69926 22.69926 28.53560 34.37194 37.29012
```

```
# Compute SSR
sum((fits-mean(y))^2)
```

```
## [1] 333.8875
```

e) The residuals e_i can be found in the table. From the table, we see that $SSE = 10.95$.

```
# Compute residuals
e <- y - fits
e
```

```
##          1          2          3          4          5          6
## -0.4038257  2.3007439 -1.6992561  0.4643996 -1.3719447  0.7098831
```

```
# Compute SSE
sum(e^2)
```

```
## [1] 10.9458
```

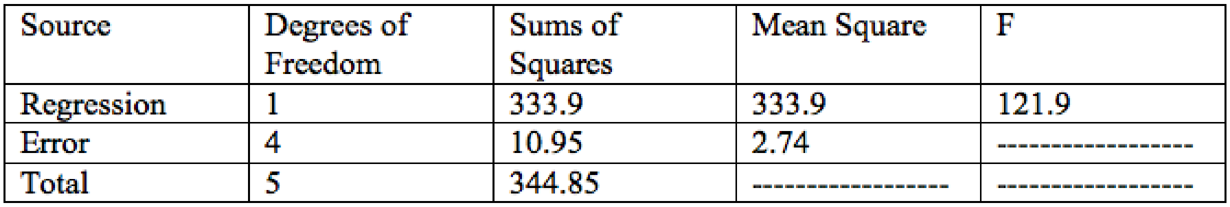
f) $r^2 = \frac{SSR}{SSTo} = \frac{SSR}{SSR+SSE} = \frac{333.9}{333.9+10.95} = \frac{333.9}{344.85} = 0.968$

Interpretation: Of all the factors that account for the differences (or variation) in number of grams of detergent solubles digested in 300 grams of feed, the percent of detergent solubles in the feed accounts for 96.8% of these differences.

Here we have made no distributional assumptions about the response y: the estimated line was obtained by calculus and the breakdown of the total sums of squares into error (SSE) and regression (SSR) was done with algebra.

Since the variation in number digested can be partitioned into explained (SSR) and unexplained (SSE) variation, we see that when we put the total amount of explained variation over the total amount of variation, the result (r^2) is the proportion (or percent) of variability explained by the predictor variable x.

g) See Figure 2 for the ANOVA table.



Source	Degrees of Freedom	Sums of Squares	Mean Square	F
Regression	1	333.9	333.9	121.9
Error	4	10.95	2.74	-------------------
Total	5	344.85	-------------------	-------------------

Figure 2: ANOVA Table

h) The linear model is $y_i = \beta_0 + \beta_1 x_i + \epsilon_i \quad$ for $i = 1, \ldots, 6$. The estimated model is $\hat{y}_i = \hat{\beta}_o + \hat{\beta}_1 x_i$.

1. y_i is random and distributes independently as $N(\beta_0 + \beta_1 x_i, \sigma^2)$.
2. β_0 is true (population) intercept and is a constant (parameter).
3. β_1 is true (population) slope and is a constant (parameter).
4. ϵ_i is random and distributes iid as $N(0, \sigma^2)$.
5. $\hat{\beta}_o$ and $\hat{\beta}_1$ are the intercept and slope estimated from the data.
6. $\hat{y}_i$, is the predicted value of y at x_i, as a function of $\hat{\beta}_o$ and $\hat{\beta}_1$.

i) The fours assumptions are (with $y_i \sim indN(\beta_0 + \beta_1 x_i, \sigma^2)$)

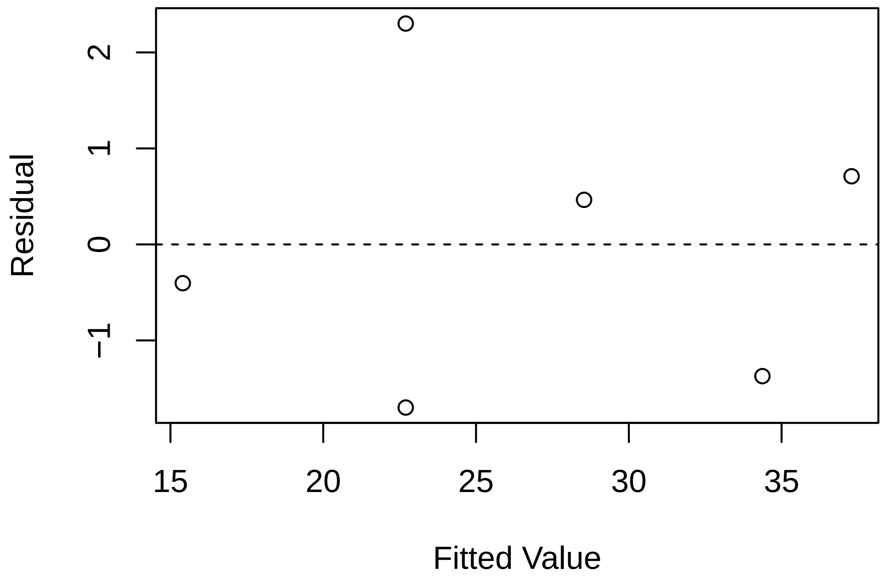
1. The data are independent: Independence would be guaranteed by random sampling, however, there is nothing in the information given which tells us this was a random sample; the sampling scheme needs to be checked.
2. Y is normally distributed at each X: Check the normal probability plot of residuals to determine if Y is normally distributed.
3. The means change linearly with X: Look at a plot of Y versus X and also plot of residuals vs fits to check linearity.
4. The variance is constant and independent of X: Look at a plot of residuals versus fitted values to check constant variance.

Note that there is always a final, implicit assumption: That outliers are not driving our conclusions. From our scatterplot we see there are no apparent outliers.

j) $\hat{\sigma} = \sqrt{MSE} = \sqrt{2.74} = 1.65$

k) See figure below.

```r
# Residuals vs. fitted values
plot(fits, resid(m), xlab="Fitted Value", ylab="Residual")
abline(h=0, lty=2)
```



This plot shows that the variance is approximately constant and that the relation between the mean value of y and x is approximately linear.

l) Since we only care one deer for one feeding, we are trying to predict a single realization of a random variable, so we use a 95% prediction interval (or 90%, 95% etc.):

$\hat{y} \pm t_{n-2,1-\frac{\alpha}{2}}\sqrt{MSE\left(1+\frac{1}{n}+\frac{(x-\bar{x})^2}{S_{xx}}\right)} = 28.53 \pm t_{4,0.975}\sqrt{2.74\times\left(1+\frac{1}{6}+\frac{(48-45.67)^2}{627.3}\right)} = 28.53 \pm 2.776 \times 4.97 = (23.56, 33.50).$

```r
# Find t-value
qt(0.975, 4)
```

```
## [1] 2.776445
```

The probability that a deer fed 300 grams of feed made up of 48% detergent-solubles will digest between 23.56 and 33.50 grams of these solubles is .95.

m) Since now we are looking at the mean value of y at a given x, we use a 95% confidence interval for the mean value of y when $x = 48$ (a 95% confidence interval for $\mu_{y|x=48}$):

$\hat{y} \pm t_{n-2,1-\frac{\alpha}{2}}\sqrt{MSE\left(\frac{1}{n}+\frac{(x-\bar{x})^2}{S_{xx}}\right)} = 28.53 \pm 2.776 \times 0.693 = (26.61, 30.45)$

So we are 95% confident that a deer fed 300 grams of feed that was 48 percent detergent-solubles would digest, on the average, between 26.61 and 30.45 grams of these solubles per day.

n) $H_0 : \beta_1 = 0$ and $H_1 : \beta_1 > 0$

T statistic: $t = \frac{b_1 - 0}{SE(b_1)} = \frac{0.73-0}{0.066} = 11.05$ where $SE(b_1) = \sqrt{\frac{MSE}{S_{xx}}} = \sqrt{\frac{2.74}{627.3}} = 0.066$. The p-value is $p = P(t_4 > 11.05) = 0.0002$

```r
# summary table
summary(m)$coef
```

```
##              Estimate Std. Error   t value     Pr(>|t|)
## (Intercept) -6.482465 3.09076994 -2.097363 0.1039623891
## x            0.729543 0.06604571 11.046032 0.0003819108
```

```r
# pull out p-value for beta_1
summary(m)$coef[2, 4]/2 # since we are doing one-tail t-test
```

```
## [1] 0.0001909554
```

Conclusion: Since $p < \alpha = 0.05$, we reject H_0 at $\alpha = 0.05$ and conclude H_1 is true, and that the slope is greater than 0, so the number of grams of detergent solubles digested increases with the percent of such solubles in the feed.