BTRY 6020 Lab II Solutions

a) $R^2 = 0.9777$. This is evidence of an extremely good fit to the line.

b) **Intercept:** the null hypothesis is: $H_0 : \beta_0 = 0$ vs $H_a : \beta_0 \neq 0$.

The test statistic is

```
tstat0 = (5.01364-0)/(3.47318)
tstat0
```

```
## [1] 1.44353
```

The pvalue is then calculated as

```
2*(1-pt(abs(tstat0), df = 44))
```

```
## [1] 0.1559538
```

At the 0.05 significance level, we do not have evidence to reject the null hypothesis that the intercept is equal to 0.

Slope: the null hypothesis is $H_0 : \beta_1 = 1$ vs $H_a : \beta_1 \neq 1$.

The test statistic is

```
tstat1 = (0.90905-1)/0.02068
tstat1
```

```
## [1] -4.397969
```

The pvalue is then calculated as

```
2*(1-pt(abs(tstat1), df = 44))
```

```
## [1] 6.833957e-05
```

At the 0.05 significance level, we have evidence to reject the null that the slope is equal to 1.

c) There is an obvious outlier, observation 16.

d) We should check to see if this value is a true outlier-first check that to see that it was entered correctly.

e) The intercept became smaller, which is closer to zero; the slope didn't change much. Notice that the Residual standard error has gone down from 12.55 to 5.61. It seems that the outlier was influencing the intercept estimate, and DRASTICALLY inflating the estimate of the variance.

f) The new residual plot shows increasing variation in residuals as the fitted values increase. The variances of the regression estimates will be wrong-too large for small fits and too small for the bigger fits. As far as the confidence intervals go, the one at x=70 will be too wide, while the CI at x=280 will be too narrow. This plot suggests that a transformation is in order. The fact that the plot looks linear but we have issues with the residuals implies we need to transform both variables.

g) We expect β_0 to be 0, and β_1 to be 1.

h) Everything looks OK except for one hairy, scary looking outlier in the upper left-hand corner.

i) Observation 11 has both extreme leverage and Cook's distance, which indicate that it might be an outlier. There are a couple of other data points having relatively higher leverage than the criteria : $2p/n = (2 * 2)/46 = 0.087$. But the differences are not big. Except observation 11, none of the points have Cook's D ≥ 1. So technically speaking, none of the points except OBS 11 are influential.

j) Since this point is influential, the regression line will be moved quite a bit towards this point compared to the fit when the point isn't included. The goal of simple linear regression is to explain the relationship

between two variables. It would be prudent to remove that data point and recalculate the regression and the ANOVA table. Keep in mind the calorimetric samples are easily contaminated.

k) (No question asked)

l) Now there is no data point having large Cook's D. There are still 3 data points having relatively large leverage, comparing to the criteria : $(hi > 2p/n) = 2 * 2/45 = 0.089$. As we say before, the differences are small, which show that there isn't crucial influence. Overall, none of these points have a lot of influence, so they shouldn't alter the model much.

m) The regression equation is estimated as: $\log(y_i) = 0.059 + 0.974 \log(x_i) + \epsilon_i$.

A confidence interval for intercept:

The formula we use is $\hat{\beta}_0 \pm t_{0.975,43} \hat{SE}(\beta_0)$. This is calculated as:

```
t_crit = qt(0.975, df = 43)
lower_int = 0.05901 - t_crit*0.05306
upper_int = 0.05905 + t_crit*0.05306

#Confidence interval is
c(lower_int, upper_int)
```

```
## [1] -0.04799569  0.16605569
```

A confidence interval for slope The formula we use is $\hat{\beta}_1 \pm t_{0.975,43} \hat{SE}(\beta_1)$. This is calculated as:

```
#t_crit is saved from the last R chunk
lower_slop = 0.97420 - t_crit*0.01092
upper_slop = 0.97420 + t_crit*0.01092

#Confidnece interval is
c(lower_slop, upper_slop)
```

```
## [1] 0.9521777 0.9962223
```

So we are 95% confident the true slope in this relationshipis between .952 and .996, that is, it is less than one. It appears the calorimetric test produces larger values than the enzymatic test.

n) When colorimetric value is 240, then $log(x_i) = log(240)$. Plugging this into the regression equation, our estimate for *log* enzymatic value is

```
lest = 0.059 + 0.974*log(240)
lest
```

```
## [1] 5.397142
```

Our 95% confidence interval for this value is $\log(\hat{y}_i) \pm t_{0.975,43} \hat{SE}(log(\hat{y}_i))$. We also have the formula for the standard error of a fitted value in linear regression of the form $y \sim x$ as: $\hat{SE}(\hat{y}) = \sqrt{MSE\,[1/n + (x - \bar{x})^2/S_{xx}]}$.

In our case the standard error of the fitted value for $\log(y_i)$ is calculated as

```
#get MSE
MSE = 0.03999^2
#get standard error of the estimate of log(y_i)
SE_logYhat = sqrt(MSE*(1/45 + (log(240)-4.83)/13.41))
SE_logYhat
```

```
## [1] 0.01063622
```

And then the 95% confience interval for $\log(\hat{y}_i)$ is given as

```
#not t_crit is still defined from above
lower_logpC = lest - t_crit*SE_logYhat
upper_logpC = lest + t_crit*SE_logYhat

#confidence interval is
c(lower_logpC, upper_logpC)
```

```
## [1] 5.375692 5.418592
```

This is on the log scale, so back on the original scale, we have the confidence interval as

```
c(exp(lower_logpC), exp(upper_logpC))
```

```
## [1] 216.0894 225.5614
```

So we are 95% confident that when the calorimetric test estimates a blood glucose level of 240, the enzymatic test will produce a level between 216.09 and 225.56. Again, the enzymetric test produces slightly lower glucose levels than the calorimetric test.

o) The formula for a 95% prediction interval is: $\log(\hat{y}_i) \pm t_{0.975,43}\hat{PD}(log(\hat{y}_i))^2$, where PD is the prediction error.

The prediction deviation has the formula in the y x linear regression as: $\hat{PD}(\hat{y}) = \sqrt{MSE\,[1 + 1/n + (x - \bar{x})^2/S_{xx}]}$. In our case it can be calculated as

```
#note MSE is defined above
pred_dev =  sqrt(MSE*(1 + 1/45 + (log(240)-4.83)/13.41))
pred_dev
```

```
## [1] 0.0413803
```

And the 95% prediction interval is calculated as

```
lower_logpP = lest - t_crit*pred_dev
upper_logpP = lest + t_crit*pred_dev

#prediction interval
c(lower_logpP, upper_logpP)
```

```
## [1] 5.313691 5.480594
```

This is on the log scale, so back on the original scale, we have the prediction interval as

```
c(exp(lower_logpP), exp(upper_logpP))
```

```
## [1] 203.0985 239.9891
```

So when the calorimetric test produces a blood glucose level of 240, the probability that the enzymetric test produces a blood glucose level between 203.10 and 239.99 is .95.