Statistics for Engineering, 4C3/6C3, 2012

Assignment 6

Kevin Dunn, dunnkg@mcmaster.ca **Due date: 12 March 2012, at 16:00**

Question 1 [3]

In a previous assignment you used an ordinary (*unpaired*) test of differences when a large sample of water was split in 22 portions, and the biochemical oxygen demand (BOD) was measured by the dilution method (11 times) and the manometric method (11 times). Here are the values again:

Dilution method	Manometric method
11	25
26	3
18	27
16	30
20	33
12	16
8	28
26	27
12	12
17	32
14	16

The confidence interval for the average difference of the two methods was calculated as: $-0.768 < \mu_{\mathrm{M}} - \mu_{\mathrm{D}} < 13.3$, showing that statistically there is no difference between the two methods. Though there is a practical difference, and the statistical lack of difference is due to an outlier.

The purpose of this question is for you to show that you can obtain the same result using a confidence interval from this least squares model:

$$y = b_0 + \gamma g_i$$

where $g_i = 0$ when the dilution method was used and $g_i = 1$ when the manometric method was used.

You may use any software to calculate the model and standard error for you, but you must show the confidence interval calculation for γ by hand.

Solution

The model and standard error were calculated with R: $y = b_0 + \gamma g_i = 16.4 + 6.3 g_i$ with standard error of 7.92 units, and 20 degrees of freedom.

```
d <- c(11, 26, 18, 16, 20, 12,  8, 26, 12, 17, 14)
m <- c(25,  3, 27, 30, 33, 16, 28, 27, 12, 32, 16)

BOD <- c(d, m)
g.d <- matrix(data=0, nrow=length(d), ncol=1)
g.m <- matrix(data=1, nrow=length(m), ncol=1)
g <- rbind(g.d, g.m)
g

model <- lm(BOD ~ g)
summary(model)
```

```
# Call:
# lm(formula = BOD ~ g)
#
# Residuals:
#     Min      1Q  Median      3Q     Max
# -19.636  -5.114   1.136   5.114  10.364
#
# Coefficients:
#             Estimate Std. Error t value Pr(>|t|)
# (Intercept)   16.364      2.387   6.856 1.16e-06 ***
# g              6.273      3.375   1.858   0.0779 .
# ---
# Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
#
# Residual standard error: 7.915 on 20 degrees of freedom
# Multiple R-squared: 0.1473,   Adjusted R-squared: 0.1046
# F-statistic: 3.454 on 1 and 20 DF,  p-value: 0.07788

# Confidence interval for dichotomous "gamma", given by "g"
confint(model)
#                  2.5 %  97.5 %
# (Intercept) 11.3852724 21.3420
# g           -0.7677425 13.3132

# Compare to the t-test:
t.test(m, d, var.equal=TRUE)
#
#   Two Sample t-test
#
# data:  m and d
# t = 1.8585, df = 20, p-value = 0.07788
# alternative hypothesis: true difference in means is not equal to 0
# 95 percent confidence interval:
#  -0.7677425 13.3131971
# sample estimates:
# mean of x mean of y
#  22.63636  16.36364
```

Using that:

- $c_t = \pm 2.085$ for the 95% confidence level (2.5% is each tail), with 20 degrees of freedom
- $\overline{\mathrm{x}} = \frac{0\times 11 + 1\times 11}{22} = 0.5$
- $\sum_j (x_j - \overline{\mathrm{x}})^2 = 11 \times (0 - 0.5)^2 + 11 \times (1 - 0.5)^2 = 2.75 + 2.75 = 5.5$
- $S_E^2(b_1) = \dfrac{S_E^2}{\sum_j (x_j - \overline{\mathrm{x}})^2} = \dfrac{7.915^2}{5.5} = 3.375^2$

The confidence interval for γ is:

$$\begin{array}{rcccl} 6.27 - c_t S_E(b_1) & \leq & \gamma & \leq & 6.27 + c_t S_E(b_1) \\ 6.27 - 2.085 \times 3.375 & \leq & \gamma & \leq & 6.27 + 2.09 \times 3.375 \\ -0.767 & \leq & \gamma & \leq & 13.3 \end{array}$$

This confidence interval agrees with the result from R's `confint(...)` function (see the code).

Question 2 [5]

In class we showed, given a x_{new} value and the linear model $y = b_0 + b_1 x$, that the prediction interval for $\hat{y}_{\text{new}}$ is:

$$\hat{y}_{\text{new}} \pm c_t \sqrt{V\{\hat{y}_{\text{new}}\}}$$

where c_t is the critical t-value, for example at the 95% confidence level.

Use the distillation column data set and with y as `VapourPressure` (units are kPa) and x as `TempC2` (units of °F) fit a linear model. Calculate and plot the prediction interval at the 95% confidence level, for vapour pressure at these temperatures: 410, 480, 530 °F. At each temperature compare the 95% prediction interval to the more conservative $2S_E$ prediction interval. Are these results expected?

Solution

Solutions with the help of Pedro Castillo

The model found was as $y = 196 - 0.33x$, with $y =$‘‘VapourPressure‘‘ [kPa] and $x =$‘‘TempC2‘‘ [°F]. For this model, $S_E = 2.989$ kPa, $n = 253$, $\sum_j (x_j - \bar{x})^2 = 86999.6$, and $\overline{x} = 480.82$.

The prediction interval is dependent on the value of $x_{\text{new, i}}$ used to make the prediction.

$$\mathcal{V}(\hat{y}_{\text{new,i}}) = S_E^2 \left(1 + \frac{1}{n} + \frac{(x_{\text{new}} - \bar{x})^2}{\sum_j (x_j - \bar{x})^2}\right)$$

We can calculate that the variance of $V\{\hat{y}_{\text{new}}\}$ is:

- $S_E^2 \left(1 + \frac{1}{n} + \frac{(x_i - \overline{\text{x}})^2}{\sum_j (x_j - \overline{\text{x}})^2}\right) = 2.989^2 \left(1 + \frac{1}{253} + \frac{(x_{\text{new}} - 480.82)^2}{86999.6}\right)$
  - For $x_{\text{new}} = 410$ °F: $V\{\hat{y}_{\text{new}}\} = 9.484\text{kPa}^2$
  - For $x_{\text{new}} = 480$ °F: $V\{\hat{y}_{\text{new}}\} = 8.969\text{kPa}^2$
  - For $x_{\text{new}} = 530$ °F: $V\{\hat{y}_{\text{new}}\} = 9.218\text{kPa}^2$

Since the degrees of freedom are $n - 2 = 251$, so the two-sided 95% confidence interval critical value is $c_t = 1.969$

- For $x_{\text{new}} = 410$ °F: $\hat{y}_{\text{new}} = 60.1 \pm 6.06$ kPa, or [54.0, 66.1] kPa.
- For $x_{\text{new}} = 480$ °F: $\hat{y}_{\text{new}} = 36.9 \pm 5.89$ kPa, or [31.02, 42.79] kPa.
- For $x_{\text{new}} = 530$ °F: $\hat{y}_{\text{new}} = 20.3 \pm 5.97$ kPa, or [14.3, 26.2] kPa.

Or, use the `predict(...)` function in R at the 95% level:

```
dist <- read.csv('http://datasets.connectmv.com/file/distillation-tower.csv')
attach(dist)
model <- lm(VapourPressure ~ TempC2)
summary(model)

new.data <- data.frame(TempC2=cbind(c(410, 480, 530)), level=0.95)
predict(model, newdata=new.data, interval="prediction")

#        fit      lwr      upr
# 1 60.11484 54.04879 66.18088
# 2 36.92152 31.02247 42.82057
# 3 20.35486 14.37472 26.33501
```

The prediction interval at the $\pm 2S_E$ level is $\hat{y}_{\text{new}} \pm 2S_E = \hat{y}_{\text{new}} \pm (2)(2.989) = \hat{y}_{\text{new}} \pm 5.978$ kPa. This prediction interval is constant over all values of x_{new}, which is obviously not correct. The further we go from the region where the model was build, the greater the level of uncertainty in the prediction. So, the $\pm 2S_E$ interval is just an estimation of the 95% confidence interval and it assumes the variance is the same at every x_{new}. Note though that the $\pm 2S_E$ interval is similar to the theoretical prediction interval calculated above.

As a comment, the values 410°F and 530°F are not within the range where the linear model was calculated; therefore the prediction intervals at these temperatures may have even greater uncertainty than given by this theoretical prediction interval.

Question 3 [12]

The percentage yield from bioreactor was investigated for a research project. Two variables were adjusted in the experiments: lactose purity and glucose purity.

Lactose	Glucose	Yield
90	85	51
94	85	79
90	90	72
94	90	94
88	87.5	50
96	87.5	99
92	82.5	49
92	92.5	89
92	87.5	73
92	87.5	74
92	87.5	64
92	87.5	65

1. Mean center the two x-variable vectors and combine them in matrix $\mathbf{X}$. Repeat the same for vector $\mathbf{y}$.
2. Now calculate $\mathbf{X}^T\mathbf{X}$ and $\mathbf{X}^T\mathbf{y}$. Give a clear interpretation for each of these.
3. Let $\mathbf{b} = [b_L, b_G]$, then calculate $\mathbf{b} = \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{y}$ *by hand*. Interpret each of these slope coefficients in $\mathbf{b}$ in the context of the linear model, $y = b_L x_L + b_G x_G$; where y is yield, x_L is the lactose concentration, and x_G is the glucose concentration.
4. Build the linear regression model, $y = b_0 + b_L x_L + b_G x_G$ (note the intercept) in R, or other statistical packages.
   (a) What is the intercept value?
   (b) Calculate this same intercept value by hand, from the values given in the answer to part 1 of this question.
   (c) Are the slope terms calculated by the software in agreement with yours (from part 3)?
5. Give confidence intervals for each of the slope coefficients at the 95% level.
6. Are the model's residuals normally distributed?

Solution

Solutions with the help of Pedro Castillo

1. The mean-centered vectors are:

```
      x1   x2
[1,]  -2 -2.5
[2,]   2 -2.5
[3,]  -2  2.5
```

```
[4,]   2  2.5
[5,]  -4  0.0
[6,]   4  0.0
[7,]   0 -5.0
[8,]   0  5.0
[9,]   0  0.0
[10,]  0  0.0
[11,]  0  0.0
[12,]  0  0.0
```

2. $\mathbf{X}^T\mathbf{X} = \begin{pmatrix} 48 & 0 \\ 0 & 75 \end{pmatrix}$ and $\mathbf{X}^T\mathbf{y} = \begin{pmatrix} 296 \\ 290 \end{pmatrix}$.

   The $\mathbf{X}^T\mathbf{X}$ matrix for centered vectors represents the degree of covariance of all combinations of vectors in the $\mathbf{X}$ matrix. The fact that there are zeros in the off-diagonals indicates that lactose purity and glucose purity are independent of each other (they have no covariance). The main diagonal values are proportional to the variance of each variable: glucose has a greater variance than lactose.

   The $\mathbf{X}^T\mathbf{y}$ vector is proportional to the covariance of each $\mathbf{x}_i$ with $\mathbf{y}$. The results indicate that y is correlated roughly the same amount with both both lactose and glucose, and that this is a positive correlation: yield and lactose purity increase and decrease together; the same for yield and glucose purity.

3. Since the $\mathbf{X}^T\mathbf{X}$ matrix is orthogonal, the inverse is easily found: $\left(\mathbf{X}^T\mathbf{X}\right)^{-1} = \begin{pmatrix} 0.0208 & 0 \\ 0 & 0.0133 \end{pmatrix}$.

   Then $\mathbf{b} = \begin{pmatrix} b_L \\ b_G \end{pmatrix} = \begin{pmatrix} 0.0208 & 0 \\ 0 & 0.0133 \end{pmatrix} \begin{pmatrix} 296 \\ 290 \end{pmatrix} = \begin{pmatrix} 6.167 \\ 3.867 \end{pmatrix}$.

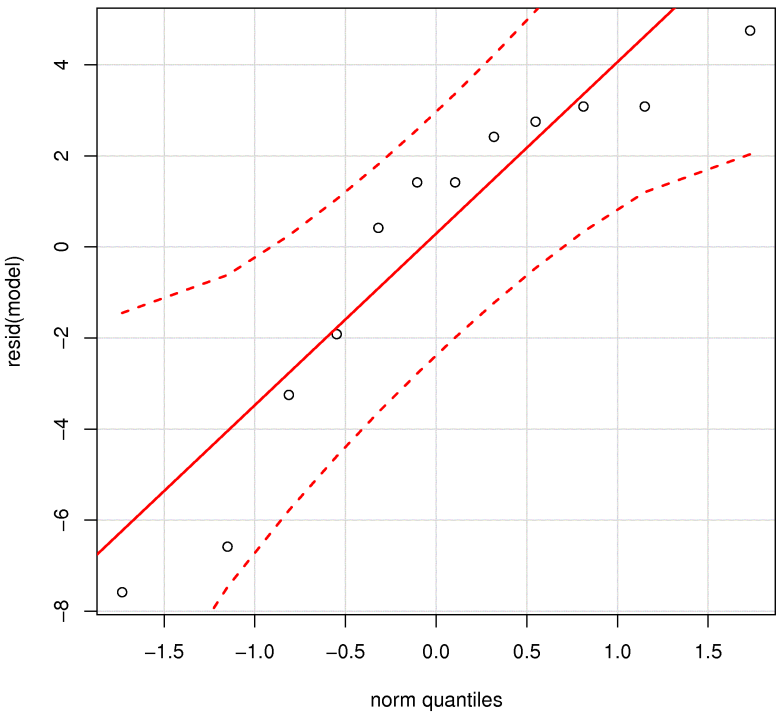
   These coefficients indicate that yield is expected to increase, on average, by 6.2 units for every percent increase in lactose purity; and that yield will increase, on average, by 3.9 units for every percent increase in glucose purity. We do not need to hold the other coefficient constant when making this interpretation, because showed that lactose and glucose are independent of each other.

4. The confidence intervals are found with the software (or by hand) as:

   - $4.7 < b_L < 7.6$
   - $2.7 < b_G < 5.0$

   None of the intervals spans zero, which means that both lactose and glucose purity have a significant effect on the yield percentage.

5. All the residuals are normally distributed, as given by the qq-plot:




```
lactose.raw <- c(90, 94, 90, 94, 88, 96, 92, 92, 92, 92, 92, 92)
glucose.raw <- c(85, 85, 90, 90, 87.5, 87.5, 82.5, 92.5, 87.5, 87.5, 87.5, 87.5)
yield.raw <- c(51, 79, 72, 94, 50, 99, 49, 89, 73, 74, 64, 65)

n <- length(lactose.raw)

x1 <- lactose.raw - mean(lactose.raw)
x2 <- glucose.raw - mean(glucose.raw)
y <- yield.raw - mean(yield.raw)
X <- cbind(x1, x2)


# Calculate:  b = inv(X'X) X'y
XTX <- t(X) %*% X    # compare this to cov(X)*(n-1)
XTY <- t(X) %*% y
XTX.inv <- solve(XTX)
b <- XTX.inv %*% XTY
b

model <- lm(yield.raw ~ lactose.raw + glucose.raw)
summary(model)
confint(model)

library(car)
bitmap('lactose-glucose-yield-residuals.png', type="png256", width=8,
       height=8, res=300, pointsize=14)
qqPlot(resid(model))
dev.off()
```

Question 4 [15]

Data on the course website were collected to predict the blending efficiency of an industrial mixer from 4 variables: particle size (x_P), mixer diameter (x_D), mixer rotational speed (x_R) and blending time (x_T): $y = b_0 + b_P x_P + b_D x_D + b_R x_R + b_T x_T$.

1. Calculate the variance-covariance matrix when the four x-variables are combined into an $\mathbf{X}$ matrix. What does this matrix tell you about this data set?
2. Fit the above linear model and determine which coefficients are significant at the 90% or higher level.
3. Interpret the slope coefficient for each variable. If these blending experiments were very expensive to conduct, based on the results from this part and the previous part 2, which variables could you disregard in the future?
4. Are the residuals normally distributed? If you identify any outliers, list the predicted value of y and the actual value of y for each outlier.
5. Omit any outliers and rebuild the model. Do any of the previous answers change? Are the residuals more normally distributed now?

Use the software tutorial on the course website to help you with this question.

Solution

Solutions with the help of Yasser Ghobara

1. The variance-covariance matrix of the raw data is:

```
> cov(data)
                   ParticleSize MixerDiameter MixerRotation BlendingTime BlendingEfficiency
ParticleSize           4.235294      0.000000      0.000000      0.00000         -15.988235
MixerDiameter          0.000000      1.882353      0.000000      0.00000           1.588235
MixerRotation          0.000000      0.000000    294.117647    -44.11765          -5.882353
BlendingTime           0.000000      0.000000    -44.117647    102.94118          26.323529
BlendingEfficiency   -15.988235      1.588235     -5.882353     26.32353          74.988235
```

   The diagonal of the matrix shows the variance of each variable and the off-diagonals represent the covariance between each variable and the others. These values represent the strength of the relationship between the variables. The fact that the off-diagonals are mostly small, relative to the diagonals, indicate that the variables are relative independent of each other.

2. The model summary from R is:

```
> summary(model)

Call:
lm(formula = data$BlendingEfficiency ~ data$ParticleSize + data$MixerDiameter +
    data$MixerRotation + data$BlendingTime)

Residuals:
    Min      1Q  Median      3Q     Max
-2.9355 -1.1542 -0.4554  1.2352  7.6645

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)        69.64838   10.98428   6.341 2.57e-05 ***
data$ParticleSize  -3.77500    0.34241 -11.025 5.74e-08 ***
data$MixerDiameter  0.84375    0.51361   1.643  0.12439
data$MixerRotation  0.01962    0.04248   0.462  0.65182
data$BlendingTime   0.26412    0.07180   3.679  0.00278 **
---
```

```
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.905 on 13 degrees of freedom
Multiple R-squared: 0.9139,     Adjusted R-squared: 0.8874
F-statistic:  34.5 on 4 and 13 DF,  p-value: 8.287e-07
```

The significant confidence intervals, at the 90% level are found using the `confint(...)` function:

```
> confint(model, level=0.9)
                             5 %        95 %
(Intercept)         50.19594077 89.10081495
data$ParticleSize   -4.38138523 -3.16861477
data$MixerDiameter  -0.06582784  1.75332784
data$MixerRotation  -0.05560599  0.09484263
data$BlendingTime    0.13696984  0.39127443
```

Only the coefficients for `ParticleSize` and `BlendingTime` do not span zero and are therefore significant at the 90% level (5% in each tail). The `MixerDiameter` coefficient could be considered as significant, since the interval spans zero but not very symmetrically.

3. The interpretations are, for example, `ParticleSize` = $b_P = -3.78$, shows an expected decrease in blending efficiency of 3.78 units for a 1 unit increase in particle size, keeping all other variables constant.

   The other slope coefficients can be interpreted in a similar way.

   Since the effect of mixer diameter and mixer rotation seem negligible, they could be omitted. I would omit mixer rotation first, then mixer diameter, and then, reluctantly, blending time.

4. The residuals do appear normally distributed in a q-q plot, except for observation 8. The predicted value is 72.1, while the actual value is 79.8. Compared with the standard error of 2.9, this is a large difference, indicating this residual is strong. I would go back to the experimental records to see what occurred.

5. Omitting point 8 and rebuilding the model:

```
> model.update <- lm(model, subset=-c(8))
> summary(model.update)
Call:
lm(formula = model, subset = -c(8))

Residuals:
    Min      1Q  Median      3Q     Max
-2.3095 -0.9788 -0.2845  0.7655  2.8213

Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)         73.94050    6.30110  11.735 6.20e-08 ***
data$ParticleSize   -3.77500    0.19480 -19.379 2.02e-10 ***
data$MixerDiameter   0.84375    0.29220   2.888   0.0136 *
data$MixerRotation  -0.02177    0.02539  -0.857   0.4081
data$BlendingTime    0.32288    0.04232   7.629 6.09e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.653 on 12 degrees of freedom
Multiple R-squared: 0.9738,     Adjusted R-squared: 0.9651
F-statistic: 111.6 on 4 and 12 DF,  p-value: 2.206e-09
```

   - The standard error has decreased quite a bit, from 2.9 to 1.65.
   - The `MixerDiameter` coefficient is now significant at the 95% level.
   - Most of the coefficients remained the same, except for `BlendingTime`, which increased from 0.26 to 0.32.

- Finally, the q-q plot shows a more normal distribution.

```
data <- read.csv('http://datasets.connectmv.com/file/blender-efficiency.csv')
summary(data)
cov(data)
model<-lm(data$BlendingEfficiency ~ data$ParticleSize + data$MixerDiameter +
                                    data$MixerRotation + data$BlendingTime)
summary(model)

confint(model, level=0.90)

library(car)
qqPlot(model, id.n=1)

# Omit point 8:
model.update <- lm(model, subset=-c(8))
summary(model.update)
qqPlot(model.update)
confint(model.update, level=0.90)
```
