Math 760

Chapter 7 HW

Gabrielle Salamanca

May 9, 2024

2. Given the data

```
##    [,1] [,2] [,3] [,4] [,5] [,6]
## z1   10    5    7   19   11   18
## z2    2    3    3    6    7    9
## y    15    9    3   25    7   13
```

for the regression model $Y_j = \beta_1 z_{j1} + \beta_2 z_{j2} + \epsilon_i$ (j = 1,2,…,6) to the standardized form of the variables y, z_1, z_2. From this fit, deduce the corresponding fitted regression equation for the original (not standardized) variables.

```
## The standardized form of the variables are:
##            z1         z2          y
## V1  0.2761327 -1.9329291  1.6567964
## V2 -0.2697022 -1.0788089  1.3485112
## V3  1.2408065 -0.6204032 -0.6204032
## V4  0.5294141 -2.4201787  1.8907646
## V5  1.2408065 -0.6204032 -0.6204032
## V6  1.5539593 -1.4429622 -0.1109971
##
## The mean of z1 is 0.7619028
## The mean of z2 is -1.352614
```

The fitted version of the equation $Y_j = \beta_1 z_{j1} + \beta_2 z_{j2} + \epsilon_i$ is $\hat{y} = \hat{\beta}_1 z_1 + \hat{\beta}_2 z_2$. And with the results above, the fitted equation is: $\hat{y} = 0.76 z_1 - 1.35 z_2$

Then, we find the values for the regular data to plug into the fitted equation. We will need the means and the square root of the diagonal of their covariance matrix.

```
## The mean for z1 is: 11.66667
## The mean for z2 is: 5
## The mean for y is: 12
## The square root covariance matrix of the predictors is:
##          z1       z2       y
## z1 5.715476 3.435113 5.692100
## z2 3.435113 2.756810 2.097618
## y  5.692100 2.097618 7.668116
```

Thus:

$$\frac{\hat{y} - 12}{7.668116} = 0.7619028\left(\frac{z_1 - 11.66667}{5.715476}\right) - 1.352614\left(\frac{z_2 - 5}{2.756810}\right)$$

Let's start with z_1

$$0.7619028\left(\frac{z_1 - 11.66667}{5.715476}\right) = 0.1333052z_1 - 1.555228$$

Then z_2

$$-1.352614\left(\frac{z_2 - 5}{2.756810}\right) = -0.4906448z_2 + 2.453224$$

Putting that all back in,

$$\Rightarrow \frac{\hat{y} - 12}{7.668116} = 0.1333052z_1 - 1.555228 - 0.4906448z_2 + 2.453224$$
$$\Rightarrow \frac{\hat{y} - 12}{7.668116} = 0.1333052z_1 - 0.4906448z_2 + 0.897996$$
$$\Rightarrow \hat{y} - 12 = 1.0222z_1 - 3.762321z_2 + 6.885937$$
$$\Rightarrow \hat{y} = 1.0222z_1 - 3.762321z_2 + 18.88594$$

And simplifying that, we get: $\hat{y} = 1.02z_1 - 3.76z_2 + 18.89$

8. Recall that that the hat matrix is defined by $H = Z(Z'Z)^{-1}Z'$ with diagonal elements $h_j j$.

(a) Show that H is an idempotent matrix [See Result 7.1 and (7-6)]

Result 7.1

Let **Z** have full rank $(r + 1 \leq n)$. The least squares estimate of β in (7-3) is given by: $\hat{\beta} = (Z'Z)^{-1}Z'y$. Let $\hat{y} = Z\hat{\beta} = Hy$ denote the fitted values of y, where $H = (Z'Z)^{-1}Z'$ is called "hat" matrix. Then the residuals $\hat{\epsilon} = y - \hat{y} = [I - (Z'Z)^{-1}Z']y = (I - H)y$ satisfy $Z'\hat{\epsilon} = 0$ and $\hat{y}'\hat{\epsilon} = 0$. Also, the residual sum of squares = $\sum_{j=1}^{n}\left(y_j - \hat{\beta}_0 - \widehat{\beta_1}z_{j1} - \ldots - \widehat{\beta_r}z_{jr}\right) = \hat{\epsilon}'\hat{\epsilon} = y'[I - Z(Z'Z)^{-1}Z']y = y'y - y'Z\hat{\beta}$

(7-6)

$[I - Z(Z'Z)^{-1}Z']'[I - Z(Z'Z)^{-1}Z'] = I - 2[Z(Z'Z)^{-1}Z'] + Z(Z'Z)^{-1}Z'[Z(Z'Z)^{-1}Z']$
$[I - Z(Z'Z)^{-1}Z']'[I - Z(Z'Z)^{-1}Z'] = [I - Z(Z'Z)^{-1}Z']$ (idempotent)

Thus, $H^2 = Z(Z'Z)^{-1}Z'[Z(Z'Z)^{-1}Z'] = Z(Z'Z)^{-1}Z' = H$

(b) Show that $0 < h_{jj} < 1$ (j =1,2,…,n), and that $\sum_{j=1}^{n} h_{jj} = r + 1$, where r is the number of independent variables in the regression model. (In fact, $\frac{1}{n} \leq h_{jj} < 1$)

Because [I-H] is an idempotent matrix, it's a positive semidefinite. Then, let a be a $nx1$ unit vector with j^{th} element 1. Then, $0 \leq a'(I - H)a = \left(1 - h_{jj}\right)$; that is, $h_{jj} \leq 1$.

$(Z'Z)^{-1}$ is a positive definite matrix. Thus $h_{jj} = b_j'(Z'Z)^{-1}b_j$, where b_j is a the j^{th} row of Z.

$$\sum_{i=1}^{r+1} h_{jj} = tr[Z(Z'Z)^{-1}Z'] = tr[(Z'Z)^{-1}(Z'Z)] = tr(I_{r+1}) = r+1$$

(c) Verify, for the simple linear regression model with one independent variable z, that the leverage h_{jj} is given by $h_{jj} = \frac{1}{n} + \frac{(z_j - \bar{z})^2}{\sum_{j=1}^{n}(z_j - \bar{z})^2}$

Using

$$(Z'Z)^{-1} = \frac{1}{n\sum_{i=1}^{n}(z_j - \bar{z})^2}\begin{bmatrix} \sum_{i=1}^{n} z_i^2 & -\sum_{i=1}^{n} z_i \\ -\sum_{i=1}^{n} z_i & n \end{bmatrix}$$

We get

$$h_{jj} = [1 \quad z_j](Z'Z)^{-1}\begin{bmatrix}1\\ z_j\end{bmatrix} = \frac{1}{n\sum_{i=1}^{n}(z_j - \bar{z})^2}\left(\sum_{i=1}^{n} z_i^2 - 2z_j\sum_{i=1}^{n} z_i + nz_j^2\right) = \frac{1}{n} + \frac{(z_j - \bar{z})^2}{\sum_{i=1}^{n}(z_j - \bar{z})^2}$$

9. Consider the following data on one predictor variable z_1 and two responses Y_1 and Y_2

```
##    [,1] [,2] [,3] [,4] [,5]
## z1   -2   -1    0    1    2
## y1    5    3    4    2    1
## y2   -3   -1   -1    2    3
```

Determine the least squares estimates of the parameters in the bivariate straight-line regression model

$$Y_{j1} = \beta_{01} + \beta_{11}z_{j1} + \epsilon_{j1}$$
$$Y_{j2} = \beta_{02} + \beta_{12}z_{j2} + \epsilon_{j2}$$

where j = 1,2,3,4,5. Also, calculate the matrices of fitted values $\hat{Y}$ and residuals $\hat{\epsilon}$ with $Y = [y_1|y_2]$. Verify the sum of squares and cross-products decomposition $Y'Y = \hat{Y}'\hat{Y} + \hat{\epsilon}'\hat{\epsilon}$

To find the least squares estimate of the $\beta's$, we use $\hat{\beta} = (Z'Z)^{-1}Z'y$. Let's find our Z's.

```
## The Z matrix is
##      [,1] [,2]
## [1,]    1   -2
## [2,]    1   -1
## [3,]    1    0
## [4,]    1    1
## [5,]    1    2
```

```
##
## The inverse of Z'Z is
##      [,1] [,2]
## [1,]  0.2  0.0
## [2,]  0.0  0.1
```

Now, let's find our $\beta's$.

```
## Our least squares estimates matrix of our parameters is
##      [,1]         [,2]
## [1,]  3.0 1.110223e-16
## [2,] -0.9 1.500000e+00
```

We'll now calculate the matrices of fitted values $\hat{Y}$, which is calculated by multiplying Z and $\hat{\beta}$.

```
##      [,1]          [,2]
## [1,]  4.8 -3.000000e+00
## [2,]  3.9 -1.500000e+00
## [3,]  3.0  1.110223e-16
## [4,]  2.1  1.500000e+00
## [5,]  1.2  3.000000e+00
```

Finally, we'll calculate the residuals $\hat{\epsilon}$ with $Y = [y_1|y_2]$. It's calculated by subtracting Y and $\hat{Y}$.

```
## The residual matrix is
##      [,1]          [,2]
## [1,]  0.2  4.440892e-16
## [2,] -0.9  5.000000e-01
## [3,]  1.0 -1.000000e+00
## [4,] -0.1  5.000000e-01
## [5,] -0.2 -4.440892e-16
```

After all that, we must verify the sum of squares and cross-products decomposition with this equation: $Y'Y = \widehat{Y}'\hat{Y} + \widehat{\epsilon}'\hat{\epsilon}$

```
## The Y'Y matrix is
##      [,1] [,2]
## [1,]   55  -15
## [2,]  -15   24
##
## The right side of the equation is
##      [,1] [,2]
## [1,]   55  -15
## [2,]  -15   24
```

12. Given the mean vector and covariance of Y, Z_1, and Z_2. Determine each of the following.

```
## mu Matrix
##      [,1]
## [1,]    4
## [2,]    3
## [3,]   -2
##
##  Sigma matrix
##      [,1] [,2] [,3]
## [1,]    9    3    1
## [2,]    3    2    1
## [3,]    1    1    1
```

(a) The best linear predictor $\beta_0 + \beta_1 Z_1 + \beta_2 Z_2$ of Y

To find β and β_0, we use these equations: $\beta = \Sigma_{ZZ}^{-1}\sigma_{ZY}$, $\beta_0 = \mu_Y - \beta'\mu_Z$

```
## Our beta matrix is
##      [,1]
## [1,]    2
## [2,]   -1
##
## Beta-0 is -4
```

Our model is: $Y = \beta_0 + \beta_1 Z_1 + \beta_2 Z_2 = -4 + 2z_1 - z_2$

(b) The mean square error of the best linear predictor

This is calculated with this equation: $\sigma_{yy} - \sigma'_{zy}\Sigma_{ZZ}^{-1}\sigma_{ZY}$

```
##      [,1]
## [1,]    4
```

(c) The population multiple correlation coefficient

This is calculated as such: $\rho_{Y(x)} = \sqrt{\frac{\sigma'_{zy}\Sigma_{ZZ}^{-1}\sigma_{ZY}}{\sigma_{YY}}}$

```
##          [,1]
## [1,] 0.745356
```

(d) The partial correlation coefficient $\rho_{YZ_1*Z_2}$

The partial correlation coefficient formula is provided by (7-56):

$$\rho_{Y_1Y_2*Z} = \frac{\sigma_{Y_1Y_2*Z}}{\sqrt{\sigma_{Y_1Y_1*Z}}\sqrt{\sigma_{Y_2Y_2*Z}}}$$

We'll first need to partition our Σ matrix and determine to covariance of $\begin{bmatrix} Y \\ Z_1 \end{bmatrix}$.

$$\begin{pmatrix} 9 & 3 & | & 1 \\ 3 & 2 & | & 1 \\ - & - & + & - \\ 1 & 1 & | & 1 \end{pmatrix}$$

```
## The covariance matrix is
##      [,1] [,2]
## [1,]    8    2
## [2,]    2    1
```

Now, $\rho_{Y_1 Y_2 * Z}$ is

```
## [1] 0.7071068
```

17. Consider the Forbes data in Exercise 1.4

```
##       order  sales profits  assets
##  [1,]     1 108.28   17.05 1484.10
##  [2,]     2 152.36   16.59  750.33
##  [3,]     3  95.04   10.91  766.42
##  [4,]     4  65.45   14.14 1110.46
##  [5,]     5  62.97    9.52 1031.29
##  [6,]     6 263.99   25.33  195.26
##  [7,]     7 265.19   18.54  193.83
##  [8,]     8 285.06   15.73  191.11
##  [9,]     9  92.01    8.10 1175.16
## [10,]    10 165.68   11.13  211.15
```

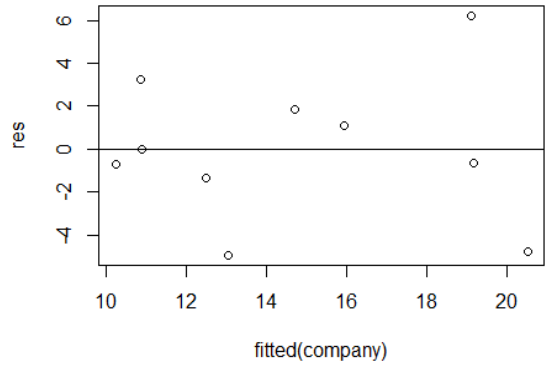
(a) Fit a linear regression model to these data using profits as the dependent variable and sales and assets as the indepedent variables.

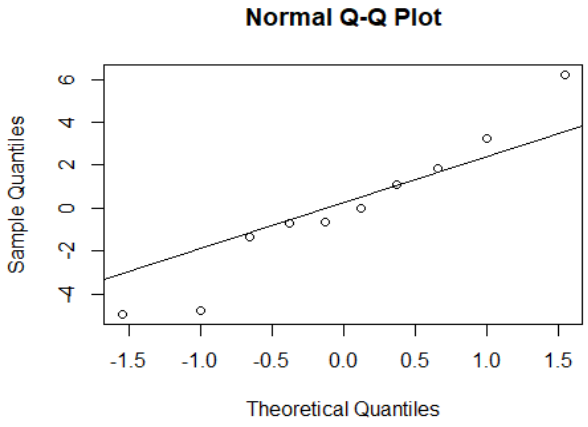
```
##
## Call:
## lm(formula = profits ~ sales + assets)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -4.954 -1.215 -0.316  1.686  6.224
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 0.013325   7.641453   0.002   0.9987
## sales       0.068058   0.027851   2.444   0.0445 *
## assets      0.005768   0.004946   1.166   0.2817
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

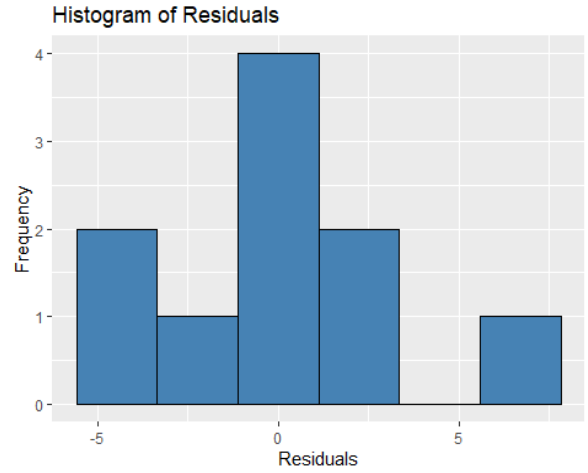
```
##
## Residual standard error: 3.863 on 7 degrees of freedom
## Multiple R-squared:  0.5569, Adjusted R-squared:  0.4303
## F-statistic: 4.399 on 2 and 7 DF,  p-value: 0.05792
```

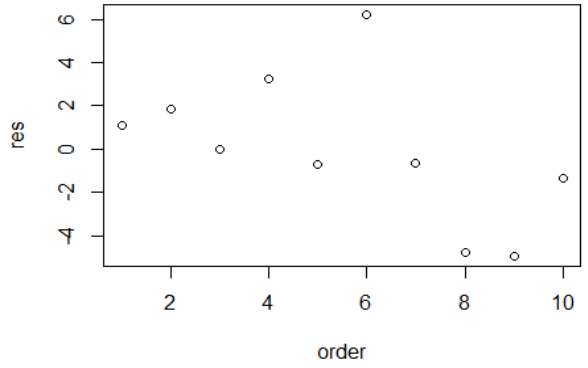
The linear regression model is: $Y = 0.013325 + 0.068058x_1 + 0.005768x_3$

(b) Analyze the residuals to check the adequacy of the model. Compute the leverages associated with the data points. Does one (or more) of these companies stand out as an outlier in the set of independent variable data points?









Given the small sample size, I would say the data is independent, but I would caution about it following a normal distribution due to 3 points being quite far from the line. However, the histogram does give the vague bell-shape curve, even if there's an empty space.

```
## The average leverage is 0.9
##    hatvalues(company)
## 1           0.6256557
## 2           0.1010997
## 3           0.2432703
```

```
## 4           0.2222081
## 5           0.2512938
## 6           0.2745888
## 7           0.2785075
## 8           0.3642405
## 9           0.2029422
## 10          0.4361935
```

We see that most of the leverage values are less than 0.9, which means there are no unusual observations.

(c) Generate a 95% prediction interval for profits corresponding to sales of 100 (billions of dollars) and assets of 500 (billion off dollars).

Here's what we know:
From (a), $Y = 0.013325 + 0.068058x_1 + 0.005768x_3$
sales, $x_1 = 100$
assets, $x_3 = 500$

The 95% prediction interval for profits is:

```
##        fit       lwr      upr
## 1 9.703207 -1.545611 20.95203
```

(d) Carry out a likelihood ratio test of $H_0: \beta_2 = 0$ with significance level of $\alpha = 0.05$. Should the original model be modified. Discuss.

$$H_0: \beta_2 = 0 \text{ vs } H_1: \beta_2 \neq 0$$

```
## Likelihood ratio test
##
## Model 1: profits ~ sales
## Model 2: profits ~ sales + assets
##   #Df  LogLik Df  Chisq Pr(>Chisq)
## 1   3 -26.808
## 2   4 -25.920  1 1.7755     0.1827
```

The χ^2 p-value is 0.1827, which is greater than $\alpha = 0.05$, so we cannot reject H_0. If we wanted to fit a model to fit our data better, we should consider only having sales as a predictor for profits.

22. Using the data on bone mineral content in Table 1.8

(a) Perform a regression analysis by fitting the response for the dominant radius bone to the measurements on the last four bones

When we perform a regression analysis of $DomRad = DomHum + Hum + DomUlna + Ulna$, our results are:

```
##
## Call:
## lm(formula = domRad ~ domHum + hum + domUln + ulna)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.131062 -0.028098  0.000606  0.035727  0.134517
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   0.1027     0.1064   0.966   0.3457
## domHum        0.2756     0.1147   2.402   0.0261 *
## hum          -0.1652     0.1381  -1.196   0.2458
## domUln        0.3566     0.1985   1.796   0.0876 .
## ulna          0.4068     0.2174   1.871   0.0760 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.06635 on 20 degrees of freedom
## Multiple R-squared:  0.7178, Adjusted R-squared:  0.6614
## F-statistic: 12.72 on 4 and 20 DF,  p-value: 2.617e-05
```

(i) Suggest and fit appropriate linear regression models.

Based on the results, it would be best to remove the humerus variable, and possibly also both ulna variables. We'll run one with both ulnas and one without them.

```
##
## Call:
## lm(formula = domRad ~ domHum + domUln + ulna)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.152610 -0.027960 -0.002006  0.027820  0.144917
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.11013    0.10728   1.027   0.3163
## domHum       0.15685    0.05802   2.704   0.0133 *
## domUln       0.36044    0.20054   1.797   0.0867 .
## ulna         0.28621    0.19453   1.471   0.1560
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.06703 on 21 degrees of freedom
## Multiple R-squared:  0.6977, Adjusted R-squared:  0.6545
## F-statistic: 16.15 on 3 and 21 DF,  p-value: 1.13e-05
##
## Call:
## lm(formula = domRad ~ domHum)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.14320 -0.05436  0.02160  0.03806  0.16288
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.34520    0.10992   3.141 0.004584 **
## domHum       0.27813    0.06059   4.590 0.000129 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.08414 on 23 degrees of freedom
## Multiple R-squared:  0.4781, Adjusted R-squared:  0.4554
## F-statistic: 21.07 on 1 and 23 DF,  p-value: 0.0001292
```

It looks like with only the dominant humerus as a predictor, the model is the best fit. Though, I will bring to attention to the R^2 values for the no-ulna model. They are notably smaller than the one where we included the ulnas.

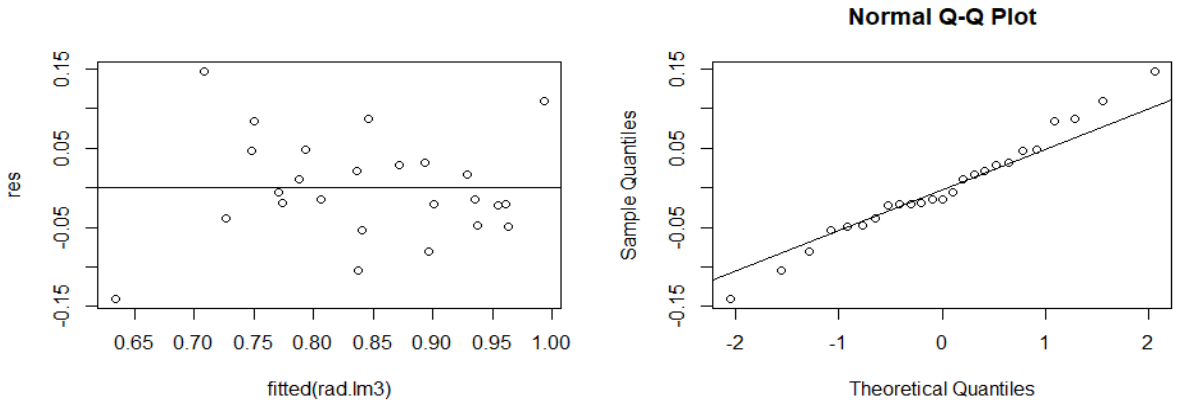
So, let's also try with one ulna each just to see if our R^2 values are better.

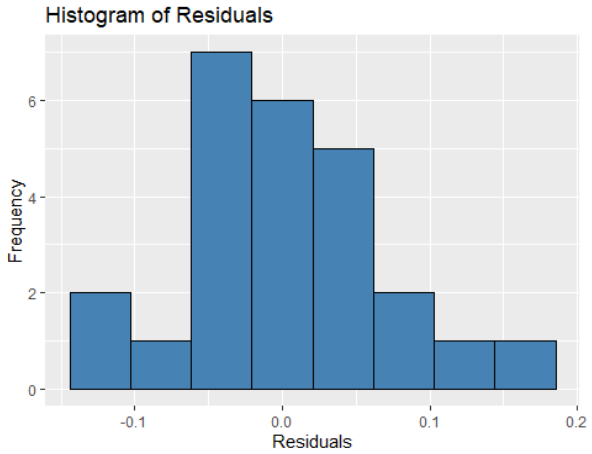
```
##
## Call:
## lm(formula = domRad ~ domHum + domUln)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.14058 -0.03802 -0.01424  0.03132  0.14739
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   0.1637     0.1035   1.581  0.12808
## domHum        0.1625     0.0594   2.735  0.01208 *
## domUln        0.5519     0.1566   3.525  0.00191 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.06878 on 22 degrees of freedom
## Multiple R-squared:  0.6665, Adjusted R-squared:  0.6362
## F-statistic: 21.98 on 2 and 22 DF,  p-value: 5.676e-06
```

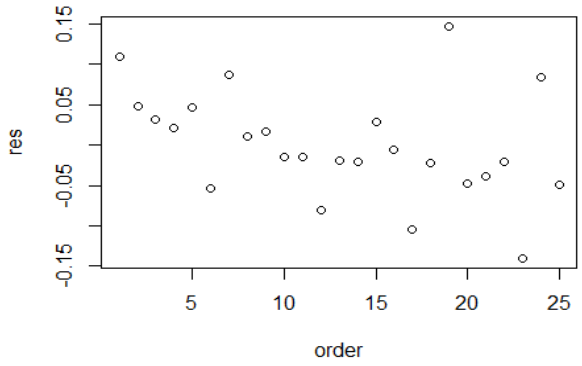
```
##
## Call:
## lm(formula = domRad ~ domHum + ulna)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.162158 -0.029020 -0.005463  0.052344  0.134185
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.13624    0.11155   1.221  0.23490
## domHum       0.19610    0.05641   3.476  0.00214 **
## ulna         0.51311    0.15532   3.303  0.00324 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.07034 on 22 degrees of freedom
## Multiple R-squared:  0.6512, Adjusted R-squared:  0.6194
## F-statistic: 20.53 on 2 and 22 DF,  p-value: 9.309e-06
```

Both of these models are better than the no-ulna model based on the R^2 values. If one had to choose, the model with the dominant bones would be the best: $DomRad = 0.1637 + 0.1625\ DomHum\ + 0.5519\ DomUlna$

(ii) Analyze the residuals







From these plots, it generally follows a normal distribution, but I would want to be careful of certain points that stray a bit from the line in the fitted values and order plots.

(b) Perform a multivariate multiple regression analysis by fitting the responses from both radius bones.

```
## Response domRad :
##
## Call:
## lm(formula = domRad ~ domHum + hum + domUln + ulna, data = bone)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.131062 -0.028098  0.000606  0.035727  0.134517
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   0.1027     0.1064   0.966   0.3457
## domHum        0.2756     0.1147   2.402   0.0261 *
## hum          -0.1652     0.1381  -1.196   0.2458
## domUln        0.3566     0.1985   1.796   0.0876 .
## ulna          0.4068     0.2174   1.871   0.0760 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.06635 on 20 degrees of freedom
## Multiple R-squared:  0.7178, Adjusted R-squared:  0.6614
## F-statistic: 12.72 on 4 and 20 DF,  p-value: 2.617e-05
##
##
## Response rad :
##
## Call:
## lm(formula = rad ~ domHum + hum + domUln + ulna, data = bone)
##
```

```
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.110436 -0.037494  0.008991  0.040042  0.089457
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.11423    0.08971   1.273   0.2175
## domHum      -0.01103    0.09676  -0.114   0.9104
## hum          0.15204    0.11649   1.305   0.2067
## domUln       0.19764    0.16743   1.180   0.2517
## ulna         0.46247    0.18333   2.523   0.0202 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.05595 on 20 degrees of freedom
## Multiple R-squared:  0.7715, Adjusted R-squared:  0.7258
## F-statistic: 16.88 on 4 and 20 DF,  p-value: 3.378e-06
```

When running the multivariate multiple regression analysis, we find that the one with radius is the better model, especially considering the R^2 values.

(c) Calculate the AIC for the model you chose in (b) and for the full model.

```
## The AIC for the model domRad = domHum + domUln is -58.0937
## The AIC for the model rad = domHum + hum + domUln + ulna is -66.79644
```

25. Amitriptyline is prescribed by some physicians as an antidepressant. However, there are also conjectured side effects that seem to be related to the use of the drug: irregular heartbeat, abnormal blood pressures, and irregular waves on the electrocardiogram, among other things. Data gathered on 17 patients who were admitted to the hospital after an amitriptyline overdose are given in Table 7.6. The two response variables are:

Y_1 = Total TCAD plasma level (TOT)

Y_2 = Amount of amitriptyline present in TCAD plasma level (AMI)

The five predictor variables are:

Z_1= Gender, where fem = 1 and male = 0 (GEN)

Z_2= Amount of antidepressants taken at time of overdose (AMT)

Z_3= PR wave measurement (PR)

Z_4= Diastolic blood pressure (DIAP)

Z_5= QRS wave measurement (QRS)

(a) Perform a regression analysis using only the first response Y_1

```
##
## Call:
## lm(formula = y1 ~ z1 + z2 + z3 + z4 + z5, data = drug)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -399.2 -180.1    4.5  164.1  366.8
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -2.879e+03  8.933e+02  -3.224 0.008108 **
## z1           6.757e+02  1.621e+02   4.169 0.001565 **
## z2           2.848e-01  6.091e-02   4.677 0.000675 ***
## z3           1.027e+01  4.255e+00   2.414 0.034358 *
## z4           7.251e+00  3.225e+00   2.248 0.046026 *
## z5           7.598e+00  3.849e+00   1.974 0.074006 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 281.2 on 11 degrees of freedom
## Multiple R-squared:  0.8871, Adjusted R-squared:  0.8358
## F-statistic: 17.29 on 5 and 11 DF,  p-value: 6.983e-05
```

When running a regression analysis on the full model with Y_1 as the response, we find that Z_5 isn't as significant at $\alpha = 0.05$, so we could drop that variable.

(i) Suggest and fit appropriate linear regression models.

Let's see an anova to see which variables to keep.

```
## Analysis of Variance Table
##
## Response: y1
##           Df  Sum Sq Mean Sq F value   Pr(>F)
## z1         1  288658  288658  3.6497  0.08248 .
## z2         1 5616926 5616926 71.0179 3.97e-06 ***
## z3         1  341134  341134  4.3131  0.06204 .
## z4         1  280973  280973  3.5525  0.08613 .
## z5         1  308241  308241  3.8973  0.07401 .
## Residuals 11  870008   79092
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

After running anova, we find that the model with only Z_2 is our best fit. Let's see if that's true.

```
##
## Call:
## lm(formula = y1 ~ z2, data = drug)
##
```

```
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1061.05  -139.23    51.19   203.25   627.51
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 462.8928   160.9090   2.877   0.0115 *
## z2            0.3065     0.0578   5.303 8.86e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 422.8 on 15 degrees of freedom
## Multiple R-squared:  0.6521, Adjusted R-squared:  0.6289
## F-statistic: 28.12 on 1 and 15 DF,  p-value: 8.861e-05
```

All are significant, but I do want to try adding Z_1 to see if our R^2 values are better.

```
##
## Call:
## lm(formula = y1 ~ z1 + z2, data = drug)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -756.05 -190.68  -59.83  203.32  560.84
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  56.72005  206.70337   0.274   0.7878
## z1          507.07308  193.79082   2.617   0.0203 *
## z2            0.32896    0.04978   6.609 1.17e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 358.6 on 14 degrees of freedom
## Multiple R-squared:  0.7664, Adjusted R-squared:  0.733
## F-statistic: 22.96 on 2 and 14 DF,  p-value: 3.8e-05
```

The R^2 values are better, even if the intercept wasn't significant. Let's try with the other variables just to make sure.

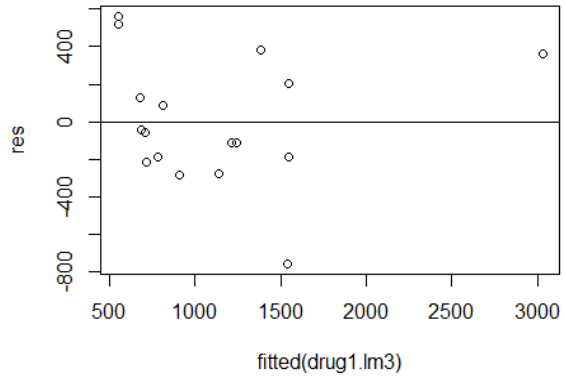
```
##
## Call:
## lm(formula = y1 ~ z1 + z2 + z3, data = drug)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -597.48 -189.26  -61.15  204.74  552.22
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1.328e+03  8.174e+02  -1.625  0.12824
## z1           5.582e+02  1.834e+02   3.044  0.00942 **
```

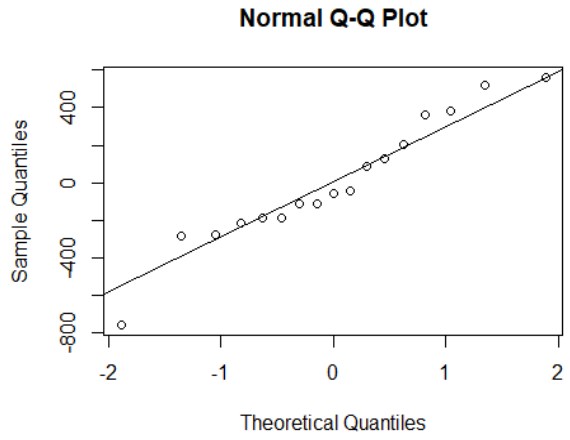
```
## z2           2.583e-01  6.169e-02   4.187  0.00106 **
## z3           8.578e+00  4.921e+00   1.743  0.10487
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 335 on 13 degrees of freedom
## Multiple R-squared:  0.8106, Adjusted R-squared:  0.7669
## F-statistic: 18.55 on 3 and 13 DF,  p-value: 5.575e-05
##
## Call:
## lm(formula = y1 ~ z1 + z2 + z3 + z4, data = drug)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -360.64 -192.74  -44.95  239.31  435.62
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -2.154e+03  9.071e+02  -2.374 0.035121 *
## z1           6.505e+02  1.800e+02   3.614 0.003555 **
## z2           3.126e-01  6.603e-02   4.735 0.000485 ***
## z3           1.049e+01  4.739e+00   2.214 0.046955 *
## z4           5.951e+00  3.518e+00   1.692 0.116499
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 313.3 on 12 degrees of freedom
## Multiple R-squared:  0.8471, Adjusted R-squared:  0.7961
## F-statistic: 16.62 on 4 and 12 DF,  p-value: 7.772e-05
##
## Call:
## lm(formula = y1 ~ z1, data = drug)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -608.6 -418.6 -137.6  192.4 2184.4
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)    918.6      314.5   2.921   0.0105 *
## z1             286.0      374.3   0.764   0.4567
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 703.2 on 15 degrees of freedom
## Multiple R-squared:  0.03746,    Adjusted R-squared:  -0.02671
## F-statistic: 0.5838 on 1 and 15 DF,  p-value: 0.4567
```

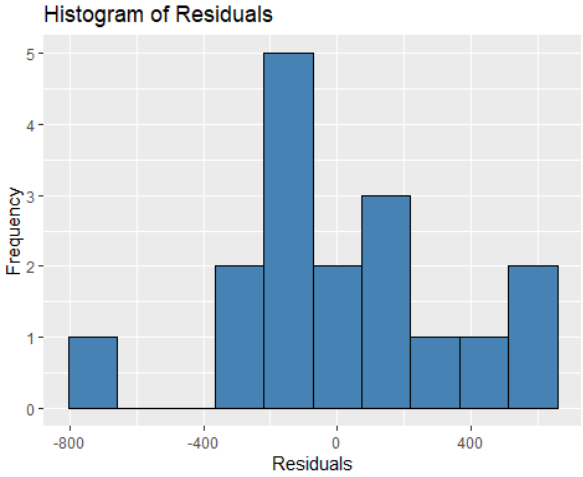
Comparing all these models, the one with 2 predictors is the best.

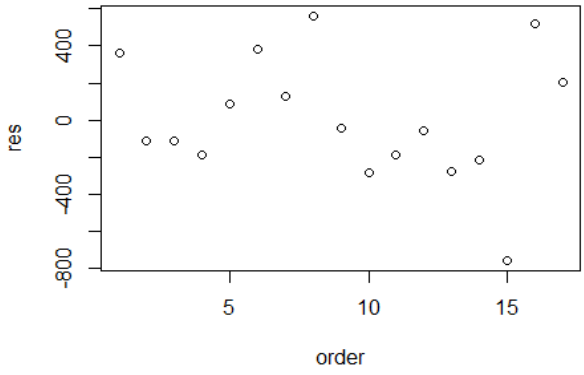
```
##
## Call:
## lm(formula = y1 ~ z1 + z2, data = drug)
##
## Coefficients:
## (Intercept)           z1           z2
##      56.720      507.073        0.329
```

(ii) Analyze the residuals









I would say the residuals are independent and follow a normal distribution, but I do spy an outlier.

(iii) Construct a 95% prediction interval for Total TCAD for $z_1 = 1$, $z_2 = 1200$, $z_3 = 140$, $z_4 = 70$, and $z_5 = 85$.

```
## The 95% prediction interval is:
##        fit      lwr      upr
## 1 958.5473 154.0402 1763.054
```

(b) Repeat (a) using the second response Y_2

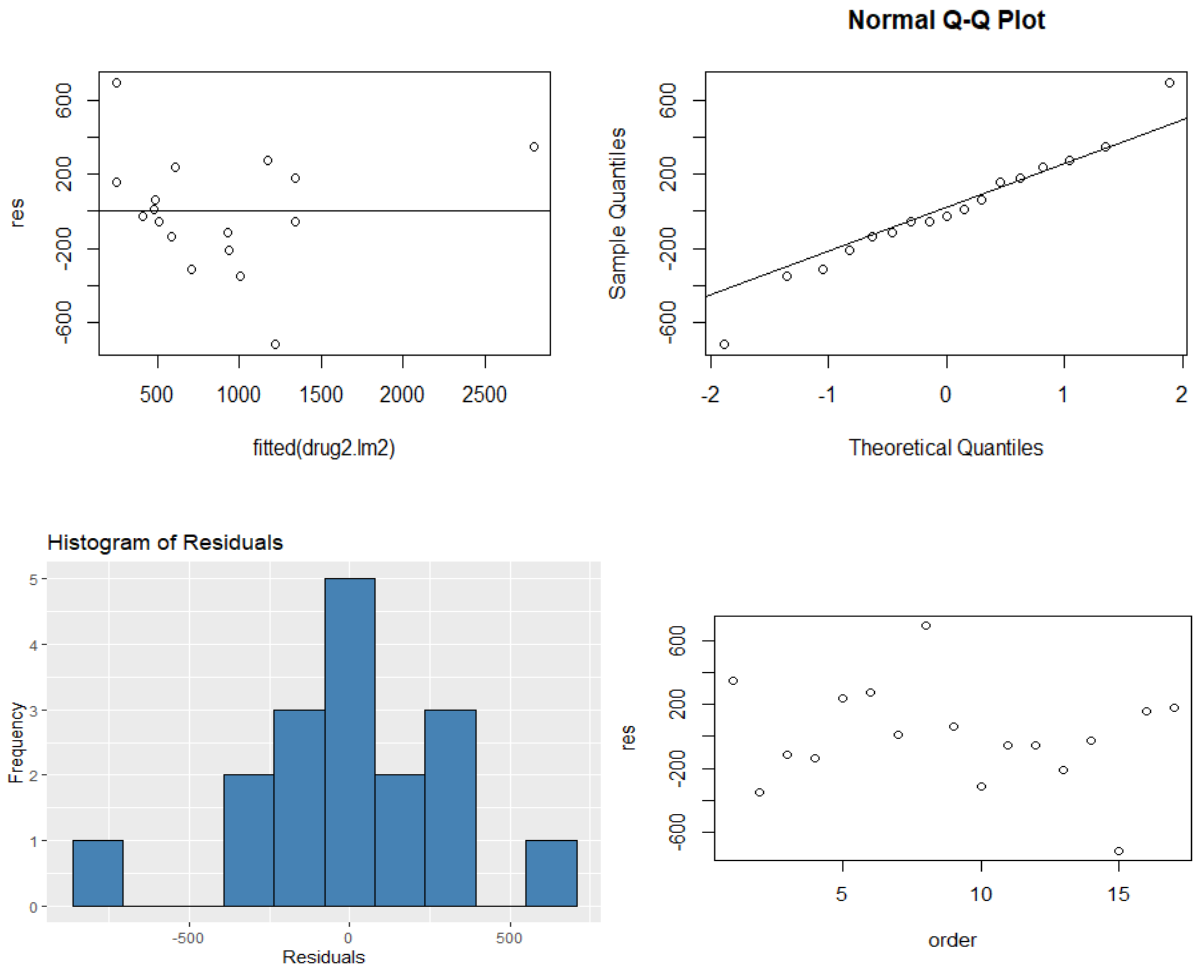
```
## Analysis of Variance Table
##
## Response: y2
##           Df  Sum Sq Mean Sq F value    Pr(>F)
## z1         1  532382  532382  6.2253   0.02977 *
## z2         1 5457338 5457338 63.8143 6.623e-06 ***
## z3         1  227012  227012  2.6545   0.13153
## z4         1  320151  320151  3.7436   0.07913 .
## z5         1  132786  132786  1.5527   0.23862
## Residuals 11  940709   85519
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Even if we change the response variable, we still find that having only two predictors, Z_1 and Z_2, is the best fit, even if the intercept isn't significant as shown below.

```
##
## Call:
## lm(formula = y2 ~ z1 + z2, data = drug)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -716.80 -135.83  -23.16  182.27  695.97
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -241.34791  196.11640  -1.231  0.23874
## z1           606.30967  183.86521   3.298  0.00529 **
## z2             0.32425    0.04723   6.866 7.73e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 340.2 on 14 degrees of freedom
## Multiple R-squared:  0.787,  Adjusted R-squared:  0.7566
## F-statistic: 25.87 on 2 and 14 DF,  p-value: 1.986e-05
##
##  Our model is
##
## Call:
## lm(formula = y2 ~ z1 + z2, data = drug)
```

```
##
## Coefficients:
## (Intercept)           z1           z2
##   -241.3479     606.3097       0.3243
```

Let's look at the residuals.



I would say the residuals are independent and follow a normal distribution, but I do spy an outlier or two.

Finally, the 95% prediction interval, where $z_1 = 1, z_2 = 1200, z_3 = 140, z_4 = 70$, and $z_5 = 85$.

```
## The 95% prediction interval is:
##        fit       lwr      upr
## 1 754.0677 -9.234071 1517.369
```

(c) Perform a multivariate multiple regression analysis using both responses Y_1 and Y_2.

```
## Response y1 :
##
## Call:
## lm(formula = y1 ~ z1 + z2 + z3 + z4 + z5, data = drug)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -399.2 -180.1    4.5  164.1  366.8
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -2.879e+03  8.933e+02  -3.224 0.008108 **
## z1           6.757e+02  1.621e+02   4.169 0.001565 **
## z2           2.848e-01  6.091e-02   4.677 0.000675 ***
## z3           1.027e+01  4.255e+00   2.414 0.034358 *
## z4           7.251e+00  3.225e+00   2.248 0.046026 *
## z5           7.598e+00  3.849e+00   1.974 0.074006 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 281.2 on 11 degrees of freedom
## Multiple R-squared:  0.8871, Adjusted R-squared:  0.8358
## F-statistic: 17.29 on 5 and 11 DF,  p-value: 6.983e-05
##
##
## Response y2 :
##
## Call:
## lm(formula = y2 ~ z1 + z2 + z3 + z4 + z5, data = drug)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -373.85 -247.29  -83.74  217.13  462.72
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -2.729e+03  9.288e+02  -2.938 0.013502 *
## z1           7.630e+02  1.685e+02   4.528 0.000861 ***
## z2           3.064e-01  6.334e-02   4.837 0.000521 ***
## z3           8.896e+00  4.424e+00   2.011 0.069515 .
## z4           7.206e+00  3.354e+00   2.149 0.054782 .
## z5           4.987e+00  4.002e+00   1.246 0.238622
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 292.4 on 11 degrees of freedom
## Multiple R-squared:  0.8764, Adjusted R-squared:  0.8202
## F-statistic:  15.6 on 5 and 11 DF,  p-value: 0.0001132
```

For Y_1, I could drop Z_5; but for Y_2, I would drop the last three variables.

(i) Suggest and fit appropriate linear regression models.

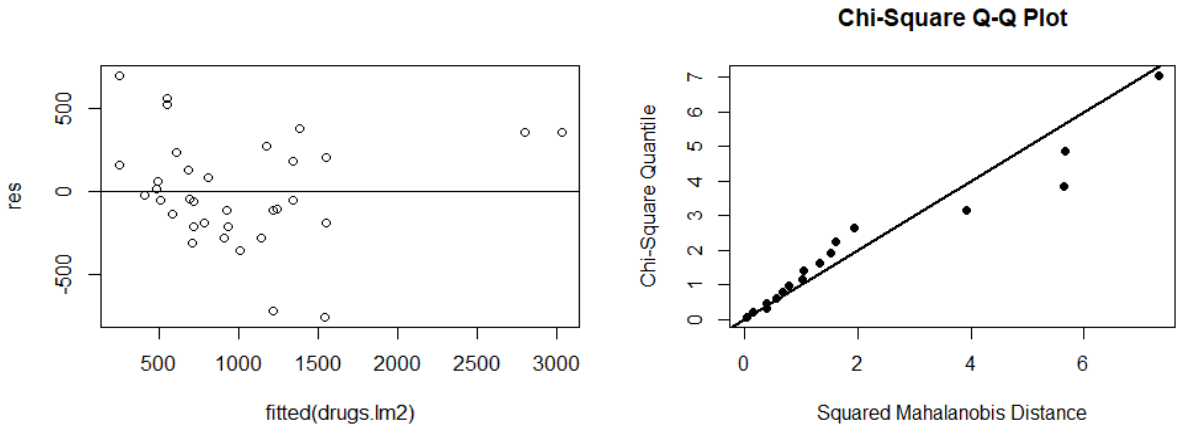
```
##
## Type II MANOVA Tests: Pillai test statistic
##    Df test stat approx F num Df den Df   Pr(>F)
## z1  1   0.65521   9.5015      2     10 0.004873 **
## z2  1   0.69097  11.1795      2     10 0.002819 **
## z3  1   0.34649   2.6509      2     10 0.119200
## z4  1   0.32381   2.3944      2     10 0.141361
## z5  1   0.29184   2.0606      2     10 0.178092
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

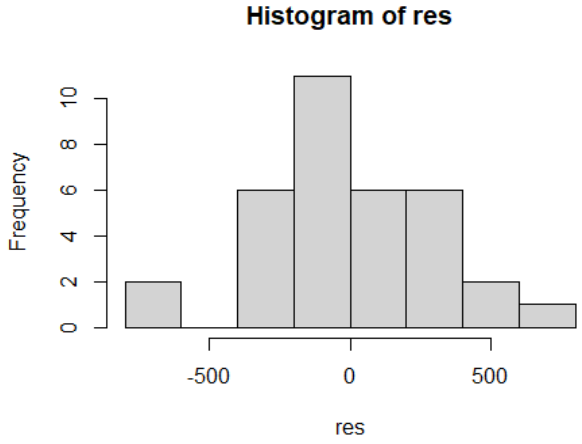
It seems we will need to keep Z_1 and Z_2 for both of them.

```
## Response y1 :
##
## Call:
## lm(formula = y1 ~ z1 + z2, data = drug)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -756.05 -190.68  -59.83  203.32  560.84
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  56.72005  206.70337   0.274   0.7878
## z1          507.07308  193.79082   2.617   0.0203 *
## z2            0.32896    0.04978   6.609 1.17e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 358.6 on 14 degrees of freedom
## Multiple R-squared:  0.7664, Adjusted R-squared:  0.733
## F-statistic: 22.96 on 2 and 14 DF,  p-value: 3.8e-05
##
##
## Response y2 :
##
## Call:
## lm(formula = y2 ~ z1 + z2, data = drug)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -716.80 -135.83  -23.16  182.27  695.97
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
```

```
## (Intercept) -241.34791  196.11640  -1.231  0.23874
## z1           606.30967  183.86521   3.298  0.00529 **
## z2             0.32425    0.04723   6.866 7.73e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 340.2 on 14 degrees of freedom
## Multiple R-squared:  0.787,  Adjusted R-squared:  0.7566
## F-statistic: 25.87 on 2 and 14 DF,  p-value: 1.986e-05
```

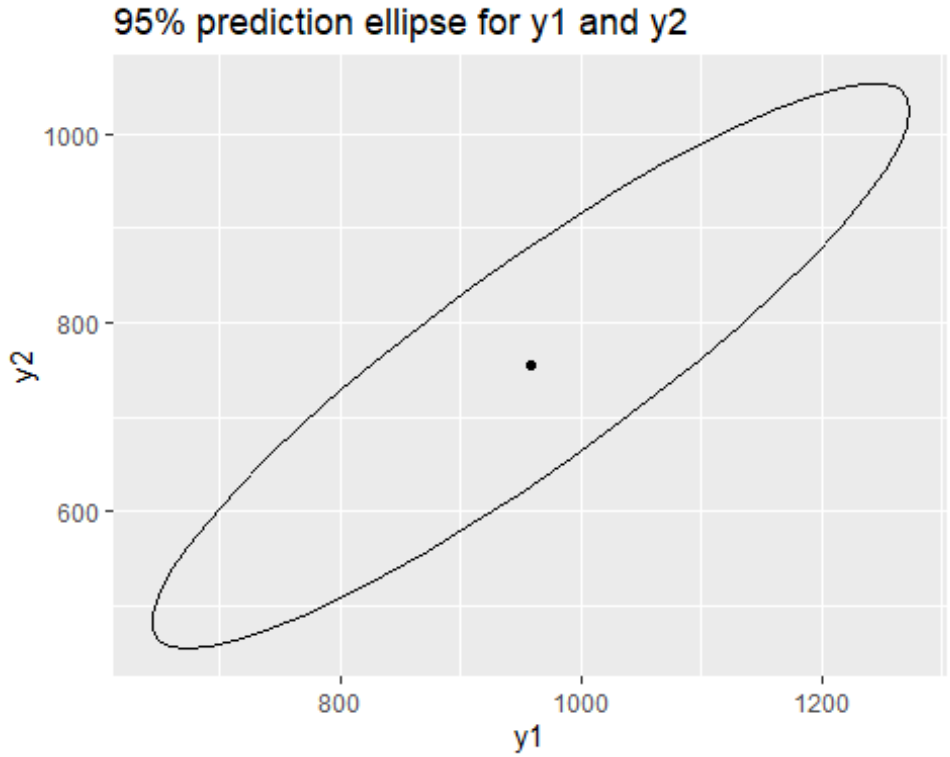
(ii) Analyze the residuals





The residuals seem to be independent and follow a normal distribution.

(iii) Construct a 95% prediction ellipse for both Ttotal TCAD and Amount of amitriptyline for $z_1 = 1$, $z_2 = 1200$, $z_3 = 140$, $z_4 = 70$, and $z_5 = 85$. Compare this ellipse with the prediction intervals in (a) and (b). Comment.



Comparing the ellipse to the prev two prediction intervals

```
##        fit      lwr      upr
## 1 958.5473 154.0402 1763.054
##        fit       lwr      upr
## 1 754.0677 -9.234071 1517.369
```

The ellipse is larger, but it is because we had to to make sure we had enough points to fit both the upper values. I can't say for sure if we did consider our extremely small lower value for (b), but it does seem the ellipse fitted within those parameters.

Code Appendix

```
knitr::opts_chunk$set(echo = FALSE)
library(car)
library(dplyr)
library(ggplot2)
library(leaps)
library(lmtest)
library(matlib)
library(MVN)
library(SIBER)
library(stats)
fullMat <- c(10,5,7,19,11,18,
             2,3,3,6,7,9,
             15,9,3,25,7,13)
full <- matrix(fullMat, nrow = 3, ncol = 6, byrow = TRUE)
rownames(full)<- c("z1", "z2", "y")
full
n <- dim(full)[1]
full1 <- as.data.frame(full)
#  function
standardize = function(x){
  z <- (x - mean(x)) / sqrt((n-1)*sd(x))
  return(z)
}
# standardize
full1 <- apply(full1, 2, standardize)
cat("The standardized form of the variables are: \n")
t(full1)
# separate
c1 <- t(full1[1,])
c2 <- t(full1[2,])
# means
mc1 <- mean(c1)
mc2 <- mean(c2)
# cat
cat("\nThe mean of z1 is", mc1, "\n")
cat("The mean of z2 is", mc2, "\n")
z1 <- t(full[1,])
z2 <- t(full[2,])
y <- t(full[3,])
# means
mean1 <- mean(z1)
mean2 <- mean(z2)
mean3 <- mean(y)
# covariance
fullCov <- cov(t(full))
squareCov <- sqrt(fullCov)
# cat
```

```r
cat("The mean for z1 is:", mean1, "\n")
cat("The mean for z2 is:", mean2, "\n")
cat("The mean for y is:", mean3, "\n")
cat("The square root covariance matrix of the predictors is: \n")
squareCov
fullMat <- c(-2,-1,0,1,2,5,3,4,2,1,-3,-1,-1,2,3)
full <- matrix(fullMat, nrow = 3, ncol = 5, byrow = TRUE)
rownames(full)<- c("z1", "y1", "y2")
full
Zmat <- c(1,1,1,1,1,
          -2,-1,0,1,2)
Z <- matrix(Zmat, nrow = 2, ncol = 5, byrow = TRUE)
cat("The Z matrix is \n")
t(Z)
# inverse
prodZ <- Z %*% t(Z)
invZ <- solve(prodZ)
cat("\nThe inverse of Z'Z is \n")
invZ
b1 <- invZ %*% Z %*% full[2,]
b2 <- invZ %*% Z %*% full[3,]
b <- cbind(b1,b2)
# print
cat("Our least squares estimates matrix of our parameters is \n")
b
Yhat <- t(Z) %*% b
Yhat
y1 <- full[2,]
y1 <- t(y1)
y2 <- full[3,]
y2 <- t(y2)
y <- rbind(y1,y2)
error <- t(y) - Yhat
# print
cat("The residual matrix is \n")
error
prodY <- y %*% t(y)
cat("The Y'Y matrix is \n")
prodY
# hats
prodHat <- t(Yhat) %*% Yhat
prodError <- t(error) %*% error
cat("\nThe right side of the equation is \n")
prodHat + prodError
muMat <- c(4,3,-2)
mu <- matrix(muMat, nrow = 3, ncol = 1, byrow = TRUE)
sigMat <- c(9,3,1,3,2,1,1,1,1)
sigma <- matrix(sigMat, nrow = 3, ncol = 3, byrow = TRUE)
cat("mu Matrix \n")
mu
```

```r
cat("\n Sigma matrix \n")
sigma
muY <- mu[1,]
muZ <- mu[2:3,]
muZ <- muZ
# sigma
sYY <- sigma[1,1]
ssZYmat <- c(3,1)
ssZY <- matrix(ssZYmat, nrow = 1, ncol = 2, byrow = TRUE)
sZY <- t(ssZY)
sZZmat <- c(2,1,1,1)
sZZ <- matrix(sZZmat, nrow = 2, ncol = 2, byrow = TRUE)
# betas
b <- inv(sZZ) %*% sZY
b0 <- muY - (t(b) %*% muZ)
# print
cat("Our beta matrix is \n")
b
cat("\nBeta-0 is", b0)
MSE <- sYY - ssZY %*% inv(sZZ) %*% sZY
MSE
pY <- sqrt((ssZY %*% inv(sZZ) %*% sZY)/sYY)
pY
sigma
# partition
s1mat <- c(9,3,3,2)
s2mat <- c(1,1)
s4 <- sigma[3,3]
# matrix
s1 <- matrix(s1mat, nrow = 2, ncol = 2, byrow = TRUE)
s2 <- matrix(s2mat, nrow = 2, ncol = 1, byrow = TRUE)
s3 <- matrix(s2mat, nrow = 1, ncol = 2, byrow = TRUE)
# calculate
mat <- s1 - (s2 %*% t(s4) %*% s3)
# print
cat("\nThe covariance matrix is \n")
mat
rho <- mat[1,2]/sqrt(mat[1,1] * mat[2,2])
rho
largeMat <- c(1, 108.28, 17.05, 1484.10,
              2, 152.36, 16.59, 750.33,
              3, 95.04, 10.91, 766.42,
              4, 65.45, 14.14, 1110.46,
              5, 62.97, 9.52, 1031.29,
              6, 263.99, 25.33, 195.26,
              7, 265.19, 18.54, 193.83,
              8, 285.06, 15.73, 191.11,
              9, 92.01, 8.10, 1175.16,
              10, 165.68, 11.13, 211.15)
large <- matrix(largeMat, nrow = 10, ncol = 4, byrow = TRUE)
```

```r
colnames(large) <- c("order", "sales", "profits", "assets")
large
large <- as.data.frame(large)
sales <- large$sales
profits <- large$profits
assets <- large$assets
order <- large$order
# fit
company <- lm(profits ~ sales + assets)
summary(company)
res <- resid(company)
# plot
plot(fitted(company), res)
abline(0,0)
# qq
qqnorm(res)
qqline(res)
# histogram
ggplot(data = large, aes(x = res)) +
    geom_histogram(bins = 6,fill = 'steelblue', color = 'black') +
    labs(title = 'Histogram of Residuals', x = 'Residuals', y = 'Frequency')
# plot
plot(order, res)
com <- large[2:4]
n <- dim(com)[1]
p <- dim(com)[2]
# leverage
avg <- 3*(p/n)
cat("The average leverage is", avg)
hats <- as.data.frame(hatvalues(company))
hats
newCo <- data.frame(sales = 100, assets = 500)
predict(company, newdata = newCo, interval = "prediction", level = 0.95)
lineCo <- lm(profits ~ sales)
lrtest(lineCo, company)
bone <- read.table("D:/Coding/R Storage/T1-8.dat", header = FALSE)
# names
domRad <- bone$V1
rad <- bone$V2
domHum <- bone$V3
hum <- bone$V4
domUln <- bone$V5
ulna <- bone$V6
radius.lm <- lm(domRad ~ domHum + hum + domUln + ulna)
summary(radius.lm)
rad.lm1 <- lm(domRad ~ domHum + domUln + ulna)
rad.lm2 <- lm(domRad ~ domHum)
summary(rad.lm1)
summary(rad.lm2)
rad.lm3 <- lm(domRad ~ domHum + domUln)
```

```r
rad.lm4 <- lm(domRad ~ domHum + ulna)
summary(rad.lm3)
summary(rad.lm4)
order <- c(1,2,3,4,5,6,7,8,9,10,11,12,13,14,15,16,17,18,19,20,21,22,23,24,25)
bones <- cbind(order, bone)
order <- bones$order
res <- resid(rad.lm3)
# plot
plot(fitted(rad.lm3), res)
abline(0,0)
# qq
qqnorm(res)
qqline(res)
# histogram
ggplot(data = bone, aes(x = res)) +
    geom_histogram(bins = 8,fill = 'steelblue', color = 'black') +
    labs(title = 'Histogram of Residuals', x = 'Residuals', y = 'Frequency')
# plot
plot(order, res)
rads.lm <- lm(cbind(domRad, rad)~ domHum + hum + domUln + ulna, data = bone)
summary(rads.lm)
rad.lm5 <- lm(rad ~ domHum + hum + domUln + ulna, data = bone)
# AIC
b <- AIC(rad.lm3)
full <- AIC(rad.lm5)
# print
cat("The AIC for the model domRad = domHum + domUln is", b, "\n")
cat("The AIC for the model rad = domHum + hum + domUln + ulna is", full)
drug <- read.table("D:/Coding/R Storage/T7-6.dat", header = FALSE)
# names
y1 <- drug$V1
y2 <- drug$V2
z1 <- drug$V3
z2 <- drug$V4
z3 <- drug$V5
z4 <- drug$V6
z5 <- drug$V7
drug1.lm1 <- lm(y1 ~ z1 + z2 + z3 + z4 + z5, data = drug)
summary(drug1.lm1)
anova(drug1.lm1)
drug1.lm2 <- lm(y1 ~ z2, data = drug)
summary(drug1.lm2)
drug1.lm3 <- lm(y1 ~ z1 + z2, data = drug)
summary(drug1.lm3)
drug1.lm4 <- lm(y1 ~ z1 + z2 + z3, data = drug)
drug1.lm5 <- lm(y1 ~ z1 + z2 + z3 + z4, data = drug)
drug1.lm6 <- lm(y1 ~ z1, data = drug)
summary(drug1.lm4)
summary(drug1.lm5)
summary(drug1.lm6)
```

```r
drug1.lm3
order <- c(1,2,3,4,5,6,7,8,9,10,11,12,13,14,15,16,17)
drugs <- cbind(order, drug)
order <- drugs$order
res <- resid(drug1.lm3)
# plot
plot(fitted(drug1.lm3), res)
abline(0,0)
# qq
qqnorm(res)
qqline(res)
# histogram
ggplot(data = drug, aes(x = res)) +
    geom_histogram(bins = 10,fill = 'steelblue', color = 'black') +
    labs(title = 'Histogram of Residuals', x = 'Residuals', y = 'Frequency')
# plot
plot(order, res)
newD <- data.frame(z1 = 1, z2 = 1200, z3 = 140, z4 = 70, z5 = 85)
pi1 <- predict(drug1.lm3, newdata = newD, interval = "prediction", level =
0.95)
cat("The 95% prediction interval is: \n")
pi1
drug2.lm1 <- lm(y2 ~ z1 + z2 + z3 + z4 + z5, data = drug)
anova(drug2.lm1)
drug2.lm2 <- lm(y2 ~ z1 + z2, data = drug)
summary(drug2.lm2)
cat("\n Our model is \n")
drug2.lm2
res <- resid(drug2.lm2)
# plot
plot(fitted(drug2.lm2), res)
abline(0,0)
# qq
qqnorm(res)
qqline(res)
# histogram
ggplot(data = drug, aes(x = res)) +
    geom_histogram(bins = 10,fill = 'steelblue', color = 'black') +
    labs(title = 'Histogram of Residuals', x = 'Residuals', y = 'Frequency')
# plot
plot(order, res)
pi2 <- predict(drug2.lm2, newdata = newD, interval = "prediction", level =
0.95)
cat("The 95% prediction interval is: \n")
pi2
drugs.lm1 <- lm(cbind(y1, y2)~ z1 + z2 + z3 + z4 + z5, data = drug)
summary(drugs.lm1)
Anova(drugs.lm1)
drugs.lm2 <- lm(cbind(y1, y2)~ z1 + z2, data = drug)
summary(drugs.lm2)
```

```r
res <- resid(drugs.lm2)
# plot
plot(fitted(drugs.lm2), res)
abline(0,0)
# qq
mvn(res, multivariatePlot = "qq")
# histogram
hist(res)
confidenceEllipse <- function(mod, newdata, level = 0.95, ggplot = TRUE){
# labels
lev_lbl <- paste0(level * 100, "%")
resps <- colnames(mod$coefficients)
title <- paste(lev_lbl, "prediction ellipse for", resps[1], "and", resps[2])
# prediction
p <- predict(mod, newdata)
# center of ellipse
cent <- c(p[1,1],p[1,2])
# shape of ellipse
Z <- model.matrix(mod)
Y <- mod$model[[1]]
n <- nrow(Y)
m <- ncol(Y)
r <- ncol(Z) - 1
S <- crossprod(resid(mod))/(n-r-1)
# radius of circle generating the ellipse
tt <- terms(mod)
Terms <- delete.response(tt)
mf <- model.frame(Terms, newdata, na.action = na.pass,
                  xlev = mod$xlevels)
z0 <- model.matrix(Terms, mf, contrasts.arg = mod$contrasts)
rad <- sqrt((m*(n-r-1)/(n-r-m)) * qf(level,m,n-r-m) *
              z0 %*% solve(t(Z)%*%Z) %*% t(z0))
 # generate ellipse using ellipse function in car package
ell_points <- car::ellipse(center = c(cent), shape = S,
                           radius = c(rad), draw = FALSE)
# ggplot2 plot
if(ggplot){
  ell_points_df <- as.data.frame(ell_points)
  ggplot2::ggplot(ell_points_df, ggplot2::aes(.data[["x"]], .data[["y"]])) +
  ggplot2::geom_path() +
  ggplot2::geom_point(ggplot2::aes(x = .data[[resps[1]]],
                                   y = .data[[resps[2]]]),
                      data = data.frame(p)) +
  ggplot2::labs(x = resps[1], y = resps[2],
                    title = title)
  } else {
    # base R plot
    plot(ell_points, type = "l",
         xlab = resps[1], ylab = resps[2],
         main = title)
```

```
    points(x = cent[1], y = cent[2])
  }
}

# ellipse
confidenceEllipse(mod = drugs.lm2, newdata = newD)
pi1
cat("\n")
pi2
```