Continuous Data: HW3

Omkar Kulkarni

Cystic Fibrosis (CF) is a multisystem disease that results from loss of function in the CF transmembrane conductance regulator (CFTR) gene. In one study, O'Neill1 and colleagues measured the Maximal static expiratory pressures (PEmax), of 25 patients with CF together with other important covariates. Their goal was to determine factors that influence the PEmax. The aim of this homework is to build a predictive model for PEmax using the weight (in kg) of the patients based on the. One way to do this is by fitting a simple linear regression model of the form:

$Y_i = \alpha + \beta X_i + \varepsilon_i$

The intercept **α** is estimated to have mean of 63.54 with 12.70 standard error and a confidence interval of 37.27 to 89.89 at 95%. The slope **β** is estimated to have mean of 1.18 with 0.3 standard error and a confidence interval of 0.564 to 1.809 at 95%. This is tabulated in table 1 of appendix.

There is no meaningful interpretation of intercept, α = 63.54, as in our setup we do not observe 0 weight. However, for β = 1.18 we can say, for per unit(kg) increase of weight among white population we observe 1.18 unit of increase in PEmax (Maximal static expiratory pressures). Refer figure 1 in appendix.

The 95% confidence interval for β is (0.56, 1.81). Thus with confidence coefficient 0.95, we estimate that the mean PEmax increases by somewhere between 0.56 and 1.81 for each addition of weight per unit (kg).

For two sided hypothesis test to test that regression coefficient is 0.

$$H_0 : \beta = 0$$

$$H_1 : \beta \neq 0$$

This can be reached by referring to 95% confidence interval of β which is (0.56, 1.81) as this interval does not include 0, we reject H_0 and say $\beta \neq 0$

Also, $| t^*| > t\ (1 - \frac{\alpha}{2}; n - 2)$, conclude H_1 i.e 3.94 > 2.069. Hence there is a linear association between PEmax and weight.

The scope of the model lies between 13 kg and 74 kg of weight. For values outside this range, no associative hypothesis can be made. Moreover, there is no mention of the Cystic Fibrosis being an adult phenomenon, if it is not the lower range of weight is questionable in the data. However, if it is Cystic Fibrosis is observable only among adults, there is no need to be alarmed with weight range. Also, figure 5 shows boxplot of weight, and no outliers or skewness are seen in weights, the predictor variable.

Randomness in residuals is evident from figure 2 and figure 3 of the appendix also the assumptions of linearity, normality and independence of errors, presence of outliers are checked. As no extreme outliers are seen, hence there is no disproportionate pull of the regression. Figure 6, depicts the boxplot, QQ plot of residuals and a decent normalcy is seen, which strengths the assumption of randomness of ε_i The variance of the residuals needs to be constant for each level of the predictor variable, we can check it with squared residuals versus the predictor values, weight, as seen in figure 4. Slight increase in R^2 is evident as weights increase.

BONUS QUESTION: By centering the predictor variables, i.e subtracting the variable's mean from each observation for that variable, the model was checked for interaction between weight and height. From table 1.1 it is clear, the estimated mean and SD of interaction term is 0.052 ± 0.018 with 95% confidence interval of 0.01 to 0.09, with p-value =0.01, p < .05

The SAS code can be seen in Appendix B.

Theoretical Part

Listing the 3 models :

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{1i}.X_{2i} + \varepsilon_i \qquad ------------- \quad 1$$

$$Y_i = \widetilde{\beta_0} + \widetilde{\beta_1} X_{1i} + \widetilde{\beta_2} X_{2i} + \widetilde{\varepsilon_i} \qquad ------------- \quad 2$$

$$Y_i = \beta_0^* + \beta_1^* X_{1i} + \varepsilon_i^* \qquad ------------- \quad 3$$

And theε_i, $\widetilde{\varepsilon_\iota}$, ε_i^* are normally distributed with mean = 0 and variance $\sigma^2, \widetilde{\sigma^2}, \sigma^{*2}$ respectively.

Where Y_i is blood pressure for individual i, X_{1i} weather person i uses new treatment =1, or not =0. X_{2i} =1 (=0) indicates presence (absence) of the gene. And 70% of the population are taking the treatment. Also,

$$P(X_{2i} = 1 \,|X_{1i} = 0) = q_0$$

$$P(X_{2i} = 1 \,|X_{1i} = 1) = q_1$$

a. Using the true data-generating model (1), write down the average causal effect of the treatment on the blood pressure for fixed levels of x2, separately for patients with the genotype and for patients without the genotype.

For Without genotype, $X_{2i} = 0$ and hence using equation 1, we can write

$$E(Y_i) = \beta_0 + \beta_1 X_{1i}$$

Hence, the slope β_1 is the average causal effect of the treatment on the blood pressure for patients without genotype.

For With genotype, $X_{2i} = 1$ and hence using equation 1, we can write

$$E(Y_i) = (\beta_0 + \beta_2\,) + (\beta_1 + \beta_3) X_{1i}$$

Hence, the slope $(\beta_1 + \beta_3)$ is the average causal effect of the treatment on the blood pressure for patients with genotype.

b. Does estimating the effect of x1 on Y given x2 based on model (2) result in unbiased estimates of these causal effects ? Explain.

If $\beta_3 = 0$ is assumed, then the model 1 and model 2 are equal, and there is no bias in parameter estimators of model 2. However, if the assumption 'β = 0' fails, then there is bias in parameter estimators of model 2, as the interaction term is missing. (With the interaction term $\beta_3$the model 1 is the least squares model, and model 2 would fail to reach the optimum least squares of residuals without the β_3).

For the assumption $\beta_3 = 0 \; to \; hold$:

Joint Probability of $X_{1i} \; and \; X_{2i}$

	$X_{2i} = 0$	$X_{2i} = 1$	Total
$X_{1i} = 0$	$0.3 * (1 - q_0)$	$q_0 * 0.3$	0.3
$X_{1i} = 1$	$0.7 * (1 - q_1)$	$q_1 * 0.7$	0.7
Total	$0.3 * (1 - q_0) + 0.7 * (1 - q_1)$	$q_0 * 0.3 + q_1 * 0.7$	1

From joint probability table we can observe ,

$$\widetilde{\beta_0} = 0.3 * (1 - q_0)$$

$$\widetilde{\beta_0} + \widetilde{\beta_1} = 0.7 * (1 - q_1)$$

$$\widetilde{\beta_0} + \widetilde{\beta_2} = q_0 * 0.3$$

$$\widetilde{\beta_0} + \widetilde{\beta_1} + \widetilde{\beta_2} = q_1 * 0.7$$

Solving for $\widetilde{\beta_0}$, $\widetilde{\beta_1}$, $\widetilde{\beta_2}$ we see, the last $\widetilde{\beta_0} + \widetilde{\beta_1} + \widetilde{\beta_2}$ equals sum of 2nd and 3rd minus the 1st above equations, and hence solving for q_1

$$q_1 = \frac{q_0 * 0.6 + 0.4}{1.4}$$

If the above relation holds, we can say $\beta_3 = 0$ and there is no interaction between the two predictors.

c. How does the variance (and hence precision) compare. Can you say something about the mean squared errors?

According to least squares criteria, the β parameters are calculated so as to minimize the sum of squared errors SSE, in model 1. In model 2, we make $\beta_3 = 0$ and hence the other parameters fail to optimally have least SSE. Hence, $SSE_1 < SSE_2$.

Also, $MSE = \frac{SSE}{n-p}$ and $E(MSE) = \sigma^2$; and 'p' in model 1 is 4 and in model 2 it is 3. Hence $MSE_1 < MSE_2$

2 . $Y_i = \beta_0^* + \beta_1^* X_{1i} + \varepsilon_i^*$

a. Write the parameter β ∗ 1 in function of the parameters of Model (model1), and interpret this parameter.

Comparing model 1 and model 3, we can write

$$\beta_1^* = \beta_1 + \beta_3 X_{2i}$$

β_1^* is the average effect of X_{1i} that is new treatment on the blood pressure for a given value of X_{2i} that is presence or absence of gene.

b. Does estimating β ∗ 1 result in an unbiased estimator of the (unconditional) average causal effect, assuming a randomized design ? Explain
In model 3, we are not considering the effect of interaction nor conditioning it on X_{2i}, hence it might be biased.

c. How does the variance (and hence the precision) of βˆ∗ 1 compare with that of βˆ 1 from model (1) ?

$$Var(\beta_1^*) = Var(\beta_1 + \beta_3 X_{2i})$$

$$Var(\beta_1^*) = Var(\beta_1) + Var(\beta_3 X_{2i}) + 2 * Cov(\beta_1 , \beta_3 X_{2i})$$

Hence, the $Var(\beta_1^*)$ will be bigger.

Appendix A – Tables and Figures

Table 1: Parameter Estimates for α and β, intercept and weight respectively.

Parameter Estimates								
Variable	**Label**	**DF**	**Parameter Estimate**	**Standard Error**	**t Value**	**Pr > |t|**	**95% Confidence Limits**	
Intercept	Intercept	**1**	63.54564	12.70163	5.00	<.0001	37.27032	89.82097
weight	weight	**1**	1.18671	0.30086	3.94	0.0006	0.56434	1.80907

Figure 1

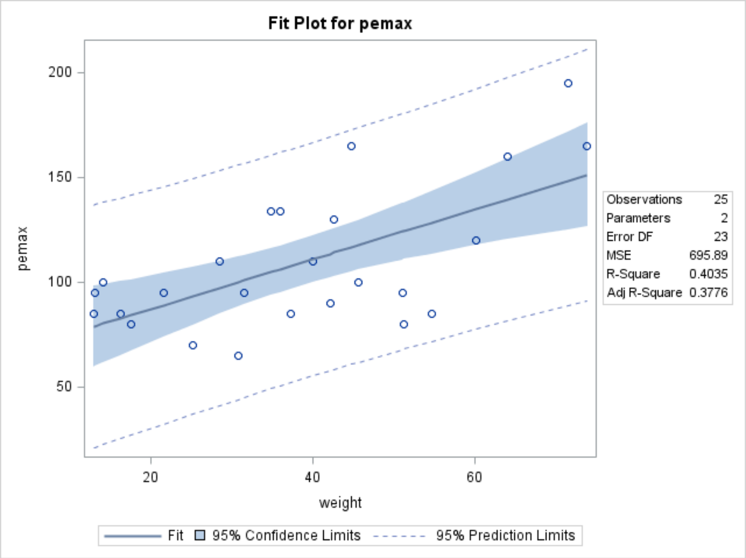


Figure 2 : Residual vs Weight

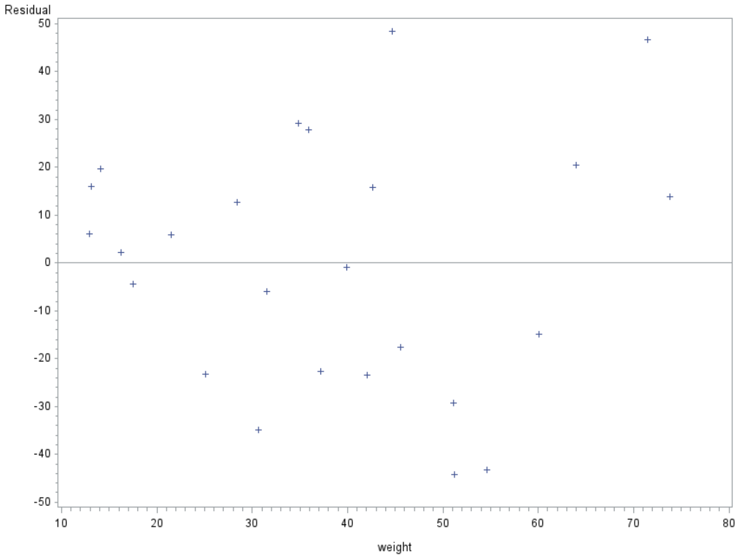


Figure 3: Residual Vs Predicted value of PEmax

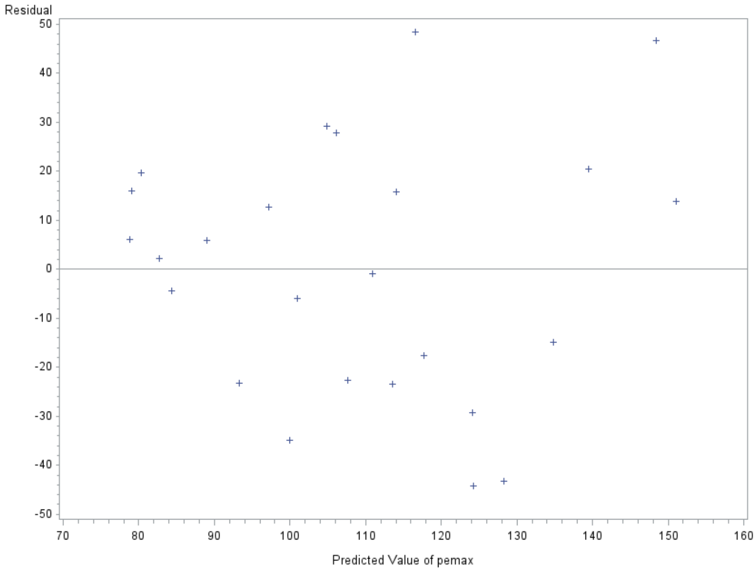


Figure 4 : Residual Square Vs Weight

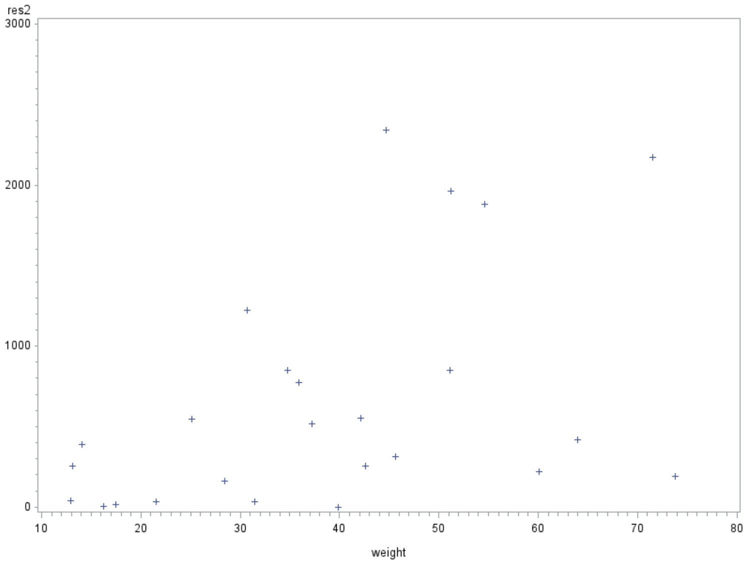


Figure 5: Histogram, boxplot, QQ plot of Weight

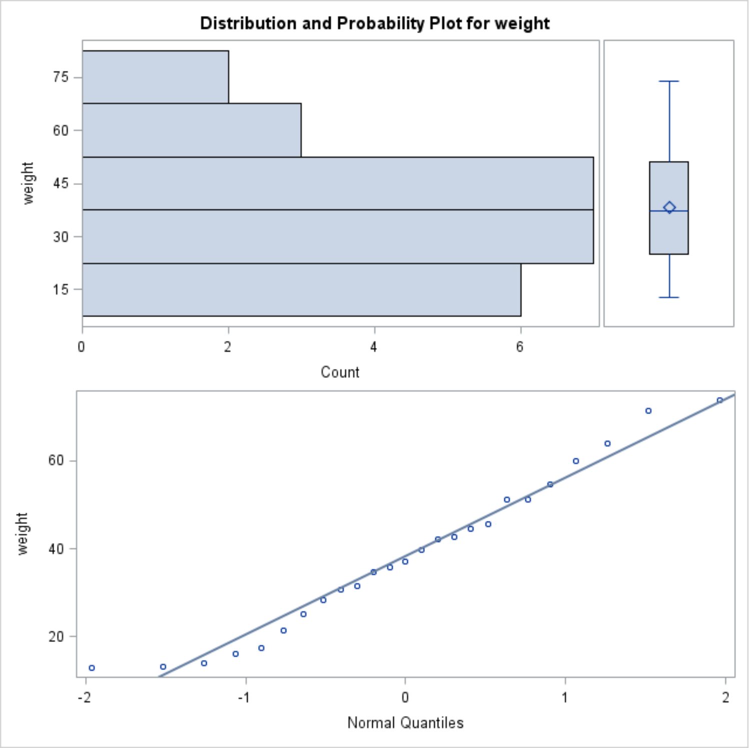
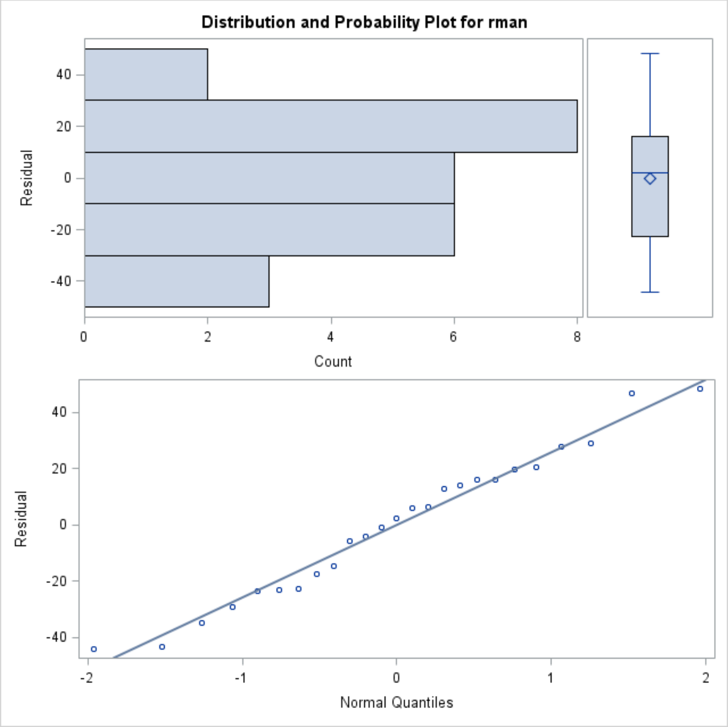


Figure 6 : Histogram, Boxplot, QQ plot of Residuals



Tabel 1.1

Parameter Estimates								
Variable	**Label**	**DF**	**Parameter Estimate**	**Standard Error**	**t Value**	**Pr > |t|**	**95% Confidence Limits**	
Intercept	Intercept	**1**	91.37170	7.94995	11.49	<.0001	74.83887	107.90454
weight	weight	**1**	-0.80201	0.95330	-0.84	0.4097	-2.78450	1.18048
height	height	**1**	1.81681	0.83177	2.18	0.0404	0.08706	3.54656
interaction		**1**	0.05218	0.01881	2.77	0.0114	0.01306	0.09130

Appendix B – SAS Code

```
/*
Author : Omkar Kulkarni
Date : 22/11/2015
Description : Analysis of continious data HW 3

*/

data Cyst;
set Conhw3.Cyst;
run;



/* Descriptive statistics */
proc univariate plot data=cyst ;
var pemax weight;
histogram pemax weight;
qqplot pemax weight;
run;

/* Exploring the relationship */
            /* 1. Proc gplot */
proc gplot data=cyst;
      plot pemax*weight;
run;

/* Linear Regression */
proc reg data = cyst;
      model pemax=weight / clb;
RUN;


proc reg data=cyst;
      model pemax=weight;
      output out=resid p=pman r=rman student=student;
run;


proc gplot data=resid;
      plot rman*weight /vref=0;
run;
proc gplot data=resid;
      plot rman*pman /vref=0;
run;

/* squared residuals versus the predictor values */
data resid2; set resid; res2 = rman*rman; run;
goptions reset=all;
proc gplot data=resid2;
plot res2*weight; run;
quit;
goptions reset=all;
proc gplot data=resid2;
plot Student*weight; run;
 quit;



/* Interaction term */
```

```
data resint;
set resid;
interaction = weight*height;
run;


proc gplot data=resint;
   plot rman*interaction /vref=0;
   run;

   proc reg data=resint;
   model pemax = weight height interaction/partial;
   run;

/*  mean-centering */

proc standard data = resint
   out = centdata
   mean = 0
   print;
  var weight height;
run;

data centdataint;
set centdata;
interaction = weight*height;
run;

proc reg data=centdataint;
   model pemax = weight height interaction/clb;
   run;
```