QUESTION II.

(1.) Define, using formulas in addition to words, the least squares criterion.

(2.) Show that the least squares criterion applied to the "intercept-only" model, i.e.,

$$y_i = \beta_o + \epsilon_i, \quad i = 1, 2, \ldots, n$$

results in the least squares estimator of β_o : $\hat{\beta}_o = \bar{y}$.

(3.) Consider the analysis of variance table for the simple linear regression model:

$$y_i = \beta_o + \beta_1 x_i + \epsilon_i, \quad i = 1, 2, \ldots, n$$

What sum of squares in the analysis of variance table for the simple linear regression model would correspond to the sum of squares error, SSE, from the "intercept-only" model in part (2.)?

(4.) With reference to your answers in the above questions, **briefly** explain in words why statisticians almost always perform a "corrected" analysis of variance, a partition of $\sum_{i=1}^{n}(y_i - \bar{y})^2$, rather than an "uncorrected" analysis of variance, a partition of $\sum_{i=1}^{n} y_i^2$

QUESTION III.

An analyst for a grocery store chain was interested in the effect of product placement on shelves (Knee, Waist, and Eye levels) and facings (or the amount of shelf space required by the products: Half or Full) on sales, the number of products sold. Three products were placed at each shelf / facing combination, for a total of 18 products. The first model fit to the data was:

$$Sales = \alpha_i + \beta_j + e_{ij}; \quad i = 1, 2, 3; \quad j = 1, 2 \qquad \textbf{(Model 1)}$$

1. Interpret the parameter estimate $\hat{\beta}_1$ from Model 1 in context.

2. If the variable Sales had been an indicator variable for whether or not a product had sold rather than the number of each product sold, would the model above have been a valid model? Why or why not?

It was also suggested that the sugar content of the products may play a part in the number sold. The second fitted model (this time with an intercept) was:

$$Sales = \beta_0 + \beta_1 Sugars + \beta_{2i} + \beta_{3j}; \quad i = 1, 2, 3; \quad j = 1, 2 \qquad \textbf{(Model 2)}$$

Output from this second model is found below and on the following page.

Model 2:

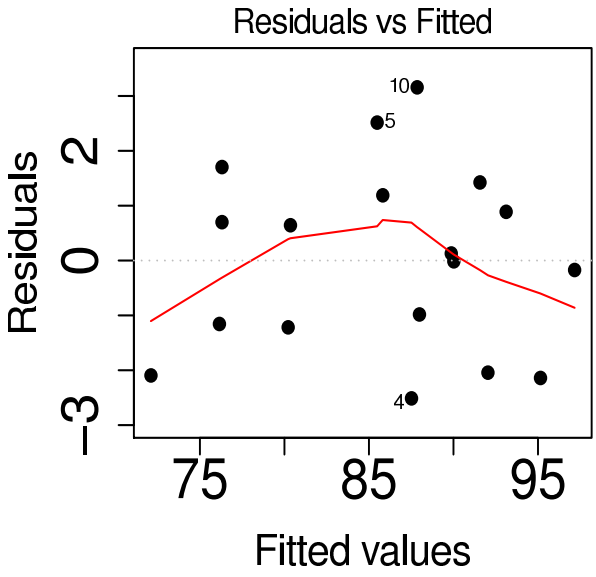
```
Coefficients:
             Estimate   Std. Error  t value      Pr(>|t|)
(Intercept)   73.774      2.875      25.661     1.60e-12 ***
FacingHalf    -7.627      1.025      -7.443     4.88e-06 ***
HeightKnee    -2.172      1.107      -1.962     0.0715
HeightWaist    3.096      1.394       2.222     0.0447 *
Sugars         2.030      0.364       5.578     8.95e-05 ***
```

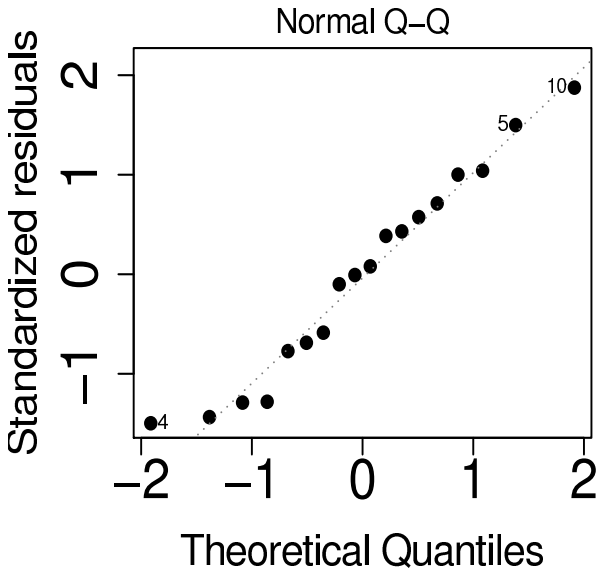
3. Is Model 2 a valid model? Discuss why or why not.

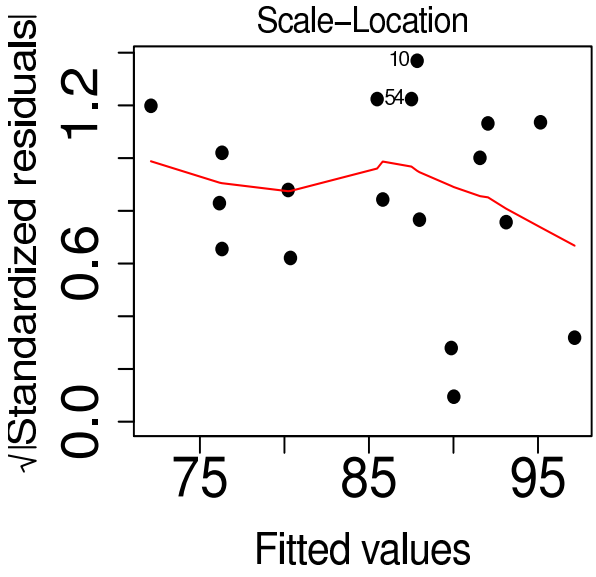
4. Regardless of your answer above, use Model 2 to test whether Sugars have an effect on Sales, after controlling for shelf placement and facing. Be sure to include the hypotheses used and your conclusion in the context of the problem.

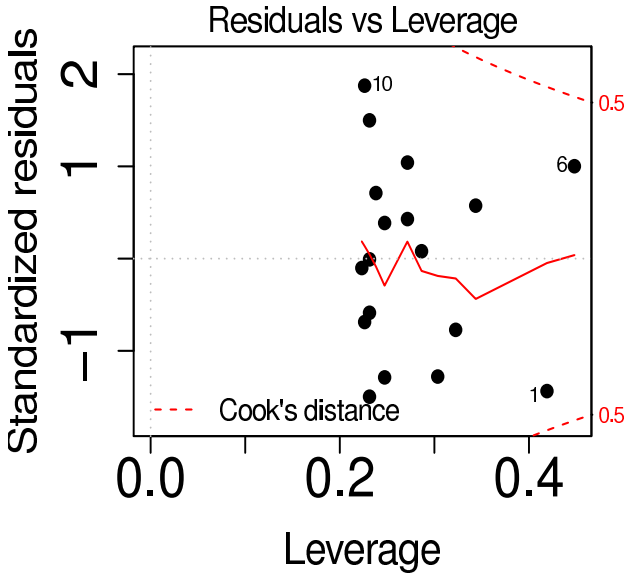
5. Suppose that a third model is fit, and an interaction between shelf placement and facing is found to be significant. How would you explain the interaction to others in the company without using statistical jargon?

Model 2: Graphical Displays









Problem V.

A book[1] on robust biostatistical methods published in 2009 considers a data set taken from Everitt (1994), *The Handbook of Statistical Analysis Using Splus*, Chapman & Hall. The data consist of the IQ scores (IQ) and behavioral problem scores (BP.score) of children of age five, labeled according to whether or not their mothers had suffered an episode of postnatal depression (state.mother = 1 if yes and 0 if no). We seek to model IQ as a function of BP.score and to determine whether the effect of BP.score differs significantly across the two groups of mothers.

The two models under consideration are as follows:

$$(1) \quad \text{IQ} = \beta_o + \beta_1 \text{BP.score} + \beta_2 \text{state.mother} + e$$

and

$$(2) \quad \text{IQ} = \beta_o + \beta_1 \text{BP.score} + \beta_2 \text{state.mother} + \beta_3 \text{BP.score} \times \text{state.mother} + e$$

A plot of the data and the two regression lines from model (2) along with a plot of the standardized residuals and leverage values from model (2) can be found below. In addition, numerical output from R for models (1) and (2) appears below.

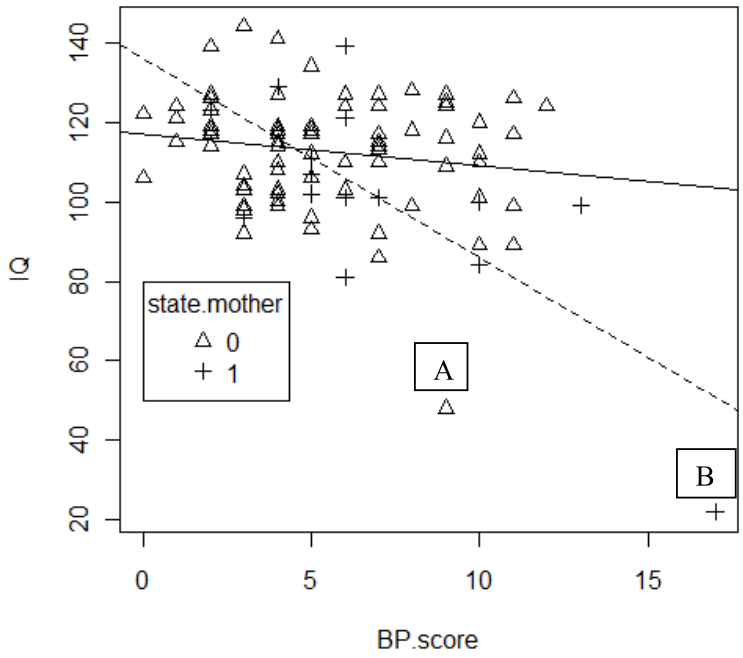
a) Two points are marked as "A" and "B" in the plot of IQ and BP.score. Two points are marked as "C" and "D" in the plot of the standardized residuals and leverage values from model (2). Match "A" and "B" with "C" and "D". Give reasons to support your choices.

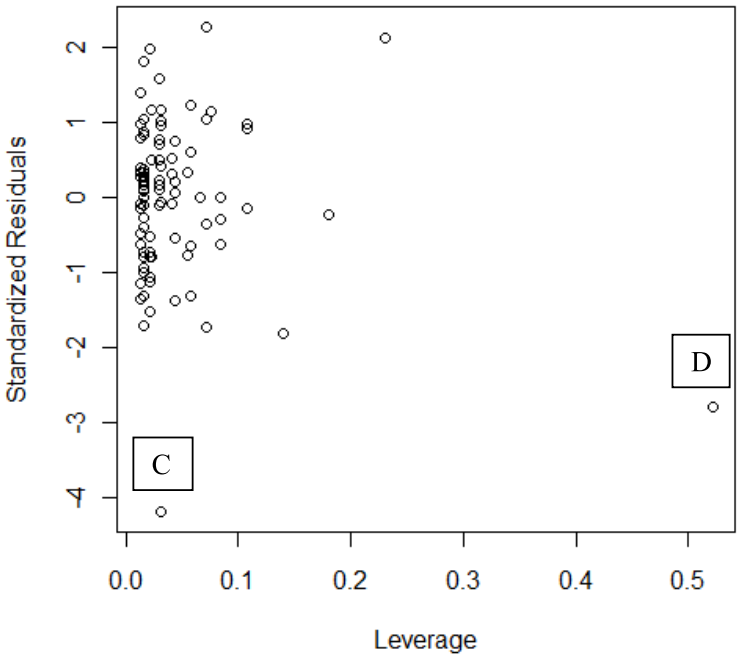
b) Using the output from R provided below, calculate the value of the F-statistic for testing the null hypothesis, $H_o : \beta_3 = 0$.

c) Briefly describe the steps you would follow in order to obtain a final model for the data on IQ and BP.score , labelled according to whether or not their mothers had suffered an episode of postnatal depression.

[1]Heritier, S., E. Cantoni, S. Copt, & M.-P. Victoria-Feser (2009) *Robust Methods in Biostatistics*. Wiley, New York

Plots associated with model (2)





Output from R for model (1)

```
Call:
lm(formula = IQ ~ BP.score + state.mother)

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  122.5832     3.3674  36.403  < 2e-16 ***
BP.score      -1.8171     0.5281  -3.441 0.000878 ***
state.mother  -8.7970     4.5782  -1.922 0.057797 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 15.9731 on 91 degrees of freedom
Multiple R-squared: 0.1699,     Adjusted R-squared: 0.1516
F-statistic: 9.312 on 2 and 91 DF,  p-value: 0.0002093
```

Edited Output from R for model (2)

```
Call:
lm(formula = IQ ~ BP.score + state.mother + BP.score:state.mother)

Coefficients:
                      Estimate Std. Error t value Pr(>|t|)
(Intercept)           117.1334     3.5034  33.434  < 2e-16 ***
BP.score               -0.8064     0.5693  -1.417 0.160063
state.mother           18.9970     8.8000   2.159 0.033531 *
BP.score:state.mother  -4.2027     ??????   ????? 0.000486 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 15.0061 on 90 degrees of freedom
Multiple R-squared: 0.2754,     Adjusted R-squared: 0.2513
F-statistic:  11.4 on 3 and 90 DF,  p-value: 2.084e-06
```

QUESTION IV:

1. The multiple regression matrix formulation is given by:

$$Y_{nx1} = X_{nx(p+1)}\beta_{(p+1)x1} + \varepsilon$$

Where:

$$\beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{bmatrix}$$

$E(\epsilon) = 0_{nx1}$ $V(\epsilon) = \sigma^2 I_{nxn}$

a. What distribution is usually assumed for ε?

b. Is it correct to test that Y has a normal distribution using a univariate test such as Sharipo-Wilks? Explain your answer.

c. What is I_{nxn}?

d. Assuming the conditions for MLR are met, how many unknown parameters are there? Display the unknown parameters.

e. If $\hat{\beta} = (X^t X)^{(-1)} X^t Y$ derive the expectation and variance of $\hat{\beta}$.

f. Is $\hat{\beta}$ an unbiased estimator of β? Explain your answer.

2. A researcher states that he has two models:

Model 1: $Y_{nx1} = X_{nx(p+1)}\beta_{(p+1)x1} + \varepsilon$ with R-squared = .7

Model 2: $\log(Y_{nx1}) = X_{nx(p+1)}\beta_{(p+1)x1} + \varepsilon$ with R-squared = .8

Note: n/p = 50.

He also states that since the R-squared for Model 2 is greater than the R-squared for Model 1, then Model 2 is better than Model 1 .

Do you agree? Explain your answer.

Problem IV:

Part A.

In a study of the percentage of raw material that responds in a reaction, researchers identified the following five factors:

- the feed rate of the chemicals (*FeedRate*), ranging from 10 to 15 liters per minute
- the percentage of the catalyst (*Catalyst*), ranging from 1% to 2%
- the agitation rate of the reactor (*AgitRate*), ranging from 100 to 120 revolutions per minute
- the temperature (*Temperature*), ranging from 140 to 180 degrees Celsius
- the concentration (*Concentration*), ranging from 3% to 6%

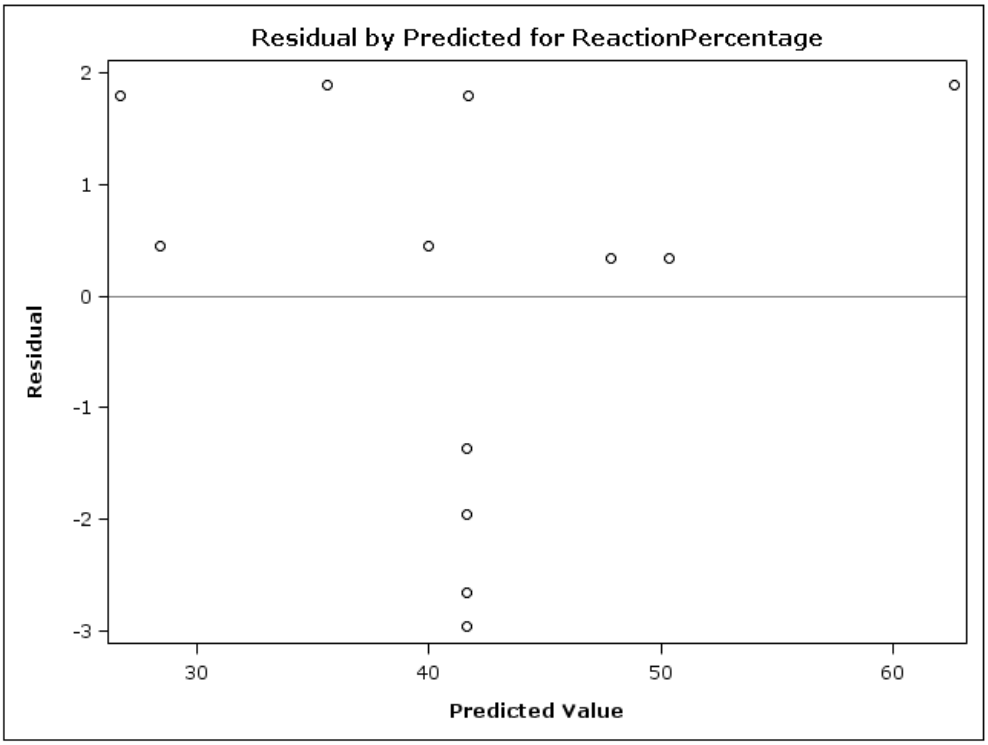
The following data set contains the results of an experiment ***designed*** to estimate main effects for all factors:

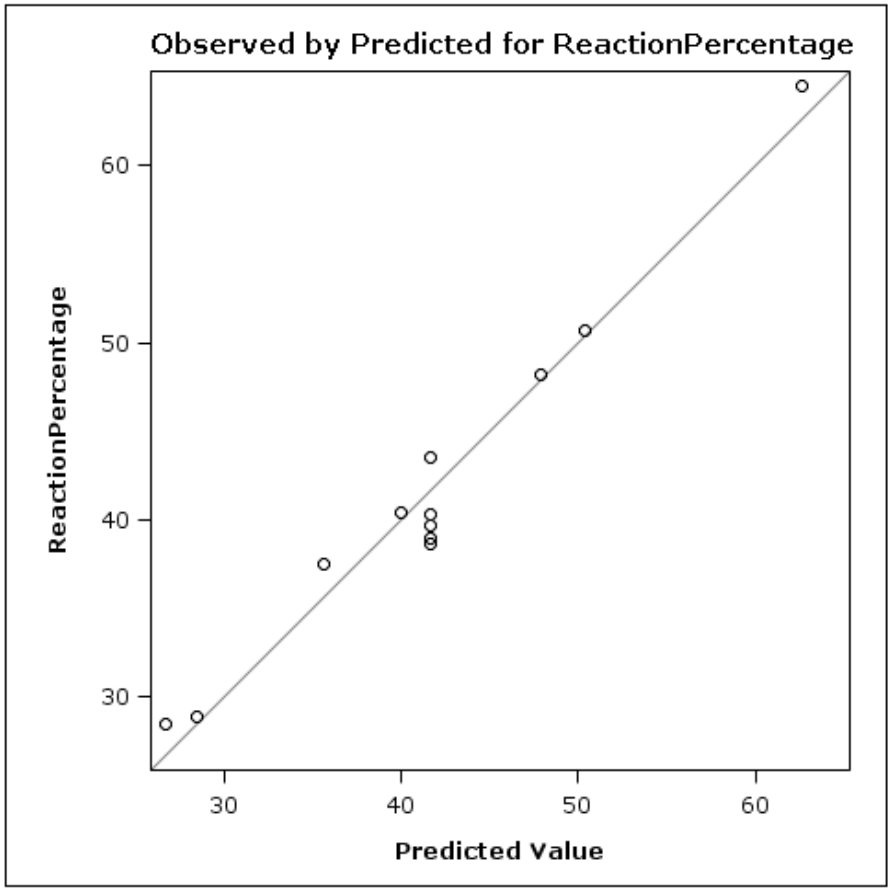
	FeedRate	Catalyst	AgitRate	Temperature	Concentration	ReactionPercentage
1	10	1	100	140	6	37.5
2	10	1	120	180	3	28.5
3	10	2	100	180	3	40.4
4	10	2	120	140	6	48.2
5	15	1	100	180	6	50.7
6	15	1	120	140	3	28.9
7	15	2	100	140	3	43.5
8	15	2	120	180	6	64.5
9	12.5	1.5	110	160	4.5	39
10	12.5	1.5	110	160	4.5	40.3
11	12.5	1.5	110	160	4.5	38.7
12	12.5	1.5	110	160	4.5	39.7

The model is :

$$\text{ReactionPercentage} = \beta_o + \beta_1 * \text{FeedRate} + \beta_2 * \text{Catalyst} + \beta_3 * \text{AgitRate} + \beta_4 * \text{Temperature} + \beta_5 * \text{Concentration} + \varepsilon$$

Some of the results of the regression analyses are given below:


Residual by Predicted for ReactionPercentage
Residual
2
1
0
-1
-2
-3
30
40
50
60
Predicted Value


Observed by Predicted for ReactionPercentage
ReactionPercentage
60
50
40
30
30
40
50
60
Predicted Value

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	5	990.27000	198.05400	33.29	0.0003
Error	6	35.69917	5.94986		
Corrected Total	11	1025.96917			

Root MSE	2.43923	R-Square	0.9652
Dependent Mean	41.65833	Adj R-Sq	0.9362
Coeff Var	5.85533		

Parameter Estimates						
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variance Inflation
Intercept	1	-43.69167	13.04097	-3.35	0.0154	0
FeedRate	1	1.65000	0.34496	4.78	0.0031	1.00000
Catalyst	1	12.75000	1.72480	7.39	0.0003	1.00000
AgitRate	1	-0.02500	0.08624	-0.29	0.7817	1.00000
Temperature	1	0.16250	0.04312	3.77	0.0093	1.00000
Concentration	1	4.96667	0.57493	8.64	0.0001	1.00000

What would you recommend to your client?

Answer the above question in terms of the following questions:

1) Is this a valid model? Why or why not?

2) If you need to add interactions and squared terms, can you just add them as a group to the model and run the analyses? Why or why not?

3) What is being tested by the F value of 33.29 in the ANOVA table? Answer in terms of the $\beta's$?

Part B.

The usual multiple linear regression model can be written as:

$$Y = X\beta + \varepsilon$$

where $V(\varepsilon) = \sigma^2 I$ and I is the (nxn) identity matrix so that $\mathrm{V(Y|X)} = \sigma^2 I$.

However, if $V(\varepsilon) = R_{nxn}$, then (Show work)

a) What is the $\mathrm{V(Y|X)}$?

B) If $\hat{Y} = X\hat{\beta} = X(X'X)^{-1}X'Y = HY$ then what is $\mathrm{V(\hat{Y}|X)}$?

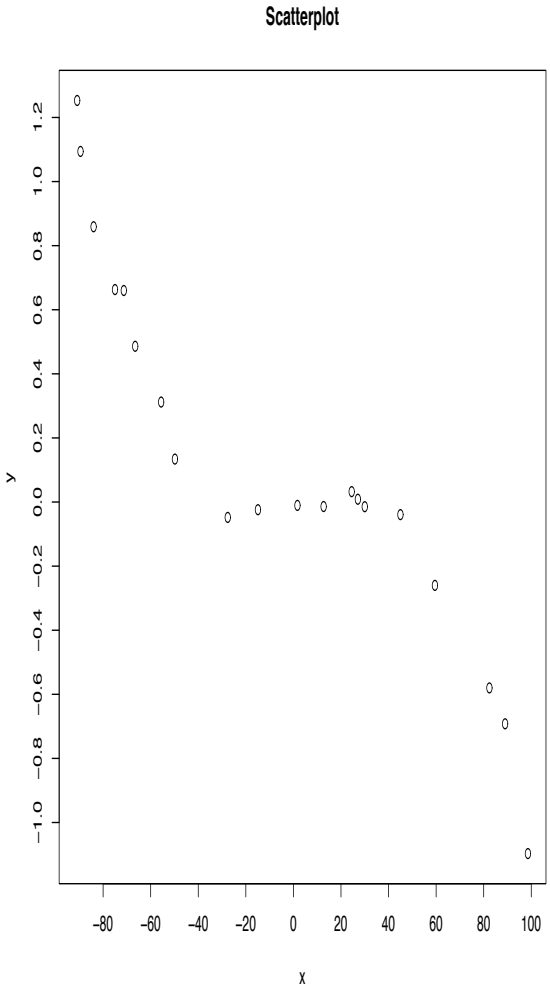
Problem III.

An engineer wants to evaluate the relationship between the standardized amount of additive, x, used in a chemical reaction and the deviation from the standard yield of a chemical reaction, y. Given the data in Table given below and the scatterplot of the data, answer the following questions. Please explain your answers. Do not make any calculations.

1. Does a reasonable model for this data satisfy the requirement for multiple linear regression?
2. If the model $y = \beta_o + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 + \epsilon$ is fit to the data, how many distinct populations are involved in the modelling?
3. If you use the model in part 2. above and you test for equality of variance, what is the null hypothesis in terms of H_o : ....?
4. If you use Anderson-Darling statistic or Shapiro-Wilks statistic to test H_o : y has a normal distribution and reject that null hypothesis, would this conclusion violate the assumptions for multiple linear regression?

Data Table

ID	x	y
1	-90.9	1.25300
2	-89.5	1.09400
3	-84.0	0.85900
4	-75.0	0.66300
5	-71.3	0.66000
6	-66.5	0.48600
7	-55.6	0.31200
8	-49.8	0.13400
9	-27.6	-0.04800
10	-14.9	-0.02400
11	1.7	-0.01050
12	12.7	-0.01400
13	24.5	0.03240
14	27.1	0.00871
15	30.0	-0.01460
16	45.0	-0.03930
17	59.5	-0.26000
18	82.4	-0.58000
19	89.0	-0.69200
20	98.6	-1.09700



QUESTION I.

A randomized trial was conducted to investigate the relationship between a continuous response y and four treatments `A`, `B`, `C`, and `D`. The sample size was $n = 200$, with 50 observations in each of the four treatment groups. Let $\boldsymbol{y}$ be the 200×1 vector of response values, ordered so that the first 50 entries are for treatment group `A`, the next 50 for `B`, then `C`, and finally `D`. The regression model $\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{\epsilon}$ was fit, where $\boldsymbol{X}$ is the 200×4 design matrix given by

$$\boldsymbol{X} = \begin{pmatrix} \mathbf{1} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{1} & \mathbf{1} & \mathbf{0} & \mathbf{0} \\ \mathbf{1} & \mathbf{0} & \mathbf{1} & \mathbf{0} \\ \mathbf{1} & \mathbf{0} & \mathbf{0} & \mathbf{1} \end{pmatrix},$$

and where each entry is a column vector of length 50. The estimated regression coefficients were

$\hat{\boldsymbol{\beta}}' = (37.5,\ -11.5,\ 1.0,\ -27.7)$, with standard errors 2.75, 3.89, 3.89, 3.89

and residual standard deviation $\hat{\sigma} = 19.45$.

(1.) Interpret each of the **four** regression parameters. As in, "the intercept, β_o, is the mean response when ...".

(2.) What assumptions are required for the regression model?

(3.) What is an approximate 95% confidence interval for the mean response in treatment group `B`?

`Hint`: If $\boldsymbol{v}$ is a 4×1 column vector, then the variance of

$\boldsymbol{v}'\hat{\boldsymbol{\beta}}$ is equal to $\hat{\sigma}^2\boldsymbol{v}'(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{v}$.

In our example,

$$(\boldsymbol{X}'\boldsymbol{X})^{-1} = \begin{pmatrix} 0.02 & -0.02 & -0.02 & -0.02 \\ -0.02 & 0.04 & 0.02 & 0.02 \\ -0.02 & 0.02 & 0.04 & 0.02 \\ -0.02 & 0.02 & 0.02 & 0.04 \end{pmatrix}$$

(4.) What is an approximate 95% confidence interval for the mean difference in response between treatment groups `B` and `A` (so, the difference $\mu_{\mathbf{B}} - \mu_{\mathbf{A}}$)?

(5.) Suppose the observations in treatment group `A` were positively correlated with those in treatment group `B`, and you correctly fit a correlated-data regression model. How would a 95% confidence interval for the mean difference in response between treatment groups `B` and `A` compare to the one reported in (4.) above, and why?

PROBLEM V.

Let

$$Y_1 = \alpha_1 + \varepsilon_1$$

$$Y_2 = 2\alpha_1 - \alpha_2 + \varepsilon_2$$

$$Y_3 = \alpha_1 + 2\alpha_2 + \varepsilon_3 \ ,$$

where $\varepsilon_1, \varepsilon_2$ and ε_3 are iid $N(0, \sigma^2)$ random variables. Derive the F statistic for testing

$$H_0 : \alpha_1 = \alpha_2 \ .$$

Note: "Derive the F statistic" means to produce an expression which is a function only of real numbers and the random variables $Y_1, Y_2,$ and Y_3. You may (and are encouraged to) introduce simplifying notation in your derivation. However, be sure to carefully define, in terms of real numbers and $Y_1, Y_2,$ and Y_3, all notation that you introduce.

Hint: Consider the usual multiple linear regression model, $Y = X\beta + \varepsilon$. Y is an $n \times 1$ vector of response variables, X is an $n \times p$ matrix (of rank p) of predictor variables, β is a $p \times 1$ vector of unknown parameters and ε is an $n \times 1$ vector of unobservable independent and identically normally distributed random variables, each with mean zero and variance σ^2. Either of two (equivalent) forms of the F statistic for testing the null hypothesis $H_0 : A\beta = c$ versus $H_1 : A\beta \neq c$, where A is a known $q \times p$ matrix of rank q, are

$$F = \frac{\left(RSS_H - RSS\right)/q}{RSS/(n-p)} = \frac{\left(A\hat{\beta} - c\right)^T \left[A\left(X^T X\right)^{-1} A^T\right]^{-1} \left(A\hat{\beta} - c\right)/q}{RSS/(n-p)} \ ,$$

where RSS and RSS_H are the residual sum of squares for the least squares fit of the unconstrained model and the model constrained by $H_0 : A\beta = c$, respectively.

QUESTION I.

You are the consulting statistician on a study of the genetic components of exercise. The PI plans to raise 30 mice in which a particular gene suspected to be involved in regulating exercise has been knocked out (you can think of it like turning off this gene), as well as 30 wild-type mice (mice in which the gene has not been knocked out). Each mouse will then be observed over a two week period, and the average number of minutes mouse i spends running on the exercise wheel per day, y_i, will be recorded. The variables in the study are

- the 60 response values $y_1, y_2, \ldots, y_{60}$
- a variable indicating comparison group ($x_i = 1$ if knockout, $x_i = 0$ if wild-type, $i = 1, 2, \ldots, 60$)
- gender ($g_i = 1$ if female, $g_i = 0$ if male, $i = 1, 2, \ldots, 60$)
- average daily food consumption (f_i, a continuous number, standardized so that a value of 0 indicates average consumption, negative values indicate less than average, positive values indicate greater than average, $i = 1, 2, \ldots, 60$).

Based on your exploratory analysis of preliminary data, you and the PI agree that a sensible model is

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 g_i + \beta_3 f_i + \beta_4 (x_i \times g_i) + \epsilon_i,$$

where the ϵ_i are i.i.d. with mean 0 and s.d. σ.

1.) Interpret each of the coefficients in the above model.

2.) Using the following `R` output construct an approximate 95% confidence interval for $\beta_1 - \beta_4$?

Based on the interval you report, comment on your conclusions regarding whether there is a genotype effect.

`Hint`: If $\boldsymbol{v}$ is a column vector of length 5, then the standard error of $\boldsymbol{v}'\hat{\boldsymbol{\beta}}$ is equal to the square root of $\hat{\sigma}^2 \boldsymbol{v}'(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{v}$, where $\hat{\sigma}$ is the estimated residual s.d., and $\boldsymbol{X}$ is the model matrix. In our example, $\hat{\sigma} = 1.716$ and

$$(\boldsymbol{X}'\boldsymbol{X})^{-1} = \begin{pmatrix} 0.07 & -0.07 & -0.07 & -0.01 & 0.07 \\ -0.07 & 0.14 & 0.07 & 0.01 & -0.13 \\ -0.07 & 0.07 & 0.13 & 0.00 & -0.13 \\ -0.01 & 0.01 & 0.00 & 0.05 & 0.00 \\ 0.07 & -0.13 & -0.13 & 0.00 & 0.27 \end{pmatrix}$$

Here's the `R` output from fitting the above model:

```
Call:
lm(formula = y ~ 1 + x + g + f + x * g)

Residuals:
    Min      1Q  Median      3Q     Max
-4.3862 -1.0618 -0.0837  1.1362  4.2091

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  10.6375     0.4463  23.835  < 2e-16 ***
x             0.8728     0.6325   1.380  0.17319
g            -0.9642     0.6265  -1.539  0.12955
f             1.1632     0.3880   2.998  0.00407 **
x:g           0.0586     0.8867   0.066  0.94755
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 1.716 on 55 degrees of freedom
Multiple R-squared: 0.221,Adjusted R-squared: 0.1643
F-statistic:   3.9 on 4 and 55 DF,  p-value: 0.007369
```

3.) An F-test of the null hypothesis that $\beta_1 = \beta_4 = 0$ returned a p-value of 0.15. Meanwhile, stepwise model selection tells you that the best-fitting model is the one without the treatment-by-gender interaction, with abbreviated `R` output:

```
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  10.6230     0.3853  27.573  < 2e-16 ***
x             0.9018     0.4404   2.048  0.04535 *
g            -0.9349     0.4392  -2.128  0.03771 *
f             1.1622     0.3843   3.025  0.00375 **
```

The PI is primarily interested in whether there is sufficient evidence of an effect of treatment, the knockout variable, on time spent exercising in mice. Based on the F-test and stepwise regression results, how would you advise the PI? Provide a detail explanation of your advice.

QUESTION II.

Consider the simple linear regression model

$$y_i = \beta_0 + \beta_1 x_i + e_i, \qquad i = 1, 2, \ldots, n$$

Suppose we observe the following values for the y_i and x_i:

y	x
0.34	1
1.28	2
1.16	3
2.11	4
2.66	5

Recall that $(\hat{\beta}_0, \hat{\beta}_1)' = \hat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}$, where $\boldsymbol{X}$ is the model matrix and that $\text{Var}(\boldsymbol{v}'\boldsymbol{Y}) = \boldsymbol{v}'\text{Var}(\boldsymbol{Y})\boldsymbol{v}$, where $\boldsymbol{v}$ is a vector of constants and $\boldsymbol{Y}$ is a random vector.

1. Compute $\hat{\boldsymbol{\beta}}$.
2. Compute $\hat{\sigma}$.
3. Compute the standard errors of $\hat{\beta}_0$ and $\hat{\beta}_1$.
4. Compute a 95% confidence interval for the mean response when $x = 3$.
5. Compute a 95% prediction interval for a new observation when $x = 3$.

QUESTION IV.

Chapter 2 of Miller (2013) Modelling Techniques in Predictive Analytics, Pearson, New Jersey makes extensive use of multiple regression to analyze attendance figures for the 81 Dodgers home games in the 2012 Major League Baseball season. Data are available on the following variables

- Attendance = home game attendance (i.e., the number of tickets sold to each game)
- Month = month in which each game was played
- Day_of_week = day of the week each game was played on
- OpponentsFromLargeMetroAreas = a dummy variable which is 1 if the opponent is the New York Mets, Chicago Cubs and White Sox, Los Angeles Angels and the Washington DC Nationals
- Temp = temperature at the stadium during the game
- Day_night = day (for day games) and night (for night games)
- BobbleheadPromotion = a dummvariable which is 1 if the game involved a bobblehead promotion

According to Miller (2013):

"Dodger Stadium, with a capacity of 56,000, is the largest ballpark in the world. From the data, we can see that Dodger Stadium was filled to capacity only twice in 2012. ... The eleven bobblehead promotions occurred on night games, six of those being Tuesday nights. ... Opponents from the large metropolitan areas (the New York Mets, Chicago Cubs and White Sox, Los Angeles Angels and Washington D.C. Nationals) are consistently associated with higher attendance. ... Explanatory graphics help us find models that might work for predicting attendance and for evaluating the effect of promotions on attendance.

Figure 2.1 shows distributions of attendance across days of the week, and

Figure 2.2 shows attendance by month.

To advise management regarding (bobblehead) promotions, we would like to know if promotions have a positive effect upon attendance, and if they do have a positive effect, how much that effect might be. To provide this advice we build a linear model for predicting attendance using month, day of the week and an indicator variable for the bobblehead promotion ... "

Given on the next few pages are Figures 2.1 and 2.2, JMP output from the least squares model fit by Miller (2013) and some additional plots. A statistics professor originally from Australia has taken a careful look at the data and found among other things that there is no evidence of significant autocorrelation in the attendance results. In other words, there is no evidence that attendance at home games on day t is statistically significantly related to attendance on days $t-1,\ t-2,\ \ldots$.

1. Describe in detail one major concern that potentially threatens the validity of the model fit by Miller (2013).

2. Explain the specific steps you would take to overcome the problem described in part (1).

3. On the basis of the plots presented what predictors would you recommend being included in the model you describe in part (2).

QUESTION IV.

Consider the usual linear regression model, written either in non-matrix notation as:

$$y_i = \beta_o + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik} + e_i, \quad i = 1, 2, \ldots, n, \tag{A}$$

where $e_1, e_2, \cdots, e_n$ are independently and identically distributed as $N(0,\ \sigma^2)$ random variables or, in matrix notation as:

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e}, \tag{B}$$

where $\mathbf{y}$ is an $n \times 1$ vector of response variables, $\mathbf{X}$ is an $n \times p$ matrix of predictor variables (with $p = k + 1$), $\boldsymbol{\beta}$ is a $p \times 1$ vector of unknown parameters and $\mathbf{e}$ is an $n \times 1$ vector of unobservable independent and identically distributed random $N(0, \sigma^2)$ variables. In what follows, you may assume that the matrix $\mathbf{X}$ is of full column rank. Using whichever notation above (A or B) that makes you more comfortable, answer the following parts to this problem. Please be concise with your answers - highly irrelevant statements may be counted against you!

1. The above model is called a **linear** regression model even though, for example, it encompasses polynomial (in the predictor variables) regression models. Explain what is **linear** about the above model.

2. Define the **least squares criterion**. That is, state what property must be satisfied for estimates of the unknown parameters of the above model to be called **least squares estimates**. Use formulas as part of your definition.

3. Specify which of the above assumptions made about the e_i's, $i = 1, 2, \cdots, n$, need **not** be true for the least squares estimators of the β_j's, $j = 0, 1, \cdots, k$, to be unbiased estimators. If all the above assumptions about the e_i's need to be true, then state so.

4. Specify which of the above assumptions made about the e_i's, $i = 1, 2, \cdots, n$, need **not** be true for the least squares estimator of σ^2 to be an unbiased estimator. If all the above assumptions about the e_i's need to be true, then state so.

5. Specify which of the above assumptions made about the e_i's, $i = 1, 2, \cdots, n$, need **not** be true for the usual least squares t tests and F tests of hypotheses about the β_j's, $j = 0, 1, \cdots, k$, to be statistically valid. If all the above assumptions about the e_i's need to be true, then state so.

Figure 2.1

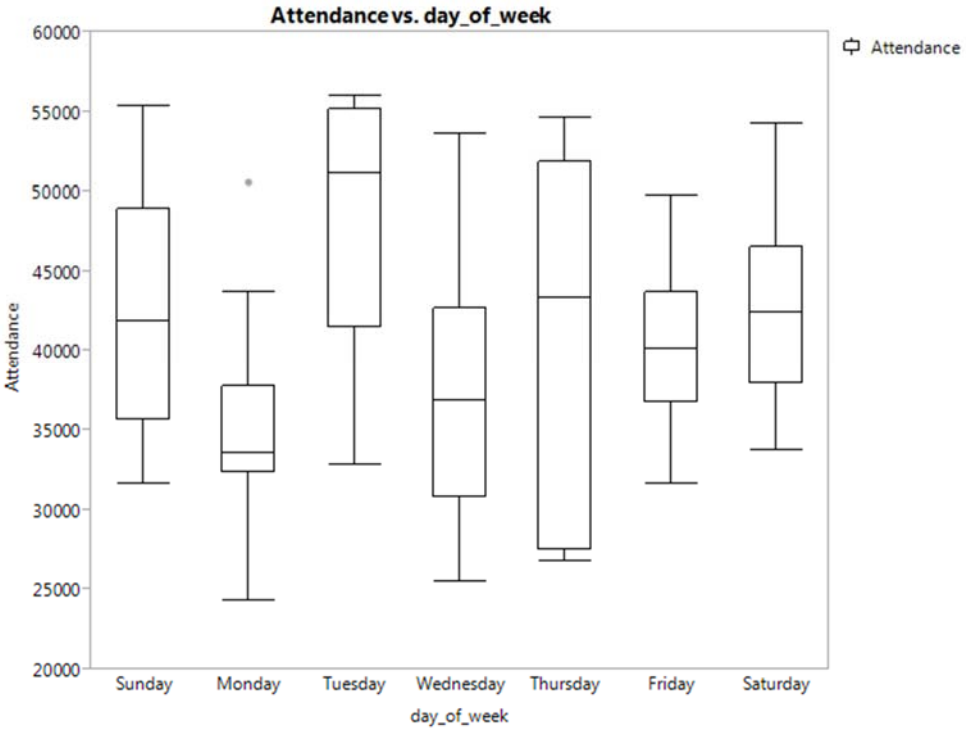
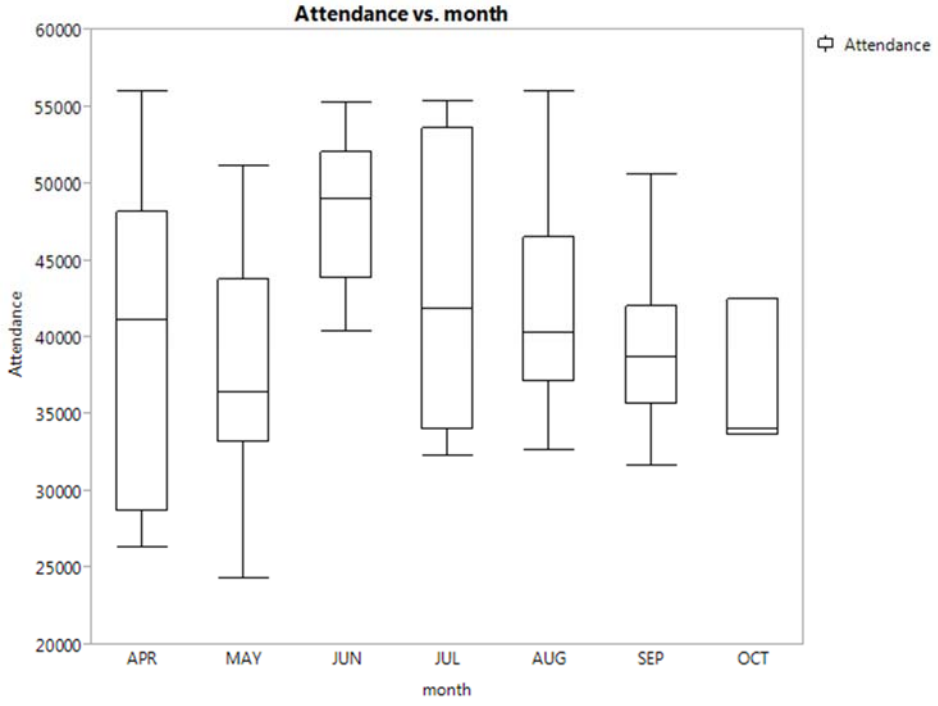
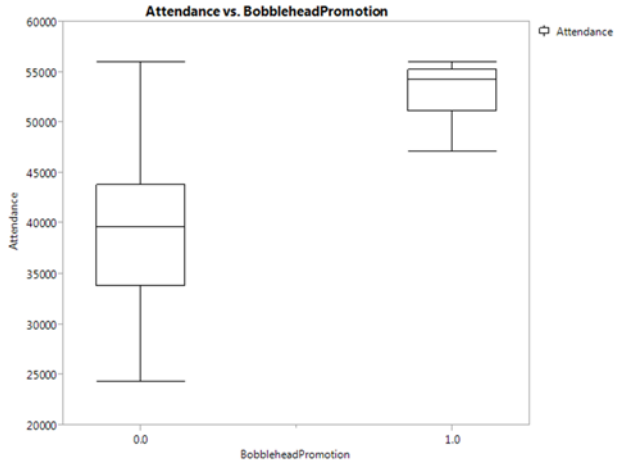
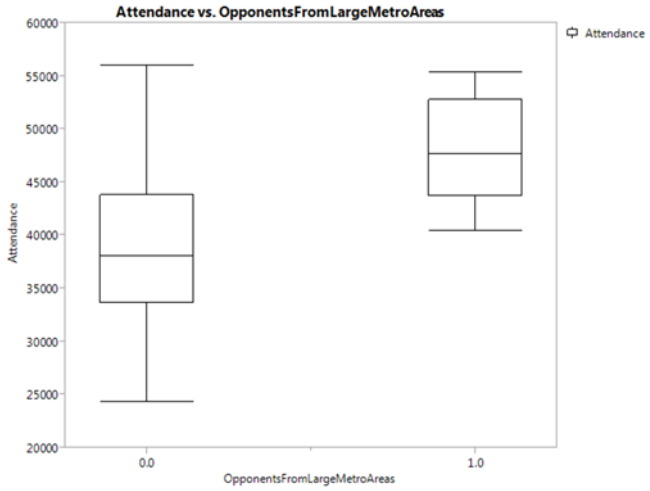


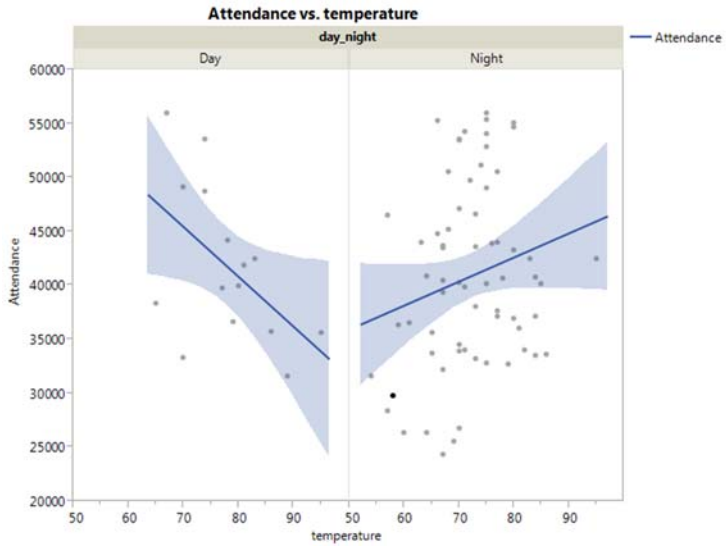
Figure 2.2



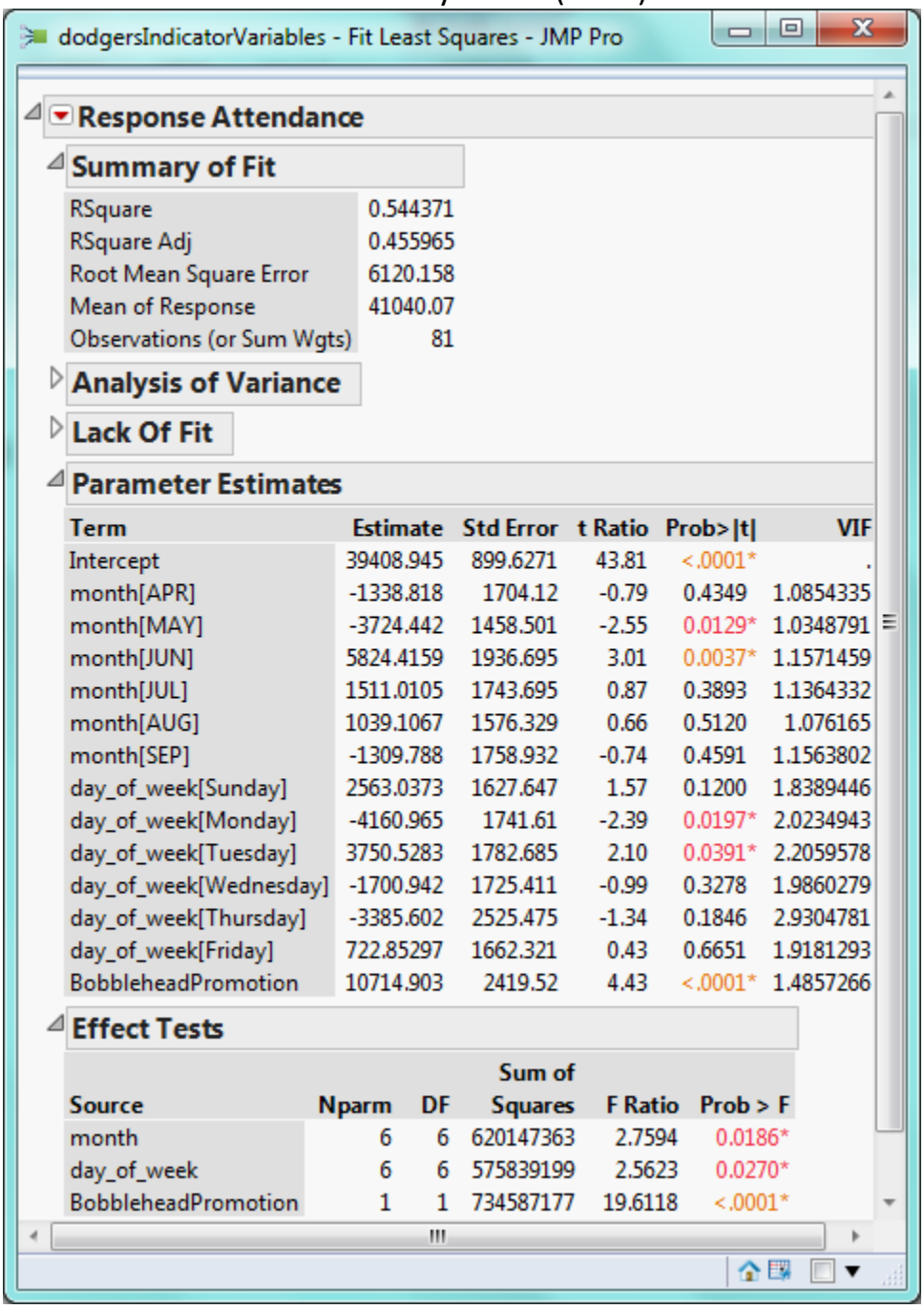
Other plots not in Miller (2013)







Model fit by Miller (2013)



dodgersIndicatorVariables - Fit Least Squares - JMP Pro

Response Attendance

Summary of Fit

RSquare	0.544371
RSquare Adj	0.455965
Root Mean Square Error	6120.158
Mean of Response	41040.07
Observations (or Sum Wgts)	81

Analysis of Variance

Lack Of Fit

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|	VIF
Intercept	39408.945	899.6271	43.81	<.0001*	.
month[APR]	-1338.818	1704.12	-0.79	0.4349	1.0854335
month[MAY]	-3724.442	1458.501	-2.55	0.0129*	1.0348791
month[JUN]	5824.4159	1936.695	3.01	0.0037*	1.1571459
month[JUL]	1511.0105	1743.695	0.87	0.3893	1.1364332
month[AUG]	1039.1067	1576.329	0.66	0.5120	1.076165
month[SEP]	-1309.788	1758.932	-0.74	0.4591	1.1563802
day_of_week[Sunday]	2563.0373	1627.647	1.57	0.1200	1.8389446
day_of_week[Monday]	-4160.965	1741.61	-2.39	0.0197*	2.0234943
day_of_week[Tuesday]	3750.5283	1782.685	2.10	0.0391*	2.2059578
day_of_week[Wednesday]	-1700.942	1725.411	-0.99	0.3278	1.9860279
day_of_week[Thursday]	-3385.602	2525.475	-1.34	0.1846	2.9304781
day_of_week[Friday]	722.85297	1662.321	0.43	0.6651	1.9181293
BobbleheadPromotion	10714.903	2419.52	4.43	<.0001*	1.4857266

Effect Tests

Source	Nparm	DF	Sum of Squares	F Ratio	Prob > F
month	6	6	620147363	2.7594	0.0186*
day_of_week	6	6	575839199	2.5623	0.0270*
BobbleheadPromotion	1	1	734587177	19.6118	<.0001*

QUESTION III.

Consider a regression model for an experiment with a continuous response Y, a treatment factor, A, with three levels, A_1, A_2, and A_3, and a continuous explanatory variable, X. Define the dummy variables,

$D_1 = 1$ if $A = A_1$ and $D_1 = 0$, otherwise,

$D_2 = 1$ if $A = A_2$ and $D_2 = 0$, otherwise.

Consider the regression model

$$E(Y) = \beta_0 + \beta_1 D_1 + \beta_2 D_2 + \beta_3 X + \beta_4 D_1 X + \beta_5 D_2 X$$

1. Obtain expressions for the mean response for each of the three treatments:
   a. A_1
   b. A_2
   c. A_3

2. Write out interpretations in terms of the mean response and its relationship to the treatments and explanatory variable for each of the parameters:
   a. β_0
   b. β_1
   c. β_2
   d. β_3
   e. β_4
   f. β_5

3. Formulate the hypotheses for a test of equal slopes for the three treatments. Explain how to carry out the test if you were provided statistical software that could fit a regression model and provide the usual analysis of variance table for the fitted regression model.

4. Does testing the hypothesis $H_0 : \beta_1 = \beta_2 = 0$ provide a test of equal effects for the three treatments in the above model? Justify your answer.

QUESTION II.

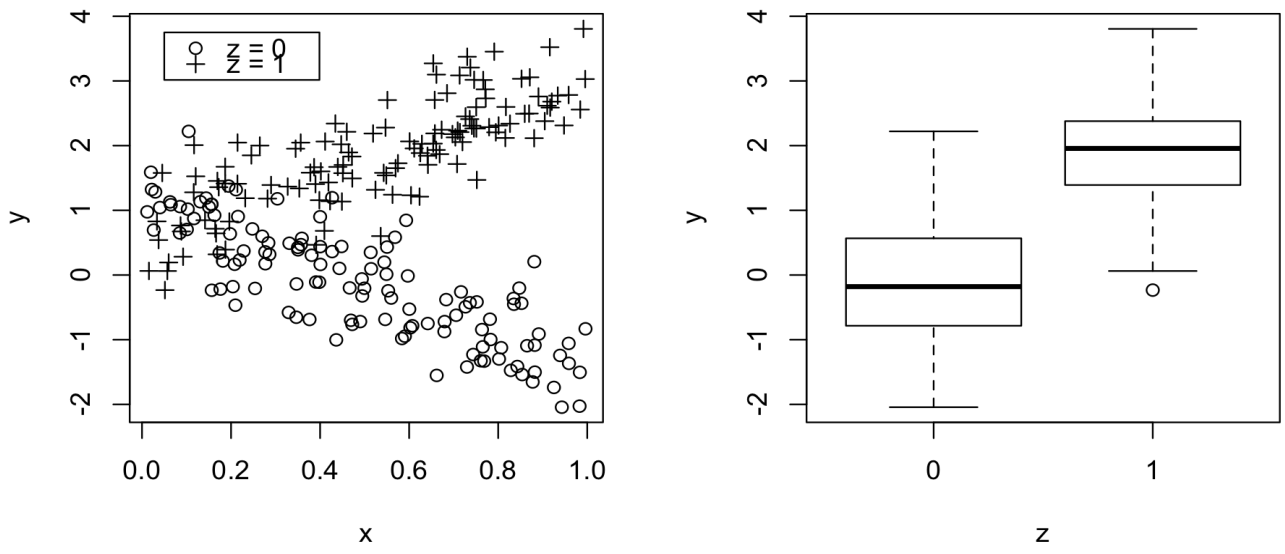
Consider the following linear model:

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 z_i + \beta_3 (x_i \times z_i) + \epsilon_i$$

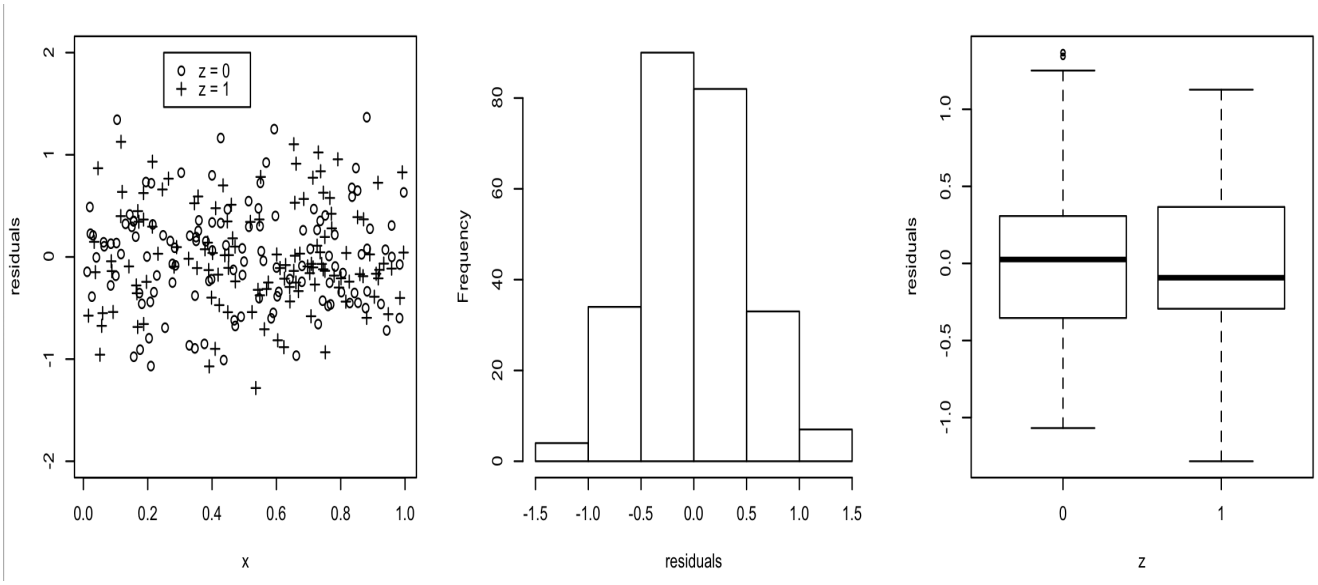
where x is continuous and z is binary. The output from fitting this model to a sample of size $n = 250$ is shown below:

```
Call:
lm(formula = y ~ x + z + x * z)

Residuals:
     Min       1Q   Median       3Q      Max
-1.28335 -0.34073 -0.04031  0.33622  1.36754

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.15235    0.08933  12.900  < 2e-16 ***
x           -2.62637    0.16003 -16.412  < 2e-16 ***
z           -0.55213    0.13378  -4.127 5.03e-05 ***
x:z          5.02318    0.23094  21.751  < 2e-16 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 0.5035 on 246 degrees of freedom
Multiple R-squared: 0.8554,Adjusted R-squared: 0.8536
F-statistic: 485.1 on 3 and 246 DF,  p-value: < 2.2e-16
```

1. Interpret each of the model coefficients (β_0, β_1, β_2, β_3) in terms of expected values.
2. Report a 95% confidence interval for the mean change in y associated with a one-unit increase in x, when $z = 0$.
3. What is the slope parameter (mean change in y associated with a one-unit increase in x) for x when $z = 1$?
4. What does the adjusted R squared measure?
5. What null and alternative hypotheses are tested using the F statistic at the bottom of the `R` output?
6. The figures below show diagnostic plots for the above model. Which of your model assumptions, if any, appear to not be met? And why?
7. One of the assumptions of our model is that the ϵ_i are *i.i.d.* realizations from the Normal distribution with mean 0 and constant variance σ^2. What does the model report as an estimate of σ?



Scatterplot of y vs. x, and side-by-side boxplots comparing y to z.



Residual plots: scatterplot of residuals vs. x; histogram of residuals; and side-by-side boxplots of residuals vs. z

QUESTION II.

1. Suppose we have n independent observations of the pair (x_i, y_i), $i = 1, 2, \ldots, n$, where y_i is binary, taking only the values 0 or 1, and x_i is continuous. Provide two reasons why linear regression of y on x, with the usual linear regression assumptions, would not be appropriate as an analysis method.

2. In a study of a particular disease in humans, researchers observed the following data.

	Female	Male	Total
Diseased	20	19	39
Healthy	20	41	61
Total	40	60	100

Let y_i be disease status (1 if diseased, 0 if healthy) and x_i the gender (1 if female, 0 if male) of the ith individual, $i = 1, 2, \ldots, 100$. Let $\pi(x)$ be the conditional probability of an individual being diseased, given its gender. Consider the logistic regression model:

$$\log\left(\frac{\pi(x_i)}{1-\pi(x_i)}\right) = \beta_0 + \beta_1 x_i.$$

(a.) Write down the likelihood function $L(\boldsymbol{\beta})$, where $\boldsymbol{\beta}' = (\beta_0, \beta_1)$.

(b.) Report the maximum likelihood estimates $\hat{\beta}_0$ and $\hat{\beta}_1$.

(c.) One way to test $H_0 : \beta_1 = 0$ is with a likelihood ratio test. The test statistic equals $-2\log\left[\frac{L_0(\hat{\beta}_0^0)}{L(\hat{\boldsymbol{\beta}})}\right]$, where L_0 is the likelihood function for the null model, the model with only the intercept, β_0^0, included:

$$\log\left(\frac{\pi_0(x_i)}{1-\pi_0(x_i)}\right) = \beta_0^0,$$

and where $\hat{\beta}_0^0$ and $\hat{\boldsymbol{\beta}}$ are the MLEs under the null and unrestricted models, respectively. Under H_0, the statistic is approximately chi-square distributed with 1 degree of freedom. Test $H_0 : \beta_1 = 0$. Some chi-square cumulative probabilities are provided.

3. A similar study to the one described above was conducted, in which the independent variable was race:

	White	Black	Hispanic	Other	Total
Diseased	5	20	15	10	50
Healthy	20	10	10	10	50
Total	25	30	25	20	100

Consider the logistic regression model

$$\log\left(\frac{\pi(r_i)}{1-\pi(r_i)}\right) = \beta_0 + \beta_1 r_{1i} + \beta_2 r_{2i} + \beta_3 r_{3i},$$

where the r_1, r_2, r_3 independent variables code for race as follows:

	r_1	r_2	r_3
White	0	0	0
Black	1	0	0
Hispanic	0	1	0
Other	0	0	1

Here is partial output from the model fit using `R` code:

```
             Estimate  Std. Error  z value    Pr(>|z|)
(Intercept)  -1.3863      0.5000   -2.773     0.00556 **
  r1          2.0794      0.6325    3.288     0.00101 **
  r2          1.7918      0.6455    2.776     0.00551 **
  r3          1.3863      0.6708    2.067     0.03878 *
```

(a.) According to the model, what is the estimated probability of a white individual being diseased? Show your calculations, in terms of the model coefficients.

(b.) Report a 95% confidence interval for the odds ratio (not *log* odds ratio) comparing Hispanic to White.

PROBLEM IV.

The World Health Organization defines low birth weight as a baby weighing less than 2500 grams at birth. This data was taken from Stat Labs: Mathematical Statistics Through Applications by Deb Nolan and Terry Speed, University of California, Berkeley. This data is available at http://www.statsci.org/datasets.html.

We want to determine if gestation, age, weight, height and the mother smoking can be use to predict low birth weight.

Variable	Description
bwt	1 if less the 2500 grams ; 0 otherwise
Gestation	Length of pregnancy in days
age	mother's age in years
height	mother's height in inches
weight	Mother's prepregnancy weight in pounds
smoke	Smoking status of mother 0=not now, 1=yes now

1. Since the dependent variable (low birth weight – bwt) is a one (1) or a zero (0), multiple linear regression is not appropriate. Please give some reasons why multiple linear regression is not appropriate.

PROBLEM III:

1. A researcher, who has run a simple linear regression through the origin, states "My residuals do not sum to 0. Thus my model assumptions have been violated. I must transform Y or X." Is this a correct statement? Must the residuals sum to 0? Please explain your answer.

2. A researcher believes that the log of the odds is a linear combination of x, x*x and log x. Write out the logit function.

3. Carefully study the results for the three models given below. What model would you recommend to your client? Explain why. Discuss the issues with validity of the models. Is there anything you can tell the client about his predictor variables?

QUESTION III

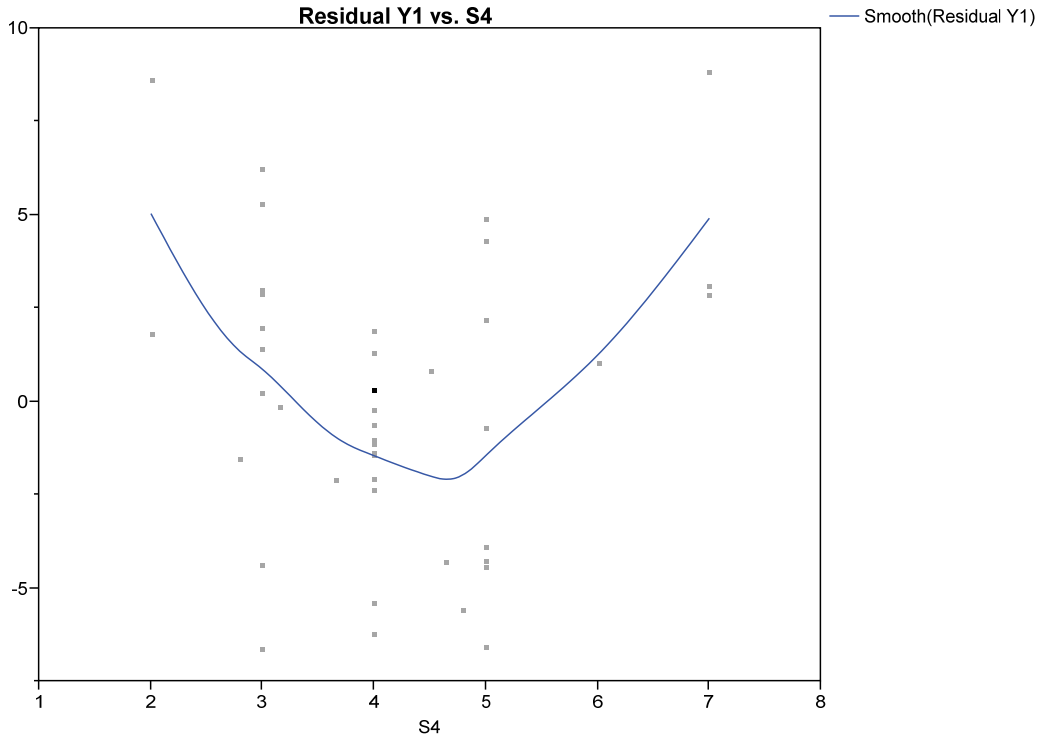
Part 1:

We have a dataset with y1 (a measure of diabetes & dependent variable) and 3 predictors BMI (Body mass index); BP (Blood pressure) and s4 a measure of glucose. We begin by considering the classical linear model relating y1 to the three explanatory variables.

Given the graph below (Residual is the "raw" residual), answer the following questions:

1. Based on the graph can you determine if the residuals are normally distributed? Explain.
2. Based on the graph below can you conclude that this is a valid model? Explain.
3. Does the graph suggest a change in the model? (That is: should you add a term; transform y1 etc.). Explain.

Graph Builder



Part 2:

Suppose that we want to predict whether a person will get the Tourette Syndrome based on Gender, HDL, Glucose & Chol. Note: Let Y = 1 if the Syndrome is detected and Y = 0 if not.

A logistic regression was run and the following output was obtained:

Gender	M	F
HDL	40	40
Glucose	125	125
Chol	190	190
Prob Y = 1	0.001	0.0001

1. What are the odds that a male with those characteristics will get the Syndrome? **Justify your answer.**
2. What are the odds that a female with those characteristics will get the Syndrome? **Justify your Answer.**
3. What is the odds ratio of a male getting the Syndrome as opposed to a female? **Justify your answer.**
4. The local newspaper front page story is: "Males are 10.01 times more likely to get Tourette Syndrome than Females. More research money is needed to help the males…"
   Is there something other than the odds ratio that needs to be considered here? Please explain your answer.

Part 3.

This part is similar to Part 1. Here we have y (a measure of diabetes & dependent variable) and 10 predictors: Age, Sex, BMI,BP, s1,s2,s3,s4,s5,s6.

A stepwise regression was run with the following results:

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	-483.9509	71.29641	-6.79	<.0001*
SEX	-31.32679	16.06386	-1.95	0.0586
BP	1.7477267	0.568201	3.08	0.0039*
S2	0.6081179	0.320221	1.90	0.0652
S5	52.425839	16.93224	3.10	0.0037*
S6	2.11306	0.72281	2.92	0.0058*

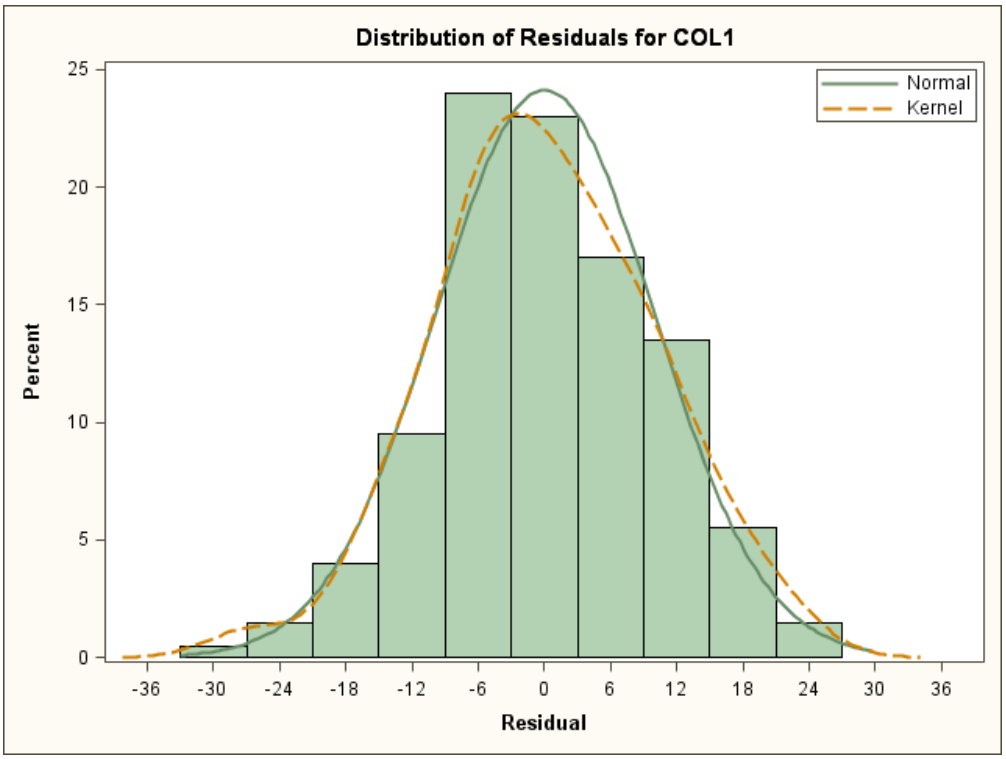
Since SEX and S2 are not significant at alpha= .05, can we remove them both from the model at the same time? Explain.

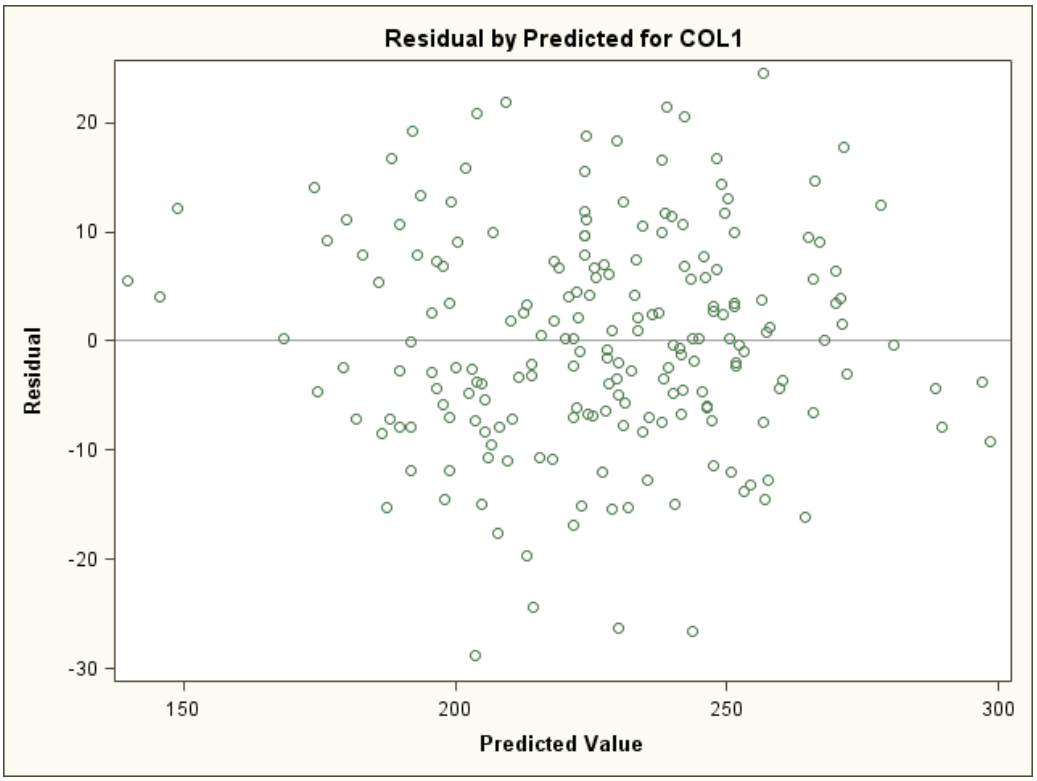
Model 1:

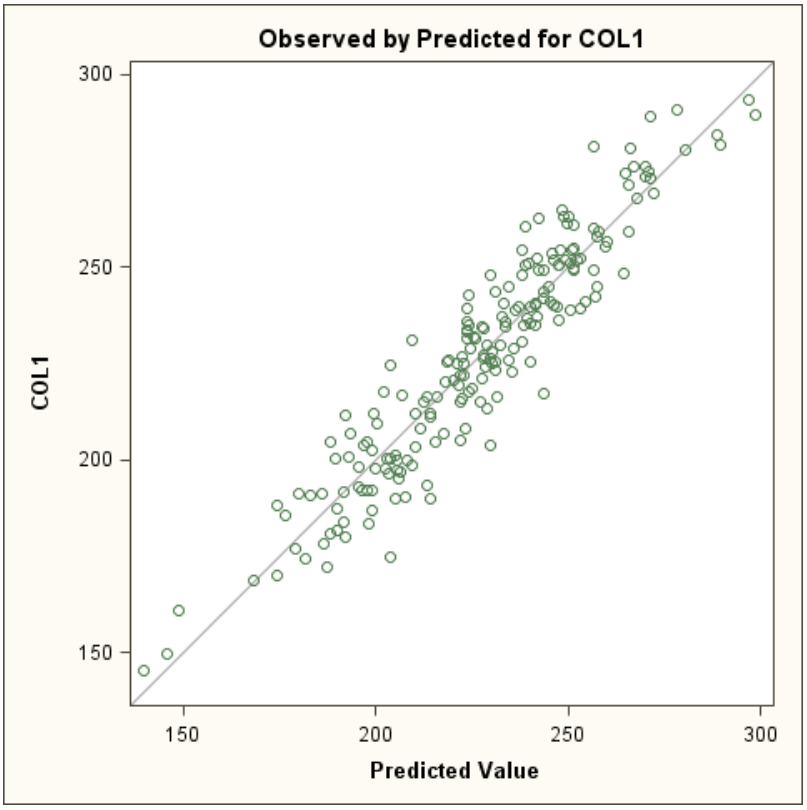
Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	9	154584	17176	166.62	<.0001
Error	190	19586	103.08628		
Corrected Total	199	174171			

Root MSE	10.15314	R-Square	0.8875
Dependent Mean	227.04725	Adj R-Sq	0.8822
Coeff Var	4.47182		

Parameter Estimates						
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variance Inflation
Intercept	1	3.81085	6.13929	0.62	0.5355	0
COL2	1	0.93090	1.53281	0.61	0.5444	12.23132
COL3	1	2.46945	0.90784	2.72	0.0071	4.27408
COL4	1	3.41763	1.17315	2.91	0.0040	8.77643
COL5	1	3.68754	0.42996	8.58	<.0001	1.05793
COL6	1	4.71660	0.41765	11.29	<.0001	1.06188
COL7	1	6.02569	0.42912	14.04	<.0001	1.03296
COL8	1	7.19138	0.45166	15.92	<.0001	1.06395
COL9	1	7.95603	0.44949	17.70	<.0001	1.05266
COL10	1	4.39536	1.49354	2.94	0.0037	19.06114




Residual by Predicted for COL1
Residual
20
10
0
-10
-20
-30
150
200
250
300
Predicted Value

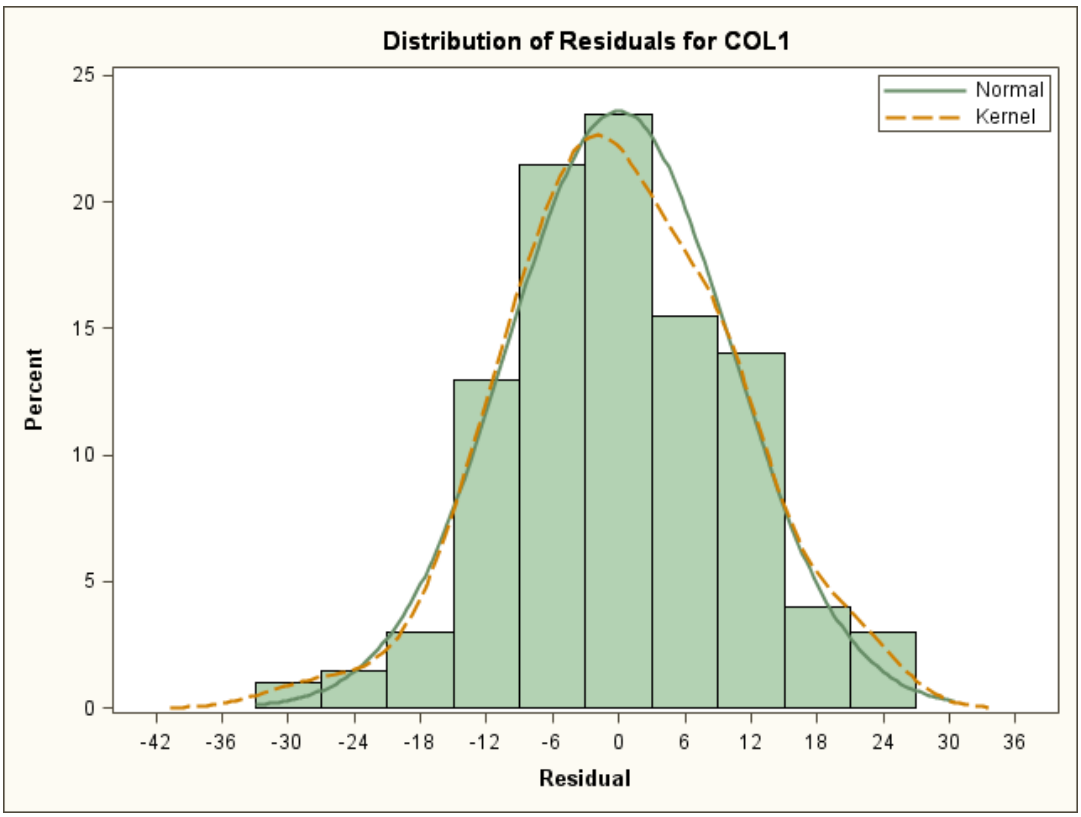

Observed by Predicted for COL1
COL1
300
250
200
150
150
200
250
300
Predicted Value

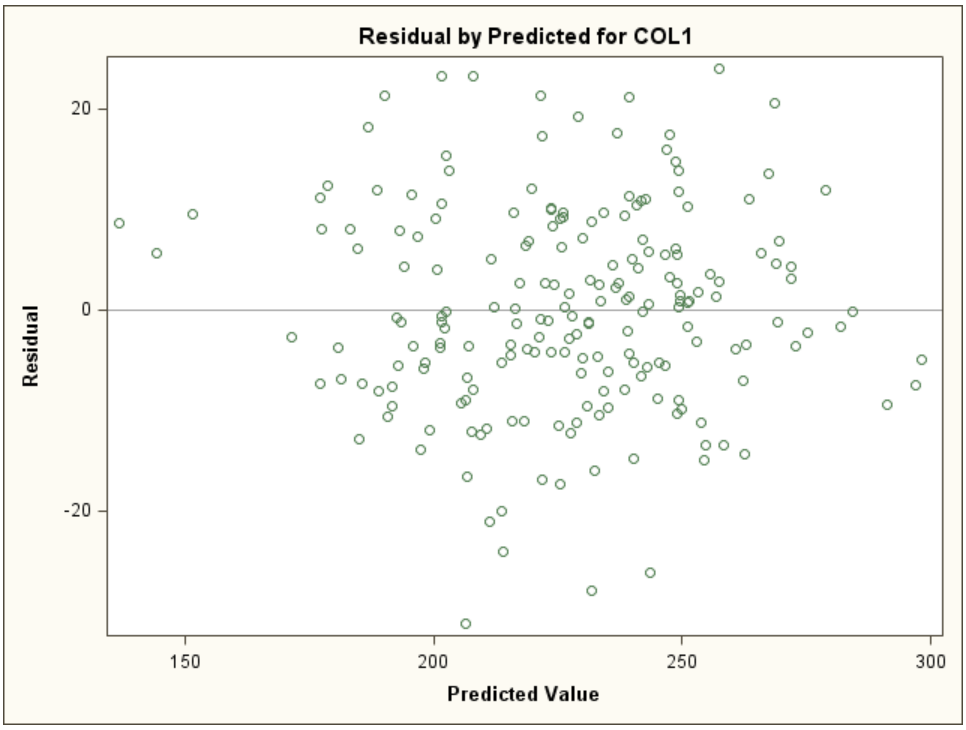
Model 2

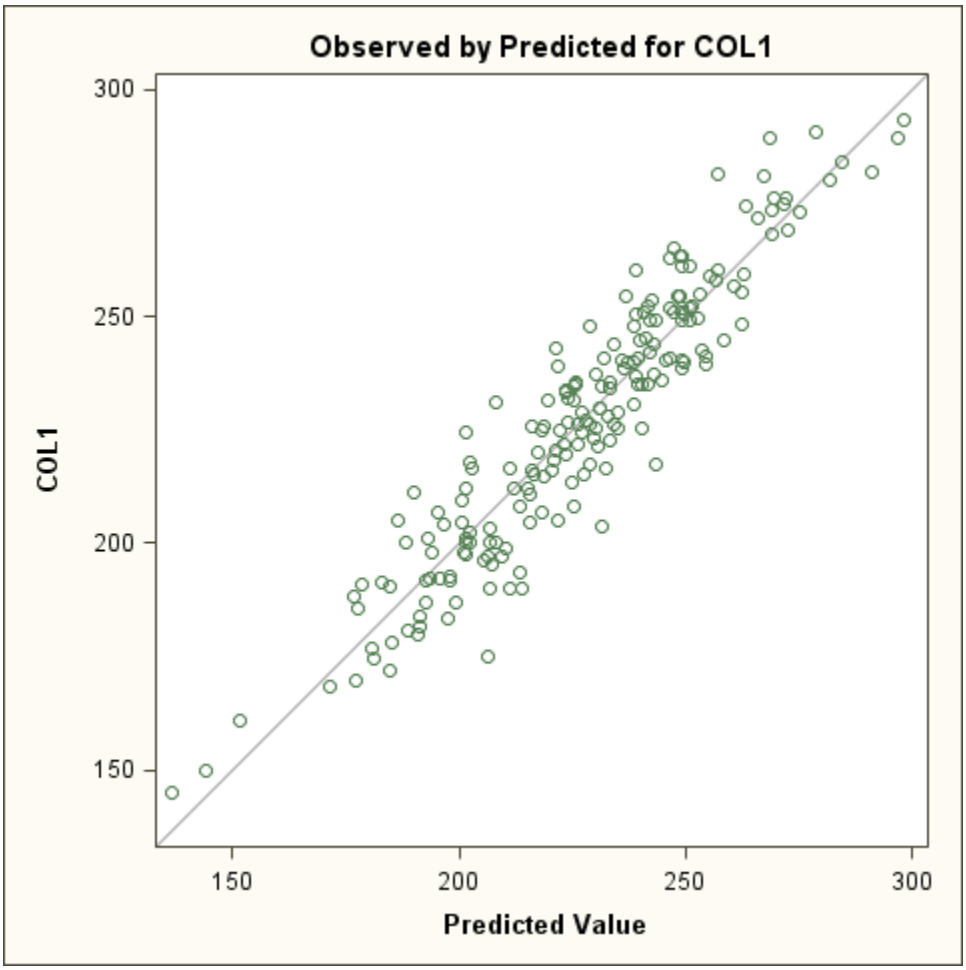
Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	8	153692	19211	179.18	<.0001
Error	191	20479	107.22094		
Corrected Total	199	174171			

Root MSE	10.35475	R-Square	0.8824
Dependent Mean	227.04725	Adj R-Sq	0.8775
Coeff Var	4.56062		

Parameter Estimates						
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variance Inflation
Intercept	1	4.08374	6.26048	0.65	0.5150	0
COL2	1	5.24085	0.46140	11.36	<.0001	1.06556
COL3	1	4.79871	0.45351	10.58	<.0001	1.02544
COL4	1	6.65044	0.41997	15.84	<.0001	1.08134
COL5	1	3.60344	0.43753	8.24	<.0001	1.05326
COL6	1	4.57544	0.42313	10.81	<.0001	1.04787
COL7	1	6.08528	0.43715	13.92	<.0001	1.03066
COL8	1	7.25810	0.46005	15.78	<.0001	1.06127
COL9	1	7.96753	0.45840	17.38	<.0001	1.05259




Residual by Predicted for COL1
Residual
20
0
-20
150
200
250
300
Predicted Value

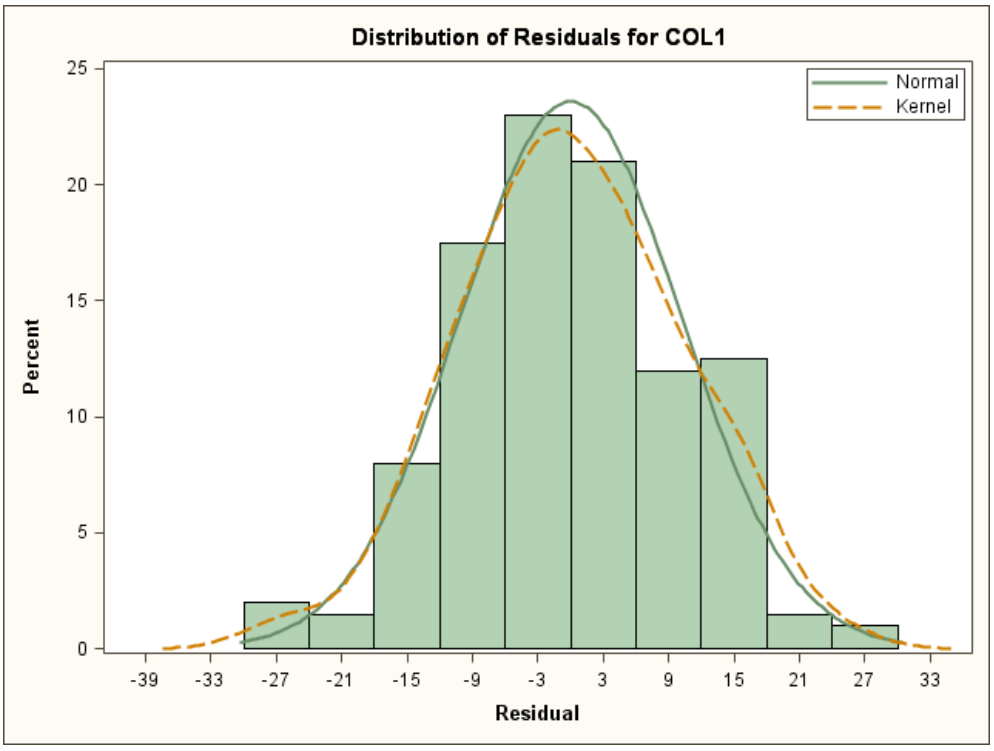

Observed by Predicted for COL1
COL1
300
250
200
150
150
200
250
300
Predicted Value

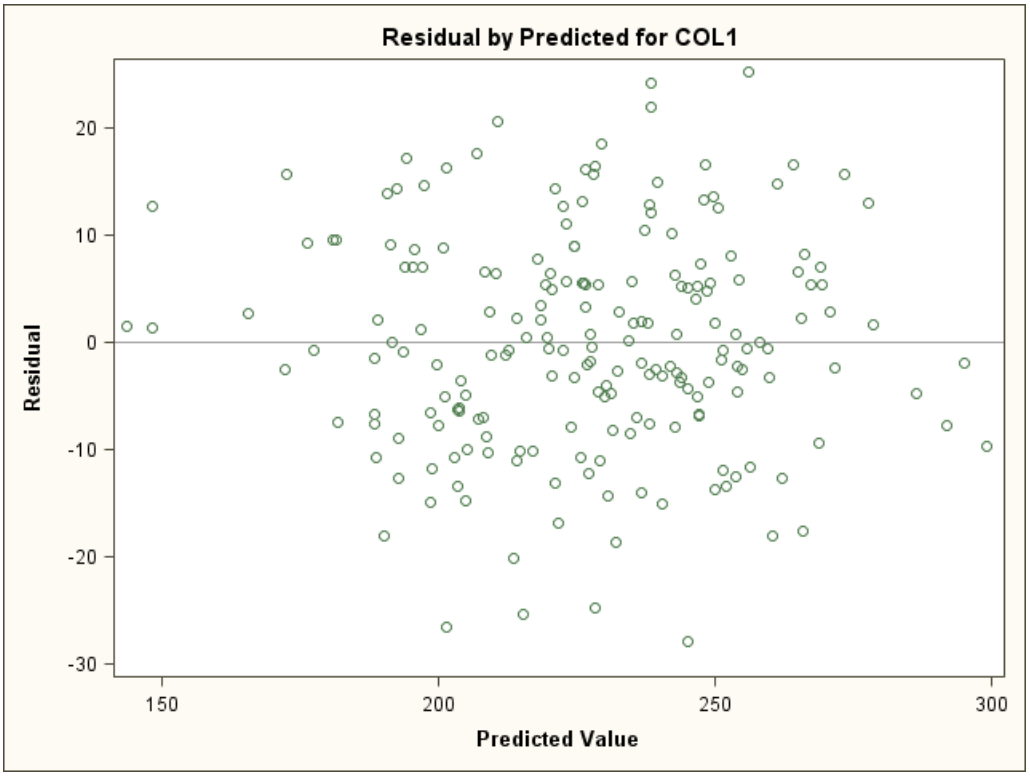
Model 3 – Stepwise

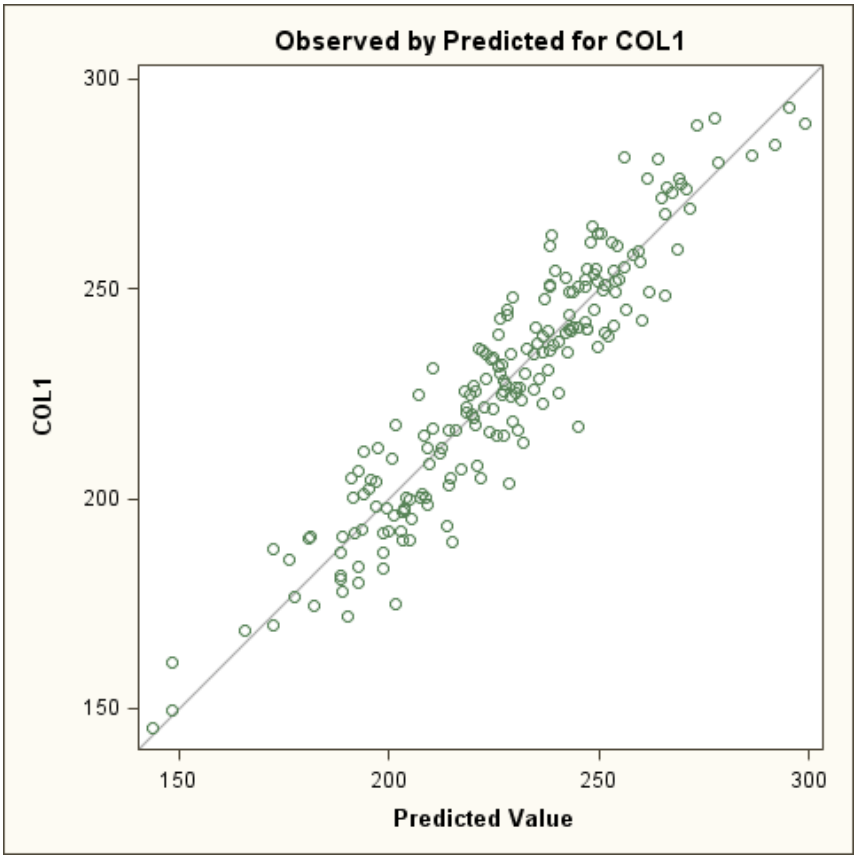
Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	7	153676	21954	205.66	<.0001
Error	192	20495	106.74650		
Corrected Total	199	174171			

Root MSE	10.33182	R-Square	0.8823
Dependent Mean	227.04725	Adj R-Sq	0.8780
Coeff Var	4.55052		

Parameter Estimates						
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variance Inflation
Intercept	1	5.19190	6.22902	0.83	0.4056	0
COL2	1	-3.29178	0.58148	-5.66	<.0001	1.69988
COL5	1	3.73159	0.43698	8.54	<.0001	1.05532
COL6	1	4.88690	0.41914	11.66	<.0001	1.03276
COL7	1	5.97025	0.43561	13.71	<.0001	1.02796
COL8	1	7.05863	0.45679	15.45	<.0001	1.05090
COL9	1	7.98222	0.45577	17.51	<.0001	1.04517
COL10	1	8.62440	0.45761	18.85	<.0001	1.72803




Residual by Predicted for COL1
Residual
20
10
0
-10
-20
-30
150
200
250
300
Predicted Value


Observed by Predicted for COL1
COL1
300
250
200
150
150
200
250
300
Predicted Value

2. What is the logit equation for the model that predicts low birth weight form the predictors given above?

3. Given the graphics on pages 10, 11, 12, and 13; the researcher decided to use a log transformation on the predictor variables. Does this seem reasonable? Please explain you answer.

4. Suppose that your model for low birth weight contains the following predictor variables:

smoke
LOG_GEST
LOG_AGE
LOG_HEIGHT
LOG_WEIGHT

Given the table below the researcher wants to delete all three of the non-significant variables all at once. Is this the correct approach? Explain your answer.

Analysis of Maximum Likelihood Estimates						
Parameter		DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept		1	109.0	18.7220	33.8971	<.0001
smoke	0	1	-1.2378	0.3058	16.3821	<.0001
LOG_GEST		1	-18.4168	2.1783	71.4834	<.0001
LOG_AGE		1	0.8558	0.6826	1.5719	0.2099
LOG_HEIGHT		1	-2.1324	3.9408	0.2928	0.5884
LOG_WEIGHT		1	-0.4369	1.0929	0.1598	0.6893

5. Suppose that your model for low birth weight has just smoke and **LOG_GEST** as predictors.

Page 14 has the marginal model plots. Do these indicate a valid model? What transformation would you suggest?

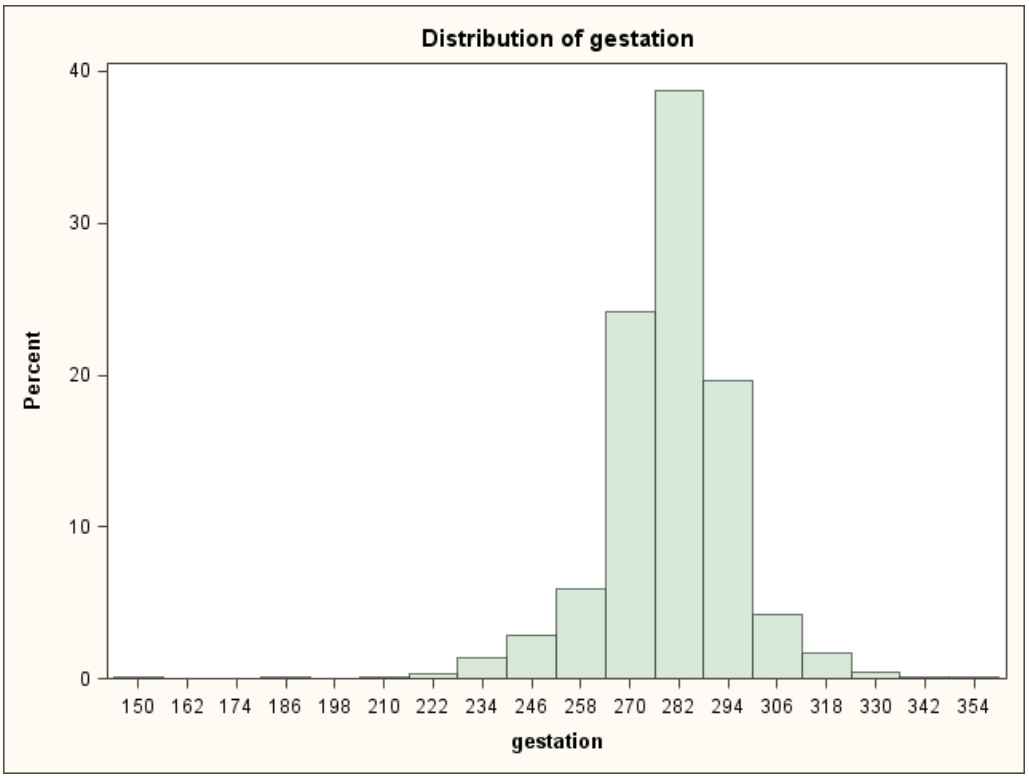
6. After an appropriate transformation, the marginal model plots on page 15 were obtained. Do these indicate a valid model? Explain your answer.

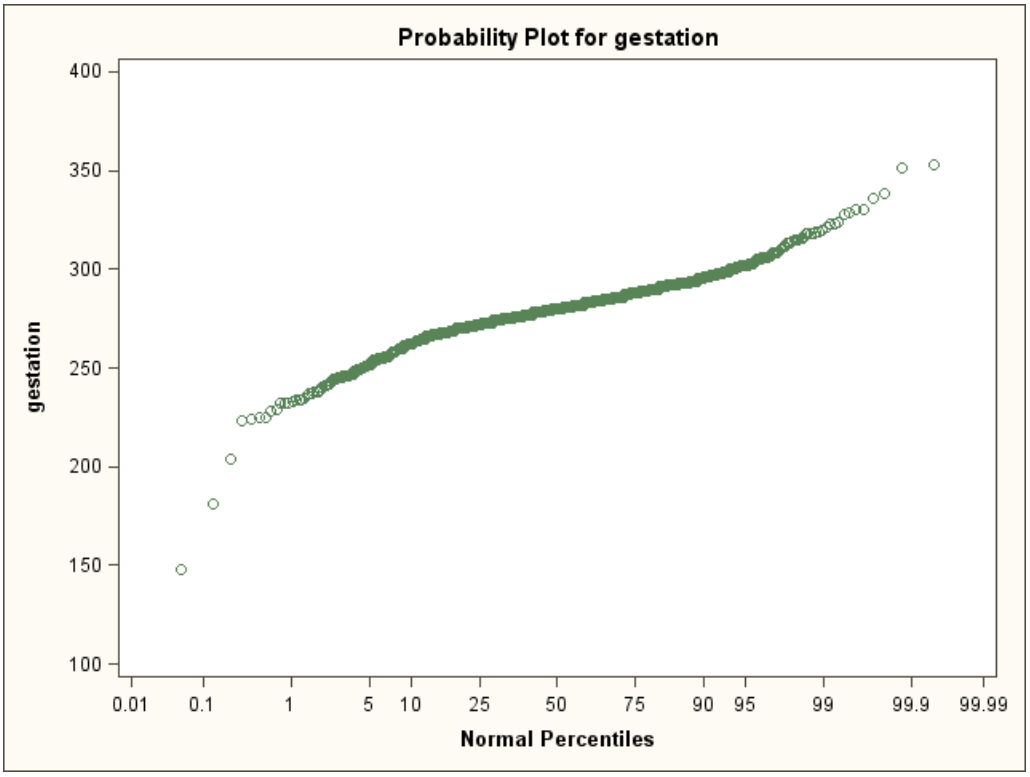
7. If the probability of a low birth weight baby is .437 for a log gestation of 5.5 for a woman who smokes and the probability of a low birth weight baby is .204 for a log gestation of 5.5 for a woman who does not smokes, what is the odds ratio of getting a low birth weight baby?

8. If the probability of a low birth weight baby is .559 for a log gestation of 5.48 for a woman who smokes and the probability of a low birth weight baby is .295 for a log gestation of 5.48 for a woman who does not smokes, what is the odds ratio of getting a low birth weight baby?

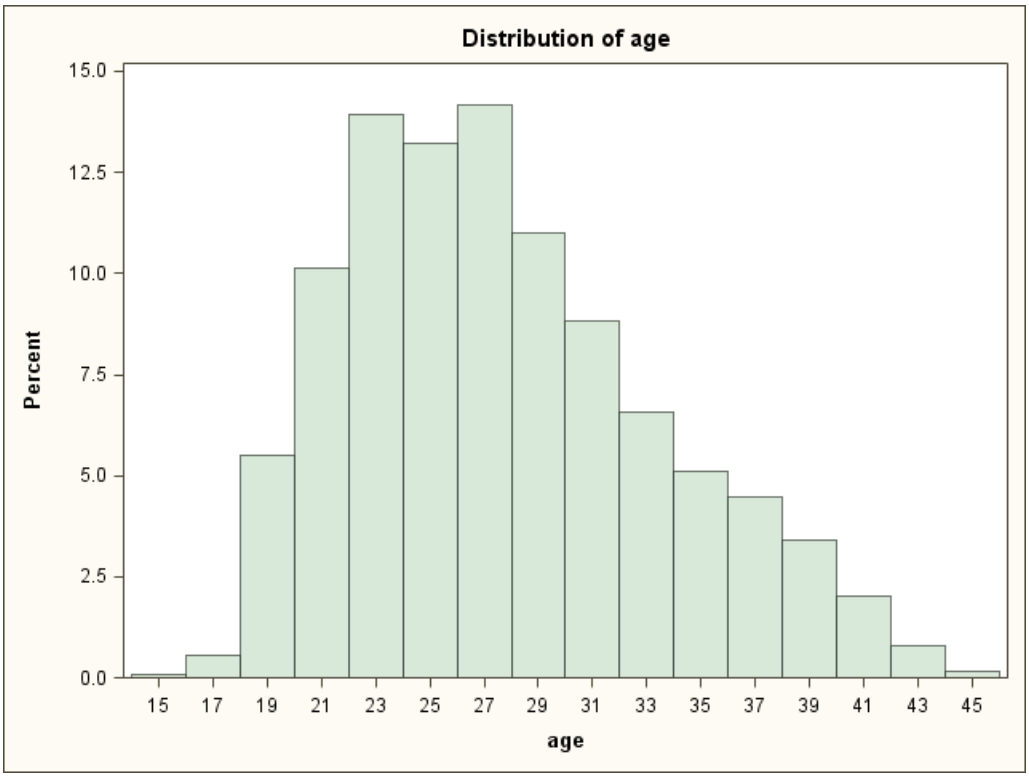
9. A researcher found that the odds ratio in parts 7 and part 8 are identical. How can you explain this phenomenon?

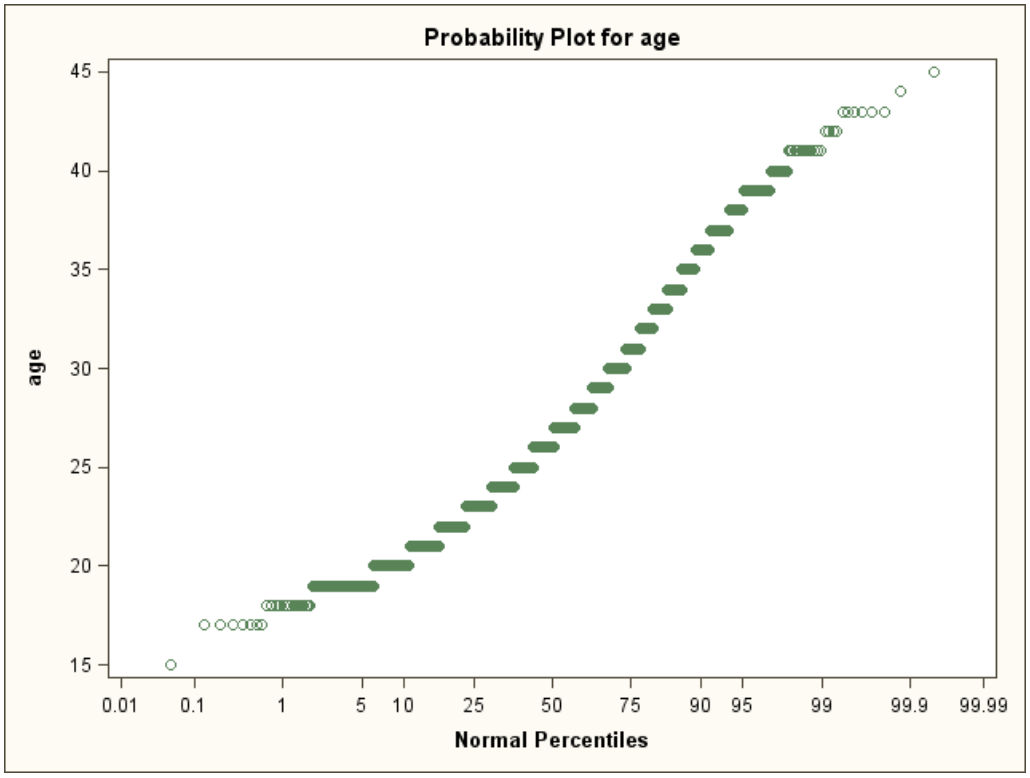
GESTATION:


Distribution of gestation
Percent
40
30
20
10
0
150 162 174 186 198 210 222 234 246 258 270 282 294 306 318 330 342 354
gestation

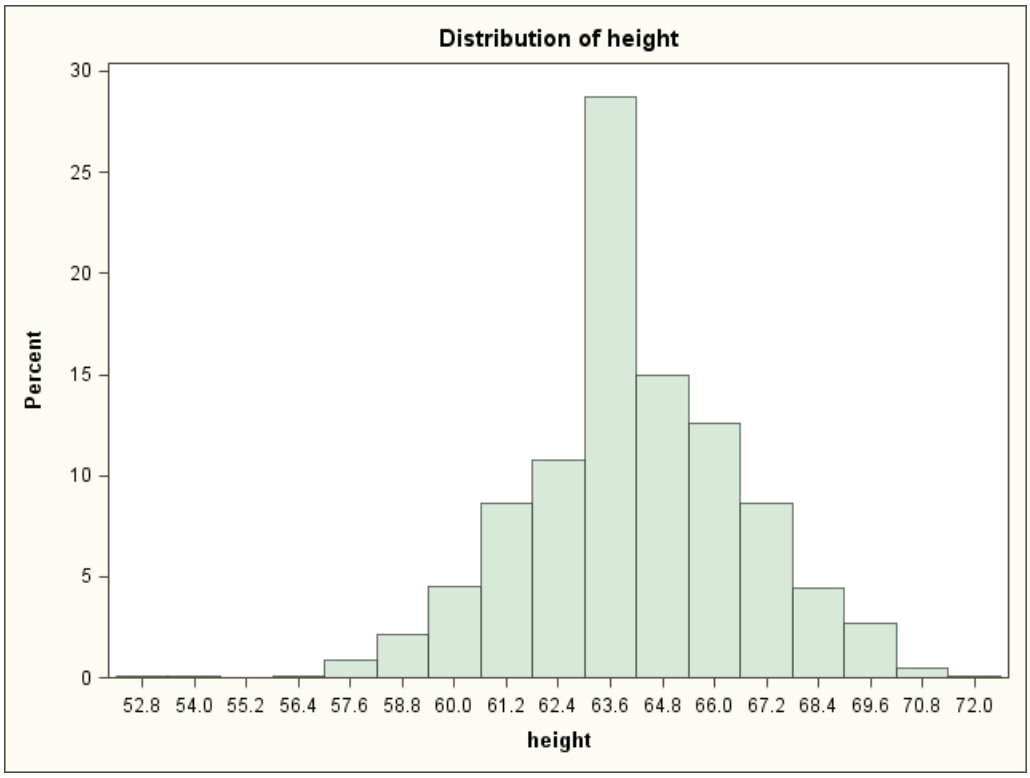

Probability Plot for gestation
gestation
400
350
300
250
200
150
100
0.01 0.1 1 5 10 25 50 75 90 95 99 99.9 99.99
Normal Percentiles

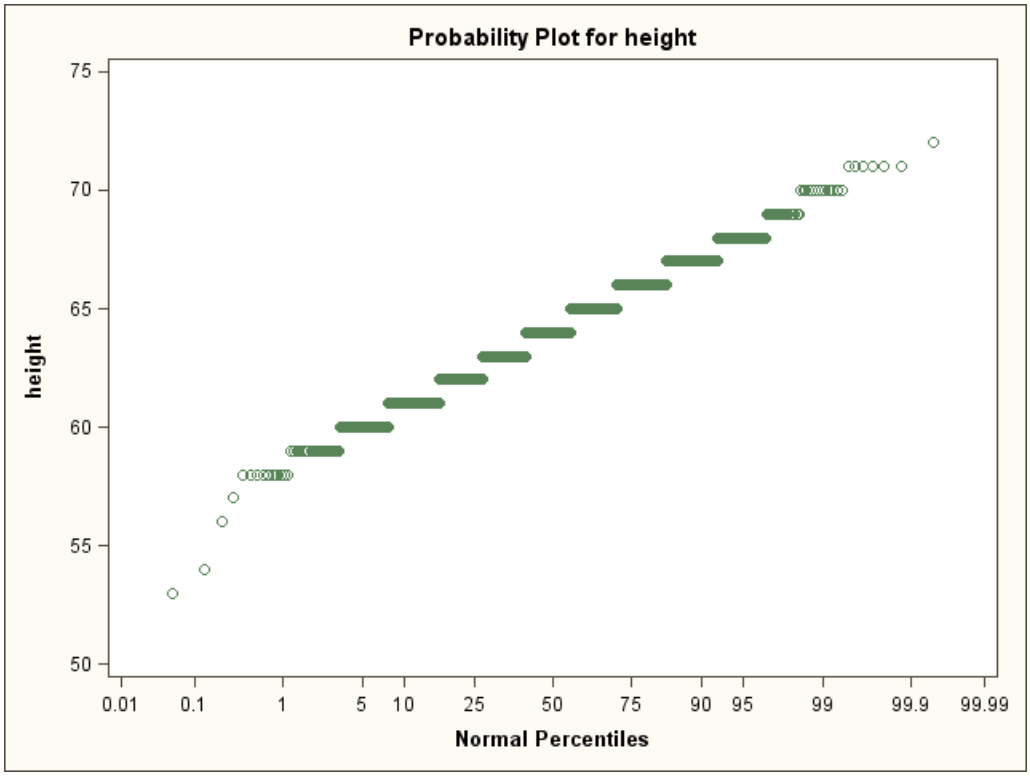
AGE:


Distribution of age
Percent
15.0
12.5
10.0
7.5
5.0
2.5
0.0
15 17 19 21 23 25 27 29 31 33 35 37 39 41 43 45
age

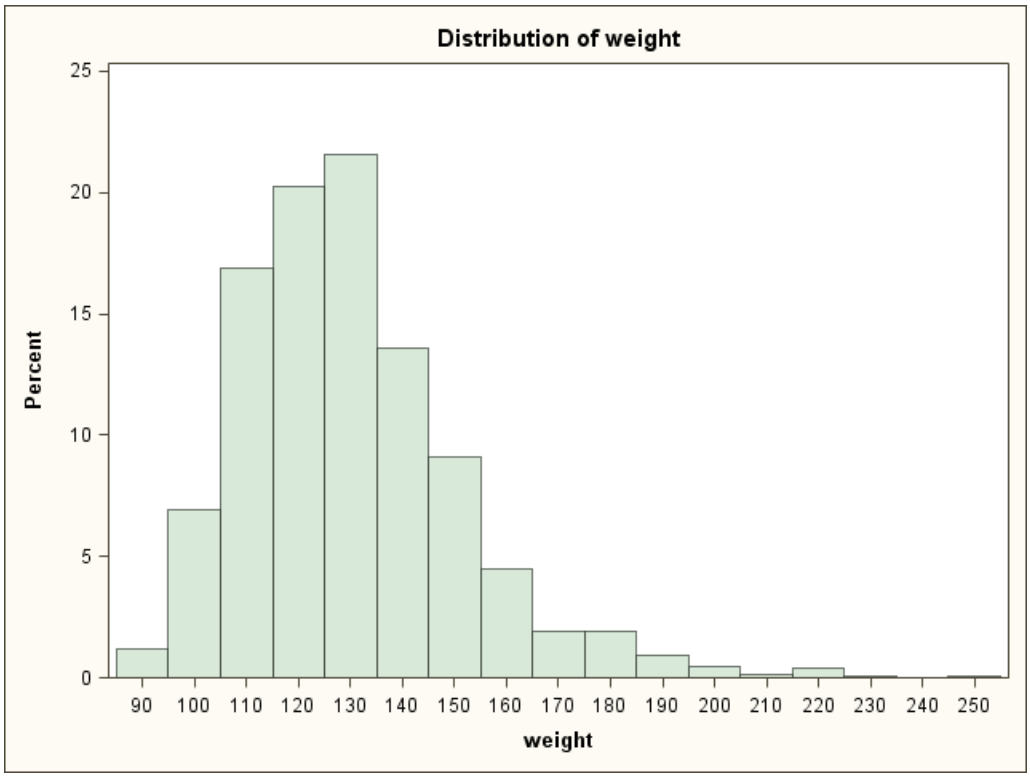

Probability Plot for age
age
45
40
35
30
25
20
15
0.01 0.1 1 5 10 25 50 75 90 95 99 99.9 99.99
Normal Percentiles

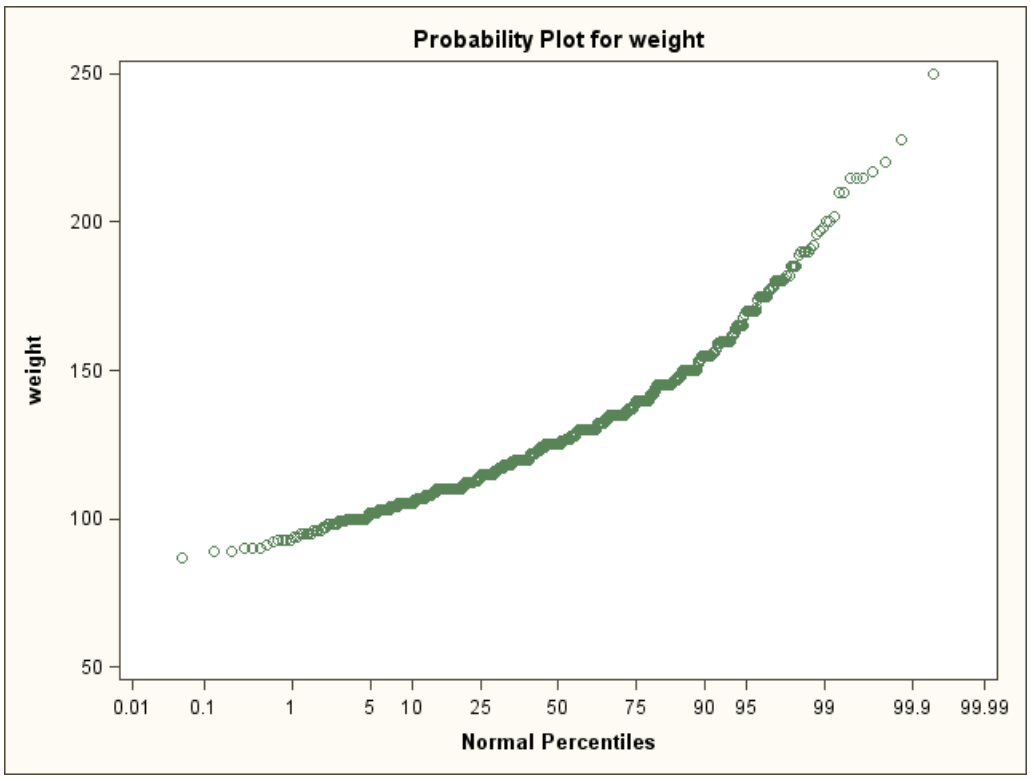
HEIGHT:

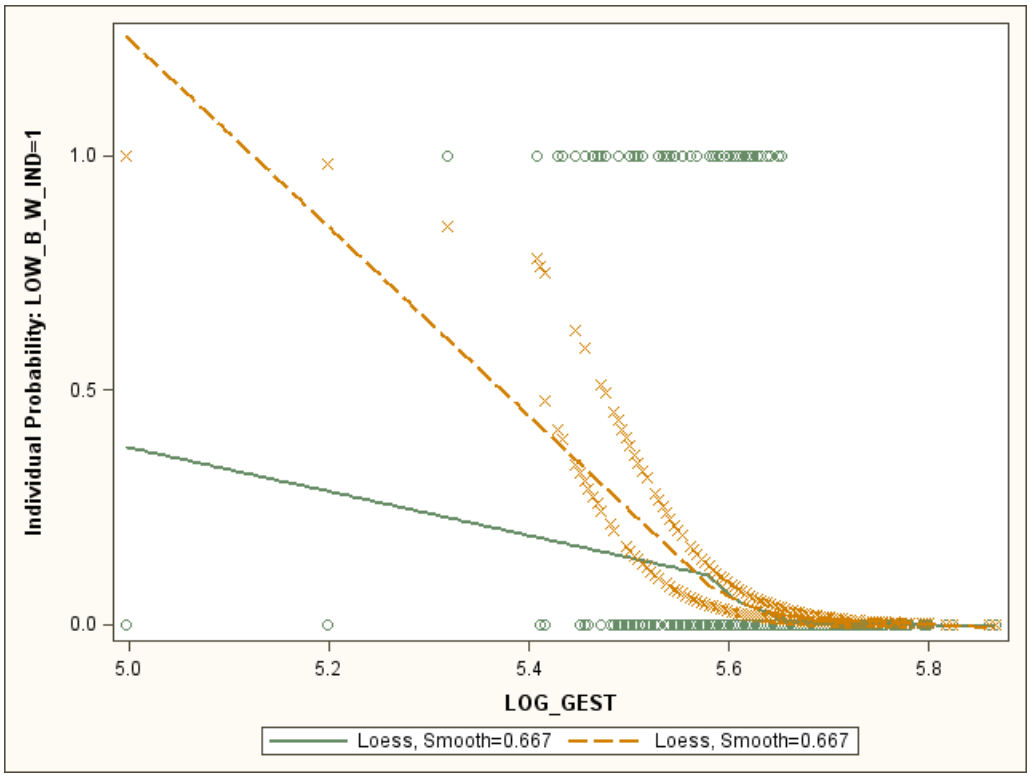

Distribution of height
Percent
30
25
20
15
10
5
0
52.8 54.0 55.2 56.4 57.6 58.8 60.0 61.2 62.4 63.6 64.8 66.0 67.2 68.4 69.6 70.8 72.0
height

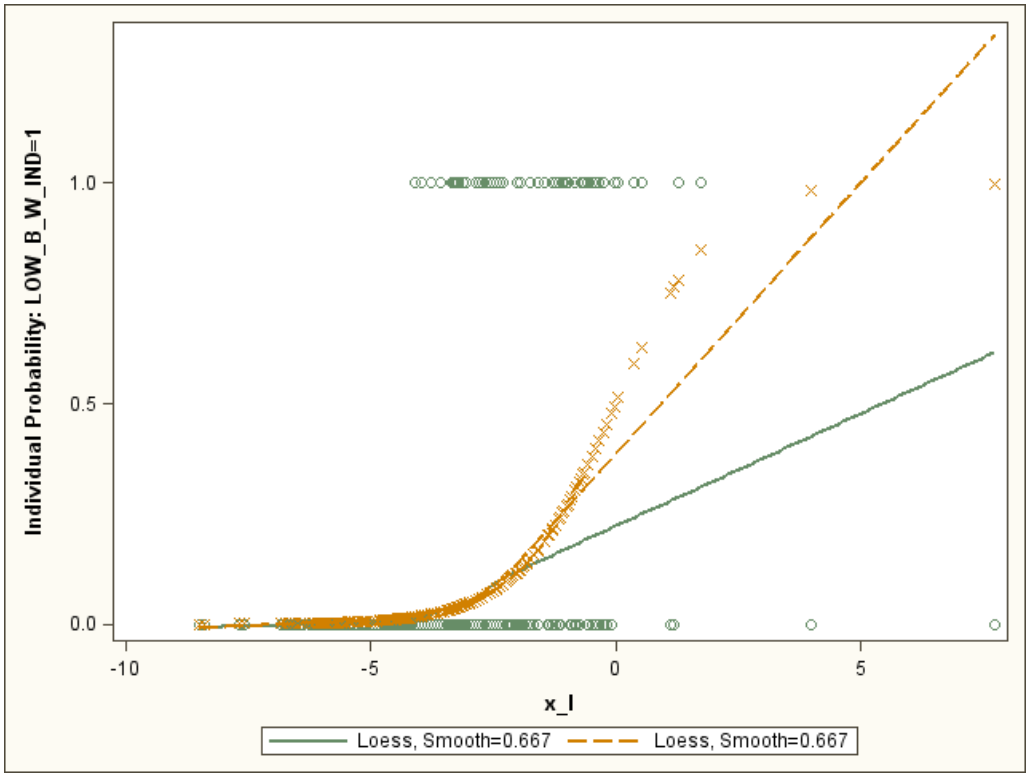

Probability Plot for height
height
75
70
65
60
55
50
0.01 0.1 1 5 10 25 50 75 90 95 99 99.9 99.99
Normal Percentiles

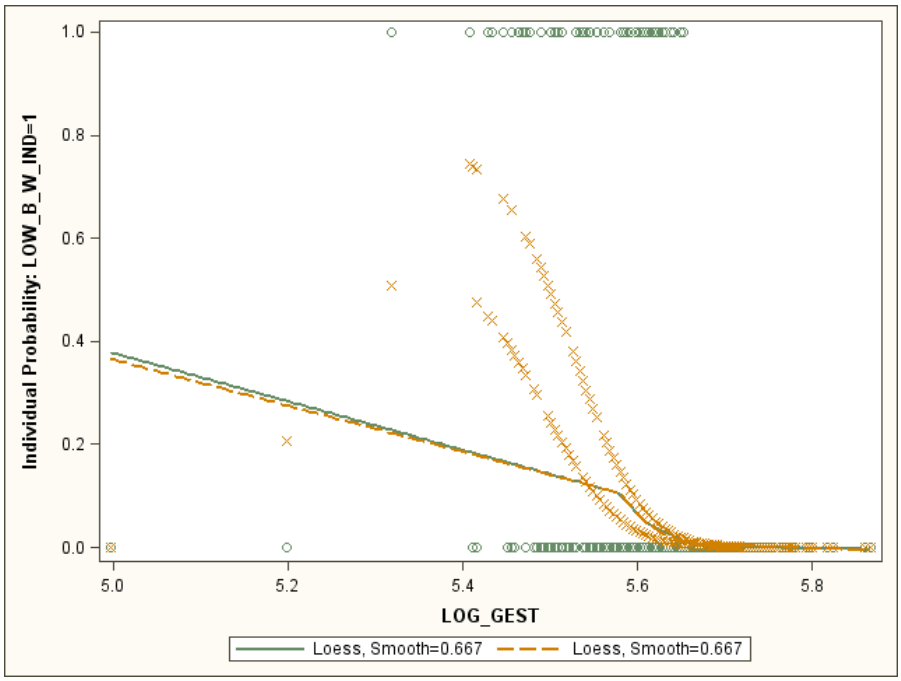
WEIGHT:

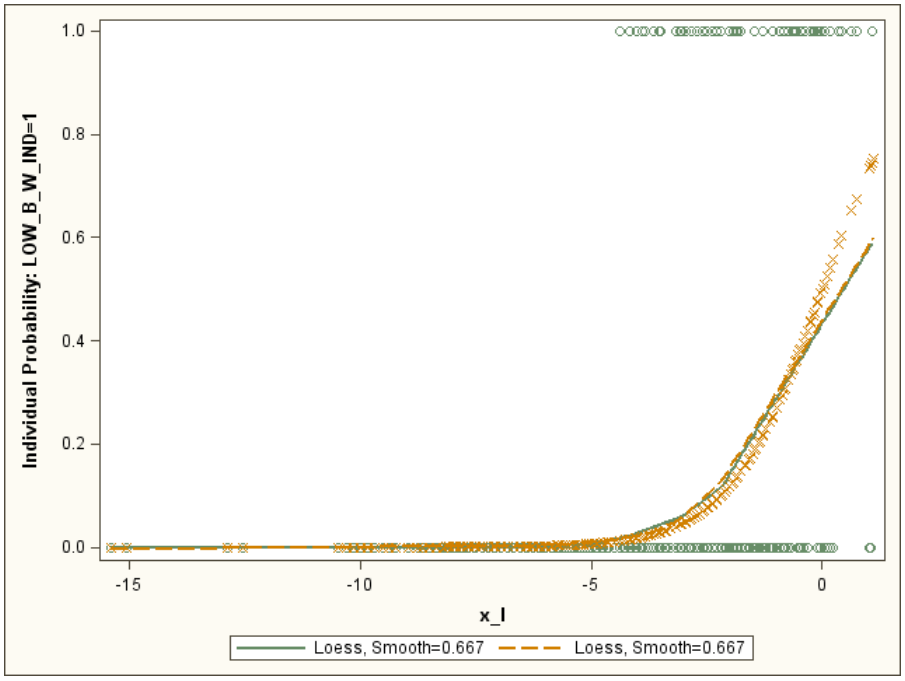

Distribution of weight
Percent
25
20
15
10
5
0
90 100 110 120 130 140 150 160 170 180 190 200 210 220 230 240 250
weight


Probability Plot for weight
weight
250
200
150
100
50
0.01 0.1 1 5 10 25 50 75 90 95 99 99.9 99.99
Normal Percentiles


Individual Probability: LOW_B_W_IND=1
1.0
0.5
0.0
5.0
5.2
5.4
5.6
5.8
LOG_GEST
Loess, Smooth=0.667
Loess, Smooth=0.667


Individual Probability: LOW_B_W_IND=1
1.0
0.5
0.0
-10
-5
0
5
x_l
Loess, Smooth=0.667
Loess, Smooth=0.667


Individual Probability: LOW_B_W_IND=1
1.0
0.8
0.6
0.4
0.2
0.0
5.0
5.2
5.4
5.6
5.8
LOG_GEST
Loess, Smooth=0.667
Loess, Smooth=0.667


Individual Probability: LOW_B_W_IND=1
1.0
0.8
0.6
0.4
0.2
0.0
-15
-10
-5
0
x_l
Loess, Smooth=0.667
Loess, Smooth=0.667