STAT 500 Homework 8

Vu Thi-Hong-Ha; NetID: 851924086

October 23, 2020

1 Question 1

(a)

- Treatment design: combinations of levels of two factors: frog species (Hyla regilla, Rana cascade, and Bufo boreas) and type of radiation filters (no filter, UV-B transmitting filter, and UV-B blocking filter).

- Experimental design: This in a randomized complete block design.

+ Experimental units: frog eggs.

+ Response variables: the percentage of eggs that failed to hatch.

+ Replications: yes. Each combination of the two factors has 4 enclosures.

+ Randomization: yes. Within each location, four enclosures were randomly assigned to each of the 9 combination of the two factors.

+ Blocking: yes. Location blocks.

(b)

(i) $\mu_{33k} = \mu + \alpha_3 + \tau_3 + (\alpha\tau)_{33} + \beta_k$. Hence, $\mu = \mu_{33k} - \beta_k$. The overall mean is the difference between the mean effects of UVB filters and species 3 and the block effects.

(ii) $\mu_{13k} = \mu + \alpha_1 + \tau_3 + (\alpha\tau)_{13} + \beta_k$. Hence, $\alpha_1 = \mu_{13k} - \mu - \beta_k = \mu_{13k} - (\mu_{33k} - \beta_k) - \beta_k = \mu_{13k} - \mu_{33k}$. The main effect of filter level 1 (i.e, no filter) is the difference between filter 1 and filter 3 when imposing on species 3.

(iii) $\mu_{32k} = \mu + \alpha_3 + \tau_2 + (\alpha\tau)_{32} + \beta_k$. Hence, $\tau_2 = \mu_{32k} - \mu - \beta_k = \mu_{32k} - (\mu_{33k} - \beta_k) - \beta_k = \mu_{32k} - \mu_{33k}$. The main effect of species 2 is the difference between species 2 and species 3 when using filter level 3.

(iv) $\mu_{12k} = \mu + \alpha_1 + \tau_2 + (\alpha\tau)_{12} + \beta_k$. Hence,

$(\alpha\tau)_{12} = \mu_{12k} - \mu - \alpha_1 - \tau_2 - \beta_k = \mu_{12k} - \mu_{33k} + \beta_k - \mu_{13k} + \mu_{33k} - \mu_{32k} + \mu_{33k} - \beta_k = (\mu_{12k} - \mu_{13k}) - (\mu_{32k} - \mu_{33k})$.

The interaction between filter level 2 and species 2 is the difference between filter 1 and 3.

(v) $\mu_{12k} = \mu + \alpha_1 + \tau_2 + (\alpha\tau)_{12} + \beta_k$. Hence, $\mu + \alpha_1 + \tau_2 + (\alpha\tau)_{12} = \mu_{12k} - \beta_k$.

(vi) $(\alpha\tau)_{12} - (\alpha\tau)_{32} - (\alpha\tau)_{13} + (\alpha\tau)_{33} = (\alpha\tau)_{12} - 0 - 0 + 0 = (\alpha\tau)_{12} = (\mu_{12k} - \mu_{13k}) - (\mu_{32k} - \mu_{33k})$.

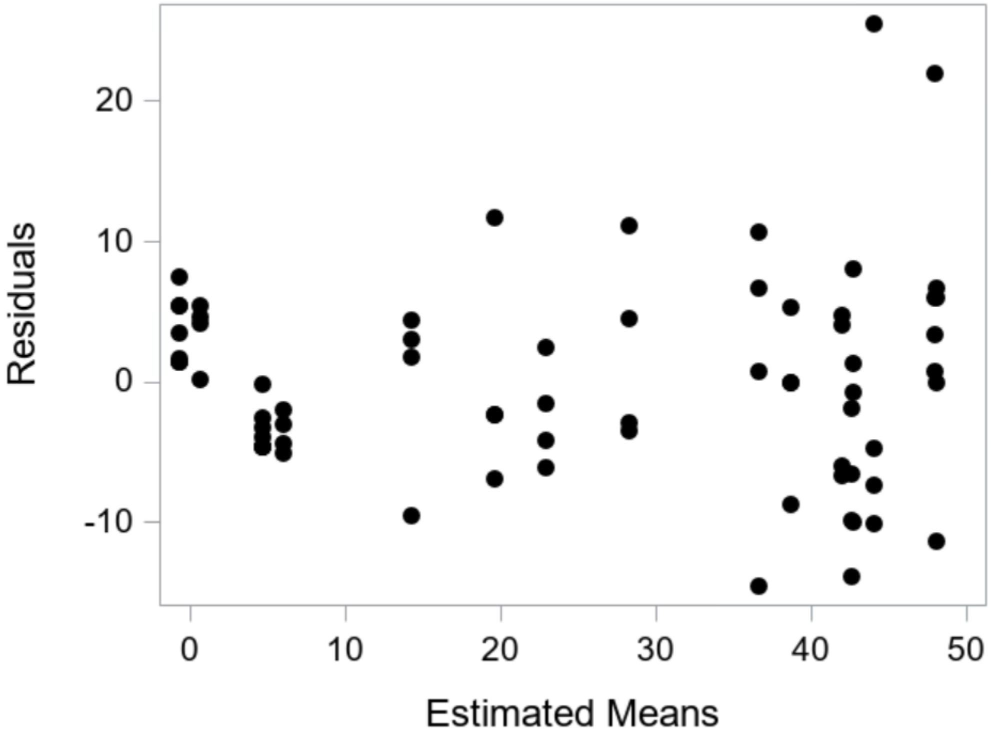
(c) The number of blocks $n = 2$, the number of levels of Factor 1 $a = 3$, the number of levels of Factor 2 $b = 3$.

Source of variation	D.f	SS	MS	F	p-value
Location	$n-1=1$	520.03125	520.03125	8.96	0.0040
Filters	$a-1=2$	3434.06194	1717.03097	29.58	$< .0001$
Species	$b-1=2$	17855.39528	8927.69764	153.82	$< .0001$
Filter*Species Interaction	$(a-1)(b-1)=4$	1914.09056	478.52264	8.24	$< .0001$
Error (residuals)	62	3598.50750	58.04044		
Corrected Total	$abN-1=71$	27322.08653			

(d) Looking at the plot, we can see that filter 2 and 3 have similar trends in hatch rates, but filter 1 is very different. The lines are not parallel, so there is interaction between filters and species. From the ANOVA table above, we can see that the test statistics when testing for the effects of filter and species interaction is $F = 8.24$, with p-value $< .0001$. At a significance level of 0.05, we reject the null that there is no effects of filter and species interaction. The conclusion from testing and the plot agree with each other.

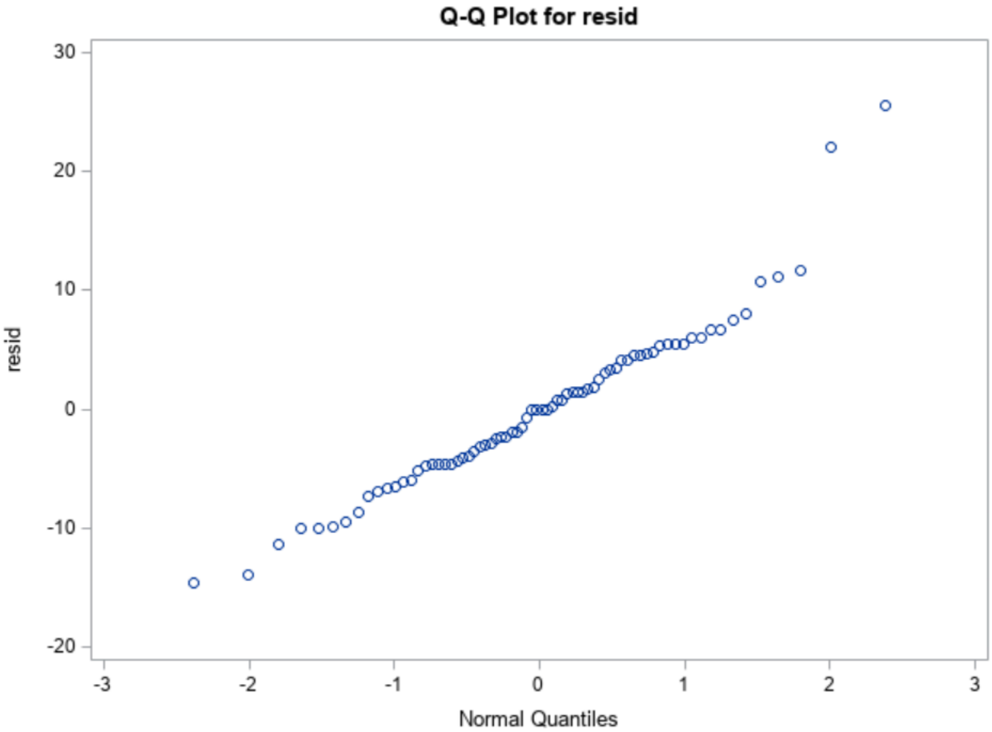
(e) From the ANOVA table above, the test statistic when testing for the effects of different filters is $F = 29.58$, with p-value $< .0001$. At a significance level of 0.05, we reject the null that there is no effects of filters.

(f)



The data points do not seem to follow any patterns, indicating homogeneous variances. There could be some residual outliers which are greater than 20.

(g)



The data points seem to follow a line, indicating that the data may follow normal distribution. There could be some outliers though.

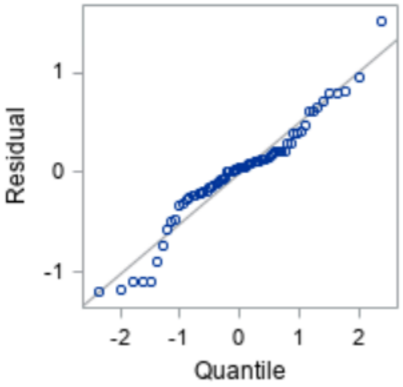
(h)

By the residual plots above, there could be some outliers, so we may need to transform data. As our data is all positive, and is not count data, we should use log transformation.

Source of variation	D.f	SS	MS	F	p-value
Location	$n-1=1$	0.3498521	0.3498521	1.16	0.2854
Filters	$a-1=2$	1.5302883	0.7651442	2.54	0.0874
Species	$b-1=2$	110.8446915	55.4223457	184.15	$<.0001$
Filter*Species Interaction	$(a-1)(b-1)=4$	3.1668982	0.7917246	2.63	0.0433
Error (residuals)	58	17.4554186	0.3009555		
Corrected Total	67	133.3471487			

With the newly transformed data, we conclude that species have significant impacts (F-statistics = 184.15, p-value $< .0001$), interaction between species and filters have significant impacts (F-statistics = 2.63, p-value 0.0433), but the filters do not have significant impacts (F-statistics = 2.54, p-value 0.0874).

However, now if we look at the new QQ plot, we can see that the newly transformed data seem to be heavy tailed.



Although the original data seems to have some outliers, the QQ plots and Shapiro-Wilk test (W = 0.953786, p-value = 0.01) indicate that the original data follow normal distribution. Therefore, in my opinion, we should not transform the data. In this case, the outliers do not put much effects, but in more serious cases, we can remove the outliers.

2 Question 2

(a)

β_0: the conditional mean of diamond price when the diamond weight is x_0.

β_1: the change in the conditional mean of diamond price when the diamond weight increases by 1 gram.

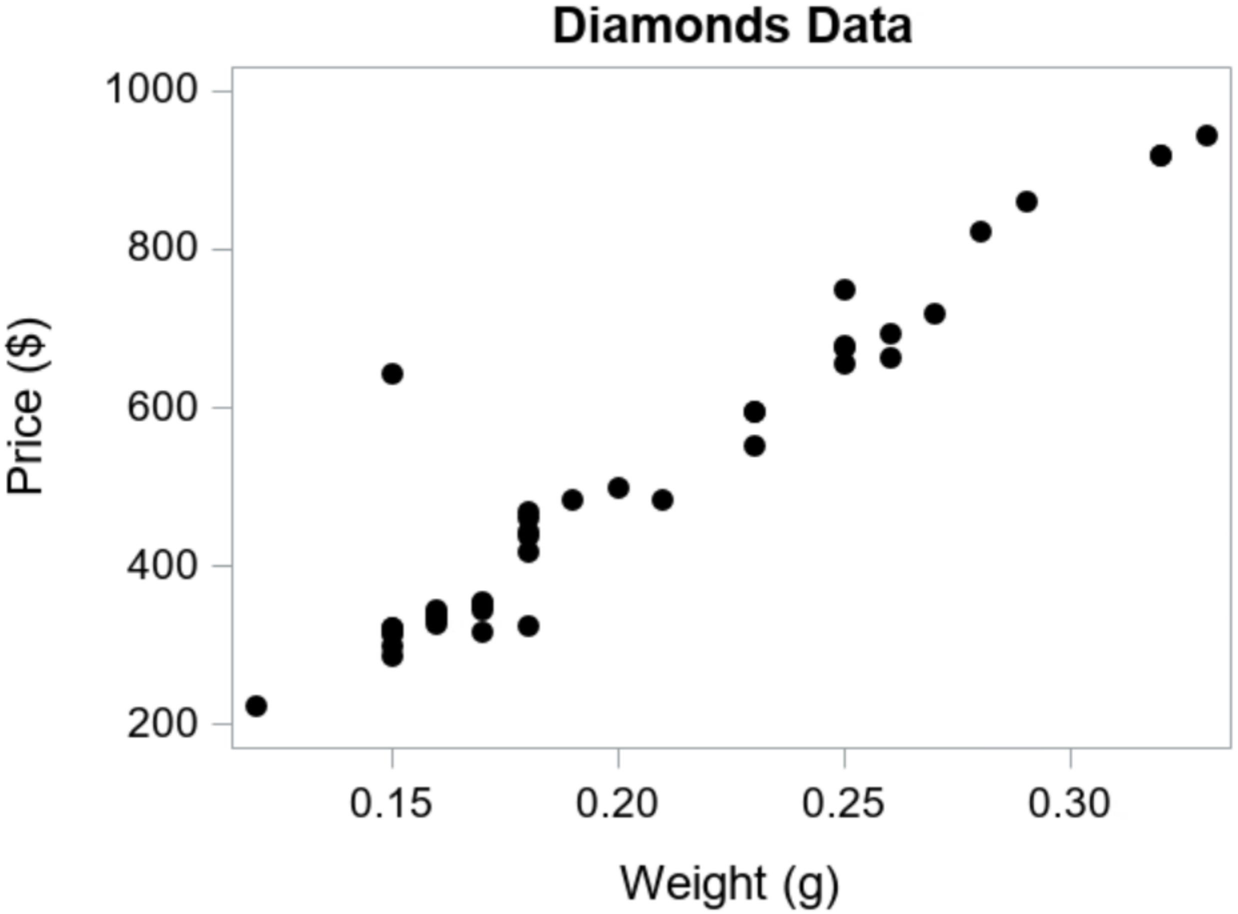
σ^2: the variation of diamond prices about the conditional mean for any specific value of diamond weight.

(b)

$$\begin{pmatrix} Y_1 \\ Y_2 \\ Y_3 \\ Y_4 \\ \dots \\ Y_{46} \end{pmatrix} = \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ 1 & x_3 \\ 1 & x_4 \\ \dots & \\ 1 & x_{46} \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix} \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \epsilon_3 \\ \epsilon_4 \\ \dots \\ \epsilon_{46} \end{pmatrix}$$

The design matrix $\mathbf{X}$ is $\begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ 1 & x_3 \\ 1 & x_4 \\ \dots & \\ 1 & x_{46} \end{pmatrix}$

(c)



From the scatter plot, it seems like there is a linear relationship between the price and weight of diamonds. They seem to have a positive correlation, meaning that the heavier the diamond is, the more expensive it is.

(d) $R^2 = 0.9130$, so the sample correlation coefficient is $\pm\sqrt{0.9130} = \pm 0.9555103$. From the plot, we can see the sample correlation coefficient should be 0.9555103, indicating that there is a strong positive correlation between the price and weight of diamonds, agreeing with part (c).

(e) $Y_{new} = \beta_0 + \beta_1 x + \epsilon$

The estimate is $\hat{Y} = b_0 + b_1 x = -216.48416 + 3543.60253x$

(f)

For $x = 0.2$: $\hat{Y} = -216.48416 + 3543.60253(0.2) = 492.2363$.

The standard error of this estimate: $\sqrt{MS_{error}(1 + \frac{1}{n} + \frac{(x-\bar{x})^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2})}$

$\sum_{i=1}^{n}(x_i-\bar{x})^2 = \frac{SS_{model}}{b_1^2} = \frac{1619759}{3543.60253^2} = 0.1289913$

$MS_{error} = 3508.60432$

$n = 46$

Then,

$$\sqrt{MS_{error}(1 + \frac{1}{n} + \frac{(x-\bar{x})^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2})} = \sqrt{3508.60432(1 + \frac{1}{46} + \frac{(0.2-0.1995652)^2}{0.1289913})} = 59.87389$$

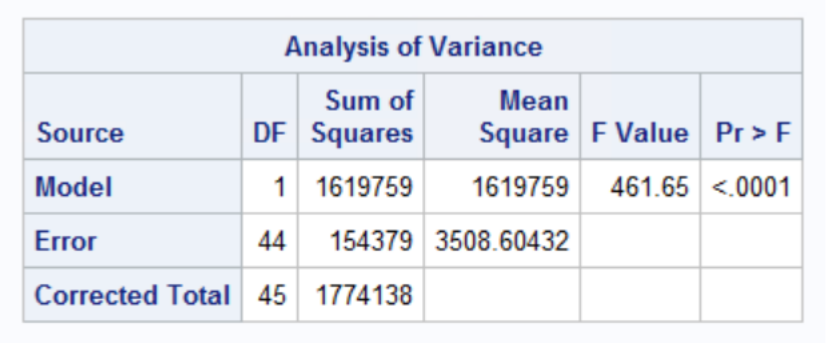
For $x = 0.28$: $\hat{Y} = -216.48416 + 3543.60253(0.28) = 775.7245$.

The standard error of this estimate:
$$\sqrt{MS_{error}(1 + \frac{1}{n} + \frac{(x-\bar{x})^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2})} = \sqrt{3508.60432(1 + \frac{1}{46} + \frac{(0.28-0.1995652)^2}{0.1289913})} = 61.32583$$
(g)

For $x = 0.2$, the residual is $498 - 492.2363 = 5.7637$

For $x = 0.28$, the residual is $823 - 775.7246 = 47.2754$

(h) ANOVA table:



Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	1619759	1619759	461.65	<.0001
Error	44	154379	3508.60432		
Corrected Total	45	1774138			

To test for the significance of the linear model, we have the test statistic $F = 461.65$ with p-value $< .0001$. Hence, at a significance level of 0.05, we reject the null that the model is not significant.

(i) $R^2 = 0.9130$. Hence, 91.3% of the variation in the diamond price can be explained by the linear regression model with the diamond weight. This indicates that our model fits quite well.

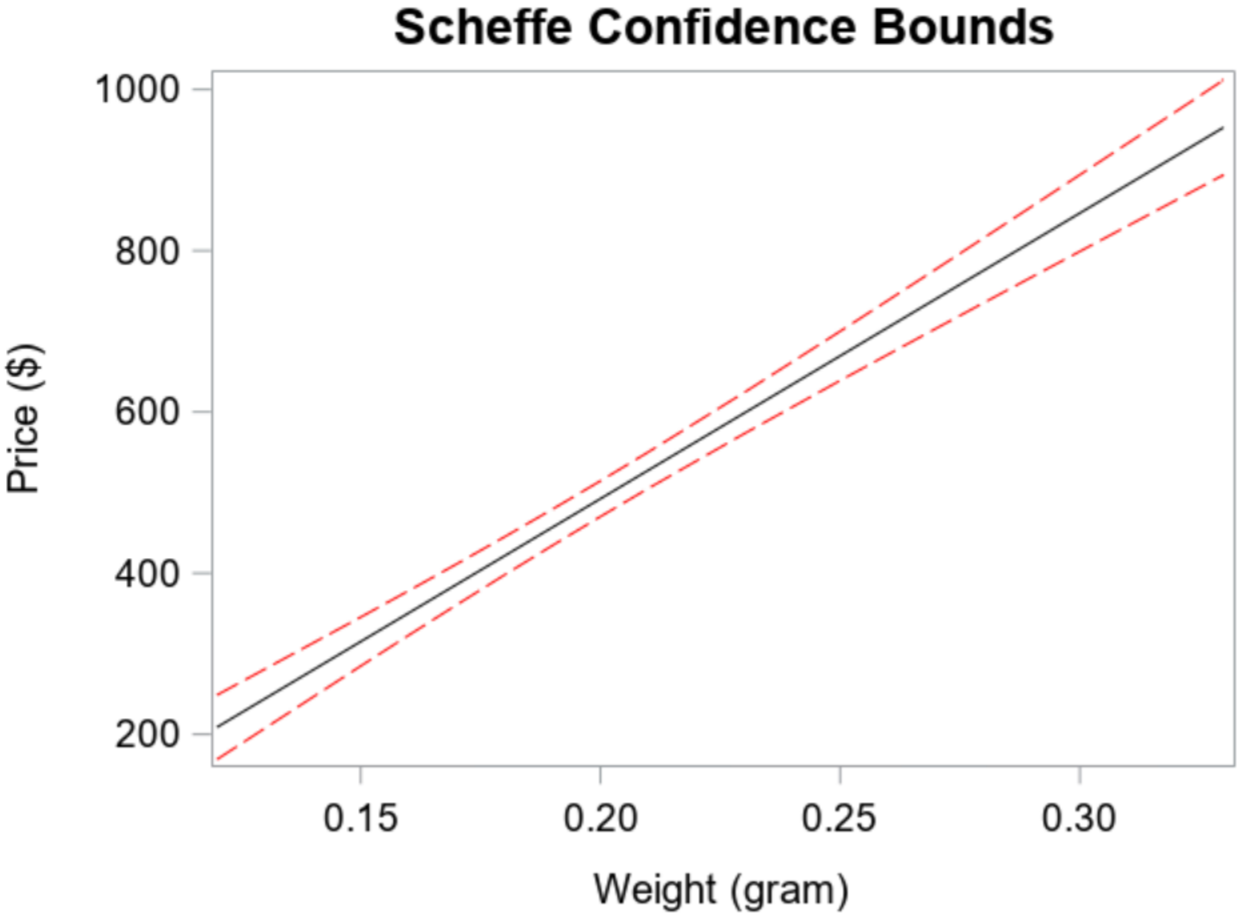
(j)

$S_{b_1} = \sqrt{MS_{error}/\sum_{i=1}^{n}(x_i-\bar{x})^2} = \sqrt{3508.60432/0.1289913} = 164.9252$.

A $100(1-\alpha)\%$ confidence interval for β_1 is $\beta_1 \pm t_{n-2,1-\alpha/2}S_{b_1} = 3543.60253 \pm 2.01537 * 164.9252 = (3211.217; 3875.988)$. This CI does not contain 0, so we reject the null that $\beta_1 = 0$ at a significance level of 0.05. If we sample for multiple times, $100(1-\alpha)\%$ times the obtained CIs will contain the true value of β_1.

(k) A 95% confidence interval for the conditional mean price of all diamonds in the population with a weight of 0.2 grams: (474.6345; 509.8382). We can be 95% confident that price of all diamonds in the population with a weight of 0.2 grams is between 474.6345 and 509.8382.

(l) Simultaneous confidence intervals for every $0.12 \leq x \leq 0.35$ using Scheffe's method: $(b_0+b_1x) \pm \sqrt{2F_{2,n-2,1-\alpha}}S_{b_0+b_1x}$, where $b_0 = -216.48416$, $b_1 = 3543.60253$, $F_{2,n-2,1-\alpha} = F_{2,44,0.95} = 3.20928$,
$S_{b_0+b_1x} = \sqrt{MS_{error}(\frac{1}{n} + \frac{(x-\bar{x})^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2})} = \sqrt{3508.60432(\frac{1}{46} + \frac{(x-0.1995652)^2}{0.1289913})}$



(m) $100(1-\alpha)\%$ prediction interval for $x = 0.28$ is

$(b_0 + b_1 x) \pm t_{44,0.975} S_{pred} = -216.48416 + 3543.60253(0.28) \pm 2.01537 * 61.32583 = (652.1303, 899.3188)$

We can be 95% confident that next new observation when the diamond weight is 0.28 is between 652.1303 and 899.3188.