HW3

Ben Tankus

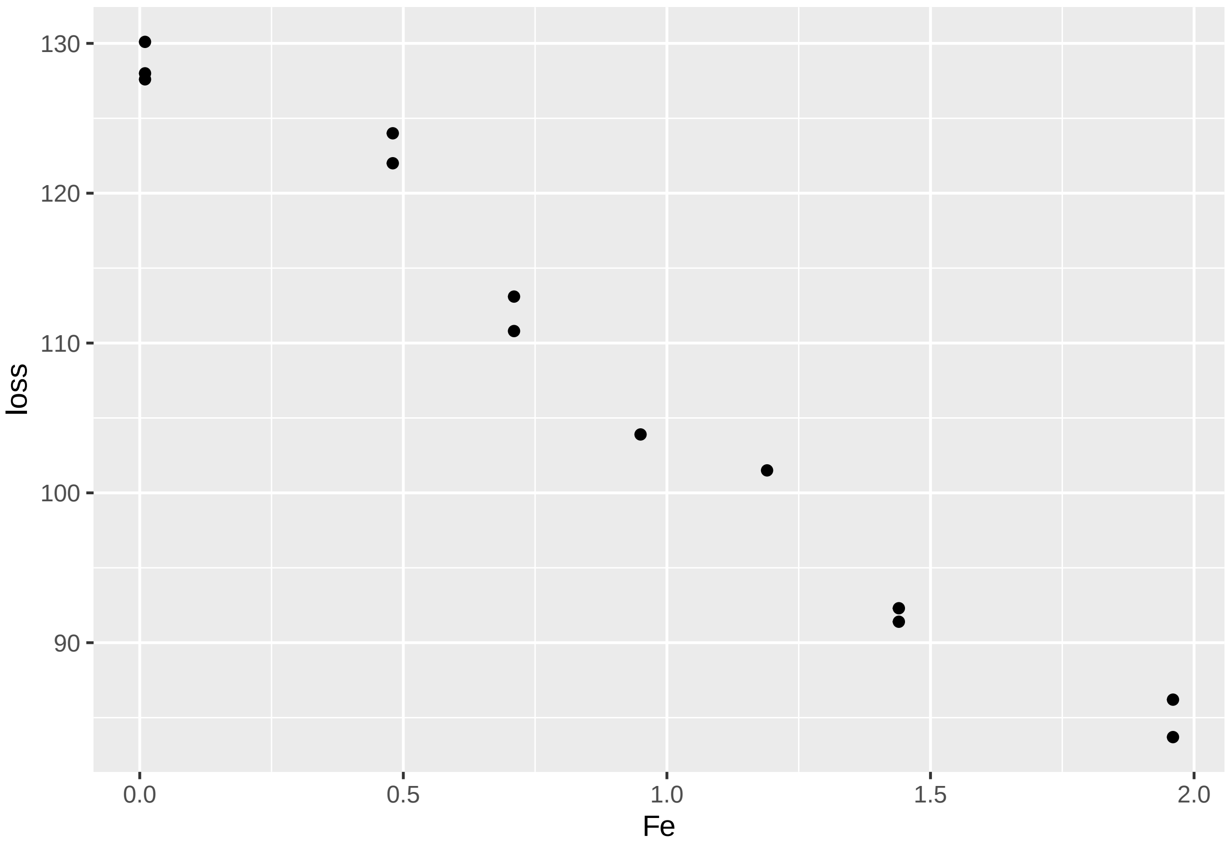
1/20/2021

(7 points) This question uses a dataset from the faraway package to perform a Lack of Fit test on a model we will fit. Use this code to load the dataset. You may need to install the faraway package if you don't already have it.

```
data(corrosion, package = "faraway") # Load data from faraway package corrosion
# Look at data
#?faraway::corrosion # Learn about dataset
```

(a) We are interested in modling the weight loss due to corrosion as a function of Iron content, that is, Iron content is the explanatory variable, and weight loss is the response. Create a scatterplot of the data and describe the relationship you see.

```
qplot(Fe, loss, data = corrosion)
```



## There is a very strong linear relationship between iron content and loss. As the iron content increases, the loss will decrease.

(b) Fit a simple linear regression model to the data. State the mathematical form of the model and report the parameter estimates β_0, β_1, and $\hat{\sigma}$.

```
fit <- lm(loss~Fe, data = corrosion)
summary(fit)
```

```
##
## Call:
## lm(formula = loss ~ Fe, data = corrosion)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.7980 -1.9464  0.2971  0.9924  5.7429
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  129.787      1.403   92.52  < 2e-16 ***
## Fe           -24.020      1.280  -18.77 1.06e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.058 on 11 degrees of freedom
## Multiple R-squared:  0.9697, Adjusted R-squared:  0.967
## F-statistic: 352.3 on 1 and 11 DF,  p-value: 1.055e-09
```

$B_0 : 129.787$ $B_1 : -24.02$ $\hat{\sigma} : 3.058$

$$loss = \beta_0 + \beta_1(Fe) + \epsilon$$

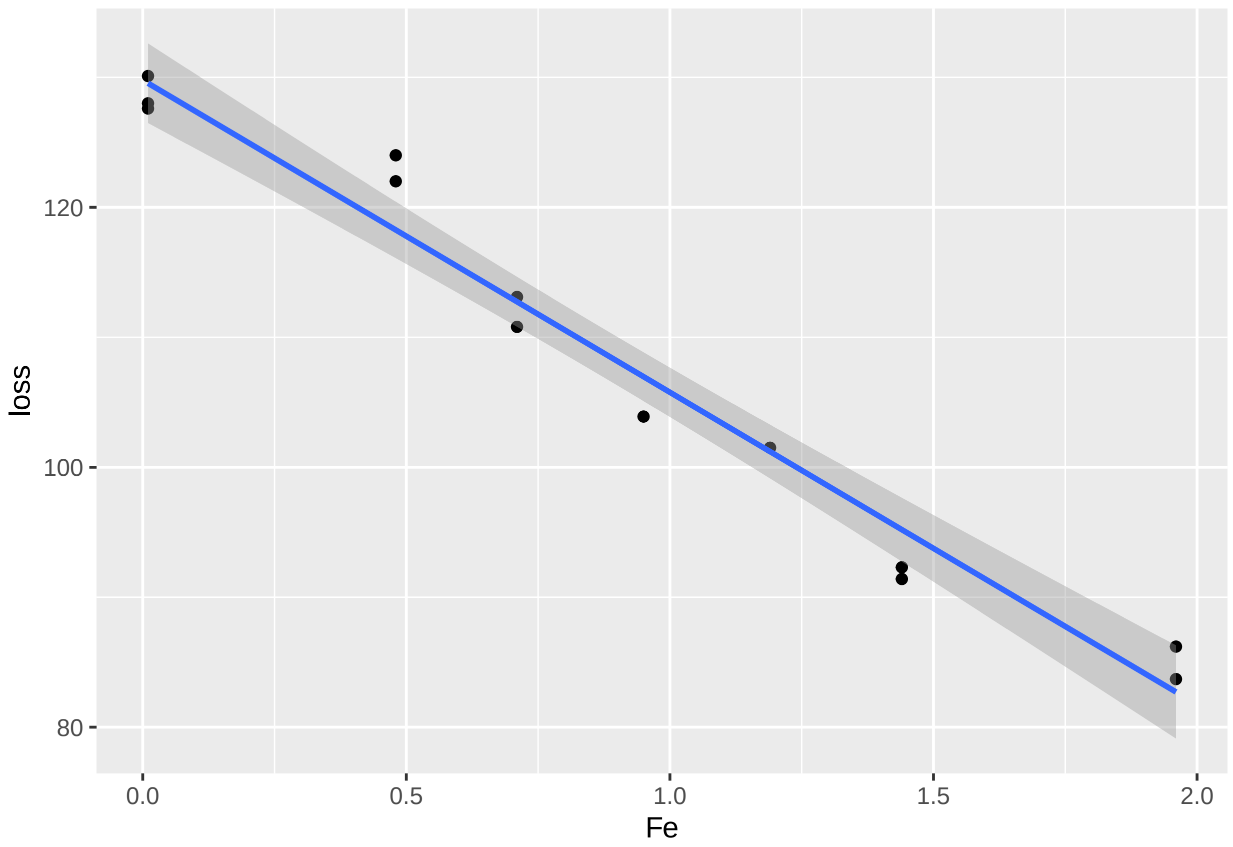
(c) In the context of the data, interpret β_1 in one sentence.

For every increase of one unit in iron content, there will be a drop of 24 units of average weight loss.

(d) Repeat part (a), but this time include the regression line and confidence bands for the mean weight loss due to corrosion.

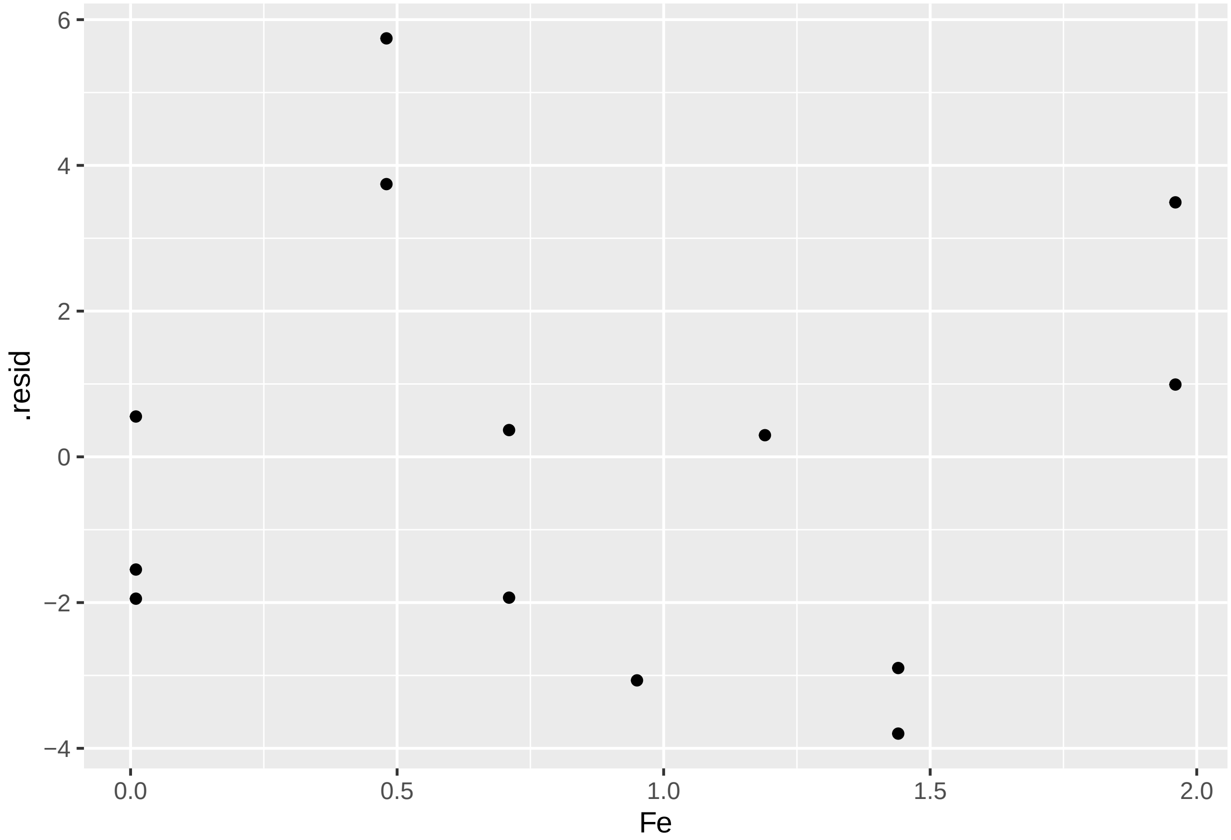
```
qplot(Fe, loss, data = corrosion) + geom_smooth(method = "lm")
```

```
## `geom_smooth()` using formula 'y ~ x'
```



(e) Now plot the residuals (y-axis) against the explanatory variable iron content (x-axis). Are the residuals centered around zero at all values of iron content?

```
fitAug <- augment(fit)
qplot(Fe,.resid, data = fitAug)
```



## The residuals are roughly centered around 0 at all values of iron. It's difficult to tell with such a small sample size.

(f) Give the null and alternative hypothesis for a Lack of Fit test.

H_o : ## Simple Linear Regression is an adequate model H_A : ## Simple Linear Regression is NOT an adequate model

(g) Perform a lack of fit test on our model. Give the F-statistic and p-value. What do you conclude? *Hint: See Lecture 6 from this module, use factor() to help fit the separate means model in lm(), and use anova() to compare the two models.

```
FitSMM <- lm(loss~factor(Fe), data = corrosion)
anova(fit, FitSMM)
```

```
## Analysis of Variance Table
##
## Model 1: loss ~ Fe
## Model 2: loss ~ factor(Fe)
##   Res.Df     RSS Df Sum of Sq      F   Pr(>F)
## 1     11 102.850
## 2      6  11.782  5    91.069 9.2756 0.008623 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Fstat: *9.28*, Pval: *0.0086*. We can conclude that SLR is not an adaquate model to fit this data. This means there is some departure from linearity in the relationship between iron content and weight loss.

(h) In one sentence, why is it necessary for the Lack of Fit test that there are independent replicate responses at some values of the explanatory variable?

For the Lack of Fit test you must calculate a mean for each group of the separate means model (SMM) to compare to the SMM to Simple Linear Regression.

2. (3 points) In the plots below, identify the SLR assumption that has been violated and explain your reasoning.

(a) Plot 1

Linearity violation. Residuals are obviously in a non-linear shape

(b) Plot 2

Constant variance violation. Varance on the left is much smaller than on the right.

(c) Plot 3

Model relevancy violation: The model looks like it consistently underestimates the response, therefore is not a good model. Model not centered around zero line.