ST517-HW3

Nora Quick

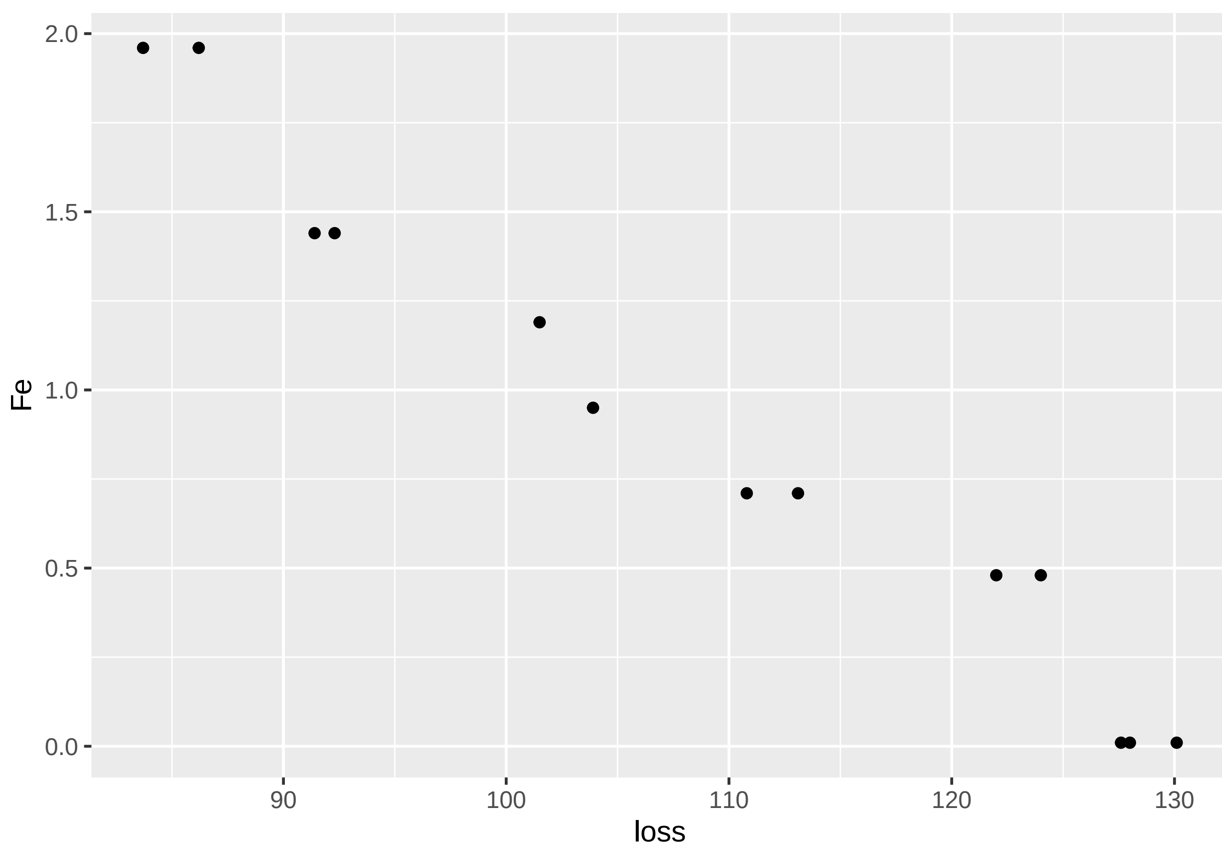
1. This question uses a dataset from the faraway package to perform a Lack of Fit test on a model we will fit. Use this code to load the dataset. You may need to install the 'faraway' package if you don't already have it.

```
data(corrosion, package = "faraway")
```

(a) We are interested in modelling 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.

```
Fe <- corrosion$Fe
loss <- corrosion$loss

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



The relationship between the Iron content and weight loss is linear. The longer it corrodes the more weight it loses in a linear fashion.

(b) Fit a simple linear regression model to the data. State the mathematical form of the model and report the parameter estimates

$\hat{}\beta 0$

,

$\hat{}\beta 1$

, 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
```

$\hat{}\beta 0$

= 129.79,

$\hat{}\beta 1$

= -24.02, and

$\hat{}\sigma$

= 3.06. The mathmatical form is y= mx + b or a linear model.

(c) In the context of the data, interpret

$\hat{}\beta 1$

in one sentence.

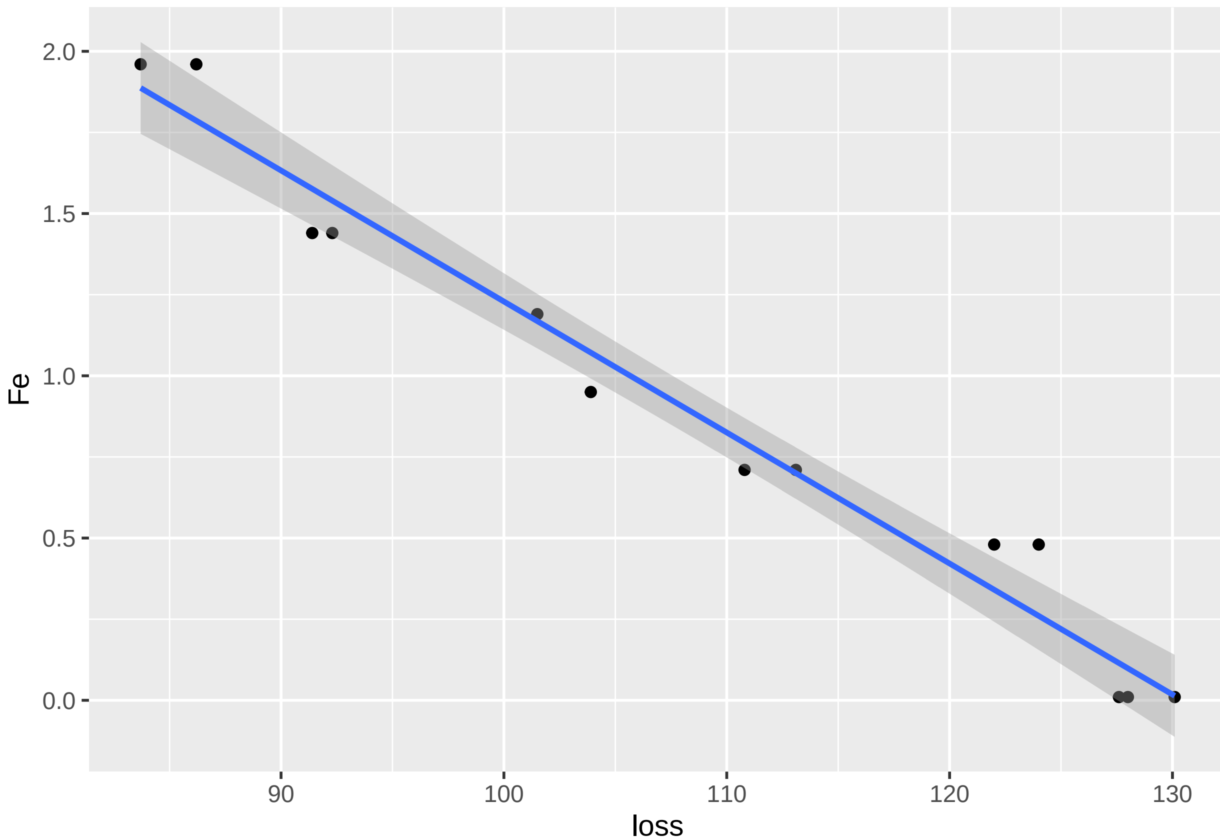
$\hat{}\beta 1$

is the slope, therefore, it is the rate in which the Iron will lose weight due to corrotion.

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

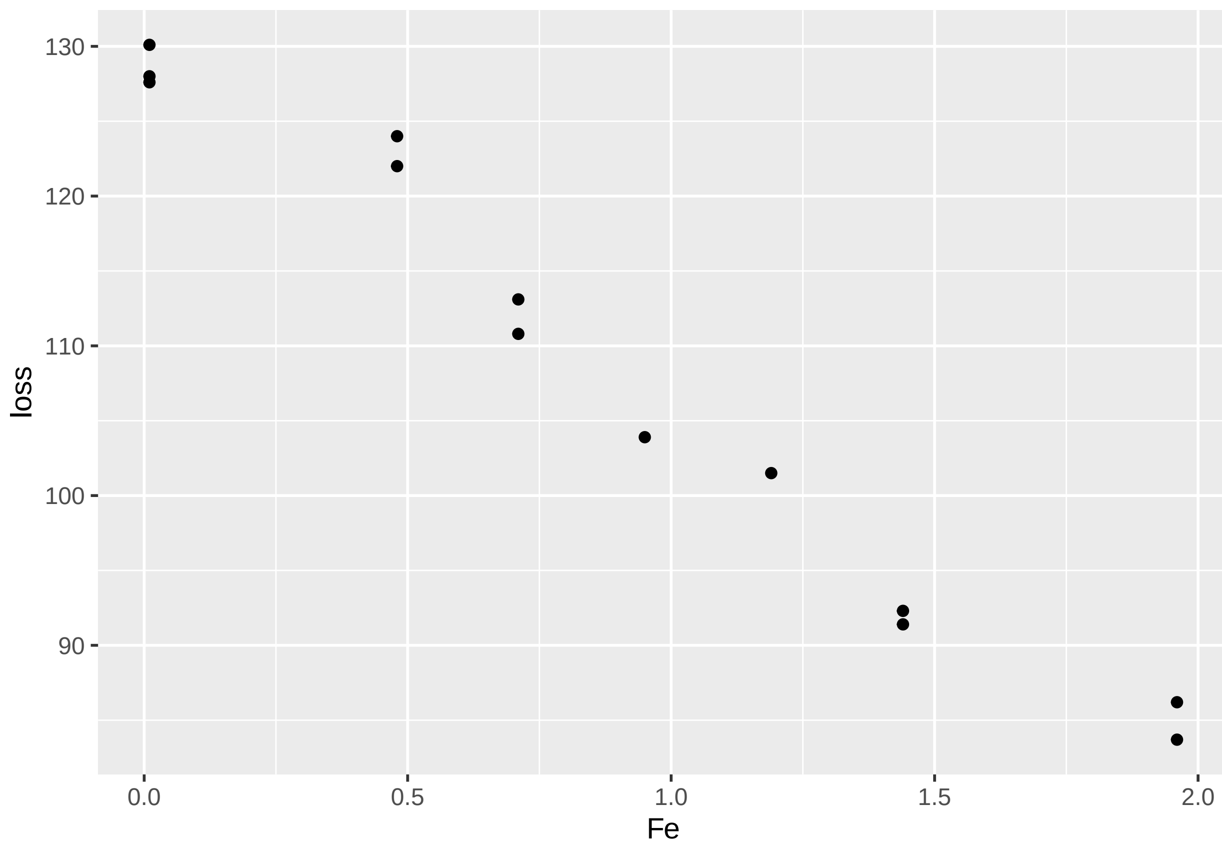
```r
qplot(loss, Fe, data = corrosion) + geom_smooth(method = "lm")
```

```
## `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?

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



Yes, it appears that all of the residuals average out to be around zero of all iron content.

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

$$H_0 : \mu = 0$$
$$H_A : \mu! = 0$$

(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.

```r
slr_corr <- lm(loss ~ Fe, data=corrosion)
sep_corr <- lm(loss ~ factor(Fe), data=corrosion)

anova(slr_corr, sep_corr)
```

```
## 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
```

After running anova() we can see that the F-statistic is 9.28 and the p-value is 0.0086. With this I conclude that the null hypothesis is true and that the data is linear. This means that there is a relationship between the weight or Iron based on the time it's corroded.

(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?

We need it for an estimate of the baseline of the Iron to make up for the random process of testing.

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

   (a) Plot 1: Fails the assumption of linearity (that Y's fall on the staright line function of the X's).

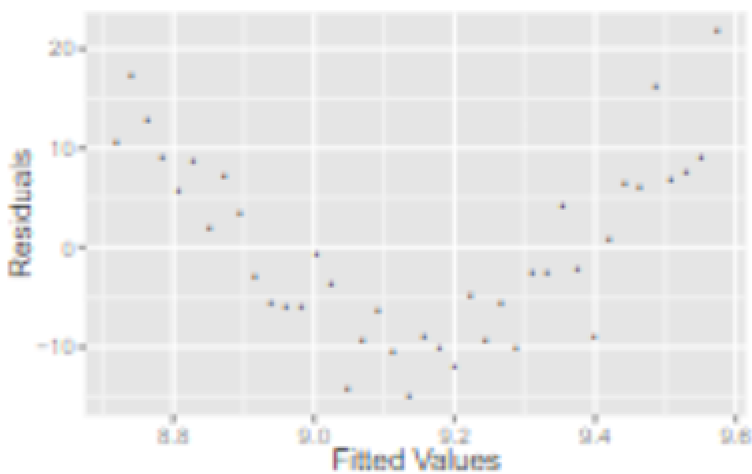


Figure 1: Plot 1

   (b) Plot 2: Fails the assumption of constant variance (that e's have the same variance).

   (c) Plot 3: Fails the assumption of normal distrubution.

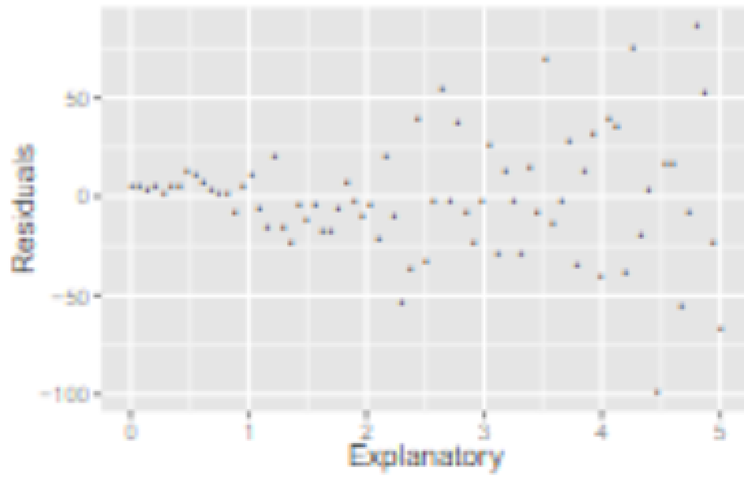


Figure 2: Plot 2

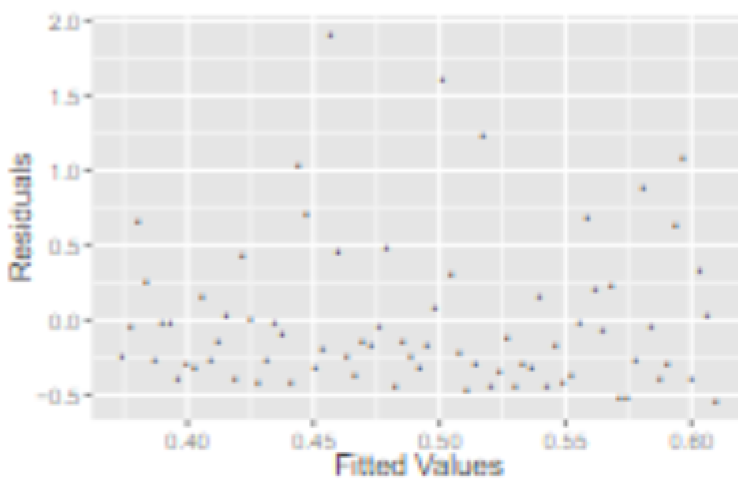


Figure 3: Plot 3