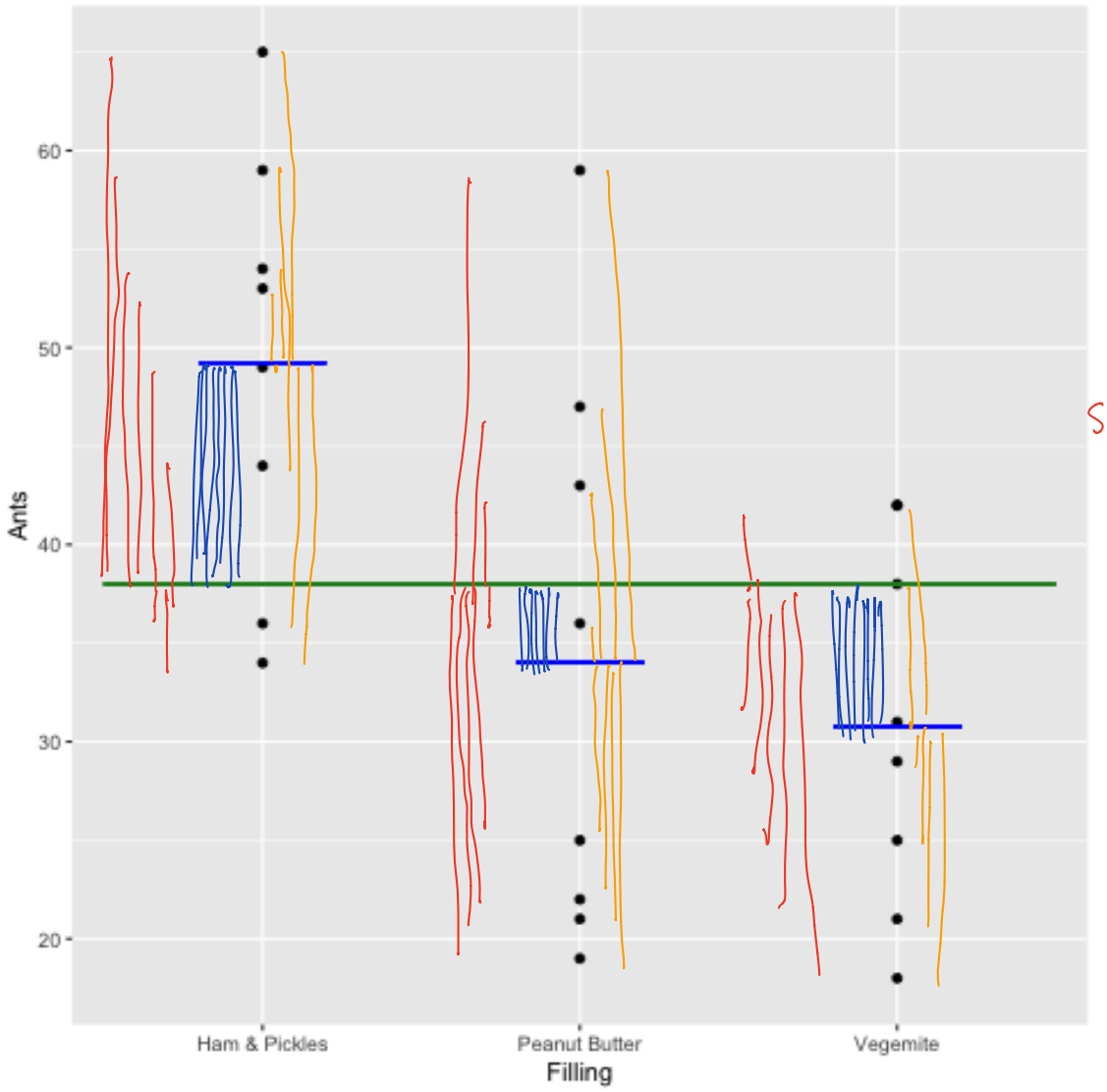
Stat 145, Fri 23-Apr-2021 -- Fri 23-Apr-2021
Biostatistics
Spring 2021

Friday, April 23rd 2021

Wk 12, Th
Topic:: Regression inference, randomization in R

Compare
- scenarios of Chapter 8 vs. Chapter 9
- Develop equivalent notion of SSG, SSE, SST
- Define R^2: happens to be the square of correlation
- ANOVA table
- Assumptions to use standard distributions (simple linear regression)

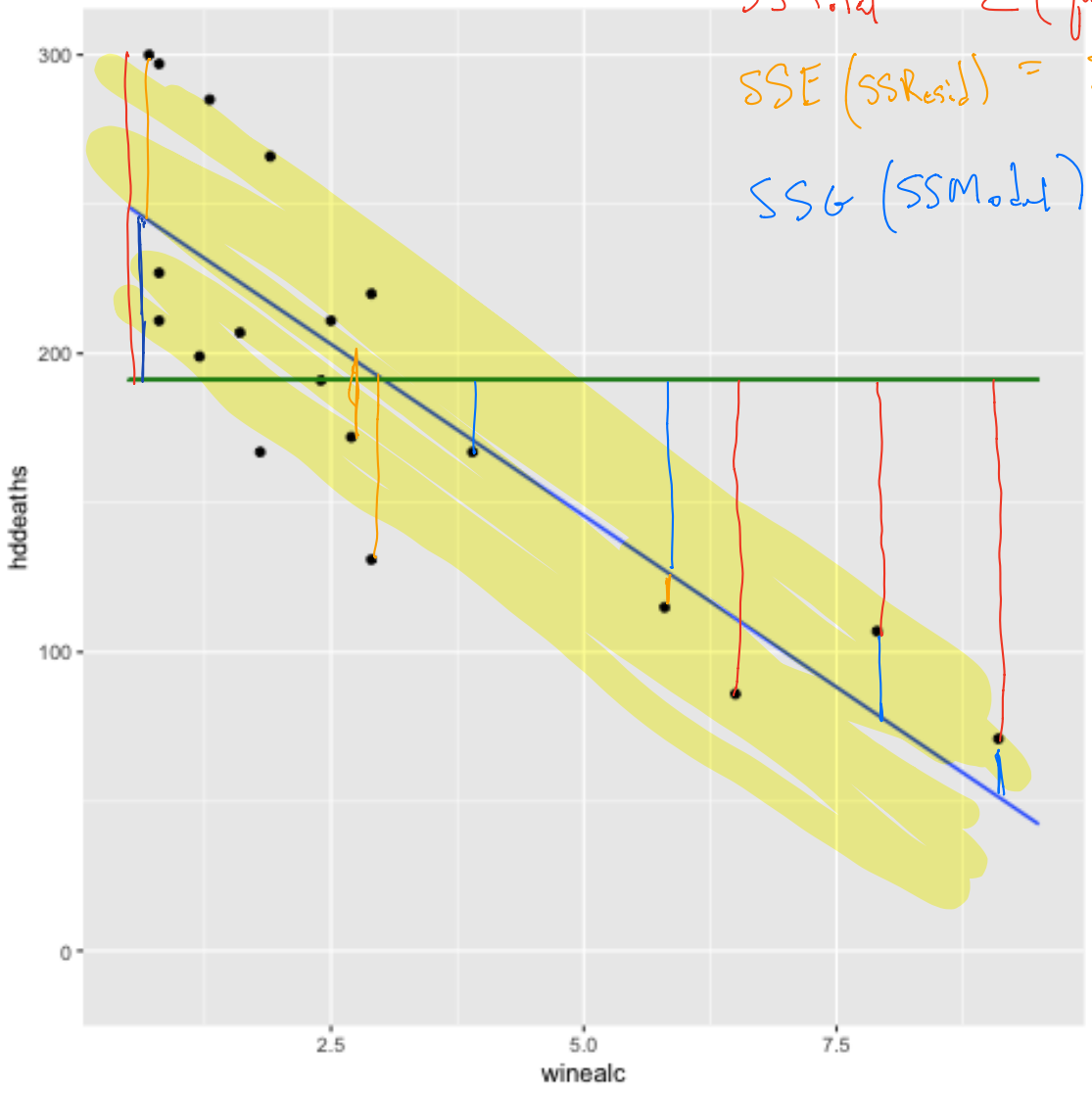
Comparison: ANOVA of Chapter 8 and ANOVA for regression (Chapter 9)



SSE = Σ squared residuals

SSG = Σ (squared lengths of blue lines)

SST = Σ (squared red lengths)



True Regression line $\beta_0 + \beta_1 x = \mu_y(x)$ = mean response at given x

Sample regression line $b_0 + b_1 x$ = best approx.

$$SST = SSG + SSE \quad \left(= SSM + SSResid \right)$$

Ratio $\frac{SSM}{SST}$ viewed as the amt. of variability in response (y) explained by the x-variable

$$R^2 = SSM/SST = (\text{correlation})^2$$

Assumptions for theoretical distribution use (Simple Linear Regression)

1. At each x, the values of y have a normal distribution
2. Spread (σ) same at each x

Your turn.

```
Spruces <- read.csv("http://scofield.site/teaching/data/csv/hesterberg/Spruce.csv")
```

Resp. var. Ht. change

Expl. var. Di. change

a) Make scatterplot

b) Check assumptions of SLR

c) ANOVA table

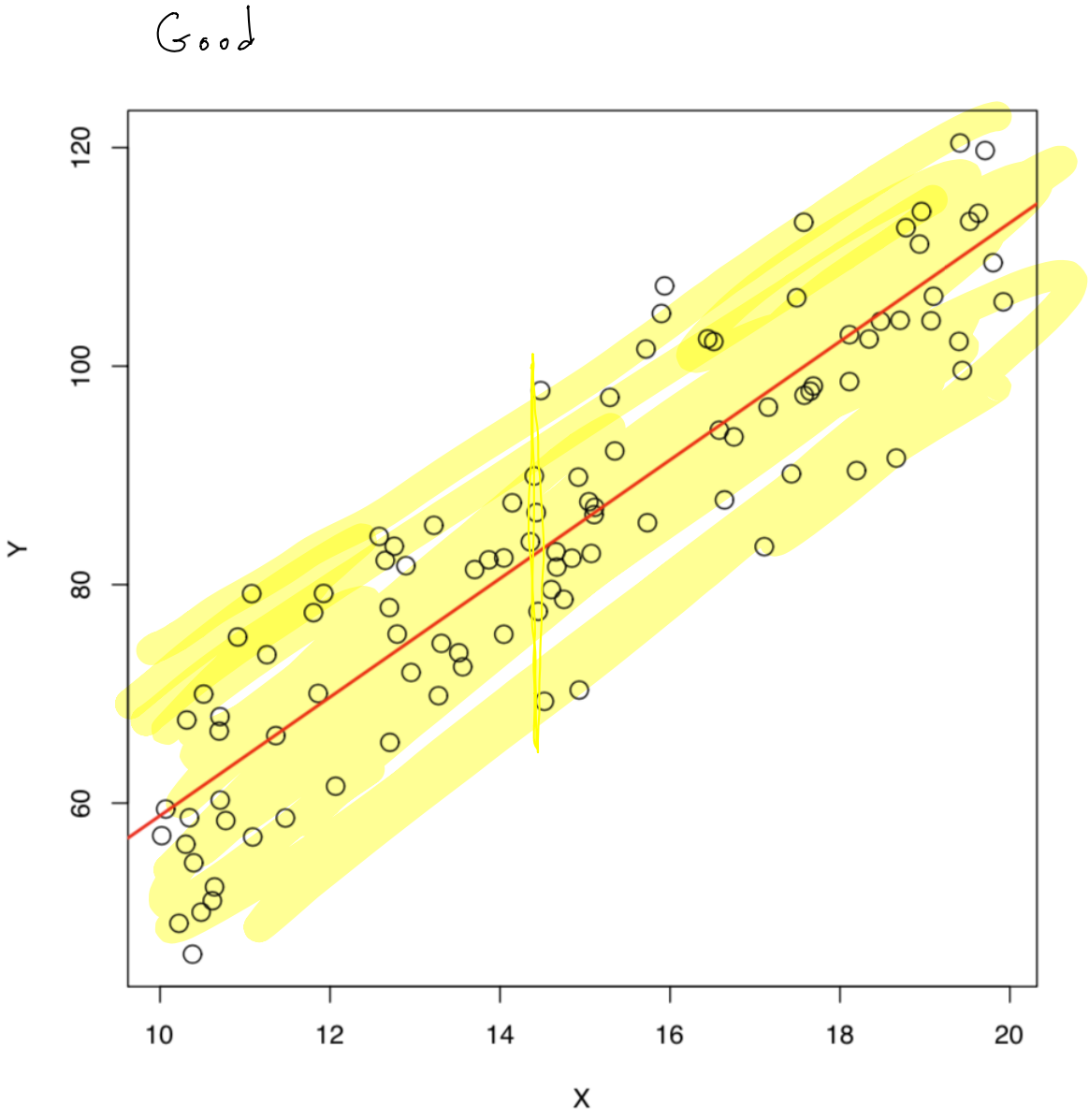


Figure 1: Good: conditions appear to be met

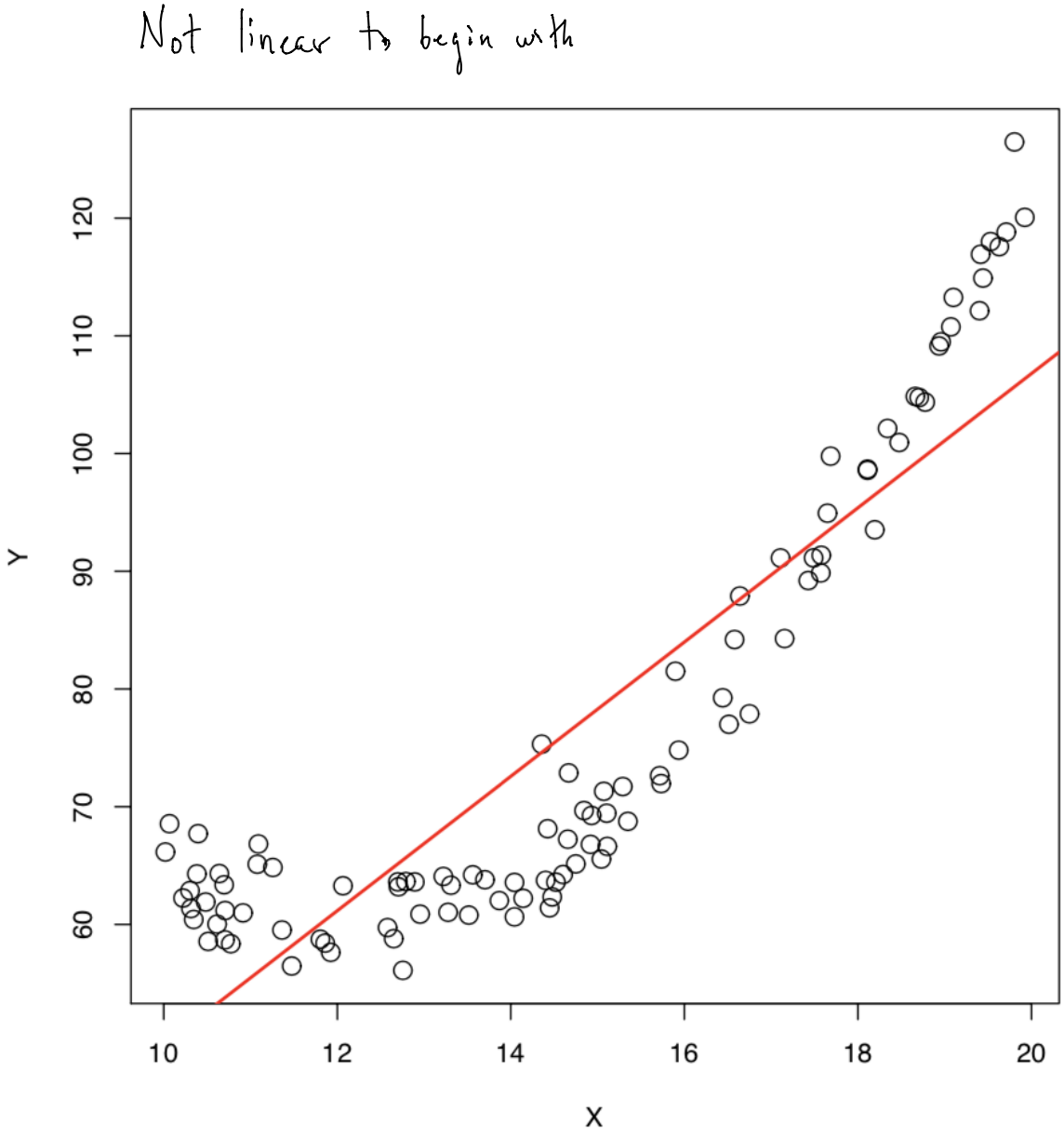


Figure 2: Bad: Not a linear association

Violates equal s.d. assumption of SLR
(data fans out as x changes)

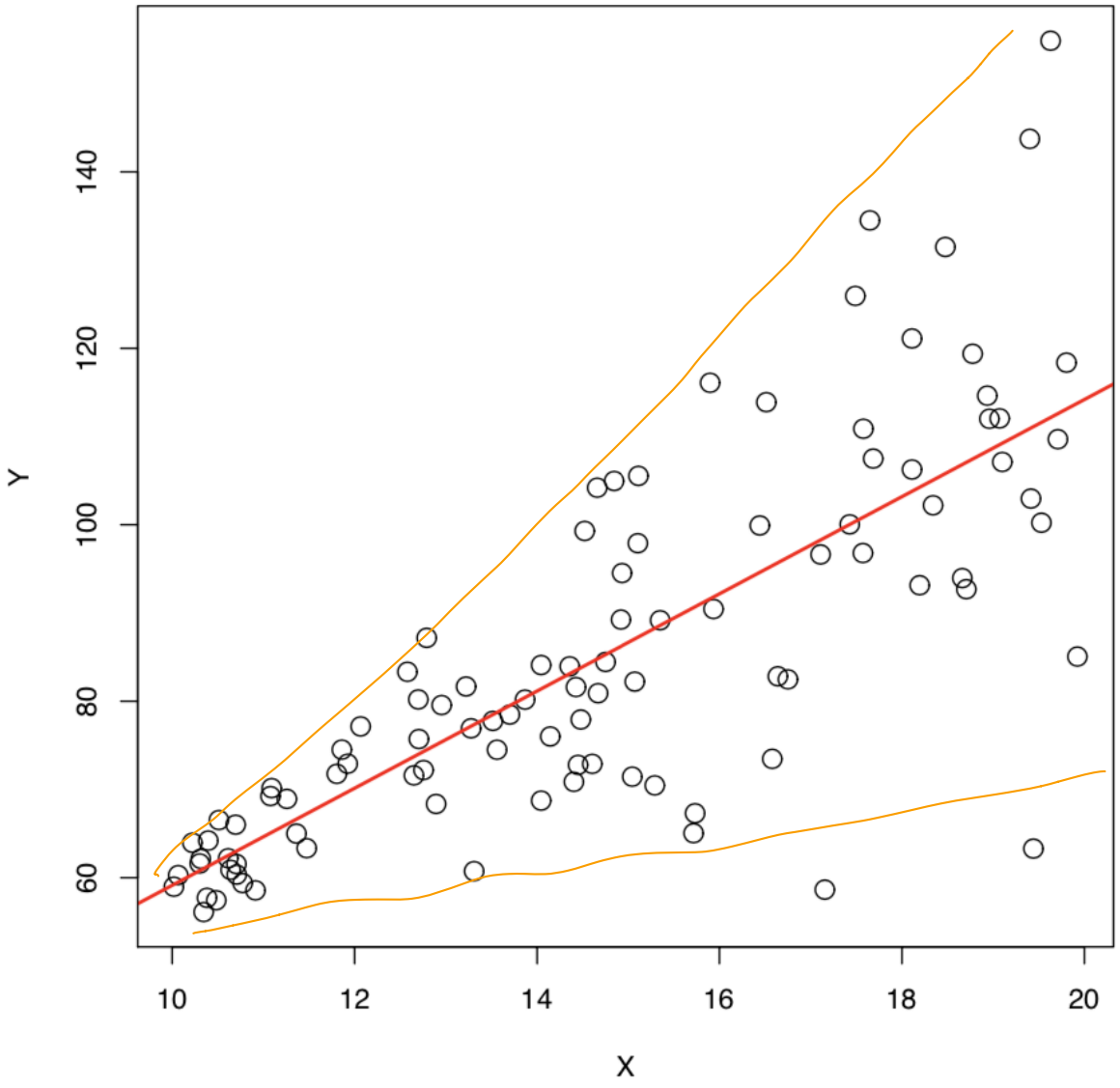


Figure 3: Bad: standard deviation changes with X

Simple Linear Regression Assumption #2 not met

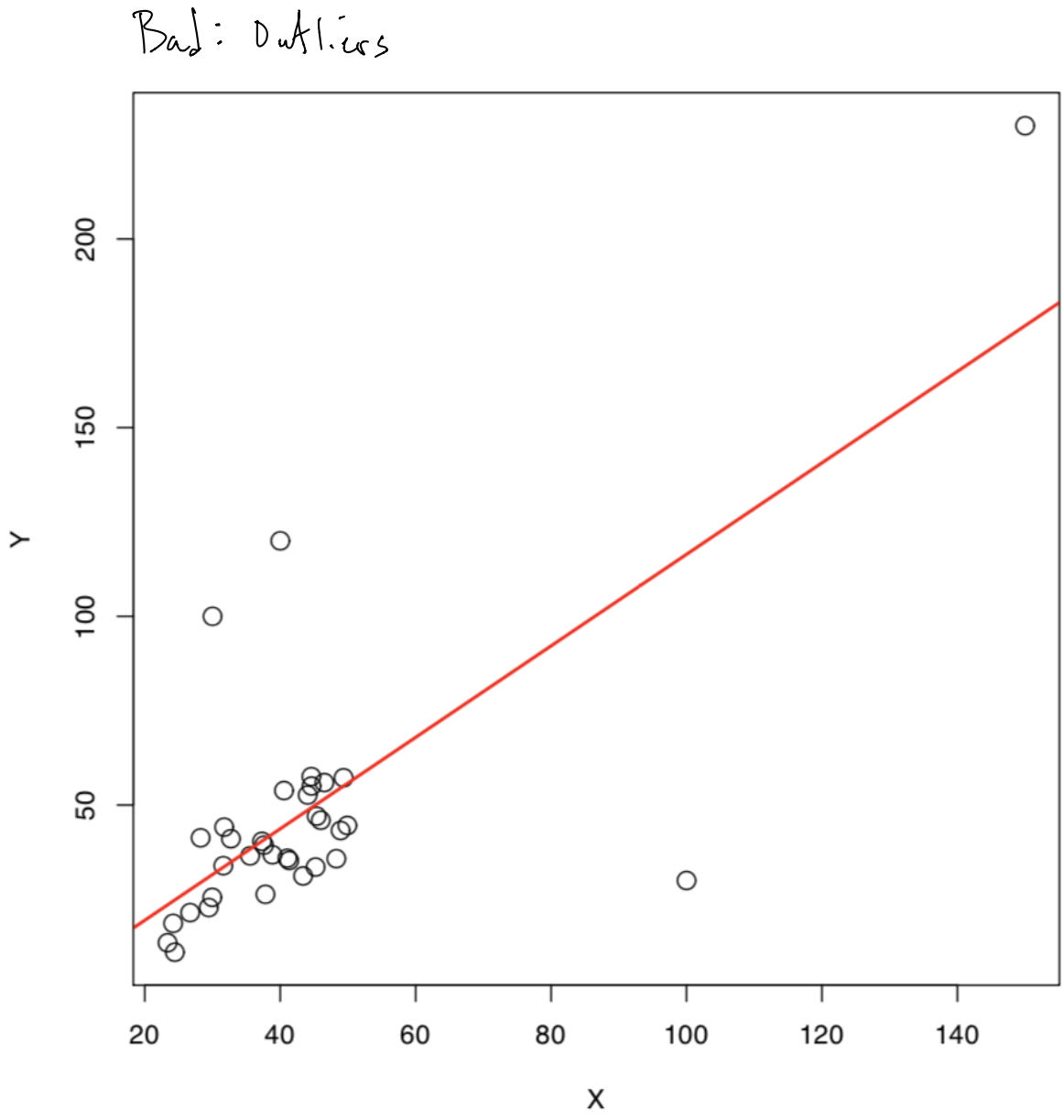


Figure 4: Bad: Outliers/influential points are present

ANOVA for regression

We can consider regression whenever we have bivariate quantitative data, a collection of n points which we can label

$$(x_1, y_1),\ \ (x_2, y_2),\ \ (x_3, y_3),\ \ \ldots\ \ (x_n, y_n).$$

Since Y is treated as a response variable, we call the y_1, y_2, …, y_n **observed** values.

One can ignore the x-values and calculate things like the mean response-value

$$\overline{y} \ = \ \frac{1}{n}(y_1 + y_2 + \cdots + y_n) \ = \ \frac{1}{n}\sum y_i,$$

or the **variance** (the square of the standard deviation)

$$s_y^2 \ = \ \frac{1}{n-1}\sum (y_i - \overline{y})^2.$$

And, since standard deviation and variance both measure how much variation there is between observed values, so does the quantity $(n-1)s_y^2$, which we again call SST, or

$$SSTotal \ = \ (n-1)s_y^2 \ = \ \sum (y_i - \overline{y})^2.$$

We can describe this relationship between SST and variance by saying that variance is SST divided by number of degrees of freedom $n-1$.

The computed regression line with intercept b_0 and slope b_1 has been used by us back in Chapter 2 to compute **predicted**, also known as **fitted**, values. We use a "hat" to distinguish these values from the observed ones:

$$\widehat{y}_1 = b_0 + b_1 x_1,\ \ \widehat{y}_2 = b_0 + b_1 x_2,\ \ \ldots,\ \ \widehat{y}_n = b_0 + b_1 x_n.$$

It is perhaps not so surprising that these fitted values have the same mean as the observed ones—that is

$$\frac{1}{n}(\widehat{y}_1 + \widehat{y}_2 + \cdots + \widehat{y}_n) \qquad \text{also equals } \ \overline{y}.$$

So, if we did the same type of sum-of-squares calculation as above, but using fitted values instead of observed ones, that would give a measure of the variability among fitted values. We will call that SSM, or

$$SSModel \ = \ \sum (\widehat{y}_i - \overline{y})^2.$$

Our fitted values all lie on the line, and the difference between an observed value y_i and its predicted value $\widehat{y}_i$ at x_i,

$$\epsilon_i \ = \ y_i - \widehat{y}_i,$$

is what we call a **residual**. You may remember that, in finding the best-fit line, we chose, out of all possible lines, the one that had the smallest possible sum-of-squares-of-residuals, sometimes called $SSResid$, or SSE (since *error* and *residual* are synonyms):

$$SSE \ = \ \sum (y_i - \widehat{y}_i)^2.$$

As when we defined similar quantities in Chapter 8, the relationship between them is

$$SSTotal = SSModel + SSResid, \qquad \text{or} \qquad SST = SSM + SSE.$$

When SSE is small in comparison with SST, that is indicative of a strong linear relationship between the variables; the line does a good job of explaining the variation in observed values. A good measure of how well the variability in response values Y is explained by the linear model $b_0 + b_1 X$ is the ratio

$$R^2 \ = \ \frac{SSM}{SST} \ = \ \frac{SST - SSE}{SST},$$

known as the **coefficient of determination**. It might have been better to use a lower-case r, and call it r^2, since the coefficient of determination is equal to the square of the correlation.

Example: `InkjetPrinters`. In the text, the Locks propose using `PPM`, the number of pages a printer can turn out per minute, to explain `Price`. The first few rows of the raw data are as follows.

```
head(InkjetPrinters)
```

```
##                                 Model PPM PhotoTime Price CostBW CostColor
## 1 HP Photosmart Pro 8500A e-All-in-One 3.9        67   300    1.6       7.2
## 2                    Canon Pixma MX882 2.9        63   199    5.2      13.4
## 3                  Lexmark Impact S305 2.7        43    79    6.9       9.0
## 4               Lexmark Interpret S405 2.9        42   129    4.9      13.9
## 5                  Epson Workforce 520 2.4       170    70    4.9      14.4
## 6                  Brother MFC-J6910DW 4.1       143   348    1.7       7.9
```

A scatterplot makes a linear relationship appear reasonable. The black points represent the data, while the purple points, lying on the line, are the *fits*. By storing the result of the `lm()` command, R can be asked to provide the fitted values (in the same order as the original data)

```
lmRes <- lm(Price ~ PPM, data=InkjetPrinters)
lmRes$fitted
```

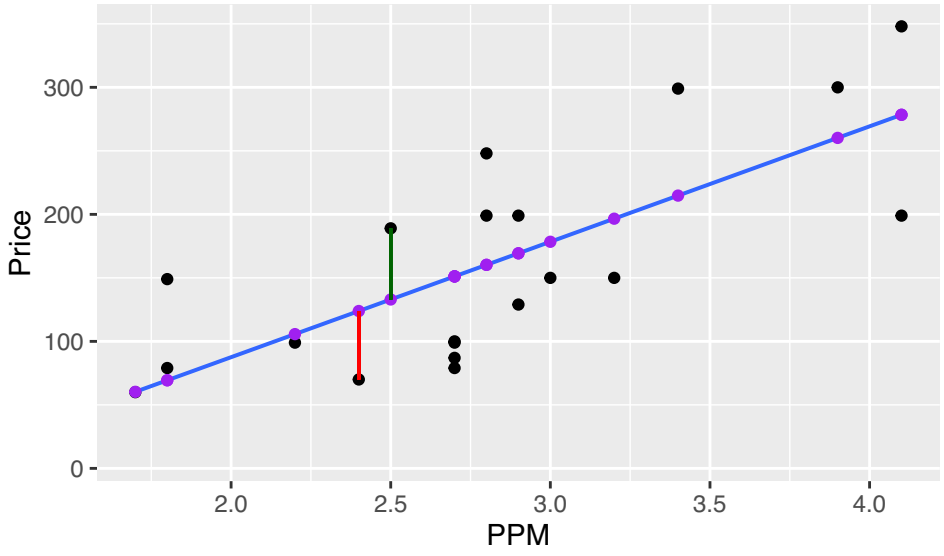
```
##         1         2         3         4         5         6         7         8
## 260.20270 169.32464 151.14902 169.32464 123.88560 278.37832 214.76367 160.23683
##         9        10        11        12        13        14        15        16
## 178.41244 196.58806 151.14902 151.14902 105.70999 132.97341 151.14902  60.27095
##        17        18        19        20
## 160.23683  69.35876  69.35876 278.37832
```

as well as the residuals

```
lmRes$residuals
```

```
##           1           2           3           4           5           6
##  39.7972965  29.6753642 -72.1490222 -40.3246358 -53.8856019  69.6216830
##           7           8           9          10          11          12
##  84.2363304  87.7631710 -28.4124425 -46.5880561 -64.1490222 -51.1490222
##          13          14          15          16          17          18
##  -6.7099884  56.0265913 -52.1490222  -0.2709545  38.7631710  79.6412387
##          19          20
##   9.6412387 -79.3783170
```

The vertical green line is from the 14th observed value, the observed price of \$189 for a Dell V715 w inkjet printer, down to its fitted price of \$132.97, a positive residual of $189 - 132.97 = 56.03$. The red vertical line is from the 5th observed value, the price of \$70 for an Epson Workforce 520 up to its fitted price of \$123.89, a negative residual of $70 - 123.89 = -53.89$.



Take a look at the output from the following commands applied to this data. First the summary from `lm()` (recall that `lmResult` stores output from `lm()`).

```
summary(lmRes)
```

```
##
## Call:
## lm(formula = Price ~ PPM, data = InkjetPrinters)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -79.38 -51.40  -3.49  43.85  87.76
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   -94.22      56.40  -1.671 0.112086
## PPM            90.88      19.49   4.663 0.000193 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 58.55 on 18 degrees of freedom
## Multiple R-squared:  0.5471, Adjusted R-squared:  0.522
## F-statistic: 21.75 on 1 and 18 DF,  p-value: 0.0001934
```

See there is a number reported at the bottom with the label `Multiple R-squared`. That is the coefficient of determination, R^2, we defined above. It says about 55% of the variability in sampled printer prices is "explained" by the variable `PPM` through the linear model

$$\widehat{\text{Price}} = -94.22 + 90.88(\text{PPM}).$$

Next look at the correlation:

```
cor(Price ~ PPM, data=InkjetPrinters)
```

```
## [1] 0.7396862
```

Squaring this correlation

$$(0.7396862)^2 \doteq 0.5471,$$

yields the same number as `Multiple R-squared` reported above.

Now look at an ANOVA table:

```
anova(lmRes)
```

```
## Analysis of Variance Table
##
## Response: Price
##           Df Sum Sq Mean Sq F value    Pr(>F)
## PPM        1  74540   74540  21.747 0.0001934 ***
## Residuals 18  61697    3428
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The reported sum-of-squares values are $SSM = 74540$ and $SSE = 61697$, which means $SST = 74540+61697 = 136237$. We defined R^2 to be the ratio

$$\frac{SSM}{SST} = \frac{74540}{136237} = 0.5471,$$

again matching `Multiple R-squared`.

Question

Consider the command

```
sum( lmRes$residuals^2 )
```

```
## [1] 61696.79
```

Can you guess what number this will give and find its value using the ANOVA table? Run the command and check that you are correct.

Looking more carefully at this ANOVA table, we see that, out of the $n = 20$ cases (printers) in the **InkjetPrinters** dataset, $n - 2 = 18$ degrees of freedom have been "assigned" to the `Residuals`, and 1 degree of freedom to `PPM`, for a total of $18 + 1 = 19 = n - 1$. In simple linear regression (i.e., regression with just 1 predictor variable), the number of degrees of freedom on the residual row is always $n - 2$. The calculations of quantities such as MSM (we called it MSG in Chapter 8), MSE and F which appear in the ANOVA table are done exactly as in 1-way ANOVA:

$$MSModel = \frac{SSModel}{1}, \qquad MSE = \frac{SSE}{n-2}, \qquad \text{and} \qquad F = \frac{MSM}{MSE},$$

and the resulting P-value, obtained with the command like

```
1 - pf(fstatistic, df1=1, df2=n-2)
```

is exactly the same as P-value from the Model Utility Test, representing another way to compute it.

Prediction and confidence intervals : Not yet discussed

If the conditions for the simple linear model are met, and if we have rejected the null hypothesis in the Model Utility Test in favor of the alternative, that the explanatory variable has some usefulness as a predictor of values of the response variable, it is typical to see the model used that way. There are two sorts of *prediction*-type questions we might ask.