SIMPLE LINEAR REGRESSION HOMEWORK SOLUTIONS

1. Suppose $z_i := \frac{x_i - \bar{x}}{s}$ for $i \in \{1, 2, \ldots, n\}$. Then we want to show

$$\sqrt{\frac{\sum_{i=1}^{n}(z_i - \bar{z})^2}{n-1}} = 1,$$

which is the same as showing

$$\sum_{i=1}^{n}(z_i - \bar{z})^2 = n-1.$$

We already saw that $\bar{z} = 0$, so we want to show

$$\sum_{i=1}^{n} z_i^2 = n-1.$$

The left-hand side here is

$$\sum_{i=1}^{n} z_i^2 = \sum_{i=1}^{n}\left(\frac{x_i - \bar{x}}{s}\right)^2 = \frac{1}{s^2}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

$$= \frac{1}{\sum_{i=1}^{n}\frac{(x_i - \bar{x})^2}{n-1}} \cdot \sum_{i=1}^{n}(x_i - \bar{x})^2 = n-1. \quad \blacksquare$$

2. (a) The equation of the least-squares regression line is

$$\widehat{\text{price}} = -12.15 * \frac{\$100}{yr} * \text{age} + 199.03 * \$100$$

Plugging in age = 4.5 yrs gives a price estimate of \$25,37050.

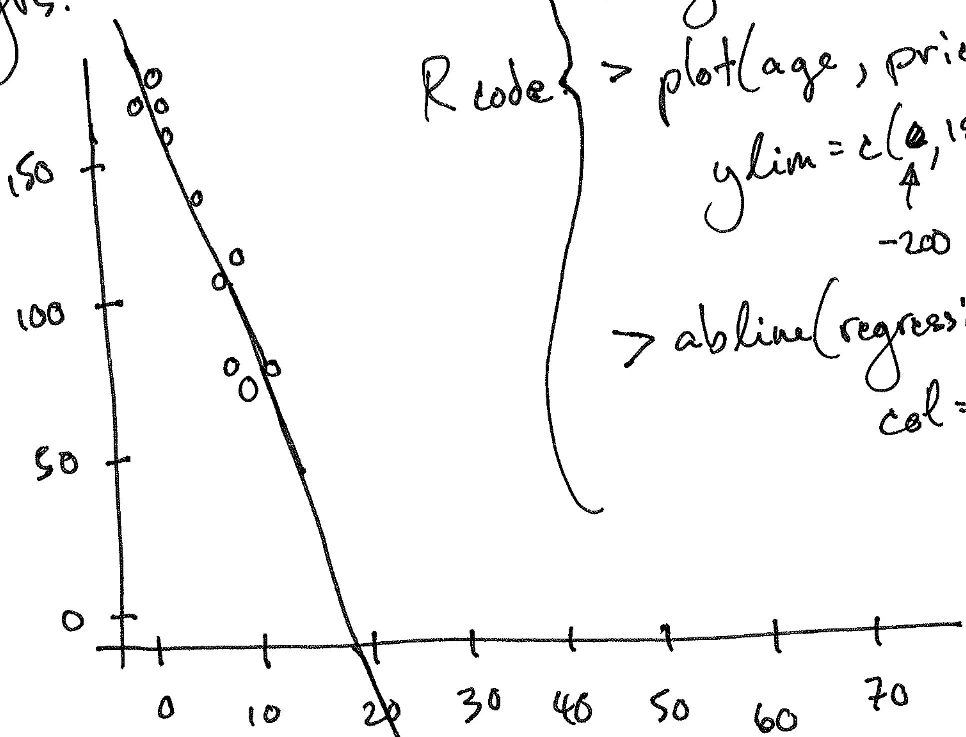
(b) The price estimate for a 60 yr. old car would be −\$55,097.00. This doesn't make any sense – it indicates a "fair" deal would be to pay someone \$55,097.00 to haul-off your 60-year old car. The problem here is an example of <u>extrapolation</u> which occurs when you attempt to make predictions of the response variable based on explanatory variable values beyond the

(#2 continued)

maximum or minimum previously observed explanatory values. You can see the problem easily when you plot the regression line and original scatterplot with the x-scale extended to 60 yrs:

R code

```
> regression <- lm(price ~ age)
> plot(age, price, xlim = c(0,65),
       ylim = c(0,180))
              -200
> abline(regression, lwd = 2,
         col = "blue")
```



It's likely that while a linear model is sufficient for prediction modeling when age $\in [0, 15]$ or so, that perhaps a quadratic model would be more appropriate over a larger time interval.

2(c)

```
> data <- used_cars
> data[11,1] <- 70
> data[11,2] <- 500
> data
   age price
1    3   172
2    6   140
3    8   112
4    4   160
5    2   165
6   11    80
7    4   155
8    7   103
9    7    84
10   9    78
11  70   500
> reg2 <- lm(data$price ~ data$age)
> reg2

Call:
lm(formula = data$price ~ data$age)

Coefficients:
(Intercept)     data$age
     93.226        5.523
```

The new regression line equation is

\hat{price} = ($552.3/yr) * age + $9322.60

You can get the r value like this:

```
> cor(data$price, data$age)
[1] 0.9060526
```

Or like this- which gives r^2… then just take the square root and attach the proper sign (r always has the same sign as the slope of the regression line):

```
> summary(reg2)

Call:
lm(formula = data$price ~ data$age)

Residuals:
   Min     1Q Median     3Q    Max
-73.98 -38.39  13.64  42.18  62.20
```

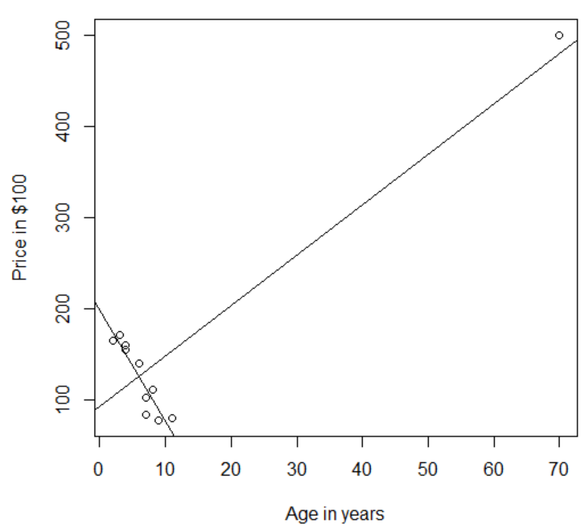
```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  93.2263    18.9534   4.919 0.000826 ***
data$age      5.5230     0.8598   6.423 0.000122 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 52.9 on 9 degrees of freedom
Multiple R-squared:  0.8209,    Adjusted R-squared:  0.801
F-statistic: 41.26 on 1 and 9 DF,  p-value: 0.0001219
```

So r^2 here is that "Multiple R-squared" value of 0.8209. So

```
> r <- sqrt(0.8209)
[1] 0.9060353
```

Note the two different methods give two slightly different answers due to round-off error. Which one is more accurate?

```
> plot(data[,1], data[,2], xlab="Age in years", ylab="Price in $100")
> abline(regression)
> abline(reg2)
```



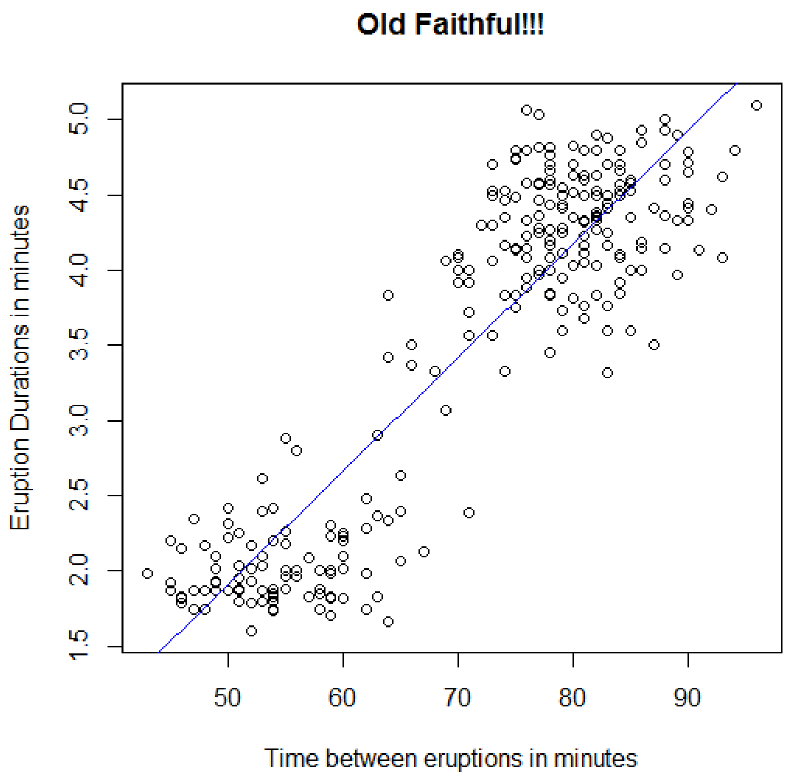
It seems the new regression line equation is not as good as the original one for predicting price based on ages of 0 to about 15 years. Note the r values are high in magnitude in each case, though. Probably to model the price behavior over age range of 0 to 70 years we need a lot more data with explanatory variable values between 10 years and 70 or more years.

This new observation way out at (70, 500) is having a huge leveraging effect on our model. It has dramatically changed our least-squares regression line equation. This point is an example of an *influential observation.* Such data points often fall far to the left or right of the rest of the data point pattern and their removal from the data set will substantially affect the model.

Note the r^2 value is high, yet we think the new model is probably not good. You can see what r^2 really measures: it measures how close, overall, data points are to the model. Even though the point pattern in the lower left-hand corner has a “down and to the left” trend, the points there are RELATIVELY CLOSE TO THE LINE… that is, relatively close with respect to the “big picture”, with age running from 0 to 70 years. This example should leave you feeling very cautious with respect to depending on r or r^2 as a measure of model goodness.

3. The “head()” command is nice… it displays the first few rows of a data set along with the variable names.

```
> head(faithful)
  eruptions waiting
1     3.600      79
2     1.800      54
3     3.333      74
4     2.283      62
5     4.533      85
6     2.883      55
> plot(faithful$waiting, faithful$eruptions, xlab="Time between eruptions in minutes",
+ ylab="Eruption Durations in minutes", main="Old Faithful!!!")
> ofreg <- lm(faithful$eruptions ~ faithful$waiting)
> abline(ofreg, col="blue")
> summary(ofreg)

Call:
lm(formula = faithful$eruptions ~ faithful$waiting)

Residuals:
     Min       1Q   Median       3Q      Max
-1.29917 -0.37689  0.03508  0.34909  1.19329

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)      -1.874016   0.160143  -11.70   <2e-16 ***
faithful$waiting  0.075628   0.002219   34.09   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.4965 on 270 degrees of freedom
Multiple R-squared:  0.8115,    Adjusted R-squared:  0.8108
F-statistic:  1162 on 1 and 270 DF,  p-value: < 2.2e-16
```



R^2 is relatively high…. 0.8115. This, along with the observation that there do not seem to be any really influential observations seems to indicate the regression line might be useful for making predictions. Note the SSE or MSE values can also be used to measure how well the line models the data (r^2 can be defined in terms of SSE, right?)

Anyway, note that there seem to be two point clusters here in the plot. What could be causing that effect? If indeed these clusters are due to, say, two levels of some unknown (or known) factor, and we separated the two groups, how good would our models be if we made a linear model for the cluster in the lower left and another for the point cluster in the upper right of the plot? Maybe not so good, right?

4. Show $b_1 = r \cdot \frac{s_y}{s_x}$.

$$\frac{d}{db_1}\sum_{i=1}^{n} e_i^2 = \frac{d}{db_1}\sum_{i=1}^{n}(y_i - b_1x_i - b_0)^2$$

$$= \sum_{i=1}^{n}\frac{d}{db_1}(y_i - b_1x_i - b_0)^2$$

The derivative of the sum of a finite # of terms is the sum of the derivatives

$$= \sum_{i=1}^{n} 2(y_i - b_1x_i - b_0)(-x_i)$$

Setting this derivative equal to zero gives

$$\sum_{i=1}^{n} x_i(y_i - b_1x_i - b_0) = 0$$

Subbing in $b_0 = \bar{y} - b_1\bar{x}$ here gives

$$0 = \sum_{i=1}^{n} x_i(y_i - b_1x_i - \bar{y} + b_1\bar{x}) = \sum_{i=1}^{n} x_iy_i - b_1\sum_{i=1}^{n}x_i^2 - \bar{y}\sum_{i=1}^{n}x_i + b_1\bar{x}\sum_{i=1}^{n}x_i.$$

$$\Longrightarrow \boxed{b_1 = \frac{n\bar{y}\bar{x} - \sum_{i=1}^{n}x_iy_i}{\bar{x}\sum_{i=1}^{n}x_i - \sum_{i=1}^{n}x_i^2}} \quad *$$

HOWEVER...

$$r \cdot \frac{S_y}{S_x} = \frac{1}{n-1}\left(\sum_{i=1}^{n} \frac{x_i - \bar{x}}{S_x} \cdot \frac{y_i - \bar{y}}{\cancel{S_y}}\right) \cdot \frac{\cancel{S_y}}{S_x}$$

$$= \frac{1}{(n-1)\,S_x^2} \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})$$

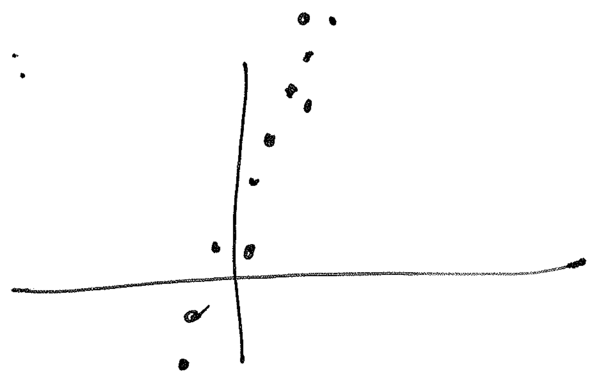
$$= \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2}$$

$$= \frac{\sum_{i=1}^{n} x_i y_i - \underbrace{\bar{x}\sum_{i=1}^{n} y_i}_{= n\bar{x}\bar{y}} - \underbrace{\bar{y}\sum_{i=1}^{n} x_i}_{= n\bar{x}\bar{y}} + n\bar{x}\bar{y}}{\sum_{i=1}^{n} x_i^2 - \underbrace{2\bar{x}\sum_{i=1}^{n} x_i}_{= 2n\bar{x}^2} + n\bar{x}^2}$$

$$= \frac{\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}}{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2}\,)$$

which is the same as (*).

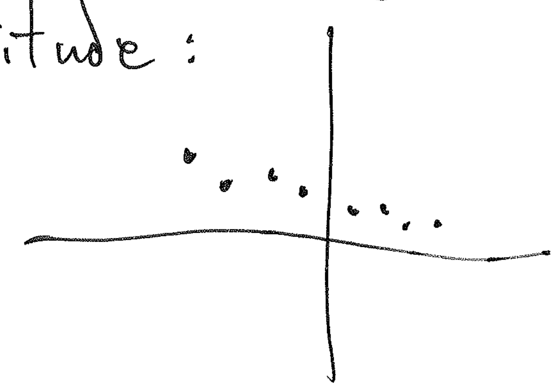
5. If $s_y > s_x$, the spread-out-ness of the data in the $\uparrow$ dimension exceeds that of the $\rightarrow$ dimension:



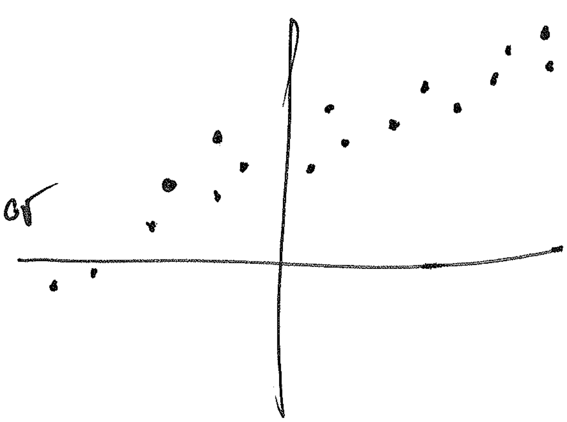
or

In such cases, the slope of the regression line b_1 is r, but multiplied by a factor > 1, so that ~~the~~ the slope is increased in magnitude.

Similarly, when $s_y < s_x$, the slope is r, but multiplied by a factor less than one in magnitude:



or



6. $b_1 = r\frac{s_y}{s_x}$, $b_0 = \bar{y} - b_1\bar{x}$.

Now, this means

$$\hat{y} = r\frac{s_y}{s_x}\cdot x + \bar{y} - b_1\bar{x}$$

Now, $\hat{z}_{y_i} = \frac{\hat{y}_i - \bar{y}}{s_y}$, $z_{x_i} = \frac{x_i - \bar{x}}{s_x}$,

So we are really being asked to show

$$\frac{\hat{y}_i - \bar{y}}{s_y} \stackrel{?}{=} r\cdot\frac{x_i - \bar{x}}{s_x}.$$

If this is true, then

$$\hat{y}_i - \bar{y} = \frac{s_y}{s_x} r x_i - \frac{s_y}{s_x} r\bar{x}$$

$$\iff \hat{y}_i = \overbrace{\frac{s_y}{s_x} r}^{\sim b_1} x_i - \overbrace{\frac{s_y}{s_x} r}^{\sim b_1}\bar{x} + \bar{y}$$

Which is the same as this. We've shown that the "best fitting" line through the standardized points has slope r and runs through the origin.

7.

Coef. of determination ~~r^2~~ $:= \dfrac{\sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2}{\sum_{i=1}^{n}(y_i-\bar{y})^2}$

little r... the correlation coeficient.

$$= \frac{\sum_{i=1}^{n}\left(\frac{rS_y}{S_x}x_i + \cancel{\bar{y}} - \frac{rS_y}{S_x}\bar{x} - \cancel{\bar{y}}\right)^2}{\sum_{i=1}^{n}(y_i-\bar{y})^2}$$

$$= \frac{\sum_{i=1}^{n}\left(\frac{rS_y}{S_x}x_i - \frac{rS_y}{S_x}\bar{x}\right)^2}{\sum_{i=1}^{n}(y_i-\bar{y})^2}$$

$$= \left(r\frac{S_y}{S_x}\right)^2 \frac{\sum_{i=1}^{n}(x_i-\bar{x})^2}{\sum_{i=1}^{n}(y_i-\bar{y})^2}$$

$$= r^2 \cdot \frac{S_y^2}{S_x^2} \cdot \frac{\sum_{i=1}^{n}(x_i-\bar{x})^2/n-1}{\sum_{i=1}^{n}(y_i-\bar{y})^2/n-1} = r^2 \cdot 1 = r^2.$$

$= S_x^2$

$= S_y^2$

8. $MSE(\hat{\theta}) := E[(\hat{\theta}-\theta)^2]$

$$= E[\hat{\theta}^2 + \theta^2 - 2\hat{\theta}\theta]$$

$$= E[\hat{\theta}^2] + \theta^2 - 2\theta E[\hat{\theta}]$$

Now, $B(\hat{\theta}) := E[\hat{\theta}] - \theta$, and so

$$(B(\hat{\theta}))^2 = (E[\hat{\theta}])^2 + \theta^2 - 2\theta E[\hat{\theta}].$$

Also, $Var(\hat{\theta}) = E[\hat{\theta}^2] - (E[\hat{\theta}])^2$.

Then $(B(\hat{\theta}))^2 + Var(\hat{\theta}) = E[\hat{\theta}^2] + \theta^2 - 2\theta E[\hat{\theta}]$

which is the same as this.

9. We want to show

$$E\left[\frac{1}{n-1}\sum_{k=1}^{n}\left(Y_k-\overline{Y}\right)^2\right] = \sigma^2 .$$

Starting with the left-hand side here ...

$$\frac{1}{n-1}E\left[\sum_{k=1}^{n}\left(Y_k-\overline{Y}\right)^2\right] = \frac{1}{n-1}E\left[\sum_{k=1}^{n}Y_k^2+\sum_{k=1}^{n}\overline{Y}^2-2\overline{Y}\sum_{k=1}^{n}Y_k\right]$$

$$= \frac{1}{n-1}E\left[\sum_{k=1}^{n}Y_k^2+n\overline{Y}^2-2n\overline{Y}^2\right]$$

$$= \frac{1}{n-1}E\left[\sum_{k=1}^{n}Y_k^2-n\overline{Y}^2\right]$$

$$= \frac{1}{n-1}\left(nE\left[Y_1^2\right]-nE\left[\overline{Y}^2\right]\right)$$

$$= \frac{n}{n-1}\left(Var(Y_1)+(EY_1)^2-Var(\overline{Y})-(E\overline{Y})^2\right)$$

$$= \frac{n}{n-1}\left(\sigma^2+\mu^2-\frac{\sigma^2}{n}-\mu^2\right) = \frac{n}{n-1}\left(\sigma^2-\frac{\sigma^2}{n}\right)$$

$$= \frac{n}{n-1}\left(\frac{\sigma^2(n-1)}{n}\right) = \sigma^2$$

Recall: $Var(X) = E[X^2]-(EX)^2$ and $E[\overline{Y}] = \mu$, $Var(\overline{Y}) = \frac{\sigma^2}{n}$

This result is the main reason we use $\frac{1}{n-1}\sum_i (Y_i-\overline{Y})^2$ to estimate σ^2 as opposed to $\frac{1}{n}\sum_i (Y_i-\overline{Y})^2$.

10. Recall the bias of an estimator $\hat{\theta}$ is defined to be

$$B(\hat{\theta}) := E[\hat{\theta}] - \theta.$$

So this means

$$B\left(\frac{1}{n}\sum_{k=1}^{n}(Y_k - \bar{Y})^2\right) = E\left[\frac{1}{n}\sum_{k=1}^{n}(Y_k - \bar{Y})^2\right] - \sigma^2$$

$$= \frac{1}{n}\cdot(n-1)\cdot E\left[\frac{1}{n-1}\sum_{k=1}^{n}(Y_k - \bar{Y})^2\right] - \sigma^2$$

We already know this is $\underline{\underline{\sigma^2}}$!!!

$$= \frac{1}{n}\cdot(n-1)\sigma^2 - \sigma^2$$

$$= \sigma^2\left(\frac{n-1}{n} - 1\right)$$

$$= \sigma^2\left(\frac{n-1-n}{n}\right)$$

$$= \frac{-\sigma^2}{n}.$$

Notice this bias disappears as n goes to infinity (as the sample size increases)

This means when we use $\frac{1}{n}\sum_k (Y_k - \bar{Y})^2$ to estimate σ^2 it will tend to fall to the left of σ^2, on the average.