Problems Sheet 5

Statistics 1

Dom Hutchinson

```
load(url("https://people.maths.bris.ac.uk/~maxca/stats1/stats1.RData"))
```

Part A

Question 1

a) Let μ_N be the mean N_2O in the new engines & μ_O be the N_2O in the old engines.
Define hypothesis test $H_0 : \mu_N = \mu_O$ & $H_1 : \mu_N < \mu_O$.
Since both populations have the same variance a pooled estimate t-test is appropriate.

b) Let μ_M be the mean systolic blood pressure in men & μ_W be the mean systolic blood pressure in women.
Define hypothesis test $H_0 : \mu_M = \mu_W$ & $H_1 : \mu_M \neq \mu_W$.
Since the two samples are independent & variances are unknown a two-sample t-test is appropriate.

c) Let μ_A be the mean reaction time after an alcholic drink & μ_B be the mean reaction before an alcoholic drink.
Define hypothesis tests $H_0 : \mu_A = \mu_B$ & $H_1 : \mu_A < \mu_B$.
Since the samples are paired & the variances are unknown a two-sample t-test is appropritate.

d) Let μ be the mean weight of a chocolate bar.
Define hypothesis test $H_0 : \mu = 50$ & $H_1 : \mu < 50$.
Since there is only one sample & the variance is unknown a one-sample t-test is approperiate.

Question 4

Let S_i denote the time of runner i at sea-level & A_i denote the time of runner i at high altitude. Define $D_i := A_i - S_i$ to be how much slower runner i was at sea-level compared to high altitude.
We assume that $W_1, \dots, W_8$ is a simple random sample from $N(\mu_w, \sigma_w^2)$ distribution with μ_w & σ_w^2 unknown.
Define hypothesis test $H_0 : \mu_w = 0$ & $H_1 : \mu_w > 0$.

```
W=runner$high-runner$sea # Sample of time differences
WBar<-mean(W)            # Sample mean
S2_W<-var(W)             # Sample variance
cat("WBar=",WBar,", S2_W=",S2_W,sep="")
```

```
## WBar=1.175, S2_W=2.290714
```

Define test statistic $T(W_1, \dots, W_8) = \sqrt{n}\dfrac{\bar{W}}{\sigma_w} \sim t_7$.
Thus

```
t_obs<-sqrt(length(W))*(WBar/sqrt(S2_W)) # Observed test statistic
cat("t_obs=",t_obs,sep="")
```

```
## t_obs=2.195823
```

Thus $p = \mathbb{P}(T > t_{obs} | H_0 \text{ true}) = \mathbb{P}(t_7 > 2.195823) = 1 - \mathbb{P}(t_7 < 2.195823)$

```r
cat("p = 1-",pt(t_obs,7)," = ",1-pt(t_obs,7),sep="") # p-value
```

```
## p = 1-0.9679374 = 0.03206257
```

Since $p = 0.0321 < 0.05 = \alpha$ there is strong evidence to reject H_0 in favor of H_1.
Thus I conclude that running at high altidude did have an affect on the runners' times, when compared to their times at sea-level.

Part B

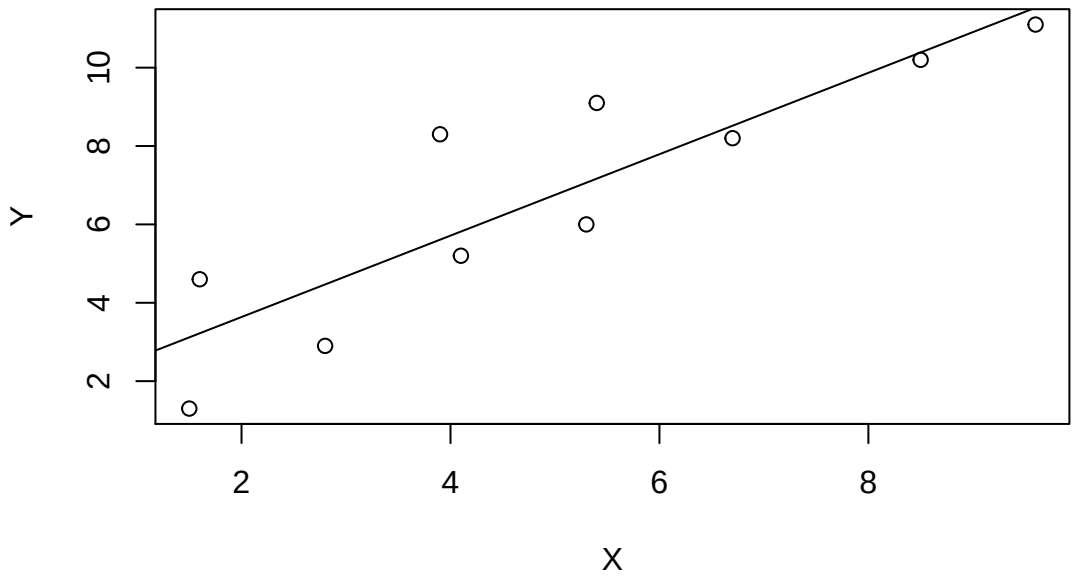
Question 4

a)

```r
X=rain$spring # Define samples
Y=rain$autumn

ss_xy<-sum(X*Y)-length(X)*mean(X)*mean(Y)
ss_xx<-sum(X^2)-length(X)*(mean(X)^2)

bHat<-ss_xy/ss_xx
aHat<-mean(Y)-bHat*mean(X)

cat("ss_xy=",ss_xy,", ss_xx=",ss_xx,", aHat=",aHat,", bHat=",bHat,sep="")
```

```
## ss_xy=69.774, ss_xx=67.184, aHat=1.559559, bHat=1.038551
```

```r
plot(X,Y)
abline(aHat,bHat)
```

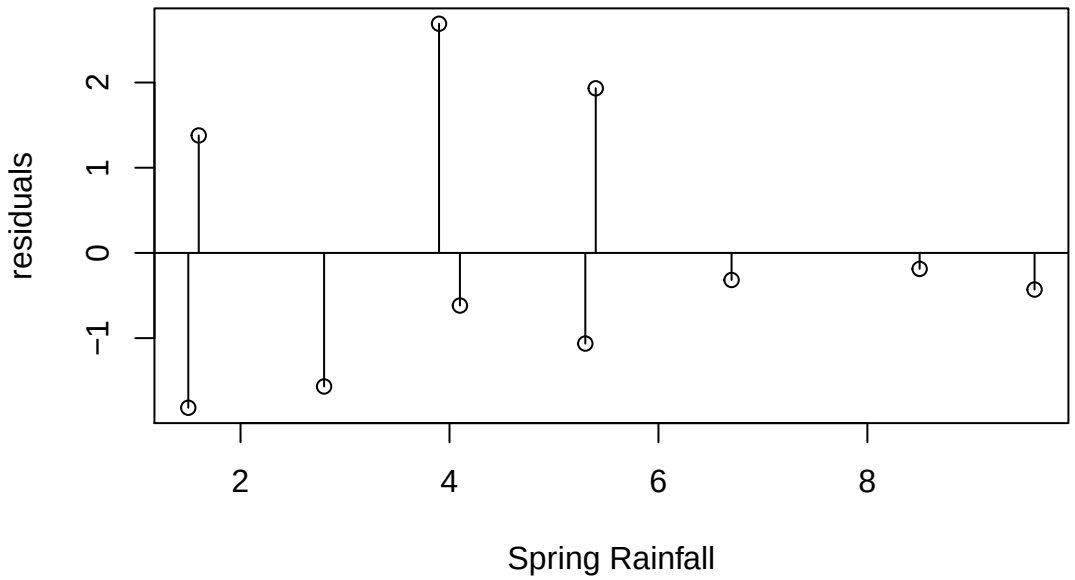


b)

```
fitted<-aHat+bHat*X
residuals<-Y-fitted
m<-matrix(c(round(fitted,2),round(residuals,2)),10,2)
colnames(m)<-c("fitted","residuals")
rownames(m)<-(1:10)
t(m)
```

```
##              1     2     3     4     5     6    7     8     9   10
## fitted    3.22  7.06  4.47 11.53  8.52  3.12 7.17 10.39  5.82 5.61
## residuals 1.38 -1.06 -1.57 -0.43 -0.32 -1.82 1.93 -0.19 -0.62 2.69
```

```
plot(X,residuals,xlab="Spring Rainfall")
segments(X,0,X,residuals)
abline(0,0)
```



From the residuals plot there is little evidence of systematic error, so the linear regression model is sufficiently good to use.

Question 5

Consider a model where the response variable Y is defined as $Y_i := \gamma x_i + e_i$ where x_i are values from the predictor variable & e_i are iid with $e_i \sim N(0, \sigma^2)$.
For this model $\mathbb{E}(Y_i|x_i) = \gamma x_i$ so the least squares estimate is the value that minimises $\sum_{i=1}^{n}(y_i - \gamma x_i)^2$.
These values can be found by setting the derivate of this sum to be 0.

$$
\begin{array}{rrcl}
 & \frac{\partial}{\partial\gamma}\sum_{i=1}^{n}(y_i-\gamma x_i)^2 & = & 0 \\
\implies & \frac{\partial}{\partial\gamma}\sum_{i=1}^{n} y_i^2-2\gamma x_i y_i+\gamma^2 x_i^2 & = & 0 \\
\implies & \sum_{i=1}^{n} -2x_i y_i+2\gamma x_i^2 & = & 0 \\
\implies & \gamma\sum_{i=1}^{n} x_i^2 & = & \sum_{i=1}^{n} x_i y_i \\
\implies & \hat{\gamma} & = & \frac{\sum_{i=1}^{n} x_i y_i}{\sum_{i=1}^{n} x_i^2}
\end{array}
$$

We estimate σ^2 as $\hat{\sigma}^2 = \frac{RSS}{n-2}$.

$$
\begin{array}{rrcl}
\text{We have} \quad RSS & \sum_{i=1}^{n}\hat{e}_i^2 & & \\
 & \sum_{i=1}^{n}(y_i-\hat{y}_i)^2 & & \\
 & \sum_{i=1}^{n}(y_i-\hat{\gamma}x_i)^2 & & \\
 & \sum_{i=1}^{n} y_i^2-2\hat{\gamma}x_i y_i+\hat{\gamma}^2 x_i^2 & & \\
 & \sum_{i=1}^{n} y_i^2-2\hat{\gamma}\sum_{i=1}^{n} x_i y_i+\hat{\gamma}^2\sum_{i=1}^{n} x_i^2 & & \\
 & \sum_{i=1}^{n} y_i^2-2\frac{\left(\sum_{i=1}^{n} x_i y_i\right)^2}{\sum_{i=1}^{n} x_i^2}+\frac{\left(\sum_{i=1}^{n} x_i y_i\right)^2\sum_{i=1}^{n} x_i^2}{\left(\sum_{i=1}^{n} x_i^2\right)^2} & & \\
 & \sum_{i=1}^{n} y_i^2-\frac{\left(\sum_{i=1}^{n} x_i y_i\right)^2}{\sum_{i=1}^{n} x_i^2} & & \\
\implies \quad \hat{\sigma^2} & \frac{1}{n-2}\left(\sum_{i=1}^{n} y_i^2-\frac{\left(\sum_{i=1}^{n} x_i y_i\right)^2}{\sum_{i=1}^{n} x_i^2}\right) & &
\end{array}
$$