Homework 1 - Solutions

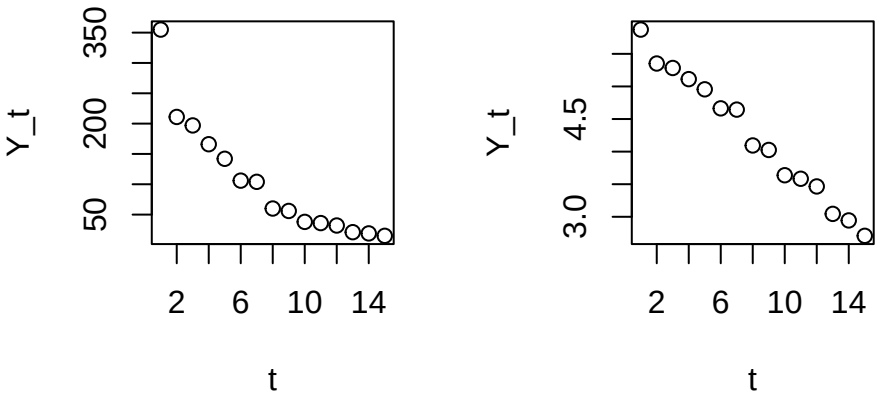
STAT 212

Problem A

(i)

(a)

```
df <- read.table("BacteriaDeath.txt", header = TRUE)
par(mfrow = c(1, 2))
time <- df$t
y <- df$Y_t
plot(time, y, xlab = "t", ylab = "Y_t")
plot(time, log(y), xlab = "t", ylab = "Y_t")
```



The log data appears more linear.

(b)

For the untransformed count:

```
lm1 <- lm(y ~ time)
lm1$coefficients
```

```
## (Intercept)        time
##   259.58095   -19.46429
```

The predictive equation is:

$$Y_t = 259.58 - 19.46t \tag{1}$$

For the log transformed count:

```
lm2 <- lm(log(y) ~ time)
lm2$coefficients
```

```
## (Intercept)        time
##   5.9731603  -0.2184253
```

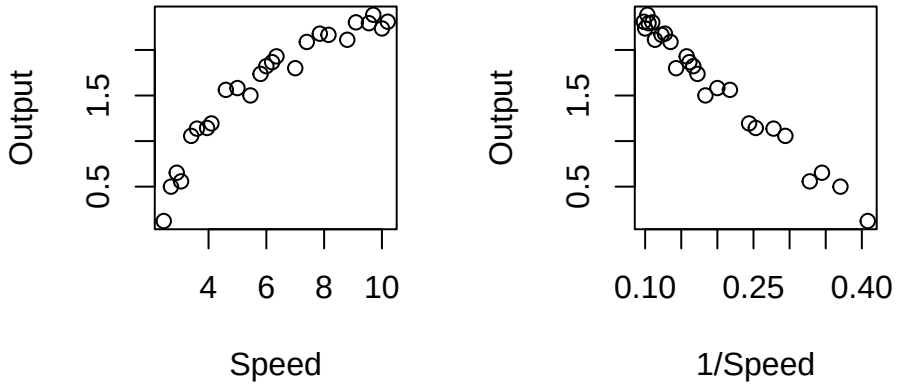
The predictive equation is:

$$Y_t = \exp(5.97 - 0.22t) \tag{2}$$

(ii)

(a)

```
df2 <- read.table("WindSpeed.txt", header = TRUE)
x <- df2$speed
y <- df2$output
par(mfrow = c(1, 2))
plot(x, y, xlab = "Speed", ylab = "Output")
plot(1/x, y, xlab = "1/Speed", ylab = "Output")
```



The plot of output versus the the reciprocal of wind speed looks more linear.

(b)

```
lm_wind <- lm(y ~ I(1/x))
lm_wind$coefficients
```

```
## (Intercept)      I(1/x)
##    2.978860   -6.934547
```

The regression line is:

$$\text{Output} = 2.98 - 6.93 \times (1/\text{Speed}) \tag{3}$$

(c)

The predicted output at speed = 8 is:

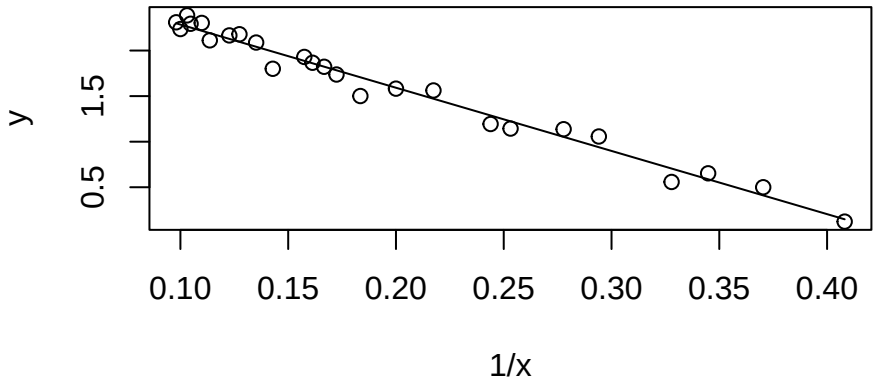
```
lm_wind$coefficients[1] + lm_wind$coefficients[2] * (1/8)
```

```
## (Intercept)
##    2.112042
```

Problem B

(i)

```
lm_wind <- lm(y ~ I(1/x))
plot(1 / x, y)
lines(1 / x, lm_wind$fitted.values)
```

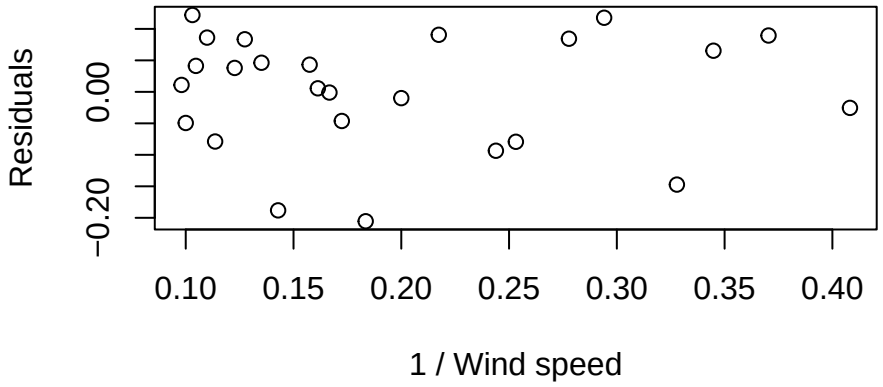


(ii)

```
resids <- lm_wind$residuals
preds <- lm_wind$fitted.values
```

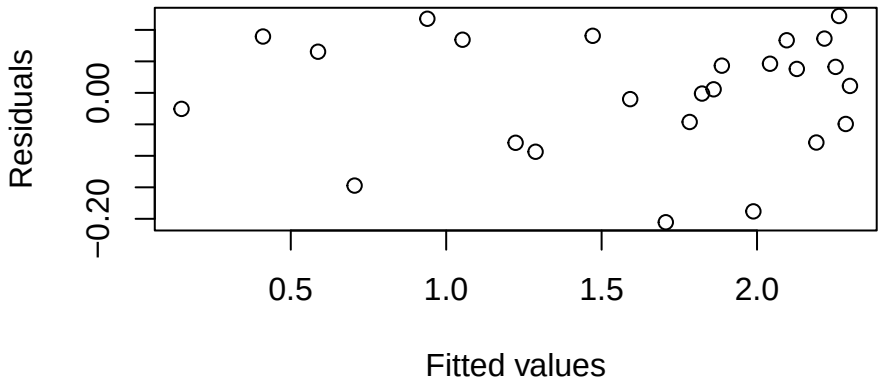
We do not see evidence of nonlinearity in the plot of residuals versus covariate:

```
plot(1 / x, resids, xlab = "1 / Wind speed", ylab = "Residuals")
```

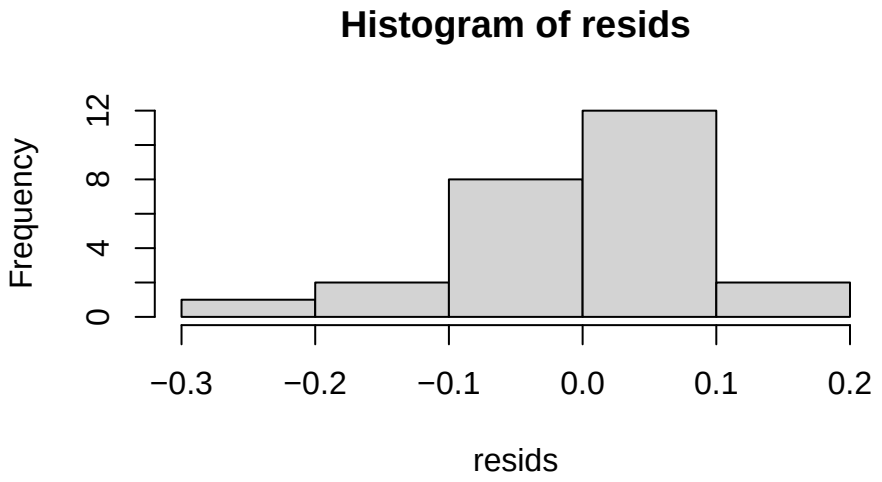


We do not see heteroskedasticity in the plot of residuals versus fitted values:

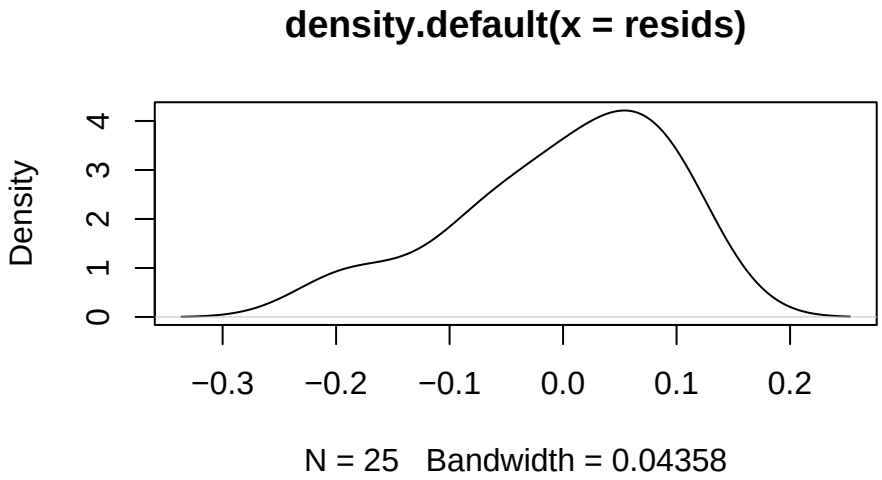
```
plot(preds, resids, xlab = "Fitted values", ylab = "Residuals")
```



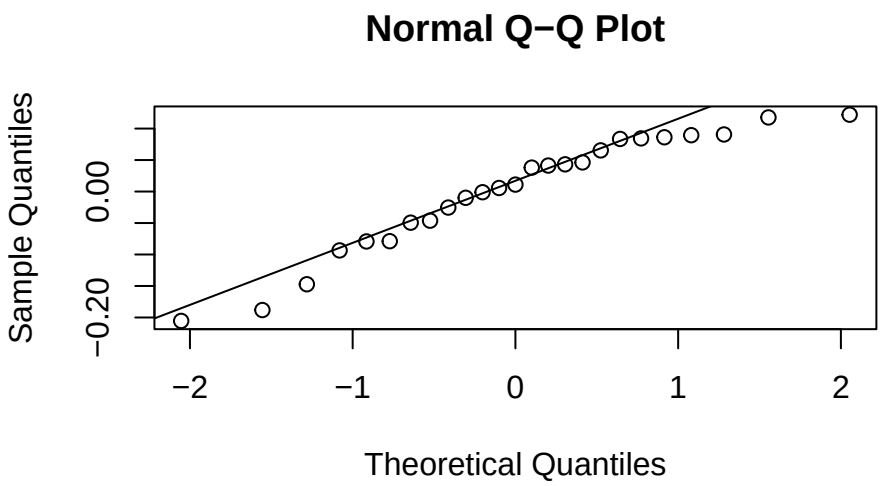

```
hist(resids, breaks = "FD")
```




```
plot(density(resids))
```




```
qqnorm(resids)
qqline(resids)
```



From the histogram, kernel density estimate, and qqplot of the residuals, we see that the distribution of residuals is skewed, with larger tails than the normal distribution.

(iii)

```
summ <- summary(lm_wind)
summ$r.squared
```

```
## [1] 0.9800249
```

98% of the variation in output is explained by the linear relationship between output and (1 / wind speed).

(iv)

```
confint(lm_wind, level = .99)
```

```
##                 0.5 %    99.5 %
## (Intercept)  2.852804  3.104916
## I(1/x)      -7.514076 -6.355019
```

The 99% confidence interval for β_1 is $[-7.51, -6.36]$.

(v)

```
new_val <- data.frame(x = 3.2)
predict(lm_wind, newdata = new_val, interval = "confidence", level = .95)
```

```
##         fit       lwr       upr
## 1 0.8118141 0.7491112 0.8745171
```

95% confidence interval for average output given wind speed is 3.2 is $[0.75, 0.87]$.

(vi)

```
new_val <- data.frame(x = 9.05)
predict(lm_wind, newdata = new_val, interval = "prediction", level = .95)
```

```
##        fit      lwr      upr
## 1 2.212612 2.010505 2.414719
```

95% prediction interval for output at this wind mill given wind speed there is 9.05 is $[2.01, 2.41]$.

Note on (v) and (vi)

In the above solutions, we used the `I()` notation for R formulas. This allowed us to enter the new observation `x` into the predict function without taking the reciprocal. If you "manually" transform `x` outside of the formula, as in

```
x_inv <- 1 / x
lm_wind_manual <- lm(y ~ x_inv)
```

then you will have to remember to transform new `x` values when making predictions:

```
predict(lm_wind_manual, newdata = data.frame(x_inv = 1 / 3.2),
        interval = "confidence", level = .95)
```

```
##         fit       lwr       upr
## 1 0.8118141 0.7491112 0.8745171
```

Problem C

(i)

Solving for b_0:

$$\begin{aligned}
\frac{\partial f(b_0, b_1)}{\partial b_0} &= -2\sum_{i=1}^{n}[y_i - (b_0 + b_1 x_i)] = 0 && \text{set derivative to 0} \\
\sum_{i=1}^{n} y_i - \sum_{i=1}^{n}(b_0 + b_1 x_i) &= 0 && \text{divide both sides by -2, split up sum} \\
n\frac{1}{n}\sum_{i=1}^{n} y_i - \sum_{i=1}^{n}(b_0 + b_1 x_i) &= 0 && \text{multiply/divide by } n \text{ to get } \bar{y} \\
n\bar{y} - nb_0 - b_1\sum_{i=1}^{n} x_i &= 0 && \text{definition of sample mean, sum of constant} \\
nb_0 &= n\bar{y} - nb_1\bar{x} && \text{rearrange, sample mean definition again} \\
b_0 &= \bar{y} - b_1\bar{x}
\end{aligned}$$

Solving for b_1:

$$\begin{aligned}
\frac{\partial f(b_0, b_1)}{\partial b_1} &= -2\sum_{i=1}^{n}[y_i - ((\bar{y} - b_1\bar{x} + b_1 x_i)]x_i = 0 \\
\sum_{i=1}^{n}(y_i x_i - \bar{y}x_i + b_1\bar{x}x_i - b_1 x_i^2) &= 0 && \text{divide both sides by } -2 \\
\sum_{i=1}^{n} y_i x_i - \bar{y}\sum_{i=1}^{n} x_i + b\bar{x}\sum_{i=1}^{n} x_i - b_1\sum_{i=1}^{n} x_i^2 &= 0 && \text{split sum} \\
b_1(\bar{x}\sum_{i=1}^{n} x_i - \sum_{i=1}^{n} x_i^2) &= \bar{y}\sum_{i=1}^{n} x_i - \sum_{i=1}^{n} x_i y_i \\
b_1 &= \frac{\sum_{i=1}^{n} x_i y_i - \bar{y}\sum_{i=1}^{n} x_i}{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2}
\end{aligned}$$

Consider the denominator first:

$$\begin{aligned}
\sum_{i=1}^{n} x_i^2 - n\bar{x}^2 &= \sum_{i=1}^{n} x_i^2 - n\bar{x}^2 + n\bar{x}^2 - n\bar{x}^2 && \text{add and subtract same term changes nothing} \\
&= \sum_{i=1}^{n} x_i^2 + n\bar{x}^2 - 2n\bar{x}^2 \\
&= \sum_{i=1}^{n} x_i^2 + n\bar{x}^2 - 2\bar{x}\sum_{i=1}^{n} x_i \\
&= \sum_{i=1}^{n} x_i^2 - 2x_i\bar{x} + \bar{x}^2 && \text{merge sums} \\
&= \sum_{i=1}^{n} (x_i - \bar{x})^2
\end{aligned}$$

Consider the numerator:

$$\begin{aligned}
\sum_{i=1}^{n} x_i y_i - \bar{y}\sum_{i=1}^{n} x_i &= \sum_{i=1}^{n} x_i y_i - \bar{y}\sum_{i=1}^{n} x_i - n\bar{x}\bar{y} + n\bar{x}\bar{y} && \text{add/subtract trick again} \\
&= \sum_{i=1}^{n} x_i y_i - \bar{y}\sum_{i=1}^{n} x_i - \bar{x}\sum_{i=1}^{n} y_i + n\bar{x}\bar{y} \\
&= \sum_{i=1}^{n} x_i y_i - x_i\bar{y} - \bar{x}y_i + \bar{x}\bar{y} \\
&= \sum_{i=1}^{n} x_i(y_i - \bar{y}) - \bar{x}(y_i - \bar{y}) \\
&= \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})
\end{aligned}$$

Finally we have

$$b_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

(ii)

$$\begin{aligned}
\hat{\beta}_1 &= \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \\
&= \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}} \frac{1}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}} \\
&= \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}} \frac{1}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}} \frac{\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}{\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}} \\
&= \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}} \frac{\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}} \\
&= \hat{\rho}\frac{\sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2}}{\sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2}} \\
&= \hat{\rho}\frac{\hat{\sigma}_Y}{\hat{\sigma}_X}
\end{aligned}$$

The estimated slope is a kind of scaled correlation. They will have the same sign. For higher R^2, $\hat{\beta}_1^2$ will be higher as well.