Question 1 – Concrete

An article on *Concrete Research* from 1989 presented data on compressive strength x and intrinsic permeability y of various concrete mixes and cures. Summary quantities are:

$$n = 15, \quad \sum_{i=1}^{15} y_i = 570, \quad \sum_{i=1}^{15} y_i^2 = 22, \quad \sum_{i=1}^{15} x_i = 45 \quad \sum_{i=1}^{15} x_i^2 = 155, \quad \sum_{i=1}^{15} x_i y_i = 1691.$$

Assume that permeability is linearly related to compressive strength.

(a) Calculate the least squares estimates of the slope and intercept.

Solution:
Substituting the summary quantities into the equations given in the lecture notes we get:

$$\begin{aligned}
\hat{\beta}_1 &= \frac{\sum y_i x_i - \frac{(\sum y_i)(\sum x_i)}{n}}{\sum x_i^2 - \frac{(\sum x_i)^2}{n}} \\
&= \frac{1691 - \frac{(570)(45)}{15}}{155 - \frac{(45)^2}{15}} \\
&= -0.95 \\
\hat{\beta}_0 &= \bar{y} - \hat{\beta}_1 \bar{x} \\
&= \frac{\sum y_i}{n} - \hat{\beta}_1 \frac{\sum x_i}{n} \\
&= \frac{570}{15} - -0.95 \frac{45}{15} \\
&= 40.85
\end{aligned}$$

(b) Use the equation of the fitted line to predict what permeability would be observed when the compressive strength is $x = 41$.

Solution: The equation of the fitted line is

$$y = 40.85 - 0.95x$$

Substiting $x = 41$ into this we get

$$\begin{aligned}
y &= 40.85 - 0.95 \times 41 \\
&= 1.9.
\end{aligned}$$

So the permeability when the compressive strength is 41 is 1.9.

Question 2 – Renewable Energy

The file "*A6-2.csv*" contains information on renewable energy in US States published by the U.S. Energy Information Administration, available on `https://dasl.datadescription.com/datafile/alternative-energy-2016/?_sfm_cases=4+59943&sf_paged=2`.
The column "*Ren.Elec.GW.h.*" refers to the percentage of renewable electricity in Gigawatt hours and the column "*Pct.Renewable.incl.Hydro*" refers to the percentage of renewable energy with Hydropower.

(a) Assuming that a simple linear regression model is appropriate, use R to obtain the least squares fit estimators relating "*Pct.Renewable.incl.Hydro*" to "*Ren.Elec.GW.h.*".

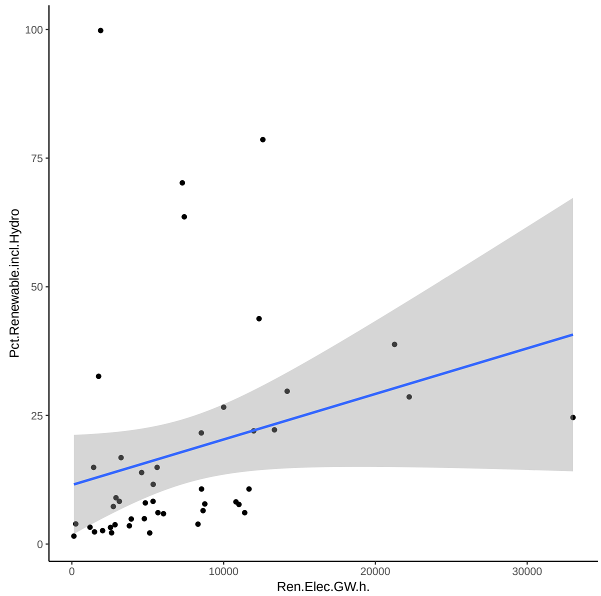
Solution:

```
> renewenergy <- read.csv("A6-2.csv")
> renewmodel <- lm(Pct.Renewable.incl.Hydro~Ren.Elec.GW.h.,data=renewenergy)
> summary(renewmodel)


Call:
lm(formula = Pct.Renewable.incl.Hydro ~ Ren.Elec.GW.h., data = renewenergy)

Residuals:
    Min      1Q  Median      3Q     Max
-16.104 -10.885  -7.849   1.594  86.643

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)    1.148e+01  4.828e+00   2.377   0.0219 *
Ren.Elec.GW.h. 8.853e-04  4.992e-04   1.773   0.0831 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 21.24 on 44 degrees of freedom
Multiple R-squared:  0.0667,Adjusted R-squared:  0.04549
F-statistic: 3.145 on 1 and 44 DF,  p-value: 0.0831
```

(b) Plot the data points in a scatter plot and add your linear regression curve. Comment on the appropriateness of the model.

Solution:

```
> ggplot(renewenergy, aes(x=Ren.Elec.GW.h.,y=Pct.Renewable.incl.Hydro))+geom_point() +g
```



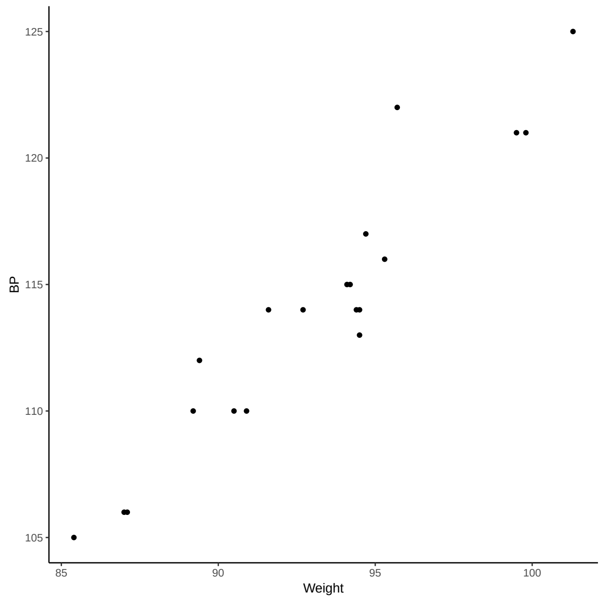
Question 3 – Blood Pressure

The data set "*A6-3.csv*" contains the blood pressure (BP) and weight (Weight) of 20 individuals.

(a) Plot a scatter diagram of the data. Does the straight-line regression model seem to be plausible?

Solution:

```
> BPdata <- read.csv("A6-3.csv")
> ggplot(BPdata,aes(x=Weight,y=BP))+geom_point() + theme_classic()
```



(b) Calculate the error sum of squares, commonly denoted by SSE. Then use this value to estimate the variance σ^2.

Solution: To calculate the error sum of squares we need to first find the coefficients of the linear model. We do this by

```
>  bloodmod<-lm(BP~Weight,data=BPdata)
>  coef(bloodmod)
```

```
(Intercept)       Weight
   2.205305     1.200931
```

This shows that $\hat{\beta}_0 = 2.2053053$ and $\hat{\beta}_1 = 1.2009313$. The calculation of SSE is then

$$\begin{aligned}
SSE &= \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \\
&= \sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2 \\
&= \sum(y_i^2 - 2\hat{\beta}_0 y_i + \hat{\beta}_0^2 - 2\hat{\beta}_1 x_i y_i + 2\hat{\beta}_0\hat{\beta}_1 x_i + \hat{\beta}_1^2 x_i^2) \\
&= \sum y_i^2 - 2\hat{\beta}_0 \sum y_i + n\hat{\beta}_0^2 - 2\hat{\beta}_1 \sum x_i y_i + 2\hat{\beta}_0\hat{\beta}_1 \sum x_i + \hat{\beta}_1^2 \sum x_i^2 \\
&= \sum y_i^2 - 2\hat{\beta}_0 n\bar{y} + n\hat{\beta}_0^2 - 2\hat{\beta}_1 \sum x_i y_i + 2\hat{\beta}_0\hat{\beta}_1 n\bar{x} + \hat{\beta}_1^2 \sum x_i^2 \\
&= 54.528016
\end{aligned}$$

Now substituting this into the equation for $\hat{\sigma}^2$

$$\hat{\sigma}^2 = MSE = \frac{SSE}{n-2} = 3.0293342$$

Question 4 – Regression without the Intercept Term

Assume that we have n pairs of data $(x1, y1)$, $(x2, y2)$, ..., (xn, yn).

(a) Suppose that the appropriate model is $Y = \beta x + \epsilon$ (no intercept). Provide an equation to estimate β.

Solution:
First define L

$$L = \sum (y_i - \beta x_i)^2$$

Differentiate to find critical point

$$\frac{\partial L}{\partial \beta} = -2 \sum (y_i - \beta x_i) x_i = 0$$

Rearrange to find β

$$\begin{aligned} 0 &= -2 \sum y_i x_i + 2\beta \sum x_i^2 \\ \beta \sum x_i^2 &= \sum y_i x_i \\ \beta &= \frac{\sum y_i x_i}{\sum x_i^2} \end{aligned}$$

(b) Do you suspect the model $Y = \beta x + \epsilon$ to fit better or worse than $Y = \beta_1 x + \beta_0 + \epsilon$ to a general data set? Explain briefly.

Solution:
The model $Y = \beta x + \epsilon$ would be a worse fit generally as it assumes that when x is equal to zero y is equal to zero. Generally there would be a base case expected to have a non-zero result and so the intercept would need to be calculated. If however the intercept was small and the range of x is very much larger it could be almost as good fit.

Question 5 – Intrinsically Linear

Decide which of the following relations between $Y > 0$ and $x > 0$ are intrinsically linear, where ϵ is a random variable (not necessarily Gaussian). If they are intrinsically linear, provide the function that transforms the equation into a linear relation.

(a) $Y = \frac{\beta_0}{\beta_1 x + \beta_2 + \beta_0 \epsilon}$

Solution:
Yes, take the reciprocal of both sides $\frac{1}{Y} = \frac{\beta_1}{\beta_0} x + \frac{\beta_2}{\beta_0} + \epsilon$.

(b) $Y = \left(e^{\beta_1 x + \beta_2 + \epsilon}\right) \beta_0$

Solution:
Yes, take the logarithm of both sides $\log(Y) = \log(\beta_0) + \beta_1 x + \beta_2 + \epsilon$.

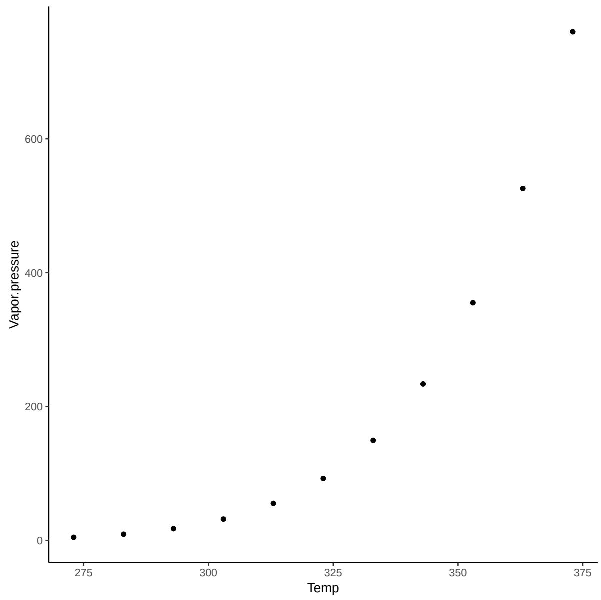
Question 6 – Water Vapor Pressure

The file "*A6-6.csv*" contains the temperature (K) and vapour pressure (mmHg) of 11 samples.

(a) Plot a scatter diagram of the data. What type of relation between the temperature and vapour pressure do you suspect?

Solution:

```
> q6data <- read.csv("A6-6.csv")
> ggplot(q6data,aes(x=Temp,y=Vapor.pressure))+geom_point() + theme_classic()
```

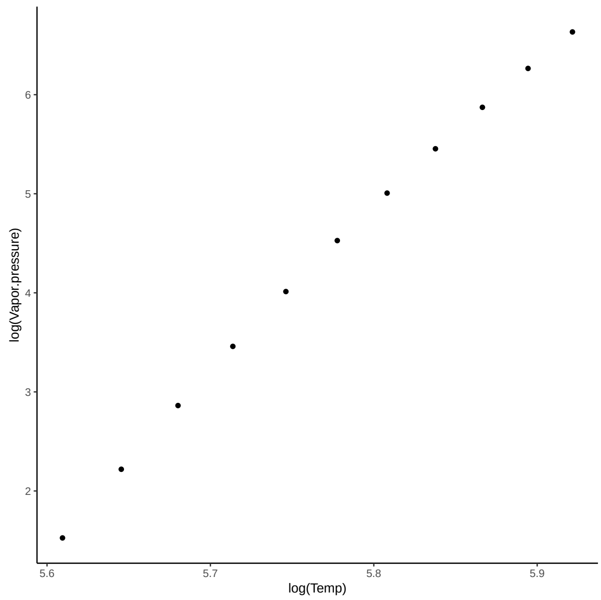


We see that the relations seems rather exponential, so a linear relation does not seem appropriate given this data set.

(b) Use an appropriate transformation to fit a linear model to the (transformed) data, relating (transformed) vapour pressure to the (transformed) temperature. Clearly state the transformation you applied as well as the resulting least square estimates $\hat{\beta}_0$, $\hat{\beta}_1$.

Solution: Given the relation, we try a logarithm transformation:

```
> ggplot(q6data,aes(x=log(Temp),y=log(Vapor.pressure)))+geom_point() + theme_classic()
```



In fact that looks pretty much linear. Let us therefore fit the model via R (for example)

```
> mod1<-lm(log(Vapor.pressure)~log(Temp),data=q6data)
> coef(mod1)

(Intercept)    log(Temp)
  -89.81008     16.31075
```

Question 7 – t-Test for Regression Models

Consider the following data on the number of pounds of steam (y) used by a chemical plant and the average temperature (x) in Fahrenheit.

Temp	21	24	32	47	50	59
Usage	185.79	214.47	288.03	424.84	454.58	539.03
Temp	68	74	62	50	41	30
Useage	621.5	675.06	562.03	452.93	369.95	273.98

Test the hypothesis $H_0 : \beta_1 = 0$ against $H_1 : \beta_1 \neq 0$ using the t-test with $\alpha = 0.05$.
Solution:

- Find $t_{1-\frac{\alpha}{2},n-2} : \hat{\Phi}\left(t_{1-\frac{\alpha}{2},10}\right) = 1 - \frac{\alpha}{2} = 0.975 \to t_{1-\frac{\alpha}{2},10} = 2.228$
- Calculate $T = \frac{\hat{\beta}_1 - \beta_1}{\sqrt{\frac{\hat{\sigma}^2}{S_{XX}}}}$. For that we need

$$S_{XX} = \sum_i x_i^2 - n\bar{x}^2 = 29256 - 12 \cdot 46.5^2 = 3309$$

$$\hat{\beta}_1 = \frac{\sum_i x_i y_i - n\bar{x}\bar{y}}{\sum_i x_i^2 - n\bar{x}^2} = \frac{265861.23 - 12 \cdot 46.5 \cdot 421.8491667}{29256 - 12 \cdot 2162.25} = 9.2080372$$

$$\hat{\sigma}^2 = \frac{\sum_i y_i^2 - n\bar{y}^2 - \hat{\beta}_1 \sum_i x_i y_i + \hat{\beta}_1 n\bar{x}\bar{y}}{n-2}$$

$$= \frac{2416081.5311 - 12 \cdot 421.8491667^2 - 9.2080372 \cdot 265861.23 + 9.2080372 \cdot 46.5 \cdot 421.8491667}{12-2}$$

$$= 3.7576343$$

$$T = \frac{9.2080 - 0}{\sqrt{\frac{3.7576}{3309}}} = 273.2488$$

- Reject H_0 if $T > 2.228$ or $T < 2.228$. Since $T = 273.2487545 > 2.228$, we do reject H_0.

```
> summary(lm(useage~temp,data=q7data))


Call:
lm(formula = useage ~ temp, data = q7data)

Residuals:
    Min      1Q  Median      3Q     Max
-2.5437 -1.2544 -0.2505  0.7965  4.0634

Coefficients:
```

```
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -6.3246     1.6639  -3.801  0.00348 **
temp          9.2080     0.0337 273.249  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.938 on 10 degrees of freedom
Multiple R-squared:  0.9999,Adjusted R-squared:  0.9999
F-statistic: 7.466e+04 on 1 and 10 DF,  p-value: < 2.2e-16
```

Question 8 – Beauty of a Proof II

Given observations $(y_1, y_2, \ldots, y_n)$ and their predictions $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$ $(i = 1, 2, ..., n)$, where x_i are observed variables $i = 1, \ldots, n$, $\hat{\beta}_0$ is the least square estimate of the intercept and $\hat{\beta}_1$ is the least square estimate of the slope.
Show that

$$\sum_{i=1}^{n} (y_i - \hat{y}_i) = 0.$$

(Hint: Use the structure of $\hat{y}_i$ and recall the equations for $\hat{\beta}_0$ and $\hat{\beta}_1$.)
Solution:

$$\begin{aligned}
\sum_{i=1}^{n} (y_i - \hat{y}_i) &= \sum_{i=1}^{n} y_i - \sum_{i=1}^{n} \hat{y}_i \\
&= n\bar{y} - \sum_{i=1}^{n} \left(\hat{\beta}_0 + \hat{\beta}_1 x_i\right) \\
&= n\bar{y} - \sum_{i=1}^{n} \hat{\beta}_0 - \sum_{i=1}^{n} \hat{\beta}_1 x_i \\
&= n\bar{y} - n\hat{\beta}_0 - n\hat{\beta}_1 \bar{x}
\end{aligned}$$

Using $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$

$$\begin{aligned}
&= n\left(\bar{y} - \bar{y} + \hat{\beta}_1 \bar{x} - \hat{\beta}_1 \bar{x}\right) \\
&= 0
\end{aligned}$$

where $\bar{y}$ is the mean of y_i.

Thank you for a great semester!
Thank you for your feedback to improve the STAT2201 lectures and my teaching style, I appreciate it very much.