TMA4255 Applied Statistics
Solution to Exercise 3

Problem 1 - Simple linear regression (theory)

It can be shown that $\mathrm{E}(B_0) = \beta_0$ and $\mathrm{E}(B_1) = \beta_1$. **a)** First, the variance of B_1

$$\begin{aligned} Var(B_1) &= \frac{1}{(n\sum_{i=1}^n x_i^2 - (\sum_{i=1}^n x_i)^2)^2}\left[n^2\sum_{i=1}^n x_i^2\sigma^2 - n\sigma^2(\sum_{i=1}^n x_i)^2\right] \\ &= n\sigma^2\left[\frac{n\sum_{i=1}^n x_i^2 - (\sum_{i=1}^n x_i)^2}{(n\sum_{i=1}^n x_i^2 - (\sum_{i=1}^n x_i)^2)^2}\right] \\ &= \frac{\sigma^2}{\sum_{i=1}^n (x_i - \bar{x})^2} \end{aligned}$$

To find the variance of B_0 we use the fact that $\bar{Y}$ and B_1 are independent, and thus $Cov(\bar{Y}, B_1) = 0$,

$$\begin{aligned} Var(B_0) &= Var(\bar{Y}) + \bar{x}^2 Var(B_1) + 2\bar{x}Cov(\bar{Y}, B_1) \\ &= \sigma^2/n + \bar{x}^2\frac{\sigma^2}{\sum_{i=1}^n (x_i - \bar{x})^2} \\ &= \frac{\sigma^2\sum_{i=1}^n x_i^2}{n\sum_{i=1}^n (x_i - \bar{x})^2} \end{aligned}$$

The covariance between B_0 and B_1 is

$$\begin{aligned} Cov(B_0, B_1) &= E(B_0B_1) - E(B_0)E(B_1) \\ &= E(\bar{Y}B_1) - E(B_1^2\bar{x}) - E(B_0)E(B_1) \\ &= E(\bar{Y})E(B_1) + \bar{x}E(B_1^2) - E(B_0)E(B_1) \\ &= E(\bar{Y})E(B_1) + \bar{x}(Var(B_1) + (E(B_1))^2) - E(B_0)E(B_1) \\ &= \frac{-\bar{x}\sigma^2}{\sum_{i=1}^n (x_i - \bar{x})^2} \end{aligned}$$

Remark: the covariance is negative.

This can also be found using matrix notation.

$$\begin{aligned}
\operatorname{Cov}\begin{pmatrix} B_0 \\ B_1 \end{pmatrix} &= \sigma^2(\mathbf{X}^T\mathbf{X})^{-1} \\
&= \sigma^2 \left(\begin{bmatrix} 1 & 1 & \cdots & 1 \\ x_1 & x_2 & \cdots & x_n \end{bmatrix} \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix} \right)^{-1} \\
&= \sigma^2 \begin{bmatrix} n & \sum x_i \\ \sum x_i & \sum x_i^2 \end{bmatrix} \\
&= \frac{\sigma^2}{n\sum x_i - (\sum x_i)^2} \begin{bmatrix} \sum x_i^2 & -\sum x_i \\ -\sum x_i & n \end{bmatrix} \\
&= \begin{bmatrix} \operatorname{Var}(B_0) & \operatorname{Cov}(B_0, B_1) \\ \operatorname{Cov}(B_0, B_1) & \operatorname{Var}(B_1) \end{bmatrix}
\end{aligned}$$

Dette gir

$$\operatorname{Var}(B_0) = \frac{\sigma^2 \sum x_i^2}{n\sum(x_i - \bar{x})^2}$$

$$\operatorname{Var}(B_1) = \frac{\sigma^2}{\sum(x_i - \bar{x})^2}$$

$$\operatorname{Cov}(B_0, B_1) = \frac{-\sigma^2 \sum x_i}{n\sum(x_i - \bar{x})^2}$$

b) <u>Confidence interval for $\mu_{Y|x=x_0}$</u>:

Use that $\hat{Y} \sim N(\mu_{\hat{Y}}, \sigma^2_{\hat{Y}})$, where

$$\begin{aligned}
\mu_{\hat{Y}} &= E[\hat{Y}|x = x_0] = E[\hat{\alpha} + \hat{\beta}x_0] = \alpha + \beta x_0 = \mu_Y \\
\sigma^2_{\hat{Y}} &= \operatorname{Var}[\hat{\alpha} + \hat{\beta}x_0] \\
&= \operatorname{Var}(\bar{Y} + \hat{\beta}(x_0 - \bar{x})) \\
&= \operatorname{Var}(\bar{Y}) + (x_0 - \bar{x})^2\operatorname{Var}(\hat{\beta}) + 2(x_0 - \bar{x})\operatorname{Cov}(\bar{Y}, \hat{\beta}) \\
&= \frac{\sigma^2}{n} + \frac{(x_0 - \bar{x})^2\sigma^2}{S_{xx}} + 0 \\
&= \sigma^2 \left[\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}} \right] \qquad \text{and} \qquad \boxed{S_{xx} = \sum_{i=1}^{n}(x_i - \bar{x})^2}
\end{aligned}$$

$S^2 = \frac{1}{n-2}\sum(Y_i - \bar{Y})^2$ is a consistent estimator for σ^2. Then

$$\hat{\sigma}^2_{\hat{Y}} = S^2 \left[\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}} \right]$$

is a consistent estimator for $\sigma^2_{\hat{Y}}$.
We then get the t-statistic

$$T = \frac{\hat{Y}_0 - \mu_{\hat{y}}}{\sqrt{\hat{\sigma}^2_{\hat{Y}}}} = \frac{\hat{Y}_0 - \mu_{\hat{y}}}{S\sqrt{\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}}} \sim t_{n-2}$$

and

$$\mu_Y \in \left[\hat{y}_0 \pm t_{\alpha/2, n-2} \cdot S\sqrt{\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}} \right]$$

Prediction interval for $Y^*_{|x_0}$:

Start with $\hat{Y}_0 - Y^*_0 \sim N(\mu_{\hat{Y}_0 - Y^*_0}, \sigma^2_{\hat{Y}_0 - Y^*_0})$, where

$$\mu_{\hat{Y}_0 - Y^*_0} = E[\hat{Y}_0 - Y^*_0] = \mu_{\hat{Y}_0} - \mu_{Y_0} = 0$$
$$\sigma^2_{\hat{Y}_0 - Y^*_0} = \cdots = \sigma^2 \left(1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}} \right)$$

Also there we get a t-statistic

$$T = \frac{(\hat{Y}_0 - Y^*_0) - \mu_{\hat{Y}_0 - Y^*_0}}{\sqrt{\hat{\sigma}^2_{\hat{Y}_0 - Y^*_0}}} = \frac{\hat{Y}_0 - Y^*_0}{S\sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}}} \sim t_{n-2}$$

and

$$Y^*_0 \in \left[\hat{y}_0 \pm t_{\alpha/2, n-2} \cdot S\sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}} \right]$$

Both interval are at their shortest when $x_0 = \bar{x}$.

Problem 2

a) Assume that

$$E(T_i) = \alpha_0 + \alpha_1 p_i. \tag{1}$$

```
Regression Analysis

The regression equation is
T = 155 + 1,90 p

Predictor        Coef       StDev          T        P
Constant      155,296       0,927     167,47    0,000
p             1,90178     0,03676      51,74    0,000

S = 0,4440      R-Sq = 99,4\%     R-Sq(adj) = 99,4\%

Analysis of Variance
```

```
Source            DF          SS          MS         F        P
Regression         1      527,82      527,82   2677,11    0,000
Residual Error    15        2,96        0,20
Total             16      530,78

Unusual Observations
Obs          p          T         Fit   StDev Fit    Residual    St Resid
 12       26,6    204,600     205,827       0,121      -1,227       -2,87R

R denotes an observation with a large standardized residual
```

From the regression analysis we get

$$E(T_i) = 155 + 1.90p_i$$

This model seems to fit reasonably well, since the observations appear to be approximately on a straight line. This plot we get by choosing Stat $\rightarrow$ Regression $\rightarrow$ Fitted Line Plot in MINITAB. (For R see source file.)

When we look at the plot of standardized residuals against the pressure we see that the model (1) does not fit that well, since the standardized residuals are dependent on the pressure. However, the standardized residuals look normally distributed.

b) Assume that

$$E(T_i) = \beta_0 + \beta_1 x_i = \beta_0 + \beta_1 100 \ln p_i \tag{2}$$

We assess this model in the same way as in **a)** by regression on $100\log(p)$ in C3.

```
Regression Analysis

The regression equation is
T = 47,9 + 0,482 logp

Predictor        Coef       StDev          T        P
Constant       47,864       2,852      16,78    0,000
logp         0,482467    0,008866      54,42    0,000

S = 0,4223      R-Sq = 99,5\%     R-Sq(adj) = 99,5\%

Analysis of Variance

Source            DF          SS          MS         F        P
Regression         1      528,11      528,11   2961,55    0,000
Residual Error    15        2,67        0,18
Total             16      530,78

Unusual Observations
Obs       logp          T         Fit   StDev Fit    Residual    St Resid
 12        328    204,600     206,102       0,118      -1,502       -3,70R
```

When studying the standardized residuals we see that this model is a better fit than (1) when plotted agains the covariate

$$E(T_i) = 47.9 + 0.482x_i,$$

where $x_i = 100 \ln p_i$. Turning to the normal plot the the standardized residuals look less normally distributed than for the model with p.

c) We will test H_0: $\beta_1 = 0$ vs. H_1: $\beta_1 \neq 0$. From the teaching material we have:

- $\hat{\beta}_1 = \dfrac{\sum T_i(x_i - \bar{x})}{\sum (x_i - \bar{x})^2} \sim N\left(\beta_i, \dfrac{\sigma^2}{\sum (x_i - \bar{x})^2}\right)$
- $S^2 = \dfrac{1}{n-2}\sum (T_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2$
- $T = \dfrac{\hat{\beta}_1 - \beta_1}{S\sqrt{\frac{1}{\sum (x_i - \bar{x})^2}}} = \dfrac{\hat{\beta}}{S}\sqrt{\sum (x_i - \bar{x})^2} \sim T_{n-2} = T_{15}$

Fitting the model gives $S^2 = 0.422^2$ and $T_{0\text{ obs}} = 54.42$. We reject H_0 when $|\hat{\beta}_1| > k$, that is, we reject H_0 when $|T_{0\text{ obs}}| > t_{\alpha/2,n-2}$.

$t_{0.005,15} = 2.95 \Rightarrow \alpha = 0.01$ gives rejection of H_0, which is in agreement with the p-value of 0.000 in MINITAB (and $< 2e-16$ in R) for the slope of $\log(p)$.

d) From the ANOVA table:

$$MS_{\text{reg}} = \frac{SS_{\text{reg}}}{1} = 528.11$$
$$MS_{\text{err}} = \frac{SS_{\text{err}}}{15} = 0.18$$

Then the F-statistic:

$$F = \frac{MS_{\text{reg}}}{MS_{\text{err}}} \sim F_{1,15}$$

Reject H_0 when $F_{0\text{ obs}} > f_{\alpha,1,15}$. $F_{0\text{ obs}} = 2962$, $f_{0.01,1,15} = 8.68$ and thus we reject H_0.

To show that the two tests are equivalent, we use that

$$F = \frac{MS_{\text{reg}}}{MS_{\text{err}}} = \frac{\sum (\bar{T} - \hat{T}_i)^2}{\frac{1}{15}\sum (T_i - \hat{T})^2}$$

$$\left.\begin{array}{l} \hat{T} = \hat{\beta}_0 + \hat{\beta}_1 x_i \\ \bar{T} = \hat{\beta}_0 + \hat{\beta}_1 \bar{x} \Rightarrow \hat{\beta}_0 = \bar{T} - \hat{\beta}_1 \bar{x} \end{array}\right\} \Rightarrow \hat{T} = \bar{T} + \hat{\beta}_1(x_i - \bar{x})$$

Thus, we get that

$$F = \frac{MS_{\text{reg}}}{MS_{\text{err}}} = \frac{\sum (\bar{T} - (\bar{T} + \hat{\beta}_1(x_i - \bar{x})))^2}{\frac{1}{15}\sum (T_i - \hat{\beta}_0 - \hat{\beta}_1 x_1)^2} = \frac{\sum [(x_i - \bar{x})\hat{\beta}_1]^2}{S^2}$$

and

$$F = \frac{\hat{\beta}_1^2}{S^2}\sum (x_i - \bar{x})^2 = T^2$$

This confirms the general result:

$$T_\nu^2 = F_{1,\nu}$$

e) Will now test H_0: $\beta_1 = 1$ vs. H_1: $\beta_1 \neq 1$. From the regression model fit:

$$\hat{\beta}_1 = 0.482$$
$$\frac{S}{\sqrt{\sum (x_i - \bar{x})^2}} = 0.008866 \quad \text{(Estimate for } SD(\hat{\beta}_1))$$

Thus, we get

$$T_{0\text{ obs}} = \frac{\hat{\beta}_1 - 1}{S/\sqrt{\sum(x_i - \bar{x})^2}} = \frac{0.482 - 1}{0.008866} = -58.43$$

We reject H_0 when $|\hat{\beta}_1 - 1| > k$, thus, we reject H_0 when $|T_{0\text{ obs}}| > t_{\alpha/2,n-2}$. $t_{0.005,15} = 2.95 \Rightarrow$ $\alpha = 0.01$ give rejection of H_0.

f) 99 % confidence interval (CI) for β_1:

$$P(\theta_1 < \beta_1 < \theta_2) \geqslant 0.99$$

$$\Rightarrow P\left(-t_{\alpha/2,n-2} < \frac{\hat{\beta}_1 - \beta_1}{S/\sqrt{\sum(x_i - \bar{x})^2}} < t_{\alpha/2,n-2}\right) \geqslant 0.99$$

$$\Leftrightarrow P\left(\hat{\beta}_1 - t_{\alpha/2,n-2}\frac{S}{\sqrt{\sum(x_i - \bar{x})^2}} < \beta_1 < \hat{\beta}_1 + t_{\alpha/2,n-2}\frac{S}{\sqrt{\sum(x_i - \bar{x})^2}}\right) \geqslant 0.99$$

$$\Rightarrow \beta_1 \in [\hat{\beta}_1 - t_{0.005,15} \cdot 0.008866, \hat{\beta}_1 + t_{0.005,15} \cdot 0.008866]$$

$$\Leftrightarrow \beta_1 \in [0.456, 0.508]$$

Reject H_0 both in **c)** and **e)** since neighter 0 nor 1 is in the interval.