STA 250: Statistics
Lab 10

In this lab, we will look into how to perform simple linear regression on R.

Highway accident data

To start with load the dataset "highway" from R package "alr3". This dataset contains observations from 39 segments of a highway about on accident rate (`Rate`) and physical characteristics of the road segment including speed limit (`Slim`), number of access points (`Acpt`), etc. We'll work with accident rate as the response and speed limit as the explanatory variable.

```
library(alr3)
data(highway)
y <- highway$Rate
x <- highway$Slim
```

If you can't load the data from the library, simple copy and paste the following

x	55, 60, 60, 65, 70, 55, 55, 55, 50, 50, 60, 50, 50, 60, 55, 60, 60, 50, 55, 60, 55, 60, 50, 60, 40, 45, 55, 55, 45, 60, 45, 55, 55, 55, 55, 50, 55, 60, 55
y	4.58, 2.86, 3.02, 2.29, 1.61, 6.87, 3.85, 6.12, 3.29, 5.88, 4.2, 4.61, 4.8, 3.85, 2.69, 1.99, 2.01, 4.22, 2.76, 2.55, 1.89, 2.34, 2.83, 1.81, 9.23, 8.6, 8.21, 2.93, 7.48, 2.57, 5.77, 2.9, 2.97, 1.84, 3.78, 2.76, 4.27, 3.05, 4.12

Task 1. Plot y against x with `xlab = "Speed limit", ylab = "Accident rate", ylim = c(0,10)`.

Least squares estimation

First do least squares estimate by using the formulas we learned.

Task 2. Get n, $\bar{x}$, s_x^2, s_{xy} and $\bar{y}$ from the data. In `R` s_x can be calculated by using `sd(x)` and s_{xy} can be calculated by using `cov(x, y)`.

Task 3. Calculate $\hat{\beta}_0 = \bar{y} - s_{xy} \cdot \bar{x}/s_x^2$ and $\hat{\beta}_1 = s_{xy}/s_x^2$ based on these numbers. Also calculate $\hat{\sigma}$. Recall $\hat{\sigma}^2 = \frac{1}{n-2}\sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2$.

Task 4. Calculate a 95% ML/reference-Bayes interval for β_1 given by $\hat{\beta}_1 \mp z_{n-2}(0.05) \cdot \hat{\sigma}_1$ where $\hat{\sigma}_1 = \hat{\sigma}/\sqrt{(n-1)s_x^2}$.

TASK 5. Calculate the p-value for testing $H_0 : \beta_1 = 0$. The p-value equals $2\{1 - \Phi_{n-2}(|\hat{\beta}_1|/\hat{\sigma}_1)\}$.

TASK 6. Get a 95% prediction interval for accident rate at speed limit $x^* = 65$. This interval is given by,

$$\hat{\beta}_0 + \hat{\beta}_1 x^* \mp \hat{\sigma} \left\{ 1 + \frac{1}{n} + \frac{(x^* - \bar{x})^2}{(n-1)s_x^2} \right\}^{1/2} \cdot z_{n-2}(0.05)$$

Doing the same with the `lm()` function

We will now perform the same tasks by using the `lm()` function of R.

TASK 7. Use

```
hwy.lm <- lm(y ~ x)
summary(hwy.lm)
```

and highlight/mark the figures that correspond to $\hat{\beta}_0$, $\hat{\beta}_1$, $\hat{\sigma}$, $\hat{\sigma}_1$, the df $= n - 2$ and the p-value for testing $H_0 : \beta_1 = 0$.

TASK 8. Check that you get the same prediction interval as in Task 6 by using

```
predict(hwy.lm, newdata = data.frame(x = 65), interval =
"prediction", level = .95)
```

Think about this...

I'd leave you one thing to think about. Why is the slope estimated negative? Should increasing the speed limit decrease highway accident rate? Or is there something more subtle going on for this dataset? (Think about which segments of a highway are likely to have lower speed limits.)