Math/Stat319 for chapter 12 : Simple Linear regression

- Observations: $(x_1, y_1), \ldots (x_n, y_n)$ where x is the explanatory/indepedent variable and the y is the response variable/dependent variable.
- Assumption: (1) Linearity; (2) error terms are independent; (3) error terms are normally distributed; (4) error terms have mean 0; (5) error terms have constant variance.
- Model: $Y_i = \beta_0 + \beta_1 X_i + \epsilon_i, \quad \epsilon_i \sim N(0, \sigma^2) \leftrightarrow Y_i \sim N(\beta_0 + \beta_1 X_i, \sigma^2)$.
- Estimation for βs, by LSE (least square estimation) or MLE (maximum likelihood estimation)

$$
\begin{aligned}
b_0 = \hat{\beta}_0 =& \bar{y} - \bar{x} b_1 \\
b_1 = \hat{\beta}_1 =& \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2} = \frac{\sum_{i=1}^{n} x_i y_i - \frac{\left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right)}{n}}{\sum_{i=1}^{n} x_i^2 - \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n}} = \frac{S_{xy}}{S_{xx}}
\end{aligned}
$$

Therefore, the estimated regression line is $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$. We call $e_i = y_i - \hat{y}_i$ the residuals.

- Interpretation: β_1 is the slope parameter, which is interpreted as the expected or true average increase in Y associated with a 1-unit increase in x. β_0 is the intercept, which is interpreted as the expected or true average value of Y when $x = 0$. $\hat{\beta}_1$: the estimated expected change associated with 1-unit increase in x is $\hat{\beta}_1$. OR 1-unit increase in x results in an $\hat{\beta}_1$ increase/decrease of Y in average.
- Estimating σ^2. The least square estimator of σ^2 is

$$
\hat{\sigma}^2 = s^2 = \frac{\sum (y_i - \hat{y}_i)^2}{n-2}
$$

- Regression and ANOVA

Table 1: ANOVA table for simple linear regression

Source of variation	df	Sum of Squares	Mean Square	f
Regression	1	$SSR = \sum(\hat{y}_i - \bar{y})^2$	SSR	$\frac{SSR}{SSE/(n-2)}$
Error	$n-2$	$SSE = \sum(y_i - \hat{y}_i)^2$	$MSE = s^2 = \frac{SSE}{n-2}$	
Total	$n-1$	$SST = \sum(y_i - \bar{y})^2$		

Now you should be able to :

- Calculate the Coefficient of Determination $r^2 = 1 - \frac{SSE}{SST}$
- Perform the F test for the model fitting. Since we only have 1 slope β_1, the F test should give you exactly the same result as the t test for $H_0 : \beta_1 = 0, H_1 : \beta_1 \neq 0$. By exactly, I mean the same p-value and of course the same result. The rejection region for the F test is $\{f \geq F_{\alpha,1,n-2}\}$

- Inference of β_1.
  - $\hat{\beta}_1 \sim N(\beta_1, \sigma^2_{\hat{\beta}_1} = \frac{\sigma^2}{S_{xx}})$
  - The estimated standard deviation of $\hat{\beta}_1$: $s_{\hat{\beta}_1} = \hat{\sigma}_{\hat{\beta}_1} = \frac{s}{\sqrt{S_{xx}}}$
  - $T = \frac{\hat{\beta}_1 - \beta_1}{S_{\hat{\beta}_1}} \sim t_{n-2}$
  - Testing hypothesis:
    - ∗ Null hypothesis: $H_o : \beta_1 = c$
    - ∗ Test statistics : $Z = \frac{\hat{\beta}_1 - c}{S_{\hat{\beta}_1}} \sim t_{n-2}$
    - ∗ Rejection Region:
      1. For $H_1 : \beta_1 > c$, we reject H_o when $t > t_{\alpha,n-2}$;
      2. For $H_1 : \beta_1 < c$, we reject H_o when $t < -t_{\alpha,n-2}$;
      3. For $H_1 : \beta_1 \neq c$, we reject H_o when $|t| > t_{\alpha/2,n-2}$.
  - A typical computer output for regression analysis:

```
The regression equation is
Kg = −2.35 + 0.00845 CO2

Predictor          Coef              SE Coef       T          P

Constant       −2.3493←β̂0         0.7966     −2.95       0.026
CO2             0.008454←β̂1       0.001261     6.70       0.001

S = 0.533964    R-Sq = 88.2%←100r²    R-Sq(adj) = 86.3%

Analysis of Variance

Source            DF        SS            MS        F        P
Regression         1    12.808        12.808    44.92    0.001
Residual Error     6     1.711←SSE     0.285
Total              7    14.519←SST
```

Figure 1: Minitab output (Fig 12.14 from book)