# STAT 306 Finding Relationships in Data

## Lab 2 - Simple Linear Regression

David Lee

University of British Columbia
Department of Statistics

January 20/22, 2015

# Using Mac machines

- To log in the Mac machines, the user name is the first 8 characters of your name (first, middle if any, last) and the password is "S" plus the first 7 digits of your student number.
- Remember to log off before you leave!
- You'll notice that there's only one key on your mouse. To achieve the effect of a right click, do "Control + click". You need this to download files from the Internet — simply clicking on the file name will sometimes open it instead.
- The shortcuts for copy and paste are "Command + C" and "Command + V" respectively (*not* Control).

# Simple linear regression

- To describe the relationship between an explanatory (predictor) variable $X$ and a response variable $Y$, we may use a linear regression model:

  $$Y = \beta_0 + \beta_1 X + \epsilon,$$

  where $\beta_0$ and $\beta_1$ are unknown constants known as regression parameters, and $\epsilon$ is a random error. The parameter $\beta_1$ tells us by how much $Y$ is *expected* to vary, given a unit change in $X$.
- Throughout this course, $X$ is assumed to be non-random. Hence the only variability of $Y$ comes from $\epsilon$.

# Regression parameter estimates

- For a given data set with predictors and responses $(x_i, y_i)$, $i = 1, \ldots, n$, the parameters $\beta_0$ and $\beta_1$ can be estimated by the least squares method:

$$(\hat{\beta}_0, \hat{\beta}_1) = \arg\min_{\beta_0, \beta_1} \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2 .$$

- The solution to this minimization can be shown to be

$$\hat{\beta}_1 = r_{xy} \frac{s_y}{s_x} = \frac{s_{xy}}{s_x^2}; \qquad \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x},$$

  where $\bar{x}$ ($\bar{y}$) is the sample mean of $x$ ($y$); $s_x$ ($s_y$) is the standard error of $x$ ($y$), and $r_{xy}$ ($s_{xy}$) is the sample correlation (covariance) between $x$ and $y$.

# Interval estimates

- Since $\hat{\beta}_0$ and $\hat{\beta}_1$ are functions of $Y$, they are random variables. Interval estimates give us an idea of their probable values.
- To obtain such intervals, we need probabilistic assumptions on $\epsilon_i$. Here we assume $\epsilon_i \overset{\text{iid}}{\sim} N(0, \sigma^2)$.
- The 95% confidence interval for $\beta_1$ is $\hat{\beta}_1 \pm t_{n-2,0.975} \times \text{se}\left(\hat{\beta}_1\right)$.
- For a given $x$, the 95% confidence interval for the mean of $Y$ is

$$(\hat{\beta}_0 + \hat{\beta}_1 x) \pm t_{n-2,0.975} \times \text{se}\left(\hat{\beta}_0 + \hat{\beta}_1 x\right),$$

  and the 95% prediction interval for a future value of $Y$ is

$$(\hat{\beta}_0 + \hat{\beta}_1 x) \pm t_{n-2,0.975} \times \text{se}\left(\hat{\beta}_0 + \hat{\beta}_1 x + \epsilon\right).$$

- Note the difference between confidence and prediction intervals.

# Hypothesis test of regression parameters

- Note the duality between confidence intervals and hypothesis tests. Suppose we want to test

$$H_0 : \beta_1 = 0 \qquad \text{vs} \qquad H_1 : \beta_1 \neq 0.$$

  The null hypothesis is rejected at significance level $\alpha$ if and only if the $100(1-\alpha)\%$ confidence interval for $\beta_1$ does not contain zero.

- Hence, there is no need to calculate the test statistic separately.
- Rejection of $H_0$ implies a statistically significant relationship between $X$ and $Y$. Failure to reject $H_0$, however, does NOT mean $\beta_1 = 0$. It simply means we don't have sufficient evidence to reject it. See this for an analogy.

# Lab question

- See WeBWorK for this week's lab question. You will need to conduct a regression analysis using a subset of the data set `time.txt`. You need to submit your response by Friday 10pm.
- The function `lsfit` implements linear regression in `R`. If you store the regression result into a variable, you will be able to explore its properties using the functions `ls.print` and `ls.diag`.
- Refer to the lab R file (and the help documents on R functions, e.g. `?lsfit`) on how to do this.