NAME: ________________

18.443 Exam 3 Spring 2015
Statistics for Applications
5/7/2015

1. **Regression Through the Origin**

   For bivariate data on n cases: $\{(x_i, y_i), i = 1, 2, \ldots, n\}$, consider the linear model with zero intercept:

   $$Y_i = \beta x_i + \epsilon_i,\ i = 1, 2, \ldots, n$$

   where ϵ_i are independent and identically distribution $N(0, \sigma^2)$ random variables with fixed, but unknown variance $\sigma^2 > 0$.

   When $x_i = 0$, then $E[Y_i \mid x_i, \beta] = 0$.

   (a). Solve for the least-squares line – $\hat{Y} = \hat{\beta} x$.

   (b). Find the distribution of $\hat{\beta}$, the slope of the least squares line.

   (c). What is the distribution of the sum of squared residuals from the least-squares fit:

   $$SS_{ERR} = \sum_{i=1}^{n} (y_i - \hat{\beta} x_i)^2$$

   (d). Find an unbiased estimate of σ^2 using your answer to (c).

2. **Simple Linear Regression**

   Consider fitting the simple linear regression model:

   $$\hat{y} = \beta_1 + \beta_2 x_i$$

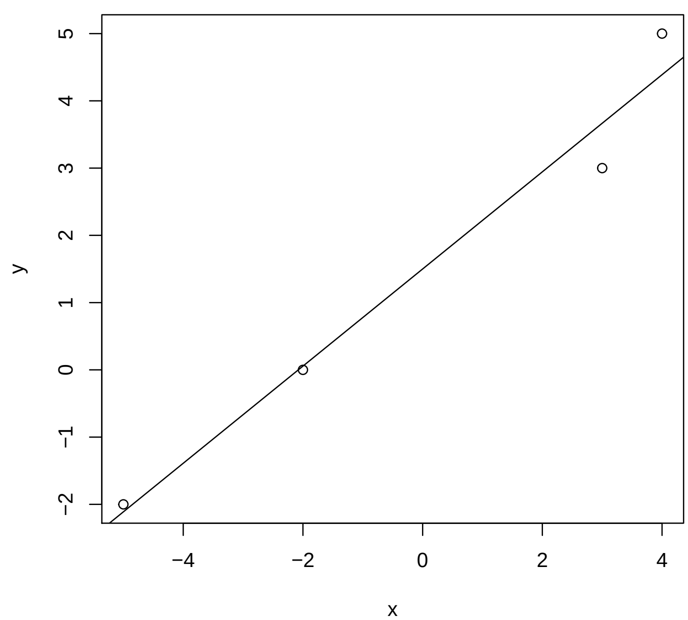
   to the following bivariate data:

i	x_i	y_i
1	-5	-2
2	-2	0
3	3	3
4	4	5

   The following code in R fits the model:

```
> x=c(-5,-2,3,4)
> y=c(-2,0,3,5)
> plot(x,y)
> lmfit1<-lm(y ~ x)
> abline(lmfit1)
> print(summary(lmfit1))

Call:
lm(formula = y ~ x)

Residuals:
```

```
       1        2        3        4 
 0.11111 -0.05556 -0.66667  0.61111 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)  
(Intercept)  1.50000    0.32275   4.648   0.0433 *
x            0.72222    0.08784   8.222   0.0145 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.6455 on 2 degrees of freedom
Multiple R-squared:  0.9713,	Adjusted R-squared:  0.9569 
F-statistic:  67.6 on 1 and 2 DF,  p-value: 0.01447
```



(a). Solve directly for the least-squares estimates of the intercept and slope of the simple linear regression (obtain the same values as in the R print summary)

(b). Give formulas for the least-squares estimates of β_1 and β_2 in terms of the simple statistics

$\overline{x} = 0$, and $\overline{y} = 1.5$

$s_x = \sqrt{S_x^2}$ =4.2426

$s_y = \sqrt{S_y^2}$ =3.1091

$r = Corr(x, y) = \frac{S_{xy}}{S_x S_y}$ =0.9855

(c). In the R print summary, the standard error of the slope $\hat{\beta}_2$ is given as $\hat{\sigma}_{\hat{\beta}_2}$ =0.0878

Using $\hat{\sigma}$ =0.65, give a formula for this standard error, using the statistics in (b).

(d). What is the least-squares prediction of $\hat{Y}$ when $X = \overline{x} = 0$, and what is its standard error (estimate of its standard deviation)?

3. Suppose that grades on a midterm and final have a correlation coefficient of 0.6 and both exams have an average score of 75. and a standard deviation of 10.

   (a). If a student's score on the midterm is 90 what would you predict her score on the final to be?

   (b). If a student's score on the final was 75, what would you guess that his score was on the midterm?

   (c). Consider all students scoring at the 75th percentile or higher on the midterm. What proportion of these students would you expect to be at or above the 75th percentile of the final? (i) 75%, (ii) 50%, (iii) less than 50%, or (iv) more than 50%.

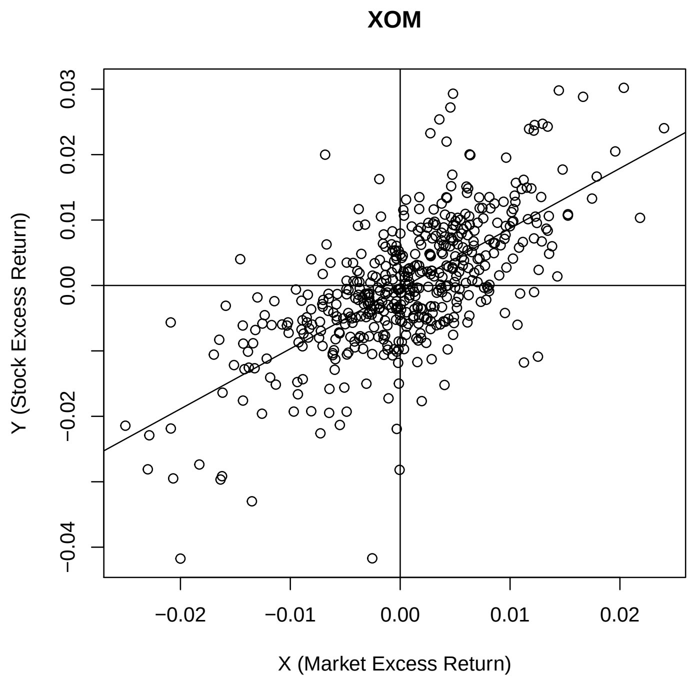
   Justify your answers.

4. **CAPM Model**

   The CAPM model was fit to model the excess returns of Exxon-Mobil (Y) as a linear function of the excess returns of the market (X) as represented by the S&P 500 Index.

   $$Y_i = \alpha + \beta X_i + \epsilon_i$$

   where the ϵ_i are assumed to be uncorrelated, with zero mean and constant variance σ^2. Using a recent 500-day analysis period the following output was generated in R:




```
> print(summary(lmfit0))

Call:
lm(formula = r.daily.symbol0.0[index.window] ~ r.daily.SP500.0[index.window],
    x = TRUE, y = TRUE)

Residuals:
      Min        1Q    Median        3Q       Max
-0.038885 -0.004415  0.000187  0.004445  0.026748

Coefficients:
                                Estimate Std. Error t value Pr(>|t|)
(Intercept)                   -0.0004805  0.0003360   -1.43    0.153
r.daily.SP500.0[index.window]  0.9190652  0.0454380   20.23   <2e-16

Residual standard error: 0.007489 on 498 degrees of freedom
Multiple R-squared:  0.451,        Adjusted R-squared:  0.4499
F-statistic: 409.1 on 1 and 498 DF,  p-value: < 2.2e-16
```

(a). Explain the meaning of the residual standard error.

(b). What does “498 degrees of freedom” mean?

(c). What is the correlation between Y (Stock Excess Return) and X (Market Excess Return)?

(d). Using this output, can you test whether the *alpha* of Exxon Mobil is zero (consistent with asset pricing in an efficient market).

$$H_0 : \alpha = 0 \text{ at the significance level } \alpha = .05?$$

If so, conduct the test, explain any assumptions which are necessary, and state the result of the test?

(e). Using this output, can you test whether the β of Exxon Mobil is less than 1, i.e., is Exxon Mobil less risky than the market:

$$H_0 : \beta = 1 \text{ versus } H_A : \beta < 1.$$

If so, what is your test statistic; what is the approximate P-value of the test (clearly state any assumptions you make)? Would you reject H_0 in favor of H_A?

5. For the following batch of numbers:

$$5, 8, 9, 9, 11, 13, 15, 19, 19, 20, 29$$

(a). Make a stem-and-leaf plot of the batch.

(b). Plot the ECDF (empirical cumulative distribution function) of the batch.

(c). Draw the Boxplot of the batch.

6. Suppose $X_1, \ldots, X_n$ are n values sampled at random from a fixed distribution:

$$X_i = \theta + \epsilon_i$$

where θ is a location parameter and the ϵ_i are i.i.d. random variables with mean zero and median zero.

(a). Give explicit definitions of 3 different estimators of the location parameter θ.

(b). For each estimator in (a), explain under what conditions it would be expected to be better than the other two.