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).

Solution:

(a). Set up the regression model with vectors/matrices:

$$Y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix},\ X = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} \text{ and } \quad e = \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix}$$

and $\beta = [\beta]$

$$Y = X\beta + e$$

The least-squares line minimizes

$$Q(\beta) = \sum_1^n (y_i - \beta x_i)^2 = (Y - X\beta)^T (Y - X\beta)$$

The least squares estimate $\hat{\beta}$ solves the first order equation: $\frac{\partial Q(\beta)}{\partial \beta} = 0$ and is given by

$$\hat{\beta} = (X^T X)^{-1} X^T Y = (\sum_1^n x_i^2)^{-1} \sum_1^n x_i y_i = \frac{\sum_1^n x_i y_i}{\sum_1^n x_i^2}$$

The least squares line is

$$y = \hat{\beta} x$$

(b). Since $\hat{\beta} = \sum_{i=1}^n w_i y_i$, where $w_i = \frac{x_i}{\sum_{j=1}^n x_j^2}$ it is a weighted sum of the independent normal random variables: $y_i \sim N(x_i \beta, \sigma^2)$. It has a normal

distribution and all we need to do is compute the expectation and variance of $\hat{\beta}$:

$$\begin{aligned} E[\hat{\beta}] &= E[\textstyle\sum_{i=1}^n w_i y_i] \\ &= \textstyle\sum_{i=1}^n w_i E[y_i] = \sum_{i=1}^n w_i \times (x_i\beta) \\ &= \textstyle\sum_{i=1}^n (\frac{x_i}{\sum_{j=1}^n x_j^2}) x_i \beta \\ &= \beta \textstyle\sum_{i=1}^n (\frac{x_i^2}{\sum_{j=1}^n x_j^2}) = \beta \end{aligned}$$

$$\begin{aligned} Var[\hat{\beta}] &= Var[\textstyle\sum_{i=1}^n w_i y_i] \\ &= \textstyle\sum_{i=1}^n w_i^2 Var[y_i] = \sum_{i=1}^n w_i^2 \sigma^2 \\ &= \sigma^2 \times \textstyle\sum_{i=1}^n (\frac{x_i^2}{(\sum_{j=1}^n x_j^2)^2}) \\ &= \sigma^2 \times \frac{\sum_{i=1}^n x_i^2}{(\sum_{j=1}^n x_j^2)^2} \\ &= \sigma^2 \times \frac{1}{(\sum_{j=1}^n x_j^2)} \end{aligned}$$

So, $\hat{\beta} \sim N(\beta, \sigma_\beta^2)$ where $\sigma_\beta^2 = \sigma^2 \times \frac{1}{(\sum_{j=1}^n x_j^2)}$

(c). For a normal linear regression model the distribution of the sum of least-squares residuals has a distribution equal to σ^2 (the error variance) times a Chi-square distribution with degrees of freedom equal to $(n-p)$, where p is the number of independent variables and n is the number of cases. In this case, $p = 1$ so

$$SS_{ERR} = \textstyle\sum_{i=1}^n (y_i - x_i\hat{\beta})^2 \sim \sigma^2 \chi^2_{df=(n-1)}.$$

(d). Since a Chi-square random variable has expectation equal to its degrees of freedom. $E[SS_{ERR}] = \sigma^2(n-1)$ so

$$\hat{\sigma}^2 = \tfrac{SS_{ERR}}{n-1}$$

is an unbiased estiamte of σ^2.

2. **Simple Linear Regression**

Consider fitting the simple linear regression model:

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

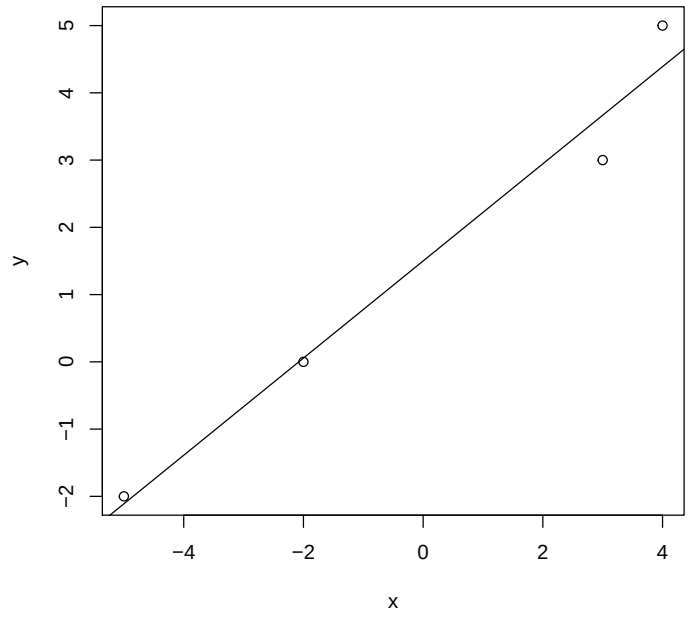
to the following bivariate data:

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)

Solution: The least-squares estimates are given by

$$\hat{\beta} = \begin{bmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \end{bmatrix} = (X^T X)^{-1} X^T Y, \text{ where}$$

$$X = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ 1 & x_3 \\ 1 & x_4 \end{bmatrix} = \begin{bmatrix} 1 & -5 \\ 1 & -2 \\ 1 & 3 \\ 1 & 4 \end{bmatrix} \text{ and } Y = \begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \end{bmatrix} = \begin{bmatrix} -2 \\ 0 \\ 3 \\ 5 \end{bmatrix}$$

Plugging in we get

$$\begin{aligned} \hat{\beta} = (X^T X)^{-1} X^T Y &= \begin{bmatrix} \sum_1^4 1 & \sum x_i \\ \sum x_i & \sum x_i^2 \end{bmatrix}^{-1} \begin{bmatrix} \sum y_i \\ \sum x_i y_i \end{bmatrix} \\ &= \begin{bmatrix} 4 & 0 \\ 0 & 54 \end{bmatrix}^{-1} \begin{bmatrix} 6 \\ 39 \end{bmatrix} \\ &= \begin{bmatrix} 6/4 \\ 39/54 \end{bmatrix} = \begin{bmatrix} 1.5 \\ 0.7222 \end{bmatrix} \end{aligned}$$

(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

Solution: We know formulas for the least-squares estimates of the slope and intercept are given by:

$$\begin{aligned} \hat{\beta}_2 &= \frac{\sum_1^n (x_i - \overline{x})(y_i - \overline{(y)})}{\sum_1^n (x_i - \overline{x})^2} = \frac{S_{xy}}{S_x^2} = r \frac{\sqrt{S_y^2}}{\sqrt{S_x^2}} = r \frac{s_y}{s_x} \\ &= (0.9855) \times \frac{3.1091}{4.2426} = 0.7222 \\ \hat{\beta}_1 &= \overline{y} - \hat{\beta}_2 \overline{x} \\ &= 1.5 - (0.7222) \times (0) = 1.5 \end{aligned}$$

(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).

Solution: We know that the variance of the slope from a simple linear regression model (where the errors have mean zero, constant variance σ^2 and are uncorrelated) is

$$Var(\hat{\beta}_2) = \sigma^2 / \sum_1^n (x_i - \overline{x})^2 = \sigma^2 / [(n-1) S_x^2]$$

The standard error of $\hat{\beta}_2$ is the square-root of this variance, plugging in the estimate $\hat{\sigma}$ for the standard deviation of the errors:

$$StErr(\hat{\beta}_2) = \hat{\sigma}/(\sqrt{(n-1)}s_x) = 0.65/(\sqrt{3}4.2526) = .088$$

(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)?

Solution: The least-squres prediction of $\hat{Y}$ when $X = \overline{x}$ must be $\overline{y}$, the mean of the dependent variable. The simple least-squares regression line always goes through the point of means: $(x, y) = (\overline{x}, \overline{y})$

The standard error of this prediction is just the estimate of the standard deviation of the sample mean $\overline{y}$ which is

$$\hat{\sigma}_{\overline{y}} = \sqrt{\frac{\hat{\sigma}^2}{n}} = \frac{\hat{\sigma}}{\sqrt{4}} = 0.65/2 = 0.325$$

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.

Solution:

(a). Let x be the midterm score and y be the final score. The least-squares regression of y on x is given in terms of the standardized values:

$$\frac{\hat{y}-\overline{y}}{s_y} = r\frac{x-\overline{x}}{s_x}$$

A score of 90 on the midterm is $(90 - 75)/10 = 1.5$ standard deviations above the mean. The predicted score on the final will be $r \times 1.5 = .9$ standard deviations above the mean final score, which is $75+(.9)\times 10 = 84$.

(b). For this case we need to regress the midterm score (x) on (y). The same argument in (a), reversing x and y leads to:

$$\frac{\hat{x}-\overline{x}}{s_x} = r\frac{y-\overline{y}}{s_y}$$

Since the final score was 75, which is zero-standard deviations above $\overline{y}$, the prediction of the midterm score is $\overline{x} = 75$.

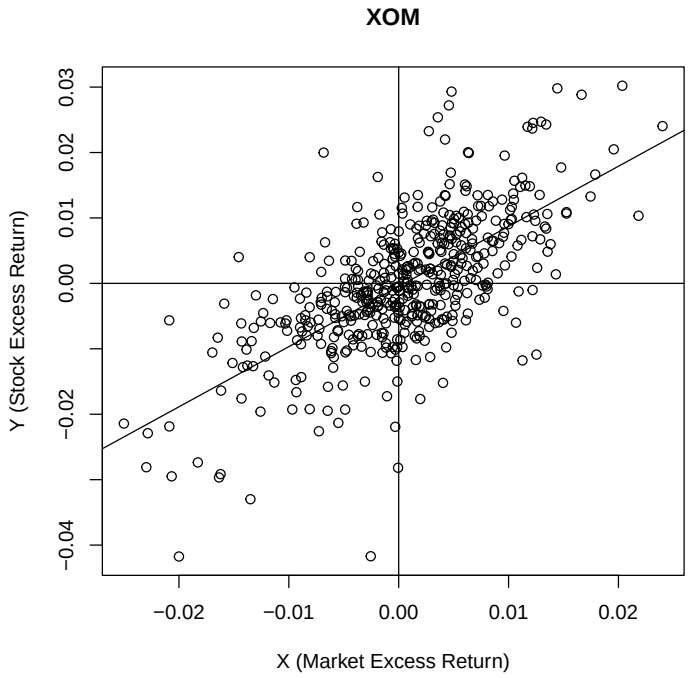
(c). By the regression effect we expect dependent variable scores to be closer to their mean in standard-deviation units than the independdent variable is to its mean, in standard-deviation units. Since the 75th percentile is on the midterm is above the mean, we expect these students to have average final score which is lower than the 75th percentile (i.e., closer to the mean). This means (*iii*) is the correct answer.

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.

Solution: The residual standard error is an estimate of the standard deviation of the error term in the regression model. It is given by

$$\hat{\sigma} = \sqrt{\frac{SS_{ERR}}{(n-p)}} = \sqrt{\frac{\sum_1^n (y_i - \hat{y}_i)^2}{n-p}}$$

It measures the standard deviation of the difference between the actual and fitted value of the dependent variable.

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

Solution: The degrees of freedom equals $(n - p)$ where $n = 500$ is the number of sample values and $p = 2$ is the number of regression parameters being estimated.

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

Solution: The correlation is $\sqrt{R - Squared} = \sqrt{.451} \approx .67$

(we know it is positive because of the positive slope coefficient 0.919)

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

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

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

Solution: Yes, apply a t-test of H_0: intercept equals 0. R computes this in the coefficients table and the statistic value is -1.43 with a (two-sided) p-value of 0.153. For a nominal significance level of .05 for the test (two-sided), the null hypothesis is not rejected because the p-value is higher than the significance level. The assumptions necessary to conduct the test are that the error terms in the regression are i.i.d. normal variables with mean zero and constant variance $\sigma^2 > 0$. If the normal distribution doesn't apply, then so long as the error distribution has mean zero and constant variance, the test is approximately approximately correct and equivalent to using a z-test for the parameter/estimate and the CLT.)

(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:

$H_0 : \beta = 1$ 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?

Solution: Yes, we apply a one-sided t-test using the statistic:

$T = \frac{\hat{\beta}-1}{stErr(\hat{\beta})} = \frac{0.919-1}{0.0454} = -.081/.0454 = -1.7841$

Under the null hypothesis T has a t-distribution with 498 degrees of freedom. This distribution is essentially the $N(0,1)$ distribution since the degrees of freedom is so high. The p-value of this statistic (one-sided) is less than 0.05 because $P(Z < -1.645) = 0.05$ for a $Z \sim N(0,1)$ so $P(T < -1.7841) \approx P(Z < -1.7841)$ which is smaller.

5. For the following batch of numbers:

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

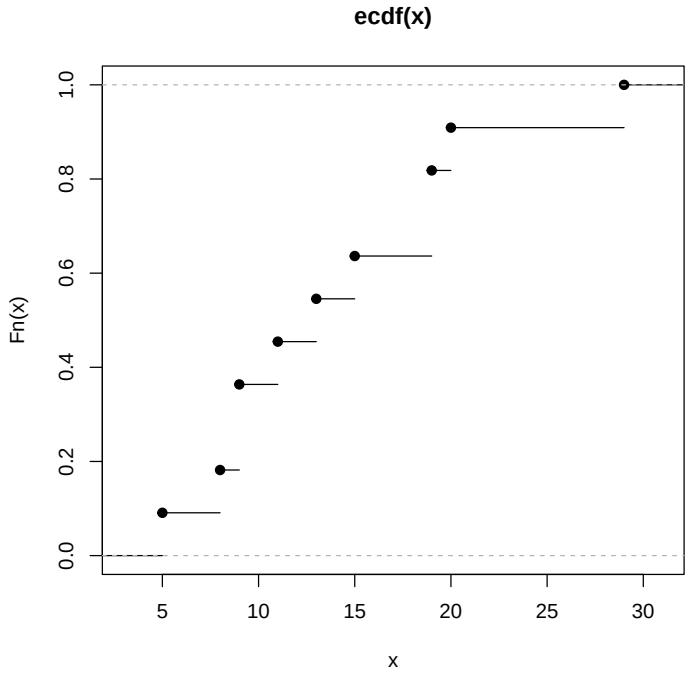
(a). Make a stem-and-leaf plot of the batch.

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

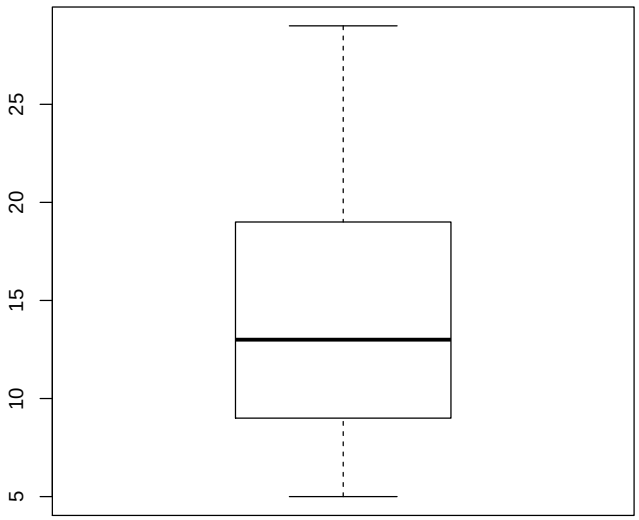
(c). Draw the Boxplot of the batch.

Solution:

```
> x=c(5,8,9,9,11,13,15,19,19,20,29)
> stem(x)

  The decimal point is 1 digit(s) to the right of the |

  0 | 5899
  1 | 13
  1 | 599
  2 | 0
  2 | 9

> plot(ecdf(x))
```




```
> median(x)

[1] 13

> quantile(x,probs=.25)

25%
  9

> quantile(x,probs=.75)

75%
 19

> boxplot(x)
```



Note that the center of the box is at the median (19), the bottom is at the 25-th percentile and top is at the 75-th percentile. The inter-quartile range is (19-9)=10, so any value more than $1.5 \times 9 = 13.5$ units above or below the box will be plotted as outliers. There are no such outliers.

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.

Solution:

(a). Consider the sample mean, the sample median, and the 10%-Trimmed mean.

$\hat{\theta}_{MEAN} = \frac{1}{n} \sum_1^n X_i$.

$\hat{\theta}_{MEDIAN} = median(X_1, X_2, \ldots, X_n)$

$\hat{\theta}_{TrimmedMean}$ = average of $\{X_i\}$ after excluding the highest 10% and the lowest 10% values.

(b). We expect the sample mean to be the best when the data are a random sample from the same normal distribution. In this case it is the MLE and will have lower variability than any other estimate.

We expect the same median to be the best when the data are a random sample from the bilateral exponential distribution. In this case it is the MLE and will hve lower variability, asymptotically than any other estimate. Also, the median is robust against gross outliers in the data resulting from the possibility of sampling distribution including a contamination component.

We expect the trimmed mean to be best when the chance of gross errors in the data are such that no more than 10% of the highest and 10% of the lowest could be such gross errors/outliers. For this estimate to be better than the median, it must be that the information in the mean of the remaining values (80% untrimmed) is more than the median. This would be the case if 80% of the data values came from a normal distribution/model.

arise from a normal distribution with

Percentiles of the Normal and t Distributions

	q_0.5	q_0.75	q_0.9	q_0.95	q_0.99	q_0.999
N(0,1)	0.00	0.67	1.28	1.64	2.33	3.09
t (df=1)	0.00	1.00	3.08	6.31	31.82	318.31
t (df=2)	0.00	0.82	1.89	2.92	6.96	22.33
t (df=3)	0.00	0.76	1.64	2.35	4.54	10.21
t (df=4)	0.00	0.74	1.53	2.13	3.75	7.17
t (df=5)	0.00	0.73	1.48	2.02	3.36	5.89
t (df=6)	0.00	0.72	1.44	1.94	3.14	5.21
t (df=7)	0.00	0.71	1.41	1.89	3.00	4.79
t (df=8)	0.00	0.71	1.40	1.86	2.90	4.50
t (df=9)	0.00	0.70	1.38	1.83	2.82	4.30
t (df=10)	0.00	0.70	1.37	1.81	2.76	4.14
t (df=25)	0.00	0.68	1.32	1.71	2.49	3.45
t (df=50)	0.00	0.68	1.30	1.68	2.40	3.26
t (df=100)	0.00	0.68	1.29	1.66	2.36	3.17
t (df=500)	0.00	0.67	1.28	1.65	2.33	3.11

18.443 Statistics for Applications

Spring 2015