THE UNIVERSITY OF SYDNEY

STAT2012 STATISTICAL TESTS

Semester 2 **Tutorial Week 11** 2015

Summary of week 10

- Regression: For the simple linear regression model:

$$\begin{aligned} Y_i &= \alpha + \beta\, x_i + \epsilon_i, \quad \text{where} \quad \epsilon_i \overset{\text{iid}}{\sim} \mathcal{N}(0,\, \sigma^2), \\ \alpha &\sim \mathcal{N}\left(\alpha, \sigma^2 \left[\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right]\right) \quad \text{and} \quad \hat{\beta} \sim \mathcal{N}\left(\beta, \frac{\sigma^2}{S_{xx}}\right) \end{aligned}$$

$$\text{Residual variance:} \quad s^2 = \frac{SSR}{n-2}, \quad SSR = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = S_{yy} - \hat{\beta} S_{xy} = SST_o - SST$$

$$\text{Coeff. of determination:} \quad r^2 = \frac{SST}{SST_o} = \frac{\hat{\beta} S_{xy}}{S_{yy}}$$

$$\text{CI for } \beta = \left(\hat{\beta} - t_{n-2,0.975}\frac{s}{\sqrt{S_{xx}}},\ \hat{\beta} + t_{n-2,0.975}\frac{s}{\sqrt{S_{xx}}}\right)$$

$$\text{CI for } \alpha = \left(\hat{\alpha} - t_{n-2,0.975}\ s\sqrt{\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}},\ \hat{\alpha} + t_{n-2,0.975}\ s\sqrt{\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}}\right)$$

- Test for regression model: To test $H_0:\ \beta = 0$ vs $H_1:\ \beta \neq 0$, the test statistic is

$$t_0 = \frac{\hat{\beta}}{s/\sqrt{S_{xx}}} \sim t_{n-2} \quad \text{or} \quad f_0 = t_0^2 = \frac{SST}{s^2} \sim F_{1,n-2}, \quad SST = \hat{\beta}^2 S_{xx} = \hat{\beta} S_{xy}.$$

- Regression: Estimation $E(\widehat{Y}_0)$ and prediction $\widehat{Y}_0$ when $x = x_0$:

$$\text{Estimated mean:} \quad E(\widehat{Y}_0) = \hat{\alpha} + \hat{\beta} x_0 \sim \mathcal{N}\left(\alpha + \beta\, x_0,\ \sigma^2\left[\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}\right]\right)$$

$$\text{Est. interval for} \quad E(\widehat{Y}_0) = \left(\hat{\alpha} + \hat{\beta} x_0 - t_{n-2,\alpha/2}\ s\sqrt{\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}},\right.$$
$$\left. \hat{\alpha} + \hat{\beta} x_0 + t_{n-2,\alpha/2}\ s\sqrt{\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}}\right)$$

$$\text{Predicted point:} \quad \widehat{Y}_0 = \hat{\alpha} + \hat{\beta} x_0 \sim \mathcal{N}\left(\alpha + \beta\, x_0,\ \sigma^2\left[1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}\right]\right)$$

$$\text{Pred. interval for} \quad Y_0 = \left(\hat{\alpha} + \hat{\beta} x_0 - t_{n-2,\alpha/2}\ s\sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}},\right.$$
$$\left. \hat{\alpha} + \hat{\beta} x_0 + t_{n-2,\alpha/2}\ s\sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}}\right)$$

Tutorial Questions

1. A statistician investigates the relationship between the amount of precipitation (in inches x) and the number of automobile accidents (y). He gathered data for 10 randomly selected days as shown below:

Day i	Number of accidents y	Amount of precipitation (in inches) x
1	5	0.05
2	6	0.12
3	2	0.05
4	4	0.08
5	8	0.10
6	14	0.35
7	7	0.15
8	13	0.30
9	7	0.10
10	10	0.20

The summary statistics are

$$\sum_{i=1}^{10} x_i = 1.5,\ \sum_{i=1}^{10} y_i = 76,\ \sum_{i=1}^{10} x_i^2 = 0.3208,\ \sum_{i=1}^{10} y_i^2 = 708,\ \sum_{i=1}^{10} x_i y_i = 14.74.$$

The following results are obtained from last week:

$$\begin{aligned} S_{yy} &= 130.4, \quad S_{xy} = 3.34, \quad S_{xx} = 0.0958, \quad \overline{y} = 7.6, \quad \overline{x} = 0.15 \\ \hat{\beta} &= 34.8643, \quad \hat{\alpha} = 2.370355 \end{aligned}$$

(a) Calculate the standard error estimate for the residuals.

(b) Calculate the 95% confidence interval for β.

(c) Test the model at a 5% level of significance.

(d) Determine the coefficient of determination and comment on the fit of the regression model.

(e) Determine the 95% estimation interval for the mean number of accidents if the precipitation of a day is 0.11 inches.

(f) Determine the 95% prediction interval for the number of accidents if the precipitation of a day is 0.11 inches.

(g) Will you use the regression model to predict the number of automobile accidents Y_0 on a stormy day when the amount of precipitation is $x_0 = 2$ inches? Explain briefly.

(h) Will you use the model to predict the number of automobile accidents Y_0 for a city which is very different from the city where the data were collected? Explain briefly.

Extra Practice Problems

1. By late 1971, all cigarette packs in the United States had to be labeled with the words, "Warning: The Surgeon General Has Determined That Cigarette Smoking Is Dangerous To Your Health." The case against smoking rested heavily on statistical, rather than laboratory, evidence. Extensive surveys of smokers and nonsmokers had revealed the former to have a much higher risks of dying from a variety of causes, most notably lung cancer and heart disease. Other types of studies, some designed with a much broader focus, painted much the same picture. Typical are the data below: 21 countries were the subjects. Recorded for each country was x, its annual cigarette consumption and y (per adult per year), its mortality rate due to coronary heart disease (per 100,000 ages 35-64) in the year of 1962.

Country	1	2	3	4	5	6	7	8	9	10	11
x	3900	3350	3220	3220	2790	2780	2770	2290	2160	1890	1810
y	256.9	211.6	238.1	211.8	194.1	124.5	187.3	110.5	233.1	150.3	124.7
Country	12	13	14	15	16	17	18	19	20	21	
x	1800	1770	1700	1680	1510	1500	1410	1270	1200	1090	
y	141.2	82.1	118.1	71.9	114.3	95.2	136.3	126.9	59.7	42.6	

The summary statistics are:

$$\sum_{i=1}^{21} x_i = 45,110,\ \sum_{i=1}^{21} y_i = 3031.2,\ \sum_{i=1}^{21} x_i^2 = 109,957,100,\ \sum_{i=1}^{21} y_i^2 = 513,248.16,\ \sum_{i=1}^{21} x_i y_i = 7,340,085.$$

(a) Test the regression model at the 0.05 level of significance.

(b) If it is known that the $X = 3000$ for an Asian country in 1962, predict the mortality rate Y for this country. Provide a 95% confidence interval for the prediction.

(c) If it is known that the $X = 3000$ for an Asian country in 1999, can we predict the mortality rate Y for this country in 1999 using the same estimated regression line? Explain.

(d) Test the significance of the sample correlation coefficient between X and Y.

(e) Do you think the linear regression model is an appropriate model for the relationship between X and Y?